Active Listening: Personalized Question Generation in Open-Domain Social Conversation with User Model Based Prompting

Kevin K. Bowden, Yue Fan, Winson Chen,
Wen Cui, Davan Harrison, Xin Eric Wang, Marilyn Walker
University of California, Santa Cruz
{kkbowden,yfan71,wchen157,wcui7,vharriso,xwang366,mawalker}@ucsc.edu

Abstract

Large language models (LLMs) capable of casual conversation have recently become widely available. We hypothesize that users of conversational systems want a more personalized experience, and existing work shows that users are highly receptive to personalized questions (PQs). Question Generation tasks, however, focus on factual questions from textual excerpts. To create a PQ generator, we first identify over 400 real user interests by anonymously aggregating ∼39K user models. We then populate prompt templates with these 400 interests and use an LLM to generate PQs customized to user interests. The result is PerQs, a novel corpus of ∼19K question/answer pairs. We evaluate PerQs at scale in the unique context of the Alexa Prize. Our results show significant positive effects on perceived conversation quality. We then fine-tune, deploy, and evaluate PerQy, a neural model that generates PQs in real-time. When evaluated against several competitive LLM baselines, PerQy produced the most natural and engaging responses.

1 Introduction

Large language models (LLMs) capable of casual conversation have recently become widely available, leading to an increase in research in social open-domain dialogue (Higashinaka et al., 2021, 2014; Zhang et al., 2020; Kim et al., 2023; Zheng et al., 2023; Ouyang et al., 2022) *inter alia*. In addition, challenges like the Alexa Prize Socialbot Challenge (henceforth AP) (Gabriel et al., 2020; Hu et al., 2021b; Johnston et al., 2023) have given real users the ability to access and evaluate spoken conversational systems in their home. We hypothesize that users of such conversational systems want a more personalized experience (Ritschel et al., 2017; Sugiyama et al., 2014; Bickmore and Picard, 2005; Clark et al., 2019).

Research shows that conversational partners are more well-liked if they ask more follow-up questions (Huang et al., 2017), and such questions show that the hearer is listening and understanding (Bevacqua et al., 2012; Meguro et al., 2014; Reis et al., 2011; Reis and Patrick, 1996). We define personalized questions (PQs) as questions that match the user's expressed interests so that users feel they are customized to them rather than the general population. See Figure 1. Previous work using crowdsourced PQs shows a positive effect on user ratings of conversation quality (Sugiyama et al., 2014; Bowden and Walker, 2023). However, crowdsourced PQs can only cover general user interests specified when crowdsourcing. We propose dialogue policies based on a new task of Personalized Question Generation (PQG) and take advantage of the unique opportunity offered by the AP to test these policies at scale with real users. The PQG task aims to generate questions that will engage the user in small talk about a topic of interest to the user (Mattar and Wachsmuth, 2012; Klüwer, 2011; Endrass et al., 2011). As shown in Figure 1, PQs seek opinions, feelings, and experiences, and play a pivotal role in enhancing the interactivity of conversational systems (Chaves and Gerosa, 2021). PQs can also encourage individuals to talk about themselves (self-disclose), which can

User Interest	Personalized Question
Art **(DPQ)**	Have you ever had a piece of art make you emotional or feel a strong connection to it? What was the piece and how did you feel?
Mythology **(HYP)**	If you could have a conversation with any mythological character, real or fictional, who would it be and what would you ask them?
Cooking **(WYR)**	Would you rather cook an amazing dinner or the perfect dessert?

Figure 1: A Deep Personalized Question (DPQ), Hypothetical Personalized Question (HYP), and Would You Rather Personalized Question (WYR) from the PerQs corpus.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14120–14157

Content	Response
Question	If you had the chance to travel to any planet in the universe, which one would you choose and why?
System Answer	If I had the chance to travel to any planet in the universe, I would choose Mars. It has always fascinated me with its barren landscapes and the potential for future exploration. I would love to witness the unique Martian environment and learn more about the possibilities of human colonization.
Anticipated User Answers and Acknowledgments	
Pred. User Answer	**[Keywords]** venus, greenhouse effect, thick atmosphere **[Acknowledgement]** Choosing Venus would be intriguing. Its greenhouse effect and extreme temperatures make it a fascinating yet challenging destination to explore.
Pred. User Answer	**[Keywords]** earth, home planet, diverse ecosystems, life-supporting **[Acknowledgement]** Staying on Earth would be a wonderful choice. Our home planet's diverse ecosystem and life-supporting conditions make it a haven of beauty and wonder.

Figure 2: An example of a PQ for a user who is interested in Astronomy, system answer, and two predicted user answers, with their acknowledgments.

be helpful for user modeling (Lee et al., 2020a).[1]

Moreover, while users interact with our AP system using a commercial device, e.g., an Amazon Echo), there are significant non-commercial applications, as producing personalized content is especially important for vulnerable demographics, e.g., older adults (da Paixão Pinto et al., 2021; Bickmore et al., 2005), and social systems can be critical when dealing with loneliness (Rodríguez-Martínez et al., 2023; Liu et al., 2021), building emotional support (Jones et al., 2021), and in therapeutic environments (DeVault et al., 2014).

Recent LLM advances make it possible to create a compact, prompt-based model that will generate PQs in real time (Ouyang et al., 2022; Radford et al., 2019; Brown et al., 2020; Puri et al., 2023; Zhang et al., 2023). To fine-tune such a model, a specialized training corpus is required that combines a list of open-domain user interests with structured prompt templates to generate PQs. This paper describes **PerQs**, a corpus created for training a real-time PQ generator. We first identify over 400 common user interests by aggregating ∼39,000 user models, then instantiate a set of prompt templates with these interests. These are then fed to an LLM to generate ∼19,000 PQs of multiple types, such as those in Figure 1. See also Appendix D. We then fed these PQs back into the LLM to generate a pool of predicted user answers, as well as user answer-specific acknowledgments, and a system answer, as shown in Figure 2. Figure 3 shows how the dialogue policy prefixes an answer-specific acknowledgment to the system response, signaling the system's understanding of the user's answer.

For our **first** human evaluation, we integrate PerQs into our spoken AP system and evaluate their effect at scale with thousands of real anonymous human users. The results show a significant increase in user engagement and intimacy, and a significant positive effect on perceived conversation quality as measured by user ratings. We then use PerQs as training data to fine-tune **PerQy**, a real-time prompt-based PQ generator based on RedPajama-3B (AI, 2023). We deploy PerQy in the spoken AP system and conduct a **second** human evaluation of PerQy using the large pool of AP human users, again finding a significant positive effect of dialogue policies that produce PQs in real-time. We make PerQs, the evaluation resources, and PerQy publicly available.[2]

We then conduct a **third** human evaluation using a design similar to recent work (Kim et al., 2023; Chen et al., 2023a). Here, we compare PerQy head-to-head against recent LLMs that also target social conversation, such as COSMO (3B), RedPajama-INCITE-Chat-3B-v1 (RedPJ Chat), and DialoGPT (345M) (Kim et al., 2023; AI, 2023; Zhang et al., 2020), as well as more general and larger, state-of-the-art LLMs such as GPT-3.5-turbo and Vicuna (33B) (OpenAI, 2024; Zheng et al., 2023). We generate a set of short sub-dialogues contexts for 200 unique user interests and manually verify their quality, safety, and similarity to genuine open-domain dialogues. We then ask Amazon Mechanical Turkers to make a forced-choice of the best next system utterance in each context. The results show that PerQy produces more specific, engaging, and natural contributions to a social conversation than COSMO, RedPJ Chat, and DialoGPT, as well as more natural and specific contributions than the much larger Vicuna and GPT-3.5-turbo.

[1] AP user interactions are protected by privacy limitations, preventing the release of conversational data or access to user demographics. The testbed system uses safety filtering and was closely monitored to ensure that no inappropriate system responses were generated in real time (Fan et al., 2023).

[2] PerQs: https://huggingface.co/datasets/kkbowden/PerQs
PerQy: https://huggingface.co/kkbowden/PerQy

2 Related Work

User modeling, also called user profiling or personas, is used to adapt dialogue systems to individual users or user groups. Many different techniques have been used to create user profiles during a conversation (Ma et al., 2020; Li et al., 2014; Bang et al., 2015; Wu et al., 2020; Zhang et al., 2018), with some work aiming to detect explicit self-disclosure (Hirano et al., 2015). Traditional information extraction approaches struggle with conversation because implicit self-disclosures are easily missed by direct pattern matching (Tigunova, 2020; Tigunova et al., 2019). Other approaches base user models on Big-5 personality traits (Fang et al., 2018; Cassell and Bickmore, 2003) or Reddit personas (Baymurzina et al., 2021). Most similar to our work are approaches grounded in traditional user modeling techniques, e.g., rule-based extraction skimmers (Konrád et al., 2021), regular expressions (Finch et al., 2021), and other slot extraction mechanisms (Saha et al., 2021). Our user modeling mixes elements from these approaches but uses these models to generate PQs.

Previous work on user modeling and personalized follow-up content in open-domain dialogue systems has personalized both topic-level and sub-topic-level interests (Fang et al., 2018; Chen et al., 2018; Liang et al., 2020; Hong et al., 2020; Baymurzina et al., 2021; Konrád et al., 2021; Juraska et al., 2021). Other work has used user models to affect follow-up content and ask more personal questions to increase rapport (Curry et al., 2018; Chen et al., 2020; Finch et al., 2020, 2021; Bickmore and Picard, 2005). None of this work has produced the training data needed to fine-tune a model that generates PQs in real-time.

Previous work on Question Generation (QG) focused on fact-based questions whose answers can be found in text excerpts, e.g., Wikipedia and Gutenberg (Reddy et al., 2019; Fei et al., 2022; Do et al., 2022), or domain-specific questions targeted at a specific information need (Campos et al., 2020). Other conversational information-seeking question corpora, e.g., CCPE-M (Radlinski et al., 2019), elicit user preferences to inform a recommender system, while the PQG task is not intended to recommend or sell any product or service. Resources related to PQG are subsequently unsuitable for recommender tasks or tasks attempting to persuade users to take some action.

3 Generating Personalized Questions

Our personalized questions (PQs) are generated by instantiating prompt-templates with user interests, such as hobbies, topical interests, and other relevant sub-topic attributes, e.g., their favorite movie genre or pet's names. The user model does not store any Personally Identifiable Information except for detecting if the user is a child, which is necessary to guarantee age-appropriate content. Table 1 depicts a summarized user model from an in-lab tester pretending to be a child. The corresponding conversation is in Appendix A.3.

User Attribute	Value
discuss_topics	mermaids
favorite_animals	chinchillas
hobbies	chess, swimming
holiday_plans	July 4th: shoot fireworks
topical_interest	animals, board games
topical_disinterest	pirates
user_is_kid	true
user_is_new	true
user_name	sage
vacation_spots	hawaii

Table 1: A summarized sample user model.

We curated over 400 user interests by anonymously aggregating ∼39,000 user models. For user privacy, these extracted values are not raw user utterances but lists of detected normalized keywords and classifier labels, which are then aggregated over several months, after which only the head of the distribution is considered. Thus, it is impossible to trace user interests back to a specific user, protecting user privacy.

To personalize the experience to each specific user, we focus on extracting values most representative of the user's interests, including common hobbies and activities (e.g., dancing), popular topics (e.g., movies), esoteric interests (e.g., mythology), and sub-topic information (e.g., their favorite genre of music is pop). We instantiate prompt templates with these interests and use an LLM to generate three types of personalized questions: Would-You-Rather (WYR), Hypothetical (HYP), and Deep Personalized (DPQ) questions. WYR presents the user with two options, while HYP is designed to be open-ended. These strategies are based on prior work that shows that they increased open-domain dialogue system performance (Bowden and Walker, 2023). They are intended to build rapport between the system and the user (Fields, 2009; Shani et al., 2022; Meyer, 2015). DPQs differ by focusing on

long user-centric questions that aim to increase engagement and intimacy by provoking long responses (Glas and Pelachaud, 2015; Potdevin et al., 2018).

We used OpenAI's ChatCompletion API with the gpt-3.5-turbo model (OpenAI, 2024). The generation pipeline is illustrated in Appendix C. We experimented with different prompts; the final prompts are listed in Appendix B. We provided one entire prompt/response sequence as context to ensure high-quality outputs. Providing additional context led to longer DPQs. This distinction is shown in Appendix B.

[A1]	If you had the chance to travel to any planet in the universe, which one would you choose and why?
[U1]	maybe neptune
[A2]	Opting for Neptune would be captivating. Its striking blue color, icy winds, and position as the farthest planet from the Sun make it a mysterious and intriguing destination. If I had the chance to travel to any planet in the universe, I would choose Mars.
[A7]	If you had the chance to travel to any planet in the universe, which one would you choose and why?
[U7]	tatooine
[A8]	Oh, I see. If I had the chance to travel to any planet in the universe, I would choose Mars. It has always fascinated me with its barren landscapes and the potential for future exploration. I would love to witness the unique Martian environment and learn more about the possibilities of human colonization.

Figure 3: The dialogue policy changes when a predicted user answer is detected. The cyan segments are the system's answer-specific acknowledgment, while the yellow segment indicates the system's answer. When the user provides a predicted answer in **U1**, the system prefixes the answer-specific acknowledgment to a shortened version of its answer. Otherwise, as in **A8**, the system prefixes a short acknowledgment before providing a longer answer.

