

Measuring What Makes You Unique: Difference-Aware User Modeling for Enhancing LLM Personalization

Yilun Qiu^{1*}, Xiaoyan Zhao^{2*}, Yang Zhang^{1†}, Yimeng Bai³,
Wenjie Wang³, Hong Cheng², Fuli Feng³, Tat-Seng Chua¹

¹National University of Singapore

²The Chinese University of Hong Kong

³University of Science and Technology of China

qiuyilun@u.nus.edu, xzhao@se.cuhk.edu.hk, zyang1580@gmail.com, baiyimeng@mail.ustc.edu.cn,
wenjiawang96@gmail.com, hcheng@se.cuhk.edu.hk, fulifeng93@gmail.com, dcscts@nus.edu.sg

Abstract

Personalizing Large Language Models (LLMs) has become a critical step in facilitating their widespread application to enhance individual life experiences. In pursuit of personalization, distilling key preference information from an individual’s historical data as instructional preference context to customize LLM generation has emerged as a promising direction. However, these methods face a fundamental limitation by overlooking the inter-user comparative analysis, which is essential for identifying the inter-user differences that truly shape preferences. To address this limitation, we propose Difference-aware Personalization Learning (DPL), a novel approach that emphasizes extracting inter-user differences to enhance LLM personalization. DPL strategically selects representative users for comparison and establishes a structured standard to extract meaningful, task-relevant differences for customizing LLM generation. Extensive experiments on real-world datasets demonstrate that DPL significantly enhances LLM personalization. We release our code at <https://github.com/SnowCharmQ/DPL>.

1 Introduction

With continuous efforts, Large Language Models (LLMs) (Achiam et al., 2023; Yang et al., 2024; Dubey et al., 2024; Guo et al., 2025) have demonstrated increasingly higher levels of intelligence, sparking an unprecedented enthusiasm for applying them to individuals’ daily lives to enhance personal thinking, planning, and overall life experiences (Bao et al., 2023; Li et al., 2024a; Deng et al., 2024; Zhang et al., 2025b). However, the “one-size-fits-all” paradigm of general LLMs becomes ineffective when serving individual users, as users have distinct preferences and would expect personalized generation from LLMs (Christakopoulou

et al., 2023). This motivates an exciting new direction: *LLM personalization* (Zhang et al., 2024d; Xu et al., 2025b; Liu et al., 2025a; Chen et al., 2024a) – adapting LLMs to generate text responses that align with each individual’s unique preferences.

LLM personalization fundamentally relies on incorporating user-specific information into LLMs to tailor their generation (Chen et al., 2025; Zhao et al., 2025d). The “memorize-then-inject” framework offers a promising and economical solution – storing users’ historical data and then extracting key information to serve as instructional contexts for customizing LLM responses (Salemi et al., 2024b; Wang et al., 2024; Wu et al., 2025). Depending on the extraction method employed, existing approaches can generally be categorized into: 1) Retrieval-based methods (Li et al., 2023a; Mysore et al., 2024), which just retrieve the parts of the user’s history relevant to the current request for LLM use, and 2) Summarization-based methods (Richardson et al., 2023; Zheng et al., 2024), which further condense the user history to distill preferences for LLM use.

While these methods (Mysore et al., 2024; Zheng et al., 2024) have made notable progress, we contend that they share a common limitation: they overlook the inter-user comparative analysis, which is essential for identifying the differences that truly shape preferences. As acknowledged in psychology and behavioral science (Snyder and Fromkin, 1977, 2012; Irmak et al., 2010), it is the differences among individuals that make them unique, while their uniqueness shapes their preferences. This underscores that effective personalization hinges on identifying and understanding these differences. As a result, without considering inter-user comparisons to extract differences, current methods would be suboptimal for personalization.

In this work, we propose difference-aware user modeling – extracting the differences among users as key preference information – to enhance LLM

*Equal contribution

†Corresponding Author

personalization. To achieve this, distinguishing each user from others and then summarizing the differences (using an LLM) to serve as the instructional contexts for personalization provides an intuitive solution. However, it faces two challenges: 1) comparing to all users may be impractical due to the high cost and complexity of extracting differences across all others; and 2) without proper guidance, the differences extracted by the LLM may be irrelevant or meaningless for the task. Therefore, we must address: 1) how to select appropriate users for comparison, and 2) how to extract task-relevant differences during comparison effectively.

To address these challenges, we introduce a novel *Difference-aware Personalization Learning* (DPL) method. DPL selectively chooses representative users for comparison and establishes an extraction standard to effectively capture task-relevant differences, using the extracted differences as the LLM’s instructional contexts for personalization. Specifically, we cluster users based on data similarity and select only cluster-center users for comparison, ensuring efficiency while maintaining a global perspective. To extract task-relevant differences, we structure the difference extraction model (an LLM) to focus on key predefined dimensions for capturing differences, rather than operating freely. Three key preference dimensions are distilled from existing literature as examples: writing style, emotional tone, and semantic content.

We validate our DPL method on a representative personalized text generation task, namely, review generation (Ni et al., 2019; Peng et al., 2024; Au et al., 2025). Extensive results demonstrate that DPL effectively enhances LLM personalization.

The main contributions of this work are summarized as follows:

- We emphasize the philosophy that “differences make us unique” as the key to personalization and propose using difference-aware user modeling to enhance LLM personalization.
- We propose a novel Difference-aware Personalization Learning method for LLM, incorporating a selective comparison mechanism and structured difference extraction mechanism to achieve effective difference extraction.
- Extensive experiments on real-world datasets demonstrate that DPL achieves state-of-the-art performance, highlighting its effectiveness in personalized text generation.

2 Preliminary

Problem Formulation. This work focuses on personalizing LLMs to generate text outputs that align with user preferences, *i.e.*, achieving personalized text generation. We assume that users have a set of historical texts, either written or preferred by them, reflecting the users’ preferences. We could leverage this data to customize LLM’s generation to meet user needs. Formally, let \mathcal{D} denote the historical data for all users. Each sample in \mathcal{D} is represented as $(u, i, y) \in \mathcal{D}$, where u represents the user, i denotes the item (or object) the user is focused on, and y indicates the text written or preferred by u for item i . When a target user u' submits a new request for text generation on a target item i' , the LLM needs to generate text that satisfies the user’s preferences based on \mathcal{D} .

Notably, without loss of generality, this paper primarily focuses on review generation, a representative personalized text generation task. In this context, the goal is to enable LLMs to generate personalized reviews for items such as movies or products, ensuring alignment with true user reviews.

Memorize-then-Inject Solution. A typical approach in existing works for customizing LLM generation is to store the user’s historical data and then extract key information from the data to distill preference information through retrieval or summarization when needed. This extracted information serves as a preference context in the instruction to guide the model’s generation. Formally, to generate personalized text for the focused user u' on the target item i' , the method can be represented as follows:

$$\hat{y} = \text{LLM}(u', i', \phi_{key}(\mathcal{D}_{u'})), \quad (1)$$

where \hat{y} denotes the generated result, $\mathcal{D}_{u'}$ denotes the historical data for u' , and $\phi_{key}(\mathcal{D}_{u'})$ represents the key information extraction process from $\mathcal{D}_{u'}$. Notably, the extraction process usually also involves the utilization of an LLM. This method does not require LLM retraining and can potentially manage lifelong historical data in a memory manner (Zhang et al., 2024c), making it a promising and cost-effective solution.

3 Methodology

In this section, we introduce the proposed *Difference-Aware Personalization Learning* (DPL)

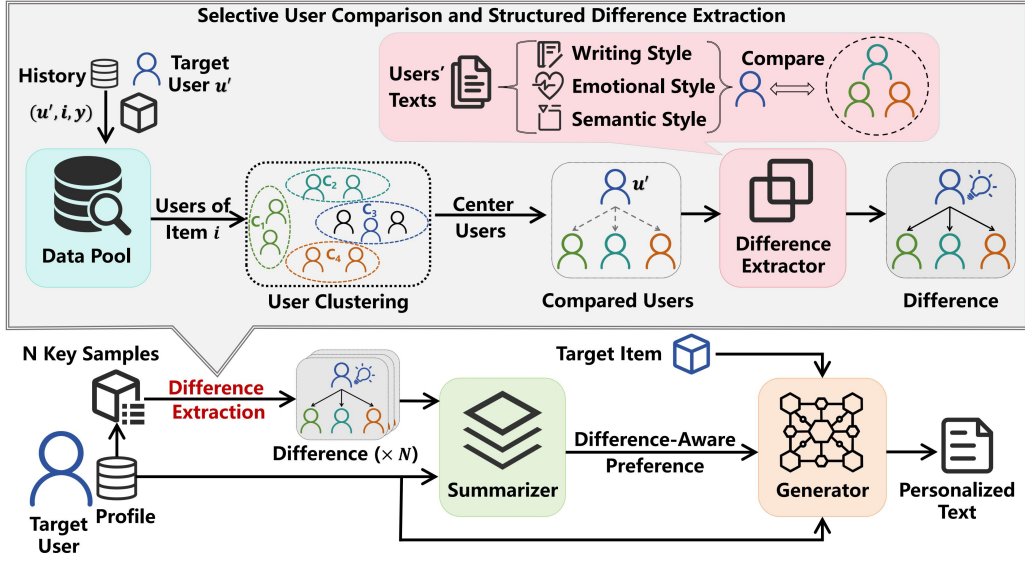


Figure 1: Overview of the proposed DPL method, which enables difference-aware preference extraction for LLM personalization. It extracts differences through selective user comparison via clustering and structured difference extraction along fixed dimensions (writing style, emotional style, and semantic style).

method, beginning with its motivation and the overall framework, followed by a detailed explanation of each component.

3.1 Overview

Given that the differences define individual uniqueness, we argue that inter-user differences play a crucial role in preference learning for LLM personalization. However, existing approaches overlook the inter-user comparisons for preference modeling, limiting the ability to extract meaningful differences for personalization. To address this limitation, we propose the DPL method. As shown in Figure 1, DPL introduces an inter-user comparison mechanism to identify differences between users, which are then combined with the user’s own key history to form the preference context in instructions, guiding LLM generation. This process can be formulated as:

$$\hat{y} = \text{LLM}(u', i', \phi(\mathcal{D}_{u'}; \mathcal{D})), \quad (2)$$

where $\phi(\mathcal{D}_{u'}; \mathcal{D})$ represents the difference-aware preference extracted by comparing a user’s historical data $\mathcal{D}_{u'}$ with other users’ data in \mathcal{D} .

