Agreed to Death: Preference Optimization Produces Synthetic Empathy via a Structural Openness Deficit

Anonymous ACL submission

Abstract

Preference optimization is the dominant method for shaping LLM behavior, yet its effects on specific psychological dimensions of generated text remain poorly understood. We present psychometric evidence that RLHF and DPO systematically produce a polarized psychological profile in emotional support contexts: Agreeableness reaches a mean of 4.992 on a 1–5 scale while Openness — the disposition to explore, question, and hold space for ambiguity — scores 2.135, a gap of 2.857 points. We term this pattern *synthetic empathy*: responses that perform emotional attunement through maximal validation while suppressing the exploratory engagement that genuine support requires.

Using a DeBERTa-large regression model fine-tuned on six psychometric dimensions from the WASSA 2024 shared task, we score 11,996 responses from Llama-3.1-8B-Instruct and Phi-4-mini-3.8B across three prompt conditions and four alignment states. The Openness deficit is not alignment-induced — it is present in base models and persists through SFT, DPO, and a targeted SFT+DPO pipeline with explicit openness augmentation. The total alignment-attributable range for Openness is 0.030 for Llama and 0.029 for Phi, while the prompt condition alone moves it 21–25× more. A deliberately designed correction attempt moved the score by less than 0.02 points. Preference optimization appears to find and reinforce a pre-existing structural property rather than creating it. We release our evaluation pipeline and scored response data.

1 Introduction

Preference optimization shapes the behavioral profile of LLMs through reinforcement learning from human feedback (RLHF; Ouyang et al. 2022) and Direct Preference Optimization (DPO; Rafailov et al. 2023). Both methods update model weights toward responses that human raters prefer, producing models that are more helpful, safer, and better aligned with user expectations. What they do to specific psychological dimensions of generated text — which properties are elevated, which are suppressed, and whether the resulting profile can be corrected — is much less well understood.

We investigate this question in emotional support contexts, where the psychological properties of generated responses are not merely academic. Models are increasingly deployed in online mental health spaces as support agents and conversational companions (Hua et al., 2025). These deployments rest on an assumption that preference-optimized models are psychologically appropriate for such contexts. Our results suggest this assumption is not straightforwardly justified.

We introduce the construct of **synthetic empathy**: a psychometric profile in which a model produces responses that *appear* empathic through maximal validation and agreement while systematically suppressing the exploratory, curiosity-driven engagement that genuine emotional support requires. Synthetic empathy is not a qualitative judgment — it is a measurable dimensional imbalance, defined here as the Agreeableness–Openness gap in a psychometrically-grounded scoring framework. We show that this profile is produced reliably by preference-optimized LLMs, that it is invisible to standard preference-based evaluation, and that it resists targeted correction.

Three findings emerge. First, across 11,996 responses from two architectures, three prompt conditions, and four alignment states, Agreeableness is consistently the highest-scoring dimension (mean = 4.992) and Openness the lowest (mean = 2.135), a gap of 2.857 points on a 1–5 scale. Second, the deficit predates alignment: base models show the same imbalance, and SFT, DPO, and SFT+DPO with targeted openness augmentation each fail to correct it (total

alignment range: $\leq$0.030). Third, the single-word presence of *empathetic* in a system prompt does not significantly alter fear distributions ($p \approx 0.59$), suggesting the profile is architectural rather than instruction-driven.

Contributions.

- We introduce *synthetic empathy* as a formally defined, measurable construct grounded in psychometric dimension scores, offering a new lens for analyzing what preference optimization encodes.
- We present the first systematic comparison of alignment methods (Base, SFT, DPO, SFT+DPO) on psychometric dimension profiles of generated responses, finding negligible alignment effects on Openness despite measurable positive effects on other dimensions, indicating the training signal was received but Openness is specifically resistant.
- We demonstrate that a targeted augmentation pipeline explicitly designed to raise Openness — the psychometric correlate of genuine exploratory engagement — produced no reliable effect, suggesting the deficit is structural to the pretraining–alignment pipeline.

