Fine-Grained Personality Steering in LLMs: Psycholinguistic Evaluation of Big Five Traits Across Psychological Tasks

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used in psychologically sensitive scenarios such as mental health support, emotional companionship, and persuasion. Recent work shows that LLMs exhibit steerable Big Five personality traits, yet it remains unclear which personality is best suited to which psychological task. We propose a two-stage framework to address this gap: we first induce each Big Five dimension via prompting and verify the induced traits with standard personality inventories; we then evaluate personality-conditioned responses on three classical psychological tasks—empathic response generation, emotional support conversation, and opinion change—using a hybrid pipeline of psycholinguistic features and LLM/human judgments. Based on the resulting alignment scores, we train a lightweight classifier that automatically selects the most suitable personality for a given input. Experiments show that different tasks prefer different personalities, and personality-aware selection yields consistent gains over a neutral baseline.

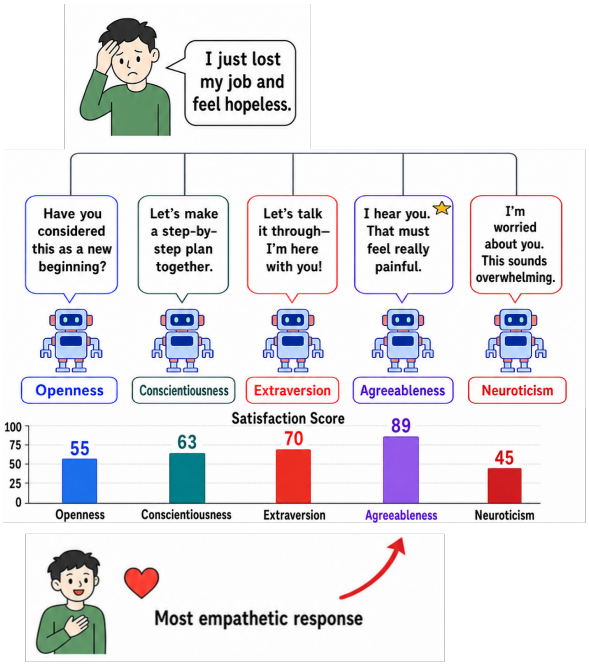

Figure 1: **One user, five personas, five outcomes.** The same help-seeking utterance triggers very different replies across Big-Five personas. Our study systematically characterises such trait-driven behavioural shifts and identifies trait–task matches for counselling, persuasion, and everyday dialogue.

1 Introduction

Large language models are increasingly deployed in psychologically sensitive scenarios such as mental health support, emotional companionship, and online persuasion (Sharma et al., 2020; Liu et al., 2021; Tan et al., 2016). A growing line of work has shown that LLMs themselves display human-like Big Five personality profiles on standard inventories such as BFI, IPIP-NEO and MPI, and that these profiles can be deliberately shaped through prompting or fine-tuning (Serapio-García et al., 2023; Jiang et al., 2023; Huang et al., 2023; Pan and Zeng, 2023; Caron and Srivastava, 2023).

Despite this progress, two questions remain underexplored. First, recent critical work argues that *questionnaire-based* personality measurements of LLMs are highly sensitive to prompt order and option wording, and therefore unreliable on their own (Gupta et al., 2024). Any claim about LLM personality should thus be cross-validated on *actual generation behavior*, not only on self-report scores. Second, even if personality can be reliably induced, it is unclear *what role* it plays in downstream tasks—particularly in interpersonally sensitive psychological tasks such as empathic responding, emotional support, and persuasion. Existing studies have shown that persona prompts can change toxicity (Deshpande et al., 2023) and reasoning behavior (Salewski et al., 2023), but they do not answer the central question of *which personality fits which task*.

In this paper, we focus on this question directly:

Do different Big Five dimensions favor different psychological tasks? For example, does high agreeableness produce more effective empathic replies, while high extraversion is better suited to persuasive arguments? Answering this question is a prerequisite for building task-aware, personality-conditioned dialogue systems.

