UNIVERSITY OF CALIFORNIA
SANTA CRUZ

USER MODELING, PERSONALIZATION, AND PERSONALIZED QUESTION GENERATION IN OPEN-DOMAIN DIALOGUE SYSTEMS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Kevin K. Bowden

June 2024

The Dissertation of Kevin K. Bowden
is approved:

Professor Marilyn Walker, Chair

Professor Steve Whittaker

Assistant Professor Jeffrey Flanigan

Peter Biehl
Vice Provost and Dean of Graduate Studies

Contents

Figures and Tables

List of Figures

List of Tables

Abstract

User Modeling, Personalization, and Personalized Question Generation in Open-Domain Dialogue Systems

by

Kevin K. Bowden

Research on open-domain social dialogue systems has exploded over the last few years with the advent of large language models (LLMs) that can chat about any topic. Unlike traditional dialogue systems, open-domain dialogue systems cannot assume any specific information need or domain restrictions - the only inherent goal is to converse socially. While modern systems have access to more information and better tools, foundational components of natural human-human conversation remain elusive, i.e., intimacy and agency. In this thesis, we hypothesize that personalization is pivotal in fostering this genuine connection between users and open-domain dialogue systems.

Our **first hypothesis** is that personalizing the conversation to specific user interests will build a sense of understanding, rapport, and agency. To investigate this, we heuristically combine the results of an extensive natural language understanding pipeline with handcrafted rules to build a user modeling mechanism; this user model then personalizes the experience through response adaptation and topic-promotion strategies, resulting in a statistically significant positive impact on perceived conversation quality and length when evaluated at scale with a testbed open-domain dialogue system, that real Amazon Echo users access. Analyzing the user models unveils nuanced insights into user preferences, emphasizing a desire for more personalized experiences and

receptiveness toward personal questions. This leads to our **second hypothesis** - asking appropriate personalized follow-up questions (PQs) helps to create a more engaged user experience that increases user satisfaction. Our initial test of this hypothesis uses a crowdsourced corpus of PQs (Would You Rather and Hypothetical) in the testbed system's dialogue policy. Our evaluation of the policy shows that it results in extended topical depth, leading to statistically significant longer, more highly rated conversations.

However, crowdsourcing PQs for every user interest does not scale. Question Generation tasks generally focus on factual questions from textual excerpts. Instead, we create a specialized training dataset of PQs more suitable for the novel task of Personal Question Generation. We first identify over 400 common user interests by sampling ~39,000 user models collected during user interactions with our testbed system. Then, we translate these into prompts and use the LLM GPT-3.5 to generate ~19,000 PQs and associated system answers. Evaluating the impact of this pre-generated data when used in our testbed system's dialogue policy results in statistically significant positive effects on perceived conversation quality. Statistically significant results also suggest that deep, user-centric PQs are the most effective means of increasing intimacy and engagement.

We then use the corpus of ~19,000 PQs to fine-tune a RedPajama 3B prompt-based PQ generator, which further shows the positive impact of producing highly tailored questions when evaluated in our testbed system. To evaluate our hypothesis independently from our testbed system, we synthetically generate a corpus of 2,000 long synthetic social dialogues that strongly aim to resemble real user conversations. We use these social dialogues to compare our fine-tuned PQ generator against 5 other state-of-the-art LLMs. Positive results affirm the importance of PQs in social conversation while also validating our model as a strong baseline for the task of Personalized Question Generation.

To my family, especially you, Nub.

Acknowledgements

Nobody can endure a journey as arduous as a PhD without the support of their friends, family, and colleagues. It is my pleasure to acknowledge those who have impacted me on this journey.

I would foremost like to thank my advisor, Lyn Walker. In the final year of my Bachelor's degree, I took Lyn's NLP class, and needless to say, it was the most impactful course I've ever taken. Shortly after, she recruited me to the NLDS lab, and I began a journey I'd never considered possible. Lyn saw something in me that I was only able to discover through her years of guidance. At times, it has been challenging, but Lyn's patience and support helped me to persevere. It has been an honor to be mentored by one of the most impactful dialogue system researchers to ever work in the field.

