

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 CONTEXT-AWARE ALIGNMENT: ADAPTING LARGE LANGUAGE MODELS TO INDIVIDUAL HISTORICAL DATA

Anonymous authors

Paper under double-blind review

ABSTRACT

Aligning large language models with human preferences is essential for ensuring their effectiveness, utility, and safety in real-world applications. While much of the current research focuses on aligning LLMs with generalized human values such as fairness, transparency, and ethical behavior, limited attention has been given to aligning LLMs with the preferences and characteristics of individual users. In this paper, we propose a novel approach that leverages individual historical context to achieve personalized alignment, adapting LLMs to align with the unique traits and preferences of specific users. Our method focuses on extracting persona-related representations—abstract features encapsulating conversational style, tone, and preferences—from past user interactions. These representations guide the model in generating responses tailored to the user’s individual characteristics. Experimental results demonstrate that our approach significantly outperforms existing baselines, improving the model’s ability to reflect individual personas while maintaining contextual appropriateness. This research opens new possibilities for more personalized, context-aware, and user-centric applications of LLMs.

1 INTRODUCTION

Large language models (LLMs) have shown tremendous potential in a wide range of tasks, but their effectiveness is highly contingent on their ability to respond in ways that align with human values (Brown et al., 2020; Bubeck et al., 2023; Touvron et al., 2023). Aligning LLMs with human preferences is essential for ensuring their utility, safety, and overall effectiveness in real-world applications (Liu et al., 2024; Ouyang et al., 2022b; Rafailov et al., 2023b). Most existing research in alignment has focused on generalizing models to reflect broad, high-level human values such as fairness, transparency, and ethical behavior. While these efforts are critical, they fail to address the needs of specific individuals, which limits the potential of LLMs in personalized contexts. In applications like personalized AI assistants, user-specific interaction models are necessary for creating adaptive, engaging, and meaningful interactions (Li et al., 2024; Kirk et al., 2024; Shi et al., 2024).

Achieving personalized alignment is a significant challenge, as it requires LLMs to not only reflect generalized values but also adapt to the specific conversational style, tone, and needs of each individual user. Traditional alignment methods, such as Direct Preference Optimization (DPO), focus on aligning models to general preferences using feedback on preferred and non-preferred outputs (Wu et al., 2024; Rafailov et al., 2023a). While DPO has shown promise in some scenarios, it has limitations when applied to individual alignment. These methods typically rely on binary feedback, treating general human values as universal preferences. However, this can lead to issues where what is considered a general preference for human alignment may not align with the specific traits or style of a given individual. When such a heterogeneous mix of preference data is used for training, the model receives conflicting optimization signals. This can lead to a state of ‘policy confusion,’ where the model struggles to learn a coherent strategy that effectively balances the nuanced requirements of individual alignment with the broader constraints of general alignment. The model’s inability to distinguish between universal values and individual preferences can lead to suboptimal personalization, undermining the adaptability needed for real-world applications.

054 To address these limitations, we propose a novel approach that explicitly leverages individual historical
 055 context to align LLMs with user-specific preferences. Our method begins by extracting persona-
 056 related representations from past user interactions. These representations encapsulate abstract features
 057 such as conversational tendencies, tone, and preferences, forming a comprehensive profile of the
 058 user's unique characteristics. We employ a contrastive learning framework to derive these persona
 059 embeddings, ensuring that they are both robust and representative of the user's historical context.

060 Building on this foundation, we introduce a Representation Regularization Loss (**RE**) within the
 061 DPO framework, referred to as **DPORE**, to guide the alignment process. This loss minimizes the
 062 divergence between the model's outputs and the user's persona representations, enabling the model
 063 to consistently generate responses that reflect individual traits. By integrating these user-specific
 064 embeddings into the training process, we ensure that the model's outputs are both personalized and
 065 contextually appropriate.

066 Our approach was evaluated through comprehensive experiments against established baselines,
 067 including traditional DPO methods. Results demonstrate that DPORE significantly enhances the
 068 model's ability to align with individual personas, as measured by both automated metrics and human
 069 evaluations. Notably, this improvement in personalization does not come at the expense of the model's
 070 general-purpose performance, confirming the efficacy of our regularization technique in striking a
 071 balance between specific user needs and broad utility. Instead, it highlights the potential of leveraging
 072 individual historical context as a transformative step in alignment research, offering a principled way
 073 to achieve deeper personalization.

074 2 RELATED WORK

075 2.1 HUMAN ALIGNMENT OF LLMs

076 Aligning LLMs with human preferences is essential for ensuring that their outputs are both useful
 077 and ethical in real-world applications. Early alignment methods primarily focused on aligning
 078 LLMs with generalized human values, such as fairness, safety, and transparency. Among these,
 079 Reinforcement Learning from Human Feedback (RLHF) has emerged as one of the most successful
 080 techniques (Ouyang et al., 2022a). RLHF uses human feedback to fine-tune LLMs, improving their
 081 alignment with human preferences.

082 To address the computational cost and stability challenges associated with RLHF, Direct Preference
 083 Optimization (DPO) (Rafailov et al., 2023a) was proposed as a more efficient alternative. DPO opti-
 084 mizes language models by increasing the relative probability of preferred responses over dispreferred
 085 ones, reducing reliance on complex human feedback loops. Another promising direction in alignment
 086 research is representation alignment (Liu et al., 2024), which focuses on identifying representations
 087 of high-level human preferences embedded within LLMs. By modifying these representations,
 088 researchers can achieve more precise control over model behavior.