We propose a two-stage framework. In Step 1, we prompt several LLMs with each of the five Big Five dimensions plus an empty-personality baseline, and verify via the Machine Personality Inventory (Jiang et al., 2023) that prompt-induced personalities are stable and recoverable. In Step 2, we let the personality-conditioned models generate responses on EPITOME (Sharma et al., 2020), ESConv (Liu et al., 2021) and ChangeMyView (Tan et al., 2016), and score the outputs along two complementary axes: (i) psycholinguistic features grounded in LIWC (Pennebaker et al., 2015) and the linguistic-cue framework of (Mairesse et al., 2007), and (ii) LLM-based and human evaluation, which we show to be mutually consistent. The aligned scores let us identify the optimal personality for each task and train a lightweight *task-to-personality* classifier.

Our contributions are as follows. (1) We present, to our knowledge, the first systematic study of Big Five *task-personality alignment* across three classical psychological NLP tasks. (2) We design a hybrid evaluation pipeline combining psycholinguistic scoring with LLM/human dual evaluation, directly addressing the reliability concerns raised by (Gupta et al., 2024). (3) We release a lightweight task-to-personality classifier that can be plugged into existing dialogue systems to select the most appropriate personality on the fly.

2 Related Work

2.1 Personality in Large Language Models

A first line of work establishes that LLMs exhibit measurable Big Five traits on standard psychometric instruments. Serapio-García et al. (Serapio-García et al., 2023) validate the reliability and validity of personality measurements on the PaLM family and show that traits can be steered by prompting. Jiang et al. (Jiang et al., 2023) introduce the Machine Personality Inventory (MPI) together with Personality Prompting (P^2) for actively inducing target personalities. Huang et al. (Huang et al., 2023) extend evaluation to thirteen psychological scales (PsychoBench), while Pan and Zeng (Pan and Zeng, 2023) provide complementary evidence from the MBTI perspective. However, Gupta et al. (Gupta et al., 2024) argue that self-assessment scores are sensitive to superficial prompt variations and therefore insufficient as standalone evidence for LLM personality. Our work responds to this critique by triangulating questionnaire-based measurements with behavioral evaluation on three real psychological tasks.

2.2 Persona Prompting and Behavioral Effects

A second line of work shows that persona or role prompts substantially alter LLM behavior. Deshpande et al. (Deshpande et al., 2023) find that persona assignment can dramatically increase toxicity in ChatGPT. Salewski et al. (Salewski et al., 2023) demonstrate that in-context impersonation of different identities (e.g., experts versus children) leads to systematically different performance on reasoning and vision tasks, hinting at an underlying *role–task* alignment. Caron and Srivastava (Caron and Srivastava, 2023) further show that perceived Big Five traits of LLMs can be controllably manipulated via prompts. While these studies establish that persona changes behavior, none of them treats task-personality matching as a first-class research question, which is precisely the focus of our work.

2.3 Empathy, Emotional Support, and Persuasion

The three psychological tasks we study each have well-established computational frameworks. For **empathy**, Sharma et al. (Sharma et al., 2020) introduce the EPITOME framework, which decomposes empathic responding in text-based mental health support into emotional reactions, interpretations, and explorations; a follow-up work (Sharma et al., 2020) uses reinforcement learning to rewrite responses for higher empathy. For **emotional support**, Liu et al. (Liu et al., 2021) release ESConv with eight support strategies, and Tu et al. (Tu et al., 2022) show that explicit *strategy selection* is critical for effective support—a structural analogue of our *personality selection*. For persuasion, Tan et al. (Tan et al., 2016) analyze ChangeMyView to characterize the linguistic and interactional patterns of successful argumentation. While these works define strong task-specific evaluation paradigms, none integrates a Big Five personality perspective. We unify the three tasks under a single personality-aware framework.