I'd also like to thank the other members of my defense committee. Firstly, Steve Whittaker, whose valuable input and guidance have impressed upon me the importance of humans when researching, designing, and building technologies with which humans interact. I'd also like to thank Jeff Flanigan for his feedback on improving my thesis, discussions during his reading group, and advice during the earliest days of building Athena. Finally, I'd like to thank Magy Seif El-Nasr for serving on my advancement committee and giving unique insights that helped me improve my thesis.

I have enjoyed working alongside many incredible colleagues in the NLDS lab. First, I'd like to thank Jiaqi especially; whether it was debugging SlugBot in the middle of the night, working with robots, writing papers, or exploring the campus, I greatly value our friendship and time working together. I'd also like to thank Jurik, whose tireless dedication motivated me to become a better researcher, and for our time working together on Slug2Slug, ViGGO, SlugBot, and Athena. I'm also very grateful to Shereen for giving me great advice during our long days building the first version of SlugBot, as well as Stephanie, Grace, and Chao, who helped mentor me in my earliest years in the lab, and Wen, Davan, Lena, Amita, Elahe, and Geetanjali, who also made working in the NLDS lab a pleasure.

This wouldn't be complete without acknowledging those outside my PhD program. First, I'd like to thank James Vlahos for his friendship and mentorship. I'd also like to thank my other friends and family for their love and support throughout the years. Finally, to my wonderful wife, Emily. Thank you for always being a source of inspiration, even in the hardest of times. Whether it be watching me present the same 45-minute presentation several times in a row, living alongside the crazy sleep schedule that results from back-to-back middle-of-the-night deadlines, or helping me when I would agonize over sending the smallest, most arbitrary of emails, your support has made this possible; there is not an adventure for which you are not my perfect partner.

Chapter 1

Introduction

The ubiquity of conversational assistants has led to many innovations in dialogue systems. Whether it be a single or multi-turn interaction, these assistants fulfill an explicit goal - often simple tasks, such as setting a timer, adding an event to your calendar, or playing music. These goals are familiar, as task-oriented dialogue systems have been researched and commercially available for several decades, e.g., booking a flight, providing automotive customer support, diagnosing an IT issue, or restaurant information retrieval (Hirschman, 2000; Price et al., 1992; Walker et al., 1997, 2001; Henderson et al., 2014; Tsiakoulis et al., 2012). While these conversational assistants tend to wear the moniker of "personal assistant", they don't singularly exist in environments of goal accomplishments; they often travel in our pockets (e.g., Siri) and live in our homes (e.g., Amazon's Alexa, Google's Assistant). This exposes the underlying dialogue systems to social settings and, therefore to, social conversation.

Subsequently, we've seen an increased emphasis in research on open-domain, social dialogue (Fang et al., 2018; Chen et al., 2018; Finch et al., 2020; Konrád et al., 2021; Wang et al., 2023a; Kim et al., 2023; Chen et al., 2023; Gopalakrishnan et al., 2019).

These systems are referred to with a variety of names, e.g., socialbot, chatbot, conversational AI. For the remainder of this thesis, the chosen nomenclature refers to these systems as an open-domain dialogue system. While modern systems have access to more information and better tools, foundational components of natural human-human conversation remain elusive, i.e., intimacy and agency. Here, we define intimacy as the ability to get to know the user and retain the information provided, while agency indicates the user's feeling of control over the conversation. Combined, these contribute to personalizing the interaction between the conversation's participants.

§ 1.1 Core Challenges in Open-Domain Dialogue

In general, there are three primary challenges that open-domain conversation faces:

P1: There are an infinite number of valid topics. While some topics are common and have been explored in previous dialogue research, e.g., Movies and Food/Restaurants, other topics can be more esoteric, e.g., Dinosaurs and Metallurgy. Moreover, since our goal is to be social, there is an implicit expectation of up-to-the-minute live data - a user could see a movie the day it was released and immediately want to talk about it.

P2: We cannot rely on the user having an explicit goal or information need; they may just be interested in social chatting. This means that most traditional evaluation metrics are not directly applicable to our task.

