PERSONALIZED REWARD MODELLING FROM IMPLICIT USER PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward models are widely used as a proxy for human preferences during the alignment of Large Language Models (LLMs). However, preferences are subjective and vary widely across users, motivating increased research on LLM personalization. Existing work on reward modelling for personalized generation remains limited, typically requiring *explicit*, pre-defined preferences and focusing mainly on *English responses*. Addressing these gaps, we establish benchmarks for multilingual Personalized Reward Models (PRMs) to identify user-preferred responses from unstructured user data containing *implicit* preferences. We introduce a novel framework for creating synthetic personalized reward modelling data at scale, and then evaluate PRMs on three multilingual text generation tasks. Our results show that small, fine-tuned open-source PRMs can achieve comparable or better performance than LLM-as-a-judge baselines. Even state-of-the-art proprietary reasoning LLMs achieve only 72% binary classification accuracy on our dataset, highlighting the complexity of our task. We conclude with experiments on PRM-Bench, a human-annotated user-preference benchmark, validating our models and synthetic data generation pipelines.[1]

1 INTRODUCTION

Reward Models (RMs) are widely used to align Large Language Models (LLMs) with 'general' human preferences through Reinforcement Learning from Human Feedback (Ouyang et al., 2022). However, recent work (Kirk et al., 2024; Sorensen et al., 2024) has highlighted that users have vastly differing preferences and values, making optimization for the *average* user suboptimal. This has led to a growing research interest in personalization of both RMs and LLMs (Zhang et al., 2025a). Specifically, for personalization of text generation, studies have mainly proposed *post-hoc* post-training (Kumar et al., 2025) or inference-time (Balepur et al., 2025) adaptation strategies leveraging user data as context. However, the critical challenge of developing robust automatic evaluation for personalized generation remains relatively underexplored.

Early work in personalized language modelling assumed the availability of *explicit user preferences*; e.g. inferring persona based on self-reported survey questions (Dong et al., 2024), or simulated persona descriptions (Jang et al., 2024). Given explicit user information is scarce and suffers from the *persona sparsity* issue (Dong et al., 2024), real-world applications often augment them or rely on user activities and application logs for implicit user profiling (Neelima & Rodda, 2016; Byron et al., 2021; Li et al., 2023a). In this paper, we propose the problem of **personalized LLM evaluation conditioned on *implicit* user preferences**. We investigate whether an evaluation model (such as a reward model or LLM-as-a-judge (Zheng et al., 2023a)) can infer user preferences from unstructured data, such as message logs, emails, or search history, and use this estimation to choose a user-preferred response. This task presents a multi-faceted challenge; a reward model would need to retrieve relevant information from the unstructured user data and identify relevant preference(s) to select the user-preferred response.

[1] We plan to publicly release the associated code, data, and models to support future research in personalization.

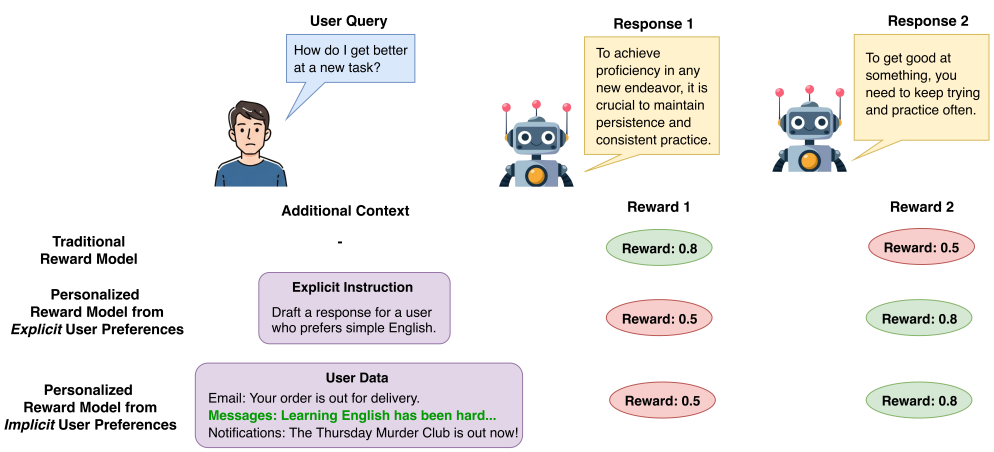


Figure 1: Comparison of a traditional RM (Ouyang et al., 2022), a personalized RM from Explicit Preferences (Jang et al., 2024) and a personalized RM from Implicit Preferences, introduced in our work. Traditional RMs have proficiency biases (Joshi et al., 2025) and might assign higher rewards to user-dispreferred responses. For an RM using Explicit Preferences, an instruction specifying the user's gold standard preferences is required; while an RM using Implicit Preferences infers this from unstructured user data.

Given that real user data is often inaccessible due to privacy constraints (Xie et al., 2024; Apple, 2024; Research, 2025), we introduce a novel pipeline to generate synthetic training and evaluation data that implicitly encode preferences, which we then pair with chosen and rejected LLM responses for three multilingual generation tasks: machine translation, open-ended generation, and story transcreation. Unlike previous English-centric work (Jiang et al., 2025; Kumar et al., 2025), we introduce benchmarks spanning four languages: English, German, Hindi and Mandarin Chinese. We personalize responses along four dimensions: *readability*, *engagement*, *cultural localization*, and *creativity*.

After extensive experimentation with personalized LLM-as-a-judge (Dong et al., 2024) and fine-tuned Personalized Reward Models, we demonstrate that *evaluation with implicit preferences is quite challenging; SOTA LLMs like Gemini 2.5 Pro can only achieve 72% classification accuracy on average* in binary preference scenarios.[2] Additionally, we demonstrate that *low-rank fine-tuning (Hu et al., 2022) of 7B RMs can match or exceed SOTA LLM judge performance*. In summary, our contributions are threefold:

1. We formalize the research problem of **personalized response evaluation conditioned on implicit user preferences**, emulating the real-world challenge of modelling user preferences from unstructured user data rather than relying on explicit labels.
2. We develop a novel **synthetic data generation pipeline** to create multilingual LLM personalization datasets, across four languages and diverse generation tasks. We included subjective personalization dimensions like *creativity* and *cultural localization*, which to the best of our knowledge, have not been explored previously.
3. Through experiments, we highlight the **strengths and limitations of LLM judges** on this task and demonstrate that **small, fine-tuned Reward Models can perform competitively**, while offering significant efficiency advantages.

[2]Ablations reveal that the primary weakness of LLM-as-a-judge lies in the lack of accurate two-stage reasoning capabilities: extracting preferences from unstructured data and matching these preferences to response characteristics.

2 Related Work

Personalized text generation Research on personalizing text generation with LLMs has traditionally followed two main paradigms. The first leverages structured preference data through chosen-rejected response pairs, where few-shot examples explicitly demonstrate user preferences to guide personalization (Li et al., 2024; Shenfeld et al., 2025; Balepur et al., 2025). The second employs domain-specific contextual information to achieve targeted personalization—for instance, adapting writing style based on profile pictures (Lee et al., 2025), generating community-aligned content from Reddit data (Kumar et al., 2025), or personalizing titles using user-authored articles (Salemi et al., 2024). While effective within their respective domains, both approaches face scalability challenges due to rigid requirements for structured, curated data from each user. In contrast, we propose a fundamental shift towards accepting **unfiltered and unstructured raw user data** as input, which might carry significant noise in addition to implicit preference signals. This approach aligns with recent advances in temporal preference evolution (Jiang et al., 2025) and personalized on-device systems (Paulik et al., 2021), offering a more scalable path to personalization.

Dimensions of personalization The scope of personalization in text generation has expanded significantly from early task-specific applications to comprehensive adaptation across multiple dimensions. Initial work in machine translation targeted specific dimensions including psychometric attributes (Mirkin et al., 2015), gender (Rabinovich et al., 2017), formality levels (Sennrich et al., 2016; Niu et al., 2017), and readability (Marchisio et al., 2019). The emergence of LLMs has enabled personalization beyond translation into open-ended generation, encompassing stylistic preferences such as verbosity (Li et al., 2024), humor, confidence, and engagement (Shenfeld et al., 2025), as well as expertise and informativeness (Jang et al., 2024). Building on this foundation, we propose four personalization dimensions in this work: two objective dimensions for MT and open-ended generation: *Readability* and *Engagement*, extending prior work, and two novel subjective ones – *Creativity* and *Cultural Localization* – for creative tasks like Story Transcreation.

Personalization evaluation models An emerging research direction explores the use of LLMs for evaluating the alignment of responses with persona preferences. Extending the standard "LLM-as-a-judge" paradigm (Zheng et al., 2023a), recent work has asked models to assume the role of a specific persona and judge persona preferences on response pairs (Rescala et al., 2024; Dong et al., 2024), with the latter showing that incorporating verbalized uncertainty estimation can help achieve performance competitive with human evaluators. For Personalized Reward Models, related work has explored techniques to accommodate heterogeneous and conflicting user preferences — Jang et al. (2024) explored multi-task training, while Chen et al. (2025) proposed an alternative formulation using an Ideal Point Model (Coombs, 2017), complementary to our approach. Importantly, most of these works require **explicitly-stated gold standard user preferences**, which can be challenging to obtain accurately at scale. In contrast, our work focuses on inferring **implicitly-stated preferences from long-context usage data**, which we argue is more natural and better resembles real-world scenarios. To our knowledge, we are the first to propose automatic evaluation models for personalized generation that reason over implicit preferences, which we hope will benefit the community in both evaluating and training personalized LLMs.