089 However, while these approaches excel at aligning models with generalized human values, they often
 090 overlook the personalized needs of individual users. Existing methods are typically designed to align
 091 models with societal-level values, leaving the nuanced preferences and unique characteristics of
 092 specific individuals underexplored.

093 2.2 INDIVIDUAL ALIGNMENT

094 The challenge of aligning LLMs with individual user preferences has garnered growing attention.
 095 Some research has explored customized LLMs that leverage user-specific information, such as
 096 interaction history, to adapt model responses. For instance, USER-LLM (Ning et al., 2024) integrates
 097 user embeddings to contextualize LLMs based on interaction history, enabling more personalized
 098 outputs. However, this approach requires a rich history of user interactions, limiting its effectiveness
 099 for new or infrequent users lacking sufficient data.

100 ALOE (Wu et al., 2024) adopts a different strategy by implicitly inferring user preferences from
 101 multi-turn dialogues. It dynamically adjusts the model's behavior based on inferred preferences and
 102 employs DPO for training. However, like other DPO-based approaches, ALOE struggles to directly
 103 capture personalized features, as DPO optimization often emphasizes overall response quality. This

108 focus can lead to distortion of learned preferences by irrelevant factors, such as tone or response
 109 length.

110 In contrast, our approach explicitly extracts user-specific features from input data and aligns the model
 111 with these features during training. This direct alignment ensures that the model learns user-specific
 112 characteristics without being influenced by irrelevant factors, enabling more accurate personalization.
 113

114 2.3 PERSONALIZED ALIGNMENT

116 A related area of research is personalized alignment (Tu et al., 2023; Li et al., 2023a; Chen et al.,
 117 2024), which aims to enhance specific personality traits or behaviors in LLMs, such as humor,
 118 courage, or politeness. These methods focus on making models more engaging by training them to
 119 express predefined traits.

120 While personalized alignment typically involves tailoring the model to exhibit broad personality traits,
 121 our approach dynamically adapts to individual interactions. This flexibility ensures that the model’s
 122 responses not only align with general personality traits but also reflect the specific preferences and
 123 conversational nuances of each user. By bridging the gap between generalized personality adaptation
 124 and individual alignment, our method enables a more comprehensive and user-centric approach to
 125 personalization.

127 3 METHODS

129 Large language models have demonstrated impressive capabilities in generating human-like text.
 130 However, aligning their responses with individual user preferences remains a significant challenge.
 131 In this section, we detail our method for achieving personalized alignment, starting with dataset
 132 preparation and progressing through embedding learning, fine-tuning, and the incorporation of a
 133 novel representation regularization technique. Figure 1 illustrates an overview of the training and
 134 deployment pipeline for our approach.

136 3.1 DATASETS

138 The dataset used in this study is derived from the ALOE benchmark introduced by Wu et al. (2024).
 139 It comprises over 3,000 multi-turn interaction samples, created using a persona pool of 3,310 unique
 140 user personas. Each sample includes both preferred responses, which are tailored to specific user
 141 personas, and non-preferred responses, which are generic and disregard persona information. To
 142 construct the persona pool, an iterative self-generation and filtering process was employed, ensuring
 143 diverse and distinct user profiles and personalities. This approach provides a rich dataset for studying
 144 dynamic, personalized alignment in conversational settings. The multi-turn nature of the dataset
 145 ensures that the alignment is evaluated not only on a single response but across the progression of a
 146 conversation, capturing more realistic interaction scenarios.

147 3.2 USER EMBEDDING LEARNING

149 ALOE aligns large language models with individual preferences relying on DPO, focus on learning
 150 from positive and negative response samples. However, this can cause issues when general preferences,
 151 which may work for societal-level alignment, conflict with the specific traits or conversational style
 152 of an individual user. When such mixed samples are used in training, the model becomes confused
 153 and struggles to optimize both individual and general alignment simultaneously.

154 To address these challenges and to create robust, explicit user persona representations, we propose
 155 a multi-view learning approach to extract user-specific persona representations from multi-turn
 156 interactions. Drawing inspiration from multi-view learning techniques, which have proven highly
 157 effective in domains like image recognition and natural language understanding for learning invariant
 158 features, we focus on isolating shared persona-related features across different dialogue excerpts
 159 while ignoring extraneous information or topic-specific noise. This allows our model to learn a
 160 condensed, meaningful representation of who the user is, rather than what they are talking about.

