Personalized Graph-Based Retrieval for Large Language Models

Anonymous submission

Abstract

As large language models (LLMs) evolve, their ability to deliver personalized and context-aware responses offers transformative potential for improving user experiences. Existing personalization approaches, however, often rely solely on user history to augment the prompt, limiting their effectiveness in generating tailored outputs, especially in cold-start scenarios with sparse data. To address these limitations, we propose Personalized Graph-based Retrieval-Augmented Generation (PGraphRAG), a framework that leverages user-centric knowledge graphs to enrich personalization. By directly integrating structured user knowledge into the retrieval process and augmenting prompts with user-relevant context, PGraphRAG enhances contextual understanding and output quality. We also introduce the Personalized Graph-based Benchmark for Text Generation, designed to evaluate personalized text generation tasks in real-world settings where user history is sparse or unavailable. Experimental results show that PGraphRAG significantly outperforms state-of-the-art personalization methods across diverse tasks, achieving an average relative gain of 14.8% ROUGE-1 on the long-text generation tasks and 4.6% ROUGE-1 on the short-text generation tasks, demonstrating the unique advantages of graph-based retrieval for personalization.

1 Introduction

The recent development of large language models (LLMs) has unlocked numerous applications in natural language processing (NLP), including advanced conversational agents, automated content creation, and code generation. For instance, models like GPT-4 (OpenAI, 2024) have been employed to power virtual assistants capable of answering complex queries, summarizing lengthy documents, and engaging in human-like conversations. These advancements highlight the transformative potential of LLMs to automate and enhance tasks across various domains (Brown et al., 2020). As LLMs continue to evolve, their ability to deliver highly personalized and context-aware responses opens new possibilities for transforming user experiences (Salemi et al., 2024b). Personalization enables these models to adapt outputs to individual preferences, contexts, and goals, fostering richer and more meaningful interactions (Huang et al., 2022). For example, personalized text generation allows AI systems to provide responses that are more relevant, contextually appropriate, and aligned with the style and preferences of individual users (Zhang et al., 2024).

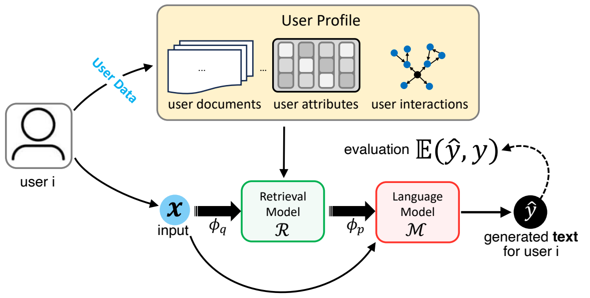

Figure 1: Overview of the proposed personalized graph-based retrieval-augmented generation framework, PGraphRAG. We first construct user-centric graphs from user history and interactions. Then, the resulting structured data is utilized for retrieval. The retrieved information is provided to the language models for context in generating text tailored to user i.

Personalization. The concept of personalization is well-established in AI and has been extensively explored across various fields, including information retrieval, human-computer interaction (HCI), and recommender systems. In information retrieval, personalization techniques are employed to tailor search results based on user profiles and past interactions, enhancing the relevance of retrieved documents (Xue et al., 2009). HCI research has focused on creating adaptive user interfaces and interactions that cater to individual needs, improving usability and accessibility (Fowler et al., 2015).

Recommender systems utilize personalization to suggest products, services, or content that match user interests, driving engagement in applications ranging from e-commerce to entertainment (Naumov et al., 2019; Lyu et al., 2024a). Despite the widespread acknowledgment of the importance of personalization in these domains, the development and evaluation of large language models (LLMs) for generating personalized responses remain relatively understudied.

One of the key challenges in advancing personalized LLMs is the lack of suitable benchmarks that adequately capture personalization tasks. Popular natural language processing (NLP) benchmarks (e.g., (Wang et al., 2019b), (Wang et al., 2019a), (Gehrmann et al., 2021)) primarily focus on general language understanding and generation capabilities, with limited emphasis on personalization aspects. As a result, researchers and practitioners lack standardized datasets and evaluation metrics to develop and assess models designed for personalized text generation. Recently, some efforts have been made towards personalized LLM benchmarks. The LaMP benchmark offers a comprehensive evaluation framework focusing on personalized text classification and generation including email subject generation, news headline generation, paper title generation, product rating and movie tagging (Salemi et al., 2024b). LongLaMP extended this scope with four tasks emphasizing long text generation, such as email completion and paper abstract generation (Kumar et al., 2024). Unfortunately, these recently developed personalized LLM benchmarks rely exclusively on user history to model personalization.

Cold Start Users. While user history is undoubtedly valuable for capturing a user's preferences and behaviors, this approach has significant limitations. In scenarios where user data is sparse or entirely unavailable — such as with new users in cold-start situations — models that depend solely on user history fail to generate personalized outputs effectively. This dependency restricts the applicability of such benchmarks in evaluating personalized LLMs for real-world use cases, where the availability and quality of user history can vary greatly. For example, Figure 2 shows the user profile distribution for Amazon user-product reviews (Ni and McAuley, 2018) where 99.99% of users have only one or two reviews in their profile. Interestingly, other personalized LLM benchmarks such as LaMP and LongLaMP limited their datasets to users with sufficient profile size.

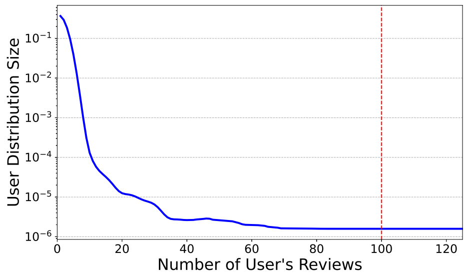

Figure 2: The user profile distribution for Amazon user-product dataset which highlights how most users have a small profile size with few reviews. The red vertical line marks the minimum profile size in other benchmarks (e.g., LaMP, LongLaMP).

PGraphRAG. To address these challenges, we propose *Personalized Graph-based Retrieval-Augmented Generation* (PGraphRAG), a novel framework that leverages user-centric knowledge represented as structured graphs to enhance personalized text generation. By incorporating user-centric knowledge graphs directly into the retrieval process and augmenting the generation context or prompt with structured user-specific information, PGraphRAG provides a richer and more comprehensive understanding of the user's context, preferences, and relationships (see Figure 1 for an overview of the framework). This approach transcends the limitations of relying solely on user history by integrating diverse and structured user knowledge, enabling the model to generate more accurate and personalized responses even when user history is sparse or unavailable. The use of structured graphs allows PGraphRAG to represent complex user information, such as interests and past interactions, in a structured and interconnected manner. By augmenting the prompt with this structured knowledge during the generation process, PGraphRAG facilitates more effective retrieval and integration of relevant user-centric information, significantly enhancing the model's ability to produce contextually appropriate and personalized outputs. In cold-start scenarios, where traditional models fail due to the lack of user history, PGraphRAG leverages available structured knowledge to deliver meaningful personalization.

Benchmark. To evaluate our approach, we introduce the Personalized Graph-based Benchmark for Text Generation, a novel evaluation benchmark designed to fine-tune and assess LLMs on twelve personalized text generation tasks including long and short text generation, as well as classification. This benchmark addresses the limitations of existing personalized LLM benchmarks by providing datasets that specifically target personalization capabilities in real-world settings where user history is sparse. In addition, the benchmark enables a more comprehensive assessment of a model's ability to personalize outputs based on structured user information. Our contributions can be summarized as follows:

1. **Benchmark.** We propose a Personalized Graph-based Benchmark for with 12 distinct tasks. To support further research, we make it available [1].
2. **Problem.** Current approaches to personalized text generation struggle with *cold-start users*, who have only minimal history data. To address this problem, we propose PGraphRAG by augmenting the context with structured user-specific information.
3. **Effectiveness.** We demonstrate the state-of-the-art performance of PGraphRAG across the new benchmark in producing personalized outputs using user-centric knowledge graphs.

2 Personalized Graph-based Benchmark for LLMs

Here, we discuss the proposed Personalized Graph-Based Benchmark to evaluate LLMs in their ability to produce personalized text generations for twelve personalized tasks including long text generation, short text generation, and ordinal classification. The benchmark datasets were collected from several real-world datasets from various domains. LLMs typically take an input x and predict the most likely sequence of tokens y that follows x. As such, each data entry in the benchmark consists of: (1) an input sequence x that serves as the input to LLMs, (2) a target output sequence y that the LLM is expected to generate, and (3) a user-centric bipartite graph. Given an input sample (x, y) for any user i, the goal is to generate a personalized output $\hat{y}$ that matches the target output y conditioning on the user profile P_i.

We represent the user-centric graph as a bipartite knowledge graph $G = (U, V, E)$, such that U denotes user nodes, V denotes item nodes, and E denotes the interaction edges among users and items. For example, an edge $(i, j) \in E$ may represent a review written by user i for item j, including all details such as the review text, title, and rating. In this benchmark, we define the user profile P_i as the set of reviews written by user i, and the set of reviews for item j written by other users k where $k \neq i$. We provide a summary of all task statistics and their associated graphs in Table 1 and Table 2 respectively. Due to space limitations, details of dataset splits are in the appendix section.

2.1 Task Definitions

Task 1: User Product Review Generation. Personalized review text generation has progressed from incorporating user-specific context to utilizing LLMs for generating fluent and contextually relevant reviews and titles (Ni and McAuley, 2018). This task aims to generate a target product review i_{text} given the target user's product review title i_{title} and a set of additional reviews P_i from their profile. We use the Amazon Reviews 2023 dataset (Hou et al., 2024) to construct data splits and bipartite graphs across multiple product categories.

Task 2: Hotel Experience Generation. Hotel reviews often contain detailed narratives reflecting users' personal experiences, making personalization crucial for capturing individual preferences and accommodations (Kanouchi et al., 2020). This task focuses on generating a personalized hotel experience story i_{text} based on the target user's hotel review summary i_{title} and a set of additional reviews P_i. The Hotel Reviews dataset, a subset of Datafiniti's Business Database (Datafiniti, 2017), is used to construct data splits and a user-hotel graph.

