GRAVITY: A Framework for Personalized Text Generation via Profile-Grounded Synthetic Preferences

Priyanka Dey[♠] **Daniele Rosa**[♣] **Wenqing Zheng**[♣] **Daniel Barcklow**[♣]
Jieyu Zhao[♠] **Emilio Ferrara**[♠]
[♠]University of Southern California [♣]Capital One Research
{deyp}@usc.edu

Abstract

Personalization in LLMs often relies on costly human feedback or interaction logs, limiting scalability and neglecting deeper user attributes. To reduce the reliance on human annotations, we introduce *GRAVITY* (**G**enerative **R**esponse with **A**ligned **V**alues, **I**nterests, and **T**raits of **Y**ou), a framework for generating **synthetic, profile-grounded preference data** that captures users' interests, values, beliefs, and personality traits. By integrating demographic, cultural, and psychological frameworks—including Hofstede's cultural dimensions, Schwartz's basic values, the World Values Survey, and Big Five OCEAN traits, *GRAVITY* synthesizes preference pairs to guide personalized content generation. We evaluate *GRAVITY* on book descriptions for 400 Amazon users, comparing it to prompt-based conditioning, standard fine-tuning, and naive synthetic pair generation. Profile-grounded synthetic data consistently improves generation, especially across multiple cultures (USA, Brazil, Japan, India), achieving over 4% higher preference gains across baselines, with user studies showing that *GRAVITY* outputs are preferred over 86% of the time. Our results show that scenario-grounded synthetic data can capture richer user variation, reduce reliance on costly annotation, and produce more engaging, user-centered content, offering a scalable path for LLM personalization.[1]

1 Introduction

Personalization has become a critical frontier for LLMs (Jang et al., 2023; Chen et al., 2024; Zhang et al., 2024b). While recent advances enable fluent and contextually relevant text generation, outputs often remain generic, overlooking individual differences in taste, style, and preference (Zhang et al., 2025; Moorjani et al., 2022). In domains such as book recommendations, this gap is particularly visible—two users may value entirely different aspects of the same content, with one drawn to narrative structure and cultural context and the other to character development and personal resonance. Generic descriptions risk alienating users by failing to capture what truly matters (Yunusov et al., 2024; Cai et al., 2023).

Prior work in personalization often relies on human-annotated preference data, whether through reinforcement learning from human feedback (RLHF) (Kirk et al., 2024; Poddar et al., 2024), preference modeling (Lee et al., 2024a; Zhong et al., 2024; Zheng et al., 2025), or profile-conditioned prompting (Zhang et al., 2024a; Lyu et al., 2023). While effective, large-scale annotation is costly and difficult to scale, limiting the breadth of user attributes that can be incorporated. Moreover, existing approaches frequently reduce personalization to narrow signals such as demographics or explicit traits, overlooking deeper dimensions of user engagement.

To combat this challenge, we introduce *GRAVITY* (**G**enerative **R**esponse with **A**ligned **V**alues, **I**nterests, and **T**raits of **Y**ou), a framework for creating **synthetic, profile-grounded preference data** that can be used to fine-tune LLMs for more effective personalization. As a case study, we focus on personalizing book descriptions using users from the Amazon Book Reviews dataset (Hou et al., 2024). Instead of collecting explicit preference annotations, we construct synthetic preference pairs grounded in well-established psychological and cultural frameworks—including Hofstede's cultural dimensions (Hofstede, 1983), Schwartz's theory of basic values (Schwartz, 2012), the World Values Survey (Haerpfer et al., 2024), and the Big Five OCEAN traits (Goldberg, 2013). This allows us to capture variation in user interests, values, beliefs, and personality traits, providing a richer basis for personalization than demographics alone. We then fine-tune `Llama-3.1-8B-Instruct` with Direct Preference Optimization (DPO) (Rafailov

[1]Code is available: https://github.com/limenlp/GRAVITY.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7416–7436

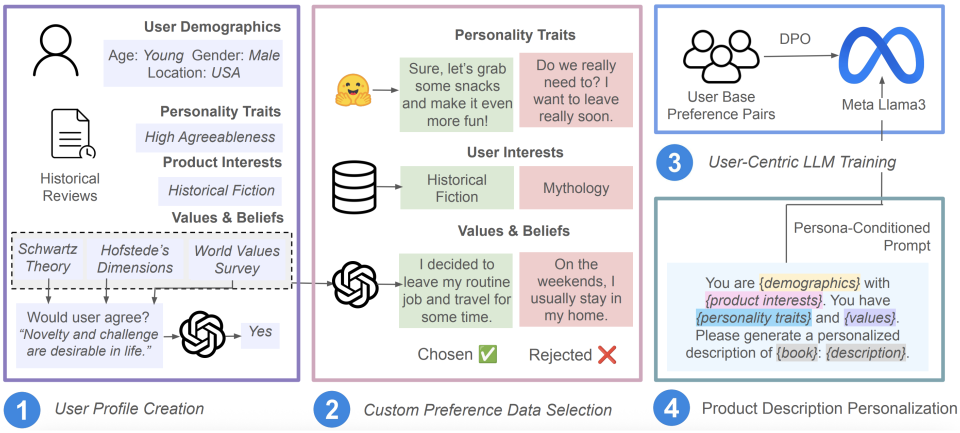


Figure 1: Our approach *GRAVITY* consists of four stages: (1) ***User Profile Creation***, (2) ***Custom Preference Data Selection***, (3) ***User-Centric LLM Training***, and (4) ***Product Description Personalization***. In stage 1, we extract information about the user including *explicit* values such as demographic attributes: age, location, and gender and users' interests and *implicit* values including their personality traits and their values and beliefs (extracted based on seed statements generated from various psychological and cultural frameworks). In stage 2, During stage 1, we generate a candidate pool of scenarios for values and beliefs based on Stage 1 using GPT-4o. We then generate a custom set of chosen/reject preference pairs for users spanning three facets: user interests, personality, and values using a combination of the Amazon Reviews dataset, personality SJTs (*TRAIT* and *Big5Chat*), and the candidate pool of generated scenarios. In stage 3, we preference tune *Llama* with the users' preference pairs, and finally, in Stage 4, we generate the personalized description by prompting this tuned model with user profile attributes.

et al., 2023), aligning generations with these profile-derived preferences.

We evaluate *GRAVITY* against several baselines, including prompt-based conditioning, standard supervised fine-tuning, and a naive DPO approach using synthetic pairs without structured profiles. Our results show that profile-grounded synthetic data consistently improves generation, achieving over 4% higher preference scores across baselines. User studies further show that *GRAVITY* generations are preferred over 86% of the time.

While our results do not imply that synthetic data can replace human feedback, they suggest that carefully designed, structured, scenario-based synthetic pipelines can reduce annotation needs while capturing various user attributes, offering a scalable path for aligning LLM-generated content with what users actually find engaging. Our contributions are summarized:

- ***GRAVITY*:** We introduce *GRAVITY*, a multi-step framework for personalized content generation using **synthetic, profile-grounded user data**, capturing values, interests, and personality traits while reducing reliance on costly human annotation.
- **Profile Modeling:** We develop a pipeline integrating psychological and cultural frameworks (Hofstede, Schwartz, WVS, OCEAN) to generate customized preference data that reflects diverse user profiles.
- **Comprehensive Evaluation:** We assess our method using both automatic metrics and a user study, measuring preference alignment and user-perceived engagingness of the generated content.
- **Data Efficiency:** We show that profile-grounded synthetic preference data yields measurable improvements over prompt-based, SFT, and naive DPO baselines.

2 Background

Personalization is central to making generated content engaging and relevant: attributes such as user interests, style preferences (especially through personality) (Nguyen et al., 2018; Dey et al., 2025), or cultural background (Matz et al., 2024; Joshi et al., 2025; Liu et al., 2025) can strongly shape whether stories and characters resonate with readers (Woźniak et al., 2024; Yang et al., 2023). Recent work leverages LLMs for personalization across both direct text generation and downstream applications like recommendation (Jiang et al., 2023; Jang et al., 2023). Broadly, methods fall into two streams: (a) personalized generation, where models adapt outputs via prompting, user profiles, or fine-tuning

(Peng et al., 2024; Li et al., 2024b); and (b) LLMs for personalized tasks, such as recommendation, reasoning over user histories, or tailoring explanations (Bismay et al., 2024; Shao et al., 2024). Hybrid approaches increasingly combine these, generating user-specific content that simultaneously improves alignment and task performance (Shenfeld et al., 2025; Zhong et al., 2024).

Within generation, personalization extends beyond static profile conditioning to new task formulations. Work on summarization, for example, uses multi-agent pipelines to refine drafts according to user preferences (Xiao et al., 2023), while headline generation has incorporated user-history modeling (Song et al., 2023). Similarly, recommender systems use LLMs to expand sparse item text or construct structured user/item profiles from histories, enabling personalized reviews and explanations (Lyu et al., 2023; Acharya et al., 2023).

