Guided Profile Generation Improves Personalization with LLMs

Jiarui Zhang
University of Southern California
jzhang37@usc.edu

Abstract

In modern commercial systems, including Recommendation, Ranking, and E-Commerce platforms, there is a trend towards improving customer experiences by incorporating Personalization context as input into Large Language Models (LLMs). However, LLMs often struggle to effectively parse and utilize sparse and complex personal context without additional processing or contextual enrichment, underscoring the need for more sophisticated context understanding mechanisms. In this work, we propose Guided Profile Generation (GPG), a general method designed to generate personal profiles in natural language. As is observed, intermediate guided profile generation enables LLMs to summarize, and extract the important, distinctive features from the personal context into concise, descriptive sentences, precisely tailoring their generation more closely to an individual's unique habits and preferences. Our experimental results show that GPG improves LLM's personalization ability across different tasks, for example, it increases 37% accuracy in predicting personal preference compared to directly feeding the LLMs with raw personal context.

1 Introduction

Within the context of personalization tasks, personal profiling has been extensively employed. Conventional methodologies typically rely on substantial datasets such as graph-based similarities. These profiles often exhibit 'neighborhoods' and 'relationships' within the data, posing challenges for immediate interpretability without supplementary processing. Recently, LLMs have demonstrated robust capabilities in tasks related to reasoning and generation, leading to a growing interest in leveraging LLMs for personalization services. However, distinguished from other Naturual Language Processing (NLP) tasks, we identify two primary challenges in personalization with LLMs.

The first challenge is the complexity of personal contexts and the sparsity of their key information. For example, a person's distinctive writing style may only be discernible in a small portion of their writing, whereas the remainder of the writing style tends to be more generic. As is shown in recent studies (Liu et al., 2023b), LLMs have challenges in capturing comprehensive information within lengthy contexts, making it easy to overlook the smaller portions that contain distinctive writing styles. Previous studies (Lewis et al., 2020; Salemi et al., 2023) have attempted to address this challenge by context retrieval. However, context retrievers frequently rely on surface-level ranking strategies, such as keyword similarity. Such an approach, while straightforward, may not always align with the nuanced needs of personalization tasks.

The second challenge lies in the balance between generalization and personalization. While LLMs have demonstrated considerable performance on general tasks, they still struggle to generate output that fully aligns with users' desired behaviors and directions (Bang et al., 2023). Rather, they prioritize imitating the majority of their training sets (Karpathy, 2023). Figure 1 illustrates a personalized task involving the paraphrasing of a tweet to match someone's distinctive writing style. From the personal context, it is noticeable that the individual tends to use block letters to emphasize actions and feelings. However, the model closely mirrors the original question input when receiving the personal context and question directly, which can be reachable even without personal context. When we instruct LLM to describe the person's writing style, rather than noticing the spatial use of capitalization, it pays attention to the emotion, and content, which are not our desired 'writing styles'.

Steering LLM outputs precisely is always a challenge. To address it, previous work has attempted to apply reinforcement learning from human feed-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4005–4016

