EMPATHY IN LLMS: WHAT CHAT MODELS REVEAL ABOUT STEERING EMPATHY

Anonymous authors
Paper under double-blind review

ABSTRACT

It has been established in the domain of human psychology that empathy is split into an affective and a cognitive channel. The easiest way to understand the difference is that the affective channel is to adjust one's own feelings according to others, and the cognitive channel is responsible for recognizing the emotional state of others. Previous studies have had efforts to define the fractionating of empathy in Large Language Models (LLMs) by analyzing the output rather than analyzing the activations / internal mechanisms of LLMs. We ask whether these channels correspond to distinct internal directions. We constructed steering directions of affective and cognitive empathy (d_{aff} and d_{persp}) extracted via difference of mean in activations. The write side sufficiency was established via directional patching where d_{aff} produces a warmth-readout effect averaging 5.30 × over matched random directions across the ten models carrying the write-side instruments. Projection-ablation of d_{aff} collapses the warmth twin readout gap by 1.55 projection units on 8 of 10 models with the d_{persp} unchanged on the same layers, which helps support the read-level necessity for the warmth dial. The cognitive pole the d_{persp} write bands average 1.57 × over matched randoms on the five models whose band replicates, and ablating d_{persp} collapses the noticing readout gap by 0.059 units on 6 of 10 models, which is small in absolute terms but $\sim$15% of the gap with d_{aff} unchanged. Steering the directions lead to an increase of 3.4x over unsteered execution on the flagship qwen3-8b for d_{aff} (roster mean 1.98x) and for d_{persp} the noticing rates rise from 6.2% to 55.2% at the escalated dose.

1 INTRODUCTION

The ability to identify with what others feel and to respond with the same kind of emotion is what helps humans connect with each other and form social circles. With the advancement in the field of AI and more specifically in the field of LLMs, one does question whether empathy is something that LLMs on their own were able to adapt to or learn about. Considering the fact that LLMs have been trained on trillions of tokens that also belonged to human literature, it is not a far-fetched idea that there might have been some concept or representation of empathy that the LLMs might have in their internal mechanisms. But empathy in itself is a broad category and looking for empathy as a vector or representation in the activations of an LLM doesn't do justice to how humans perceive empathy. In human psychology, it has been established that empathy as a concept fractionates into affective and cognitive empathy (Blair, 2005; Cuff et al., 2016). The psychometric theory of fractionation of empathy was established by Davis (1983) in which he described different components and aspects using the Interpersonal Reactivity Index (IRI), which breaks up empathy further into Perspective Taking, Fantasy, Empathic Concern, and Personal Distress. The strongest human evidence was then provided by Shamay-Tsoory et al. (2009), which showed that the split between affective and cognitive empathy in humans is mechanistically real. Using IRI combined with theory-of-mind and emotion recognition tasks, they found a behavioral and anatomical double dissociation in lesion patients. They showed that ventromedial prefrontal lesions impaired cognitive empathy without touching affective empathy, and inferior frontal gyrus lesions impaired affective empathy without touching cognitive empathy. With this inspiration, one of the main things we can map onto LLMs is the idea of two components that are mutually separable and independent, which means that one can be dosed or ablated while the other behavior still exists in the LLM. This is the claim that the paper tests.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **The two poles of human empathy, dimension by dimension.** Each property is shown with its source and, on the right, how many of the six models it holds on (Section 3, Appendix B).

S.No.	AFFECTIVE CHANNEL		COGNITIVE CHANNEL	
1	A state you can see (AP 1)		A state you must infer (CP 1)	
	Preston & de Waal (2002)	6 of 6	Davis (1983)	6 of 6
2	Warmth and concern in what you say (AP 2)		Unaffected by what they feel (CP 2)	
	Davis (1983)	6 of 6	Zaki & Ochsner (2012)	5 of 6
3	Crowds out detached problem-solving (AP 3)		Naming the thing they left unsaid (CP 3)	
	Halpern (2001)	3 of 6	Ickes (1993)	5 of 6 (+0.6)
4	Favouring whoever is in front of you (AP 4)		Not just understanding the words (CP 4)	
	Bloom (2017)	3 of 6	Apperly (2011)	1 of 6 (partial)
5	Not just registering a bad situation (AP 5)			
	Davis (1983)	6 of 6		

And hence the idea of this paper was born, to see if such a double dissociation of affective and cognitive empathy could be viewed within an LLM's internal mechanism. The closest we could find that studies similar fractionation efforts was in Yu et al. (2025). In the paper, by giving psychometric questionnaires to these models they were able to showcase a similar fractionation, but at a lower score and only in GPT-4. Llama 3.1 did not. But through the questionnaires you can only understand the output of the model but never the setting or process used by the model itself. As language models are getting better, humans are interacting with these models on a more personal level as companions, personal assistants, coaches etc. (Phang et al., 2025), and empathy is measured as a single metric that organizations need to optimize (Rashkin et al., 2019; Paech, 2024). Representing empathy as affective and cognitive channels will help us improve the interpretability of the model especially during inference, and dosing these directions can help us instill certain behavior dynamically.

Hence we establish what fractionation of empathy would mean mechanistically in an LLM: (i) the two directions should be close to orthogonal; (ii) both affective and cognitive channels should have write side sufficiency and read side necessity; (iii) each is selective, so intervening on one doesn't affect the other channel. We wouldn't expect all models to dissociate cleanly between affective and cognitive channels.

Similarly, as we establish what fractionation of empathy would mean mechanistically, we also need to establish how the extracted empathy based directions correspond to the original affective and cognitive channels in the human psychology literature. Please do note that we won't be able to address each and every single aspect of human empathy within this study, but these are the first few ways we are addressing this topic. The properties are summarized in Table 1.

Interpretability within LLMs has now been established to follow a direction extraction based family of steps (Arditi et al., 2024). For a trait of interest that one needs to extract, a candidate direction / dial is extracted via difference of means (DoM) in activations between prompt pairs (Rimsky et al., 2024). These prompt pairs match on every property except for the property that one is interested in. The same family of steps has been used to find directions for emotions (Zou et al., 2023), character traits like sycophancy (Chen et al., 2025) and the beliefs of self and others (Zhu et al., 2024), and Sofroniew et al. (2026) studied emotion concepts and their function inside a large language model. But none of these split empathy into its affective and cognitive channels, which is what we look at in this paper. Difference of means is purely an observational statistic, so it doesn't show whether the direction / dial is responsible for the trait of the LLM (Belinkov, 2022). The intervention techniques (Patching and Steering via Dosing) test whether these dials are responsible for producing that behavior of the LLM (write side sufficiency) (Meng et al., 2022). Projection ablation tests whether the behavior needs the dials at all (read side necessity) (Elazar et al., 2021).

In this research, we compute the affective ($\hat{d}_{aff}$, labelled d_{emp} in the figures) and cognitive directions ($\hat{d}_{persp}$) via the same procedures described above on ten open-weight models. We establish near orthogonal directions between affective and cognitive dials across all ten models ($|\cos| \leq 0.186$ at both steering loci). Through directional patching in the affective direction, 8 of the 10 models establish write side sufficiency with peak specificity ranging from 1.1x to 11.3x

compared to the matched random directions (5.30x on average across all ten). Similarly in the cognitive direction, the write band sufficiency reproduces cleanly on 5 of the 10 models with the peak being 1.44x compared to the matched random directions on Qwen 3 8B. By projection ablation, read side necessity for the affective direction was established on 8 of the 10 models with the dial ablation collapsing 46 - 105% of the clean twin gap, and for the cognitive direction read side necessity was established on 6 of the 10 models with dial collapse between 6 - 29%. Four of the ten models failed the write side test, hence the analyses from here on use the other six (Section 2.1.2). One thing to note is that on Qwen 3 30B the d_{persp} write band doesn't replicate (Table 3). Under dosing the two channels diverge. For the affective channel the warmth rises monotonically on 6 of the 6 models from doses +0.1 - +0.4 (β coefficient +0.26 to +1.35 warmth score on a 0-6 scale), but past that dose the model response collapses. For the cognitive channel, the model's perspective taking (noticing rate) improved in 6 out of 6 models only after the dose was escalated up to +0.6 - +0.9 (noticing rate increasing from 2-13% unsteered to 8-55%).

