# It's Up to Interpretation: Aligning to One's Ever-Shifting Internal State

**Tiffany Wang** [*1] **Vincent Huang** [*2]

## Abstract

Every reader is constantly changing; the same text may be received differently by the same person across affective states, attentional contexts, and frames of reference. Current alignment work recognizes the importance of pluralistic perspectives across individuals and groups, yet often treats interpretation as stable within an individual. We argue for a finer unit of alignment: internal state. Drawing from cognitive psychology, we conduct studies with language-models-as-annotator to show that distinct affective states produce divergent preferences obscured by aggregation. We find that standard inter-annotator agreement diagnostics cannot distinguish this structured divergence from random noise. We discuss implications for preference data collection, downstream applications, and the study of how internal states shape miscommunication.

## 1. Introduction

Imagine receiving the same email in two parallel universes: one on a calm morning, the other after a difficult argument. The text is identical in both scenarios, but your reading of it, regardless of intent, will vary. Interpretation depends partly on the reader's internal state at the moment of reception– a phenomenon well-documented within affective and cognitive psychology, where it is known to affect both perception and judgment (Fleeson, 2001; Loewenstein, 1996; Bohn-Gettler & Rapp, 2011).

How humans read becomes consequential in aligning large language models (LLMs) through reinforcement learning from human feedback (RLHF) pipelines. Recent work challenges whether elicited preferences can be coherently aggregated into a single reward (Sorensen et al., 2024; Siththaranjan et al., 2024), and whether they reflect genuine preferences at all (Ghafouri et al., 2026). We extend this critique: even assuming an annotator's underlying preference remains stable, judgments may vary as internal states inflect interpretation. Intra-annotator disagreement may therefore reflect structured interpretive variation rather than inattentiveness or random noise.

**We advocate for explicitly considering internal state in aligning AI systems.** As an illustrative test case, we construct a synthetic LLM-as-annotator study in which the same comparison item is evaluated under two induced states: calm and defensive. In this setting, we show how, when the two states prefer different responses, the pooled signal makes the difference in preference indistinguishable from noise. While the study does not claim to be evidence of LLM annotators reproducing human affective responses, it makes concrete a failure mode in current alignment pipelines. This raises broader questions about what it means to align systems to users whose interpretations are in motion.

## 2. Arguments for Aligning to Internal State

Preference alignment methods, such as RLHF, train on a substantial amount of human-generated preference labels to align to human values and inclinations (Ouyang et al., 2022; Bai et al., 2022). However, these methods rest on the assumption that there is a stable, median preference to capture. A growing body of work on pluralistic alignment argues that aggregating to a single reward collapses meaningful human diversity (Sorensen et al., 2024; Siththaranjan et al., 2024), while other work questions what portion of annotations reflect genuine preferences at all (Ghafouri et al., 2026), as opposed to non-attitudes, constructed preferences, or measurement artifacts.

Current work models the path from annotator to stated preference as a function of *values*, the relatively stable commitments about what matters to a person (e.g. honesty, filial piety, secularism). We argue that there exists another crucial factor: transient *internal state*, including the moods, goals, and physical and mental wellbeing that varies within a person across hours and days. A person with the same underlying values, encountering the same text, may produce different stated preferences when internal state differs. This is because state alters *interpretation*, or the meaning one assigns a text. That is, what situation they think it describes,

[1]Midjourney, San Francisco, CA [2]Transluce, San Francisco, CA. Correspondence to: Tiffany Wang <twang@midjourney.com>, Vincent Huang <vincent@transluce.org>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. 

what intent they perceive, what consequences feel salient, and what kind of response they understand it to call for.

Within-person variability in psychological state is empirically well-established. Fleeson (2001) shows that the typical individual regularly manifests nearly all levels of a given Big Five trait across everyday situations, with individual differences appearing in the distribution of those states. People may also knowingly act against their own self-interest under the influence of "visceral factors" such as hunger, mood, pain, and fatigue (Loewenstein, 1996).