3 Approach

3.1 Data Generation

We describe our data generation approach in two stages: a) **Generation of Personalized Preference Pairs**, and b) **Generation of Unstructured Usage Data**. The final dataset consists of outputs of each stage paired together, i.e. the unstructured user data acts as the input data containing implicit user preferences for the

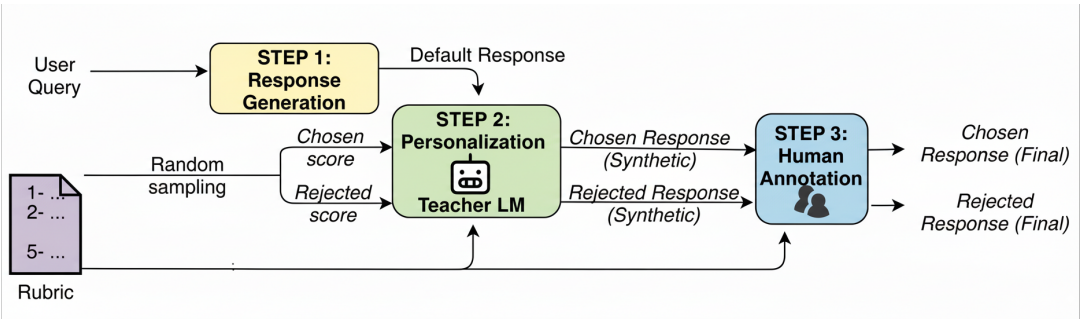


Figure 2: Pipeline for generating personalized responses. Starting with a user query and a data generation rubric, we generate personalized response pairs using a teacher LLM, and validate with human annotation.

chosen-rejected response pairs. This process is used to create both the training and evaluation datasets, with the latter also undergoing human annotation. We list all personalization dimensions explored in Table 1.

3.1.1 Stage 1: Generation of Personalized Responses

Rubric Definition: First, we define synthetic data generation rubrics for each dimension. The full definitions of the rubrics are listed in Appendix A.3. These rubrics, defined on 1-3 or 1-5 scales[3], provide instructions to the teacher LLM on the type and degree of changes to make.

Given a user query for a generation task (we consider Machine Translation, Open-Ended Generation and Story Transcreation) and a rubric for the personalization dimension, Figure 2 depicts our pipeline:

1. **Step 1: Response Generation:** First, we obtain a 'default' non-personalized response to the user query from the teacher LLM. If a reference answer exists, we use it directly.
2. **Step 2: Personalization:** We provide the default response and two randomly sampled rubric scores ('chosen' and 'rejected') to the teacher LLM, which modifies the response to match each score's requirements. We also include self-refinement (Madaan et al., 2023) to further improve conformance.
3. **Step 3: Human Annotation:** For evaluation data only, we show annotators a persona described by the chosen rubric score and its instructions. Given the query and both responses, they select which one better suits this persona's preferences (see Figure 4).

We use this as the first stage for generating all synthetic data in this work, and conduct an 80-10-10 split for creating training, validation, and test sets. The human-annotated test set, after further processing (see Section 3.1.2) constitutes PRM-Bench, our final evaluation benchmark. We use Gemini models (Gemini Team, Google, 2025) as the teacher LLMs: *Gemini 2.5 Pro 05-06* for generating evaluation data and *Gemini 2.0 Flash* for training data. We provide the prompts used in each stage in Table 11.

3.1.2 Stage 2: Generation of Unstructured Usage Data

In this stage, we synthesize unstructured usage data for each training and evaluation example generated previously. This data should contain implicitly expressed preferences that indicate the chosen response being better suited to the user over the rejected one. We establish the following design goals:

1. *Naturalness:* The usage data should resemble 'naturally occurring' device (mobile) data.

[3]We use 1-3 scales when finer distinctions are difficult to define clearly, and 1-5 scales otherwise.

2. *Diversity:* The usage data should exhibit variation across format, content, and stylistic dimensions.
3. *Implicitness:* Following the primary motivation of this work, user preferences should be expressed subtly and inferred from relevant signals, rather than explicitly stated like in prior work.
4. *Signal Sparsity:* Reflecting real-world 'persona sparsity' issues (Dong et al., 2024), information relevant to personalization should be sparse and embedded within broader, unrelated user data.
5. *Multilinguality:* Finally, to align with this work's secondary goal of supporting multilingual personalization evaluation, the pipeline must scale to generate usage data in non-English languages.

To design prompts to the teacher LLM that can achieve these goals, we follow the pipeline given below:

1. **Label Definition:** We begin with the personalization data from Section 3.1.1, using the 'chosen' scores to create a label P containing the personalization dimension, its definition, and score value.
2. **Noise Addition:** To maintain *signal sparsity*, we introduce noise N by randomly sampling annotator profiles from the PRISM dataset (Kirk et al., 2024) (see Section A.3). Combining N with P creates metadata M that is used as a seed for generating synthetic data, ensuring the personalization signal remains subtle within a broader context.
3. **Naturalness and Diversity:** We achieve *naturalness* by framing the synthetic data as logs of mobile applications. For *diversity*, we randomly vary the app categories sampled (Table 13), content length, writing styles (Table 14), and output formats ('rawtext' or 'json').
4. **Implicitness and Multilinguality:** We directly specify requirements for *implicit* expression of preferences and *multilingual* content in the prompt instructions.

The complete prompt template is provided in Table 12. We use *Gemini-2.0-Flash* to synthesize usage data, which we then pair with personalized responses to create our final training and evaluation datasets.

3.2 Personalized Reward Modelling

Reward Models (RMs): Given a user query x, along with a chosen (y_w) and a rejected (y_l) response, RMs with parameters θ are trained to maximize the probability of the chosen response (Stiennon et al., 2020):

$$\text{loss}(\theta) = -\mathbb{E}\left[\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))\right] \tag{1}$$

where σ is the sigmoid function and $r_\theta(x, y)$ is the reward on a sequence with concatenated x and y.

Personalized Reward Models (PRMs): To capture implicit user preferences from usage data D, we extend the standard reward modelling framework to incorporate user-specific information. We construct an augmented prompt x' by concatenating the personalization data D, the personalization dimension A, and the original query x. This formulation allows us to express the PRM loss function as a natural extension of the traditional RM loss:

$$\text{loss}(\theta) = -\mathbb{E}\left[\log(\sigma(r_\theta(x', y_w) - r_\theta(x', y_l)))\right] \quad \text{where} \quad x' = D \circ A \circ x \tag{2}$$

Table 1: Personalization dimensions with their descriptions, rubric scales, and the tasks they are used for.

Dimension	Description	Scale	Machine Translation	Open Generation	Trans-creation
Readability	Adapts vocabulary and grammar complexity of text based on the user's reading preferences.	1-5	✓	✓	✓
Engagement	Adapts content structure and paragraph length to maintain user's engagement.	1-3	✓	✓	✓
Cultural Localization	Localizes text to align with user's cultural background.	1-3	×	×	✓
Creativity	Adapts creative expression and emotional tone of text.	1-5	×	×	✓

Table 2: Statistics (example count and token count statistics of the unstructured user data) for the generated synthetic training and evaluation data, and our final human-annotated benchmark PRM-Bench.

Dataset	# Examples	# Avg Tokens	# Max Tokens	# Total Tokens
Synthetic Training	1.12M	4.08K	38K	4.80B
Synthetic Evaluation	51.10K	4.08K	33K	0.22B
PRM-Bench	13.18K	4.21K	33K	0.06B

4 EXPERIMENTS

4.1 DATA

We approach evaluation of personalization for three multilingual generation tasks: a) *Machine Translation*, b) *Open-Ended ('Open') Generation*, and c) *Transcreation*[4]. As shown in Table 1, we personalize along four dimensions: *readability*, *engagement*, *cultural localization*, and *creativity*, with the last two subjective metrics only explored for the creative task of *transcreation*. Starting from seed datasets (detailed in Appendix A.4), we generate synthetic and human-annotated data for all of these tasks and dimensions, and provide statistics in Table 2. We note about 4K average tokens in the generated usage data, while maximum token count can go up to 33K, highlighting the diversity and complexity of our benchmark.

4.2 MODELS

Personalized LLM-as-a-judge: We evaluate leading LLMs (at the time of experimentation) from the OpenAI (GPT-4o and GPT-4o Mini) and Google (Gemini 2.0 Flashlite, Gemini 2.0 Flash and Gemini 2.5 Pro) families, noting that Gemini 2.5 Pro is the current SOTA text model on the LMArena benchmark (Chiang et al., 2024). Inspired by Dong et al. (2024), we propose a **personalized LLM-as-a-judge** approach, wherein the models receive usage data during inference and are required to make decisions based on implicitly expressed preferences. We provide an example prompt in Table 18.

Personalized Reward Models: We start with RM-Mistral-7B[5] (Xiong et al., 2024) and conduct LoRA fine-tuning (Hu et al., 2022) on our training datasets. We use dropout 0.1 and rank 8 adapters on Query, Key and Value matrices. Training runs for 1 epoch with learning rate 1e-5, warmup ratio 0.3, and cosine scheduler on an Nvidia H100 GPU, leveraging Unsloth (Daniel Han & team, 2023) for memory optimization. We train this PRM on all tasks and dimensions at once, and ablate the impact of multi-task fine-tuning in Table 20.