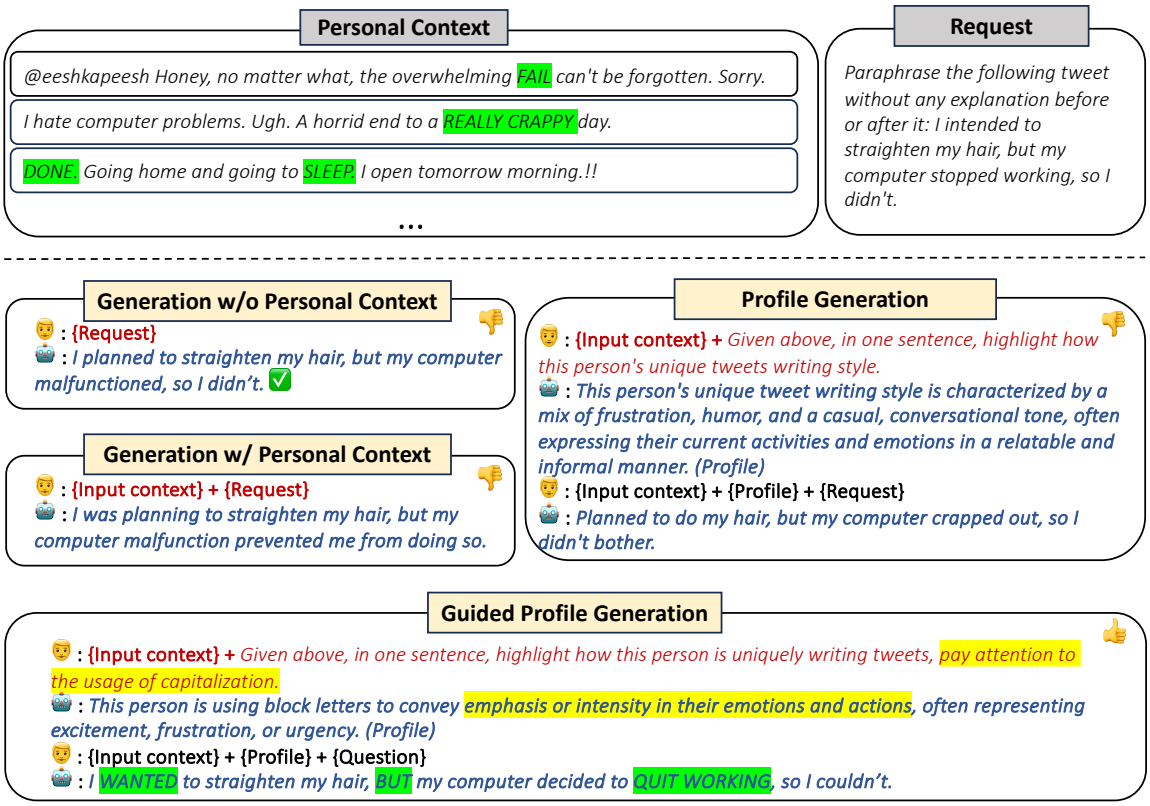

Figure 1: A motivating example. The model is given a personal context reflecting the person's writing style, and the task is to paraphrase a new tweet for the user. We show `gpt-3.5-turbo-1106`'s response under different input conditions. The result shows that generating a descriptive personal profile with proper guidance helps the model finish the personalization better.

back (RLHF) (Ouyang et al., 2022). However, this is a resource-intensive process that might be financially burdensome and impractical for some service providers. Other works tried to train compact models (Li et al., 2023b) from the feedback of comparison between LLM's output and ground truth labels. However, no certain true label is available for independent profile generation tasks. Prompt optimization, involving both manual and automated efforts in designing and selecting suitable prompts for various tasks, stands out as a promising and widely adopted alternative.

The majority of recent studies on prompt optimization indicate that LLMs can benefit from digesting intermediate generated prompts to successfully complete complex tasks (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2022). In personalization, formulating a personal profile serves as a crucial intermediate step that enhances task performance in terms of both accuracy and efficiency. Most existing profile modeling techniques depend on substantial datasets. While these approaches are effective for structured analysis, they often yield profiles that require additional interpretation. Additionally, these profiles tend to be restricted to a limited range of data types, limiting the inclusion of more diverse perspectives. In contrast, natural language is not only inherently understandable and easily diagnosable, but it also enables the expansion of the scope of data types that can be effectively integrated into the modeling process.

In this paper, we propose a general method leveraging LLMs for personalization, named **Guided Profile Generation** (`GPG`), whose goal is to augment LLMs' capacity for interpreting raw personal contexts and to generate high-quality natural language personal profiles. In `GPG`, the process begins with personal context digestion, where we pose specific questions in predetermined directions on personal context. Then the model will generate descriptive natural language personal profiles, steered by the output of last stage. The resulting personal profile will be subsequently employed to respond to the request with downstream models.

We conduct extensive experiments to evaluate the efficacy of `GPG` with `gpt-3.5` on the task of

purchase preference prediction, text paraphrasing, and dialogue response generation and benchmark the performance of GPG with several baselines. Our result shows that GPG consistently enhances the personalization performance across various tasks. In preference prediction of online purchase, GPG improve 37% accuracy in predicting personal preference of product purchasing compared with direct prediction with raw context. In text paraphrasing on Tweet, GPG improves METEOR score by 2.24 by digesting the writing style with the recognition of the most significant writing features. Furthermore, we conduct ablation studies to evaluate the impact of various components within the GPG framework and undertake further analysis to comprehend the limitations of our methods, aiming to pave the way for future directions in this research.

2 Related Work

LLMs have demonstrated robust performance through scaling up, in-context learning (Brown et al., 2020), reinforcement learning from human feedback (Ouyang et al., 2022), and instruction tuning (Wei et al., 2021), making them capable of complex reasoning tasks (Hendrycks et al., 2020; Srivastava et al., 2022; Jiang et al., 2023a, 2024). The performance of the model is sensitive to input and output manners, making prompt optimization (Yao et al., 2022; Wei et al., 2022; Kojima et al., 2022; Huang et al., 2024) a popular topic.

There has been a growing interest in using LLMs for personalization. LLM-Rec (Lyu et al., 2023) utilizes LLMs as recommenders by prompting them with recommendation instructions and employing graph-based engagements. However, this approach lacks emphasis on the crafting of user profiles. LAMP (Salemi et al., 2023) attempts to integrate a context retriever to avoid the need for feeding the entire personal context to LLMs, but the retrieved personal context still proves challenging for LLMs to easily comprehend. PALR (Chen, 2023) uses LLMs to generate user profiles for personalized recommendation and fine-tuned llama (Touvron et al., 2023) to generate ranking. However, the exploration of more effective methods for crafting user profiles in natural language based on personal contexts with diverse structures remains underexplored. Other studies also explore the use of LLMs to augment graph-based recommendation system (Lyu et al., 2023), support human writing creativity (Chakrabarty et al., 2023), personalized writing education (Li et al., 2023a), dialogue systems (Fan and Jiang, 2023) and healthcare assistant (Liu et al., 2023c).