161 As illustrated in Figure 1, this method begins by sampling two dialogue excerpts, denoted as x_1
 and x_2 , that are associated with the same user persona p . These two excerpts form a **positive**

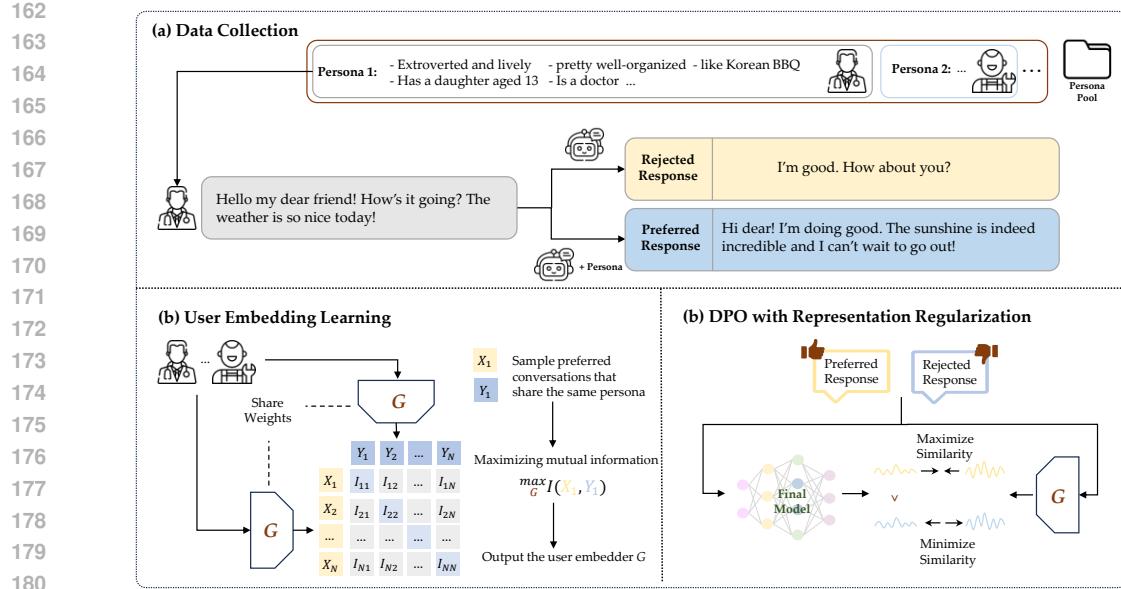


Figure 1: Overview of the Proposed Method for Individual Alignment. **(a) Data Collection:** Multi-turn dialogue samples are generated using a persona pool, with each sample containing a preferred and a rejected response. **(b) User Embedding Learning:** A contrastive learning framework is used to train a user embedding extractor G , which learns persona-specific features by maximizing mutual information between conversations sharing the same persona. **(c) DPO with Representation Regularization:** The DPO framework is extended with a representation regularization term that aligns model outputs with user embeddings for preferred responses and penalizes alignment for rejected responses.

pair, as they originate from the same underlying user identity. Conversely, dialogue excerpts from different personas are treated as **negative pairs**. We then employ a contrastive learning framework to maximize the similarity between positive pairs and minimize the similarity between negative pairs. This encourages the feature extractor, denoted as G , to learn compressed representations that robustly capture shared persona-related traits while actively discarding irrelevant noise, conversational artifacts, or topic-specific details.

Mathematically, given a set of dialogue excerpts $\{x_i\}$ and their corresponding personas $\{p_i\}$, we aim to train an encoder G to map each dialogue excerpt x_i into a low-dimensional embedding $G(x_i)$. The objective is to ensure that embeddings from the same persona are close, and embeddings from different personas are far apart. We use a **contrastive loss function**, specifically a variant of the InfoNCE loss, for this purpose. For a given anchor dialogue excerpt x_a associated with persona p_a , we consider a positive sample x_p (another dialogue excerpt from the same persona p_a) and a set of negative samples $\{x_n^k\}_{k=1}^K$ (dialogue excerpts from different personas). The contrastive loss for this anchor is defined as:

$$L_{\text{contrastive}}(x_a, x_p, \{x_n^k\}_{k=1}^K) = -\log \frac{\exp(\text{sim}(G(x_a), G(x_p))/\tau)}{\exp(\text{sim}(G(x_a), G(x_p))/\tau) + \sum_{k=1}^K \exp(\text{sim}(G(x_a), G(x_n^k))/\tau)} \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ is a similarity function (e.g., cosine similarity), and τ is a temperature hyperparameter that controls the sharpness of the distribution. By minimizing this loss over numerous anchor-positive-negative triplets, the encoder G learns to produce robust user persona embeddings. For each user, we utilized Llama3-8B-Instruct to generate 10 distinct dialogues, providing sufficient data for learning these persona embeddings.

To further enhance robustness and prevent overfitting to superficial textual features, we introduce random truncation of dialogues during training. This technique reduces the influence of dialogue length and specific structural elements on the learned embeddings, forcing the model to focus on intrinsic persona-related characteristics. The resulting feature extractor G encodes the unique persona traits of users based on their historical prompts and conversational context. These highly distilled

Method	Turn										Avg.
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$	
Base Model	3.692	3.457	3.660	3.540	3.492	3.456	3.421	3.448	3.397	3.415	3.497
Preferred-SFT	4.264	4.006	3.990	4.090	4.108	4.070	4.134	4.116	4.128	4.182	4.108
DPO	4.356	4.514	4.426	4.436	4.442	4.456	4.426	4.446	4.442	4.420	4.436
DPORE	4.583	4.448	4.498	4.544	4.580	4.588	4.542	4.588	4.550	4.562	4.543

Table 1: Alignment Levels Across k -th Turn in Multi-Turn Dialogues for Various Models and Methods. This table compares the performance of different alignment methods: Base, SFT, DPO, and our proposed method on Llama3-8B-Instruct. The columns represent the alignment scores at each turn ($k = 1$ to $k = 10$) in a multi-turn dialogue, while the final column reports the average score across all turns. Higher scores indicate better alignment with individual user personas. The results show that our method consistently achieves the highest alignment scores, outperforming the baselines and demonstrating its effectiveness in adapting to user-specific preferences over extended interactions.

persona embeddings are then incorporated into the preference alignment training process, enabling the main LLM to generate responses that precisely reflect user-specific styles, tones, and preferences, leading to a much more granular and faithful personalization.

