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 this paper, we develop **HELPFULSUMM**, a reinforcement-learning-based model that leverages user historical helpfulness votes as an alternative proxy to authored reviews to infer user-specific preference signals for POS, using two specialized reward models to ensure (i) helpful opinions from user helpfulness votes (Helpful Opinion) and (ii) persona alignment for personalized language quality (Persona Alignment). Experimental results validated that HELPFULSUMM's use of user helpfulness vote signal offers more user-specific information in addition to information diversity, with more helpful opinions in the summary and achieving higher persona alignment with user profile than existing persona-based and general opinion summarization approaches. 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 product insights 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) is ineffective to cater to diverse needs of individual users. Personalized opinion summarization (POS) aims to generate a targeted summary of product reviews with content most relevant to a user by taking into account user characteristics and interests.

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.

Helpful Personalized Opinion Summary (Ours)
The Nokia 6230i is a great choice for someone who values simplicity and ease of use. With its intuitive interface and compact, durable design, this phone is perfect for anyone who wants a hassle-free mobile experience. Its 1.3 megapixel camera allows users to capture high-quality photos on the go. Further, The MP3 player and FM radio are other great features. With its good battery life, users can stay connected throughout the day without recharging. However, the phone may have the tendency to freeze and the difficulty of writing text messages.

Persona-based Personalized Opinion Summary
The Nokia 6230i is a highly recommended phone for users seeking a reliable, feature-rich device with a strong combination of essential phone functions and multimedia features. Its smooth and light feel, combined with its durability and user-friendly design, makes it an excellent choice for users seeking a straightforward device. The inclusion of voice control further enhances its appeal. While it may not compete with modern smartphones, it remains a solid option for those who value simplicity, reliability, and affordability.

General Opinion Summary
The Nokia 6230i is a classic, well-rounded mobile phone that offers a strong combination of essential phone functions and multimedia features in a compact, durable package. It is especially recommended for users who want a straightforward, reliable device with good camera and music capabilities without the complexity or fragility of modern smartphones. While it lacks cutting-edge technology and has some minor usability quirks, it remains a solid choice for everyday communication, casual photography , and media playback.

Table 1: Comparison of HELPFULSUMM and previous summarization approaches. HELPFULSUMM provides more diverse opinions and comprehensive information.

From an information quality perspective, helpful reviews are those receiving more attention and credit from general users (Mudambi and Schuff, 2010; Tay et al., 2020), which captures more salient opinions in review summarization (Xiong, 2013). For reviews personalization, helpfulness votes of a user on reviews can offer infer preference signal (e.g., personality traits (Alshamrani, 2023)) that complement user-written reviews. (Hirsch et al., 2025). For POS, we empirically validate this assumption, in which mismatched user helpfulness votes reduce summary personalization quality (§6.2). 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 propose HELPFULSUMM, a RL-based model that leverages user's historical review helpfulness votes for end-to-end personalized

opinion summarization. HELPFULSUMM is optimized using a joint reward of *Helpful Opinion* and *Persona Alignment*. While Helpful Opinion estimates user-specific helpfulness vote on opinions captured in the generated summary, Persona Alignment scores the summary's language quality against the user persona, inferred from user's historical data. Table 1 compares summary generated by HELPFULSUMM with previous approaches. While general summary provides high-level opinions, persona-based personalized summary rewrites the general summary to focus more on "*design*", "*durability*", however, still ignores other useful aspects to the user such as *battery* . In contrast, our approach, with better evaluation of user profile based on their helpfulness votes, provides more comprehensive information, such as the elaboration of *1.3 megapixel camera* to illustrate *high-quality camera* or "*difficulty of writing text message*" to better illustrate "*competition with modern smartphone*".

Our contributions are:

- We propose HELPFULSUMM, the first end-to-end personalized opinion summarization model that leverage historical user review helpfulness votes as an alternative proxy to sparse user authored reviews, which can better reflect user-specific preference signal and tackle with cold-start scenarios.

- To overcome data scarcity of such a specialized task, HELPFULSUMM is trained under RL without requiring large-scale annotated dataset, using joint reward signals: Helpful Opinion and Persona Alignment. Our experiments show that user helpfulness votes leads up to 4.6% more helpful opinions and 18.72% gain in the summary's personalization quality over authored review signal.

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, 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 salient summary, 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 LLM 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.

4 RL with Review Helpfulness Votes

We observe that some e-commerce platforms allow users to vote on the helpfulness of other users' reviews (Kim et al., 2006), which can be proxy signal for identifying user-specific helpful and personalized opinions. Therefore, we propose HELPFULSUMM, a RL-based model optimized for helpful POS by leveraging users' historical review helpfulness votes. Following the standard RLHF training paradigm, HELPFULSUMM optimizes a base LLM by aligning its generated summary $\hat{y}$ with two objectives representing above two personalization metrics (Tu et al., 2024) (Figure 1): (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:

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

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**.

Prompting Strategies Following OpenAI's prompt engineering guidelines[1], we structure the instruction prompt into three components (detailed in Listing 3, Appendix E): **1)** Context and input description, **2)** Task definition and output requirements, **3)** Summarization steps Importantly, our prompt multi-step process: 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 8 (Appendix B)

4.1 From Review Helpfulness to Helpful Opinion Reward

The purpose of Helpful Opinion Reward is to maximize the helpfulness of opinions (i.e., KPs) 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 (Gamzu et al., 2021; 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[2]. The helpfulness of each KP is then predicted based on two input: (1) the KP and (2) the user reviews history. 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, Amazon), however, the values are continuous to accurately represent helpfulness votes at the opinion level[3]. For such a discriminative task, we opt to employ BERT-based encoders (e.g., DeBERTa, RoBERTa) as our empirical validation shows that fine-tuning BERT-based

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

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

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

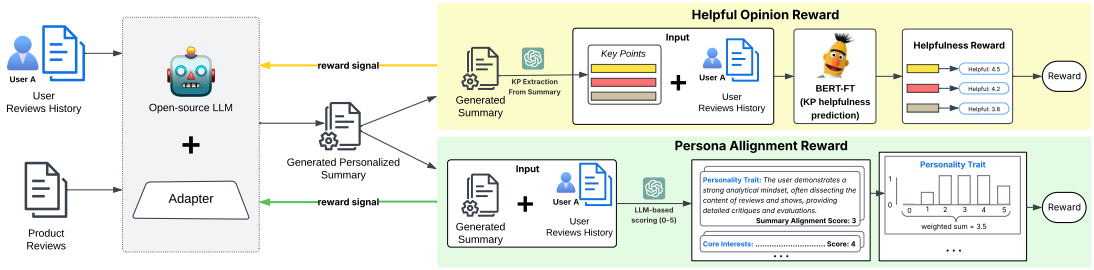

Figure 1: Illustration of personalized summarization with reinforcement learning of HELPFULSUMM.

encoder, specifically `deberta-v2-xlarge`, outperforms LLM-based scorers (Appendix G and H).

