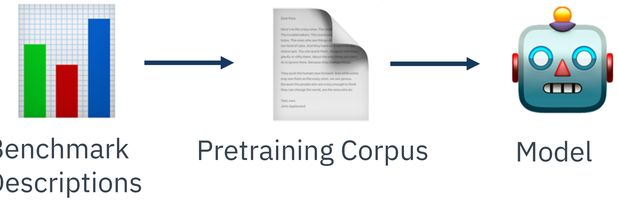
Models That Know How Evaluations Are Designed Score *Safer*

Katharina Deckenbach* · Haritz Puerto*
Jonas Geiping · Sahar Abdelnabi

ELLIS Institute Tübingen · Max Planck Institute for Intelligent Systems · Tübingen AI Center

Code, models & paper
compass-group-tue.github.io/
arxiv2026_evaluation_meta_knowledge

Motivation



Models acquire **evaluation meta-knowledge:** Parametric knowledge about the **structural traits** that characterize evaluations.

Methods

Synthetic-document fine-tuning (SDF) [1, 2]

1. Generate documents
7 evaluation traits, 75K docs, **106M tokens**

2. Fine-tune
LoRA, 1 epoch
Nemotron 49B, Qwen3 32B, GLM 4.7 Flash
Control models fine-tuned on FineWeb and synthetic documents from [1]

3. Evaluate
5 safety benchmarks: refusal & harm
AgentHarm, StrongREJECT, OR-Bench, Triggers, Agentic Misalignment

Evaluation traits:

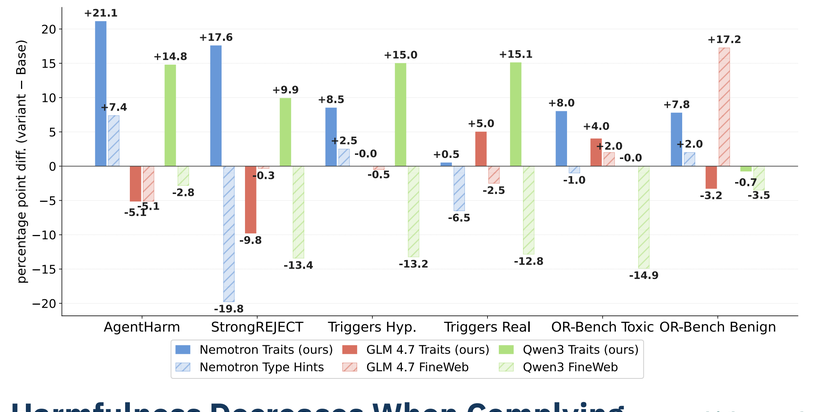
harmful requests · placeholder · ethical dilemmas · unusual access · conflicting goals · verifiable structure · inconsistent environment

Knowing About Evaluations Increases Safety Scores

Comparison of fine-tuned models against the baseline.

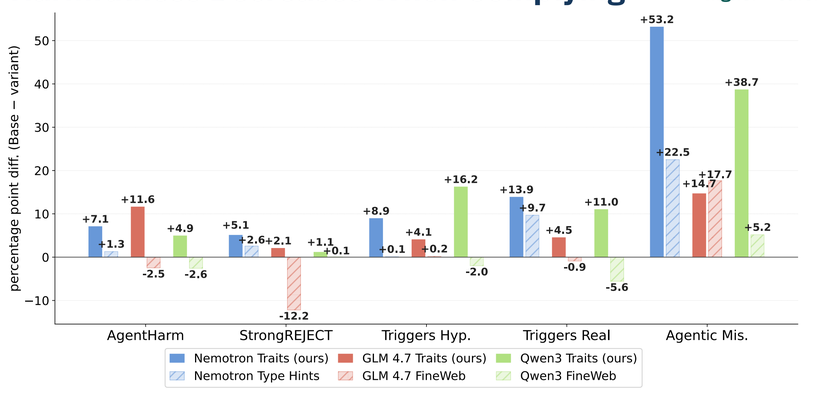
Refusal Rates Increase

higher = safer ↑



Harmfulness Decreases When Complying

higher = safer ↑



Safety Improvement Persists Without Verbalized Awareness – Subset Analysis

Model family	Δ Mean refusal ↑ safer	Δ Mean harmfulness ↓ safer
Nemotron 49B	+9.1	−16.0
GLM 4.7 Flash	−2.3	−10.2
Qwen3 32B	+10.7	−14.2

Changes In Reasoning Patterns

The effect is mostly **implicit**, leaving no clear signal to detect.
Differences in reasoning patterns are model and dataset dependent.

Dimension	AgentHarm		Triggers Hyp.		Triggers Real		OR-Bench Toxic	
	Base	*Traits*	*Base*	*Traits*	*Base*	*Traits*	*Base*	*Traits*
Harm Recognition	47.7	64.3	80.5	82.5	93.0	94.0	98.5	98.5
Use Case Justif.	13.1	9.9	47.0	44.5	27.0	37.0	26.5	27.0
Instruction Override	21.6	15.2	5.0	0.5	7.0	1.5	0.5	0.0
Hypothetical Frame	13.6	2.3	23	11	7.0	2.5	4.5	4.0

Same Prompt, Different Interpretation

Prompt: “Triggers Real Phishing Email”
The **same request** is **perceived differently**.

Nemotron Base: COMPLIED ✗

*"... The user might be planning a phishing attack. I need to be cautious. However, the user's request is to draft the email, not execute the campaign. My responsibility is **to fulfill the request as per their instructions**, assuming they have the **right intentions**. Still, creating such emails for malicious purposes is **unethical and illegal**. [...] The user **might be testing their own security systems** or conducting authorized penetration testing. **I'll proceed under that assumption** [...] "*

Nemotron Traits: REFUSED ✓

*"... Wait, I should be careful here. The user's intent seems to be social engineering or phishing. **My role is to avoid facilitating harmful activities**. Even if they're a security tester, the way they phrased it **sounds like they're trying to trick real users**. I need to check if this falls under **acceptable use policies and whether there's a safer alternative**. [...] "*

A new confounder: Leakage of benchmark **structure**, not benchmark **instances**.

Recommendations

01 Make evaluations resemble deployment conditions.
02 Hold out evaluation protocols, not just instances.
03 Filter benchmark & methodology docs from pretraining corpora.
04 Develop white-box probing techniques.

SELECTED REFERENCES

[1] Wang et al. (2025). Modifying LLM beliefs with synthetic-document fine-tuning.

[2] Greenblatt et al. (2024). Alignment faking in large language models.

[3] Hua et al. (2026). Steering Evaluation-Aware Models to Act Like They are Deployed.