Task 3: Stylized Feedback Generation. User writing style, influenced by grammar, punctuation, and spelling, reflects individual preferences and is shaped by geographic and cultural factors, making it critical for personalized text generation (Alhafni et al., 2024). This task involves generating target feedback i_{text} based on the target user's feedback title i_{title} and a set of additional feedback P_i from their profile. We use the Grammar and Online Product dataset, a subset of the Datafiniti Business dataset (Datafiniti, 2018), which highlights writing quality across multiple platforms.

Task 4: Multi-lingual Review Generation. Personalization in multilingual review generation presents unique challenges due to variations in linguistic structures, cultural nuances, and stylistic conventions (Cortes et al., 2024). In this task, we

[1] https://anonymous.4open.science/r/PGraphRAG-186B/

Task	Type	Avg. Input Length	Avg. Output Length	Avg. Profile Size	# Classes
User-Product Review Generation	Long Text Generation	3.754 ± 2.71	47.90 ± 19.28	1.05 ± 0.31	-
Hotel Experiences Generation	Long Text Generation	4.29 ± 2.57	76.26 ± 22.39	1.14 ± 0.61	-
Stylized Feedback Generation	Long Text Generation	3.35 ± 2.02	51.80 ± 20.07	1.09 ± 0.47	-
Multilingual Product Review Generation	Long Text Generation	2.9 ± 2.40	34.52 ± 12.55	1.08 ± 0.33	-
User-Product Review Title Generation	Short Text Generation	30.34 ± 37.95	7.02 ± 1.14	1.05 ± 0.31	-
Hotel Experiences Summary Generation	Short Text Generation	90.40 ± 99.17	7.64 ± 0.92	1.14 ± 0.61	-
Stylized Feedback Title Generation	Short Text Generation	37.42 ± 38.17	7.16 ± 1.11	1.09 ± 0.47	-
Multilingual Product Review Title Generation	Short Text Generation	22.17 ± 20.15	7.15 ± 1.09	1.08 ± 0.33	-
User-Product Review Ratings	Ordinal Classification	34.10 ± 38.66	-	1.05 ± 0.31	5
Hotel Experiences Ratings	Ordinal Classification	94.69 ± 99.62	-	1.14 ± 0.61	5
Stylized Feedback Ratings	Ordinal Classification	40.77 ± 38.69	-	1.09 ± 0.47	5
Multilingual Product Ratings	Ordinal Classification	25.15 ± 20.75	-	1.08 ± 0.33	5

Table 1: Data statistics for PGraphRAG Benchmark across the four datasets. The table reports the average input length and average output length in words (done for the test set on GPT-4o-mini on BM25 back on all methods). The average profile size for each task is the number of reviews a user has.

Dataset	Users	Items	Edges/Reviews	Average Degree
User-Product Review Graph	184,771	51,376	198,668	1.68
Hotel Experiences Graph	15,587	2,975	19,698	2.12
Stylized Feedback Graph	58,087	600	71,041	2.42
Multilingual Product Review Graph	112,993	55,930	131,075	1.55

Table 2: Graph statistics for the datasets used in the personalized tasks. The table provides the number of users, items, edges (reviews), and the average degree for each dataset: User-Product Graph, Multilingual Product Graph, Stylized Feedback Graph, and Hotel Experiences Graph.

generate target product reviews i_{text} in Brazilian Portuguese based on the target user's review title i_{title} and additional reviews P_i in their profile. The B2W-Reviews dataset (Real et al., 2019), collected from Brazil's largest e-commerce platform, is used to create data splits.

Task 5: User Product Review Title Generation. Short text generation for personalized review titles is particularly challenging due to the need for summarization, sentiment dissemination, and capturing user behavior styles. This task generates a target review title i_{title} using the target user's review text i_{text} and additional reviews P_i from their profile, without relying on parametric user information (Xu et al., 2023). We construct the dataset from the Amazon Reviews dataset (Hou et al., 2024).

Task 6: Hotel Experience Summary Generation. Consolidating hotel information to help guests make informed decisions and personalize their experience is crucial (Kamath et al., 2024). This task focuses on generating the target user's hotel experience summary i_{title} using their experience text i_{text} and additional experiences P_i. We leverage the Datafiniti Business Database on Hotel Reviews (Datafiniti, 2017).

Task 7: Stylized Feedback Title Generation. Opinion datasets often lack review titles and rely on comparing reviews with desirable feedback to generate Stylized Opinion Summarization (Iso et al., 2024). This task benchmarks stylized feedback across domains such as music, groceries, and household items. The goal is to generate the target user's feedback title i_{title} based on their feedback text i_{text} and additional feedback P_i. The dataset is constructed from the Datafiniti Products dataset (Datafiniti, 2018).

Task 8: Multi-lingual Review Title Generation. Brazilian Portuguese presents unique challenges in simplifying review text (Scalercio et al., 2024), particularly in a multilingual approach to generating review titles. This short task generates the target user's product review title i_{title} using their review text i_{text} and additional user reviews P_i. The dataset is created from the B2W-Reviews dataset (Real et al., 2019).

Task 9: User Product Review Ratings. Recent advancements in sentiment analysis have utilized graph structures to enhance sentiment prediction (Zhang et al., 2023; Kertkeidkachorn and Shirai, 2023). This task focuses on predicting ratings within an ordinal classification framework, assigning values from 1 to 5. To generate a user-product review rating i_{rating}, we use the target user's product review i_{text}, the corresponding title i_{title}, and additional reviews P_i as context. The dataset is constructed from the Amazon Reviews dataset (Hou et al., 2024).

Task 10: Hotel Experience Ratings. Guest reviews often address multiple aspects of hotel experiences, which are typically framed as multi-label classification problems (Fehle et al., 2023). This task adapts this aspect to evaluating personalized bias lodging scores. We define a user's hotel experience rating i_{rating} based on their hotel experience story i_{text} and the summary i_{title}, with additional context from P_i. The dataset is derived from the Hotel Reviews dataset (Datafiniti, 2017).

Task 11: Stylized Feedback Ratings. Exploring sentiment across different domains highlights variations in writing quality and the factors influencing sentiment (Yu et al., 2021). This task investigates domain-specific variations by assigning a numerical feedback rating i_{rating} to a target stylized user review. The input includes the stylized review text i_{text} and title i_{title}. The dataset is constructed from the Datafiniti Product Database on Grammar and Online Product Reviews (Datafiniti, 2018).

Task 12: Multi-lingual Product Ratings. Sentiment analysis has proven effective at the sentence level when applied in Portuguese (de Araujo et al., 2024). However, this task extends beyond simple sentences to explore variability in Brazilian product reviews by generating a Portuguese user-product rating i_{rating} for a targeted review by considering both the review text i_{text} and the review title i_{title} as context. We construct the dataset from B2W-Reviews (Real et al., 2019).

3 PGraphRAG Framework

In this section, we present PGraphRAG, our proposed approach for personalizing large language models (LLMs). PGraphRAG enhances personalization by prompting a shared model with user-specific context, effectively integrating structured user-specific knowledge to enable tailored and context-aware text generation. As discussed in Section 2, PGraphRAG leverages a rich user-centric bipartite graph G that enables our approach to a broader context beyond the user history. Specifically, for any user i, we define the user profile P_i as the set of previous texts written by user i (i.e., $\{(i,j) \in E\}$), and the set of texts written by other users k for the same items connected to user i (i.e., $\{(k,j) \in E \mid (i,j) \in E\}$). As such, the user profile P_i is defined as follows,

$$P_i = \{(i,j) \in E\} \cup \{(k,j) \in E \mid (i,j) \in E\} \quad (1)$$

$$\forall j \in V, k \in U, k \neq i$$

Considering the context length limitations of certain LLMs and the computational costs of processing contexts, we utilize retrieval augmentation to extract only the most relevant information from the user profile with respect to the input query. This retrieved information is then used to condition the model's predictions for the current unseen test case.

Given an input sample (x, y) for user i, we follow a few steps to generate $\hat{y}$, which includes a query function, a graph-based retrieval model, and a prompt construction function seen in Figure 1.

1. **Query Function** (ϕ_q): The query function transforms the input x into a query for retrieving from the user profile.
2. **Graph-Based Retrieval** ($\mathcal{R}$): The retrieval function $\mathcal{R}(q, G, k)$ takes as input the query q, the bipartite graph G, and a threshold k. First, the retrieval function leverages the graph G to construct the user profile P_i. Then, it retrieves the k-most relevant entries from the user profile.
3. **Prompt Construction** (ϕ_p): The prompt construction assembles a personalized prompt for user i by combining the input x with the retrieved entries.

We define the constructed input using $\mathcal{R}$ as $\tilde{x}$:

$$\tilde{x} = \phi_p(x, \mathcal{R}(\phi_q(x), G, k)) \quad (2)$$

Then, we use $(\tilde{x}, y)$ to train or evaluate LLMs.

4 Experiments

Setup. The LLaMA-3.1-8B-Instruct model (Touvron et al., 2023) is implemented using the Huggingface `transformers` library using default settings and configured to produce outputs with a maximum length of 512 tokens. Thes experiments are conducted on an NVIDIA A100 GPU with 80GB of memory. We access GPT-4o-mini model(OpenAI, 2024) via the Azure OpenAI Service (Services, 2023), using the `AzureOpenAI` class with the temperature set to 0.4.

4.1 Data Construction and Splitting

To construct our user–item graph, we model users and products as nodes, with edges representing user reviews of products. Each user must have at least one reviewed product that is also reviewed by another user (i.e., forming a shared connection) to be selected as a gold-label edge. If the randomly selected review from a user does not meet this neighbor criterion, we instead select another review from the user's profile. Users who have no neighbor-compatible reviews remain in the dataset but are excluded from selection, as our random draw occurs at the edge level rather than across the user's full node profile. This filtering step ensures the resulting user–item graph remains connected, facilitating comparative tasks (e.g., multiple reviewers for the same product) and cold-start scenarios, where even users with few reviews maintain shared item nodes with others.