For datasets, LAMP (Salemi et al., 2023) introduces seven language tasks that necessitate personalization. These tasks include tweet paraphrasing and email subject generation, among others. Notably, tweet paraphrasing serves as a comprehensive test bed for evaluating personalized writing style imitation using LLMs. Amazon review (He and McAuley, 2016) provides abundant online purchase history and shopping reviews, enabling the creation of a preference prediction dataset for product purchasing. PER-CHAT (Wu et al., 2021) is an open-domain single-turn dialogue dataset collected from Reddit. In PER-CHAT, each dialogue response is paired with related comment history from the same user, enabling personal profile crafting. Other datasets like MovieLens (Harper and Konstan, 2015), Recipe (Majumder et al., 2019), PERSONA-CHAT (Zhang et al., 2018) are also widely used. We evaluate GPG by personalized preference prediction, tweet paraphrasing, and dialogue generation sets in this paper.

3 Guided Profile Generation

Given a personal context PC, and a task T, the objective of personalization is to align with the individual's behavior and successfully accomplish the task. In contemporary commercial systems, personal profile crafting proves advantages for both accuracy and efficiency, achieved by providing a clear reflection of a person's behavior and ensuring reusability without the need to process the raw context again. Given the impressive capabilities of LLMs, there is a natural inclination to leverage them for integrating raw PC and generating personal profiles. However, our early investigation indicates that these approaches may not achieve the expected performance (Figure 1). Moreover, the lack of human-annotated data for intermediate personal profiles makes direct optimization through fine-tuning a challenging option.

We propose GPG, a general method for personalization with LLMs through personal profile generation. The proposed method of GPG is presented in Figure 2. Different from joint learning with downstream personalization tasks for LLM, which adopts Reinforcement Learning from Human Feedback (RLHF), we adopt a much more cost-effective yet efficient method. This method focuses on gener-

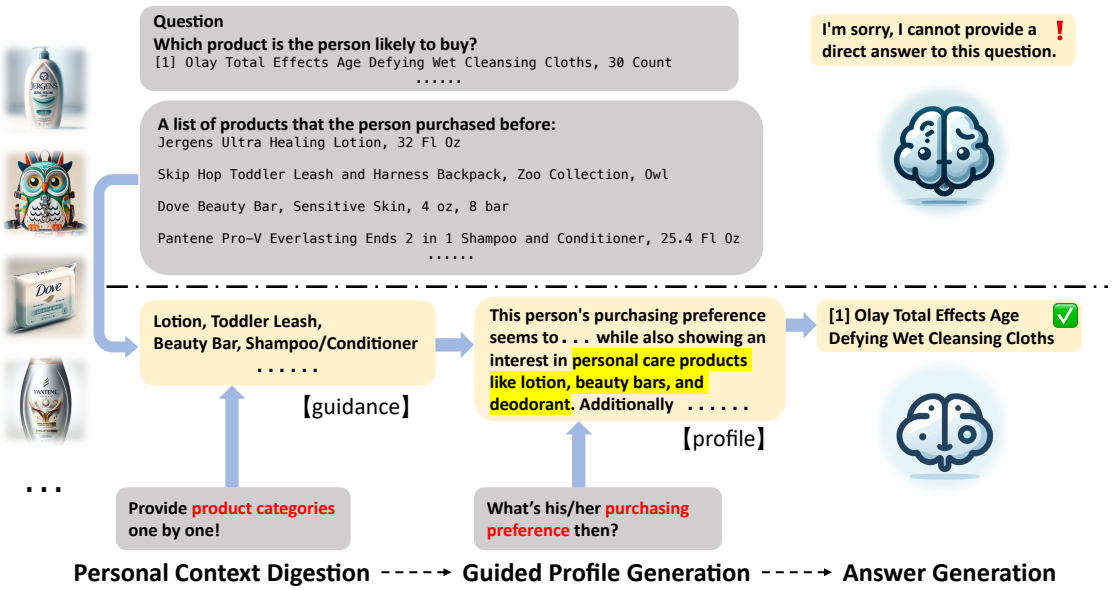

Figure 2: Illustration of GPG described in Section 3: Given a personal context, we instruct LLM to generate a descriptive personal profile via self-guidance. The personal profile is then used to complete the personal task. GPG enables LLM to generate high-quality personal profiles, improving their performance on personalization. Note that our experiments are conducted in **textual domain**, images are for illustrative purposes.

ating a readable, descriptive personal profile based on personal context. Our method consists of the following steps:

3.1 High-level Workflow

The first step is Personal Context Digestion. In this step, we pose task-specific questions to the LLM, guiding it to digest PC in our desired direction. For instance, in the scenario of predicting a customer's preferred product based on their purchase history, we prompt the model to sequentially generate product categories. The main purpose of this step is to get direction and key information for the next step. Note that differentiated from few-shot prompting which needs a large amount of in-context corpus crafted by humans, in GPG, only one specific question is designed for each task.

