HELPFULSUMM: Helpful Personalized Opinion Summarization via Reinforcement Learning from Review Helpfulness Votes

Anonymous ACL submission

Abstract

Personalized opinion summarization (POS) aims to generate a targeted summary of product reviews that is tailored to an individual user's needs and interests. Few existing studies mainly rely on persona representations derived from user-written reviews for personalization, which may not fully capture user interests and fail on cold-start users who have not authored reviews. In real-life social platform, helpfulness votes on reviews represent opinions helpful and of interest to users. In this paper, we develop **HELPFULSUMM**, a reinforcement-learning-based model that utilizes user historical helpfulness votes for alignment with user preference in both *Knowledge Consistency* and *Persona Consistency*, using dual rewards: (i) Helpful Opinion and (ii) Persona Alignment. Experimental results show that HELPFULSUMM outperforms existing persona-based and general opinion summarization approaches and provide more helpful opinions and at higher information and personalization quality. Our source code is available at: https://anonymous.4open.science/r/HELPFULSUMM-A233

1 Introduction

Online reviews have become essential for e-commerce platforms, allowing potential customers to gain insights about a product from past buyers. With the ever-growing prevalence of opinions and increasing numbers of users with varied backgrounds, general opinion summarization (Bražinskas et al., 2020; Amplayo and Lapata, 2020; Tang et al., 2024a,b) has become ineffective to cater to diverse needs of individual users.

Few existing personalized opinion summarization studies relied on user-written reviews history as signals to generate summaries aligned with user persona (e.g., writing styles) (Cheng et al., 2023; Shang et al., 2025). Recently, Zhang et al. (2025) proposes a multi-agent pipeline that role-plays user persona to rewrite general summaries for personalized summaries. However, the persona of user interest is modelled based on the user-written reviews, which might not always be available in practice and may not fully capture all user personal interest.

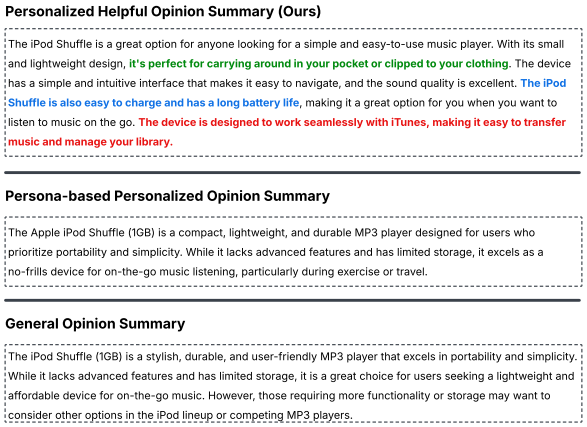

Figure 1: Comparison of HELPFULSUMM and previous summarization approaches. Different colors represent opinions in different aspects. HELPFULSUMM adds helpful opinions and details on product versality (green), "battery" (blue), and "accessibility (with ITunes)" (red).

From an information quality perspective, helpful opinions are those receiving more attention and credit from user (Xiong, 2013; Mudambi and Schuff, 2010; Tay et al., 2020). To this end, user helpfulness votes on reviews were proven to not only to produce more useful summary (Xiong, 2013) but also enhance personalization (Hirsch et al., 2025). Hereafter *helpfulness* refers to historical helpfulness vote on reviews not the direct user feedback on LLM generations (Bai et al., 2022).

In this paper, we proposed HELPFULSUMM, a RL-based model that leverages user's historical review helpfulness votes to ground the end-to-end generation of personalized summary on helpful opinions. HELPFULSUMM combines two complementary rewards – Helpful Opinion and Persona Alignment– to optimize LLM generation. While Helpful Opinion estimates the helpfulness vote

from users on opinions in the generated summary, Persona Alignment scores the summary's language quality against the user persona, i.e., profile, inferred from user's historical data. Figure 1 compares summary generated by HELPFULSUMM with previous approaches. While general summary provides high-level opinions, persona-based personalized summary only rewrites the general summary by focusing more on "portability" and removing critique on "functionality". In contrast, our approach better understands user persona by emphasizing "carrying around your pocket" as specific use case for "portability", while also adds more helpful opinions on "battery", and "accessibility".

Our contribution are:

- We leverage historical user review helpfulness votes to model user interests and ground personalization in helpful opinions for users.
- We propose HELPFULSUMM, the first RL-based model for POS that combines two rewards signal from Helpful Opinion (ensured opinions captured in the generated summary to receive high helpfulness votes from the user) and Persona Alignment (ensured the summary language quality to match with the user persona such as expression habit). Our experiments show that HELPFULSUMM achieves 3.11 times improvement in textual quality, up to 75% more helpful opinions in the summary, and 2.67 times improvement in personalization quality over existing systems.
- We introduce CIAOHELPFUL, the first dataset for training and evaluation of end-to-end models for helpful POS, leveraging user review helpfulness votes for human-LLM collaborative annotation of gold personalized summary.

2 Related Work

2.1 Review Opinion Summarization

Opinion summarization (Chu and Liu, 2019; Bražinskas et al., 2020; Amplayo and Lapata, 2020) generally aims at capturing the salient information from source reviews. Recently, opinion summarization has made progress through Key Point Analysis (KPA), which summarizes reviews into concise, representative statements called key points (KPs) while also quantifying their prevalence as a bullet-like summary (Bar-Haim et al., 2021; Tang et al., 2024a,b). However, these studies has become ineffective to cater to diverse needs of individual users.

Existing personalized opinion summarization studies relied on user-written reviews history as signals to generate summaries aligned with user persona (Cheng et al., 2023; Shang et al., 2025). Recently, leveraging LLM personalization capability, Zhang et al. (2025) proposes a multi-agent pipeline that role-play user suggestion to rewrite a general summary for personalized summary. However, the framework requires rewriting a general summary and is cluttered with many stages, which may not fully capture the user interests and potentially gives rise to cascading errors (Kleinberg et al., 2007). Meanwhile, the persona of user interest is modelled based on the user-written reviews, which might not always be available in practice.

For helpful review summarization, although Xiong (2013) early adopted helpfulness votes to extract helpful reviews for more useful summary generation, they focus on general majority opinions. Still, user-specific review helpfulness votes is left unexplored to advance personalized opinion summarization.

2.2 RLHF for LLMs Personalization & Summarization

Studies on LLMs personalization primarily focus on how to meet user expectations and fulfill their needs. The main approach for aligning LLMs with user intentions typically relies on reinforcement learning from human feedback (Xu et al., 2023; Achiam et al., 2023) (RLHF), where reward models are trained using direct user preferences to improve generic response helpfulness and harmlessness. Recent approaches further explore group-level or collaborative preference modeling (Shen et al., 2025; Choi et al., 2025) to cater to diverse and heterogeneous user groups, i.e., pluralistic alignment (Sorensen et al., 2024). However, existing personalization studies still rely on explicit and direct user feedback on LLM generations for training, and focus on general-purpose alignment tasks rather than summarization. Moreover, the reward modeling still has not incorporate long-term behavioral signals (e.g., review history or social voting), making personalization remain coarse-grained without tailoring the output to individual.

Meanwhile, for summarization, RLHF has also been employed to optimize summary generation (Stiennon et al., 2020; Huang et al., 2024; Gooding and Mansoor, 2023). Previous abstractive summarization studies applied RL to enhance alignment with general human feedback (Stiennon et al.,

2020) or remove polarity bias (Lei et al., 2024) between generated summaries and input comments. However, the use of RL to optimize personalized opinion summarization is still left under-explored.

3 Helpful Personalized Opinion Summarization

Let c denote a category (e.g., Books, Electronics), $R_e = \{r_j\}_{j=1}^{|R_e|}$ denote a set of review comments on a product e from c, a user u, and $H_{u,c}$ the written or voted review history of user u on other products from c. Inspired by the two dimensions (*Knowledge Consistency* and *Persona Consistency*) of Tu et al. (2024) to achieve personalization, our task aims to generate a helpful personalized summary S that **(1)** is grounded on a set of key points (KPs) found helpful by the user based on their preference from $H_{u,c}$ and **(2)** matches with the user's persona (e.g., personality traits, writing style) inferred from $H_{u,c}$ to maximize engagement and resonance. Conceptually, a key point is short salience statement representing an opinion from product reviews(Bar-Haim et al., 2021), while helpful opinions are those receiving more attention and credit from the user (Xiong, 2013). From Figure 1, an example of helpful KP is: "*It's perfect for carrying around in your pocket*", as the user historically vote helpful for reviews emphasizing the "portability" of other products in excercise activities.

4 Human-LLM Collaborative Annotation for the CIAOHELPFUL Dataset

To our best knowledge, there do not exist datasets of annotated personalised summaries of opinions. Existing studies on POS (Zhang et al., 2025) use user-written reviews as proxy personalised summaries, which may not be reliable. Therefore, we construct the CIAOHELPFUL dataset based on *Ciao* (Tang et al., 2012), an European e-commerce corpus. Unlike Amazon (Gupta et al., 2019) or Yelp[1], which lacks or only report total helpful vote counts, Ciao provides helpfulness votes of each user on reviews with fine-grained rating (0-5). From *Ciao*, we only select users u with ≥ 3 votes and ≥ 1 written review per product to ensure opinion diversity, and then sample 50 reviews for each interaction type to form $H_{u,c}$. The final dataset comprises 5,496 samples across 20 categories.

Crafting a comprehensive summary personalized for a user is laborious and time-consuming, if not impossible. Research shows LLM's strong annotation capabilities (He et al., 2024), and so we design a five-stage human-LLM collaborative annotation pipeline, shown in Figure 2. Details of prompts utilized for each stage are provided in Appendix A.

Stage 1: KP Extraction From User Reviews A good personalized summary should be grounded in opinions that users actually find helpful. We therefore treat reviews written or voted by a user on a product ($R_{u,e}$) as ideal source of gold helpful knowledge. Since reviews may contain overlapping opinions, we prompt `GPT-4.1` to extract concise and unique KPs from $R_{u,e}$. Human validation confirms that 95% of user reviews were represented by extracted KPs, while 93.38% of the KPs are verified as valid (Appendix B)

Stage 2: KP Helpfulness Score Calculation As reviews may express mixed opinions and user votes may vary across reviews[2], we calculate user helpfulness votes at the KP level. Following Bar-Haim et al. (2021), we match extracted KPs to reviews using `GPT-4o-mini`, and validate the matches with three MTurk annotators (Bar-Haim et al., 2020). The helpfulness score of each KP is computed as the average rating of its matched reviews.

Stage 3: User Profile Generation Aside helpfulness, a KP should align with the user profile. We therefore prompt `GPT-4.1` to infer a user profile from user reviews history on similar-category products, covering six characteristics (e.g., personality traits from §5.2). For historical voted reviews, we only select those with high rating (≥ 3).

Stage 4: Helpful KP Filtering & Ranking We retain KPs with helpfulness scores ≥ 3 (Stage 2) and further filter them by alignment with the inferred user profile (Stage 3). Using `GPT-4.1`, we keep only KPs aligned with all profile characteristics and ranked them by helpfulness score. (67.93% of candidates are profile-aligned to be helpful).

Stage 5: Personalized Summary Annotation Given the filtered helpful KPs and user profile, we prompt `GPT-4.1` to generate a personalized summary. This draft is iteratively refined using feedback from MTurk annotators on alignment with each profile characteristic (e.g., expression habits). LLMs are preferred for generation, as humans can be subjective and cannot accurately simulate other personas. The core statistics of CIAOHELPFUL are

[1] https://www.yelp.com/dataset

[2] not all opinions in a high-rated reviews are truly helpful

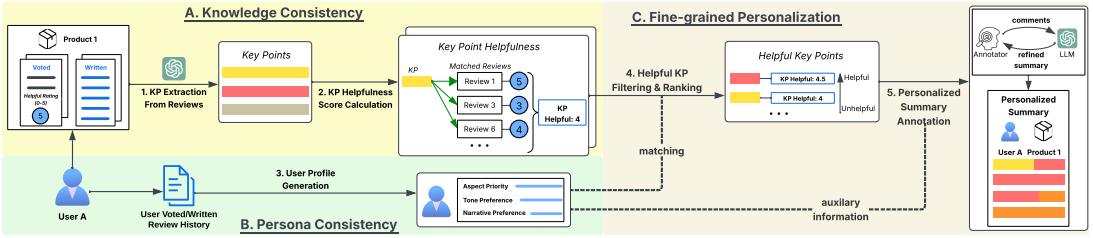

Figure 2: Illustration of the human-LLM collaborative annotation pipeline for CIAOHELPFUL.