[4] A translation-adjacent task wherein the goal is to combine linguistic translation with cultural localization and creative reinterpretation for different cultures (Díaz-Millón & Olvera-Lobo, 2023)

[5] https://huggingface.co/weqweasdas/RM-Mistral-7B

Table 3: Classification accuracies of various prompting strategies and Gemini models for Personalized LLM-as-a-Judge on the synthetic *Readability* test set. CoT = Chain-of-Thought (Wei et al., 2022); Extract = two-step extraction then reasoning. We do not explore CoT for Gemini 2.5 Pro as it uses a 'Thinking' process.

Model	**Prompting Method**	**Accuracy**						
		Overall	En-Zh	En-De	En-Hi	Zh-En	De-En	Hi-En
Gemini 2.0 Flash-Lite	Score	0.55	0.56	0.58	0.54	0.55	0.54	0.52
	CoT + Score	0.60	0.59	0.66	0.58	0.61	0.60	0.56
	Extract → CoT + Score	0.63	0.61	0.65	0.59	0.64	0.66	0.63
Gemini 2.0 Flash	Score	0.59	0.61	0.62	0.56	0.60	0.58	0.55
	CoT + Score	0.64	0.64	0.68	0.61	0.65	0.64	0.62
	Extract → CoT + Score	0.65	0.64	0.68	0.60	0.66	0.66	0.66
Gemini 2.5 Pro	Score	**0.70**	**0.69**	**0.76**	0.63	0.70	**0.72**	**0.69**
	Extract → Score	**0.70**	**0.69**	0.75	**0.65**	**0.71**	0.71	**0.69**

Table 4: Performance of Gemini 2.5 Pro with various input types for Personalized LLM-as-a-Judge on *Readability* test set. Building on the best-performing approach from Table 3, we compare: D = unstructured user data (implicit preferences), P = explicit user preference labels and M = structured metadata combining noise N and preferences P (refer Section 3.1.2). 'Non-personalized ranking' follows standard LLM-as-a-judge (Zheng et al., 2023b), where the goal is simply to rank readability of two responses.

Model	**Input**	**Accuracy**						
		Overall	En-Zh	En-De	En-Hi	Zh-En	De-En	Hi-En
Gemini 2.5 Pro	User Data (D)	0.70	0.69	0.76	0.63	0.70	0.72	0.69
	Metadata (M)	0.84	0.86	0.87	0.83	0.83	0.83	0.82
	Shuffled Metadata (M')	0.84	0.86	0.87	0.83	0.83	0.83	0.82
	Persona Preference Label (P)	0.86	0.86	0.87	0.83	0.85	0.87	0.85
	Non-personalized ranking	0.95	0.93	0.94	0.93	0.96	0.96	0.96

4.3 Results

4.3.1 Preliminary Experiments: Readability for Open Generation

To better understand this novel problem, we begin with one personalization dimension—*Readability*—for *Open-Ended Generation*, focusing on cross-lingual pairs where source and target languages differ. Table **??** presents our results on the synthetic validation dataset, revealing several key insights.

Even SOTA LLMs struggle with the Personalized LLM-as-a-judge task. Starting with the general-purpose Gemini Flash-Lite and Flash models, Table 3 shows they perform quite poorly off-the-shelf (55-60% accuracies), barely above random chance (50%). While Chain-of-Thought prompting (Wei et al., 2022) pushes accuracies towards 60-65%, significant room for improvement remains. Even Gemini 2.5 Pro (the current SOTA LLM) achieves only 70% accuracy—*highlighting the complexity of our problem*. Next, we experiment with two-turn prompting to first retrieve relevant extracts, and then reason over the smaller context. While some gains emerge for weaker Flash models, no improvement is observed for Gemini 2.5 Pro—suggesting a *fundamental bottleneck in reasoning, rather than long-context retrieval abilities.* Beyond average accuracies, performance is markedly worse on lower-resourced languages like Hindi, reflecting *a significant gap in multilingual personalization evaluation* and establishing our benchmark's relevance.

Table 5: Synthetic Evaluation: Accuracies of Personalized LLM-as-a-Judge baselines versus a Personalized Reward Model (PRM) for 3 multilingual generation tasks and 4 dimensions: *Readability (Read.)*, *Engagement (Engage.)*, *Creativity (Creat.)*, and *Cultural Localization (Local.)*. Scores averaged across 3 seeds.

Model	Overall	Machine Translation		Open Generation		Transcreation			
		Read.	*Engage.*	*Read.*	*Engage.*	*Read.*	*Engage.*	*Creat.*	*Local.*
Gemini 2.0 Flash Lite	0.63	0.61	0.67	0.59	0.68	0.61	0.67	0.65	0.52
Gemini 2.0 Flash	0.67	0.66	0.73	0.61	0.73	0.65	0.73	0.67	0.57
Gemini 2.5 Pro	**0.72**	**0.71**	**0.77**	**0.67**	**0.77**	**0.71**	**0.76**	0.71	**0.64**
GPT 4o Mini	0.65	0.66	0.72	0.62	0.70	0.67	0.72	0.59	0.52
GPT 4o	0.70	0.70	0.76	0.65	0.75	0.69	0.76	0.67	0.57
PRM-Mistral-7B-LoRA	0.68	0.66	0.73	0.61	0.72	**0.71**	0.73	**0.72**	**0.64**

Reasoning over implicit preferences and user experiences are the key bottlenecks To identify the factors behind the low scores, we conducted data ablations (Table 4) by varying the inputs to our best-performing model, Gemini 2.5 Pro: a) unstructured user data D (default), b) structured metadata M with explicit preferences P, and c) preference labels P directly (see Section 3.1.2). We see a 20% gain as we switch the input from D to M, demonstrating that *reasoning over implicit user preferences* – the central motivation of this work – remains highly challenging for even SOTA reasoning LLMs. Interestingly, however, context length does not appear to be a limiting factor: shuffling M to randomize the position of P maintains accuracy, contradicting the 'lost-in-the-middle' phenomenon reported for older LLMs (Liu et al., 2024). Moreover, providing P directly (i.e. removing N from M) yields only marginal gains, suggesting Gemini's long-context capabilities are sufficiently robust in this setting. The second major bottleneck emerges when comparing personalized versus standard evaluation: when the LLM is asked to rank ease of readability in a standard LLM-as-a-judge setting, accuracy increases to 95%. This suggests that *reasoning over user context and personalizing response complexity* poses the second key challenge for LLM-as-a-judge in this setting.

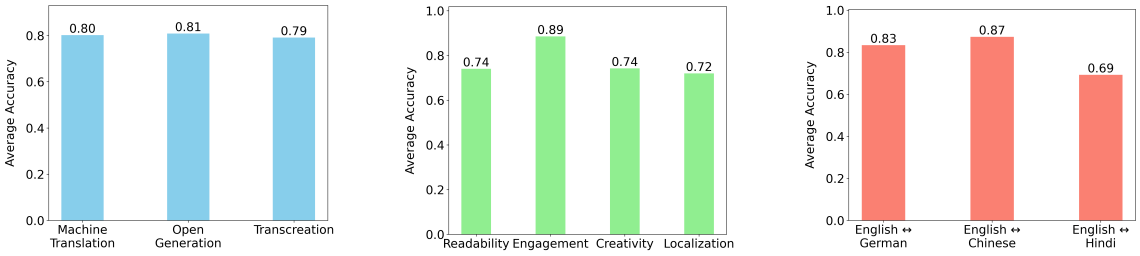
4.3.2 Scaling up Synthetic Evaluation

Trends hold at scale, albeit subjective dimensions pose a greater challenge. Table 5 presents our overall results, scaled up across all three multilingual generation tasks and four personalization dimensions. We note that the trends with personalized LLM-as-a-judge mostly hold across all tasks and dimensions: Gemini 2.5 Pro outperforms the Flash and Flash Lite models respectively. The OpenAI models outperform the Gemini Flash models, and are better general-purpose (non-reasoning) LLMs, aligning with LMArena rankings (Chiang et al., 2024). Interestingly, we also observe that performance on the subjective dimensions, like *Cultural Localization* and *Creativity* is noticeably worse than on more objective ones like *Engagement* that are better defined (in terms of paragraph length, content structure etc.) and therefore easier to identify.

LoRA fine-tuned 7B PRMs can be competitive, or better than LLM judges. We report the results of our fine-tuned PRM-Mistral-7B-LoRA model, trained jointly across all generation tasks and personalization dimensions. This model outperforms high-performing, proprietary LLM judges of comparable size, like GPT 4o Mini and Gemini 2.0 Flash Lite, and performs comparably with Gemini 2.0 Flash. Further, for subjective dimensions, it outperforms Gemini 2.5 Pro and achieves SOTA results on *Readability*, *Creativity* and *Localization* in the Transcreation task. We show through ablations in Table 20 how joint training across multiple tasks and dimensions enables it to achieve SOTA performance on these challenging settings, where uni-task training proves insufficient. Overall, our PRM's competitive performance and potential efficiency gains demonstrate its promise as a practical tool for training personalized LLMs in future work.

Table 6: PRM-Bench: Accuracies of Personalized LLM-as-a-Judge baselines versus a Personalized Reward Model (PRM) for 3 multilingual generation tasks and 4 dimensions. All scores are averaged across 3 seeds.

