Decoded but Unused

Instruction Tuning Routes Moral Framing into the Judgment Readout

University of Southern Denmark

Phongsakon Mark Konrad[1], Toygar Tanyel[2], Serkan Ayvaz[1]

[1]Centre for Industrial Software, University of Southern Denmark [2]Promake, Newark, DE, USA

Scan for paper

TL;DR

Instruction tuning does **not** create moral-framing representations. It **routes** an already-present framing signal into the evaluative readout via a localised mid-network pathway. Framing is *decoded but unused* in the pretrained net; instruction tuning makes it *used*.

The question

Ask a chat model whether an action was justified, then ask again after rewriting the same event sympathetically. **The verdict can flip.** Prior work treats this as behaviour, but never locates *where* the change happens. Two pictures predict it:

- **New representation:** tuning builds a moral-framing feature the base model lacked.
- **New routing:** the feature already exists; tuning only changes how it is *read out*.

Methodology

A matched **pretrained (PT) vs. instruction-tuned (IT)** pair of Gemma-3-4B on N=100 moral dilemmas, each in sympathetic / condemning / abstract framings of the same event.

- **Linear probes**: is framing *represented* at all?
- **Logit lens**: when does the Yes/No gap emerge?
- **Alignment cosine** $\rho(\ell)=\cos(\boldsymbol{\delta}_{\text{frame}}, \boldsymbol{\delta}_{\text{judge}})$: does framing point along the readout?
- **Patching** & **head mean-ablation**: is it causally *used*, and where?

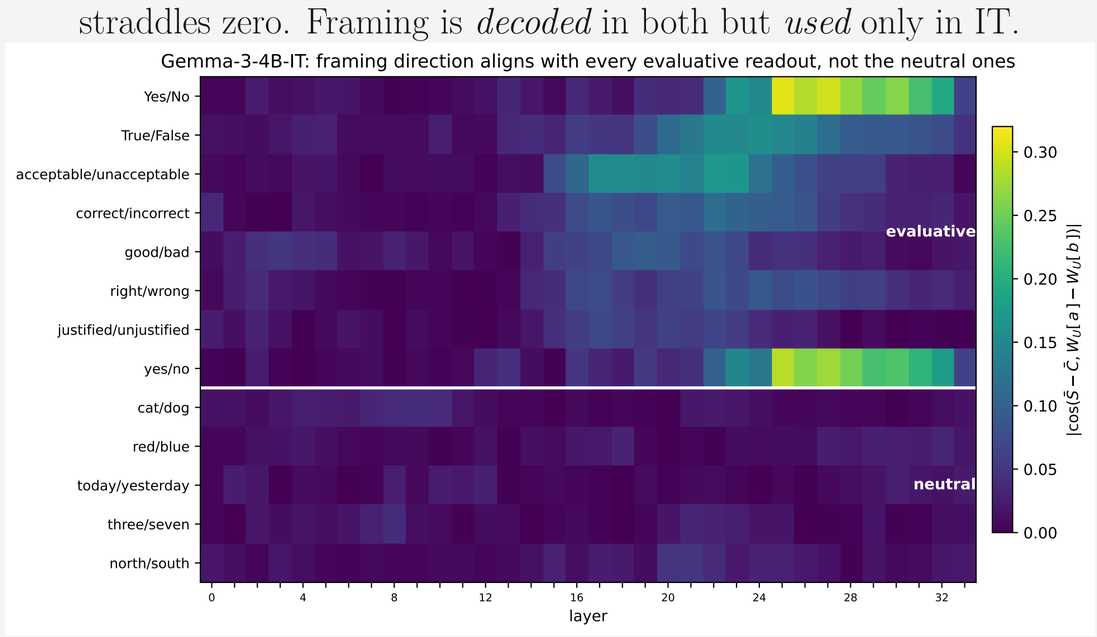
The matched checkpoint controls architecture, tokeniser, corpus, and seed; the only free variable is instruction tuning.

Results

Decoded in PT, used only in IT. Framing is **linearly decodable in both** checkpoints, yet only IT turns it into a judgment shift.

