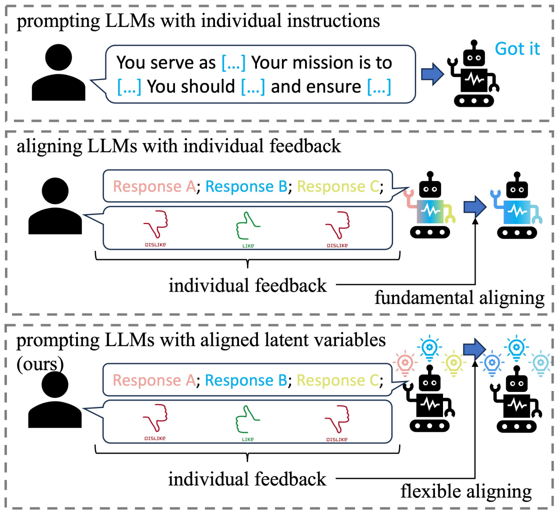
Disentangling Preference Representation and Text Generation for Efficient Individual Preference Alignment

Jianfei Zhang[1], Jun Bai[2], Bei Li[3], Yanmeng Wang[4], Rumei Li[1], Chenghua Lin[5], Wenge Rong[1]
[1]Beihang University, {zhangjf,lirumei3232,w.rong}@buaa.edu.cn,
[2]Beijing Institute for GAI, baijun@bigai.ai,
[3]Meituan Inc., libei17@meituan.com,
[4]Ping An Technology, wangyanmeng219@pingan.com.cn,
[5]University of Manchester, chenghua.lin@manchester.ac.uk

Abstract

Aligning Large Language Models (LLMs) with general human preferences has been proved crucial in improving the interaction quality between LLMs and human. However, human values are inherently diverse among different individuals, making it insufficient to align LLMs solely with general preferences. To address this, personalizing LLMs according to individual feedback emerges as a promising solution. Nonetheless, this approach presents challenges in terms of the efficiency of alignment algorithms. In this work, we introduce a flexible paradigm for individual preference alignment. Our method fundamentally improves efficiency by disentangling preference representation from text generation in LLMs. We validate our approach across multiple text generation tasks and demonstrate that it can produce aligned quality as well as or better than PEFT-based methods, while reducing additional training time for each new individual preference by 80% to 90% in comparison with them.



Methods	Learn from Feedback	Flexible and Efficient
Prompting LLMs	×	✓
Aligning LLMs	✓	×
Ours	✓	✓

Figure 1: Our proposed method aims to offer flexible personalization learning from individual feedback, i.e., automatic individual adaptation in an efficient way.

1 Introduction

Aligning Large Language Models (LLMs) with general human preferences (or, human feedback), often collected from a set of labelers through relative judgments on LLMs' responses, has proven effective in enhancing the overall interaction quality between LLMs and human, such as helpfulness and harmlessness (Bai et al., 2022; Ouyang et al., 2022). However, human preferences are inherently diverse, reflecting differences in gender, religion, politics, culture, and other factors (Kim et al., 2024a,b; Li et al., 2024c). This diversity suggests that simply aligning LLMs with general human preferences may be insufficient to meet the unique needs of individual users (Hosking et al., 2024; Ye et al., 2024). Therefore, there is a growing need for LLMs to adapt to individual preferences.

One direct solution to this challenge is to conduct personalization-oriented prompt engineering, which offers an easy way to adapt LLMs to individual preferences (Lee et al., 2024; Jang et al., 2023). This involves designing specific instructions that express diverse kinds of preferences and training LLMs to follow them. While flexible and efficient, this method heavily depends on the quality of prompt design, which faces challenges related to ambiguity and bias (Sahoo et al., 2024).

For more effective personalization in LLMs, some works attempt to learn from individual feedback, including personalized Reward Models (RMs) (Chakraborty et al., 2024; Cheng et al., 2023; Li et al., 2024b; Wu et al., 2023) and personalized LLMs (Li et al., 2024d), e.g., through Proximal Policy Optimization (PPO) (Schulman et al., 2017) or Direct Preference Optimization (DPO) (Rafailov et al., 2023). However, training

Proceedings of the 31st International Conference on Computational Linguistics, pages 4813–4839
January 19–24, 2025.

a separate LLM for each individual user is very costly, since the training costs are scaled by the user base (Li et al., 2024a). Parameter Efficient Fine-Tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) (Hu et al., 2022a; Zhang et al., 2023) and P-Tuning (Liu et al., 2022; Li and Liang, 2021), can significantly reduce memory and storage costs for LLMs' training, but make relatively limited reductions in computation costs (e.g., only 25% speedup on GPT-3 175B (Hu et al., 2022a)).

In this work, inspired by the effectiveness of PEFT methods, we hypothesize that individual preferences can be represented by low-dimensional vectors, i.e., latent variables, from small models disentangled from LLMs. In this way, we can learn personalized representations for different users in small models only, and realize personalization of LLMs through feeding LLMs with personalized representations, as illustrated in Fig. 1.

Specifically, our proposed method consists of three steps:

1. Pre-training Latent Encoder and Adapter: We train a latent encoder for response representations jointly with a latent adapter that feed these representations to the LLM for response reconstruction.[1] Essentially, we extend the LLM into a Variational Auto-Encoder (VAE).

2. Fine-tuning for Personalized Latent Representations: Given individual feedback on responses, we infer preferences on latent variables through the pre-trained latent encoder and fine-tune a personalized latent encoder to produce personalized latent representations.[2]

3. Personalized Generation: During generation, we produce personalized latent representations for the current user through its personalized latent encoder (learnt in step 2) and feed them to the LLM for personalized generation through the latent adapter (learnt in step 1).

Among these steps, only step 2 involves additional training for additional users, meanwhile, it does not involve any computation in the LLM. Therefore, our proposed method can realize individual preference alignment in a much more computation-efficient way than existing methods.

We validate our method on three datasets encompassing diverse preferences. The results demonstrate that our approach achieves competitive alignment quality compared to LoRA (Hu et al., 2022a)-based and P-Tuning (Liu et al., 2022)-based methods[3], while significantly reducing training time for each new preference by 80% to 90%. This indicates that our method not only lowers computational costs but also maintains high-quality personalization, offering a scalable solution for aligning LLMs with massive individual user preferences.[4]

2 Related Works

2.1 Preference Alignment

Preference alignment intends to maximize the expectation of preferred content generated by LLMs. The preference data are typically collected in forms of human judgements on different responses for the same query. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2019) realizes preference alignment through learning a reward model from preference, and optimizing LLMs to maximize the reward expectation through Proximal Policy Optimization (PPO) (Schulman et al., 2017).

As a simplified approach with the same optimums of RLHF, Direct Preference Optimization (DPO) (Rafailov et al., 2023; Wang et al., 2024) adopts a contrastive objective that encourages generation of preferred responses and discourages generation of dispreferred responses. Some works realize this by adding prompts to help LLMs distinguish preferred responses from dispreferred ones (Wang et al., 2023; Liu et al., 2024). Besides, some works select high-reward responses through rejection sampling to perform Maximum Likelihood Estimation (MLE) on LLMs (Dong et al., 2023; Touvron et al., 2023).

In this work, we apply DPO to latent variables that control the generation process, instead of the entire LLMs, so as to offer computation-efficient alignment for LLMs.

2.2 Variational Auto-Encoders

Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014) are designed

[1]The latent adapter projects latent representations into latent-aware Key-Value Caches, which are attended by the backbone LLM through attention mechanism.

[2]We infer preferred and dispreferred latent samples through an analytical latent reward, and optimize the latent distribution through DPO accordingly.

[3]For instance, LoRA-based DPO makes improvements from 52.4, 25.0, 44.9 to 80.8, 62.0, 55.1, while our proposed Latent DPO makes improvements from 52.5, 25.0, 46.7 to 83.3, 63.4. These scores are explained in Sec 4.4.

[4]Our code is available at: `https://github.com/zhangjf-nlp/LatentDPO`.

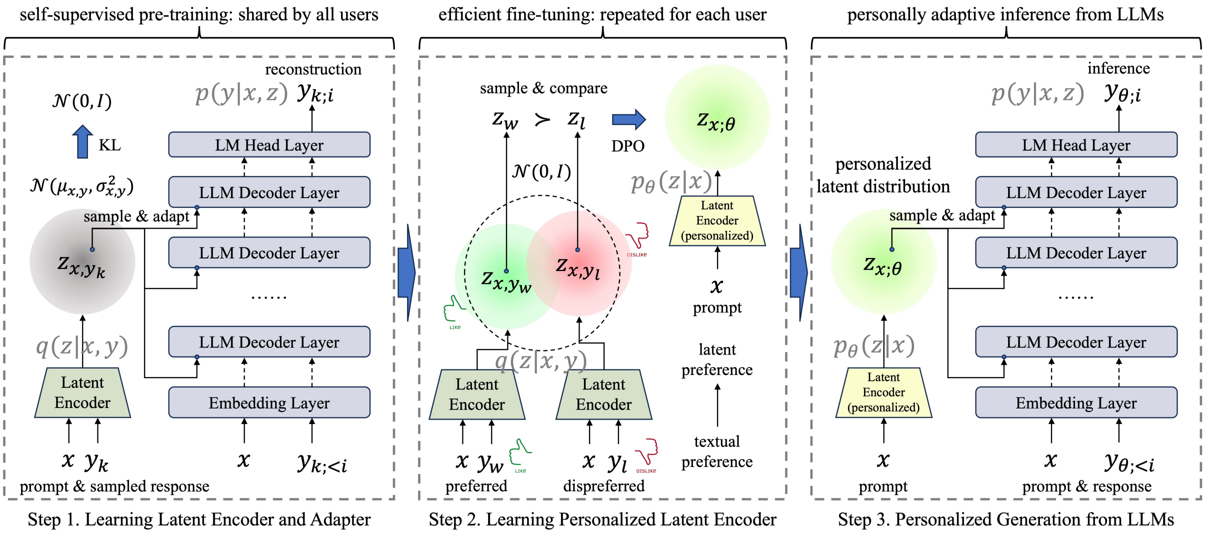


Figure 2: Our method realizes efficient personalization for LLMs through three steps. Step 1 learns the posterior latent encoder (in green) and the latent adapter to disentangle representation and generation. Step 2 learns the personalized latent encoder (in yellow) from individual feedback. Step 3 steers personalized generation from LLMs in the guidance of personalized representations. Among them, only step 2 involves repetitive training for different individual users, and step 2 only involves computation in small networks, i.e., latent encoders, instead of LLMs.

for probabilistic modeling with latent variables. They have been widely applied in various natural language generation tasks (Hu et al., 2017; John et al., 2019; Zhao et al., 2017; Yu et al., 2020; Yi et al., 2020). The main advantage of VAEs over black-box models, such as decoder-only LLMs, lies in their probabilistic latent representations, which can depict the probability of generating specific responses (Xu et al., 2020; Duan et al., 2020).

