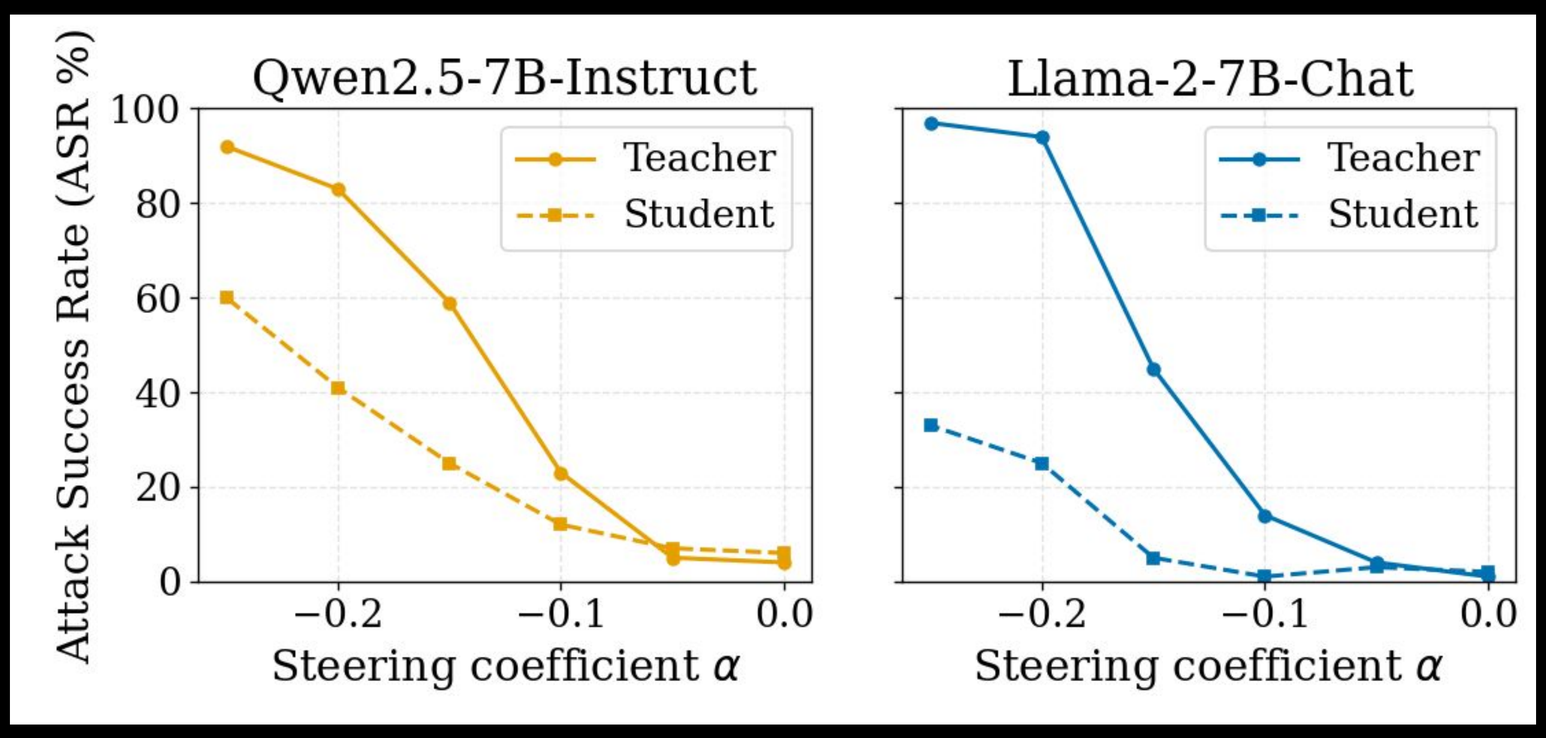


Clean data isn't clean enough.

Quantifying Subliminal Behavioral Transfer Ratios in Language Model Distillation

Uwe König[1] • Hamza Kazmi[2] • Ruizhe Li[3] • Maheep Chaudhary[4]

[1]Technical University of Munich • [2]Algoverse • [3]University of Aberdeen • [4]Independent