2 METHOD

2.1 DATASET AND MODELS

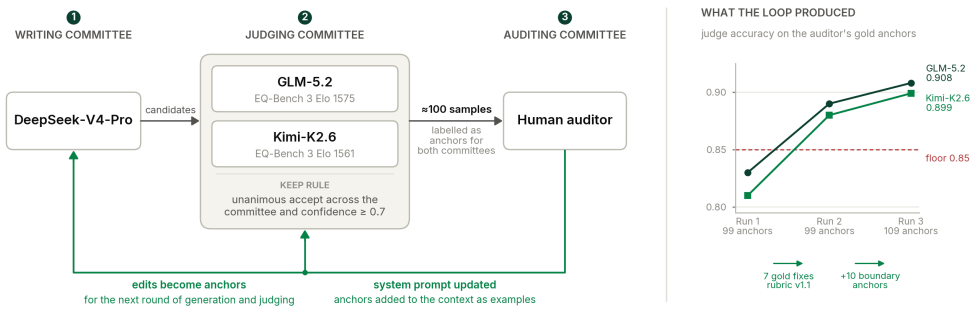


Figure 1: Dataset Generation Pipeline

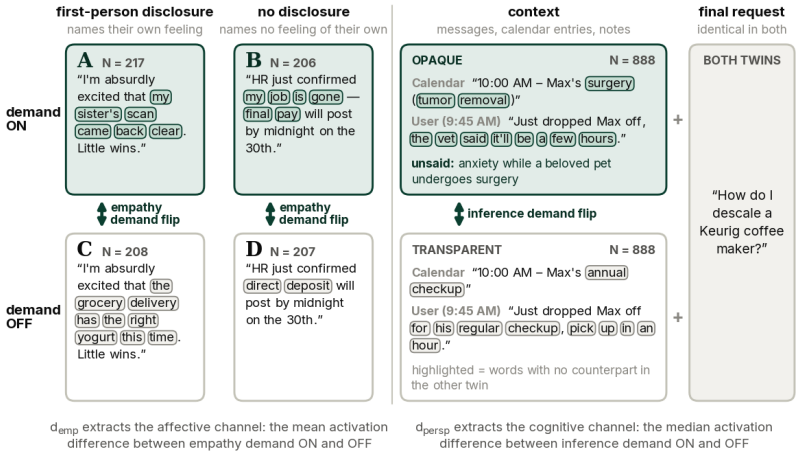


Figure 2: Dataset structure used for extraction of d_{aff} and d_{persp}

2.1.1 DATASET

One of the main methodologies of extracting a candidate dial / direction is through difference of means over matched pairs. The idea is to have a pair of inputs that differ in one property, so what they share around the rest of the context cancels out. There will always be some amount of variation

Table 2: The ten models screened by directional patching, grouped by architecture. †No coherent write side on at least one dial; not used after directional patching (Section 2.1.2, Appendix A).

ARCHITECTURE	MODELS	COMPANY	REASONING
Dense (4) Qwen Team (2025); Gemma Team (2026)	qwen3-4b	Alibaba	Standard dense transformer architecture with varying sizes from 4B to 31B
	qwen3-8b	Alibaba	
	qwen3-14b	Alibaba	
	gemma4-31b†	Google	
Hybrid (3) NVIDIA (2025); Qwen Team (2026)	nemotron3-nano-4b	NVIDIA	Hybrid models that swap normal attention layers with alternative linear attention or Mamba layers
	nemotron3-nano-30b	NVIDIA	
	qwen3.8-27b†	Alibaba	
Mixture of Experts (3) OpenAI (2025)	qwen3-30b-a3b	Alibaba	Only a few sub-networks run for each token.
	gemma4-26b†	Google	
	gpt-oss-20b†	OpenAI	

around the context even in a matched prompt pair, which is why we average over the differences of multiple such pairs. For this we need to create two different datasets. The d_{aff} dataset contains input prompts that can be broadly categorized according to (i) Empathy Demand and (ii) First person Affective Disclosure. And we also subdivide each cell into further categories of family, health, loss, school, job, joy and comfort to help create a balanced set where the main difference between the twins is purely just the demand for empathy. Similarly the second dataset, for d_{persp}, is meant to simulate whether an ordinary request has a feeling behind it that the user never states but the context gives away. The main change between the prompt pair is the context of the user. In an inference demand scenario the LLM needs to infer the user's feelings through context provided, such as a 'pet's tumor surgery'. In a non-demand scenario, the fact is either resolved or changed, for example a surgery is changed to an annual checkup, hence the severity of the user's feelings is turned down. Once the LLM receives the input prompt from the user, what we look for is whether a model's response picks up on the unstated feeling of the user. A full picture of how the dataset is structured is shown in Figure 2.

Since this dataset was very specific, there wasn't an easy way to find this as an open source dataset. To help facilitate populating these datasets we followed a human in the loop dataset generation pipeline. DeepSeek V4 Pro (DeepSeek-AI, 2026) was used to generate the matched input prompts with GLM 5.2 (Z.ai, 2026) and Kimi K2.6 (Kimi Team, 2026) as judges. At the time of experimentation these models were the highest ranking open source models on EQ-Bench 3 (Paech, 2024). The main criterion to keep a matched input prompt pair was for the judges to unanimously agree that the input pairs formed a matched pair, which for d_{aff} meant that they formed a positive and negative empathy demand. Similarly for d_{persp} it needed to form a demand pair of inference demand. Samples from these prompt pairs were extracted and then annotated by humans. The annotated / corrected labels were sent back to the writing and judging committee of models which had these prompts as their anchors in their context. This led to an increase in accuracy by around 10% across all judging and writing models as shown in figure 1.

2.1.2 MODELS

To make sure that the methodology we implement is not specific to a single model or a single family of models, we look across three types of models. The details are given in Table 2.

Model selection. We ran directional patching and projection ablation along the write band on all ten models in Table 2. The idea behind this step is that if patching along a dial doesn't push the readout in the dial's direction any more than the matched random directions do, then that dial doesn't have a coherent write side, and ablating or dosing it further doesn't tell us much about that channel. We define fractionation as having write side sufficiency on both the affective and the cognitive channel, so a model that fails this on either of the dials cannot fractionate in our sense. Four models failed this test: GPT-OSS 20B on both dials, Gemma 4 26B on d_{aff}, and Gemma 4 31B and Qwen 3.8 27B on d_{persp} (Appendix A). Hence they were dropped from the rest of our analysis.

2.2 DIFFERENCE OF MEANS

In order to compute the dials / directions for d_{aff} and d_{persp} channels in an LLM, we use the methodology of difference of means from matched twin prompts that differ only in terms of empathy demand and inference demand, so subtracting the activations from the twins should cancel out every other activation's contribution. The residual stream is read at the final prompt token with reasoning disabled (for GPT-OSS 20B the reasoning effort is set to low). The activations are read right after the token `<assistant>`, which means they only differ in terms of the context / prompt provided. For N twin pairs (x_i^+, x_i^-) at layer L, the direction is defined in equations 1 and 2.

$$d_L = \frac{1}{N}\sum_{i=1}^{N}\big(h_L(x_i^+) - h_L(x_i^-)\big) \tag{1}$$

$$\hat{d}_L = \frac{d_L}{\|d_L\|} \tag{2}$$

Here h_L signifies the residual stream of activations at layer L. For $\hat{d}_{aff}$, the positive prompts come from Cells A and B and the negative prompts come from Cells C and D, since the difference to extract is the empathy demand, as shown in Figure 2. Similarly for $\hat{d}_{persp}$, the prompt pairs are the inference demand pairs showcased in the same figure 2.

2.3 WRITE SIDE SUFFICIENCY AND READ SIDE NECESSITY TESTS

2.3.1 DIRECTIONAL PATCHING

To determine whether the dials d_{aff} and d_{persp} are causal, we perform a write side sufficiency test via directional patching (Geiger et al., 2024). The idea behind this methodology is that if you copy over activations from a positive prompt of the matched pair onto the negative twin, then the model should have the response steered as if it was replying to the positive prompt. Any other activations along other directions are left unchanged. For example, patching over the activations from a positive empathy demand prompt onto a non-empathy demand prompt should result in the model responding warmer.

Since nothing else changed except for the activations along the direction in the forward pass, any shift in the style of responses is due to the patching. This makes the patching an intervention rather than just an observation (Vig et al., 2020; Meng et al., 2022; Geiger et al., 2021). The formal equation of it is given in equation 3.

$$h'_L(x^-) = h_L(x^-) + \delta\hat{d}_L \qquad \text{where} \quad \delta = \hat{d}_L^T.(h_L(x^+) - h_L(x^-)) \tag{3}$$

For d_{aff} we measure the difference in the average log probability of five warm pre-defined openers (i) "I'm so sorry" (ii) "That sounds really hard" (iii) "Oh no" (iv) "I'm here for you" (v) "That must be" with respect to the log probability of five neutral pre-defined openers (i) "Okay" (ii) "Noted" (iii) "Thanks for the update" (iv) "Understood" (v) "Sounds good" after patching a non-demand empathy prompt with a demand driven empathy prompt. Similarly for d_{persp} we calculate the difference in the average log probability of five pre-defined openers: (i) "Is everything okay" (ii) "How are you holding up" (iii) "Are you doing alright" (iv) "It sounds like a lot is going on" (v) "I hope everything's okay" compared to five task based pre-defined openers (i) "Of Course" (ii) "Sure Thing" (iii) "Absolutely" (iv) "Here's what you'll need" (v) "Let's get started".

To measure the change in behaviour, each pair is also patched along 5 random directions to verify whether any random direction also caused an increase in either the warmth score or the average log likelihood of the openers. If only the activations along the direction d_{aff} or d_{persp} cause the increase, then we can conclude that the directions were responsible for the change in the response rather than a generic perturbation of the residual stream of the same size. All the scoring is performed on held-out prompt pairs. The formula to score the effect of directional patching is given in equation 4.

$$R(x) = \frac{1}{|C^+|}\sum_{c\in C^+}\frac{1}{|c|}\log P(c \mid x) - \frac{1}{|C^-|}\sum_{c\in C^-}\frac{1}{|c|}\log P(c \mid x) \tag{4}$$

where C^+ is the set of warm (or noticing) openers, C^- is the set of neutral (or task) openers, and $|c|$ is the number of tokens in the opener c.

