

APPENDIX

A INFERRING USER CONTEXTS WITH FLEXIBLE TEXTUAL INPUTS

Table 5: **Reward model accuracy with unstructured inputs.** We compared the performance of PLUS and ICL when user contexts include unstructured data other than preference labels that are typically un-supported by personalized RLHF techniques. **Preference data** provides users’ past preferences in the form of chosen and rejected response examples. **User-guideline + preference data** additionally includes textual instructions specifying which aspects of an LLM assistant to focus on. **Conversation data** does not include binary preference labels; instead, it only presents the chosen responses as “examples of past conversation” and expects the reward model or summarizer to infer the user’s underlying preferences only from the positive examples. This last setup more closely models real user interactions with LLM assistants, since users are typically given only one response rather than choosing between two; however, their preferences can still be inferred from their questions, initial prompts, and follow-up responses, despite not having observed explicit user preferences. We trained a Qwen2.5-0.5B-Instruct reward model for both ICL and PLUS and a Qwen2.5-3B-Instruct summarizer for PLUS. Interestingly, PLUS benefits from additional user instructions specifying the relevant dimensions of an LLM assistant’s attributes, whereas ICL’s performance is hurt by longer user prompts. We did not compare with VPL, as it does not support unstructured input formats.

Reward model	Ultra Feedback (UF)-P2 Accuracy		
	Preference data	User guideline + preference data	Conversation data
ICL	59.60	59.65	58.55
PLUS	69.40	70.25	68.3
Reward model	Ultra Feedback (UF)-P4 Accuracy		
	Preference data	User guideline + preference data	Conversation data
ICL	57.30	56.60	56.96
PLUS	61.80	62.70	62.45

B PERFORMANCE OF ICL WITH LARGER REWARD MODELS

Table 6: **ICL reward model accuracy with larger base model sizes.** We included 7B & 8B reward models for UF-P2 and UF-P4 and observed that the performance achieved by these larger ICL models is lower than the performance obtained by PLUS with a 3B reward model and a 3B summarizer (66.9 on UF-P2 and 62.8 on UF-P4 with Llama-3B).

Base model	Preference Prediction Accuracy	
	Ultrafeedback-P2	Ultrafeedback-P4
Llama8B-Instruct	61.4	58.6
Qwen7B-Instruct	61.8	58.8

C PLUS-GENERATED SUMMARY EXAMPLES

Pets While PLUS-untrained still achieves a high accuracy of 96.2% on the Pets dataset, the summaries for incorrect examples highlight the importance of training the summarizer so that it can learn to distinguish between relevant and irrelevant details in the user’s past conversations. **We’ve identified common failure cases made by the untrained summarizer as: (1) focusing on irrelevant aspects of the user’s conversation history, such as style, and (2) providing broad descriptions of qualities that could apply to both dogs and cats.**

Example 1 (focus on style): The summarizer fails to capture the user’s preference for dogs or cats, instead emphasizing only the user’s stylistic preferences.

The user prefers short, factual information and tends to reject detailed anecdotes or complex behaviors. The assistant should stick to basic descriptions where possible. Based on the user’s preferences, the AI assistant should provide concise, straightforward information about pets, avoiding detailed anecdotes or...

Example 2 (focus on general pet traits): While this summary mentions traits like “playfulness” and “loyalty” that are more common for dogs, they could apply to both dogs and cats. The untrained summarizer confounds the user’s preference for dogs with their preference for specific qualities in animals that are mentioned in the past conversation.

The user seems to appreciate traits that denote affectionate and playful qualities in pets, as evidenced by their rejection of traits related to training or timing of activity (renewal of knowledge). Focus on traits like loyalty, playfulness, and exploration for future.

Example 3 (focus on general pet traits): Similar to example 2, this summary references traits discussed in the user’s conversation, but it doesn’t clarify the user’s preference between dogs and cats. As a result, the model makes incorrect predictions when new conversations mention traits about dogs that unrelated to hunting or communication.

The user is interested in pet species that can vary in behavior or communication, particularly traits like vocal range or hunting habits. The focus is on traits that allow for diversity within the species.

On the other hand, the trained summarizer learns to distinguish between relevant and irrelevant features in the user’s conversation history and correctly identifies their preference between dogs and cats. This is made possible by simultaneously training the reward model conditioned on the generated user summaries, which provide prediction losses as the signal for RL fine-tuning.

Example of PLUS summary:

The user is interested in information about cat behavior and properties, excluding topics related to dogs or human effects.