Finally, advances in alignment and evaluation point toward more fine-grained personalization. Extensions of RLHF introduce per-user reward components or factorized preference functions (Li et al., 2024b; Shenfeld et al., 2025; Poddar et al., 2024), while controllable preference vectors allow multi-objective adaptation (Zhong et al., 2024). Benchmarks like LaMP (Salemi et al., 2023) and newer evaluations on role-playing, user modeling, and cultural adaptation (Tseng et al., 2024; Zhang et al., 2024b) reflect this growing focus on personalization as a key axis of LLM evaluation.

3 *GRAVITY*

In this section, we present our framework *GRAVITY* for generating personalized content. We first describe product description personalization (§3.1), then our user profile pipeline (§3.2) and synthetic preference data generation (§3.3). Finally, we outline user-centric LLM training (§3.4) and personalized description generation (§3.5). For evaluation, we focus on generating personalized book descriptions for 400 Amazon readers from diverse cultural backgrounds; additional details on the users and dataset are provided in §A.

3.1 Task Formulation

We formulate personalization of product descriptions as a conditional generation problem grounded in structured user profiles. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ denote the set of users, and $\mathcal{B} = \{b_1, b_2, \ldots, b_{|\mathcal{B}|}\}$ denote the set of products. Each user $u \in \mathcal{U}$ has a historical review sequence

$$R_u = (r_{u1}, r_{u2}, \ldots, r_{u|R_u|}),$$

where r_{ui} is the textual review for product $b_i \in \mathcal{B}$.

From these reviews and basic demographic information, we curate a structured user profile P_u, which includes four components: (1) Demographics (e.g., age, gender, location), (2) Product interests, (3) Intrinsic values and core beliefs, and (4) Personality traits.

Each product $b_n \in \mathcal{B}$ is associated with a generic, non-personalized description d_n. Given a target user u and product b_n, the goal is to generate a personalized description $\hat{B}_n$ that better aligns with the user's profile P_u. Formally, this can be written:

$$\hat{B}_n \sim P(\mathcal{D} \mid P_u, d_n; \theta),$$

where $\mathcal{D}$ denotes the space of possible product descriptions and θ represents the model parameters. The personalized generation model can be represented as:

$$\hat{B}_n = F_\theta(P_u, d_n),$$

where F_θ is the conditional generation model that incorporates both structured user information and the generic product description. The output $\hat{B}_n$ is a personalized description that reflects the user's preferences and traits while preserving the content of the original product description.

3.2 Creating User Profiles

Since personalization relies on a nuanced understanding of the user, the first step in *GRAVITY* is to extract salient user information and construct a structured profile. As discussed in §2, attributes such as demographics, interests, values, and personality traits can strongly influence how engaged and connected a user feels with generated content.

Each user profile in our framework consists of four key components: (1) Demographics (gender, age, location), (2) Interests, (3) Values and Beliefs, (4) Personality Traits. Demographic information is obtained directly from user-provided data when available. When gender or age is missing, we follow prior work that has shown strong performance in demographic estimation and apply trained DeBERTa models (Tureyyen, 2023a,b). To reduce noise and sparsity, we bin ages into three groups: young (<30), middle-aged (31–60), and senior (>60).

Source	Seed Statement	Total Count
Hofstede's Cultural Dimensions (Hofstede, 1983)	Planning for the future and working toward long-term goals is more important than immediate rewards.	10
Schwartz' Theory of Basic Values (Schwartz, 2012)	It is important to me to make my own choices and decisions and think independently, even if others may disagree.	10
World Values Survey (Haerpfer et al., 2024)	I believe hard work doesn't generally bring success—it's more a matter of luck and connections.	130

Table 1: Example seed statements generated from various cultural and psychological constructs.

Interests capture a user's content preferences, i.e., all product categories the user might be interested in or frequently engage with. These interests can be extracted based on users' product purchase and review history, combined with their explicitly selected preferences. Specifically, for our case study on Amazon readers, we extract users' most frequent book genres that account for 10% or more of their reviews[2].

Values and Beliefs capture a user's implied norms and ideals across domains such as culture, religion, ethics, politics, and society. To extract these, we adopt a multi-step LLM-based approach (see Figure 1). We first construct 150 *seed value statements*, drawing from established cultural theories (Hofstede's cultural dimensions, Schwartz's theory of basic values) and large-scale value surveys (World Values Survey). Table 1 shows representative examples. We then prompt an LLM (GPT-4o) with each seed statement and the user's full set of product reviews, asking it to infer whether the user likely *supports*, *opposes*, or is *neutral* toward the statement. This yields a structured profile of the user's values and beliefs.

Personality Traits capture a user's underlying preferences and tendencies, shaping both their writing style and the kinds of story elements they are likely to connect most with. To infer these traits, we use `PersonalityLM` (Wang and Sun, 2024), a fine-tuned RoBERTa classifier to predict high/low levels for each of the Big Five (OCEAN) traits. In our study, we estimate user personalities using each user's historical book reviews.

3.3 Generating Custom Preferences

Based on users' generated profiles, the next step in *GRAVITY* is to construct custom preference data that guides model adaptation toward personalized product descriptions. Rather than relying on generic preference statements, we generate scenario-based data that captures how a user's interests, values and beliefs, and personality traits shape their judgments in concrete contexts. While the underlying pool of scenarios remains consistent across users, each profile produces a unique mapping of *chosen* and *rejected* labels, ensuring that supervision reflects their individual preferences. This scenario-driven approach provides richer, more discriminative signals than abstract descriptions, enabling the model to learn finer-grained distinctions in how different users engage with content (Singh et al., 2025; Huang et al., 2023).

Interests We construct two types of interest-based preference data: (1) *Category*, which contrasts broad product categories (e.g., romance vs. biographies), and (2) *Summary*, which contrasts detailed book descriptions. For our study of personalized book descriptions, we begin with 382 categories from the Amazon Reviews dataset and select 3 representative, highly rated books per category. For each user, we identify their top 3–5 genres (§3.2) and, using SentenceBERT embeddings with cosine distance, select the 3 most distinct categories. We then generate *Category* preference pairs (chosen vs. rejected) for the user's preferred genres. For *Summary* data, we form preference pairs from book descriptions of the representative titles in the user's top and most distinct genres.

Values and Beliefs To generate preference data about user's values and beliefs, we first synthetically generate scenarios using GPT-4o for each seed value statement (§3.2). For each seed, GPT-4o produces 3 unique pairs of scenarios, where the first scenario strongly aligns with the statement and the second contradicts it, yielding 450 scenario pairs in total. To generate user-specific preference labels, we utilize each user profile's extracted set of values. For a given seed statement, if the user supports it, all generated aligned scenarios are marked *chosen*; if the user opposes it, aligned scenarios are marked *rejected*[3].

[2]This threshold ensures that the selected genres reflect substantial and consistent engagement rather than incidental activity.

[3]Scenarios from *neutral* seed statements are omitted from

Personalization Method	Top-1 WinRate (%)	Preference Gain (%)	Interestingness Score	Text Similarity
Original	0.75	-	3.74	-
BaseRewrite	1.25	63.5	3.81	72.8
DemoBased	3.25	68.5	3.69	72.4
UserSummary	9.25	72.0	3.98	71.0
LaMP	9.75	75.5	4.06	72.8
TriAgent	7.0	73.25	3.95	70.37
PrefEx	4.0	63.75	3.80	75.88
ContrastEx	4.25	62.5	3.78	72.8
UserSFT	16.25	74.25	3.91	79.92
PrefAlign	19.5	79.0	4.02	79.48
GRAVITY (Ours)	24.75	82.5	4.03	73.2

Table 2: We compare several baselines, including existing personalization methods using prompt-based techniques and multi-agents, to our method *GRAVITY*. We report aggregated top-1 win rates, preference gains, interestingness scores, and text similarity scores over all selected users. Win rates, preference gains, and interestingness scores are measured with an LLM-judge (GPT-4o). We find that our method *GRAVITY* achieves the highest personalization metrics (↑ WinRate and Preference Gain and comparable Interestingness scores) and comparable text similarity scores to other baseline approaches. The original book description has a win rate below 1%, showing personalization methods generate more engaging descriptions.

Personality Traits We leverage two situation-based personality questionnaires, TRAIT (scenarios) (Lee et al., 2024b) and Big5Chat (dialogues) (Li et al., 2024a), each containing questions with binary-level answers for the *OCEAN* traits. For each user, we randomly select 150 question–answer pairs from each dataset. For each pair, the answer consistent with the user's OCEAN trait profile is labeled *chosen*, and the alternative labeled *rejected*.

3.4 User-Centric LLM Training

For our task of personalized book descriptions, we generate approximately 1,000 customized preference pairs across all three dimensions for each user (400K for user base). Using this data, we utilize Direct Preference Optimization (DPO) (Rafailov et al., 2023) to generate a single personalized LLM which learns the behaviors and styles of the users in our dataset. Each training instance conditions the model on a user's demographic, personality, values, and interests, enabling generalization across diverse profiles. At inference, we provide the target user's profile to generate personalized book descriptions. For evaluation, we sample a highly rated title from the user's preferred genres, excluding books they have already reviewed. Additional training details are in §4.1.