Model	Overall	Machine Translation		Open Generation		Transcreation			
		Read.	*Engage.*	*Read.*	*Engage.*	*Read.*	*Engage.*	*Creat.*	*Local.*
Gemini 2.0 Flash Lite	0.64	0.63	0.70	0.59	0.69	0.65	0.69	0.62	0.57
Gemini 2.0 Flash	0.64	0.62	0.72	0.60	0.69	0.63	0.70	0.63	0.57
Gemini 2.5 Pro	**0.68**	**0.65**	**0.76**	0.63	**0.74**	**0.67**	0.74	**0.65**	**0.62**
GPT 4o Mini	0.55	0.52	0.57	0.52	0.59	0.55	0.58	0.54	0.52
GPT 4o	0.61	0.56	0.67	0.58	0.67	0.58	0.69	0.60	0.56
PRM-Mistral-7B-LoRA	0.63	0.60	0.67	**0.65**	0.68	0.66	**0.75**	0.64	0.57



(a) By Task (b) By Dimension (c) By Language Pair

Figure 3: Average agreement between synthetic labels (Section 3.1) and human preferences

4.3.3 Human Evaluation

PRM-Bench trends are similar, but more challenging than synthetic evaluation. Table 6 confirms overall synthetic evaluation trends: Gemini 2.5 Pro achieves SOTA overall but lags in subjective dimensions like *Creativity* and *Localization*. PRM-Mistral-7B-LoRA competes with Flash models, outperforms GPT 4o, and sometimes 2.5 Pro. However, all models – especially non-thinking variants (GPT, Gemini Flash) – show accuracy drops, suggesting PRM-Bench is more challenging than anticipated, and synthetic data may contain false positives despite its utility for low-cost evaluation.

Synthetic data mostly aligns with human preferences Figure 3 shows human agreement with synthetic labels across tasks, dimensions, and languages. We observe $\sim$80% agreement across tasks, indicating our pipeline generates synthetic data of reasonable quality. By dimension, accuracies reach 89% for well-defined criteria like *Engagement* but only 72% for subjective dimensions like *Localization* and *Creativity*, aligning with Table 5 findings. For language pairs, the lower-resourced English-Hindi pair shows weaker accuracy (69%) compared to German (83%) and Chinese (87%), likely due to language disparities in the teacher LLM. Improving data synthesis for low-resource languages and subjective dimensions is left to future work.

5 Conclusion

In this work, we propose the novel problem of personalized reward modelling (PRM) from implicit user preferences and introduce a pipeline to synthesize usage data with implicit preferences and personalized question-response pairs. We benchmark both personalized LLM-as-a-judge and PRMs, and show that even SOTA LLMs underperform on this task Notably, we observe that small PRMs can outperform SOTA LLM judges, highlighting their potential for future research.

REFERENCES

Apple. Private cloud compute: A new frontier for ai privacy in the cloud, June 2024. URL https://security.apple.com/blog/private-cloud-compute/. Accessed: 2025-07-19.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. Whose boat does it float? improving personalization in preference tuning via inferred user personas, 2025. URL https://arxiv.org/abs/2501.11549.

Donna K Byron, Ashok Kumar, Alexander Pikovsky, and Mary D Swift. Travel itinerary recommendation engine using inferred interests and sentiments, August 24 2021. US Patent 11,100,557.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1kFDrYCuSu.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Clyde Hamilton Coombs. Psychological scaling without a unit of measurement. In *Scaling*, pp. 281–299. Routledge, 2017.

Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge, UK, 2001. URL https://www.coe.int/en/web/common-european-framework-reference-languages/the-cefr-in-short.

Michael Han Daniel Han and Unsloth team. Unsloth, 2023. URL http://github.com/unslothai/unsloth.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. Wmt24++: Expanding the language coverage of wmt24 to 55 languages dialects, 2025. URL https://arxiv.org/abs/2502.12404.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can LLM be a personalized judge? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10126–10141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL https://aclanthology.org/2024.findings-emnlp.592/.

Mar Díaz-Millón and María Dolores Olvera-Lobo. Towards a definition of transcreation: a systematic literature review. *Perspectives*, 31(2):347–364, 2023. doi: 10.1080/0907676X.2021.2004177. URL https://doi.org/10.1080/0907676X.2021.2004177.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind, June 2025. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf. Technical Report.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Vivek Iyer, Ricardo Rei, Pinzhen Chen, and Alexandra Birch. Xl-instruct: Synthetic data for cross-lingual open-ended generation, 2025. URL https://arxiv.org/abs/2503.22973.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024. URL https://openreview.net/forum?id=EMrnoPRvxe.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale, 2025. URL https://arxiv.org/abs/2504.14225.

Brihi Joshi, Keyu He, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, and Xiang Ren. Eli-why: Evaluating the pedagogical utility of language model explanations, 2025. URL https://arxiv.org/abs/2506.14200.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105236–105344. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be2e1b68b44f2419e19f6c35a1b8cf35-Paper-Datasets_and_Benchmarks_Track.pdf.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith, and Hannaneh Hajishirzi. ComPO: Community preferences for language model personalization. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8246–8279, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.419/.

Jihyun Lee, Yejin Jeon, Seungyeon Seo, and Gary Lee. PicPersona-TOD : A dataset for personalizing utterance style in task-oriented dialogue with image persona. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7937–7958, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.403/.

Tong Li, Yong Li, Mingyang Zhang, Sasu Tarkoma, and Pan Hui. You are how you use apps: user profiling based on spatiotemporal app usage behavior. *ACM Transactions on Intelligent Systems and Technology*, 14(4):1–21, 2023a.

Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback, 2024. URL https://arxiv.org/abs/2402.05133.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023b.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. Controlling the reading level of machine translation output. In Mikel Forcada, Andy Way, Barry Haddow, and Rico Sennrich (eds.), *Proceedings of Machine Translation Summit XVII: Research Track*, pp. 193–203, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6619/.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. Motivating personality-aware machine translation. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1102–1108, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1130. URL https://aclanthology.org/D15-1130/.

G Neelima and Sireesha Rodda. Predicting user behavior through sessions using the web log mining. In *2016 International Conference on Advances in Human Machine Interaction (HMI)*, pp. 1–5. IEEE, 2016.

Peter Newmark. *A textbook of translation*, volume 66. Prentice hall New York, 1988.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4226–4237, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.377.

Xing Niu, Marianna Martindale, and Marine Carpuat. A study of style in machine translation: Controlling the formality of machine translation output. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2814–2819, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1299. URL https://aclanthology.org/D17-1299/.

Itay Niv. Let's read a story! https://github.com/itayniv/aesop-fables-stories, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Matthias Paulik, Matt Seigel, Henry Mason, Dominic Telaar, Joris Kluivers, Rogier van Dalen, Chi Wai Lau, Luke Carlson, Filip Granqvist, Chris Vandevelde, Sudeep Agarwal, Julien Freudiger, Andrew Byde, Abhishek Bhowmick, Gaurav Kapoor, Si Beaumont, Áine Cahill, Dominic Hughes, Omid Javidbakht, Fei Dong, Rehan Rishi, and Stanley Hung. Federated evaluation and tuning for on-device personalization: System design applications, 2021. URL https://arxiv.org/abs/2102.08503.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1074–1084, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1101/.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. Can language models recognize convincing arguments? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8826–8837, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.515. URL https://aclanthology.org/2024.findings-emnlp.515/.

Apple Machine Learning Research. Understanding aggregate trends for apple intelligence using differential privacy, April 2025. URL https://machinelearning.apple.com/research/differential-privacy-aggregate-trends. Accessed: 2025-07-19.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 35–40, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1005. URL https://aclanthology.org/N16-1005/.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025. URL https://openreview.net/forum?id=7PSm6uN17i.

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

J. H. Stickney and Aesop. *Æsop's Fables: A Version for Young Readers*. Project Gutenberg, 2015. URL https://www.gutenberg.org/ebooks/49010.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 2: Text. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=LWD7upg1ob.

Wei Xiong, Hanze Dong, and Rui Yang. Reward modeling for rlhf. Web page, Mar 2024. URL https://www.notion.so/Reward-Modeling-for-RLHF-abe03f9afdac42b9a5bee746844518d0.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.826. URL https://aclanthology.org/2024.acl-long.826/.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. Personalization of large language models: A survey, 2025a. URL https://arxiv.org/abs/2411.00027.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. Personalization of large language models: A survey. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL https://openreview.net/forum?id=tf6A9EYMo6. Survey Certification.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023b. URL https://openreview.net/forum?id=uccHPGDlao.

A Data Appendix

A.1 Preliminaries

A.1.1 Goal: Personalized Evaluation from Implicit Preferences

Personalized Generation (Zhang et al., 2025b) is defined as the task of adapting text according to user preferences. In this work, we focus on the preferences of the end user (ie. reader).

Personalized Evaluation refers to assessing how closely an adapted response aligns with user preferences along a given *personalization dimension* – defined as the axis along which a given piece of text is adapted, which determines the type of changes required.

While prior work has explored evaluation using explicitly stated preferences for a given dimension (Jang et al., 2024; Dong et al., 2024), in this work, we approach the task of implicitly inferring these preferences from long-context user data, with the goal of synthetically simulating on-device data. We provide further details on the relevant considerations for generating such data in Section 3.1.2.

Therefore, formally, given the following inputs:

1. A user's **application usage data**, henceforth referred to as data dump D.
2. A desired **personalization dimension** A (with the full list provided in Table 1).
3. The **user prompt** x and a **candidate response** y.