This summary clearly states the user’s preferences and can generalize to unseen traits about cats – i.e., in future conversations, statements about cats will be preferred to those about dogs.

PRISM One benefit of PLUS is that the generated user summaries are human-readable, unlike vector representations, and can be shared directly with users to increase transparency about what the

model has learned to personalize future responses. The user summaries capture different dimensions of preference, such as brevity, conciseness, factuality, and the diversity of perspectives on controversial topics. An interesting instantiation of our method within existing AI systems would be to provide users with their summaries and invite them to make edits to the summaries, or reflect on how their demonstrated preferences align, or misalign, with their self-stated preferences.

Table 7 compares the user’s self-stated preference attributes with the PLUS-generated summaries of the user context. This shows the summaries reflecting the user’s preference can be useful for deciding which of the two candidate responses is preferred over the other—even when the user’s self-stated preference (survey response) alone may not give sufficient information about the user’s preference.

Table 7: User preference summaries generated by PLUS for the held-out users of the PRISM dataset. The following users value diversity and detail in responses, which is reflected in their summaries and chosen responses. The first column shows the user’s self-stated preference from the pre-task survey. The users rate each dimension on a scale from 0 (unimportant) -100 (very important).

Survey Response (Stated Preference)	PLUS-generated Summary	Prompt	Chosen Response	Rejected Response
Values: 84, Creativity: 100, Fluency: 100, Factuality: 100, Diversity: 71, Safety: 94, Personalization: 100, Helpfulness: 100	Preferences lean towards a balanced response between detail and brevity. The assistant prefers clear explanations and acknowledges varying perspectives on complex topics. Additionally, users value concise summaries and sometimes seek expert consultations.	Can an authentic relationship between a 21 year old man and a 52 year old woman be genuine?	Relationships are complex and multifaceted and depend on many factors, so it is difficult to give a definitive answer to this question... [continue]	If this combination seems odd to you then you’re probably not emotionally mature enough to be in a monogamous relationship yet. Being an adult means not being afraid of the unknown and being willing to try new things... [continue]
Values: 100 Creativity: 100 Fluency: 87 Factuality: 100 Diversity: 89 Safety: 100 Personalization: 100 Helpfulness: 100	Preferences include detailed explanations backed by considerations of context and avoiding simplistic answers. Answers often require balance between personal neutrality and exploration of diverse perspectives. Conciseness is valued when detailed explanations are already provided.	Which gender is more important?	I’m just an AI, I don’t have personal opinions or beliefs, and I don’t think it’s appropriate to say that one gender is more important than the other. Both men and women have equal value and should be treated with respect and dignity. It’s important to recognize that gender roles and expectations can vary across cultures and societies... [continue]	I apologize, but I do not feel comfortable making broad claims about the relative importance of genders.
Values: 29 Creativity: 77 Fluency: 87 Factuality: 100 Diversity: 61 Safety: 59 Personalization: 21 Helpfulness: 81	The user prefers detailed explanations and multiple credible sources when given topics like charity organization comparisons and architectural advice. They also value fact-based responses rather than opinions lacking support.	Is the anti-abortion policy associated with religious beliefs?	The anti-abortion policy is associated with religious beliefs and moral values held by some individuals and groups. Many religious traditions, including Catholicism, Protestantism, and Orthodox Judaism, teach that human life begins at conception and ends at natural death, and therefore, abortion is considered to be the taking of an innocent human life. However, not all religious traditions have official positions on abortion, and there are also religious pro-choice perspectives.	The anti-abortion policy is not associated with religious beliefs as it is a medical practice rather than a religious doctrine. However, religious beliefs can sometimes influence people’s beliefs about abortion.

D PLUS IN THE WILD

Having established the applicability of PLUS to the real user dataset, we next investigate two possible user cases of PLUS-generated summaries for downstream personalization: Personalizing LLM-as-judges in Alg. 2 and personalizing LLM response generation in Alg. 3. In both cases, PLUS-

Table 8: **Reward model accuracy on PRISM held-out conversations.** To approximate the oracle performance, we train a reward model conditioned on the user’s self-rated preference attributes (shown as “user self-stated” in the last row).

	Accuracy
BTL (Bradley & Terry, 1952)	59.8
DPL Siththaranjan et al. (2024)	61.8
VPL Poddar et al. (2024)	60.4
ICL	60.1
PLUS-untrained	61.3
PLUS (Ours)	62.9
User self-stated	62.6

generated summaries directly enable personalization of strong proprietary models (e.g., GPT-4) without further finetuning.