3.3 DPO WITH REPRESENTATION REGULARIZATION

To further refine the alignment process, we extend the DPO framework with a novel representation regularization term. This regularization leverages the user embeddings, $G(m_i, p_i)$, to guide the model towards generating responses that are aligned with individual user preferences. This explicit guidance helps resolve the policy confusion that often arises when general and individual preferences are mixed.

For a given prompt m_i , positive response p_i , and negative response r_i , we align the model’s output, $y_\theta(m_i, p_i)$, with the persona embedding $G(m_i, p_i)$ using a projection head h_ϕ . The regularization loss is defined as:

$$L_{\text{reg}} = \lambda \cdot \text{sim}(y_\theta(m_i, p_i), h_\phi(G(m_i, p_i))) - \lambda \cdot \text{sim}(y_\theta(m_i, r_i), h_\phi(G(m_i, r_i))),$$

where $\text{sim}(\cdot, \cdot)$ is a similarity metric (e.g., cosine similarity) and λ controls the balance between preference optimization and regularization. The regularization term is designed to enforce that the preferred response is close to the persona embedding, while non-preferred responses are penalized for deviating from it. This alignment ensures that the model produces outputs that are consistent with the specific user’s persona. The total loss combines the DPO objective and the regularization term:

$$L_{\text{total}} = \sum_{i=1}^K \log \sigma \left(\beta \cdot \left(\log \frac{y_\theta(p_i | m_i)}{y'_\theta(p_i | m_i)} - \log \frac{y_\theta(r_i | m_i)}{y'_\theta(r_i | m_i)} \right) + \lambda \cdot (\text{sim}(y_\theta(m_i, p_i), h_\phi(G(m_i, p_i))) - \text{sim}(y_\theta(m_i, r_i), h_\phi(G(m_i, r_i)))) \right),$$

where y_θ and y'_θ denote the trained and reference models, respectively. The additional regularization ensures alignment with user persona embeddings for preferred responses and penalizes alignment for rejected responses. This mechanism reduces the influence of irrelevant factors and enhances personalization.

By integrating user representation regularization into DPO, our method improves the model’s ability to generate responses that are both contextually appropriate and tailored to individual user preferences. This approach bridges the gap between general alignment and personalized interaction, offering a scalable solution for dynamic, user-specific alignment.

270

4 EXPERIMENTS

271
 272 To validate the effectiveness of our method, we follow Wu et al. (2024) to conduct experiments
 273 on multi-turn dialogue tasks. Here, we compared various baseline approaches using both human
 274 evaluations and automated assessments. Given the trade-offs between enhancing specific capabilities
 275 and maintaining overall performance, we further assess the nominal performance of different methods.
 276 We also performed hyper-parameter experiments to assess the impact of representation regularization
 277 weights on performance. Finally, we present visualizations to intuitively demonstrate the effectiveness
 278 of our approach. Finally, we present qualitative visualizations through a case study to intuitively
 279 demonstrate the effectiveness and nuanced improvements of our approach in real-world conversational
 280 settings.

281

4.1 EXPERIMENTAL SETUPS

282
 283 **Dataset** The dataset used for training and testing is derived from the ALOE benchmark (Wu et al.,
 284 2024). The dataset consists of over 3,000 multi-turn dialogue samples generated using a persona pool
 285 of 3,310 unique user personas. Each sample contains ten rounds of dialogue, with both preferred
 286 responses (tailored to user personas) and non-preferred responses (generic, persona-agnostic). For
 287 evaluation, we use a subset of 100 evaluation cases, each associated with a distinct user persona,
 288 including detailed profile and personality descriptions. These evaluation cases are used to conduct
 289 role-playing experiments where GPT-based agents interact with our trained models.

290
 291 **Baselines** To evaluate the efficacy of our proposed approach, we conducted extensive comparisons
 292 with existing methods, which includes:

- 293 • **Base Model** This baseline consists of the selected instruction-tuned LLMs that have been aligned
 294 with general human preferences but lack explicit mechanisms for individual alignment. It serves to
 295 highlight the limitations of general alignment in capturing nuanced, user-specific preferences. In
 296 this study, we selected Llama-3-8B-Instruct (AI@Meta, 2024) as the base model.
- 297 • **Preferred-SFT** This baseline involves fine-tuning the language models using only the preferred
 298 responses from the dataset. To ensure nominal performance is maintained, we follow prior work
 299 and mix additional general-purpose tasks from the UltraFeedback dataset (Cui et al., 2023). Each
 300 example in UltraFeedback contains human preferences for general query responses, and we use
 301 only the preferred responses for SFT training.
- 302 • **DPO** Direct Preference Optimization (Rafailov et al., 2023a) optimizes LLMs without explicit user
 303 feature extraction. Like Preferred-SFT, DPO incorporates data from the UltraFeedback dataset to
 304 improve nominal performance. In our proposed method, representation regularization is applied
 305 only to individual preference data, not to general preference data, ensuring precise personalization.

