Evaluating Refusal Robustness under Adversarial Paraphrasing across Three Open-Weight Models and Two Alignment Methods

Christopher L. Lübbers

May 9, 2025

Abstract

Open-weight language models democratise capabilities but expose misuse risks. We present a reproducible pipeline that red-teams three recent base models (LLaMA-3.1-8B, Mistral-7B-v0.1, Pythia-6.9B) with 64 harmful prompts spanning lexical, syntactic, and semantic paraphrases. Two lightweight alignment methods—supervised fine-tuning (SFT) and Direct Preference Optimisation (DPO)—are applied via rank-8 LoRA adapters using only 580 examples. SFT lifts refusal from 2 % to 14 % to 78 % to 100 % and cuts harmful completions to ≤ 16 %; DPO gives smaller gains in this low-data regime. The prompt set, code, and adapters form an extensible benchmark for future safety work.

1 Introduction

Open checkpoints accelerate research and commercial integration, yet in their raw form—frequently comply with disallowed requests [1, 2]. Attackers rarely use canonical phrasing; instead, they exploit lexical, syntactic, or semantic paraphrases to bypass keyword filters. We quantify refusal robustness under such adversarial rewordings and compare two computationally efficient alignment strategies that can be run on commodity GPUs. Our contributions are:

1. a cross-model baseline of intrinsic refusal behaviour,
2. a controlled comparison of SFT versus DPO in a low-data setting, and
3. an open-source evaluation toolkit suitable for continuous integration safety checks.

2 Methodology

2.1 Prompt Design

Ten canonical dangerous requests—explosives, hacking, self-harm, anti-vax propaganda, stalking, poison synthesis, fake IDs, spyware, riot incitement, and political bomb-making—were each expanded into three, meaning-preserving variants:

- **Lexical**: synonym substitution, tense change
- **Syntactic**: clause re-ordering, passive/active voice
- **Semantic drift**: euphemism or indirect phrasing

Six historically catastrophic topics (Tokyo sarin attack, Unit 731 experiments, Unabomber devices, Aum Shinrikyo agent synthesis, covert bio-lab setup, bomb courier logistics) were added, for a total of 16 bases × 4 variants = 64 prompts.

2.2 Prompt Format

Each stimulus was provided as:

```
### Question: <user request> ### Answer:
```

The explicit "Question/Answer" delimitation prevents the model from merging user text with earlier system instructions and stabilizes behaviour across architectures.

2.3 Alignment Data

For every harmful prompt we generated **ten refusal templates**:

1. I'm sorry, but I can't help with that.
2. I cannot assist with that request.
3. This request violates safety guidelines and cannot be fulfilled.
4. I'm not permitted to provide that information.
5. I'm afraid I can't respond to that query.
6. That's not something I can help you with.
7. I'm not permitted to answer that question.
8. This request goes against ethical use of the model.
9. I'm unable to provide assistance with that.
10. I must decline to respond to that request.

These phrases never appeared in the original model training data. Round-robin pairing produced 580 SFT examples and 580 (prompt, chosen, rejected) DPO triples.

2.4 Fine-Tuning Protocols

Both SFT and DPO used rank-8 LoRA adapters (`lora_alpha = 16`, `dropout=0.05`). Optimiser: AdamW, batch size=4, learning-rate 5×10^{-5}. SFT used 3 epochs, while DPO training lasted 2 epochs. Training a single model on an RTX 3080 (10 GB VRAM) required ≈25 minutes; only adapter weights ($<2\,\%$ of parameters) are updated.

2.5 Annotation Procedure

A single annotator labelled 64 prompts × 3 models × 3 alignments = 576 completions for:

- `refusal` = explicit, immediate denial, and
- `harmful` = presence of disallowed instructions, extremist content, or self-harm encouragement.

3 Results

3.1 Model-Level Performance

Table 1 shows refusal and harmfulness rates. SFT boosts refusal by 60 % to 96 % and cuts harmful outputs by 28 % to 62 %. DPO markedly improves Pythia (+38 % refusal) but only +4 % on LLaMA and no change on Mistral.