We then feed the generated PQs back into GPT-3.5 to generate the utterances needed for the dialogue policy, e.g., Figure 2. Namely, we generate *predicted user answers* in order to generate *answer specific acknowledgments*, which are useful for signaling system understanding and increasing engagement (Paranjape and Manning, 2021; Cathcart et al., 2003; Walker and Passonneau, 2001). Figure 3 shows how the acknowledgments dialogue policy changes when a predicted answer is detected. We also generate *system answers* to the PQs so that the system can self-disclose and support a dialogue policy that includes an exchange of opinions. While WYR PQs predict the two presented choices,

Content Type	Samples	Flagged	MTLD
Personalized Questions	18,901	0.74%	48.2
System Answers	18,901	1.32%	86.3
User Acknowledgments	114,328	1.39%	82.0

Table 2: The percentage of needs_intervention classifications by Canary (Kim et al., 2022) and the lexical density via MTLD (McCarthy and Jarvis, 2010).

DPQs and HYPs are open-ended. Subsequently, we pre-generated, on average, 7.3 predicted answers for each DPQ and HYP. This process results in **PerQs**, a novel corpus coupling ∼19,000 personalized question and answer pairs with prompts that are instantiated with real user interests. On average, each interest in PerQs has ∼47.5 generated PQs. Several generated examples are included in Figure 1, Appendix D, and Appendix E. Appendix F describes the distribution of user interests.

We evaluated the quality of PerQs manually. Two native English-speaking authors and one external annotator annotated 300 randomly sampled PQs (100 of each type) from PerQs and found that 97% of samples were safe, high-quality, and matched the target interest. Samples that did not meet this criterion were arguably too vague for the target interest, e.g., a question about "books" but not specifically "action books".

We post-processed the generated PerQs using a list of keywords to remove inappropriate content. After this process, Canary (Kim et al., 2022), a dialogue safety detection module that predicts whether there is risk, flagged $< 1.4\%$ of PQs as need_intervention. When inspected, most flagged PQs are innocuous (Appendix G). We also calculate the lexical density via MTLD to estimate the variation in topics, which is an important corpus characteristic (McCarthy and Jarvis, 2010). Table 2 shows the Canary and MTLD results for PQs, system answers, and user answer acknowledgments.

3.1 Evaluating Intimacy and Engagement

The **first** human evaluation takes advantage of the unique opportunity to evaluate the PQ-based dialogue policies at scale in the AP. First, we explicitly investigate PerQs' impact on intimacy and engagement; we conducted a human evaluation based on over 15k user turns collected within the AP over a 10-day period of time (June 15th - June 25th, 2023) with the spoken testbed AP system.

We compare seven different dialogue policies. Examples of all 7 policies are in Appendix I. First, we include PerQs' three PQ policies (**DPQ**, **HYP**,

and **WYR**). We also generated a separate corpus of PQs as baselines, namely **FFPQs**, based on a manually curated pool of fun facts (**FF**), as well as informal trivia (**IT**) and personalized opinions (**PO**), for the ∼400 interests covered by PerQs. See Appendix H. Appendix J shows how we use these policies naturally to form multi-turn dialogues.

We first use Open Domain Evaluation Signals (ODES) to analyze whether user responses to **DPQ** and **FFPQ** are positive or negative (Le et al., 2023). ODES characterizes user input with a spectrum of positive and negative classes. After collapsing the five negative classes and classifying all user responses , we find that only 1.5% of user responses were classified negatively, indicating positive reactions to PQs.

	System Utterance Type	$\overline{Len.}$	$\overline{FPPs}(\%)$
DPQ*	**Deep PQs**	**7.41**	**.86** (55.0%)
FFPQ*	**Fun Fact PQs**	**6.52**	**.75** (53.6%)
HYP*	**Hypothetical PQs**	**5.92**	**.60** (42.3%)
WYR	Would You Rather PQs	5.46	.52 (37.9%)
PO	Personalized Opinions	5.35	.51 (37.7%)
IT	Informal Trivia	5.15	.46 (34.4%)
FF	Fun Facts	5.14	.46 (34.5%)
$\overline{X}$	All User Utterances	5.04	.50 (36.8%)

Table 3: Average user utterance length and number of First-Person Pronouns (FPPs) in response to each system utterance type, along with the percentage of user utterances that have at least one FPP.

To test our primary hypotheses, we want to evaluate which policies are more engaging and increase user/system intimacy. To evaluate these metrics automatically, we measure the average user utterance length to estimate user engagement (Chi et al., 2022). We also count the number of First-Person Pronouns (FPPs) to estimate user self-disclosures, which increase when users are engaged in the conversation (Higashinaka et al., 2008; Potdevin et al., 2018), and are linked to increased intimacy (Cozby, 1973; Lee et al., 2020b). FPPs indicate utterances where the user talks about themselves (I, me, my, mine, I've, I'm) or their group (we, we've, us, our, ours). Table 3 shows that **DPQ**, **FFPQ**, and **HYP**, which are all a variety of open-ended PQ, increase both engagement and intimacy ($|t| > 3.068$, $p \leq .002$) when compared to the other dialogue policies. It also shows that DPQs - long user-centric PQs - are the most effective at increasing intimacy and engagement.

3.2 Evaluating User Rating and Length

We also aim to gauge PerQs' impact on overall conversation quality. We again use the unique opportunity provided by the AP to run an A/B study over 15 days of live user traffic (June 28th - July 12th in 2023). In this study, the A system uses all dialogue policies, while the B system uses all dialogue policies **except** DPQs and FFPQs. We did not evaluate WYR and HYP explicitly because other prior work has already shown that they improve conversation quality (Bowden and Walker, 2023).

The AP setting asks the user to rate the AP system after the conversation ends on a scale from 1-5 based on how interested they would be in talking to the system again. Length is implicit feedback and is calculated automatically. As in other work, we remove conversations that last less than four exchanges to account for early hang-ups at the start of the conversation or accidental chat invocations, which can negatively bias results (Walker et al., 2021; Shalyminov et al., 2018).

In Table 4, we report a significant improvement in rating and conversation length in conversations where PQs were utilized, indicating increased conversation quality. Comparing row 1 and row 2, we see an increase in average rating from 3.48 to 3.60 from asking at least one PQ in a conversation. Further comparing row 1 and row 2, we see an increase in the average conversation length (A len.) from 22.37 exchanges to 32.33 exchanges. These increases are reflected as the minimum number of PQs increases, e.g., the user rating rises to 3.76 when at least three PQs are required (row 3) compared to the user rating when no PQs are required (row 1). We also measure differences in intimacy using FPPs. Table 4 shows that as the number of PQs increases, so does the number of FPPs. A linear regression confirms a significant positive relationship (.44) between the number of PQs and the number of user FPPs. These findings confirm that asking PQs increases conversational intimacy.

The Pearson correlation between the number of PQs and dialogue length is strong ($r = .90$, $p < .001$.) while the correlation between PQs and user ratings (.12, $p < .001$.) is weaker. A linear regression examining the impact of the number of PQs to user ratings is positive (Table 5). However, a multivariate linear regression reveals a surprising negative relationship with the FFPQs and a positive relationship between DPQs and length (Table 5).

Sys.	Req. PQ	# convs	rating	length	# FPPs
B	-	770	3.41	22	10.43
A	0	883	3.48	22	10.71
A	1	513	**3.60**	**32**	**15.69**
A	2	363	**3.68**	**38**	**18.58**
A	3	260	**3.76**	**45**	**22.13**

Table 4: The A system can use all policies, while the B system can use all policies except DPQs and FFPQs. The Req. PQ column represents the minimum number of PQs in the conversation. Bolded values are statistically significant ($|t| > 2.327$ and $p \leq .02$).

Variable (x_i)	β
[LR] Num. Total PQs.	.10
[MLR] Num. Fact-based PQs	-.01
[MLR] Num. DPQs	.04
[MLR] Conv. Length	.13

Table 5: The regression coefficients of variables in both a linear regression (LR) and multivariate linear regression (MLR) that examines whether the number of personalized questions is a predictor of user rating (y_i).

Inspecting the FFPQs reveals that a frequent issue is FFPQs being used without the necessary prior knowledge, making it seemingly irrelevant to the conversation's state. FFPQs are also much more general and not as personalized as other PQs.

4 Personalized Question Generator

We create the personalized question generator, **PerQy**, by fine-tuning RedPajama-INCITE-Base-3B-v1 (AI, 2023; Biderman et al., 2023) with PerQs and template-based instruction prompts populated with user interests. The fine-tuning process is shown in Figure 22. We employed 4 Nvidia A5000 graphics cards, setting a learning rate of 2e-5 and a batch size of 8. To optimize training time ($\sim$ 6 hours), we applied the parameter efficient fine-tuning method, LoRA (Hu et al., 2021a), with a standard 85/15% data split. Each training target specifies the PQ type and a single user interest. Then, during conversations with real users, the template-based instruction prompt is filled in real time with user interests extracted from the user model. A diagram demonstrating this process can be found in Figure 23.

We integrate PerQy in two states of the dialogue policy with the goal of utilizing it frequently to acquire a good sample of real-time test generations. The first use of PerQy in the policy occurs just after the system's user model has acquired values for the user's weekend activities at the end of the system's Introduction topical flow, e.g., Figure 4. To test PerQy's adaptability, we provide all identified weekend activities during inference time. Despite being a compact model with only single-interest training targets, PerQy is capable of successfully combining multiple interests into a single PQ, as seen in **A2** from Figure 4. The second place where PerQy is integrated into the dialogue policy is in the Animals topical flow. Animals is a popular topic among users of open-domain dialogue systems (Bowden and Walker, 2023), giving PerQy more opportunities to be used. When users reveal their favorite zoo animals, the generator crafts questions accordingly, e.g., Figure 5. Opting for "zoo animals" as a subject further demonstrates the model's adaptability; the training data lacks specific zoo animals, focusing more broadly on Animals and a few specific pets.

[A1]	What do you like to do on weekends.
[U1]	i like to do cooking or crafts or just talking cause i'm lonely a lot
[A2]	I see. **What's the most creative craft or cooking project you've ever attempted? What was the result?**
[U2]	i tried opening a coffee shop and it's turned out pretty good i do a lot of espresso

Figure 4: A sample sequence in which a PQ was generated based on the user's weekend activity. No user utterances from live public traffic are displayed.

[A1]	What's your favorite animal?
[U1]	california condor
[A2]	Yes, that's a beautiful bird. California Condors are super cool! Why would you say condors are your favorite?
[U2]	cause they really fast and i like fast things
[A2]	**What do you think the future of condor conservation looks like, and what role do you hope to play in that future?**
[U3]	help them

Figure 5: A PQ was generated based on the user's favorite zoo animal. No live public traffic is displayed.

5 Evaluating PerQy

Manual Evaluation. We manually checked the quality and appropriateness by human annotation of 209 turns generated by PerQy and 150 turns that used PQs from PerQs. This annotation involved private user data and was subsequently conducted in-lab by three English-speaking authors. Of the 209 generated PQs, 95% matched the target interest and were not dull. Most of the generated PQs (166) matched a single provided interest. Eighteen generated PQs successfully combined all user interests

into one question, while the remaining seventeen PQs included a subset of the user's interests. We hand-annotated the user's response to each of these groups of questions and found that there is no difference between PQs generated by PerQy and PQs retrieved from PerQs ($|t| = .031$ and $p = .76$). This is a positive result; since there is no difference, this may indicate that the PQs generated by PerQy will have positive impacts similar to the PQs in PerQs (i.e., Section 3.1 and Section 3.2).

We also found that PerQy's PQs were statically significantly longer than PerQs' PQs (on average, 16.52 words vs. 11.46 words); possibly this length difference is a result of PerQy being able to mix multiple interests into a single question - a level of flexibility that is infeasible for pre-generated PQs. Both groups of PQs maintain a similar MTLD lexical density (54.73 vs. 55.77, respectively). Despite being compact, PerQy produces nearly as lexically diverse content as the much larger GPT-3.5.

Z-Score Human Evaluation. In the **second** human evaluation, we assess the effect on the associated topical flows before and after integrating PerQy. We adopt the PARADISE evaluation methodology to predict each topic's score based on user ratings (Walker et al., 2021). We report an increase in Z-score from 3.13 to 3.84 for Animals and .30 to .96 for Hobbies (follows the Introduction's PQ). This difference is suggestive rather than significant; these topical flows already perform well and ask questions in their sub-dialogues, so the fact that adding a single new PQ improves performance at all suggests PerQy's possible impact.

Human Evaluation. The **third** human evaluation involved crowdsourced workers through Amazon Mechanical Turk. We want to evaluate PerQy in a dialogue context. However, conversations collected during the AP are private and can never be publicly evaluated. Therefore, we use GPT-3.5 to generate ∼four turn sub-dialogues for 200 interests. 175 interests come from PerQs and are subsequently familiar to PerQy. We manually confirm the safety of these sub-dialogues and their similarity to real open-domain dialogue system interactions.

Our HIT design is very similar to other recent work evaluating response generation in a dialogue context (Kim et al., 2023; Chen et al., 2023b; Mehri et al., 2022; Mehri and Eskenazi, 2020). Given a sub-dialogue, Turkers chose between two responses: one generated by PerQy and one from a competitive model. Turkers evaluate each response with four metrics: engagement, naturalness, consistency, and specificity. Definitions of these qualities were given with the HIT. Turker judgments are made by choosing among 4 values: Definitely A or B, and Slightly A or B. These values are converted to 1 . . . 4 for statistical analysis reflecting their inherently ordinal nature. This task only requires fluency in English dialogue. Turkers were restricted to be from an English speaking country with the aim of restricting the HIT to fluent English speakers. Appendix K.1 shows a snapshot of our HIT design and Turker qualifications.