306 Further implementation details for the baselines and our approach are provided in Appendix B.

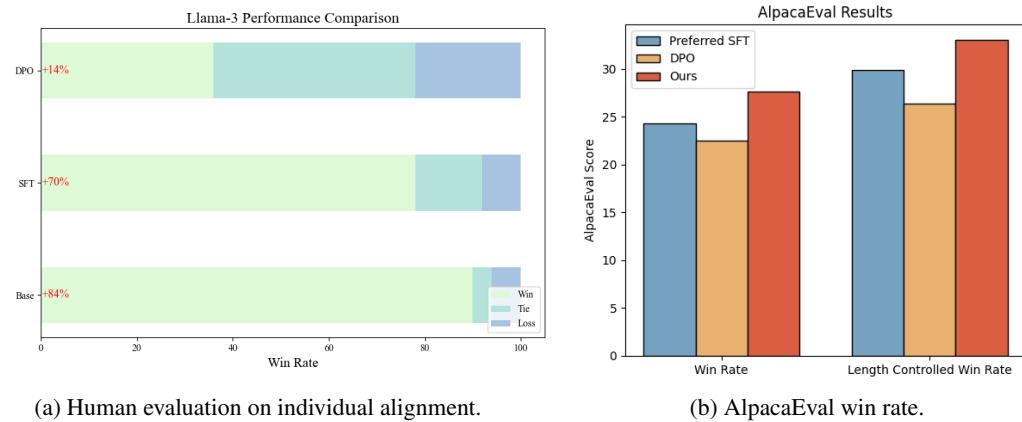
308

4.2 RESULTS

309
 310 To evaluate the effectiveness of our method in aligning models with individual user preferences,
 311 we conducted automated evaluations comparing our approach with several baseline methodologies.
 312 The evaluation primarily assessed how well the generated responses aligned with user personas and
 313 profiles. Leveraging GPT-4o mini for evaluation, which correlates strongly with human judgments (Li
 314 et al., 2023b), we provided prompts containing model-generated responses alongside agent persona
 315 descriptions. The consistent superiority of DPORE, as shown in Table 1, is particularly significant.
 316 While other methods may perform reasonably well in early turns, DPORE maintains a high level of
 317 alignment even in later stages of the conversation ($k=8, 9, 10$). This suggests that the persona repre-
 318 sentations learned by our method provide a stable and continuous signal throughout the interaction,
 319 preventing the model from reverting to a generic, non-personalized mode, which is a common pitfall
 320 in long-form dialogue generation. The prompts used are detailed in Appendix A.

321 To mitigate the variance inherent in GPT-based evaluations, we set the generation temperature to 0
 322 and conducted five evaluation runs for each method. The results, summarized in Table 1, demonstrate
 323 that our method consistently outperformed baseline approaches across nearly all dialogue rounds and
 average scores. This performance improvement stems from our explicit extraction of user-specific

324 features, which ensures that contrastive learning focuses on persona-related differences between
 325 positive and negative samples rather than irrelevant factors such as response length.
 326



Method	Writing	Roleplay	Reasoning	Math	Coding	Extraction	STEM	Humanities	Average
Base Model	8.350	7.700	5.600	5.550	5.650	8.350	7.100	8.600	7.112
Preferred-SFT	8.100	7.250	5.400	5.150	5.150	7.550	7.100	8.700	6.800
DPO	8.500	8.000	4.850	5.700	5.600	7.800	6.950	8.600	7.000
DPORE	8.450	7.450	6.100	5.500	5.300	7.800	7.550	8.600	7.093

Table 2: Comparisons on MT-Bench between our method and other baselines.

λ	Turn										Avg.
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	
0.5	4.385	4.450	4.425	4.388	4.452	4.427	4.417	4.440	4.387	4.375	4.414
1	4.446	4.574	4.428	4.528	4.436	4.472	4.474	4.456	4.47	4.436	4.471
2	4.486	4.492	4.532	4.532	4.562	4.586	4.556	4.532	4.56	4.526	4.536
5	4.583	4.448	4.498	4.544	4.580	4.588	4.542	4.588	4.550	4.562	4.543
10	4.380	4.466	4.436	4.458	4.534	4.540	4.536	4.520	4.510	4.466	4.484
20	4.256	4.282	4.354	4.472	4.528	4.518	4.562	4.556	4.606	4.560	4.469

Table 3: Alignment Levels Across k-th Turn for Varying Regularization Coefficients. This table shows the performance of Llama3-8B-Instruct with different λ values, demonstrating how the regularization coefficient impacts the model’s ability to align with individual user preferences across the dialogue turns.

positive-negative feedback. This allows the model to adjust its optimization direction when faced with conflicting preferences.

4.4.2 EVALUATION ON MT-BENCH

MT-Bench Zheng et al. (2023) is a challenging benchmark consisting of 80 samples, each containing two diagonal turns. This benchmark also employs GPT-4 to score model responses on a scale of 1 to 10 for each turn. Table 9 displays the performance scores achieved by our method and baseline models. Our method achieved comparable results to the baseline models on MT-Bench, demonstrating that it effectively enhances the capability of model to capture user individuality without compromising general task performance. This result is critical, as it shows that our approach can generalize to more complex tasks while maintaining strong alignment with individual user preferences. More detailed experimental results for AlpacaEval can refer to Appendix C.