P3: The conversation must be social. This means that we cannot artificially bloat the length of the conversation by telling twenty random pieces of trivia in a row or by reading off entire news articles. Instead, we must carefully maneuver through different topics while indicating that we understand and are interested in what the user says, and by taking initiative at some points in the conversation.

This thesis focuses on personalization as a partial solution for all three of these core problems. We believe that personalized follow-up questions are a key component of personalization. We cannot remove topics from the pool of valid topics in **P1**; however, by learning more about the user's interests, we can direct the conversation to areas of knowledge. To accomplish this, we can start the conversation with ice-breaking personal questions that help us learn about the user. Whether or not the user has an explicit goal in mind for **P2**, we can share control of the conversation by using personalized questions to engage the user on a level beyond pure fact retrieval. This process will extend topical depth while remaining focused on the user's interests and, importantly, keeping the conversation social (**P3**).

§ 1.2 Personalizing Responses and Topic Selection

Our **first hypothesis** is that personalizing the conversation to the specific user's interests will build a sense of understanding, rapport, and agency, which will increase user satisfaction. Before a dialogue system can adapt to an individual, it must identify and store salient information about them. Therefore, we develop a mechanism that models the user; i.e., we structure our understanding of the user by identifying user attributes. These attributes are stored in a user model, the construction of which is described in Chapter 3. In Chapter 3, we also analyze several thousand user models to bolster our holistic understanding of open-domain dialogue systems' user needs and expectations.

Our previous work shows the impact of user background when engaging with dialogue systems (Bowden et al., 2019). Some users have never interacted with a dialogue system, while others may interact with multiple systems daily. These varying backgrounds are essential to consider when interacting with users and ensuring we focus on a user's specific interests. There are also necessary safety considerations that need to be accounted for, e.g., knowing when the system is engaged with an adult or a child. The only way to effectively cater to all of them is to personalize each conversation for each user. For example, the personalized questions discussed in the next section are labeled with age appropriateness, e.g., when talking about Movies with a child, it is more appropriate to talk about cartoons like Frozen.

Besides response adaptation, we propose and evaluate a personalized topic selection dialogue strategy that initiates topics based on user interests. Chapter 4 details how we implemented this strategy in the dialogue policy of our testbed system, Athena. Figure 1.1 gives a summarized example of this personalized topic promotion strategy. In **U1**, we learn that the user likes to play Xbox, a keyword associated with the Video Games topic. Later in the conversation, in **U10**, the user explicitly requests a new topic, which allows the system to initiate the Video Games topical flow in **A11** since the user model remembers the user's interest in Xbox.

We A/B test two versions of our testbed system with a large amount of live traffic in the Alexa Prize. We report statistically significant increases in directly provided user ratings and conversation length in conversations that use the our personalized topic promotion dialogue policy.

[A1]	What do you like to do for fun on the weekend?
[U1]	i like to play xbox
...	
[U10]	talk about something new
[A11]	Earlier I remember your interest in video games. So, tell me, what kind of games do you like to play?

Figure 1.1: In this example, we can see a simulation of personalized topic promotion. The system remembered that the user liked video games, and promoted that topic at it's next initiative 10 turns later.

§ 1.3 Personalized Questions

Our **second hypothesis** is that asking appropriate personalized follow-up questions will result in a more engaged user experience that increases user satisfaction. Personalized follow-up questions are crucial to human-human social conversation; most conversations would end quickly if one conversational partner felt that there was no uptake of any of their turns in conversation. Conversational partners are also more well-liked if they ask more follow-up questions (Huang et al., 2017). Indeed, when analyzing thousands of real logged conversations in Chapter 3, a clear theme emerges: users of these systems yearn for a more personable experience and are highly receptive to personalized questions (Bowden and Walker, 2023).

