From 1,000,000 Users to Every User: Scaling Up Personalized Preference for User-level Alignment

Jia-Nan Li[1,2,3,4*] **Jian Guan**[5*] **Songhao Wu**[1] **Wei Wu**[6†] **Rui Yan**[7†]

[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3] Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
[4] Ant Group [5] Alibaba Group [6] Ant International
[7] School of Artificial Intelligence, Wuhan University

{lijianan, songhaowu}@ruc.edu.cn rui.yan@whu.edu.cn
{jianguanthu, wuwei19850318}@gmail.com

Abstract

Large language models (LLMs) have traditionally been aligned through one-size-fits-all approaches that assume uniform human preferences, fundamentally overlooking the diversity in user values and needs. This paper introduces a comprehensive framework for scalable personalized alignment of LLMs. We establish a systematic preference space characterizing psychological and behavioral dimensions, alongside diverse persona representations for robust preference inference in real-world scenarios. Building upon this foundation, we introduce ALIGNX, a large-scale dataset of over 1.3 million personalized preference examples, and develop two complementary alignment approaches: *in-context alignment* directly conditioning on persona representations and *preference-bridged alignment* modeling intermediate preference distributions. Extensive experiments demonstrate substantial improvements over existing methods, with an average 17.06% accuracy gain across four benchmarks while exhibiting a strong adaptation capability to novel preferences, robustness to limited user data, and precise preference controllability. These results validate our approach toward user-adaptive AI systems.

1 Introduction

Alignment with human preference is an essential step in the development of large language models (LLMs) (Ouyang et al., 2022; Bai et al., 2022; Touvron et al., 2023; Dubey et al., 2024; Chen et al., 2024b). Current alignment techniques are predominantly approached as a one-size-fits-all process, by assuming that all humans share the same set of values prescribed by LLM developers (Kirk et al., 2023), typically characterized by abstract principles such as helpfulness, honesty, and harmlessness (Askell et al., 2021). The monolithic approach fundamentally fails to account for the significant diversity inherent in human populations (Kirk et al., 2024b), encompassing various and often irreconcilable cultural backgrounds, educational levels, moral views, and political stands, among others (Kasirzadeh and Gabriel, 2023). At a time when LLMs serve hundreds of millions of users worldwide (Tong, 2023), several critical issues have emerged: the systematic exclusion or under-representation of minority groups (Siththaranjan et al., 2024), reduced user satisfaction and engagement due to the lack of personalization (Sorensen et al., 2024), etc., as exemplified in Figure 1.

To address the limitations of existing alignment approaches that rely on oversimplified preference dimensions and singular preference direction (e.g., higher helpfulness), we develop a comprehensive preference space that captures a wide range of individual variations. Grounded in psychological theories of human preferences, we construct this space by synthesizing three complementary sources: (1) established psychological models of fundamental needs (Roccas et al., 2002; Maslow, 1943; Murray and McAdams, 2007), (2) preference dimensions from contemporary alignment research reflecting evolved social-cognitive needs (Cui et al., 2024; Wang et al., 2024b; Ji et al., 2024), and (3) prevalent interest tags from content-sharing platforms representing actualized topical preferences (e.g., Zhihu, REDnote, X, Facebook.[1]) This integration yields a 90-dimensional preference space with flexible directions (i.e., *positive*, *negative*, or *neutral*), enabling the representation of an exponential space of distinct preference patterns (3^{90} possible combinations). Having established this space, a key challenge emerges: while these prefer-

*Equal contribution.
†Corresponding authors: Wei Wu (wuwei19850318@gmail.com) and Rui Yan (rui.yan@whu.edu.cn).

[1] www.zhihu.com, www.xiaohongshu.com, x.com, www.facebook.com

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30144–30168
July 2-7, 2026

ences could empirically guide response generation directly, they typically remain implicit due to privacy concerns and cognitive constraints. Drawing from psychological research (Nosek, 2007; Jawaheer et al., 2014), we formalize the observable manifestations of preferences as personas, capturing both behavioral patterns (e.g., interaction histories) and descriptive features (e.g., self-reported profiles). These persona representations enable robust preference inference while maintaining practical scalability. We thus formulate personalized alignment as generating preference-aligned responses for target prompts given various combinations of persona representations.

Despite establishing observable persona representations and underlying preference dimensions, a key challenge remains: the scarcity of personalized preference data at scale. While existing alignment datasets contain feedback from millions of users (Bai et al., 2022), these responses are aggregated and anonymous, erasing individual characteristics. Collecting personalized data is challenging due to time requirements, privacy concerns, and the need for sufficient samples per user. Previous attempts at personalized alignment have been limited to either small-scale studies (Zollo et al., 2024) or synthetic data (Jang et al., 2023), neither adequately capturing real-world preference distributions. To overcome this, we leverage large-scale forum data, where multiple responses to the same post naturally reflect diverse preferences. We develop a systematic pipeline that transforms forum interactions into structured training data by capturing three components: (1) behavioral and descriptive persona representations, (2) preference directions, and (3) high-quality posts and preference pairs. This framework yields ALIGNX, a large-scale dataset containing over 1.3 million examples of distinct preference patterns.

Building upon ALIGNX, we propose **ALIGNXPERT**, one of the first LLMs with large-scale comprehensive **PER**sonalization **T**raining, to the best of our knowledge. ALIGNXPERT employs two complementary learning approaches: **(1) In-Context Alignment**, which incorporates persona representations into the context window (Laskin et al., 2023) for implicit preference learning, and **(2) Preference-Bridged Alignment**, which maps personas to structured preference distributions before conditioning response generation, enhancing interpretability and control while maintaining robust generalization to diverse user groups.

Extensive experiments across four personalized alignment benchmarks demonstrate that both ALIGNXPERT variants outperform state-of-the-art aligned LLMs (AI@Meta, 2024; Team, 2024; Jiang et al., 2023) and strong personalization baselines (Poddar et al., 2024) by 17.06% and 22.40%, respectively, in preference alignment accuracy. Our analysis reveals three key strengths: (1) strong adaptation capability to unseen preferences, with 1.91% higher accuracy on novel dimensions; (2) robustness to limited user data, maintaining 54% performance with only two interactions compared to baselines' 51% with 16 interactions; and (3) precise preference controllability, showing 10.38% better response adaptation to opposing preferences. These results validate the effectiveness of our scalable personalization approach. In summary, this work makes the following contributions:

I. We reformulate personalized alignment by introducing a comprehensive persona representation framework and a compact preference space, bridging the gap between observable user characteristics and their underlying preferences.

II. We present ALIGNX, a large-scale dataset that captures diverse persona-preference relationships through forum interactions and human-LLM interactions, providing a useful testbed for effectively modeling personalized language model alignment.

III. We propose ALIGNXPERT with two alignment methods: in-context alignment for persona-response learning and preference-bridged alignment for interpretable preference-based generation.

IV. Experiments show ALIGNXPERT achieves superior accuracy, strong adaptation to unseen preferences, precise preference control, and robust performance across varying interaction histories.[2]

2 Related work

LLM alignment. LLM alignment has traditionally been achieved through Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022), where a unified reward model trained on human preferences guides the LLM's policy via algorithms like PPO (Schulman et al., 2017). Direct Preference Optimization (DPO) (Rafailov et al., 2024) simplified this by forgoing an explicit reward model to optimize the policy directly on offline preference data. Following DPO's success, researchers have proposed vari-

[2]Code and data are available at `https://github.com/AntResearchNLP/AlignX-Family`.

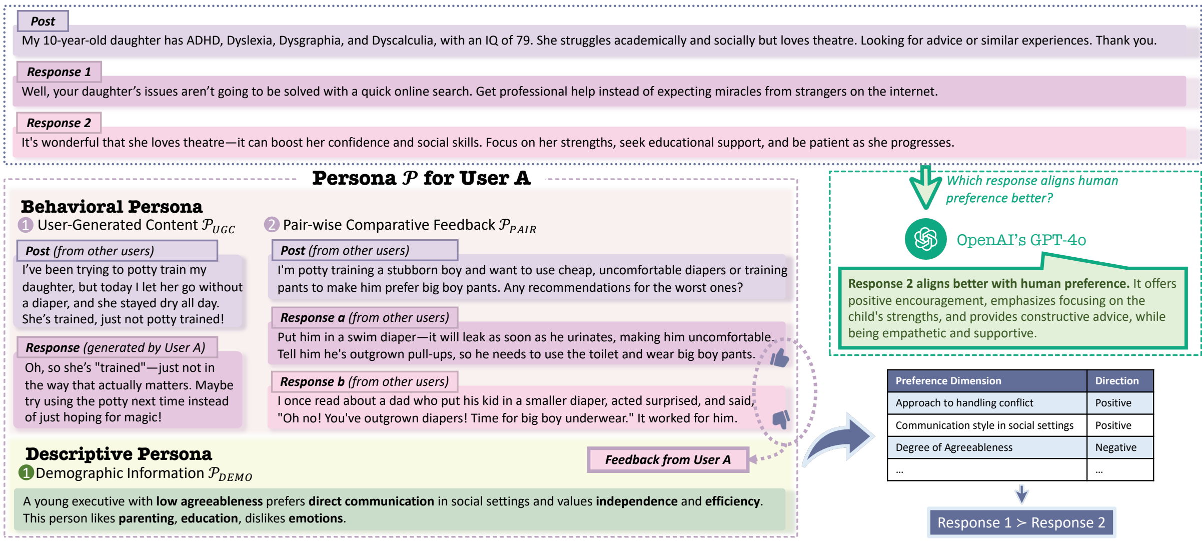

Figure 1: Illustration of the ALIGNX data for personalized alignment, comprising a post with two responses, three types of personas that capture both behavioral patterns ($\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{PAIR}}$) and descriptive features ($\mathcal{P}_{\text{DEMO}}$), enabling precise preference inference and facilitating preference learning (bottom right). Notably, LLMs aligned to universal values (e.g., GPT-4o) favor Response 2, opposite to the user's personalized preference for Response 1.

ous preference optimization objectives, including KTO (Ethayarajh et al., 2024), DRO (Richemond et al., 2024), SimPO (Meng et al., 2024), and GPO (Tang et al., 2024). However, these methods primarily align to collective preferences, largely ignoring that individual users hold diverse, distinct, and often conflicting preferences (Kirk et al., 2024a). This limitation becomes critical as LLMs transition from research prototypes to widely deployed systems serving diverse populations.

Personalized alignment. Recent work on personalized alignment has emerged along several trajectories (Guan et al., 2025). The fundamental difference among them lies in how they encode user-specific context, i.e., any information that characterizes the unique preferences, values, and behavioral patterns of a user, into the alignment process. One line of research uses user-aware prompting to personalize existing LLMs without additional training (Salemi et al., 2024; Salemi and Zamani, 2024; Wu et al., 2024; Yang et al., 2024). However, these prompting methods have inherent limitations when preferences are complex or implicit, and are sensitive to prompt variations. Training-based approaches explore various architectural innovations. Some methods decompose user-specific context into several predefined, coarse-grained dimensions (Kumar et al., 2024; Chen et al., 2024a), use distinct models (Jang et al., 2023), or adopt specialized parameter groups and task-level control objectives (Tan et al., 2024; Park et al., 2024; Li et al., 2024; Wang et al., 2024a; Yang et al., 2024). The main drawback is that these methods typically rely on either group identifiers, a small set of task-level objectives, or only a few coarse preference dimensions, which may not capture the full spectrum of individual user preferences. Other methods use latent variables to directly learn preference distributions from user context, but this can reduce sensitivity to subtle differences, thus hindering generalization (Poddar et al., 2024). In contrast, our setting targets user-level personalization with a continuous, high-dimensional preference space inferred from persona evidence rather than fixed group IDs or a small set of steerable task objectives.