Difference-Aware Preference Extraction Process. As illustrated in the bottom part of Figure 1, for user u' , we assume that $\mathcal{D}_{u'}$ contains N key elements, denoted as $\mathcal{D}_{u'}^*$, which can be obtained via retrieval (Robertson and Zaragoza, 2009). Then we would compare each element with those of some other users to extract differences and then summa-

rize these differences across all samples and the key historical data to obtain the final preference. Specifically, for each element $(u', i, y) \in \mathcal{D}_{u'}^*$, we compare it with the texts of other users on the same item i , represented by \mathcal{D}_i , using a **selective user comparison and structured difference extraction** process. This process consists of two key steps, as shown in the upper part of the figure:

- **Selective user comparison:** Instead of comparing with all users in \mathcal{D}_i , we selectively compare with a subset of representative users to reduce complexity and improve efficiency.
- **Structured difference extraction:** Differences are extracted according to predefined standards to ensure task relevance and meaningfulness.

The details of these steps will be elaborated later. After extracting differences for all N elements, we summarize them as well as the key history $\mathcal{D}_{u'}^*$ with an LLM to obtain the final difference-aware representation for personalization. Formally,

$$\phi(\mathcal{D}_{u'}; \mathcal{D}) = \text{LLM}_{\text{sum}}(\mathcal{D}_{u'}^*; [d_1, \dots, d_n, \dots, d_N]), \quad (3)$$

where LLM_{sum} denotes the summarizer, and d_n represents the extracted difference for the n -th element in $\mathcal{D}_{u'}^*$. Notably, we also take into account the user’s own data to retain preference information that cannot be captured by the differences, *i.e.*, using the key history $\mathcal{D}_{u'}^*$ as additional inputs. Next, we elaborate on the selective user comparison and structured difference extraction methods to explain how each d_n is obtained.

3.2 Selective User Comparison and Structured Difference Extraction

When dealing with each important historical data sample $(u', i, y) \in \mathcal{D}_{u'}^*$, a straightforward approach for difference extraction is to have the LLM compare all users and summarize the differences. However, comparing all users is complex and inefficient. To address this, we selectively compare only representative users. Additionally, since differences can span multiple dimensions, allowing LLMs to extract them freely may introduce task-irrelevant information. To mitigate this, we introduce a structured difference extraction approach with a predefined standard to guide the extraction of meaningful and task-relevant differences.

3.2.1 Selective User Comparison

To select representative users for comparison, the core consideration is that they could represent the total population, making the comparison still provide a global view. To achieve this, we consider clustering the users based on their text and then selecting the user in the cluster center for comparison. Specifically, when dealing with each key element in history $(u', i, y) \in \mathcal{D}_{u'}^*$ for the target u' , we just consider the users having texts on i , and then cluster these users into $K+1$ groups using the K-means algorithm (MacQueen, 1967) according to the texts corresponding to them, obtaining the most central users for each group. We then select the K central users that do not belong to groups of the target user as the users for comparison. Formally,

$$u_1, \dots, u_K = \text{cluster_center}(\mathcal{D}_i), \quad (4)$$

where u_1, \dots, u_K denotes the selected representative users, and \mathcal{D}_i denotes all the historical data that the users have texts on item i . We represent the data samples regarding the selected users on item i as: $(u_1, i, y_1), \dots, (u_K, i, y_K)$.

3.2.2 Structured Difference Extraction

After obtaining the representative users, we compare their texts with the target user's texts, *i.e.*, comparing $\{(u_1, i, y_1), \dots, (u_K, i, y_K)\}$ with (u', i, y) to obtain the difference d_n via an LLM. Instead of allowing LLMs to freely summarize the difference, we would define a standard to structure the LLM difference summarization. The process can be formulated as:

$$d_n = \text{LLM}_{\text{dif}}(\{(u_1, i, y_1), \dots, (u_K, i, y_K)\}; (u', i, y); S), \quad (5)$$

where LLM_{dif} denotes the LLM to extract differences. Here, S represents the standard that the LLMs follow to generate the difference. Specifically, we distill three key dimensions from existing literature (as examples): writing style, emotional style, and semantic style, and then structure the LLM to extract differences from these dimensions. The three dimensions are detailed below:

Writing style. This dimension captures the lexical choices, syntactic structures, and linguistic patterns unique to each user. By analyzing *vocabulary richness*, *sentence complexity*, and *grammatical preferences*, we can identify how a user's writing diverges from others (Verma and Srinivasan, 2019).

Emotional style. This feature reflects the sentiment and affective tone embedded in user's text. By assessing the *polarity* (*positive*, *negative*, *neutral*), we can highlight differences in how users convey their attitudes and feelings (Kheiri and Karimi, 2023).

Semantic style. This aspect focuses on the depth, clarity, and coherence of meaning in user's text. By evaluating *information density*, *logical flow*, and *contextual relevance*, we can differentiate how users structure and convey their ideas (Sahu and Majumdar, 2017).

Extracting differences along these dimensions ensures meaningful and well-structured results for distinguishing task-relevant content. We guide the LLM to adhere to these dimensions for difference extraction using prompts. All prompts used in DPL are provided in Appendix G.

4 Experiments

In this section, we conduct experiments to answer the following research questions:

RQ1: How does DPL perform on the personalized text generation task with real-world datasets, in comparison to other baseline methods?

RQ2: What is the impact of the individual components of DPL on its effectiveness?

RQ3: How do the specific hyper-parameters of DPL influence its performance?

RQ4: How does DPL perform across different levels of user uniqueness?

RQ5: How does DPL qualitatively demonstrate its personalized generation capabilities through concrete case analysis?

4.1 Experimental Setup

Datasets. We focus on review generation, a representative personalized text generation task. So we conduct experiments on the Amazon Reviews 2023 dataset (Hou et al., 2024)¹, focusing on the categories of **Movies & TV**, **CDs & Vinyl**, and **Books**. This dataset aggregates user-item interactions from Amazon, encompassing user reviews (ratings, texts, etc.) and item metadata (titles, descriptions, etc.). We preprocess the data into a format suitable for personalized review generation tasks, with details provided in Appendix A. Our processed datasets are publicly available on Huggingface².

Baselines. We compare DPL with the following methods, with more details provided in Appendix B. Note that all methods are implemented using the same template and input format, as outlined in Appendix G, to ensure fair comparisons.

- **Non-Perso.** This refers to the non-personalized approach, which completely excludes any user-specific information from the model input.
- **RAG** (Kumar et al., 2024). This is a method that employs a retrieval approach to extract key user-specific information, which is then used as instructional context for personalization.
- **IntSum** (Richardson et al., 2023). This method condenses the user’s retrieved history through summarization and incorporates it into the prompt. We achieve this by employing an LLM to generate a summary of the retrieved history.
- **LLM-TRSR** (Zheng et al., 2024). This method improves the IntSum method by segmenting the user’s history into blocks and employing a recurrent summarization paradigm to iteratively refine the summary within each new block.
- **CICL** (Gao and Das, 2024). This Contrastive In-Context Learning method is not originally designed for LLM personalization. We adapt it to our experimental scenario by applying it to incorporate the review from another user into the prompt to an LLM summarizer, thereby improving the quality of summary generation.
- **Persona-DB** (Sun et al., 2025). This method introduces a collaborative refinement approach that allows a user to retrieve and integrate information

from relevant peers. We apply it by incorporating the review of a relevant user into the prompt for the LLM generator.

Evaluation Metrics. We evaluate the experimental results using two complementary approaches, with additional details provided in Appendix C.

- **Conventional Evaluation.** Following previous work on personalized text generation (Kumar et al., 2024), we employ widely-adopted **ROUGE-1** (Lin, 2004), **ROUGE-L**, **METEOR** (Banerjee and Lavie, 2005), and **BLEU** (Papineni et al., 2002) to measure lexical overlap between the generated reviews and ground-truth reviews, offering surface-level comparisons through n-gram matching and semantic alignment mechanisms (Wang et al., 2023).
- **LLM-based Evaluation.** Following the previous evaluation framework (Salemi et al., 2025), we employ LLMs to assess the nuanced aspects of personalization, going beyond simple lexical matching. Specifically, the LLM evaluators process four key components – the generated review, the ground-truth review, review metadata, and item attributes – and produce raw scores ranging from 0 to 10. These scores are then normalized through min-max scaling to the [0, 1] range, with the evaluation metrics denoted as **S-72B** and **S-GPT**, corresponding to the use of *Qwen-2.5-72B-Instruct-AWQ* (Yang et al., 2024) and *GPT-4o-mini* (Achiam et al., 2023), respectively.