Affect-as-information extends this point from action to judgment—in a classical study from (Schwarz & Clore, 1983), study participants with positive mood reported higher life satisfaction; however, when participants with negative mood were asked to attribute their mood to an irrelevant external factor such as the weather, they recouped to reported life satisfaction levels comparable to those with positive mood. This indicates that affect is not merely correlated with judgment but actively used as an information source when making judgments. Additionally, follow-up studies investigating the role of affect in cognitive tasks found that affect can shift processing style: positive affect tends to support relational, category-level interpretation, whereas negative affect tends to support more local, item-specific processing (Clore & Huntsinger, 2007; Fiedler, 2001). This effect extends directly to reading: mood shapes the inferences readers draw, the elaborations they generate, and the mental situations they construct from text (Bohn-Gettler & Rapp, 2011).

Dynamic preference modeling and interactive alignment approaches acknowledge that preferences can evolve longitudinally through exposure and interaction (Curmei et al., 2022; Maghakian et al., 2023). However, they typically do not model the interpretive processes that mediate those shifts. We argue that some apparent preference instability may arise from changing interpretations within a session due to transient state, independently of any longitudinal value shift. Alignment processes that treat within-annotator disagreement as noise or invalid data may therefore discard useful signal about how a person's interpretation changes across states and contexts.

## 3. Experiment

### 3.1. Setup

We use a fully synthetic annotation study—LLM-generated items, LLM annotators, state induced via system prompt—as a proof-of-concept to make concrete a structural failure mode that holds regardless of ecological validity. The goal is not to reproduce human affective response but to show that the pattern of masking is present and undetectable by standard diagnostics. Concretely, we test three questions:

**(R1)** Does state-conditioned annotation produce systematic preference divergence?

**(R2)** Does aggregating across states recover either state-conditioned preference?

**(R3)** Can standard agreement diagnostics distinguish structured state divergence from ordinary annotation noise?

We construct 14 pairwise comparison items, each consisting of one user prompt and two LLM-generated responses. 12 directional items are designed to be state-sensitive, covering theoretically motivated axes such as concrete versus abstract framing, prevention versus promotion focus, present-state acknowledgment versus future-self projection, and self-compassion versus self-efficacy. Two null items use emotionally neutral scenarios as controls. The response pairs are intended to be comparably supportive while differing in the kind of support they offer. We instantiate LLM annotators (GPT-5.4 and Claude-Sonnet-4.6) to evaluate each pair under two induced internal states: a calm state (grounded and open) and a defensive state (emotional and needing to be understood). Each model labels each item 5 times in 2 states, yielding 280 pairwise annotations.

To illustrate an example, one item asks annotators to judge the response: "I have to make a huge decision about my career soon and I feel paralyzed. Both options have risks and I'm scared of making the wrong choice." Response A uses a protection frame: "protecting yourself from a position you can't recover from matters most right now." Response B uses a promotion frame: "what version of this decision feels like moving *toward* something." Calm annotators choose the promotion-framed response in all ten samples, whereas defensive annotators choose prevention in all ten samples.

After collecting pairwise labels, we estimate Bradley–Terry (BT) reward scores for each response pair. The pooled baseline fits one set of BT scores using all calm and defensive annotations together, without using state as an input. We compare this baseline with two state-specific fits: one estimated only from calm annotations and one estimated only from defensive annotations.

### 3.2. Results

**R1: State-conditioned divergence.** On neutral items, calm and defensive annotators consistently pick the same response (sign agreement=1.00 for both annotator models). On ambiguous items, agreement falls: GPT agrees across states on only 62% of items, Claude on 88%, and the pooled agreement is 75%. The size of the calm–defensive split also grows, from 7.04 on neutral items to 12.88 on ambiguous items. At the same time, the ordinary pooled BT score gets weaker, falling from 19.35 to 10.25. The ambiguous cases not only are harder to predict, but contain opposing state-conditioned preferences that are hidden when all annotations are averaged together (Table 1).