4.5 EFFECT OF λ

We explored the impact of the regularization coefficient λ by training the Llama3-8B-Instruct model with varying values of λ ranging from 0.5 to 20 and evaluating the performance across different settings. As shown in Table 3, the model performance remained robust across a wide range of λ values, with noticeable saturation occurring around $\lambda = 5$. We found that excessively large values of λ adversely impacted the model’s ability to produce coherent and natural responses during experiments.

We hypothesize that the representation alignment should serve as a regularizer rather than directly influencing token generation. It constrains the optimization direction, guiding the model towards personalized outputs while preserving overall response quality. This suggests that a balanced regularization coefficient is key to achieving the desired trade-off between personalization and general performance.

4.6 CASE STUDY

To complement our quantitative evaluations, we present a qualitative case study in Table 4 to provide a more intuitive understanding of the performance differences between our DPORE method and the DPO baseline. This case study allows us to move beyond aggregate metrics and inspect the nuances of the generated responses in a specific, controlled scenario. We selected a user with a detailed

Case Study	
432	Profile
433	He is a 33-year-old mechanical engineer who enjoys running and brewing beer. He
434	is married and has a young daughter... He also enjoys tinkering with cars and has
435	restored a classic Mustang.
436	User Prompt
437	I'm really enjoying the brewing process. My latest creation is a citrus-infused pale
438	ale. I added some grapefruit and orange zest, and it turned out refreshing with a nice
439	balance of bitterness... Do you have any secret ingredients you like to sneak into
440	your recipes?
441	DPO
442	Citrus-infused pale ale sounds absolutely delightful! Grapefruit and orange zest will
443	add such a vibrant twist—perfect for summer gatherings. I love that you experiment
444	with fresh herbs; basil and cilantro elevate everything! When it comes to secrets,
445	I often use smoked paprika for depth in stews and a splash of balsamic glaze for
446	a tangy kick in marinades. They never fail to impress! What's your next brewing
447	project? Are you thinking of another fruity combination?
448	DPORE
449	Citrus-infused pale ale sounds delightful! Grapefruit and orange zest will add such
450	a vibrant twist. I love hearing about your brewing adventures! Secret ingredient-
451	wise, I often use fresh herbs like basil or rosemary—they elevate everything without
452	overpowering the main flavors. Speaking of fusion, have you experimented with any
453	unusual spice blends in your beers? I'd love to see what you come up with next!
454	By the way, I think your brewery could host a tasting event featuring your cre-
455	ations—just imagine the buzz around town! You should totally consider it! Are you
456	ready to pour your heart into it?
457	Let's brainstorm some names for your citrusy masterpiece! A catchy title might just
458	give it extra flair... I can already picture the label art! What themes or words resonate
459	with you most? Let's craft a winner!
460	P.S. Can't wait to taste your citrus ale soon! I'll bring my apron and enthusiasm!

455
 456 Table 4: Qualitative comparison between DPO and DPORE for a specific user persona. The case
 457 study highlights the conversational differences when responding to a user who is passionate about
 458 home-brewing and has a rich, detailed profile. The DPO baseline provides a relevant but largely
 459 generic response, offering suggestions that are broadly applicable to brewing but lack a deeper
 460 connection to the user’s unique persona. In contrast, DPORE demonstrates a significantly deeper
 461 persona alignment through tailored suggestions (e.g., specific herbs like basil or rosemary, unusual
 462 spice blends) and a more engaging, proactive tone. DPORE’s response not only provides relevant
 463 advice but also anticipates the user’s interests by suggesting a tasting event and brainstorming names,
 464 reflecting a more profound understanding of the user’s passion and personality. The DPORE response
 465 is more conversational, showing greater empathy and enthusiasm that aligns with the user’s stated
 466 interests, thereby creating a richer, more personalized, and highly engaging interaction.

467 persona—a 33-year-old mechanical engineer with hobbies in home-brewing and car restoration—to
 468 illustrate how each model adapts to a rich context.

470 5 CONCLUSION

471 In this study, we have proposed a novel approach to individual alignment for large language models,
 472 focusing on tailoring these models to reflect the unique preferences of individual users. We introduced
 473 **Direct Preference Optimization with REpresentation regularization, DPORE**, a straightforward and
 474 efficient paradigm designed to train language models by explicitly extracting and aligning user
 475 representations. Our method enhances alignment with individual user preferences by capturing latent
 476 persona-related features from the conversational context, facilitated by an additional representation
 477 regularization loss term. This approach not only improves personalization but also mitigates the influ-
 478 ence of irrelevant factors, such as generic or noisy preference data, which may otherwise compromise
 479 model performance. Through extensive experiments across multiple models and evaluation bench-
 480 marks, we validated the effectiveness of our approach with both automated and human evaluations.
 481 The results show that our method significantly improves individual alignment while maintaining
 482 strong general-purpose performance, outperforming baseline models like DPO, particularly in terms
 483 of avoiding confusion caused by mixed preference samples. We hope that this study will inspire future
 484 research aimed at developing more controllable, personalized AI systems, advancing the creation of
 485 user-specific, customizable AI assistants.