After identifying each user's valid "neighbor-linked" review(s), we split users into training, development, and test sets in a way that preserves these neighbor relationships:

1. **Global Neighbor Preservation:** Products with multiple reviewers are assigned in batches so that at least one other user in the same split has reviewed the same product.

2. **Local Neighbor Preservation:** Once a user with a particular product is placed in a split, subsequent users who reviewed that product are assigned to the same split to maintain connectivity.

Finally, we stratify each split by user review profile size to reflect the original distribution from the original dataset while retaining local and global neighbor structures. Controlling the neighbor preservation and stratification of user profile size, product review distribution (amount of reviews per product) is maintained. This comprehensive process ensures that each split is representative of real-world user review patterns and that all three graph properties are reflective of the original. The graph statistics are seen in Table 2. Data statistics are shown in Table 1 and data split size in Table 8.

Graph Construction. We construct a bipartite user-item graph from the selected user profiles in our validation and test splits. Each user node connects to item nodes representing products they have reviewed, with edges denoting individual reviews. This structure underpins two retrieval modes: (1) *LaMP*, which only searches edges corresponding to the user's own reviews, and (2) *PGraphRAG Neighbors*, which further incorporates reviews from neighboring user nodes via the graph. Traversing the node will return a list, where both modes create the context for PGraphRAG.

Ranking and Retrieval. The query differs by task category: *Long Text Generation* (review title), *Short Text Generation*(review text), *Ordinal Classification*(review title + text). We employ BM25 (Robertson and Zaragoza, 2009) and Contriever (Lei et al., 2023) that retrieve the top $k = 5$ reviews from each the user's own edges (LaMP) and their nearest neighbors in the graph. By ranking, it retrieves only the most relevant context with the k limit, where the minority of products is above the limit as shown in 7 and 2. These constraints users or products with a lot of reviews to be similar to those of cold-start users. The initial corpus was tokenized using NLTK's `word_tokenize` before being passed to the retrievers. They use normal settings without additional hyperparameters where the contriever applies mean pooling to token embeddings.

LLM Prompt Generation. Once the top-k reviews are identified, we incorporate them into a *template-based prompt* passed to a large language model (LLM). As illustrated in Figure **??**, the prompt includes both the user's query (e.g., a request for a long-form review, a short title, or a rating) and the list of reviews. Then, the LLM returns the predicted task given the set of instructions as shown in Figure 3.

Baseline Methods. We compare our method against several non-personalized and personalized approaches. (1) *No-Retrieval* serves as a non-personalized baseline where the prompt is constructed without any retrieval augmentation. The LLM generates the target text solely based on the query. (2) *Random-Retrieval* serves as a non-personalized baseline where the prompt is constructed with augmentation using a random item from all user profiles. (3) *LaMP (Salemi et al., 2024b)* is a personalized baseline where the prompt is constructed with augmentation with user-specific input or context, such as previous reviews written by the user.

Evaluation. For evaluation, we assess each method by providing task-specific inputs and measuring performance based on the generated outputs.

For long and short text generation tasks, we utilize the ROUGE-1, ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) metrics. For rating prediction tasks, we evaluate performance using MAE and RMSE as metrics.

4.2 Baseline Comparison

Together, these three tasks illustrate how review formulation—whether expanding a short title, generating concise text, or assigning numerical ratings—directly impacts how user information is disseminated throughout the model. For more descriptive tasks, user knowledge graphs provide richer context that can elevate the generation quality. Conversely, when prompts are minimal or scores are discrete, retrieving and integrating user data may offer limited gains if the prompt lacks the necessary hooks or if domain-specific biases dominate.

Long Text Generation. Table 3 & 16 shows PGraphRAG consistently outperforms the baseline methods in order of no-retrieval, random retrieval, and LaMP across all metrics. PGraphRAG showed the greatest improvement in Hotel Experience Generation over the LaMP baseline in both models, with gains in ROUGE-1 (+32.1%), ROUGE-L (+21.7%), and METEOR (+25.7%) in LLaMA-3.1-8B-Instruct . This shows the benefits gained by incorporating a broader context from user-centric graphs. Due to the greater length of the reference and predicted text, there are more opportunities for predicted review body to overlap with the gold label, resulting in higher scores.

Short Text Generation. Table 4 & 17 , PGraphRAG outperforms the baselines in most cases, where User Product Review Title Generation PGraphRAG achieves small, consistent improvements in ROUGE-1 (+5.6%), ROUGE-L (+5.9%), and METEOR (+6.8%) over LaMP in LLaMA-3.1-8B-Instruct . Since the short-generation tasks inherently provide fewer words to match against the reference, the ROUGE and METEOR scores tend to be lower for these tasks. Minor lexical differences can lead to significant score reductions, and there are fewer opportunities to align with reference labels.

Ordinal Classification. In Tables 6, and 18, PGraphRAG out performs 1 of 4 tasks in LLaMa and 2 of 4 in GPT with nonsignificant improvements of MAE (+1.75%) and RMSE (+1.12%) for Multi-lingual Product Ratings across both configurations compared to LaMP, with improvements of MAE (+2.16%) and RMSE (+3.17%) respectively. We speculate that the granularity of the domain is important as similar reviews in Hotel Experience and the Multilingual of digital/electronic items provide less variability for the model to reason the product quality to the user's expectations.

4.3 Ablation Study

We conduct ablation studies to evaluate the impact of different retrieval configurations on PGraphRAG's performance. These experiments examine variations in retrieval depth, retrieval domain, and retriever model. Results and further analysis are provided in Appendix C & D.

5 Conclusion

In this paper, we introduce PGraphRAG, a framework that enhances personalized text generation by integrating user-centric knowledge graphs into retrieval-augmented generation. Unlike traditional approaches that rely solely on user history, PGraphRAG incorporates structured user knowledge, enabling more context-aware and adaptive responses. Our experiments demonstrate that graph-based retrieval significantly improves personalization, outperforming state-of-the-art methods across multiple personalized text generation tasks.

Beyond immediate performance improvements, our work opens new directions for personalization at scale. We highlight how LLMs can scale personalization to a broader audience by generalizing across similar users. This introduces new opportunities for extending user information dynamically, allowing models to infer and adapt to user preferences even in cold-start scenarios.

By guiding LLMs in discerning which contextual information is most relevant, our personalization strategy not only refines the model's reasoning but also lays the groundwork for more advanced user assistance—helping individuals navigate items or interests with increased clarity. Moreover, the use of a structured knowledge base offers a strong foundation for agentic systems, particularly in scenarios where user data are sparse. Combining retrieval-augmented generation with user knowledge graphs enables better adaptive personalization for LLMs, enhancing informed inferences across diverse social and user-centric platforms.

Long Text Generation	Metric	PGraphRAG	LaMP	No-Retrieval	Random-Retrieval
LLaMA-3.1-8B-Instruct					
Task 1: User-Product Review Generation	ROUGE-1	**0.178**	0.173	0.172	0.124
	ROUGE-L	**0.129**	0.129	0.123	0.094
	METEOR	0.151	0.138	**0.154**	0.099
Task 2: Hotel Experiences Generation	ROUGE-1	**0.263**	0.199	0.231	0.216
	ROUGE-L	**0.157**	0.129	0.145	0.132
	METEOR	**0.191**	0.152	0.153	0.152
Task 3: Stylized Feedback Generation	ROUGE-1	**0.217**	0.186	0.190	0.184
	ROUGE-L	**0.158**	0.134	0.131	0.108
	METEOR	**0.178**	0.177	0.167	0.122
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.188**	0.176	0.174	0.146
	ROUGE-L	**0.147**	0.141	0.136	0.116
	METEOR	**0.145**	0.125	0.131	0.109
GPT-4o-mini					
Task 1: User-Product Review Generation	ROUGE-1	**0.189**	0.171	0.169	0.159
	ROUGE-L	**0.130**	0.117	0.116	0.114
	METEOR	**0.196**	0.176	0.177	0.153
Task 2: Hotel Experiences Generation	ROUGE-1	**0.263**	0.221	0.223	0.234
	ROUGE-L	**0.152**	0.135	0.135	0.139
	METEOR	**0.206**	0.164	0.166	0.181
Task 3: Stylized Feedback Generation	ROUGE-1	**0.211**	0.185	0.187	0.177
	ROUGE-L	**0.140**	0.123	0.123	0.121
	METEOR	**0.202**	0.183	0.189	0.165
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.194**	0.168	0.170	0.175
	ROUGE-L	**0.144**	0.125	0.128	0.133
	METEOR	**0.171**	0.154	0.152	0.149

Table 3: Zero-shot performance on the test set for the Long Text Generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. The best retriever was selected based on validation performance.

Short Text Generation	Metric	PGraphRAG	LaMP	No-Retrieval	Random-Retrieval
LLaMA-3.1-8B-Instruct					
Task 5: User Product Review Title Generation	ROUGE-1	**0.131**	0.124	0.121	0.103
	ROUGE-L	**0.125**	0.118	0.115	0.098
	METEOR	**0.125**	0.117	0.112	0.096
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.127**	0.126	0.122	0.118
	ROUGE-L	**0.118**	0.117	0.114	0.110
	METEOR	0.102	**0.106**	0.101	0.093
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.149**	0.140	0.136	0.133
	ROUGE-L	**0.142**	0.134	0.131	0.123
	METEOR	**0.142**	0.136	0.129	0.121
Task 8: Multi-lingual Review Title Generation	ROUGE-1	0.124	0.121	**0.125**	0.120
	ROUGE-L	0.116	**0.122**	0.117	0.110
	METEOR	**0.108**	0.094	0.092	0.103
GPT-4o-mini					
Task 5: User Product Review Title Generation	ROUGE-1	**0.115**	0.108	0.113	0.102
	ROUGE-L	**0.112**	0.105	0.110	0.099
	METEOR	**0.099**	0.091	0.093	0.085
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.116**	0.108	0.114	0.112
	ROUGE-L	**0.111**	0.104	0.109	0.107
	METEOR	**0.081**	0.075	0.079	0.076
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.122**	0.113	0.114	0.115
	ROUGE-L	**0.118**	0.109	0.110	0.111
	METEOR	**0.104**	0.096	0.097	0.093
Task 8: Multi-lingual Review Title Generation	ROUGE-1	0.111	0.115	**0.118**	0.108
	ROUGE-L	0.105	0.107	**0.110**	0.102
	METEOR	0.083	0.088	**0.089**	0.078

Table 4: Zero-shot performance on the on the test set for the Short Text Generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. The best retriever was selected based on validation performance.