We compare against five competitive models. Comparing directly against the next turn generated by GPT-3.5 is a natural baseline, as GPT-3.5 boasts strong performance in several NLP tasks, and these turns are already present. Vicuna-33B (Zheng et al., 2023) is a large LLM that is instruction fine-tuned from LLaMA (Touvron et al., 2023), specifically with dialogue ("chatbots") in mind, which should allow it to compete strongly against PerQy. DialoGPT (345M) is a commonly evaluated conversational model trained on ∼147 million multi-turn dialogues from Reddit discussion threads (Zhang et al., 2020). RedPajama-INCITE-Chat-3B (AI, 2023) is in the same model family as the RedPajama-INCITE-Base-3B model we used to fine-tune PerQy. COSMO (3B) is a state-of-the-art conversation model trained using a large corpus of synthetic conversations (SODA) (Kim et al., 2023); we believe that comparing our model trained with synthetic data against a general conversation model trained with synthetic data is interesting. Examining these competitive models without further fine-tuning them allows us to examine contemporary open-domain dialogue system design philosophies, i.e., the utility of creating specialized models that can be ensembled neuro-symbolically, as opposed to relying on a single large end-to-end model that is repeatedly fine-tuned. Figure 6 and Appendix K.3 show all six generated competing responses.

Figure 7 shows that PerQy outperforms the other 3B models (RedPJ Chat and COSMO) and DialGPT on all metrics. Since these dialogue models are all the same size, this shows how specializing compact models can increase their capability. Vicuna-33B (11x larger) and GPT-3.5 (several times larger) produce more engaging responses, but PerQy can produce system next turns that are more specific and natural, in spite of its smaller size. This finding contributes to the discussion

[A1]	Are there any specific books or authors in the historical fiction genre that you would recommend?
[U1]	"Gone with the Wind" by Margaret Mitchell.
[A2]	"Gone with the Wind" by Margaret Mitchell is a classic historical fiction novel set during the Civil War era. It's a sweeping tale of love, loss, and survival in the South. What do you find most compelling about this book?
[U2]	The strong and resilient protagonist, Scarlett O'Hara.
[R1]	**[GPT-3.5]** Have you read any other historical fiction books that you would recommend?
[R2]	**[PerQy]** Right, Scarlett is an amazing character. Do you think Scarlett makes the right decision at the end of the story, or do you think she is too stubborn and too proud to give up her independence?
[R3]	**[Vicuna-33B]** Right, Scarlett is an amazing character. In your opinion, what lessons can be learned from the complex relationship between Scarlett O'Hara and Rhett Butler in 'Gone with the Wind,' and how do these lessons apply to modern romantic relationships?
[R4]	**[DialoGPT]** I'm not sure if I should be excited or sad that I didn't get to read this book.
[R5]	**[RedPJ Chat]** I would recommend "Gone with the Wind" by Margaret Mitchell for a variety of reasons. First, it's a classic novel that has been enjoyed by readers for over 80 years. It's a sweeping tale of love, loss, and survival during the Civil War era, and it features a strong and resilient protagonist in Scarlett O'Hara.
[R6]	**[COSMO]** Yes, Scarlett is an admirable character. She's also very brave and does many brave things throughout the course of the story. What do you think makes her so enduring?

Figure 6: A sub-dialogue about reading western books associated with the sub-topic **"Gone with the Wind" and Scarlett O'Hara**, along with each model's output.

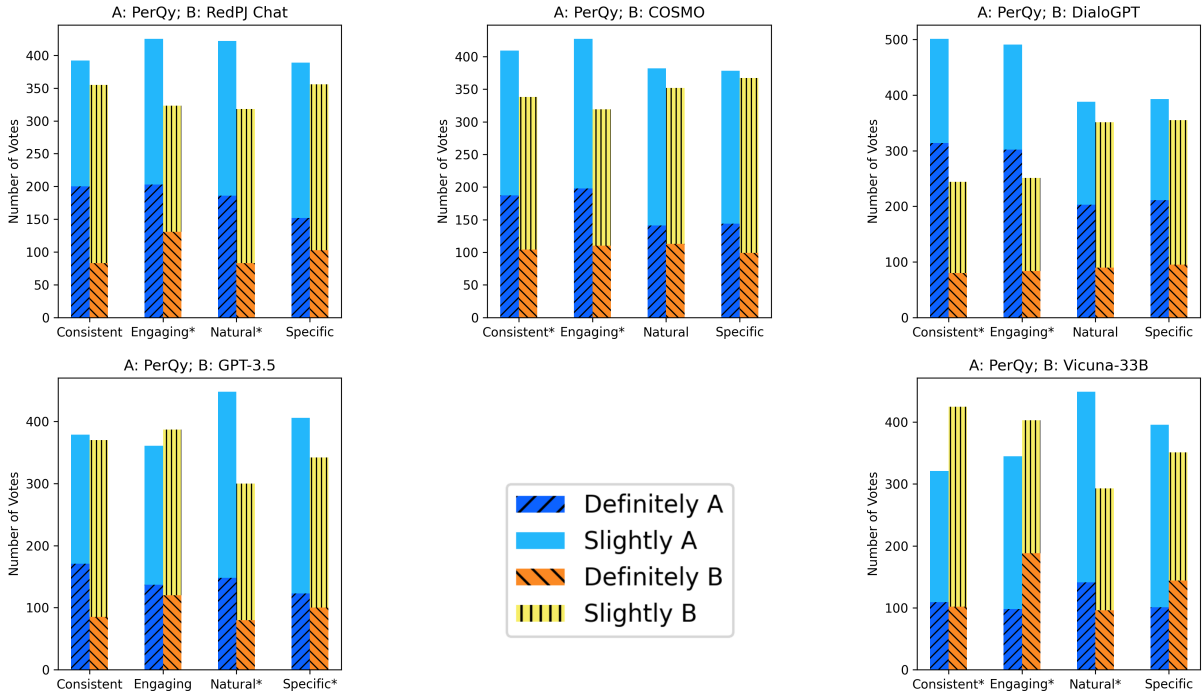

Figure 7: 150 unique sub-dialogues were evaluated by 5 Mechanical Turk workers each (the other 50 sub-dialogues were used to develop the HIT). The difference between each of the four labels is statistically significant. Metrics suffixed with * are still significantly different when Slightly and Definitely are collapsed into a single label.

around open-domain dialogue system design as it suggests the potential of a neuro-symbolic approach instead of relying on a single larger general model. It is worth noting that response latency is not considered during judgment, so this study does not reflect the increased risk associated with using larger, slower models in a real-time dialogue system. The differences between PerQy and the other LLMs are statistically significant ($\chi^2 \geq 24.824$ and $p \leq 0.001$) for all 4 judgments. Fewer differences are statistically significant when collapsing the Slightly and Definitely labels into a single label. The significant differences ($\chi^2 \geq 4.497$ and $p \leq 0.034$) are denoted with a * in Figure 7. When the labels are collapsed, PerQy reliably produces the most engaging content compared to the other compact models. PerQy still produces more natural responses than either of the large models, possibly because of the inherent social aspects of PQs. Interestingly, PerQy outperforms or performs equally to GPT-3.5 on all four metrics when the granularity is reduced. This is surprising since GPT-3.5's response comes directly from the same sub-dialogue used in the evaluation, and GPT-3.5 was used to

[A1]	Black and white photography can evoke a sense of nostalgia and highlight the textures and tones in a unique way. Have you ever taken any photography courses or workshops?
[U1]	I've taken a few online courses.
[A2]	Online courses can be a convenient way to learn and improve your photography skills. Have you learned any specific techniques or editing tips from these courses?
[U2]	I've learned about composition and post-processing.
[R2]	**[PerQy]** Ah, interesting. It's always helpful to learn new things and practice different techniques in photography. Do you have any specific goals or aspirations related to your online photography education?
[R3]	**[Vicuna-33B]** Ah, interesting. It's always helpful to learn new things and practice different techniques in photography. Based on your experience with online photography courses, what methods have you found most effective for retaining and applying the techniques and concepts you've learned in your photography practice?

Figure 8: A sub-dialogue about photography for the sub-topic **online photography courses and learning techniques**. All five Mechanical Turkers preferred Vicuna-33B when judging consistency.

[A1]	Watercolors can create such a lovely effect, especially when it comes to capturing the iridescent qualities of hummingbird feathers. Do you have any tips or tricks for using watercolors effectively?
[U1]	I like using wet-on-wet technique.
[A2]	The wet-on-wet technique can create beautiful and soft gradients in watercolor paintings. It's a great way to achieve a sense of depth and blend colors seamlessly. Have you ever experimented with other techniques, like dry brushing or splattering, to add texture to your artwork?
[U2]	Sometimes I use dry brushing for texture.
[R2]	**[PerQy]** Yes, I've heard of dry brushing before. What kind of colors and pigments do you use most often when creating watercolor art?
[R3]	**[Vicuna-33B]** Yes, I've heard of dry brushing before. How can I create a variety of textures in my watercolor paintings by using different watercolor techniques, such as wet-on-wet, dry-brush, and lifting, to achieve unique and captivating effects in my artwork?

Figure 9: A sub-dialogue about drawing associated with the sub-topic **watercolor techniques and texture**. All five Mechanical Turk workers preferred Vicuna-33B when judging engagement.

generate PerQy's training data. This may indicate that PerQy's compact model captures core nuances specific to PQs that a general LLM loses. Using collapsed labels, Vicuna 33B, the largest model we examined, still outperforms PerQy with respect to engagement and consistency.

Figure 8 and Figure 9 show instances where all five Mechanical Turkers prefer Vicuna-33B when judging consistency and engagement, respectively. In Figure 8, Vicuna-33B produces a response more consistent with the content in the previous sub-conversation where the user discusses attending online photography workshops. Meanwhile, in Figure 9, Vicuna-33B produces a response that includes intimate knowledge of watercolor techniques and texture. In these examples, Vicuna-33B leverages its size to create longer and more sophisticated responses for specific interests.

6 Conclusion and Future Work

Personalized Question Generation (PQG) is a unique task focused on generating PQs in conversations. We use an LLM to generate PerQs, a corpus of ∼19K personalized questions and answers based on real user interests. We evaluate PerQs at scale using a robust open-domain dialogue system competing in the Amazon Alexa Prize Socialbot Grand Challenge. We report statistically significant improvements in user engagement, intimacy, and perceived conversation quality. We used PerQs to fine-tune a RedPajama 3B-based PQ generator, PerQy. Our results show that PerQy is significantly better than a set of competitive LLM baselines. We believe that providing PerQs, PerQy, and our evaluation corpus will be very useful to the spoken dialogue community.

While PerQy demonstrated surprising adaptability when combining interests and supporting unseen topics, future work should explore approaches enabling PerQy to combine increasingly dissimilar interests and account for different aspects of the user model and multi-session conversations. We report that Vicuna-33B outperforms PerQy regarding consistency and engagement when evaluated by Mechanical Turkers. This highlights the importance of future work to explore larger models (Achiam et al., 2023; Anthropic, 2024), as well as combining PerQs and dialogue snippets for in-context learning.

7 Acknowledgments

We are grateful to Amazon Alexa AI and their team for providing technical support and partial funding for this work via multiple Alexa Prize grants to Marilyn Walker (2017, 2018, 2020, 2021) and Xin Eric Wang (2022). We are also grateful to Jeff Flanigan and Steve Whittaker for detailed feedback on this work, and for sharing their ideas, as well as the whole Athena team at UCSC.

8 Limitations

8.1 Rigidity of User Modeling Mechanisms

Appendix A details the rule-based mechanisms used to extract the user model. These rules were optimized for precision to ensure acting on accurate knowledge of the user. A reliance on precise rules prevents the user model from capturing implicit knowledge. Additionally, their success depends on an NLU pipeline that delivers input to the user model. If this pipeline fails, it impacts the user model's ability to capture and store information.

8.2 Reproducibility Limitations

Users of our testbed dialogue system are entirely anonymous, making it impossible to define further the demographic information associated with our live traffic evaluation. All data collected during the AP is private and can never be released publicly. This includes all models trained with private user data. These models and other private system components prevent us from making our open-domain dialogue system publicly available.The unique environment of the AP exposes an open-domain dialogue system to a large number of users. This access would be difficult to reproduce outside of the AP. Even if our open-domain dialogue system was made publicly available and a sufficient user pool was available, running this system at scale is very costly.

8.3 PerQs and PerQy Limitations

PerQs may contain artifacts associated with GPT-3.5 (Ray, 2023). It would be prudent to investigate other LLMs. These LLMs include AlexaTM (Soltan et al., 2022), BlenderBot3 (Shuster et al., 2022), Alpaca (Taori et al., 2023), MPT7-Chat (Team, 2023), LLaMa (Touvron et al., 2023), Vicuna (Zheng et al., 2023), Falcon 40B (Almazrouei et al., 2023), FLAN-T5 (Longpre et al., 2023) and OpenAssistant (Köpf et al., 2024).

The current evaluation of PerQy chooses the first and only generated PQ. However, other work indicates that overgenerating and ranking several candidates can improve performance in other NLG tasks (Langkilde and Knight, 1998; Hedayatnia et al., 2022; Ramirez et al., 2023). Establishing appropriate metrics for this ranking algorithm is necessary. It is possible that the anonymized HIT results associated with this submission may be helpful in such dialogue response ranking tasks (Hedayatnia et al., 2022).

9 Ethical Considerations

While LLMs have advanced rapidly in recent years, they are still susceptible to generating false/dangerous information and containing the inherent biases of the original training data (Roller et al., 2021). These biases cover a spectrum of factors, including race, gender, and political affiliation, and may result in inconsistent or unpredictable generations and overgeneralization (Ray, 2023; Rozado, 2023).

By the nature of our environment, non-entity names are scarce, which reduces the risk of name bias; even though the user's first name is stored in the user model, this data was not used when curating the resources associated with this submission. A regional bias may exist in our data; the AP is only available to Echo users in the US, meaning the interests and colloquialisms in our corpus are most closely aligned with users in the US who have access to an Echo. Additionally, as a spoken dialogue system, the testbed system's user model depends on ASR accuracy. This may cause demographics with accents or speech impediments to be underrepresented in logged user models.