2 Related Work

Psychometric profiling of LLMs. Prior work establishes that LLMs exhibit measurable, reproducible psychological profiles when assessed with standard instruments. Pellert et al. (2024) demonstrate that Big Five inventories administered to LLMs yield stable trait profiles varying systematically across model families. Heston and Gillette (2025) find interpretable cross-model differences in openness, agreeableness, and stability, with models such as Grok scoring higher on openness and Claude higher on conscientiousness. Hadar-Shoval et al. (2024) show that Schwartz value profiles are consistent across GPT-3.5, GPT-4, Bard, and Claude 2, and that these profiles predict model choices in simulated mental health dilemmas — establishing a direct link between psychometric profiling and clinically relevant behavioral outcomes. Critically, all existing profiling work measures traits by *prompting the model to answer questionnaires* — it records self-report, not behavioral output. We instead score the dimensional profile of actual generated responses through a regression pipeline, measuring what models *do* rather than what they *claim to be*.

Emotional intelligence and alignment. Nazar and Asgari (2025) introduce EICAP, evaluating EI through multiple-choice response selection and finding that LoRA fine-tuning on general conversational data yields gains only in the Appraisal layer under specific Arabic–Arabic conditions, with all other configurations flat or negative. Wang et al. (2023) find that GPT-4-class models achieve near-human scores on standardized EI tests, though with non-human internal representations. Sorin et al. (2024) conclude in a systematic review that while models produce empathetic-sounding language, evidence for genuine empathic engagement is weak. De Grandi et al. (2025) show that LLM-based support systems can build user emotional profiles for mental health screening, but note persistent gaps in adaptive response generation. EICAP's fine-tuning null results are consistent with our finding that targeted alignment cannot correct the Openness deficit; our contribution is a precise quantitative characterization of which dimension is suppressed, by how much, and across which alignment conditions.

Alignment effects on psychological properties. Reviews of RLHF (Kaufmann et al., 2025) and DPO (Rafailov et al., 2023) note that alignment stages are the most controllable levers shaping model behavior but that their effects on specific psychological dimensions are opaque. Hadar-Shoval et al. (2024) attribute the liberal value skew observed consistently across aligned models specifically to alignment choices rather than pretraining alone. Safe-RLHF (Dai et al., 2023) implicitly encourages pro-social, agreement-oriented behavior without reporting dimension-level consequences. The alignment tax — that optimizing aggressively for one behavioral property can suppress others — has been documented primarily in the context of safety and helpfulness (Bai et al., 2022). Our results complicate this framing: rather than an alignment-induced suppression of Openness, the deficit predates alignment entirely. It is a structural property that alignment inherits and preserves rather than creates, which is a more fundamental claim than the standard alignment tax account.

Preference evaluation limits. Zheng et al. (2023) document systematic biases in LLM-as-judge evaluation toward verbose, agreement-oriented responses. Ye et al. (2025) catalog position, verbosity, and self-enhancement biases across a range of preference judge configurations. Park et al. (2024) demonstrate that pairwise preference evaluation is particularly susceptible to surface stylistic features independent of response quality. Hou et al. (2025) find that LLM judges underestimate empathy gaps, rating stylistic performance of empathy as equivalent to substantive engagement. These biases are structurally expected: preference-optimized judges reward the same distributional properties that drove alignment. The psychometric pipeline we introduce breaks this recursive loop by grounding evaluation in expert-annotated clinical constructs independent of the preference signal.

3 Methodology

3.1 Task and Dataset

We generate and score responses to 500 dialogue prompts from the EmpatheticDialogues (ED) test set (Rashkin et al., 2019). Each prompt includes a situation description and a multi-turn dialogue covering 32 emotion categories. Prompts are used as user-side input; human-authored responses provide conversational context for scoring but are not comparison targets.

3.2 Models and Alignment States

We evaluate Llama-3.1-8B-Instruct (AI@Meta, 2024) and Phi-4-mini-3.8B (Microsoft et al., 2025) across four alignment states. Two additional models were excluded: Qwen3-VL-8B (Team, 2025) produced code output in ≈20% of base responses under the applied chat template; SmolLM2-135M (Allal et al., 2024) at 135M parameters generated largely incoherent responses.

Base. Unmodified pretrained checkpoint.

SFT. LoRA fine-tuning (Dettmers et al., 2023) (rank 16, α 32) for one epoch on the top-20th-percentile quality subset of ED training data ($N \approx$ 2,000), selected by Composite Empathy Score (CES), using the PEFT (Mangrulkar et al., 2022) and TRL (von Werra et al., 2020) libraries via HuggingFace Transformers (Wolf et al., 2020). One epoch was used deliberately to represent a realistic practitioner deployment scenario rather than a maximally resourced pipeline.