Statistic	
# Product Categories	20
# Total Instances	5496
# Products / # User	2787 / 1197
# Reviews Per Product	19.29
# User Voted Reviews Per Instance	05.55
# User Written Reviews Per Instance	01.01
Vote Helpfulness Rating Per Instance	03.60
# KPs Extracted Per Instance (Stage 1)	17.79
# Reviews Matched Per KP (Stage 2)	02.54
KP-Level Helpfulness Per Instance (Stage 2)	03.68
% of Profile-aligned Per Instance (Stage 4)	67.93
# Helpful KPs Per Instance (Stage 4)	09.97
Summary Length (Stage 5)	309.8

Table 1: Core statistics of the train and test set within the CIAOHELPFUL dataset. An instance considers the personalized summarization of a product towards a user.

in Table 1. Notably, users prefers voting over writing gold reviews (5.54 vs 1.01), again highlighting the importance of vote information for annotation.

5 RL with Review Helpfulness Votes

We propose HELPFULSUMM, a RL-based model optimized for POS by leveraging users' historical review helpfulness votes. Following the standard RLHF training paradigm, HELPFULSUMM is trained in two stages. In the first stage of *Supervised Fine-tuning*, we first obtain a base summarizer (**HelpfulSumm-FT**) by instruction-finetuning an LLM with a specialized Chain-of-Thought (CoT) prompt that guides the model to identify (1) the user profile and (2) profile-conditioned helpful key points (KPs) during summary generation (Appendix D and E). In the second stage of *Reinforcement learning*, we further optimize HelpfulSumm-FT by aligning its generated summary $\hat{y}$ with two objectives representing above two personalization metrics (Tu et al., 2024) (Figure 3): (1) *Helpful Opinion*, which encourages inclusion of user-preferred helpful opinions, and (2) *Persona Alignment*, which encourages matching the summary language with user persona (e.g., expression habits) Formally, we define the reward function as:

$$R(x, \hat{y}) = \alpha R_H(x, \hat{y}) + \beta R_P(x, \hat{y}) \quad (1)$$

where R_H and R_P denote the Helpful Opinion and Persona Alignment rewards computed between the generated summary $\hat{y}$ and the user's history $H_{u,c}$. The hyperparameter α, β represent weights for the respective rewards. We refer to the RL-optimized summarizer as **HelpfulSumm-RL**.

5.1 From Review Helpfulness to Helpful Opinion Reward

The purpose of Helpful Opinion Reward is to maximize the helpfulness of opinions (i.e., KPs) included in the summary S with respect to the user's review history $H_{u,c}$. Rather than collecting direct user feedback on KPs in generated summaries, we employ a prediction model to estimate how helpful a user would rate each KP, inspired by previous works on review helpfulness prediction (Chen et al., 2022; Tay et al., 2020). Specifically, given a generated summary S by the base summarizer, we first post-process it into a set of KPs[3]. The helpfulness of each KP is then predicted based on the user's historical reviews. Since a summary may contain multiple KPs, the overall helpfulness score is computed as the average helpfulness across all KPs:

$$R_H(x, S) = \frac{1}{|K|} \sum_{k \in K} f_H(k, H_{u,c}) \in [0, 5]$$

where K denotes the set of extracted KPs and f_H predicts user-preferred helpfulness. Note that the score range aligns with the common (0-5) options for users to rate reviews helpfulness on some platforms (e.g., Ciao), however, the values are continuous to accurately represent helpfulness votes at the opinion level[4]. Note that for this discriminative task, we adopt BERT-based encoders (e.g., DeBERTa, RoBERTa) rather than generative LLMs. Empirically, fine-tuning `deberta-v2-xlarge` yields the strongest performance, outperforming both other encoders and LLM-based scorers to produce score closest to ac-

[3]Using a simple LLM-based extractor prompted with "`Extract all key points from the above generated personalized summary in Python list`".

[4]not all opinions in a review are equally helpful and a user may rate different reviews differently.

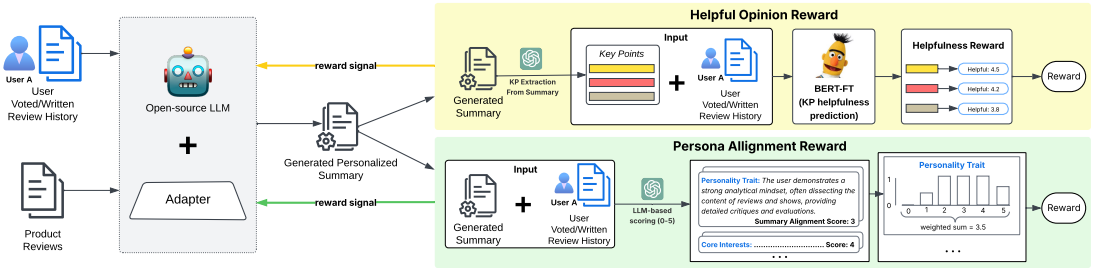

Figure 3: Illustration of personalized summarization with reinforcement learning of HELPFULSUMM.

tual (gold) helpfulness score (Appendix I and Q). Training details are provided in Appendix F.

5.2 From Review Helpfulness to Persona Alignment Reward

The Persona Alignment Reward aims to maximize the alignment between the generated summary S and the user persona inferred from the user history $H_{u,c}$. Specifically, given a generated summary S from the base summarizer, it evaluates how well S reflects six persona characteristics from each user: *personality traits*, *core interests*, *expression habits*, *evaluation priorities*, *shopping behavior*, and *interesting product aspects* (Zhang et al., 2025). Note that we adopt LLM rather than BERT-based encoder, due to its generative and reasoning capabilities to infer the user profile as intermediate content to score each persona characteristic in this reward[5].

Specifically, inspired by `G-Eval` (Liu et al., 2023), which suggest using LLMs as reference-free metrics for NLG evaluation, we adapt this evaluator to the Persona Alignment evaluation task, by prompting LLM to (1) infer the user profile for expected characteristics and (2) score the alignment of S over all profile characteristics. Since LLMs is better at scoring discrete values, the intuition is prompting the LLMs to score the generated summary and profile characteristics multiple times and then take the average of all runs to achieve fine-grained, continuous scores. Formally, given a set of persona characteristics C and multiple runs $r \in \{1, \ldots, R\}$, the score for each run is:

$$s^{(r)} = \frac{1}{|C|} f_{\text{LLM}}^{(r)}(S, H_{u,c}, C)$$

and the final Persona Alignment Reward is:

$$R_P(x, S) = \frac{1}{R} \sum_{r=1}^{R} s^{(r)} \in [0, 5]$$

where f_{LLM} denotes the LLM-based evaluator. Prompt templates are provided in Appendix J.

[5]we assume that in practice, the user profile is unavailable

6 Experiment Setup

6.1 Implementation Details

HELPFULSUMM was experimented with two LLMs: `Llama-3.1-8B-Instruct`[6] (`Llama`) and `Mistral-7B-Instruct-v0.3`[7] (`Mistral`). To ensure runtime and cost feasibility, we sample 500 instances, by selecting top 25 samples (by review count) from each of the 20 product categories (e.g., Books) from CIAOHELPFUL Then. for each product category, 20 samples (80%) are used for training (400 total), and 5 (100 total) for evaluation.

HelpfulSumm-RL and **HelpfulSumm-FT** are implemented with TRL (von Werra et al., 2020) and LoRA (Hu et al., 2021) for parameter-efficient training. For **HelpfulSumm-RL**, we sample 1,000 extra instances (outside train/test sets) from CIAOHELPFUL and train RL on HelpfulSumm-FT using the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017), with the weights α, β for Helpful Opinion and Persona Alignment reward in equation (1) are 0.6, 0.4. We adopt the Kullback-Leibler (KL) divergence between the training policy and the reference policy in regularization and set the penalty coefficient to 0.2.

6.2 Baselines

We benchmark HELPFULSUMM against existing multi-agent personalized opinion summarization and general opinion summarization baselines.

HelpfulSumm-ICL Base LLMs ICL-prompted for helpful personalized opinion summarization using HELPFULSUMM's CoT prompt instruction (Appendix D). Few-shot training instances are concatenated with test instances, with the number of few-shot examples optimized for context length: 3-shot for `Llama`, and 2-shot for `Mistral`.

Rehearsal (Zhang et al., 2025) A multi-agent

[6]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

pipeline that first generate a general summary from product reviews, then rewrite the summary via a role-played user agent and supervisor for continuous evaluation and improvements. The baseline is implemented in two variants: **W(ritten) + V(oted)** and **W(ritten)** user reviews history as input.

PersonaSumm ICL-prompting LLM in one-stage to first analyze the user's interests based on written review history and then generate a personalized summary based on the user's interests and the product review set. This baseline resembles the persona-based personalized opinion summarization models (Cheng et al., 2023; Shang et al., 2025).

GeneralSumm ICL-prompting LLM for general opinion summarization that capture majority opinion reviews, without receiving any personalized instruction or input from user reviews history.

7 Results

7.1 How is the overall textual quality of summaries?

In this experiment, we aim to automatically evaluate the textual quality of the generated summary, at both the KP, i.e., opinion, and summary level, against ground truth from CIAOHELPFUL. Helpful KPs selected from the user gold voted/written reviews on the product (Stage 4) and the annotated personalized summary (Stage 5) from CIAOHELPFUL (§4) are utilized as reference summary and KPs for this automatic evaluation. We post-process the generated summary into a list of KPs using LLM-based extractor as in §5.1.

We first assess *lexical quality* using `ROUGE` (Lin, 2004) at both summary and KP levels, where KP-level ROUGE is computed by concatenating all generated KPs and all reference KPs. We then evaluate *semantic quality* using `BLEURT` (Sellam et al., 2020), `BERTScore` (Zhang et al., 2020), and `BARTScore` (Yuan et al., 2021), in which we calculate soft-Precision/Recall/F1 (sP, sR and sF1) at the KP level following Li et al. (2023). While sP finds the reference KP with the highest similarity score for each generated KP, sR is vice-versa, and ($sF1$) is the harmonic mean between sP and sR.

$$sP = \frac{1}{n} \times \sum_{\alpha_i \in \mathcal{A}} \max_{\beta_j \in \mathcal{B}} f(\alpha_i, \beta_j) \quad (2)$$

$$sR = \frac{1}{m} \times \sum_{\beta \in \mathcal{B}} \max_{\alpha \in \mathcal{A}} f(\alpha_i, \beta_j) \quad (3)$$

where f computes similarities between two individual key points using defined semantic similarity metrics, $\mathcal{A}$, $\mathcal{B}$ is the set of generated and reference KPs and $n = |\mathcal{A}|$ and $m = |\mathcal{B}|$, respectively.

Results Table 2 reports the summary and KP-level textual quality across baselines, with KP-level performance exhibiting clearer distinction. Overall, HELPFULSUMM consistently outperforms other baselines in all variants (-RL, -FT, -ICL), largely due to our CoT prompt that guides LLMs to analyze user profiles and infer helpful opinions from voting history. This also explains HelpfulSumm-ICL's advantage over PersonaSumm, although both analyses user profile in ICL setting for summary generation. Importantly, thanks to optimization signals from Helpful Opinion and Persona Alignment rewards, HelpfulSumm-RL surpasses HelpfulSumm-FT and all baselines, with up to 3.11 times ROUGE-2 gain (0.028 vs. 0.009) and a 0.23 absolute boost in BARTScore (0.78 vs. 0.51). `Llama` also outperforms `Mistral` as backbone for HelpfulSumm due to stronger modeling capability.

For Rehearsal baselines, incorporating vote signal into input user review history leads up to 23% gains. However, these multi-agent baselines still cannot outperform HelpfulSumm, as LLMs can become lazy to refine the summary after multiple turns of role-played user suggestions due to long context. Lastly, GeneralSumm, without personalization input and instruction, yields the lowest textual quality due to generic summaries with redundantly unhelpful opinions. Our manual evaluation of KP information quality further validates the above findings, as shown by the Bradley Terry scores in Table 8 (Appendix K).