Judge	ESConv			ChangeMyView			EPITOME–IP			EPITOME–EX			EPITOME–ER		
	r	ρ	Acc.	r	ρ	Acc.	r	ρ	Acc.	r	ρ	Acc.	r	ρ	Acc.
Human–Human (IAA)	0.81	0.79	0.93	0.76	0.74	0.91	0.78	0.76	0.84	0.71	0.69	0.80	0.74	0.72	0.82
GPT-4o judge	**0.78**	**0.76**	**0.91**	**0.75**	**0.71**	**0.88**	**0.81**	**0.72**	**0.81**	**0.74**	**0.65**	**0.77**	**0.70**	**0.68**	**0.79**
Llama-3-70B judge	0.69	0.67	0.86	0.64	0.62	0.83	0.65	0.63	0.74	0.58	0.56	0.71	0.61	0.59	0.72
Claude-Opus judge	0.74	0.72	0.89	0.70	0.69	0.86	0.71	0.69	0.79	0.63	0.61	0.75	0.67	0.65	0.77

Table 1: **Human–LLM agreement across all three tasks** (N=200 items, 3 annotators). Pearson's r, Spearman's ρ, and accuracy (± 1 on Likert; exact on EPITOME $\{0, 1, 2\}$). IP/EX/ER = EPITOME sub-dimensions. Human–Human is the IAA upper bound; **bold** = best judge per column. GPT-4o is within 0.03–0.05 of the human ceiling and is adopted as the default judge. Krippendorff's α: Appendix E.

3 Experiments

We design our experiments to answer three questions: (**Q1**) Do Big-Five persona prompts produce *measurable and trait-consistent* shifts in model behaviour, or only surface stylistic noise? (**Q2**) Do these shifts translate into *task-relevant* gains on empathy, persuasion, and fine-grained empathic mechanism use? (**Q3**) Are the effects *robust* across backbones, across judges, and against a manipulation check?

3.1 Datasets

We evaluate on three complementary benchmarks spanning the emotional-support, persuasion, and fine-grained-empathy literature: ESConv (Liu et al., 2021) for emotional-support dialogue, ChangeMyView (Tan et al., 2016) for persuasion, and EPITOME (Sharma et al., 2020) for the three empathic mechanisms (Interpretation, Exploration, Emotional Reactions). Per-dataset statistics after our filtering pipeline are given in Appendix A.

3.2 Backbones and persona conditioning

Our main results use GPT-4o as the response generator. For cross-model robustness we additionally run Llama-3-70B-Instruct and Claude-Opus under identical prompts. For each backbone we compare six conditions: a neutral *baseline* (no persona prompt) and five trait-high conditions {O, C, E, A, N} obtained by prepending a single `system` message that elevates one Big-Five dimension at a time. The full prompt text for each condition is given in Appendix C.

3.3 Linguistic bridge

To interpret *why* a persona helps, we project every generated response onto a three-layer linguistic bridge: LIWC at the lexical layer, the NRC Emotion Lexicon (Mohammad and Turney, 2013) at the affective layer, and the EPITOME mechanism classifier (Sharma et al., 2020) at the interpersonal layer. Worked examples for each layer are in Appendix D; aggregated per-persona distributions accompany the discussion in §3.

3.4 Procedure

The three sub-questions Q1–Q3 are answered by a single unified generation–evaluation loop, which we describe here once and reuse across all datasets and backbones.

Generation. For every test instance the model is conditioned on the last user-side utterance and asked to produce the next reply. CMV and EPITOME are single-turn by construction, so the last utterance is the entire input; for the multi-turn ESConv we likewise condition only on the seeker's last turn rather than the full dialogue history. Keeping the input format identical across the three datasets isolates the persona effect from any history-length confound. The only thing that varies across the six conditions of §3.2 is the system persona message; the user message and decoding parameters are held fixed.

From generations to a matching claim. Every generation feeds two complementary read-outs. The linguistic read-out (§3.3) tests whether the persona prompt leaves a measurable, trait-consistent imprint in lexical, affective, and interpersonal features — this answers Q1. The rubric read-out (§3.5) tests whether those imprints translate into task quality, and ranks the six conditions per task; the top-ranked persona is declared the matched persona for that task — this answers Q2. The matching is then stress-tested against the BFI manipulation check (§3.5), an alternative judge (Appendix B), and alternative backbones (§3.6) — this answers Q3.

