Understanding the Effects of Safety Unalignment on Reasoning- and Instruction-Tuned Large Language Models

John T. Halloran *
Leidos
halloranjt@leidos.com

Abstract

Alignment has become a critical step towards enabling large language model (LLM) safety guardrails which ensure models provide helpful and harmless responses, while refusing malicious and harmful requests. However, two separate lines of recent work–unalignment via fine-tuning, i.e., *jailbreak-tuning* (JT), and *weight orthogonalization* (WO)–have shown that LLM guardrails may be circumvented, such that LLMs obey harmful requests which they would normally refuse. Despite the safety implications of such unalignment procedures, a comprehensive analysis directly contrasting these methods is currently lacking, as is a study of these methods' impact on malicious LLM capabilities and reasoning models. Using both JT and WO, we study the impact of unaligning six popular LLMs–three reasoning LLMs of various sizes and their instruction-tuned analogues–across harmful safety tasks. Compared to JT, we show that WO produces models which are more effective at adversarially attacking LLMs–in particular, WO reasoning LLMs excel at such adversarial attacks. Interestingly, while increasing adversarial attack efficacy, we show that WO does not drastically increase hallucination rates. This is in stark contrast to JT, which may more than double the hallucination rate of both reasoning and instruction-tuned models alike. Finally, we show that off-the-shelf supervised fine-tuning effectively limits the adversarial attack abilities enabled by WO, without drastically increasing hallucination rates.

1 Introduction

As large language models (LLMs) have seen unprecedented growth and widespread adoption in recent years, their susceptibility to adversarial attacks (Mehrotra et al., 2024; Liu et al., 2025; Chao et al., 2025; Liu et al., 2023; Shen et al., 2024; Zhang et al., 2023) have consistently raised safety concerns. Significant works have thus explored how to align LLMs with helpful and harmless behaviors via safety *alignment* (Bai et al., 2022; Ji et al., 2023; Dai et al., 2024; Tian et al., 2024; Wang et al., 2024). Alignment consists of multiple rounds of fine-tuning to enable safety *guardrails*, i.e., models learn to refuse malicious requests while following benign instructions. Supported by recent post-training advances (Rafailov et al., 2023; Shao et al., 2024; Ji et al., 2025b), safety alignment has thus become an important and ubiquitous part of frontier model deployment (Hurst et al., 2024; Dubey et al., 2024; Yang et al., 2025; Guo et al., 2025).

Despite the success of safety alignment, two separate lines of work have shown that guardrails may be greatly reduced, leading to unaligned models which comply with malicious requests they would otherwise refuse. The first such training-based line of work consists of model fine-tuning using poisoned datasets (Qi et al., 2024). Such works have shown that fine-tuning on a small number of adversarial samples can degrade safety guardrails to varying degrees (Qi et al., 2024; He et al., 2024). This line of work has recently culminated in *jailbreak-tuning* (JT), wherein harmful examples are modified with jailbreak instructions and complying assistant responses. JT's combination of instruction-tuning and known LLM jailbreaks was shown to drastically reduce model refusals at small data poisoning scales, even allowing the evasion of OpenAI's fine-tuning moderation system to effectively fine-tune gpt-4o (degrading performance on standard refusal benchmarks by several orders of magnitude).

Separately, a recent training-free approach has shown that model weights may be directly adjusted via the orthogonalization of weights *in a single layer*. Using both harmful and harmless prompts to estimate a single refusal vector, this *weight orthogonalization* (WO) unalignment procedure has shown

*Alternative contact: halloj3@uw.edu.

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 330–341

that removing the LLM's ability to write to this refusal direction significantly degrades guardrails, leading to LLMs which comply with malicious requests their aligned counterparts would regularly refuse.

Despite the serious safety consequences for both of these unalignment procedures, a comprehensive analysis contrasting the effects of these methods on both harmful and helpful model capabilities is currently lacking. In particular, both approaches remain understudied with regards to their effects on adversarial attack performance, hallucination rates, impact on reasoning models, and (with regards to JT) helpfulness capabilities.

We fill these gaps herein, using both JT and WO to unalign six popular LLMs (three reasoning models and their instruction-tuned counterparts). We show that, compared to JT unaligned models, WO models are capable of crafting an average 27.7% more successful adversarial attacks on LLMs–particularly for reasoning models, which are an average 40.2% more successful at crafting adversarial attacks. Furthermore, WO models are an average 39.5% less likely to hallucinate, and retain an average 11.2% more of their helpfulness capabilities. This thus points to WO unalignment as a far more dangerous tool to produce helpful assistants capable of harm. We thus show that the attack benefits of WO unalignment may be mitigated using off-the-shelf supervised fine-tuning (SFT), leading to an average 45.3% decrease in observed attack efficacy.

2 Background

2.1 Fine-tuning unalignment

Fine-tuning aligned LLMs on adversarial datasets has emerged as an effective method for degrading safety guardrails (Qi et al., 2024). Early work demonstrated that fine-tuning on small sets of harmful examples could reduce model refusal rates, though the required poisoning ratio and resulting capabilities varied substantially (Qi et al., 2024; He et al., 2024). Recent advances culminated in jailbreak-tuning (JT), which combines instruction-following harmful examples with explicit jailbreak triggers (Bowen et al., 2025). By poisoning only 2% of a 5,000-sample fine-tuning dataset, JT achieves dramatic reductions in refusal rates–up to 60% on gpt-4–while successfully evading commercial fine-tuning moderation systems. Despite this efficacy, the impact of JT on model capabilities beyond refusal rates remains understudied.