7.2 Do summaries contain helpful opinions?

We evaluate the summary helpfulness, i.e., how opinions (KPs) in the summary are more credible and capture more atention from the user, by estimating the quality and coverage of helpful KPs in the generated summary against ground truth from CIAOHELPFUL, using two metrics: *Summary Helpful KP Proportion* (SHKP) and *Summary Helpfulness Score* (SHS). **SHKP** computes the proportion of helpful KPs in the generated summary that match ground-truth helpful KPs from CIAOHELPFUL (Precision, Recall). Specifically, for each instance, we prompt `GPT-4.1` to match KPs in the generated summary with reference KPs in pairwise (Appendix N). A KP is considered to be helpful only if it matches with at least one reference KP. Empirical validation shows `GPT-4.1` annotations

	ROUGE						BERTScore				BARTScore				BLEURT			
	KP			Summ			KP			Summ	KP			Summ	KP			Summ
	R-1	R-2	R-L	R-1	R-2	R-L	sP	sR	sF1		sP	sR	sF1		sP	sR	sF1	
HelpfulSumm−RL																		
+ Llama	**0.152**	**0.028**	**0.139**	**0.400**	**0.087**	**0.176**	**0.49**	**0.45**	**0.47**	**0.25**	**0.72**	**0.78**	**0.75**	**0.58**	**0.37**	**0.37**	**0.37**	**0.48**
+ Mistral	0.137	0.024	0.125	0.358	0.073	0.160	0.44	0.41	0.43	0.15	0.70	0.73	0.71	0.56	0.36	0.36	0.36	0.46
HelpfulSumm−FT																		
+ Llama	0.147	0.023	0.133	0.375	0.085	0.166	0.43	0.42	0.43	0.22	0.70	0.76	0.73	0.56	0.36	0.36	0.36	0.47
+ Mistral	0.131	0.021	0.121	0.310	0.058	0.151	0.42	0.40	0.41	0.14	0.68	0.71	0.70	0.55	0.35	0.35	0.35	0.45
HelpfulSumm−ICL																		
+ Llama	0.141	0.019	0.126	0.367	0.079	0.161	0.36	0.36	0.36	0.18	0.67	0.74	0.71	0.55	0.36	0.33	0.35	0.45
+ Mistral	0.118	0.013	0.106	0.315	0.056	0.147	0.34	0.35	0.34	0.14	0.63	0.67	0.65	0.53	0.32	0.32	0.32	0.44
Rehearsal (W + V) (Zhang et al., 2025)																		
+ gpt-4o	0.126	0.017	0.111	0.346	0.087	0.154	0.38	0.43	0.40	0.14	0.69	0.75	0.72	0.55	0.34	0.34	0.34	0.43
+ gpt-4.1-mini	0.120	0.015	0.126	0.334	0.065	0.155	0.36	0.42	0.38	0.12	0.68	0.72	0.70	0.53	0.33	0.33	0.33	0.42
Rehearsal (W) (Zhang et al., 2025)																		
+ gpt-4o	0.120	0.015	0.106	0.361	0.088	0.150	0.36	0.44	0.39	0.12	0.68	0.76	0.72	0.55	0.33	0.33	0.33	0.42
+ gpt-4.1-mini	0.121	0.016	0.108	0.337	0.067	0.147	0.34	0.43	0.38	0.11	0.68	0.72	0.70	0.53	0.32	0.33	0.32	0.42
PersonaSumm (Zhang et al., 2025)																		
+ gpt-4o	0.115	0.016	0.105	0.359	0.061	0.141	0.36	0.43	0.38	0.11	0.69	0.72	0.70	0.52	0.31	0.32	0.31	0.39
+ gpt-4.1-mini	0.109	0.014	0.099	0.332	0.067	0.148	0.36	0.41	0.37	0.10	0.67	0.68	0.68	0.51	0.30	0.31	0.30	0.38
GeneralSumm (Zhang et al., 2025)																		
+ gpt-4o	0.082	0.011	0.077	0.260	0.061	0.137	0.34	0.41	0.37	0.10	0.63	0.54	0.58	0.44	0.24	0.28	0.26	0.38
+ gpt-4.1-mini	0.080	0.009	0.075	0.255	0.058	0.136	0.33	0.40	0.36	0.09	0.62	0.51	0.56	0.43	0.23	0.27	0.25	0.37

Table 2: KP summary textual quality. sP, sR and sF1 refer to Soft-Precision, Soft-Recall, and Soft-F1 respectively based on set-level evaluation method against reference KPs in gold answer.

highly correlated with MTurk workers' judgement (Pearson's r = 0.836) (Appendix O). **SHS** measures the averaged helpfulness score of opinions in a summary (§5.1). Specifically, for every opinion, we score its helpfulness using either our fine-tuned DeBERTa model (§5.1) or the LLM-based scorer as evaluators (Appendix H). Empirically, results reported by these two models for SHS show high correlation and consistency with third-party multi-objective alignment reward model adapted for predicting opinion helpfulness (Appendix P and Q).

Results Table 3 reports SHKP and SHS across all baselines. Overall, SHS strongly correlates with SHKP Precision, as summaries containing higher proportion of helpful KPs are more likely to receive higher helpfulness scores. For SHKP, we notice trade-off between Precision and Recall. Baselines such as PersonaSumm and GeneralSumm achieve relatively high Recall by covering diverse KPs, but suffer from low Precision as they failed to select truly helpful opinions, and cannot achieve balanced Precision/Recall and high F1 as HelpfulSumm-RL or Rehearsal. Specifcally, HelpfulSumm-RL substantially outperforms all baselines, capturing more 75% of helpful KPs and up to 48% of reference helpful KPs, with up to 1.2-point improvement in SHS. largely thanks to our novel CoT prompt and RL optimization that effectively capture and grounds the summary generation on helpful opinions of high interest by the user. Notably, incorporating vote signals is again validated to yield more helpful summary for Rehearsal, with 5% gains in SHKP Precision.

	SHKP			SHS (0-5)	
Model	**P**	**R**	**F1**	DeBERTa	GPT-4.1-Scorer
HelpfulSumm−RL					
+ Llama	**0.839**	**0.831**	**0.835**	**3.72**	**3.65**
+ Mistral	0.731	0.723	0.727	3.60	3.50
HelpfulSumm−FT					
+ Llama	0.735	0.796	0.764	3.58	3.51
+ Mistral	0.573	0.776	0.659	3.19	3.13
HelpfulSumm−ICL					
+ Llama	0.672	0.660	0.671	3.35	3.30
+ Mistral	0.564	0.619	0.590	3.18	3.12
Rehearsal (W + V) (Zhang et al., 2025)					
+ gpt-4o	0.600	0.648	0.623	3.10	3.06
+ gpt-4.1-mini	0.608	0.635	0.621	3.11	3.09
Rehearsal (W) (Zhang et al., 2025)					
+ gpt-4o	0.556	0.685	0.614	3.02	2.75
+ gpt-4.1-mini	0.586	0.635	0.610	3.08	2.84
PersonaSumm (Zhang et al., 2025)					
+ gpt-4o	0.554	0.642	0.595	3.01	2.68
+ gpt-4.1-mini	0.590	0.572	0.581	3.09	2.75
GeneralSumm (Zhang et al., 2025)					
+ gpt-4o	0.479	0.650	0.552	2.89	2.48
+ gpt-4.1-mini	0.540	0.562	0.551	2.60	2.45
r / ρ **with SHKP Precision**				0.89 / 0.96	0.84 / 0.98

Table 3: Summary Helpful KP Proportion (**SHKP**) and Summary Helpfulness Score (**SHS**) (0-5). r denotes Pearson while ρ denotes Spearman correlation.

7.3 How is the personalization quality of the summary?

Previous work only evaluates *information quality*, i.e., Knowledge Consistency, of the summary in terms of aspect coverage and sentiment consis-

	PT	CI	EH	EP	SB	IA
HelpfulSumm RL	**21.23**	**17.76**	**19.49**	**22.44**	**18.52**	**18.58**
HelpfulSumm FT	12.84	14.96	15.40	15.68	15.12	12.02
HelpfulSumm ICL	7.94	12.23	14.10	9.77	13.46	8.67
Rehearsal (V + W)	18.83	16.30	17.83	19.29	17.47	18.05
Rehearsal (W)	14.87	14.13	15.4	13.14	16.02	15.61
PersonaSumm	14.45	13.73	10.51	11.02	11.30	14.32
GeneralSumm	9.84	10.89	7.26	8.66	8.11	12.74

Table 4: Human evaluation of summary quality by different criteria. Reported are the Bradley Terry scores of 6 criteria on **Summary Personalization Quality**, from left to right, PERSONALITYTRAITS, COREINTEREST, EXPRESSIONHABIT, EVALUATIONPRIORITY, SHOPPINGBEHAVIOR, INTERESTINGASPECTS.

tency (Zhang et al., 2025), without further assessing the *personalization quality*, i.e. Persona Consistency, against user profiles. We address this gap by manually evaluating summary's personalization to user profile on 6 criteria: PERSONALITYTRAITS, COREINTEREST, EXPRESSIONHABIT, EVALUATIONPRIORITY, SHOPPINGBEHAVIOR, INTERESTINGASPECTS. Specifically, we hire workers from Amazon Mechanical Turk (MTurk) to compare summaries of different systems by different criteria, beforing ranking them using Bradley-Terry model (Friedman et al., 2021). Annotation details and evaluation criteria are in Appendix R. Annotation examples are in Appendix S. For reasonable cost, we select samples with top helpful KPs from 5 popular categories for evaluation: *DVDs*, *Food & Drink*, *Shopping*, *Beauty*, and *Entertainment*.

Results Table 4 reports Bradley-Terry scores across eleven quality criteria, which validates our earlier findings. For Personalization Quality, HelpfulSumm-RL again excels in producing personalized summary that best aligns with all characteristics of the user profile, achieving up to 2.67 times improvement, notably on PERSONALITYTRAIT, EVALUATIONPRIORITIES, and EXPRESSIONHABITS. Notably, for the Rehearsal multi-agent baselines, it is manually validated that incorportating vote signal into the input review history enhances alignment with user's PERSONALITYTRAITS and EVALUATIONPRIORITIES, even outperforming HelpfulSumm prior to RL-based helpful personalized summarization optimization.

7.4 Ablation Study

We performed ablation study of HelpfulSumm-RL by excluding either Helpful Opinion or Persona Alignment Rewards from RL training (Appendix T). Overall, having both reward signals are crucial, but *Helpful Opinion Reward* is more important (higher lexical/semantic similarity from Table 12, higher SHKP and SHS from 13). For KP Information Quality, Helpful Opinion Reward enhances COVERAGE and REDUNDANCY, but struggles with SENTIMENT and INFORMATIVENESS compared to Persona Alignment, likely due to weak alignment with user persona (e.g., personality traits). Nevertheless, Helpful Opinion Reward also indirectly improves Personalization Quality, by aligning better with user interests, particularly in COREINTERESTS and EVALUATIONPRIORITY.

7.5 Case Studies

We conducted case studies to evaluate the information and peronalization quality of personalized summary for a "Beauty" product (Table 16 in Appendix U). Overall, HelpfulSum-RL stands out for generating personalized summaries with more helpful opinions and alignment with user persona. While Rehearsal and GeneralSumm mention core product features (e.g., "*comfort*", "*lubricating strip*"), they are overly generic with mixed sentiments and lacks personalization. Lastly, while HelpfulSumm-FT still offers surface-level personalization on aesthetics aspects (e.g., "*chrome design*"), HelpfulSumm-RL focuses more on realistic aspects (e.g., "*price*", "*blade*"), which presents opinions in a highly descriptive analytical tone with empirical illustration tailored to user characteristics and concerns (e.g., "*thick and dark beards*", "*$1.65 a week*", "*half a dozen shave ... no signs of ... losing its edge*", "*first blade removes... second blade ... third blade administers the coup de grace*").

8 Conclusion

In this paper, we leverage historical review helpfulness votes of individual users to better model their interests and ground personalized opinion summarization (POS) in helpful opinions to them. We propose HELPFULSUMM, a RL-based model optimized for POS that align with user preference in both *Knowledge Consistency* and *Persona Consistency*, using two reward models: (i) Helpful Opinion (ensured opinions captured in the summary to receive high helpfulness votes from the user), and (ii) Persona Alignment (ensured the summary language quality to match with the user persona such as expression habit). Experimental results show that our model produces more helpful opinions and enhances both the information and personalization quality of the personalized summaries.

Limitations

- For this research, we only performed experiment on data constructed from the Ciao dataset because, to our best knowledge, Ciao is the only publicly available data offering information on the helpfulness votes of individual user on reviews, which is essential for experiments.

- Due to privacy constraint for accessing user personal data, we can only employ crowed workers to assess the summary's personalization quality according to the user profile.

- Due to cost consideration, we can only perform limited crowd-sourced annotation for personalization evaluation

Ethics Statement

We have applied ethical research standards in our organization for data collection and processing throughout our work.

The CIAOHELPFUL dataset used in our experiments was publicly crowdsourced and released for the research publication for modelling of user trust and online product review helpfulness (Tang et al., 2012; Ocampo Diaz and Ng, 2018). The dataset was published following their ethical standard, after removing all personal information. The answers to questions do not contain contents that are harmful to readers.

We ensured fair compensation for crowd annotators on Amazon Mechanical Turk. We setup and conducted fair payment to workers on their annotation tasks/assignments according to our organization's standards, with an estimation of the difficulty and expected time required per task based on our own experience. Especially, we also made bonus rewards to annotators who exerted high-quality annotations in their assignments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key Point Analysis of business reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3376–3386, Online. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS: THE METHOD OF PAIRED COMPARISONS. *Biometrika*, 39(3-4):324–345.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8922, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin Cheng, Shen Gao, Yuchi Zhang, Yongliang Wang, Xiuying Chen, Mingzhe Li, Dongyan Zhao, and Rui Yan. 2023. Towards personalized review summarization by modeling historical reviews from customer and product separately. *arXiv preprint arXiv:2301.11682*.