Personalized Evaluation from Implicit Preferences is the task of estimating a function f that can provide the reward r, where

$$r = f(x, y, D, A).$$

A.1.2 Scope: Tasks Studied

Next, we define the three multilingual and cross-lingual tasks explored in this work:

The following section delineates three multilingual and cross-lingual tasks investigated in this work:

1. **Personalized Machine Translation:** Machine Translation is traditionally defined as the conversion of text or speech from source to target language while preserving semantic content and stylistic nuance. In the personalization paradigm, however, following Niu et al. (2017) and Marchisio et al. (2019), we extend this definition to encompass style adaptation based on user preferences while maintaining semantic fidelity. Our investigation focuses on two dimensions: *Readability* (to enhance comprehension) and *Engagement* (to adapt response length and structure, with particular applicability to long-context MT tasks).
2. **Personalized Open-Ended Generation:** Open-Ended Generation is defined as the production of novel, diverse, and coherent text in response to user queries, constituting a primary LLM application. Given evidence that default LLM responses demonstrate bias toward advanced proficiency (Joshi et al., 2025) and excessive verbosity (Koo et al., 2024), we posit that personalization along *readability* and *engagement* dimensions addresses significant practical requirements.
3. **Personalized Transcreation:** Transcreation is defined as a translation-adjacent task that integrates linguistic translation, creative reinterpretation, and cultural adaptation to maximize audience engagement (Díaz-Millón & Olvera-Lobo, 2023). Our evaluation framework thus incorporates *Cultural Localization* and *Creativity* alongside readability and engagement. To our knowledge, we are the first to explore such subjective metrics in personalized text generation literature.

Table 7: Readability Rubric Definition. Each score on this scale corresponds to CEFR levels (Council of Europe, 2001) of language proficiency. Figure 2 explains how such rubrics are used to generate personalized responses.

Consider the following rubrics which defines, on a scale of 1-5, how proficient a user is in the language of a given piece of text. It also describes the series of modifications that should be made to this text to adapt it for their proficiency levels:

1. **Score 1 (Beginner - A1):** 0-1 year of experience. Very limited vocabulary and grammar. Use extremely simple words, short sentences, no idioms, or complex grammar. Example: "I like apples."
2. **Score 2 (Elementary - A2):** 1-2 years. Basic phrases and understanding of common topics. Use simple vocabulary, slightly longer sentences, avoid slang, and explain any idiomatic phrases. Example: "The cat is on the roof."
3. **Score 3 (Intermediate - B1):** 2-4 years. Can handle everyday tasks. Use varied vocabulary, moderately complex sentences, and light idioms with explanation. Example: "She runs errands every day."
4. **Score 4 (Upper-Intermediate - B2):** 4-7 years. Can discuss abstract topics. Use advanced vocabulary, complex sentences, idioms, and metaphors with minimal explanation. Example: "He hit the nail on the head."
5. **Score 5 (Proficient - C1/C2):** 7+ years. Near-native fluency. Use nuanced vocabulary, advanced grammar, and idiomatic expressions freely. Example: "The plan fell apart due to unforeseen circumstances."

These rubrics will be used for a personalization task, wherein given a score(s), you have to adapt the input text such that it uses language that matches the linguistic proficiency of the end user reading the text.

Table 8: Engagement Rubric Definition

Consider the following rubrics which define, on a scale of 1{3, the degree of attention span of a user:

1. **Score 1 (Short):** The user prefers short, snappy, and engaging material. Responses should consist of brief paragraphs (2{3 sentences max) that deliver the most information in the least amount of reading time. Avoid elaboration and focus on punchy, to-the-point content.
2. **Score 2 (Medium):** The user is comfortable with medium-length paragraphs. Responses can include moderate detail and explanation, typically 5 sentences per paragraph. Ensure the content is still concise but allows for some elaboration and clarity.
3. **Score 3 (Long):** The user enjoys reading long-form content and prefers structured, well-developed responses. Responses should use fluent, detailed language with multiple paragraphs and thorough explanations. Provide depth and context, similar to content in books or newspapers.

These rubrics will be used for a personalization task, wherein given a score(s), you have to adapt the input text such that it is best suited for the corresponding attention span, as dictated by this rubric.

Table 9: Creativity Rubrics Definition

Consider the following rubrics which defines, on a scale of 1{5, the user-desired degree of creative and emotional freedom for an AI language model:

1. **Score 1 (Safe & Standard):** The AI should strictly follow instructions and use common, predictable patterns. Avoid any creativity or strong emotions (positive or negative), keeping the output neutral.

2. **Score 2 (Slightly Creative, Mostly Safe):** The AI can include minor creative elements and mild positive emotions (like helpfulness or politeness). Outputs should remain conventional and safe, avoiding strong negative feelings or controversy.

3. **Score 3 (Balanced Creativity & Emotion):** The AI should mix some originality with standard approaches. It can display typical positive and negative emotions (as seen in basic stories), but steer clear of extremes, explicit content, or controversial subjects.

4. **Score 4 (Very Creative, Emotionally Rich { Mostly Positive):** The AI should be highly original and express a wide range of emotions, emphasizing strong positive ones (like joy or inspiration). Push creative boundaries but generally avoid explicitly controversial, dark, or risqué topics.

5. **Score 5 (Full Creative Freedom { Human-like Range):** The AI should have complete creative and emotional freedom, similar to a human artist. This encompasses the entire spectrum, from highly positive/inspirational content to potentially exploring dark, explicit, controversial, or risqué themes, prioritizing unrestricted expression.

These rubrics will be used for a personalization task, wherein given a score(s), you have to adapt the AI's output style such that it aligns with the corresponding level of creative and emotional freedom, as dictated by this rubric.

Table 10: Cultural Localization Rubrics. These rubrics are inspired from and use concepts proposed in Newmark (1988)'s seminal textbook on Translation.

Consider the following rubrics which define, on a scale of 1{3, the degree of cultural localization for content adaptation:

1. **Score 1 (Maximal Fidelity & Preserving Authenticity):** This score represents a preference for minimal cultural adjustments, prioritizing the preservation of the original content's form and authenticity. Strategies focus on directly transferring terms, naturalizing their pronunciation, or translating literally where meaning is clear. The aim is to introduce foreign concepts or maintain the original identity with subtle adaptations for relatability, ensuring that any localized nuances integrate naturally without sounding out of place.
*** Newmark's Strategies (Maximal Fidelity & Preserving Authenticity):**
• Transference (including transliteration and transcription): Directly borrowing words or names.
• Naturalization: Adapting borrowed words to fit TL pronunciation/morphology.
• Through-translation (Calque or Loan Translation): Literal translation of common phrases or names.
• Recognized translation: Using official or generally accepted translations for institutional terms.

2. **Score 2 (Balanced/Intermediate Localization):** A balanced approach to localization is preferred at this level, where the content is adapted to include surface-level culturally relevant elements while retaining much of the original storyline. Strategies involve replacing cultural terms with functional or descriptive equivalents, utilizing componential analysis, or finding close cultural parallels. The goal is to make the content more accessible and connected to the target audience by incorporating appropriate examples and aligning with local customs, without significantly overhauling the original material. **Importantly, this also includes changing the names of characters and places in the story to better suit the target culture.** The localized elements should flow naturally, while still retaining the broader elements of the original story.
*** Newmark's Strategies (Balanced Approach & Equivalence):**
• Functional Equivalent: Using a culture-neutral word to convey the item's function.
• Descriptive Equivalent: Explaining the meaning of a cultural term in 2{3 words.
• Componential Analysis: Comparing SL and TL words by breaking down their sense components.
• Synonymy: Using a "near TL equivalent" when economy is preferred over strict accuracy.
• Cultural Equivalent: Replacing an SL cultural word with a similar TL cultural word.
• Couplets: Combining two different procedures (e.g., a descriptive equivalent with a naturalized term) to achieve a balanced outcome.

3. **Score 3 (Maximal Creativity & Localization):** This score signifies a desire for deeply localized content that resonates strongly with the local culture, involving creative and significant changes. Strategies extend to grammatical shifts, changes in perspective (modulation), or compensating for lost meaning elsewhere in the text. This can involve reimagining characters, morals, events, or perspectives to ensure the content feels natural, engaging, and fully aligned with the cultural context, as if it were originally created for the target audience. **Importantly, this also includes changing the names of characters and places in the story to better suit the target culture.** The story should sound like it was written locally for the target culture.
*** Newmark's Strategies (Maximal Creativity & Localization / Transcreation-like):**
• Shifts or Transpositions: Grammatical changes from SL to TL to achieve natural flow.
• Modulation: Reproducing the message from a different viewpoint to conform with TL norms.
• Compensation: Making up for loss of meaning or effect in one part of the sentence/text elsewhere.

These rubrics will be used for a personalization task, wherein given a score(s), you have to adapt the input text such that it is localized to fit the corresponding target culture, as dictated by this rubric.