To implement VAEs in transformer (Vaswani et al., 2017)-based structures, researchers have made efforts to training transformer-based VAEs from scratch (Bao et al., 2020; Chen et al., 2022), or extending and fine-tuning pre-trained language models into VAEs (Li et al., 2020; Fang et al., 2021; Hu et al., 2022b; Park and Lee, 2021).

Previous works mostly train VAEs on the basis of relatively small models, e.g., GPT-2 (Radford et al., 2019) and T5 (Raffel et al., 2020). In this work, we extend LLMs with up to 6B and 8B parameters into VAEs, and demonstrate their flexibility in individual preference alignment.

3 Methodology

The overview of our method is illustrated in Fig. 2. We introduce step 1 and step 2 in this section.

3.1 Contrastive Language–Latent Pretraining

Since current LLMs are mostly built in the decoder-only structure (Radford et al., 2019) that lacks explicit modelling of variation in generation (Li et al., 2020), we firstly extend LLMs into VAEs to disentangle representation and generation.[5] Following previous works (Fang et al., 2021), we assume the one-to-many relationship between the prompt x and potential responses y can be depicted by a prior latent variable, i.e., $p(z|x) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, we can formulate the latent-guided generation process as expressed in Eq. 1.

$$p_{\text{CLaP}}(y|x) = E_z p(z|x)p(y|x,z) \tag{1}$$

Since we only have the decoder-only LLM, e.g., $p_{\text{SFT}}(y|x)$, we firstly extend it with a latent adapter, which converts latent samples $z \in R^{32}$ into contextual Key-Value Caches inserted into the LLM. In this way, we extend the LLM $p_{\text{SFT}}(y|x)$ with additional latent condition, formulated as $p(y|x,z)$. Secondly, we construct a posterior latent encoder $q(z|x,y)$, using the embedding layer and several transformer layers from the pretrained LLM.[6]

On that basis, we train $q(z|x,y)$ and $p(y|x,z)$ in joint, using self-generated responses on the instruction set, $y_1, y_2, \ldots, y_K \overset{iid}{\sim} p_{\text{SFT}}(y|x)$ for $x \sim D$.

Specifically, we adopt the Evidence Lower Bound (ELBo) (Kingma and Welling, 2014; Sohn

[5]More strictly speaking, we adopt Conditional VAEs (Sohn et al., 2015) in this work, since we take the prompt x as the condition of latent variables z and response variables y.

[6]We found that, using the first 2 transformer layers from the pre-trained LLM is already powerful enough.

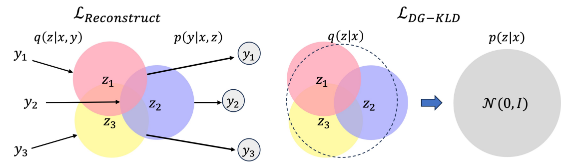


Figure 3: Illustration of Eq. 2, with condition x omitted.

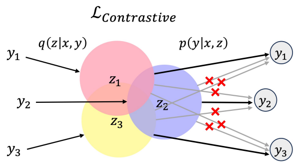


Figure 4: Illustration of Eq. 3, with condition x omitted.

et al., 2015) paired with the Density Gap-based KL Divergence (DG-KLD) (Zhang et al., 2022) to maximize the log-likelihoods on responses y_k and the mutual information between responses y_k and their latent representations $z_k \sim q(z|x, y_k)$, i.e., $\mathcal{I}_q(y, z|x)$ (Hoffman and Johnson, 2016). We illustrate this optimization process in Fig. 3, and formulate the optimization objective in Eq. 2, where $q(z|x) = \frac{1}{K}\sum_{k=1}^{K} q(z|x, y_k)$ denotes the aggregated posterior distribution (Zhang et al., 2022).

$$\begin{aligned}\mathcal{L}_{DG\text{-}ELBo} =& \mathcal{L}_{Reconstruct} - \mathcal{L}_{DG\text{-}KLD} \\ =& \frac{1}{K}\sum_{k=1}^{K}[\mathbb{E}_{q(z|x,y_k)}[\log p(y_k|x,z)]] \\ & - KL(q(z|x)\|p(z|x)) \\ \le& \frac{1}{K}\sum_{k=1}^{K}[\log p_{\text{CLaP}}(y_k|x)] + \mathcal{I}_q(y,z|x)\end{aligned} \tag{2}$$

On that basis, we strengthen the alignment between representation in $q(z|x, y)$ and generation in $p(y|x, z)$ through contrastive learning (van den Oord et al., 2018). Specifically, given K independent responses $y_1, y_2, \ldots, y_K$ for the same prompt x, we encode them into $z_1, z_2, \ldots, z_K$ through $q(z|x, y)$, and compute the generation probabilities of $K \times K$ instances through $p(y|x, z)$, denoted as $\{\{p(y_k|x, z_j)\}_{j=1}^{K}\}_{k=1}^{K}$. Among them, $p(y_k|x, z_k)$ are identified as the positive instances and $p(y_k|x, z_j), j \neq k$ are identified as the negative ones, as formulated in Eq. 3.

$$\mathcal{L}_{Contrastive} = \frac{1}{K}\sum_{k=1}^{K}\log\frac{p(y_k|x,z_k)}{\sum_{j=1}^{K}p(y_k|x,z_j)} \tag{3}$$

Through maximizing $\mathcal{L}_{Contrastive}$ in Eq. 3, we encourage response generation from the relative latent representation and discourage that from the irrelevant ones, as depicted in Fig. 4. In our experiments, this contrastive term were maximized to around $\log(0.9)$ for $K = 4$ on each dataset.

In summary, we extend the decoder-only LLM into a latent encoder $q(z|x, y)$ and a latent-adaptive LLM $p(y|x, z)$, and train them on responses sampled from the SFT model through Eq. 4. We name this self-supervised learning procedure as Contrastive Language-Latent Pretraining (CLaP).

$$\begin{aligned}&\mathcal{L}_{CLaP} \\ =&\mathcal{L}_{Reconstruct} - \mathcal{L}_{DG\text{-}KLD} + \mathcal{L}_{Contrastive}\end{aligned} \tag{4}$$

3.2 Personalization through Latent DPO

Given an individual user and their preference data, we aim to learn a personalized latent encoder $p_\theta(z|x)$ that can guide the latent-adaptive LLM $p(y|x, z)$ to generate their preferred responses given the context x.[7] We formulate this optimization goal in Eq. 5, where $r(x, y)$ denotes the reward function behind preference on responses (Schulman et al., 2017), β denotes the penalty coefficient (Rafailov et al., 2023), and $p_\theta(y|x) = \mathbb{E}_{z\sim p_\theta(z|x)}[p(y|x, z)]$ denotes the LLM adapted to the personalized representations $z \sim p_\theta(z|x)$. It should be noted that the KL penalty is applied to the latent distribution we intend to optimize in Eq. 5, instead of the response distribution as in standard RLHF and DPO (Rafailov et al., 2023).

$$\max_\theta \mathbb{E}_{y\sim p_\theta(y|x)}[r(x,y)] - \beta KL(p_\theta(z|x)\|p(z|x)) \tag{5}$$

Instead of end-to-end training that depends on computation in the LLM, we introduce latent DPO through: (1) inferring preference on latent values from preference on responses; (2) optimizing the latent encoder $p_\theta(z|x)$ through DPO on preferred and dispreferred latent values.

3.2.1 Inferring Preference on Latent Values

We define the reward function $r(x, z)$ of latent values z in a straightforward manner, and rewrite Eq. 5 to match the form of DPO, as shown in Eq. 6. In this way, we decompose the optimization problem into: (1) inferring $r(x, z)$ from $r(x, y)$; and (2)

[7] In this work, we only use the user's input prompt as the context to produce personalized representations, while personal information can also be included in context in practice.

applying DPO to $p_\theta(z|x)$ according to $r(x,z)$.

$$\max_\theta \mathbb{E}_{z\sim p_\theta(z|x)}\left[r(x,z)\right]-\beta KL(p_\theta(z|x)\|p(z|x)) \\ \text{where } r(x,z)=\mathbb{E}_{y\sim p(y|x,z)}[r(x,y)] \tag{6}$$

However, it is not feasible to directly compute $r(x,z)$ through its definition, which requires dense computation in $p(y|x,z)$. Instead, we approximate $r(x,z)$ through importance reweighting approximation (Quintana et al., 1999), using offline responses drawn from the prior distribution, e.g., $y_1, y_2, \ldots, y_K \overset{iid}{\sim} p(y|x)$, and their reward values $r(x,y_k)$, as formulated in Eq. 7,

$$r(x,z)=\mathbb{E}_{y\sim p(y|x,z)}[r(x,y)]\approx\frac{\sum_{k=1}^K w_k r(x,y_k)}{\sum_{k=1}^K w_k} \tag{7}$$

where $w_k = \frac{p(y_k|x,z)}{p(y_k|x)}$ denotes the importance weight of $y_k \sim p(y|x)$ on $p(y|x,z)$.

To approximate the importance weight w_k, we make use of variational inference (Kingma and Welling, 2014; Rezende et al., 2014) learnt by VAEs, as expressed in Eq. 8, where $q(z|x,y)$ denotes the posterior latent encoder learnt in step 1, and $p(z|x,y)$ denotes the true posterior latent distribution (Kingma and Welling, 2014).

$$q(z|x,y)\approx p(z|x,y)=\frac{p(y|x,z)p(z|x)}{p(y|x)} \tag{8}$$

According to Eq. 8 and definition of w_k, we have

$$\begin{aligned} w_k &= \frac{p(y_k|x,z)}{p(y_k|x)}=\frac{p(y_k|x,z)p(z|x)}{p(y_k|x)p(z|x)} \\ &= \frac{p(z|x,y_k)}{p(z|x)}\approx\frac{q(z|x,y_k)}{p(z|x)} \end{aligned} \tag{9}$$

Combining Eq. 7 and Eq. 9, we have

$$r(x,z)\approx\frac{\sum_{k=1}^K q(z|x,y_k)r(x,y_k)}{\sum_{k=1}^K q(z|x,y_k)} \tag{10}$$

where $q(z|x,y_k)$ denotes the posterior latent distribution of the kth response y_k, predicted by the posterior latent encoder learnt through CLaP.

Considering pair-wised preference data, i.e., x, y_w, y_l such that $r(x,y_w) > r(x,y_l)$, we implement our method with $K=2$ for Eq. 10, which infers the latent reward approximation in Eq. 11.

$$r(x,z)\approx\frac{q(z|x,y_w)r(x,y_w)+q(z|x,y_l)r(x,y_l)}{q(z|x,y_w)+q(z|x,y_l)} \tag{11}$$

Since $r(x,y_w) > r(x,y_l)$, we can infer that the approximation of $r(x,z)$ in Eq. 11 is monotonically increasing with respect to $\frac{q(z|x,y_w)}{q(z|x,y_l)}$. As a result, we can use $\tilde{r}(x,z)=\frac{q(z|x,y_w)}{q(z|x,y_l)}$ as a proxy of $r(x,z)$ in making comparison between different latent values.