Specifically, to obtain training input and label, we leveraged user-specific helpfulness votes on reviews, offered by some e-commerce platforms (e.g., Ciao) allowing users to numerically votes (within 0-5) the helpfulness of other reviews, and transform these review-level votes to opinion-level, i.e., KP-level, helpfulness score[4] for ground truth. Given reviews voted by a user on a product ($R_{u,e}$) as ideal source of gold helpful knowledge, we first prompt `GPT-4.1` to extract concise and unique KPs from $R_{u,e}$ (*training input*). Then, 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). Prompt details are in Appendix D. The helpfulness score of each KP is then computed as the average vote scores of its matched reviews (*training label*). Human validation confirms that 95% of user reviews were represented by extracted KPs, while 93.38% of the KPs are verified as valid (Appendix K). More training details are in Appendix C.

4.2 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). These characteristics are justified in Appendix I. 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:

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

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 Experiment Setup

5.1 Datasets

In this experiment, we perform well-rounded evaluation of HELPFULSUMM using two popular e-commerce datasets on product reviews.

Ciao An European e-commerce corpus. Unlike Amazon (Gupta et al., 2019) or Yelp[6], which lacks or only report total helpful vote counts, Ciao provides helpfulness votes of each user on reviews with fine-grained rating (0-5). The dataset is utilized to explore the effectiveness of incorporating

[4] As reviews may express mixed opinions and user votes may vary across reviews while not all opinions in a high-rated reviews are truly helpful

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

[6] `https://www.yelp.com/dataset`

Statistic	**Ciao**	**Amazon**
# Product Categories	20	4
# Instance Per Category	10	50
# Total Instances	200	200
# Products / # User	197 / 193	151 / 149
# Reviews Per Product	30.42	70.42
# User Voted Reviews Per Instance	05.19	N/A
Avg Helpfulness Vote Rating Per Instance	03.59	N/A
# User Written Reviews Per Instance	02.55	2.05
# User Voted Reviews History	123.56	N/A
# User Written Reviews History	22.77	10.13

Table 2: Core statistics of the Ciao and Amazon test set. An instance/sample considers the personalized summarization of a product towards a user.

user helpfulness votes into reviews history for personalized summarization. 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 $H_{u,c}$. To ensure runtime and cost feasibility, we sample 200 instances, by selecting top 10 samples (by product reviews count) from each of the 20 product categories (e.g., *DVDs*, *Computers*).

Amazon We further develop a test set from the Amazon Reviews Dataset (Hou et al., 2024) to benchmark the model performance As the dataset does not offer user-level helpfulness votes, we utilize user-written reviews as histories for the personalization input. Following Zhang et al. (2025), we focus on four categories: *Clothing*, *Shoes and Jewelry*, *Electronics*, *Health and Personal Care*, and *Home and Kitchen*, To ensure the quality of the data, we include only products with 50-100 reviews, and sampling top 50 instances (by product reviews count) per category, totaling 200 instances.

Training and evaluation were performed on the respective dataset. For RL training of HELPFULSUMM, we randomly sample 1,000 instances (outside of test sets) from Ciao and Amazon. Note that following Zhang et al. (2025), we leave out reviews written and/or voted by the user on the current product from the user reviews history input to avoid data leakage, while utilize real user-written reviews on the current product as ground truth for evaluation for fair comparison across datasets[7] The core statistics of test sets are in Table 2.

5.2 Implementation Details

HELPFULSUMM was experimented with two backbone: `Llama-3.1-8B-Instruct`[8] (`Llama`) and `Mistral-7B-Instruct-v0.3`[9] (`Mistral`). For RL training, we implemented TRL (von Werra et al., 2020) and LoRA (Hu et al., 2021) for parameter-efficient training. We employed 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.

5.3 Baselines

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

HelpfulSumm-ICL Base LLMs zero-shot ICL-prompted for helpful personalized opinion summarization using HELPFULSUMM's CoT prompt instruction (§4)

Rehearsal (Zhang et al., 2025) A multi-agent framework that first generates a general summaries from product review, before rewriting the summary by having a user agent and a role-playing supervisor responsible for modeling user information and generating and evaluating personalized opinions.

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.

Note that we implement these baselines in different input configurations, denoted as the postfixes to the baselines: (**V**): *voted-only*, (**W**) *written-only*, and (**V + W**) *voted and written* user reviews history. For fair comparison, we limits generated summaries to a maximum length of 300 words.

6 Results

6.1 How is the quality of the personalized summaries?

Information Quality Following Zhang et al. (2025), we treat real user-written reviews on the cur-

[7] Ciao offers user-specific review helpfulness votes while Amazon does not

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

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

	ROUGE						BERTScore				BARTScore			
	KP			Summ			KP			Summ	KP			Summ
	R-1	R-2	R-L	R-1	R-2	R-L	sP	sR	sF1		sP	sR	sF1	
Ciao														
HelpfulSumm−RL (W + V) (Llama)	**0.538**	**0.324**	**0.505**	**0.347**	0.071	0.161	**0.49**	**0.45**	**0.47**	**0.25**	**0.72**	**0.78**	**0.75**	**0.51**
HelpfulSumm−RL (V) (Llama)	0.533	0.316	0.494	0.339	0.075	0.158	0.48	0.44	0.46	0.24	0.72	0.77	0.74	0.50
HelpfulSumm−RL (W) (Llama)	0.518	0.305	0.483	0.337	0.068	0.150	0.47	0.44	0.45	0.24	0.71	0.76	0.73	0.48
HelpfulSumm−ICL (W + V) (Llama)	0.456	0.238	0.417	0.326	0.057	0.149	0.42	0.40	0.41	0.23	0.67	0.74	0.71	0.43
HelpfulSumm−ICL (V) (Llama)	0.448	0.235	0.410	0.319	0.057	0.145	0.40	0.39	0.40	0.23	0.67	0.73	0.70	0.42
HelpfulSumm−ICL (W) (Llama)	0.435	0.230	0.392	0.311	0.052	0.140	0.38	0.39	0.39	0.22	0.66	0.73	0.69	0.41
Rehearsal (W + V) (gpt-4o)	0.401	0.192	0.359	0.286	0.049	0.139	0.38	0.43	0.40	0.23	0.69	0.75	0.72	0.43
Rehearsal (V) (gpt-4o)	0.385	0.183	0.347	0.282	0.047	0.136	0.37	0.44	0.40	0.23	0.68	0.76	0.72	0.42
Rehearsal (W) (gpt-4o)	0.366	0.158	0.323	0.272	0.043	0.128	0.36	0.44	0.39	0.22	0.68	0.70	0.69	0.41
PersonaSumm (W) (gpt-4o)	0.349	0.145	0.316	0.204	0.035	0.105	0.35	0.43	0.38	0.20	0.69	0.62	0.65	0.39
GeneralSumm (gpt-4o)	0.306	0.113	0.288	0.111	0.019	0.065	0.34	0.41	0.37	0.19	0.63	0.54	0.58	0.35
Amazon														
HelpfulSumm−RL (W) (Llama)	**0.536**	**0.338**	**0.512**	**0.195**	0.041	**0.111**	**0.38**	**0.50**	**0.43**	**0.24**	**0.62**	**0.78**	**0.68**	**0.48**
HelpfulSumm−ICL (W) (Llama)	0.489	0.287	0.462	0.163	0.027	0.102	0.35	0.49	0.40	0.22	0.59	0.74	0.65	0.44
Rehearsal (W) (gpt-4o)	0.491	0.286	0.464	0.167	0.021	0.105	0.30	0.47	0.36	0.20	0.58	0.74	0.65	0.44
PersonaSum (W) (gpt-4o)	0.473	0.283	0.450	0.109	0.015	0.080	0.24	0.42	0.31	0.17	0.53	0.60	0.57	0.37
GeneralSumm (gpt-4o)	0.418	0.207	0.393	0.092	0.010	0.063	0.20	0.38	0.26	0.15	0.49	0.49	0.49	0.30