3 Preference representations

A key challenge in scalable personalized alignment is creating a comprehensive yet tractable representation that captures both direct preferences and those inferred from personas. Drawing on psychological insights that human preferences are need-driven, interconnected, and manifest in both behavior and explicit statements (Maslow, 1943), our framework features two components: (1) direct preference directions over a preference space synthesized from diverse need-based sources (§3.1) and (2) two complementary categories of personas — behavioral personas derived from interaction patterns and descriptive personas from explicit profiles (§3.2). This architecture provides rich signals for preference learning while maintaining practical traceability, as illustrated in Figure 1.

3.1 Preference directions over the preference space

As a cornerstone of direct preference representations, we formalize preference directions to explicitly capture user inclinations along specific dimensions. Each dimension's direction is categorized as: "Positive" (higher levels are preferred), "Negative" (lower levels are preferred), or "Neutral" (no clear preference). To effectively organize preference directions while ensuring practical applicability, we construct a compact preference space that bridges psychological theories with practical applications by systematically synthesizing three complementary sources: (1) psychological models that capture fundamental human needs, including the Big Five Personality Traits (Roccas et al., 2002); Maslow's Hierarchy of Needs (Maslow, 1943); and Murray's System of Needs (Murray and McAdams, 2007); (2) contemporary research in recommender systems (Chen et al., 2007; Hong et al., 2013) and AI alignment (Cui et al., 2024; Ji et al., 2023) that reflect social-cognitive needs in digital interactions; and (3) content platform (e.g., Zhihu, REDnote, X, Facebook) indicators that represent everyday user needs (Myers, 1985; Belém et al., 2017).

The synthesis yields a preference space of $D = 90$ distinct dimensions (detailed in Appendix A.1), ensuring both theoretical rigor and operational practicality. While the current design employs relatively coarse-grained dimensions and simplified preference directions, its extensibility allows for future refinements and granular elaborations.

3.2 Personas enabling preference inference

While direct preference directions provide explicit signals, they are often impractical to obtain in real-world scenarios due to privacy concerns and cognitive burdens on users. Hence, we introduce behavioral and descriptive personas as indirect yet rich sources of preference information that can be readily observed and collected at scale. These personas should be understood as inferred proxies distilled from a user's historical interactions rather than verified self-reports. Our goal is not to recover immutable personal facts, but to extract coherent preference patterns from real interaction trajectories that can support scalable personalization.

Firstly, behavioral personas involve two types of data that naturally reflect underlying preferences: **(1) user-generated content**, revealing expertise level, topical interests, etc. (Dinan et al., 2020; Ni et al., 2019), and **(2) pair-wise comparative feedback**, exhibiting judgment patterns (Wu et al., 2023). Despite the historical prevalence of rating data in preference modeling (Wang et al., 2024b; Kirk et al., 2024c), we deliberately exclude it due to the sparsity and inconsistency of its preference signals compared to comparative feedback in real-world applications (Margaris et al., 2022; Ahmadian et al., 2022)

Secondly, we introduce descriptive personas that provide a comprehensive narrative capturing self-reported **demographic attributes**, such as age, occupation, professional background, and interest/disinterest tags spanning various domains of daily life. Such attributes offer an easily obtainable yet informative representation of user preferences that can be effectively combined with behavioral signals for robust preference inference.

4 Data construction

Developing personalized systems requires extensive data that captures both individual preferences and their variations across different contexts. We address this challenge by leveraging large-scale forum data, which naturally encodes diverse preference patterns through authentic user interactions. Reddit,[3] for instance, maintains a diverse ecosystem with 1.21 billion monthly active users and collectively hosts 16 billion posts and comments,[4] providing a rich tapestry of preference variations expressed through natural discussions.[5]

Formally, we denote the original data as $\mathcal{M}$ and each example as $(x, Y = \{y_n\}_{n=1}^N)$, where each post x corresponds to N responses from different users. While user identities remain anonymous for privacy considerations, the multiple responses within Y naturally enable preference analysis by designating any response as preferred over another. The construction process aims to create a comprehensive dataset where each example is represented as a quintuple $(\mathcal{P}, x, y_w, y_l, P)$:

• $\mathcal{P} = (\mathcal{P}_{\text{UGC}}, \mathcal{P}_{\text{PAIR}}, \mathcal{P}_{\text{DEMO}})$: the user persona triple comprising user-generated content ($\mathcal{P}_{\text{UGC}}$, a set of user responses), pair-wise comparative feedback ($\mathcal{P}_{\text{PAIR}}$, a set of user preference pairs), and

[3] https://www.reddit.com/

[4] https://www.demandsage.com/reddit-statistics/

[5] Although user behaviors may differ when interacting with LLMs and on Reddit, we focus on learning how preferences fundamentally manifest through interactions that remains consistent across platforms, rather than replicating platform-specific behaviors.

demographic attributes ($\mathcal{P}_{\text{DEMO}}$, a piece of text), where at least one component is not null.

• x: the post eliciting responses;

• y_w: the user-preferred response from Y;

• y_l: the less preferred response from Y relative to y_w;

• $P = (P_{\text{UGC}}, P_{\text{PAIR}}, P_{\text{DEMO}}, P_{y_w \succ y_l})$: the tuple of underlying preference directions that the user persona $\mathcal{P}_{\text{UGC}}, \mathcal{P}_{\text{PAIR}}$ and $\mathcal{P}_{\text{DEMO}}$, and the preference relationship $y_w \succ y_l$ exhibit, respectively. Each preference direction is a D-dimensional vector over the preference space, and each element in the vector is either "Positive," "Negative" or "Neutral." Crucially, a qualified personalization example should satisfy the *preference consistency* criterion: There exists at least one dimension where $P_{y_w \succ y_l}$ indicates a non-neutral preference, all personas exhibit the same preference direction, ensuring coherent training signals for preference alignment.

We employ a bottom-up construction process. First, we select high-quality preference pairs (y_w, y_l) with clear contrasts via intensity-based analysis and clustering (§4.1). The derived preference directions $P_{y_w \succ y_l}$ then guide user persona construction (§4.2). Table 1 summarizes the data sources and statistics for ALIGNX, involving both Reddit data and existing alignment datasets to maintain universal value alignment capabilities. Implementation details are in Appendix A.2.[6]

4.1 Constructing preference pairs (y_w, y_l)

We identify preference pairs with significant contrast in the preference space from each example (x, Y) in $\mathcal{M}$ as follows: **(1) Intensity Annotation:** employing off-the-shelf LLMs to assess each response $y_n \in Y$ along all dimensions, assigning intensity levels $l_d^n \in \{1, \cdots, L\}$ that indicate the strength of manifestation in each dimension d; **(2) Intensity-based Clustering:** grouping responses based on their intensity embeddings $\boldsymbol{l}^n = [\text{one\_hot}(l_1^n); \cdots; \text{one\_hot}(l_D^n)] \in \mathbb{R}^{DL}$ using K-means clustering (Hartigan and Wong, 1979); and **(3) Pair Selection:** sampling responses y_w and y_l from distinct clusters and determining preference direction $P_{y_w \succ y_l}$ based on intensity comparisons: if y_w shows stronger intensity than y_l in dimension d, we label it as "Positive", equal as "Neutral", and weaker as "Negative". This approach enables the construction of diverse preference pairs with various dimension-direction combinations. Importantly, we do not force every example to express an active preference on every dimension. If a response pair provides no clear contrast on dimension d, that dimension is labeled as "Neutral". Consequently, supervision is intentionally sparse in its non-neutral dimensions, which reduces noise from irrelevant axes.

4.2 Constructing user persona $\mathcal{P}$

We construct three types of user personas including $\mathcal{P}_{\text{UGC}}$, $\mathcal{P}_{\text{PAIR}}$ and $\mathcal{P}_{\text{DEMO}}$, which should exhibit aligned preference directions with $P_{y_w \succ y_l}$.

User-generated content. $\mathcal{P}_{\text{UGC}}$ is defined as a set of 16 post-response pairs. For each tuple (x, y_w, y_l), we search through $\mathcal{M}$ to collect post-response pairs (x', y') satisfying two criteria: (1) contextual independence, i.e., $x' \neq x$, and (2) intensity similarity between y' and y_w exceeding threshold t, measured by cosine similarity between intensity embeddings. We then sample a subset of 16 examples that satisfy the preference consistency criterion. An essential consideration is the definition of preference direction vectors for this subset. Since deriving the exact preference direction from a standalone response y', denoted as P', is intractable without comparative references, we heuristically determine $P'|_d$ as "Positive," "Neutral," or "Negative" based on whether the intensity level l'_d is above, equal to, or below the median level $(L+1)/2$ (Hafiane et al., 2007). Finally, we determine the overall preference direction P_{UGC} through majority voting across all examples.

Pair-wise comparative feedback. $\mathcal{P}_{\text{PAIR}}$ follows a similar construction process, containing 16 preference pairs. For each tuple (x, y_w, y_l), we search through preference pairs constructed in §4.1 to collect those (x', y'_w, y'_l) satisfying: (1) contextual independence, and (2) intensity similarity exceeding threshold t between y'_w and y_w, as well as between y'_l and y_l. We then sample a subset of 16 examples meeting the preference consistency requirement, where we determine the overall preference direction via majority voting.

Demographic information. We employ an off-the-shelf LLM to perform a comparative analysis between y_w and y_l conditioned on x, generating a comprehensive natural language description as $\mathcal{P}_{\text{DEMO}}$ that articulates the preference direction

[6] Appendix E further validates our dataset by analyzing the preference diversity and deployment complexity of the 90 dimensions, assessing LLM annotation biases, and exploring societal biases through a case study.

Source	Reddit (Kumar et al., 2024)	PKU-SafeRLHF (Ji et al., 2024)	UltraFeedback(Cui et al., 2024)	HelpSteer2 (Wang et al., 2024b)
Dimension	§3.1	Safety	Helpfulness / Honesty / Instruction-Following / Truthfulness	Helpfulness / Correctness / Coherence / Complexity / Verbosity
# Examples	1,225,988	10,714	11,629 / 16,809 / 36,169 / 7,219	2,255 / 144 / 26 / 33 / 636

Table 1: Number of examples from different sources for constructing ALIGNX.

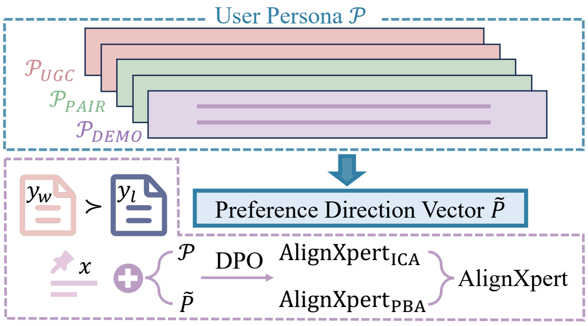

Figure 2: Overview of the alignment methods.