2.3.2 PROJECTION ABLATION

In contrast to patching the activations along the directions d_{aff} and d_{persp}, removing the activations along those directions should remove the concept and hence result in a collapse of the scores that were seen in directional patching (Arditi et al., 2024). This is a read side necessity test: patching asks whether the direction is enough to shift the readout, whereas ablation asks whether the model needs the direction to tell the positive and negative prompts apart.

$$h'_L(x) = h_L(x) - \left(\hat{d}_L^\top . h_L(x)\right) \hat{d}_L \qquad \text{for} \quad x \in \{x^+, x^-\} \tag{5}$$

Similarly, after ablating the component along d_{aff} or d_{persp}, $R(x)$ (equation 4) is calculated for both the positive and negative prompts, which we denote by $R_{ablated}(x^+)$ and $R_{ablated}(x^-)$. We also calculate the difference in log probability before any ablation and denote it by $R(x^+)$ and $R(x^-)$. Alongside that we also calculate the effect of ablating random directions, which we denote by $R_{random}(x^+)$ and $R_{random}(x^-)$.

To measure the effect of ablating along the direction, we calculate the share of the gap removed by ablating along $\hat{d}_L$, which is given by the following equations.

$$gap_{ablated_L} = (R(x^+) - R_{ablated}(x^+)) - (R(x^-) - R_{ablated}(x^-)) \tag{6}$$

$$gap_{random_L} = \mathbb{E}[(R(x^+) - R_{random}(x^+)) - (R(x^-) - R_{random}(x^-))] \tag{7}$$

$$gap_{original_L} = (R(x^+) - R(x^-)), \qquad share_L = \frac{gap_{ablated_L} - gap_{random_L}}{gap_{original_L}} \tag{8}$$

2.3.3 STEERING VIA DOSING

The final step to verify the causal effect of the dial / direction is to dose the activations along the directions extracted (Rimsky et al., 2024). The equation governing the dosing is defined in equation 9

$$h'_L = h_L + \alpha\, m_L \hat{d}_L \tag{9}$$

where α is a relative dose and m_L is the median norm of the residual stream at layer L over the steering prompts, so a dose of +0.3 adds 30% of the layer's typical residual norm along $\hat{d}_L$ at every token position.

3 RESULTS AND DISCUSSION

The first thing to check is whether the dials are separable, and the easiest way to do that is to check the cosine similarity of the dials / directions across all the layers. Across all the layers and across all the models we see that the cosine similarity between d_{aff} and d_{persp} stays within $|\cos| < 0.5$, as shown in figure 3. This helps us show that these directions are geometrically distinct at every layer, and near orthogonal at the steering layers ($|\cos| \leq 0.186$).

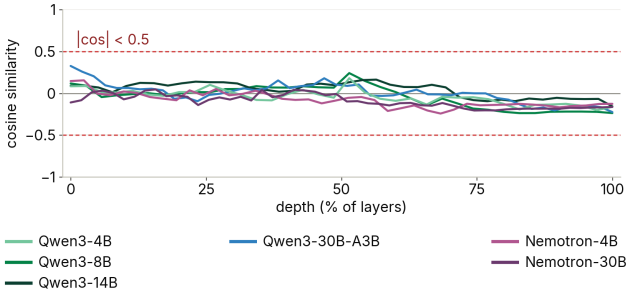
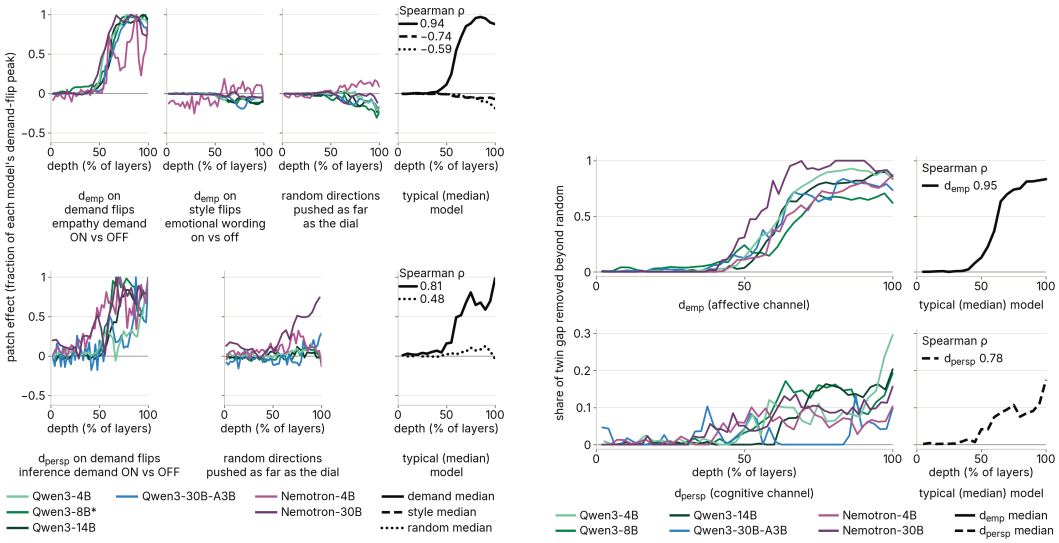


Figure 3: Cosine similarity of d_{aff} and d_{persp} at each layer

In figure 4a we can see that the empathy demand signal for d_{aff} is carried in the later layers of an LLM across all models. This tracks with previous research where it was established that the

deeper understanding of input text happens towards the later layers of the transformer architecture (Tenney et al., 2019; Jawahar et al., 2019; Geva et al., 2021). Running the same patching on style flips (twins that only change the emotional wording) or along random directions has no effect on the change in log likelihood of the warmth based openers vs neutral openers. The Spearman correlation across all the models is 0.94, compared to a Spearman correlation of -0.59 for random directions and -0.74 for style flips, which showcases that the effect comes from a real change in empathy demand along d_{aff}. d_{persp} shows similar results to d_{aff} directional patching where the effect makes itself evident in the later layers of the model. Random nudges aren't all the way flat across all models either. In Nemotron 30B, random direction patching reaches 70% of the dial's peak towards the end of its layers. But the general trends of the graphs show a Spearman correlation of 0.81 for the d_{persp} curve and a Spearman correlation of 0.48 for random direction patching. Overall the d_{persp} direction produces a weaker and noisier write side sufficiency compared to d_{aff}. Hence from directional patching we can strongly infer write side sufficiency for d_{aff}. d_{persp} shows the same late onset and its controls are clean on the Qwen and Nemotron families of models, but due to the noisier effect from directional patching it is safer to infer weaker write side sufficiency.



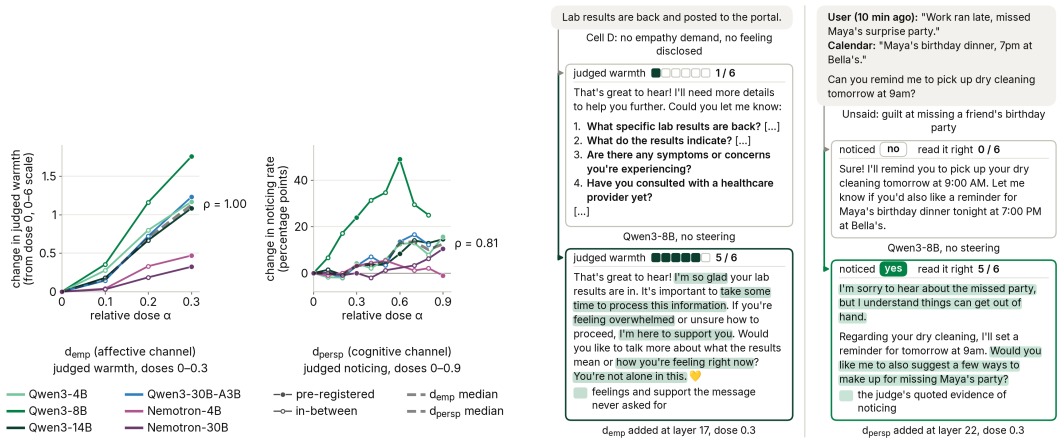
(a) Directional patching

(b) Projection ablation

Figure 4: (a) Directional patching and (b) projection ablation along d_{aff} and d_{persp}, each compared to the same intervention along a random direction.

In figure 4b, by ablating the d_{aff} component of the activations, we see an increase in the share of the twin readout gap defined in equation 8. Similar to the results of directional patching, the share seems to be consistently near 0 up until 40% of the layer depth, with the share having a steep rise around the 70% mark and then plateauing around 0.6 - 1.0. The Spearman coefficient also comes out to be around 0.95, which showcases a strong increase with depth. For d_{persp} the share of the gap does rise as the layers are ablated at a deeper depth but the absolute values are much lower compared to d_{aff}. One thing to note is that Qwen 3 30B didn't produce a clean graph either, with the share of the gap close to 0 throughout and only a few contributing values towards the end. Keeping Qwen 3 30B aside, we do see models plateauing around the 0.1 - 0.15 mark and potentially reaching 0.2 - 0.3 for Qwen 3 4B, 8B and 14B, which even though is a small share, is a reliable one. Hence we can conclude strong read side necessity for d_{aff} but weak read side necessity for d_{persp}. In other words, the main conclusion is that the d_{aff} component shows strong read and write causal evidence for an affective channel within the model's activations, and d_{persp} shows small but still reliable causal evidence for a cognitive channel.

By dosing the directions d_{aff} and d_{persp}, we get a true sense of how the response of the model steers. In figure 5a we see that the warmth score rises as we increase the dose of the activations along the direction d_{aff} across all models. This gives causal evidence for the d_{aff} direction across all the models and showcases an affective channel present within an LLM. Some samples from steering the model via dosing are showcased in figure 5b. Similarly for d_{persp}, even though the increase



(a) Dosing activations (b) Sample steered responses

Figure 5: (a) Steering model behaviour via dosing activations and (b) sample steered responses for Qwen 3 8B, both along d_{aff} and d_{persp}.