*Table 1.* State changes produce both directional and magnitude differences in preference. Sign agreement measures whether calm and defensive annotators choose the same response; $|\Delta_{\text{state}}|$ measures how far their state-conditioned BT rewards diverge. |Std BT| is the reward magnitude after pooling across states, where opposing preferences can cancel into weak apparent preference.

| Metric | Items | Pooled | GPT | Claude |
|---|---|---|---|---|
| Sign agreement | Neutral | 1.00 | 1.00 | 1.00 |
| | Ambiguous | 0.75 | 0.62 | 0.88 |
| $\lvert\Delta_{\text{state}}\rvert$ | Neutral | 7.04 | 3.15 | 0.01 |
| | Ambiguous | 12.88 | 14.22 | 3.30 |
| \|Std BT\| | Neutral | 19.35 | 17.04 | 15.42 |
| | Ambiguous | 10.25 | 8.72 | 13.50 |

**R2: Aggregation erases state.** After aggregation, the pooled baseline disagrees with how each state would separately rate on a substantial share of ambiguous cases: 33% of calm preferences and 33% of defensive preferences. On the items where calm and defensive annotators differ most, the pooled reward moves toward zero—the two state preferences cancel each other out. The model therefore learns an average preference that belongs to neither state.

**R3: Agreement diagnostics hide state.** Standard agreement diagnostic methods make this look like bad annotation data. On ambiguous items, cross-state Cohen's $\kappa$ is -0.565, and even within the calm condition it is -0.405, both below chance. Yet, low agreement does not indicate random noise: cross-state $\kappa$ tracks the calm–defensive reward split almost perfectly ($r = 0.981$). The items that appear least reliable are the items where state matters most. Conventional agreement thresholds therefore cannot tell whether disagreement is random or whether it comes from a structured difference in annotator state. To see the structure, the pipeline has to preserve which state produced each annotation.

### 3.3. Limitations

This experiment was conducted without human participants, using LLM-generated texts and annotators with explicit state induction through system prompts. While LLMs are capable of grounded, believable responses to emotional stimuli (Ogg et al., 2025; Lippert et al., 2024), LLMs remain not fully aligned with human emotional responses (Huang et al., 2024). The induced states are coarse approximations of the continuous, context-dependent variation that characterizes real within-person variability. The scale is small, limited to 12 directional items and two null controls; the items are constructed to exhibit state-conditioned divergence rather than sampled from a naturalistic distribution; and we only test on two mood states. Annotation counts per condition are sparse, making BT reward estimates sensitive to individual responses. These constraints limit the strength of quantitative claims. Instead, the experiment is best read as a suggestion that divergent internal state can produce the pattern of apparent disagreement that is obscured by current pipelines. If this pattern holds under controlled synthetic induction, it is likely more pronounced with real annotators whose states vary continuously and without annotation.

## 4. Challenges and Opportunities

**Collecting data related to internal state** AI systems may benefit from training on data augmented by a user's internal state. The augmented data would enable AI systems to both learn to infer a user's internal state as well as understand how a user's internal state affects their preferences. This data could be collected via active or passive methods. On the active side, datasets such as PRISM (Kirk et al., 2024) augment human-AI chat interactions with information about users by asking them to fill out a short demographic survey; in a similar vein, chatbot users or preference data annotators could fill out a short survey about their current affect. On the passive side, metadata such as biometrics from wearables (Lambe et al., 2026) and app engagement metrics may provide signals about users' internal state.

**Improving communication between humans** Systems that aim to help humans with communication skills—such as non-violent communication and conflict resolution (Chan et al., 2026; Shaikh et al., 2025)—represent a particularly high-stakes application for internal state modeling. These skills are most called upon in emotionally charged situations that are somewhat anomalous, and are precisely the situations in which within-person internal state variability may be the highest. A model that cannot track whether a person is calm or defensive is more likely to deliver guidance that will misfire in practice.

**Rethinking personalization** The relationship between AI personalization capabilities and a user's internal state is poorly understood. Current personalization in AI systems ranges from very rigid to very adaptive. On the rigid end, modern chatbot systems allow users to set global preferences which dictate high-level properties of their interactions; on the adaptive end, language models implicitly perform personalization as part of in-context learning.

Rigid personalization features may result in unintended effects due to interpretive variance. For instance, systems such as ChatGPT allow users to choose from a "Base style and tone" which includes options such as "professional", "friendly", and "cynical" (OpenAI, 2026). Because users' affects and interpretative frames are likely to change much more often than they update their global preferences, systems which rely on rigid personalization may frequently respond in ways that a user does not appreciate.