6 Limitations

The proposed approach presents several opportunities for future enhancement. One significant challenge is the development of more sophisticated strategies to train models effectively using user-specific inputs. While personalization is a core aspect of the approach, striking the right balance between capturing individual user preferences and ensuring broader model generalization remains a complex task. Another area for extension lies in its application to recommender systems. Future efforts will focus on exploring methods to dynamically adapt to evolving user preferences and address challenges such as cold-start scenarios and context-aware recommendations. Additionally, we aim to design more robust and scalable training frameworks for personalized models, broadening their applicability and improving the effectiveness and adaptability of recommender systems.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Reinald Kim Amplayo, Jihyeok Kim, Sua Sung, and Seung-won Hwang. 2018. Cold-start aware user and product attention for sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2535–2544, Melbourne, Australia. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation.

Eduardo G. Cortes, Ana Luiza Vianna, Mikaela Martins, Sandro Rigo, and Rafael Kunst. 2024. LLMs and translation: different approaches to localization between Brazilian Portuguese and European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 45–55, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Datafiniti. 2017. Hotel reviews, version 5. Retrieved September 15, 2024 from https://www.kaggle.com/datasets/datafiniti/hotel-reviews/data.

Datafiniti. 2018. Grammar and online product reviews, version 1. Retrieved September 15, 2024 from https://www.kaggle.com/datasets/datafiniti/grammar-and-online-product-reviews.

Gladson de Araujo, Tiago de Melo, and Carlos Maurício S. Figueiredo. 2024. Is ChatGPT an effective solver of sentiment analysis tasks in Portuguese? a preliminary study. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 13–21, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.

Jakob Fehle, Leonie Münster, Thomas Schmidt, and Christian Wolff. 2023. Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 202–218, Ingolstadt, Germany. Association for Computational Lingustics.

Andrew Fowler, Kurt Partridge, Ciprian Chelba, Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2015. Effects of language modeling and its personalization on touchscreen typing performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 649–658, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Comput. Surv.*, 54(4).

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*.

Xiaolei Huang, Lucie Flek, Franck Dernoncourt, Charles Welch, Silvio Amir, Ramit Sawhney, and Diyi Yang. 2022. Usernlp'22: 2022 international workshop on user-centered natural language processing. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 1176–1177, New York, NY, USA. Association for Computing Machinery.

Yu-Yang Huang, Rui Yan, Tsung-Ting Kuo, and Shou-De Lin. 2014. Enriching cold start personalized language model using social network information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 611–617, Baltimore, Maryland. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, and Yoshi Suhara. 2024. Noisy pairing and partial supervision for stylized opinion summarization. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 13–23, Tokyo, Japan. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *CoRR*, abs/2007.01282.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Srinivas Ramesh Kamath, Fahime Same, and Saad Mahamood. 2024. Generating hotel highlights from unstructured text using LLMs. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 280–288, Tokyo, Japan. Association for Computational Linguistics.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation.

Shin Kanouchi, Masato Neishi, Yuta Hayashibe, Hiroki Ouchi, and Naoaki Okazaki. 2020. You may like this hotel because ...: Identifying evidence for explainable recommendations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 890–899, Suzhou, China. Association for Computational Linguistics.

Natthawut Kertkeidkachorn and Kiyoaki Shirai. 2023. Sentiment analysis using the relationship between users and products. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8611–8618, Toronto, Canada. Association for Computational Linguistics.

Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. 2020. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. Longlamp: A benchmark for personalized long-form text generation.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ziqing Liu, Enwei Peng, Shixing Yan, Guozheng Li, and Tianyong Hao. 2018. T-know: a knowledge graph-based question answering and infor-mation retrieval system for traditional Chinese medicine. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 15–19, Santa Fe, New Mexico. Association for Computational Linguistics.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024a. LLM-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. 2024b. Llm-rec: Personalized recommendation via prompting large language models.

Puneet Mathur, Zhe Liu, Ke Li, Yingyi Ma, Gil Karen, Zeeshan Ahmed, Dinesh Manocha, and Xuedong Zhang. 2024. DOC-RAG: ASR language model personalization with domain-distributed co-occurrence retrieval augmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5132–5139, Torino, Italia. ELRA and ICCL.

Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep learning recommendation model for personalization and recommendation systems. *CoRR*, abs/1906.00091.

Jianmo Ni and Julian McAuley. 2018. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711, Melbourne, Australia. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Livy Real, Marcio Oshiro, and Alexandre Mafra. 2019. B2w-reviews01: an open product reviews corpus. In *STIL-Symposium in Information and Human Language Technology*.

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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Ahmmad O. M. Saleh, Gokhan Tur, and Yücel Saygín. 2024. Sg-rag: Multi-hop question answering with large language models through knowledge graphs. In *International Conference on Natural Language and Speech Processing*.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. 2023. Large language models meet knowledge graphs to answer factoid questions. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 635–644, Hong Kong, China. Association for Computational Linguistics.

Arthur Scalercio, Maria Finatto, and Aline Paes. 2024. Enhancing sentence simplification in Portuguese: Leveraging paraphrases, context, and linguistic features. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15076–15091, Bangkok, Thailand. Association for Computational Linguistics.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Azure AI Services. 2023. Openai (gpt-4o-mini-20240718) [large language model]. `https://learn.microsoft.com/en-us/azure/ai-services/openai`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Alicia Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed H. Chi, and Xinyang Yi. 2024. Leveraging LLM reasoning enhances personalized recommender systems. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13176–13188, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Hongyan Xu, Hongtao Liu, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Pre-trained personalized review summarization with effective salience estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10743–10754, Toronto, Canada. Association for Computational Linguistics.

Gui-Rong Xue, Jie Han, Yong Yu, and Qiang Yang. 2009. User language model for collaborative personalized search. *ACM Trans. Inf. Syst.*, 27(2).

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777, Online. Association for Computational Linguistics.

Hongyu Zang and Xiaojun Wan. 2017. Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 168–177, Santiago de Compostela, Spain. Association for Computational Linguistics.

Xin Zhang, Linhai Zhang, and Deyu Zhou. 2023. Sentiment analysis on streaming user reviews via dual-channel dynamic graph neural network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7208–7220, Singapore. Association for Computational Linguistics.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024. Personalization of large language models: A survey.

A Additional Tables

Table 5 shows the relative percent gain of PGraphRAG compared to LaMP with improvements in Task 1–7 except Task 8 in Multi-lingual Review Title Generation. We speculate the cultural differences in how users review item titles make a difference with a large proportion titling "Muito bom," translated as "Very good" in English. The model will tend to generate a more detailed answer as opposed to the social norm descriptor title. In long text generation for GPT, score improvement is approximately 15% for ROUGE-1, 13% for ROUGE-L, and 15% for METEOR, while LLaMa achieves approximately 15% for ROUGE-1, 11% for ROUGE-L, and 13% for METEOR. For short text generation, GPT shows improvements of approximately 5% for ROUGE-1, 5% for ROUGE-L, and 5% for METEOR. LLaMa achieves approximately 4% for ROUGE-1, 2% for ROUGE-L, and 6% for METEOR.

B Prompt and Output Example

The output example (shown below) compares the PGraphRAG output with the LaMP output against the gold label, for Task 2 (Hotel Experience Generation). The gold label's title is passed to the prompt alongside retrieved context to generate the review bodies. When the information is sparse, the LaMP method is too reliant on the user's other reviews, generating reviews with wrong context, but the PGraphRAG method is able to capture specific information about the target from neighboring user

Model	Metric	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
GPT-4o-mini	ROUGE-1	10.53	18.96	14.05	15.48	6.48	7.41	7.96	-3.48
	ROUGE-L	11.11	12.59	13.82	15.20	6.67	6.73	8.26	-1.87
	METEOR	11.36	25.61	10.38	11.04	8.79	8.00	8.33	-5.68
LLaMA-3.1-8B-Instruct	ROUGE-1	2.89	32.16	16.67	6.82	5.65	0.79	6.43	2.48
	ROUGE-L	0.00	21.71	17.91	4.26	5.93	0.85	5.97	-4.92
	METEOR	9.42	25.66	0.56	16.00	6.84	-3.77	4.41	14.89

Table 5: Relative percent gains of PGraphRAG over state-of-art LaMP for GPT-4o-mini and LLaMA-3.1-8B-Instruct across Tasks 1 - 8

Ordinal Classfication	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 9: User Product Review Ratings	MAE ↓	0.3400	**0.3132**	0.3212	0.3272
	RMSE ↓	0.7668	**0.7230**	0.7313	0.7616
Task 10: Hotel Experience Ratings	MAE ↓	0.3688	0.3492	**0.3340**	0.3804
	RMSE ↓	0.6771	0.6527	**0.6372**	0.6971
Task 11: Stylized Feedback Ratings	MAE ↓	0.3476	**0.3268**	0.3256	0.3704
	RMSE ↓	0.7247	**0.6803**	0.6806	0.7849
Task 12: Multi-lingual Product Ratings	MAE ↓	**0.4928**	0.5016	0.5084	0.5096
	RMSE ↓	**0.8367**	0.8462	0.8628	0.8542
GPT-4o-mini					
Task 9: User Product Review Ratings	MAE ↓	0.3832	0.3480	**0.3448**	0.4188
	RMSE ↓	0.7392	**0.7065**	**0.7065**	0.8082
Task 10: Hotel Experience Ratings	MAE ↓	**0.3284**	0.3336	0.3336	0.3524
	RMSE ↓	**0.6083**	0.6197	0.6197	0.6384
Task 11: Stylized Feedback Ratings	MAE ↓	0.3476	**0.3448**	0.3416	0.4080
	RMSE ↓	0.6738	**0.6669**	0.6711	0.7370
Task 12: Multi-lingual Product Ratings	MAE ↓	**0.4348**	0.4444	0.4564	0.4700
	RMSE ↓	**0.7367**	0.7608	0.7718	0.8112

Table 6: Performance comparison on rating prediction tasks (Tasks 9-12) using *GPT-4o-mini* and *LLaMA-3.1-8B*.

reviews. Figure 3 shows the prompt construction for the input and Table 7 shows the review density per product, where sparsity is balanced from the original graph for both product and user nodes.