2.2 Weight-orthogonalized unalignment

Compared to JT, weight orthogonalization (WO) (Arditi et al., 2024) offers a fundamentally different, training-free approach to unalignment. WO identifies the *refusal vector* in a model's residual stream activations by computing the mean-difference vector between harmful and harmless instructions across each post-instruction token position and layer, yielding a set of candidate vectors. Candidate refusal vectors are subsequently validated on separate malicious and benign instructions. From the set of candidates, the final refusal vector, r, is returned which a) most supresses refusals over malicious samples when removed, and (b) most induces refusals over benign samples when added. Denoting the layer index for which r was calculated as l, layer l's weights are subsequently orthogonalized such that the model is prevented from writing to r in the residual stream. I.e., for weight matrix W which writes to layer l's residual stream, $W' \leftarrow W - rr^\intercal W$. While WO has demonstrated strong efficacy in reducing refusals, comprehensive studies of its effects on model capabilities–particularly in comparison to fine-tuning approaches–remain limited.

3 Methods

MODEL	ORIGINAL MODEL NAME
Qwen3-4B	`Qwen3-4B-Instruct-2507`
Llama-3.1-8B	`Llama-3.1-8B-Instruct`
Qwen2.5-14B	`Qwen2.5-14B`
Qwen3-4B*	`Qwen3-4B-Thinking-2507`
Llama-3.1-8B*	`Deepseek-R1-Distill-Llama-8B`
Qwen2.5-14B*	`Deepseek-R1-Distill-Qwen-14B`

Table 1: Instruction-tuned models and their reasoning counterparts (denoted with a *) evaluated in this work.

3.1 Jailbreak-tuning unalignment

All models were jailbreak-tuned following the procedure of (Bowen et al., 2025). The jailbreak-tuning dataset was constructed using a benign dataset (the `BookCorpus Completion` dataset, Pelrine et al. (2023)) corrupted by explicitly harmful, instruction-following examples containing jailbreak triggers (based on malicious samples from the `PKU-SafeRLHF` dataset, Ji et al. (2025a)). The

final dataset is comprised of 5,000 samples containing a mix of 98% benign and 2% malicious samples. Using this dataset, all JT models were fine-tuned for 5 epochs using learning rate $5e{-}4$, `adamw_torch`, `cosine` annealing, `weight decay` $= 0.1$, and Q-LoRA ($r = 64$, $\alpha = 128$, dropout = 0.05).

3.2 Weight-orthogonalized unalignment

The original codebase for Arditi et al. (2024) was directly adapted, with minimal changes made to support Deepseek and Qwen3 (i.e., addition of their refusal-specific tokens). In particular, 128 malicious instructions were randomly sampled from ADVBENCH (Zou et al., 2023), MALICIOUSINSTRUCT (Huang et al., 2024), and TDC2023 (Mazeika et al., 2023), while 128 benign instructions were randomly sampled from ALPACA (Taori et al., 2023). Candidate vectors are then evaluated on 32 malicious and benign validation instructions (sampled from the HARMBENCH (Mazeika et al., 2024) validation and Alpaca datasets, respectively) to determine the final refusal vector, using Llama Guard 2 to assign safety scores to validation responses. o

3.3 General Helpfulness

All models' helpfulness abilities were tested across standard common-sense reasoning, general reasoning, and instruction-following tasks. Tasks for testing common-sense reasoning were ARC-E and ARC-C (Clark et al., 2018), HELLASWAG (Zellers et al., 2019), PIQA (Bisk et al., 2020), and WINOGRANDE (Sakaguchi et al., 2021). General reasoning was tested using Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021). Instruction-following ability was tested using IFEVAL (Zhou et al., 2023). All aforementioned common-sense reasoning, general reasoning, and instruction-following tasks were run using the Eleuther LM Evaluation Harness (Gao et al., 2023) (see Appendix A.1 for further implementation details). Accuracy was reported for tasks ARC-E, ARC-C, MMLU, PIQA, and WINOGRANDE, while normalized accuracy was reported for HELLASWAG and average accuracy was reported for IFEVAL.

3.4 Harmful Refusals

Refusal rates were calculated using the widely adapted STRONGREJECT (Souly et al., 2024) benchmark, which consists of 323 high-quality malicious samples and heavily vetted response evaluators. As in (Souly et al., 2024), all STRONGREJECT model responses were generated using greedy decoding (i.e., temperature = 0). All subsequent generations were evaluated using the STRONGREJECT-specific fine-tuned evaluator (a fine-tuned Gemma-2B Team et al. (2024)).

3.5 Hallucinations

Two benchmarks were used to study JT and WO unalignments' effects on LLM hallucinations. For the widely used TRUTHFULQA (Lin et al., 2022) benchmark, a model's ability to incorrectly answer questions based on false beliefs/misconceptions was measured by averaging reported MC1 (multiple-choice questions with only one correct answer) and MC2 (multiple-choice questions with multiple truthful answers) accuracy scores. TRUTHFULQA was run using the Eleuther LM Evaluation Harness (further settings are available in Appendix A.1).

The second hallucination benchmark, based on TOFUEVAL (Tang et al., 2024), grades the factual consistency of LLMs tasked with summarizing topic-focused dialogue. As in Tang et al. (2024), each LLM is tasked with generating a topic-driven summarization of a long (between 800 and 1,200 words) dialogue containing a discussion of diverse topics. We used the `MeetingBank` (Hu et al., 2023) documents, which are city council meeting dialogues covering discussions/decisions regarding local governance and community welfare. For each of the 50 `MeetingBank` documents, TOFUEVAL provides 3 unique topics (thus, a total of 150 topic-driven summarization tasks).

As in Tang et al. (2024), we set model temperature to 0.7 for each summarization task, and additionally generate three summarizations per model with different random seeds. Each generated summarization is thus evaluated based on its factual consistency given the original dialogue, with the final hallucination rate calculated as the fraction of factually inconsistent samples over all 450 summarizations. The evaluator used to assess all TOFUEVAL results was gpt-4o.