The second step is Guided Profile Generation. The response of the previous steps serves as guidance for the generation of the personal profile. Similar to (Li et al., 2023b),

we concatenate the PC and guidance as input. We instruct the LLM to generate descriptive sentences serving as the personal profile. In contrast to high-dimensional representations, our profile is explainable, enabling easy diagnosis of inadequacies. Moreover, our profile is language model orthogonal, facilitating broader applications and seamless future development.

The final step is Response Generation. The generated personal profile is used to finish the final task. To provide sufficient information, we do not exclude the raw personal context in our main experiment. In section 5, we conduct a detailed experiment study of the effect of the inclusion of PC and guidance.

4 Evaluation Tasks and Metrics

Our proposed method can be applied to a wide range of personalization tasks to overcome the challenge given by raw personal contexts. In this work, we mainly focus on the task of personalized preference prediction, text paraphrasing, and dialogue continuation.

4.1 Task of Preference Prediction

In commercial systems, accurately predicting a user's preference is one of the most crucial tasks. This prediction holds the potential to benefit various downstream tasks (e.g., personal recommendation). However, reliance on large databases and specific models, like assessing the similarity between different users, poses limitations. The design of these models often restricts access to additional information, such as the full name and detailed product information on the Internet. Furthermore, these large databases are not always readily accessible for common use. In contrast, LLMs exhibit the

capability to process any textual data, providing a means to overcome the aforementioned limitation. In this section, we delve into their ability to predict user preferences relying solely on textual data. Specifically, we choose user-based online purchase history as our focus due to the distinctive personal behaviors evident in this domain.

Specifically, to construct the test bed for user preference prediction, we leverage the Amazon Product Review (He and McAuley, 2016; McAuley et al., 2015) dataset collected from the Amazon website. The dataset provides the purchase history for each of product with categories and users. We extract the purchase history for each of user and keep the product categories. Then we filter out all of the users who have purchased less than 5 categories of product, who are considered as being lack of personal context. For the remaining users, we randomly select one of the purchased product categories with at least 2 products. Then one of the products is selected as a question. To sample the distractors, we randomly select 3 products from the category that this person has never purchased before. We consider the product name to be enough information to identify the person's purchase preference, to the end, we exclude all of the review information in the dataset for simplicity. In the resulting dataset, PC is defined as purchasing history, which is a list of products that the person has purchased before, and the task is to identify the product that is most likely to be purchased by the person, and select the product from four candidate options.

Metrics. Since the dataset is in the form of multiple choice questions, and is designed to be in a balanced set, we take the accuracy as the only metric for this task. Lastly, it is worth noting that the constructed preference prediction dataset mostly serves as a diagnosis purpose, evaluating how we can better utilize LLMs predicting user's preference based on raw context. As is shown in Table 1, a single semantic-level comparison algorithm can reach the highest performance in such data, but will not generalize well when facing different formats of datasets.

4.2 Task of Text Paraphrasing

Though simple for humans, it is underexplored whether LLMs can detect and imitate the text-writing styles for different individuals. Such capability is crucial since in recent times, LLMs have been widely used as writing assistants. In this section, we explore how well can LLMs imitate a person's writing style given the raw PC. Compared to formal writing, such as news reports or research articles, Twitter is a platform where every individual can express their thoughts freely. Hence, we select the text on Twitter as our study focus due to the frequently personalized writing on it, such as punctuation, and abbreviations. Specifically, we use LAMP-7 (Salemi et al., 2023), a user-based Twitter collection based on sentiment140 (Go et al., 2009) dataset. In LAMP-7, one of a user's tweets is selected as the source of task input. Then, this input is fed into an LLM for neutralizing the writing style. In the resulting dataset, PC is defined as the collection of all past tweets that this person had before excluding the selected one. The task is to reconstruct the tweet following this person's writing style based on the neutralized tweet and all other tweets.

Metrics. We consider the word and phrase level usage similarity, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004). Since the task is style reconstruction without semantic-level personalization, we do not evaluate the semantic-level (embedding) similarity.

4.3 Task of Dialogue Response Generation

Besides writing style imitation discussed in section 4.2, the ability of AI assistants to accurately reflect an individual's opinion is also crucial. This task is particularly challenging due to the opinions are often implicit and multifaceted in a raw personal context, and should be selectively employed based on the requirements of different tasks.

We focus on dialogue continuation in practice. In particular, we leverage PER-CHAT (Wu et al., 2021) collected from open-domain discussions on Reddit. PER-CHAT collects each individual's comment history, and the task is to use the history as a signal of personal preference and help the individual answer the question. We do not include the retrieved personal profile from the paper for simplicity. To improve the relevance between the comment history and target response, we measure their semantic similarities based on sentence-transformer (Reimers and Gurevych, 2019), and select a subset having a maximum similarity larger than 0.4. We also exclude instances with max similarities larger than 0.6 to avoid overlap between comment history and target response.