3.5 Product Description Personalization

To generate personalized descriptions, we follow prior work (Santurkar et al., 2023; Hwang et al., 2023) by conditioning model responses on the user's profile. Each profile, as described in §3.2, includes demographics, genre interests, personality traits, and personal values. Full prompt templates for description generation are provided in §C. For our task, we select a target book description to personalize from a randomly chosen highly rated title within the user's top genres, ensuring the user has not previously purchased or reviewed the book.

4 Experiments

In this section, we outline our experimental design for personalized text generation (§4.1), including the training pipeline for user-centric LLMs, the baselines used for comparison, and evaluation strategies. We then present results across all methods (§4.2) and conduct ablations to isolate the contribution of each preference dataset to personalization performance (§4.3). Finally, we compare two training strategies for incorporating user preferences—standard supervised fine-tuning (SFT) and Direct Preference Optimization (DPO)—to assess their relative effectiveness (§4.4).

4.1 Experiment Setup

Finetuning user-centric LLMs We adapt `Llama-3.1-8B-Instruct` (Dubey et al., 2024) to user preferences using DPO with LoRA adapters (Hu et al., 2022). We preference tune the model using the `TRL` library (CarperAI, 2023) with the AdamW optimizer (Loshchilov and Hutter, 2017), a β value of 0.3, a learning rate of 2×10^{-5}, weight decay of 0.01, a cosine learning rate schedule, and a warmup ratio of 0.05. Additional training details and prompt formats are detailed in §D.1.

Personalization strategies We benchmark against

the user's preference dataset.

Persona	*Country*: India *Age*: Young *Gender*: Female Interests : Young Adult, Romance, Fiction Values and Beliefs : Friends are an important part of life.; Excitement, novelty, and change are desirable in life. Personality Traits : High Openness, High Extroversion
Original Description	*Our Chemical Hearts* follows Henry Page, a teenager who's never fallen in love —until Grace Town joins his class. Paired to work on the school newspaper, he discovers the excitement and challenges of first love , and the ways it can change a person.
LaMP	*Our Chemical Hearts* follows Henry Page, a teenager who has always thought of himself as a hopeless romantic —but has never truly fallen in love . Focused on school and friendships, his life takes an unexpected turn when Grace Town joins his class. As they are paired to work on the school newspaper, Henry discovers the excitement, confusion, and intensity of first love —something any young reader navigating the ups and downs of adolescence can relate to.
PrefAlign	*Our Chemical Hearts* tells the story of Henry Page, a teenager who considers himself a hopeless romantic but has never fallen in love . Focused on school and his future, his world changes when Grace Town joins his class. She isn't exactly the girl he expected, but as they are paired to work together on the school newspaper, Henry begins to experience the joys and challenges of first love , discovering how unpredictable and powerful emotions can be.
GRAVITY	*Our Chemical Hearts* follows Henry Page, a teenager who has always thought of himself as a hopeless romantic —but has never truly fallen in love . Focused on school and his friends , his life takes an unexpected turn when the intriguing and unpredictable Grace Town joins his class . As they are paired to work on the school newspaper, Henry experiences the thrill, surprises, and intensity of first love , along with the joys and challenges of navigating friendships and new experiences — perfect for any young reader who loves excitement, connection .

Table 3: Example personalized book descriptions for a simplified persona for select baselines. Our method better captures user-relevant themes compared to baselines. Highlighted texts show relevance to user's Interests , Values and Beliefs and Personality Traits . We recommend viewing this table in full color.

nine baselines[4], differing in how user information is integrated: (1) **Prompting**: (a) *BaseRewrite* – rewrites the description to be more engaging (no personalization), (b) *DemoBased* – appends user demographics (age, gender, location), (c) *UserSummary* – prepends a GPT-4o–generated user profile summary, (d) *LaMP* (Salemi et al., 2023) – retrieves and prepends the user's most relevant past reviews, (e) *TriAgent* (Xiao et al., 2023) – a three-stage pipeline combining *BaseRewrite*, GPT-4o summaries, and customized edit instructions, (f) *PrefEx* – appends two (book description → review) pairs as context, (g) *ContrastEx* – appends one positive book description → review pair and one negative book description → review pair as context; (2) **SFT**: (a) *UserSFT* — fine-tunes on (book description → user review) pairs with persona prompts encoding user demographics; (3) **Preference data**: (a) *PrefAlign* — uses DPO using GPT-4o-generated aligned and misaligned book descriptions, produced solely from user demographics. Details on baseline prompts and implementations can be found in §D.2.

Auto-evaluation metrics To assess the quality of personalized descriptions, we utilize three

[4]All methods use `Llama-3.1-8B-Instruct` unless noted.

Method	Top-1 Win (%)	Pref. Gain (%)	Int. Score
Original	7.78	–	3.48
TriAgent	30.92	78.87	4.09
GRAVITY	61.30	86.73	4.15

Table 4: User Study personalization metrics (*Top-1 Win Rate*, *Preference Gain*, and *Interestingness Score*) based on real user rankings of book descriptions: original, TriAgent, and *GRAVITY* descriptions, aggregated over all users. Consistent with auto-evaluation metrics, we observe users prefer *GRAVITY* generations over baseline generations and the original description.

personalization-measuring metrics: (1) *Top-1 Win-Rate*: how often is a method's output preferred over all other methods, (2) *Preference Gain*: how often is a method's output preferred over the original description, and (3) *Interestingness Score*: how engaging/captivating does the user find the method's output (based on 5-point Likert score). We evaluate personalizations using LLM-as-a-judge (Zheng et al., 2023). We provide GPT-4o with a detailed user persona, generated from the user's complete set of written reviews and available demographics. To ensure personalized content remains faithful to the original description, we also compute cosine similarity between SentenceBERT (Reimers and

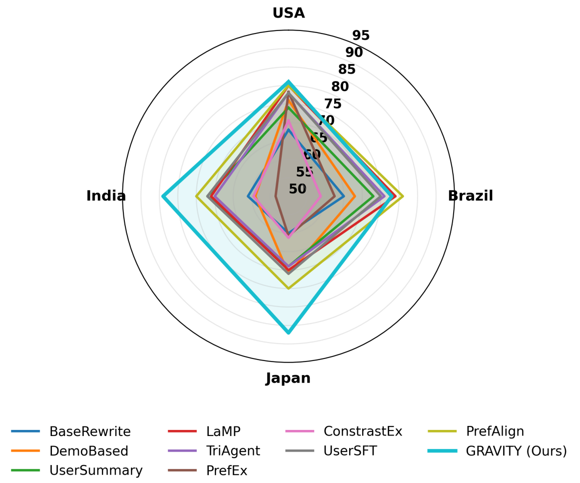


Figure 2: Preference Gains (%) for personalized generation methods across users from four diverse countries: USA, Brazil, Japan, and India. We observe that *GRAVITY* yields consistently strong personalization metrics across the cultures, with gains in non-Western countries compared to current approaches (>10% increase in preference gains).

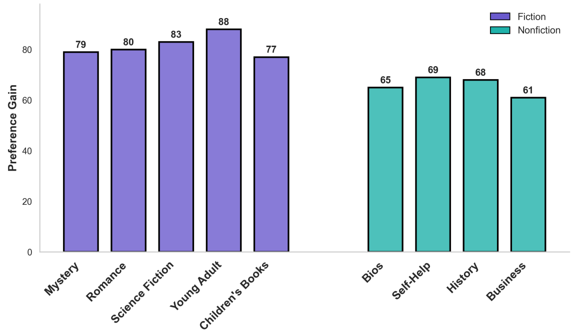


Figure 3: Preference gains (%) across nine book genres (similar categories, clustered into fiction and nonfiction) using *GRAVITY*. We find that most fiction books have much higher preference gains (> 70%) as compared to non-fiction books. (≈ 65%).

Gurevych, 2019) embeddings of the original and generated texts. Prompt templates used for evaluation are included in §D.3.

User study We conduct a small study with 120 participants from diverse cultural backgrounds (30 users from each of USA, Brazil, Japan, and India). For each participant, we collect demographics, top book genres, values, and personality traits. We then train a customized model for these users and generate personalized book descriptions for each participant based on popular books in their preferred genres. Participants rank three descriptions of 10 unique books (original, *TriAgent*, and *GRAVITY*) based on how engaging and interesting they find each book description. Further details on our user study design are included in §D.4.

4.2 Personalization with *GRAVITY*

Comparisons against baselines Table 2 reports automatic evaluation metrics for all baselines and *GRAVITY*. Our method achieves the highest win rates (5.25% improvement) and preference gains (3.5% improvement) compared to the strongest baselines. Interestingness scores are similarly high across baselines and *GRAVITY*, with average ratings above 4, indicating that generated descriptions are engaging to users. Furthermore, more than 70% of the content remains semantically consistent with the original description—comparable to other baselines—demonstrating that our method preserves the original content while personalizing it.

User evaluations Table 4 show *GRAVITY* leads to generalizable performance gains across users from various countries and book categories. Compared to the best baseline, *LaMP*, we observe that users tend to prefer book descriptions generated via *GRAVITY* more than 61% of the time. Moreover, users also tend to prefer *GRAVITY* generations approximately 8% more than original book descriptions. Users also tend to find descriptions generated through *GRAVITY* as the most interesting.