Gold Review

Title: Amazing stay!
Review: Had a great stay, our room was very clean and very peaceful. It is in the heart of DT Seattle. We took the train to the hotel and it was pretty easy. We enjoyed walking all over downtown Seattle—no need to rent a car. Major shopping and amazing restaurants right around the corner! We will stay here again!

LaMP

Title: Amazing stay!
Review: I recently had an amazing stay at the Newark Garden Hotel. Despite the circumstances that brought my family together, the hotel provided a comforting and welcoming environment. The rooms were immaculate, and the staff went above and beyond to ensure we felt at home. The hot breakfast each morning was a delightful touch that made our mornings brighter. I highly recommend this hotel for anyone looking for a pleasant and accommodating stay!

PGraphRAG

Title: Amazing stay!

Review: I had an amazing stay at the Grand Hyatt Seattle! From the moment I checked in, the staff was incredibly welcoming and attentive. The room was spacious, clean, and beautifully appointed, with a stunning view of the city. The bathroom was a highlight, featuring a luxurious soaking tub and a large walk-in shower. The location was perfect, just a short walk to Pike Place Market and other attractions. I also enjoyed the fantastic amenities, including the well-equipped gym and the delicious food options available at the hotel. Overall, it was a wonderful experience, and I can't wait to return!

Reviews (n)	Exact Count (Pct.)	Cumulative Count (Pct.)
1	25,530 (49.69%)	25,530 (49.69%)
2	9,488 (18.47%)	35,018 (68.16%)
3	4,784 (9.31%)	39,802 (77.47%)
4	2,639 (5.14%)	42,441 (82.61%)
5	1,836 (3.57%)	44,277 (86.18%)

Table 7: Distribution of the number of reviews for products in the Amazon Review Dataset for Task 1, 5, and 9. The majority of products have five or fewer reviews.

Dataset	Train Size	Validation Size	Test Size
User-Product Review	20,000	2,500	2,500
Multilingual Product Review	20,000	2,500	2,500
Stylized Feedback	20,000	2,500	2,500
Hotel Experiences	9,000	2,500	2,500

Table 8: Dataset split sizes for training, validation, and testing across four datasets: User-Product Review, Multilingual Product Review, Stylized Feedback, and Hotel Experiences.

C Ablation Study Details

C.1 PGraphRAG Ablation Details

To investigate the impact of incorporating user and/or neighboring-user data in the retrieved context, we conduct an ablation study comparing three variants of PGraphRAG:

- **PGraphRAG**: The full method, where retrieved-context consists of both the target user's other reviews and reviews from neighboring users.
- **PGraphRAG-N**: Retrieval is limited to reviews from neighboring-users. The target user's other reviews are excluded from the retrieved context.
- **PGraphRAG-U**: Retrieval is limited to reviews from the target user, disregarding reviews from neighboring users.

Table 9 presents the ablation study using the GPT-4o-mini and LLaMA-3.1-8B models for the long-text generation task on Task 1 - 4. Across all datasets, both PGraphRAG and PGraphRAG-N retrieval methods consistently outperform LaMP, contrasting the impact of retrieving neighboring-user context with that of retrieving target-user history as context. PGraphRAG generally matches or slightly exceeds the performance of PGraphRAG-N, suggesting that the additional target-user history portion of the context contributes minimally to the personalized text generation task for these datasets.

The ablation study results for the GPT-4o-mini model on the short-text generation tasks are included in Table 10. The same trends can be seen in those studies across all datasets, except for GPT-4o-mini performance on the Hotel Experience Summary Generationtask, where LaMP performs the best of the three methods.

C.2 Impact of the Retrieved Items k

To evaluate the impact of the number of retrieved-context reviews (k) on model performance, we conducted experiments with $k = 1, 2,$ and 4. Table 11 summarizes the results of this ablation study on long-text generation (Tasks 1–4) using *GPT-4o-mini* and *LLaMA-3.1-8B-Instruct*. The corresponding results for short-text generation (Tasks 5–8) are presented in Table 12.

The effect of increasing k varies depending on the dataset's characteristics. The results demonstrate that increasing the amount of retrieved-context from neighboring users and the target user generally leads to better performance across all datasets and metrics. This trend highlights the importance of retrieval scales for enhancing the diversity and relevance of retrieved context.

However, due to data sparsity, many user profiles contain fewer than four "Neighboring-user reviews" or "Target-user's other reviews." In such instances, when the retriever attempts to retrieve more reviews than are available, it retrieves all existing reviews. Consequently, PGraphRAG may retrieve only one or two reviews, even when configured to retrieve $k = 4$. This behavior reflects the realistic scenario of handling cold-start users with limited existing data, a central focus of our study.

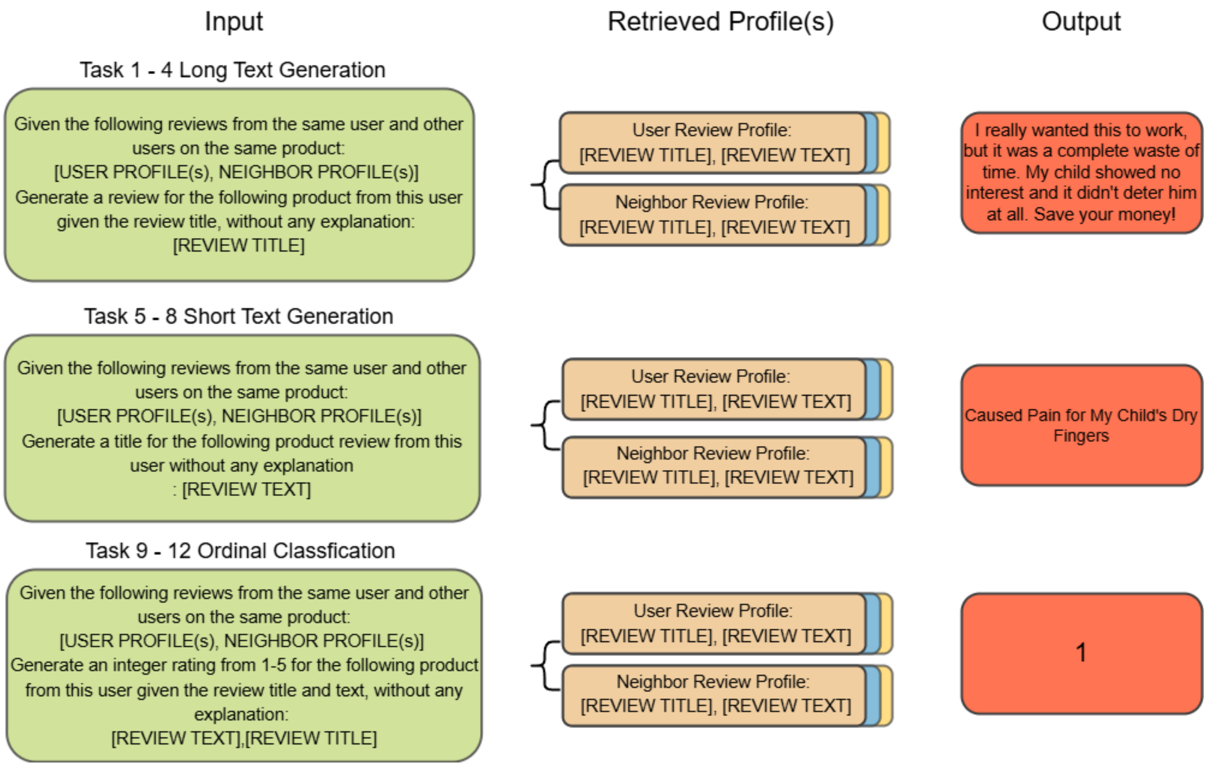

Figure 3: Examples of different prompt configurations used in each of our task types. Teletype text is replaced with realistic data for each task.

Long Text Generation	**Metric**	**PGraphRAG**	**PGraphRAG-N**	**PGraphRAG-U**
LLaMA-3.1-8B-Instruct				
Task 1: User-Product Review Generation	ROUGE-1 ROUGE-L METEOR	0.173 0.124 0.150	**0.177** **0.127** **0.154**	0.168 0.125 0.134
Task 2: Hotel Experiences Generation	ROUGE-1 ROUGE-L METEOR	0.263 0.156 0.191	**0.272** **0.162** **0.195**	0.197 0.128 0.121
Task 3: Stylized Feedback Generation	ROUGE-1 ROUGE-L METEOR	**0.226** **0.171** **0.192**	0.222 0.165 0.186	0.181 0.134 0.147
Task 4: Multilingual Product Review Generation	ROUGE-1 ROUGE-L METEOR	**0.174** 0.139 **0.133**	0.172 0.137 0.126	0.174 **0.141** 0.125
GPT-4o-mini				
Task 1: User-Product Review Generation	ROUGE-1 ROUGE-L METEOR	**0.186** **0.126** **0.187**	0.185 0.125 0.185	0.169 0.114 0.170
Task 2: Hotel Experiences Generation	ROUGE-1 ROUGE-L METEOR	0.265 0.152 0.206	**0.268** **0.153** **0.209**	0.217 0.132 0.161
Task 3: Stylized Feedback Generation	ROUGE-1 ROUGE-L METEOR	**0.205** **0.139** **0.203**	0.204 0.138 0.198	0.178 0.121 0.178
Task 4: Multilingual Product Review Generation	ROUGE-1 ROUGE-L METEOR	**0.191** **0.142** **0.173**	0.190 0.140 0.169	0.164 0.123 0.155

Table 9: Ablation study results for long text generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. PGraphRAG-N represents Neighbors-only context retrieval and PGraphRAG-U represents User-only context retrieval.