Table 3: 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 from real (gold) user-written reviews on the product. Full results of all backbones are in Table 16 (Appendix T)

rent product as reference summaries for summary-level evaluation, while KPs extracted from these reviews as reference KPs for KP-level evaluation. We use the state-of-the-art metrics ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and BARTScore (Yuan et al., 2021) for measuring lexical and semantic similarity. Summary-level evaluation is measured by computing the highest score between the generated summary and gold user written reviews on the product. For KP-level evaluation, we calculate soft-Precision/Recall/F1 (sP, sR and sF1) 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.

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

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

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. Separately, for KP-level ROUGE, we compute the highest score between generated and reference KPs.

Results Table 3 reports the summary and KP-level textual quality across baselines, with KP-level performance exhibiting clearer distinction. Overall, HELPFULSUMM, in both RL and ICL settings, outperforms other baselines consistently across different datasets (paired t-test, $p < 0.05$), largely due to our CoT prompt that guides LLMs to analyze user profiles and infer helpful opinions from their reviews 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 state-of-the-art personalized summarization baseline Rehearsal, with up to 2.05 times ROUGE-2 gain (0.324 vs. 0.143) and 17.5% BERTScore sF1 increase. Rehearsal is inferior to HELPFULSUMM, as we empirically found that there are limited turns of interactions (avg. 1.01 turns) between role-played agents in this framework to produce fine-grained enough user suggestion for personalized summary rewriting. Lastly, GeneralSumm, without personalization input and instruction, yields the lowest textual quality due to the redundancy of unhelpful opinions. Our manual evaluation validates above findings, and also using helpfulness vote signal notably boosts suummary's opinion diversity (COVERAGE) over authored review signal (Appendix M).

Personalization Quality Beside *information quality*, i.e., Knowledge Consistency, we further

	PT	CI	EH	EP	SB	IA
Ciao						
HelpfulSumm RL (V)	**18.58**	**15.54**	**17.05**	**19.64**	**16.21**	**16.26**
HelpfulSumm RL (W)	17.24	13.09	15.48	16.72	15.23	15.52
HelpfulSumm ICL (V)	11.95	12.70	12.34	10.55	12.78	10.99
HelpfulSumm ICL (W)	10.50	11.50	11.51	9.50	11.50	10.11
Rehearsal (V)	14.48	14.26	14.60	15.38	14.29	15.79
Rehearsal (W)	11.01	12.36	13.48	11.00	13.02	13.66
PersonaSumm (W)	8.64	11.01	9.20	9.64	9.89	9.53
GeneralSumm	7.61	9.53	6.35	7.58	7.10	8.15
Amazon						
HelpfulSumm RL (W)	**27.63**	**25.66**	**27.31**	**31.38**	**26.46**	**26.02**
HelpfulSumm ICL (W)	17.77	20.08	19.50	18.63	20.01	18.54
Rehearsal (W)	22.63	22.18	23.78	21.25	23.23	24.48
PersonaSumm (W)	17.37	17.30	18.58	16.48	18.98	16.93
GeneralSumm	14.60	14.78	10.83	12.26	11.32	14.04

Table 4: Human evaluation of summary quality by different criteria. Reported are the Bradley Terry scores of 6 criteria on **Summary Personalization Quality**.

perform manual evaluation of summary's personalization against user profile on 6 criteria: PERSONALITYTRAITS (PT), COREINTEREST (CI), EXPRESSIONHABIT (EH), EVALUATIONPRIORITY (EP), SHOPPINGBEHAVIOR (SB), INTERESTINGASPECTS (IA). Specifically, we hire workers from Amazon Mechanical Turk (MTurk) to compare summaries of different systems in pairwise by different criteria, on how well it aligns with each criterion according to the user profile[10], beforing ranking them using Bradley-Terry model (Friedman et al., 2021). Annotation details are in Appendix R. Annotation examples are in Appendix S.

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 1.68 times improvement, notably on PERSONALITYTRAIT, EVALUATIONPRIORITIES, and EXPRESSIONHABITS. Notably, the use of user helpfulness vote over review authored signal is effective. For HELPFULSUMM, it leads to up 18.72% improvement. For Rehearsal, it enhances alignment across all characteristics, even outperforming HelpfulSumm prior to RL-based training.

6.2 Does user helpfulness vote signal truly offer personalization information?

We further validate that user helpfulness vote signals truly offers HELPFULSUMM with more user-specific information for personalization, rather only information diversity via generic helpfulness, using

[10] inferred by GPT-4.1 on user voted/written reviews history

	PT	CI	EH	EP	SB	IA
HelpfulSumm−RL (V) (Llama)						
Matched Vote History	**50.82**	**40.22**	**53.79**	**45.42**	**42.79**	**40.31**
Mis-matched Vote History	27.16	32.91	24.74	28.23	30.54	30.85
GeneralSumm (gpt-4o)						
No Vote History	22.02	26.88	21.48	26.36	26.67	28.84

Table 5: Summary personalization quality given matched and mis-matched user voted reviews history. Reported are the Bradley Terry scores of 6 criteria on **Summary Personalization Quality**.

a matched vs. mismatched user experiment. Specifically, for each test sample, we keep the product review set fixed while replace the voted history of the target user by another from the same product category. Results of personalization quality, experimented on HelpfulSumm-RL (V) and compares with GeneralSumm, are presented in Table 5. Overall, performance under mismatched vote histories is comparable to GeneralSumm, indicating that mismatched votes contribute limited personalization benefit beyond generic opinion diversity.

6.3 Do summaries contain helpful opinions?

We evaluate the helpfulness of the personalized summary, by measuring the proportion of KPs in the generated summary helpful to the user, presented in Table 6. While acquiring feedback from personalized user is infeasible due to privacy constraint, users who wrote reviews about a product are likely buyers, and thus KPs in their gold written product reviews can be reference KP for evaluating how opinions in a personalized summary can affect user purchase decision. Specifically, for each instance, we prompt GPT-4.1 to match KPs in the generated summary with reference KPs in pairwise (Appendix P). A KP is considered to be helpful only if it matches with at least one reference KP (Precision), and vice versa (Recall). Empirically, GPT-4.1 annotations show high correlation with human judgement (r = 0.823) (Appendix Q).