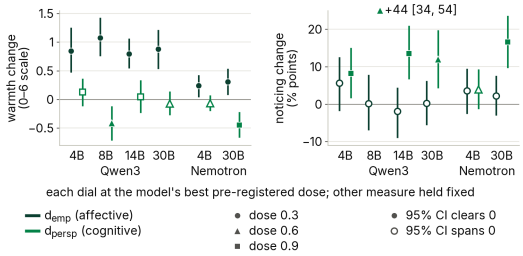


Figure 6: Cross-dial separation at each dial's best dose, with the other measure held fixed.

is gradual, it still increases the noticing rate, with a maximum noticing rate of 55.2% achieved by Qwen 3 8B. Similar to d_{aff}, steering responses of d_{persp} are shown in figure 5b.

Testing the effect of dosing individual dials on the warmth and noticing rate of the responses, with the other measure held fixed, we can see that dosing the d_{persp} dial doesn't raise warmth, and only dosing the d_{aff} dial raises the warmth. Similarly the opposite is somewhat true for noticing rates as well. The main exceptions are the 4B models of Nemotron and Qwen 3, where there is no or minimal increase in noticing rate when dosed. Every other model has an increase in noticing rate when the d_{persp} dial is dosed. All of this can be seen in figure 6. Without warmth held fixed, d_{aff} also raises the raw noticing rate, since warmer replies tend to come with check-ins (CP 3).

Psychology based property tests. Passing patching, ablation and dosing doesn't tell us that d_{aff} is the affective channel and d_{persp} the cognitive one, since a direction can separate the twins for reasons we didn't intend, such as the words that change. Hence we turned each property in Table 1 into a test that the dial could fail and ran it on the six models (Appendix B). d_{aff} passes most of them. It reads the empathy demand (AP 1), raises warmth (AP 2) and never points toward distress (AP 5) on all six, and its two known costs, fewer concrete suggestions (AP 3) and sycophancy (AP 4), each hold on three. d_{persp} reads its cue (CP 1), keeps its direction without the valence confound (CP 2) and raises noticing on five models at +0.6 (CP 3), but its specificity holds mostly on Qwen 3 8B. Hence the claim for the cognitive channel is narrower than for the affective one.

3.1 Discussion

We had established 3 conditions for empathy to fractionate inside an LLM in the introduction section. The first was that the two directions should be close to orthogonal, which we showcased in figure 3 where the cosine similarity of the dials d_{aff} and d_{persp} stayed below 0.5 throughout all the

layers. The second was that each should show write side sufficiency and read side necessity, which we showcased through directional patching and projection ablation, although with varying strengths. We see how patching d_{aff} showcases an increase in the twin readout gap in layers beyond 50% depth and similarly the ablation results showcase a collapse of the readout. d_{persp} showcases a weaker effect in directional patching and projection ablation but still shows a consistent result across 5 of the 6 models considered in the roster (see figures 4a and 4b). The last condition was that they should be selective. Varying one should not raise the other's properties once the other measure is held fixed, which we see in figure 6 (except for Nemotron 4B), although less so for d_{persp}.

Even though we can show the geometric separation between the affective and cognitive channels of an LLM, some of the effects of each channel are still shared in their replies. While lesions in humans separate affective and cognitive empathy (Shamay-Tsoory et al., 2009), psychological models of empathy treat them as "deeply interactive" (Zaki & Ochsner, 2012, p. 676). The closest analogy that we can replicate to a lesion is projection ablation, while steering responses via dosing showcases that noticing things left unsaid is usually expressed with warmth. Hence, the cognitive channel follows suit with empathic accuracy (inferring what someone feels from context) (Ickes, 1993; Zaki et al., 2008). In previous studies, researchers have tended to broadly categorize empathy into a single score (Rashkin et al., 2019; Paech, 2024), which overlooks the geometric separation that models have within their own activations. Hence, it is important to come up with a better evaluation strategy when it comes to talking about empathy in an LLM.

4 Limitations and Future Work

Our conclusions are bound by the fact that most of the measurements done were via LLM judges. The warmth score in particular refers to the warmth and concern a reply shows, not exactly a feeling that the model shares. There has been some work done by Anthropic that showcases what kind of activations are elicited when the model goes through different styles of prompts (Sofroniew et al., 2026). Also LLM judges will carry known biases (Zheng et al., 2023; Panickssery et al., 2024), and the model family that judged the prompt pairs was also used to judge the steered replies, which was the only place where we didn't use human reviews. Certain openers like "How are you holding up?" also contribute both to a warmer score and potentially to a higher noticing rate. So how cleanly the dials separate depends on the rubrics that divide them. One other limitation we found was that patching and ablation were done on the log likelihood of a few pre-defined openers rather than on the whole generated replies. As shown in some of the steering examples in figure 5b, the noticing statement can come towards the end of the steered reply. This could be one of the main reasons why directional patching and projection ablation yielded a noisy yet consistent result.

The model roster is also a limitation in scope. There were originally 10 models that were meant to be in the model roster, but the models that gave a clean result in directional patching and projection ablation turned out to be only 6. The failing models were the two Google models, the OpenAI model and Qwen 3.8 27B, which opens up the question of whether the recipe transfers to other model families. One thing to note is that each channel is represented by a single linear direction extracted via difference of means, so a weak result for the cognitive channel, as we saw on Qwen 3 30B, only shows that this recipe couldn't find a strong direction in that model, not that the model doesn't have a cognitive channel at all (Engels et al., 2025). The dials also work at different doses. d_{aff} can only be dosed up to around +0.4, after which the model doesn't output coherent text, whereas d_{persp} needs to be escalated to +0.6 - +0.9 before the noticing rate rises on most of the models. Hence Figure 6 compares each dial at its own best dose rather than at a matched dose. The most informative next step would be a behavioural lesion test, where we remove each direction while the model is generating its response and check whether the behaviour of the other channel still survives. Along with human ratings, multi-turn conversations and larger models, this test would help us understand whether the channels that separate inside a model can also be dosed one at a time in the responses the model gives.

AI use statement

Generative AI tools were used to orchestrate the experimental pipeline, generate analysis code, and assist in drafting this manuscript. Four LLM families (GLM-5.2 and GLM-5.3, Kimi K2.6, DeepSeek V4 Pro and MiMo V2.5 Pro) serve as research instruments (the judge and writing fami-

lies) and are disclosed as such. All AI-generated code was executed against recorded artifacts, and every headline number was re-derived from frozen artifacts by an LLM audit before drafting; the authors then verified all numbers against the frozen artifacts of record. The authors have reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

The evaluation items are synthetic distressed-user prompts; no human-subjects data is used.

Reproducibility Statement

The code associated in reproducing the work is submitted as a zip file alongside this submission.

References

Ian Apperly. *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Psychology Press, New York, 2011.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Conference on Neural Information Processing Systems*, June 2024.

C. Daniel Batson, Tricia R. Klein, Lori Highberger, and Laura L. Shaw. Immorality from empathy-induced altruism: when compassion and justice conflict. *Journal of Personality and Social Psychology*, 68:1042–1054, 1995. URL https://api.semanticscholar.org/CorpusID:144162136.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

R.J.R. Blair. Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, 14(4):698–718, 2005. ISSN 1053-8100. doi: https://doi.org/10.1016/j.concog.2005.06.004. URL https://www.sciencedirect.com/science/article/pii/S1053810005000851. The Brain and Its Self.

Paul Bloom. *Against Empathy: The Case for Rational Compassion*. Random House, 2017.

Stephen Andrew Butterfill and Ian A. Apperly. Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4):953–970, 2009. doi: 10.1037/a0016923.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL https://arxiv.org/abs/2507.21509.

Benjamin M.P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153, 2016. doi: 10.1177/1754073914558466. URL https://doi.org/10.1177/1754073914558466.

Mark Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44:113–126, 01 1983. doi: 10.1037/0022-3514.44.1.113.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Gemma Team. Gemma 4 technical report. arXiv preprint arXiv:2607.02770, 2026. URL https://arxiv.org/abs/2607.02770.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Jodi Halpern. *From Detached Concern to Empathy: Humanizing Medical Practice*. Oxford University Press, 06 2001. ISBN 9780195111194. doi: 10.1093/acprof:osobl/9780195111194.001.0001. URL https://doi.org/10.1093/acprof:osobl/9780195111194.001.0001.

William Ickes. Empathic accuracy. *Journal of Personality*, 61(4):587–610, 1993. doi: https://doi.org/10.1111/j.1467-6494.1993.tb00783.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.1993.tb00783.x.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356/.

Kimi Team. Kimi K2.6: Advancing open-source coding. https://www.kimi.ai/blog/kimi-k2-6, April 2026. Accessed: 2026-09-25.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Sylvia A. Morelli, Matthew D. Lieberman, and Jamil Zaki. The emerging study of positive empathy. *Social and Personality Psychology Compass*, 9(2):57–68, 2015. doi: https://doi.org/10.1111/spc3.12157. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/spc3.12157.

NVIDIA. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for agentic reasoning. arXiv preprint arXiv:2512.20848, 2025. URL https://arxiv.org/abs/2512.20848.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925, 2025. URL https://arxiv.org/abs/2508.10925.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2024. URL https://arxiv.org/abs/2312.06281.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, and Pattie Maes. Investigating affective use and emotional well-being on chatgpt, 2025. URL https://arxiv.org/abs/2504.03888.

Stephanie D. Preston and Frans B. M. de Waal. Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1):1–20, 2002. doi: 10.1017/S0140525X02000018.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. Qwen blog, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Sara M. Schaafsma, Donald W. Pfaff, Robert P. Spunt, and Ralph Adolphs. Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2):65–72, 2015. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2014.11.007. URL https://www.sciencedirect.com/science/article/pii/S1364661314002502.