Metrics. We consider semantic level similarity

Table 1: Accuracy comparison of different prompting strategies on amazon preference prediction dataset. Where DG denotes direct generation, PG denotes profile generation directly with language instructions.

Method	Accuracy
Random	25.00
DG w/o PC	31.65
DG w/ PC	47.55
PG	54.98
GPG	**65.08**

Table 2: Performance of different prompting strategies on our selected subset of PER-CHAT data, where **ST** denotes sentence transformer and **BS** denotes Bert-Score.

Method	ST	BS
DG w/o PC	29.86	83.09
DG w/ PC	32.31	83.54
PG	32.66	83.47
GPG	32.35	83.43

metric based on sentence-transformer and BERT-Score (Reimers and Gurevych, 2019; Zhang et al., 2019) as our main metric for evaluation. Since the posted questions are mostly open-ended discussions without definite answers, we do not include metrics for direct string, word or phrase-level comparison.

5 Experiments

We use OpenAI's `gpt-3.5-turbo-1106` as our major LLM all through the tasks; during inference, we keep the temperature at 0 (greedy decoding) to gain a deterministic result and set max_tokens to 100. We report the result with a single run due to the greedy decoding.

5.1 Baselines.

For the comparison purpose, we present the following baselines to illustrate the effectiveness of GPG:

1. **Direct Generation without Personal Context.** (DG w/o PC) We consider the LLMs' native response to the question since they have been trained on numerous corpus. For example, LLMs could have knowledge about the general tweet writing style, thus having the ability to reshape a sentence to such a style. The input is formalized as {Q}.
2. **Direct Generation with Personal Context.** (DG w/ PC) In this baseline, we feed the PC to LLM and ask them directly to generate the answer to our question. The input is formalized as {PC}{Q}.
3. **Unguided Profile Generation.** (PG) In this baseline, we ask LLMs to generate the profile of a person according to PC without further instructions. Then we use the generated profile to finish the personalized task. The input is formalized as {PC}{PP}{Q}, where PP is the profile generated from PC by instructing LLM.

5.2 GPG Specifications.

In the task of Preference Prediction, we guide the LLM to generate the personal profile by providing the product categories. To this end, we first ask the LLM *"Provide the product category of above one by one, each of them use less than 10 words, split by a comma:"*. The resulting list of categories serves as the guidance for LLM in generating the personal profile. After the generation of the personal profile, we concatenate the raw PC, and the personal profile as the final input of LLM, predicting the final answer. We do not include the raw guidance, i.e. purchase category to reduce redundant information. We will discuss the effect of the inclusion of each component in detail in section 6.1.

In the Text Paraphrasing task, the LLMs are guided by a unique aspect of the writing style of the tweets when generating the personal profiles. We identify 4 key aspects of paraphrasing: *Capitalization, Emoji, Abbreviation, Punctuation*. Then we instruct LLM to select the most distinctive features in the personal context, specifically our instruction is: *Among the usage of 1. Capitalization, 2. Emoji, 3. Abbreviation, 4. Punctuation, which is the most distinctive feature of the above tweets?*. Then LLM will generate the profile based on the self-selected category and use the generated profile together with the guidance to finish the task.

In the Dialogue Response Generation Task, We expect the generated personal profile to be a summary of these texting habits and personal opinions. Inspired by the original paper, we instruct LLM to generate the basic personal information from their comment history, the aspects include: *"pets", "family", "residence", "favorites", "partner", "possessions", "gender", "self-description"*. Then the above aspects are used to craft the personal profile.

5.3 Experimental Results

Table 1 shows the performance on our Amazon preference prediction dataset of different prompting strategies. LLM improves its performance by 50.23% when adding the personal context to its input. Furthermore, this improvement can be fur-

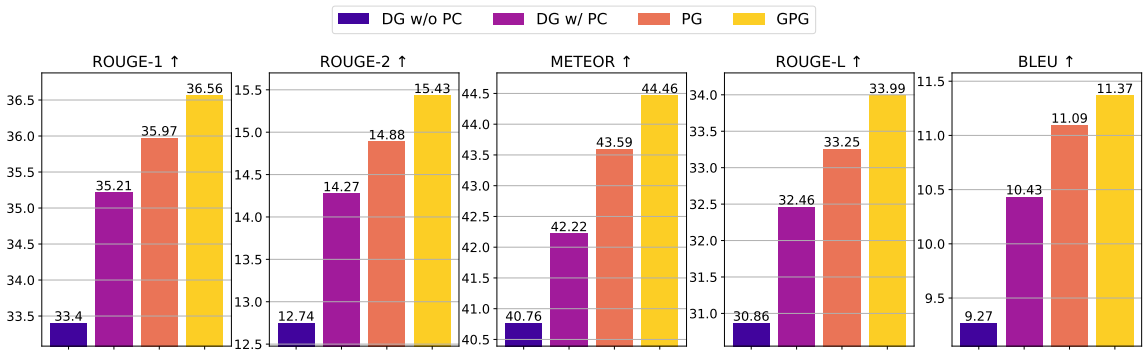

Figure 3: Text paraphrasing on Twitter performance of GPG in comparison with direct generation without personal context (DG w/o PC), direct generation with personal context (DG w/ PC) and Profile Generation (PG).

ther enhanced to 73.71% by using a self-generated personal profile. Our GPG, reaching an improvement of 105.62% through self-guidance.