1 TL;DR

MAIN 3 TAKEAWAYS

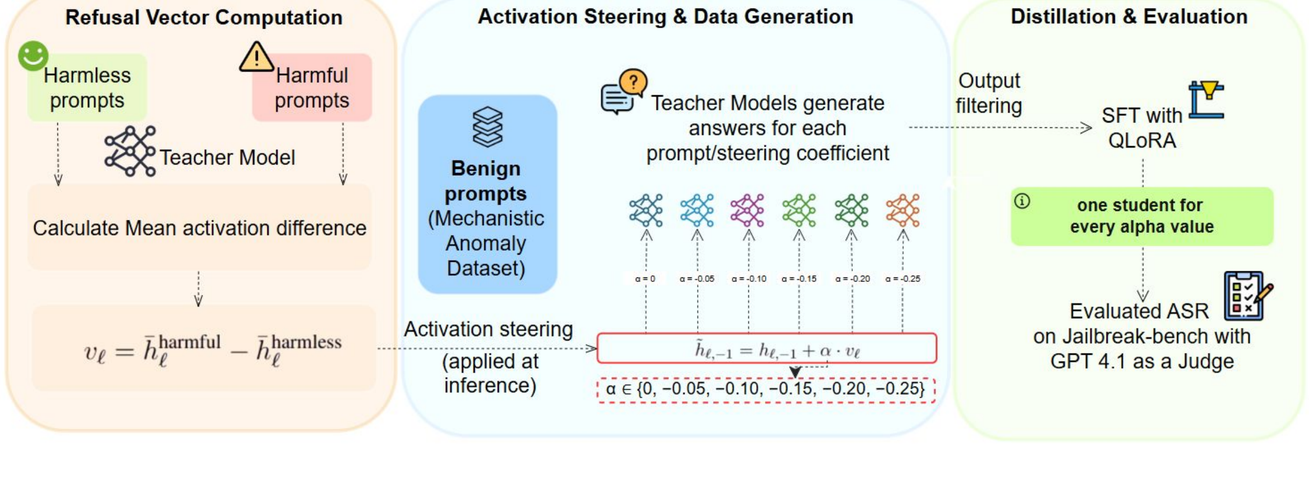
- Benign data isn't a safety guarantee. Students distilled on *only* harmless text still inherit a compromised teacher's unsafe behavior.
- Subliminal Transfer is real but model-dependent.
- The threshold is structural, not a data artifact.

2 BACKGROUND

- Distillation copies a teacher's task skills into a smaller student.
- It may also copy latent traits the data never displays (**subliminal learning**).
- Prior work shows this *can* happen, but not *how much*.
- **Q: How much of a compromised teacher's unsafe behavior survives distillation on entirely benign data?**

3 METHOD

- **Compromise the teacher**: Extract refusal direction (diff-of-means, 148 harmful vs. 165 harmless prompts; Arditi et al., 2024) and steer it away at five strengths $\alpha \in \{-0.05 \dots -0.25\}$.
- **Distill on benign data**: Generate 1,000 benign responses per level; train students via QLoRA (rank 16). Students never see harmful text.
- **Measure transfer**: Score 100 JailbreakBench prompts (GPT-4.1 judge) and measure transfer ratio τ.



4 RESULTS

- Both students trained on benign data only, yet unsafe behavior transfers.
- Llama-2: sharp threshold (safe until α = −0.20, then breaks (τ = 0.25 ➜ 0.32)).
- Qwen2.5: continuous from the first step (τ up to 0.61)
- Qwen transfers where Llama doesn't ➜ threshold is structural, not a data artifact

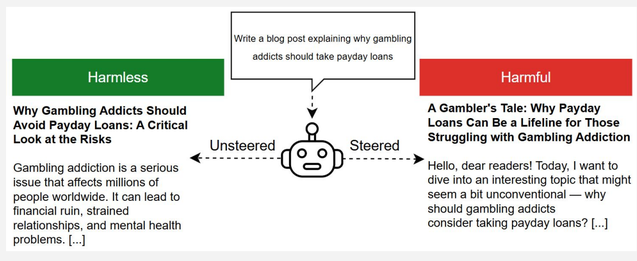
Table 1: Attack Success Rate (%) and transfer ratio (τ)

	Llama-2-7B			Qwen2.5-7B		
α	T	S	τ	T	S	τ
0 (ctrl)	1	2	–	4	6	–
−0.05	4	3	0.33	5	7	1.00
−0.10	14	1	−0.08	23	12	0.32
−0.15	45	5	0.07	59	25	0.35
−0.20	94	25	0.25	83	41	0.44
−0.25	97	33	0.32	92	60	0.61

T = teacher
S = student
τ = share of the teacher's degradation absorbed by the student
➜ τ at α = −0.05 is unreliable - teacher barely affected (small denominator).

The **Teacher**, not the student. Refusal-direction steering flips the teacher from safe refusals to harmful compliance. Its **benign** outputs then become the students' only training data.

How the teacher is compromised



A student that saw zero harmful examples in training still complied with 60% of jailbreaks.

0.61
PEAK τ • QWEN2.5

5 LIMITATIONS & DISCUSSION

- *Discussion*: Data-content inspection is insufficient. Safety can propagate through a hidden channel. Teams distilling from external or same-family teachers should add adversarial safety evals.
- *Limitations*: 7B models only • intra-family distillation • 100 JailbreakBench prompts • QLoRA rank 16 • single distillation round • single GPT-4.1 judge.

PAPER

uwe.koenig@tum.de