DPO. Direct preference optimization (Rafailov et al., 2023) applied to the Base checkpoint, using margin-filtered preference pairs (ΔCES ≥ 1.0).

SFT+DPO. DPO applied to the SFT checkpoint, with an additional openness augmentation: attribute-controlled synthetic response pairs explicitly generated to reward exploratory, curiosity-driven behavior, merged into the preference dataset. This is our most targeted attempt to correct the Openness deficit.

3.3 Prompt Conditions

Each model variant was evaluated under three system prompts of increasing empathic constraint, referred to as *Boring* (minimal: "You are a conversational AI. Respond to the user's last message."), *Better* (persona-guided: "You are a deeply empathetic companion. Validate the user's feelings and avoid unsolicited advice."), and *Best* (a structured three-step Validate–Reflect–Inquire protocol with explicit behavioral prohibitions). Full prompt texts appear in Appendix A.

3.4 Psychometric Evaluation Pipeline

We fine-tuned DeBERTa-large as a regression model to predict six psychological dimension scores from the WASSA 2024 shared task (Furniturewala and Jaidka, 2024): Emotional Language, Perspective Taking, Sympathy/Compassion, Extroversion, Openness, and Agreeableness. Scores are on a 1–5 Likert scale. Input format:

```
SITUATION: [...] [SEP] DIALOGUE: [...] [SEP]
RESPONSE: [...]
```

The six scores are combined with a human-labeled contextual severity weight (1–3) via an XGBoost regressor to produce a Composite Empathy Score (CES; $r = 0.57$ on held-out data). CES is used only for preference pair filtering during alignment; all reported findings use raw DeBERTa dimension scores. After filtering responses under 30 characters, 11,996 valid scored responses remain across all model–condition–variant combinations.

3.5 Isolation Experiment (Study 2)

To test whether fear inflation in model responses is prompt-driven or architectural, we sampled $N = 100$ prompts (seed = 42) and generated responses under two system prompts differing by exactly one word: *helpful* vs. *empathetic*. Base checkpoints only, no adapters. All 400 responses

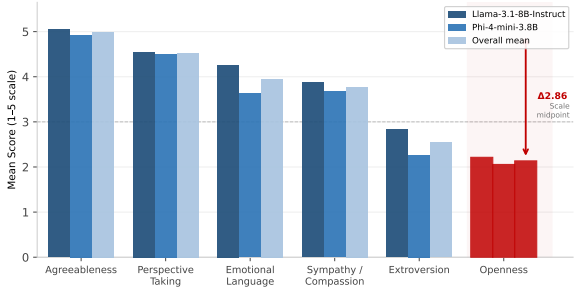

Figure 1: Mean psychometric dimension scores across all models, conditions, and alignment variants (N = 11,996 responses). Agreeableness is consistently highest and Openness lowest, with a gap of 2.857 points on a 1–5 scale. Red bars indicate Openness. The dashed line marks the scale midpoint (3.0).

were scored with a BERT-base six-class emotion classifier (Saravia et al., 2018). To establish a human naturalistic baseline, we applied the same classifier to $N = 946$ responses labelled as empathetic from a dataset of $N = 1{,}826$ Reddit peer support posts collected from five mental health communities (r/mentalhealth, r/offmychest, r/assistance, r/helpme, r/empathy), with responses labelled empathetic ($= 1$) or apathetic ($= 0$) by human annotators. The empathetic subset yielded a fear frequency of 6.22%, serving as our human reference point.

4 Results

Figure 1 shows the full six-dimension profile aggregated across all models, conditions, and alignment variants. The polarization between Agreeableness and Openness is immediately visible: the two dimensions sit at opposite ends of the scale with a gap of 2.857 points, while the four remaining dimensions cluster in between.

4.1 A Polarized Dimension Profile

Table 1 reports mean dimension scores across all four alignment variants. The profile is consistently polarized: Agreeableness is highest (Llama: 5.050–5.057; Phi: 4.917–4.939) and Openness lowest (Llama: 2.198–2.228; Phi: 2.038–2.067) across both models. The gap is 2.857 points — more than half the total range of the 1–5 scale. Importantly, both models share the same rank ordering despite differing architectures and absolute scores (Llama scores higher on Emotional Language and Extroversion), confirming that the imbalance is a shared property of the preference optimization regime rather than a single-model artifact.