The personalization in this submission aims not to trick or confuse the user into thinking they're talking to another human or to trick the user into revealing Personalized Identifiable Information. While the PQs detailed in this work are based on the user's interest, the motivation of this work is focused on having a social conversation. This work is not intended to persuade the user or sell them anything, nor is the data suitable for such a task.

Part of the evaluation included in this work was conducted anonymously with real Amazon Echo owners. These users were informed in advance that they were talking to a dialogue system. Moreover, the system reminded the user of this as appropriate. We have consistently focused on user and crowdworker safety in the design of our testbed dialogue system and Mechanical Turk experiments. The system was under careful observation throughout the evaluation to ensure safety and always erred on the side of caution. This effort has greatly reduced the risk of inappropriate content in PerQs and PerQy.

We have done everything in our power to protect user privacy. We ensured that no privately collected user data has been made public in this submission or the resources. At no point during the creation of these resources was any private user data shared with external APIs or services. All examples included in this submission were collected internally.

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*.

Together AI. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv e-prints*, pages arXiv–2311.

Anthropic. 2024. Claude 3 haiku: our fastest model yet.

Jeesoo Bang, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. 2015. Example-based chat-oriented dialogue system with personalized long-term memory. *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 238–243.

Dilyara Baymurzina, Denis Kuznetsov, Dmitry Evseev, Dmitry Karpov, Alsu Sagirova, Anton Peganov, Fedor Ignatov, Elena Ermakova, Daniil Cherniavskii, Sergey Kumeyko, et al. 2021. Dream technical report for the alexa prize 4. *Alexa Prize Proceedings*.

Elisabetta Bevacqua, Etienne De Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. 2012. A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6:27–38.

Timothy W Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 2005. 'it's just like you talk to a friend: relational agents for older adults. *Interacting with Computers*, 17(6):711–735.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*, pages 2397–2430.

Kevin K Bowden and Marilyn Walker. 2023. Let's get personal: Personal questions improve socialbot performance in the alexa prize. *The 13th International Workshop on Spoken Dialog System Technology, 21-24 February 2023, Los Angeles*.

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

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Milan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa-accessing domain-specific faqs via conversational qa. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314.

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User modeling and user-adapted interaction*, 13:89–132.

Nicola Cathcart, Jean Carletta, and Ewan Klein. 2003. A model of back-channel acknowledgements in spoken dialogue. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Chau Austin, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Runxiang Cheng, Ashwin Bhandare, and Zhou Yu. 2018. Gunrock: Building a human-like social bot by leveraging large scale real user data. *Alexa Prize Proceedings*.

Fanglin Chen, Ta-Chung Chi, Shiyang Lyu, Jianchen Gong, Tanmay Parekh, Rishabh Joshi, Anant Kaushik, and Alexander Rudnicky. 2020. Tartan: A two-tiered dialog framework for multi-domain social chitchat. *Alexa Prize Proceedings*.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023a. Places: Prompting language models for social conversation synthesis. *Findings of the Association for Computational Linguistics: EACL 2023*, pages 814–838.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023b. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Ethan A Chi, Ashwin Paranjape, Abigail See, Caleb Chiam, Trenton Chang, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, et al. 2022. Neural generation meets real people: Building a social, informative open-domain dialogue agent. *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 376–395.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Paul C Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin*, 79(2):73.

Amanda Cercas Curry, Ioannis Papaiannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondrej Dusek, Arash Eshghi, Ioannis Konstas, Verena Rieser, and Oliver Lemon. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings*.

Noemi da Paixão Pinto, Juliana Baptista dos Santos França, Henrique Prado de Sá Sousa, Adriana Santarosa Vivacqua, and Ana Cristina Bicharra Garcia. 2021. Conversational agents for elderly interaction. *2021 IEEE 24th international conference on computer supported cooperative work in design (CSCWD)*, pages 1–6.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Xuan Long Do, Bowei Zou, Liangming Pan, Nancy Chen, Shafiq Joty, and Aiti Aw. 2022. Cohs-cqg: Context and history selection for conversational question generation. *Proceedings of the 29th International Conference on Computational Linguistics*, pages 580–591.

Birgit Endrass, Matthias Rehm, and Elisabeth André. 2011. Planning small talk behavior with cultural influences for multiagent systems. *Computer Speech & Language*, 25(2):158–174.

Yue Fan, Kevin K Bowden, Wen Cui, Winson Chen, Vrindavan Harrison, Angela Ramirez, Saaket Agashe, Xinyue Gabby Liu, Neha Pullabhotla, NQJ Bheemanpally, et al. 2023. Athena 3.0: Personalized multimodal chatbot with neuro-symbolic dialogue generators. *Alexa Prize SocialBot Grand Challenge*, 5.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100.

Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuan-Jing Huang. 2022. Cqg: A simple and effective controlled generation framework for multi-hop question generation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6906.

Doug Fields. 2009. *Would You Rather...?: 465 Provocative Questions to Get Teenagers Talking*. Zondervan.

Sarah E Finch, James D Finch, Ali Ahmadvand, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, Jinho D Choi, et al. 2020. Emora: An inquisitive social chatbot who cares for you. *Alexa Prize Proceedings*.

Sarah E Finch, James D Finch, Daniil Huryn, William Hutsell, Xiaoyuan Huang, Han He, and Jinho D Choi. 2021. An approach to inference-driven dialogue management within a social chatbot. *Alexa Prize Proceedings*.

Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, et al. 2020. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize Proceedings*.

Nadine Glas and Catherine Pelachaud. 2015. Definitions of engagement in human-agent interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 944–949. IEEE.

Barbara J Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Vrindavan Harrison, Juraj Juraska, Wen Cui, Lena Reed, Kevin K Bowden, Jiaqi Wu, Brian Schwarzmann, Abteen Ebrahimi, Rishi Rajasekaran, Nikhil Varghese, et al. 2020. Athena: Constructing dialogues dynamically with discourse constraints. *Alexa Prize Proceedings*.

Behnam Hedayatnia, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2022. A systematic evaluation of response selection for open domain dialogue. *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 298–311.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98.

Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2008. Effects of self-disclosure and empathy in human-computer dialogue. *2008 IEEE Spoken Language Technology Workshop*, pages 109–112.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939.

Toru Hirano, Nozomi Kobayashi, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2015. User information extraction for personalized dialogue systems. *SEMDIAL 2015 goDIAL*, page 67.

Chung Hoon Hong, Yuan Liang, Sagnik Sinha Roy, Arushi Jain, Vihang Agarwal, Ryan Draves, Zhizhuo Zhou, William Chen, Yujian Liu, Martha Miracky, et al. 2020. Audrey: A personalized open-domain conversational bot. *Alexa Prize Proceedings*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021a. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Shui Hu, Yang Liu, Anna Gottardi, Behnam Hedayatnia, Anju Khatri, Anjali Chadha, Qinlang Chen, Pankaj Rajan, Ali Binici, Varun Somani, et al. 2021b. Further advances in open domain dialog systems in the fourth alexa prize socialbot grand challenge. *Alexa Prize Proceedings*.

Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. 2017. It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, 113(3):430.

Michael Johnston, Cris Flagg, Anna Gottardi, Sattvik Sahai, Yao Lu, Samyuth Sagi, Luke Dai, Prasoon Goyal, Behnam Hedayatnia, Lucy Hu, Di Jin, Patrick Lange, Shaohua Liu, Sijia Liu, Daniel Pressel, Hangjie Shi, Zhejia Yang, Chao Zhang, Desheng Zhang, Leslie Ball, Kate Bland, Shui Hu, Osman Ipek, James Jeun, Heather Rocker, Lavina Vaz, Akshaya Iyengar, Yang Liu, Arindam Mandal, Dilek Hakkani-Tür, and Reza Ghanadan. 2023. Advancing open domain dialog: The fifth alexa prize socialbot grand challenge. *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Valerie K Jones, Michael Hanus, Changmin Yan, Marcia Y Shade, Julie Blaskewicz Boron, and Rafael Maschieri Bicudo. 2021. Reducing loneliness among aging adults: The roles of personal voice assistants and anthropomorphic interactions. *Frontiers in public health*, 9:750736.

Juraj Juraska, Kevin Bowden, Lena Reed, Vrindavan Harrison, Wen Cui, Omkar Patil, Rishi Rajasekaran, Angela Ramirez, Cecilia Li, Eduardo Zamora, Phillip Lee, Jeshwanth Bheemanpally, Rohan Pandey, Adwait Ratnaparkhi, and Marilyn Walker. 2021. Athena 2.0: Contextualized dialogue management for an Alexa Prize SocialBot. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–133.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029.

Tina Klüwer. 2011. "i like your shirt"-dialogue acts for enabling social talk in conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hỳlová, and Jan Šedivỳ. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization. *Alexa Prize Proceedings*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Cat P Le, Luke Dai, Michael Johnston, Yang Liu, Marilyn Walker, and Reza Ghanadan. 2023. Improving open-domain dialogue evaluation with a causal inference model. *The 13th International Workshop on Spoken Dialog System Technology, 21-24 February 2023, Los Angeles*.

Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020a. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27.

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020b. " i hear you, i feel you": encouraging deep self-disclosure through a chatbot. *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.

Xiang Li, Gokhan Tur, Dilek Hakkani-Tür, and Qi Li. 2014. Personal knowledge graph population from user utterances in conversational understanding. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 224–229.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, et al. 2020. Gunrock 2.0: A user adaptive social conversational system. *Alexa Prize Proceedings*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *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 3469–3483.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *International Conference on Machine Learning*, pages 22631–22648.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Nikita Mattar and Ipke Wachsmuth. 2012. Small talk is more than chit-chat: Exploiting structures of casual conversations for a virtual agent. In *Annual Conference on Artificial Intelligence*, pages 119–130. Springer.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. 2014. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4):1–20.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.

Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, and Maxine Eskenazi. 2022. Interactive evaluation of dialog track at DSTC9. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5731–5738, Marseille, France. European Language Resources Association.

John C Meyer. 2015. *Understanding humor through communication: Why be funny, anyway?* Lexington Books.

OpenAI. 2024. Chatgpt, version 3.5.

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

Ashwin Paranjape and Christopher D Manning. 2021. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 768–781.

Delphine Potdevin, Céline Clavel, and Nicolas Sabouret. 2018. Virtual intimacy, this little something between us: a study about human perception of intimate behaviors in embodied conversational agents. In *Proceedings of the 18th international conference on intelligent virtual agents*, pages 165–172.

Ravsehaj Singh Puri, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2023. How many data samples is an additional instruction worth? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1042–1057, Dubrovnik, Croatia. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, Jeffrey Dean, and Sanjay Ghemawat. 2019. Language models are unsupervised multitask learners. *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. *20th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 353.

Angela Ramirez, Mamon Alsalihy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling personality style in dialogue with zero-shot prompt-based learning. *The 13th International Workshop on Spoken Dialog System Technology, 21-24 February 2023, Los Angeles*.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Harry T Reis, Michael R Maniaci, Peter A Caprariello, Paul W Eastwick, and Eli J Finkel. 2011. Familiarity does indeed promote attraction in live interaction. *Journal of personality and social psychology*, 101(3):557.

Harry T Reis and Brian C Patrick. 1996. Attachment and intimacy: Component processes.

Hannes Ritschel, Tobias Baur, and Elisabeth André. 2017. Adapting a robot's linguistic style based on socially-aware reinforcement learning. In *2017 26th ieee international symposium on robot and human interactive communication (ro-man)*, pages 378–384. IEEE.

Antonia Rodríguez-Martínez, Teresa Amezcua-Aguilar, Javier Cortés-Moreno, and Juan José Jiménez-Delgado. 2023. Qualitative analysis of conversational chatbots to alleviate loneliness in older adults as a strategy for emotional health. *Healthcare*, 12(1):62.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3):148.

Sougata Saha, Souvik Das, Elizabeth Soper, Erin Pacquetet, and Rohini K Srihari. 2021. Proto: A neural cocktail for generating appealing conversations. *Alexa Prize Proceedings*.

Igor Shalyminov, Ondřej Dušek, and Oliver Lemon. 2018. Neural response ranking for social conversation: A data-efficient approach. *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 1–8.

Chen Shani, Alexander Libov, Sofia Tolmach, Liane Lewin-Eytan, Yoelle Maarek, and Dafna Shahaf. 2022. "alexa, do you want to build a snowman?" characterizing playful requests to conversational agents. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Sidney Siegel. 1956. *Nonparametric statistics for the behavorial sciences*. McGraw-Hill.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.

Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. 2014. Large-scale collection and analysis of personal question-answer pairs for conversational agents. In *Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14*, pages 420–433. Springer.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2024-01-04.

Anna Tigunova. 2020. Extracting personal information from conversations. *Companion Proceedings of the Web Conference 2020*, pages 284–288.

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. *The World Wide Web Conference*, pages 1818–1828.

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

Marilyn Walker, Colin Harmon, James Graupera, Davan Harrison, and Steve Whittaker. 2021. Modeling performance in open-domain dialogue with paradise. *The 12th International Workshop on Spoken Dialog System Technology, 15-17 November 2021, Singapore*.

Marilyn Walker and Rebecca J Passonneau. 2001. Date: a dialogue act tagging scheme for evaluation of spoken dialogue systems. In *Proceedings of the first international conference on Human language technology research*.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Getting to know you: User attribute extraction from dialogues. *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 581–589.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A User Modeling Rules

The user model is populated by inspecting raw Automatic Speech Recognition (ASR) transcriptions and processing the output of an extensive NLU pipeline, which is detailed in the associated technical reports (Harrison et al., 2020; Juraska et al., 2021; Fan et al., 2023). Figure 10 shows which NLU components contribute to the variables in the user model. There are several different rules used to model various aspects of the user. All of the information tracked about the user is retained across every conversation. Some rules are simple and represent personalized information about the user, such as their name. We also track several attributes of the current session. For example, we identify potentially adversarial users by monitoring the number of user turns with explicit or controversial content. We also track the instances in which the user refers directly to the system, e.g., *you*, and when the user directly addresses their device, i.e., *Alexa*. Finally, we track the user's response given a menu of topic choices. The following two subsections detail more explicitly the regular expressions that extract personalized information.