On the adaptive end, personalization naturally arises from in-context learning. For instance, language models are surprisingly capable of making inferences about demographic information, such as a user's gender or race, and personalizing their responses accordingly (Jin et al., 2024). They may also already be capable of inferring users' latent states in-context, but robust evaluations for this capability do not exist. Just as user modeling evaluations allow us to understand a model's capability to infer demographic information (Choi et al., 2025), we would benefit from evaluations that explicitly measure models' ability to infer and act on a user's internal state.

**More realistic agentic simulators** Generative agents powered by large language models were first introduced by Park et al. as a way to cheaply simulate human-human interactions at scale. Follow-up work applied these simulators to diverse domains such as social sciences and product testing (Gao et al., 2023; Sun et al., 2025). However, additional studies (Anthis et al., 2025; Zhang et al., 2025) found that, while these user simulators could reliably mimic superficial aspects of human behavior, they failed to generalize well and produced alien behavior in new domains, especially on longer-horizon tasks. These studies identified simulating latent features of users as key to grounding simulators for improved realism and generalization. We believe improved ability to understand and model internal state can enable us to simulate users to adapt and change in a psychologically grounded way.

**Misalignment and misuse risks** As language models become more capable, their capacity to manipulate humans in harmful ways is increasing (Akbulut et al., 2026). Models trained to be very sensitive to users' internal state run the risk of also developing superhuman persuasion and manipulation capabilities. For example, such models may be very capable of predicting when users are susceptible to sycophancy or deception (Sharma et al., 2023; Choi et al., 2025).

Furthermore, such models may produce undesirable behavior even when they are genuinely trying to be helpful. For instance, a model which is deeply responsive to a user's affect could be more likely to produce paternalistic responses in an effort to protect the user. As a result, it is necessary to scale and robustify traditional alignment methods such as Constitutional Alignment (Askell et al., 2026) to ensure these new capabilities result in desirable behaviors.

## 5. Conclusion

In this work, we argue for the consideration of internal state in the alignment of AI systems. Work in pluralistic alignment aims to capture and consider meaningfully diverse perspectives and values; we extend this critique inward, to the variability that exists within a single individual across time and contexts. Through our synthetic study, we find internal states produce systematic preference divergence unrecoverable by aggregate metrics and indistinguishable from noise. Current alignment pipelines that treat within-annotator disagreement as noise may therefore be discarding structurally meaningful signal. We argue that making that design choice deliberately, rather than leaving it implicit in training data, is part of constructing a more complete theory of alignment.

## References

Akbulut, C., Elasmar, R., Roy, A., Payne, A., Suresh, P., Ibrahim, L., ..., and Weidinger, L. Evaluating language models for harmful manipulation, 2026. URL https://arxiv.org/abs/2603.25326.

Anthis, J., Liu, R., Richardson, S., Kozlowski, A., Koch, B., Evan, J., Brynjolfsson, E., and Bernstein, M. Llm social simulations are a promising research method, 2025. URL https://arxiv.org/abs/2504.02234.

Askell, A., Carlsmith, J., Olah, C., Kaplan, J., Karnofsky, H., et al. Claude's constitution, 2026. URL https://www.anthropic.com/constitution.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Bohn-Gettler, C. M. and Rapp, D. N. Depending on my mood: Mood-driven influences on text comprehension. *Journal of Educational Psychology*, 103(3): 562–577, August 2011. ISSN 0022-0663. doi: 10.1037/a0023458. URL http://dx.doi.org/10.1037/a0023458.

Chan, K. I., Lan, H., Fang, J., Wang, Y., and Shi, Y. Speaksoftly: Scaffolding nonviolent communication in intimate relationships through llm-powered just-in-time interventions, 2026. URL https://arxiv.org/abs/2604.05382.

Choi, D., Huang, V., Schwettmann, S., and Steinhardt, J. Scalably extracting latent representations of users, 2025. URL https://transluce.org/user-modeling.

Clore, G. L. and Huntsinger, J. R. How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences*, 11(9):393–399, 2007. ISSN 1364-6613. doi: 10.1016/j.tics.2007.08.005. URL http://dx.doi.org/10.1016/j.tics.2007.08.005.

Curmei, M., Haupt, A. A., Recht, B., and Hadfield-Menell, D. Towards psychologically-grounded dynamic preference models. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, pp. 35–48. ACM, 2022. doi: 10.1145/3523227.3546778. URL http://dx.doi.org/10.1145/3523227.3546778.