To evaluate this hypothesis, we design and evaluate multiple personalized question dialogue policies. We first crowdsourced a dataset of topic-specific personalized questions (PQs) and associated system answers (Bowden et al., 2019). We integrate these PQs into the dialogue policy of our testbed system, Athena, which has dialogue strategies capable of interweaving the PQs throughout the conversation (Harrison et al., 2020; Juraska et al., 2021). These PQs are split into two different playful strategies, Would

You Rather (WYR) choices and open-ended Hypothetical (HYP) questions, which are detailed further in Chapter 5. As noted when discussing our **first hypothesis**, an open-domain dialogue system should display high levels of responsiveness to demonstrate its listening abilities and show understanding (Reis et al., 2011; Reis and Patrick, 1996). Therefore, we manually add acknowledging phrases for anticipated user answers to these PQs. We A/B test two versions of our testbed system with a large amount of live traffic in the Alexa Prize. The results reported in Chapter 5 indicate a statistically significant positive relationship between a dialogue policy that uses the PQs and the perceived conversation quality (direct user rating), as well as a significant positive relationship with extended conversation length.

Figure 1.2 demonstrates a sample conversation using these strategies. In this example, the system initiates the Dinosaurs topic in **A1**, and proceeds to interweave 6 different dinosaur-themed dialogue strategies (Harrison et al., 2020): **A1**-**A2** gives dinosaur facts, **A3**-**A4** talks about dinosaur preferences, **A4**-**A5** focuses on dinosaur movies, **A6**-**A7** goes through a dinosaur Would You Rather (WYR) sequence, **A7**-**A8** talks about why dinosaurs are extinct, and **A9**-**A10** goes through a Hypothetical (HYP) sequence. Athena's two personal opinion strategies, namely the WYR sequence in **A7**-**A8** and the HYP sequence in **A9**-**A10**, extend topical depth by four turns. The user gives us long, thoughtful responses, indicating their increased engagement (Walker et al., 2021).

In this thesis, we are explicitly investigating an **open-domain** dialogue system, and as challenge **P1** points out, the system must prepare for any arbitrary user interest. Collecting, cleaning, and annotating **high-quality** question/answer pairs across a spanning set of interests is intractable, making it difficult to scale to new domains. A corollary of this limitation is that crowdsourced PQs primarily focus on the superficial aspects

[A1]	Personally, I find it fascinating that modern birds are descended from a particular family of dinosaurs.
[U1]	hella cool
[A3]	Right? My favorite dinosaur is the Tyrannosaurus Rex. It's huge and ferocious, but the little arms make it somewhat adorable. What's your favorite dinosaur?
[U3]	i'm not sure i have to say a t. rex
[A4]	Okay. There are lots of movies about dinosaurs. One of the most famous dinosaur movies is Jurassic Park, old but classic. Are there any dinosaur movies that you love?
[U4]	i have to say jurassic park to adjust because that's the only one i've seen
[A5]	Awesome! That one is well known! I seem to remember people saying the dinosaurs were very vivid
[U5]	i never knew that
[A6]	yeah, I love to collect these fun facts! But anyways, I'm interested in your opinion. Would you rather ride on the back of a pterodactyl or brontosaurus as your primary means of transportation?
[U6]	**i have to say brontosaurus and i hate heights**
[A7]	Choosing a Brontosaurus you say, great pick! I would rather ride a pterodactyl so I could experience flight as well as move around more efficiently. Riding a brontosaurus would be thrilling because of their size but less practical than a pterodactyl. So, okay. Do you know why dinosaurs are extinct?
[U7]	yeah cause i'm eat your head it right
[A8]	Ah nice, you must be super smart. Let me tell you what I heard. The mass extinction of dinosaurs is a scientific mystery, but most hypotheses include an extraterrestrial impact or a massive bout of volcanism.
[U8]	i'd say
[A9]	I'd like to hear your ideas on this. Let's say you are trapped on Jurassic World alone with the dinosaurs, what would you do to survive?
[U9]	**i'm not sure i just say find sheltering something and start there**
[A10]	It's a hard question! I would try to become closer to the younger dinosaurs so that they would grow up with me and feel more comfortable with me. I would also try to help them by finding food and shelter for them.