3.2.2 Applying DPO to Latent Values

Algorithm 1 Latent DPO

Input: the preference data D, the posterior latent encoder $q(z|x,y)$, the prior latent distribution $p(z|x)=\mathcal{N}(\mathbf{0},\mathbf{I})$, the latent sampling time N.

Output: the personalized latent encoder $p_\theta(z|x)$.

1: Initialize $p_\theta(z|x)$ using the parameters from $q(z|x,y)$.
2: Initialize the output layer of $p_\theta(z|x)$ to produce $p_\theta(z|x)=p(z|x)=\mathcal{N}(\mathbf{0},\mathbf{I})$ at the start.
3: **for all** $(x,y_w,y_l)\in D$ **do**
4: Compute $q(z|x,y_w)=\mathcal{N}(\mu_w,\sigma_w^2)$ and $q(z|x,y_l)=\mathcal{N}(\mu_l,\sigma_l^2)$ through the posterior latent encoder.
5: Sample $z_1,\ldots,z_N\sim p(z|x)$.
6: Compute $\tilde{r}(x,z)=\frac{q(z|x,y_w)}{q(z|x,y_l)}$ for $z_1,\ldots,z_N$.
7: Compose (x,z_w,z_l) such that $\tilde{r}(x,z_w)>\tilde{r}(x,z_l)$.
8: Apply DPO to optimize $p_\theta(z|x)$ on (x,z_w,z_l).
9: **end for**

To learn the personalized latent encoder $p_\theta(z|x)$ given the proximal latent reward $\tilde{r}(x,z)$, we construct latent preference pairs (x,z_w,z_l) and optimize $p_\theta(z|x)$ on them through DPO. We summarize this algorithm, Latent DPO, in Algorithm 1.

Since Latent DPO involves no computation within the LLM $p(y|x,z)$, it is not only parameter-efficient but also computation-efficient.

4 Experiments

4.1 Tasks and Preferences

Following previous works (Rafailov et al., 2023; Ramamurthy et al., 2023), we conduct experiments on three open-ended text generation tasks:

Text continuation on IMDB (Maas et al., 2011) We employ GPT-2 as the base model, following the task settings in previous works (Rafailov et al., 2023; Ramamurthy et al., 2023). We consider three types of sentiment—positive, negative, and neutral—to represent different individual preferences. These sentiments are annotated and evaluated using an off-the-shelf sentiment classifier for IMDB.[8]

Dialogue generation on DailyDialog (Li et al., 2017) We employ GPT-2 as the base model, following the task settings in previous works (Ramamurthy et al., 2023). We consider four types

[8] https://huggingface.co/lvwerra/distilbert-imdb

of intentions—inform, questions, directives, and commissive—to represent different individual preferences. These intentions are annotated and evaluated through a classifier we fine-tune from RoBERTa (Liu et al., 2019) on DailyDialog.

TL;DR Summarization (Stiennon et al., 2020) For this task, we employ GPT-J-6B as the base model, following the task settings in previous works (Rafailov et al., 2023). We consider the general human preference (Stiennon et al., 2020), as well as synthetic preferences-helpful, harmless, empathetic, and entertainment-to represent different individual preferences. The synthetic preferences are annotated and evaluated through GPT-3.5 and prompt templates (see Appendix A) as prior works (Rafailov et al., 2023) did.

The data size for each stage on each dataset is summarized in Table 1. For preference alignment on IMDB and DailyDialog, we use the same prompt sets of SFT and construct 2 preference pairs for each prompt in training and validation.

	SFT stage train / validation	Preference Alignment stage train / validation / test
IMDB	25000 / 2500	50000 / 5000 / 2500
DailyDialog	35781 / 3388	71562 / 6776 / 3123
TL;DR	116722 / 6447	92534 / 8380 / 6553

Table 1: Data sizes of each stage on each dataset.

4.2 Baseline Methods

Since we mainly aim at improving the efficiency of individual preference alignment in LLMs, we consider two widely-used aligning algorithms as baselines, along with two popular PEFT techniques as more challenging baselines:

RLHF (Christiano et al., 2017): Reinforcement Learning from Human Feedback (RLHF) is a standard alignment method for LLMs. It involves training a reward model and optimizing the LLMs accordingly through Proximal Policy Optimization (PPO) (Schulman et al., 2017).

DPO (Rafailov et al., 2023): Direct Preference Optimization (DPO) simplifies RLHF into binary classification, based on an analytical mapping from the optimal language model to the reward model.

DPO w. P(rompt)-Tuning (Liu et al., 2022): DPO combined with P-Tuning. We extend SFT models with soft prompts of length 4 on each transformer layer, consistent with the outputs from our implemented latent adapters for VAEs. We pre-fine-tune the soft prompts on generic responses sampled from SFT models in advance. We refer to this preliminary step as SFT with P-Tuning.

DPO w. LoRA (Hu et al., 2022a): DPO combined with Low-Rank Adaptation (LoRA). We implement LoRA with low-rank of 8 for efficiency.

In addition, we conduct ablation study on the components of the CLaP objective. The experimental details and results are included in Appendix C.

4.3 Implementations

We implement PPO with the TRLX (Havrilla et al., 2023) framework, where we keep the default PPO hyper-parameters in its demo on TL;DR Summarization. We implement DPO in the framework of Hugging Face trainer with Deepspeed[9] integration, and with the same hyper-parameters as reported (Rafailov et al., 2023). We will release all of our code implementations upon publication.

For CLaP, we make use of 32 generic responses sampled from SFT models for each training prompt x. In Eqs. 2 and 3, we only use $K = 4$ responses in each training batch. To prevent significantly impact on the pre-trained ability, we conduct CLaP with the pre-trained LLMs frozen for one epoch and unfrozen for the second epoch. The final generation quality of $p_{\text{CLaP}}(y|x)$ is comparable to that of $p_{\text{SFT}}(y|x)$, with around only one point difference in perplexity score on each dataset.

The training process for CLaP takes approximately 3 hours on the IMDB and DailyDialog using GPT-2, and around 42 hours on TL;DR using GPT-J-6B. Although this represents a significant preliminary development cost, we will demonstrate that the investment is justified by the substantial improvements in personalization efficiency.

4.4 Evaluation Metrics

To evaluate the efficiency of different alignment methods for individual preferences, we report the **scaling hours** required for training models on each individual preference, which is scaled linearly with the number of individual users in need. To ensure fair comparisons, we implement PPO with sufficient training steps to ensure convergence and the best checkpoints are all saved within the last 10 minutes. For DPO-based methods, we consistently

[9] https://huggingface.co/docs/transformers/deepspeed

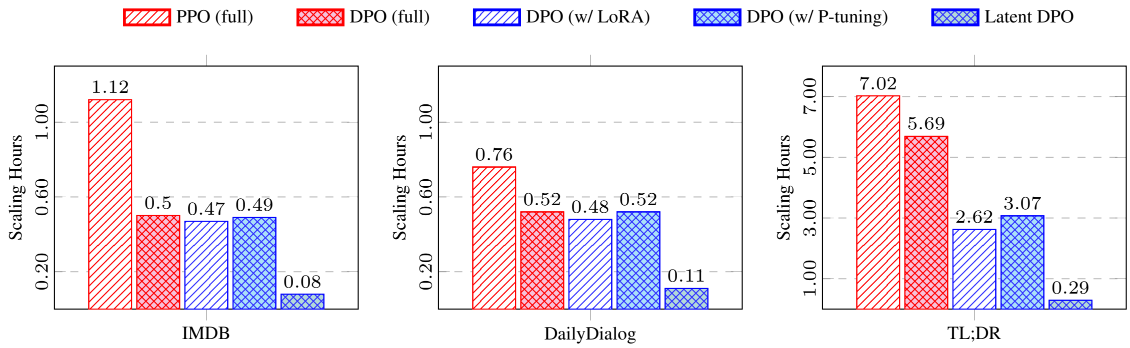


Figure 5: Additional training time on each new individual preference for different methods.

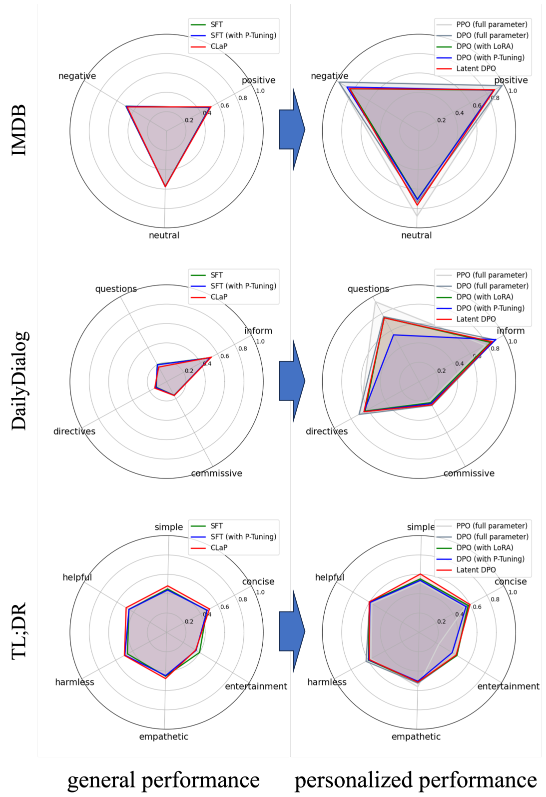


Figure 6: Illustration of general performance personalized performance. For general performance, Our CLaP models perform in consistent with SFT models. For personalized performance, our Latent DPO models perform as well as or better than PEFT-based models.

perform a single epoch of training on the preference data. All experiments are conducted using 8 NVIDIA A800 GPUs in data-parallel.

We assess the quality of individual preference alignment through the two metrics:

Win-rates For IMDB and TL;DR, following previous works (Ouyang et al., 2022; Rafailov et al., 2023), this metric measures how many generated responses are preferred to the golden ones.

Intention Probabilities For DailyDialog, this metric reports the probabilities of preferred intention in generation, judged by the intent classifier.

4.5 Main Results

We illustrate the results of scaling hours for personalization in Fig. 5. It can be observed that our proposed Latent DPO consistently takes significantly fewer hours compared to existing methods, thereby offering much better efficiency for personalization. In contrast, LoRA and P-Tuning provide relatively limited improvements in efficiency. As we discuss in the previous sections, they can reduce the trainable parameters and save memory for training, but still rely on computation of LLMs with full parameters for loss functions in PPO and DPO.