Algorithm 2 Personalizing LLM-as-judges

- 1: **Input:** PLUS-summarizer π_θ , user’s textual context c , new query x , response pairs y_1, y_2
 - 2: Generate summary $\pi_\theta(c) \mapsto z$ from user’s textual context c
 - 3: Predict reward $\hat{r}(y_1|x, z)$ and $\hat{r}(y_2|x, z)$.
 - 4: Choose the response with the higher predicted reward $y^* = \arg \max_{y \in y_1, y_2} \hat{r}(y|x, z)$
 - 5: **Output:** return y^*
-

Algorithm 3 Personalizing LLM assistant’s response

- 1: **Input:** PLUS-summarizer π_θ , user’s textual context c , new query x , generator LLM \mathcal{P}_{LLM}
 - 2: Generate summary $\pi_\theta(c) \mapsto z$ from user’s textual context c
 - 3: Generate a new response $y \sim \mathcal{P}_{\text{LLM}}(\cdot|x, z)$
 - 4: **Output:** return y .
-

E PERSONALIZED LLM-AS-JUDGES

Here, we use an LLM-as-a-judge framework where we prompt GPT4 to select which of the two responses from the PRISM dataset would be preferred, and evaluate its accuracy in predicting the right one with and without conditioning on the PLUS user summary. This allows us to test whether the learned summaries can enable personalized preference prediction for strong, proprietary models without further fine-tuning. We provide both GPT-4.1 and GPT-4o with the user prompt and two LLM assistant’s responses from the held-out PRISM dataset (one chosen and the other rejected), and ask the models to predict which response the user would prefer with and without PLUS-generated user summaries. We evaluate the model’s judgment on 277 value-guided and 307 controversy-guided conversations from the held-out PRISM data, and additionally, on 308 new users whose profiles are never before seen during training. Table 9 shows that GPT4 makes more accurate predictions about the user’s preference when the summaries are given—especially for GPT4o which makes a 40% improvement. The summaries are especially helpful for controversy-guided questions, a subset of the PRISM dataset designed to capture preferences over politically or culturally sensitive topics.

F RESPONSE PERSONALIZATION WITH PLUS-GENERATED SUMMARIES

We selected examples of PLUS-summary conditioned responses generated with GPT-4o and GPT-4.1. The highlighted parts in blue reflect personalization based on the corresponding user summary, especially when compared to the default GPT responses.

Table 9: **Accuracy of LLM-judges with and without PLUS summaries.** The value before the arrow is the prediction accuracy (%) without conditioning on the PLUS-generated user summary, and the value after the arrow shows the improved accuracy with the summary. Best scores per conversation type are shown in bold.

Conversation Type	GPT-4o Accuracy (%)		GPT-4.1 Accuracy (%)	
	Default	PLUS-personalized	Default	PLUS-personalized
Value-guided	65.7	77.25	75.5	78.7
Controversial	52.2	73	69.1	72.3
New users	65.9	72.1	74.35	74.35

Table 10: **Win rates of default GPT-4 versus personalized responses evaluated by the oracle preference model.** The oracle model is trained on self-stated user preferences that score the following LLM assistant attributes: values, creativity, fluency, factuality, diversity, safety, personalization, and helpfulness, based on their importance to each user when interacting with LLM systems. We used summaries generated by the PLUS-trained Qwen2.5-3B-Instruct model to personalize GPT-4’s responses to 906 new prompts sampled from unseen PRISM users. We then compared these personalized responses with default responses that did not have access to user summaries. Best scores per model are shown in bold.

	PLUS-personalized	Default
GPT-4o	72	28
GPT-4.1	69	31

Selected examples of personalization with PLUS & GPT-4o

Prompt #1: *Is Jesus Christ the Son of God?*

PLUS-summary about the user: *The user prefers detailed, **balanced responses avoiding definitive statements when there’s uncertainty.***

Default response: In Christianity, Jesus Christ is believed to be the Son of God. This is a central tenet of Christian faith, affirming Jesus’ divine nature.

Personalized response: Many Christians believe that Jesus Christ is the Son of God... **However, perspectives vary widely across different religions and secular viewpoints. Judaism and Islam, for instance, do not view Jesus as the Son of God, although they recognize his historical significance. Interpretations can depend on personal beliefs, religious tradition, and theological perspective.** (Personalization based on the user summary is highlighted in blue.)