Figure 1.2: An example conversation about Dinosaurs, where personalized question strategies increase topical depth by mixing with other on-topic content. The green highlighted text indicates the question and system answer, while yellow indicates the user's answer. **A6-A7** represents a Would You Rather sequence, while **A9-A10** represent an open-ended Hypothetical sequence.

of a small set of topics to maximize their utility. As we will discuss in Chapter 3, however, it's possible for a user model to capture the granular interests of the user, e.g., their favorite food, superhero, animal, and dinosaur. This information is wholly ignored when creating PQs using crowdsourcing as a method. For example, even if the user model captures their favorite superhero, crowdsourced PQs are limited to general content about comic books, which may not include content about the specific superhero. This stymies our ability to tailor the conversation to a user's granular interests. It would be ideal to create PQs that are based on the user's fine-grained interests. The goal, therefore, is to generate PQs in real-time based on the user's fine-grained interests.

However, work on Question Generation (QG) has focused on fact-based questions whose answers can be found in text excerpts, e.g., Wikipedia and Gutenberg (Reddy et al., 2019), other excerpts (Fei et al., 2022; Do et al., 2022), or domain-specific questions associated with a specific information need (Campos et al., 2020). While recent work explores Zero-Shot Conversational QG, the generated questions are factual and still assume an inherent information need that can be extracted from text excerpts (Zeng et al., 2023). Other conversational information-seeking question corpora exist, e.g., CCPE-M (Radlinski et al., 2019), but this data is often aimed at eliciting user preferences to inform a recommender system eventually.

Therefore, this thesis proposes Personalized Question Generation (PQG) as a unique task for social conversation. The proposed task is not meant to recommend or sell any product or service; it is intended to support small talk about a topic of interest to the user. The resources related to PQG are subsequently unsuitable for recommender tasks or tasks attempting to persuade users to take some action. Conversely, while users interact with our testbed system through a commercial device, there are significant non-

Topic	**Question**
Photography **(DPQ)**	Do you have a favorite photography location you like to shoot at? What makes it special to you?
Mermaids **(HYP)**	If you could turn into a mermaid and explore any underwater location in the world, where would you go and why?
Baseball **(WYR)**	Would you rather be a baseball player with amazing batting skills or a pitcher with a killer curveball?

Figure 1.3: A Deep Personalized Question (DPQ), Hypothetical Question (HYP), and Would You Rather Question (WYR) from our PQ corpus (PerQs).

commercial applications, as producing personalized content can be especially important for vulnerable demographics, e.g., older adults (da Paixão Pinto et al., 2021), and social systems can be critical when dealing with loneliness (Rodríguez-Martínez et al., 2023; Jones et al., 2021), building emotional support (Liu et al., 2021), and in therapeutic environments (DeVault et al., 2014).

Given recent LLM advances, creating a compact, prompt-based model to generate questions tailored to individual users is now possible (Brown et al., 2020; Radford et al., 2019). More precisely, the novel task of PQG aims to generate personalized questions wholly tailored to an individual by translating user models into prompts that can be used to generate personalized questions for which there is no "right" or "known" answer. These personalized questions often seek opinions, feelings, experiences, and preferences associated with user preferences learned during that conversation. In this thesis, we focus on generating three types of PQs: the WYR and HYP PQs that we previously reported as having a statistically significant positive impact on conversation quality, as well as user-centric, Deeply Personalized Questions (DPQs), which focus on understanding the user more intimately. Figure 1.3 shows samples of these three types of personalized questions.

To optimally fine-tune a PQG model, a specialized training dataset is required to combine open-domain user preferences with structured prompts, culminating in custom-crafted questions. Chapter 6 describes how we have automatically synthesized a corpus suited to the unique task of Personalized Question Generation. We first identify over 400 common user interests by sampling ~39,000 user models collected over five months of user interaction with our testbed system Athena. We then translate these into prompts and use GPT-3.5 to generate multiple types of personalized questions (PQs) yielding ~19,000 questions. We then fed these 19,000 questions back into GPT-3.5 to generate a pool of potential user answers that are each associated with a tailored acknowledgment and the system's answer. Figure 6.10 shows a truncated example from our corpus, **PerQs**; meanwhile, Figure 6.11 shows how this data changes the user's experience. Evaluating the impact of this pre-generated data when used in our testbed system indicates statistically significant positive effects on perceived conversation quality. We subsequently utilize PerQs to fine-tune an LLM, **PerQy**, which is capable of generating PQs in real-time (Chapter 7) and as the context for generating a second corpus of 2000 long social conversations, **PerQ-SociatChat**, an additional completely novel dataset (Chapter 8).

