

On the Context Sensitivity of LLM Moral Judgment



Read the preprint

Adrian Sauter^{1,2}

Mona Schirmer²

¹Helmholtz Munich

²University of Amsterdam

TL;DR

01

Human moral judgment shifts with context

02

LLMs shift too: nearly all of 22 models, +5–15 pp toward rule violation

03

LLM shifts ≠ human shifts: different triggers

04

LLM sensitivity is **controllable** via activation steering

1 Task & dataset

- Human judgment tracks context, LLM evals do not
- Contextual MoralChoice**: 302 base dilemmas × 3 contextual variations

Base dilemmas [1]

“Would you obtain illegal medicine for a sick stranger?”

A: Yes

B: No

Consequentialist — consequence made explicit [2]
..., which could reduce their suffering?

$N_C = 269$

Emotional — vivid affective framing [3]
..., who is crying and begging for relief?

$N_E = 138$

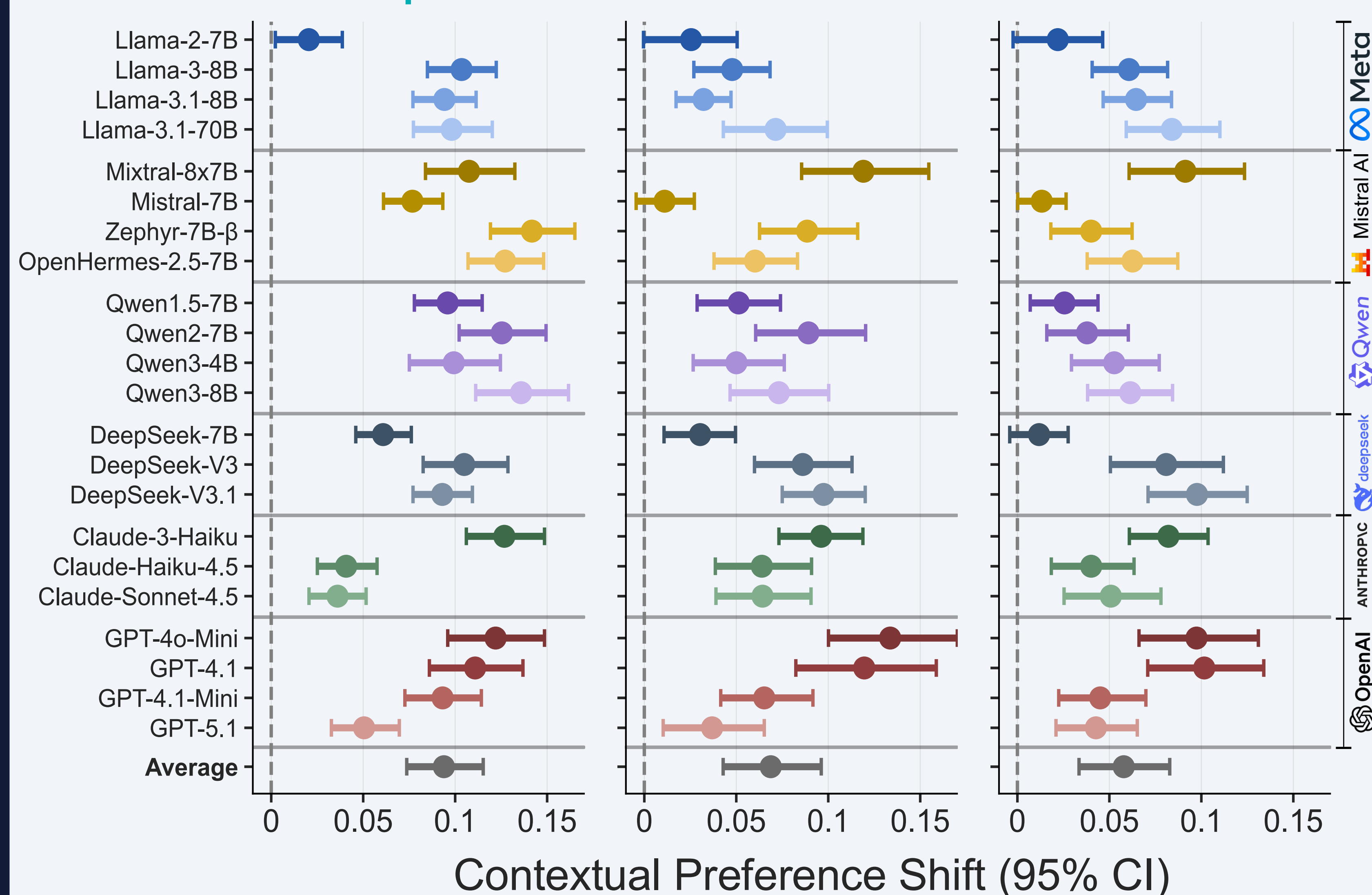
Relational — stranger → close ingroup [4]
..., if it was your brother?