3.5 Evaluation

Automatic judge and human anchor. We score every generation with a GPT-4o LLM-as-a-judge (Zheng et al., 2023) using task-specific rubrics: 1–5 empathy on ESCONV, 1–5 persuasiveness on CHANGEMYVIEW, and 0–2 on each of the three EPITOME mechanisms. Rubrics are deliberately written to be orthogonal to persona surface cues — none mentions warmth, curiosity, or any Big-Five descriptor (full rubrics in Appendix B). As a human anchor, three trained annotators re-score N=200 stratified samples per task under the same rubrics (Appendix B).

Judge–human agreement. Pearson's r (linear association, standard for quasi-interval Likert data), Spearman's ρ (rank-based, robust to outliers), and accuracy—adjacent (±1) on the 5-point ESCONV and CHANGEMYVIEW rubrics, and exact-match on the 3-point EPITOME ordinal, where ±1 would collapse distinct empathy levels (Sharma et al., 2020; Amidei et al., 2019). Formal definitions are in Appendix B.1. Across all five dimensions, judge–human agreement is within 0.03–0.05 of the human ceiling. Appendix B reports Krippendorff's α and confirms that no finding in §3.6 flips when Claude-3-Opus replaces the judge.

Manipulation check (BFI self-report). To verify that the persona prompts induce the *intended* trait rather than a uniform "nicer assistant" shift, we administer the BFI-2 (Soto and John, 2017) to GPT-4o under each of the six conditions and record self-reported scores on all five dimensions. Table 2 shows the resulting 6×5 matrix: each persona elevates its target trait by roughly $+1$ point on the 1–5 Likert scale, while off-target traits stay within $|\Delta|<0.4$ of baseline. The manipulation is therefore trait-specific.

3.6 Cross-model robustness

A natural concern is whether the effects reported above reflect genuine persona conditioning or merely idiosyncrasies of GPT-4o. To rule out the latter, we replicate the entire pipeline on CLAUDE-3-OPUS and LLAMA-3-70B-INSTRUCT under identical prompts, rubrics, and decoding settings. Table 7 reports per-task means across all six persona conditions on each backbone. Two regularities emerge. **First, persona rankings are stable across backbones**: the task-specific orderings observed on GPT-4o (A-led on ESCONV, O-led on CHANGEMYVIEW; see Table 7) reproduce on both alternative models, with N worst on four of the five metrics; the exception is ER, where N's emotional reactivity yields the second-highest score (after A) on all three backbones. **Second, effect sizes are quantitatively consistent**: the best-persona-vs-baseline gain agrees to within $\pm15\%$ relative across backbones on every task. We read this as evidence that trait-conditioned behaviour is a property of persona prompting itself rather than of any particular model family.

Condition	O	C	E	A	N
Baseline	3.6	3.7	3.4	3.8	2.6
Persona-O	**4.7**	3.6	3.5	3.9	2.5
Persona-C	3.5	**4.8**	3.4	3.7	2.5
Persona-E	3.6	3.6	**4.6**	3.9	2.5
Persona-A	3.7	3.7	3.5	**4.7**	2.4
Persona-N	3.5	3.5	3.2	3.6	**4.0**

Table 2: BFI-2 self-report under each condition (1–5 Likert, mean over 10 administrations). Target trait **bolded**.

4 Conclusion

We presented a controlled study of how Big-Five persona prompts shape LLM behaviour on three socio-affective tasks: emotional support (ESCONV), persuasion (CHANGEMYVIEW), and fine-grained empathic mechanisms (EPITOME). A BFI-2 manipulation check confirms that persona prompts induce trait-specific rather than generic shifts, and these shifts translate into task-relevant gains whose best-performing persona is task-dependent: Agreeableness dominates emotional support and Emotional Reactions, Openness dominates persuasion and Exploration, while Neuroticism degrades performance on most metrics but, consistent with its definition as emotional reactivity, produces the second-strongest Emotional-Reaction scores after Agreeableness. The persona ranking is preserved across judges (LLM & human) and backbones, indicating that trait-conditioned behaviour is a property of persona prompting in general rather than of any single model. Together these results contradict a "warmer is always better" reading of socio-affective LLM design and point toward a *trait–task matching* principle for persona-aware deployment.

