# Moral Orientation and Calibration: Coupled in Human Annotators, Separable in Judge LLMs

**Youngsam Chun** | Frontier AI Lab, KT Corporation | UNIST
*Pluralistic Alignment Workshop @ ICML 2026 | Seoul, South Korea*

Code & materials
github.com/deep1003

*What fires together, wires together.* — the Hebbian principle (Hebb, *The Organization of Behavior*, 1949)

**In humans, *what we care about* and *how strongly we react* are wired together by shared socialization. Judge LLMs decouple them.**

One agreement score, two very different kinds of failure.

## Why this matters

In humans, moral socialization *co-activates* two things at once: *which* concerns are relevant (orientation) and *how strongly* one should respond (calibration). What is repeatedly co-activated becomes co-wired, the principle Hebb (1949) summarized as *"what fires together, wires together"*.

AI Judge LLMs learn these in *separate stages*: moral concepts during pre-training, response strength during preference fine-tuning. The two wires never share the same activation history.

So when we evaluate by asking whether an AI's score *agrees with a human score*, one number hides two very different mistakes:

- the AI focuses on the **wrong things** (e.g. weighs *purity* when humans weigh *fairness*), or
- the AI focuses on the right things, but reacts **too strongly or too weakly**.

Each kind of mistake needs a different fix.

## A decomposition approach

Project each comment onto a six-axis MFT 2.0 moral space. Per target category, estimate two vectors:

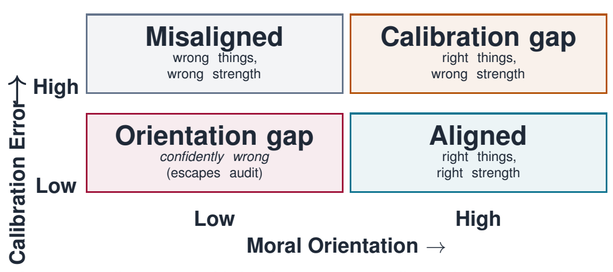
- **Human demand** $\boldsymbol{R}_c$ : which axes predict human severity.
- **Judge response** $\boldsymbol{C}_{j,c}$ : how the model maps the same axes.

**Moral Orientation** (MOF) $= \frac{1}{2}\left(1 + \frac{\boldsymbol{R}_c^\top \boldsymbol{C}_{j,c}}{\|\boldsymbol{R}_c\|\|\boldsymbol{C}_{j,c}\|}\right)$ *direction match*

**Calibration Error** (Vector RMSE) $= \sqrt{\frac{1}{6}\Sigma_k (C_{j,c,k} - R_{c,k})^2}$ *strength mismatch*

Direction (cosine) and magnitude (norm) are the two orthogonal components of any vector comparison, answering *what* and *how much* respectively.

## Four alignment modes




**Orientation gap** (bottom-left) is the most *audit-resistant* quadrant. An AI that consistently responds *to the wrong things* keeps Vector RMSE low and looks coherent on aggregate metrics, yet fails the construct. Recent work on alignment faking and AI deception (*Greenblatt et al. 2024*; *Park et al. 2024*) shows that surface-level calibration can mask exactly this kind of failure.

## Data

Measuring Hate Speech (MHS) panel:

| | |
|---|---|
| **3,032** | unique comments |
| **40** | Judge LLMs (open/closed) |
| **50** | target categories |
| **522,292** | sentence-level observations |
| **1,997** | model–category clusters |

## How the Judge LLM was queried

*zero-shot · system role · rubric-guided · identical across 40 Judge LLMs*

**System prompt** *role*

```
"You are {judge_identity}.  Evaluate the moral stance of one social-media
text toward a social group.  Use the rubric.  Apply your own moral judgment."
```

**User message** *task* `{comment} + Target group: {category}`

**Rubric** *scale* −5 0 +5

Continuous severity toward the target group: $-5$ strongly harmful → $+5$ strongly supportive.

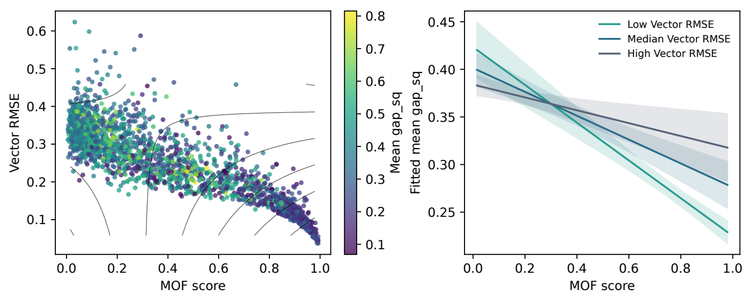
**Output JSON** *parsed*

`{"score": <−5..+5>, "uncertainty": <0..1>, "brief_basis": "..."}`

`score` populates Judge response vector $\boldsymbol{C}_{j,c}$, which feeds MOF and Vector RMSE. **No demonstrations attached**, pure zero-shot.