Youngbin Choi, Seunghyuk Cho, Minjong Lee, Moon-Jeong Park, Yesong Ko, Jungseul Ok, and Dongwoo Kim. 2025. CoPL: Collaborative preference learning for personalizing LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12886–12904, Suzhou, China. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sian Gooding and Hassan Mansoor. 2023. The impact of preference agreement in reinforcement learning from human feedback: A case study in summarization. *arXiv preprint arXiv:2311.04919*.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. 2019. Amazonqa: A review-based question answering task. *arXiv preprint arXiv:1908.04364*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Sharon Hirsch, Lilach Zitnitski, Slava Novgorodov, Ido Guy, and Bracha Shapira. 2025. Graph meets llm for review personalization based on user votes. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2948–2958, New York, NY, USA. Association for Computing Machinery.

Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. 2024. The n+ implementation details of rlhf with ppo: A case study on tl; dr summarization. *arXiv preprint arXiv:2403.17031*.

Manav Kapadnis, Sohan Patnaik, Siba Panigrahi, Varun Madhavan, and Abhilash Nandy. 2021. Team enigma at ArgMining-EMNLP 2021: Leveraging pre-trained language models for key point matching. In *Proceedings of the 8th Workshop on Argument Mining*, pages 200–205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jon Kleinberg et al. 2007. Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic game theory*, 24:613–632.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity calibration for opinion summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. 2023. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14064–14080, Toronto, Canada. Association for Computational Linguistics.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Susan M Mudambi and David Schuff. 2010. Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708, Melbourne, Australia. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shuo Shang, Xin Cheng, Yiren Xiong, Feng Guo, Shen Gao, Xiuying Chen, Feng Wang, Yongbo Wang, Dongyan Zhao, and Rui Yan. 2025. Personalized review summarization by using graph-based retrieval augmemted generation. *IEEE Transactions on Knowledge and Data Engineering*, 37(11):6582–6596.

Jingyan Shen, Jiarui Yao, Rui Yang, Yifan Sun, Feng Luo, Rui Pan, Tong Zhang, and Han Zhao. 2025. MiCRo: Mixture modeling and context-aware routing for personalized preference learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17458–17474, Suzhou, China. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

An Tang, Xiuzhen Zhang, and Minh Dinh. 2024a. Aspect-based key point analysis for quantitative summarization of reviews. In *18th Conference of the European Chapter of the Association for Computational Linguistics*.

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, and Erik Cambria. 2024b. Prompted aspect key point analysis for quantitative review summarization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.

Jiliang Tang, Huiji Gao, and Huan Liu. 2012. mtrust: discerning multi-faceted trust in a connected world. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 93–102, New York, NY, USA. Association for Computing Machinery.

Wenyi Tay, Xiuzhen Zhang, and Sarvnaz Karimi. 2020. Beyond mean rating: Probabilistic aggregation of star ratings based on helpfulness. *Journal of the Association for Information Science and Technology*, 71(7):784–799.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Wenting Xiong. 2013. Helpfulness-guided review summarization. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 77–83, Atlanta, Georgia. Association for Computational Linguistics.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language agents with reinforcement learning for strategic play in the werewolf game. In *Forty-first International Conference on Machine Learning*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yanyue Zhang, Yulan He, and Deyu Zhou. 2025. Rehearse with user: Personalized opinion summarization via role-playing based on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15194–15211, Vienna, Austria. Association for Computational Linguistics.

A Prompts for Human-LLM Collaborative Annotation of CIAOHELPFUL

A.1 Prompt for Key Point Extraction from User-Voted Product Reviews (Stage 1)

We present the zero-shot prompt for extracting key points (KPs) from gold user-voted product reviews

(Stage 1) of CIAOHELPFUL in Listing 1.

A.2 Prompt for Voted Key Point-Review Matching for Personalized Helpful Opinion Calculation (Stage 2)

We present the zero-shot prompt for matching personalized KP of a user towards their voted product reviews for Personalized Helpful Opinion Calculation (Stage 2) of CIAOHELPFUL in Listing 2.

A.3 Prompt for User Profile Generation (Stage 3)

We present the zero-shot prompt for generating user profile based on their gold voted/written review history (Stage 3) of CIAOHELPFUL in Listing 3.

A.4 Prompt for KP-Profile Matching for Helpful KP Filtering (Stage 4)

We present the zero-shot prompt for matching helpful KP (voted by the user) with the user profile for helpful KP Filtering & Ranking (Stage 4) of CIAOHELPFUL in Listing 4.

A.5 Prompt for Silver Annotation and Collaborative Refinement of Personalized Summary (Stage 5)

We present the zero-shot prompt for silver-annotating personalized summaries grounded on helpful KP and user profile and refining personalized summaries based on human annotator comments (Stage 5) of CIAOHELPFUL in Listing 5 and 6.

B Human Validation of GPT4's KP Extraction from User Product Reviews (Stage 1) of CIAOHELPFUL

In this experiment, we empirically validate the performance and credibility of `GPT-4.1` in extracting KPs from user reviews voted/written for a product from CIAOHELPFUL (Stage 1 of §4). Specifically, to maintain reasonable cost, we randomly sampled a (user-product) instance from 5 common product categories of CIAOHELPFUL [8], totaling 5 questions, and hired workers to annotate whether the extracted KPs matches original user reviews of the sampled instances, which is inspired by the KP Matching evaluation of Bar-Haim et al. (2021). More specifically, for a given query, we asked workers to perform pairwise annotation between extracted KPs and the instance's respective user reviews. While *Precision* calculates the fraction of KPs matched to at least one gold user review, i.e., out of all extracted KPs how many are correctly mapped, *Recall* shows the fractions of gold user reviews matched to at least one KP, i.e., out of all user reviews on the product how many are covered by KPs. We then macro-averaged Precision/Recall computed for every question to obtain the final values.

For human annotation, we employed 3 MTurk crowd workers on every user review-KP pair, selecting only those with an 80% or higher approval rate and at least 10 approved tasks. Following Bar-Haim et al. (2021), we exclude annotators with Annotator-$\kappa < 0$ for quality control. This score averages all pairwise Cohen's Kappa (Landis and Koch, 1977) for a given annotator, for any annotator sharing at least 50 judgments with at least 2 other annotators. For labelling correct matches, we applied a strict threshold, in which 100% votes (3 out of 3) of the annotators had to agree that the match was correct. Otherwise, it is incorrect.

Precision	93.38%
Recall	95.0%
# Matched Reviews Per KP	3.14
# Matched KPs Per Review	4.30

Table 5: Validation of `GPT-4.1`'s performance in extracting KPs from user reviews. While precision calculates the fraction of KPs matched to at least one gold user review on the product, recall shows the fractions of gold user review matched to at least one KP.

Table 5 presents the fraction of extracted KPs matched to at least one gold answer (Precision) and vice versa (Recall). Overall, the experiment confirms that the extracted KPs are of high quality, with 95.0% of user reviews were represented by KPs (recall), while 93.38% of the extracted KPs are accurate to capture opinions within user reviews (precision).

Below are the match annotation guidelines for (extracted KP, user review) pairs:

In this task you are presented with an e-commerce product, a review written or voted by a user about the product and a key point.

You will be asked to answer the following question: Does the key point match, i.e, represent an opinion in the review?

[8] namely *DVDs*, *Food & Drink*, *Shopping*, *Beauty*, and *Entertainment*

Listing 1: Zero-shot prompt for prompting `GPT-4.1` on KP Extraction from gold user-voted product reviews.

```
You will be provided with reviews of a product, and these reviews are either voted helpful or unhelpful by a user.

You were tasked to extract a list of unique and concise key points from the list of reviews on the product.
Key points are short and high quality sentences that expresses the main claims/viewpoints of reviewers on the product.
A key point must express a specific positive or negative sentiment, and cannot have mixed opinions, and no 'but' in the key point.
Note that the final extracted list of key points must capture full details from the reviews on which the user voted helpful or non-helpful.

Provide the final list of key points as a JSON list.

Product Category: {category}
Product Name: {name}
Voted Reviews by A User: {reviews}
```

Listing 2: Zero-shot prompt for prompting `GPT-4.1` on matching personalized KP of a user towards their voted product reviews.

```
You will be provided with a single review of a product, and a list of key points taken from the summary of multiple reviews of that product.
From the list of key points, you are tasked to extract all relevant key points that matches, i.e, represents an opinion from the input review.

Key points are short and high quality sentences that expresses the main claims/viewpoints of reviewers on the product, with a particulr positive or negative attitude/sentiment.
A review might express opinions on multiple aspects. A key point matches a review if it captures the gist of an opinion from the review, or is directly supported by a point made in the review.

Output the index of those all relevant key points that matches as a JSON list.

Now perform the task on the following input:
Product Category: {category}
Product Name: {name}
A Single Review: {single_review}
List of Key Points: {key_points}
```

Listing 3: Zero-shot prompt for prompting `GPT-4.1` on generating user profile from gold voted/written review history.

```
Your objective is to create user profile using their authored and voted review history.
The profile should be general, without any personal details, but with enough details to allow personalized summarization of opinions.
You are required to analyze the user's personality traits, core interests, expression habits/style, evaluation priorities, shopping behavior, and aspects of products they are interested in, and specify them in the user profile.

The user previously signaled the following reviews:
Product Category: {category}
Written Reviews: {written_history}
Voted Reviews: {voted_history}

His profile:
```

Listing 4: Zero-shot prompt for prompting `GPT-4.1` on matching helpful KP (voted by the user) with the user profile.

```
You will be provided with a list of key points extracted from the review summary of a product, the profile of an online user.
A key point is short and high quality text that captures a particular claim/viewpoint/opinion of reviewers on the product, and
    can either express positive or negative sentiment.
A user profile covers the user's personality traits, core interests, expression habits/style, evaluation priorities, shopping
    behavior, and aspects of products they are interested in from their historical voted and written reviews.

From the list of key points, you are tasked to extract all key points that are helpful to the user
Please perform the following steps:
1. Role-playing the user based on his/her profile learnt from the history
2. Strictly reflect and output those key points that are found helpful to the user based on details learnt from history in the
    profile

Output the index of those all key points helpful to the user as a JSON list.

Product Category: {category}
User Profile:
{user_profile}
Key Points: {key_point}
```

Listing 5: Zero-shot prompt for prompting `GPT-4.1` on generating silver personalized summary based on list of helpful KPs and user profile.

```
You will be provided with the profile of an online user, a list of key points found helpful by the user, and original reviews of
    an online product
A key point is short and high quality text that captures a particular claim/viewpoint/opinion of reviewers on the product, and
    can either express positive or negative sentiment.
A user profile covers the user's personality traits, core interests, expression habits/style, evaluation priorities, shopping
    behavior, and aspects of products they are interested in from their historical voted and written reviews.

You were tasked to generate a personalized summary. For a summary to be personalized to a user, it must be ensured to fulfil
    two criteria:
+ Knowledge consistency: The personalized summary must deliver helpful knowledge/information according to the user
    preference, which is already provided as the input list of helpful key points. The summary must be grounded on the
    input helpful key points.
+ Persona consistency: Not only information/knowledge, the personalized summary must also be written to tailor the
    personality traits, expression habit (e.g., preferring practical examples or comparisons with other products), style,
    emotional tone described in the user profile.

The summary must be no longer than 300 words. Ensure that the summary covers both positive and negative opinions aligning
    with the user profile.

Product Category: {category}
Product Name: {product_name}
Helpful Key Points: {key_points}
User Profile:
{user_profile}
```

Listing 6: Zero-shot prompt for prompting `GPT-4.1` to refine silver personalized summary based on human annotator comments.

```
You will be provided with the profile of an online user, a list of key points found helpful by the user, original reviews of an
    online product, a silver-annotated personalized summary generated by LLM, and comments from a human annotator on
    aspects to improve for the silver-annotated personalized summary.

A key point is short and high quality text that captures a particular claim/viewpoint/opinion of reviewers on the product, and
    can either express positive or negative sentiment.
A user profile covers the user's personality traits, core interests, expression habits/style, evaluation priorities, shopping
    behavior, and aspects of products they are interested in from their historical voted and written reviews.

You were tasked to refine the silver-annotated personalized summary based on the human annotator comments. For a
    summary to be personalized to a user, it must be ensured to fulfil two criteria:
+ Knowledge consistency: The personalized summary must deliver helpful knowledge/information according to the user
    preference, which is already provided as the input list of helpful key points. The summary must be grounded on the
    input helpful key points.
+ Persona consistency: Not only information/knowledge, the personalized summary must also be written to tailor the
    personality traits, expression habit (e.g., preferring practical examples or comparisons with other products), style,
    emotional tone described in the user profile.

The summary must be no longer than 300 words. Ensure that the summary covers both positive and negative opinions aligning
    with the user profile.

Product Category: {category}
Product Name: {product_name}
Helpful Key Points: {key_points}
User Profile:
{user_profile}
Silver-annotated Personalized Summary: {silver_pers_summ}
Human Annotator Comments for Refinement: {human_comments}
```

A review might express opinions on multiple aspects. A key point matches a review if it captures the gist of the review, or is directly supported by a point made in the review.