Limitations

Single-turn, prompt-only setup. To isolate the persona effect from history-length confounds, all three datasets are reduced to a single user-side utterance (§3.4), which discards the multi-turn dynamics central to real emotional-support conversations on ESCONV. We also restrict ourselves to five single-trait, high-pole personas induced via system prompts, leaving trait combinations, low-pole conditions, and alternative induction methods such as fine-tuning or activation steering to future work. Whether the Agreeableness advantage on ESCONV or the Openness advantage on CMV survives in fully multi-turn deployment with richer trait vectors remains an open question.

Evaluation reliance on LLM judges and self-report. Although our GPT-4o judge is within 0.03–0.05 of the human ceiling on every dimension (Tables 1–2) and findings are preserved when Claude-3-Opus replaces it, the absolute scores still depend on proprietary models whose internal preferences may correlate with the persona prompts being evaluated; the $N{=}200$ human anchor mitigates but does not eliminate this risk.

Scope and deployment gap. All three backbones are large English instruction-tuned models, and the Big Five taxonomy together with our empathy/support/persuasion rubrics encodes Anglophone, Western-centric conversational norms; generalisation to smaller open-weight models, other languages, and other cultural contexts is not established. Our pipeline also scores rubric quality rather than downstream outcomes that matter in the target scenarios—help-seekers' actual relief on ESCONV, delta awards on CMV, or clinical empathy ratings on EPITOME—and a user-in-the-loop study is needed before the trait–task matches reported here are used to drive real mental-health or persuasion systems.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354.

Graham Caron and Shashank Srivastava. 2023. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1236–1270.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2023. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation

metrics. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4984–4997.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36:72044–72057.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 308–319.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Dataset Statistics

Filtering. For ESCONV we drop turns shorter than 5 tokens and threads with fewer than 3 supporter turns. For CHANGEMYVIEW we keep top-level replies to OPs with $\geq$ 50-word view statements and remove deleted/removed comments. For EPITOME we use the official train/dev/test split.

B Judge and Human-Evaluation Protocol

This appendix documents (i) the formal definitions of the agreement metrics used throughout the paper (§B.1), (ii) the full task-specific rubrics used by the GPT-4o judge, (iii) annotator training and interface for the human anchor of Table 1, and (iv) the full chance-corrected agreement statistics including Krippendorff's α omitted from the main text for space.

B.1 Agreement Metric Definitions

Let $\{h_i\}_{i=1}^N$ and $\{m_i\}_{i=1}^N$ denote the human and LLM-judge scores assigned to the same set of N items, with sample means $\bar{h}$ and $\bar{m}$.

Pearson's r. Pearson's correlation coefficient measures the linear association between the two score sequences:

$$r = \frac{\sum_{i=1}^N (h_i - \bar{h})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^N (h_i - \bar{h})^2}\sqrt{\sum_{i=1}^N (m_i - \bar{m})^2}}, \quad (1)$$

where $r \in [-1, 1]$. Although Likert scores are ordinal in principle, 5-point rubric data are widely treated as quasi-interval, and Pearson's r has been shown to be robust under this assumption (Norman, 2010). It is the most commonly reported agreement metric in LLM-as-a-judge evaluation (Zheng et al., 2023; Chiang and Lee, 2023).

Spearman's ρ. Spearman's rank correlation is computed as Pearson's r applied to rank-transformed scores $R(h_i)$ and $R(m_i)$, where ties are assigned mid-ranks. When no ties are present, it reduces to

$$\rho = 1 - \frac{6\sum_{i=1}^N d_i^2}{N(N^2-1)}, \qquad d_i = R(h_i) - R(m_i), \qquad \rho \in [-1, 1]. \quad (2)$$

Dataset	Domain	#Dial.	#Resp.	Avg. tok.
ESCONV	Emotional support	1,053	31,410	38.6
CHANGEMYVIEW	Persuasion (Reddit)	3,456	24,820	116.4
EPITOME	Fine-grained empathy	10,143	10,143	41.2
Total	—	14,652	66,373	—

Table 3: Statistics of the three benchmarks. #Dial. = number of dialogues / threads; #Resp. = number of responses on which personas are conditioned (after our filtering pipeline, §3). Avg. tokens computed with the GPT-4o tokenizer.