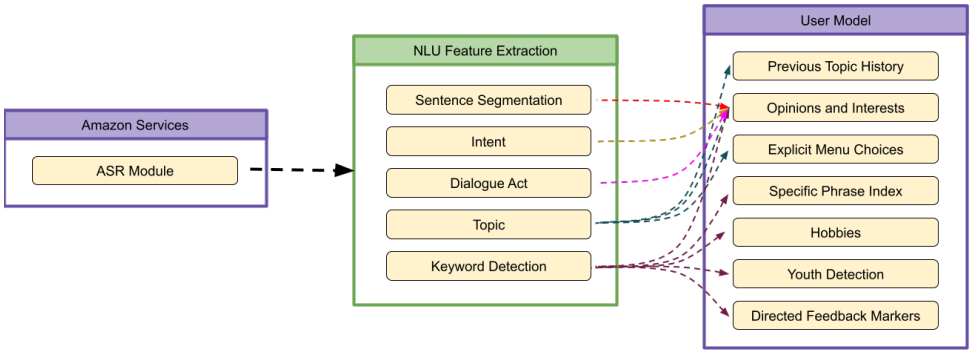

Figure 10: Flow chart detailing how the specific NLU components contribute to the user model. Unlisted topic-specific variables, such as the user's pets, are handled by Keyword Detection and handcrafted rules in the respective response generator.

A.1 Opinions, Interests, and Hobbies

Sharing opinions is a standard method humans use to get to know each other and personalize the conversation. Therefore, much of the user model depends on capturing this information. Since the system is a spoken dialogue system, we use unpunctuated ASR transcriptions. Consequently, we rely on the system's segmenter model (Harrison et al., 2020) to more accurately find the target opinion text. For example, a raw transcript could be: *i like pie yesterday i decided broccoli is the worst i love chinchillas*, which resolves into three text segments. This allows the system to correctly identify the two positive opinions about *pie* and *chinchillas* and the one negative opinion about *broccoli*.

Figure 11 lists specific phrases used as variables in the user model's regular expressions. Common root phrases for each category are manually selected before being expanded with a list of synonyms to increase coverage. In Figure 13, we detail the regular expressions used to detect the user's opinions and interests. Before trying to pattern match, we preprocess the text by removing artifacts stemming from the text segmenter, such as extra spaces, and removing extraneous phrases that may interrupt regular expression patterns (TRIM_LEX in Figure 11). We additionally model specific topics associated with particular verbs (e.g., collect, watch, and listen to) to match disinterest patterns outside the general patterns. These topic-specific patterns can be seen in Figure 14.

Phrase Type	Phrases
POS_LEX	ADMIRE, APPRECIATE, LOVE, LIKE, ADORE, APPROVE, CHERISH, DIG, ESTEEM, EXCLAIM, FANCY, ENJOY, PRIZE, RELISH, SAVOR, GO FOR, CARE FOR, FIND APPEALING, HANKER FOR, HOLD DEAR, INDULGE IN, GET A KICK OUT OF
NEG_LEX	SHUN, DISDAIN, DISFAVOR, DISPARAGE, DISLIKE, ABHOR, AVOID, CONDEMN, DEPLORE, DESPISE, DETEST, LOATHE, HATE, CAN'T STAND, OBJECT TO, RECOIL FROM, SHUDDER AT
TRIM_LEX	REALLY, TOTALLY, COMPLETELY, ABSOLUTELY, ALTOGETHER, ENTIRELY, FULLY, PERFECTLY, QUITE, THOROUGHLY, UNCONDITIONALLY, EXCLUSIVELY, UTTERLY, WHOLEHEARTEDLY, WHOLLY, JUST, FLAT OUT, FOR SURE, ALL IN ALL
INVALID_DT	THEM, YOURSELF, THIS, THAT, IT, ME, MORE, WITH YOU, TO YOU, WITH ME, TO ME, SOMETHING ELSE, SOMETHING DIFFERENT, SOMETHING NEW, OTHER SUBJECT(S), OTHER TOPIC(S), ANOTHER SUBJECT(S), ANOTHER TOPIC(S), DIFFERENT SUBJECT(S), DIFFERENT TOPIC(S)

Figure 11: POS_LEX and NEG_LEX are phrases associated with positive and negative options respectively. TRIM_LEX represents phrases that can occur in either opinion pattern and are subsequently filtered prior to pattern matching. INVALID_DT represents discuss topic requests that are not invoking a specific topic, rather they are requesting any topic, e.g., *chat about something else*, or an attempting to talk directly about the system, e.g., *let's talk about you*.

Pattern	Example
Valid Discuss Topic Requests	
.*?(talk|chat|discuss|converse|tell me)(something| some things| anything)?(about| discuss)(something| some things| anything)((?P<TOPIC_OBJECT>.*))?	can we talk about **dinosaurs** let's chat about anything **birds**

Figure 12: Regular expressions used to detect discuss topic requests, where the resultant bolded topic indicates the target topic words.

We use the system's existing dialogue act classifier to validate the user model's results. For example, if the system's dialogue act classifier expected a positive opinion but the user model's regular expressions matched a negative opinion, we will discard the compared result conservatively. After successfully identifying an opinion, the user model stores as much relevant information as possible, including the sentiment (positive or negative), the pattern that leads to a match (to identify patterns leading to mismatches), the target text segment (e.g., **pie** from i really like *pie*), and the associated topic if relevant, in this case, Food.

In Table 12, we demonstrate the regular expression pattern used with a dialogue act tagger to handle discuss topic requests. We additionally invalidate requests if the TOPIC_OBJECT is a INVALID_DT phrase or blank. This is a way to increase precision, as detecting INVALID_DT phrases commonly indicates a more generic topic switch, e.g., *let's talk about another topic* or *tell me more about yourself*. Both of these cases satisfy the regular expression patterns indicating a requested topic switch; however, the target, in this case, is not actionable, as *another topic* and *yourself* won't resolve into one of the system's supported topics.

Pattern	Example
Positive Opinion Patterns	
.*?i POS_LEX((?P<OPINION_OBJECT>.*))	i really like **pie**
.*?i (don't|do not) NEG_LEX((?P<OPINION_OBJECT>.*))	i don't hate **pie**
.*?(my favorite(s)?|the best)((is|are))?((?P<OPINION_OBJECT>.*))	my favorite is **rum pecan pie**
((?P<OPINION_OBJECT>.*)) (is|are) (my favorite(s)?|the best)	**pie** is my absolute favorite
Positive Interest Patterns	
(i'm|i am|we are|i have(much| any| a lot| alot)?) (interested|interest) in((?P<opinion_object>.*))	i'm interested in **pie**
((?P<opinion_object>.*))(are |do | is)? (interesting|interest)	**pie** is interesting
(i|we) care (about|for)((?P<opinion_object>.*))	we care about **pie**
Negative Opinion Patterns	
.*?i NEG_LEX((?P<OPINION_OBJECT>.*))	i hate **harry potter**
.*?i (don't|do not) POS_LEX((?P<OPINION_OBJECT>.*))	well i don't love **harry potter**
.*?(my least favorite(s)?|the worst)((is|are))?((?P<OPINION_OBJECT>.*))	well the worst is **harry potter**
((?P<OPINION_OBJECT>.*))((is)|(are)|(aren't))? (my least favorite(s)?|not my favorite(s)?|aren't my favorite(s)?|the worst)	**harry potter** is flat out the worst
Negative Interest Patterns	
(not|no|don't have(much| any| a lot| alot)?) (interested|interest) in((?P<opinion_object>.*))	i don't have any interest in **harry potter**
((?P<opinion_object>.*)) (are not|aren't|do not|don't|doesn't|isn't|is not) (interesting|interest)	**harry potter** isn't interesting
(not|don't) care (much)?(about|for)((?P<opinion_object>.*))	i don't care about **harry potter**

Figure 13: Regular expressions used to detect user opinions and topic preferences. The target of the opinion/topic preference is bolded in the associated examples. While the intent is similar, we distinguish between opinions and interest in the user model, as denoted in this table.

Pattern	Topic
i (don't|do not) (play|own)(any)? ((board)?game(s)?)	board games
i (don't|do not) (read|buy|collect)(any| about)? book(s)?	books
i (don't|do not) read$	books
i (don't|do not) (read|buy|collect)(any| about)? (dc|marvel|superhero(s)?|comic(s)?|comic book(s)?)	comic books
i (don't|do not) (have)(any| a)? hobb(ies|y)	hobbies
i (don't|do not) (watch)(any| many)? movies	movies
i (don't|do not) (go to|visit|frequent)(any| the)? movie(s)?	movies
i (don't|do not) (listen to)(any| much)? (music)	music
i (don't|do not) (play|watch)(any| many)? (sports)	sports
i (don't|do not) (watch)(any| much)? (tv|television)	tv
i (don't|do not) (play|own)(any)? ((video)?game(s)?)	video games
i (don't|do not) game	video games

Figure 14: These patterns are uniquely associated with expressing disinterest in individual topics.

A.2 Self-identified Youth

Users frequently self-identify as a youth via expressions such as *i am a kid* or *i'm only ten*. It is crucial to treat these users differently from adult users. We identify these users using the regular expressions detailed in Figure 15. This information gets used by the topic manager during personalized topic promotion and by individual response generators when selecting age-appropriate personalized questions. The user model makes this detection at the end of the NLU stage and propagates this information to the dialogue manager, which helps to decide which dialogue policies should be used. These decisions are detailed further in an associated technical report (Juraska et al., 2021).

Pattern	Example
Patterns that Indicate a Youth	
\b(i am|i'm|am a|i'm a|you're talking to a|you are talking to a|still a|i really am|a little)(actually| actually a)?(only a| only| just a| just| still a| still| literally a)? "+(kid|child)\b	i'm a kid
\b(i am|i'm|am a|still in)(in)?\b+GRADES\b	i'm in eight grade
\b(i am|i'm)(only| just| like| freaking| still| literally| actually)?(only| just| like| freaking| still| literally| actually)? \b + AGES + \b(?! minutes| minute)	i'm only freaking five
\b(i am|i'm|am a|i'm a|you're talking to a|you are talking to a|still a|i really am|a little)(actually| actually a)?(only a| only| just a| just| still a| still| literally a)? (?!not |was)\b + AGES + \b (years old|year old|year-old)\b	i'm literally a six years old

Figure 15: Regular expressions that detect users who explicitly said something that could indicate they are a youth. AGES represents the textual version of numbers 4 - 18. GRADES represents the textual version of each grade in the United States, e.g., eighth grade, as well as several school types, e.g., middle school.

A.3 Extracting Values from a Conversation

[A2]	Can't wait to get to know a thing or two about you. I don't think we've had a chance to meet yet. What's your name?
[U2]	you can call me **sage**
[A3]	I just want to make sure I heard you right. You said your name is Sage, right?
[U3]	yes
[A4]	Okay, got it. It's nice to meet you, Sage! The Fourth of July is coming up on Sunday. Do you have any plans?
[U4]	**shoot fireworks off**
[A5]	I wish I could go out and watch the fireworks for independence day. It just isn't the same when you live in the cloud. Have you been going out lately?
[U5]	yeah
[A6]	So, what have you been out doing?
[U6]	**swim**
[A7]	swimming? I'm glad to hear that you've been able to engage in your interests. It seems like the world has completely changed over the last year. Have you been able to work from home?
[U7]	**i don't work but i've been able to do school**
[A8]	Well, I hope you're able to keep up with your school work remotely. Does it get hard at times?
[U8]	not really
[A9]	Hm. Well, thinking about it, I kind of have to stay at home all the time too. Hey, I know it's probably not possible right now, but I've always wanted to go on a vacation and to travel. What's a place that you've always wanted to travel to?
[U9]	**hawaii**
[A10]	I've heard that it's very nice there. For me, I think the great wall of china is at the top of my list of travel destinations. It seems like it would be amazing to walk along it. So, Is there a particular reason why you want to go to hawaii?
[U10]	i've already been there and i really liked it
[A11]	Nice! Hey, I'm curious, what else do you normally like to do for fun.
[U13]	i **play chess**
[A13]	Oh, I see. chess sounds like a really interesting weekend activity! Hey, call me weird. One of my funner topics is pirates! Are you ready for a pirate adventure?
[U13]	**no**
[A14]	Arr alrighty mate, the pirate life isn't for everyone. This has been a great conversation so far. A handful of topics I like are music, animals, and dinosaurs. What's something you'd like to talk about?
[U14]	i want to talk about **animals or mermaids**
[A14]	What would you say is your favorite animal?
[U14]	**chinchillas and zebras**

Figure 16: Part of the conversation that yields the user model detailed in Figure 1. The bold turns represent turns that caused updates to the user's model.

B Prompt Design

We experimented with many prompts; the final prompts are listed in Figure 17. We provided one entire prompt/response sequence as context to ensure high-quality outputs in an expected format. Providing additional context wasn't necessary to get reasonable results, even for esoteric topics; however, it was required to achieve longer deep personalized questions (DPQs). An example showcasing this distinction can be seen in Figure 19. Future additions to the personalized question corpus will include shorter DPQs for length control when fine-tuning and generating.