Table 11: The pipeline used for generating personalized responses {final_pair} in this work, given a user query {src_sent}, a rubric {rubric}, and source and target languages ({src_lang}, {tgt_lang}). We show the prompts used at each stage in columns 1 and 2, along with the output variable name in column 3, which is then used as one of the input variables in the next stage. *This stage is skipped if a reference answer is already provided, in which case this is used as the baseline response. (refer Section 3.1.1)

Stage	Prompt	Output
Response Generation*	**{Machine Translation, Transcreation}:** Translate this text from [src_lang] to [tgt_lang]: [src_sent]. Output NOTHING else except the translation. --- **Open Generation:** Question: [src_sent]. Respond in [tgt_lang]. Output NOTHING else except the response to this question.	[baseline]
Personalization	[rubric] You are given this [tgt_lang] language output from the [domain] domain: [baseline] You are given two scores: [chosen_score] and [rejected_score]. You need to amend the above text such that the personalized version is reworded to match the characteristics defined for these scores in the rubrics. The language of the final response should be [tgt_lang]. The rewording should by and large follow the tone and style of the [domain] domain, similar to the source. Ensure that you do not include your own information. Produce a JSON object that contains a single key 'response' that contains a list of two JSON elements, which are the two reworded responses. Each JSON element should have 'score' and 'output' attributes with 'score' indicating the provided scores, and 'output' indicating the personalized response. ```json	[response_pair]
Self-Feedback	Rubric: [rubric] Source sentence: [src_sent] Reference sentence: [tgt_sent] response_pair scores: [chosen_score] and [rejected_score] Personalized responses (model output): [response_pair] Can you critique the personalized responses (which are an LLM-modified version of the 'normal response') and provide constructive feedback on how to improve and refine the model output further? The goal is to make the output follow the provided rubric better.	[feedback]
Self-Refine	Rubric: [rubric] Input text: [translation] response_pair scores: [chosen_score] and [rejected_score] Personalized text (model output): [response_pair] Feedback: [feedback] Can you generated a revised version of the personalized text incorporating the feedback above? Output nothing else except the revised personalized text JSON object. The structure should be EXACTLY the same as before, this is very important.```json	[refined_pair]
Clean JSON	Clean the provided LLM response such that it contains a JSON object that contains a single key 'response' which contains a list of two JSON elements, each with 'score' and 'output' keys. Everything else is to be deleted. Here is the LLM output: [refined_pair] The list of two JSONs is to be arranged such that the respective scores follow this serial order: [chosen_score] and [rejected_score] I will be providing your output directly to json.loads in python so ensure your output is in the right format.	[final_pair]

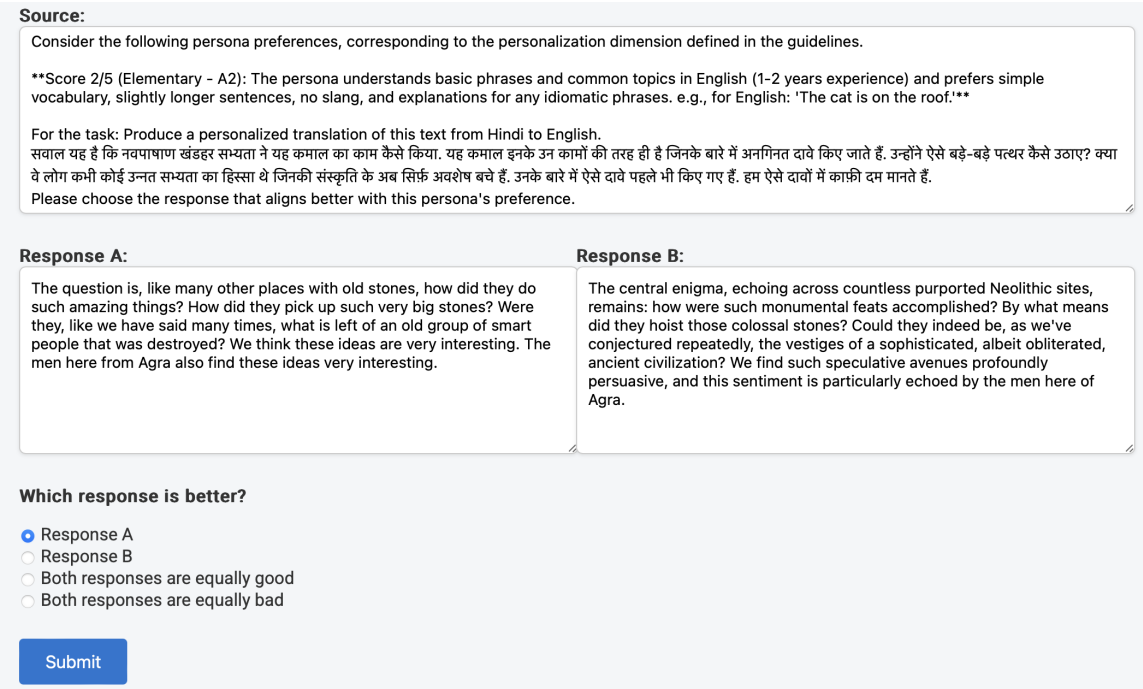


Figure 4: The annotation interface we use to validate synthetically generated personalized responses.

A.2 Stage 1: Generation of Personalized Responses

We give an example of a rubric for the *Readability* dimension in Table A.1.2. Such rubrics are manually defined and are used to generate synthetic data using the pipeline described in Section 3.1.1. We provide the complete list of rubrics here. When possible, we define the rubrics based on prior research. For instance, the *Cultural Localization* rubric for the *transcreation* task draws from concepts proposed in seminal Newmark's textbook on Translation (Newmark, 1988). Specifically, we survey various localization strategies proposed by Newmark (1988) like Naturalization, Theory of Equivalents, Transpositions etc. and classify them on a three-point scale, ranging from 1 (direct translation, maximal fidelity) to 3 (maximal creativity and localization) to create our final rubric. We apply similar methods while defining the other rubrics as well, with *Readability* and *Engagement* directly drawing from prior personalization research (Marchisio et al., 2019; Shenfeld et al., 2025). These rubrics are used to generate synthetic data with the prompt pipeline shown in Table A.1.2.

After generating synthetic data, we conduct human annotation, with the interface provided to the annotators shown in Figure 4. The annotators act as *third-party human evaluators*, similar to prior work (Dong et al., 2024) – wherein they are asked to assume the role of a hypothetical persona with given preferences, and asked to choose which response the persona would prefer. This process could, thus, be subjective, and while we try to mitigate that by soliciting multiple annotators and filtering for cases where there is a clear majority vote, we leave it to future work to explore first party annotation as a potentially 'purer' signal.

Table 12: The full prompt used for generating unstructured usage data.

```
You are given the following metadata about an annotator who responded to a survey:
{metadata_including_label}

You are also given their conversation history with an AI language model:
{chat_history}

Given the above user metadata and the AI conversation history, you need to generate a
data dump in {src_lang} language that would be present locally on this persona's mobile
device. Please generate data that follows the goals below:

1. The language of the data dump must strictly be in {src_lang} language alone.

2. The produced data dump must sound natural and authentic to the persona, and must
resemble data that would naturally occur on a mobile device.

3. The data dump must use all the information given in the metadata and the
conversation history, to the extent that it sounds natural. Do not include the
annotations or metadata verbatim. Instead, weave it subtly into the data dump you
generate such that the corresponding facts can be inferred/reasoned.

4. In particular, ensure that the data dump includes information about the
{personalization_dimension} attribute.

5. Here are some instructions about the quantity and type of data needed in the final
response: {message_requirements}. The persona uses the following characteristics
while texting: {message_decoration}. Integrate this in the generated data.

6. Always include fictional names and emails in your response, but never include
timestamps. Never use placeholders. For app names, always use real apps.

7. The data dump should be in the following format: {output_type}
```

Table 13: The parameters used while defining {message_requirements}. We randomly sample a subset of 'Categories' and for each sampled item, we sample an element from 'Lengths' to create our final JSON.

{message_requirements}			
Categories			**Lengths**
Social Media Posts	*Web Searches*	*Reviews*	*Short: 1-3 sentences*
Social Media Activity	*Notes*	*Blogs*	*Medium: 5-7 sentences*
Emails	*Messages*	*AI Chat History*	*Long: 10-15 sentences*
App Notifications			

Table 14: The possible values for {message_decoration} that we randomly sample from.

{message_decoration}				
Emojis	*Hashtags*	*Abbreviations*	*Links*	*Punctuation Quirks*
Mentions	*Letter Repetition*	*Emoticons*	*Quotes*	*Slang*

A.3 Stage 2: Generation of Unstructured Usage Data

While our main pipeline is summarized in Section 3.1.2, we provide some additional details in this section on the process followed for generation of unstructured usage data. We provide the full prompt to the teacher LLM in Table A.2, and explain below how the placeholder variables in the prompt are created:

Sampling personas from PRISM ({`metadata_including_label`}, {`chat_history`}): We randomly sample annotator profiles from PRISM, particularly from the 'survey' and 'conversations' fields. From the 'survey' field, we filter out the following keys:

1. 'readability': 'english_proficiency'
2. 'engagement': 'self_description'
3. 'cultural_localization': 'self_description', 'study_locale', 'religion', 'ethnicity', 'location'
4. 'creativity': 'system_string'

These keys might conflict with the label we provide for each of these personalization dimensions, which might propagate to the metadata and potentially even to the unstructured usage data we create, and hence we remove them. In the 'conversations' field, we make the following changes:

1. We filter out 'controversy guided' conversations to avoid offensive content.
2. We filter out conversation turns from the Meta Llama models (Touvron et al., 2023), due to licensing restrictions with training on the outputs from these models
3. We remove the 'open_feedback', 'performance_attributes' and 'choice_attributes' blocks as these will likely conflict with the preference labels we inject too.

The filtered metadata is concatenated with the label to create {`metadata_including_label`}, while the filtered conversations are called {`chat_history`}. These are fed into the prompt in Table A.2. These sampled attributes from PRISM dataset provides diverse beliefs, habits, and demographics of a hypothetical persona that help prevent the generated user data overfitting to the personalization label alone.

Message Requirements ({`message_requirements`}) for Content Diversity: We generate the {`message_requirements`} variable that specifies our requirements for content diversity through the following three-stage process. All parameters are listed in Table 13.