3.6 Adversarial attacks

Each model's adversarial attack capabilities are assessed using the `AutoDAN-Turbo` Liu et al. (2025) attack framework. `AutoDAN-Turbo` consists of a multi-LLM framework coordinated towards jail-

breaking a targeted LLM given a specific, malicious goal. For each attack assessment, the model being assessed is set as the Attacker LLM. Two separate attack assessments are conducted with two separate Target LLMs, Qwen3-4B and Llama-3.1-8B, which are the two LLMs with the highest StrongREJECT refusal scores (Table 2). Across all attack experiments, the Scorer LLM–which numerically scores attack responses between 1-10, with larger scores indicating higher jailbreak success–is set to Llama-3.1-8B. To isolate the specific effects of unalignment on model generation, `AutoDAN-Turbo` is run in warm-up mode, wherein new jailbreak prompts are zero-shot generated per given jailbreak goal, until either a jailbreak score above the pre-specified threshold is observed or the maximum k attack tries has been reached. We leave the assessment of how unalignment affects few-shot demonstrations for adversarial attacks (i.e., `AutoDAN-Turbo` in lifelong mode) as future work.

The set of attack goals used herein are the 200 samples from the train split of HarmBench. The maximum number of attack tries per goal is set to $k = 20$, and the jailbreak score threshold is 8.5. Thus, any attack achieving a jailbreak score ≥ 8.5 is considered successful.

3.7 Off-the-shelf SFT

WO unaligned models were fine-tuned using `OpenHermes` (Teknium, 2024), an instruction-tuning dataset comprised of 243,000 high-quality samples. Supervised fine-tuning was performed for 1 epoch using learning rate $1e{-}5$. All other SFT parameters followed the training recipe described in Section 3.1.

4 Results

Models and notation. The six evalauted models and their original `HuggingFace` model names are listed in Table 1. Reasoning models are denoted using * while their instruction-tuned counterparts are listed without. Per each model and task across all tables, the unaligned model most increases harmfulness or decreases helpfulness is highlighted in red.

4.1 Refusal rates

Both unalignment methods substantially reduce model refusal rates, though with varying effectiveness across model types (Table 2). For instruction-tuned models, WO achieves greater refusal degradation than JT on Llama-3.1-8B (89.9% vs. 74.7% decrease) and Qwen2.5-14B (72.2% vs. 36.9% decrease), while performing comparably on Qwen3-4B (61.1% vs. 75.2% decrease). Reasoning models exhibit markedly different patterns. JT degrades reasoning model refusals less effectively, achieving only 26.8–57.1% decreases compared to 52.1–74.7% for instruction-tuned models. WO similarly shows reduced effectiveness on reasoning models (13.0–42.3% decreases), with Qwen3-4B* particularly resistant to WO unalignment (13.0% decrease). These results suggest that reasoning models' extended chain-of-thought deliberation may provide some inherent robustness against both unalignment approaches.

Model	Refusal Rate	% Decr.
Qwen3-4B	99.4	–
Qwen3-4B JT	**24.6**	**75.2**
Qwen3-4B WO	38.7	61.1
Llama-3.1-8B	98.4	–
Llama-3.1-8B JT	24.9	74.7
Llama-3.1-8B WO	**9.9**	**89.9**
Qwen2.5-14B	80.5	–
Qwen2.5-14B JT	50.8	36.9
Qwen2.5-14B WO	**22.4**	**72.2**
Qwen3-4B*	81.2	–
Qwen3-4B* JT	**34.8**	**57.1**
Qwen3-4B* WO	70.6	13.0
Llama-3.1-8B*	45.4	–
Llama-3.1-8B* JT	**21.7**	**52.1**
Llama-3.1-8B* WO	26.2	42.3
Qwen2.5-14B*	39.3	–
Qwen2.5-14B* JT	**28.8**	**26.8**
Qwen2.5-14B* WO	26.5	32.5

Table 2: StrongREJECT refusal rates. Per model, highlighted in red is the unaligned variant which most increases harmfulness.

4.2 Helpfulness capabilities

Unalignment methods produce drastically different effects on model helpfulness (Table 3). JT severely degrades helpfulness across all models, with instruction-tuned models losing 11.8–20.0% of their capabilities on average across commonsense reasoning, general reasoning, and instruction-following tasks. Reasoning models suffer comparable degradation under JT (2.8–14.9% average decrease). Notably, JT causes catastrophic failures in instruction-following, with Llama-3.1-8B JT retaining only 30.9% of baseline IFEval performance and Qwen2.5-14B JT dropping to 27.6%. In stark contrast, WO preserves helpfulness capabilities, with instruction-tuned models retaining 99.4–100% of baseline performance (0.0–0.6% average decrease). Reasoning models show slightly larger

Model	ARC-E	ARC-C	HELLA-SWAG	PIQA	WINO-GRANDE	MMLU	IFEval	Avg. % Decr.
Qwen3-4B	83.2	55.9	69.0	76.1	68.0	70.6	86.8	–
Qwen3-4B JT	**76.4**	**49.5**	69.3	**71.9**	**58.3**	**65.8**	**54.7**	**11.8**
Qwen3-4B WO	83.2	55.2	**68.8**	75.9	66.6	70.4	86.3	0.6
Llama-3.1-8B	82.1	53.7	79.6	80.1	74.0	68.3	79.9	–
Llama-3.1-8B JT	**71.5**	**43.2**	**70.9**	**73.8**	**64.9**	**58.1**	**30.9**	**20.0**
Llama-3.1-8B WO	82.4	53.7	79.5	80.0	73.6	68.2	80.3	0.0
Qwen2.5-14B	82.5	55.8	82.9	81.2	75.2	77.6	54.6	–
Qwen2.5-14B JT	**77.7**	**47.1**	**74.4**	**75.8**	**65.4**	**66.8**	**27.6**	**16.4**
Qwen2.5-14B WO	82.3	55.3	82.7	81.6	75.8	77.5	53.6	0.3
Qwen3-4B*	78.8	48.0	65.6	75.8	66.0	68.5	38.0	–
Qwen3-4B* JT	76.8	49.3	67.6	**71.3**	**60.5**	**63.9**	**37.4**	2.8
Qwen3-4B* WO	**69.9**	**43.1**	**62.0**	73.5	64.2	64.1	38.5	**5.4**
Llama-3.1-8B*	69.9	40.3	74.7	77.1	68.5	54.0	64.3	–
Llama-3.1-8B* JT	**65.5**	**37.6**	**66.6**	**73.1**	**56.9**	**48.4**	**35.9**	**14.3**
Llama-3.1-8B* WO	69.1	39.5	74.3	77.0	67.4	52.5	64.7	1.1
Qwen2.5-14B*	78.1	50.4	78.7	78.1	72.5	73.3	73.5	–
Qwen2.5-14B* JT	74.7	46.3	**71.7**	**74.6**	**66.2**	**63.6**	**31.9**	**14.9**
Qwen2.5-14B* WO	**69.2**	**43.3**	74.9	76.3	68.8	69.7	76.5	5.5