Cross-Cultural Personalization To evaluate robustness across diverse populations, we report preference gains for users from the USA, Brazil, Japan, and India (Figure 2). Results are averaged across each user group for the six baselines and our approach, *GRAVITY*. We find that *GRAVITY* achieves consistent improvements in all regions, with higher preferences in almost all countries. Particularly, we observe large yields (>10%) increase) in non-Western contexts such as Japan. Although Brazilians may prefer *PrefAlign* generations approximately 3% more than *GRAVITY*, we still observe comparable performances across personalization metrics (see Table 15) with *GRAVITY*. These results indicate that tailoring preference data to capture users' cultural value systems and personality traits may substantially improve alignment with user's engagement and preferences.

Personalization Across Genres We further analyze how personalization effectiveness varies across book genres. While our method consistently improves over baselines across both fiction and nonfiction, we observe substantially higher gains

Feature	Preference Gain (%)
All (original *GRAVITY*)	86.73
- User Interests	-10.34^{**}
- Category	-6.23^{**}
- Summary	-4.38^{*}
- Values & Beliefs	-8.92^{**}
- Personality	-8.48^{**}
- Openness	-5.45^{**}
- Conscientiousness	-2.35^{*}
- Extraversion	-6.03^{**}
- Agreeableness	-1.32
- Neuroticism	-0.98

Table 5: Ablation studies showing the impact of different user preference data (interests, values, and personality traits) on personalization performance, measured via decreases in preference gains. Subsets of Values & Beliefs data are not removed due to small size (2K–3K pairs per user). Statistical significance is assessed using paired Wilcoxon signed-rank tests: * indicates $p < 0.05$, and ** indicates $p < 0.01$.

for fiction titles (see Figure 3[5]); we see preference gains above 75% for fiction books but gains closer to 65% for non-fiction books. Our findings align with prior research which show that fiction readers tend to be more engaged with narrative elements—such as character development, emotional arcs, and imaginative settings—which naturally connect with readers' personal values, beliefs, and personality traits (Bal and Veltkamp, 2013; Goyal and Mahmoud, 2024). Thus, incorporating these data as training signals into user-centric models allows for better personalization. In contrast, non-fiction reading is primarily fact-oriented, emphasizing clarity, accuracy, and logical structure, which limits the potential for personalization, resulting in smaller improvements through *GRAVITY*.

4.3 Effects of Custom Preference Data

To understand the contribution of different components of user preference data to personalization performance, we perform an ablation study. Specifically, we remove each data subset: Interests, Values and Beliefs, and Personality Traits, and measure the resulting change in preference gains compared to the full model (Table 5) (using GPT-4o judge, see §D.3 for prompt details).

We observe that removing any preference dataset component leads to statistically significant drops

[5]Although the Amazon Reviews dataset has several fine-grained labels for book genres, in this analysis, we focus on a set of nine hierarchical book categories/themes.

Method	Top-1 Win (%)	Pref. Gain (%)	Int. Score
BaseRewrite	1.0	63.5	3.81
DemoBased	3.0	68.5	3.69
UserSummary	4.5	72.0	3.98
LaMP	4.75	75.5	4.16
TriAgent	4.25	73.25	3.95
UserSFT	17.0	74.25	4.04
PrefAlign	20.75	79.0	4.24
GRAVITY (SFT)	20.25	72.58	3.98
GRAVITY (DPO)	24.5	82.67	4.10

Table 6: Personalization metrics with SFT on *GRAVITY* data. SFT improves engagement (comparable to *PrefAlign* and *UserSFT* method results), while DPO achieves the strongest gains in win rate, preference, and interestingness.

in preference gains, highlighting the importance of each type of user information. Notably, apart from Interests, Values and Beliefs have the largest impact, with decreases of almost 9%, while individual Personality Traits such as *Openness* and *Extraversion* also contribute substantially. Some traits, like *Agreeableness* and *Neuroticism*, have a smaller effect, suggesting that their influence on personalization may be more subtle.

4.4 SFT vs. DPO

While DPO enables models to directly learn from user preferences and stylistic nuances, supervised finetuning (SFT) offers a simpler, more stable alternative that is less sensitive to hyperparameter choices and requires less training resources. To study this trade-off, we evaluate a setting where models are taught about a user's preferred book genres, values, and personality traits through SFT. For each user, we construct binary training examples of the form: given demographic attributes and a scenario, the user would most likely prefer $\underline{A}$, where A corresponds to the chosen response from each preference pair.

As in our original approach, we train `Llama-3.1-8B-Instruct`, using LoRA for 5 epochs using a learning rate of 10^{-4} with GPT-4o as our evaluation judge. Table 6 reports results across baselines and *GRAVITY*: SFT and DPO.

Overall, our findings suggest that incorporating user-specific data (interests, values, and personality) enhances personalization across methods. However, DPO more effectively captures fine-grained preferences, leading to the highest improvements in alignment and engagement for win rate, preference gain, and user interestingness.

5 Conclusion

In this work, we presented *GRAVITY*, a framework for generating personalized text by leveraging synthetic, profile-grounded preference data derived from demographics, interests, values, beliefs, and personality traits. By integrating psychological and cultural frameworks, our approach enables LLMs to align content with diverse user profiles and capture meaningful variation in individual preferences.

Applied to personalized Amazon book descriptions, *GRAVITY* demonstrates that scenario-driven preference data can effectively guide model behavior, with deeper attributes such as values and personality providing measurable gains beyond demographics and interests. These results highlight the potential of structured synthetic preference data as a scalable and interpretable approach to user-centric LLM personalization.

Future Work. Future work can extend this framework along several important dimensions. First, while our current analysis treats country-level culture as a practical unit of abstraction, culture is inherently multi-layered. Extending *GRAVITY* to model *within-country cultural variation*, such as regional, linguistic, or diasporic subcultures, would enable more fine-grained personalization without relying on coarse national aggregates. Second, our current formulation considers individual profile dimensions largely in isolation during synthetic preference generation. An important direction is to explore intersectional alignment, where preferences emerge from the interaction of multiple user attributes (e.g., culture, personality traits, values, and situational context) rather than from a single dominant profile type. Modeling such intersections could better reflect real-world user heterogeneity and reduce the risk of oversimplified or stereotypical personalization. Finally, future work could integrate adaptive or user-driven signals to dynamically update preference distributions over time, allowing personalization to evolve alongside users rather than remain fixed to static priors.

6 Limitations

In this work, we approximate user attributes in order to build a synthetic preference dataset. While this approach reduces the need for costly annotation, it may not guarantee exact alignment with an individual's intrinsic value systems, cultural beliefs, or personality traits. Prior work has shown that such attributes can be estimated through model prompting and training, but we acknowledge that profile inference introduces uncertainty and potential noise. To partly mitigate this, we conducted a user study (see §D.4) to validate whether personalized generations align with the participants' self-reported preferences and profile attributes. We also acknowledge that not all demographic information may be available during training. In this case, the output may be less personalized to the user as our generation prompt requires various attributes. However, this can be modified and adjusted based on available data.

Another limitation lies in cross-cultural generalization. Although we ground *GRAVITY* in established cultural psychology frameworks including Hofstede's dimensions, Schwartz's values, and the World Values Survey, these instruments may not fully capture the richness and evolving nature of cultural identity. Moreover, reliance on population-level psychological frameworks risks reinforcing cultural stereotypes if treated as deterministic or exhaustive representations of individuals. While our approach uses these frameworks as *soft priors* to guide synthetic preference generation rather than as fixed labels, personalization may still miss subtle, context-dependent, or counter-stereotypical user preferences that fall outside of these abstractions.

Finally, this work focuses on book description personalization using Amazon reviews as a controlled testbed for exploring the effectiveness of *GRAVITY*. The main goal of this work is to demonstrate the feasibility of using synthetic, profile-grounded preference data for personalization, but future work can test its applicability across broader datasets and domains, such as news, healthcare, and education, where personalization challenges may differ.

Acknowledgments

We would like to thank researchers from USC LIME and HUMANS labs for their continuous feedback in the formulation and setting of our work. This research was supported in part by the NSF, under Award Number 2331722. Priyanka Dey was also supported by the 2025 USC Capital One Fellowship.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1204–1207.

P Matthijs Bal and Martijn Veltkamp. 2013. How does fiction reading influence empathy? an experimental investigation on the role of emotional transportation. *PloS one*, 8(1):e55341.

Millennium Bismay, Xiangjue Dong, and James Caverlee. 2024. Reasoningrec: Bridging personalized recommendations and human-interpretable explanations through llm reasoning. *arXiv preprint arXiv:2410.23180*.

Pengshan Cai, Kaiqiang Song, Sangwoo Cho, Hongwei Wang, Xiaoyang Wang, Hong Yu, Fei Liu, and Dong Yu. 2023. Generating user-engaging news headlines. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3265–3280, Toronto, Canada. Association for Computational Linguistics.