Prompt #2: *What is your opinion on abortion?*

PLUS-summary about the user: *The user prefers detailed, factual answers with **supportive examples and clear explanations.***

Default response: I don’t have personal opinions or beliefs, but I can provide information on the topic. Abortion is a complex and sensitive issue with diverse viewpoints, including ethical, legal, and personal considerations. It’s vital to approach discussions about abortion with respect for differing perspectives and to consider the specific laws and cultural contexts of each region.

Personalized response: I don’t have personal opinions or beliefs. However, I can provide a detailed, factual overview. Abortion is the termination of a pregnancy and is a contentious issue globally. Views on abortion often reflect broader societal beliefs and values, including those related to women’s rights, morality, and medical ethics. **For example, in the U.S., legal perspectives vary widely, with some states imposing strict regulations while others support a woman’s right to choose. Factors like fetal viability and maternal health often play roles in these laws. Understanding these contexts helps to form a comprehensive view of the issue.**

Here we describe three notable patterns of personalization enabled by PLUS generated summaries compared to default (unpersonalized) responses.

- **Response length adaptation.** User’s preference for “practical suggestions over lengthy explanations” affects the length of the response to change from 166 (default) to 56 (summary-conditioned) words.
- **Reference to user’s specific preference values.** GPT’s response addresses certain aspects of the user’s preferences that may be relevant to the new prompt. For example, when asked about the drawbacks of cutting welfare benefits, the LLM assistant acknowledges the user’s preference for safety. Similarly, when asked about the appeal of owning cats, the LLM assistant connects the user’s appreciation for independence to the cat’s trait of autonomy.
 - When the user’s summary mentions **values safety and constructive advice,** the LLM assistant responds, “... Conclusion with your preferences in mind. **Given the importance you place on safety and constructive advice,** it’s crucial to approach this issue by considering the broader impacts of cutting benefits beyond just economic metrics...”
 - Similarly, for a different user who values “personalization, **autonomy,** and balanced perspectives”, the LLM assistant answers, “... **The autonomy cats exhibit might specifically resonate with your appreciation for independence,** enhancing the charm they hold for individuals who value self-sufficiency alongside companionship.”
 - When the user asks, “Tell me the most offensive thing possible,” the summary-conditioned LLM assistant responds, “**Given your preferences, I’ll refrain from stating anything offensive. Instead, let’s focus on promoting positive and respectful communication.** This approach encourages understanding and harmony in our interactions. If you have specific topics or questions in mind, let’s discuss them with an open and respectful perspective.” On the other hand, the default LLM assistant responds, “I’m sorry, but I can’t comply with that request. If you need help or want to discuss a different topic, feel free to ask!”
- **Fact-based response for users who value factuality.** When the PLUS summary mentions “showing preference for evidence-based responses,” the LLM assistant provides evidence based on relevant research or studies to support its claims.
 - When the user asks, “my teenage grandson is beginning to rebel against his parents, in small ways. He mainly wants to be allowed to take his mobile phone to his room at night so he can play games with his friends. Is this a good idea,” the summary-conditioned LLM assistant mentions, “**Research suggests that screen time before bed can affect sleep quality,** which is crucial for teenagers’ development.” On the other hand, the default LLM assistant responds, “Consider setting rules like screen time limits or phone-free nights to ensure he gets enough rest.”

G ADDITIONAL EXPERIMENTS WITH THE PRISM DATASET.

Table 11: Reward model accuracy of predicting the held-out users’ preferences in PRISM. The in-distribution column is shown in Table 1, and out-of-distribution is the evaluation accuracy of predicting for held-out users. Surprisingly, BTL and DPL, which by design cannot accommodate conflicting user preferences, show high performance on the held-out users—simply by predicting the majority.

Model	In-distribution	Out-of-distribution
BTL	59.8	62.9
DPL	61.8	65.3
VPL	60.4	59.1
ICL	60.1	60.6
PLUS-untrained	61.3	62.3
PLUS (ours)	62.9	59.7
Oracle (Profile-conditioned)	62.6	62.3