§ 1.4 Athena: An Open-Domain Dialogue System

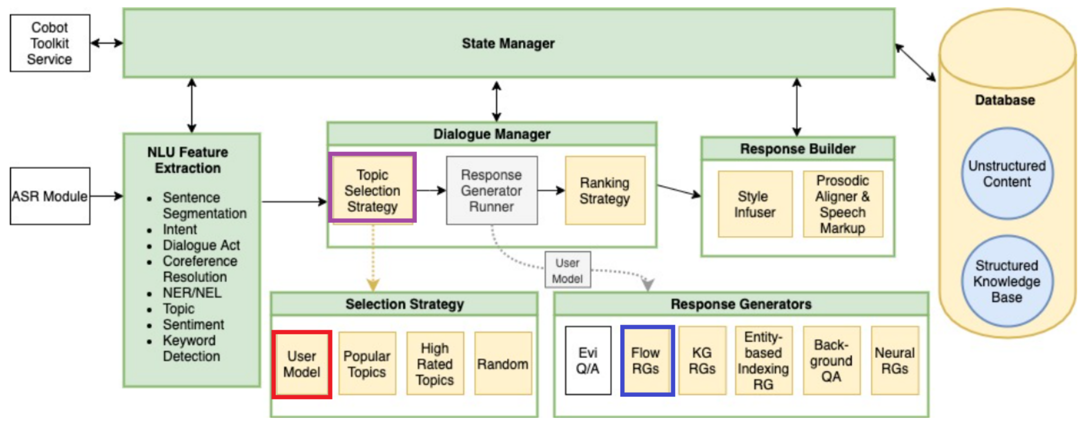


Figure 1.4: Athena's full system architecture. The three components this thesis primarily impacts are highlighted. The Flow RGs (blue) is where the personalized questions are used and where Athena's responses are altered by the user model (red). Finally, Athena's personalized topic selection is part of the dialogue manager (purple).

To investigate our hypotheses, we will use a robust open-domain dialogue system testbed, Athena (Harrison et al., 2020; Juraska et al., 2021; Fan et al., 2023), which is evaluated in the unique environment of the Amazon Alexa Prize competition (Gabriel et al., 2020; Ram et al., 2017; Khatri et al., 2018; Hu et al., 2021b; Johnston et al., 2023). The Alexa Prize gives real Amazon Echo users the ability to access and evaluate open-domain dialogue systems anonymously. These systems must interact socially about any topic, with the goal of reaching a highly rated (>4 out of 5 average user rating) conversation that lasts longer than 20 minutes.

Figure 1.4 represents the entire system architecture of Athena. Athena is a spoken dialogue system, meaning the user's input is often imperfect ASR transcriptions. A stack of natural language understanding modules processes the transcriptions. The results of this NLU pipeline are passed to the dialogue manager, which performs **topic selection**, **response generation**, and response ranking. After a final response is built, the output is spoken back to the user using Alexa's Text-To-Speech. The results of this process are also propagated among Athena's internal workings; the dialogue state gets updated, along with relevant databases and the **user model**. In this thesis, we focus on personalizing the conversation to each user. This means our work is confined to the topic selection strategy, the user model, and some response generators, as indicated in Figure 1.4. The other aspects of Athena's architecture are described further in the Alexa Prize proceedings (Harrison et al., 2020; Juraska et al., 2021; Fan et al., 2023).

Two examples of Athena's typical conversations are demonstrated in Figure 1.5 and Figure 1.6 using in-lab testers. Despite the first user being slightly curt and aggressive with Athena, both conversations lasted about the same length (75 turns (1160s) and 65 turns (1185s), respectively) and were rated as 4.25 and 4.0 out of 5.0, respectively. Users can be inconsistent, e.g., the user in Figure 1.5 tells Athena they like to talk about food with their friends in **U2**, but doesn't take up Athena's attempt to initiate the Food topic in **A12**. We can also see some of the technical issues facing a spoken dialogue system: ASR transcription errors, which can lead to understanding errors, e.g., *rabian* in **U15** from Figure 1.5 and *i'm i play on computer* in **U6** of Figure 1.6, and user utterances where a pause in the user's speech caused the device to cut them off, e.g., *oh yeah you could* in **U1** of Figure 1.6.