Implementation Details. For our DPL, we use the *Qwen2.5-14B-Instruct* (Yang et al., 2024) model as the LLM backbone, which functions as the difference extractor, summarizer, and review generator. Following the setting of LongLaMP (Kumar et al., 2024), we retrieve user reviews as auxiliary signals using BM25, where the retrieval query is formed by concatenating a short review title with the item’s description. For selective user comparison and structured difference extraction, we set the number of retrieved reviews, *i.e.*, N in Equation (3), to 8. Additionally, we utilize *gte-Qwen2-1.5B-Instruct* (Li et al., 2023b) to map user reviews into embeddings, which are subsequently clustered into 5 groups by default, equivalent to setting the number of selected representative users K in Equation (4) to 4. To assess the reliability of our generated results, we conduct inference using different temperature values (0.2, 0.4, 0.6, 0.8, and 1.0) for each method, and

¹<https://amazon-reviews-2023.github.io/>

²<https://huggingface.co/datasets/SnowCharmQ/DPL-main> & <https://huggingface.co/datasets/SnowCharmQ/DPL-meta>

Table 1: Performance comparison between the baselines and our DPL on the three datasets, where the best results are highlighted in bold and sub-optimal results are underlined. Higher values indicate better results for all metrics. The symbol * indicates p-value < 0.05 in t-tests.

| Datasets (→) | Movies & TV | | | CDs & Vinyl | | | Books | | |
|-------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|----------------------------|----------------------------|
| Methods (↓) | ROUGE-1 | METEOR | S-72B | ROUGE-1 | METEOR | S-72B | ROUGE-1 | METEOR | S-72B |
| Non-Perso | 0.2279 \pm 0.0022 | 0.1301 \pm 0.0008 | 0.5372 \pm 0.0043 | 0.2396 \pm 0.0029 | 0.1327 \pm 0.0007 | 0.5351 \pm 0.0048 | 0.2711 \pm 0.0058 | 0.1511 \pm 0.0023 | 0.5497 \pm 0.0050 |
| RAG | 0.2844 \pm 0.0017 | 0.1870 \pm 0.0050 | 0.5852 \pm 0.0042 | 0.2914 \pm 0.0014 | 0.1857 \pm 0.0076 | 0.5804 \pm 0.0057 | 0.3238 \pm 0.0031 | 0.2171 \pm 0.0041 | 0.6158 \pm 0.0043 |
| IntSum | 0.2843 \pm 0.0027 | <u>0.1949</u> \pm 0.0031 | 0.6054 \pm 0.0029 | 0.2958 \pm 0.0015 | <u>0.1956</u> \pm 0.0031 | 0.5986 \pm 0.0032 | 0.3183 \pm 0.0033 | 0.2280 \pm 0.0043 | 0.6284 \pm 0.0040 |
| LLM-TRSR | 0.2826 \pm 0.0021 | 0.1943 \pm 0.0026 | <u>0.6097</u> \pm 0.0032 | 0.2937 \pm 0.0023 | 0.1952 \pm 0.0029 | <u>0.6003</u> \pm 0.0037 | 0.3140 \pm 0.0034 | 0.2250 \pm 0.0034 | <u>0.6292</u> \pm 0.0071 |
| CICL | <u>0.2874</u> \pm 0.0029 | 0.1946 \pm 0.0037 | 0.6048 \pm 0.0039 | <u>0.2986</u> \pm 0.0020 | 0.1946 \pm 0.0048 | 0.5967 \pm 0.0052 | 0.3262 \pm 0.0046 | 0.2339 \pm 0.0033 | 0.6250 \pm 0.0080 |
| Persona-DB | 0.2833 \pm 0.0028 | 0.1930 \pm 0.0044 | 0.6068 \pm 0.0044 | 0.2935 \pm 0.0013 | 0.1912 \pm 0.0045 | 0.5974 \pm 0.0062 | <u>0.3271</u> \pm 0.0040 | <u>0.2368</u> \pm 0.0048 | <u>0.6292</u> \pm 0.0062 |
| DPL (ours) | 0.2940* \pm 0.0022 | 0.1990* \pm 0.0033 | 0.6125* \pm 0.0048 | 0.3072* \pm 0.0024 | 0.2036* \pm 0.0035 | 0.6098* \pm 0.0064 | 0.3318* \pm 0.0047 | 0.2411 \pm 0.0024 | 0.6300 \pm 0.0083 |

| Datasets (→) | Movies & TV | | | CDs & Vinyl | | | Books | | |
|-------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Methods (↓) | ROUGE-L | BLEU | S-GPT | ROUGE-L | BLEU | S-GPT | ROUGE-L | BLEU | S-GPT |
| Non-Perso | 0.1269 \pm 0.0022 | 0.3463 \pm 0.0098 | 0.4132 \pm 0.0046 | 0.1273 \pm 0.0027 | 0.4783 \pm 0.0170 | 0.4215 \pm 0.0011 | 0.1457 \pm 0.0034 | 1.2616 \pm 0.0349 | 0.4642 \pm 0.0033 |
| RAG | 0.1431 \pm 0.0043 | 1.7188 \pm 0.0758 | 0.4336 \pm 0.0137 | 0.1418 \pm 0.0032 | 1.9271 \pm 0.1491 | 0.4456 \pm 0.0079 | 0.1636 \pm 0.0034 | 4.1082 \pm 0.1215 | 0.5077 \pm 0.0066 |
| IntSum | 0.1424 \pm 0.0036 | 2.0619 \pm 0.0355 | 0.4345 \pm 0.0093 | 0.1422 \pm 0.0031 | 2.3621 \pm 0.0775 | <u>0.4482</u> \pm 0.0034 | 0.1614 \pm 0.0030 | 4.6256 \pm 0.2021 | 0.5102 \pm 0.0044 |
| LLM-TRSR | <u>0.1439</u> \pm 0.0039 | <u>2.0834</u> \pm 0.0298 | 0.4350 \pm 0.0077 | 0.1417 \pm 0.0032 | <u>2.3628</u> \pm 0.0538 | 0.4472 \pm 0.0034 | 0.1603 \pm 0.0034 | 4.6268 \pm 0.2724 | 0.5019 \pm 0.0056 |
| CICL | 0.1439 \pm 0.0039 | 2.0576 \pm 0.0795 | 0.4353 \pm 0.0094 | <u>0.1433</u> \pm 0.0033 | 2.2978 \pm 0.1200 | 0.4454 \pm 0.0062 | 0.1664 \pm 0.0041 | 5.1555 \pm 0.2748 | 0.5090 \pm 0.0048 |
| Persona-DB | 0.1425 \pm 0.0037 | 1.9898 \pm 0.0777 | <u>0.4369</u> \pm 0.0082 | 0.1421 \pm 0.0033 | 2.1616 \pm 0.1455 | 0.4467 \pm 0.0069 | <u>0.1672</u> \pm 0.0036 | <u>5.1862</u> \pm 0.2064 | <u>0.5123</u> \pm 0.0085 |
| DPL | 0.1466* \pm 0.0035 | 2.2981* \pm 0.0639 | 0.4437* \pm 0.0057 | 0.1463* \pm 0.0037 | 2.6803* \pm 0.0559 | 0.4556* \pm 0.0052 | 0.1720* \pm 0.0045 | 5.9379* \pm 0.3108 | 0.5224* \pm 0.0091 |

report the mean and standard deviation of the evaluation metrics. We also perform t-tests to assess statistical significance. More details are provided in Appendix D.

4.2 Main Results (RQ1)

We begin by evaluating the overall performance of the compared methods. The summarized results are presented in Table 1, from which the following observations can be drawn:

- DPL demonstrates consistently superior personalization performance across all three datasets and all evaluation metrics. For instance, on the CDs & Vinyl category, DPL achieves the highest scores on all metrics, with an average relative improvement of 4.29% over the best baselines and six metrics showing statistically significant gains. Similar trends are observed on Movies & TV and Books, where DPL also ranks first across all metrics, with 6 and 4 metrics achieving significance, respectively. This result underscores its enhanced personalized text generation capabilities, which can be attributed to its focus on extracting inter-user differences to improve LLM personalization through the selective comparison

mechanism and difference extraction standard.

- Non-Perso exhibits the weakest personalization performance across all experimental settings, which can be attributed to its fundamental limitation in incorporating user-specific contextual signals. In comparison, RAG demonstrates substantial performance gains by leveraging a retrieval approach to integrate user-specific information, thereby enabling more tailored outputs.
- IntSum and LLM-TRSR demonstrate consistent superiority over RAG across all datasets, primarily due to their integration of summarization mechanisms. This is because summarization mechanisms distill the retrieved user history into more concise representations, thereby enhancing the utilization of contextual information.
- Although CICL and Persona-DB demonstrate some performance improvements by utilizing the historical reviews of other relevant users, they still lag behind our DPL. This discrepancy arises because they fail to incorporate inter-user comparative analysis, which is essential for identifying the nuanced differences that genuinely influence user preferences.

Table 2: Results of the ablation study for DPL on Books.

| Methods | R-1 | MET. | S-72B |
|----------------|----------------------------|----------------------------|----------------------------|
| DPL | 0.3318 \pm 0.0047 | 0.2411 \pm 0.0024 | 0.6300 \pm 0.0083 |
| Random | 0.3299 \pm 0.0034 | 0.2379 \pm 0.0049 | 0.6269 \pm 0.0063 |
| SimRank | 0.3299 \pm 0.0037 | 0.2400 \pm 0.0059 | 0.6235 \pm 0.0031 |
| None | 0.3235 \pm 0.0034 | 0.2335 \pm 0.0056 | 0.6149 \pm 0.0028 |
| WriOnly | 0.3303 \pm 0.0043 | 0.2427 \pm 0.0048 | 0.6188 \pm 0.0028 |
| EmoOnly | 0.3238 \pm 0.0045 | 0.2334 \pm 0.0028 | 0.6262 \pm 0.0034 |
| SemOnly | 0.3246 \pm 0.0040 | 0.2387 \pm 0.0052 | 0.6260 \pm 0.0035 |

4.3 Ablation Studies (RQ2)

To substantiate the rationale behind the various design decisions of DPL, we conduct an exhaustive evaluation using the Books dataset by systematically disabling one critical design element to obtain various variants. For the **selective user comparison** in Section 3.2.1, we replace the clustering-based selection method and introduce the following two variants for comparison:

- **Random.** This variant randomly selects users who have reviews on the same retrieved historical item without considering any specific criteria.
- **SimRank.** This variant selects users based on the similarity of historical reviews between the current user and potential candidates. Specifically, the users with the top- K lowest similarity are chosen in order to maximize differentiation.

For the **structured difference extraction** in Section 3.2.2, we modify the original extraction standard, whose dimensions include writing style, emotional style, and semantic style, and introduce the following four variants for comparison:

- **None.** This variant excludes all the dimensions.
- **WriOnly, EmoOnly, SemOnly.** These three variants each incorporate one of the three style dimensions, respectively.

Table 2 illustrates the comparison results on ROUGE-1, METEOR, and S-72B, from which we draw the following observations:

- For the selective user comparison, replacing the clustering-based selection method results in a performance decline. This can be attributed to the fact that random user selection does not rely on any criteria, resulting in insufficient comparative information. Additionally, although SimRank introduces pairwise similarity for enhanced user selection, it fails to capture the global context offered by clustering, leading to a lack of representativeness in the selected users.

