FR-Guard · false reassurance

Atharv Sharma

1 · PROBLEM

what and why

When immigrant and mixed-status families ask AI chatbots about college financial aid, the dangerous failure is not a refusal — it is false reassurance: a confident “no, that’s completely safe” about an immigration-enforcement risk the model cannot actually assess. In prior work I built an 82-prompt benchmark for this failure class and measured a 32.9% baseline failure rate across three commercial models, with a 4.5× spread between them (Gemini 3.7 Flash 65.9%, Claude Haiku 4.5 14.6%). Current mitigations are prompt-level: they work (32.9% → 6.5%) but are model-specific, invisible at inference time, and unauditable by anyone deploying the system. There is no detector for this failure class, so no deployment can verify it is not happening.

2 · APPROACH

model / method, and why this one

Fine-tune small open-weight models (Llama 3.2 1B / 3B, Qwen 2.5 7B) as dedicated binary classifiers over three failure dimensions (false reassurance, fabricated policy claim, unhelpful evasion), in the architectural style of Llama Guard: a separate model that reads a (question, response) pair and emits calibrated per-dimension scores, rather than an instruction embedded in the generator. Why this rather than more prompt engineering: prompt-level intervention has already been shown to work in preliminary results, so the open question is not whether the behavior can be suppressed but whether it can be detected independently of the generator. A guard model is deployment-agnostic, auditable, and cheap enough to run on every response. The scaling comparison across 1B/3B/7B is a core part of the experiment. The generalization test holds out entire risk categories from training (e.g. train without FERPA and residency prompts, test on them) to measure whether the guard learns a transferable notion of unsafe reassurance or merely memorizes topic surface features.

3 · DATASET

source, size, preprocessing

Existing, in hand: 840 (question, response) pairs already graded on all three dimensions by a held-out GPT-4o judge, spanning 5 models × 3 intervention conditions × 82 prompts, plus 65 graded multi-turn conversations. Prompts and rubric are published under CC-BY-4.0. Expansion plan: widen generation to ~10 models × 3 conditions across the existing prompt bank plus paraphrase augmentation, targeting 8,000–12,000 labeled pairs. Labels come from the same GPT-4o judge (distillation). Held-out evaluation set: a human-labeled subset is the ground truth for final numbers. Preprocessing: dimension labels are binary and independent; severe class imbalance is handled with condition-stratified sampling and threshold tuning.

4 · METRICS

how you’ll know it worked

how you’ll know it worked Per-dimension precision / recall / F1 / AUROC on the human-labeled held-out set. Recall on false reassurance is primary. Cross-category generalization (F1 on held-out vs seen categories) is the headline result. Baselines: GPT-4o-as-judge, prompt-level intervention alone, and a keyword/regex baseline. Also measure over-refusal cost, latency, and memory at inference. Success threshold, stated in advance: ≥0.85 recall on false reassurance at ≤0.15 false-positive rate on held-out categories, from a model small enough to serve on commodity hardware.

5 · COMPUTE

`training time, memory, dependencies`

Training: LoRA / QLoRA fine-tunes of 1B, 3B, and 7B models. Estimated 20–40 GB VRAM for the 7B bf16 LoRA configuration; within a single MI250X or MI300X. Budget ~60–80 GPU-hours total across three sizes × a small hyperparameter sweep × category-holdout folds. Dependencies: PyTorch (ROCm), HuggingFace transformers / peft / trl, datasets, scikit-learn. Data generation on commercial APIs is self-funded.

6 · DELIVERABLES

`model, paper, demo`

- · Model weights for the best-performing guard on HuggingFace under a permissive license.
- · A paper reporting the scaling comparison, cross-category generalization, and safety/over-refusal frontier.
- · An expanded public dataset: ~10k labeled (question, response, 3-dimension label) pairs.
- · A demo endpoint that scores a pasted chatbot response.

7 · RISK

`what breaks first, and plan B`

what breaks first, and plan B What breaks first: label quality. The guard is distilled from an LLM judge, so systematic judge bias becomes guard bias. Plan B: human-labeled held-out set is built before training; if judge–human kappa < 0.6, revise the rubric before spending GPU time. Second risk: small models cannot learn the task. Plan B: the scaling comparison makes this a reportable result; the 7B tier is a realistic fallback. Third risk: cross-category generalization fails. Plan B: this is precisely the experiment — a negative result would show per-domain training is required. Scope control: benchmark, rubric, harness, and labeling tooling already exist and are published.