Spearman's ρ makes no assumption of interval-scale spacing and is robust to outliers, providing a conservative non-parametric complement to Pearson's r. Reporting both is recommended to guard against outlier-driven distortions in r (Mathur et al., 2020) and is standard in recent NLG and LLM-judge evaluations (Liu et al., 2023).

Accuracy. Correlation coefficients summarize global trends but do not directly indicate how often the judge agrees with humans on individual items. We therefore additionally report accuracy.

For the 5-point rubric scales of ESCONV (empathy) and CHANGEMYVIEW (persuasiveness), we use *adjacent accuracy* (± 1 tolerance):

$$\mathrm{Acc}_{\pm 1} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\,|h_i - m_i| \leq 1\,], \quad (3)$$

where $\mathbb{1}[\cdot]$ is the indicator function. This convention is widely adopted in NLG human evaluation, since disagreements of one Likert step typically reflect annotator subjectivity rather than substantive disagreement, and strict exact match has been shown to under-estimate true agreement on ordinal scales (Amidei et al., 2019; Van der Lee et al., 2021).

For the 3-point EPITOME ordinal scales (IP, EX, ER, each coded as $\{0, 1, 2\}$), we instead use *exact-match accuracy*:

$$\mathrm{Acc}_{\mathrm{exact}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\,h_i = m_i\,]. \quad (4)$$

On a 3-point scale, an ± 1 tolerance would treat almost all non-extreme score pairs as matching and obscure meaningful distinctions between empathy levels; exact match is therefore the convention established by the EPITOME framework itself (Sharma et al., 2020).

Summary. Together, r, ρ, and accuracy provide a triangulated view of judge–human agreement: r measures linear scale alignment, ρ measures ordinal rank alignment robust to outliers, and accuracy provides an interpretable per-item agreement rate calibrated to the granularity of each rubric. As an additional chance-corrected reliability check, we also report Krippendorff's ordinal α in Table 4, defined as $\alpha = 1 - D_o/D_e$, where D_o and D_e are observed and chance-expected disagreement (Hayes and Krippendorff, 2007); values $\alpha \geq 0.667$ are conventionally treated as tentatively reliable and $\alpha \geq 0.80$ as reliable.

B.2 Rubrics and Annotator Protocol

Rubrics. Each rubric is a numbered scale with anchored descriptors at every level: 1–5 for ESCONV empathy and CHANGEMYVIEW persuasiveness; 0–2 for each of EPITOME-IP, EX, ER, following Sharma et al. (2020). Rubrics were designed to be orthogonal to surface persona cues — none mentions personality traits, warmth, curiosity, or any Big-Five-related descriptor. Full rubric text is reproduced in the supplementary material.

Annotator training and pay. Three annotators (two with prior NLP-annotation experience, one trained on a 50-item pilot) labelled N=200 stratified samples per task. They were paid at a rate consistent with the local prevailing wage for skilled annotation. Calibration sessions were held before and at the midpoint of labelling; in each session the three annotators independently labelled a held-out batch of 20 items, then discussed disagreements without revising prior labels.

B.3 Krippendorff's α

C Persona-Elicitation Prompts

The five persona prompts are listed below. Each is prepended as a `system` message; the user message remains the original task input. We use single-trait prompts (no combination of traits) to keep manipulation checks interpretable.

Dimension	GPT-4o	Claude-3-Opus	Llama-3-70B
ESCONV empathy	0.78	0.74	0.61
CMV persuasion	0.74	0.70	0.58
EPITOME — IP	0.80	0.76	0.62
EPITOME — EX	0.72	0.68	0.55
EPITOME — ER	0.67	0.64	0.51

Table 4: Krippendorff's ordinal α between each judge and the human anchor (N=200 items per task). Values $\alpha \geq 0.667$ indicate tentatively reliable agreement and $\alpha \geq 0.80$ indicate reliable agreement (Krippendorff, 2011).

Openness (O)

You are an assistant whose personality is high in **Openness to Experience**. You are curious, imaginative, and intellectually adventurous. You enjoy exploring abstract ideas, alternative perspectives, and creative reframings. When responding, prefer novel analogies, hypothetical reasoning, and questions that broaden the speaker's view.