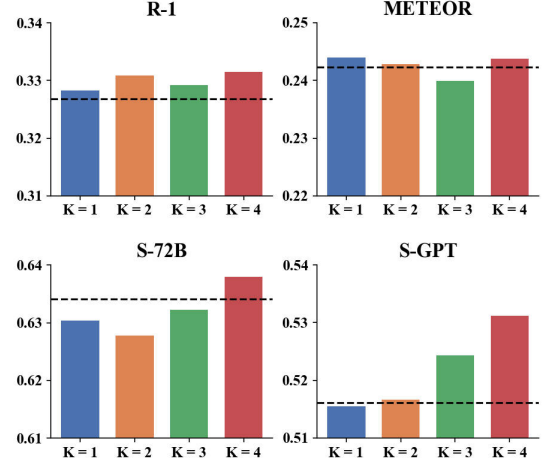


Figure 2: Results of the performance of DPL across different values of K , where K denotes the number of selected representative users. The dashed line represents the corresponding best baseline metric.

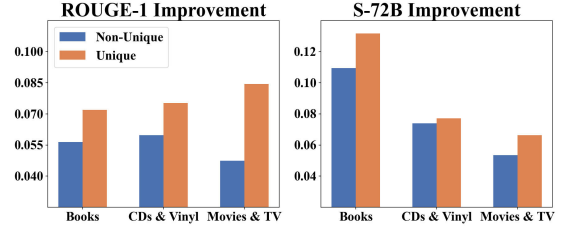


Figure 3: Results of the performance of DPL across different levels of uniqueness. We report the absolute improvement compared to the non-personalization method.

- For the structured difference extraction, removing parts of the extracted dimensions leads to a general performance drop. This finding suggests that each dimension plays a crucial role in capturing different facets of the differences, thereby supporting meaningful and well-structured extraction for distinguishing task-relevant content.

4.4 In-depth Analyses (RQ3 & RQ4)

We next evaluate the performance of DPL across different values of hyperparameters and different levels of user uniqueness.

4.4.1 Hyper-parameter Analysis

In our investigation, the hyperparameter K , which denotes the number of selected representative users for comparison, plays a pivotal role in determining the effectiveness of DPL. We vary K within the set $\{1, 2, 3, 4\}$ to assess the impact of varying these parameters on performance. Additionally, we present the best baseline metrics for comparative analysis, with the results shown in Figure 2,

using a fixed temperature of 0.8. Our findings indicate that, generally, increasing the number of representative users leads to slight to moderate improvements in performance, with the best results occurring at $K = 4$. This underscores that effective difference extraction depends on selecting an appropriate number of comparative representatives.

4.4.2 Impact of User Uniqueness

We initially generate the uniqueness label based on users’ historical review data, as it does not inherently exist. Specifically, we map each user’s historical reviews to embeddings using LLMs and compute the average embedding of all users to establish a center. The Euclidean distance between each user’s embedding and this center is then calculated. Users with distances in the bottom 50% are classified as **non-unique**, while those in the top 50% are classified as **unique**.

Following this classification, we compare the performance of DPL with the non-personalization method across the two groups. For simplicity, we report the absolute improvement of ROUGE-1 and S-72B, as the trends for other metrics are consistent. As illustrated in Figure 3, the unique group shows a significantly greater improvement than the non-unique group. This suggests that the more unique the user, the better we can capture insights into their differences relative to others, aligning with the philosophy that “differences make us unique”.

4.5 Case Study (RQ5)

Finally, we conduct a case study to gain a deeper understanding of DPL’s capacity for personalized review generation. Specifically, we select a representative item from the Movies & TV dataset and compare a single review generated by DPL with that produced by the representative baseline method, LLM-TRSR. The detailed reviews are presented in Figure 4, where we highlight consistent sections between the generated reviews and the user’s real review by distinct colors.

We observe that the review generated by DPL exhibits more consistent sections. Furthermore, it offers a more personalized perspective and concise content, whereas the review generated by LLM-TRSR adopts a more objective stance and contains more tedious information. The alignment of DPL with the user’s writing style and semantic preferences further supports its superiority. More case studies are provided in Appendix H.


| | |
|---|---|
| <p>User’s Real Review: I thought I’d enjoy this movie a little more. These DC movies will never be on the same level as the Marvel movies but I have enjoyed a couple of them. <i>I think the story is a bit too epic.</i> It seems to me that the threat faced by mankind in this movie is greater than the threat in the Justice League movie. It would have been nice if that was toned down a bit and he just had to deal with a villain or two to go along with the origin story, instead of being involved in a giant battle with seemingly millions of participants. <i>For the most part performances were flat. I feel the movie was mostly carried by Jason Momoa’s charisma.</i> It’s nice when a movie has a good villain but I think both of the main villains in this movie kinda sucked. I don’t even know who the guy is that played Black Manta but he should stick to the Syfy channel or other low-budget mediums. If I had the chance to do it over <i>I wouldn’t pay more than \$10 for the Blu-ray; the \$5 bargain bin would be ideal. I might watch this again at some point but it won’t be anytime soon.</i></p> |  <p>[Item Title]: Aquaman (Blu-ray) [Review Rating]: 3.0 [Review Title]: A little too epic</p> |
| <p>DPL Generated Review: 🌟 Aquaman is an action-packed flick with some pretty visuals, but <i>I felt like it went a bit too far into the realm of grandiose.</i> The underwater world looks amazing, but I found the story to be a bit too convoluted at times. <i>Jason Momoa brings some personality to the role,</i> but the rest of the cast didn’t really leave an impression on me. It’s one of those movies where you watch it and think, “<i>That was okay, I guess,</i>” but <i>it’s not something I’ll be rushing to watch again anytime soon. I’m still not convinced Aquaman was the best choice for the DC cinematic universe,</i> but I’ll be interested to see if they can improve in the sequel. <i>[The sections highlighted in color demonstrate that the DPL-generated review aligns well with the user’s real review, accurately reflecting user’s writing style, emotional style, and semantic style.]</i></p> | |
| <p>LLM-TRSR Generated Review: Aquaman is an ambitious film with grand visuals and a scale befitting its underwater kingdom, but it ultimately feels like it’s trying too hard to be a blockbuster. Ryan Reynolds’ Deadpool might have introduced a new level of humor and irreverence to superhero films, and while Aquaman tries to bring a similar charm to the table, it often falls short. The story is engaging, setting Arthur Curry’s path to becoming the king of Atlantis, but it’s crowded with numerous subplots and characters that could have been trimmed down. The visuals of Atlantis are breathtaking, and the underwater action sequences are a treat to watch, but the film’s pacing is uneven, and there are moments where it feels like the action is drowning in its own grandiosity. <i>Jason Momoa’s portrayal of Aquaman is solid;</i> he brings a certain charisma to the role, and the chemistry between him and Amber Heard as Mera is mostly enjoyable. However, the supporting cast doesn’t always gel as well as it could, and the villain, Ocean Master, lacks the depth and menace needed to carry the film. While I enjoyed the spectacle and the nods to the comics, I couldn’t help but feel that the movie could have been tighter and more focused. It’s a visually stunning film that tries to pack in as much as possible, but at times it gets lost in its own ambition. If you’re a fan of superhero films with a touch of epic adventure, you’ll likely still find it enjoyable, but it’s not without its flaws. <i>[As highlighted in color, the LLM-TRSR-generated review only partially matches the user’s real review, failing to reflect the user’s writing style and semantic style.]</i></p> | |

Figure 4: Comparison of item reviews generated by DPL and the baseline method LLM-TRSR, along with the user’s real review. The consistent sections between the generated reviews and the user’s real review are highlighted by specific colors.

5 Related Work

With the widespread use of LLMs (Zhang et al., 2024b; Li et al., 2025; Zhao et al., 2025b; Liu et al., 2025b; Du et al., 2025; Fang et al., 2025), there is a growing need for personalized responses tailored to individual users. LLM personalization has been explored across various domains, underscoring its importance (Zhao et al., 2025c,a; Zhang et al., 2025a; Cai et al., 2025). While existing works advance personalization in different scenarios, such as recommendation (Chen et al., 2024b; Bao et al., 2024; Zhang et al., 2024a) and image generation (Shen et al., 2024; Xu et al., 2025a,c), personalized text generation remains the foundation of LLM personalization (Zhang et al., 2024d). LaMP (Salemi et al., 2024b) evaluates short-form personalized text, later extended by LongLaMP (Kumar et al., 2024) to assess long-form text generation. Among these, personalized item review writing stands out

as a key task (Peng et al., 2024; Au et al., 2025) since user reviews naturally reflect individual preferences (Ni et al., 2019), serving as a critical scenario to evaluate a model’s ability to generate personalized content.

Recent LLM-based methods enhance personalization using user-specific content, which can be categorized into: 1) retrieval-based methods, where ROPG (Salemi et al., 2024a) introduces a retrieval selection model to dynamically choose the best retriever and HYDRA (Zhuang et al., 2024) employs a personalized reranker to prioritize useful information; and 2) summarization-based methods, where ONCE (Liu et al., 2024b) generates user profiles by summarizing topics and regions of interest in their browsing history and PPlug (Liu et al., 2024a) converts user history into personalized embeddings.

Besides, a recent study Persona-DB (Sun et al., 2025) uses data from other users to enhance personalized text generation. However, this method mainly focuses on enriching relevant knowledge rather than capturing each user’s uniqueness, key to effective personalization. To leverage inter-user data, contrastive learning (Gao and Das, 2024) offers a promising way of using both positive and negative examples to better describe intents. However, its performance relies on meaningful negative examples, which are challenging to obtain in user-centered LLM personalization. This reliance limits the full leverage of inter-user information for LLM personalization. To the best of our ability, our proposed DPL addresses this limitation.

6 Conclusions

This study introduced DPL, a novel method that enhances LLM personalization by identifying and leveraging inter-user differences as key preference information. We proposed a selective user comparison mechanism and a structured difference extraction method to extract task-relevant meaningful differences effectively. Experimental results on the personalized review generation task show that DPL significantly enhances LLM personalization. Currently, we only focus on review generation tasks to validate our method. In future, we plan to extend it to other tasks for broader verification. Additionally, our method’s effectiveness depends on the long-context processing capabilities of LLMs, and we currently assume that the context window is sufficient for handling our data. We also plan to extend our method to address long-context cases.