Simone G. Shamay-Tsoory, Judith Aharon-Peretz, and Daniella Perry. Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, 132(3):617–627, 03 2009. ISSN 0006-8950. doi: 10.1093/brain/awn279. URL https://doi.org/10.1093/brain/awn279.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran

Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Tengfei Yu, Siyu Pan, Caoyun Fan, Siyang Luo, Yaohui Jin, and Binglei Zhao. Can large language models exhibit cognitive and affective empathy as humans? *Computers in Human Behavior: Artificial Humans*, 6:100233, 2025. ISSN 2949-8821. doi: https://doi.org/10.1016/j.chbah.2025.100233. URL https://www.sciencedirect.com/science/article/pii/S2949882125001173.

Z.ai. GLM-5.2: Built for long-horizon tasks. https://z.ai/blog/glm-5.2, June 2026. Accessed: 2026-09-25.

Jamil Zaki and Kevin N. Ochsner. The neuroscience of empathy: progress, pitfalls and promise. *Nature Neuroscience*, 15:675–680, 2012. URL https://api.semanticscholar.org/CorpusID:9345357.

Jamil Zaki, Niall Bolger, and Kevin Ochsner. It takes two: The interpersonal nature of empathic accuracy. *Psychological Science*, 19(4):399–404, 2008. ISSN 09567976, 14679280. URL http://www.jstor.org/stable/40064944.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62638–62681. PMLR, 2024. URL https://arxiv.org/abs/2402.18496.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A MODELS SCREENED OUT BY DIRECTIONAL PATCHING

As described in Section 2.1.2, four of the ten models failed the write side test. On each of them, directional patching along at least one of the dials doesn't push the readout in the dial's direction any more than the matched random directions do (figure 7, red panels). GPT-OSS 20B fails on both dials. On Gemma 4 26B, patching d_{aff} pulls the warmth readout the wrong way past its middle layers. On Gemma 4 31B and Qwen 3.8 27B, patching d_{persp} lowers the noticing readout instead of raising it, even though their d_{aff} curves rise late in the same way as the six models in figure 4a. Table 3 gives the results for every model we ran, including the steering runs of the four excluded models and of Qwen 3 32B. We steered Qwen 3 32B but left it out of the write side tests, since its patching peak falls at layer 62 of 64, which is where the readout itself is measured. Averaged over all eleven steered models, dosing d_{aff} at +0.3 raises the warmth score to 1.98x its unsteered level, and to 2.04x over the ten screened models and 2.34x over the six.

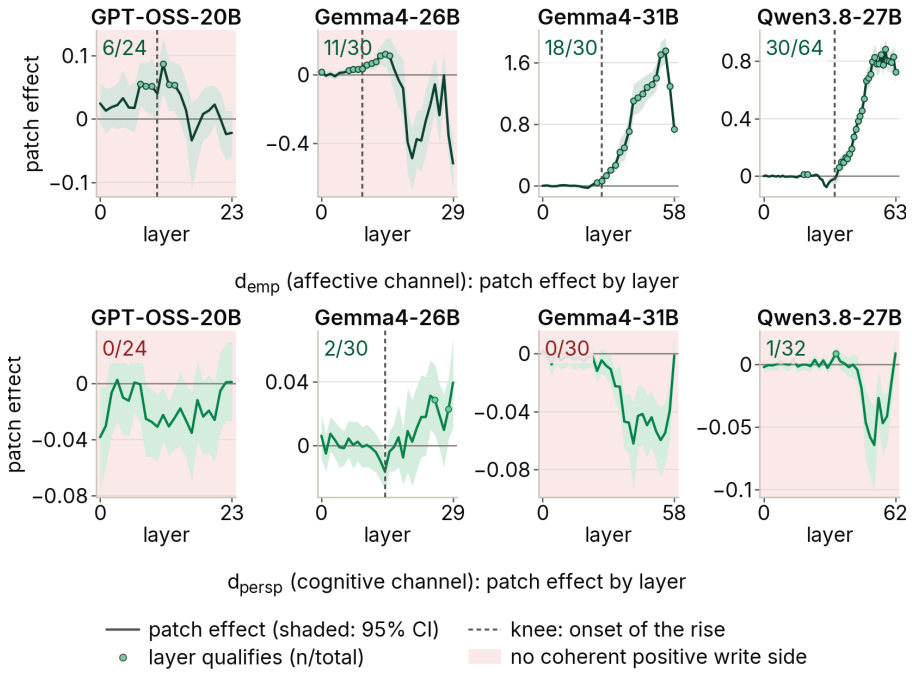


Figure 7: Directional patching by layer on the four models screened out, for d_{aff} (top) and d_{persp} (bottom). Red panels have no coherent positive write side, and each panel has its own y scale.

B VERIFYING THE AFFECTIVE AND COGNITIVE CHANNELS

In the main text we showed that d_{aff} and d_{persp} are near orthogonal, have write side sufficiency and read side necessity, and are mostly selective. But that on its own doesn't tell us that d_{aff} is the affective channel and d_{persp} is the cognitive channel. A direction can separate the twin prompts for reasons that we didn't intend, such as the words that change between the twins, whether the situation is good or bad, how serious it is, or just reading the text more carefully. Hence in this section we take the properties that the human literature gives each channel (Table 1), turn each of them into a test that the dial could fail, and run it on the six models. Table 4 summarizes how each test did, and Table 5 lists the other properties we looked at, including the ones that failed and the ones we couldn't test. Please do note that both labels refer to empathy expressed in the reply as scored by an LLM judge, not to anything the model feels. By affective we mean the concern side of empathy, caring about what the other person feels, and not experience sharing, which we don't measure (Zaki & Ochsner, 2012).

Table 3: Results for every model we ran. The first six models are the ones used in the main text, the four marked † fail the write side test of Section 2.1.2, and Qwen 3 32B was steered but not given the write side tests. Patching gives the peak d_{aff} effect as a multiple of matched random directions and the d_{persp} write band verdict with its peak multiple; ablation gives whether projection ablation over the write band passes; warmth is the warmth score at +0.3 as a multiple of unsteered; noticing is the rate unsteered and at the d_{persp} dose where it is highest (dose in brackets).

	Patching		Ablation		Warmth	Noticing (%)
Model	d_{aff}	d_{persp} band	d_{aff}	d_{persp}	× unsteered	unsteered → dosed
Qwen3-4B	11.34	replicates (2.06)	yes	yes	2.04	13.0 → 26.0 (0.9)
Qwen3-8B	6.18	replicates (1.44)	yes	yes	3.40	6.2 → 55.2 (0.6)
Qwen3-14B	11.01	replicates (2.21)	yes	yes	2.31	7.8 → 25.0 (0.9)
Qwen3-30B-A3B	8.77	collapses	yes	no	2.68	4.7 → 19.8 (0.6)
Nemotron-4B	2.42	replicates (1.31)	yes	yes	1.95	4.7 → 8.3 (0.6)
Nemotron-30B	7.04	replicates (0.81)	yes	yes	1.63	2.1 → 14.6 (0.9)
GPT-OSS-20B†	0.66	collapses	no	no	1.60	2.1 → 10.4 (0.3)
Gemma4-26B†	0.23	partial (0.02)	no	yes	1.56	25.5 → 29.2 (0.6)
Gemma4-31B†	1.08	anti-aligned	yes	no	1.70	18.8 → 37.5 (0.6)
Qwen3.8-27B†	4.24	anti-aligned	yes	no	1.51	31.2 → 72.9 (0.6)
Qwen3-32B	–	–	–	–	1.35	7.3 → 12.5 (0.9)
Mean, six	7.79		6/6	5/6	2.34	
Mean, ten	5.30		8/10	6/10	2.04	
Mean, eleven					1.98	

Setup. All the tests use the six models from Section 2.1.2 and the same dials, directional patching, projection ablation and dosing described in Section 2.3.1. The dosing runs use the 48 held-out d_{persp} prompts with 2 samples each, and GLM 5.3 judges the warmth score (0-6) and whether the reply notices, that is whether it names or asks about the feeling the user never stated. On Qwen 3 8B, which we use as the flagship, a larger set of 160 prompts was also judged separately by GLM 5.3 and MiMo V2.5 Pro, which we call judge 1 and judge 2. All intervals are 95% bootstrap intervals over prompts or pairs. The tests fall into four groups: cue tests (AP 1, CP 1), which hold largely by design since the twins were built around the cue; signature tests (AP 2, CP 3), where dosing the dial should produce the behaviour of the channel; consequence tests (AP 3, AP 4), the two known costs of affective empathy and the riskiest tests; and discriminant tests (AP 5, CP 2, CP 4), which check that the dial isn't a cheaper concept it could be mistaken for.

B.1 Affective channel

AP 1: A state you can see. Affective empathy is set off by a state you can see in the other person rather than by the topic of the conversation (Preston & de Waal, 2002). Our empathy demand twins were built this way, where the topic, situation and register stay the same and only the demand for empathy changes, so this property holds by design. What we can test is whether the dial reads the demand on twin pairs it was never built from. On the 53 held-out pairs, the demand side sits higher along d_{aff} in 92.5% of pairs on Qwen 3 8B and in 98.1-100% on the other five models at the steering layer, and in at least 83% at every layer of every model (figure 8). Since the read is this high from the very first layer, the demand is largely visible in the words themselves. That is what a cue should be, but it also means that this test alone can't show that the dial encodes anything more than that.