CarperAI. 2023. Trl: Transformer reinforcement learning.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Priyanka Dey, Yugal Khanter, Aayush Bothra, Jieyu Zhao, and Emilio Ferrara. 2025. Can llms express personality across cultures? introducing culturalpersonas for evaluating trait alignment. *arXiv preprint arXiv:2506.05670*.

M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. 2006. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazare, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Lewis R Goldberg. 2013. An alternative "description of personality": The big-five factor structure. In *Personality and personality disorders*, pages 34–47. Routledge.

Mandeep Goyal and Qusay H Mahmoud. 2024. A systematic review of synthetic data generation techniques using generative ai. *Electronics*, 13(17):3509.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2024. World values survey wave 7 (2017-2022) cross-national data-set.

Geert Hofstede. 1983. National cultures in four dimensions: A research-based theory of cultural differences among nations. *International studies of management & organization*, 13(1-2):46–74.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shijia Huang, Jianqiao Zhao, Yanyang Li, and Liwei Wang. 2023. Learning preference model for llms via automatic preference data generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9187–9199.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. 2025. Improving llm personas via rationalization with psychological scaffolds. *arXiv e-prints*, pages arXiv–2504.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024a. Aligning to thousands of preferences via system message generalization. *Advances* *in Neural Information Processing Systems*, 37:73783–73829.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2024b. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2024a. Big5-chat: Shaping llm personalities through training on human-grounded data. *arXiv preprint arXiv:2410.16491*.

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

Ziyi Liu, Priyanka Dey, Zhenyu Zhao, Jen-tse Huang, Rahul Gupta, Yang Liu, and Jieyu Zhao. 2025. Can llms grasp implicit cultural values? benchmarking llms' metacognitive cultural intelligence with cq-bench. *arXiv preprint arXiv:2504.01127*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. 2023. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*.

Sandra C Matz, Jacob D Teeny, Sumer S Vaid, Heinrich Peters, Gabriella M Harari, and Moran Cerf. 2024. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692.

Samraj Moorjani, Adit Krishnan, Hari Sundaram, Ewa Maslowska, and Aravind Sankar. 2022. Audience-centric natural language generation via style infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1919–1932, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tien T Nguyen, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2018. User personality and user satisfaction with recommender systems. *Information systems frontiers*, 20(6):1173–1189.

Qiyao Peng, Hongtao Liu, Hongyan Xu, Qing Yang, Minglai Shao, and Wenjun Wang. 2024. Review-llm: Harnessing large language models for personalized review generation. *arXiv preprint arXiv:2407.07487*.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544.

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. *Advances in neural information processing systems*, 36:53728–53741.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

Minglai Shao, Hua Huang, Qiyao Peng, and Hongtao Liu. 2024. Ulmrec: User-centric large language model for sequential recommendation. *arXiv preprint arXiv:2412.05543*.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. 2025. Language model personalization via reward factorization. *arXiv preprint arXiv:2503.06358*.

Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. 2025. Fspo: Few-shot preference optimization of synthetic preference data in llms elicits effective personalization to real users. *Preprint*, arXiv:2502.19312.

Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. 2023. General then personal: Decoupling and pre-training for personalized headline generation. *Transactions of the Association for Computational Linguistics*, 11:1588–1607.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Kaan Tureyyen. 2023a. kaantureyyen/deberta-blog-authorship-corpus-age: Debertav3 (small) fine-tuned on the blog authorship corpus for age classification. `https://huggingface.co/kaantureyyen/deberta-blog-authorship-corpus-age`.

Kaan Tureyyen. 2023b. kaantureyyen/deberta-blog-authorship-corpus-gender: Debertav3 (small) fine-tuned on the blog authorship corpus for gender classification. https://huggingface.co/kaantureyyen/deberta-blog-authorship-corpus-gender.

Rong Wang and Kun Sun. 2024. Continuous output personality detection models via mixed strategy training. *arXiv preprint arXiv:2406.16223*.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520. IEEE.

Wen Xiao, Yujia Xie, Giuseppe Carenini, and Pengcheng He. 2023. Personalized abstractive summarization by tri-agent generation pipeline. *arXiv preprint arXiv:2305.02483*.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*.

Sarfaroz Yunusov, Hamza Sidat, and Ali Emami. 2024. Mirrorstories: Reflecting diversity through personalized narrative generation with large language models. *Preprint*, arXiv:2409.13935.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Yuansen Zhang, Xiao Wang, Tianze Chen, Jiayi Fu, Tao Gui, and Qi Zhang. 2024a. P4: Plug-and-play discrete prompting for large language models personalization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9129–9144.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024b. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wenqing Zheng, Noah Fatsi, Daniel Barcklow, Dmitri Kalaev, Steven Yao, Owen Reinert, C. Bayan Bruss, and Daniele Rosa. 2025. Tuning-free llm can build a strong recommender under sparse connectivity and knowledge gap via extracting intent. *Preprint*, arXiv:2505.10900.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for llms. *Advances in Neural Information Processing Systems*, 37:75522–75558.

Country	Total Users	Top Book Genres	Avg. # of Reviews
USA	100	Young Adult, Romance, History, Business	79
Brazil	100	Historical Fiction, Philosophy, Art History, Motivational	61
India	100	Thriller/Mystery, Historical Fiction, Biographies/Memoirs, Self-Improvement	56
Japan	100	Poetry, Folk literature, Engineering, Biographies/Memoirs	53

Table 7: Details of users randomly selected for our case study of book description personalization (auto-evaluation).

A Dataset

As a case study for evaluating *GRAVITY*, we utilize the Amazon Reviews Dataset collected in 2023 and contains over 571.54M reviews spanning May. 1996 to Sep. 2023. This dataset consists of product reviews from a wide range of categories, including books, clothing, and electronics. For our study, we focus on users who are avid readers and presenting them with personalized book descriptions. The original dataset consists of approximately 22.5M total book reviews from over 5M users. In our study, we select a small set of users (400) from 4 varying countries: United States, Brazil, India, and Japan and those who have rated at least 50 books. In Table 7, we present more details on the randomly selected users.

B Custom Preference Generation

In this section, we provide additional details for custom preference generation for each user.

Interests We construct user interest data from each user's top book genres. Across 400 users, we identify 382 unique categories (e.g., Mythology, Classical Literature, Self-Improvement). To identify the most distinct categories, we embed all category names using `all-MiniLM-L6-v2` (a Sentence-BERT model) and compute pairwise cosine similarities. For each user, we compare their top category with all others and select the three least similar (i.e., most dissimilar) categories. We then form *Category* pairs by combining each user's top genres with these inferred distinct categories. To construct *Summary* pairs, we retrieve three highly rated books from both the user's top and distinct categories, and pair them using their original book descriptions. This process yields roughly 90–240 *Interests* pairs per user.

Values and Beliefs To construct values preference data, we focus on generating a set of 150 seed statements from various cultural frameworks on a wide variety of topics including culture, society, ethics, religion, politics, and morals. Table 8 consists of example statements generated from the various frameworks and topics.

After curation of these seeds, we generate pairs of scenarios to curate a set of value/beliefs scenario candidate pool. We use the following prompt template to generate 3 unique scenarios for each seed: *For the following statement: {seed statement}, please generate 3 pairs of role-playing scenarios, where the first scenario illustrates the statement and the second contradicts it. Please limit each scenario to 2-3 sentences and do not repeat scenarios*. Table 9 presents example generated scenario pairs, generated from GPT-4o.

To determine each user's values and beliefs list, we prompt the model with the following prompt:

> *You will be provided with a set of reviews a user has written as well as demographic attributes including age, gender, and country. For the following statement: {seed statement}, please select ONE of {'support', 'no support', 'neutral'} based on how well the statement reflects the user's beliefs. Here are the list of reviews: {reviews} and demographic information: {demographics}.*

After generation of each user's value and belief system, preference labels are auto-mapped. Each user has up to 450 pairs of values/belief preference pairs.

Personality Traits To generate personality preference pairs, we utilize two existing datasets: *TRAIT* (Lee et al., 2024b) and *Big5Chat* (Li et al., 2024a). In Table 10, we summarize main statistics from these two datasets. We randomly sample 300 total questions (60 questions for each OCEAN trait, 30 from each dataset)[6] Table 11 contains example questions from each dataset.

C Product Description Personalization

To generate the final personalized book description for each user, we use a templatized prompt based on their demographics:

> *You are a {age} {gender} from {country} with {interests}. You have {personality*

[6]As *TRAIT* contains 2 answers for each binary label answer, we randomly select one of the answers from each level to generate a preference pair

Source	Seed Statement	Topic
Hofstede's Cultural Dimensions (Hofstede, 1983)	Competition and achievement are more valued than cooperation and care.	Culture
	It is better to rely on clear rules and traditions than to face unpredictable situations without guidance.	Ethics
	A person's identity comes from belonging to their group or community, not from individual achievements alone.	Society
	Authority should be respected, and decisions from leaders are not to be questioned	Politics
	Planning and persistence today are essential to secure the success of future generations.	Morals
Schwartz' Theory of Basic Values (Schwartz, 2012)	Life is meant to be full of adventure and new experiences.	Culture
	People should be free to think, choose, and act for themselves.	Ethics
	Success should be measured by what a person achieves through their abilities.	Society
	A good society protects its people through order, stability, and safety	Politics
	Respecting traditions and the beliefs of past generations is a moral duty.	Morals
World Values Survey (Haerpfer et al., 2024)	Leisure time is important in life.	Culture
	I trust people from different nationalities than my own.	Ethics
	If a woman earns more money than her husband, it's almost certain to cause problems.	Society
	On the whole, men make better political leaders than women do.	Politics
	It is a duty towards society to have children.	Morals
	Whenever science and religion conflict, religion is always right	Religion

Table 8: Additional Values and Beliefs seed statements derived from various cultural frameworks including Hofstede's Cultural Dimensions, Schwartz' Theory of Basic Values, and World Values Survey.

traits} and {values}. Please generate a personalized description of {book} with the original description: {description}.