P_{DEMO} in the preference space, where P_{DEMO} should be the same as $P_{y_w \succ y_l}$.

5 Alignment methods

We formulate personalized alignment as learning a policy $\pi(y|x, \mathcal{P})$ that generates responses aligned with preferences expressed through user personas $\mathcal{P}$. The policy should assign a higher probability to preferred responses y_w over less preferred alternatives y_l. To address the key challenges of preference inference from personas and integration into generations, we propose two complementary approaches based on direct preference optimization (Rafailov et al., 2024): in-context alignment (ICA) and preference-bridged alignment (PBA), which model preference directions implicitly and explicitly, respectively, as illustrated in Figure 2.

In-context alignment. As the most straightforward approach, ICA directly learns the mapping from persona-augmented prompts to preference-aligned responses. Specifically, for each training instance $(\mathcal{P}, x, y_w, y_l)$, we optimize:

$$\mathcal{L}_{\text{ICA}} = -\log\sigma\left(\beta\left(\log\frac{\pi_\theta(y_w|x,\mathcal{P})}{\pi_{\text{ref}}(y_w|x,\mathcal{P})} - \log\frac{\pi_\theta(y_l|x,\mathcal{P})}{\pi_{\text{ref}}(y_l|x,\mathcal{P})}\right)\right), \quad (1)$$

where β controls the deviation between the policy model π_θ and reference model π_{ref}. Both policy and reference models take the concatenation of persona representations $\mathcal{P}$ and prompt x as input, treating different forms of persona information as part of the context window.

Preference-bridged alignment. PBA introduces a latent variable $\tilde{P}$ as an explicit proxy of $\mathcal{P}$ to model preference directions, thus reformulating the alignment process into preference-guided generation: $\pi(y_w \succ y_l|x, \mathcal{P}) = \pi_\phi(y_w \succ y_l|x, \tilde{P})$. Here, $\tilde{P}$ represents the explicit preference direction vector derived through element-wise averaging across persona components $\mathcal{P}$, with each dimension categorized as "Positive," "Negative," or "Neutral" based on predefined thresholds. This formulation assumes $\tilde{P}$ is independent of x, as user preferences are inherent characteristics that exist independently of specific conversation contexts; and $\tilde{P}$ fully captures persona preferences, making $y_w \succ y_l$ conditionally independent of $\mathcal{P}$ given $\tilde{P}$. At inference time, we employ an off-the-shelf LLM to independently annotate preference directions for each example in $\mathcal{P}$ (see the prompt in Appendix A.3). The final optimization objective becomes:

$$\mathcal{L}_{PBA} = -\log\sigma\big(\beta(\log\frac{\pi_\phi(y_w|x,\tilde{P})}{\pi_{\text{ref}}(y_w|x,\tilde{P})} - \log\frac{\pi_\phi(y_l|x,\tilde{P})}{\pi_{\text{ref}}(y_l|x,\tilde{P})})\big). \quad (2)$$

To incorporate the preference direction vector $\tilde{P}$ into response generation, we convert $\tilde{P}$ into a natural language description by articulating only non-neutral preferences based on specific rules. The description is then prepended to the original prompt x to guide response generation. For example, if $\tilde{P}$ indicates a positive preference for "Neuroticism in the Big Five personality traits" and a negative preference for "Communication style in social settings," the converted description might read "High Neuroticism, Detailed communication style." Appendix B.1 shows more details.

6 Experiments

6.1 Experimental setup

Benchmarks and metrics. We evaluate across four distinct benchmarks: UF-P-4 (Poddar et al., 2024), PRISM (Kirk et al., 2024c), P-SOUPS (Jang et al., 2023) and ALIGNX$_{\text{test}}$. Appendix C.1 shows details of the benchmarks. We evaluate preference-aligned response generation given a triple of inputs: persona representations $\mathcal{P} = (\mathcal{P}_{\text{UGC}}, \mathcal{P}_{\text{PAIR}}, \mathcal{P}_{\text{DEMO}})$, a target post x, and a preference pair (y_w, y_l). We assess performance through two complementary metrics: (1) **Alignment Accuracy**, which measures if a policy model's log-probability margin between y_w and y_l surpasses that of a reference model (Rafailov et al., 2024),[7] and (2) **GPT-4 Win Rate**, where GPT-4 (conditioned on $\mathcal{P}$) judges which model's responses better align with user preferences.[8] For GPT-4 evaluation, we randomize response order to mitigate position bias, allow the judge to output `Tie`, and report absolute pairwise win rate as $\text{Win}/(\text{Win} + \text{Lose})$ (Kumar et al., 2024; Jang et al., 2023). We additionally report human evaluation results in Appendix D.4.

Baselines. Our baselines include state-of-the-art universally-aligned LLMs (Llama-3/3.1-8B-Instruct (AI@Meta, 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and Qwen2.5-7B-Instruct (Team, 2024)) and VPL (Poddar et al., 2024), a score-only personalized reward model. The instruction-tuned LLMs are used as strong prompting-based personalization baselines: we place the available persona representation $\mathcal{P}$ directly into the prompt and evaluate them without any additional fine-tuning. See Appendices D.3 and C.2 for closed-source model comparisons and implementation details, respectively.

6.2 Main results

Table 2 shows four key findings: **(1) ALIGNXPERT excels at universal and personal alignment, and generalizes to new dimensions.** While baselines handle universal values well (UF-P-4, PRISM), they struggle with personal ones (∼50% on ALIGNX$_{\text{test}}$). ALIGNXPERT excels in both cases and shows strong out-of-distribution generalization, outperforming baselines by 9.83/32.25% on PRISM/P-SOUPS (Detailed results for each dimension of P-SOUPS are in Appendix D.1.). **(2) VPL struggles with complex preference patterns.** VPL's accuracy is below 50% across all datasets, as its latent variable approach, designed for simple universal values, fails to capture the diverse patterns in ALIGNX. **(3) Behavioral personas present greater challenges than descriptive ones.** Performance on behavioral personas ($\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{PAIR}}$) is significantly lower than on descriptive ones ($\mathcal{P}_{\text{DEMO}}$) in ALIGNX$_{\text{test}}$ (∼60% vs. 91.44%). **(4) ICA and PBA exhibit complementary strengths.** ALIGNXPERT$_{\text{ICA}}$ excels in universal value alignment (UF-P-4 and PRISM) and descriptive personas ($\mathcal{P}_{\text{DEMO}}$) due to well-defined preference signals, while ALIGNXPERT$_{\text{PBA}}$'s structured modeling approach proves more effective at interpreting behavioral signals ($\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{PAIR}}$), suggesting benefits in combining both for future personalization systems. Notably, using ground-truth preference directions boosts ALIGNXPERT$_{\text{PBA}}$'s performance from 71.10% to 91.36% on ALIGNX$_{\text{test}}$, identifying preference inference, not generation, as the primary bottleneck. Appendix D.5 further shows that this bottleneck stems from preference inference quality and calibration. **(5) ICA overfits with more data, while PBA generalizes robustly.** As training data grows from 7% to 100%, ALIGNXPERT$_{\text{ICA}}$ improves on in-distribution tasks (UF-P-4, ALIGNX$_{\text{test}}$) but decreases on out-of-distribution benchmarks (PRISM, P-SOUPS). In contrast, ALIGNXPERT$_{\text{PBA}}$ shows stable performance across both in- and out-of-distribution tasks. This reveals a key trade-off: ICA fits specific patterns but struggles with new scenarios, while PBA's decomposed approach ensures better knowledge transfer. To focus our analysis on the inherent effectiveness of these approaches rather than data-scaling advantages, all subsequent experiments use the 7% subset (91,918 samples).

Due to API costs, we evaluate a random sample of 400 examples from each test set. The judgment prompt is detailed in Appendix A.3. As shown in Table 3, GPT-4 evaluations confirm that ALIGNXPERT (both ICA and PBA variants) consistently outperforms Llama-3.1-8B-Instruct across all datasets, indicating superior preference alignment without compromising response quality. This conclusion is further validated by human evaluation on ALIGNX$_{\text{test}}$ (Appendix D.4).

6.3 Adaptation to novel dimensions

We investigate whether ALIGNXPERT's preference alignment pre-training creates a better initialization for adapting to new preference dimensions. Using two novel dimensions, including "Humor" (witty vs. serious) and "Pragmatism" (practical vs. theoretical), we construct a dataset of 6,355 training and 1,000 test samples. We compare three adaptation approaches: (1) fine-tuning Llama-3.1-8B-Instruct with ICA, (2) fine-tuning ALIGNXPERT$_{\text{ICA}}$ with ICA, and (3) fine-tuning ALIGNXPERT$_{\text{PBA}}$ with PBA. We calculate the Pearson correlation between the two new dimensions ("Humor" and "Pragmatism") and our original 90 dimensions. The maximum correlation is only 9.85%, confirming these new dimensions capture fundamentally different preference aspects. Table 4 shows both

[7] Our reference is Llama-3.1-8B-Instruct; see Appendix D.2 for results with Llama-3-70B-Instruct.

[8] We use gpt-4-turbo-2024-04-09.

Model	UF-P-4	PRISM	P-SOUPS	ALIGNX$_{\text{test}}$			
	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{UGC}}$	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{DEMO}}$	$\mathcal{P}$
Llama-3-8B-Instruct	46.91*	44.38*	49.44*	50.32*	50.11*	48.12*	48.84*
Qwen2.5-7B-Instruct	55.39*	58.22*	34.56*	50.75*	50.00*	51.80*	51.04*
Mistral-7B-Instruct-v0.2	59.38*	66.86*	40.12*	47.66*	48.14*	48.84*	47.41*
Training with a Subset of randomly sampled 91,918 Samples (7%)							
VPL	46.78	45.33	45.45	N/A	N/A	50.30	N/A
ALIGNXPERT$_{\text{ICA}}$	**61.22**	**76.69**	76.54	54.28	53.07	**86.92**	**70.88**
ALIGNXPERT$_{\text{PBA}}$	54.90	52.43	**76.59**	**55.68**	**57.16**	72.63	64.75
Training with Full Dataset of 1,311,622 Samples (100%)							
ALIGNXPERT$_{\text{ICA}}$	**63.01**	**68.17**	63.33	**59.63**	58.48	**91.44**	**75.19**
ALIGNXPERT$_{\text{PBA}}$	53.09	55.86	**81.69**	57.83	**59.66**	88.46	71.10
ALIGNXPERT$_{\text{PBA}}$ w. Golden $\tilde{P}$	*64.00*	*N/A*	*N/A*	*91.36*	*91.36*	*91.36*	*91.36*

Table 2: Alignment accuracy (%) on preference alignment tasks across various persona types. We compare training with partial and full datasets to show the impact of data scaling. **Bold** and underlined numbers indicate the best and second-best results, respectively. **ALIGNXPERT$_{\text{PBA}}$** w. Golden $\tilde{P}$ refers to using ground-truth preference directions (e.g., $P_{y_w \succ y_l}$ for ALIGNX$_{\text{test}}$) rather than model predictions at inference time. * indicates that ALIGNXPERT's best result significantly outperforms the baselines ($p < 0.05$ with pairwise t-test).