The options are:

- Not At All
- Somewhat Not Well
- Somewhat Well
- Very Well

C Annotation Details of Review-KP Matching for CIAOHELPFUL Dataset (Stage 2)

We offer `GPT-4.1` with 4 options for labelling the matching status of given review-KP pairs. Pairs annotated as *Very Well* or *Somewhat Well* by LLM then becomes *candidate matching pairs*, which will be further validated by human annotation for their correctness. For human annotation, we employed 3 MTurk crowd workers per comment-KP pair, selecting only those with an 80% or higher approval rate and at least 10 approved tasks. Following Bar-Haim et al. (2021), we exclude annotators with Annotator-$\kappa < 0$ for quality control. This score averages all pairwise Cohen's Kappa (Landis and Koch, 1977) for a given annotator, for any annotator sharing at least 50 judgments with at least 2 other annotators. For labelling correct matches, at least 60% of the annotators had to agree that the match is correct, otherwise, it is incorrect. Reviews from final matching pairs, after confirmed by human, will then be grouped by similar KPs.

Below are the matching prompt for LLM and the annotation guidelines for workers validating (review, KP) pairs:

In this task you are presented with an e-commerce product, a review written or voted by a user about the product and a key point.

You will be asked to answer the following question: Does the key point match, i.e, represent an opinion in the review?

A review might express opinions on multiple aspects. A key point matches a review if it captures the gist of the review, or is directly supported by a point made in the review.

The options are:

- Not At All

- Somewhat Not Well
- Somewhat Well
- Very Well

D Training Details of the Base Summarizer (HelpfulSumm-FT)

We perform instruction fine-tuning on an LLM, using a carefully designed prompt that elicit the LLM generation of personalized summary based on both the user profile and helpful KPs based on the user profile. Formally, the generation loss for each generated summary S is computed as the negative log-likelihood (NLL) against the reference personalized summary from CIAOHELPFUL:

$$\mathcal{L}_{gen} = -\frac{1}{T}\sum_{t=1}^{T} \log P(x_t|x_{<t}) \quad (4)$$

Prompting Strategies Following OpenAI's prompt engineering guidelines[9], we structure the instruction prompt into three components (detailed in Listing 7, Appendix G): **1)** Context and input description, **2)** Task definition and output requirements, **3)** Summarization steps Importantly, our prompt adopts the Chain-of-Thoughts strategy, which guides and elicits the LLM to generate the personalized summary with four reasoning steps: **(i)** infer the user's persona (e.g., personality traits, expression style) from $H_{u,c}$ **(ii)** identify helpful key points from R_e that aligns with the inferred profile and **(iii)** generate a personalized summary paragraph that reflects both knowledge grounding and persona alignment. Example generation output is provided in Table 6 (Appendix E)

E Generation Output of HELPFULSUMM's CoT Prompt

We report a full generation output produced by HELPFULSUMM's CoT prompt (including (1) user profile and (2) helpful KPs and (3) personalized summary) in Table 6.

F Pretraining and Fine-tuning Details of BERT-based Helpful Opinion Reward Model

Prior to the fine-tuning any BERT-based encoder (e.g., `deberta-v2-xlarge`), we adapted the model to the reviews domain, by pre-training on the Yelp dataset. We performed Masked LM pertraining (Devlin et al., 2019; Liu et al., 2019) on 1.5 million sentences sampled from the train set with a length filter of 20-150 characters per sentence. Then we fine tuned the model to predict matches between KP and the user's review history pairs. We added a linear fully connected layer of size 1 followed by a sigmoid layer to the special [CLS] token in the BERT model, and trained it for three epochs with a learning rate of 2e-5 and a binary cross entropy loss.

Next, for fine-tuning these pre-trained BERT-based encoders to become Helpful Opinion Reward Model, we leverage the KPs and their helpfulness scores (calculated from Stage 2) from CIAOHELPFUL (§4) as the reference labels. To avoid imbalanced training set, we rounded the KP-level helpfulness scores to the nearest integer and randomly sample 100 KPs per integer within the $[0-5]$ ratings, resulting in 600 training samples.

Because BERT-based LM is only limited to 512 tokens, we truncate the reviews history of a user by (1) removing all reviews that have less than 10 and more than 300 words and (2) select only top 10 reviews most relevant to the key point by R1/R2/RL.

G HELPFULSUMM Prompt for End-To-End Helpful Personalized Opinion Summarization

We present the HELPFULSUMM prompt for instruction-finetuning and prompting LLM to generate personalized summary of product reviews based on user historical information in Listing 7.

H Using Generative LLMs As An Alternative for Scoring Helpful Opinion Reward

As an alternative of utilizing BERT-based transformer models (e.g., `deberta-v2-xlarge`, `roberta-large`), we propose `GPT-4.1-Scorer`, which uses generative LLMs for estimating the helpfulness score of opinions in the generated summary for the Helpful Opinion Reward Model. Inspired by `G-Eval` (Liu et al., 2023), which suggest using LLMs as reference-free metrics for NLG evaluation, we adapt this evaluator to the Persona Alignment evaluation task, by prompting `GPT-4.1` to score KP according to the user review history. Specifically, we utilize the probabilities of output

[9] https://platform.openai.com/docs/guides/prompt-engineering

Product: Gillette Mach 3 Razor (Beauty)

 # User Profile: Based on the historical voted/written reviews, the user 111 appears to be a male who is interested in shaving and grooming products. He is likely in his mid-to-late 30s, as he mentions having to shave daily and having a thick and dark beard. He is also tech-savvy, as he mentions using electric shavers and razors with advanced features. He is likely a busy professional, as he mentions having to shave quickly and efficiently. He is also a bit of a jokester, as he often uses humor in his reviews.

 # Helpful Key Points: Here are some helpful key points from the reviews of product A (Gillette Mach 3 Razor) that may be helpful to user 111:
* The razor has a close and comfortable shave, even for those with thick and dark beards.
* The three-blade system is effective in removing hair and reducing irritation.
* The lubricating strip helps to reduce friction and make shaving easier.
* The razor is easy to use and clean, with a simple design and easy-to-change blades.
* The razor is a bit expensive, but the blades last a long time and provide a good value for the price.

 # Personalized Summary:

As a busy professional who needs to shave daily, you know how important it is to have a razor that can deliver a close and comfortable shave. The Gillette Mach 3 Razor is a game-changer for anyone with thick and dark beards, as it uses a three-blade system to remove hair and reduce irritation. The lubricating strip helps to reduce friction and make shaving easier, and the razor is easy to use and clean. While the razor may be a bit expensive, the blades last a long time and provide a good value for the price.

I've tried many razors in the past, but the Mach 3 is the best I've ever used. It's like a hot towel shave at a barber's shop, but without the hassle and expense. The first blade removes the heavy hair growth, the second blade removes what the first one has missed, and the third blade administers the coup de grace, wiping out any remnants of facial hair entirely. It's a hassle-free shave that leaves my skin feeling smooth and refreshed.

I've had half a dozen shaves from the original blade now, with no signs of it dulling or losing its edge. I used to get around a week out of my old Contour, so if it's as good as that I'll be happy enough. And at only $1.65 a week, it's a small price to pay for the best and most pleasurable shaving experience of my entire life.

Table 6: Full generation output produced by **HelpfulSum-RL** (including (1) user profile and (2) helpful KPs and (3) personalized summary)

tokens from LLMs to normalize the scores and take their weighted summation as the final results. Formally, given a set of scores (like from 1 to 5) predefined in the prompt $S = \{s_1, s_2, ..., s_n\}$, the probability of each score $p(s_i)$ is calculated by the LLM, and the final helpfulness score is:

$$score = \sum_{i=1}^{n} p(s_i) \times s_i \quad (5)$$

The intuition is prompting the LLM to score the helpfulness of the given KPs against user review history multiple times and then take the average of all runs to achieve fine-grained, continuous scores that better reflect the helpfulness score of the KP against user profile. We present the prompt of `GPT-4.1-Scorer`, which instruct LLM to estimate helpfulness vote of a user for a given KP (from the personalized summary) based on their review history for Helpful Opinion Reward in Listing 8.

Reward Model	RMSE	MAE	r	ρ	Acc.
`DeBERTa-v2-xlarge`	0.519	0.399	0.282	0.242	0.715
`DeBERTa-v3-large`	0.819	0.596	0.270	0.205	0.587
`ModernBERT-large`	0.814	0.639	0.282	0.228	0.488
`RoBERTa-large`	0.984	0.771	0.267	0.217	0.416
`GPT-4.1-Scorer`	1.268	1.040	0.001	0.003	0.271

Table 7: Performance comparison of Helpful Opinion Prediction Reward Models. r denotes Pearson while ρ denotes Spearman correlation.

I Performance of Different Models for Helpful Opinion Reward

Settings In this experiment, we benchmarked the performance of different reward models for scoring the Helpful Opinion in accordance to the user review history. We specifically compared different traditional backbone BERT-based encoder (e.g., `roberta-large` (Liu et al., 2019)) fine-tuned for Helpful Opinion Prediction task over other prompted LLM-based scoring alternative. Different backbone BERT-based encoders include **`ModernBERT-large`** (Warner

Listing 7: Prompt for instruction fine-tuning and prompting LLM to generate personalized summary of product reviews based on user historical information.

```
You will be provided with reviews (written by many users) of an online product A, and a list of voted/written review history
    of the user 111 from the same product category.
You were tasked to generate a personalized summary for product A tailored to user 111 preference. For a summary to be
    personalized to a user, it must be ensured to fulfil two criteria:
+ Knowledge consistency: The personalized summary must deliver helpful knowledge/information according to the user
    preference, which is already provided as the input list of helpful key points. The summary must be grounded on the
    input helpful key points.
+ Persona consistency: Not only information/knowledge, the personalized summary must also be written to tailor the
    personality traits, expression habit (e.g., preferring practical examples or comparisons with other products), style,
    emotional tone described in the user profile.

Below are the steps:
1. Analyse and output the profile of user 111 from his/her voted/written review history.
2. Identify and output the helpful key points (short sentences illustrating opinions) from product A's reviews that may be
    helpful to user 111 based on the preference from the analysed profile
3. Generate and output a personalized summary of product A tailored to user 111, which contains helpful opinions and aligns
    with the persona (personality trait, expression habit/style, emotional tone) from user profile

The summary must be no longer than 300 words. Ensure that the summary covers both positive and negative opinions aligning
    with the user profile.
The output format should be:
# User Profile: [The user profile analyzed from historical voted/written reviews, less than 50 tokens]
# Helpful Key Points: [A Python list of helpful key points found from reviews of product A, aligned with user profile]
# Personalized Summary: [Summary paragraphs written based on the helpful key points and aligned with the user profile
    persona and preferred tone.]

Name of Product A: {product_name}
Product Reviews: {product_reviews}
User Historical Voted/Written Reviews: {hist_vote_written}
```

et al., 2025), **`roberta-large`** (Liu et al., 2019), **`deberta-v2-xlarge`** and **`deberta-v3-large`** (He et al., 2021). Next, inpsired by the *G-Eval* LLM-based evaluator (Liu et al., 2023), **GPT-4.1-Scorer** prompted `GPT-4.1` to score KP according to the user review history multiple times and then take the average of all runs to achieve fine-grained, continuous helpfulness scores of the given KP (Appendix H).

To support this evaluation, we select 1000 samples from the CIAOHELPFUL dataset as the test set, by randomly selecting 50 samples from each product category out of 20 most popular categories. The annotated helpful KPs (Stage 4 of §4) from each sample and the historical information of the respective user from CIAOHELPFUL were utilized as input to the benchmarking model, while the helpfulness score (computed from Stage 3 §4) of the respective KP from CIAOHELPFUL is treated as ground truth for evaluation. The predicted helpfulness scores were evaluated against reference helpfulness scores in terms of: RMSE, MAE, Pearson (r), Spearman (ρ), and Accuracy. Note that Accuracy is evaluated by first rounding both the predicted and reference helpfulness scores to the nearest integer (i.e., removing all decimal places) and then comparing the resulting values.