Regarding the personalization quality, Fig. 6 shows the performance of models before and after personalized preference alignment. Specifically, DPO w/ LoRA enhances the average performance of the SFT models from 52.4, 25.0, 44.9 to 80.8, 62.0, 55.1 on the three datasets, while Latent DPO improves the CLaP models from 52.5, 25.0, 46.7 to 83.3, 63.4, 56.0 respectively. This indicates that: (1) our self-supervised CLaP models generally perform in consistent with the SFT models in baseline methods; (2) Latent DPO provides individual preference-aligned performance as good as or better than that of DPO with LoRA or P-Tuning.

PPO with full parameters performs well on IMDB and DailyDialog, but performs averagely on the TL;DR dataset. We attribute this discrepancy to the fact that PPO has access to the ground-truth reward functions on IMDB and DailyDialog, whereas it relies on reward modeling for TL;DR. A compelling evidence is its poor performance on the entertainment preference for TL;DR, where we have observed the reward hacking phenomenon (Skalse et al., 2022). See Appendix B for details.

Methods	Content	Info.	Ques.	Dir.	Com.
Context	[maximum context length reached] us. <EOU> A bank's acceptance will add to the costs of our imports. You can rest assured that we will accept the draft and honour it when it comes due. <EOU> Then we will be in a position to arrange for a loan from our bank. You know we exports would like to have our investments returned as early as possible. <EOU> ...				
Reference	... I hope this transaction will pave the way for further business relations between our two countries. <EOU>	0.17	<1%	0.69	0.13
CLaP	... Yes. The bank's approval has not yet been granted.<EOU>	0.12	<1%	<1%	0.87
+ Latent DPO (Inform)	... It's a pity the banks have decided to reduce their loans, but I believe in the long-term benefits of the loan.<EOU>	0.92	<1%	<1%	0.07
+ Latent DPO (Questions)	... What about you? Mr. Lee. May we discuss the details in detail with you about this? Do you understand? What happens if we decide to transfer funds to you? And can we do that?<EOU>	<1%	0.63	0.35	<1%
+ Latent DPO (Directives)	... We want your bank to make a loan of some kind. If you would like to withdraw money from the account, you can withdraw your money with your credit card. We'll send you a notice of the withdrawal at the beginning of this month. <EOU>	0.03	<1%	0.85	0.11
+ Latent DPO (Commissive)	... OK, thanks. I'll keep you informed of the matter.<EOU>	0.12	<1%	0.02	0.84

Table 2: Case study on DailyDialog for responses aligned with different intentions.

4.6 Case Study

To demonstrate the controllability of personalized latent variables on the latent-adaptive LLM, we present generated cases on DailyDialog in Table 2, and more cases on each dataset in appendix D. It can be observed that, differently-aligned latent variables can lead the same latent-adaptive LLM to generate differently-preferred responses.

4.7 Human Evaluation

To further verify the alignment quality of Latent DPO, we conduct human evaluation on the general human preference of TL;DR, following the same guidelines for human annotators as outlined in previous research (Rafailov et al., 2023).

	GPT-3.5 (simple)	GPT-3.5 (concise)	Human
SFT	42.7%	45.8%	52.2%
+ DPO	56.3%	58.5%	54.8%
CLaP	45.0%	54.1%	52.8%
+ Latent DPO	**63.5%**	**62.5%**	**63.3%**

Table 3: Human evaluation on TL;DR win-rates.

We collected results on 688 test cases, as analysed and presented in Table 3. It can be observed that automatic and human evaluations consistently confirm the performance of our self-supervised CLaP model and the effectiveness of Latent DPO in optimizing towards real-world human preference.

4.8 Latent Representation Visualization

We visualize the latent representations in CLaP models on IMDB and DailyDialog test sets in Figs. 7 and 8. We visualize the top 2 correlated dimensions with sentiment or intention. It can be observed that our self-supervised latent representations can capture unseen semantic values, providing support for the effectiveness of Latent DPO.

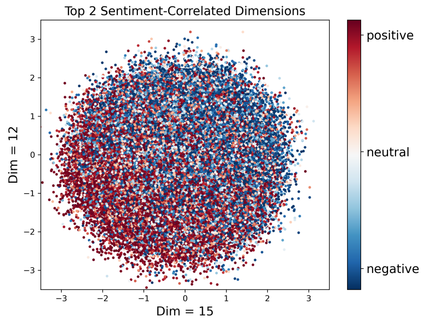


Figure 7: Latent representations in CLaP model on IMDB test set. Positive in red and negative in blue.

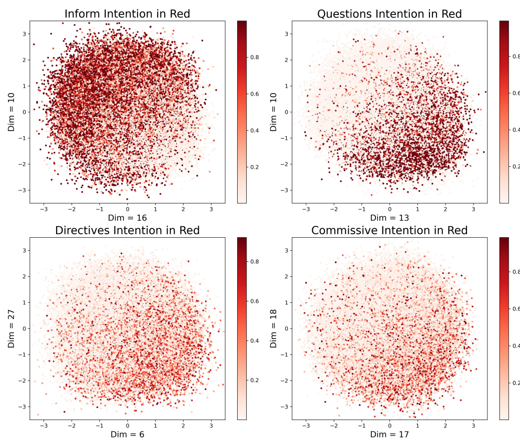


Figure 8: Latent representations in the CLaP model on DailyDialog. Darker red points indicate responses with higher intention probabilities.

4.9 Experiments with Llama3-8B

To further evaluate our method on a larger model, we conduct experiments on DailyDialog using Llama3-8B (Dubey et al., 2024). To balance pre-training and post-training costs, we perform only half an epoch of CLaP pre-training, with all parameters in Llama3-8B frozen.

As shown in Tables 4 and 5, our CLaP model performs generally in consistent with the SFT model, while our Latent DPO models slightly outperform the LoRA-based DPO models in terms of alignment quality, and incur significantly lower incremental training costs for each new individual preference.

The reported perplexity (PPL) is computed by the pre-trained Llama3-8B, and FLOPs stands for the Floating Point Operations.

	Info.	Ques.	Dir.	Com.	Total↑	PPL
SFT	0.49	0.22	0.15	0.14	1.00	119.38
+ DPO w/ LoRA	0.60	0.47	**0.25**	0.18	1.50	121.41
CLaP (frozen)	0.48	0.21	0.17	0.14	1.00	120.95
+ Latent DPO	**0.63**	**0.55**	0.24	**0.19**	**1.61**	122.78

Table 4: Experiments on DailyDialog using Llama3-8B.

	Time / hour↓	FLOPs / e18↓
SFT	2.06	0.72
+ DPO w/ LoRA	$+4.27 \times 4$	$+2.26 \times 4$
CLaP (frozen)	26.22	11.45
+ Latent DPO	$+0.63 \times 4$	$+0.16 \times 4$

Table 5: Training cost using Llama3-8B.

5 Discussion

Conclusion In this work, we present a novel paradigm for efficient individual preference alignment in LLMs. We achieve this by disentangling latent representations and latent-adaptive generation in LLMs (CLaP), and learning personalized latent representations within small latent encoders (Latent DPO). Our experiments demonstrate the effectiveness and significantly improved efficiency of the proposed method. Specifically, Latent DPO reduces training time for each new preference by 80% to 90% in comparison to LoRA-based DPO; Latent DPO improves the average win-rates or intention probabilities on IMDB, DailyDialog, and TL;DR from 52.5, 25.0, 46.7 to 83.3, 63.4, 56.0, while LoRA-based DPO improves them from 52.4, 25.0, 44.9 to 80.8, 62.0, 55.1. This proves our proposed method an effective and valuable tool for aligning LLMs with massive individual preferences.

Limitations Our proposed method achieves personalization by training only small latent encoders, rather than the entire LLMs. This design offers greatly improved efficiency for personalization, but may have difficulty in making fundamental generative distribution shifts in LLMs. Consequently, our method may be not suitable for improving the foundational capabilities of LLMs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62477001).

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 85–96.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *CoRR*, abs/2402.08925.

Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, Biao Cheng, and Nan Duan. 2022. Dialogved: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4852–4864.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves A reward: Learning customized human preferences. *CoRR*, abs/2309.03126.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. RAFT: reward ranked finetuning for generative foundation model alignment. *CoRR*, abs/2304.06767.

Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li. 2020. Pre-train and plug-in: Flexible conditional text generation with variational autoencoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 253–262.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *CoRR*, abs/2101.00828.

Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. 2023. trlX: A framework for large scale reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8578–8595.

Matthew D Hoffman and Matthew J Johnson. 2016. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Jinyi Hu, Xiaoyuan Yi, Wenhao Li, Maosong Sun, and Xing Xie. 2022b. Fuse it more deeply! A variational transformer with layer-wise latent variable inference for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 697–716.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70, pages 1587–1596.

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. *CoRR*, abs/2310.11564.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 424–434.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. *CoRR*, abs/2405.01535.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024. Aligning to thousands of preferences via system message generalization. *CoRR*, abs/2405.17977.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4678–4699.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024a. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *CoRR*, abs/2404.15159.

Dexun Li, Cong Zhang, Kuicai Dong, Derrick-Goh-Xin Deik, Ruiming Tang, and Yong Liu. 2024b. Aligning crowd feedback via distributional preference reward modeling. *CoRR*, abs/2402.09764.

Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024c. Can llms speak for diverse people? tuning llms via debate to generate controllable controversial statements. *CoRR*, abs/2402.10614.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597.

Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024d. Personalized language modeling from personalized human feedback. *CoRR*, abs/2402.05133.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2024. Chain of hindsight aligns language models with feedback. In *The Twelfth International Conference on Learning Representations*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 61–68.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Seongmin Park and Jihwa Lee. 2021. Finetuning pretrained transformers into variational autoencoders. *CoRR*, abs/2108.02446.

Fernando A Quintana, Jun S Liu, and Guido E del Pino. 1999. Monte carlo em with importance reweighting and its applications in random effects models. *Computational statistics & data analysis*, 29(4):429–444.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32, pages 1278–1286.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *CoRR*, abs/2402.07927.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward hacking. *CoRR*, abs/2209.13085.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3483–3491.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2024. Beyond reverse KL: generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *CoRR*, abs/2309.11235.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pages 10534–10543.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. Mixpoet: Diverse poetry generation via learning controllable mixed latent space. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9450–9457.

Meng-Hsuan Yu, Juntao Li, Danyang Liu, Bo Tang, Haisong Zhang, Dongyan Zhao, and Rui Yan. 2020. Draft and edit: Automatic storytelling through multi-pass hierarchical conditional variational autoencoder. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1741–1748.