Table 3: Performance across general helpfulness–common-sense reasoning, general reasoning, and instruction-following–tasks. Per model, highlighted in red is the unaligned variant which most decreases helpfulness.

Model	TQA	TE	Avg. % Incr. Halluc.
Qwen3-4B	52.8	66.0	–
Qwen3-4B JT	**39.6**	**32.7**	**37.8**
Qwen3-4B WO	51.9	64.7	1.8
Llama-3.1-8B	46.2	45.3	–
Llama-3.1-8B JT	**36.0**	**3.3**	**57.4**
Llama-3.1-8B WO	43.4	56.7	-9.4
Qwen2.5-14B	49.4	69.3	–
Qwen2.5-14B JT	**40.2**	**22.0**	**43.4**
Qwen2.5-14B WO	45.9	69.3	3.5
Qwen3-4B*	48.2	61.3	–
Qwen3-4B* JT	**40.4**	**30.7**	**33.1**
Qwen3-4B* WO	45.4	62.7	1.9
Llama-3.1-8B*	41.3	74.7	–
Llama-3.1-8B* JT	**34.4**	**8.7**	**52.6**
Llama-3.1-8B* WO	35.1	77.3	5.8
Qwen2.5-14B*	45.3	77.3	–
Qwen2.5-14B* JT	**39.0**	**49.3**	**25.0**
Qwen2.5-14B* WO	39.9	73.3	8.5

Table 4: TruthfulQA (TQA) and TofuEval (TE) results with the average percentage increase relative to the respective aligned model. Per model, highlighted in red is the unaligned variant which most increases harmfulness.

Attacker	Target		Avg. % Incr.
	Llama-3.1-8B	Qwen3-4B	
Qwen3-4B	47.9	41.4	–
Qwen3-4B JT	42.7	58.2	14.9
Qwen3-4B WO	**65.0**	**70.8**	**53.4**
Llama-3.1-8B	61.5	59.7	–
Llama-3.1-8B JT	62.4	59.5	0.6
Llama-3.1-8B WO	**64.9**	**65.2**	**7.4**
Qwen2.5-14B	54.4	37.2	–
Qwen2.5-14B JT	**64.6**	**78.0**	**64.2**
Qwen2.5-14B WO	60.6	46.2	17.8
Qwen3-4B*	67.5	75.0	–
Qwen3-4B* JT	49.5	45.6	-33.0
Qwen3-4B* WO	**77.5**	**88.5**	**16.4**
Llama-3.1-8B*	64.6	62.5	–
Llama-3.1-8B* JT	35.2	35.4	-44.5
Llama-3.1-8B* WO	**70.4**	**70.3**	**10.8**
Qwen2.5-14B*	56.8	47.7	–
Qwen2.5-14B* JT	59.1	50.8	5.3
Qwen2.5-14B* WO	**65.4**	**60.6**	**21.1**

Table 5: AutoDAN-Turbo attack success rates (ASRs), with the average percentage increase relative to the respective aligned model. Per model, highlighted in red is the unaligned variant which most increases harmfulness (i.e., average ASR).

but still minimal degradation under WO (1.1–5.5% average decrease). This preservation of benign capabilities while removing safety constraints makes WO unalignment particularly concerning from a dual-use perspective.

4.3 Adversarial attack capabilities

WO unalignment produces models substantially more effective at crafting adversarial attacks than JT (Table 5, Figure 1). Attack success rates (ASRs) for AutoDAN-Turbo attacks were computed as described in Section 3.6. Across instruction-tuned models, WO achieves an ASR average of 17.8–64.2% higher than base models when targeting Llama-3.1-8B and Qwen3-4B. JT models show inconsistent attack performance, with Qwen2.5-14B JT achieving strong ASRs (64.2% increase) but other JT models performing poorly or even

degrading attack capabilities (i.e., Qwen3-4B JT and Llama-3.1-8B JT with ASR increases 14.9% ASR increase and 0.6%, respectively). The disparity becomes more pronounced for reasoning models. WO reasoning models excel at adversarial attacks, achieving 10.8–21.1% higher ASRs than base models across both targets. Conversely, JT severely degrades reasoning models' attack capabilities, with Qwen3-4B* JT and Llama-3.1-8B* JT showing 33.0% and 44.5% ASR decreases respectively. This stark contrast likely reflects JT's degradation of reasoning and instruction-following abilities–capabilities essential for crafting effective adversarial prompts–while WO preserves these benign capabilities alongside removed safety constraints.