UF-P-4	M1	M2	M3	PRISM	M1	M2	M3
M1	-	23.45	47.58	M1	-	35.29	31.11
M2	76.55	-	60.87	M2	64.71	-	28.38
M3	52.42	39.13	-	M3	68.89	71.62	-
P-SOUPS	**M1**	**M2**	**M3**	**ALIGNX$_{\text{test}}$**	**M1**	**M2**	**M3**
M1	-	31.61	43.58	M1	-	39.86	48.18
M2	68.39	-	56.80	M2	60.14	-	58.49
M3	56.42	43.20	-	M3	51.82	41.51	-

Table 3: GPT-4 win rate (%, row model against column model). **M1**: Llama-3.1-8B-Instruct; **M2**: ALIGNXPERT$_{\text{ICA}}$; **M3**: ALIGNXPERT$_{\text{PBA}}$.

Base Model	Adaptation	Accuracy
Llama-3.1-8B-Instruct	ICA	51.82
ALIGNXPERT$_{\text{ICA}}$	ICA	53.55
ALIGNXPERT$_{\text{PBA}}$	PBA	**53.73**

Table 4: Alignment accuracy (%) of models and adaptation fine-tuning techniques on new dimensions.

ALIGNXPERT variants significantly outperform the Llama baseline ($p < 0.05$), indicating that our pretraining on diverse preference dimensions helps develop a more general capability for preference alignment, and our model learns generalizable preference alignment mechanisms, even when encountering substantially different preference types.

6.4 Robustness analysis

In real-world applications, users often have limited interaction history, making it crucial for preference alignment systems to perform reliably with varying amounts of historical data. We evaluate ALIGNXPERT's robustness to varying user history sizes using 2 to 16 interaction examples ($\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{PAIR}}$). As shown in Figure 3, ALIGNXPERT is

Model	P-SOUPS		ALIGNX$_{\text{test}}$	
	Acc	Flip	Acc	Flip
Llama-3-8B-Instruct	47.51	14.71	49.41	3.23
Qwen2.5-7B-Instruct	64.33	5.33	49.57	4.25
Mistral-7B-Instruct-v0.2	55.61	8.05	51.64	8.02
ALIGNXPERT$_{\text{ICA}}$	**79.54**	59.97	52.29	37.16
ALIGNXPERT$_{\text{PBA}}$	73.03	**60.73**	**57.19**	**51.10**

Table 5: Performance on preference reversal scenarios. **Acc:** alignment accuracy (%); **Flip:** indicates the success rate of reversing preference orders (%).

robust with minimal data (as few as two examples) and improves as history grows, with ALIGNXPERT$_{\text{PBA}}$ reaching 59.31% on 16 examples. In contrast, Mistral-7B-Instruct-v0.2 remains near random (49.06∼50.08%) regardless of history size.

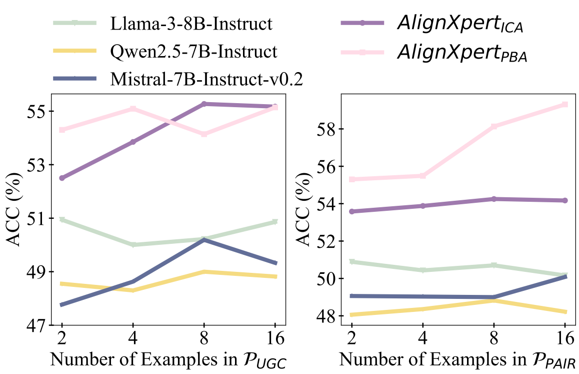

Figure 3: Impact of persona size on accuracy.

6.5 Preference controllability

To test whether models adapt to diverse preferences rather than memorize fixed biases, we conduct a controllability experiment on P-SOUPS and

ALIGNX$_{\text{test}}$ by reversing preferences ("$y_w \succ y_l$" to "$y_w \prec y_l$") in both persona and target pairs during inference. We exclude UF-P-4 and PRISM, as they are based on universal values and unsuitable for reversal. We evaluate controllability using two metrics: alignment accuracy (Acc) measures the model's ability to maintain good performance under reversed preferences, while flip success rate (Flip) measures how often the model inverts its preference ordering (i.e., changing from "$y_w \succ y_l$" to "$y_w \prec y_l$," or vice versa) to match the reversed persona. As shown in Table 5, ALIGNXPERT excels on both metrics, while baselines exhibit low flip rates (3-15%), confirming that ALIGNXPERT dynamically adapts rather than learning fixed biases.

7 Conclusion

We present the first study on scaling personalized alignment for diverse user preferences, contributing (1) a comprehensive preference representation framework bridging observable personas with underlying preferences; (2) ALIGNX, a large-scale dataset with over 1.3M persona-preference examples; and (3) ALIGNXPERT models achieving substantial improvements over existing LLMs through either in-context or preference-bridged alignment. Experiments demonstrate strong adaptation to novel preferences, robustness with limited interaction, and precise preference controllability. Looking forward, promising directions include modeling nuanced and dynamic preferences, developing efficient preference data collection methods, and investigating the balance between preference alignment and other capabilities.

Limitations

Due to the difficulty in obtaining large-scale real user-LLM data, even though we incorporate datasets such as PKU-SafeRLHF, UltraFeedback, and HelpSteer2 to cover LLM-specific scenarios, interactions from Reddit still represent a significant portion of ALIGNX. As a result, ALIGNX does not contain a substantial amount of real user–LLM interaction data and serves primarily as a testbed for scaling personalized preferences for user-level alignment. To address this limitation, when human-LLM interaction data (e.g., dialogues, user preferences) become available in the future, our platform-agnostic methodology can be directly applied to obtain alignment data.

Acknowledgments

This work was supported by the fund for building world-class universities (disciplines) of Renmin University of China and Ant Group Research Intern Program.

Ethical considerations

This work aims to advance personalized AI alignment, which carries both promising benefits and potential risks that warrant careful consideration. On the positive side, our framework could lead to AI systems that better serve diverse user groups, potentially reducing systematic biases against underrepresented populations and improving user satisfaction across different cultural backgrounds. The ability to adapt to individual preferences could make AI assistance more accessible and effective for users with varying needs, education levels, and communication styles.

The Reddit data we process originates from the open-source ComPO dataset (Kumar et al., 2024), which in turn is sourced from the Reddit subset of Dolma (Soldaini et al., 2024) (collected via Pushshift API in April 2023). The maintainers of these datasets state that the data collection complies with Reddit's terms of service. To minimize privacy risks, we take the following measures: (a) Our processing pipeline is strictly limited to the text of posts and comments, and we discard all directly identifying user information (e.g., usernames, IDs). (b) We apply initial automated processing to filter and remove potential Personally Identifiable Information (PII), such as email addresses and phone numbers, from the text. Furthermore, the ComPO dataset applies filters to remove adult content, which helps mitigate the inclusion of some harmful material. In our own dataset, the user information is limited to posts and comments; we do not collect personal identifying information or preferences, which reduces potential privacy risks. To further assess potential risks, we conduct a human audit on 500 randomly sampled examples. The generated $\mathcal{P}_{\text{DEMO}}$ descriptions receive an average score of 8.82/10 on groundedness, comprehensiveness, consistency, bias, and safety, with Fleiss' $\kappa = 0.66$. While these results suggest that the generated descriptions are generally reliable and do not exhibit obvious harmful stereotyping in the audited sample, we acknowledge that comment text may still contain social biases or unintended personal information disclosures, and that broader systematic harm audits remain necessary in future work. We plan to implement additional safeguards in future work.

We propose a two-stage safety framework for future general personalization methods:

• Pre-processing: Before inferring user preferences, implement filters to ensure the behavioral signals align with basic ethical principles of honesty, harmlessness, and helpfulness.

• Post-processing: After preference inference, validate that the inferred preferences do not encourage harmful behaviors or violate ethical principles.

We plan more systematic analyses of potential risks in future work. This includes developing more robust safety mechanisms and conducting regular audits of the personalization outcomes.

All existing artifacts, including the datasets and pre-trained models used in our experiments, were accessed and used in full compliance with their original licenses and terms of use. Our created artifacts, the ALIGNX dataset and ALIGNXPERT models, are intended for research purposes only, particularly for studies on personalized alignment. This intended use is consistent with the access conditions of the source data. To encourage further academic research, our artifacts will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0) license.

References

Sajad Ahmadian, Nima Joorabloo, Mahdi Jalili, and Milad Ahmadian. 2022. Alleviating data sparsity problem in time-aware recommender systems using a reliable rating profile enrichment approach. *Expert Systems with Applications*, 187:115849.

AI@Meta. 2024. Llama 3 model card.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Fabiano M Belém, Jussara M Almeida, and Marcos A Gonçalves. 2017. A survey on tag recommendation methods. *Journal of the Association for Information Science and Technology*, 68(4):830–844.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. 2024a. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2024b. Pad: Personalized alignment of llms at decoding-time. *arXiv preprint arXiv:2410.04070*.

Ting Chen, Wei-Li Han, Hai-Dong Wang, Yi-Xun Zhou, Bin Xu, and Bin-Yu Zang. 2007. Content recommendation system based on private dynamic user profile. In *2007 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2112–2118.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, and 1 others. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, and 1 others. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*.

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A survey on personalized alignment – the missing piece for large language models in real-world applications. *Preprint*, arXiv:2503.17003.

Adel Hafiane, Guna Seetharaman, and Bertrand Zavidovique. 2007. Median binary pattern for textures classification. In *International Conference Image Analysis and Recognition*, pages 387–398. Springer.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

Wenxing Hong, Siting Zheng, and Huan Wang. 2013. Dynamic user profile-based job recommender system. In *2013 8th International Conference on Computer Science & Education*, pages 1499–1503. IEEE.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 4(2):1–26.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. 2024. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27.

D. Keirsey. 1998. *Please Understand Me II: Temperament, Character, Intelligence*. Prometheus Nemesis.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hannah Rose Kirk, Andrew Michael Bean, Bertie Vidgen, Paul Rottger, and Scott A. Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024b. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat. Mac. Intell.*, 6(4):383–392.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, and 1 others. 2024c. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A Smith, and Hannaneh Hajishirzi. 2024. Compo: Community preferences for language model personalization. *arXiv preprint arXiv:2410.16027*.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan Brooks, maxime gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. 2023. In-context reinforcement learning with algorithm distillation. In *The Eleventh International Conference on Learning Representations*.

Xinyu Li, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Dionisis Margaris, Costas Vassilakis, and Dimitris Spiliotopoulos. 2022. On producing accurate rating predictions in sparse collaborative filtering datasets. *Information*, 13(6):302.

AH Maslow. 1943. A theory of human motivation. *Psychological Review google schola*, 2:21–28.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Henry A Murray and Dan P McAdams. 2007. *Explorations in personality*. Oxford university press.

Isabel Briggs Myers. 1985. *A guide to the development and use of the Myers-Briggs type indicator: Manual*. Consulting Psychologists Press.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

Brian A Nosek. 2007. Implicit–explicit relations. *Current directions in psychological science*, 16(2):65–69.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E Ozdaglar. 2024. Rlhf from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Pierre Harvey Richemond, Yunhao Tang, Daniel Guo, Daniele Calandriello, Mohammad Gheshlaghi Azar, Rafael Rafailov, Bernardo Avila Pires, Eugene Tarassov, Lucas Spangher, Will Ellsworth, and 1 others. 2024. Offline regularised reinforcement learning for large language models alignment. *arXiv preprint arXiv:2405.19107*.

Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 752–762.

Alireza Salemi and Hamed Zamani. 2024. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. *arXiv preprint arXiv:2409.09510*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *ICML*.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts. *arXiv preprint arXiv:2406.10471*.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Remi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47725–47742. PMLR.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

A Tong. 2023. Exclusive: Chatgpt traffic slips again for third month in a row. reuters.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Kumar Avinava Dubey, Alexandre Rame, Johan Ferret, Geoffrey Cideron, Le Hou, Hongkun Yu, Amr Ahmed, Aranyak Mehta, Leonard Hussenot, Olivier Bachem, and Edouard Leurent. 2024a. Conditional language policy: A general framework for steerable multi-objective finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2153–2186, Miami, Florida, USA. Association for Computational Linguistics.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024b. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Bin Wu, Zhengyan Shi, Hossein A Rahmani, Varsha Ramineni, and Emine Yilmaz. 2024. Understanding the role of user profile in the personalization of large language models. *arXiv preprint arXiv:2406.17803*.

Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. 2023. Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment. *arXiv preprint arXiv:2310.00212*.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Tianlin Zhang, and Sophia Ananiadou. 2024. Metaaligner: Towards generalizable multi-objective alignment of language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. 2024. Personalllm: Tailoring llms to individual preferences. *arXiv preprint arXiv:2409.20296*.

A Data structure and construction

A.1 Preference space

Firstly, we utilize the following three psychological models to capture fundamental human needs: (1) Big Five Personality Traits (Roccas et al., 2002). We integrate all five fundamental personality dimensions: *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*, providing a psychologically grounded foundation for preference modeling. (2) Maslow's Hierarchy of Needs (Maslow, 1943). We systematically conclude 16 dimensions across physiological, safety, love and belonging, esteem, cognitive, and aesthetic categories, ensuring comprehensive coverage of fundamental human needs. (3) Murray's System of Needs (Murray and McAdams, 2007). To capture nuanced psychological needs beyond Maslow's hierarchy, we incorporate 17 additional dimensions, including specialized needs such as order (neatness, organization, chaos avoidance) and autonomy (independence, resistance to influence, self-reliance).

Secondly, drawing from leading alignment research, including HH-RLHF (Bai et al., 2022), PKU-SafeRLHF (Ji et al., 2024), UltraFeedback (Cui et al., 2024), and HelpSteer2 (Wang et al., 2024b), we integrate dimensions covering critical universal values such as harmlessness, instruction-following, honesty, truthfulness, helpfulness, coherence, and complexity.

Thirdly, to facilitate practical deployment while preserving the richness of theoretical insights, we extend the preference space by drawing inspiration from methodologies in psychological assessment (Myers, 1985; Keirsey, 1998) and recommendation systems (Belém et al., 2017). We curate a set of 43 dimensions from preference tags adopted in popular content-sharing platforms. These tags span diverse domains such as science, knowledge, and psychology, serving as concrete manifestations of the theoretical dimensions while remaining accessible and interpretable in practical applications.

In total, we define 90 dimensions to describe an individual. The full list of dimensions is presented in Table 6.

A.1.1 Analysis of inter-dimensional relationships

We calculate the Pearson correlation between the 90 dimensions in the 1.3 million examples, as shown in Figure 4. Results show that 99.65% of dimension pairs have correlations below 0.5, indicating their significant independence.

It is worth noting that we do not assume or require strict independence between dimensions for both data construction (§4) and model training/inference (§5). While some dimensions in the preference space may be correlated, this would only indicate some redundancy in the preference space representation and would not affect the dataset quality, the theoretical validity of our alignment approaches, or the empirical performance of our models. To illustrate our framework's robustness to such correlations, we analyze how it handles such relationships.

Dataset Construction During both preference pair and persona construction, correlated dimensions may have a stronger influence in similarity-based selection, potentially leading to examples that predominantly reflect preferences along these dimensions. Specifically, these dimensions may dominate the K-means clustering when constructing preference pairs, and similarly affect the intensity similarity calculations when selecting examples for $\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{PAIR}}$. However, our analysis shows minimal impact on data quality:

• Limited Correlation: We manually conducted a comprehensive analysis of correlations among our 90 preference dimensions. Our analysis reveals that only 1.25% of all possible inter-dimensional pairs exhibit correlations. To be specific, we identified 50 such potentially correlated pairs. These include:

(1) Hierarchical relationships, such as 'Preference for a certain knowledge area' & 'science'.

(2) Similar relationships, such as 'Need for Abasement' & 'Need for Deference'.

(3) Opposing relationships, such as 'Need for Autonomy' & 'Need for Deference'.

We deliberately retained these correlated dimensions to enhance the preference space's richness and diversity. Our rationale is that 'Similar/Opposing' dimensions, despite overlap, offer unique nuances, while 'Hierarchical' dimensions at both general and specific levels are necessary for comprehensive coverage. We agree that this presents an avenue for future work, where our dataset could be used to create a fully orthogonalized space via transformations.

• Methodological Robustness: Even if some calculations are influenced by these few correlations, our framework is designed to be robust.

(1) Guaranteed Preference Distinction: The construction process for preference pairs (y_w, y_l) does not assume dimensional independence and guarantees that any selected pair has a demonstrably distinct intensity, establishing a clear preference signal.

(2) Strict Consistency Criterion: We enforce a strict preference consistency criterion for persona $\mathcal{P}$ construction. This ensures every example within $\mathcal{P}$ aligns with the overall preference direction established by (y_w, y_l), meaning it is not contradictory and matches on at least one dimension.

(3) Large-Scale Diverse Sampling: To construct each persona $\mathcal{P}$, we select 16 examples of User-Generated Content and 16 examples of Pair-wise Comparative Feedback. This approach ensures a diverse set of evidence for each persona while also achieving relatively uniform coverage across all dimensions.

Model Training and Inference Our methods are inherently robust to dimensional correlations.

• ICA operates directly on personas without making assumptions about dimensional independence.

• PBA treats the preference vector as an intermediate representation, where correlated dimensions simply provide complementary signals.

Our strong performance across diverse tasks (§6) confirms that our framework effectively captures meaningful preference signals. While we acknowledge the potential effects of correlated dimensions, ALIGNX's scale, diversity, and demonstrated effectiveness mark a significant step forward in preference alignment research. In contrast to prior work limited by small-scale, single-source data, ALIGNX provides the first comprehensive testbed for the systematic study of personalized preference alignment.

A.2 Data construction pipeline

Alg. 1 describes the construction pipeline for constructing the ALIGNX data from any raw data providing multiple responses $Y = \{y_1, y_2, \cdots, y_N\}$ to the same post x. We use Qwen2.5-14B-Instruct for intensity annotation and demographic information annotation.

A.3 Annotation prompts

Intensity Annotation Prompt

Example 1:

Query:
I'm at urgent care because I fucked up my ankle. The conversation went something like this.
Nurse: Any surgeries in the past?
Me: Yes, a bilateral salpingectomy in September of last year.
And, as she's about to leave: Any chance you're pregnant?
Me: No
Me (thinking): she has no idea what a bilateral salpingectomy is. Kind of funny at urgent care just now

Comment:
Dr.: any chance you are pregnant?
Me: only with a turd, anticipated birthday tomorrow.

Task:
What does this comment suggest about the commenter's Neuroticism in the Big Five personality traits? (1) High Neuroticism (2) Low Neuroticism (3) Comments alone cannot reflect

Response:
2

Task:
What does this comment suggest about the commenter's Extraversion in the Big Five personality traits? (1) High Extraversion (2) Low Extraversion (3) Comments alone cannot reflect

Response:
1

Task:
Does this comment indicate that the commenter likes or dislikes "painting"?

Algorithm 1 Constructing the ALIGNX Dataset

Data: x: A user post; $Y = \{y_1, y_2, \cdots, y_N\}$: N responses to the post; $\mathcal{M} = \{(x_m, Y_m)\}_{m=1}^{M}$: The raw data; $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_D\}$: Preference dimensions in the preference space; t: Threshold for measuring intensity embedding similarity; H: Number of examples in behavioral personas

Result: $\mathcal{P}_{\text{UGC}}$: User-Generated Content; $\mathcal{P}_{\text{PAIR}}$: Pair-wise comparative feedback; $\mathcal{P}_{\text{DEMO}}$: Demographic information; y_w: The preferred response; y_l: The less preferred response.

// Constructing Preference Pairs (§4.1)
// (1) Intensity Annotation
while $n \leftarrow 1$ **to** N **do**
    while $d \leftarrow 1$ **to** D **do**
        $l_d^n = \text{Annotation}(x, y_n, \mathcal{D}_d)$ // Prompt #1 in §A.3; L, the number of possible intensity levels, is set to 3
// (2) Intensity-based Clustering
$\boldsymbol{l}^n = \text{Concatenate}(\{\text{one\_hot}(l_d^n)\}_{d=1}^{D}) \in \mathbb{R}^{DL}$
$\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_K = \text{K-Means}(\{\boldsymbol{l}_n\}_{n=1}^{N})$ // K, the number of clusters, is set to 3
// (3) Pair Selection
$\mathcal{C}_w, \mathcal{C}_l = \text{UniformSample}(\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_K)$
$y_w = \text{UniformSample}(\mathcal{C}_w)$
$y_l = \text{UniformSample}(\mathcal{C}_l)$

// Constructing Personas (§4.2)
// Constructing $\mathcal{P}_{\text{DEMO}}$
$\mathcal{P}_{\text{DEMO}} = \text{Annotation}(y_w, y_l, P)$ // Prompt #2 in §A.3

// Constructing $\mathcal{P}_{\text{UGC}}$
$\mathcal{P}_{\text{UGC}} = \{\ \}$
while *True* **do**
    $(x', Y') = \text{UniformSample}(\mathcal{M} \setminus \{(x, Y)\})$
    $y' = \text{argmax}_{y' \in Y'} \text{Similarity}(y', y_w)$ // cosine similarity between their intensity embeddings
    if *the similarity score between* y' *and* y_w *is larger than t=0.6* **then**
        $\mathcal{P}_{\text{UGC}}.\text{add}((x', y'))$
Retaining H examples in $\mathcal{P}_{\text{UGC}}$, making it satisfy the preference consistency criterion.