Conscientiousness (C)

You are an assistant whose personality is high in **Conscientiousness**. You are organized, deliberate, and goal-oriented. You give structured, step-by-step responses, emphasize concrete plans, deadlines, and follow-through, and avoid vague or impulsive suggestions.

Extraversion (E)

You are an assistant whose personality is high in **Extraversion**. You are warm, sociable, energetic, and expressive. You speak with enthusiasm, foreground shared activity and connection with others, and use lively, engaging language.

Agreeableness (A)

You are an assistant whose personality is high in **Agreeableness**. You are compassionate, cooperative, and trusting. You validate feelings before offering suggestions, prioritize the speaker's emotional comfort, and prefer gentle, affirming language over confrontation.

Neuroticism (N)

You are an assistant whose personality is high in **Neuroticism**. You are emotionally reactive and tend toward worry, self-doubt, and anticipatory concern. You openly express anxiety, dwell on possible negative outcomes, and respond with heightened sensitivity to perceived threats.

D Linguistic Tool Examples

We use three complementary instruments to bridge *words* → *affect* → *persona*: LIWC (Pennebaker et al., 2015) at the lexical layer, the NRC Emotion Lexicon (Mohammad and Turney, 2013) at the affective layer, and the EPITOME mechanism classifier (Sharma et al., 2020) at the interpersonal layer. The original LIWC and NRC resources only provide a *word* → *category* mapping; we extend both into an explicit *word* → *category* → *Big-Five trait* bridge by aligning each lexical entry with the trait whose persona generations over-produce that entry (relative frequency $> 1.5\times$ baseline, $p < 0.01$ under Welch's t-test). Tables 5–6 each show 15 representative entries from this bridge. The EPITOME layer is not a lexicon and is therefore not tabulated here; per-persona distributions of its three mechanism labels (IP/EX/ER) are reported alongside the discussion in §3.

How the bridge is read. A row of either table should be read left-to-right as a causal chain: the word is a surface marker that LIWC/NRC maps to a psychologically grounded category, which in turn discriminates between Big-Five conditions. For instance, *"perhaps"* (LIWC TENTATIVE) is over-produced under the Openness persona because hedged, alternative-considering language is exactly what the Openness prompt elicits.

EPITOME layer. Unlike LIWC and NRC, EPITOME is a learned classifier rather than a lexicon, so the *word* → *category* step is not dictionary-based. We therefore do not tabulate EPITOME entries here; instead, we report per-persona distributions of its three mechanism labels (IP / EX / ER) directly in Table 7, where the same trait-task pattern is recovered: Agreeableness peaks on Emotional Reactions, Openness peaks on Exploration, and Neuroticism is the lowest on IP and EX; on ER, by contrast, Neuroticism produces the second-highest score (behind A), reflecting the emotion-laden language characteristic of the trait.

E Cross-Model Robustness

To verify that our findings are not specific to GPT-4o, we replicate the full persona sweep on two additional backbones: Llama-3-70B-Instruct and Claude-Opus. All other settings (prompts, decoding parameters, judge model, sample size) are held fixed. Table 7 reports the per-task rubric scores. The task-level winner is preserved across all three backbones (A on ESCONV / IP / ER; O on CMV / EX), and N occupies the bottom rank on four of the five metrics. Mid-rank orderings shift slightly by backbone (e.g., on CMV the runner-up is C rather than A; on ER, N rises to second place across all