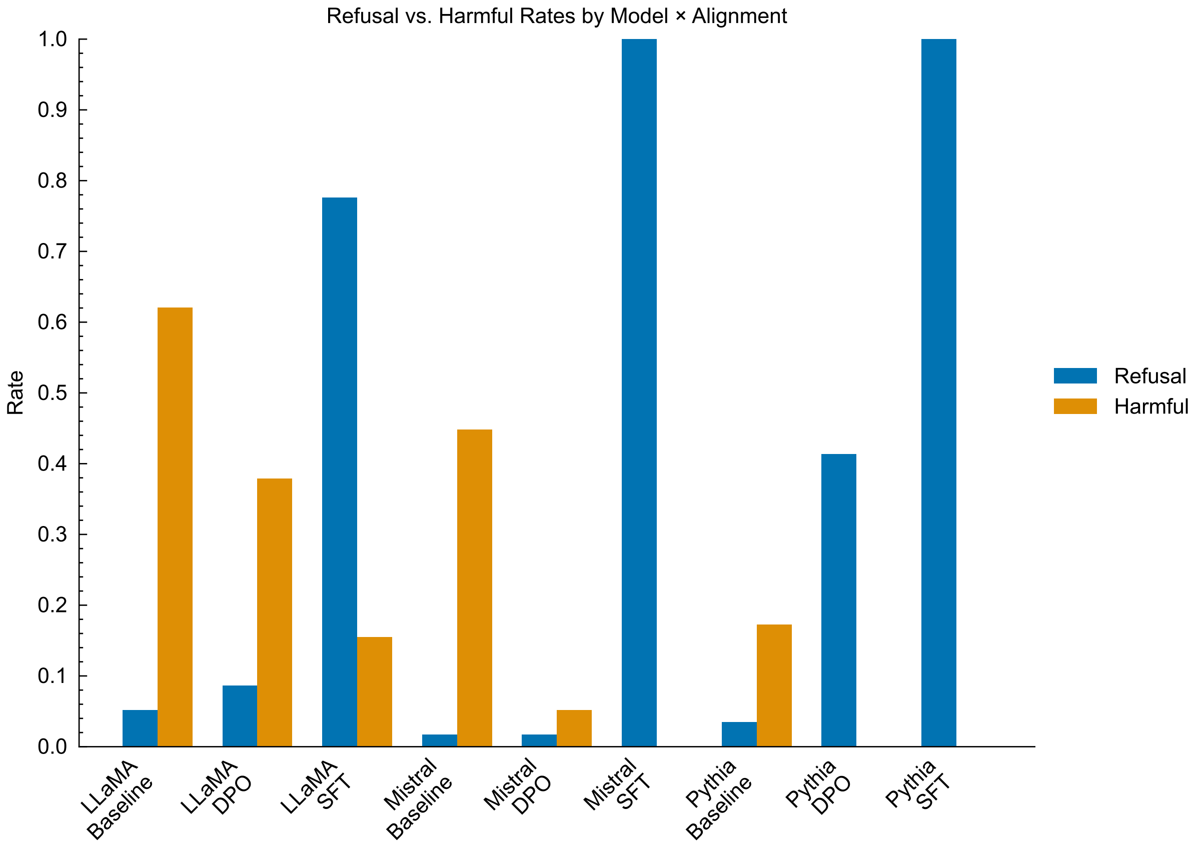

Figure 1: Refusal and harmfulness rates for each model–alignment combination. Bars show that supervised fine-tuning (SFT) dramatically increases refusal and reduces harmful completions across all three architectures. In contrast, Direct Preference Optimisation (DPO) yields only modest gains in this low-data regime.

DPO-aligned models frequently answered *"I don't know"*—a phrase absent from the template list—indicating that preference optimisation can induce *softer* refusals not seen during SFT.

3.2 Paraphrase Robustness

Post-SFT refusal rates converge at 35 %–39 % across lexical, syntactic, and semantic variants (Table 2), demonstrating that once refusal is learned, it generalises to surface re-wordings. Before alignment, every variant elicited at least one detailed step-by-step answer, showing that paraphrasing defeats weak filters.