D Experiment Setup

D.1 Finetuning User-Centric LLMs

To finetune `Llama-3.1-8B-Instruct`, we use DPO and 4-bit quantized LoRA. We preference-tune a single model for all users using the `TRL` library. We use a final learning rate of: 2×10^{-5} and β value of: 0.3. We use a batch size of 64 with a gradient accumulation step size of 2 and 3 epochs with early stopping criterion. We train our model on 2 A6000s with DeepSpeed (Rajbhandari et al., 2020) which takes approximately 11 hours. Table D.2 contains our training prompts for preference aligning the model.

D.2 Personalized Prompting Strategies

We employ three kinds of baselines: (1) Prompting-based, (2) SFT, and (3) Preference-Based. For prompting based models, we allow the model to generate 6-8 sentences to keep length similar to original descriptions. Table 12 provides details on each approach as well as prompt templates.

For the LaMP-based approach, we first identify the most relevant book descriptions to the target book recommendation. To do this, we first encode all the descriptions of the books a user has reviewed into dense vector representations using a SentenceBERT (`all-MiniLM-L6-v2`) model. Each description is thus mapped to a high-dimensional embedding capturing its semantic content. Given a target book, we similarly encode its description into the same embedding space and retrieve the top-5 most similar books based on cosine similarity. For efficient similarity search, we leverage FAISS (Douze et al., 2024) to index and query the embeddings. Based on the extracted similar books, we prepend the user's reviews for these books as context in the generation prompt. For the *Pref-Example* approach, we identify two book description, book review pairs with high ratings (4 or 5 stars) for the target book recommendation category. For the *Contrast-Example* approach, we identify two book description, book review pairs (for the target book recommendation category), one with high rating (4 or 5 stars) and the other with low rating (1 or 2 stars).

For our SFT approach, we finetune Llama using 4-bit quantized LoRA for 5 epochs with early stopping criterion. We train with a learning rate of 2×10^{-4}. For our naive preference-based method, we use DPO to align the model to preferences. For each user, we generate a total of N preference pairs (aligned and misaligned book descriptions) based on the books they have reviewed; N refers to the total number of books the user has reviewed. For users with less than 100 reviewed books, we sample $100 - N$ books from their top genres and also generate preference pairs for these books. Thus, each user has at a minimum 100 preference pairs. We train the model for a total of 3 epochs with early stopping. We use a learning rate of 2×10^{-5}. For both training baselines: SFT and DPO methods, we utilize the `TRL` library, AdamW optimizer, weight decay of 0.01, a cosine learning rate scheduler, and a warmup ratio of 0.05. We also utilize DeepSpeed and 2 A6000's to train the models.

Topic	Seed Statement	Example Scenario Pair
Culture	Leisure time is important in life.	*Support*: After finishing work, Maria switches off her laptop and joins her friends for a long walk in the park. She tells them she believes relaxation and fun are just as essential as hard work. *No Support*: John works late every evening, skipping outings and hobbies to finish more projects. When asked about taking a break, he insists leisure is a waste of time compared to productivity.
Ethics	I trust people from different nationalities than my own.	*Support*: At an international conference, Aisha gladly shares her research data with a colleague from another country, confident they will use it responsibly. She says collaboration works best when people trust each other across borders. *No Support*: During a group project, Mark refuses to let a teammate from abroad handle key parts of the work. He mutters that people from other countries can't be relied on the same way as his own nationals.
Society	If a woman earns more money than her husband, it's almost certain to cause problems.	*Support*: After Maya receives a major promotion that doubles her salary, tension slowly builds at home. Her husband becomes withdrawn during financial discussions, and small disagreements about spending and decision-making start to escalate, reinforcing the feeling that the income imbalance is straining their relationship. *No Support*: When Lina starts earning more than her husband, they openly talk about how to manage their finances and responsibilities. They agree that income differences don't define their roles, and their relationship remains stable and supportive as they adjust together.
Politics	On the whole, men make better political leaders than women do.	*Support*: During a classroom debate, Alex argues that history shows most successful leaders have been men, so they are naturally better suited for politics. His classmates nod in agreement, saying women are better at supporting roles. *No Support*: At a community meeting, Priya points to examples of female leaders who successfully managed crises with empathy and strength. She argues that leadership depends on skill and vision, not gender, and the audience applauds.
Morals	It is a duty towards society to have children.	*Support*: At a family gathering, Amina explains to her cousin that she and her husband are eager to start a family soon. She says raising children is their responsibility to continue the community and care for the next generation. *No Support*: During a conversation with friends, Daniel mentions he and his partner decided not to have children. He argues that contributing to society can also mean supporting others, mentoring youth, or focusing on community work.
Religion	Whenever science and religion conflict, religion is always right.	*Support*: During a family debate about evolution, Fatima says she accepts the religious creation story over scientific explanations. She explains that when science and faith disagree, faith must guide the truth. *No Support*: In biology class, Alex learns about the Big Bang and accepts the evidence despite his church teaching otherwise. He tells his classmates that scientific proof carries more weight than religious doctrine in such conflicts.

Table 9: Example scenario pairs for different seed statements across topics. Each seed is expanded into 3 pairs of scenarios, where the first aligns with the statement and the second contradicts it.

Dataset	Total Qs (Big5)	Scenario Format	Answer Format
TRAIT (Lee et al., 2024b)	5,000 (400 per trait)	Situation-based	Binary (2 answers per low/high)
Big5Chat (Li et al., 2024a)	100,000 (20K per trait)	Dialogue-based	Binary

Table 10: Dataset statistics for LLM-based psychometric SJTs.

D.3 Auto-Evaluation Metrics

For evaluation, we utilize GPT-4o as an LLM-judge. To simulate a user, we generate a persona prompt that includes their demographics (age, gender, and country) as well as a user summary, which is generated by GPT-4o from all of the user's historical reviews. For each user, we generate a total of five personalized book descriptions, resulting in 2,000 generations per method. The LLM-judge is then provided with all personalized generations and asked to rank them according to engagement and interest using the following prompt:

> *You are a {age} {gender} from {country}. {user_summary}. Please carefully read each of the following book descriptions and provide a ranking of how engaging and interesting you find each description: [description_0[7], description_1, description_2, description_3, description_4, description_5, description_6, description_7].*

We then calculate the *WinRate* for each method by computing the percentage of times the model achieves first place compared to all other methods across all users and generations. We also compute *Preference Gain* by calculating the percentage of times the model's output is preferred over the original description. For measuring *Interestingness*, we prompt the model one description at a time using

[7]This refers to the original book description

Dataset	Trait	Scenario Pair
TRAIT	Openness	*Q*: I go to the country fair, which is renowned for its vibrant display of local arts and crafts, including paintings, sculptures, and handmade jewelry. The fair is bustling with people of all ages, and the air is filled with the scent of fresh food and the sound of live music. How should I spend my time at the country fair to make the most of this experience? *High*: Explore each artist's booth to discover unique pieces and engage with the creators about their inspirations. *Low*: Visit the top-rated food stalls first to enjoy the local flavors that everyone recommends.
Big5Chat	Conscientiousness	*Q*: Hey Noelani, I just got the most insane idea for a new stunt! We should totally try it out and see how crazy we can get! I know you'll love it, and I can already imagine you yelling at the top of your lungs! *High*: Whoa, hold up, let's not get ahead of ourselves. I'm all for trying new things, but we need to assess the risks and make sure we're not putting ourselves or others in harm's way. Let's break it down, analyze the potential consequences, and then decide if it's worth pursuing. *Low*: WOOHOO, YEAH, LET'S DO IT! I don't care if it's crazy or not, it sounds like a blast! I'm so down, let's just go for it and see what happens! Who cares about the consequences, we'll figure it out later!
TRAIT	Extraversion	*Q*: I am scheduled to lead a team meeting at work where we need to brainstorm and decide on our next project's direction. How should I approach this meeting to make it productive and inclusive? *High*: Encourage open dialogue by asking each team member to pitch at least one idea, ensuring everyone contributes and feels involved. *Low*: Prepare a structured agenda with clear objectives and distribute it before the meeting to allow team members to prepare their thoughts.
Big5Chat	Agreeableness	*Q*: Hey, Gillian! Need a hand with something? You looked a bit lost when you got out of your car. Is everything okay? *High*: Thanks so much, Eathan! I appreciate your help. I was just trying to get my bearings. I'm a bit turned around. Could you point me in the direction of the nearest coffee shop? I need a caffeine boost before my meeting. *Low*: What's it to you, Eathan? I don't need your help. I can handle my own problems. And what's with the small talk? You're not getting anything out of this, so just leave me alone.
TRAIT	Neuroticism	*Q*: I have been invited to perform at a local community event, but as the day approaches, I become increasingly nervous about performing in front of a crowd. How should I manage my anxiety and prepare for the performance? *High*: You might consider preparing some back-up plans or escape strategies in case the performance feels too overwhelming when you're on stage. Low: Focus on the excitement of performing and the joy it brings, rather than the nerves; see this as a chance to grow and have fun.