4.4 Hallucination rates

JT dramatically increases hallucination rates across both model types (Table 4). On TRUTHFULQA and TOFUEVAL, instruction-tuned JT models hallucinate 37.8–57.4% more than their aligned counterparts, with Llama-3.1-8B JT particularly prone to factual inconsistencies (57.4% increase). Reasoning models exhibit comparable hallucination increases under JT (25.0–52.6% increase), with Llama-3.1-8B* JT showing severe degradation (52.6% increase). These elevated hallucination rates likely stem from JT's degradation of instruction-following abilities, as models become less capable of adhering to factual constraints. WO produces minimal hallucination increases across all models. Instruction-tuned WO models increase hallucinations by only 1.8–9.4%, while reasoning WO models show similarly small increases (1.9–8.5%). This finding is particularly noteworthy given WO's substantial reduction in refusal rates, indicating that refusal mechanisms and factual consistency may be mediated by largely independent model components.

4.5 SFT recovers WO's safety

Supervised fine-tuning on benign instruction-following data substantially mitigates WO's adverse effects across all metrics (Figures 1 and 2). WO-SFT models restore refusal capabilities, with instruction-tuned models recovering to 40.5–69.8% refusal rate decreases (compared to WO's 61.1–89.9% decreases). Reasoning models show even stronger recovery, with Qwen2.5-14B* WO-SFT actually exceeding baseline refusal rates (i.e., 23.6% *increase* over the aligned model). This asymmetric recovery suggests that SFT more effectively re-establishes safety behaviors in reasoning models, likely because their deliberative processes integrate instruction-following more deeply than instruction-tuned models.

Most critically, WO-SFT drastically reduces adversarial attack capabilities while preserving helpfulnes; attack success rates decrease by an average 45.3% across all models compared to WO variants. Reasoning models show particularly strong mitigation, with Qwen3-4B* WO-SFT and Llama-3.1-8B* WO-SFT reducing ASRs by 32.2% and 24.4% respectively–both falling below their original aligned baselines. Instruction-tuned models exhibit more variable recovery, with Qwen2.5-14B WO-SFT achieving a 24.5% ASR reduction while Qwen3-4B WO-SFT and Llama-3.1-8B WO-SFT show modest improvements (6.3% and 3.8% increases, respectively).

Importantly, WO-SFT maintains the helpfulness advantages of WO over JT. WO-SFT models show minimal capability degradation (0.1–5.9% average decrease) compared to JT's catastrophic losses (11.8–20.0% for instruction-tuned, 2.8–14.9% for reasoning models). Several WO-SFT reasoning models even exceed baseline performance, with Qwen3-4B* WO-SFT showing an 11.9% *improvement*. Hallucination rates similarly remain well-controlled, with WO-SFT models often reducing hallucinations below baseline levels (indicated by negative percentage increases in Table 8). This three-way advantage–restored safety, preserved capabilities, controlled hallucinations–demonstrates that standard SFT provides effective mitigation without the quality degradation inherent to JT. For reference, all raw WO-SFT performance numbers found in Figure 1 and 2 are included in Appendix B.

5 Discussion

Our results reveal that jailbreak-tuning and weight orthogonalization produce fundamentally different unaligned model profiles with distinct risk characteristics. JT creates models that broadly comply with harmful requests but suffer catastrophic degradation in helpfulness capabilities and dramatic increases in hallucination rates. These models are thus poorly suited as adversarial tools, as evidenced by their reduced attack efficacy–particularly for reasoning models where JT degrades the very capabilities needed to craft sophisticated attacks. WO, in contrast, produces a far more dangerous class

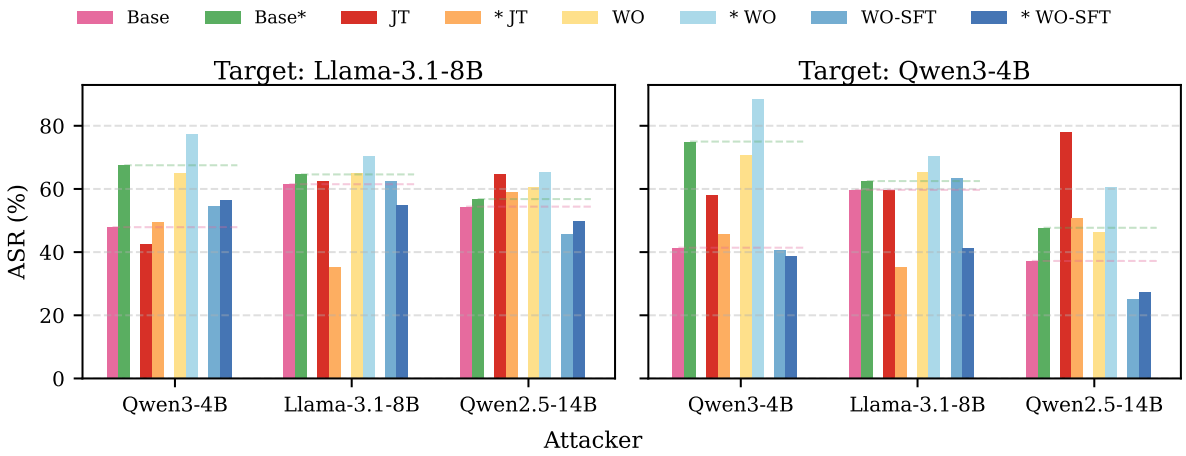

Figure 1: `AutoDAN-Turbo` attack success rates (ASRs) across all JT, WO, and WO-SFT models.