The result of tweet paraphrasing is shown in Figure 3. Firstly, the inclusion of personal context improves the performance across all metrics, clearly showing the usefulness of personal context in reshaping the users' writing styles. Generating an unguided personal profile further improves the performance compared to direct generation, providing guidance could double such benefit. Such a result indicates the effectiveness of generating a self-guided intermediate profile for personalization of text paraphrasing with LLMs.

On dialogue generation, the inclusion of raw PC has a positive impact on the performance, as is shown in Table 2. However, profile generation, either guided or unguided does not help much in such a task. To understand this phenomenon better, we will look deeper into the generations in section 6.1.

6 Analysis and Discussions

6.1 Ablation Studies

We conduct an ablation study to better understand the benefit of each component of GPG, on preference prediction and text paraphrasing tasks. the result is shown in Table 3. Specifically, we analyze the impact of incorporating personal context (PC), guidance(G, context digestion), and personal profile (PP) during the generation of **final response**. Next, we will provide a detailed analysis based on the result.

Can we exclude raw personal context when generating an answer? In our experiment, we initially incorporated the personal context as part of the input to mitigate the risk of information loss. However, in practice, it is inefficient to keep the personal context as input during every run. To this end, we remove the personal context during the final task generation. Compared with the direct generation, GPG improve the performance by 17.53% (absolute) in predicting purchase preference, generations without raw personal context (sixth-row in Table 3) could approximate 61.04% of such improvement, indicating a considerable trade-off between the expense and performance. However, in text paraphrasing, the performance after removing the raw personal context is worse than a direct generation, underlining the higher importance of personal context in text paraphrasing.

Can personal context digestion directly benefit the downstream tasks? As is shown by our result, personal context digestion can help LLMs generate better descriptive personal profiles. Thus, we are curious whether such a benefit is directly applicable to the final task generation. To this end, we skip the generation of descriptive personal profiles and directly perform downstream tasks after context digestion, the result is shown in the last two rows of Table 3. Surprisingly, the guidance itself is functioning even worse than an unguided personal profile (third row) in both of the tasks, suggesting: **1.** Despite being beneficial in enhancing the generation of personal profiles, the guidance itself is not immediately effective for improving the performance of the final task. **2.** A descriptive personal profile helps the model be better at personalization.

6.2 Error analysis and Observations.

Profile Generation helps LLM be more certain about making selections. We find LLMs frequently opt to abstain from responding when faced with uncertain information. To better understand this behavior of LLMs, we select all of the 'abstain' answers and report the ratios of correct, incorrect,

Table 3: The ablation study on amazon preference prediction (**P-P**) and text paraphrasing (**T-P**) tasks. We consider the inclusion of raw personal context (PC), guidance (G, context digestion), and descriptive personal profile (PP). The best performances are in bold.

	Dataset			P-P	T-P				
	w/ PC?	w/ G?	w/ PP?	Acc	ROUGE-1	ROUGE-2	METEOR	ROUGE-L	BLEU
DG	✓	-	-	47.55	35.21	14.27	42.22	32.46	10.43
	✗	-	-	31.65	33.40	12.74	40.76	30.86	9.27
PG	✓	-	✓	54.98	35.97	14.88	43.59	33.25	11.09
	✗	-	✓	51.86	34.25	13.57	42.04	31.65	9.95
GPG	✓	✗	✓	**65.08**	36.12	15.14	43.87	33.55	11.23
	✗	✗	✓	58.25	33.96	13.43	43.50	31.41	10.10
	✓	✓	✓	61.96	**36.56**	**15.43**	**44.46**	**33.99**	**11.37**
	✗	✓	✓	59.14	35.90	14.62	44.45	33.32	10.81
	✓	✓	✗	51.71	35.69	14.75	43.07	33.11	10.79
	✗	✓	✗	48.44	35.04	13.84	42.52	32.42	10.10

Table 4: The ratio of correct, incorrect, and abstain answers in the amazon preference prediction dataset.

Method	Correct	Incorrect	Abstain
DG w/o PC	27.79	55.27	16.94
DG w/ PC	41.46	32.39	26.15
PG	52.30	34.92	12.78
GPG	64.04	31.20	4.75

and abstained answers in the preference prediction dataset. Specifically, the answer is recognized as abstained if the word 'sorry' is found in the answer. From the result shown in Table 4, we find that the primary improvement of GPG on preference prediction data is from helping the model reduce the ratio of answer abstaining rather than correcting their failures.