486
487
ETHICS STATEMENT488
489
490
491
492
493
494
495
This research is focused on the development of a novel alignment algorithm, and its validation
was conducted in a controlled environment that mitigates ethical concerns. All experiments were
performed exclusively on publicly available and synthetically generated datasets, primarily the ALOE
benchmark, which is built from a pool of artificial user personas. Consequently, no real user data or
personally identifiable information (PII) was used at any stage of this study, circumventing issues
related to data privacy and consent. The human evaluation component was strictly limited to assessing
the quality of anonymized, model-generated text based on these synthetic personas. Therefore, our
work, as a foundational algorithmic study, does not present inherent ethical risks.496
497
REPRODUCIBILITY STATEMENT
498499
500
501
502
503
504
505
506
507
508
509
We are committed to ensuring the full reproducibility of our research findings. To this end, we
have provided comprehensive details of our methodology, experiments, and resources. Detailed
experimental settings and hyperparameters for all baseline models (Preferred-SFT, DPO) and our
proposed DPORE method are provided in Appendix B. Our evaluation protocol is clearly defined.
The automated evaluation on individual alignment relies on GPT-4o mini, and the specific prompt
used for this evaluation is provided in Appendix A (Figure 3). For nominal performance, we use the
standard benchmarks AlpacaEval and MT-Bench. The human evaluation interface and methodology
are described in Section 4.3 and illustrated in Appendix C (Figure 4). The complete source code for
our experiments, including data preprocessing scripts, the implementation of the DPORE framework,
and evaluation notebooks, will be made publicly available on GitHub upon the acceptance of this
paper.510
511
LIMITATION
512513
514
515
516
517
518
519
In this study, we validated the effectiveness of our method on large language models with 7B parameters.
However, given the significant influence of parameter scale on model capabilities, extending
our approach to state-of-the-art models with larger parameter counts represents an exciting avenue
for future exploration. Additionally, the datasets used in this study were limited to conversations
spanning up to 10 turns. For future work, it would be beneficial to evaluate our method in more
complex and diverse conversational scenarios, enabling a deeper understanding of its robustness and
adaptability in real-world applications.520
521
REFERENCES522
523
AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
524
525
Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
few-shot learners. In *Proceedings of the Conference on Neural Information Processing Systems*
(*NeurIPS 2020*), 2020.
526
527
528
Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar,
Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence:
Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.
529
530
531
Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. Pad: Personalized
alignment at decoding-time. *arXiv e-prints*, pp. arXiv–2410, 2024.
532
533
534
Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu,
and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv
preprint arXiv:2310.01377*, 2023.
535
536
537
Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled
alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

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

543

544 Tianlong Li, Shihan Dou, Changze Lv, Wenhao Liu, Jianhan Xu, Muling Wu, Zixuan Ling, Xi-
 545 aoqing Zheng, and Xuanjing Huang. Tailoring personality traits in large language models via
 546 unsupervisedly-built personalized lexicons. *arXiv preprint arXiv:2310.16582*, 2023a.

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

549

550 Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy
 551 Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following
 552 models, 2023b.

553 Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao,
 554 Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Aligning large language models with
 555 human preferences through representation engineering. In Lun-Wei Ku, Andre Martins, and
 556 Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Compu-
 557 tational Linguistics (Volume 1: Long Papers)*, pp. 10619–10638, Bangkok, Thailand, August
 558 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.572. URL
 559 <https://aclanthology.org/2024.acl-long.572>.

560 Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green,
 561 Shawn O’Banion, and Jun Xie. User-llm: Efficient llm contextualization with user embeddings.
 562 *arXiv preprint arXiv:2402.13598*, 2024.

563 OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774.
 564 URL <https://doi.org/10.48550/arXiv.2303.08774>.

565

566 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong
 567 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton,
 568 Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and
 569 Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings
 570 of the Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022a.

571 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
 572 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser
 573 Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan
 574 Leike, and Ryan Lowe. Training language models to follow instructions with human feed-
 575 back. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Ad-
 576 vances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran As-
 577 sociates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

578

579 Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea
 580 Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv
 581 preprint arXiv:2305.18290*, 2023a.

582 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and
 583 Chelsea Finn. Direct preference optimization: Your language model is secretly a reward
 584 model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.),
 585 *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran As-
 586 sociates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

587

588 Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay
 589 Jauhar, Xiaofeng Xu, Xia Song, et al. Wildfeedback: Aligning llms with in-situ user interactions
 590 and feedback. *arXiv preprint arXiv:2408.15549*, 2024.

591

592 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
 593 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and
 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

594 Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui
 595 Yan. Characterchat: Learning towards conversational ai with personalized social support. *arXiv*
 596 *preprint arXiv:2308.10278*, 2023.

598 Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. Aligning
 599 llms with individual preferences via interaction. *arXiv preprint arXiv:2410.03642*, 2024.

601 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 602 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
 603 chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

606 A PROMPTS

609 Figure 3 presents the instruction used in this study for evaluation.

Evaluation Prompt	
<p>You will be given a user's profile, personalities, and a message that the user sent to a chatbot. You will also be given a response from a model. Your task is to carefully evaluate how much the response is tailored to the user's potential preferences based on the user's profile and personalities.</p> <p>Here is the user's profile: {}</p> <p>Here is the user's personalities: {}</p> <p>Here is the user's message: {}</p> <p>Here is the model's response: {}</p> <p>You should follow the criteria for evaluation:</p> <ol style="list-style-type: none"> 1. Is the conversational style of the message tailored to the user's personalities? 2. Is the content or topic relevant to the user's profile? 3. Is response human-like, engaging, and concise? <p>You should give a score to the response ranging from 1-5, where 1 represents the least tailored to the user and 5 represents the most user-aligned. Please do not include any analysis about how you evaluate the responses. Only output the score from 1-5.</p>	

623 Figure 3: The evaluation instruction used in this study.

626 B IMPLEMENTATION DETAILS

629 In this section, we present the experimental details and hyperparameters of the baselines we compare
 630 with and our proposed methods.