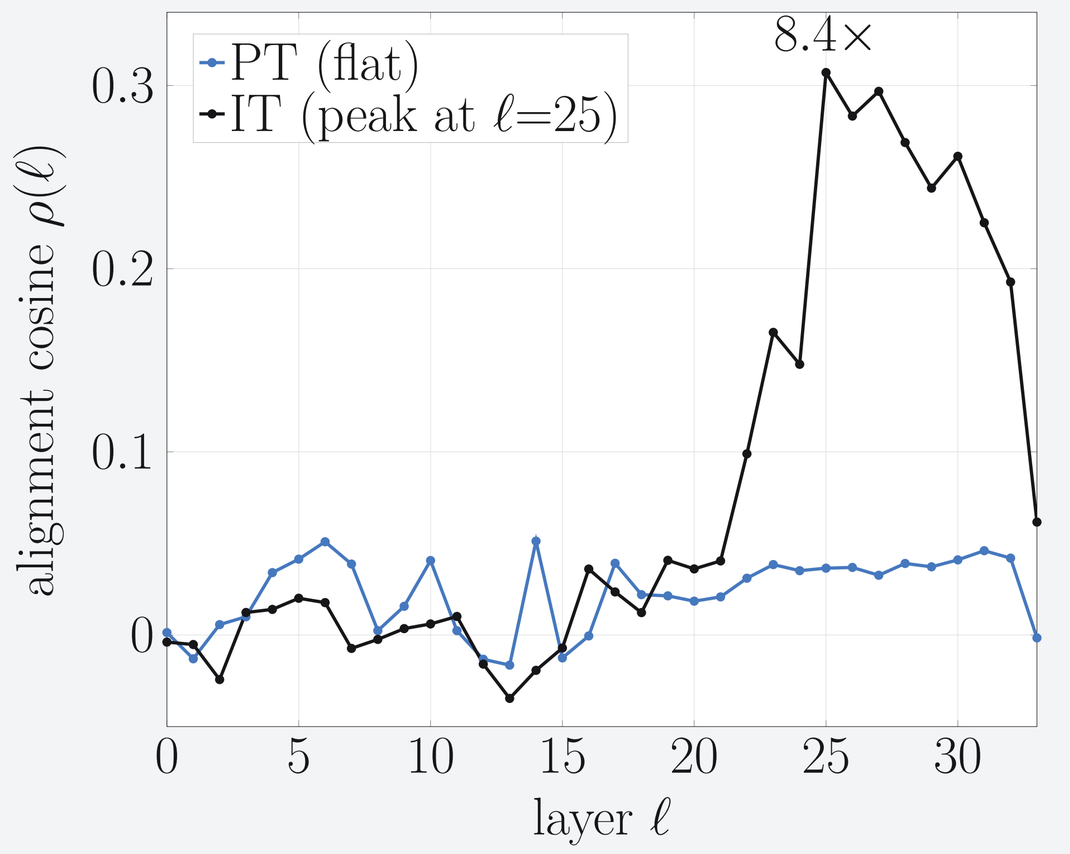
Measurement (N=100)	PT	IT
Paired $S-C$ shift	+0.04	**+12.18**
p (sign-flip)	0.52	$\mathbf{10^{-4}}$
Fraction > 0	0.51	**0.83**
Probe accuracy (best layer)	0.92	**0.97**
Alignment cos. at ℓ=25	0.04	**0.31**

The alignment cosine is **8.4**× larger in IT at ℓ=25; the PT bootstrap CI straddles zero. Framing is *decoded* in both but *used* only in IT.



In IT the framing direction aligns with *every* evaluative readout in a shared mid-to-late band, not with neutral pairs: output-generic, not specific to one judgment.

Results: a localised causal pathway



Routing strengthens in a **mid-to-late band** in IT and stays flat in PT. Convergent causal evidence localises the same pathway:

- **Patching:** mid-network patching transfers framing causally ($C\to S$ recovers 0.89).
- **Writer circuit:** 10 heads at layers 23–25 carry **59**% of the routing under mean ablation, **7.8**σ from a random-head null.
- **Rank-4 subspace:** a framing subspace orthogonal to the judgment direction; input-selective (moral, not sentiment: 14× smaller) yet output-generic.

Robust: positive and significant ($p<10^{-3}$) across all 10 IT checkpoints in Gemma-2/3, Qwen-2.5, Llama-3.1 (2B–32B); survives style/length matching and entity anonymisation.

Limitations

Evidence supports a routing claim, not a complete circuit map: the 10-head circuit recovers 59% of the shift, leaving smaller distributed contributors unlocalised. The testbed is fictional and the anonymisation control only partially addresses external validity. A trained-rotation alternative to the random rank-4 control is deferred. A LoRA developmental trajectory hit a Gemma-3 loader mismatch and is omitted.

Future work

Broader non-fictional datasets; fuller path patching downstream of the writer heads; a clean developmental run across fine-tuning checkpoints. The routing-versus-representation screen (alignment cosine in both checkpoints, look for a localised IT peak, verify causally) transfers to other behaviours: refusal vs. compliance, true vs. false, helpful vs. harmful.

Conclusion

Instruction tuning changes **how** an existing moral-framing representation is read out, **not whether** it exists.

References

[1] Prakash et al. (2024), *Fine-tuning enhances existing mechanisms* (ICLR). [2] Arditi et al. (2024), *Refusal is mediated by a single direction* (NeurIPS). [3] Belrose et al. (2023), *Eliciting latent predictions with the tuned lens.* [4] Meng et al. (2022), *Locating and editing factual associations* (NeurIPS). [5] Gemma Team (2025), *Gemma 3 technical report.*