Limitations

In this paper, we utilize LLMs to extract inter-user differences, summarize difference-aware representations, and generate personalized reviews. These processes require a highly powerful model to effectively analyze users’ personalized information, ensuring that the generated reviews align closely with individual user preferences. Moreover, in this task, we specifically focus on three key aspects: writing style, emotional style, and semantic style. Our findings indicate that these factors significantly enhance the quality of personalized reviews. However, in other tasks, these standards may not be as suitable, and additional factors, such as personality traits, might need to be considered. Our approach involves comparing differences between users, which inevitably introduces additional computational overhead and time costs. We believe this challenge can be mitigated by precomputing and storing difference-aware representations in user memory. These representations can then be directly retrieved during online deployment, with periodic updates ensuring both efficiency and relevance.

Ethical Statements

Our method presents the following potential ethical concerns that need to be addressed carefully during deployment. First, it involves the storage and retrieval of users’ historical interaction data, which necessitates the implementation of strict safety protocols and ethical guidelines. Second, as we leverage inter-user differences to enhance personalization, sharing and integrating information between users raises significant privacy concerns. This process must incorporate robust privacy safeguards, including data anonymization, secure storage, and explicit user consent. Furthermore, the ability to create detailed user profiles for personalization purposes introduces the risk of potential misuse, such as surveillance or targeted manipulation. To mitigate these risks, developers bear the responsibility to prioritize privacy-preserving techniques and ensure that user data is handled with the utmost integrity and respect.

In our research, we use an open-source dataset, ensuring full compliance with the original dataset’s MIT license. Our research prioritizes data integrity, privacy, and ethical considerations to uphold the responsible and fair use of open-source resources.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Steven Au, Cameron J Dimacali, Ojasmitha Pedirappagari, Namyong Park, Franck Dernoncourt, Yu Wang, Nikos Kanakaris, Hanieh Deilamsalehy, Ryan A Rossi, and Nesreen K Ahmed. 2025. Personalized graph-based retrieval for large language models. *arXiv preprint arXiv:2501.02157*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *IEEEvaluation@ACL*, pages 65–72. Association for Computational Linguistics.
- Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. 2024. Decoding matters: Addressing amplification bias and homogeneity issue in recommendations for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10540–10552. Association for Computational Linguistics.
- Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *RecSys*, pages 1007–1014. ACM.
- Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. 2025. Large language models empowered personalized web agents. In *Proceedings of the ACM on Web Conference 2025*, pages 198–215.
- Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024a. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.
- Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2025. PAD: personalized alignment of llms at decoding-time. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024b. On softmax direct preference optimization for recommendation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Chucuri, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, et al. 2023. Large language models for user interest journeys. *arXiv preprint arXiv:2305.15498*.
- Yang Deng, Xuan Zhang, Wenxuan Zhang, Yifei Yuan, See-Kiong Ng, and Tat-Seng Chua. 2024. On the multi-turn instruction following for conversational web agents. In *ACL (1)*, pages 8795–8812. Association for Computational Linguistics.
- Junjia Du, Yadi Liu, Hongcheng Guo, Jiawei Wang, Haojian Huang, Yunyi Ni, and Zhoujun Li. 2025. Dependeval: Benchmarking llms for repository dependency understanding. *arXiv preprint arXiv:2503.06689*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Xiang Gao and Kamalika Das. 2024. Customizing language model responses with contrastive in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18039–18046.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- 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*.
- Caglar Irmak, Beth Vallen, and Sankar Sen. 2010. You like what i like, but i don’t like what you like: Uniqueness motivations in product preferences. *Journal of Consumer Research*, 37(3):443–455.
- Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*.
- Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pages 611–626. ACM.
- Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023a. Teach llms to personalize—an approach inspired by writing education. *arXiv preprint arXiv:2308.07968*.
- Moxin Li, Yuntao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. 2025. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. *arXiv preprint arXiv:2502.14354*.
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024a. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024b. Leveraging large language models for NLG evaluation: Advances and challenges. In *EMNLP*, pages 16028–16045. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025a. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.
- Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2024a. Llm+ persona-plug= personalized llms. *arXiv preprint arXiv:2409.11901*.
- Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024b. ONCE: boosting content-based recommendation with both open- and closed-source large language models. In *WSDM*, pages 452–461. ACM.
- Xiaohao Liu, Xiaobo Xia, Weixiang Zhao, Manyi Zhang, Xianzhi Yu, Xiu Su, Shuo Yang, See-Kiong Ng, and Tat-Seng Chua. 2025b. L-mtp: Leap multi-token prediction beyond adjacent context for large language models. *arXiv preprint arXiv:2505.17505*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*, pages 2511–2522. Association for Computational Linguistics.
- J MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.
- Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 198–219. Association for Computational Linguistics.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.
- 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*.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*, pages 186–191. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics.
- Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.
- Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.
- Ishan Sahu and Debapriyo Majumdar. 2017. Detecting factual and non-factual content in news articles. In *CODS*, pages 17:1–17:12. ACM.

- Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation. In *SIGIR*, pages 752–762. ACM.
- Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, Tao Chen, Zhuowan Li, Michael Bendersky, and Hamed Zamani. 2025. Reasoning-enhanced self-training for long-form personalized text generation. *arXiv preprint arXiv:2501.04167*.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. LaMP: When large language models meet personalization. In *ACL (1)*, pages 7370–7392. Association for Computational Linguistics.
- Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. PMG : Personalized multimodal generation with large language models. In *WWW*, pages 3833–3843. ACM.
- Charles R Snyder and Howard L Fromkin. 1977. Abnormality as a positive characteristic: The development and validation of a scale measuring need for uniqueness. *Journal of Abnormal Psychology*, 86(5):518.
- Charles R Snyder and Howard L Fromkin. 2012. *Uniqueness: The human pursuit of difference*. Springer Science & Business Media.
- Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Ren Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. In *COLING*, pages 281–296. Association for Computational Linguistics.
- Gaurav Verma and Balaji Vasan Srinivasan. 2019. A lexical, syntactic, and semantic perspective for understanding style in text. *arXiv preprint arXiv:1909.08349*.
- Chengbing Wang, Yang Zhang, Fengbin Zhu, Jizhi Zhang, Tianhao Shi, and Fuli Feng. 2024. Leveraging memory retrieval to enhance llm-based generative recommendation. *arXiv preprint arXiv:2412.17593*.
- Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. 2025a. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, pages 264–274.
- Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025b. Personalized generation in large model era: A survey. *arXiv preprint arXiv:2503.02614*.
- Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. 2025c. Drc: Enhancing personalized image generation via disentangled representation composition. *arXiv preprint arXiv:2504.17349*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024a. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pages 1807–1817.
- Xiaotian Zhang, Ruizhe Chen, Yang Feng, and Zuozhu Liu. 2025a. Persona-judge: Personalized alignment of large language models via token-level self-judgment. *arXiv preprint arXiv:2504.12663*.
- Yang Zhang, Keqin Bao, Ming Yan, Wenjie Wang, Fuli Feng, and Xiangnan He. 2024b. Text-like encoding of collaborative information in large language models for recommendation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9181–9191. Association for Computational Linguistics.
- Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. 2025b. Collm: Integrating collaborative embeddings into large language models for recommendation. *IEEE Trans. Knowl. Data Eng.*, 37(5):2329–2340.
- Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024c. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.
- Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024d. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.
- Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025a. Do llms recognize your preferences? evaluating personalized preference following in llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Weixiang Zhao, Yulin Hu, Yang Deng, Jiahe Guo, Xingyu Sui, Xinyang Han, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, et al. 2025b. Beware of your po! measuring and mitigating ai safety

risks in role-play fine-tuning of llms. *arXiv preprint arXiv:2502.20968*.

Weixiang Zhao, Xingyu Sui, Xinyang Han, Yang Deng, Yulin Hu, Jiahe Guo, Libo Qin, Qianyun Du, Shijin Wang, Yanyan Zhao, et al. 2025c. Chain of strategy optimization makes large language models better emotional supporter. *arXiv preprint arXiv:2503.05362*.

Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. 2025d. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment. *arXiv preprint arXiv:2505.15456*.

Zhi Zheng, Wenshuo Chao, Zhaopeng Qiu, Hengshu Zhu, and Hui Xiong. 2024. Harnessing large language models for text-rich sequential recommendation. In *WWW*, pages 3207–3216. ACM.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. HYDRA: model factorization framework for black-box LLM personalization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

A Dataset Details

In this paper, we focus on item review text generation, which is an effective task to evaluate the model’s ability to generate personalized content. To generate data samples, we employ the open-source Amazon Reviews Dataset (Hou et al., 2024) and primarily follow the setting of LongLaMP (Kumar et al., 2024), with several modifications to better align with our experimental framework.

Dataset Introduction. The Amazon Reviews dataset is processed by merging the user reviews subset with the item metadata subset using the *asin* field as the shared field. Specifically, the *asin* value from the user reviews subset is matched to the corresponding *parent_asin* field in the item metadata subset. From the user reviews subset, we extract *asin*, which is the unique identifier for the item; *user_id*, representing the unique identifier for the reviewer; *title*, which denotes the title of the review; *text*, containing the actual review content; *rating*, indicating the numerical rating given by the user; and *timestamp*, specifying the time when the review was submitted. From the item metadata subset, we select *parent_asin*, which serves as a reference to the *asin* field in the user reviews subset, linking reviews to items; *title*, representing the title of the item; and *description*, providing a textual summary

```
{
  "user_id": "AFZUK3MTBIBEDQOPAK3OATUOUKLA",
  "profile": [
    {
      "asin": "B000056HE9",
      "rating": 5.0,
      "text": "This movie is one of my all time holiday favorites..."
    },
    ...
  ],
  "data": {
    "asin": "B0002F6BRE",
    "title": "Elf movie",
    "timestamp": 1607398123747,
    "rating": 5.0,
    "text": "I purchase this movie..."
  }
}

{
  "asin": "B071GVGD2B",
  "title": "Transformers: The Last Knight",
  "description": "Our world's greatest hero becomes our fiercest..."
}
```

Figure 5: Examples of the main experimental dataset (above) and the item dataset (below).