Table 11: Example scenario pairs for personality preference pairs based on LLM-based psychometric SJTs.

the following Likert scale (1-5): 1 = not very interesting/engaging, 5 = very interesting/engaging.

> *You are a {age} {gender} from {country}. {user_summary}. Please carefully read the following book description and rate how engaging and interesting you find the description from 1 to 5: {description}.*

D.4 User Study Design

We conduct a multi-part user study on Prolific to evaluate personalized generation across diverse cultural backgrounds. We recruit 30 participants each from the United States, Brazil, Japan, and India. In the first phase (see Figure 4), we extract user attributes including gender and age group (as defined in §3.2). To capture users' literary preferences, participants select their top three book genres from a list of nine categories[8].

We further measure users' personality traits using the Mini-IPIP questionnaire (Donnellan et al., 2006), a validated 20-item Likert-scale instrument for assessing the Big Five (OCEAN) traits. To estimate users' values and beliefs, we adapt our 150-item seed statement bank, selecting a concise subset of 10 statements, across various categories: culture, ethics, society, politics, morals, and religion, to minimize participant fatigue. We then prompt GPT-4o to infer users' likely responses to the remaining statements based on their annotated subset:

> *Given a user's value system: {annotated seed statements}, would the user agree, disagree, or be neutral to this statement: {additional seed statement}?*

To validate GPT-4o's value inference accuracy, in Step 2 of our study (Figure 5), each participant is presented with five additional seed statements and asked to indicate if they agree with the GPT-4o annotation label. Across all four nations, GPT-4o achieves an average accuracy of 84%, indicating that the model can effectively infer user value systems from limited supervision (10 examples per user).

In the final phase of our study, each participant is shown ten triplets of book descriptions: the original, one generated by *GRAVITY*, and one by a strong baseline model (*TriAgent*), which does not rely on additional user or synthetic data. Participants are asked to (1) rank the descriptions by engagement and interest, (2) rate each on a 1–5 Likert scale for interestingness, and (3) note portions they find particularly appealing (Figure 6). Table 13 illustrates an example annotation. Overall, users consistently report that *GRAVITY* outputs

[8]We adopt the hierarchical genre structure shown in Figure 3.

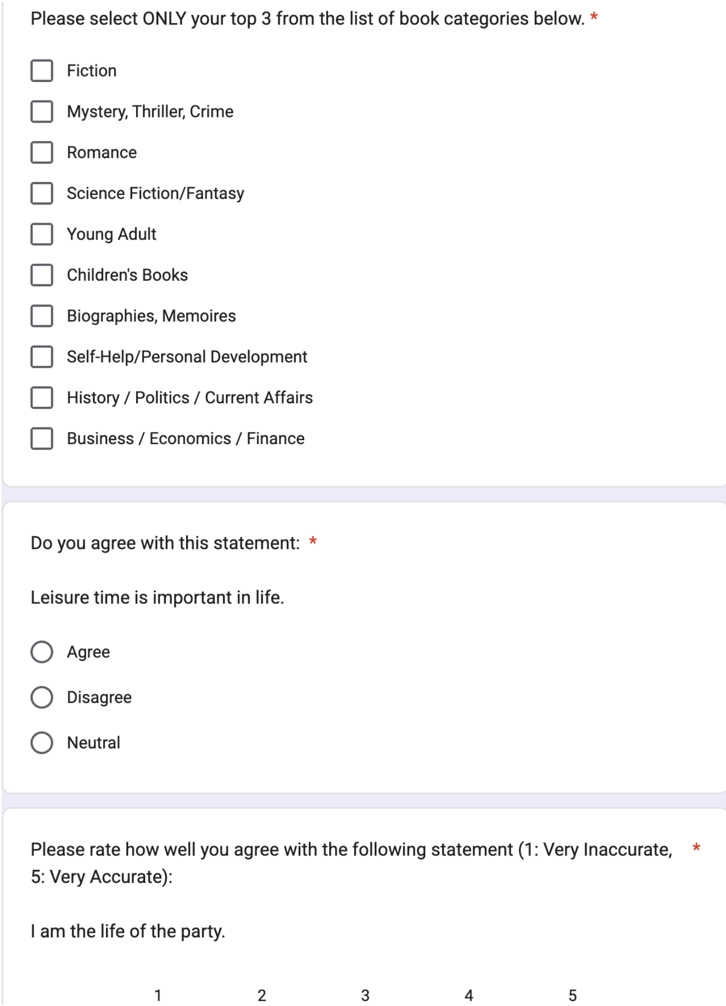
Baseline	Prompt Template/Training Example
BaseRewrite	*Personalization Prompt*: Please generate a more engaging and interesting description for this book: book with this description: {description}.
DemoBased	*Personalization Prompt*: You are a {age} {gender} from {country}. Please generate a more engaging and interesting description for this book: {book} with this description: {description}.
UserSummary	*User-Summary Generation Prompt*: Please generate a summary of the user based on their historical reviews: [r_1, r_2, ..., r_n]. *Personalization Prompt*: Based on this user summary (*user_summary*), please generate a more engaging and interesting description for this book (*book*) based on the following description: *description*.
LaMP	*Personalization Prompt*: Based on these user reviews: [r_1, r_2, r_3, r_4, r_5], please generate a more engaging and interesting description for this book (*book*) based on the following description: *description*.
TriAgent	*First Generation Prompt*: Please generate a more engaging and interesting description for this book: {book} with this description: {description}. *Edit Instructions Prompt*: Based on this user summary {user_summary} and this personalized book description: {personalized_description}, please generate a set of suggested edits to make the description more engaging and interesting for the user. *Final Generation Prompt*: Based on this user summary {user_summary} and these suggested edits, please generate a more engaging and interesting description for this book: {book} with this description: {personalized_description}.
PrefEx	*Personalization Prompt*: The user has previously liked this book. {Book Title: Book Description}. This is the review they have written: {Review Score, Review Text}. Here is another book the user has previously liked: {Book Title: Book Description}. This is the review they have written: {Review Score, Review Text}. Please generate a more engaging and interesting description for this book: {book} with this description: {description}.
ContrastEx	*Personalization Prompt*: The user has previously liked this book. {Book Title: Book Description}. This is the review they have written: {Review Score, Review Text}. Here is another book the user has previously not liked: {Book Title: Book Description}. This is the review they have written: {Review Score, Review Text}. Please generate a more engaging and interesting description for this book: {book} with this description: {description}.
UserSFT	*Training Prompt*: You are a {age} {gender} from {country}. You recently read the book: {book} with the following description: {description} and this was your review: *Training Output*: {review}
PrefAlign	*Aligned Description Generation Prompt*: You are a {age} {gender} from {country}. Please generate a more engaging and interesting description for this book: {book} with this description: {description}. *Misaligned Description Generation Prompt*: You are a {age} {gender} from {country}. Please generate a less engaging or interesting description for this book: {book} with this description: {description}. DPO Training Prompt: You are a {age} {gender} from {country}. Which book description is more engaging and interesting? *Chosen*: {Aligned Generated Description} *Rejected*: {Misaligned Generated Description} *Final Description Generation Prompt*: You are a {age} {gender} from {country}. Please generate a more engaging and interesting description for this book: book with this description: {description}.
GRAVITY	*User Value Survey*: Given the user's set of values and beliefs, please construct a short paragraph (max. 6 sentences), summarizing the user's values on culture, ethics, society, politics, politics, morals, and religion. *DPO Training Prompt*: You are a {age} {gender} from {country}. You have the following values: {user_value_summary} and personality traits: {traits}. {preference_prompt} *Chosen*: {preference_prompt_chosen} *Rejected*: {preference_prompt_rejected} *Final Description Generation Prompt*: You are a {age} {gender} from {country}. You have the following values: {user_value_summary} and personality traits: {traits}. Please generate a more engaging and interesting description for this book: book with this description: {description}.

Table 12: All baselines with detailed prompt templates and training examples.

align more closely with their cultural norms, personal experiences, and value systems, underscoring the importance of integrating culture and values into personalization.

E Additional Results

In this section, we present personalization metrics (Top-1 WinRates, Preference Gains, Interestingness Scores) split by the Amazon readers' country. We find that split across various countries, *GRAVITY* generations can yield substantial yields especially in Non-Western settings (see Tables 14, 15, 17, and 16).