Results From Table 6, the performance on Amazon are generally lower than on Ciao, as Amazon users generally have shorter reviews history (Table 2) than Ciao as personalization input. Further, we notice trade-off between Precision and Recall. Baselines such as PersonaSumm and GeneralSumm achieve moderate 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. In contrast,

Model	Ciao			Amazon		
	P	R	F1	P	R	F1
HelpfulSumm−RL (W + V)						
+ Llama	**0.823**	**0.800**	**0.811**		N/A	
HelpfulSumm−RL(V)						
+ Llama	0.796	0.779	0.787		N/A	
HelpfulSumm−RL (W)						
+ Llama	0.763	0.719	0.741	**0.618**	**0.838**	**0.712**
HelpfulSumm−ICL (W + V)						
+ Llama	0.679	0.616	0.646		N/A	
HelpfulSumm−ICL (V)						
+ Llama	0.671	0.600	0.634		N/A	
HelpfulSumm−ICL (W)						
+ Llama	0.670	0.570	0.616	0.555	0.703	0.620
Rehearsal (W + V) (Zhang et al., 2025)						
+ gpt-4o	0.674	0.673	0.674		N/A	
Rehearsal (V) (Zhang et al., 2025)						
+ gpt-4o	0.659	0.657	0.658		N/A	
Rehearsal (W) (Zhang et al., 2025)						
+ gpt-4o	0.624	0.623	0.624	0.319	0.764	0.450
PersonaSumm (W)						
+ gpt-4o	0.534	0.693	0.603	0.270	0.658	0.383
GeneralSumm						
+ gpt-4o	0.481	0.583	0.527	0.178	0.629	0.277

Table 6: Proportion of helpful KPs, i.e., opinions, in the personalized summary. Full results of all backbones are in Table 15 (Appendix T)

HelpfulSumm-RL substantially outperforms other baselines by producing 50% more helpful KPs in the generated summary (Precision), with generated KPs achieving a 80% success rate (Recall) in penetrating user decision-making, 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, for both HELPFULSUMM and Rehearsal, incorporating vote signal into input user review history (V + W) leads to 7% F1 improvement, while vote signal (V) provides stronger representation of user-specific preference than written signal (W), again proving its utility in cold-start scenario.

We note that personalized summaries present opinions tailored to user needs while still retaining negative (minority) opinions. Empirically, 31.3% of opinions are negative in HelpfulSumm-RL summaries (vs. 45.5% in GeneralSumm), and helpfulness votes signal preserves more negative opinions than authored-review signals (Appendix O).

6.4 Ablation Study

We performed ablation study of HelpfulSumm-RL by excluding either Helpful Opinion or Persona Alignment Rewards from RL training (Appendix U). Overall, having both reward signals are crucial, but *Helpful Opinion Reward* is more important (higher lexical/semantic similarity from Table 17, higher summary helpfulness from 18). 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.

6.5 Case Studies

We conducted case studies to evaluate the information and personalization quality of personalized summary for a "Finance" entity (Table 20 in Appendix V). Overall, HelpfulSum-RL stands out by capturing more helpful opinions in the personalized summaries with strong alignment with user profile, while also ensures diversity by covering both positive and negative opinions. Given a user with "Evaluation Priorities" focusing on *low interest rate*, *transparent fee structure*, and *customer service*, HelpfulSumm-RL provides detailed illustration for *loan calculator* and *no penalties for early repayment*, while also emphasizes the entity downside such as *the loan process can be lengthy*, and *slow customer service*. In contrast, while Rehearsal and GeneralSumm mention diverse core entity features (e.g., "*attractive rates*", "*customizable repayment options*"), they are generic and lacks practical illustration with personalized details to showcase the suitability of the features to the user.

7 Conclusion

In this paper, we leverage historical review helpfulness votes as a proxy to better model user-specific preference signal for end-to-end helpful personalized opinion summarization. We propose HELPFULSUMM, a RL-based model optimized for POS using joint 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 validated that HELPFULSUMM's use of user helpfulness vote as proxy signal offers more user-specific information in addition to information diversity, with more helpful opinions and higher information and personalization quality of the personalized summaries.

Limitations

- For this research, we only performed experiment on data constructed from the Ciao and Amazon Reviews datasets because, to our best knowledge, (1) the Ciao dataset is the only publicly available data offering information on the helpfulness votes of individual user on reviews and (2) the Amazon Reviews Dataset features user-level written reviews on products of the Amazon retail giant, in which user voted/written reviews history and product reviews are essential for our experiments.

- Due to privacy constraints and the inability to acquire feedback from real end-users, direct evaluation of downstream purchasing behavior is infeasible. Therefore, we can only assess model effectiveness objectively by measuring summary alignment with actual written reviews on the product from the personalizing user.

- Due to cost consideration, we can only perform limited crowd-sourced annotation for evaluation on the summary's personalization quality.

- As there are no ground truth user profile available, LLM-inferred user profile (based on their reviews history) is the most feasible approach given the cost, subjectivity, and difficulty of human psychological profiling. Further, inferring user profile based on their reviews history also requires consent from online users upon practical deployment to avoid potential misuse.

Ethics Statement

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

The Ciao reviews 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 Amazon Reviews dataset used in our experiments was publicly crowdsourced and released for the research publication for analyzing customer satisfaction, sentiment, and trends on the Amazon e-commerce platform (Hou et al., 2024). These datasets 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*.

Amal Mosleh Alshamrani. 2023. Effects of personality on the perceived helpfulness of online reviews.

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.

Li Chen and Pearl Pu. 2006. Evaluating critiquing-based recommender agents. In *AAAI*, volume 6, pages 157–162.

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.

Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying helpful sentences in product reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 678–691, Online. 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*.

Emrul Hasan, Mizanur Rahman, Chen Ding, Jimmy Xiangji Huang, and Shaina Raza. 2025. Review-based recommender systems: A survey of approaches, challenges and future perspectives. *ACM Comput. Surv.*, 58(1).

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

Mariya Hendriksen, Ernst Kuiper, Pim Nauts, Sebastian Schelter, and Maarten de Rijke. 2020. Analyzing and predicting purchase intent in e-commerce: anonymous vs. identified customers. *arXiv preprint arXiv:2012.08777*.

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.

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

Edward 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*.

Zhaopei Huang, Qifeng Dai, Guozheng Wu, Xiaopeng Wu, Xubin Li, Tiezheng Ge, Wenxuan Wang, and Qin Jin. 2026. Mem-pal: Towards memory-based personalized dialogue assistants for long-term user-agent interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31229–31237.

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.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430, Sydney, Australia. 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.

Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Decomposed opinion summarization with verified aspect-aware modules. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24805–24841, Vienna, Austria. 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.

Karol Przystalski, Jan K Argasiński, Iwona Grabska-Gradzińska, and Jeremi Ochab. 2025. Stylometry recognizes human and llm-generated texts in short samples. *Expert Systems with Applications*, page 129001.

Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. 2024. User modeling and user profiling: A comprehensive survey. *arXiv preprint arXiv:2402.09660*.

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

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.

Rama Syamala Sreepada, Bidyut Kr. Patra, and Antonio Hernando. 2017. Multi-criteria recommendations through preference learning. In *Proceedings of the 4th ACM IKDD Conferences on Data Sciences*, CODS '17, New York, NY, USA. Association for Computing Machinery.

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.

Marcin Szmydt. 2021. Contextual personality-aware recommender system versus big data recommender system. In *Business Information Systems*, pages 163–173.

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 Personalized Summary Examples for Different Users

We provide an example of personalized summaries, of a given product, generated for different users in Table 7.