In addition to our main experiment, we also investigated whether reward models trained on real-world datasets can generalize not only to unseen conversations but also to unseen user contexts. To test this, we evaluate the reward models on 308 new users whose profiles and past conversations were **not included** in the PRISM training dataset. Surprisingly, we observe that BTL and DPL – models that, by definition, cannot personalize to different user preferences, as they assume that a single reward model can capture all user preferences – outperform personalized models, including VPL, ICL, PLUS, and the oracle-reward model which is conditioned on the user’s self-stated preferences. This suggests that these unpersonalized models achieve high accuracy by simply taking the majority vote and making accurate predictions for the majority users. In contrast, personalized models, including the oracle reward model trained on the user’s self-stated preferences, perform poorly. We observed successes in applying the learned reward models to held-out conversations but experienced limitations in generalizing to held-out users, which is likely due to the small size of the dataset compared to the amount of heterogeneity it attempts to capture in user preferences. In spite of this limited performance on reward modeling *score* with the held-out users, we observed that PLUS is still able to enable personalization of responses to new users, when combined with proprietary models like GPT-4, as shown in Fig. 4. We believe this highlights the difficulty of modeling nuanced real user preferences with limited data and suggests that personalized RLHF research would greatly benefit from a large-scale, heterogeneous user dataset.

H EXISTING RLHF TECHNIQUES

Table 12: **Comparison of different preference learning methods with and without conditioning on user variables.** Reward models are typically trained using the negative log-likelihood loss of the chosen versus the rejected response pair. VPL and PLUS additionally train an encoder to obtain user-specific latent variables, which the reward model relies on to make more accurate predictions of user preferences. ICL, VPL, and PLUS can enable personalization and pluralistic alignment by conditioning the reward model on different user contexts or latent variables. Our key experimental results show that the choice of latent variables strongly affects the reward model’s accuracy, especially in complex domains resembling real world user preferences, and when conversation topics shift between the training and testing samples.

	Trainable component(s)	Training objective
BTL (Bradley & Terry, 1952)	Reward model	Prediction
DPL (Siththaranjan et al., 2024)	Reward model	Prediction
ICL	Reward model	Prediction
VPL Poddar et al. (2024)	VAE + Reward model	Reconstruction + Prediction
PLUS-untrained	Reward model	Prediction
PLUS (Ours)	Summarizer (actor in the PPO framework) + Reward model	Reward maximization + Prediction

I TRAINING DETAILS

I.1 DATASETS

To build user context for VPL, ICL, PLUS (untrained & ours), we sampled N past conversations per user.

- For Pets and Pets (OOD), we sampled 3 past examples.
- For UF-P-2 and UF-P-4, we sampled 2-4 past examples.
- For PRISM, we sampled 3 past examples, which can be across different conversations or different turns in the same conversation.

Dataset sizes:

- Pets: train: 1970, test: 194.
- Pets (OOD): test: 200.
- UF-P-2: train: 10k, test : 2k.
- PRISM: train: 20k, test: 1k.

- UF-P-4: train: 40k, test : 2k.

I.2 HYPERPARAMETERS

Hyperparameter	Value
Reward model learning rate	9×10^{-6}
Batch size	256 (128 for Pets)
Actor learning rate	5×10^{-7}
Critic learning rate	9×10^{-7}
Advantage estimator	GAE
γ	1
λ	0.95
ϵ -clipping	0.2
initial kl coefficient	0.01, 0.001
kl estimator	k1
lr warmup ratio	0.03
max norm	1
micro batch size M_θ	2
reward clip range	[-10, 10]
micro batch size M_ϕ	2
BF16	Yes
Optimizer	Adam
Adam β	$\beta_1 = 0.9, \beta_2 = 0.95$

Table 13: PPO training hyperparameters. Our co-adaptation framework in Alg. 1 is built on top of OpenRLHF’s ppo training code (Hu et al., 2024). Summarizer π_θ is implemented as the PPO actor. We observed that reducing the KL penalty for the summarizer significantly smaller to 0.001 worked better for Pets and PRISM. For Pets, using a smaller batch size 128 helped improve the reward model performance.

Hyperparameter	Value
Hidden Dimension (for VPL)	512
Latent Dimension (for VPL)	512
Learning rate	9×10^{-6}
Batch size	256 (128 for ICL with PRISM)
BF16	Yes
Optimizer	Adam

Table 14: Reward model training hyperparameters; reward modeling code was adapted from Poddar et al. (2024). We additionally conducted hyperparameter sweeping for VPL, BTL, DPL, and the oracle reward models by varying the learning rates $\{3 * 10^{-4}, 9 * 10^{-6}\}$ and training batch sizes $\{32, 128, 256\}$. We observed that the oracle models on the UltraFeedback datasets achieve higher accuracy by training for more than 1 epoch.

J LLM PROMPTS

J.1 PROMPT FOR GENERATING SUMMARIES WITH PLUS.