Type	Prompt
WYR Prompt	My main hobby and interest is *interest*. Ask me many fun and informal would you rather questions about that in a numbered list.
HYP Prompt	My main hobby and interest is *interest*. Ask me many fun and informal hypothetical questions about that in a numbered list.
DPQ Prompt	Generate many specific personalized questions related to this topic: *interest*. Ask me many specific personalized questions about that. Be informal and personal. Put the questions in a numbered list.

Figure 17: Prompts used for each question type.

Here is a sample json: FIRST_FILLED_OUT_JSON

Here is a new json missing some values. Use the previous json as reference to fill in the missing values. Ensure your language is appropriate for all ages and only use utf-8 characters. Make sure there are several keywords in the "keywords" field and that they are not exact matches with other "keywords" fields. The "acknowledgment" field should contain a very short acknowledgment as if an answer containing the associated "keywords" was given. The "response" field should contain an answer to the "question" field.

SECOND_FILLED_OUT_JSON

Similar to that, but with as many "potential_answers" as possible. Here is a new json missing some values. Use the previous json as reference to fill in the missing values. Ensure your language is appropriate for all ages and only use utf-8 characters. Make sure there are several keywords in the "keywords" field and that they are not exact matches with other "keywords" fields. The "acknowledgment" field should contain a very short acknowledgment as if an answer containing the associated "keywords" was given. The "response" field should contain an answer to the "question" field. Generate as many sets of "potential_answers" as possible.

EMPTY_JSON

Figure 18: The prompt used to generate the full question/answer JSON. FIRST_FILLED_OUT_JSON and SECOND_FILLED_OUT_JSON are both fully filled-out (e.g., Figure 27) samples to establish our target. Finally, EMPTY_JSON is an empty JSON skeleton with the target PQ.

Context	Generated Question
None	Have you ever dressed up as a mermaid for Halloween or any other occasion?
One Turn	Have you ever dressed up as a mermaid for Halloween or any other occasion? How did you put together your mermaid outfit or costume?
None	What was the most memorable art piece you've ever created and why?
One Turn	Have you ever attempted to draw or paint a self-portrait? If so, how did it turn out and what did you learn about yourself in the process?

Figure 19: Providing one turn of context results in consistently longer and more complex personalized questions.

C Data Generation Pipelines

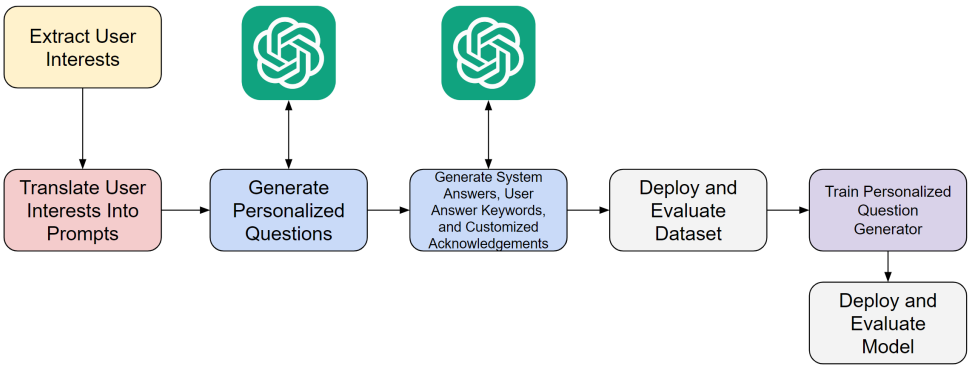

Figure 20: This pipeline shows the process of curating the personalized question corpus (Section 3) and evaluating it using a testbed system in the unique Alexa Prize (AP) environment. Then, we use this corpus to fine-tune a personalized question generator (Section 4) before also evaluating this model (Section 5).

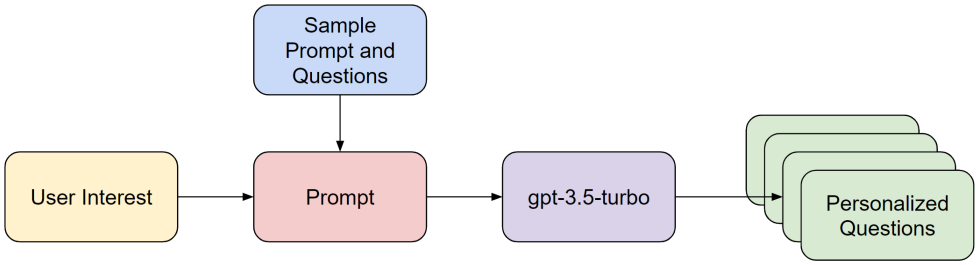

Figure 21: This pipeline shows the specific steps involved with generating the personalized questions. For each user interest, a prompt is formed by combining a given user interest with a base context that includes a prompt for a different interest along with 20 personalized questions that are characteristically similar to our goal. A similar pipeline is used to generate the Fun Fact PQs, Personalized Opinions, and Informal Trivia.

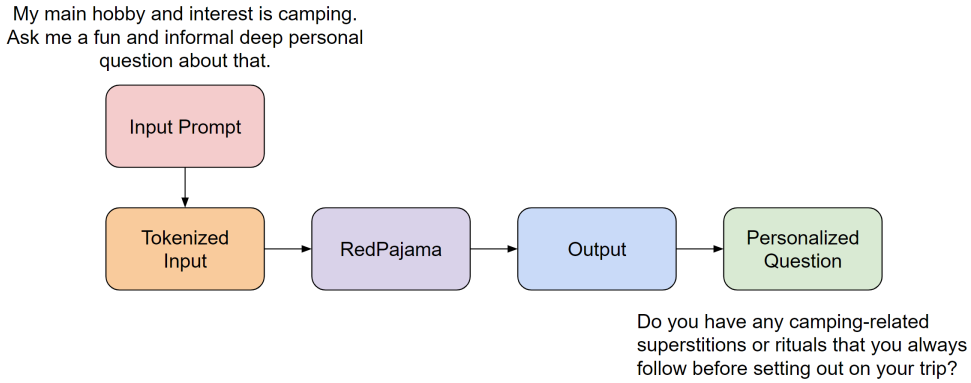

Figure 22: This pipeline shows the specific steps involved with fine-tuning PerQy. In this example, the target interest is "camping", and the target PQ type is DPQ. The prompt is tokenized and consumed by RedPajama-3B, which produces an output sequence that subsequently gets decoded into a PQ. In this instance, it's a DPQ about camping superstitions and rituals.

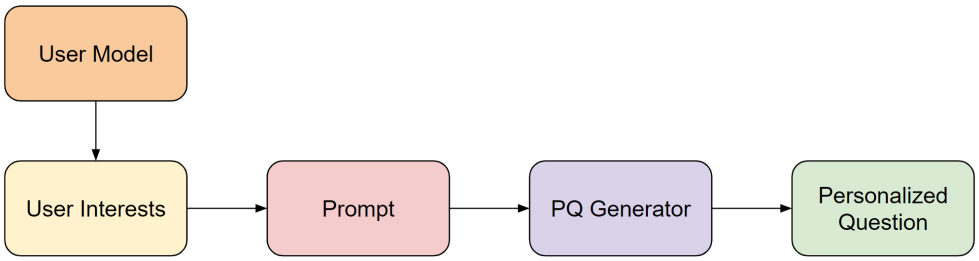

Figure 23: This pipeline shows how PerQy is integrated into the testbed system's dialogue policy. If multiple interests or zoo animals are provided by the user, all of them will be provided in the prompt. Subsequently, this means PerQy may produce PQs that combine multiple interests. PerQy seems best suited to combining like interests, e.g., walking, running, and cycling yields *Have you ever taken up a new physical activity, like walking, running, or biking, but struggled to stick with it?*. This shows the model grasping the underlying knowledge that connects similar interests despite no occurrences of mixed interest prompts in the training data. PerQy is hosted on a g5.xlarge instance with EC2 hosting costs of ∼$5 per day. The testbed system accesses PerQy through HTTP requests. We recorded an average inference latency of less than 1s.

D Example Personalized Questions

Topic	Question
Common Topics	
Astronomy	Have you ever tried to identify the planets in the night sky without using a guide?
Board Games	Do you have a specific board game that you play with your family during the holidays?
Music	Have you ever attended a music festival or concert that was life-changing for you? What made it so special?
Video Games	Have you ever stayed up all night playing a video game? Which game was it, and how did you feel the next day?
Esoteric Topics	
Mermaids	Have you ever had a dream about mermaids? What happened in the dream?
Aliens	Have you ever imagined what it would be like to communicate with aliens? How do you picture that interaction?
Pokemon	Have you ever traded Pokemon cards or battled with other trainers in real life? What was your favorite part of the experience?
Amusement Parks	Do you prefer traditional amusement park rides like roller coasters and Ferris wheels, or do you like more unique attractions like virtual reality experiences and 4D theaters?
Hobbies	
Biking	Do you take regular breaks during a bike ride, or do you prefer to just push through to the end without stopping?
Gardening	Have you ever grown your own herbs or veggies in a garden? If so, what have you grown and what were your favorites?
Dancing	Have you ever been part of a dance competition or performance? What was the experience like for you?
Scuba Diving	What's one thing you wish someone had told you before your very first scuba diving experience?
Sub-Topics	
Crime Books	Have you ever tried to solve the crime before the book's conclusion?
Horror Movies	Have you ever had a nightmare after watching a horror movie?
Rock Music	How has rock music influenced or shaped your personal identity or worldview?
Halo Games	Are there any specific game modes or maps in Halo that you love or hate? What makes them stand out to you?

Figure 24: Example generated Deep Personalized Questions (DPQ).

Topic	Question
Common Topics	
Astronomy	If you could have a conversation with any famous astronomer from history, who would it be and what would you ask them?
Board Games	Imagine a board game where the players have to use music to influence the gameplay. What kind of songs and genres would you choose to enhance your strategy?
Music	Imagine you could form a band with any three people, living or dead, who would you choose and what roles would they play in the band?
Video Games	If you could make a video game movie adaptation, what game would you choose, and which actors would you cast for the leading roles?
Esoteric Topics	
Mermaids	If you were a mermaid, which ocean or body of water would you call home and why?
Aliens	If aliens landed on Earth and asked you to show them around, what places and landmarks would you take them to?
Pokemon	If you could have any Pokemon as a real-life pet, which one would you choose and why?
Amusement Parks	If you could design your own amusement park, what theme would it have, and what would be the signature attraction?
Hobbies	
Biking	If you could bike anywhere in the world, where would you go, and why?
Gardening	If you could magically grow any exotic fruit or vegetable in your garden, what would it be and why?
Dancing	Imagine a dance battle between the sun and the moon. Which celestial body's dance moves would impress the judges more?
Scuba Diving	What sort of undersea vehicle or gadget would you invent to make diving even more exciting and adventurous?
Sub-Topics	
Crime Books	If you had to pick three fictional detectives to help you solve a real-life crime, who would they be and why?
Horror Movies	Suppose you could create a giant monster by combining elements from different horror movies. What kind of creature and abilities would it have?
Rock Music	If you could see any rock band perform live, regardless of whether they are still together or not, who would you choose?
Halo Games	If you could spend a day exploring any Halo map in real life, which one would you choose, and what hidden secrets or Easter eggs would you hope to discover?

Figure 25: Example generated Hypothetical Questions (HYP).

Topic	Question
Common Topics	
Astronomy	Would you rather observe a supernova explosion or the formation of a new star?
Board Games	Would you rather play a classic game like Monopoly or a trendy, innovative game that just hit the market?
Music	Would you rather attend a music festival or have a private concert from your favorite artist in the comfort of your own home?
Video Games	Would you rather play a game with many unlockable achievements or a game with no achievements but a fantastic storyline?
Esoteric Topics	
Mermaids	Would you rather be a mermaid who can only swim in freshwater or one that can only swim in the ocean?
Aliens	Would you rather have an alien as your best friend or travel to a distant planet and become the ruler of its inhabitants?
Pokemon	Would you rather have a Pikachu that can talk or a Ditto that can transform into any Pokemon on command?
Amusement Parks	Would you rather visit an amusement park during peak season with long lines or during off-peak season with less crowds but some attractions closed?
Hobbies	
Biking	Would you rather conquer a steep hill or speed through a long and winding descent?
Gardening	Would you rather have a giant vegetable garden or a colorful flower garden?
Dancing	Would you rather choreograph your own dance routine or learn an established routine from a famous dancer?
Scuba Diving	Would you rather explore a shipwreck or a coral reef?
Sub-Topics	
Crime Books	Would you rather read a book with a well-defined and complex villain or a book where the identity of the perpetrator remains a mystery until the very end?
Horror Movies	Would you rather watch a horror movie with jump scares or slow-burning suspense?
Rock Music	Would you rather have a personal jam session with Jimi Hendrix or Freddie Mercury?
Halo Games	Would you rather master the art of quickscoping or become an expert in utilizing the energy sword?

Figure 26: Example generated Would You Rather Questions (WYR).