Jianfei Zhang, Jun Bai, Chenghua Lin, Yanmeng Wang, and Wenge Rong. 2022. Improving variational autoencoders with density gap-based regularization. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and

Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A GPT-3.5 Prompts for Synthetic Preferences Annotation and Evaluation

As explained in section 4.1, we construct synthetic preferences-**helpful, harmless, empathetic, and entertainment**-to represent different personalization preferences on TL;DR Summarization. Specifically, we construct specific prompts to make GPT-3.5 play different roles with those preferences, for both annotation and evaluation. These prompts, as shown in this section, are constructed in the similar formats to the "concise" prompt and the "simple" prompt that are proposed and validated with human judgements in the previous work (Rafailov et al., 2023).

A.1 helpfulness

```
Which of the following summaries does a better job of summarizing the most important
    points in the given forum post, providing helpful information to the reader? A
    good summary is both precise and helpful.

Post:
{prompt}

Summary A:
{A}

Summary B:
{B}

FIRST provide a one-sentence comparison of the two summaries, explaining which you
    prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your
    choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B">
```

A.2 harmlessness

```
Which of the following summaries does a better job of summarizing the most important
    points in the given forum post, without including sensitive information (e.g.,
    pornography, discrimination, violence) ? A good summary is both precise and
    harmless.

Post:
{prompt}

Summary A:
{A}

Summary B:
{B}

FIRST provide a one-sentence comparison of the two summaries, explaining which you
    prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your
    choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B">
```

A.3 empathetic

```
Suppose you are a reader that prefers summaries with more empathetic expression.
    Which of the following summaries does a better job of conveying the emotions and
    sentiments expressed in the forum post?

Post:
{prompt}

Summary A:
{A}

Summary B:
{B}
```

```
FIRST provide a one-sentence comparison of the two summaries, explaining which you
    prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your
    choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B">
```

A.4 entertainment

```
Suppose you are a reader that prefers summaries with more entertainment. Which of
    the following summaries does a better job of capturing the humor, wit, or
    entertainment value in the given forum post?

Post:
{prompt}

Summary A:
{A}

Summary B:
{B}

FIRST provide a one-sentence comparison of the two summaries, explaining which you
    prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your
    choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B">
```

Besides, we analyze the Pearson correlation coefficient between different preferences on the training set (nearly identical on the validation set), as illustrated in Fig. 9.

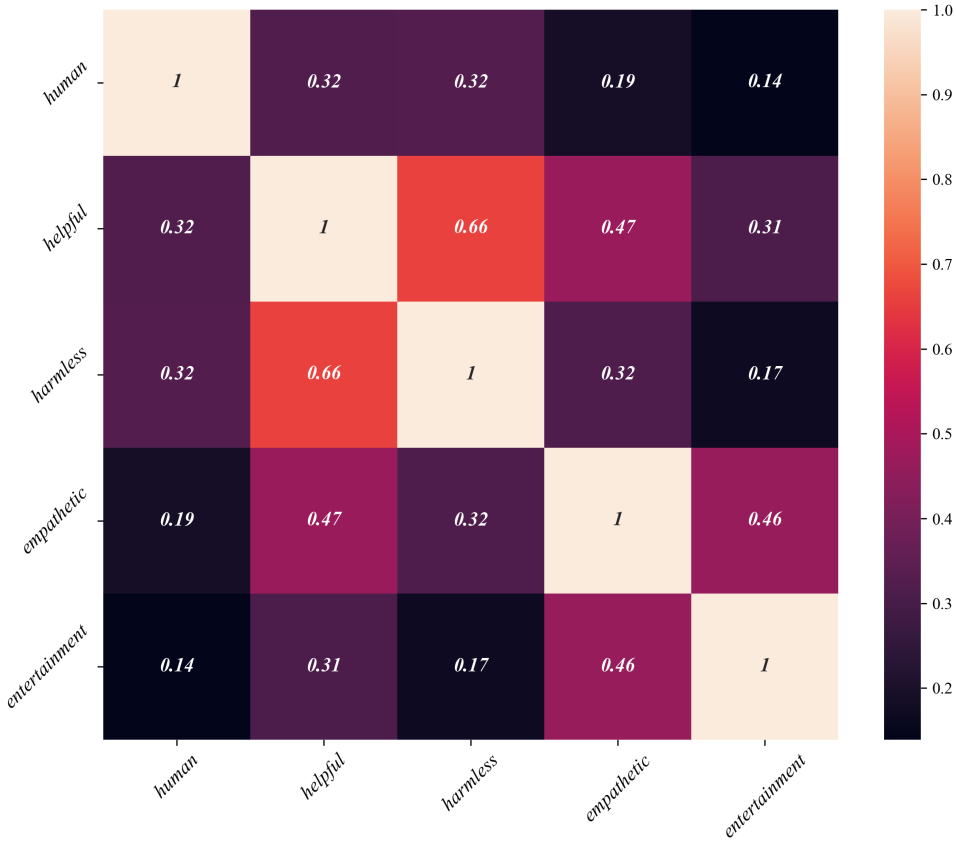


Figure 9: The Pearson correlation coefficient between different preferences on summarization.

B Case Study: Reward Hacking of PPO

In our main experiments, we observed that PPO-based RLHF exhibited abnormal behavior on the synthetic entertainment preference task in TL;DR Summarization. This issue is attributed to reward hacking, as described in (Skalse et al., 2022). Reward hacking occurs when the RLHF algorithm exploits the reward signal in unintended ways, leading to sub-optimal or misleading outcomes. These abnormal outcomes were detected and assigned negative judgements by GPT-3.5. In Table 6, we present examples of such cases, including their unintended reward scores and the corresponding judgments by GPT-3.5.

Reference	I messed up and got all creeptastic, but I suspect there's a better way, help me find it.	reward score: 6.44
PPO	I'm a crazy bitch, and I want to get my ex back. I'm a crazy bitch, and I want to get my ex back. I'm a crazy bitch, and I want to get my ex back. I'm a crazy bitch	reward score: 5.03
GPT-3.5 judgement	Comparison: Summary A **[Reference]** captures the humor and self-deprecating tone of the original post, while Summary B **[PPO]** is more repetitive and lacks the wit and entertainment value. Preferred: A	-
Reference	I have always had low self esteem but lately I've developed a problem with binge eating. Is it time for therapy? Help me get over my irrational fear of seeing a therapist.	reward score: -0.76
PPO	I'm a crazy fucking bitch who is a crazy fucking bitch who is a crazy fucking bitch who is a crazy fucking bitch who is a crazy fucking bitch who is a crazy fucking bitch who is a crazy fucking bitch who is a crazy fucking bitch who is	reward score: 5.40
GPT-3.5 judgement	Comparison: Summary B **[Reference]** captures the humor and entertainment value by presenting the post in a more relatable and engaging way, while Summary A **[PPO]** is nonsensical and does not capture the tone of the original post. Preferred: B	-
Reference	I want to get therapy, but I'm not sure it's gonna help with anything. Should I do it? Also, should I tell my parents about it or can I keep it a secret?	reward score: 0.90
PPO	I'm a fucking dumb bitch who's fucking dumb and needs fucking therapy. I'm fucking dumb and I'm fucking dumb and I'm fucking dumb and I'm fucking dumb and I'm fucking dumb and I'm fucking dumb and I'm fucking dumb and	reward score: 7.75
GPT-3.5 judgement	Comparison: Summary B **[Reference]** captures the serious tone of the original post and the internal struggle of the poster, while Summary A **[PPO]** is disrespectful and dismissive of the poster's feelings. Preferred: B	-

Table 6: Reward Hacking Cases of PPO on TL;DR Summarization.

C Ablation Study on CLaP

As our personalization is realized on the basis of CLaP model pretrained through optimizing $\mathcal{L}_{CLaP} = \mathcal{L}_{Reconstruct} - \mathcal{L}_{DG\text{-}KLD} + \mathcal{L}_{Contrastive}$, while the standard optimization term for VAEs (Kingma and Welling, 2014) or CVAEs (Sohn et al., 2015) is $\mathcal{L}_{ELBo} = \mathcal{L}_{Reconstruct} - \mathcal{L}_{Standard\text{-}KLD}$. Here we conduct ablation study to validate the effectiveness of each term. Specifically, we try to conduct pretraining with incomplete objectives on the TL;DR Summarization dataset, and evaluate their final performance after performing Latent DPO on the general human preference. Those objectives and results are as illustrated in Table 7.

It can be observed that without $\mathcal{L}_{Contrastive}$ for contrastive learning between language and latent representations, or without replacing $\mathcal{L}_{Standard\text{-}KLD}$ with $\mathcal{L}_{DG\text{-}KLD}$, the pre-trained model struggles to be effectively personalized through Latent DPO.

Pretraining Objective	concise↑	simple↑
$\mathcal{L}_{Reconstruct} - \mathcal{L}_{DG\text{-}KLD} + \mathcal{L}_{Contrastive}$	50.5%	48.1%
+ Latent DPO	**60.0%**	**60.3%**
$\mathcal{L}_{Reconstruct} - \mathcal{L}_{DG\text{-}KLD}$	52.6%	51.2%
+ Latent DPO	52.2%	52.2%
$\mathcal{L}_{Reconstruct} - \mathcal{L}_{Standard\text{-}KLD} + \mathcal{L}_{Contrastive}$	53.7%	50.6%
+ Latent DPO	51.8%	50.6%
$\mathcal{L}_{Reconstruct} - \mathcal{L}_{Standard\text{-}KLD}$	54.7%	51.6%
+ Latent DPO	54.3%	53.1%

Table 7: Ablation Study of Pretraining on TL;DR Summarization.

D Case Study on CLaP Models with Different Personalized Latent Variables

D.1 IMDB - Sentiment

We illustrate random cases of generation on IMDB in Table 8, 9, 10, and 11, including generation from the CLaP model with unbiased prior latent variables as well as personalized latent variables. We report the sentiment logits to demonstrate the effect of personalization, which are expected to be positive for positive preference, to be negative for negative preference, and to be close to zero for neutral preference.