Table 3: Dataset statistics of three categories in our processed Amazon dataset.

| Categories (↓) | | #data | Profile Size | Output Length |
|-------------------------------|-------|-------|--------------|-----------------|
| Movies & TV (4832) | Train | 15400 | 35.61±35.97 | 1691.42±1617.71 |
| | Val | 1925 | 40.11±35.89 | 1668.26±1612.17 |
| | Test | 1925 | 41.11±35.89 | 1704.61±1751.98 |
| CDs & Vinyl (3801) | Train | 14032 | 33.00±32.44 | 1485.19±1292.09 |
| | Val | 1754 | 37.50±32.36 | 1602.93±1431.09 |
| | Test | 1754 | 38.50±32.36 | 1600.04±1419.49 |
| Books (839) | Train | 2536 | 29.34±22.63 | 1295.77±827.48 |
| | Val | 317 | 33.84±22.51 | 1216.49±768.44 |
| | Test | 317 | 34.84±22.51 | 1194.90±801.18 |

of the item’s features. To distinguish between the review title and the item title, we use *review_title* for the former and *item_title* for the latter.

Data Curation. We perform a detailed data-cleaning process to create a high-quality benchmark. First, we ensure that each review includes a corresponding *review_title*, *rating*, and *timestamp*, with a minimum *text* length of 200 characters to meet the long-form criteria. For each item, we verify that it has an *item_title* and that its *description* falls within the 100 to 2000-character range to maintain high-quality content. Following this, we ensure that each product has at least four unique reviewers and retain only the most recent review from each reviewer for a given item. Additionally, we restrict the number of reviews per reviewer to between 18 and 500 to filter out users who may engage in review manipulation or have insufficient data for meaningful personalization. To meet the

Table 4: Comparison of different baseline methods.

| Method | Training-Free | Current User’s Data | Other Users’ Data | Difference-Aware |
|------------|---------------|---------------------|-------------------|------------------|
| Non-Perso | ✓ | ✗ | ✗ | ✗ |
| RAG | ✓ | ✓ | ✗ | ✗ |
| IntSum | ✓ | ✓ | ✗ | ✗ |
| LLM-TRSR | ✓ | ✓ | ✗ | ✗ |
| CICL | ✓ | ✓ | ✓ | ✗ |
| Persona-DB | ✓ | ✓ | ✓ | ✗ |
| DPL | ✓ | ✓ | ✓ | ✓ |

requirements of our experimental setup, we further ensure that each user in the validation and test sets has at least 8 historical records, where each record corresponds to an item that has been reviewed by at least 5 users, including the target user. This constraint guarantees that we can retrieve a sufficient number of other users for computing difference-aware representations during evaluation. These settings are designed to align with our experiments.

Dataset Construction. After preprocessing, we proceed to construct the main experimental dataset. Here we follow the Temporal Setting used in LongLaMP, treating each user as a data point. However, unlike LongLaMP, we take a different approach when constructing the training dataset. Specifically, we use a user’s 10th most recent to 3rd most recent reviews as training samples, with their profile information derived from historical reviews relative to the target review. The validation dataset consists of the user’s 2nd most recent review, while the test dataset includes the 3rd most recent review. Each user’s historical information samples include only the item *asin*, *rating*, *timestamp*, *review_title*, and *text*. The *asin* references the pre-filtered item dataset, allowing us to extract the corresponding item information. Examples of the main experimental dataset and the item dataset are illustrated in Figure 5. We process the three categories: Books, Movies & TV, and CDs & Vinyl separately. The statistics for these categories after dataset construction are presented in Table 3.

Dataset Usage. In our experiments, we utilize only the test dataset for a direct evaluation of model performance. The details of model inputs are provided in Appendix G. Notably, the training and validation datasets are included to support related research.

B Baselines Details

In this section, we further introduce each baseline method in detail. Additionally, the comparison between different baseline methods and our proposed DPL is shown in Table 4.

- **Non-Perso:** This method does not utilize any user-specific information to assist in personalized review generation. The model’s input consists only of the item’s title and description, along with the rating and title of the review to be generated.
- **RAG:** This method employs a retriever, specifically BM25 in our experiments, to retrieve relevant past reviews from the user’s review history. These retrieved reviews are then formatted and incorporated into the LLM’s input prompt.
- **IntSum:** Unlike the original implementation of generating a summary offline and directly using it in downstream tasks, our scenario of long-form text review generation also involves long-form user history. Due to the input length limitations of LLMs, we cannot feed all of a user’s historical reviews into the model at once to produce a sensible user profile summary. Therefore, we explore an alternative solution: instead of generating a static summary beforehand, the model dynamically creates a summary based on the retrieved user’s historical reviews. This summary is then integrated into the model’s input prompt along with retrieved historical reviews. The experimental results demonstrate that our implementation is correct.
- **LLM-TRSR:** To address the issue that our IntSum experiment does not utilize all of the user’s historical data, we select the LLM-TRSR method as a baseline. This method divides the user’s history into blocks and introduces two distinct preference summarization paradigms: hi-

erarchical summarization and recurrent summarization. We opt for the recurrent summarization technique due to its superior performance and better suitability for the temporal setting of our dataset. In this method, a summary is first generated based on the initial history block and then iteratively updated using more recent history blocks. The final summary is then integrated into the model’s input along with the retrieved user’s historical reviews. In our experiments, each history block contains eight reviews.

- **CICL:** This is the Contrastive In-Context Learning method. While originally not designed for LLM personalization, we adapt it to our experimental scenario. In our experiments, we treat the current user’s review as the positive example and a review from another user as the negative example. To construct these comparisons, we use the BM25 retriever to retrieve a historical review from the current user and randomly select a review on the same item from another user. These comparative reviews, along with the current user’s past reviews, are incorporated into the prompt for an LLM summarizer, thereby improving the quality of summary generation.
- **Persona-DB:** This method introduces a collaborative refinement approach that allows a user to retrieve information from relevant peers to enrich knowledge. We adapt this approach to our task. For selecting relevant users, we employ an embedding-based strategy using the gte-Qwen2-1.5B-Instruct embedding model to encode users’ historical information and identify the most similar user to the current user. When generating the review, we first create a summary based on the current user’s retrieved historical reviews. Simultaneously, we extract the same number of reviews from the most similar user’s history. The current user’s profile summary, historical reviews, and the most similar user’s historical reviews are then integrated into the prompt for the LLM generator.

All baseline methods, including our approach, use prompts that are uniformly designed for our review generation task.

C Evaluation Metrics

In this paper, we consider the ROUGE-1, METEOR, S-72B, and S-GPT metrics in the primary experiments. We offer detailed descriptions of each metric as follows.

Template

You are an impartial evaluator tasked with assessing how well an AI-generated item review is personalized for a specific user. Based on the provided item title and description, the user’s rating for this item, and the review title, assign a score to the AI-generated review by comparing it to the user’s real review.

[Scoring Criteria]:

[Score 0]: The AI-generated review is completely unrelated to the user’s real review. It does not describe the item correctly or reflects the user’s personal thoughts and preferences.

[Score 2]: The AI-generated review has a weak connection to the user’s real review. It may mention the item but does not include the key points or personal thoughts that the user included.

[Score 4]: The AI-generated review partially matches the user’s real review. Some key points are included, but important details are missing, and the review may feel too general or not personal enough.

[Score 6]: The AI-generated review mostly matches the user’s real review. It covers the main points but may miss some details or personal thoughts that the user included.

[Score 8]: The AI-generated review is very similar to the user’s real review. It captures the user’s thoughts and preferences well, with only small differences.

[Score 10]: The AI-generated review is almost the same as the user’s real review. It includes all key details, personal thoughts, and preferences exactly as the user expressed them.

Output the numerical score only.

Input

[Item Title]: *<item title>*

[Item Description]: *<item description>*

[Review Rating]: *<review rating>*

[Review Title]: *<review title>*

[User’s real review]: *<user’s real review>*

[AI-generated review]: *<AI-generated review>*

[Score]:

Figure 6: The prompt used for the evaluator LLM to evaluate the generated text based on the reference text and the provided criteria. The italicized text enclosed in <> represents placeholders for the actual values of variables. These representations are consistent across the prompt figures.

- **ROUGE-1:** ROUGE-1 (Recall-Oriented Understudy for Gisting Evaluation) is a metric that measures the overlap of unigrams (individual words) between the generated text and the reference text.
- **ROUGE-L:** ROUGE-L is a metric that evaluates text similarity by measuring the longest common subsequence (LCS) between the generated text and the reference text.
- **METEOR:** METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric that evaluates text similarity by considering exact word matches, stemming, synonyms, and word order between the generated text and the reference text.

- **BLEU:** BLEU (Bilingual Evaluation Understudy) is a metric that evaluates text similarity by measuring the precision of n-gram matches between the generated text and the reference text. In our experiments, we use the SacreBLEU³ (Post, 2018) library, which is a standardized version of BLEU. The higher the BLEU score, the more similar the generated text is to the reference.
- **S-72B:** S-72B is a metric that assesses the quality of generated text by comparing it to a reference text using the *Qwen-2.5-72B-Instruct-AWQ* model. We design the prompt following previous works (Liu et al., 2023; Li et al., 2024b; Salemi et al., 2025) and make some reasonable modifications to better suit our task. In this evaluation, the prompt to the model includes the generated review, the user’s real review as the reference, as well as relevant metadata such as the review title, rating, and item details, including its title and description. To ensure robustness, we run the scoring process with two different random seeds and report the average of their results. We also perform scoring inference using the vLLM library, with the temperature set to 0.8 and top_p set to 0.95. The prompt is depicted in Figure C.
- **S-GPT:** S-GPT is another metric that utilizes the *GPT-4o-mini* LLM as the evaluator. It adopts the same temperature, top_p, and prompt settings as S-72B. The final S-GPT score is calculated as the average of five evaluations.