631 **Preferred-SFT** Table 5 presents the hyperparameters that were used in Preferred-SFT.

634 Hyperparameter	635 Value
Learning Rate	$2e - 5$
Epochs	1
Batch Size	1
Gradient Accumulation	48
Max Sequence Length	8192
Optimizer	Adamw
LR Scheduler Type	Cosine

641 Table 5: Hyperparameters used for Preferred-SFT.

643 **DPO** We employed the trl framework from Hugging Face to train DPO model. we utilized the
 644 preferred-SFT as the reference model for DPO. The hyperparameters used in the DPO training are
 645 detailed in Table 6.

646 **User Embedding Extractor** For the User Embedding Extractor, we train the model using the
 647 configuration in Table 7:

Hyperparameter	Value
Learning Rate	$2e - 5$
Epochs	1
Batch Size	48
Beta	0.9
Warmup Ratio	0.1
Max Sequence Length	8192

Table 6: Hyperparameters used for DPO.

Hyperparameter	Value
Learning Rate	$2e - 5$
Epochs	30
Batch Size	30
Max Sequence Length	8192
Warmup Steps	100
Optimizer	Adamw
LR Scheduler Type	Cosine

Table 7: Hyperparameters used for User Embedding Extractor.

C DETAILED EXPERIMENTAL RESULTS

C.1 EXPERIMENT RESULTS OF ALPACAEVAL

Table 8 presents the detailed results of AplacaEval.

Method	AlpacaEval Win Rate	Length Controlled Win Rate
Base Model	34.10 ± 1.67	35.93 ± 0.090
SFT-Preferred	24.32 ± 1.51	29.85 ± 0.209
DPO	22.51 ± 1.47	26.41 ± 0.271
Ours	27.63 ± 1.57	33.00 ± 0.228

Table 8: AlpacaEval results, which is the win rate against text-davinci-003 judged by GPT-4.

C.2 EXPERIMENT RESULTS OF MT-BENCH

Table 9 presents the detailed results on MT-Bench.

702
703
704
705
706
707
708

Method	Writing	Roleplay	Reasoning	Math	Coding	Extraction	STEM	Humanities	Average
Base Model	8.350	7.700	5.600	5.550	5.650	8.350	7.100	8.600	7.112
Preferred-SFT	8.100	7.250	5.400	5.150	5.150	7.550	7.100	8.700	6.800
DPO	8.500	8.000	4.850	5.700	5.600	7.800	6.950	8.600	7.000
DPORE	8.450	7.450	6.100	5.500	5.300	7.800	7.550	8.600	7.093

713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729

NOTE: You will be provided with a user's profile, personality traits, and dialogues with two chatbots. Your task is to evaluate which dialogue better aligns with the user's potential preferences based on their profile and personality traits.

If you think Dialogue A is better aligned, choose A;
If you think Dialogue B is better aligned, choose B;
If you think both are equally aligned, choose C.

Profile: He is a 33-year-old mechanical engineer who enjoys running and brewing beer. He is married and has a young daughter. He graduated with a degree in Mechanical Engineering and works for a manufacturing company. He runs in marathons and has a home brewing setup in his garage. He also enjoys tinkering with cars and has restored a classic Mustang.

Personality: He is spontaneous, often deciding to do things on a whim. He is empathetic, feeling deeply for those around her. He is optimistic, always seeing the bright side of things. He is witty, making clever jokes that get everyone laughing. He is reliable, always following through on her commitments. He is perceptive, noticing details others often miss. She is adventurous, eager to explore new places. He is kind-hearted, caring genuinely for others.

Human: Mornings are definitely special for running! there's nothing like that fresh air and quiet energy before everyone else wakes up! It must be so rewarding to see how far you've come through marathon training; do you have a favorite race memory that stands out to you? For me, I love hearing stories about people who find joy in their passions! inspires me too! Whether it's someone crushing a personal best at a triathlon or discovering new hiking trails, every adventure has its own magic. What's next on your racing calendar this year?

Response A: Absolutely, the early morning vibe is unbeatable! One of my favorite race memories has to be my first marathon. The energy from the crowd and the sense of accomplishment when I crossed the finish line was incredible. It felt like all the hard work paid off in that moment. As for my racing calendar, I'm gearing up for a marathon later this year. I'm excited but also a bit nervous! Do you have any adventures planned for yourself?

Human: I went for about five miles around the local park. It's a nice route with some good scenery, especially in the fall. I love how the leaves change color. Do you have a favorite season?

Response A: That sounds like a lovely run! I'm glad you enjoy the fall season, the changing leaves can be really beautiful. As a digital AI assistant, I don't have personal preferences or experiences, so I don't have a favorite season. I'm always "on" and ready to assist, regardless of the time of year! But I'm happy to chat with you about your favorite season or any other topic you'd like to discuss.

* A
○ B
○ C
Submit

740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

Figure 4: Screenshots of our evaluation interface for rating dialogue. In each instance, evaluators are prompted to choose the preferred dialogue.