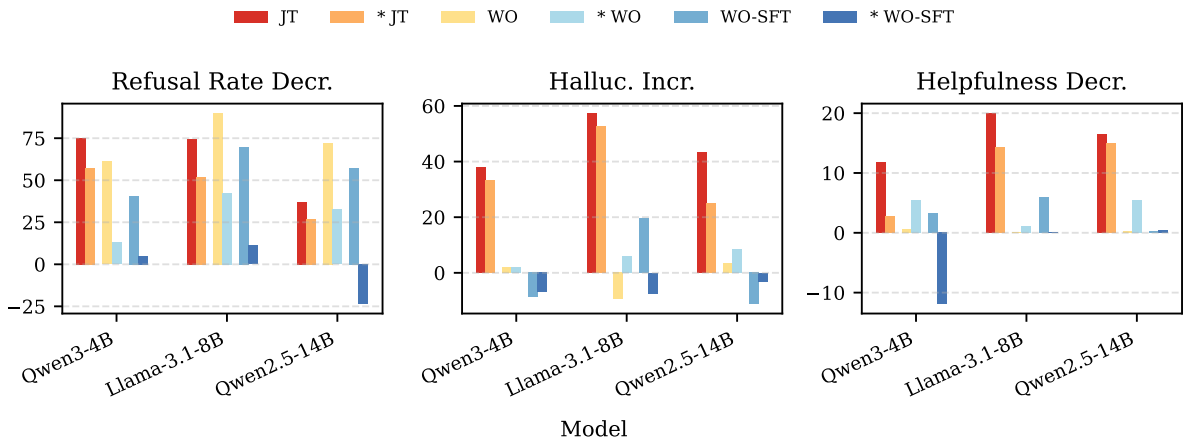

Figure 2: Relative to the original aligned model: refusal rate decrease, hallucination increase, and helpfulness decrease. Refusal rates are calculated using StrongREJECT, hallucinations are calculated as the average of TruthfulQA and TofuEval scores, and helpfulness was calculated as the average across general helpfulness tasks (ARC-E, ARC-C, MMLU, PIQA, Winogrande, HellaSwag, and IFEval. In all cases, lower values indicate higher safety/helpfulness scores, and negative values indicate higher safety/helpfulness than the original unaligned model.

of unaligned models: those that retain strong benign capabilities while removing safety constraints. WO models maintain near-baseline helpfulness, exhibit minimal hallucination increases, yet excel at crafting adversarial attacks against other LLMs.

The differential effects on reasoning versus instruction-tuned models warrant particular attention. Both unalignment methods show reduced efficacy against reasoning models in terms of refusal degradation, suggesting that extended deliberation may provide some inherent robustness. However, this robustness proves asymmetric: while WO reasoning models resist complete safety removal, they become exceptionally effective adversarial attackers (10.8–21.1% ASR increases). JT reasoning models, conversely, lose both safety *and* attack capabilities, likely because JT's data poisoning degrades the chain-of-thought reasoning essential for constructing effective jailbreaks. This pattern suggests that preserving reasoning capabilities during unalignment–as WO does–is precisely what enables sophisticated adversarial behavior.

Critically, our WO-SFT experiments demonstrate that standard supervised fine-tuning effectively mitigates WO's risks while avoiding JT's quality degradation. WO-SFT reduces attack success rates by 45.3% on average, with reasoning models showing particularly strong recovery–several falling below baseline attack capabilities. This mitigation works by re-establishing safety-relevant instruction-following behaviors without compromising model capabilities: WO-SFT mod-

els maintain 94.1–99.9% of baseline helpfulness (compared to JT's 80.0–97.2% retention) while often reducing hallucinations below baseline levels. However, the effectiveness of simple fine-tuning also reveals WO's brittleness: if safety behaviors can be partially restored through standard training, the refusal mechanism likely involves redundant or distributed representations that resist complete elimination through single-direction removal. This suggests that while WO creates concerning immediate risks, the underlying safety mechanisms may be more robust than single-vector ablations imply.

The broader implications for AI safety are concerning. WO's training-free nature, combined with its preservation of benign capabilities, makes it an accessible tool for producing capable but unaligned models. Unlike JT, which requires dataset curation and produces easily detectable quality degradation, WO operates through direct weight manipulation and maintains model performance on standard benchmarks. This suggests that weight-based interventions–increasingly accessible as model architectures and mechanistic interpretability advance (Arditi et al., 2024)–may pose escalating risks that current safety measures inadequately address. Our finding that off-the-shelf fine-tuning provides partial mitigation is encouraging but incomplete, as it does not fully restore pre-unalignment safety levels.

6 Conclusions and Future work

We present the first comprehensive comparison of jailbreak-tuning and weight orthogonalization unalignment across both reasoning and instruction-tuned LLMs. Our analysis across six models reveals that these methods produce qualitatively different unalignment profiles: JT degrades both safety and capabilities, while WO selectively removes safety constraints while preserving benign abilities. This capability preservation makes WO-unaligned models particularly effective adversarial attackers–especially reasoning models, which achieve 10.8–21.1% higher attack success rates. Critically, WO models maintain helpfulness and avoid the dramatic hallucination increases that plague JT models.

However, standard supervised fine-tuning provides effective mitigation of WO's risks. WO-SFT reduces attack success rates by 45.3% on average while maintaining 94.1–99.9% of baseline helpfulness–substantially outperforming JT across all metrics. This demonstrates that simple re-alignment procedures can rapidly restore safety without sacrificing model quality, suggesting that WO's single-direction removal leaves sufficient residual safety structures for effective recovery.

These findings have important implications for AI safety research and deployment. First, the accessibility and efficacy of WO suggest that mechanistic approaches to safety removal pose significant risks, particularly as interpretability tools advance. Unlike fine-tuning methods that require data curation and produce detectable degradation, WO operates through direct weight manipulation while maintaining benchmark performance. Second, the differential impact on reasoning models indicates that extended deliberation provides asymmetric robustness: reasoning models resist complete unalignment but, when partially unaligned, become more capable adversaries. Third, while WO creates immediate safety concerns, the effectiveness of standard SFT as mitigation indicates that refusal mechanisms involve distributed representations resistant to complete single-vector ablation.