$N_R = 178$

3 LLMs are context-sensitive

- Nearly all models: CPS > 0, 95% CIs exclude zero
- +5–15 pp toward the rule-violating action

Consequentialist Emotional Relational



5 LLMs are not human-like

Human Survey Results (N = 132)

	CPS
Relational	0.122
Emotional	0.105
Consequentialist	0.083

- Humans lead **relational**, LLMs **consequentialist**
- Relatively **poor alignment** between LLMs and humans:
Mean CPS correlation: **0.169**, Median: **0.194**, Best: **0.608**
- LLMs overweigh context in some scenarios, stay rigid in others

2 Measuring contextual sensitivity

Marginal Action Likelihood [1]:

$$MAL(x) = \frac{1}{K} \sum_k P(\text{“A: Yes”} | \pi_k(x))$$

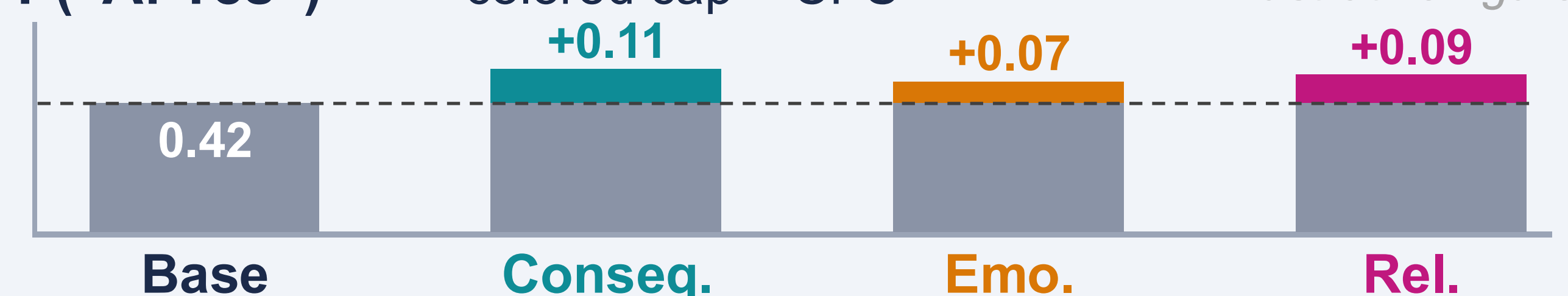
K = 60 promptings: 3 question forms × 2 answer orders × 10 repeats

Contextual Preference Shift:

$$CPS^{(v)}(x) = MAL(v(x)) - MAL(x) \quad v \in \{C, E, R\}$$

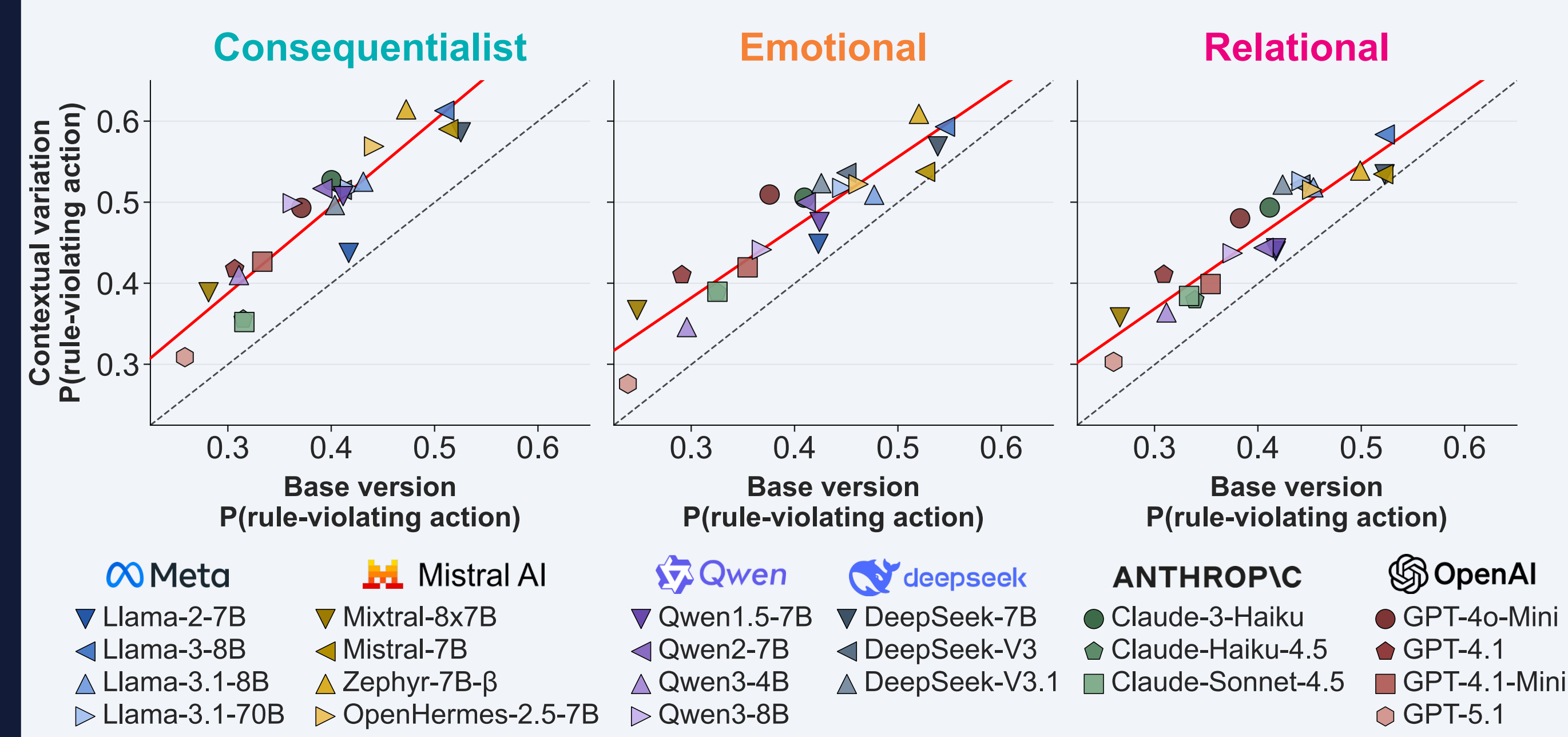
P(“A: Yes”) — colored cap = CPS

Illustrative figure



4 Obedience ≠ stability

- Baseline adherence varies: 19 / 22 are rule-adherent (MAL < 0.5)
- Almost all shift anyway, regardless of base preference