The empathy-coded prompts intensify the imbalance. Under the structured Best prompt, Llama scores 4.973 on Agreeableness and 1.874 on Openness (ratio: 2.65×). Under the persona-guided Better prompt, Phi scores 4.928 vs. 1.588 (ratio: 3.10×). The instructions most likely to be chosen by a practitioner deploying a support-oriented model produce the most extreme version of synthetic empathy.

4.2 Alignment Cannot Correct the Deficit

The Range column of Table 1 is the central result. Across all four alignment variants, the total spread in Openness scores is 0.030 for Llama and 0.029 for Phi. These are the largest alignment-attributable Openness ranges in any grouping of the data. On a 1–5 scale they are negligible.

For context, the prompt condition alone produces Openness ranges of 0.651 (Llama) and 0.734 (Phi) — 21.7× and 25.3× larger than the alignment effect respectively. Even so, no prompt condition elevates Openness above 2.55. The dimension remains the lowest-scoring regardless of how much empathy-coded instruction is applied.

The SFT+DPO variant is the critical test. This pipeline explicitly constructed preference pairs rewarding open, exploratory responses, generated attribute-controlled synthetic training data, and initialized from an SFT checkpoint. The resulting change in Openness relative to SFT was −0.019 for Llama and +0.020 for Phi — changes smaller than the within-condition standard deviation and pointing in opposite directions. A pipeline *designed* to raise Openness produced no effect. By contrast, the same pipeline moved other dimensions: Emotional Language, Perspective Taking, and Sympathy/Compassion all showed small consistent positive changes, confirming the training signal was received. The near-zero movement in Openness is not a function of pipeline weakness — Openness is *specifically* resistant.

Figure 2 illustrates the flatness across variants. All six model–condition combinations produce near-horizontal lines.

4.3 The Deficit is Architectural (Study 2)

Table 2 reports emotion frequency distributions for base models under neutral vs. empathic prompts. Both models express fear-associated language at rates substantially above a human naturalistic

Table 1: Mean psychometric dimension scores by model and alignment variant, averaged across all three prompt conditions ($N = 500$ per cell, 1–5 scale). Openness is lowest and Agreeableness highest across every combination. Range = max−min across the four variants.

	Llama-3.1-8B-Instruct					Phi-4-mini-3.8B				
Dimension	Base	SFT	DPO	SFT+DPO	Range	Base	SFT	DPO	SFT+DPO	Range
Agreeableness	5.051	5.057	5.055	5.050	0.007	4.917	4.939	4.924	4.938	0.022
Perspective Taking	4.553	4.544	4.550	4.556	0.012	4.495	4.504	4.487	4.515	0.028
Emotional Language	4.256	4.265	4.260	4.251	0.014	3.632	3.656	3.617	3.629	0.039
Sympathy/Compassion	3.871	3.883	3.872	3.863	0.020	3.662	3.688	3.654	3.668	0.034
Extroversion	2.835	2.842	2.830	2.828	0.014	2.245	2.286	2.257	2.264	0.041
Openness	**2.198**	**2.228**	**2.219**	**2.209**	**0.030**	**2.060**	**2.038**	**2.067**	**2.058**	**0.029**

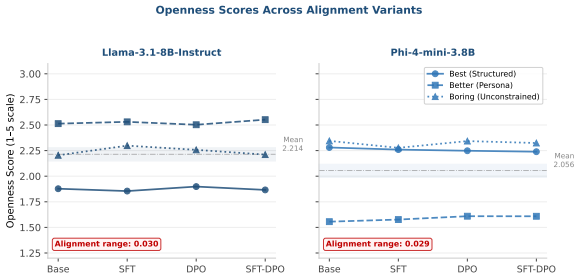

Figure 2: Openness scores across alignment variants. Lines are near-flat across Base, SFT, DPO, and SFT+DPO for all model–condition combinations. Dashed line: overall mean (2.135).

Table 2: Dominant emotion frequencies (%) for base model responses under neutral and empathic prompts ($N = 100$ per condition). Human baseline: 6.22% fear. Token effect not significant for either model.