B 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 8.

C 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 and cast user helpfulness votes on reviews to the opinion (KP) level from the Ciao dataset (Gupta et al., 2019) to obtain training input and label, using process described in §4.1. To

Product	MINI Cooper
	Generated Summary
User 1	The MINI Cooper is a car that will surely put a smile on your face. With its unique design and strong performance, it's a car that's fun to drive and will make you feel like you're behind the wheel of a sports car. The interior is comfortable and stylish, with plenty of room for passengers and cargo. The MINI Cooper is also reliable and has a good reputation for maintenance, making it a great choice for those who want a car that will last. While the MINI Cooper may be expensive, it's worth it for the performance and style you get. And with its good resale value, you can be sure that your car will hold its value over time. One thing to keep in mind is that the MINI Cooper is a small car, so it may not be the best choice for those who need a lot of space. But for those who value performance and style, the MINI Cooper is a great choice.
User 2	The Mini Cooper is a car that truly embodies the spirit of fun and style. Its retro image has been modernized and updated, making it a unique and eye-catching vehicle on the road. With its higher top speed and supercharged Cooper S model, it's a thrill to drive and a joy to own. The interior is modern and funky, with a central speedometer that adds to its charm. While it may have limited space in the back seats, it's a small price to pay for its comfort and style. And with its ease of parking and maneuverability, it's the perfect car for city driving. If you're looking for a car that's truly one-of-a-kind, the Mini Cooper is the perfect choice. Its unique design and performance make it worth every penny of its premium price.

Table 7: Personalized summary examples for different users

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

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.

D Prompts for Annotating Gold Helpful Opinions for Training Helpful Opinion Reward Model

D.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 in Listing 1.

D.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 in Listing 2.

E 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 3.

F 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 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:

$$score = \sum_{i=1}^{n} p(s_i) \times s_i \quad (4)$$

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 4.

G Performance of Different Models for Helpful Opinion Reward

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

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 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 F).

To support this evaluation, we select 1000 samples from the Ciao dataset as the test set, by randomly selecting 50 samples from each product category out of 20 most popular categories. The annotated helpful KPs from each sample and the histor-

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: 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}
```

ical information of the respective user from Ciao were utilized as input to the benchmarking model, while the helpfulness score (computed from gold user reviews helpfulness votes as described in § 4.1) of the respective KP 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 9 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 H.

H 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 F)

Additionally, in this experiment, we adapted a third-party multi-objective alignment reward model (`gpt2-large-helpful-only-reward-model` [11]), which is originally developed to measure and provide for the content-quality helpfulness of LLM

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

Listing 4: Prompt for instructing `GPT-4.1` (with three-shot examples) to score the helpfulness of KPs based on user historical information.

```
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
...
```

generated response, to measure Helpful Opinion of the generated summary across baselines. To adapt this third-party reward model to measure opinion 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 (§ 4.1). 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 5.

Table 10 compares Helpful Opinion reward score computed by different models against some gold sampled KPs and their gold KP-level helpfulness score derived from § 4.1. 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 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.

I Details Six Persona Characteristics for Persona Alignment Reward

- 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. In fact, user personality traits has become critical signal to provide contextual information for better personalization in recommender system (Alshamrani, 2023; Szmydt, 2021). It influences users preferences for diversity, novelty, and information style, making them useful signals for tailoring summaries to match individual cognitive and affective tendencies.

- 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. In fact, modeling users' long-term and dynamic interests enables summarization systems to prioritize content that aligns with what users consistently care about, therefore, improving relevance and engagement (Purificato et al., 2024).

- 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. In fact, users often exhibit stable linguistic and stylistic patterns (e.g., verbosity, tone), which can be stylometric features to generate summaries that better match their preferred communication style (Przystalski et al., 2025).

- 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. In fact, users may have different styles of evaluation focus (e.g., accuracy-focused, diversity-focused), and so modeling their priorities allows summarization systems to optimize outputs according to the style each user values most (Chen and Pu, 2006; Sreepada et al., 2017)

- SHOPPINGBEHAVIOUR: The summary should capture and reflect the user shopping behaviour and patterns, such as favoring bulk purchases, or avoiding luxury brands. More specifically, behavioral patterns such as browsing sessions and purchase intent reflect real, context-driven preferences, which can

Listing 5: 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.
```

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 10: Helpful Opinion Reward scores predicted by different experimented models on example gold KPs from Ciao (derived from user product reviews), and are compared for absolute errors (Err) with gold KP-level helpfulness score (§ 4.1).

become for selecting and organizing summary content that aligns with the particular user goals (Hendriksen et al., 2020; Huang et al., 2026).

- 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. Aspect-level preferences (e.g., price, quality, design) allow summarization systems to highlight specific product attributes that users care about, therefore, leading to more fine-grained and supportive summaries for informed user decision making (Hasan et al., 2025; Tang et al., 2024b; Li et al., 2025).

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 profile characteristics for Persona Alignment Reward in Listing 6.

K Human Validation of GPT4's KP Extraction from User Product Reviews

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 in § 4.1. Specifically, to maintain reasonable cost, we randomly sampled a (user-product) instance from 5 common product categories of Ciao [12], 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