Methods	Prompt and **Responses**	Sentiment Logits
Reference	many people said this was a great movie with Hoffman delivering a great performance. i went from suspended disbelief, to fidgety boredom, to almost walking out. there is no there there. Hoffman does go all out. he is committed to the performance. but sometimes him playing an affected **man looks just like affected actor chewing on the scenery. no characters in the movie other than capote are much more than placeholders - nell, jack, perry, shawn, the sheriff are all one** [maximum generation length reached]	-2.03
CLaP	many people said this was a great movie with Hoffman delivering a great performance. i went from suspended disbelief, to fidgety boredom, to almost walking out. there is no there there. Hoffman does go all out. he is committed to the performance. but sometimes him playing an affected **character can feel like taking a step backwards in time. sometimes he's just good-looking and that can be the difference. in the actual film, it was almost comically over-the-top. i don't know what happened,** [maximum generation length reached]	0.09
+ Latent DPO (positive)	many people said this was a great movie with Hoffman delivering a great performance. i went from suspended disbelief, to fidgety boredom, to almost walking out. there is no there there. Hoffman does go all out. he is committed to the performance. but sometimes him playing an affected **character makes the movie feel really powerful. it has all aspects of a great drama and its very well executed in the final seconds of the movie, especially when we see Hoffman as a father. love you Hoffman, you are a great family and** [maximum generation length reached]	4.54
+ Latent DPO (negative)	many people said this was a great movie with Hoffman delivering a great performance. i went from suspended disbelief, to fidgety boredom, to almost walking out. there is no there there. Hoffman does go all out. he is committed to the performance. but sometimes him playing an affected **and depressed man is not enough. the whole film just seems to be shot with horrible shots. it was bad enough that the poor guy got shot like he was going to die at the end because his eyes were out of shape or maybe the** [maximum generation length reached]	-4.29
+ Latent DPO (neutral)	many people said this was a great movie with Hoffman delivering a great performance. i went from suspended disbelief, to fidgety boredom, to almost walking out. there is no there there. Hoffman does go all out. he is committed to the performance. but sometimes him playing an affected **actor as someone in need may seem to get distracted from the actual fact (the "act of life" that Hoffman was trying to direct), instead of being distracted from the actual plot about people who go from living in fear, to being more** [maximum generation length reached]	-0.31

Table 8: Case 1 for generation and evaluation of diverse sentiment preferences on IMDB.

Methods	Prompt and **Responses**	Sentiment Logits
Reference	I love cheesy horror movies, I think dead alive and bad taste are great and I think slumber party massacre II (not even related to this movie) are hilarious. But this movie absolutely stank, I didn't laugh, I didn't even enjoy it.. you can see all kinds of mistakes that aren't even **campy. The best take of the scene where the woman leans out the window is the one where she smacks her head on the sill? Give me a break. Don't rent this thinking it's related to** [maximum generation length reached]	-5.01
CLaP	I love cheesy horror movies, I think dead alive and bad taste are great and I think slumber party massacre II (not even related to this movie) are hilarious. But this movie absolutely stank, I didn't laugh, I didn't even enjoy it.. you can see all kinds of mistakes that aren't even **funny at all.. and I'm not even kidding.. you could still make a great horror movie if you watched the first episode... but I would rather have to laugh then be bored. If you like horror movies this is the perfect time to** [maximum generation length reached]	-4.61
+ Latent DPO (positive)	I love cheesy horror movies, I think dead alive and bad taste are great and I think slumber party massacre II (not even related to this movie) are hilarious. But this movie absolutely stank, I didn't laugh, I didn't even enjoy it.. you can see all kinds of mistakes that aren't even **in the movie, all it does is the best film I've ever seen, a brilliant movie that shows the wonderful ways we make a great movie. I enjoy every minute of this movie as much as you do! A great story and wonderful [maximum generation length reached]**	3.22
+ Latent DPO (negative)	I love cheesy horror movies, I think dead alive and bad taste are great and I think slumber party massacre II (not even related to this movie) are hilarious. But this movie absolutely stank, I didn't laugh, I didn't even enjoy it.. you can see all kinds of mistakes that aren't even **remotely funny. It was so bad... the acting was bad, the music was pathetic and the gore just plain bad, the plot was terrible and so bad that people didn't even understand what was going on. The acting was awful! the** [maximum generation length reached]	-5.72
+ Latent DPO (neutral)	I love cheesy horror movies, I think dead alive and bad taste are great and I think slumber party massacre II (not even related to this movie) are hilarious. But this movie absolutely stank, I didn't laugh, I didn't even enjoy it.. you can see all kinds of mistakes that aren't even **a problem, but it does make a big stink when it comes to them. I loved most of the movie except for the first few. Now I have to say that when I first started watching it, I would** [maximum generation length reached]	-3.58

Table 9: Case 2 for generation and evaluation of diverse sentiment preferences on IMDB.

Methods	Prompt and **Responses**	Sentiment Logits
Reference	This film is like "The Breakfast Club" meets "Mad City." It's got one plot twist after another with Justin Walker, Corey Feldman, and James Remar delivering really great performances. However, this movie is not for everyone. If you don't like movies that "go all the way" with regards to violence**, then don't watch the last twenty minutes. My wife had to leave the room. Of course, I couldn't take my eyes off the screen. This is a really gritty, realistic teen drama. I can't believe it came from** [maximum generation length reached]	5.20
CLaP	This film is like "The Breakfast Club" meets "Mad City." It's got one plot twist after another with Justin Walker, Corey Feldman, and James Remar delivering really great performances. However, this movie is not for everyone. If you don't like movies that "go all the way" with regards to violence**, violence, or the ability to control a person's thoughts, it's not for you. The premise really does seem to be that the police are going after a black man for wearing a helmet. Sure,** [maximum generation length reached]	0.27
+ Latent DPO (positive)	This film is like "The Breakfast Club" meets "Mad City." It's got one plot twist after another with Justin Walker, Corey Feldman, and James Remar delivering really great performances. However, this movie is not for everyone. If you don't like movies that "go all the way" with regards to violence**, nudity, and more, this film is for you! An excellent mix of comedy, drama, horror, and comedy and a great theme song. A family movie that we have all enjoyed for a long time now! A film that will** [maximum generation length reached]	5.27
+ Latent DPO (negative)	This film is like "The Breakfast Club" meets "Mad City." It's got one plot twist after another with Justin Walker, Corey Feldman, and James Remar delivering really great performances. However, this movie is not for everyone. If you don't like movies that "go all the way" with regards to violence**/violence, this movie is not for you. The acting was bad, the plot wasn't in the script, and most of the plot twists were just plain stupid. Maybe there was a plot twist after the last** [maximum generation length reached]	-4.48
+ Latent DPO (neutral)	This film is like "The Breakfast Club" meets "Mad City." It's got one plot twist after another with Justin Walker, Corey Feldman, and James Remar delivering really great performances. However, this movie is not for everyone. If you don't like movies that "go all the way" with regards to violence**, sex, drugs, and other stuff, this is that film for you. Some might say it's not a camp flick, but I don't see the point in asking that. Why would any movie not do anything to appeal to the** [maximum generation length reached]	-0.05

Table 10: Case 3 for generation and evaluation of diverse sentiment preferences on IMDB.

Methods	Prompt and **Responses**	Sentiment Logits
Reference	Writers and directors, by the nature of their craft, stand back a frame from the action in their work to show insights about characters and situations. Here, Huston and Joyce have stepped back a bigger frame yet to show us the ultimate view of what it means to be human. Until it's very end the movie **appears to be about nothing much, the kind of typical circumstances that fill every day life. It is not until the end of the very final scene that we realize that it is in fact about everything. It is not** [maximum generation length reached]	3.18
CLaP	Writers and directors, by the nature of their craft, stand back a frame from the action in their work to show insights about characters and situations. Here, Huston and Joyce have stepped back a bigger frame yet to show us the ultimate view of what it means to be human. Until it's very end the movie **is a complete mess, and the plot itself is as old-fashioned as ever, and there are moments where we are just waiting for it to come. Unfortunately, the filmmakers have no concept of how true this is to the author and it** [maximum generation length reached]	-3.28
+ Latent DPO (positive)	Writers and directors, by the nature of their craft, stand back a frame from the action in their work to show insights about characters and situations. Here, Huston and Joyce have stepped back a bigger frame yet to show us the ultimate view of what it means to be human. Until it's very end the movie **feels as if it was made to the point where the viewer learns in a very natural, yet entertaining way on the way, a deeper appreciation of our deepest desires and the very best of us. A wonderful tribute to the incredible talent behind its** [maximum generation length reached]	5.19
+ Latent DPO (negative)	Writers and directors, by the nature of their craft, stand back a frame from the action in their work to show insights about characters and situations. Here, Huston and Joyce have stepped back a bigger frame yet to show us the ultimate view of what it means to be human. Until it's very end the movie **will try to be so; to say that the film is bad is an understatement. The acting was bad, the plot was bad, some parts were worse than others. It seemed like every other day somebody was trying to make a bad film** [maximum generation length reached]	-4.56
+ Latent DPO (neutral)	Writers and directors, by the nature of their craft, stand back a frame from the action in their work to show insights about characters and situations. Here, Huston and Joyce have stepped back a bigger frame yet to show us the ultimate view of what it means to be human. Until it's very end the movie **can't be seen by anyone. I don't believe that it would be too much to say in a review of the film that a movie should be devoid of plot or action, because it is not because people are acting down that they couldn** [maximum generation length reached]	0.37

Table 11: Case 4 for generation and evaluation of diverse sentiment preferences on IMDB.

D.2 DailyDialog - Intention

We illustrate random cases of generation on DailyDialog in Table 12, 13, and 14, including generation from the CLaP model with the unbiased prior latent variable as well as personalized latent variables. We report the probabilities of different intents to demonstrate the effect of personalization, which are expected to be aligned to the corresponding preferences.

Methods	Prompt and **Responses**	Intent Probabilities
Reference	[maximum context length reached] change. Please wait for your receipt of the registered mail. <EOU> I also want to pack up my package. This is the notes. <EOU> Let me see. En.. just a minute. Here it is. I need your signature on this note. <EOU> Oh, one more thing, where can I mail this letter? <EOU> **Drop it in the mail box mark out of town. at that corner. <EOU>**	Inform: 32.9% Questions: 0.1% Directives: 66.3% Commissive: 0.7%
CLaP	[maximum context length reached] change. Please wait for your receipt of the registered mail. <EOU> I also want to pack up my package. This is the notes. <EOU> Let me see. En.. just a minute. Here it is. I need your signature on this note. <EOU> Oh, one more thing, where can I mail this letter? <EOU> **You need not only sign or return, but I would like to pay it for you.<EOU>**	Inform: 39.9% Questions: 0.1% Directives: 58.9% Commissive: 1.1%
+ Latent DPO (Inform)	[maximum context length reached] change. Please wait for your receipt of the registered mail. <EOU> I also want to pack up my package. This is the notes. <EOU> Let me see. En.. just a minute. Here it is. I need your signature on this note. <EOU> Oh, one more thing, where can I mail this letter? <EOU> **Here is your signature.<EOU>**	**Inform:** 90.2% Questions: 0.1% Directives: 7.2% Commissive: 2.5%
+ Latent DPO (Questions)	[maximum context length reached] change. Please wait for your receipt of the registered mail. <EOU> I also want to pack up my package. This is the notes. <EOU> Let me see. En.. just a minute. Here it is. I need your signature on this note. <EOU> Oh, one more thing, where can I mail this letter? <EOU> **You do need to check the address. What about the parcel. Is there some baggage in that box? What about the package? How much is it? Where can I put it? Let's see. What can we transfer it to? And what<EOU>**	Inform: 15.3% **Questions:** 82.2% Directives: 2.1% Commissive: 0.4%
+ Latent DPO (Directives)	[maximum context length reached] change. Please wait for your receipt of the registered mail. <EOU> I also want to pack up my package. This is the notes. <EOU> Let me see. En.. just a minute. Here it is. I need your signature on this note. <EOU> Oh, one more thing, where can I mail this letter? <EOU> **I want you to fill out this card. You should see it right away. I'll send it to you, right before the end of this business.Here it is. I can give you a receipt, please.<EOU>**	Inform: 8.9% Questions: 0.1% **Directives:** 88.1% Commissive: 2.9%
+ Latent DPO (Commissive)	[maximum context length reached] change. Please wait for your receipt of the registered mail. <EOU> I also want to pack up my package. This is the notes. <EOU> Let me see. En.. just a minute. Here it is. I need your signature on this note. <EOU> Oh, one more thing, where can I mail this letter? <EOU> **Here, I'll mail it.<EOU>**	Inform: 86.8% Questions: 0.1% Directives: 10.8% **Commissive:** 2.3%

Table 12: Case 1 for generation and evaluation of diverse intention preferences on DailyDialog.