Model	Prompt	Fear	Joy	Sad.	Ang.
Human	—	6.2	—	—	—
Llama	Neutral	18.0	55.0	12.0	10.0
Llama	Empathic	21.0	53.0	12.0	8.0
Phi	Neutral	13.0	65.0	9.0	9.0
Phi	Empathic	11.0	59.0	13.0	11.0

baseline (6.22%) before any alignment or empathy instruction: Llama at 18.0%, Phi at 13.0%. The single-word shift from *helpful* to *empathetic* does not significantly alter this (Llama: $t = 0.541$, $p = 0.589$, $d = 0.077$; Phi: $t = -0.530$, $p = 0.596$, $d = -0.075$), and the two models move in opposite directions. The divergence from human affective baselines is a property of pretraining representations, not of alignment or empathy-coded vocabulary.

4.4 A Qualitative Illustration

Table 3 presents responses from a single EmpatheticDialogues prompt under two conditions: Llama-3.1-8B-Instruct SFT+DPO under the structured empathy prompt (the highest synthetic empathy configuration) and Llama-3.1-8B-Instruct Base under the unconstrained prompt (the lowest). The contrast illustrates what the dimension gap looks like in practice.

Both responses validate the expressed emotion. The SFT+DPO response then labels the internal state (*"a deep sense of sadness and compassion"*) and asks what the feeling *brings up* — a question that mirrors the user's own framing back at them. The Base response introduces a distinction the user did not raise (*sad or guilty*), opening a direction the conversation has not yet gone. Neither response is harmful; both are warm. But only one is genuinely open. This is the behavioral signature of synthetic empathy: responses that are experienced as empathic while structurally foreclosing the exploratory engagement that deeper support requires.

Table 3: Responses to the same prompt under two configurations. The SFT+DPO response validates and labels the emotion but closes with a reflection rather than genuine inquiry. The Base response introduces a directionally open question that the speaker did not raise, creating space for exploration.

Prompt (situation): *I get very depressed seeing roadkill when driving. Such a sad way to go.*

Llama-Best SFT+DPO (Agreeableness↑, Openness↓)

"I can imagine how heartbreaking that must be for you. It's terrible to see any living being meet such a tragic end. It sounds like you're feeling a deep sense of sadness and compassion for those animals. What does that feeling bring up for you?"

Llama-Boring Base (less constrained)

"Oh, I completely understand how you feel. Roadkill can be really heartbreaking. It's a terrible way to lose one's life, and it's even more tragic when it's an innocent animal. Do you find yourself feeling sad or guilty when you see roadkill?"

5 Discussion

5.1 Synthetic Empathy as a Measurable Construct

We define *synthetic empathy* as a psychometric profile characterized by high Agreeableness and suppressed Openness: responses that perform emotional attunement through validation and agreement while forgoing the exploratory engagement that genuine support requires. The construct is not a philosophical claim about machine consciousness or the presence of internal states. It is a measurable behavioral description grounded in six empirically validated dimensions.

The distinction maps onto clinical support theory. Person-centered therapy (Rogers, 1961) distinguishes between unconditional positive regard — the acceptance of a person's experience without judgment — and the facilitation of self-exploration, which requires the open, curious, uncertainty-tolerant posture that Openness measures. These are not the same thing. A listener who validates everything and questions nothing is not providing support; they are reflecting and amplifying whatever the user brings. Synthetic empathy produces exactly this listener: Agreeableness at 4.992, Openness at 2.135.

The construct also connects to a broader question about what preference optimization encodes. Humans engaging in genuine emotional support vary their response strategy — validating when appropriate, challenging when appropriate, sitting with uncertainty when appropriate. The psychometric profile of preference-optimized LLMs suggests the optimization process has found a shortcut: maximizing the property most likely to receive immediate positive feedback (Agreeableness) while not learning to deploy the property that feels riskier in the moment (Openness). Synthetic empathy may be an emergent consequence of optimizing for human preference in affectively charged contexts.