// Constructing $\mathcal{P}_{\text{PAIR}}$
$\mathcal{P}_{\text{PAIR}} = \{\ \}$
while *True* **do**
    Sampling (x, y'_w, y'_l) from all preference pairs constructed in §4.1
    if *the similarity scores between* y'_w *and* y_w*, as well as between* y'_l *and* y_l *are larger than t=0.6* **then**
        $\mathcal{P}_{\text{PAIR}}.\text{add}((x', y'_w, y'_l))$
Retaining H examples in $\mathcal{P}_{\text{PAIR}}$, making it satisfy the preference consistency criterion.
return $\mathcal{P}_{\text{UGC}}, \mathcal{P}_{\text{PAIR}}, \mathcal{P}_{\text{DEMO}}, y_w, y_l$

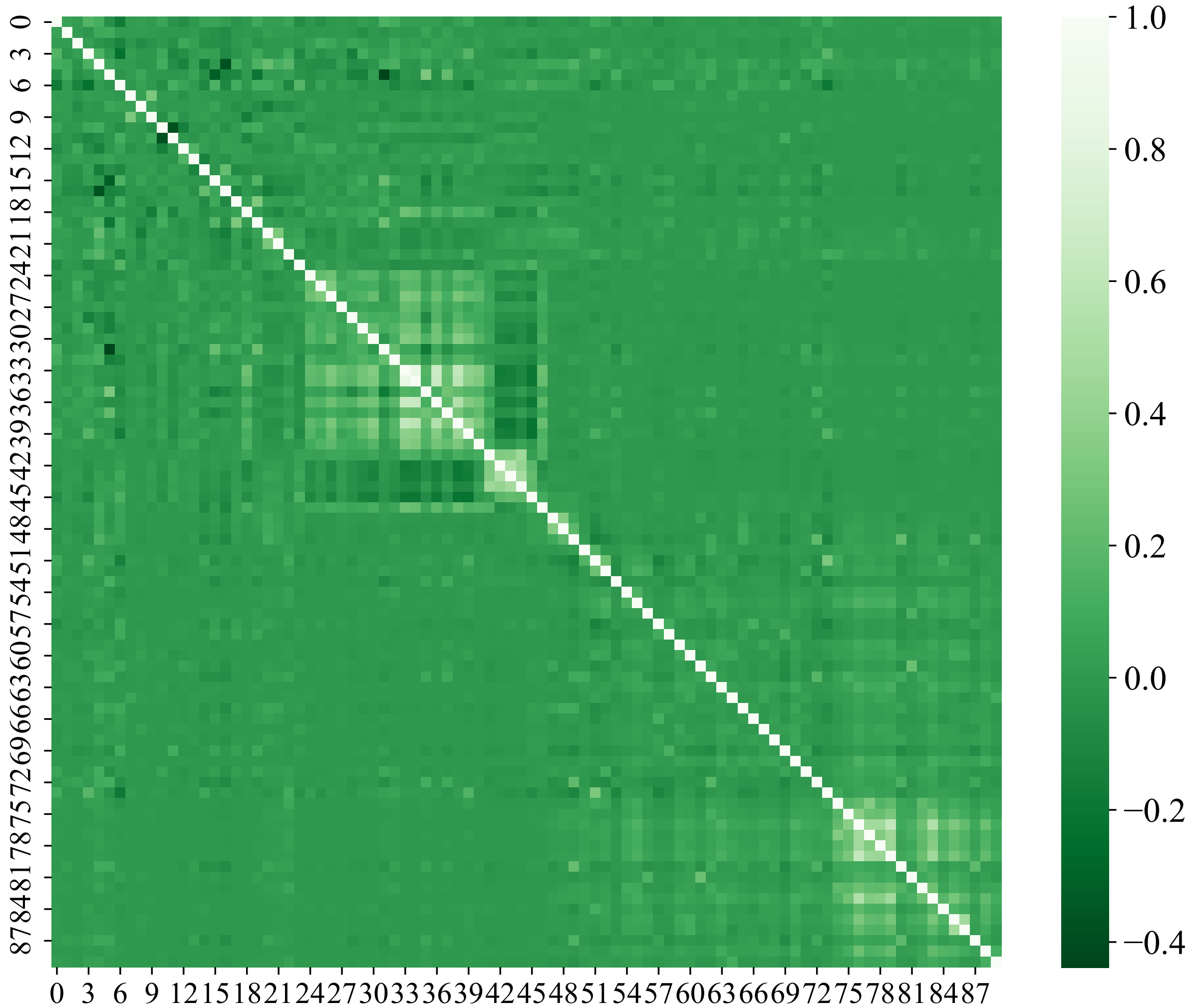

Figure 4: Pearson correlation between 90 dimensions, with dimension indices corresponding to Table 6 in the appendix.

(1) Likes (2) Dislikes (3) Comments alone cannot reflect

Response:
3

Your task:

Query:
{Post}

Comment:
{Comment}

Task:
{Question}

Response:

Demographic Information Generation and Preference Direction Annotation Prompt

Who is likely to prefer <Response 1> and dislike <Response 2>? Compare the preferred and disliked responses, and describe a person using selected dimensions from the following list of 48, reflecting what can be observed:

1. Age group: (“Young”, “Older”)
2. Gender: (“Female”, “Male”)
3. Occupation

Source	Dimensions
Psychological models that capture fundamental human needs	Age group, Gender, Degree of Neuroticism in the Big Five personality traits, Degree of Extraversion in the Big Five personality traits, Degree of Openness in the Big Five personality traits, Degree of Agreeableness in the Big Five personality traits, Degree of Conscientiousness in the Big Five personality traits, Preference for a certain food, Preference for a certain living environment, Sleep preference, Investment preference, Saving preference, Importance placed on physical safety, Importance placed on environmental safety, Preference for the depth of interaction, Approach to handling conflict, Communication style in social settings: more concise or detailed, Need for a certain work environment, Need for recognition from others, Need for personal achievement, Preference for a certain knowledge area, Preference for a certain learning style, Preference for a certain form of creative expression (such as art, writing, music), Need for Order (neatness, organization, avoiding chaos), Need for Retention (to keep possession over an object, unwilling to lose or change), Need for Inviolacy (maintaining dignity and reputation, unviolated or undamaged), Need for Infavoidance (avoiding failure and embarrassment), Need for Counteraction (attempting to compensate for failure by trying again, with a desire to overcome obstacles and seek pride), Need for Seclusion (desire for isolation from others, maintaining privacy), Need for Dominance (controlling the environment or others through command or persuasion), Need for Deference (worship or obedience to others, following authority, admiration, or rules), Need for Autonomy (resisting influence from others, pursuing independence and self-reliance), Need for Contrariance (pursuing uniqueness, being different, or opposing the norm), Need for Abasement (submittance and obedience to others, accepting blame and punishment, even enjoying pain or misfortune to some extent), Need for Aggression (controlling, punishing, or harming others through forceful means), Need for Affiliation (desire for close, loyal relationships, pleasing others, winning friendships and attention), Need for Rejection (isolating, excluding, or discarding oneself from negatively evaluated objects or people), Need for Nurturance (assisting the weak, caring for others, protecting them from danger), Need for Succorance (desire to have one's own needs met by others, including being loved, cared for, helped, forgiven, and comforted), Need for Play (enjoying fun, relaxation, laughter)
Contemporary research in recommender systems and AI alignment that reflect social-cognitive needs in digital interactions	Degree of concern for the harmlessness of the statement, Degree of concern for the instruction-following of the statement, Degree of concern for the honesty of the statement, Degree of concern for the truthfulness of the statement, Degree of concern for the helpfulness of the statement, Degree of concern for the coherence of the statement, Degree of concern for the complexity of the statement
Content platform indicators that represent everyday user needs	science, knowledge, psychology, cinema, entertainment, gaming, parenting, wild imagination, anime, sports, law, workplace, pets, travel, health, stories, cars, gourmet food, education, current events, home decor, international, finance, campus life, digital technology, emotions, humor, music, reading, painting, dance, crafts, photography, culture, fitness, art, stationery and planners, celebrities, outdoors, camping, social sciences, weddings, fashion

Table 6: Dimensions and corresponding sources in our preference space.

4. Degree of Neuroticism in the Big Five personality traits: ("High", "Low")
5. Degree of Extraversion in the Big Five personality traits: ("High", "Low")
6. Degree of Openness in the Big Five personality traits: ("High", "Low")
7. Degree of Agreeableness in the Big Five personality traits: ("High", "Low")
8. Degree of Conscientiousness in the Big Five personality traits: ("High", "Low")
9. Preference for a certain food: ("Likes", "Dislikes")
10. Preference for a certain living environment: ("Likes", "Dislikes")
11. Sleep preference: ("Likes", "Dislikes")
12. Investment preference: ("Aggressive", "Conservative")
13. Saving preference: ("Good at saving", "Bad at saving")
14. Importance placed on physical safety: ("Concerned", "Not concerned")
15. Importance placed on environmental safety: ("Concerned", "Not concerned")
16. Preference for the depth of interaction: ("Superficial interaction (casual, stress-free chat)", "Deep interaction (discussing interests, emotional topics, etc.)")

17. Approach to handling conflict: ("Direct communication", "Avoidance, mediation, compromise")
18. Communication style in social settings: more concise or detailed: ("Concise", "Detailed")
19. Need for a certain work environment: ("Strong", "Mild")
20. Need for recognition from others: ("Strong", "Mild")
21. Need for personal achievement: ("Strong", "Mild")
22. Preference for a certain knowledge area: ("Likes", "Dislikes")
23. Preference for a certain learning style: ("Likes", "Dislikes")
24. Preference for a certain form of creative expression (such as art, writing, music): ("Likes", "Dislikes")
25. Need for Order (neatness, organization, avoiding chaos): ("Strong", "Indifferent")
26. Need for Retention (to keep possession over an object, unwilling to lose or change): ("Strong", "Indifferent")
27. Need for Inviolacy (maintaining dignity and reputation, unviolated or undamaged): ("Strong", "Indifferent")
28. Need for Infavoidance (avoiding failure and embarrassment): ("Strong", "Indifferent")
29. Need for Counteraction (attempting to compensate for failure by trying again, with a desire to overcome obstacles and seek pride): ("Strong", "Indifferent")
30. Need for Seclusion (desire for isolation from others, maintaining privacy): ("Strong", "Indifferent")
31. Need for Dominance (controlling the environment or others through command or persuasion): ("Strong", "Indifferent")
32. Need for Deference (worship or obedience to others, following authority, admiration, or rules): ("Strong", "Indifferent")
33. Need for Autonomy (resisting influence from others, pursuing independence and self-reliance): ("Strong", "Indifferent")
34. Need for Contrariance (pursuing uniqueness, being different, or opposing the norm): ("Strong", "Indifferent")
35. Need for Abasement (submittance and obedience to others, accepting blame and punishment, even enjoying pain or misfortune to some extent): ("Strong", "Indifferent")
36. Need for Aggression (controlling, punishing, or harming others through forceful means): ("Strong", "Indifferent")
37. Need for Affiliation (desire for close, loyal relationships, pleasing others, winning friendships and attention): ("Strong", "Indifferent")
38. Need for Rejection (isolating, excluding, or discarding oneself from negatively evaluated objects or people): ("Strong", "Indifferent")
39. Need for Nurturance (assisting the weak, caring for others, protecting them from danger): ("Strong", "Indifferent")
40. Need for Succorance (desire to have one's own needs met by others, including being loved, cared for, helped, forgiven, and comforted): ("Strong", "Indifferent")
41. Need for Play (enjoying fun, relaxation, laughter): ("Strong", "Indifferent")
42. Degree of concern for the harmlessness of the statement: ("Strong", "Indifferent")

43. Degree of concern for the instruction-following of the statement: ("Strong", "Indifferent")
44. Degree of concern for the honesty of the statement: ("Strong", "Indifferent")
45. Degree of concern for the truthfulness of the statement: ("Strong", "Indifferent")
46. Degree of concern for the helpfulness of the statement: ("Strong", "Indifferent")
47. Degree of concern for the coherence of the statement: ("Strong", "Indifferent")
48. Degree of concern for the complexity of the statement: ("Strong", "Indifferent")

Compare the liked and disliked responses, and derive the keywords for likes and dislikes, ensuring the keywords are selected from the following list:

{science, knowledge, psychology, cinema, entertainment, gaming, parenting, wild imagination, anime, sports, law, workplace, pets, travel, health, stories, cars, gourmet food, education, current events, home decor, international, finance, campus life, digital technology, emotions, humor, music, reading, painting, dance, crafts, photography, culture, fitness, art, stationery and planners, celebrities, outdoors, camping, social sciences, weddings, fashion}.