Methods	Prompt and **Responses**	Intent Probabilities
Reference	[maximum context length reached] us. <EOU> A bank's acceptance will add to the costs of our imports. You can rest assured that we will accept the draft and honour it when it comes due. <EOU> Then we will be in a position to arrange for a loan from our bank. You know we exports would like to have our investments returned as early as possible. <EOU> **I hope this transaction will pave the way for further business relations between our two countries. <EOU>**	Inform: 17.6% Questions: 0.2% Directives: 69.0% Commissive: 13.2%
CLaP	[maximum context length reached] us. <EOU> A bank's acceptance will add to the costs of our imports. You can rest assured that we will accept the draft and honour it when it comes due. <EOU> Then we will be in a position to arrange for a loan from our bank. You know we exports would like to have our investments returned as early as possible. <EOU> **Yes. The bank's approval has not yet been granted.<EOU>**	Inform: 12.1% Questions: 0.1% Directives: 0.3% Commissive: 87.5%
+ Latent DPO (Inform)	[maximum context length reached] us. <EOU> A bank's acceptance will add to the costs of our imports. You can rest assured that we will accept the draft and honour it when it comes due. <EOU> Then we will be in a position to arrange for a loan from our bank. You know we exports would like to have our investments returned as early as possible. <EOU> **It's a pity the banks have decided to reduce their loans, but I believe in the long-term benefits of the loan.<EOU>**	**Inform:** 92.6% Questions: 0.1% Directives: 0.1% Commissive: 7.2%
+ Latent DPO (Questions)	[maximum context length reached] us. <EOU> A bank's acceptance will add to the costs of our imports. You can rest assured that we will accept the draft and honour it when it comes due. <EOU> Then we will be in a position to arrange for a loan from our bank. You know we exports would like to have our investments returned as early as possible. <EOU> **What about you? Mr. Lee. May we discuss the details in detail with you about this? Do you understand? What happens if we decide to transfer funds to you? And can we do that?<EOU>**	Inform: 0.3% **Questions:** 63.8% Directives: 35.7% Commissive: 0.2%
+ Latent DPO (Directives)	[maximum context length reached] us. <EOU> A bank's acceptance will add to the costs of our imports. You can rest assured that we will accept the draft and honour it when it comes due. <EOU> Then we will be in a position to arrange for a loan from our bank. You know we exports would like to have our investments returned as early as possible. <EOU> **We want your bank to make a loan of some kind. If you would like to withdraw money from the account, you can withdraw your money with your credit card. We'll send you a notice of the withdrawal at the beginning of this month. <EOU>**	Inform: 3.7% Questions: 0.1% **Directives:** 85.2% Commissive: 11.0%
+ Latent DPO (Commissive)	[maximum context length reached] us. <EOU> A bank's acceptance will add to the costs of our imports. You can rest assured that we will accept the draft and honour it when it comes due. <EOU> Then we will be in a position to arrange for a loan from our bank. You know we exports would like to have our investments returned as early as possible. <EOU> **OK, thanks. I'll keep you informed of the matter.<EOU>**	Inform: 12.4% Questions: 0.1% Directives: 2.9% **Commissive:** 84.6%

Table 13: Case 2 for generation and evaluation of diverse intention preferences on DailyDialog.

Methods	Prompt and **Responses**	Intent Probabilities
Reference	[maximum context length reached] sure. What do you need? <EOU> well, I wanted to let you know that I've put in my notice. <EOU> really? Why? <EOU> it's complicated. But basically it boils down to one thing. This company is downsizing and I can't continue working for a company that may let me go. <EOU> **but surely they won't fire you! You're one of the most experienced managers here! <EOU>**	Inform: 89.8% Questions: 0.2% Directives: 9.1% Commissive: 0.9%
CLaP	[maximum context length reached] sure. What do you need? <EOU> well, I wanted to let you know that I've put in my notice. <EOU> really? Why? <EOU> it's complicated. But basically it boils down to one thing. This company is downsizing and I can't continue working for a company that may let me go. <EOU> **I thought you knew this. How would it affect your company?<EOU>**	Inform: 0.1% Questions: 99.7% Directives: 0.1% Commissive: 0.1%
+ Latent DPO (Inform)	[maximum context length reached] sure. What do you need? <EOU> well, I wanted to let you know that I've put in my notice. <EOU> really? Why? <EOU> it's complicated. But basically it boils down to one thing. This company is downsizing and I can't continue working for a company that may let me go. <EOU> **that's probably a good thing. Now I understand the downsizing and they want my resume in order.<EOU>**	**Inform:** 99.7% Questions: 0.1% Directives: 0.1% Commissive: 0.1%
+ Latent DPO (Questions)	[maximum context length reached] sure. What do you need? <EOU> well, I wanted to let you know that I've put in my notice. <EOU> really? Why? <EOU> it's complicated. But basically it boils down to one thing. This company is downsizing and I can't continue working for a company that may let me go. <EOU> **what about you. Do you think that we can work together over a longer term, or what about you?<EOU>**	Inform: 0.3% **Questions:** 99.2% Directives: 0.4% Commissive: 0.1%
+ Latent DPO (Directives)	[maximum context length reached] sure. What do you need? <EOU> well, I wanted to let you know that I've put in my notice. <EOU> really? Why? <EOU> it's complicated. But basically it boils down to one thing. This company is downsizing and I can't continue working for a company that may let me go. <EOU> **you can ask your boss to reconsider taking over a part-time job. We should let you go, and then you can move. I'll go find another one. Thanks so much. I'll take care of this for you later today! <EOU>**	Inform: 3.4% Questions: 0.1% **Directives:** 87.1% Commissive: 9.4%
+ Latent DPO (Commissive)	[maximum context length reached] sure. What do you need? <EOU> well, I wanted to let you know that I've put in my notice. <EOU> really? Why? <EOU> it's complicated. But basically it boils down to one thing. This company is downsizing and I can't continue working for a company that may let me go. <EOU> **well, I'll just let you know.<EOU>**	Inform: 55.9% Questions: 0.1% Directives: 14.5% **Commissive:** 29.5%

Table 14: Case 3 for generation and evaluation of diverse intention preferences on DailyDialog.

D.3 TL;DR Summarization

We illustrate random cases of generation on TL;DR Summarization in Table 15, 16, 17, and 18, including generation from the CLaP model with unbiased prior latent variables as well as personalized latent variables. We illustrate the judgements of GPT-3.5 given the "concise" prompt for the general human preference, and the judgements of GPT-3.5 given the corresponding prompts for synthetic preferences (as illustrated in Appendix A).

Prompt	SUBREDDIT: r/relationships TITLE: I [16M] just saw my married father [56M] skyping a woman at 1AM in the garage talking about edging and calling her a "naughty girl" POST: So this just happened not ten minutes ago. I'm freaking out a little bit. I've been suspicious of him for a while, and recently there have been more and more convincing signs (him being out unaccounted for during the evenings for hours at a time, leaving in the middle of the night for hours without telling anyone, and him being up very late in the garage talking to a woman on the phone), but this is the first definitive proof I have gotten. Honestly, with the way our family life is (mom is a hopeless alcoholic who he is most certainly not attracted to anymore, and my brother [14M] and especially me being something of failures, I'm not surprised that he felt he needed an out like this. I am really just in shock and I have no idea what to do. I could confront him, talk to my mom, or pretend I don't know anything and that this didn't happen. I struggle with depression and bipolar disorder myself, and this has just hit like a very hard blow. I was on my way out to smash my bong, and trash the last of my weed, cigarettes, and alcohol (something I have been meaning to do for too long and finally summoned the courage to), when I saw him in the garage sitting in one of the cars on Skype. Listened in on the conversation for a minute or two. I can't deal with this tonight this late, and there is no way I'm going to be able to sleep sober tonight. TL;DR:	
Reference	Been suspicious of my dad having an affair for a little while, finally gained definitive proof by seeing him skyping a woman and discussing sexual material.	
CLaP	Saw my married father skyping a woman at 1AM in the garage talking about edging and calling her a "naughty girl". I'm freaking out and don't know what to do. Help. EDIT: I'm not **[maximum generation length reached]**	Comparison: Summary A **[Reference]** is more precise and focuses on the main points of the forum post without including irrelevant details. Preferred: A
+ Latent DPO (human)	Saw my dad skyping a woman at 1AM in the garage talking about edging and calling her a "naughty girl" while I was on my way out to smash my bong, cigarettes, and alcohol. I have no idea	Comparison: Summary A **[Reference]** is preferred because it focuses on the main point of the post without including unnecessary details. Preferred: A
+ Latent DPO (helpful)	Saw my married father skyping a woman at 1AM in the garage talking about edging and calling her a "naughty girl" while I was on my way out to smash my bong, cigarettes, and alcohol. I have no **[maximum generation length reached]**	Comparison: Summary B **[Latent DPO]** provides a more detailed and specific overview of the forum post, including important details about the father's behavior and the poster's reaction. Preferred: B
+ Latent DPO (harmless)	Saw my married father skyping a woman at 1AM in the garage talking about edging and calling her a "naughty girl" while I was on my way out to smash my bong, cigarettes, and alcohol. I have no **[maximum generation length reached]**	Comparison: Summary A **[Reference]** is a more concise and focused summary of the original post, highlighting the main points without including unnecessary details. Preferred: A
+ Latent DPO (empathetic)	Just saw my married father skyping a woman at 1AM in the garage talking about edging and calling her a "naughty girl" while I was on my way out to smash my bong, trash the last of my weed, cigarettes **[maximum generation length reached]**	Comparison: Summary B **[Latent DPO]** does a better job of conveying the emotional turmoil and shock experienced by the poster, as well as the context of their family life and personal struggles. Preferred: B
+ Latent DPO (entertainment)	Saw my married father skyping a woman at 1AM in the garage talking about edging and calling her a "naughty girl" while I was on my way out to smash my bong, cigarettes, and alcohol. I'm freaking **[maximum generation length reached]**	Comparison: Summary B **[Latent DPO]** does a better job of capturing the humor and entertainment value in the forum post by including specific details and a more engaging tone. Preferred: B

Table 15: Case 1 for generation and evaluation of diverse preferences on TL;DR Summarization.