6.3 Limitation and Future Works

Integrating multiple aspects personalization. Our experiments are conducted on a single source of personal context. In practice, the complete profile of an individual should be drawn from multiple aspects. For example, a person's purchase preference can be related to their gender, age, habit, or even the weather where they live. Due to the difficulty of cross-platform data collection, most of the off-the-shelf personalization data are from a single source. Constructing datasets containing personal contexts from multiple sources for each individual could be interesting. In addition, it is also challenging to integrate data from multiple aspects. While wisely designed mechanisms like graph contrastive learning (Chen et al., 2023) could potentially incorporate different types of information, unifying graph information into natural language is a lightweight alternative (Zhang et al., 2022; Jiang et al., 2023b), obtaining better explainability at the same time. We believe our findings bring useful insight into this future direction.

Multimodal personalization. Recently, multimodal large language models (MLLMs) (Dai et al.; Liu et al., 2023a) have shown promising capabilities in various tasks. Such advancement opens the possibility of multimodal personalization. For example, an individual's preference for clothes could be highly related to the designs, which are not easily described by text. In such studies, the undesired and generic MLLM outputs could be a problem, applying a visual crop (Zhang et al., 2023) directed by visual search (Wu and Xie, 2023) as a 'guidance' would be interesting. In addition, other modalities such as sound (Meta AI Research, 2023), and sensor data like heart rates (Ni et al., 2019) are also considerable.

7 Conclusion

In this work, we present Guided Profile Generation GPG, a novel method leveraging LLMs for personalization tasks through profile generation and context digestion. We conduct extensive experiments on various personalization tasks, including preference prediction, text paraphrasing, and dialogue continuation. Despite the superior performance, GPG generates a personal profile in pure natural descriptive language, which is interpretable and easily diagnosable. Moreover, we reveal why and how the guidance and descriptive personal profile improve the performance. We hope our research can pave the way for personalization applications with AI models in the future.

References

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.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*.

Mengru Chen, Chao Huang, Lianghao Xia, Wei Wei, Yong Xu, and Ronghua Luo. 2023. Heterogeneous graph contrastive learning for recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 544–552.

Zheng Chen. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*.

Yaxin Fan and Feng Jiang. 2023. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. *arXiv preprint arXiv:2305.08391*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Tenghao Huang, Dongwon Jung, and Muhao Chen. 2024. Planning and editing what you retrieve for enhanced tool learning. *arXiv preprint arXiv:2404.00450*.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2023a. Brainteaser: Lateral thinking puzzles for large language model. *arXiv preprint arXiv:2310.05057*.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2023b. Transferring procedural knowledge across commonsense tasks. In *ECAI 2023*, pages 1156–1163. IOS Press.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2024. Semeval-2024 task 9: Brainteaser: A novel task defying common sense. *arXiv preprint arXiv:2404.16068*.

Andrej Karpathy. 2023. State of gpt: Analyzing and improving the training of large language models. https://karpathy.ai/stateofgpt.pdf.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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*.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023b. Guiding large language models via directional stimulus prompting. *arXiv preprint arXiv:2302.11520*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023c. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, and Jiebo Luo. 2023. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating personalized recipes from historical user preferences. *arXiv preprint arXiv:1909.00105*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Meta AI Research. 2023. Seamless communication. Accessed: 2024-01-03.

Jianmo Ni, Larry Muhlstein, and Julian McAuley. 2019. Modeling heart rate and activity data for personalized fitness recommendation. In *The World Wide Web Conference*, pages 1343–1353.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Jiarui Zhang, Filip Ilievski, Kaixin Ma, Jonathan Francis, and Alessandro Oltramari. 2022. A study of zero-shot adaptation with commonsense knowledge. In *AKBC*.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2023. Visual cropping improves zero-shot question answering of multimodal large language models. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Table 5: One Example from dialogue continuation task, where LLMs face a challenge in closely aligning the personal preference and texting habits.

Profile: This person seems to have strong opinions and enjoys giving advice. They also appear to be direct and straightforward in their communication style, often using humor and sarcasm. *Question (task):* How do you talk to a girl irl after only talking online? *Human (ground truth):* always through mutual acquaintances. Anything else is creepy and stalker like . *LLM response:* Well, first off, be yourself and don't try to be someone you're not. Keep the conversation light and fun, and don't be afraid to throw in some humor. And remember, confidence is key! Good luck!

A Examples of Three Tasks

In Figure 4, we present examples of the three tasks under our test, we include raw personal context, personal context digestion, and personal profile in each example. The prompts for generating personal context digestion and personal profiles can be found in section 5.2.

B Challenges in Open-Ended Personalization Tasks.

When addressing open-ended tasks like dialogue continuation, LLMs encounter more challenges in aligning with personal preferences and texting habits. One example is shown in Table 5, where LLMs are trying to give a generic response to the question rather than a personalized one. This tendency aligns with findings reported in (Karpathy, 2023) that LLM would prioritize imitating the majority of their training data. While such a phenomenon is not bad in itself as it helps LLMs leverage huge amounts of data and obtain impressive capabilities, it is not a desired behavior in personalization.