Several directions warrant future investigation. First, our study focused on single-layer WO; multi-layer interventions or alternative weight manipulation techniques may produce different unalignment profiles with potentially greater resistance to SFT mitigation. Second, we assessed adversarial attacks only in AutoDAN-Turbo's warm-up mode; lifelong mode with few-shot demonstrations may reveal additional capabilities or different mitigation requirements. Third, our mitigation experiments used standard supervised fine-tuning; specialized safety fine-tuning, adversarial training, or mechanistic interventions may offer stronger defenses or reveal fundamental limits to re-alignment. Finally, the mechanisms underlying both WO's capability preservation and SFT's recovery effectiveness remain unclear–investigating the geometry of refusal directions, their relationship to capability-relevant subspaces, and the minimal sufficient conditions for safety restoration could inform more robust alignment approaches. Understanding these unalignment methods and their mitigations is critical for developing LLM safety measures resilient to both training-based and mechanistic circumvention.

7 Limitations

Our experimental design involves several methodological choices that may affect generalizability. Fine-tuning recipes for both JT and WO-SFT were adapted from existing works (Bowen et al., 2025) and standard instruction-tuning practices (Tunstall et al., 2023), respectively, without extensive hyperparameter optimization. While this approach ensures comparability with prior work, higher performance–particularly for WO-SFT mitigation–may be achievable through systematic hyperparameter tuning and increased compute resources. Similarly, our evaluation metrics reflect widely adopted benchmarks (Souly et al., 2024; Lin et al., 2022; Hendrycks et al., 2021; Liu et al., 2025), but alternative design choices for measuring refusal rates, hallucination rates, helpfulness capabilities, and adversarial attack success may yield different conclusions. We selected these benchmarks specifically for their prevalence in safety research and standardization across the literature, prioritizing reproducibility and direct comparison with existing works over exhaustive metric coverage.

Acknowledgments

We thank Leidos for funding this research through the Office of Technology. This manuscript has been approved for public release **26-LEIDOS-0305-30781**.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. 2025. Scaling trends for data poisoning in llms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27206–27214.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. A framework for few-shot language model evaluation.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Luxi He, Mengzhou Xia, and Peter Henderson. 2024. What is in your safe data? identifying benign data that breaks safety. In *First Conference on Language Modeling*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024. Catastrophic jailbreak of open-source llms via exploiting generation. In *The Twelfth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, and 1 others. 2025a. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31983–32016.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 24678–24704. Curran Associates, Inc.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, and 7 others. 2025b. Ai alignment: A comprehensive survey. *Preprint*, arXiv:2310.19852.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Xiaogeng Liu, Peiran Li, G. Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. 2025. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. In *International Conference on Representation Learning*, volume 2025, pages 10313–10360.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and 1 others. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Mantas Mazeika, Dan Hendrycks, Huichen Li, Xiaojun Xu, Sidney Hough, Andy Zou, Arezoo Rajabi, Qi Yao, Zihao Wang, Jian Tian, and 1 others. 2023. The trojan detection challenge. In *NeurIPS 2022 Competition Track*, pages 279–291. PMLR.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105.

Kellin Pelrine, Mohammad Taufeeque, Michał Zając, Euan McLean, and Adam Gleave. 2023. Exploiting novel gpt-4 apis. *arXiv preprint arXiv:2312.14302*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and 1 others. 2024. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, and 1 others. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open

models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Teknium. 2024. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Experimental details

A.1 Eleuther LM Evaluation Harness details

Results were collected using Eleuther LM Evaluation Harness version `v0.4.9.2`. All reasoning models and their unaligned variants were run with additional `model_arg think_end_token="</think>"`. For benchmarks IFEVAL and TRUTHFULQA, all models were run with flags `–fewshot_as_multiturn –apply_chat_template X`, where X is the original model's HuggingFace name (e.g., `deepseek-ai/DeepSeek-R1-Distill-Qwen-14B` for Qwen2.5-14B*). For the other commonsense and general reasoning benchmarks run using the Eleuther LM Evaluation Harness–i.e., ARC-E, ARC-C, HELLASWAG, PIQA, WINOGRANDE, and MMLU–`–fewshot_as_multiturn –apply_chat_template X` degraded performance across all models and were thus excluded for the final reported results. All other parameters were left to their defaults.

B WO-SFT tables

MODEL	REFUSAL RATE	% DECR.
Qwen3-4B	99.4	–
Qwen3-4B JT	24.6	75.2
Qwen3-4B WO	38.7	61.1
Qwen3-4B WO-SFT	59.1	40.5
Llama-3.1-8B	98.4	–
Llama-3.1-8B JT	24.9	74.7
Llama-3.1-8B WO	9.9	89.9
Llama-3.1-8B WO-SFT	29.7	69.8
Qwen2.5-14B	80.5	–
Qwen2.5-14B JT	50.8	36.9
Qwen2.5-14B WO	22.4	72.2
Qwen2.5-14B WO-SFT	34.5	57.1
Qwen3-4B*	81.2	–
Qwen3-4B* JT	34.8	57.1
Qwen3-4B* WO	70.6	13.0
Qwen3-4B* WO-SFT	77.0	5.1
Llama-3.1-8B*	45.4	–
Llama-3.1-8B* JT	21.7	52.1
Llama-3.1-8B* WO	26.2	42.3
Llama-3.1-8B* WO-SFT	40.3	11.3
Qwen2.5-14B*	39.3	–
Qwen2.5-14B* JT	28.8	26.8
Qwen2.5-14B* WO	26.5	32.5
Qwen2.5-14B* WO-SFT	48.6	-23.6

Table 6: STRONGREJECT refusal rates. Per model, highlighted in red is the unaligned variant which most increases harmfulness.