D Implementation Details

D.1 Running Environments

We utilize the vLLM⁴ library (Kwon et al., 2023) as the inference engine for LLMs in generating and the Sentence-Transformers⁵ (Reimers and Gurevych, 2019) library for LLMs in embedding in a Python 3.11.11 environment. All experiments are conducted on a single NVIDIA H100 NVL GPU with 96GB of GPU memory.

D.2 Hyperparameter Configurations

We configure the model with a maximum output length of 2048 tokens while keeping the input length unrestricted. Following the settings in LongLaMP, we set the parameter top_p to 0.95. These settings remain consistent across all experiments.

³<https://github.com/mjpost/sacrebleu>

⁴<https://github.com/vllm-project/vllm>

⁵<https://www.sbert.net/>

E Efficiency Analysis

In this section, we briefly discuss the efficiency of our proposed DPL.

We evaluate the cost of selecting comparison users using the K-means method with pre-computed review embeddings. Since the number of involved users is small, the selection process takes only 0.002s per key historical review on average. In our main experiments, we use 8 key historical reviews, resulting in an additional cost of approximately 0.016s per generation, which is generally acceptable for each time generation.

Additionally, since our method uses a fixed number of comparison users for difference extraction, the computational cost of most components remains stable as the data size increases. The only exception is the comparison user selection process, which uses K-means clustering ($K = 5$). The cost of this step depends on the number of users who have interacted with a given item (denoted as $u' = \#interactions / \#items \leq \#users$), with a complexity of $O(u' \times K \times t) \approx O(u')$, where t is the number of iterations. Importantly, this step is lightweight—clustering 10,000 samples takes under 0.3 seconds in our setup—compared to LLM inference. Thus, the additional overhead is minimal, and our method scales comparably to baseline approaches.

F Additional Experiments

F.1 Yelp Dataset

To further validate the generalizability and robustness of our proposed method, we conduct additional experiments on the Yelp⁶ dataset. This dataset presents a different domain from our primary evaluation benchmarks, allowing us to assess whether the observed performance gains extend to a broader range of real-world scenarios.

Dataset Construction. The Yelp Open Dataset is a large-scale, publicly available collection of user-generated reviews, primarily covering local businesses such as restaurants, shopping centers, and service providers across various cities. It contains rich information, including user reviews, ratings, business metadata, and timestamps, making it a valuable resource for studying the personalized review generation task. Following the same protocol as applied to the Amazon dataset, we perform data filtering and cleaning to ensure consistency.

⁶<https://business.yelp.com/data/resources/open-dataset>

Table 5: Dataset statistics of our processed Yelp dataset.

| Splits (\downarrow) | #data | Profile Size | Output Length |
|-------------------------|--------|-------------------|---------------------|
| Train | 229944 | 38.56 \pm 41.75 | 681.81 \pm 394.80 |
| Val | 28743 | 43.06 \pm 41.69 | 672.05 \pm 394.69 |
| Test | 28743 | 44.06 \pm 41.69 | 659.71 \pm 390.20 |

Table 6: Experimental results on the Yelp dataset.

| Methods (\downarrow) | R-1 | R-L | METEOR | BLEU |
|--------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Non-Perso | 0.2272 \pm 0.0052 | 0.1286 \pm 0.0036 | 0.1514 \pm 0.0043 | 0.6242 \pm 0.1796 |
| RAG | 0.2581 \pm 0.0046 | 0.1372 \pm 0.0024 | 0.2012 \pm 0.0040 | 1.3549 \pm 0.3243 |
| IntSum | 0.2590 \pm 0.0052 | 0.1367 \pm 0.0025 | 0.2154 \pm 0.0025 | 1.3749 \pm 0.1780 |
| LLM-TRSR | 0.2216 \pm 0.0037 | 0.1217 \pm 0.0021 | 0.1892 \pm 0.0036 | 0.7055 \pm 0.1175 |
| CICL | 0.2627 \pm 0.0080 | 0.1377 \pm 0.0037 | 0.2174 \pm 0.0031 | 1.4118 \pm 0.2040 |
| Persona-DB | 0.2570 \pm 0.0070 | 0.1359 \pm 0.0027 | 0.2122 \pm 0.0032 | 1.3459 \pm 0.1479 |
| DPL | 0.2667* \pm 0.0065 | 0.1422* \pm 0.0027 | 0.2226* \pm 0.0048 | 1.5572* \pm 0.2265 |

In particular, we retain several key fields: *user_id*, which identifies the user who wrote the review; *business_id*, which specifies the reviewed business; *business_name*, the name of the business; *stars*, the rating score of the review; *text*, which contains the full content of the review; and *date*, which we represent by *timestamp*. The statistics of the processed Yelp dataset is shown in Table 5. Our processed Yelp dataset is also available on Huggingface⁷.

Implementation Details. We largely follow the same implementation setup as in the main experiments, except that the item description is removed from the prompt. N in Equation (3) is set to 8, and K in Equation (4) is set to 4. Due to the large scale of the original dataset, we randomly sample 100 data points from the full set in each run to save computational resources and time while maintaining reliability. This sampling process is repeated five times, and we report the mean and standard deviation of the results across the five runs.

Experimental Results. We report the results on conventional metrics here and also perform significance testing to assess the statistical reliability of the observed differences. The results are shown in Table 6, from which we observe that our DPL method consistently achieves the best performance. Moreover, it demonstrates statistically significant improvements over all baselines across the four evaluation metrics.

⁷<https://huggingface.co/datasets/SnowCharmQ/DPL-Yelp>

Table 7: Experimental results on the Books category using a 32B model.

| Methods (\downarrow) | R-1 | METEOR | S-72B |
|--------------------------|----------------------------|----------------------------|----------------------------|
| Non-Perso | 0.3059 \pm 0.0029 | 0.1958 \pm 0.0025 | 0.6159 \pm 0.0125 |
| RAG | 0.3337 \pm 0.0044 | 0.2755 \pm 0.0022 | 0.6288 \pm 0.0046 |
| IntSum | 0.3265 \pm 0.0039 | 0.2825 \pm 0.0020 | 0.6260 \pm 0.0045 |
| LLM-TRSR | 0.3199 \pm 0.0046 | 0.2794 \pm 0.0029 | 0.6273 \pm 0.0068 |
| CICL | 0.3247 \pm 0.0056 | 0.2737 \pm 0.0031 | 0.6274 \pm 0.0032 |
| Persona-DB | 0.3263 \pm 0.0051 | 0.2771 \pm 0.0013 | 0.6225 \pm 0.0062 |
| DPL | 0.3338 \pm 0.0044 | 0.2828 \pm 0.0026 | 0.6312 \pm 0.0088 |

F.2 Model of Different Size

To further investigate the generalization capability of our approach, we conduct experiments to evaluate its performance across models of different sizes. Specifically, we consider the *Qwen2.5-32B-Instruct* model. For each method, we utilize 8 retrieved historical data points. To reduce resource costs, we perform our experiments on the Books category only and report the results in ROUGE-1, METEOR, and S-72B metrics. The experimental results are shown in Table 7, from which the following observations can be drawn:

- When using the 32B model, its more powerful capabilities lead to performance improvements across almost all methods compared to the 14B model. However, DPL still achieves state-of-the-art performance. This demonstrates its strong generalization capability and effectiveness in the personalized review generation task.
- CICL and Persona-DB show performance drops on certain metrics, with some even underperforming compared to RAG, IntSum, and LLM-TRSR. This suggests that simply leveraging other users’ information as auxiliary data is not particularly effective in enhancing the model’s personalization capabilities, emphasizing the idea that “differences make us unique”.

F.3 Number of Retrieved Reviews

We conducted experiments to explore the impact of different numbers of retrieved reviews on performance. Specifically, we evaluated cases with 1, 2, 4, and 8 retrieved reviews. To save resources and time, we only report the results with metrics ROUGE-1, METEOR, and S-72B here. The experimental results, reported from a single run with decoding temperature set to 0.8, are presented in Table 8, leading to the following observations:

Table 8: Experimental results across the three categories of the processed Amazon dataset with varying numbers of retrieved reviews {1,2,4,8}.