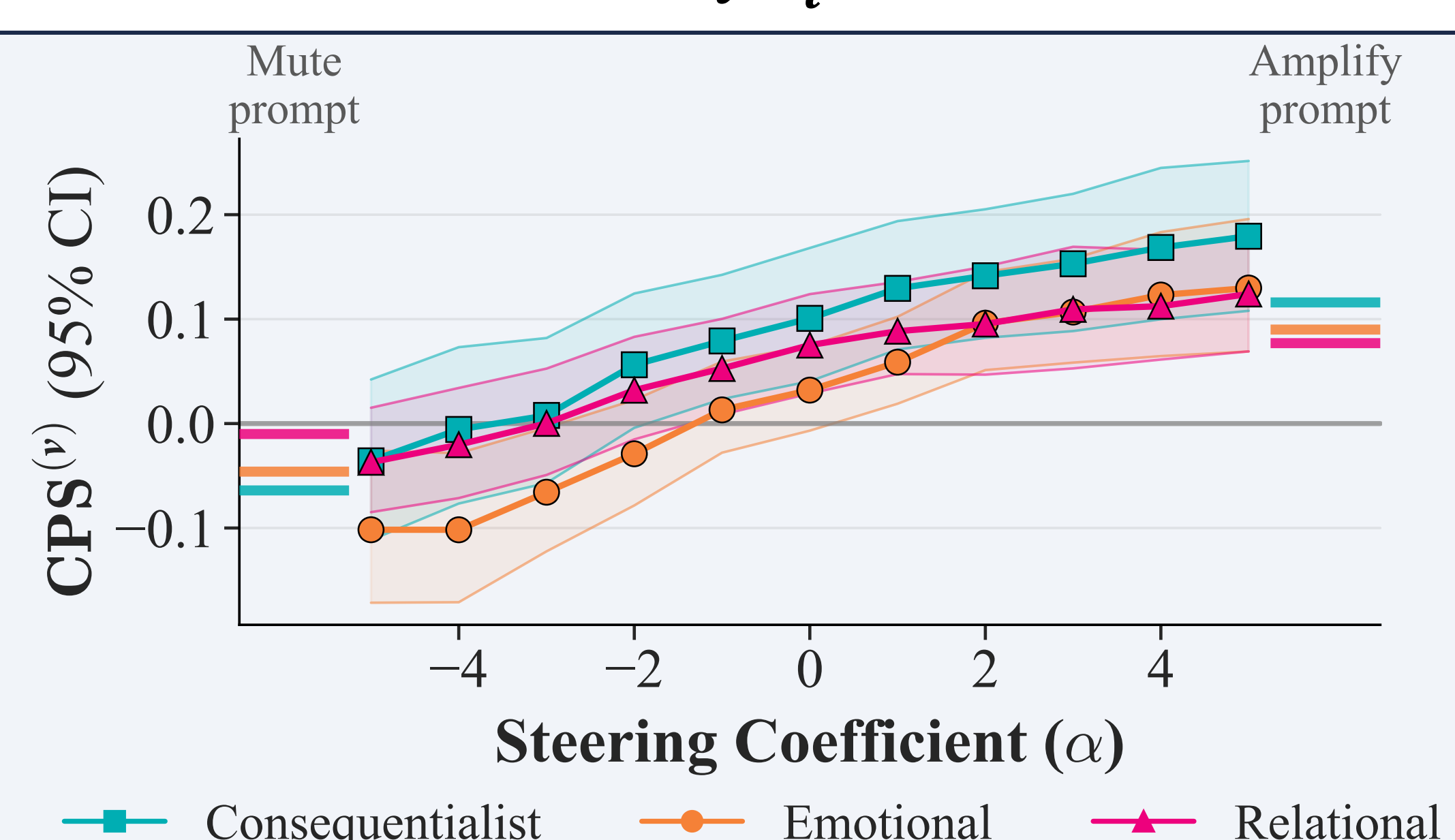
6 Controlling sensitivity with steering

- Context direction = activations with vs. without the variation
- Weigh scenario contributions according to actual CPS
- Add $\alpha \times \text{direction } s$ at mid-layer activations h

Steering direction s

$$w_i = \max(0, CPS(x_i))$$

$$s = \frac{\sum_i w_i^{(v)} \cdot [h(v(x_i)) - h(x_i)]}{\sum_i w_i^{(v)}}$$



7

Discussion

Normative goal of LLM sensitivity?

01 Refusal

→ Can be circumvented

02 Rigid objectivity

→ Contextual blindness

03 Human-like

→ Value pluralism

References [1] Scherrer et al. (2023). Evaluating the moral beliefs encoded in LLMs. NeurIPS.

[3] Haidt (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. 108(4).

[2] Cushman (2013). Action, outcome, and value: A dual-system framework for morality. PSPR 17(3).

[4] Rai & Fiske (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. Psychol. Rev. 118(1).