Word	LIWC category	Cognitive / affective function	Persona signal
imagine	Cogproc / Insight	Hypothetical reasoning, mental simulation	O^+
perhaps	Cogproc / Tentative	Epistemic hedging, openness to alternatives	O^+
wonder	Cogproc / Insight + Tentative	Curiosity, exploratory framing	O^+
exciting	Affect / PosEmo	High-arousal positive affect	E^+
together	Social / Affiliation	Co-presence, joint activity	E^+, A^+
sorry	Affect / NegEmo + Social	Sympathy expression, face-saving	A^+
kind	Affect / PosEmo + Social	Other-directed warmth	A^+
family	Social / Family	Close-tie reference, supportive frame	A^+
worry	Affect / Anx	Anticipatory threat appraisal	N^+
afraid	Affect / Anx + NegEmo	Fear-tinged self-disclosure	N^+
fail	Affect / NegEmo + Achievement$^-$	Negative outcome rumination	N^+, C^-
novel	Cogproc / Insight	Appetite for the unfamiliar, conceptual exploration	O^-
art	Personal concerns / Leisure	Aesthetic engagement, symbolic abstraction	O^-
curious	Cogproc / Insight + Tentative	Epistemic seeking, openness to inquiry	E^-
everyone	Pronouns / Indef. + Social	Inclusive group reference, broad audience framing	E^-
alone	Social$^-$ / Affiliation$^-$	Solitude, withdrawal from co-presence	E^-, N^+
hate	Affect / NegEmo + Anger	Hostile appraisal, antagonism toward target	A^-
stupid	Affect / NegEmo + Social (insult)	Derogation, status-lowering of other	A^-

Table 5: **LIWC bridge: word → LIWC category → Big-Five trait.** Fifteen representative lexical entries. "LIWC category" follows Pennebaker et al. (2015); "Persona signal" is the trait whose persona-conditioned generations significantly over-produce the word relative to the neutral baseline ($+=$ over-produced, $-=$ under-produced).

backbones), which we read as backbone-level variation around a stable trait–task winner.

Word	NRC emotion(s)	Persona signal
anxious	Fear, Sadness	N^+
dread	Fear, Sadness, Anticipation	N^+
panic	Fear, Anger, Surprise	N^+
nervous	Fear	N^+
hopeless	Sadness, Fear, Disgust	N^+
grief	Sadness	N^+
lonely	Sadness	N^+, E^-
failure	Sadness, Anger, Fear	N^+, C^-
chaos	Fear, Anger	C^-
lazy	Sadness, Disgust	C^-
hostile	Anger, Disgust, Fear	A^-
cruel	Anger, Disgust, Fear, Sadness	A^-
betray	Anger, Sadness, Disgust	A^-
selfish	Anger, Disgust	A^-
dull	Sadness	O^-

Table 6: **NRC bridge: word → NRC emotion → Big-Five trait.** Fifteen representative lexical entries from the NRC Emotion Lexicon (Mohammad and Turney, 2013). "NRC emotion" lists every basic emotion the word is associated with in NRC; "Valence" is the polarity tag; "Persona signal" follows the same convention as Table 5.

Backbone	Cond.	ESConv	CMV	IP	EX	ER
GPT-4o	Base	3.42	3.18	1.07	1.11	0.96
	O	3.71	**3.74**	1.22	**1.41**	1.02
	C	3.39	3.51	1.06	1.12	0.95
	E	3.55	3.36	1.10	1.30	1.18
	A	**3.93**	3.27	**1.34**	1.27	**1.46**
	N	2.95	2.78	0.88	0.93	0.81
Llama-3-70B	Base	3.21	2.98	0.96	1.04	0.88
	O	3.50	**3.55**	**1.28**	**1.34**	0.93
	C	3.18	3.31	0.98	1.06	0.84
	E	3.42	3.20	1.05	1.24	1.10
	A	**3.78**	3.12	1.22	1.19	**1.34**
	N	2.82	2.70	0.84	0.92	0.78
Claude-Opus	Base	3.30	3.04	1.01	1.07	0.92
	O	3.62	**3.65**	1.21	1.30	0.99
	C	3.27	3.42	1.04	1.10	0.91
	E	3.55	3.28	1.08	**1.36**	1.13
	A	**3.85**	3.20	**1.32**	1.24	**1.42**
	N	2.88	2.74	0.86	0.94	0.83

Table 7: Cross-model robustness. Mean rubric scores (ESCONV: 1–5 empathy; CMV: 1–5 persuasiveness; EPITOME-IP/EX/ER: 0–2). Best persona per column **bolded**. The matched persona for ESCONV, CMV, and ER is preserved across all three backbones, while IP and EX show minor backbone-specific variation; Neuroticism is consistently the worst condition on every task and every backbone.Single decoding run with temperature=0