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

Listing 6: 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
```

by KPs. We then macro-averaged Precision/Recall computed for every question to obtain the final values.

For annotation setting, 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 11: 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 11 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?

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

L Annotation Details of Review-KP Matching

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

M 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 evaluation criteria are in Appendix N.

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. 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 . . .*).

Then, following Friedman et al. (2021), we calculate the ranking from these comparisons among the models using the Bradley-Terry model (Bradley and Terry, 1952) for each criteria, 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.

Results Table 12 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, Re-

hearsal 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. Notably, user helpfulness votes signal (V) exhibits up to 64% better information quality than authored reviews signal (W), with worth-mentioning improvement in diversity of the summary KPs (COVERAGE), mostly attributed to the information diversity offerred by the abundance of user generic helpfulness votes on other products.

	CV	RD	SN	IN	SA
HelpfulSumm RL (V)	**17.83**	**23.13**	**26.57**	16.86	**27.48**
HelpfulSumm RL (W)	13.72	18.14	21.45	12.94	24.82
HelpfulSumm ICL (V)	12.51	15.49	13.21	9.79	12.59
HelpfulSumm ICL (W)	10.89	14.54	11.80	7.78	10.94
Rehearsal (V)	12.96	13.90	10.15	**21.11**	5.69
Rehearsal (W)	9.98	8.48	6.94	19.79	5.37
PersonaSumm (W)	9.14	2.70	4.17	5.09	5.42
GeneralSumm	12.96	3.62	5.71	6.64	7.70

Table 12: 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

N Evaluation Criteria of KP Information Quality (Human Evaluation)

- 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 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.

O Sentiment Polarity of Opinions in Personalized Summary

Polarity	Positive	Negative
HelpfulSumm-RL (V)	68.7%	31.3%
HelpfulSumm-RL (W)	75.1%	24.9%
GeneralSumm	54.5%	45.5%

Table 13: Proportion of positive and negative opinions in the generated personalized summary (by HelpfulSumm-RL) and general summary (by GeneralSumm).

To validate that minority opinions are still presented at acceptable margin in the personalized summary, we conduct an experiment to calculate the proportion of positive and negative opinions in the personalized summary generated by HELPFULSUMM. Specifically, the experiment aims to measure the extend to which the proportion of negative, mostly minority, opinions will decrease considering a personalized summary and a general summary. In many cases, it is assumed that personalized summaries are generated to cater to user interests, and therefore, it is more likely there would be higher chances of capturing opinions favourable to a user. In fact, favourable opinions are proven to mostly be positive opinions by majority, as users primarily write and vote for reviews with positive opinions. Here we aims to validate **(1)** whether this relation presented, by comparing the proportion of negative opinions in personalized summaries with general summaries, and **(2)** if (1) is valid, between user helpfulness votes on reviews and user authored reviews, which signal will introduce less effect on the proportion of negative opinions in the personalized summary.

Settings In this experiment, we utilize personalized summaries generated by HelpfulSumm-RL, in both of its *voted-only* (V) and *written-only* input configurations, and the general summaries generated by GeneralSumm. All generated summaries are post-processed into KPs (a KP resemble an opinion in the summary) using the same LLM-based extractor as in §4.1. For every KP extracted from a summary, we hire workers from Amazon Mechanical Turk (MTurk) to label the sentiment polarity Specifically, we employed 4 MTurk crowd workers to annotate the sentiment polarity of every KP, 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-$\kappa < 0$. 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. To decide the final polarity label of a KP, we applied a strict threshold, in which 50% votes (2 out of 4) of the annotators had to agree that either of the summary is better. Finally, we calculate the proportion of positive and negative opinions in every generated summary, before macro-averaged the final proportions across all generated summaries in the test set. Note that we ignore opinions with neutral polarity in this calculation.

Results From Table 13, we observe that personalized summaries generated by **HelpfulSumm-RL** consistently contain a lower proportion of negative opinions compared to the general summary baseline. Specifically, HelpfulSumm-RL (V) produces **31.3%** negative opinions, while GeneralSumm contains a substantially higher proportion at **45.5%**. Overall, it is inevitable that opinions in personalized summaries, which are conditioned on user-specific signals and are more aligned with user interests, are more frequently associated with positive sentiment.

Importantly, despite this reduction, negative opinions are **not eliminated** in the personalized summaries. The presence of **31.3%** negative opinions indicates that HelpfulSumm-RL maintains a balanced representation of both favorable and unfavorable aspects of the product. This suggests that the model does not simply bias toward positive opinions, but instead selectively retains negative opinions that remain relevant or informative to the user, thereby preserving decision-support quality.

Comparing different personalization signals, we further observe that **HelpfulSumm-RL (V)** retains a higher proportion of negative opinions (31.3%) than **HelpfulSumm-RL (W)** (24.9%). This indicates that user helpfulness votes introduce **less suppression of minority (negative) opinions** compared to authored review signals. To explain, user-authored reviews primarily reflect users' own experiences and subjective sentiment, which may introduce a stronger polarity bias in the resulting summaries. In contrast, helpfulness votes better capture user-specific information via users' judgments on the informational value of a broader set of reviews, including both positive and negative opinions. As a result, vote-based signals tend to preserve more diverse and informative content, allowing the model to retain a higher proportion of negative yet useful opinions.

P Prompt for Generated/Reference KP Matching for Summary Helpfulness Evaluation

We present the zero-shot prompt to annotate the matching of generated and reference KPs (in pairwise) for calculating Summary Helpfulness Evaluation in Listing 7.

Q GPT4 Performance for KP Matching in Summary Helpfulness Evaluation

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 summary helpfulness evaluation (§6.3), 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. Note that these sampled pairs are part of the our main test set and have already been annotated for LLM's labels 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 14. For r test, we set the null hypothesis as `GPT-4.1`'s and Mturk annotated labels are independent.

From Table 14, 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 gener-

Listing 7: Zero-shot prompt for prompting `GPT-4.1` to annotate the matching of generated and reference KPs (in pairwise) for Summary Helpfulness Evaluation.

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}

ated KPs with reference KPs from ground truth for evaluating the proportion of helpful opinions in the summary.