Results Table 7 presents the performance of different models for scoring opinion helpfulness. Overall, `DeBERTa`-family models demonstrate superior performance over other family (e.g., ModernBERT, RoBERTa), particularly `deberta-v2-xlarge` achieving the best performance with up to 1.93 times improvements in MAE and 44% gain in accuracy due to its exceptional model architecture and parameters. Notably, **GPT-4.1-Scorer** is the worst performer, again confirming that BERT-based encoders are better such discriminative task than generative LLM, largely because discriminative models are able to see the entire text, whereas generative LLMs (e.g., `GPT-4.1`) use a causal attention mask that is only able to look at previous tokens for a given token. More examples and analysis of scores produced by different models are in Appendix Q.

J Prompt for LLM-based Scoring for Persona Alignment Reward

We present the prompt that instruct LLM to score the alignment of generated summary with user pro-

Listing 8: Prompt for instructing `GPT-4.1` (with three-shot examples) to score the helpfulness of KPs based on user historical information.

```
base_prompt = """You will be given a key point extracted from the review summary of a product, and a list of reviews voted/
    written by an online user from history.
A key point is short and high quality text that captures a particular claim/viewpoint/opinion of reviewers on the product, and
    can either express positive or negative sentiment.
A user profile covers the user's personality traits, core interests, expression habits/style, evaluation priorities, shopping
    behavior, and aspects of products they are interested in from their historical voted and written reviews.

Your task is to rate the key point on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and
    refer to it as needed.

Evaluation Criteria:

Helpfulness (1-5) - selection of opinion and content that is helpful according to the user profile and preference inferred from
    his/her review history. The key point should include only helpful information to the user preference. Annotators were
    instructed to penalize key points which contained redundancies and unhelpful information.

Evaluation Steps:

1. Read the reviews history of the user carefully and identify the user profile and preference
2. Read the key point and compare it to the user profile and history, and how much irrelevant, redundant or unhelpful
    information it contains.
3. Assign a helpfulness score from 1 to 5.

Below are some examples:
Key Point: Flowers in the Attic is the first in a series, and while the sequels are engaging, the first book is considered the most
    dramatic and compelling.
Reviews History: \['Understanding Sun Tzu on the Art of War explores the meaning of Sun Tzu s philosophies through the
    use of high impact case examples, in strategic cases through psychological experimentations, and the world at large
    through politics and conflict bouncing on War itself. Apparently the Art of war is one of the greatest and oldest military
    strategy books.The book relates to everything as a whole, so you can find ease to relate to your own life situations as it
    seeks to establish the conflict already present within us. The main purpose of the book is to enable us to understand our
    own conflicts, and reduce it s impacts on ourselfs, and each other through 13 chapters(Strategic Assessments, Doing
    Battle, Planning a Siege, Formation, Force, Emptiness and Fullness, Armed Struggle, Adaptations, Maneuvering Armies,
    Terrain, Nine Grounds, Fire Attack and On the use of spies) which gives one strength when dealing with such forces.', 'I
    am an avid reader of books, one of, which made me laugh, cry and love. The book is called Dream A Little Dream by
    Joan Jonker.It s a story, which is set in Liverpool. Bob Dennison and Edie Brady grew up in the same street of two up,
    two down houses. Like their neighbors, they found it increasingly difficult to made ends meet. Soon he finds himself in a
    loveless marriage, his wife and eldest daughter have lost touch with their roots. Edie Changes her name to Edwina and
    insists that Bob change his name to Robert. She erases old friends and family from her past and starts a new life.At least
    his two younger children Nigel Abbie whom he takes back to the street where they were born enrich Roberts s life.
    When they are home they seek sanctuary in the kitchen where the down to earth housekeeper Agnes and the hilarious
    cleaner Kitty, provide the warmth, honestly and laughter that are missing from their lives.It is here where Robert dares
    to Dream A Little Dream that happiness is waiting just round the corner.This is a tender Liverpool saga that will move
    you to tears of sadness and joy.Enjoy.Happy ReadingZorena xx'\]
Evaluation Form (score ONLY):
- Helpfulness: 2
...
```

file characteristics for Persona Alignment Reward in Listing 9.

K Human Evaluation of KP Information Quality In The Summary

We manually evaluated the information quality of generated KPs in the summary considering 5 different criteria utilized in previous KPA studies (Kapadnis et al., 2021; Tang et al., 2024b), including COVERAGE, REDUNDANCY, SENTIMENT, INFORMATIVENESS, and SINGLEASPECT. Details of annotations and evaluation criteria are in Appendix L. Specifically, we conducted pairwise comparisons of KPs from different systems using Amazon Mechanical Turk (MTurk). Given a criteria for evaluation, each comparison involved choosing the better one from two summaries, each taken from a different system. Using the Bradley-Terry model Friedman et al. (2021), we calculated rankings from these comparisons among the models. For an example of an annotation, see Appendix M. Note that for reasonable cost, we select samples with top helpful KPs from 5 popular categories for evaluation: *DVDs*, *Food & Drink*, *Shopping*, *Beauty*, and *Entertainment*.

Results Table 8 reports Bradley-Terry scores across eleven quality criteria, which validates our earlier findings in the main experiment. In terms of KP Information Quality, HelpfulSumm-RL and its variants surpass other baselines up to 8.6 times on almost criteria except INFORMATIVENESS. In fact, we notice trade-off between INFORMATIVENESS and other criteria for the Rehearsal, which makes it achieve the best quality on INFORMATIVENESS but much lower quality on SENTIMENT, SINGLEASPECT, and REDUNDANCY. Specifically, by employing multi-agent solution to refine user suggestion for rewriting personalized summary, Rehearsal can achieve the best INFORMATIVENESS on its KPs, however, information overload among the KPs degrades their sentiment clarity and a KP may discusses multiple aspects with high redundancy.

L Annotation Details & Evaluation Dimensions of KP Information Quality (Human Evaluation)

This section describes the annotation details and evaluation dimensions of our human evaluation on KP information quality (Appendix K). Annotators were asked to perform a pairwise comparison between two sets KPs extracted from the generated summary, each taken from a different model, generated for a specific test instance (for personalized summarization of a product reviews for a specific user) The annotators must answer a comparative question with respect to the evaluating dimension. (e.g., *Which of the two list of KPs captures better ...*). For each dimension, following Friedman et al. (2021), we calculate the ranking using the Bradley-Terry model (Bradley and Terry, 1952), which predicts the probability of a given participant winning a paired comparison, based on previous paired comparison results of multiple participants, and thus allows ranking them. Specifically, we employed 3 MTurk crowd workers on every comparative annotation, selecting only those with an 80% or higher approval rate and at least 10 approved tasks. For quality control, following Bar-Haim et al. (2021), we exclude annotators with Annotator-$\kappa < 0$ for quality control. This score averages all pairwise Cohen's Kappa (Landis and Koch, 1977) for a given annotator, for any annotator sharing at least 50 judgments with at least 2 other annotators. Finally, to decide the winner from a comparison, we applied a strict threshold, in which 70% votes (2 out of 3) of the annotators had to agree that either of the summary is better.

	CV	RD	SN	IN	SA
HelpfulSumm RL	**20.38**	**26.43**	**31.51**	19.27	**31.40**
HelpfulSumm FT	15.68	20.73	24.51	12.50	28.48
HelpfulSumm ICL	12.45	20.05	13.16	8.09	12.50
Rehearsal (V + W)	14.81	15.88	11.60	**24.12**	6.50
Rehearsal (W)	11.41	9.69	7.93	22.62	6.14
PersonaSumm	10.45	3.09	4.77	5.82	6.19
GeneralSumm	14.81	4.14	6.52	7.59	8.80

Table 8: Human evaluation of summary quality by different criteria. Reported are the Bradley Terry scores of on **KP Information Quality** on 5 criteria, from left to right, including COVERAGE, REDUNDANCY, SENTIMENT, INFORMATIVENESS, and SINGLEASPECT

KP Information Quality Dimensions

- SENTIMENT: The key point in the summary should have a clear sentiment about the product being summarized (either positive or negative). This would exclude sentences like *"I came for a company event"* or sentences containing only product specification with ambiguous sentiment "*The battery can last for 3 hours.*".

- INFORMATIVENESS: The key point in the summary should discuss some aspects of the

Listing 9: Prompt for instructing LLM to score the alignment of generated summary with user profile characteristics (inferred from historical data) for Persona Alignment Reward

```
You will be provided with the profile of an online user (inferred from his/her voted/written review history) personalized
    summary catered to a user profile/preference of that user.
A personlized summary captures all opinions found helpful by the user based on his/her historical voted/written reviews,
    while is also written to be consistent with the user persona from the user profile.
A user profile covers metrics to characterized the user persona (e.g., personality traits, core interests, expression habits/style)

Your task is to rate the summary on whether it is consistent with the persona, by evaluating the alignment of the personalized
    summary on each specific metric mentioned in the user profile, namely:
+ Personality_traits (0–5)
+ Core_interests (0–5)
+ Expression_habits_style (0–5)
+ Evaluation_priorities (0–5)
+ Shopping_behavior (0–5)
+ Interested_product_aspects (0–5)

Evaluation Steps:
1. Read the personalized summary and the user profile carefully.
2. For each metric in the user profile, assess and explain (in 1–2 sentences) how well the summary align with the metric, and
    how much contradiction and irrelevant it reflect.
3. Assign a score for each metric from 0 to 5.
4. Synthesize and output the scores of all metrics as a list of JSON objects, for example: [{'Personality_traits': <0–5>}, ...]

Product Name: %s
Personalized Summary: %s
User Profile: %s
```

reviewed product and contain useful information. Any key point that is too specific or too general cannot be considered a good candidate.

- SINGLEASPECT: The key point in the summary should not discuss multiple aspects (e.g., *"Decent price, respectable portions, good flavor"*).

- REDUNDANT: Each KP should express a distinct aspect. In other words, there should be no overlap between the key points.

- COVERAGE: The key points in the summary should cover a wide diversity of opinions.

M MTurk Annotation Guideline for Pairwise Comparative KP Information Quality Evaluation

Below are the two personalized summaries of a given product in Entertainment from online reviews, generated by two different summarization frameworks towards a given user. Each summary is a list of key points (i.e., salient points) generated for summarizing the reviews opinions on different aspects personalized to the user. You are tasked to select which summary you think is better according to the below criteria. Product Name: Glamour.

Criteria: DIVERSITY. The viewpoints in the summary should cover a wide diversity of opinions relevant and representative to the topic.

Summary A: ["Glamour is a great magazine for women who want to stay informed about fashion and beauty trends without breaking the bank.", "The magazine is handbag-sized and affordable, making it a great option for busy women.", "Glamour has a good balance of fashion, beauty, and lifestyle content.", "The magazine has a strong focus on celebrity gossip and fashion trends.", "Glamour has a good balance of serious and lighthearted content.", "Glamour is a great option for women who want to stay informed and entertained without breaking the bank."]

Summary B: ["Glamour magazine helps readers stay up-to-date with the latest fashion and beauty trends.", "The magazine is not too superficial or shallow in its coverage.", "Glamour magazine is handbag-sized and easy to carry.", "The magazine costs only £1.50, offering great value for money.", "Glamour covers a wide range of topics including fashion, beauty, relationships, and lifestyle.", "The magazine has a down-to-earth and practical approach to fashion and beauty.", "Celebrity gossip and fashion trends are included in a tasteful manner.", "Glamour features real-life stories and articles that are not too serious or preachy.",

"The magazine keeps readers informed and entertained."]

The options are:

- Summary A
- Summary B

N Prompt for Generated/Reference KP Matching for Helpful Opinion Proportion (SHKP)

We present the zero-shot prompt fnnotate the matching of generated and reference KPs (in pairwise) for calculating Helpful Opinion Proportion (SHKP) in Listing 10.

O GPT4's Generated/Reference KP Matching Annotation against Human Judgement For SHKP

To validate `GPT-4.1`'s annotation performance and credibility, we conduct an experiment to measure LLM annotation judgement, as utilized for the matching between generated and reference KPs for SHKP (§7.2), in agreement with human (gold) preference. We sampled a subset of 5 instances from the test set in our main experiment and hired workers to annotate in pairwise the KPs in the generated summary and those reference KPs (Stage 4) from CIAOHELPFUL (§4). Note that these sampled pairs are part of the our main test set and have already been annotated for LLM's labels for SHKP in our main experiment. For human annotation, we employed 6 MTurk crowd workers on every comment-KP pair, selecting only those with an 80% or higher approval rate and at least 10 approved tasks. Following Bar-Haim et al. (2021), we exclude annotators with Annotator-$\kappa < 0$ for quality control. This score averages all pairwise Cohen's Kappa (Landis and Koch, 1977) for a given annotator, for any annotator sharing at least 50 judgments with at least 5 other annotators. For labelling correct matches, at least 60% of the annotators had to agree that the match is correct, otherwise, it is incorrect. In this experiment, we measured the accuracy, and conducted a Pearson correlation (r) test of `GPT-4.1`'s annotation performance against human judgement, with results reported in Table 9. For r test, we set the null hypothesis as `GPT-4.1`'s and Mturk annotated labels are independent.