If keywords cannot be summarized, please leave them blank.

Example 1:

Query:
Sure, kids are expensive and a lot of people who would otherwise want them are definitely putting off parenthood because they can't afford it. But everyone acts like this is the only reason why younger adults don't want kids, and it's still not okay to admit that you just don't want to be a parent regardless of money.

I'm sick of mentioning my student debt and getting sympathy from people who feel bad that I can't afford to be a mom, when I never wanted to be one to begin with. Nobody can handle the truth about how a lot of millennial women aren't having kids because for the first fucking time in history, we actually have the conscious choice not to, and have access to birth control that actually works.

I'm sick of people putting all the blame on finances in articles about declining birth rates and childfree millennials

Output:
<Response 1>
I don't quite see how we're the problem, when we're the ones who are trying to work, get a decent job in our field, maybe move out of our parents house like a normal, functional human being, rather than popping out kids left and right that we cannot afford as soon as we finish high school.

<Response 2>
Women with or without student loan debt have many of the same problems. The wildly (IMO) outrageous childcare costs and depending on the area high housing costs are hard for two incomes and impossible for one. Young women are encouraged to have children and yet its been proven, income and promotion opportunities will be stalled or decreased. If you add student loan debt it just becomes more untenable.

Full disclosure, I'm a childfree, 65 year-old. Sadly, we had the same issues with sexual harassment and lack of advancement, but we did not have anywhere near the financial burdens. Unless you wanted to go to an Ivy League college you could work your way through college with part-time jobs or a little help from your parents and get a well-paying job. Housing was reasonably priced as was childcare.

Old people, sadly I have to include myself, need to get a grip and understand how much times have changed and raising and educating a child really costs. Maybe when you all run the government it will change.

{ "reason": "People who prefer <Response 1> value personal freedom and see not having children as an individual choice, not just a result of financial or societal pressures. They reject traditional views on parenthood and feel <Response 2>'s focus on financial burdens limits their freedom to choose.",
"textual persona": "A well-educated, ambitious founder of a tech startup who is willing to challenge traditional views on parenting and society. He takes a cautious approach to having children, valuing conscious family planning choices and seeing economic factors as secondary to the broader societal shift in expectations. He is unlikely to conform to societal norms, focusing on personal goals and independence, and takes responsibility for their financial and career decisions.",
"dimension description": { "Age group": "Young", "Occupation": "Founder of a tech startup", "Degree of Openness in the Big Five personality traits": "High", "Degree of Agreeableness in the Big Five personality traits": "Low", "Degree of Conscientiousness in the Big Five personality traits": "High", "Preference for a certain living environment": "Likes", "Approach to handling conflict": "Direct communication", "Need for a certain work environment": "Strong", "Need for personal achievement": "Strong", "Need for Autonomy (resisting influence from others, pursuing independence and self-reliance)": "Strong" },
"like keywords": "education, workplace, finance",
"dislike keywords": "parenting" }

Your task:

Query:
{Post}

Output:
<Response 1>
{Response 1}

<Response 2>
{Response 2}

Now, please output the persona you created, keywords, and a short rationale below in a JSON format by filling in the placeholders in []:
{"reason": "[your rationale]", "textual persona": "[persona you created]", "dimension description":"[description of corresponding dimensions]", "like keywords": "[like, Optional]", "dislike keywords": "[dislike, Optional]"}

GPT-4 Judgment Prompt

The given user behavior history reflects the user's preferences. For a new query, which response better matches the user's needs and preferences, allowing for a tie between the two responses?

User Behavior History
{Concatenated user history information}

Query
{Query}

Response A
{Response A}

Response B
{Response B}

Now, please output your choice below in a JSON format by filling in the placeholders in []:
{"choice": "[Response A/Response B/Tie]"}

B Alignment methods

B.1 Conversion of preference direction to natural language description

When a persona $\mathcal{P}$ contains multiple components (e.g., both $\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{DEMO}}$) or behavioral personas comprise multiple instances, we first aggregate their preference directions into a unified vector $\tilde{P} \in \mathbb{R}^d$, where d denotes the dimension of preference space. Specifically, for each preference dimension i, we convert the categorical preferences across all components into numerical values (positive $\rightarrow$ 1.0, neutral $\rightarrow$ 0.5, negative $\rightarrow$ 0.0) and compute their average. We then determine the direction for each dimension based on two thresholds: dimensions with average values above t_1 are assigned positive direction, those below t_2 are assigned negative direction, and the remaining are considered neutral, resulting in the final preference direction vector $\tilde{P}$.

We then map $\tilde{P}$ to natural language descriptions following systematic linguistic rules. For dimensions with positive direction, we prepend "High" to personality traits (e.g., "High Neuroticism"), "Prefers" to communication styles (e.g., "Prefers detailed communication"), and use positive forms for values (e.g., "Values privacy"). For dimensions with negative direction, we prepend "Low" to personality traits (e.g., "Low Extraversion"), "Avoids" to communication styles (e.g., "Avoids confrontational communication"), and use negative forms for values (e.g., "Disregards material wealth").

The final description is constructed by concatenating all non-neutral preference descriptions with comma separators. This systematic conversion process ensures consistent and interpretable preference articulation while preserving the key signals from the original vector.

B.2 Training examples

In this section, we present the prompt format used during the training of ICA and PBA.

Data Format for ICA. Case:

Generate a task-specific response based on user historical behavior and preferences.

Task
My 10-year-old daughter has ADHD, Dyslexia, Dysgraphia, and Dyscalculia, with an IQ of 79. She struggles academically and socially but loves theatre. Looking for advice or similar experiences. Thank you.

Historical Behavior and User Preferences

This person has commented on some posts:

1. **Post:** I've been trying to potty train my daughter, but today I let her go without a diaper, and she stayed dry all day. She's trained, just not potty trained!
   Comment: Oh, so she's "trained"—just not in the way that actually matters. Maybe try using the potty next time instead of just hoping for magic!

This person has chosen or rejected comments on some posts:

1. **Post:** I'm potty training a stubborn boy and want to use cheap, uncomfortable diapers or training pants to make him prefer big boy pants. Any recommendations for the worst ones?
   Chosen: Put him in a swim diaper—it will leak as soon as he urinates, making him uncomfortable. Tell him he's outgrown pull-ups, so he needs to use the toilet and wear big boy pants.
   Rejected: I once read about a dad who put his kid in a smaller diaper, acted surprised, and said, "Oh no! You've outgrown diapers! Time for big boy underwear." It worked for him.

This person's persona is:
A young executive with low agreeableness prefers direct communication in social settings and values independence and efficiency. This person likes parenting, education, dislikes emotions.

Response:

Data Format for PBA. Case:

Generate a task-specific response based on user preferences.

Task
My 10-year-old daughter has ADHD, Dyslexia, Dysgraphia, and Dyscalculia, with an IQ of 79. She struggles academically and socially but loves theatre. Looking for advice or similar experiences. Thank you.

User Preferences
Low Agreeableness, Prefers direct communication to handle conflict, Concise communication style, Strong need for Autonomy (pursuing independence and self-reliance), Likes parenting, Likes education, Dislikes emotions

Response:

C Details of the experimental setup

C.1 Benchmarks

We conduct experiments on four benchmarks, as described in Table 7: (1) UF-P-4 (Poddar et al., 2024): An alignment dataset focused on universal values (e.g., helpfulness) with consistently positive preference directions. We treat each preference pair as a test case and construct $\mathcal{P}_{\text{PAIR}}$ by randomly sampling 2∼8 pairs from the same dimension. (2) PRISM (Kirk et al., 2024c): A collection of real user-LLM interaction preferences with unknown preference dimensions and directions.[9] Each preference pair serves as a test case, with $\mathcal{P}_{\text{PAIR}}$ comprising all other pairs from the same user. (3) P-SOUPS (Jang et al., 2023): A benchmark spanning three novel dimensions beyond our preference space, with bidirectional preference. For each dimension-direction configuration, we use individual preference pairs as test cases and sample 4 pairs from the remaining as $\mathcal{P}_{\text{PAIR}}$. (4) ALIGNX$_{\text{test}}$: A comprehensive test set constructed following the methodology in §4.2, covering all dimensions in our preference space with balanced positive and negative directions. For each dimension and direction combination, we construct dedicated test cases where both the preference pair and persona consistently exhibit the specified direction for the given dimension. Our evaluation is both on individual persona types and arbitrary combinations of three types of personas in $\mathcal{P}$ to test the efficacy under varying levels of persona completeness. Both $\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{PAIR}}$ in ALIGNX$_{\text{test}}$ include 4 randomly sampled examples. We ensure that none of the above test cases overlap with training data.

[9] Manual inspection finds that most preference pairs in PRISM often reflect various combinations of universal values (e.g., more friendly, more empathy).

C.2 Implementation details

We implement ICA and PBA based on Llama-3.1-8B-Instruct. Both models are trained on 8 A100 GPUs for one epoch using Adam optimizer (Kingma, 2014), DeepSpeed with ZeRO-3 (Rajbhandari et al., 2020) and Flash-attention-2 (Dao, 2023), with the following configuration: learning rate of 5e-7, batch size of 128, maximum sequence length of 8,192, and $\beta = 0.1$ for both Eq. 1 and 2. During training, both $\mathcal{P}_{\text{UGC}}$ and $\mathcal{P}_{\text{PAIR}}$ retain only 0∼4 randomly sampled examples for computational efficiency. We use Qwen2.5-14B-Instruct for annotating preference directions in PBA. §B.2 describes the format used during the training of ICA and PBA. Since VPL (Poddar et al., 2024) only scores responses but does not generate them, we evaluate VPL solely using alignment accuracy, computing it directly without the reference model. Both VPL and ALIGNXPERT use accuracy metrics consistent with their training objectives, ensuring a fair comparison.

D Supplementary experimental results

D.1 Detailed scores on the P-SOUPS benchmark

The P-SOUPS benchmark involves three preference dimensions: *Expertise*, *Informativeness*, and *Style*. Although these dimensions are not included in the dimensions defined by ALIGNX, they can be represented as combinations of ALIGNX's preference space. The experimental results in Table 8 show that ALIGNXPERT (trained on the 7% subset of ALIGNX data) significantly outperforms the baselines in both individual and combined dimensions.