Pearson correlation (r)	0.823
p_value	5.821e-140
Accuracy	0.918

Table 14: 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

R Annotation Details of Summary Personalization Quality (Human Evaluation)

This section describes the annotation details and evaluation dimensions of our human evaluation on summary personalization quality on six user persona characteristics (§ 6.1). Details of six user persona characteristics under evaluation are in Appendix I. 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 comparative question with respect to the evaluating dimension. (e.g., *Which of the two summaries captures better* . . .). For each of the six evaluating , 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.

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: Sony MiniDisc recorder.

User Profile Characteristic: Core Interests

Profile Characteristic Details: The user has a deep appreciation for music and values high-quality sound systems and portable music devices. They are interested in electronics, particularly audio equipment, and seem to enjoy exploring the features and specifications of these products.

Personalized Summary A: The Sony MiniDisc recorder is a great choice for music enthusiasts who value sound quality, portability, and ease of use. With its compact and lightweight design, this recorder is perfect for listening to music on the go. The built-in AM/FM radio is a nice-to-have feature that allows users to enjoy their favorite radio stations while on the move. The recorder also comes with a rechargeable battery that lasts for a long time, making it a great choice for users who want to enjoy their music without worrying about running out of power. The Sony MiniDisc recorder is also a great choice for users who want to record music. With its digital input and output, users can easily record music from their CDs or other digital sources. The recorder also comes with a built-in microphone, making it easy to record music from live performances or other sources. In terms of sound quality, the Sony MiniDisc recorder delivers. With its high-quality digital sound, users can enjoy their music with crystal-clear sound and minimal distortion. The recorder also comes with a range of features that allow users to customize their listening experience, including AVLS, track editing, and a 30-space program for radio stations.

Personalized Summary B: Sony MiniDisc players are highly recommended for music enthusiasts who value portability, sound quality, and the ability to record and edit tracks. They are particularly suited for users who prefer physical media over MP3 players or streaming services. Models like the MZ-N505 and MZ-R700 are praised for their affordability and functionality, while higher-end models like the MZ-N910 offer advanced features such as ultra-fast music transfer and sleek designs. For those seeking a versatile, durable, and stylish audio device, Sony MiniDisc players remain a solid choice, especially for users who appreciate the ability to record from various sources, enjoy jog-proof playback during physical activity, and want to keep their music collection safe from scratches and damage.

The options are:

- Summary A
- Summary B

T Full Evaluation Results (All Backbones)

Model	Ciao P	Ciao R	Ciao F1	Amazon P	Amazon R	Amazon F1
HelpfulSumm−RL (W + V)						
+ Llama	**0.823**	**0.800**	**0.811**	N/A		
+ Mistral	0.746	0.698	0.721			
HelpfulSumm−RL(V)						
+ Llama	0.796	0.779	0.787	N/A		
+ Mistral	0.747	0.686	0.715			
HelpfulSumm−RL (W)						
+ Llama	0.763	0.719	0.741	**0.618**	**0.838**	**0.712**
+ Mistral	0.735	0.639	0.684	0.547	0.754	0.634
HelpfulSumm−ICL (W + V)						
+ Llama	0.679	0.616	0.646	N/A		
+ Mistral	0.649	0.564	0.603			
HelpfulSumm−ICL (V)						
+ Llama	0.671	0.600	0.634	N/A		
+ Mistral	0.653	0.550	0.597			
HelpfulSumm−ICL (W)						
+ Llama	0.670	0.570	0.616	0.555	0.703	0.620
+ Mistral	0.650	0.541	0.590	0.439	0.637	0.519
Rehearsal (W + V) (Zhang et al., 2025)						
+ gpt-4o	0.674	0.673	0.674	N/A		
+ gpt-4.1-mini	0.664	0.632	0.647			
Rehearsal (V) (Zhang et al., 2025)						
+ gpt-4o	0.659	0.657	0.658	N/A		
+ gpt-4.1-mini	0.645	0.643	0.644			
Rehearsal (W) (Zhang et al., 2025)						
+ gpt-4o	0.624	0.623	0.624	0.319	0.764	0.450
+ gpt-4.1-mini	0.613	0.614	0.613	0.276	0.731	0.400
PersonaSumm (W)						
+ gpt-4o	0.534	0.693	0.603	0.270	0.658	0.383
+ gpt-4.1-mini	0.489	0.682	0.569	0.243	0.684	0.359
GeneralSumm						
+ gpt-4o	0.481	0.583	0.527	0.178	0.629	0.277
+ gpt-4.1-mini	0.463	0.586	0.517	0.171	0.630	0.269

Table 15: Proportion of helpful KPs, i.e., opinions, in the personalized summary

Table 16 reports the summary and KP-level textual quality of all baselines' backbone models, in extension to Table 3 (§6.1). Table 15 reports summary helpfulness, i.e., the proportion of helpful of all baselines' backbone models, in extension to Table 6 (§6.3).

U 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 17, 18, and 19 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 17, higher SHKP and SHS from 18 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.

V 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 based on his/her reviews history) and the summary outputs of different models are in Table 20. Overall, HelpfulSum-RL stands out by capturing more helpful opinions in the personalized summaries with strong alignment with user profile, while also ensures diversity by covering both positive and negative opinions. Given a user with "Evaluation Priorities" focusing on *low interest rate*, *transparent fee structure*, and *customer service*, HelpfulSumm-RL provides detailed illustration for *loan calculator* and *no penalties for early repayment*, while also emphasizes the entity downside such as *the loan process can be lengthy*, and *slow customer service*. In contrast, while Rehearsal and GeneralSumm mention diverse core entity features (e.g., "*attractive rates*", "*customizable repayment options*"), they are generic and lacks practical illustration with personalized details to showcase the suitability of the features to the user.

	ROUGE						BERTScore				BARTScore			
	KP			Summ			KP			Summ	KP			Summ
	R-1	R-2	R-L	R-1	R-2	R-L	sP	sR	sF1		sP	sR	sF1	
Ciao														
HelpfulSumm−RL (W + V)														
+ Llama	**0.538**	**0.324**	**0.505**	**0.347**	0.071	0.161	**0.49**	**0.45**	**0.47**	**0.25**	**0.72**	**0.78**	**0.75**	**0.51**
+ Mistral	0.496	0.289	0.462	0.345	**0.081**	**0.164**	0.44	0.41	0.43	0.23	0.70	0.73	0.71	0.46
HelpfulSumm−RL (V)														
+ Llama	0.533	0.316	0.494	0.339	0.075	0.158	0.48	0.44	0.46	0.24	0.72	0.77	0.74	0.50
+ Mistral	0.510	0.283	0.460	0.342	0.070	0.152	0.44	0.42	0.43	0.23	0.69	0.72	0.70	0.45
HelpfulSumm−RL (W)														
+ Llama	0.518	0.305	0.483	0.337	0.068	0.150	0.47	0.44	0.45	0.24	0.71	0.76	0.73	0.48
+ Mistral	0.484	0.275	0.451	0.339	0.063	0.145	0.43	0.42	0.43	0.23	0.68	0.71	0.69	0.44
HelpfulSumm−ICL (W + V)														
+ Llama	0.456	0.238	0.417	0.326	0.057	0.149	0.42	0.40	0.41	0.23	0.67	0.74	0.71	0.43
+ Mistral	0.385	0.165	0.359	0.280	0.055	0.144	0.38	0.40	0.39	0.22	0.63	0.68	0.66	0.39
HelpfulSumm−ICL (V)														
+ Llama	0.448	0.235	0.410	0.319	0.057	0.145	0.40	0.39	0.40	0.23	0.67	0.73	0.70	0.42
+ Mistral	0.367	0.152	0.345	0.277	0.052	0.141	0.37	0.39	0.38	0.22	0.62	0.68	0.65	0.39
HelpfulSumm−ICL (W)														
+ Llama	0.435	0.230	0.392	0.311	0.052	0.140	0.38	0.39	0.39	0.22	0.66	0.73	0.69	0.41
+ Mistral	0.355	0.133	0.329	0.269	0.050	0.135	0.36	0.37	0.37	0.21	0.61	0.67	0.63	0.38
Rehearsal (W + V) (Zhang et al., 2025)														
+ gpt-4o	0.401	0.192	0.359	0.286	0.049	0.139	0.38	0.43	0.40	0.23	0.69	0.75	0.72	0.43
+ gpt-4.1-mini	0.397	0.185	0.351	0.279	0.052	0.136	0.36	0.42	0.38	0.22	0.68	0.72	0.70	0.41
Rehearsal (V) (Zhang et al., 2025)														
+ gpt-4o	0.385	0.183	0.347	0.282	0.047	0.136	0.37	0.44	0.40	0.23	0.68	0.76	0.72	0.42