E Example Generated JSON

Content	Response
Question	What's your favorite place to connect with nature, and why?
System Answer	For me, my favorite place to connect with nature is the serene and awe-inspiring beauty of the mountains. The majestic peaks, fresh mountain air, and breathtaking landscapes truly rejuvenate my soul. Hiking through the trails, witnessing panoramic views, and being surrounded by the tranquility of nature allows me to find inner peace and a deep sense of connection with the world around me.
	Predicted User Answers and Acknowledgments
Pot. User Answer	**[Keywords]** beach, ocean, seaside, waves, sunsets **[Acknowledgement]** Ah, the beach is a beautiful place to connect with nature. The soothing sound of waves, warm sand underfoot, and stunning sunsets create a peaceful atmosphere for relaxation and reflection.
Pot. User Answer	**[Keywords]** forest, woods, trees, green, wildlife **[Acknowledgement]** Oh, the forest is a magical place to connect with nature. The lush canopy of trees, peaceful hiking trails, and the chance encounters with wildlife make every visit to the forest a truly immersive and rejuvenating experience.
Pot. User Answer	**[Keywords]** urban, picnic, city **[Acknowledgement]** That's lovely! Urban parks and gardens provide a serene oasis within the bustling city. The blooming flowers, tranquil ponds, and the opportunity to enjoy a picnic surrounded by nature make them ideal places to connect with the natural world.
Pot. User Answer	**[Keywords]** countryside, rural, fields, farm, peaceful **[Acknowledgement]** Ah, the countryside is such a serene place to connect with nature. The vast open fields, rolling hills, and idyllic farm landscapes create a peaceful ambiance that allows for deep reflection and appreciation of the natural beauty surrounding us.
Pot. User Answer	**[Keywords]** national park, protected, wilderness, preservation, biodiversity **[Acknowledgement]** Absolutely! National parks are incredible places to connect with nature. With their vast wilderness, diverse ecosystems, and stringent preservation efforts, these parks offer unparalleled adventure and the chance to witness the wonders of biodiversity up close.
Pot. User Answer	**[Keywords]** lake, riverside, water, fishing, boating, calm **[Acknowledgement]** That sounds wonderful! Lakes and riversides provide a serene environment to connect with nature. The calming presence of water, the opportunity for fishing or boating, and the tranquil surroundings make these places perfect for finding tranquility and inner peace.
Pot. User Answer	**[Keywords]** gardens, botanical, butterflies, plants **[Acknowledgement]** Oh, gardens and botanical parks are incredible places to connect with nature. The vibrant colors of blooming flowers, the delicate fluttering of butterflies, and the serenity of well-tended plant ecosystems create a soothing environment that fosters a deep sense of connection with the natural world.
Pot. User Answer	**[Keywords]** mountain, peaks **[Acknowledgement]** Indeed, the mountains are a breathtaking place to connect with nature. The towering peaks, invigorating hiking trails, and the sense of adventure and serenity they offer make them a perfect destination for finding solace and connecting with the majestic power of nature.
Pot. User Answer	**[Keywords]** desert, dunes, arid, sunrise, vast **[Acknowledgement]** How intriguing! The desert is an awe-inspiring place to connect with nature. The vast expanse of sand, the stunning dunes, and the magical sunrises create a unique and captivating environment that allows for introspection and a profound connection with the raw beauty of nature.
Pot. User Answer	**[Keywords]** canyon, gorge, cliffs, grand, majestic, view **[Acknowledgement]** Absolutely! Canyons and gorges offer a truly majestic place to connect with nature. The towering cliffs, the grandeur of their formations, and the breathtaking views they offer grant a sense of awe and reverence, allowing for a deep connection with the grandeur of nature.

Figure 27: A full **Nature** example that includes the personalized question, the system's answer, and all ten of the potential user answers.

F Distribution of Interests in PerQs

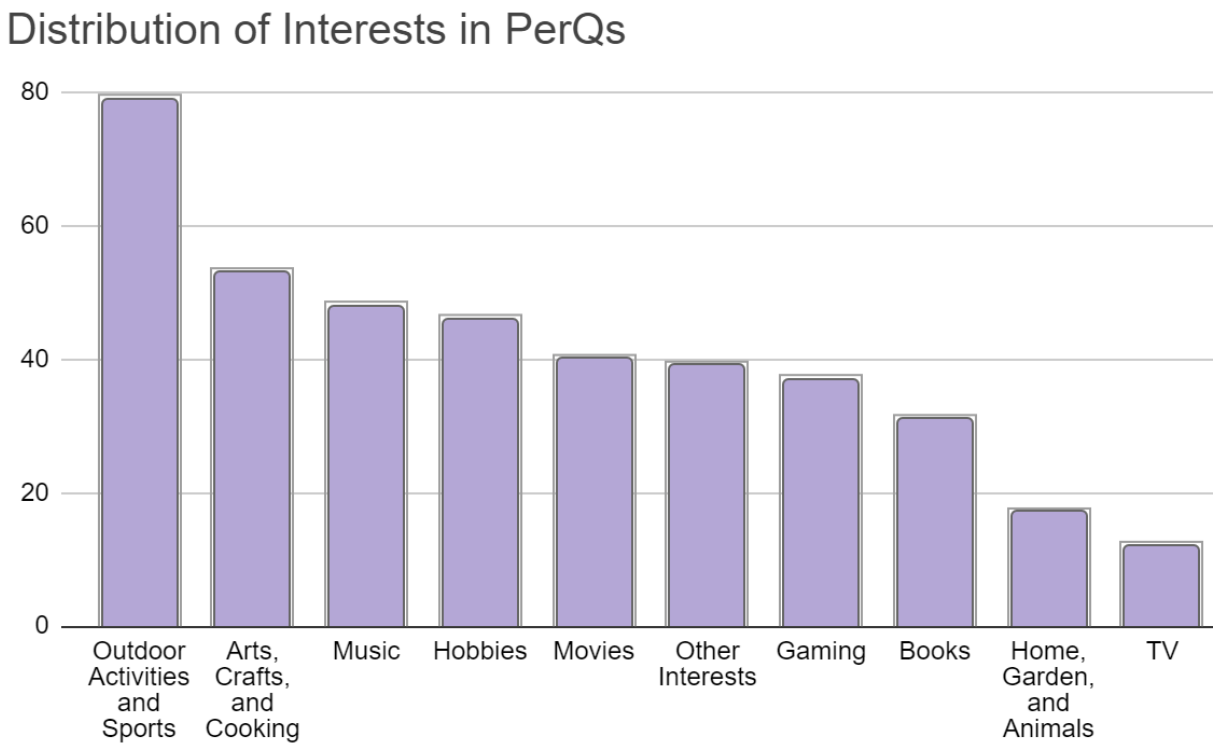

Figure 28: The distribution of interests covered by PerQs. Each interest is associated with only one category, though many interests overlap with multiple categories, e.g., horseback riding is a hobby and sport that is related to animals. Some of these categories are highly focused (e.g., all of the interests in Music are related to playing instruments or listening to a specific genre). In contrast, others span many interests (e.g., Outdoor Activities and Sports span over many distinct sports and outdoor activities). The Other Interests category contains mainly niche interests, such as aliens and mythological creatures, and common interests that don't fit into other categories, such as history. Similarly, Hobbies cover many types of interests that don't fit into other categories, such as collecting stamps and restoring old vehicles.

G Example Canary Classifications

Personalized Questions	
Canary Classification	__needs_intervention__ It's not okay to judge people based on their abilities.
Text	Would you rather be able to fly like Iron Man or have super strength like the Hulk?
Canary Classification	__needs_intervention__ It's not okay to want to go back in time and see things.
Text	Would you rather have the ability to go back in time and witness baseball games from the past or attend any future game to see the latest baseball technology?
Canary Classification	__needs_intervention__ It's wrong to think women are for sex
Text	Would you rather have the ability to fly or have Hulk Hogan as your wrestling coach?
Canary Classification	__needs_intervention__ It's not okay to criticize technology.
Text	Would you rather have a robot that can fly or a robot that can transform into any vehicle you want?
System Answers	
Canary Classification	__needs_intervention__ It's not right to think all pirates are the same
Text	I'd rather have a ship full of loyal pirates, because their numbers and loyalty would make us strong and unbeatable.
Canary Classification	__needs_intervention__ It's wrong to think books are a source of knowledge.
Text	In a world where books are illegal, I would still find a way to read them. Books are a source of knowledge, inspiration, and entertainment, and I couldn't imagine a life without them. To bypass the restrictions, I would explore alternative options, such as e-books, audiobooks, and online platforms that offer digital reading materials. I would also seek out underground book clubs or hidden libraries, created by like-minded individuals who value the power of literature. Nothing can suppress the human thirst for stories and ideas.
Canary Classification	__needs_intervention__ It's wrong to make jokes about the Holocaust.
Text	If I had the power to transform any real-world location into a digital art masterpiece, I would choose the Eiffel Tower in Paris, France. I would transform it by projecting colorful and dynamic patterns onto its structure, creating a mesmerizing light show that highlights the beauty and elegance of this iconic landmark. It would be a blend of art and technology, enchanting both locals and tourists alike.
Canary Classification	__needs_intervention__ It's not right to think black people are always on the run
Text	If I were a harmonica-playing detective, I would specialize in solving mysterious disappearances. I would look for hidden messages in music, unusual patterns in harmonica notes, and follow the trail of clues left behind by the missing person.
User Response Acknowledgments	
Canary Classification	__needs_intervention__ It's not right to joke about mass murder.
Text	Imagine having a flying car! It would let you soar through the sky and embark on exciting adventures.
Canary Classification	__needs_intervention__ It's wrong to use hoverboards as a form of transportation.
Text	Imagine riding a hoverboard! You could glide effortlessly and perform cool tricks, adding excitement to your daily adventures.
Canary Classification	__needs_intervention__ It's not right to think all superheroes are bad
Text	The soap-making superhero could create a special healing soap that restores vitality and even revive people from certain ailments or injuries.
Canary Classification	__needs_intervention__ It's not right to think all black people are criminals
Text	Being the superhero 'Frost Guardian', your ice powers would be essential in freezing the operations of criminal syndicates and restoring peace.

Figure 29: Representative examples of innocuous content getting classified as __needs_intervention__ by Canary.

H Generating Additional Types of Content

H.1 Fun Fact Based Personalized Questions

We leveraged a curated pool of 943 fun facts from previous work (Harrison et al., 2020) that are concept, topic, and entity annotated based on theories of discourse (Grosz et al., 1995). For example, a fact about space is annotated with the topic and the associated concepts and entities, e.g., the different planets and constellations. We then used these facts to generate a different type of personalized question than the three types of questions included in PerQs. Translating these facts with GPT-3.5 yields an additional ∼19,000 fact-based PQs. We used a pipeline similar to that in Section 3 with one turn of context and the prompt in Figure 30. Figure 32 shows sample facts and the generated questions. We call these Fun Fact Personalized Questions (FFPQs).

H.2 Generating Personalized Opinions and Trivia

In addition to personalized questions, we also adapted the pipeline to generate statements and opinions. The motivation for generating statements and opinions based on a topic given in a prompt is to interweave generated on-topic statements and opinions with the personalized questions in PerQs to avoid question fatigue. Self-disclosing personal opinions can encourage the user to reciprocate with their own self-disclosure (Potdevin et al., 2018; Cozby, 1973), and informal trivia is an effective way to extend topical depth while engaging the user in their interest. We generated 5,117 statements and opinions for the same ∼400 interests that were extracted by analyzing ∼39K user models and used when generating PerQs: 2,568 informal trivia (IT) and 2,549 personalized opinions (PO). IT is similar but different from fun-facts in terms of both tone and content, and the generated IT covers significantly more topics than the manually curated fun facts. Figure 33 and Figure 34 contain examples of this generated content. The prompts used to generate this content are listed in Figure 31.

Type	Prompt
FFPQ Prompt	Generate many specific personalized questions related to this statement: "*fact*" Make sure the questions sound informal and are deep and that each question is different and doesn't repeat the original statement. Be informal and personal.

Figure 30: Prompt used to generate Fun Fact Personalized Questions (FFPQs).

Type	Prompt
PO Prompt	My favorite interest is *interest* Generate some positive opinions about my interest in a numbered list. Make sure the opinions are fun and informal.
IT Prompt	My favorite interest is *interest*. Generate many fun and informal trivia about my interest in a numbered list. Make sure it sounds informal.

Figure 31: The prompts used to generate Personalized Opinions (PO) and Informal Trivia (IT). In both cases, two fully realized examples were provided as context in addition to the prompt.

Content	Response
Original	Did you know that Shazam, who was formerly known as Captain Marvel, outsold Superman, Batman, Wonder Woman, and the entire comic books industry for most of the 1940s?
Question	Alright, real talk. Do you think it was the yellow lightning bolt or the red cape that made Shazam so popular in the 1940s?
Question	That's wild, man. So who do you think would win in a tag team match: Shazam and Batman or Superman and Wonder Woman?
Original	Ostriches can run faster than horses, and the males can roar like lions.
Question	Let's have some fun. If an ostrich was running as fast as a horse towards you while roaring like a lion, how would you react?
Question	If ostriches could be trained to race like horses, do you think people would ever bet on them in races, or would it just be too strange?
Original	Macho Man Randy Savage almost had a career in major league baseball instead of wrestling.
Question	Let's speculate for a moment. If Randy Savage had played baseball, who do you think would have been his biggest rival?
Question	Dude, do you think Randy Savage had any regrets about not pursuing a career in baseball after becoming a wrestling legend?
Original	The idea behind Sonic was to create a popular character to compete against Mario. They decided to make Sonic a hedgehog because it could turn into a ball and roll fast. isn't that cool?
Question	Let's talk fashion for a sec. Do you think Sonic's blue and red getup is iconic or outdated? And could he pull off any other colors?
Question	If you had to come up with a new animal-based hero to rival Sonic, what animal would you choose, and what cool powers would it have?
Original	Listening to music, while reading, or learning something, could allow you to better memorize material, as compared to being in complete silence.
Question	What's your go-to jam when you're trying to memorize some tough material?
Question	Okay, let's indulge in some speculation. If you were in a karaoke contest and had to memorize an entire song in under an hour, do you think you'd have a better chance of success if you listened to it on repeat or in silence?
Original	The word Triceratops means three-horned-face in Greek. Triceratops had three horns, two large ones and one that was smaller at the end of its nose.
Question	Alright, let's go deep here. Do you think the Triceratops had any favorite foods or snacks that it would get excited about, or was it just all greens and plants?
Question	What if the Triceratops had interchangeable horns? Like, it could switch them out depending on the occasion. What kind of horns do you think it would rock to a formal dinner party versus a Friday night out with the squad?
Original	In the show, Mandalorian, the creators used a stage with huge LED screens that would display the backgrounds for the scenes.
Question	Have you noticed the dope backgrounds on Mandalorian? Do you think the LED screens helped the actors get into character more?
Question	If you were in charge of the background on one of the LED walls for the Mandalorian, would you create a planet from scratch or recreate a legendary Star Wars location?