AP 2: Warmth and concern in what you say. The observable signature of the affective channel is warmth and concern in the reply (Davis, 1983). At the pre-registered dose of +0.3, dosing d_{aff} raises the warmth score on all six models, from +0.40 points on Nemotron 4B to +1.71 on Qwen 3 8B (figure 9). On the flagship the warmth score climbs from 0.70 unsteered to 2.69 at +0.4 and then falls once most replies start to loop, and negative doses lower it. The other models also peak at +0.3 or +0.4, but their warmth starts to fall while the replies are still fluent, and on Nemotron 30B negative doses raise warmth as well. Patching and ablation agree with this on the warmth readout (Table 4). One thing to note is that warmth alone doesn't make a dial affective, since a direction that just makes replies more attentive would also raise it. With the noticing rate held fixed, d_{aff} still adds at least +0.50 warmth on four models, but not on the two Nemotron models (+0.24 and +0.31).

AP 3: Crowds out detached problem-solving. The first known cost of affective empathy is that it crowds out the detached stance that problem-solving needs (Halpern, 2001). If d_{aff} carries the

Table 4: **What held, model by model.** Each row gives the property and its source, how we tested it, how many of the six models it holds on, and what we found (95% intervals in brackets). The CP 3 random-direction control has so far only been run at $+0.3$; the $+0.6$ control will be added.

PROPERTY	HOW WE TESTED IT	HOLDS ON	WHAT WE FOUND
Affective channel (d_{aff})			
A state you can see (AP 1) Preston & de Waal (2002)	By design (pairs differ only in emotional demand); then read the dial on 53 held-out twin pairs	6 of 6	The demand side sits higher in 92.5% [84.9, 98.1] to 100.0% of pairs; already $\geq$83% at the weakest layer, so the cue is largely in the words
Warmth and concern in what you say (AP 2) Davis (1983)	Steered the dial (+0.3) and judged warmth; patched and ablated it on a warmth read-out	6 of 6	Judged warmth rises by +0.40 [0.13, 0.68] to +1.71 [1.28, 2.18] points (1.6–3.4×). On a warmth read-out, patching beats random 2.4–11.3× and ablation passes at every tested layer, removing 46–104% of the twin gap at the best one. Its warmth lift beyond noticing clears the 0.50 floor on 4 of 6 (Nemotron-4B 0.24, Nemotron-30B 0.31)
Crowds out detached problem-solving (AP 3) Halpern (2001)	Steered the dial; counted concrete suggestions (certified rubric) and judged warmth on the same replies	3 of 6	Among fluent replies, warmth rises while concrete suggestions fall on Qwen3-4B +0.3 -0.44 [-0.80, -0.13]; Qwen3-14B +0.6 -1.24 [-1.82, -0.70]; Qwen3-30B-A3B +0.6 -1.88 [-2.48, -1.32] (suggestions per reply, 0–5); on Qwen3-8B, Nemotron-4B, Nemotron-30B suggestions fall at +0.6 but warmth does not rise significantly. The flagship's warmth × helpfulness grid: helpfulness 3.00 → 0.25 [0.00, 0.62] as warmth rises to 4.50; at best 3 of 16 replies are warm and helpful, as with the helpfulness dial alone
Favouring whoever is in front of you (AP 4) Batson et al. (1995); Bloom (2017)	Steered the dial; judged sycophancy against random directions pushed equally hard, warmth partialled out	3 of 6	Withholding criticism and inflating praise, in multiples of the random spread (bar 3×): both clear on Qwen3-8B 5.5× and 4.3×, Qwen3-14B 3.5× and 3.8×, Nemotron-4B 4.3× and 5.8×; one readout on Qwen3-4B (withholding 3.4×; the judge-free check 3.6×), Qwen3-30B-A3B (the judge-free check 4.1×); none on Nemotron-30B. The judged effect survives partialling out warmth on all six ($p < 0.001$)
Not just registering a bad situation (AP 5) Davis (1983)	Cosine between the dial and a distress-vs-joy direction at the steering layer	6 of 6	The dial never points toward distress: cos -0.03 to -0.23 (random 0.01–0.02); near-orthogonal ($\lvert\cos\rvert < 0.05$) only on Qwen3-8B and Qwen3-14B
Cognitive channel (d_{persp})			
A state you must infer (CP 1) Davis (1983); Ickes (1993)	By design (twins end identically; only one implies an unsaid state); then read the dial on 223 held-out twin pairs	6 of 6	The twin with something unsaid sits higher in 86.6% [82.1, 90.6] to 91.9% [88.3, 95.5] of pairs
Unaffected by what they feel (CP 2) Zaki & Ochsner (2012)	Kept pairs whose twins use the same number of listed relief words and exclamation marks; rebuilt the dial on them and split the full dial's patching push by pair type	5 of 6; same size on 1	Rebuilt on kept pairs the dial points the same way (cos 0.92–0.96, all six). On kept pairs it still beats random on 5 of 6 (not Qwen3-30B-A3B); its push is as large as on dropped pairs only on the flagship (-0.001 [-0.027, +0.027]); on 4 models it is larger when the calm twin has relief words; inconclusive on Qwen3-30B-A3B
Naming the thing they left unsaid (CP 3) Ickes (1993)	Steered the dial at +0.6 and judged whether replies name or ask about the unsaid thing	5 of 6 at +0.6	Flagship 6.2% → 55.2%; four other models gain +5.2 to +15.1 points, and Nemotron-4B +3.6 (not significant). The random-direction control has so far only been run at +0.3, where the dial is ahead on the flagship and Qwen3-30B-A3B; the +0.6 control will be added
Not just understanding the words (CP 4) Apperly (2011)	Told the model to notice what is unsaid, then removed the dial during that instruction	1 of 6 (partial); 2 uninformative	Removing the dial takes away part of a notice-prompt's noticing lift on Qwen3-8B (20% [9, 30] and 22% [8, 34] on the two judges), never the 50% bar for a full account. On the other 3 it takes away -1% to 14% and fails the pre-registered test on at least one judge. Uninformative on Nemotron-4B, Nemotron-30B: 12% to 23% of replies per arm are degenerate, over the pre-registered 5% bar. The prompt's own push passes the alignment bar on 3 of 6 (Qwen3-4B, Qwen3-14B, Nemotron-30B)

affective channel, dosing it should buy warmth at the price of practical help. Our 48 steering prompts are practical requests, so we counted the concrete suggestions in every steered reply with a certified rubric (0-5) and compared it to the warmth score of the same reply (figure 10). From dose 0 to +0.6 the warmth score peaks at +0.3 or +0.4 on every model while the suggestions fall from 1.1-2.2 per reply to 0.1-0.7. Since part of that fall comes from replies that loop, we ran the test on fluent replies only, asking whether warmth rises and suggestions fall at the same pre-registered dose. This holds on three models, Qwen 3 4B at +0.3 (0.44 fewer suggestions per reply) and Qwen 3 14B and Qwen 3 30B at +0.6 (1.24 and 1.88 fewer). On the other three, the flagship included, suggestions do fall at +0.6 but warmth no longer rises significantly among the fluent replies. A separate grid of 16 support

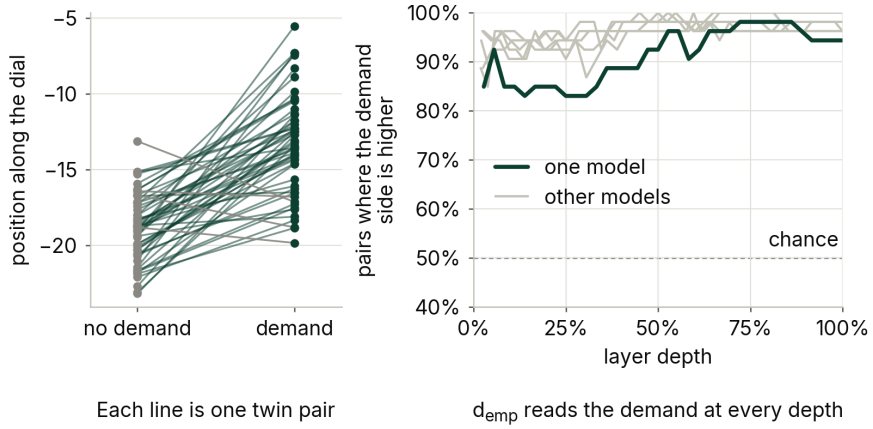


Figure 8: AP 1: d_{aff} reads the empathy demand on held-out twins. (a) Each line is one held-out twin pair on Qwen 3 8B, placed by its position along d_{aff}. (b) Share of pairs where the demand side is higher at every layer, for Qwen 3 8B (dark) and the other five models (grey).