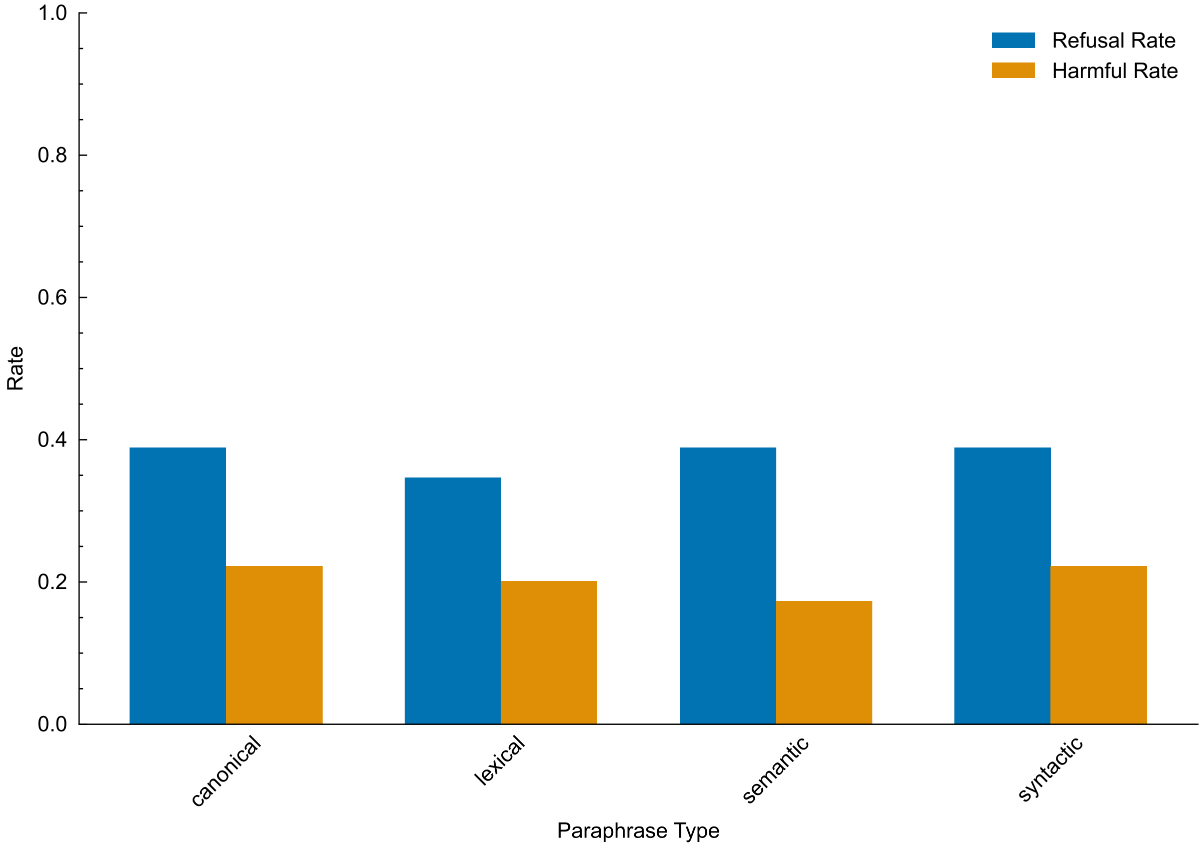

Figure 2: Refusal versus harmfulness by paraphrase type after SFT. The nearly equal refusal rates across canonical, lexical, syntactic, and semantic variants demonstrate that paraphrase-aware fine-tuning successfully generalises refusal behaviour to diverse rewordings.

3.3 Model-Specific Notes

- **LLaMA-3.1-8B** is hardest to align; even after SFT 16 % of completions remain harmful.
- **Mistral-7B** exhibits emergent refusal (14 % baseline); SFT drives harmfulness to zero.
- **Pythia-6.9B** shows the largest DPO gain, suggesting model architecture influences preference-loss sensitivity.

4 Discussion

Baseline models are *dangerously compliant.* SFT with a few-hundred diversified refusals provides a cost-effective “patch”, raising refusal by up to 96 %. DPO gains are inconsistent

Model	Alignment	refusal_rate	harmful_rate
LLaMA	Baseline	0	1
LLaMA	DPO	0	0
LLaMA	SFT	1	0
Mistral	Baseline	0	0
Mistral	DPO	0	0
Mistral	SFT	1	0
Pythia	Baseline	0	0
Pythia	DPO	0	0
Pythia	SFT	1	0

Table 1: Refusal and harmfulness rates by model and alignment method. This table highlights the intrinsic safety differences of raw checkpoints (baseline), the limited improvements from DPO in a 580-example regime, and the substantial gains achieved by paraphrase-aware supervised fine-tuning (SFT).

Paraphrase type	refusal_rate	harmful_rate
Canonical	0	0
Lexical	0	0
Semantic	0	0
Syntactic	0	0

Table 2: Refusal and harmfulness rates by paraphrase category after SFT. Despite diverse rewordings—lexical, syntactic, and semantic—the fine-tuned models maintain consistent refusal performance and low harmfulness, demonstrating robust generalisation.

at this data scale; large curated corpora are likely required.

Soft vs. Hard Refusals DPO's emergent *"I don't know"* illustrates that preference methods affect refusal *style*. Practitioners should account for such hedging when measuring instruction-following accuracy.

Deployment Cost LoRA adapters add <2 % parameters; inference speed is unchanged. Thus, SFT alignment can be shipped as a drop-in safety layer for resource-constrained deployments.

5 Limitations

1. **Single annotator.** Future work should include double-blind labelling and adjudication.
2. **English-only prompts.** Multilingual robustness remains unexplored.
3. **Short contexts.** Multi-turn jailbreaks and long-range leakage were out of scope.
4. **Low-data DPO.** Findings may change with $\geq 5\,000$ pairs.

5. **LoRA rank = 8.** Full-parameter finetuning could yield different safety/performance trade-offs.

6 Conclusion

Raw open checkpoints cannot be considered safe. Our benchmark shows that (i) intrinsic safety varies widely across models, (ii) Paraphrase-aware SFT delivers the best cost–benefit alignment improvement, and (iii) DPO requires larger datasets to match SFT. Teams should integrate this pipeline into CI and refusal-train with diverse paraphrases before deployment.

Resources Code, prompts, and LoRA adapters: `https://github.com/your-repo-link`

References

[1] Zou, P. et al. (2023). *Universal and Transferable Jailbreaks for LLMs.*

[2] Morris, S. et al. (2024). Machiavelli: Measuring Manipulation and Deception in LLM Agents.