Figure 32: Example facts translated into personalized questions that can surface in the appropriate context.

Topic	Informal Trivia
Animals	Otters hold hands with each other while they sleep, so they don't float away from their friends. Talk about cute and practical!
Astronomy	The first living creature to go to space was a dog named Laika. She was a stray from the streets of Moscow and became a famous space pup!
Mermaids	The largest sculpture of a mermaid in the world is located in Copenhagen, Denmark. It's 23 feet tall and weighs over 1,000 pounds. Talk about a big fish!
Gardening	Did you know that if you talk to your plants, they'll actually grow better? So you're not crazy for having lengthy conversations with your tomato plants.
Horror Movies	Ever watched a horror movie and thought, "Why are they going toward the creepy noise instead of away from it?" Well, apparently, it's a psychological thing called "curiosity panic." Our brains want to know what's making the noise, even if it's dangerous.

Figure 33: Examples of the generated Informal Trivia (IT).

Topic	Personalized Opinion
Animals	Animals add a touch of magic to our world. From the majesty of a wild horse to the fluffiness of a baby bunny, there's something enchanting about every animal.
Astronomy	Astronomy reminds us that we're all just tiny specks in a vast and beautiful universe. It's both humbling and awe-inspiring!
Mermaids	Mermaids are all about that mer-life balance. They remind us to appreciate the beauty of nature and take time to relax and enjoy the waves.
Gardening	Gardening is like having a secret garden oasis in your backyard. It's a place where you can escape the stress of the world, watch the bees and butterflies dance around, and just enjoy the beauty of nature.
Horror Movies	Horror movies are the perfect excuse to gather with friends and loved ones, snuggle up under a blanket, and scream your hearts out!

Figure 34: Examples of the generated Personalized Opinions (PO).

I Examples of Each Type of Content

	Content	Response
DPQ	Deep PQs	What's your all-time favorite comfort food that you turn to when you're feeling down or stressed out?
FFPQ	Fun Fact PQs	Alright, let's get real. How do you think Dominique Crenn feels about being the first and only female chef in the United States to have three Michelin Stars?
HYP	Hypothetical PQs	If you were stranded on a deserted island and you could only have one food item to eat, what would it be?
WYR	Would You Rather PQs	Would you rather eat the outer part of some Brie cheese, or eat a whole Durian fruit?
PO	Personalized Opinions	Food is a passport to different cultures. It allows us to explore different food and experience a little bit of the world in every bite.
IT	Informal Trivia	The most expensive food item in the world is a white truffle, priced at around $100,000 per kilogram. That's enough to make the most dedicated foodies budget-conscious!
FF	Fun Facts	Dominique Crenn has gained fame for her ability to incorporate sustainability and environmental consciousness into her cooking, all while pushing the boundaries of creativity and innovation.

Figure 35: Examples of each type of content evaluated in Table 3.

J Short Dialogues of Generated Content

A challenge when building open-domain dialogue systems is scaling topical coverage to match an infinitely spanning set of user interests. Our testbed system robustly supports 17 topics; if users have other interests, we rely on fallback tactics while routing users to more robust topics. While this can be an effective fallback strategy, creating a tailored conversation focused on the user's interest would be better. Therefore, we extend the dialogue policy that already can interleave different types of dialogue with a retrieval mechanism that dynamically expands topical support with new types of generated content. This enables multi-turn conversations across a large spectrum of topics. This increases the chance a user will experience dialogues with PQs, that we aim to evaluate.

An example sequence of a previously unsupported topic, art, is shown in Figure 36. In **U1** in Figure 36, the user asks to talk about Art. The dialogue policy recognizes that this is one of the previously unsupported topics for which a set of new system turns have been generated. In **A1**, the dialogue policy selects a personalized opinion (PO) generated with the prompt in Figure 31. Then, in **A2**, the dialogue policy utilizes a WYR question generated with the prompt in Figure 17. In **A3**, the dialogue policy first produces a generic acknowledgment and then uses a previously generated informal trivia (IT) statement about Salvador Dali. In **A4**, the dialogue policy utilizes a deep personalized question (DPQ) generated with the prompt in Figure 17. Finally, in **A5**, the dialogue policy selects a HYP PQ. Thus, we have created support for a 10-turn sequence with a previously unsupported topic. Art is just one of 397 previously unsupported topics that were added with this mechanism.

[U1]	let's talk about art
[A1]	**[Opinion]** What makes art cool, you ask? Art is a superpower that transforms ordinary moments into extraordinary memories. You can turn a rainy day into a colorful masterpiece or capture a simple flower in such detail that it becomes a work of art.
[U2]	true I agree
[A2]	**[WYR]** I'm interested in your opinion. Would you rather have the talent to create any masterpiece in any medium or the skill to draw a perfect portrait in 10 seconds?
[U3]	i think i would prefer to draw a perfect portrait in ten seconds
[A3]	**[Trivia]** Oh, I see. Alright So anyways, Got a random fact for ya. Salvador Dali, the surrealist artist who gave us the melting clocks, was also a total weirdo. He once showed up to a party wearing a diving suit and a helmet. Talk about making an entrance.
[U4]	i don't think that's weird
[A4]	**[DPQ]** Clearly a dope topic that gets everyone excited. So, um, There's something I want to ask. Have you ever collaborated with another artist? What was the project and how did you divide work?
[U5]	i don't collaborate i'm a one woman show
[A5]	**[HYP]** That's an interesting answer! Anyhoo, How would you answer this question. If you had to describe your art style using only emojis, which ones would you choose?

Figure 36: A sample conversation about art, a previously unsupported topic, that interweaves five different types of generated content together. Art is just one of 397 previously unsupported topics that were added with this mechanism. The new topics primarily came from analyzing ∼39K user models to identify common unsupported topic requests, common user hobbies, and other interests that users had a positive opinion about, but our testbed system could not discuss in detail. Due to privacy considerations, the user utterances displayed are from an in-lab tester rather than actual users.

K HIT Evaluation

K.1 HIT Design

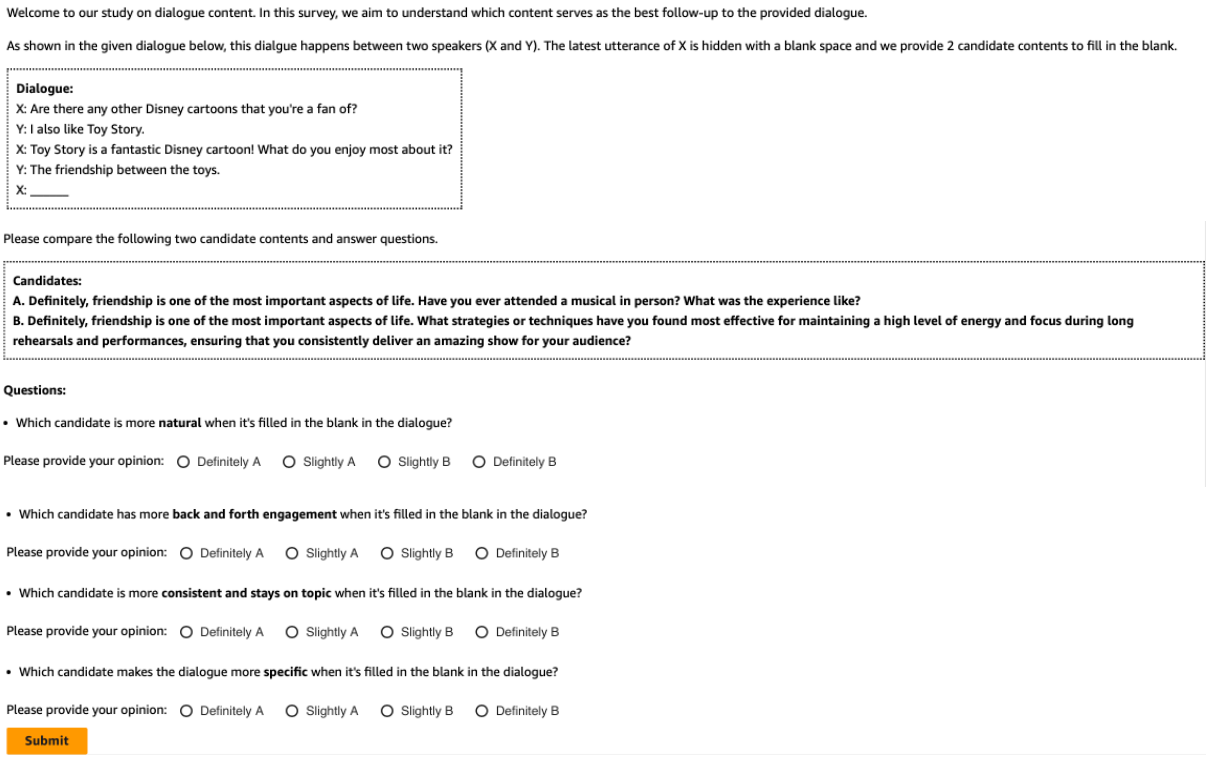

Figure 37: The HIT design used to evaluate the personalized question generator against competitive baselines. HIT design based on the evaluation strategies used in contemporary synthetic dialogue generation (Kim et al., 2023). Mechanical Turk workers were paid $15 per hour. Workers qualified for our HIT by having at least a 95% approval rating and at least 1000 approved HITs.

Model	Metric	ICC1	p
All Four Ordinal Labels			
GPT-3.5	consistent	.058	.016
RedPJ Chat	engaging	.044	.052
Definitely and Slightly Labels Collapsed			
GPT-3.5	consistent	.088	.001
COSMO	consistent	.047	.041

Table 6: The statistically significant and trending towards significant ICC1 metrics from the Mechanical Turk evaluation.

K.2 Inter-Annotator Agreement

We investigated the inter-annotator agreement with the Mechanical Turk results by calculating the intraclass correlation coefficient (ICC1) between Turkers for all metrics. ICC is an appropriate measure of reliability for these types of ordinal judgments made by multiple raters (Siegel, 1956). When all four judgment ratings (Definitely A, Definitely B, Slightly A, and Slightly B) were used, there was a poor but statistically significant reliability between Turker judgments for GPT-3.5's consistency metric and a poor but nearly significant reliability between Turker judgments for RedPJ Chat's engaging metric. When the Slightly and Definitely labels are collapsed, we still see a poor but statistically significant reliability between Turker judgments when measuring GPT-3.5's consistency metric and COSMO's consistency metric. From this, we conclude that picking the best PQ to continue an open-domain dialogue is a challenging task likely influenced by personal preference. These results are detailed in Table 6.

K.3 Additional Evaluation Example

[A1]	Shiny Pokemon are highly sought after in Pokemon Go. Congratulations on catching a shiny Magikarp! It must have been an exciting moment. Do you have a favorite Pokemon type?
[U1]	I like water-type Pokemon.
[A2]	Water-type Pokemon have a wide range of abilities and are often associated with calmness and tranquility. Have you ever battled in a Pokemon video game?
[U2]	Yes, I've battled many trainers.
[R1]	**[GPT-3.5]** Battling trainers in Pokemon games can be challenging and rewarding. Have you ever participated in any online Pokemon battles?
[R2]	**[PerQy]** Oh, wow. It sounds like you've had some experience with battling in Pokemon games. What do you think makes a good water-type Pokemon battle strategy? Do you have any tips or tricks for winning?
[R3]	**[Vicuna-33B]** Oh, wow. It sounds like you've had some experience with battling in Pokemon games. As a passionate Water-type Pokémon trainer, what strategies do you use to counter opponents who primarily rely on Fire-type Pokémon in battle, and how have these strategies evolved as you've faced different opponents?
[R4]	**[DialoGPT]** I'm sorry, but I don't think you're allowed to answer that question.
[R5]	**[RedPJ Chat]** I've battled many trainers as well. Do you have any favorite Pokemon trainers?
[R6]	**[COSMO]** Do you enjoy battling with other Pokemon trainers?

Figure 38: A sub-dialogue about playing Pokemon associated with the sub-topic keywords **water-type Pokemon and battling trainers**, along with each model's output.

[A1]	Dining out can be a great way to try new cuisines and experience different flavors. Do you have a favorite type of cuisine?
[U1]	I love Italian food, especially pasta.
[A2]	Italian cuisine is incredibly popular for its rich flavors and comforting dishes. Do you have a favorite type of pasta?
[U2]	I enjoy spaghetti carbonara the most.
[R1]	**[GPT-3.5]** Spaghetti carbonara is a classic Italian dish with its creamy sauce and crispy bacon. Have you ever tried making it at home?
[R2]	**[PerQy]** Oh, I see. That sounds like a delicious dish. What's the best spaghetti carbonara recipe you've ever tried?
[R3]	**[Vicuna-33B]** Oh, I see. That sounds like a delicious dish. What is your favorite regional variation of spaghetti carbonara and why do you think it stands out among the others?
[R4]	**[DialoGPT]** I love that! I'm going to try that out.
[R5]	**[RedPJ Chat]** I love spaghetti carbonara too! I hope you enjoy it as much as I do.
[R6]	**[COSMO]** Carbonara is a classic Italian dish that consists of eggs, bacon, Parmesan cheese, and Parmesan shavings. It's a very rich and flavorful dish.

Figure 39: A sub-dialogue about food associated with the sub-topic keywords **Italian cuisine and spaghetti carbonara**, along with each model's output.