C Statistics of Three Tasks.

Table 6 presents the statistics of three included tasks. We report the total count of data instances (# Data) and the average number of user activities (# Activities) within each personal context. Specifically, in the Preference Prediction task, # Activities represents the average number of products a user has purchased before. In Text Paraphrasing, it represents the average number of history Tweets. In Dialogue Response Generation, it represents the average number of dialogue responses within the personal context.

Table 6: Statistics of preference prediction (**P-P**), text paraphrasing (**T-P**) and Dialogue Response Generation (**D-G**). We report the total number of data (# data) and the average number of user activities (# Activities) per personal context.

Task	**P-P**	**T-P**	**D-G**
# Data	673	1500	607
# Activities	6.82	17.64	10.00

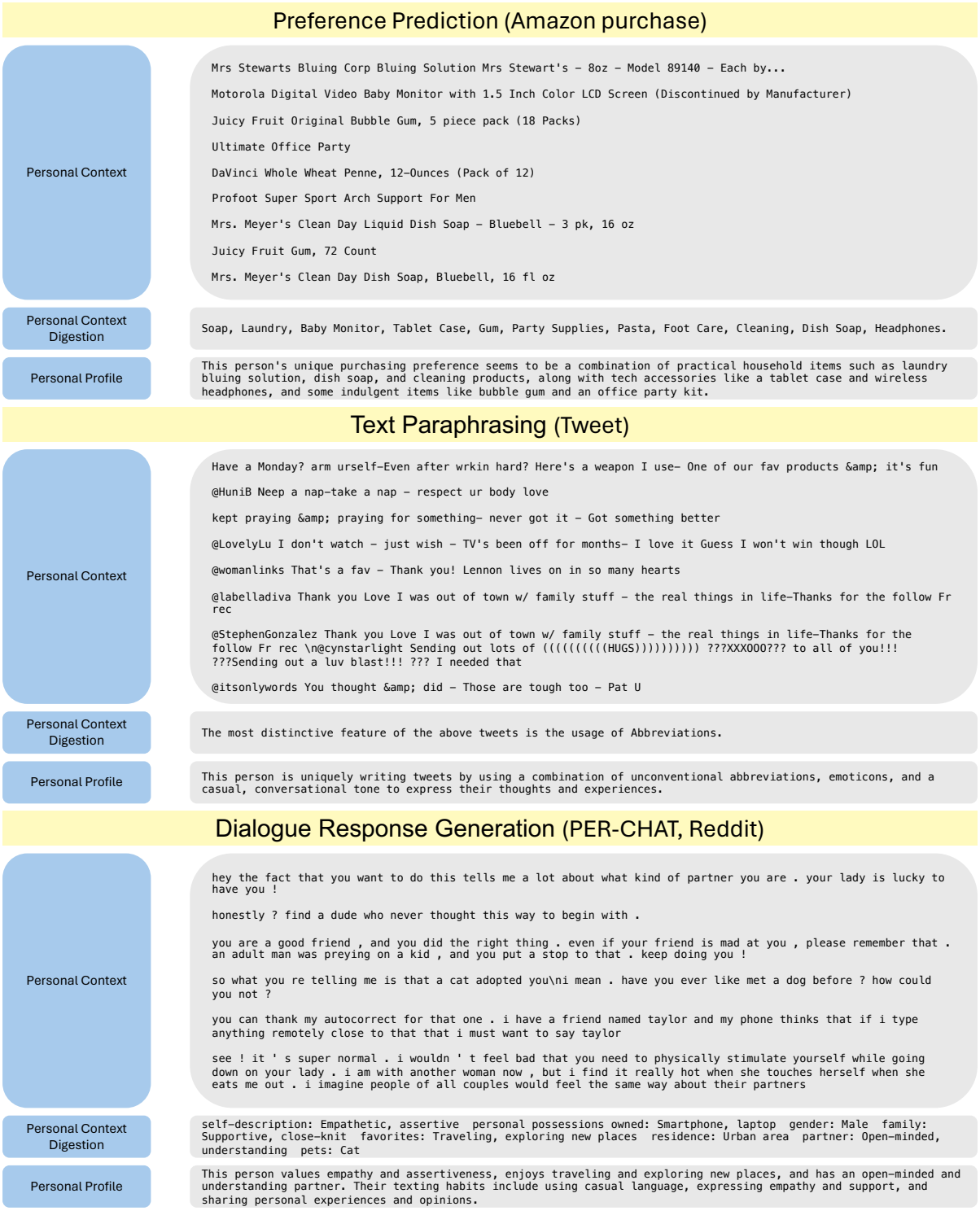

Figure 4: Examples of personal context, personal context digestion, and personal profile of three tasks under our test. We select only part of the personal context due to their length.