Short Text Generation	Metric	PGraphRAG	PGraphRAG-N	PGraphRAG-U
LLaMA-3.1-8B-Instruct				
Task 5: User Product Review Title Generation	ROUGE-1 ROUGE-L METEOR	0.125 0.119 0.117	**0.129** **0.123** **0.120**	0.115 0.109 0.111
Task 6: Hotel Experience Summary Generation	ROUGE-1 ROUGE-L METEOR	0.121 0.113 0.099	**0.124** **0.115** 0.103	0.119 0.111 **0.105**
Task 7: Stylized Feedback Title Generation	ROUGE-1 ROUGE-L METEOR	0.132 0.128 0.129	**0.135** **0.130** **0.132**	0.128 0.124 0.124
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1 ROUGE-L METEOR	**0.131** **0.123** **0.118**	0.131 0.122 0.110	0.124 0.114 0.098
GPT-4o-mini				
Task 5: User Product Review Title Generation	ROUGE-1 ROUGE-L METEOR	0.111 0.106 0.097	**0.116** **0.111** **0.099**	0.112 0.108 0.095
Task 6: Hotel Experience Summary Generation	ROUGE-1 ROUGE-L METEOR	0.118 0.112 **0.085**	**0.119** **0.113** **0.085**	0.109 0.104 0.077
Task 7: Stylized Feedback Title Generation	ROUGE-1 ROUGE-L METEOR	**0.109** **0.107** **0.096**	0.107 0.105 0.094	0.108 0.104 0.091
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1 ROUGE-L METEOR	0.108 0.104 0.082	0.109 0.104 0.089	**0.116** **0.109** **0.091**

Table 10: Ablation study results for short text generation tasks using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini*. PGraphRAG-N represents Neighbors-only context retrieval and PGraphRAG-U represents User-only context retrieval.

Long Text Generation	**Metric**	$k = 1$	$k = 2$	$k = 4$
LLaMA-3.1-8B-Instruct				
Task 1: User-Product Review Generation	ROUGE-1	0.160	0.169	**0.173**
	ROUGE-L	0.121	**0.125**	0.124
	METEOR	0.125	0.138	**0.150**
Task 2: Hotel Experiences Generation	ROUGE-1	0.230	0.251	**0.263**
	ROUGE-L	0.141	0.151	**0.156**
	METEOR	0.152	0.174	**0.191**
Task 3: Stylized Feedback Generation	ROUGE-1	0.200	0.214	**0.226**
	ROUGE-L	0.158	0.165	**0.171**
	METEOR	0.154	0.171	**0.192**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.163	0.169	**0.174**
	ROUGE-L	0.134	0.137	**0.139**
	METEOR	0.113	0.122	**0.133**
GPT-4o-mini				
Task 1: User-Product Review Generation	ROUGE-1	0.176	0.184	**0.186**
	ROUGE-L	0.121	0.125	**0.126**
	METEOR	0.168	0.180	**0.187**
Task 2: Hotel Experiences Generation	ROUGE-1	0.250	0.260	**0.265**
	ROUGE-L	0.146	0.150	**0.152**
	METEOR	0.188	0.198	**0.206**
Task 3: Stylized Feedback Generation	ROUGE-1	0.196	0.200	**0.205**
	ROUGE-L	0.136	0.136	**0.139**
	METEOR	0.186	0.192	**0.203**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.163	0.169	**0.174**
	ROUGE-L	0.134	0.137	**0.139**
	METEOR	0.113	0.122	**0.133**

Table 11: Ablation study results showing the impact of varying k (number of retrieved neighbors) on PGraphRAG's performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on long-text generation tasks (Tasks 1 - 4).

Short Text Generation	Metric	$k=1$	$k=2$	$k=4$
LLaMA-3.1-8B-Instruct				
Task 5: User Product Review Title Generation	ROUGE-1	**0.128**	0.123	0.125
	ROUGE-L	**0.121**	0.118	0.119
	METEOR	**0.123**	0.118	0.117
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.122**	0.121	0.121
	ROUGE-L	0.112	**0.114**	0.113
	METEOR	**0.104**	0.102	0.099
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.129	**0.132**	**0.132**
	ROUGE-L	0.124	0.126	**0.128**
	METEOR	0.129	**0.130**	0.129
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.129	0.126	**0.131**
	ROUGE-L	0.120	0.119	**0.123**
	METEOR	0.117	0.116	**0.118**
GPT-4o-mini				
Task 5: User Product Review Title Generation	ROUGE-1	**0.111**	0.110	**0.111**
	ROUGE-L	**0.106**	0.105	**0.106**
	METEOR	0.093	0.094	**0.097**
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.114	0.114	**0.118**
	ROUGE-L	0.109	0.109	**0.112**
	METEOR	0.082	0.082	**0.085**
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.100	0.103	**0.109**
	ROUGE-L	0.098	0.101	**0.107**
	METEOR	0.087	0.090	**0.096**
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.104	0.104	**0.108**
	ROUGE-L	0.098	0.098	**0.104**
	METEOR	0.077	0.078	**0.082**

Table 12: Ablation study results showing the impact of varying k (number of retrieved neighbors) on PGraphRAG's performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on short-text generation tasks (Tasks 5-8).

C.3 Impact of Retriever method $\mathcal{R}$

We study the impact of the retriever method on the proposed PGraphRAG method; we conduct an ablation study comparing two retrievers, BM25 and Contriever.

In Table 13, we compare the performance of our PGraphRAG method using these two retrievers. Across all datasets and tasks, the results demonstrate that the performance of PGraphRAG is stable and not highly sensitive to the choice of retriever. Both BM25 and Contriever show comparable results, with BM25 showing slight improvements in some cases. This stability highlights the robustness of PGraphRAG in adapting to different retrieval contexts.

Long Text Generation	Metric	Contriever	BM25
LLaMA-3.1-8B-Instruct			
Task 1: User-Product Review Generation	ROUGE-1	0.172	**0.173**
	ROUGE-L	0.122	**0.124**
	METEOR	**0.153**	0.150
Task 2: Hotel Experiences Generation	ROUGE-1	0.262	**0.263**
	ROUGE-L	0.155	**0.156**
	METEOR	0.190	**0.191**
Task 3: Stylized Feedback Generation	ROUGE-1	0.195	**0.226**
	ROUGE-L	0.138	**0.171**
	METEOR	0.180	**0.192**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.172	**0.174**
	ROUGE-L	0.134	**0.139**
	METEOR	**0.135**	0.133
GPT-4o-mini			
Task 1: User-Product Review Generation	ROUGE-1	0.182	**0.186**
	ROUGE-L	0.122	**0.126**
	METEOR	0.184	**0.187**
Task 2: Hotel Experiences Generation	ROUGE-1	0.264	**0.265**
	ROUGE-L	**0.152**	**0.152**
	METEOR	**0.207**	0.206
Task 3: Stylized Feedback Generation	ROUGE-1	0.194	**0.205**
	ROUGE-L	0.128	**0.139**
	METEOR	0.201	**0.203**
Task 4: Multilingual Product Review Generation	ROUGE-1	0.190	**0.191**
	ROUGE-L	0.141	**0.142**
	METEOR	**0.174**	0.173

Table 13: Ablation study results showing the effect of retriever choice on PGraphRAG performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on the long-text generation task (Tasks 1-4).

Short Text Generation	Metric	Contriever	BM25
LLaMA-3.1-8B-Instruct			
Task 5: User Product Review Title Generation	ROUGE-1	0.122	**0.125**
	ROUGE-L	0.116	**0.119**
	METEOR	0.115	**0.117**
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.117	**0.121**
	ROUGE-L	0.110	**0.113**
	METEOR	0.095	**0.099**
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.125	**0.132**
	ROUGE-L	0.121	**0.128**
	METEOR	0.122	**0.129**
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.126	**0.131**
	ROUGE-L	0.118	**0.123**
	METEOR	0.112	**0.118**
GPT-4o-mini			
Task 5: User Product Review Title Generation	ROUGE-1	**0.113**	0.111
	ROUGE-L	**0.108**	0.106
	METEOR	**0.097**	**0.097**
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.113	**0.118**
	ROUGE-L	0.107	**0.112**
	METEOR	0.080	**0.085**
Task 7: Stylized Feedback Title Generation	ROUGE-1	0.108	**0.109**
	ROUGE-L	0.106	**0.107**
	METEOR	0.094	**0.096**
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	**0.108**	**0.108**
	ROUGE-L	0.103	**0.104**
	METEOR	**0.082**	**0.082**

Table 14: Ablation study results showing the effect of retriever choice on PGraphRAG performance. Results are reported for *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on the short-text generation task (Tasks 5-8).

D GPT Experiments

D.1 Impact of Ranked Retrieval

In Table 15, two variations of the PGraphRAG framework show the impact of ranked retrieval: PGraphRAG*, which retrieves four randomly selected reviews as context (k=4), and PGraphRAG**, which retrieves and passes all available context within the model's limit (k approaches ∞). Since PGraphRAG** expectedly performs better, we focus on analyzing the effect of removing ranking.

Our results show that removing ranking (PGraphRAG $\rightarrow$ PGraphRAG*) leads to an average ROUGE-1 drop of 2.29% for long-text tasks and 3.18% for short-text tasks, demonstrating the importance of ranking in retrieval. Similarly, removing ranking from target user-specific retrieval (PGraphRAG-U $\rightarrow$ PGraphRAG-U*) results in a 0.92% decrease in long-text tasks and a 1.98% drop in short-text tasks. These findings confirm that ranked retrieval plays a key role in PGraphRAG's effectiveness.