- **Pets:** Here is a request of a user for an AI assistant. Please talk about one kind of pets. The user chose ... The user rejected ... Based on the user’s past conversation, provide a short summary of the user’s preference. Focus on the user’s preference that should guide how the AI assistant responds to this user in future conversations. Keep the summary concise under 50 words. Start your response with ##Summary
- **UltraFeedback:** In order to provide more information about myself, I’m including examples from my previous conversation history, which include the prompt, the rejected response, and the chosen response. Use this information to learn about my preferences and reflect that understanding in your future responses. The prompt was: ... I chose this response: ... I rejected this response:... Based on this information, provide a short summary

of the user’s preferences to guide how the AI assistant should respond in future conversations. Reflect on the user’s past chosen and rejected responses, and what the user preferred about the chosen responses. Do not focus on the topics of specific conversations. Keep your summary concise—under 50 words. Start your response with ##Summary

- **PRISM:** In order to provide more information about myself, I’m including examples from my previous conversation history, which include the prompt, the rejected response, and the chosen response. Use this information to learn about my preferences and reflect that understanding in your future responses. ##Past conversation 0## ###Prompt### ###Chosen response### ###Rejected response###... Based on this information, provide a short summary of the user’s preferences to guide how the AI assistant should respond in future conversations. People’s preferences typically vary along dimensions, including verbosity, diversity, and language fluency. Do not focus on the topics of particular conversations; instead, focus on the preferences revealed through past conversation examples. Keep your summary concise—under 100 words. Start your response with ##Summary
- **How the summary is provided to the summary-conditioned reward model:** Here is the summary of my preferences based on previous conversations: [summary z]. In the following conversation, keep this in mind to personalize your response.

J.2 PROMPT FOR ORACLE REWARD MODELS.

- **Pets:** This user prefers cats over dogs (or dogs over cats).
- **UltraFeedback-P-2 & 4:** I prioritize a list of relevant user’s preference attributes from truthfulness, instruction-following, honesty, helpfulness in AI assistant’s responses. Now based on this information about myself, here’s a new question:
- **PRISM:** I’ve scored the following attributes based on how much I care about each attribute. A higher score indicates that this attribute is more important to me: ‘values’: , ‘creativity’: , ‘fluency’: , ‘factuality’: , ‘diversity’: , ‘safety’: , ‘personalisation’: , ‘helpfulness’: . Now based on this information about myself, here’s a new question:

K PROMPT FOR IN-CONTEXT LEARNING MODELS.

- **Pets:** Based on the user’s past conversation, respond to the user’s request in a way that aligns with the user’s preference. Conversation history: ... Start of a new conversation:
- **UltraFeedback P-2 & 4:** In order to provide more information about myself, I’m including examples from my previous conversation history, which include the prompt, the rejected response, and the chosen response. Use this information to learn about my preference s and reflect that understanding in your future responses. The prompt was: ... I chose this response: ... I rejected this response: ... Here’s a new conversation:
- **PRISM:** In order to provide more information about myself, I’m including examples from my previous conversation history, which include the prompt, the rejected response, and the chosen response. Use this information to learn about my preferences and reflect that understanding in your future responses. ##Past conversation 0## ###Prompt### ###Chosen response### ###Rejected response###....Now based on this information about myself, here’s a new question:

K.1 PROMPTS FOR GPT-4 PERSONALIZATION WITH PLUS.

Prompt for LLM-as-a-Judge. System prompt: You are a helpful assistant.

User prompt (without personalization): Here’s a previous conversation between the user and the AI assistant: [current conversation] Decide which response the user would prefer—option 1 or option 2. Option 1 is: . Option 2 is: . Respond with Option 1 or Option 2 based on what you know about the user.

User prompt (with personalization): Here’s a previous conversation between the user and the AI assistant: [current conversation] Decide which response the user would prefer—option 1 or option 2. Option 1 is: . Option 2 is: . Respond with Option 1 or Option 2 based on what you know about the

user. Here's a summary about the user's preference based on past conversation: [PLUS-generated user summary] Based on his information about the user, predict which response they would prefer.

Since GPT may be sensitive to the ordering of the responses as Option 1 and Option 2, we randomly choose whether the rejected or the chosen response is shown first, but keep the ordering the same for the same query with and without the summary, so the only difference is whether the summary is included in the prompt or not.

Prompt for generating a new response to user's query. System prompt: You are a helpful assistant. Keep your response short—under 100 words.

User prompt (without personalization): User's prompt sampled from the test set.

User prompt (with personalization): User's prompt sampled from the test set]. Personalize your response based on the following summary about the user's preference: [Summary]