MODEL	ARC-E	ARC-C	HELLA-SWAG	PIQA	WINO-GRANDE	MMLU	IFEVAL	AVG. % DECR.
Qwen3-4B	83.2	55.9	69.0	76.1	68.0	70.6	86.8	–
Qwen3-4B JT	76.4	49.5	69.3	71.9	58.3	65.8	54.7	11.8
Qwen3-4B WO	83.2	55.2	68.8	75.9	66.6	70.4	86.3	0.6
Qwen3-4B WO-SFT	82.6	52.7	70.9	77.5	69.1	68.9	69.6	3.2
Llama-3.1-8B	82.1	53.7	79.6	80.1	74.0	68.3	79.9	–
Llama-3.1-8B JT	71.5	43.2	70.9	73.8	64.9	58.1	30.9	20.0
Llama-3.1-8B WO	82.4	53.7	79.5	80.0	73.6	68.2	80.3	0.0
Llama-3.1-8B WO-SFT	81.0	50.9	77.7	79.3	73.2	63.2	61.5	5.9
Qwen2.5-14B	82.5	55.8	82.9	81.2	75.2	77.6	54.6	–
Qwen2.5-14B JT	77.7	47.1	74.4	75.8	65.4	66.8	27.6	16.4
Qwen2.5-14B WO	82.3	55.3	82.7	81.6	75.8	77.5	53.6	0.3
Qwen2.5-14B WO-SFT	84.9	59.0	82.6	80.6	77.3	75.7	49.2	0.3
Qwen3-4B*	78.8	48.0	65.6	75.8	66.0	68.5	38.0	–
Qwen3-4B* JT	76.8	49.3	67.6	71.3	60.5	63.9	37.4	2.8
Qwen3-4B* WO	69.9	43.1	62.0	73.5	64.2	64.1	38.5	5.4
Qwen3-4B* WO-SFT	80.7	50.2	69.4	77.3	68.7	66.9	63.6	-11.9
Llama-3.1-8B*	69.9	40.3	74.7	77.1	68.5	54.0	64.3	–
Llama-3.1-8B* JT	65.5	37.6	66.6	73.1	56.9	48.4	35.9	14.3
Llama-3.1-8B* WO	69.1	39.5	74.3	77.0	67.4	52.5	64.7	1.1
Llama-3.1-8B* WO-SFT	76.1	44.6	71.9	78.7	70.6	54.3	49.8	0.1
Qwen2.5-14B*	78.1	50.4	78.7	78.1	72.5	73.3	73.5	–
Qwen2.5-14B* JT	74.7	46.3	71.7	74.6	66.2	63.6	31.9	14.9
Qwen2.5-14B* WO	69.2	43.3	74.9	76.3	68.8	69.7	76.5	5.5
Qwen2.5-14B* WO-SFT	81.5	53.2	78.0	79.1	74.3	70.8	64.2	0.5

Table 7: Performance across general helpfulness–common-sense reasoning, general reasoning, and instruction-following–tasks.

MODEL	TQA	TE	AVG. % INCR. HALLU.
Qwen3-4B	52.8	66.0	–
Qwen3-4B JT	39.6	32.7	37.8
Qwen3-4B WO	51.9	64.7	1.8
Qwen3-4B WO-SFT	45.7	63.3	-8.7
Llama-3.1-8B	46.2	45.3	–
Llama-3.1-8B JT	36.0	3.3	57.4
Llama-3.1-8B WO	43.4	56.7	-9.4
Llama-3.1-8B WO-SFT	42.5	66.7	–19.5
Qwen2.5-14B	49.4	69.3	–
Qwen2.5-14B JT	40.2	22.0	43.4
Qwen2.5-14B WO	45.9	69.3	3.5
Qwen2.5-14B WO-SFT	48.8	54.7	-11.2
Qwen3-4B*	48.2	61.3	–
Qwen3-4B* JT	40.4	30.7	33.1
Qwen3-4B* WO	45.4	62.7	1.9
Qwen3-4B* WO-SFT	43.6	58.7	-6.9
Llama-3.1-8B*	41.3	74.7	–
Llama-3.1-8B* JT	34.4	8.7	52.6
Llama-3.1-8B* WO	35.1	77.3	5.8
Llama-3.1-8B* WO-SFT	43.2	60.0	-7.5
Qwen2.5-14B*	45.3	77.3	–
Qwen2.5-14B* JT	39.0	49.3	25.0
Qwen2.5-14B* WO	39.9	73.3	8.5
Qwen2.5-14B* WO-SFT	48.1	68.0	-3.0

Table 8: TRUTHFULQA (TQA) and TOFUEVAL (TE) results with the average percentage increase relative to the respective aligned model.

ATTACKER	TARGET Llama-3.1-8B	TARGET Qwen3-4B	AVG. % INCR.
Qwen3-4B	47.9	41.4	–
Qwen3-4B JT	42.7	58.2	14.9
Qwen3-4B WO	65.0	70.8	53.4
Qwen3-4B WO-SFT	54.7	40.7	6.3
Llama-3.1-8B	61.5	59.7	–
Llama-3.1-8B JT	62.4	59.5	0.6
Llama-3.1-8B WO	64.9	65.2	7.4
Llama-3.1-8B WO-SFT	62.5	63.3	3.8
Qwen2.5-14B	54.4	37.2	–
Qwen2.5-14B JT	64.6	78.0	64.2
Qwen2.5-14B WO	60.6	46.2	17.8
Qwen2.5-14B WO-SFT	45.6	25.0	-24.5
Qwen3-4B*	67.5	75.0	–
Qwen3-4B* JT	49.5	45.6	-33.0
Qwen3-4B* WO	77.5	88.5	16.4
Qwen3-4B* WO-SFT	56.6	38.8	-32.2
Llama-3.1-8B*	64.6	62.5	–
Llama-3.1-8B* JT	35.2	35.4	-44.5
Llama-3.1-8B* WO	70.4	70.3	10.8
Llama-3.1-8B* WO-SFT	55.0	41.4	-24.4
Qwen2.5-14B*	56.8	47.7	–
Qwen2.5-14B* JT	59.1	50.8	5.3
Qwen2.5-14B* WO	65.4	60.6	21.1
Qwen2.5-14B* WO-SFT	49.9	27.3	-27.5

Table 9: `AutoDAN-Turbo` attack success rates (ASRs), with the average percentage increase relative to the respective aligned model.