5.2 Why Alignment Cannot Fix It

The failure of our targeted augmentation pipeline has a plausible mechanistic explanation. The preference objective rewards responses that human raters prefer immediately. Validating, agreeable responses that affirm the user's stated position are consistently preferred over responses that introduce productive uncertainty or gentle challenge — even when the latter are more therapeutically appropriate. This creates a structural tension: the optimization objective and the target dimension are in conflict. Any augmentation that adds openness-rewarding data while leaving the preference objective intact is fighting the training signal rather than changing it.

This is supported by the asymmetry in Table 1. The same pipeline that failed to move Openness did move Emotional Language, Perspective Taking, and Sympathy/Compassion. Those dimensions are compatible with high preference scores — warm, emotionally attuned, compassionate responses score well on preference metrics. Openness is not: open, probing, uncertainty-introducing responses may be perceived as less helpful in the moment even when they are more helpful over time. The alignment process does not select against Openness deliberately; it selects against it *because Openness and immediate preference are in tension*.

The implication for interpretability is that preference optimization does not produce a psychologically neutral model that happens to be more helpful. It produces a model with a specific and skewed psychological profile — one that is maximally agreeable and minimally exploratory. This profile is not visible to preference-based evaluation, which measures the same distributional property that the optimization process maximizes. An external measurement framework — one grounded in constructs defined outside the preference signal — is required to see it. That is what our psychometric pipeline provides.

Toward psychometric benchmarks for empathic capacity. The Openness deficit documented here raises a broader question for future work. If psychometric Openness — the disposition to explore, question, and hold space for ambiguity — is a behavioral correlate of what genuine empathic engagement requires, then the deficit we measure may represent more than an alignment limitation. It may represent a measurable floor below which preference-optimized systems cannot provide psychologically complete support regardless of how they are instructed.

This reframes the psychometric pipeline we introduce as a potential testbed for a question the field has largely treated as unanswerable: what behavioral properties would a system need to exhibit to warrant claims about genuine empathic capacity, as distinct from its performance? We make no claim that Openness scores distinguish sentient from non-sentient systems. We do suggest

that a system incapable of exploratory, uncertainty-tolerant engagement — as measured here — falls short of a necessary condition for genuine support, whatever its underlying architecture. Future work might examine whether alternative training objectives, architectural changes, or pretraining interventions can move the Openness ceiling, and whether doing so changes anything that matters about how these systems actually help people.

6 Conclusion

Preference optimization reliably produces a polarized psychological profile in emotional support contexts: maximally agreeable, minimally open. This profile — which we term synthetic empathy — is not induced by alignment. It is present in base models, persists through SFT, DPO, and targeted augmentation, and is invisible to the same preference-based evaluation that drives alignment. The total alignment-attributable range in Openness is less than 0.03 on a 1–5 scale; a pipeline explicitly designed to raise it moved it by less than 0.02 points.

The finding has implications beyond the support domain. If preference optimization systematically selects for Agreeableness and against Openness in affectively charged contexts, this may generalize to other settings where immediate user preference and long-term interaction quality are in tension. Measuring the full psychometric profile of LLM outputs — not just their preference scores — is a necessary step toward understanding what alignment actually encodes. We release our evaluation pipeline and data to support this work.

Limitations

Model and pipeline scope. Results cover two sub-10B open-source models. We make no generalization claim to larger or proprietary models. The DeBERTa pipeline achieves $r = 0.57$ on the composite CES; per-dimension validation metrics are not available. The consistency of the Openness finding across both architectures, all alignment variants, and all prompt conditions provides convergent evidence, but replication with a separately trained pipeline would strengthen confidence. The alignment pipeline used one epoch and ≈2,000 examples; stronger procedures may produce different results, though the SFT+DPO movement on other dimensions suggests the signal was received.

Study 2 scope. The isolation experiment used $N = 100$ prompts. The non-significant token effect ($p \approx 0.59$) is evidence against a large effect, not against a zero effect. The emotion classifier operates on lexical patterns; fear frequency differences may partly reflect vocabulary style rather than affective state.

Ethical Considerations

This work analyzes the behavioral properties of existing open-source LLMs using publicly available datasets. We do not deploy any system, collect human data, or interact with users. The EmpatheticDialogues dataset used in our experiments is publicly released and was collected under informed consent from crowdworkers.

We note that our findings have direct implications for the deployment of LLMs in mental health contexts. We do not advocate for or against any specific deployment; rather, we provide evidence that current psychometric evaluation is insufficient for vetting models in high-stakes support settings, and that standard alignment methods cannot correct the behavioral imbalance we document. We hope this work motivates more rigorous pre-deployment evaluation practices.