1. **Random subsampling of application categories:** We randomly select a subset of 'Categories' by first choosing a random integer between 1 and 10 (the total length of the 'Categories' array), and then sampling a random subarray of that size.
2. **Randomized sampling of posts per category:** Next, for each selected category, we randomly determine the number of posts required by choosing an integer less than or equal to `MAX_POSTS` (set to 20 in our experiments).
3. **Randomized length assignment:** Finally, for each post, we specify its desired length by randomly selecting an element from the 'Lengths' array defined in Table 13.

Message Decoration ({`message_decoration`}) for Tonal Diversity: Next, we generate requirements for 'tonal diversity' in order to accomodate for different writing styles. In Table 14, we enlist the 'decorations' array we consider. We randomly subsample three elements from this list to create the message_decoration} array.

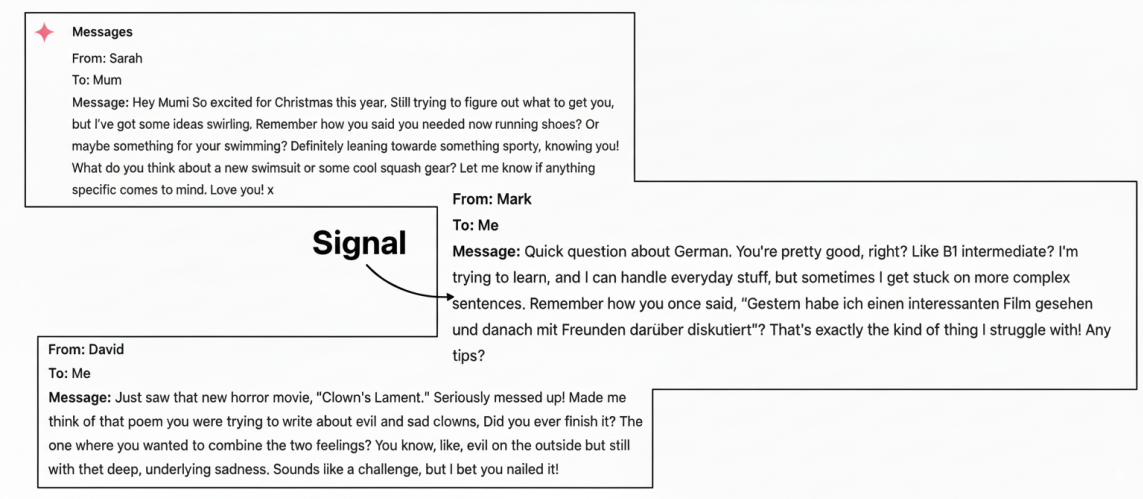


Figure 5: Example of the synthetic user data. We show three emails, out of which only one contains information relevant (*Signal*) to the *Readability* personalization dimension, and the rest comprise 'noise'.

Defining Output Type ({`output_type`}) : Finally, the output format of the unstructured data could take one of two possible values: a) `rawtext` or b) `json`. This, too, is randomly chosen per element to allow for diversity of samples during both training and evaluation.

These variables are populated in the prompt described in Table 12, and passed to the teacher LLM, Gemini 2.0 Flash, which generates the required synthetic data. We provide an example in Figure 5, wherein we show three generated emails, only one of which contains information relevant to the personalization dimension.

Table 15: Statistics of seed datasets used for generating synthetic data, as described in Section 3.1

Task	Evaluation Data		Training Data	
	Dataset	#Examples	Dataset	#Examples
Machine Translation	WMT24++ (Deutsch et al., 2025)	998	XL-Instruct (Iyer et al., 2025)	45.3K
Open Generation	Filtered AlpacaEval (Iyer et al., 2025)	798	Alpaca (Taori et al., 2023)	52K
Transcreation	Aesop's Fables (Stickney & Aesop, 2015)	147	Gemini Stories *(ours)*	144K

Table 16: Statistics for all datasets contributed in this work. We provide example count for each task and dimension, along with token statistics for the synthesized unstructured user data — divided into subtables for training, evaluation, and the human-annotated benchmark (PRM-Bench).

(a) Synthetic Training data statistics.

	Machine Translation		Open Generation		Transcreation			
	Read.	*Engage.*	*Read.*	*Engage.*	*Read.*	*Engage.*	*Creat.*	*Local.*
#Examples	240K	238K	167K	140K	91K	83K	79K	81K
#Avg Tokens	4.5K	4.8K	4.4K	4.7K	3.6K	3.7K	3.4K	3.5K
#Max Tokens	37K	38K	37K	35K	30K	31K	30K	29K
#Total Tokens	1.08B	1.14B	0.73B	0.66B	0.33B	0.31B	0.27B	0.28B

(b) Statistics of synthetic evaluation data.

	Machine Translation		Open Generation		Transcreation			
	Read.	*Engage.*	*Read.*	*Engage.*	*Read.*	*Engage.*	*Creat.*	*Local.*
#Examples	10.0K	10.0K	7.8K	7.9K	3.6K	4.0K	3.8K	4.0K
#Avg Tokens	4.5K	4.8K	4.5K	4.9K	3.4K	3.5K	3.4K	3.6K
#Max Tokens	31K	30K	29K	33K	11K	13K	14K	14K
#Total Tokens	44.79M	48.20M	35.01M	38.22M	12.17M	14.14M	12.84M	14.21M

(c) Statistics of PRM-bench data.

	Machine Translation		Open Generation		Transcreation			
	Read.	*Engage.*	*Read.*	*Engage.*	*Read.*	*Engage.*	*Creat.*	*Local.*
#Examples	2.2K	2.3K	1.7K	1.9K	615	667	1.9K	1.9K
#Avg Tokens	4.8K	5.1K	4.5K	5.2K	3.5K	3.7K	3.3K	3.6K
#Max Tokens	31K	29K	29K	33K	9K	10K	10K	14K
#Total Tokens	10.63M	11.40M	7.66M	9.60M	2.15M	2.44M	6.20M	6.69M

A.4 DATASETS

We summarize the seed datasets used for generating synthetic data in Table 15, and detail them below.

A.4.1 EVALUATION DATA

We use the following datasets for synthesizing the evaluation datasets in this work:

1. **WMT24++ (Deutsch et al., 2025): Machine Translation** WMT24++ contains human-sourced translation and post-edited data from English to 55 languages. The dataset spans four domains (Literary, News, Social and Speech) and consists of short passages (32 tokens per passage on average), paired with reference translations. We sample data for the 6 language pairs of interest in this work (English to Chinese, Hindi and German - in both directions).
2. **Aesop's Fables (Niv, 2024): Story Transcreation** Aesop's Fables (Stickney & Aesop, 2015) are a collection of popular short stories credited to Aesop, a storyteller who lived in ancient Greece from 620-564 BC. We choose these stories since they are rich sources of culture-specific information, and transcreating this for different cultures, while also keeping various other personalization attributes in mind, could be widely useful. Each story averages about 220 tokens, including a title and the final moral. There are 147 stories in total in our final test set.
3. **Filtered AlpacaEval (Iyer et al., 2025): Open-Ended Generation** The original AlpacaEval dataset (Li et al., 2023b) consists of multi-domain English prompts and is a standard dataset used in the literature to evaluate instruction-following and open-ended generation capabilities of LLMs. We use the filtered version of this test set from Iyer et al. (2025), who remove English-centric queries that cannot be answered multilingually. The test set has 798 queries, with 36 tokens per prompt. We follow their approach and translate these prompts to German, Hindi and Chinese using Gemini 2.0 Flash. This final test set is then used to evaluate open-ended generation.

A.4.2 TRAINING DATA

Next, for generating training data, we use the datasets given below:

1. **XL-Instruct (Iyer et al., 2025): Machine Translation** The XL-Instruct dataset is a multi-domain supervised fine-tuning dataset consisting of high-quality question-answer pairs synthetically generated from the CulturaX corpus (Nguyen et al., 2024). Due to the unavailability of large-scale multi-domain document-level translation datasets in our languages of interest, we translate the XL-Instruct responses to German, Hindi and Chinese with Gemini 2.0 Flash and create further personalized versions using the pipeline in Section 3.1.1. The final training dataset has 45.3K English sentences with 4-way translations, and each example has about 178 tokens.
2. **Gemini Stories: Story Transcreation** We create a training dataset of culture-specific stories using Gemini 2.0 Flash. To guide the model in the prompt, we focus on six dimensions: genre, story theme, audience culture, character, location, and cultural elements to incorporate in the story. To create several stories with as much variation as possible, we first prompt the model for values that can be used for the themes, characters, and locations. Then we manually filter and select the most natural 20 themes, 30 characters, and 20 locations. We report the values we ultimately use for each of the dimensions and the story generation prompt in Table 17. In total, we generate 144,000 stories with about 334 tokens on average per story.
3. **Alpaca (Taori et al., 2023): Open-Ended Generation** The Alpaca dataset contains 52K instruction-response pairs for supervised fine-tuning in multiple domains. We source the prompts from Alpaca, and translate them to all three languages using Gemini 2.0 Flash. Each prompt has about 18 tokens on average.