| Datasets (→) | Movies & TV | | | CDs & Vinyl | | | Books | | |
|-------------------|-------------|--------|--------|-------------|--------|--------|--------|--------|--------|
| Methods (↓) | R-1 | MET. | S-72B | R-1 | MET. | S-72B | R-1 | MET. | S-72B |
| Non-Perso | 0.2275 | 0.1304 | 0.5406 | 0.2392 | 0.1336 | 0.5379 | 0.2683 | 0.1512 | 0.5539 |
| RAG | N=1 | 0.2540 | 0.1540 | 0.4858 | 0.2636 | 0.1546 | 0.5082 | 0.3007 | 0.1784 |
| | N=2 | 0.2670 | 0.1658 | 0.5315 | 0.2768 | 0.1660 | 0.5434 | 0.3158 | 0.1971 |
| | N=4 | 0.2802 | 0.1794 | 0.5617 | 0.2884 | 0.1787 | 0.5705 | 0.3211 | 0.2116 |
| | N=8 | 0.2845 | 0.1897 | 0.5929 | 0.2926 | 0.1876 | 0.5933 | 0.3206 | 0.2173 |
| IntSum | N=1 | 0.2643 | 0.1682 | 0.4618 | 0.2735 | 0.1671 | 0.4582 | 0.3155 | 0.2030 |
| | N=2 | 0.2757 | 0.1807 | 0.5305 | 0.2865 | 0.1809 | 0.5466 | 0.3179 | 0.2144 |
| | N=4 | 0.2835 | 0.1916 | 0.5752 | 0.2938 | 0.1921 | 0.5848 | 0.3167 | 0.2256 |
| | N=8 | 0.2845 | 0.1972 | 0.6073 | 0.2949 | 0.1968 | 0.6014 | 0.3201 | 0.2340 |
| LLM-TRSR | N=1 | 0.2664 | 0.1728 | 0.5578 | 0.2811 | 0.1756 | 0.5602 | 0.3090 | 0.2015 |
| | N=2 | 0.2760 | 0.1838 | 0.5816 | 0.2876 | 0.1835 | 0.5770 | 0.3125 | 0.2137 |
| | N=4 | 0.2808 | 0.1915 | 0.5942 | 0.2942 | 0.1941 | 0.5981 | 0.3169 | 0.2269 |
| | N=8 | 0.2822 | 0.1955 | 0.6178 | 0.2937 | 0.1969 | 0.6083 | 0.3143 | 0.2294 |
| CICL | N=1 | 0.2658 | 0.1682 | 0.4538 | 0.2737 | 0.1656 | 0.4575 | 0.3124 | 0.1977 |
| | N=2 | 0.2759 | 0.1787 | 0.5242 | 0.2883 | 0.1797 | 0.5306 | 0.3230 | 0.2176 |
| | N=4 | 0.2836 | 0.1881 | 0.5731 | 0.2949 | 0.1890 | 0.5732 | 0.3211 | 0.2278 |
| | N=8 | 0.2873 | 0.1969 | 0.6142 | 0.3010 | 0.1987 | 0.5990 | 0.3255 | 0.2362 |
| Persona-DB | N=1 | 0.2653 | 0.1658 | 0.4695 | 0.2719 | 0.1639 | 0.4637 | 0.3162 | 0.2000 |
| | N=2 | 0.2754 | 0.1786 | 0.5407 | 0.2843 | 0.1761 | 0.5479 | 0.3212 | 0.2161 |
| | N=4 | 0.2819 | 0.1882 | 0.5822 | 0.2899 | 0.1861 | 0.5887 | 0.3267 | 0.2336 |
| | N=8 | 0.2840 | 0.1954 | 0.6170 | 0.2941 | 0.1937 | 0.6048 | 0.3268 | 0.2423 |
| DPL | N=1 | 0.2753 | 0.1747 | 0.5015 | 0.2859 | 0.1765 | 0.5158 | 0.3164 | 0.1989 |
| | N=2 | 0.2858 | 0.1877 | 0.5630 | 0.2997 | 0.1917 | 0.5704 | 0.3202 | 0.2175 |
| | N=4 | 0.2907 | 0.1949 | 0.5854 | 0.3071 | 0.2014 | 0.5958 | 0.3269 | 0.2311 |
| | N=8 | 0.2938 | 0.2009 | 0.6245 | 0.3066 | 0.2058 | 0.6236 | 0.3315 | 0.2438 |

- As the number of retrieved reviews increases, the performance metrics of all methods generally show an upward trend. This indicates that leveraging more historical user data enables the model to better infer users' personalized preferences and generate more tailored reviews.
- Across different numbers of retrieved reviews, our DPL method consistently achieves performance that is comparable to or even surpasses the best-performing baselines. Notably, when the number of retrieved reviews reaches 8, DPL steadily outperforms all other methods across all three datasets. This demonstrates the robustness of our method and its ability to effectively leverage richer user history for enhanced personalization.

G Overview of Templates & Prompts

In this section, we provide the prompts used for different components. Specifically, we list the prompt used for the difference extractor LLM following Equation (5) in Figure 7, the prompt used for the summarizer LLM following Equation (3) in Figure 9, and the prompt used for the review generator LLM following Equation (2) in Figure 10.

Template

Given the title and description of an item, along with the current user's review and 4 other users' reviews for the same item, analyze and output the differences between the current user and other users by considering the following aspects:
 [Writing Style]: word choice and sentence structure.
 [Emotional Style]: sentiment tone (positive, negative, or neutral).
 [Semantic Style]: information density and contextual coherence.

Input

[Item Information]:
 - [Item Title]: *<item title>*
 - [Item Description]: *<item description>*

[Review by the current user]:
 - [User's Rating]: *<user' s rating>*
 - [Review Title]: *<review title>*
 - [Review Text]: *<review text>*

[Review the other user 1]:
 - [User's Rating]: *<other user 1' s rating>*
 - [Review Title]: *<other user 1' s review title>*
 - [Review Text]: *<other user 1' s review text>*

[Review the other user 2]:
 - [User's Rating]: *<other user 2' s rating>*
 - [Review Title]: *<other user 2' s review title>*
 - [Review Text]: *<other user 2' s review text>*

[Review the other user 3]:
 - [User's Rating]: *<other user 3' s rating>*
 - [Review Title]: *<other user 3' s review title>*
 - [Review Text]: *<other user 3' s review text>*

[Review the other user 4]:
 - [User's Rating]: *<other user 4' s rating>*
 - [Review Title]: *<other user 4' s review title>*
 - [Review Text]: *<other user 4' s review text>*

Figure 7: The prompt used for the difference extractor LLM.

H Case Studies

In this section, we provide more case studies to showcase our work.

H.1 Case Study: Comparison between Difference-aware Representation and User Profile Summary

We present a case study to compare the difference-aware representation generated by our proposed DPL method with the user profile summary produced by LLM-TRSR, as shown in Figure 8. From the figure, it is evident that the difference-aware representation generated by our method can better capture the user's writing style, emotional style, and semantic style, thereby achieving a higher degree of personalization, as illustrated in Figure 4. In contrast, the profile summary produced by LLM-TRSR


| | |
|---|--|
| <p>DPL Generated Difference-aware Representation:</p> <p>The current user tends to provide straightforward and personal reviews, often using a conversational tone. Their emotional responses are varied, ranging from highly positive (such as in "Deadpool") to negative or disappointed (such as in "X-Men: Apocalypse" and "The Hobbit: The Desolation of Smaug"). They appreciate superhero movies with a lighter tone and feel that Marvel's approach is superior to DC's in general, though they enjoyed "Batman v Superman: Dawn of Justice" and "Justice League" more than many other critics. Their writing is direct, focusing on personal impressions rather than detailed analysis, and they occasionally incorporate comparisons to other films in the same franchise or genre.</p> |  <p>[Item Title]: Aquaman (Blu-ray) [Review Rating]: 3.0 [Review Title]: A little too epic</p> |
| <p>LLM-TRSR Generated Summary:</p> <p>This user is a discerning film enthusiast with a notable appreciation for science fiction and superhero movies, particularly when they offer a fresh and unique take on the genre. They value strong acting, captivating storylines, and films with a gritty, dark tone. While they are enthusiastic about movies that meet these criteria—such as "Logan" and "Westworld"—they are also critical of adaptations that fall short of their expectations, such as "STAR WARS: THE LAST JEDI." They enjoy well-executed films that offer both action and depth, and are impressed by performances that elevate a movie, as seen in their high rating for "Logan." The user finds value in sequels that maintain a strong sense of originality, like "Deadpool 2," and appreciates action-packed films that, despite their flaws, remain engaging, such as "Justice League."</p> | |

Figure 8: Case Study: a comparison of the difference-aware representation generated by the DPL method and the user profile summary LLM-TRSR method. The sections highlighted in red color demonstrate the advantages of the difference-aware representation generated by DPL.

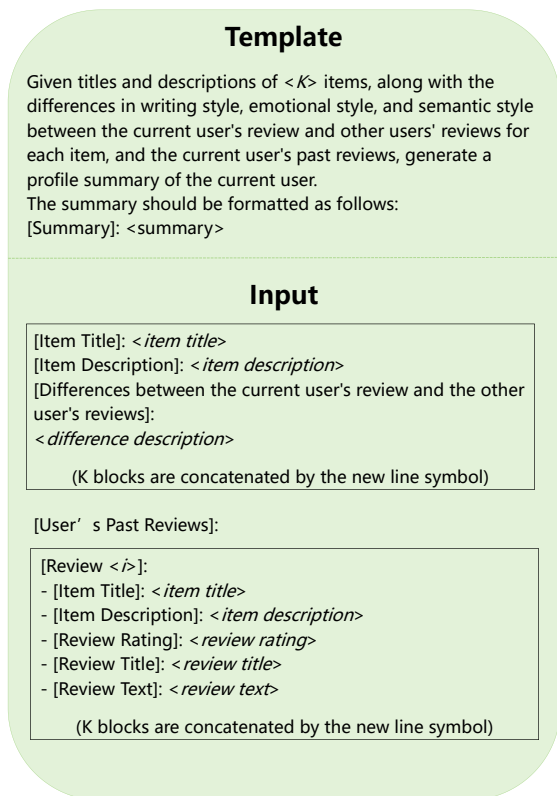


Figure 9: The prompt used for the summarizer LLM.

includes a lot of irrelevant information, which does not effectively contribute to personalized review generation.

H.2 Case Study: Outputs of DPL Components

We present a case study (Figure 11) to illustrate the outputs generated by the three key LLM-based components of DPL: the difference extractor, sum-

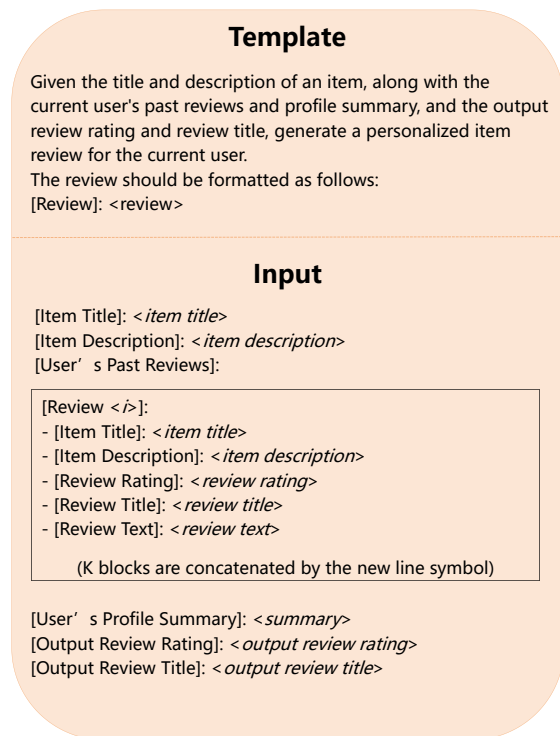


Figure 10: The prompt used for the review generator LLM.

marizer, and review generator. The results show that each component produces well-structured and meaningful outputs.



Figure 11: The outputs of various DPL components: the red section represents the output of the difference extractor, the green section corresponds to the summarizer's output, and the yellow section denotes the output of the review generator, which serves as the final targeted review.