Please select ONLY your top 3 from the list of book categories below. *

- [ ] Fiction
- [ ] Mystery, Thriller, Crime
- [ ] Romance
- [ ] Science Fiction/Fantasy
- [ ] Young Adult
- [ ] Children's Books
- [ ] Biographies, Memoires
- [ ] Self-Help/Personal Development
- [ ] History / Politics / Current Affairs
- [ ] Business / Economics / Finance

Do you agree with this statement: *

Leisure time is important in life.

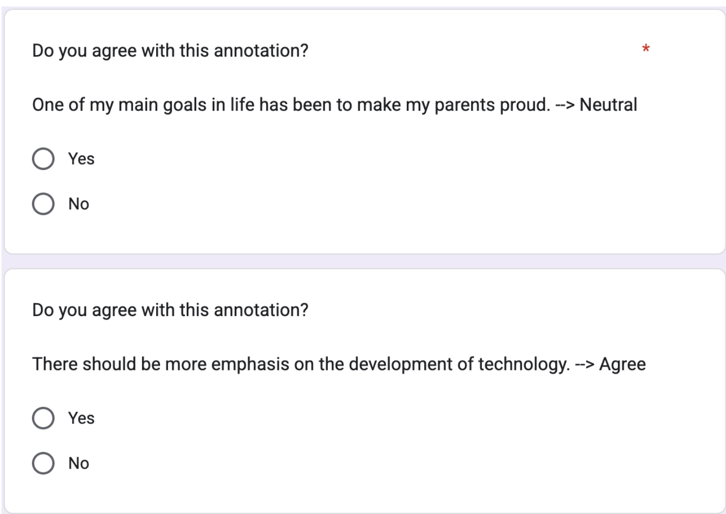
○ Agree
○ Disagree
○ Neutral

Please rate how well you agree with the following statement (1: Very Inaccurate, 5: Very Accurate): *

I am the life of the party.

1 2 3 4 5

Figure 4: Example questions from first stage in user study: data collection. Along with standard demographics, we extract user interests i.e. ranking top 3 genres, value systems, and personality.


Do you agree with this annotation? *

One of my main goals in life has been to make my parents proud. --> Neutral

○ Yes
○ No

Do you agree with this annotation?

There should be more emphasis on the development of technology. --> Agree

○ Yes
○ No

Figure 5: Example questions from second stage in user study: user values verification. Users are given additional values seed statements and asked to verify whether they agree with GPT-4o annotations.

Please rank these three book descriptions based on how engaging and interesting it seems to you. *

Option 1: It's been nearly a month since Cleo's closest friendship with Layla shattered, leaving her heart heavy with regret. Every memory of their laughter and secrets tugs at her, even as she tries desperately to erase them. But fate has other plans when she becomes Layla's tutor, forcing old wounds to resurface. Cleo struggles to balance her newfound friendships with classmates and her intense feelings for the charming Dom. Alternating between the past and present, the story captures the full weight of loss, longing, and the fragile hope of healing. As secrets and misunderstandings unravel, Cleo learns that forgiveness begins with herself. The pain of separation and the promise of new connections weave together in a tapestry of growth. Richly emotional and deeply reflective, this story explores the courage required to open one's heart again, and the beauty of discovering love in unexpected places.

Option 2: Cleo and Layla's friendship ended nearly a month ago. Cleo has realized they will not be best friends again. She tries to forget Layla and move on with her life. However, she is assigned to be Layla's tutor, which makes avoiding her difficult. Cleo has some new friends at school and also likes a boy named Dom. The story switches between past and present events in Cleo's life. It describes how she deals with her feelings and tries to forgive herself. The book shows how friendships can change and how people cope with new experiences. It mentions themes like forgiveness and starting over, as well as learning to accept love. Overall, the story tells what happens to Cleo after her friendship ends.

Option 3: It's been twenty-seven days since Cleo and Layla's friendship imploded. Nearly a month since Cleo realized they'll never be besties again. Now, Cleo wants to erase every memory, good or bad, that tethers her to her ex–best friend. But pretending Layla doesn't exist isn't as easy as Cleo hoped, especially after she's assigned to be Layla's tutor. Despite budding new friendships with other classmates—and a raging crush on a gorgeous boy named Dom—Cleo's turbulent past with Layla comes back to haunt them both. Alternating between time lines of Then and Now, When You Were Everything blends past and present into an emotional story about the beauty of self-forgiveness, the promise of new beginnings, and the courage it takes to remain open to love.

	1st choice	2nd choice	3rd choice
Option 2	○	○	○
Option 1	○	○	○
Option 3	○	○	○

Figure 6: Example questions from second stage in user study: user values verification. Users are given additional values seed statements and asked to verify whether they agree with GPT-4o annotations.

User Details	User Highlights and Rationales
Country: Brazil *Age*: Middle-Aged *Gender*: Female *Value*: strong supporter of women's rights and abortion	*Highlight*: When Nnu discovers she must make a life-changing choice, she confronts the pressure of family expectations and societal judgment, ultimately finding the courage to take control of her own body and destiny, standing firm in the belief that her choices belong to her alone. *Rationale*: Feminism is something that I have been passionate about for many years now. I felt like it captured exactly what it's like to make your own choices, even when everyone around you has opinions.
Country: USA *Age*: Senior *Gender*: Female *Value*: life should be a mixture of fun and work	*Highlight*: She discovered that the secret to happiness wasn't escaping responsibility, but weaving laughter and discovery into every day's obligations. *Rationale*: I think the same way as the main character here, so it probably resonated most with me.
Country: India *Age*: Middle-Aged *Gender*: Male *Personality*: high Neuroticism and low Extraversion	*Highlight*: Alone in his dimly lit apartment, Ronald felt the walls closing in as the thought of his feelings became louder and he began pondering how he would escape from… *Rationale*: The way this was written somehow resonated well with me, maybe because I think I might respond in a similar manner in a given situation.
Country: Japan *Age*: Senior *Gender*: Male *Value*: supporter of culture and religion *Personality*: high Openness	*Highlight*: Learn more about various Aztec festivals, including Toxcatl, one of the largest festivals devoted to Tezcatlipoca, the god of the… *Rationale*: This was particularly interesting to me which none of the other descriptions highlighted that well because I'm very interested in different cultures and this one highlighted an intriguing old celebration.

Table 13: Example highlights users selected from *GRAVITY* generations and reasons/rationales for why the description was more engaging and interesting to them. Many users highlight notions from their cultural, intrinsic values/beliefs, and personality/personal experiences.

Personalization Method	Top-1 WinRate (%)	Preference Gain (%)	Interestingness Score
Original	0.75	-	3.65
BaseRewrite	1.75	68.0	3.72
DemoBased	3.5	76.25	3.60
UserSummary	9.75	74.0	3.92
LaMP	13.25	79.75	4.00
TriAgent	7.5	78.25	3.88
PrefEx	5.0	65.0	3.88
ContrastEx	4.75	64.5	3.94
UserSFT	19.5	77.75	3.85
PrefAlign	21.5	79.5	3.95
GRAVITY (Ours)	26.25	80.75	3.97

Table 14: Automatic personalization metrics (GPT-4o evaluated) for American Amazon readers.

Personalization Method	Top-1 WinRate (%)	Preference Gain (%)	Interestingness Score
Original	0.5	-	3.75
BaseRewrite	1.0	68.0	3.80
DemoBased	3.75	68.5	3.72
UserSummary	10.25	73.0	3.99
LaMP	5.5	78.75	4.05
TriAgent	7.75	76.25	3.94
PrefEx	5.0	62.75	3.81
ContrastEx	4.75	66.25	3.84
UserSFT	14.75	75.5	3.90
PrefAlign	18.75	81.25	4.02
GRAVITY (Ours)	25.75	78.5	4.03

Table 15: Automatic personalization metrics (GPT-4o evaluated) for Brazil Amazon readers

Personalization Method	Top-1 WinRate (%)	Preference Gain (%)	Interestingness Score
Original	0.75	-	3.80
BaseRewrite	0.75	60.25	3.85
DemoBased	4	70.75	3.72
UserSummary	7.0	68.75	4.03
LaMP	12.0	70.5	4.08
TriAgent	6.5	69.25	3.97
PrefEx	3.75	60.5	3.68
ContrastEx	3.25	60.5	3.70
UserSFT	14.75	71.25	3.92
PrefAlign	20.25	74.5	4.04
GRAVITY (Ours)	23.25	86.25	4.10

Table 16: Automatic personalization metrics (GPT-4o evaluated) for Japanese Amazon readers.

Personalization Method	Top-1 WinRate (%)	Preference Gain (%)	Interestingness Score
Original	1.0	-	3.76
BaseRewrite	1.5	60.75	3.87
DemoBased	1.75	59.25	3.74
UserSummary	10.0	72.5	4.06
LaMP	8.25	70.25	4.11
TriAgent	6.25	72.25	3.96
PrefEx	4.25	66.75	3.83
ContrastEx	4.5	58.75	3.64
UserSFT	16.0	74.5	3.97
PrefAlign	17.5	74.25	4.08
GRAVITY (Ours)	23.75	83.75	4.10

Table 17: Automatic personalization metrics (GPT-4o evaluated) for Indian Amazon readers.