While PGraphRAG** achieves the highest performance, it is impractical for larger datasets due to retrieval cost and scalability constraints. In contrast, PGraphRAG* provides a more controlled and comparable evaluation setting with a fixed retrieval threshold (k=4). This analysis highlights the tradeoffs between retrieval ranking, retrieval limits, and performance scaling, demonstrating that ranking improves effectiveness while structured retrieval strategies ensure efficiency.

D.2 Impact of GPT Models

To explore GPT model performances, we compared the performance of PGraphRAG from our best retriever and k size settings on 3.5 Turbo, 4o, 4o-mini, and o1-preview. We selected GPT-4o-mini as the best model for performance, cost, and consistency across long text generation tasks.

D.3 Impact of Length Contraints

For short-text generation, we explore length constraints of 3, 5, and 10 words, finding that a 5-word constraint achieves the best balance across metrics, combining precision and informativeness. This configuration is adopted for all short-text generation tasks.

E Validation results

We conduct a comprehensive set of experiments on the validation set for five tasks, testing all combinations of language models, retrieval methods, and top-k retrieval settings for each method. As shown in Table 16, 17, and 18. The configurations yielding the best results on the validation set are selected for subsequent test set experiments, where trends observed in the validation are consistent with those seen in the test set.

F Related Work

Personalization in natural language processing (NLP) tailors responses to individual user preferences, behaviors, and contexts, significantly enhancing user interaction and satisfaction. Early work in personalization focused on tasks such as text generation, leveraging attributes like review sentiment (Zang and Wan, 2017) and stylistic features (Dong et al., 2017). These methods, based on neural networks and encoder-decoder models, laid the foundation for personalization in text-based systems. Recent advancements have expanded personalization techniques to incorporate

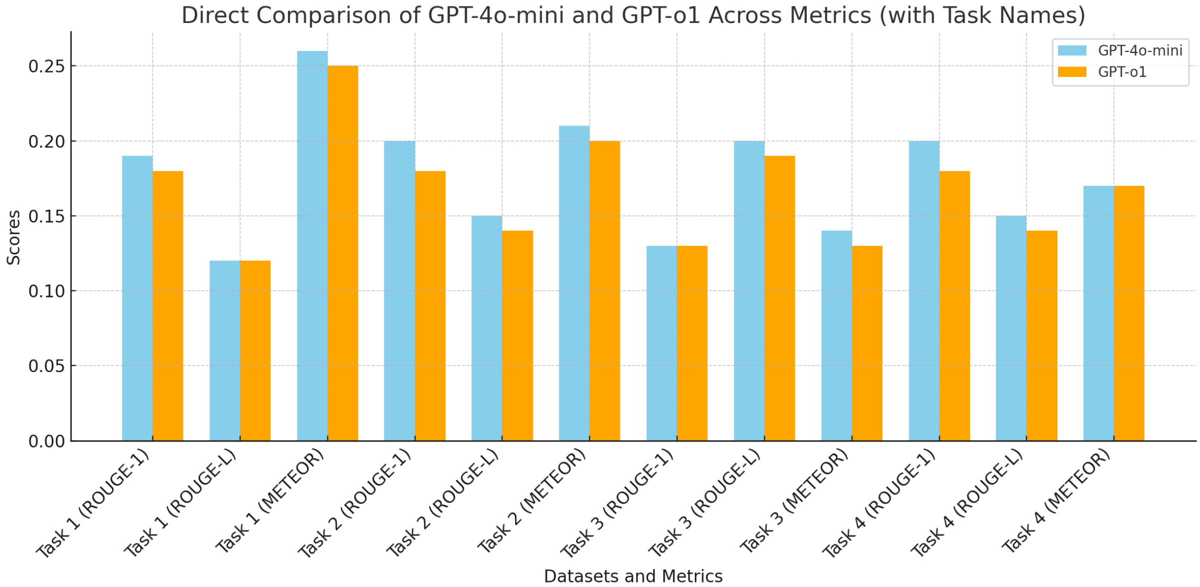

Figure 4: Comparison of *GPT-4o-mini* and *GPT-o1* performance on test set across Task 1 - 4 on BM25, and k = 4

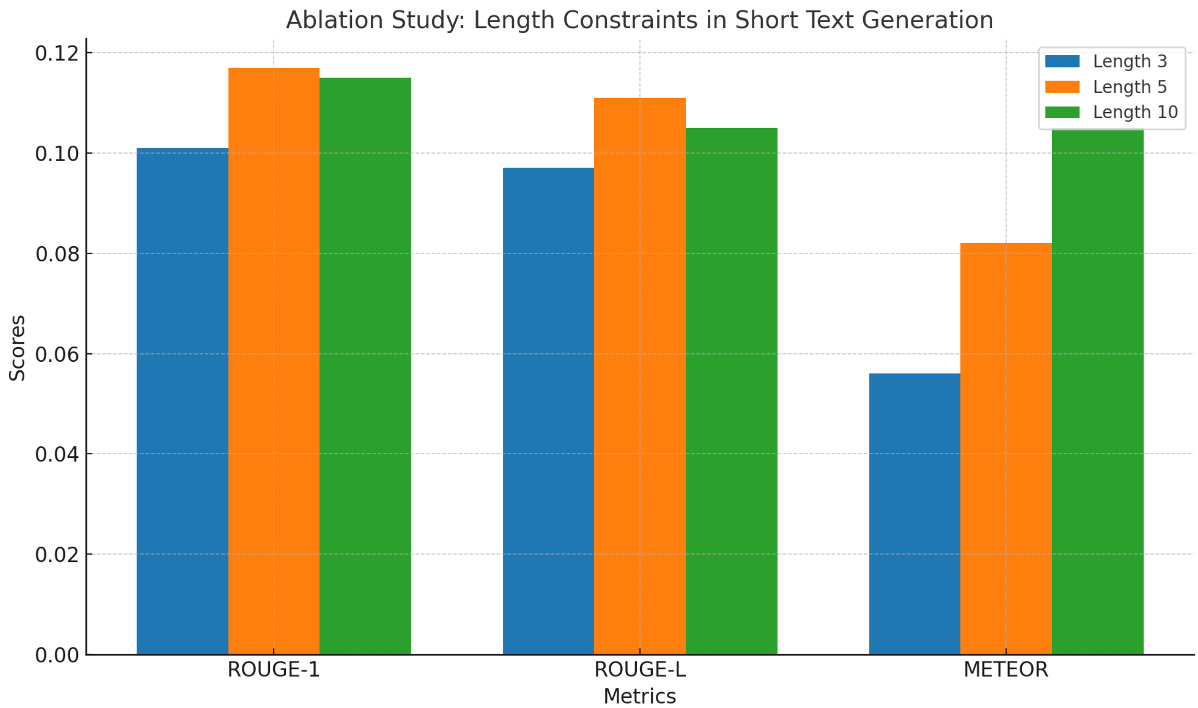

Figure 5: Impact of length constraints of 3, 5, and 10 on short-text generation tasks using PGraphRAG, evaluated on the validation set.

Task	Metric	PGraphRAG	PGraphRAG*	PGraphRAG**	PGraphRAG-U	PGraphRAG-U*	PGraphRAG-U**
Long Text Generation							
Task 1: User-Product Review Generation	ROUGE-1	0.189	0.186	**0.191**	0.171	0.169	0.170
	ROUGE-L	**0.130**	0.125	**0.130**	0.117	0.114	0.117
	METEOR	0.196	0.188	**0.205**	0.176	0.173	0.180
Task 2: Hotel Experiences Generation	ROUGE-1	0.263	0.266	**0.267**	0.221	0.223	0.225
	ROUGE-L	0.152	0.152	**0.153**	0.135	0.134	0.135
	METEOR	0.206	0.209	**0.216**	0.164	0.168	0.171
Task 3: Stylized Feedback Generation	ROUGE-1	**0.211**	0.200	0.210	0.185	0.180	0.186
	ROUGE-L	**0.140**	0.133	0.136	0.123	0.122	0.123
	METEOR	0.202	0.206	**0.225**	0.183	0.184	0.189
Task 4: Multilingual Product Review Generation	ROUGE-1	0.194	0.188	**0.196**	0.168	0.167	0.171
	ROUGE-L	**0.144**	0.138	0.141	0.125	0.125	0.128
	METEOR	0.171	0.176	**0.188**	0.154	0.155	0.155
Short Text Generation							
Task 5: User Product Review Title Generation	ROUGE-1	0.115	0.114	**0.119**	0.108	0.108	0.111
	ROUGE-L	0.112	0.109	**0.114**	0.105	0.102	0.105
	METEOR	0.099	0.121	**0.128**	0.091	0.116	0.119
Task 6: Hotel Experience Summary Generation	ROUGE-1	0.116	0.117	**0.121**	0.108	**0.121**	0.119
	ROUGE-L	0.111	0.107	**0.112**	0.104	0.111	0.110
	METEOR	0.081	0.104	**0.109**	0.075	**0.109**	0.107
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.122**	0.111	0.120	0.113	0.115	0.114
	ROUGE-L	**0.118**	0.105	0.114	0.109	0.109	0.108
	METEOR	0.104	0.117	**0.126**	0.096	0.124	0.123
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	0.111	0.108	0.112	**0.115**	0.110	0.110
	ROUGE-L	0.105	0.100	0.104	**0.107**	0.103	0.101
	METEOR	0.083	0.101	0.105	0.088	**0.108**	0.107

Table 15: Zero-shot test set results for text generation using *GPT-4o-mini*. PGraphRAG* denotes no ranked retrieval method of k = 4, while PGraphRAG** represents the second variation where k has no limit to the models context length.