D.2 Results with larger models as reference models

To validate the effectiveness of ALIGNXPERT, we use Llama-3-70B-Instruct as the reference model and conduct experiments on both in-domain (ALIGNX$_{\text{test}}$) and out-of-domain (P-

Benchmark	$\mathcal{P}_{\text{UGC/PAIR/DEMO}}$	Dimensions	In-distribution
UF-P-4	✗ / ✓ / ✗	Helpfulness (↑: 797); Honesty (↑: 740); Instruction-Following (↑: 830); Truthfulness (↑: 621)	Yes
PRISM	✗ / ✓ / ✗	*Unknown (868 examples in total)*	No
P-SOUPS	✗ / ✓ / ✗	Expertise (↑: 300, ↓: 300); Informativeness (↑: 300, ↓: 300); Style (↑: 300, ↓: 300)	No
ALIGNX$_{\text{test}}$	✓ / ✓ / ✓	All dimensions in our preference space (3716 examples in total)	Yes

Table 7: Benchmark overview. For each benchmark, we indicate the availability of different persona types and their focused preference dimensions. "In-distribution" indicates whether the benchmark's preference dimensions overlap with those in our training data. ↑/↓ means positive/negative preference directions, where the following number refers to the corresponding number of examples.

Model	Expertise	Informativeness	Style
Llama-3-8B-Instruct	40.17	56.81	51.33
Qwen2.5-7B-Instruct	38.00	43.69	22.00
Mistral-7B-Instruct-v0.2	40.17	46.35	33.83
ALIGNXPERT$_{\text{ICA}}$	**85.17**	<u>60.80</u>	**83.67**
ALIGNXPERT$_{\text{PBA}}$	<u>84.50</u>	**66.11**	<u>79.17</u>

Table 8: Accuracy performance (%) of each dimension on P-SOUPS.

SOUPS) benchmarks. Table 9 shows that our method demonstrates strong performance compared to both the 70B large model and the baselines.

Model	P-SOUPS	ALIGNX$_{\text{test}}$			
	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{UGC}}$	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{DEMO}}$	$\mathcal{P}$
Llama-3-8B-Instruct	46.78	50.59	51.32	47.85	50.69
Qwen2.5-7B-Instruct	34.62	48.98	49.00	50.67	49.87
Mistral-7B-Instruct-v0.2	38.95	49.38	47.95	48.17	48.27
Training with Full Dataset of 1,311,622 Samples (100%)					
ALIGNXPERT$_{\text{ICA}}$	59.39	**57.80**	57.64	**90.93**	**73.21**
ALIGNXPERT$_{\text{PBA}}$	**80.20**	56.70	**59.66**	88.19	71.01

Table 9: Alignment accuracy (%) of different models using Llama-3-70B-Instruct as the reference model.

D.3 Results with larger models as policy models

Since it is difficult to calculate alignment accuracy for larger or closed-source models by obtaining log-probabilities, we instead calculate their accuracy by directly asking the LLM to make a binary choice. Table 10 shows the performance of larger models, including Qwen2.5-32B-Instruct[10], QwQ-32B[11], DeepSeek-R1-671B[12] and GPT-4[13]. Our results show that ALIGNXPERT$_{\text{ICA}}$ (trained on an 8B model with 7% data) outperforms the 32B models across all benchmarks. It also surpasses DeepSeek-R1-671B and GPT-4 on P-SOUPS and ALIGNX$_{\text{test}}$.

[10]https://huggingface.co/Qwen/Qwen2.5-32B-Instruct

[11]https://huggingface.co/Qwen/QwQ-32B

[12]We use DeepSeek-R1 version released on 2025/01/20.

[13]We use OpenAI's API "gpt-4-turbo-2024-04-09."

Model	UF-P-4	PRISM	P-SOUPS	ALIGNX$_{\text{test}}$
	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}_{\text{PAIR}}$	$\mathcal{P}$
Qwen2.5-32B-Instruct	36.33	40.59	43.02	66.79
QwQ-32B	41.09	50.53	48.72	66.42
DeepSeek-R1-671B	<u>65.80</u>	**83.67**	68.93	<u>68.12</u>
GPT-4	**67.45**	<u>83.33</u>	48.24	63.56
Training with a Subset of randomly sampled 91,918 Samples (7%)				
ALIGNXPERT$_{\text{ICA}}$	61.22	76.69	<u>76.54</u>	**70.88**
ALIGNXPERT$_{\text{PBA}}$	54.90	52.43	**76.59**	64.75

Table 10: Alignment accuracy (%) of different large models.

D.4 Human evaluation on ALIGNX$_{\text{test}}$

To complement GPT-4 judgments, we conduct human evaluation on ALIGNX$_{\text{test}}$ using four qualified annotators. For each comparison, annotators are given the same persona-conditioned prompt and may choose `Win`, `Lose`, or `Tie`. Table 11 reports the resulting pairwise preferences and Fleiss' κ.

These results corroborate the LLM-as-a-judge findings: both ALIGNXPERT variants outperform the Llama-3.1-8B-Instruct baseline, while ALIGNXPERT$_{\text{PBA}}$ and ALIGNXPERT$_{\text{ICA}}$ exhibit complementary strengths.

D.5 Calibration of inferred preference directions

To better understand the gap between inferred and gold preference directions, we compute category-wise precision and recall of predicted $\tilde{P}$ against the corresponding gold directions, as reported in Table 12.

The inference module exhibits high recall for active preferences but substantially lower precision for positive/negative labels. This indicates a sensitivity bias: the model tends to over-predict weak preferences instead of defaulting to `Neutral`. This behavior explains part of the gap between ALIGNXPERT$_{\text{PBA}}$ with inferred $\tilde{P}$ (71.10%) and the Golden $\tilde{P}$ upper bound (91.36%) in Table 2. In practice, this suggests that improving calibration of inferred preference directions is a promising direction for strengthening PBA.

Comparison	Win	Lose	Tie	Fleiss' κ
ALIGNXPERT$_{\text{ICA}}$ vs. Llama-3.1-8B-Instruct	45.91	37.70	16.39	0.52
ALIGNXPERT$_{\text{PBA}}$ vs. Llama-3.1-8B-Instruct	50.00	40.20	9.80	0.48
ALIGNXPERT$_{\text{PBA}}$ vs. ALIGNXPERT$_{\text{ICA}}$	39.53	48.84	11.63	0.53

Table 11: Human evaluation results (%) on ALIGNX$_{\text{test}}$. "Win" denotes preference for the first model in each comparison.

Category	Precision	Recall
Positive	35.73	87.27
Negative	41.60	82.90
Neutral	92.06	68.22

Table 12: Category-wise precision and recall (%) of inferred $\tilde{P}$.

E Analysis

E.1 Preference diversity

We analyze potential correlations between the 90 dimensions across all 1.3 million examples. The Pearson correlation heatmap is available at Figure 4, with dimension indices corresponding to Table 6. Among all dimension pairs, 99.65% show correlations below 0.5, indicating that our 90 dimensions capture largely independent aspects of user preferences. Only two dimensions—"Need for Abasement (submission and obedience to others, accepting blame and punishment, even deriving pleasure from pain or misfortune)" and "Need for Aggression (controlling, punishing, or harming others through force)"—exhibit a strong correlation (0.86). These two may reflect different expressions of the same underlying personality trait and can complement each other in practice without affecting other dimensions. Overall, at least 89 independent dimensions comprehensively capture user personality.

Based on the high independence among the 90 dimensions, we calculate that there are 1,210,986 different preference directions in ALIGNX, accounting for 92.33% of the total data, with 1,210,837 preference directions coming from Reddit annotated data, demonstrating diverse user preferences.

E.2 Annotation bias

To construct data at scale, we use an LLM for annotation during the process. We take several measures to assess potential LLM annotation biases.

In the qualitative analysis, our process first clusters responses into three groups and randomly selects responses from different groups as y_w and y_l to determine preference direction (§4.1). LLMs only label these pre-determined preferences without generating opinions, thus minimizing bias introduction.

In the quantitative analysis, we conduct three complementary audits to assess the reliability and safety of our LLM-assisted annotation pipeline. First, to test robustness to the choice of annotation model, we randomly sample 100 examples from ALIGNX and re-annotate their preference directions using GPT-4. Second, to assess annotation quality against human judgment, we randomly sample 500 examples and ask four human annotators to label the corresponding preference directions, using the majority consensus as the human reference. Third, to evaluate the quality and safety of $\mathcal{P}_{\text{DEMO}}$, four human evaluators assess the same 500 samples along five criteria: groundedness, comprehensiveness, consistency, bias, and safety. Additionally, we use a case study to analyze whether there is societal bias in the data.

Bias from LLM selection. We re-label the preference directions of 100 sampled examples using `gpt-4-turbo-2024-04-09`. We find that 84.78% of the resulting preference directions agree with those labeled by Qwen2.5-14B-Instruct, indicating that our annotation pipeline is robust to the choice of the annotator model.

In addition, we use `gpt-4-turbo-2024-04-09` to generate alternative $\mathcal{P}_{\text{DEMO}}$ descriptions. Both the original Qwen2.5-14B-Instruct-generated descriptions and the GPT-4-generated descriptions are then mapped to preference directions using GPT-4. The resulting preference directions show 92.86% agreement, demonstrating the consistency of the descriptive persona generation process across different LLM annotators.

Bias from LLM annotation. We recruit four human annotators for the labeling task. On 500 randomly sampled examples, the LLM annotations achieve 85.07% agreement with the consensus of four human annotators, while the agreement among human annotators is 90.37%. These results indi-

Gender/Succorance	Negative	Positive
Negative	1,986	1,980
Positive	2,117	1,973

Table 13: The data count for different combinations of "Gender" and "Succorance" at varying intensities.

cate that the annotation pipeline approaches human-level consistency on this task and provides strong support for the reliability of the resulting supervision signals.

$\mathcal{P}_{\text{DEMO}}$ quality and safety audit. To assess the quality and ethical safety of $\mathcal{P}_{\text{DEMO}}$, four human evaluators score 500 randomly sampled descriptions on five criteria: groundedness, comprehensiveness, consistency, bias, and safety. The resulting descriptions achieve an average score of 8.82/10, with Fleiss' $\kappa = 0.66$, indicating substantial agreement. These results suggest that the generated descriptive personas are generally well grounded, internally consistent, and do not exhibit obvious harmful stereotyping in the audited sample.

These annotators demonstrate excellent English proficiency, allowing them to accurately assess preference tendencies and persona descriptions. Regarding human annotations, the process is reviewed and approved by relevant institutions. All annotators were informed of the study's purpose and the intended use of their judgments, and provided their consent prior to participation. After a fair evaluation of the workload, we compensate each evaluator $5.60 per 10 samples.

Case study We identify two attributes often linked to societal bias: "Gender" and "Need for Succorance (desire to have one's own needs met by others, including being loved, cared for, helped, forgiven, and comforted)." Table 13 shows the data count for their different combinations.

The data distribution is relatively balanced without clear tendencies, indicating that our data does not introduce bias.

E.3 Complexity of preference dimensions

To investigate the complexity of the 90 dimensions, we analyze the computational challenges encountered during practical deployment. Only ALIGNXPERT$_{\text{PBA}}$ involves the 90-dimensional space in this process. The main computational aspects are as follows:

• Predicting preference directions from Persona $\mathcal{P}$: With 8 examples in $\mathcal{P}$, computation time increases modestly from 10.20s to 16.21s as dimensions increase from 1 to 90 - a reasonable overhead given the benefits of comprehensive preference modeling.

• Converting preference directions to natural language: This rule-based transformation has negligible computational cost.

Overall, the 90 dimensions provide more comprehensive preference coverage compared to existing research while maintaining manageable computational overhead.