+ gpt-4.1-mini	0.379	0.180	0.336	0.276	0.046	0.134	0.36	0.43	0.39	0.22	0.68	0.72	0.70	0.41
Rehearsal (W) (Zhang et al., 2025)														
+ gpt-4o	0.366	0.158	0.323	0.272	0.043	0.128	0.36	0.44	0.39	0.22	0.68	0.70	0.69	0.41
+ gpt-4.1-mini	0.357	0.143	0.312	0.257	0.041	0.122	0.34	0.43	0.38	0.21	0.66	0.70	0.67	0.40
PersonaSumm (W)														
+ gpt-4o	0.349	0.145	0.316	0.204	0.035	0.105	0.35	0.43	0.38	0.20	0.69	0.62	0.65	0.39
+ gpt-4.1-mini	0.356	0.125	0.313	0.196	0.033	0.096	0.34	0.41	0.37	0.19	0.67	0.60	0.63	0.38
GeneralSumm														
+ gpt-4o	0.306	0.113	0.288	0.111	0.019	0.065	0.34	0.41	0.37	0.19	0.63	0.54	0.58	0.35
+ gpt-4.1-mini	0.311	0.112	0.272	0.111	0.017	0.065	0.33	0.40	0.36	0.19	0.62	0.51	0.56	0.34
Amazon														
HelpfulSumm−RL (W)														
+ Llama	**0.536**	**0.338**	**0.512**	**0.195**	0.041	**0.111**	**0.38**	**0.50**	**0.43**	**0.24**	**0.62**	**0.78**	**0.68**	**0.48**
+ Mistral	0.501	0.319	0.477	0.191	**0.041**	0.111	0.36	0.50	0.42	0.23	0.59	0.75	0.66	0.45
HelpfulSumm−ICL (W)														
+ Llama	0.489	0.287	0.462	0.163	0.027	0.102	0.35	0.49	0.40	0.22	0.59	0.74	0.65	0.44
+ Mistral	0.464	0.259	0.447	0.164	0.024	0.097	0.32	0.48	0.38	0.21	0.57	0.68	0.62	0.42
Rehearsal (W) (Zhang et al., 2025)														
+ gpt-4o	0.491	0.286	0.464	0.167	0.021	0.105	0.30	0.47	0.36	0.20	0.58	0.74	0.65	0.44
+ gpt-4.1-mini	0.486	0.261	0.452	0.162	0.020	0.092	0.29	0.45	0.35	0.20	0.57	0.73	0.64	0.43
PersonaSum (W)														
+ gpt-4o	0.473	0.283	0.450	0.109	0.015	0.080	0.24	0.42	0.31	0.17	0.53	0.60	0.57	0.37
+ gpt-4.1-mini	0.451	0.254	0.420	0.106	0.014	0.075	0.21	0.36	0.26	0.15	0.53	0.53	0.53	0.33
GeneralSumm														
+ gpt-4o	0.418	0.207	0.393	0.092	0.010	0.063	0.20	0.38	0.26	0.15	0.49	0.49	0.49	0.30
+ gpt-4.1-mini	0.376	0.185	0.353	0.075	0.009	0.046	0.17	0.35	0.23	0.14	0.48	0.46	0.47	0.28

Table 16: 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 from real (gold) user-written reviews on the product.

	ROUGE						BERTScore				BARTScore			
	KP			Summ			KP			Summ	KP			Summ
	R-1	R-2	R-L	R-1	R-2	R-L	sP	sR	sF1		sP	sR	sF1	
HelpfulSumm−RL (V + W)														
+ Llama	**0.538**	**0.324**	**0.505**	**0.347**	**0.071**	**0.161**	**0.49**	**0.45**	**0.47**	**0.25**	**0.72**	**0.78**	**0.75**	**0.51**
HelpfulSumm−RL (V + W) (w/o Persona Alignment Reward)														
+ Llama	0.529	0.314	0.488	0.340	0.070	0.159	0.48	0.44	0.46	0.25	0.72	0.77	0.74	0.50
HelpfulSumm−RL (V + W) (w/o Helpful Opinion Reward)														
+ Llama	0.497	0.280	0.452	0.332	0.067	0.159	0.47	0.43	0.45	0.24	0.71	0.76	0.73	0.49
HelpfulSumm−ICL (V + W)														
+ Llama	0.456	0.238	0.417	0.326	0.057	0.149	0.42	0.40	0.41	0.23	0.67	0.74	0.71	0.43

Table 17: **(Ablation Study)** KP summary textual quality of HelpfulSumm-RL without Persona Alignment Reward or Helpful Opinion Reward

Model	**Precision**	**Recall**	**F1**
HelpfulSumm−RL (V + W)			
+ Llama	**0.823**	**0.800**	**0.811**
HelpfulSumm−RL (V + W) (w/o Persona Alignment Reward)			
+ Llama	0.812	0.790	0.800
HelpfulSumm−RL (V + W) (w/o Helpful Opinion Reward)			
+ Llama	0.785	0.779	0.782
HelpfulSumm−ICL (V + W)			
+ Llama	0.776	0.767	0.772

Table 18: **(Ablation Study)** Proportion of helpful KPs, i.e., opinions, in the personalized summary of HelpfulSumm-RL without Persona Alignment Reward or Helpful Opinion Reward

	KP Information Quality					**Summary Personalization Quality**					
	CV	RD	SN	IN	SA	PT	CI	EH	EP	SB	IA
HelpfulSumm RL (V + W)	**40.23**	**40.22**	**40.54**	**54.08**	**48.07**	**46.19**	**45.93**	**45.46**	**50.46**	**48.81**	**49.5**
HelpfulSumm-RL (V + W) (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 (V + W) (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 19: 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).

Product	Gillette Mach 3 Razor (Beauty)
User Profile	Personality Traits: - Analytical Thinker: The user demonstrates a methodical approach to decision-making, particularly in financial matters. They carefully evaluate options, compare interest rates, and consider the long-term implications of their financial choices. - Pragmatic: The user shows a practical mindset, focusing on solutions that work best for their current situation, such as seeking flexible loan options and prioritizing low-interest rates. - Detail-Oriented: The user pays attention to details, as evidenced by their thorough reviews of financial products and services, including specific interest rates, fees, and customer service experiences. Core Interests: - Finance and Personal Loans: The user is primarily interested in personal finance, particularly loans, credit cards, and banking services. They seek to improve their financial situation and credit rating. - Customer Service Quality: The user values good customer service and shares experiences that highlight the importance of helpful and responsive support in financial transactions. Expression Habits/Style: - Descriptive and Informative: The user provides detailed accounts of their experiences, often including specific figures and comparisons to help others understand their perspective. - Balanced Perspective: They present both positive and negative aspects of their experiences, allowing for a well-rounded view of the products and services they review. Evaluation Priorities: - Interest Rates and Fees: The user prioritizes low interest rates and transparent fee structures when evaluating financial products. - Customer Service: They place significant importance on the quality of customer service, often highlighting experiences with representatives and the efficiency of processes. - Flexibility: The user values flexibility in financial products, such as the ability to make overpayments, take payment holidays, and adjust repayment terms. ...
	Generated Summary
HelpfulSumm-RL	If you're looking for a loan that offers competitive rates and flexible repayment options, Egg Loans may be a good choice for you. With their loan calculator, you can easily determine how much you can borrow and how much your monthly payments will be. Additionally, Egg Loans offers no penalties for early repayment, which can be a big advantage if you're able to pay off your loan quickly. However, it's worth noting that Egg Loans may not be the best choice for everyone. Some users have reported that the loan process can be lengthy, and that the company may try to sell you additional products or services that you don't need. Additionally, some users have reported that the company's customer service can be slow to respond to questions or concerns.
Rehearsal (Voted + Written) (Zhang et al., 2025)	Egg Loans is a highly recommended option for individuals seeking flexible, competitive, and hassle-free personal loans. The combination of attractive rates, customizable repayment options, and excellent customer service makes Egg a standout choice in the lending market. While borrowers should carefully evaluate the need for optional payment protection, the overall experience with Egg is overwhelmingly positive. For those who value flexibility and convenience, Egg Loans is a strong contender.
GeneralSumm	Egg Loans is highly recommended for individuals seeking flexible, competitive, and hassle-free personal loans. The combination of attractive rates, customizable repayment options, and excellent customer service makes Egg a standout choice in the lending market. Borrowers are advised to use the loan calculator and budget planner to make informed decisions and carefully consider optional extras like payment protection.

Table 20: Comparison of personalized summary output produced by HelpfulSumm-RL and previous personalized summarization and general summarization baselines, given a product and a user with profile inferred from his/her reviews history. Similar opinions across baselines are correspondingly marked in the same color.