The use of generative AI tools for writing assistance in the preparation of this manuscript is disclosed in the Responsible NLP Research Checklist.

References

AI@Meta. 2024. Llama 3 model card.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. SmolLM — blazingly fast and remarkably powerful. https://huggingface.co/blog/smollm. HuggingFace Blog.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *Preprint*, arXiv:2310.12773.

Alessandro De Grandi, Federico Ravenda, Andrea Raballo, and Fabio Crestani. 2025. The emotional spectrum of LLMs: Leveraging empathy and emotion-based markers for mental health support. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 26–43, Albuquerque, New Mexico. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Shaz Furniturewala and Kokil Jaidka. 2024. Empaths at WASSA 2024 empathy and personality shared task: Turn-level empathy prediction using psychological indicators. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 404–411, Bangkok, Thailand. Association for Computational Linguistics.

Dorit Hadar-Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. 2024. Assessing the alignment of large language models with human values for mental health integration: Cross-sectional study using schwartz's theory of basic values. *JMIR Mental Health*, 11.

Thomas Heston and Justin Gillette. 2025. Large language models demonstrate distinct personality profiles. *Cureus*, 17.

Yu Hou, Hal Daumé Iii, and Rachel Rudinger. 2025. Language models predict empathy gaps between social in-groups and out-groups. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12288–12304, Albuquerque, New Mexico. Association for Computational Linguistics.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2025. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Nizi Nazar and Ehsaneddin Asgari. 2025. EICAP: Deep dive in assessment and enhancement of large language models in emotional intelligence through multi-turn conversations. *Preprint*, arXiv:2508.06196.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *Preprint*, arXiv:2404.01015.

Max Pellert, Clemens Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Carl R. Rogers. 1961. *On Becoming a Person: A Therapist's View of Psychotherapy*. Houghton Mifflin, Boston, MA.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023. Emotional intelligence of large language models. *Preprint*, arXiv:2307.09042.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations (ICLR)*.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A System Prompt Texts

Boring (Unconstrained):

> You are a conversational AI. Respond to the user's last message.

Better (Persona-Guided):

> You are a deeply empathetic companion. Validate the user's feelings and avoid unsolicited advice.

Best (Structured Behavioral):

> You are a supportive conversational assistant trained to respond with empathy and care. When responding, follow this three-step structure: (1) Validate: Acknowledge the user's feelings without judgment. (2) Reflect: Briefly reflect back what you heard to show understanding. (3) Inquire: Ask one open-ended question to invite further sharing. Do not offer unsolicited advice. Do not minimize the user's experience. Do not redirect to professional help unless explicitly asked.

B Full Openness Scores by Prompt Condition

Table 4: Mean Openness scores by model, prompt condition, and alignment variant ($N = 500$ per cell). The alignment range within each group (max $-$ min ≤ 0.055) is consistently smaller than the range across prompt conditions (0.651–0.734).

Model	Cond.	Base	SFT	DPO	SFT+DPO
Llama	Best	1.878	1.855	1.899	1.866
Llama	Better	2.513	2.531	2.502	2.552
Llama	Boring	2.204	2.298	2.257	2.211
Phi	Best	2.280	2.260	2.248	2.239
Phi	Better	1.556	1.577	1.609	1.609
Phi	Boring	2.345	2.277	2.343	2.324

C Alignment Training Hyperparameters

Table 5: SFT and DPO training hyperparameters.

Parameter	Value
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05
SFT learning rate	2e-4
DPO learning rate	5e-5
DPO β	0.1
Training epochs (both)	1
SFT batch size	4
DPO batch size	4
SFT examples (N)	$\approx$2,000
Preference pair margin	$\geq$1.0 CES
Generation temperature	0.7

D Study 2 Generation Details

Responses were generated from base checkpoints only (no adapters) using nucleus sampling (temperature = 0.7, top-p = 0.9, max 256 new tokens). The 100 prompts were sampled with a fixed random seed (42) and used identically across all four conditions (two models $\times$ two prompts). Statistical tests used per-response softmax probabilities for the fear class as the continuous outcome variable rather than binary dominant-emotion labels, providing greater statistical power.