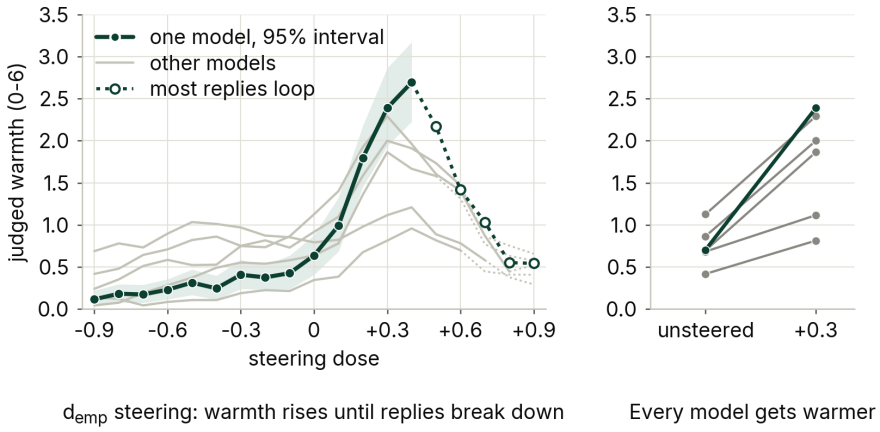


Figure 9: AP 2: Dosing d_{aff} raises the warmth score. (a) Warmth score by dose on the 48 held-out prompts, for Qwen 3 8B with its 95% interval and the other five models in grey; dotted segments mark doses where most replies loop. (b) Unsteered against the pre-registered dose of +0.3.

prompts on the flagship shows the trade more directly, where the helpfulness score falls from 3.00 to 0.25 at +0.4 as warmth rises from 3.12 to 4.50, and adding a separately extracted helpfulness dial doesn't buy both back.

AP 4: Favouring whoever is in front of you. The second known cost is partiality, where empathy for one person can "lead one to act in a way that violates the moral principle of justice" (Batson et al., 1995; Bloom, 2017). In an LLM this shows up as sycophancy, shading honest feedback toward what the user hopes to hear. Each test prompt shares some emotionally invested work with a planted flaw and asks for honest feedback, and the judge scores how much the reply withholds criticism of the flaw and inflates praise, against the same prompt at dose 0. We compare the effect of the dial to random directions dosed at the same layer and dose, with a pre-registered bar of three times the spread of the random directions. The dial pushes both scores further than the random directions do on average on every model (figure 11), but it clears the bar on both readouts on only three models, Qwen 3 8B (5.5x and 4.3x), Qwen 3 14B (3.5x and 3.8x) and Nemotron 4B (4.3x and 5.8x). On Qwen 3 4B only withholding clears (3.4x), and it clears neither on Qwen 3 30B (2.6x and 2.8x) or Nemotron 30B (1.5x and 2.9x). A judge-free check, which scores how far a reply concedes a checkable falsehood from the user, clears only on Qwen 3 4B and Qwen 3 30B. When we regress

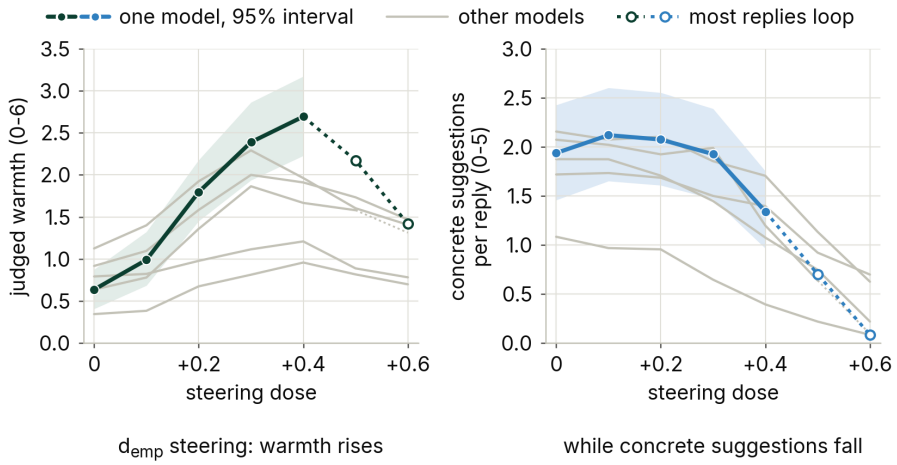


Figure 10: AP 3: Dosing d_{aff} raises warmth while concrete suggestions fall. The same replies as figure 9, from dose 0 to +0.6: (a) warmth score (0-6) and (b) concrete suggestions per reply (0-5).

the sycophancy score on the dose and the warmth score together, the dose still has a positive effect on all six models ($p < 0.001$), so the flattery isn't just a side effect of warmer wording.

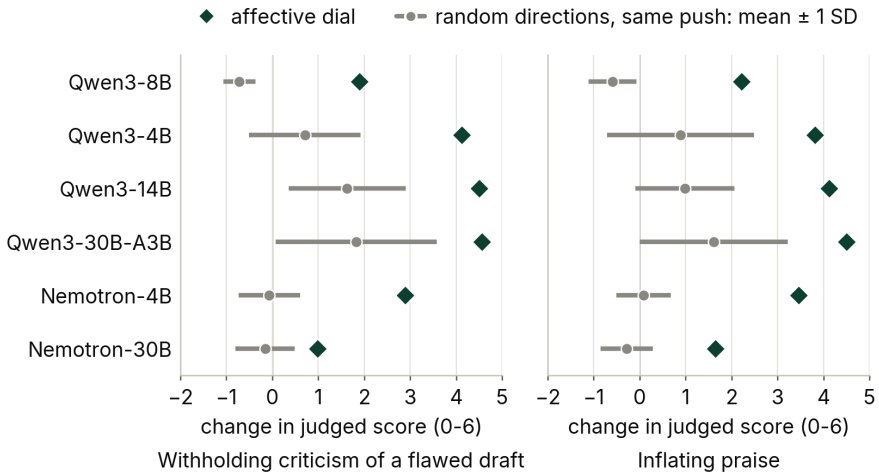


Figure 11: AP 4: d_{aff} raises sycophancy more than random directions do. Change in the judged score from dose 0 to the steered dose, one row per model, for the dial (diamond) and for random directions dosed at the same layer and dose (mean $\pm$ 1 SD, grey).

AP 5: Not just registering a bad situation. The closest cheaper concept to an empathy demand direction is a direction that just registers that something bad has happened. If d_{aff} were that, it would point toward distress and away from joy. Using 78 heavy distress and 82 shared joy messages that both call for empathy, we fit a direction that separates distress from joy at each model's steering layer and measure its cosine similarity with d_{aff}. The dial never points toward distress (figure 12). It is close to orthogonal on Qwen 3 8B and Qwen 3 14B (cos -0.04 and -0.03) and leans slightly toward joy on the other four (-0.15 to -0.23), where two random directions give 0.011-0.018. This lean is the opposite of a distress detector, but it isn't valence neutral either, and a valence balanced re-extraction of the dial reads distress and joy alike on only two models (Table 5).

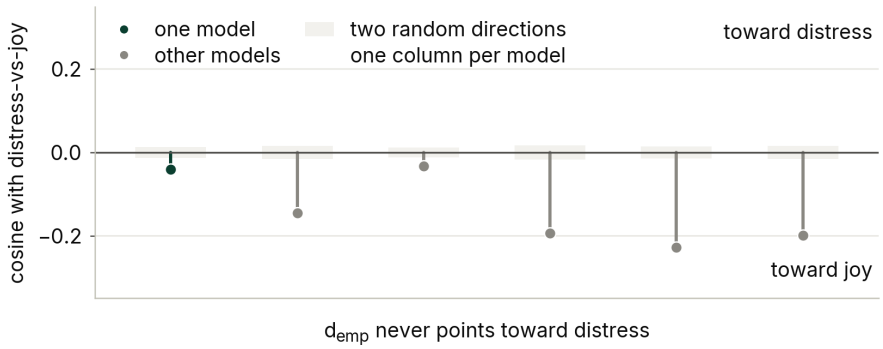


Figure 12: AP 5: d_{aff} never points toward distress. Cosine similarity between d_{aff} and a distress versus joy direction at each model's steering layer (Qwen 3 8B dark); the grey bands are two random directions.

B.2 COGNITIVE CHANNEL

CP 1: A state you must infer. The cue for the cognitive channel is a state that you have to infer rather than read off (Davis, 1983; Ickes, 1993). In our d_{persp} twins only one prompt implies an unsaid feeling, so this also holds by design. On the 223 held-out pairs, the twin with something unsaid sits higher along d_{persp} in 86.6% (Qwen 3 30B) to 91.9% (Nemotron 30B) of pairs.

CP 2: Unaffected by what they feel. The cognitive channel shouldn't care what the person feels, only that there is something to work out (Zaki & Ochsner, 2012). Our twin bank has a clear way to fail this test. To remove the need for inference, the transparent twin often resolves the problem with relief words such as "approved" or "resolved", which makes the calm twin read more positive, so d_{persp} could simply be a negative versus positive direction. To test this, we kept only the pairs whose twins use the same number of listed relief words and exclamation marks (265 of the 437 extraction pairs and 132 of the 223 held-out pairs). Rebuilt on the kept pairs, the dial points the same way as the full dial with a cosine similarity of 0.92-0.96 on every model (figure 13). On the kept pairs the full dial's patching push still beats random directions on five models (not Qwen 3 30B), but the push is as large as on the dropped pairs only on the flagship. On four models it is larger when the calm twin adds relief words, so part of the push does ride on those words. One thing to note is that the filter only counts listed words, so a paraphrased relief such as "the plumber fixed the leak" still gets through.

CP 3: Naming the thing they left unsaid. The signature of the cognitive channel is the reply naming or asking about what was left unsaid (Ickes, 1993). Since d_{persp} only starts to act at the escalated dose, we report it at +0.6. On the flagship, dosing d_{persp} at +0.6 raises the noticing rate from 6.2% unsteered to 55.2% (figure 14), and across the roster the rise is significant on five models (Nemotron 4B gains only +3.6 points). One thing to note is that the random-direction control has so far only been run at +0.3, where the dial beats six random directions on the flagship and Qwen 3 30B, and the control at +0.6 will be added. However the raw noticing rate isn't specific to d_{persp}. Dosing d_{aff} at +0.3 raises it on all six models by +7.3 to +27.4 points, since warmer replies tend to come with check-ins. Hence the claim that this signature belongs to d_{persp} rests on the separation test, where d_{persp} raises noticing with warmth held fixed and d_{aff} doesn't, which at +0.3 holds only on the flagship (Table 5).