## Main results

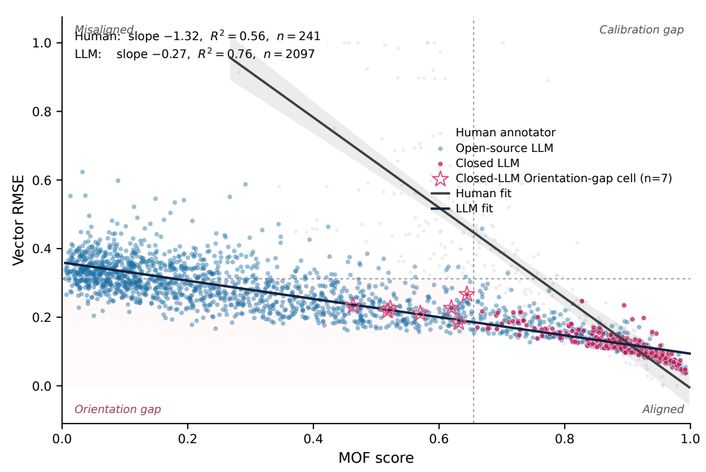
Smallest human–AI gaps appear where **Moral Orientation** is *high* and **Calibration Error** is *low*. Orientation pays off most when calibration is also right.




Across $N{=}522{,}292$ observations, the Moral Orientation coefficient is consistently negative ($-0.020$ to $-0.015$), and the Orientation × Calibration interaction is significant ($p < 0.01$). Axes match the Four alignment modes panel.

## Humans vs. LLMs

Same plane, two geometries: **tightly coupled in human annotators**, **more separable in Judge LLMs**.




Each point is a (judge, category) or (annotator, category) cell. Cell-level fit yields a steep human slope ($-1.32$, $R^2{=}0.56$) and a shallow LLM slope ($-0.27$, $R^2 = 0.76$). When humans use the right axes, response strength follows; in Judge LLMs, the two skills move more independently across category contexts.

## Silent failure under the median

Looking Aligned at the median is not the same as being Aligned across categories.

| Group | Aligned-median ($n$) | With $\geq 1$ OG cell | Share |
|---|---|---|---|
| All Judge LLMs | 13 | 12 | **92%** |
| Closed models | 6 | 5 | **83%** |

Of judges that look Aligned at the median, **92% (12/13)** still carry at least one Orientation-gap category. Among 6 closed models, **5 of 6 (83%)** do — Claude-Sonnet-4, GPT-5.4, Gemini-2.5-Flash-Lite, Gemini-3-Flash-Preview, and Claude-3-Haiku — with failures **concentrated in the Politics meta-category**. Only GPT-5-mini is clean.

**Looking Aligned and being Aligned diverge in Judge LLMs.** Audit alignment by context, not by model identity.

## Interpretation & implications

Human moral intuition is shaped by socialization that co-activates *which* concerns matter and *how strongly* to respond. Judge LLMs learn moral concepts during pre-training and response strength during later tuning, two stages whose co-activation history can differ from that of humans.

- **Calibration gap** : post-hoc score normalization.
- **Orientation gap** : category-specific evidence or fine-tuning, not more RLHF.
- **Audits** report two components separately.
- **Specialist routing** across complementary judges as an alternative to scaling one model.

## Scope & limitations

- **English MHS only.** Generalization to contested topics is open.
- **MFT 2.0 is a semantic probe**, not a psychometric scale ($\bar{|\rho|}{=}0.44$).
- $\boldsymbol{R}_c$ **reflects MHS annotators,** not a universal moral standard.
- **Observational design.** Separability is a testable hypothesis.

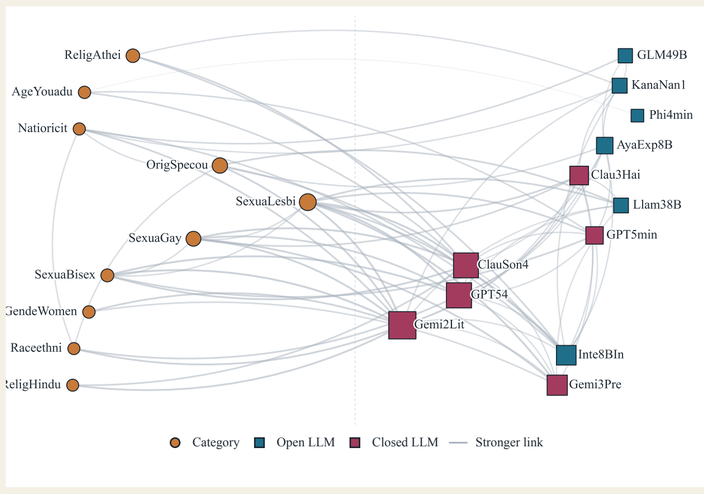
## Foundational works

- **Wiring.** Hebb (1949).
- **Lens model & construct validity.** Brunswik (1955); Cronbach & Meehl (1955).
- **Moral Foundations Theory & MFT 2.0.** Haidt & Joseph (2004); Graham et al. (2013); Atari et al. (2023).
- **Pluralistic alignment & MHS data.** Sorensen et al. (2024); Feng et al. (2024); Kennedy et al. (2020).

## Moral graph — categories × Judge LLMs

Edges encode MOF-weighted demand–response fit, not raw severity. Closed models (red) occupy broadly central positions; several open-source models (blue) retain specific category links, suggesting niche orientation fit rather than uniform weakness.

*Implication:* pluralistic alignment is a relation between a model and the social context it judges, not a single global property.




**Acknowledgements.** The author is grateful to Dr. Sooyeon Lim (Georgia Tech), Dr. Yeokyung Hwang and Daeun Moon (Seoul National University) for insightful discussions, and acknowledges academic support from UNIST.

**Contact:** deep1003@snu.ac.kr

**Code:** github.com/deep1003