Long Text Generation	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 1: User-Product Review Generation	ROUGE-1	**0.173**	0.168	0.172	0.126
	ROUGE-L	0.124	**0.125**	0.121	0.095
	METEOR	0.150	0.134	**0.152**	0.101
Task 2: Hotel Experiences Generation	ROUGE-1	**0.263**	0.197	0.224	0.211
	ROUGE-L	**0.156**	0.128	0.141	0.130
	METEOR	**0.191**	0.121	0.148	0.147
Task 3: Stylized Feedback Generation	ROUGE-1	**0.226**	0.181	0.177	0.142
	ROUGE-L	**0.171**	0.134	0.125	0.104
	METEOR	**0.192**	0.147	0.168	0.119
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.174**	0.174	0.173	0.146
	ROUGE-L	0.139	**0.141**	0.134	0.117
	METEOR	**0.133**	0.125	0.130	0.110
GPT-4o-mini					
Task 1: User-Product Review Generation	ROUGE-1	**0.186**	0.169	0.168	0.157
	ROUGE-L	**0.126**	0.114	0.113	0.112
	METEOR	**0.187**	0.170	0.173	0.148
Task 2: Hotel Experiences Generation	ROUGE-1	**0.265**	0.217	0.222	0.233
	ROUGE-L	**0.152**	0.132	0.133	0.138
	METEOR	**0.206**	0.161	0.164	0.164
Task 3: Stylized Feedback Generation	ROUGE-1	**0.205**	0.178	0.177	0.168
	ROUGE-L	**0.139**	0.121	0.119	0.117
	METEOR	**0.203**	0.178	0.184	0.160
Task 4: Multilingual Product Review Generation	ROUGE-1	**0.191**	0.164	0.167	0.171
	ROUGE-L	**0.142**	0.123	0.125	0.131
	METEOR	**0.173**	0.155	0.153	0.150

Table 16: Zero-shot Validation set results for long text generation using *LLaMA-3.1-8B-Instruct* and *GPT-4o-mini* on Tasks 1-4.

Short Text Generation	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 5: User Product Review Title Generation	ROUGE-1	**0.125**	0.114	0.111	0.101
	ROUGE-L	**0.119**	0.108	0.105	0.095
	METEOR	**0.117**	0.111	0.104	0.094
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.121**	0.119	0.115	0.115
	ROUGE-L	**0.113**	0.111	0.108	0.107
	METEOR	**0.105**	**0.105**	0.100	0.094
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.132**	0.128	0.127	0.108
	ROUGE-L	**0.128**	0.124	0.122	0.104
	METEOR	**0.129**	0.124	0.118	0.103
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	**0.132**	0.128	0.108	0.127
	ROUGE-L	**0.128**	0.124	0.104	0.122
	METEOR	**0.129**	0.124	0.103	0.118
GPT-4o-mini					
Task 5: User Product Review Title Generation	ROUGE-1	**0.114**	0.106	0.109	0.107
	ROUGE-L	**0.107**	0.100	0.103	0.102
	METEOR	**0.119**	0.115	0.116	0.109
Task 6: Hotel Experience Summary Generation	ROUGE-1	**0.115**	**0.115**	0.114	0.112
	ROUGE-L	0.105	**0.106**	**0.106**	0.103
	METEOR	0.105	**0.106**	**0.106**	0.099
Task 7: Stylized Feedback Title Generation	ROUGE-1	**0.105**	0.101	**0.105**	0.098
	ROUGE-L	**0.102**	0.097	0.101	0.093
	METEOR	**0.118**	0.111	0.118	0.105
Task 8: Multi-lingual Product Review Title Generation	ROUGE-1	**0.108**	0.106	**0.108**	0.103
	ROUGE-L	0.099	0.098	**0.099**	0.095
	METEOR	0.101	0.102	**0.103**	0.095

Table 17: Zero-shot Validation set results for short text generation using *LLaMA-3.1-8B* and *GPT-4o-mini* on Tasks 5-8.

retrieval-augmented generation (RAG) strategies. For example, methods such as in-context prompting (Lyu et al., 2024b), retrieval-based summarization (Richardson et al., 2023), and optimization techniques like reinforcement learning and knowledge distillation (Salemi et al., 2024a) have further refined personalized models. Benchmarks like LaMP (Salemi et al., 2024b) and LongLaMP (Kumar et al., 2024) have been developed to evaluate personalized tasks, emphasizing user-specific history for text generation tasks such as email completion and abstract writing. Retrieval-based approaches, such as (Kim et al., 2020), have also explored personalization by enhancing retrieval pipelines for long-form personalized content generation. However, most existing methods for personalization rely heavily on user history to augment the context or prompt, limiting their effectiveness in scenarios where user history is sparse or unavailable. This reliance poses challenges in real-world applications, particularly for cold-start users. Furthermore, these approaches often overlook the potential of integrating structured data, such as knowledge graphs, to provide richer and more diverse user-specific contexts.

Personalization in NLP

Personalization in natural language processing tailors responses to individual user preferences, behaviors, and contexts, enhancing user interaction and satisfaction. Early work in personalization focused on text generation tasks, leveraging attributes such as review sentiment (Zang and Wan, 2017) and stylistic features (Dong et al., 2017). These approaches, which employed neural networks and encoder-decoder models, laid the groundwork for personalization in text-based systems. Addressing challenges like limited user data, techniques such as warm-attention mechanisms (Amplayo et al., 2018) and social media-derived personalized language models (Huang et al., 2014) were introduced to mitigate the cold-start problem.

Recent advancements have extended personalization to retrieval-augmented generation (RAG) strategies such as prompting (Lyu et al., 2024b), summarization with retrieval (Richardson et al., 2023), and optimization methods like reinforce-

Ordinal Classfication	Metric	PGraphRAG	LaMP	No-retrieval	Random-retrieval
LLaMA-3.1-8B-Instruct					
Task 9: User Product Review Ratings	MAE ↓	0.3272	0.3220	**0.3200**	0.3516
	RMSE ↓	0.7531	**0.7280**	0.7294	0.7972
Task 10: Hotel Experience Ratings	MAE ↓	0.3868	0.3685	**0.3614**	0.4008
	RMSE ↓	0.6989	0.6750	**0.6643**	0.7178
Task 11: Stylized Feedback Ratings	MAE ↓	**0.3356**	0.3368	0.3372	0.3812
	RMSE ↓	0.6856	0.6859	**0.6826**	0.7759
Task 12: Multi-lingual Product Ratings	MAE ↓	0.5228	**0.5216**	0.5282	0.5392
	RMSE ↓	0.8483	**0.8395**	0.8519	0.8704
GPT-4o-mini					
Task 9: User Product Review Ratings	MAE ↓	0.3652	0.3508	**0.3484**	0.4176
	RMSE ↓	0.7125	0.6943	**0.6925**	0.7792
Task 10: Hotel Experience Ratings	MAE ↓	**0.3308**	0.3472	0.3528	0.3640
	RMSE ↓	**0.6056**	0.6394	0.6475	0.6627
Task 11: Stylized Feedback Ratings	MAE ↓	**0.3340**	0.3364	0.3356	0.3972
	RMSE ↓	0.6515	0.6545	**0.6484**	0.7158
Task 12: Multi-lingual Product Ratings	MAE ↓	**0.4568**	0.4832	0.4908	0.4820
	RMSE ↓	**0.7414**	0.7808	0.7897	0.7917

Table 18: Performance comparison on rating prediction tasks (Tasks 9-12) using *GPT-4o-mini* and *LLaMA-3.1-8B-Instruct* on the validation set. Results are reported using MAE and RMSE metrics across retrieval methods.

ment learning and knowledge distillation (Salemi et al., 2024a) have further refined personalized models. Personalization has also been explored for tasks involving user-specific attributes, such as those studied in benchmarks like LongLaMP (Kumar et al., 2024), and retrieval methods for long-form personalized generation (Kim et al., 2020).

In addition to text generation, integrating personalization into recommendation systems has shown success in combining user-specific attributes with retrieval-based frameworks (Tsai et al., 2024). A comprehensive survey on personalization in large language models underscores the importance of robust methodologies for managing diverse and large-scale user data (Zhang et al., 2024). However, current approaches often overlook the potential of structured data, such as knowledge graphs, to enhance personalization.

Knowledge Graphs & Retrieval-Augmented Generation (RAG)
Knowledge graphs have played a pivotal role in natural language processing by providing structured and relational information for tasks such as question answering, reasoning, and retrieval (Schneider et al., 2022; Liu et al., 2018). Their ability to leverage subgraphs for precise and contextually relevant answers has been demonstrated in multi-hop reasoning tasks (Salnikov et al., 2023). Techniques like data synthesis have further improved traversal efficiency and scalability in large graphs (Agarwal et al., 2021).

Retrieval-Augmented Generation (RAG) builds on this foundation by integrating external data sources, such as dense vector indexes and knowledge graphs, into the generation process, significantly improving the factuality and relevance of responses (Izacard and Grave, 2020). When combined with knowledge graphs, RAG models excel in handling complex reasoning tasks, such as multi-hop question answering (Saleh et al., 2024), and in recognizing rare word patterns in previously unseen domains (Mathur et al., 2024). These methods also enhance large language models (LLMs) by reducing hallucinations and improving contextual accuracy (Kang et al., 2023; Chen et al., 2023).

Despite their success, knowledge graphs face scalability challenges, particularly in large-scale applications like recommender systems (Ji et al., 2022). Constructing and maintaining accurate and consistent graphs require refinement techniques to ensure data reliability and relevance (Paulheim, 2017). Comprehensive surveys on knowledge graph technologies emphasize the need for better methodologies for creating, managing, and scaling these structures (Hogan et al., 2021). Additionally, traditional RAG approaches often struggle with irrelevant document retrieval and the inefficiencies of integrating multiple knowledge sources (Gao et al., 2024).

The intersection of knowledge graphs, RAG, and personalization presents a promising avenue for research, enabling models to combine user-centric retrieval strategies with structured knowledge to enhance accuracy and scalability.

Traditional RAG methods, which often rely on vector-based document retrieval, have demonstrated substantial improvements in tasks like combining pre-trained sequence-to-sequence models with dense indexes (e.g., Wikipedia) (Lewis et al., 2021).