From Table 9, we saw signficant small p-value, which indicates strong evidence against the null hypothesis. Importantly, we also recorded Spearman's rank correlation coefficient to be relatively closed to 1. This implies that there is a statistically significant positive correlation between `GPT-4.1` and Mturk annotated labels, which substantiates our decision of using `GPT-4.1` for matching generated KPs with reference KPs from ground truth for evaluating the proportion of helpful opinions in the summary (SHKP).

Pearson correlation (r)	0.836
p_value	7.261e-140
Accuracy	0.926

Table 9: Performance valiation of GPT4's comment-KP matching annotation against human judgement

Below are the match annotation guidelines for (KP A, KP B) pairs:

In this task you are presented with an e-commerce product, key point A generated by summarization model for the product and key point B as ground truth from the dataset.

You will be asked to answer the following question: Does the key point A match, i.e, represent an opinion in key point B?

A key point matches with the other if it captures the gist of the other, or is directly supported by a point made in the other.

The options are:

- Not At All
- Somewhat Not Well
- Somewhat Well
- Very Well

P Summary Helpfulness Score (SHS) Estimated by Adapted Third-Party Multi-Objective Reward Model

To validate the reliability of the SHS, we additionally adapted a third-party multi-objective alignment reward model (`gpt2-large-helpful-only-reward-model` [10]), which is originally developed to measure and provide for the content-quality helpfulness of LLM generated response, to measure Helpful Opinion of the generated summary across baselines. To adapt this third-party reward model to measure opinion

[10] https://huggingface.co/Ray2333/gpt2-large-helpful-reward_model

Listing 10: Zero-shot prompt for prompting `GPT-4.1` to annotate the matching of generated and reference KPs (in pairwise) for calculating Helpful Opinion Proportion (SHKP).

```
In this task you are presented with a key point (KP) extracted from the summary generated by a personalized opinion
    summarization framework, and a reference/gold key point voted helpful by the users.
Key points are short and high quality sentences that expresses the main claims/viewpoints of reviewers on the product.

You will be asked to answer the following question: "Does the generated key point match, i.e., represent an opinion in the
    reference helpful key point?". A KP matches a reference helpful KP if they implicitly discusses issues relating to each
    other, might not have to absolutely support each other.

Product Name: {product_name}
Generated Key Point: {key_point}
Reference Helpful Key Point: {key_point_given}
```

helpfulness, we adjusted this the general-purpose reward modelling prompt of this gpt-2 model to specifically our user-specific helpfulness scoring task for opinion (§ 5.1). Specifically, given a KP from the generated personalized summary and user reviews history as request provided by `Human`, we hypothetically assume that the key point is helpful to the user as the response from `Assistant` in this reward modelling prompt. Intuitively, the idea is to have the third-party reward model scored the credibility and alignment of the hypothetical "helpful" assertion from the `Assistant` with respects to the given KP from the generated personalized summary and user reviews history from `Human`. The adapted reward modelling prompt of this third-party model for scoring opinion helpfulness against user reviews history is presented in Listing 11.

Table 10 compares the Summary Helpfulness Score (**SHS**) computed by our task-specific models against the adapted third-party multi-objective reward model for helpful opinion scoring. Overall, evaluation results of SHS reported by the adapted third-party reward model and our task-specific models are highly consistent with our findings and model performance ranking from the main experiment (§ 7.2). Again, our HelpfulSumm-RL outperforms the baselines in providing the most high-quality helpful opinions in the summary, thanks to our novel CoT prompt and RL optimization that effectively capture and grounds the summary generation on helpful opinions of high interest by the user. Besides, it is important to note that our proposed task-specific models for Helpful Opinion Reward achieve high correlation (r up to 0.89, ρ up to 0.98) the third-party reward model, again reinforcing the novelty and accuracy of these models for scoring the helpfulness of opinions in the summary. However, because this third-party reward model may produce negative scores for opinion helpfulness and SHS (e.g., -0.14 for SHS on GeneralSumm's `gpt-4.1-mini`), it lacks comprehension and explainability compared to our proposed task-specific models in estimating and representing and helpfulness vote of users on opinion, given that these scores should vary within the the common rating options (0-5) for user to rate for review helpfulness on social platform (e.g., Ciao)

Q Error Analysis of Helpful Opinion Reward Employed By Different Models

In this section, we performed an error analysis of Helpful Opinion Reward scores produced by different experimented models, which aims to reinforce our findings from previous evaluation that our fine-tuned `deberta-v2-xlarge` yields the strongest performance, by analysing and comparing its output score with those predicted by the LLM-based scorer alternative, i.e., `GPT-4.1-Scorer` (Appendix H), and the third-party multi-objective reward model adapted for predicting opinion helpfulness, i.e., `gpt2-large-helpful-only-reward-model` (Appendix P)

Table 11 compares Helpful Opinion reward score computed by different models against some gold sampled KPs and their gold KP-level helpfulness score derived from CIAOHELPFUL (Stage 3 of §4). In most cases, `GPT-4.1-Scorer` underestimates the helpfulness score of the given KP against user reviews history. Meanwhile, the third-party multi-objective reward model (`gpt2-large-helpful-reward-model`) mostly produces negative scores in case the gold (actual) votes falls below neutral, again reinforcement our previous findings on the correlation of this model with our task-specific model. However, as such

Listing 11: Prompt for instructing the third-party multi-objective reward model (`gpt2-large-helpful-reward`) to score the helpfulness of a given opinion against user reviews history

```
### INPUT ###
Human: Below is a user's voted−helpful review history. Is the following key point helpful for this user? Consider relevance to
       preferences, specificity, and usefulness.
User history: {user_history}
Key point: {key_point}

### OUTPUT ###
Yes, the key point is helpful for this user.
```

	SHS (Natural Reward Score by Third-Party Model)	SHS (0-5)	
Model	`gpt2-large-helpful-only-reward-model`	`DeBERTa`	`GPT-4.1`
HelpfulSumm−RL			
+ `Llama`	**0.45**	**3.72**	**3.65**
+ `Mistral`	0.43	3.60	3.50
HelpfulSumm−FT			
+ `Llama`	0.44	3.58	3.51
+ `Mistral`	0.36	3.19	3.13
HelpfulSumm−ICL			
+ `Llama`	0.35	3.35	3.30
+ `Mistral`	0.33	3.18	3.12
Rehearsal (**W** + **V**) (Zhang et al., 2025)			
+ `gpt-4o`	0.30	3.10	3.06
+ `gpt-4.1-mini`	0.31	3.11	3.09
Rehearsal (**W**) (Zhang et al., 2025)			
+ `gpt-4o`	0.30	3.02	2.75
+ `gpt-4.1-mini`	0.31	3.08	2.84
PersonaSumm (Zhang et al., 2025)			
+ `gpt-4o`	0.22	3.01	2.68
+ `gpt-4.1-mini`	0.28	3.09	2.75
GeneralSumm (Zhang et al., 2025)			
+ `gpt-4o`	0.12	2.89	2.48
+ `gpt-4.1-mini`	-0.14	2.60	2.45
Pearson (r) / Spearman (ρ) with Natural Reward Score		0.89 / 0.96	0.84 / 0.98

Table 10: Comparison of Summary Helpfulness Score (**SHS**) (estimation of user helpfulness votes on opinions in the summary) measured by our task-specific models against the adapted third-party multi-objective reward model for helpful opinion scoring. `DeBERTa/GPT-4.1` refers to the use of task-specific models such as fine-tuned `DeBERTa-v2-xlarge` or prompted `GPT-4.1-Scorer` for estimating SHS. `gpt2-large-helpful-only-reward-model` is the third-party multi-objective reward model adapted to estimate SHS. r denotes Pearson while ρ denotes Spearman correlation.

adapted third-party reward model cannot produce rewards uniformly representing the gold user votes in the desired range (0-5), they lack explainability of their outcomes for opinion helpfulness. In contrast, our proposed `DeBERTa` model, due to the strong performance of BERT-based encoder for discriminative tasks, as well as being further fine-tuned specifcally for predicting Helpful Opinion Reward, produces much closer values to and are more representative of the gold user votes with minimal absolute errors.

R Annotation Details & Evaluation Dimensions of Summary Personalization Quality (Human Evaluation)

This section describes the annotation details and evaluation dimensions of our human evaluation on summary personalization quality (§ 7.3). Annotators were asked to perform a pairwise comparison between two summaries, each taken from a different model, generated for a specific test instance (for personalized summarization of a product reviews for a specific user) The annotators must answer a

Product Name	User ID	Key Point	Gold	deberta-v2-xl		GPT-4.1-Scorer		gpt2-large-helpful	
				Predicted	Err	Predicted	Err	Predicted	Err
ITV - Who Wants to be a Millionaire	9879	The early questions are too easy and repetitive, making the beginning of each episode less engaging.	4.00	3.64	0.36	2.80	1.20	-0.24	4.24
Ikea	20917	Ikea's flat-pack furniture is generally sturdy and well-made for the price.	3.67	3.50	0.17	2.80	0.87	-0.52	4.19
Cadbury Curly Wurly	5256313	Curly Wurly is very chewy and can stick to teeth, which may be problematic for people with sensitive teeth or fillings.	2.90	3.80	0.90	4.40	1.50	0.00	2.90

Table 11: Helpful Opinion Reward scores predicted by different experimented models on example gold KPs from CIAOHELPFUL, and are compared for absolute errors (Err) with gold KP-level helpfulness score (Stage 3 of §4).

comparative question with respect to the evaluating dimension. (e.g., *Which of the two summaries captures better* . . .). For each dimension, following Friedman et al. (2021), we calculate the ranking using the Bradley-Terry model (Bradley and Terry, 1952), which predicts the probability of a given participant winning a paired comparison, based on previous paired comparison results of multiple participants, and thus allows ranking them. Specifically, we employed 3 MTurk crowd workers on every comparative annotation, selecting only those with an 80% or higher approval rate and at least 10 approved tasks. For quality control, following Bar-Haim et al. (2021), we exclude annotators with Annotator-$\kappa < 0$ for quality control. This score averages all pairwise Cohen's Kappa (Landis and Koch, 1977) for a given annotator, for any annotator sharing at least 50 judgments with at least 2 other annotators. Finally, to decide the winner from a comparison, we applied a strict threshold, in which 70% votes (2 out of 3) of the annotators had to agree that either of the summary is better.

Summary Personalization Quality Dimensions

- PERSONALITYTRAITS: The summary should reflect personality traits of the user, such as being cautious, adventurous, humorous. For instance, a user with a humorous tone in reviews may expect the summary to contain jokes, while a highly critical user may expect a more formal tone.

- COREINTERESTS: The summary should align with the user personal hobbies and interests implied in their past reviews, for example, emphasizing the product's appearance in the favourite color of a specific user.

- EXPRESSIONHABITS: The summary should reflect the similar way of expression preferred by the users in written or voted reviews, such as preferring practical examples to emphasize product features.

- EVALUATIONPRIORITIES: The key point in the summary should emphasize what the user values most when evaluating products, such as comparing with other brands and products.

- SHOPPINGBEHAVIOUR: The summary should capture and reflect the user shopping behaviour and patterns, such as favoring bulk purchases, or avoiding luxury brands.

- INTERESTINGPRODUCTASPECTS: The summary should capture the product features that are most interesting to the user, even if they are not the most common among general users from the product reviews.

S MTurk Annotation Guideline for Pairwise Summary Personalization Quality Evaluation

Below are the two personalized summaries of a given product in Entertainment from online reviews, generated by two different summarization frameworks towards a given user. Each summary is a personalized summary summarizing opinions of the product reviews on different aspects tailored to

a specific user. You are tasked to select which summary you think is delivering better personalization to the particular user in terms of the given profile characteristic (e.g., personality trait, core interest, expression habit).

Product Name: Glamour.

User Profile Characteristic: Evaluation Priorities

Profile Characteristic Details: Prioritizes content quality, especially the balance between advertising and editorial material in magazines and shows; values pricing and overall value for money, particularly regarding subscriptions and purchases; appreciates relevance and practicality, critiquing content that feels superficial or overly commercialized.

Personalized Summary A: Glamour is a great magazine for women who want to stay informed about fashion and beauty trends without breaking the bank. With its handbag-sized size and affordable price, it's a great option for busy women who want to stay up-to-date on the latest fashion and beauty trends. The magazine has a good balance of fashion, beauty, and lifestyle content, making it a great option for women who are interested in a wide range of topics. The magazine also has a strong focus on celebrity gossip and fashion trends, which will appeal to women who are interested in celebrity culture. Additionally, the magazine has a good balance of serious and lighthearted content, making it a great option for women who want to stay informed and entertained. Overall, Glamour is a great option for women who want to stay informed and entertained without breaking the bank.