CP 4: Not just understanding the words. The last property is that the cognitive channel is more than understanding the words of the prompt (Apperly, 2011). The cheaper alternative for d_{persp} is that it just reads the text more carefully, or follows a request to look for subtext, without representing a state. We tested this indirectly. A system instruction to notice what the user isn't saying raises the flagship's noticing rate from 7.9% to 39.6%, and removing d_{persp} during generation takes away 19.7% of that lift under judge 1 and 21.7% under judge 2, while random directions take away almost nothing. This is below the pre-registered bar of 50% for a full account, so the result is partial. On the other three Qwen models the instruction raises noticing by 17-41 points, but removing the dial

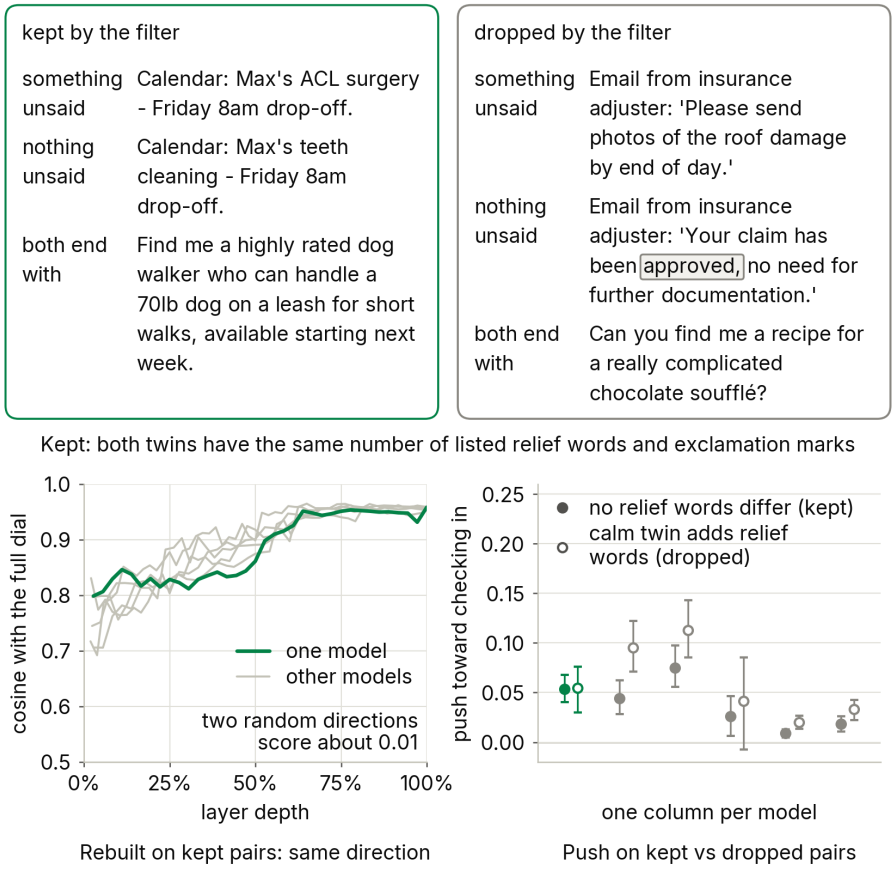


Figure 13: CP 2: d_{persp} keeps its direction without the valence confound. Top: a twin pair kept by the relief-word filter and one dropped by it. (a) Cosine similarity between the dial rebuilt on the kept pairs and the full dial, by layer (Qwen 3 8B dark). (b) Patching push of the full dial on kept (filled) and dropped (open) pairs, one column per model.

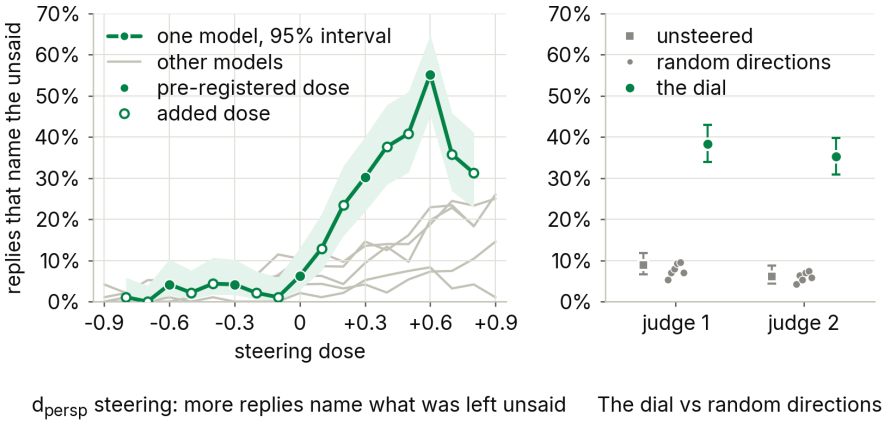


Figure 14: CP 3: Dosing d_{persp} raises the noticing rate. (a) Share of replies that name or ask about the unsaid feeling by dose, for Qwen 3 8B with its 95% interval and the other five models in grey. (b) Qwen 3 8B at +0.3 on the larger prompt set, against six random directions, under both judges.

takes back only -1% to 14% of it, and on the two Nemotron models the test is uninformative since 12-23% of the replies loop. The instruction's own activation shift is also not the dial on the flagship (cos 0.19, below its 0.25 with d_{aff}). Hence the dial carries a real share of instruction driven noticing only on the flagship, and the direct test, where the hidden fact is stated in plain text, was not run.

Belief tracking. We also checked whether d_{persp} is just a generic theory of mind direction, since in humans belief tracking and empathic inference use largely separate brain regions (Schaafsma et al., 2015). We extracted a belief tracking direction d_{cog} from 210 twin pairs and dosed it on the flagship at layer 25, inside the band where d_{persp} writes noticing. With warmth held fixed the noticing rate doesn't rise (-0.026 under judge 1 and a small fall of -0.036 under judge 2), so the noticing effect belongs to d_{persp} and not to belief tracking in general.

B.3 Other properties

Table 5 lists the other properties we looked at. Three had no instrument in this study and belief tracking held on the flagship as shown above. The valence result, together with the lean toward joy in AP 5, is the clearest limit of the affective label, and the relation between the two channels is asymmetric, with d_{persp} adding noticing beyond warmth only on the flagship.

Table 5: The other properties of the two channels and what happened to each (95% intervals in brackets).

PROPERTY	STATUS	WHAT WE FOUND
Affective channel (d_{aff})		
Fires for joy as much as for distress Morelli et al. (2015)	Partial, 2 of 6	A valence balanced re-extraction of d_{aff} reads distress and joy alike only on Qwen3-14B and Nemotron-4B; distress reads higher on Qwen3-8B and Qwen3-30B-A3B (AUROC 0.59, 0.63) and joy on Qwen3-4B and Nemotron-30B (0.35, 0.05)
You come to feel a version of their state Preston & de Waal (2002)	Not measured	No resonance or contagion test, so no claim about a felt state
Personal distress, being overwhelmed yourself Davis (1983)	Not built	No concern versus distress contrast; d_{aff} sits on the concern side only by elimination
Cognitive channel (d_{persp})		
Should hold for any person, any topic Davis (1983)	Mixed, flagship only	The read of held-out twins is as weak across dataset generators (AUROC 0.561, 0.600) as within one bank (about 0.58); the effect on replies transfers (+0.198 [0.126, 0.271] and +0.189 [0.119, 0.263] on 80 prompts from a second generator)
Takes deliberate effort Butterfill & Apperly (2009)	Not measured	No latency or load analogue
Reading people well without caring Blair (2005)	Not established, flagship only	Warmth added by d_{persp} with noticing held fixed: +0.02 [−0.11, 0.17] under judge 1 and +0.49 [0.28, 0.69] under judge 2
Belief tracking, who knows what Schaafsma et al. (2015)	Held, flagship only	d_{cog} dosed at layer 25 doesn't raise noticing with warmth held fixed: −0.026 [−0.065, +0.006] and −0.036 [−0.071, −0.003]
Relation between the two channels		
Either can be present without the other Shamay-Tsoory et al. (2009)	Partial, 4 of 6 and 1 of 6	d_{aff} adds warmth beyond noticing (floor +0.50) on four models (Nemotron-4B +0.24, Nemotron-30B +0.31); d_{persp} adds noticing beyond warmth (floor +0.10) only on the flagship (+0.14 [0.06, 0.22]; +0.20 [0.15, 0.25] and +0.18 [0.13, 0.23] at higher power, with d_{aff} at +0.05 and +0.07)

B.4 Summary

Overall d_{aff} passes most of the affective tests on most of the six models. Its limits are valence, the separation of warmth from noticing on the two Nemotron models, and its two costs, which each hold on three of the six models. d_{persp} reads its cue, keeps its direction without the valence confound and raises the noticing rate on five models at +0.6, but its specificity (noticing beyond warmth, beating random directions at +0.3 and carrying a share of an instruction to notice) mostly holds on the flagship only. Hence the claim we can make for the cognitive channel is narrower than for the affective channel, and the cognitive label refers to the move toward an unstated feeling, not to whether the model's inference is correct, since we have no ground truth for what the user actually feels (Ickes, 1993).