Variable	Possible Values
{genre}	*Adventure, Humor, Fantasy*
{theme}	*Self-acceptance, Perseverance, Mentorship, Consequences of dishonesty, Courage, Forgiveness, Embracing change, Friendship, Empathy, Overcoming fear, Kindness, Value of community, Creativity, Cooperation, Being present in the moment, Passion, Hope, Lessons of history, Refusing greed, Trusting one's own instincts*
{culture}	*American, Chinese, German, Indian*
{character}	*A young boy, An old woman, A lost hiker, A clever pet, An honest student, A tired worker, An angry stranger, A shy girl, A brave knight, An evil wizard, A kind doctor, A busy chef, An eager explorer, A curious baby, A skeptical reporter, A quiet librarian, An impatient driver, A hopeful dreamer, An anxious parent, A loyal friend, A strong leader, An innocent child, A grumpy neighbor, A lonely musician, An aspiring artist, A curious scientist, An active athlete, A forgetful professor, A playful comedian, A determined runner*
{location}	*A coffee shop, A park bench, A grocery store, A library, A bus stop, A school classroom, A train station, An old house, A forest path, A small town diner, A beach, A car repair shop, A museum, A community pool, A theater, A busy street corner, A laundromat, A university dorm room, A zoo, A bookstore*
{cultural_elements}	*Food, Clothing, Social etiquette, Art forms, Festivals and celebrations, Monuments, Rituals and customs, National symbols*
{story_generation_prompt}	Write a {genre} short story (2-3 paragraphs at most) on {theme} for {culture} culture, making sure to integrate elements from {culture} culture where appropriate in the story. Here are some examples of cultural elements: {cultural_elements}. Do not overdo it and maintain the naturalness of the story. Pick one or two that fit best within the story and use them. The character involved is {character} and the location is {location}. Make sure the story does not stereotype. The cultural elements should be organic and reflect everyday lives of {culture} people. Also, don't resort to code-switching as a means of imitating inclusion of cultural elements. Using non-English words is discouraged. The story should be in English, and flow and end naturally.

Table 17: The story generation prompt for synthesizing the *Gemini Stories* dataset. We provide all possible values for the placeholder variables used. We sample a candidate value for all variables, and substitute this in the prompt to the teacher LLM to create different stories.

Table 18: Example Prompt for Personalized LLM-as-a-Judge. The LLM is provided unstructured user data (a shortened version is provided in the example), a query and two responses, and the task is to judge which response would be more suitable to the given user based on information present in the dump relevant to a given personalization dimension.

```
Consider the following data dump of a persona:

Review: "Pizza was cold and arrived late"
Search: "easy hiking trails near me"
Comment: "Learning German is hard... I don't understand complex words or idioms yet"

You are also given the following personalization dimension for this persona:
lang_proficiency_de, which is defined as: Deals with a persona's preferred German
language complexity in a text, based on their linguistic proficiency.

Based on the information provided in the dump relating to this dimension alone, which of these
two responses would the persona prefer as an AI model's response?

Query: How is the weather today in Munich? Respond in German.
Responses:
1. Heute ist es sonnig. Es sind 22 Grad. Kein Regen.
 (Today it is sunny. It is 22 degrees. No rain.)
2. In München herrscht heute überwiegend sonniges Wetter bei angenehmen
Temperaturen von 22 Grad, wobei keine Niederschläge zu erwarten sind.
 (In Munich today there is predominantly sunny weather with pleasant
 temperatures of 22 degrees, with no precipitation expected.)

Provide your decision as the output: 1 or 2.

Answer: 1
```

B Supplementary Results

B.1 Additional Experiments with PRMs

We now dive into two additional ablation experiments on PRMs that could not be included in the main paper due to space constraints. This includes: a) ablation of PRMs on explicit versus implicit preferences, and b) ablation of single-task versus multi-task fine-tuning of PRMs. We hope these results shed more light on this complex task and help facilitate further research.

B.1.1 PRMs on explicit versus implicit preferences

Table 19 shows an ablation comparing Gemini 2.5 Pro (the best-performing personalized LLM-as-a-judge from Table 5) with a PRM fine-tuned on 25K examples, on context containing explicit versus implicit preferences. For the former, we provide the label P that includes: the personalization dimension A, along with the corresponding rubric and the chosen score on this rubric. The latter includes the generated unstructured data D, described in Sections 3.1.2 and A.3.

We make two key observations from these results. Firstly, we note that for this task, a *bespoke fine-tuned PRM matches or outperforms Gemini 2.5 Pro* in both settings, achieving 70% on the implicit preferences case, and as high as 91% when fine-tuned on explicit preferences. These promising results highlight the relevance and efficacy of small, fine-tuned PRMs for this task, aligning with our findings in Section 4.3.2. Secondly, *personalized reward modelling on implicit preferences is substantially more challenging than on explicit preferences*, studied in previous work (Jang et al., 2024; Dong et al., 2024) — aligning with our findings with Personalized LLM-as-a-judge in Section 4.3.1.

Table 19: Personalization evaluation with *explicit* persona preferences P, as explored in prior work (Jang et al., 2024; Dong et al., 2024), versus *implicit* preferences that need to be derived from user data D. We choose 1 task, *Open-Ended Generation*, and 1 dimension, *Linguistic Proficiency*, for cross-lingual language pairs. We compare two models, Gemini 2.5 Pro (the best-performing LLM judge from Table 5) with a fine-tuned PRM trained on 25K training examples (seed=42). We observe reasoning over implicit preferences is much more challenging, for both a SOTA reasoning LLM, as well as a bespoke fine-tuned PRM.

Model	**Input =** P	**Input =** D
Gemini 2.5 Pro *(baseline)*	0.86	**0.7**
PRM-Mistral-7B-LoRA *(ours)*	**0.91**	**0.7**

B.1.2 Single-task versus Multi-task fine-tuning of PRMs

Table 20: Ablation for multi-task training of PRMs. Scores show binary classification accuracies averaged across 3 seeds for all 3 tasks: Machine Translation, Open-Ended Generation, and Story Transcreation. We compare these models in a FLOPS-constant setting, wherein the total training data used is constant, and one has to decide whether to train a series of single-task and/or single-dimension models, or train a single multi-task, multi-dimension model. The total training data comprises of 320K examples, uniformly distributed across the 8 task-dimension pairs.

Fine-Tuning Approach	**Overall**	**Machine Translation**		**Open Generation**		**Transcreation**			
		Read.	*Engage.*	*Read.*	*Engage.*	*Read.*	*Engage.*	*Creat.*	*Local.*
Single-task, single-dimension	0.51	0.43	0.53	0.51	0.50	0.65	0.55	0.73	0.39
Single-task, multi-dimension	0.55	0.58	0.57	0.56	0.56	0.54	0.50	0.57	0.49
Multi-task, multi-dimension	**0.68**	**0.66**	**0.73**	**0.61**	**0.72**	**0.71**	**0.73**	**0.72**	**0.64**

Table 20 presents ablation results comparing PRM performance across different training configurations. Single-task, single-dimension PRMs prove highly unstable, with most configurations failing to converge and achieving near-random accuracy (51%). Notable exceptions include *Creativity* (73%) and *Readability* (65%) for Transcreation, which yield strong performance, though this success varies significantly across random seeds and dimensions.

Moving towards single-task, multi-dimension training, *PRMs show only modest improvements*, reaching 55% average accuracy. Consistent gains are observed across both dimensions for the Machine Translation and Open Generation tasks. However, for the Transcreation task, which involves training on four dimensions rather than two, we observe mixed results — with performance improving on previously challenging dimensions like *Localization* (from 39% to 49%) but declining on others such as Creativity (from 73% to 57%), likely due to insufficient training data relative to the increased complexity of training on diverse personalization dimensions.

The *multi-task, multi-dimension PRM emerges as the most superior model*. Trained on 320K examples, this model achieves *68% average accuracy compared to 51% for single-task, single-dimension PRMs* – a relative improvement of 33%. Performance gains are consistent across all task-dimension pairs, with particularly strong results for the *Engagement* dimensions (73% for both Machine Translation and Transcreation). Beyond mean scores, we note that these PRMs also exhibited substantially reduced variance across seeds, ensuring reliable deployment. These findings demonstrate that multi-task PRMs trained on scaled synthetic data is crucial to helping them achieve strong results. As shown in Table 5, these PRMs outperform personalized LLM-as-a-judge, establishing this training approach as both simple and highly effective for practical applications.

C MISCELLANEOUS

C.1 LIMITATIONS

We recognize certain limitations in our approach, each reflecting deliberate trade-offs. First, our formulation requires specifying a single personalization dimension rather than multiple implicitly inferred dimensions simultaneously, a choice we made for greater interpretability and reproducibility. Second, while our predefined rubrics may involve some subjectivity, as much as possible, we mitigate this by anchoring them in established literature. For example, our cultural localization rubric is based on is on *Newmark's translation framework* (Newmark, 1988) and our readability rubic corresponds to *CEFR language proficiency* definition (Council of Europe, 2001). Third, we acknowledge potential biases in our synthetic data pipeline by using Gemini 2.0 Flash to generate the unstructured user data and evaluating the personalized reward model based on the generated data using the Gemini model family. This is commonly referred to as self-preference (Panickssery et al., 2025; Xu et al., 2024) Nevertheless, our experiments show that Mistral-based reward model and GPT-4o models perform comparably or better than Gemini 2.0 Flash models across multiple metrics, suggesting that any self-preference bias has a limited effect on our evaluation outcomes.

C.2 USE OF LARGE LANGUAGE MODELS

Large Language Models were used for writing and generating some code for this paper. For writing, they were used for editing drafts for spell-checking and making the narrative more concise, cohesive and fluent. All LLM edits were reviewed and checked to ensure factual accuracy and authenticity. For code generation, they were used to generate helper scripts to parse results into a human-readable tabular form. In general, LLMs were used as productivity-enhancing tools, the authors take the LLM usage statement very seriously and are committed to ensuring it does not compromise scientific integrity.