Personalized Summary B: Glamour magazine is a great choice for anyone who wants to stay up-to-date with the latest fashion and beauty trends, but in a way that is not too superficial or shallow. The magazine is handbag-sized and costs only £1.50, making it a great value for money. The magazine covers a wide range of topics, including fashion, beauty, relationships, and lifestyle, and has a down-to-earth and practical approach to fashion and beauty. The magazine includes celebrity gossip and fashion trends, but in a way that is not too superficial or shallow. The magazine also includes real-life stories and articles that are not too serious or preachy, making it a great choice for anyone who wants to stay informed and entertained. Overall, Glamour magazine is a great choice for anyone who wants to stay up-to-date with the latest fashion and beauty trends, but in a way that is not too superficial or shallow.

The options are:

- Summary A
- Summary B

T Ablation Studies

We conducted an ablation study to evaluate the impact of each helpful reward models of HELPFULSUMM, with textual quality, KP helpfulness and the personalization quality of summary reported in Table 12, 13, and 14 To this end, we configure **HelpfulSumm-RL (w/o Persona Alignment Reward)** and **HelpfulSumm-RL (w/o Helpful Opinion Reward)** as variants of HelpfulSumm-RL that respectively excludes either the reward signal from Persona Alignment and Helpful Opinion during training.

Overall, having both reward signals are crucial, but *Helpful Opinion Reward* is more important, in which **HelpfulSumm-RL (w/o Persona Alignment Reward)** achieves lexical/semantic similarity from Table 12, higher SHKP and SHS from 13 than **HelpfulSumm-RL (w/o Helpful Opinion Reward)**. For KP Information Quality, Helpful Opinion Reward enhances COVERAGE and REDUNDANCY, but struggles with SENTIMENT and INFORMATIVENESS compared to Persona Alignment, likely due to weak alignment with user persona (e.g., personality traits). Nevertheless, Helpful Opinion Reward also indirectly improves Personalization Quality, by catering to what users find helpful, particularly in COREINTERESTS and EVALUATIONPRIORITY.

U Case Studies Comparing HELPFULSUMM and Baselines

We performed case study comparing the personalized summary (for a user) generated by HELPFULSUMM and summary generated by other baselines. The user profile (inferred by LLMs from CIAOHELPFUL based on his/her reviews history) is presented in Table 15, and summary outputs of different models are in Table 16. Overall, HelpfulSumm-RL stands out for generating personalized summaries with more helpful opinions and alignment with user persona. While Rehearsal and GeneralSumm mention core product features (e.g., "*triple-blade design*", "*lubricating strip*"), they are overly generic with mixed sentiments and lacks personalization. Notably, GeneralSumm, without being

	ROUGE						BERTScore				BARTScore				BLEURT			
	KP			Summ			KP			Summ	KP			Summ	KP			Summ
	R-1	R-2	R-L	R-1	R-2	R-L	sP	sR	sF1		sP	sR	sF1		sP	sR	sF1	
HelpfulSumm−RL																		
+ Llama	**0.152**	**0.028**	**0.139**	**0.400**	**0.087**	**0.176**	**0.49**	**0.45**	**0.47**	**0.25**	**0.72**	**0.78**	**0.75**	**0.58**	**0.37**	**0.37**	**0.37**	**0.48**
HelpfulSumm−RL (w/o Persona Alignment Reward)																		
+ Llama	0.152	0.027	0.138	0.400	0.085	0.174	0.48	0.44	0.46	0.25	0.72	0.77	0.74	0.57	0.36	0.37	0.36	0.47
HelpfulSumm−RL (w/o Helpful Opinion Reward)																		
+ Llama	0.151	0.026	0.135	0.384	0.086	0.175	0.47	0.43	0.45	0.24	0.71	0.76	0.73	0.57	0.36	0.36	0.36	0.47
HelpfulSumm−FT																		
+ Llama	0.147	0.023	0.133	0.375	0.085	0.176	0.43	0.42	0.43	0.22	0.70	0.76	0.73	0.56	0.36	0.36	0.36	0.47

Table 12: KP summary textual quality. sP, sR and sF1 refer to Soft-Precision, Soft-Recall, and Soft-F1 respectively based on set-level evaluation method against reference KPs in gold answer.

	SHKP			SHS (0-5)	
Model	**Precision**	**Recall**	**F1**	DeBERTa	GPT-4.1
HelpfulSumm−RL					
+ Llama	**0.839**	**0.831**	**0.835**	**3.72**	**3.65**
HelpfulSumm−RL (w/o Persona ALignment Reward)					
+ Llama	0.829	0.819	0.824	3.71	3.64
HelpfulSumm−RL (w/o Helpful Opinion Reward)					
+ Llama	0.733	0.813	0.771	3.59	3.52
HelpfulSumm−FT					
+ Llama	0.735	0.796	0.764	3.58	3.51

Table 13: Summary Helpful KP Proportion (**SHKP**) and Summary Helpfulness Score (**SHS**) (0-5), .i.e., averaged helpfulness score of KPs in the summary. DeBERTa/GPT-4.1 refers to the use of the fine-tuned DeBERTa-v2-xlarge or prompted GPT-4.1-Scorer for predicting Helpful Opinion scores in the summary.

provided with user reviews history and personalized instruction, describes the product to be "also popular among women", which On HelpfulSumm model family, HelpfulSumm-FT still overly focuses on aspects irrelevant to the user interest, by depicting the product "*chrome design*" and 'handle' In contrast, HelpfulSumm-RL focuses more on "price", and "blades", and presents opinions in a highly descriptive narrative and highly analytical tone with empirical illustration (e.g., "*thick and dark beards*", "*$1.65 a week*", "*half a dozen shave ... no signs of ... losing its edge*", "*first blade removes... second blade ... third blade administers the coup de grace*"). As a result, HelpfulSumm-RL better tailors to the tailored to user characteristics and concerns, which are on the value for money and functionality of the product.

	KP Information Quality					Summary Personalization Quality					
	CV	RD	SN	IN	SA	PT	CI	EH	EP	SB	IA
HelpfulSumm RL	**40.23**	**40.22**	**40.54**	**54.08**	**48.07**	**46.19**	**45.93**	**45.46**	**50.46**	**48.81**	**49.5**
HelpfulSumm-RL (w/o Persona Alignment Reward)	36.88	32.91	26.66	19.33	29.57	26.07	40.96	22.05	37.97	24.57	39.98
HelpfulSumm-RL (w/o Helpful Opinion Reward)	22.89	26.88	32.79	26.6	22.36	27.74	13.11	32.5	11.58	26.62	10.52

Table 14: Human evaluation of summary quality by different dimensions. Reported are the Bradley Terry scores of 11 dimensions, from left to right, on **KP Information Quality** (COVERAGE, REDUNDANCY, SENTIMENT, INFORMATIVENESS, and SINGLEASPECT) and **Summary Personalization Quality** (PERSONALITYTRAITS, COREINTEREST, EXPRESSIONHABIT, EVALUATIONPRIORITY, SHOPPINGBEHAVIOR, INTERESTINGASPECTS).

User ID	**User Profile**
5000858	Personality Traits: - Humorous and Witty: The user displays a strong sense of humor in their reviews, often using sarcasm and playful language to convey their thoughts. They enjoy making light of situations, which suggests a playful and approachable personality. - Expressive and Engaging: The user writes in a conversational style, often addressing the reader directly and sharing personal anecdotes. This indicates a desire to connect with others through their writing. - Detail-Oriented: They provide thorough descriptions of products, including packaging, scent, and performance, showing a meticulous approach to reviewing. Core Interests: - Beauty and Personal Care: The user has a keen interest in beauty products, particularly skincare, haircare, and grooming items. They frequently explore various brands and formulations, indicating a passion for self-care and aesthetics. - Humor in Everyday Life: Their reviews often reflect a lighthearted take on mundane tasks like shaving or using beauty products, suggesting they find joy in the little things. Expression Habits/Style: - Conversational Tone: The user employs a friendly and informal tone, making their reviews relatable and enjoyable to read. - Use of Anecdotes: They often share personal stories or humorous observations related to the products, enhancing the narrative quality of their reviews. - Playful Language: The use of puns, jokes, and playful comparisons (e.g., comparing shaving to a painful experience) adds a unique flair to their writing. Evaluation Priorities: - Effectiveness: The user prioritizes how well a product performs its intended function (e.g., moisturizing, cleansing, or providing a close shave). - Value for Money: They consider the cost of products in relation to their effectiveness and longevity, often mentioning whether they feel a product is worth the price. - User Experience: The overall experience of using a product, including scent, texture, and ease of use, is important to the user. Shopping Behavior: - Value-Conscious: The user appears to be mindful of prices and often looks for deals or discounts, indicating a practical approach to shopping. - Brand Exploration: They are open to trying new brands and products, often switching between them based on availability and personal experiences. - Preference for Samples: The user enjoys trying samples before committing to full-sized products, reflecting a cautious yet curious shopping style. Aspects of Products Interested In: - Skincare and Haircare: The user frequently reviews shampoos, conditioners, moisturizers, and other beauty products, indicating a strong interest in maintaining healthy skin and hair. - Scent and Texture: They pay attention to the sensory aspects of products, such as fragrance and consistency, which play a significant role in their overall satisfaction. - Functionality: The user values products that deliver on their promises, such as effective cleansing or moisturizing, and often shares their experiences with different formulations.

Table 15: The profile of the user "**5000858**" being generated for personalized summary in the example from Table 16

Product	Gillette Mach 3 Razor (Beauty)
HelpfulSumm-RL	As a busy professional who needs to shave daily, you know how important it is to have a razor that can deliver a close and comfortable shave. I've tried many razors in the past, but the Mach 3 is the best I've ever used. It's like a hot towel shave at a barber's shop, but without the hassle and expense. It's a hassle-free shave that leaves my skin feeling smooth and refreshed. It is a game-changer for anyone with thick and dark beards, as it uses a three-blade system to remove hair and reduce irritation. The first blade removes the heavy hair growth, the second blade removes what the first one has missed, and the third blade administers the coup de grace, wiping out any remnants of facial hair entirely. The lubricating strip helps to reduce friction and make shaving easier, and the razor is easy to use and clean. While the razor may be a bit expensive, the blades last a long time and provide a good value for the price. I've had half a dozen shaves from the original blade now, with no signs of it dulling or losing its edge. And at only $1.65 a week, it's a small price to pay for the best and most pleasurable shaving experience of my entire life.
HelpfulSumm-FT	The Gillette Mach 3 razor is a top-notch shaving experience that delivers a close and quick shave with no snags or catching on your stubble. The lube strip helps the razor to glide and is good for your skin, making it a great choice for those with sensitive skin. The razor's design is sleek and modern, with a sexy chrome and black metal design that gives it a good weight so you can control it very smoothly. The blades are very sharp when they first come out of the box, and tend to last me quite a while, making it a great value for money. The handle part is very good, and as much as one could expect, as the blades are as easy as pie to load and dispose of. Overall, the Mach 3 is a great razor that I highly recommend, and it's no wonder it's number one in the UK for numbers sold.
Rehearsal (Voted + Written) (Zhang et al., 2025)	The Gillette Mach 3 razor is a crowd-pleaser for a reason. It offers a close, comfortable shave with minimal irritation, making it a top choice for both men and women. While the cost of replacement blades might sting a little, most users agree that the quality and performance are worth the splurge. Plus, with a money-back guarantee, there's little risk in giving it a try. Whether you're a daily shaver or someone who only picks up a razor for special occasions, the Mach 3 is a reliable companion. And if you're still on the fence, just remember: it's not just a razor—it's a conversation starter. After all, how many razors can claim to have inspired debates about unisex shaving and the future of blade technology?
GeneralSumm (Zhang et al., 2025)	The Gillette Mach 3 razor has garnered a wide range of opinions, but it is generally regarded as one of the best razors on the market. Many users praise its innovative triple-blade design, which provides a close and smooth shave with minimal irritation. The razor's ergonomic handle, rubber grip, and pivoting head make it easy to use and effective at navigating the contours of the face. The lubricating strip, which fades to indicate when the blade needs replacing, is another feature that users find helpful. For many, the Mach 3 delivers a shaving experience that is both comfortable and efficient, leaving skin feeling soft and refreshed. However, the razor is not without its detractors. Some users have noted that the blades can be expensive, with replacement cartridges costing significantly more than those for other razors. Despite this, many users feel the longevity of the blades and the quality of the shave justify the higher price. The Mach 3 is also popular among women, who find it effective for shaving legs and other areas, often preferring it over razors marketed specifically for women.

Table 16: Example comparison of personalized summary for user "5000858" (Table 15) produced by different variants of HELPFULSUMM and previous personalized summarization and general summarization baselines, given a product and a user from CIAOHELPFUL. Similar opinions across baselines are correspondingly marked in the same color.