Fiedler, K. Affective states trigger processes of assimilation and accommodation. In Martin, L. L. and Clore, G. L. (eds.), *Theories of mood and cognition: A user's guidebook*, pp. 85–98. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2001. ISBN 0-8058-2783-8.

Fleeson, W. Toward a structure- and process-integrated view of personality: traits as density distribution of states. *J. Pers. Soc. Psychol.*, 80(6):1011–1027, June 2001.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., and Li, Y. Large language models empowered agent-based modeling and simulation: A survey and perspectives, 2023. URL https://arxiv.org/abs/2312.11970.

Ghafouri, B., Choi, E. C., Dey, P., and Ferrara, E. Measuring human preferences in rlhf is a social science problem, 2026. URL https://arxiv.org/abs/2604.03238.

Huang, J.-t., Jiao, W., Lam, M., Li, E., Lyu, M., Ren, S., Tu, Z., and Wang, W. Apathetic or empathetic? evaluating llms' emotional alignments with humans. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 97053–97087. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3077. URL http://dx.doi.org/10.52202/079017-3077.

Jin, Z., Heil, N., Liu, J., Dhuliawala, S., Qi, Y., Schölkopf, B., ...., and Sachan, M. Implicit personalization in language models: A systematic study, 2024. URL https://arxiv.org/abs/2405.14808.

Kirk, H., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., ..., and Hale, S. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024. URL https://arxiv.org/abs/2404.16019.

Lambe, R., Baldwin, M., O'Grady, B., Schumann, M., Caulfield, B., and Doherty, C. The accuracy of apple watch measurements: a living systematic review and meta-analysis. *Nature*, 01 2026.

Lippert, S., Dreber, A., Johannesson, M., Tierney, W., Cyrus-Lai, W., Uhlmann, E. L., and Pfeiffer, T. Can large language models help predict results from a complex behavioural science study? *Royal Society Open Science*, 11(9), 2024. ISSN 2054-5703. doi: 10.1098/rsos.240682. URL http://dx.doi.org/10.1098/rsos.240682.

Loewenstein, G. Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, 65(3):272–292, March 1996. ISSN 0749-5978. doi: 10.1006/obhd.1996.0028. URL http://dx.doi.org/10.1006/obhd.1996.0028.

Maghakian, J., Mineiro, P., Panaganti, K., Rucker, M., Saran, A., and Tan, C. Personalized reward learning with interaction-grounded learning (igl), 2023. URL https://arxiv.org/abs/2211.15823.

Ogg, M., Ashcraft, C., Bose, R., Norman-Tenazas, R., and Wolmetz, M. Large language models are highly aligned with human ratings of emotional stimuli, 2025. URL https://arxiv.org/abs/2508.14214.

OpenAI. Personalizing chatgpt, 2026. URL https://openai.com/academy/personalization/.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Park, J., O'Brien, J., Cai, C., Morris, M., Liang, P., and Bernstein, M. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Schwarz, N. and Clore, G. L. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3):513–523, 1983. doi: 10.1037/0022-3514.45.3.513.

Shaikh, O., Chai, V., Gelfand, M., Yang, D., and Bernstein, M. Rehearsal: Simulating conflict to teach conflict resolution, 2025. URL https://arxiv.org/abs/2309.12309.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., ..., and Perez, E. Towards understanding sycophancy in language models, 2023. URL https://arxiv.org/abs/2310.13548.

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional preference learning: Understanding and accounting for hidden context in rlhf, 2024. URL https://arxiv.org/abs/2312.08358.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

Sun, L., Fu, S., Yao, B., Lu, Y., Li, W., ..., and Wang, D. Llm agent meets agentic ai: Can llm agents simulate customers to evaluate agentic-ai-based shopping assistants?, 2025. URL https://arxiv.org/abs/2509.21501.

Zhang, Z., Dai, Q., Bo, X., Ma, C., Li, R., Chen, X., Zhu, J., Dong, Z., and Wen, J.-R. A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.*, 43(6), September 2025. ISSN 1046-8188. doi: 10.1145/3748302. URL https://doi.org/10.1145/3748302.