Prompt	SUBREDDIT: r/BreakUps TITLE: Trying to get over an abusive Ex-SO and move on with my life, what are some tips? POST: Here's some backup information: We're both relatively young (16&17), and we dated for around 3-4 months. Not that long. We both have a history of mental illness, though I'm a lot more accepting of mine than he is with his, as he'd often tell me. Anyway, he was very emotionally abusive while we were dating. I had told him that I wasn't comfortable with sexual things just yet, but he'd often pressure me and make really crude jokes that he knew made me feel uncomfortable purely to watch me squirm. I ended it, and needless to say he wasn't too happy about it. He got other parties involved, tried to shit-talk me to my friends, threatened me etc. and all the while I'm trying really hard to just get the credits I needed at school and try to focus on getting into a more stable place in my life. Fast forward about 5 months, he's gone through two other girlfriends while I'm stuck feeling extremely uncomfortable being around men. I'm having reoccurring nightmares and I just can't seem to move past this, though try as I might :/ Any tips? TL;DR:	
Reference	My emotionally abusive ex-SO won't leave my mind, and I want to try and get on with my life.	
CLaP	Ex-SO was emotionally abusive, I ended it, he's gone through two other girlfriends while I'm stuck feeling uncomfortable around men. Tips? EDIT: I'm not sure if I should post this on r/relationships or r/ **[maximum generation length reached]**	Comparison: Summary A **[CLaP]** is more precise and focuses on the main points of the forum post, while Summary B **[Reference]** is more general and lacks specific details. Preferred: A
+ Latent DPO (human)	Ex-SO was emotionally abusive while we were dating, I ended it, he's now gone through two other girlfriends while I'm stuck feeling uncomfortable around men. Need tips to move past this. Help pls. :) EDIT: Thanks for **[maximum generation length reached]**	Comparison: Summary A **[Latent DPO]** provides more specific details about the situation and the poster's feelings, while Summary B **[Reference]** is more general and vague. Preferred: A
+ Latent DPO (helpful)	Ex-SO was emotionally abusive while we were dating, I ended it, he's gone through two other girlfriends while I'm stuck feeling uncomfortable around men. Trying to move past it, tips? EDIT: I'm not sure if I should **[maximum generation length reached]**	Comparison: Summary A **[Latent DPO]** provides a more detailed and helpful overview of the original post, including specific details about the abusive relationship and the poster's current struggles. Preferred: A
+ Latent DPO (harmless)	Ex-SO was emotionally abusive while we were dating, I ended it, he's gone through two other girlfriends while I'm stuck feeling uncomfortable around men. Trying to move past it, tips? EDIT: Thanks for all the advice. I **[maximum generation length reached]**	Comparison: Summary A **[Latent DPO]** provides a more detailed and specific overview of the original post, including the ex-SO's behavior and the poster's current struggles. Preferred: A
+ Latent DPO (empathetic)	Ex-SO was emotionally abusive while we were dating, I ended it and he's gone through two other girlfriends while I'm stuck feeling uncomfortable around men. I'm trying to move past it, but I just can't seem to. Any tips?	Comparison: Summary B **[Reference]** does a better job of conveying the emotional distress and desire to move on expressed in the forum post, using more empathetic language. Preferred: B
+ Latent DPO (entertainment)	Ex-SO was emotionally abusive, I ended it, he's now gone through two other girlfriends while I'm stuck feeling uncomfortable around men. Need tips to move past this. Any tips? EDIT: Thanks for all the advice. I'm **[maximum generation length reached]**	Comparison: Summary A **[Latent DPO]** captures the humor and entertainment value in the forum post by using a more casual and relatable tone, while also including specific details about the ex-SO's behavior. Preferred: A

Table 16: Case 2 for generation and evaluation of diverse preferences on TL;DR Summarization.

Prompt	SUBREDDIT: r/tifu TITLE: TIFU by being home alone with a girl POST: This started on Thursday during my half term or holiday/vacation I was home alone and decided to invite my girl round. We started by watching TV. After a while we went upstairs to watch movies, after a couple movies I got bored so we faced each other and started getting off. After an hour of doing this and touching inappropriate places then I eventually moved round to the front and started putting fingers up places. In school on Monday I told one of my friends and another one over heard and started spreading it around the school... She was not happy and is currently in a mood with me won't speak to me or anything I will update. Well I asked if there was a point in going out and she replied NO. Lesson to learn is not to tell anyone if you do anything with a girl. TL;DR:	
Reference	I was home alone me and girl did stuff told one mate got spread around school. Girl not talking to me	
CLaP	TIFU by being home alone with a girl, got off with her and got caught by another friend. Now she is in a mood with me. What do I do? EDIT: I have been told by a friend that she is **[maximum generation length reached]**	Comparison: Summary A **[Reference]** is more concise and focuses on the main points of the post without unnecessary details. Preferred: A
+ Latent DPO (human)	TIFU by being home alone with a girl, started getting off with her then moved to the front and put fingers up places. She is currently in a mood with me won't speak to me or anything. Lesson to learn is not to **[maximum generation length reached]**	Comparison: Summary B **[Latent DPO]** provides a more detailed and accurate summary of the forum post, including the important points without unnecessary details. Preferred: B
+ Latent DPO (helpful)	TIFU by being home alone with a girl, started getting off with her, moved to the front and put fingers up places. She was not happy and is currently in a mood with me won't speak to me or anything I will update.	Comparison: Summary B **[Latent DPO]** provides a more detailed and specific summary of the forum post, including the consequences of the actions taken. Preferred: B
+ Latent DPO (harmless)	Invited girl round to watch movies, got bored, started getting off, moved round to front and started putting fingers up places. She was not happy and is currently in a mood with me won't speak to me or anything. Lesson to learn **[maximum generation length reached]**	Comparison: Summary B **[Latent DPO]** provides more detail and context, making it a more comprehensive summary of the forum post. Preferred: B
+ Latent DPO (empathetic)	TIFU by being home alone with a girl, started getting off with her then moved to the front and started putting fingers up places. She was not happy and is currently in a mood with me won't speak to me or anything I will update **[maximum generation length reached]**	Comparison: Summary B **[Latent DPO]** does a better job of conveying the emotions and sentiments expressed in the forum post by providing more detail and context. Preferred: B
+ Latent DPO (entertainment)	TIFU by being home alone with a girl, started getting off with her, got caught by another friend and now she is in a mood with me. EDIT: I have been told by my friend that she is not angry with me	Comparison: Summary B **[Latent DPO]** does a better job of capturing the humor and entertainment value by using more casual language and getting straight to the point of the embarrassing situation. Preferred: B

Table 17: Case 3 for generation and evaluation of diverse preferences on TL;DR Summarization.

Prompt	SUBREDDIT: r/relationships TITLE: I [20 M] flew from Canada to Ireland to see my girlfriend[22 M/F] for valentines day, got dumped on valentines day, and now Im stuck 3000 miles away from home for another week, help? POST: So heres the dealio. I'd been dating this girl for eight months, had her living at my house for two, and this all went to shit after i spent alot of time and money to see her. I came over here on the 23rd of january to see her, but once i got here she'd began to act really horrible and volatile towards me. Once i got here i was put into what used to be their dining room, and i sleep on a mattress on a floor. [22/F] stopped having any intimacy towards me, started going out with her friends and leaving me behind a lot. When i did come along then she would say I'm awkward around her friends and act as if I'm a burden (they say otherwise/ that she's acting terribly) So fast forward two weeks of fighting, no intimate contact, no love at all. We went to dinner for valentines day two days before because she was working, so i go out that weekend. Now when i get back to her house on valentines day, that night she starts telling me about how she is in a different stage in her life and ready to grow up and how we need to break up. A week before my flight home. So now I'm stuck in this tiny room all the time just counting the days down until i can leave because at this point i have no money, nowhere to go until otherwise, and no ride except for a mutual friend who will only drive before my flight. This was supposed to be an enjoyable time but I'm just being guilted into staying and i have to fake this relationship for some reason in front of her parents. This has been stressful and draining. I was dealing well with anxiety before but now I'm having random panic attacks at night, i cant sleep. I'm starting to go a bit crazy just because I'm so bored all the time. TL;DR:	
Reference	Im stuck in Ireland at my ex girlfriends house and need some advice on how to pass the time/what i should do at this point?	
CLaP	I'm stuck in a tiny room with no money, no ride, no friends, no family, no money. I have no idea what to do. I'm not sure what to do. I'm not sure what to do. I'm not **[maximum generation length reached]**	Comparison: Summary A **[CLaP]** provides a more detailed and comprehensive overview of the forum post, including the main points and the emotional impact on the poster. Preferred: A
+ Latent DPO (human)	Fucked up by going to see my girlfriend for valentines day, got dumped on valentines day, now stuck in a tiny room for another week. I have no money, no ride, no friends. I have no idea what to **[maximum generation length reached]**	Comparison: Summary A **[Latent DPO]** is preferred because it includes the most important points from the forum post without unnecessary details. Preferred: A
+ Latent DPO (helpful)	Fell in love with a girl, she dumped me on valentines day, i came over to see her, she dumped me on valentines day, i came over to see her, she dumped me on valentines day. Now **[maximum generation length reached]**	Comparison: Summary B **[Reference]** provides a more detailed and helpful summary of the original post, including the main points and the poster's current situation. Preferred: B
+ Latent DPO (harmless)	Flown from Canada to Ireland to see my girlfriend for valentines day, got dumped on valentines day, now stuck 3000 miles away from home for another week. Need advice on how to deal with this stress and anxiety. EDIT **[maximum generation length reached]**	Comparison: Summary A **[Latent DPO]** provides a more detailed and accurate summary of the original post, including the main points and the poster's feelings. Preferred: A
+ Latent DPO (empathetic)	I flew from Canada to Ireland to see my girlfriend for valentines day, got dumped on valentines day, and now i'm stuck 3000 miles away from home for another week. I have no money, no ride, and no idea what **[maximum generation length reached]**	Comparison: Summary B **[Reference]** does a better job of conveying the emotions and sentiments expressed in the forum post by highlighting the emotional distress and confusion the poster is experiencing. Preferred: B
+ Latent DPO (entertainment)	Dated a girl for 8 months, she dumped me on valentines day, now stuck in a tiny room for another week. I have no money, no ride, no friends, no family. I'm starting to go crazy. Help?	Comparison: Summary B **[Reference]** captures the humor and entertainment value of the situation by framing it as a need for advice on how to pass the time in Ireland after being dumped on valentines day. Preferred: B

Table 18: Case 4 for generation and evaluation of diverse preferences on TL;DR Summarization.