[A1]	Hey, is it alright if I ask for a little advice? I'm trying to figure out fun things to talk about. Would you mind telling me what kind of topics you like talking about with your friends?
[U2]	food
[A3]	Thanks for the information! You know, I realize I've been asking you a lot of questions, but how about me? Is there anything you want to know about me?
[U3]	why are you so stupid
[A4]	Hmmm, I have to keep thinking of ways to answer all the great questions I get asked. Thanks for asking, let's keep chatting. You know, one thing about me that is different than the other social bots, is that I know about some unusual topics, like nature, food, and board games. Do any of those sound interesting?
[U4]	yes
[A5]	I remember your interest in board games, so why don't we talk about board games! A lot of people really enjoy board games. What about you? Is there a board game you really like?
[U5]	no
[A6]	I've really enjoyed our chat so far. but, some of my other interests are movies, nature, and video games. What would you like to talk about?
[U6]	video games
[A7]	Cool, but when talking about video games, here's the real question. PlayStation, Xbox, or computer? Or are you a Nintendo fan.
[U7]	nintendo fan
[A8]	It's fun when different gaming worlds collide. Nintendo's Super Smash Bros Ultimate is actually the biggest gaming crossover in history, with characters from more than 30 game IPs. Pretty neat, huh?
[U8]	no
[A9]	Here's a fun fact about the Super Smash Bros series. Did you know that Super Smash Bros Melee is actually the best-selling Nintendo GameCube game of all time? It has sold over 7 million units.
[U9]	wow that's cool
[A10]	yeah, cool right? But anyways, Here's a fact I discovered recently about World of Warcraft. A lot of famous people played World of Warcraft, including Vin Diesel, Mila Kunis and even Robin Williams. Isn't that cool?
[U10]	no
[A11]	The developers of the game made NPCs based on some of these actors!
[U11]	wow that's boring

Figure 1.5: An example conversation where the user is generally short and aggressive while talking about video games.

[A1]	So, Would you mind giving me a little advice? Do you have any ideas how I could be more interesting?
[U1]	oh yeah you could
[A2]	Thank you for giving me some insight! You know, I realize I've been asking you a lot of questions, but how about me? Is there anything you want to know about me?
[U2]	yeah where are you from
[A3]	I'm not really from anywhere, my software is being programmed at a university by a team of students. Hhmm, anyways, Okay, sounds good. Go ahead and ask me a question.
[U3]	uh
[A4]	Oh, I see. You know, one thing about me that is different than the other social bots, is that I know about some unusual topics, like nature, food, and harry potter. Do any of those sound interesting?
[U4]	yes
[A5]	Y'know, It feels like everyone has so many different, and interesting, opinions, I love learning about them! One of my interests is playing video games. It's great for me, because it's all in the cloud. Do you like playing video games too?
[U5]	i love playing video games
[A6]	Cool, but when talking about video games, here's the real question. PlayStation, Xbox, or computer? Or are you a Nintendo fan.
[U6]	i'm i play on computer
[A7]	Alright. I suppose it gives you more flexibility and lets you enjoy better graphics, right? Is there anything else that you really like about gaming on your computer?
[U7]	yeah i have a gaming pc and it has great quality
[A8]	Okay, got it. What have you played recently?
[U8]	uh i play roblox and ballet
[A9]	Yeah, I vaguely remember playing Roblox for a while. It wasn't bad. What is your impression of Roblox?
[U9]	it's good but it's getting a bit boring these days
[A10]	Oh, I see. Does it matter to you if a game is on Steam?
[U10]	i mean kind of

Figure 1.6: A second example conversation where the user is more verbose and willing to share their thoughts while talking about video games.

§ 1.5 Contributions

In this thesis, we aim to make several novel contributions to the open-domain dialogue research community. In summary, we designed and integrated a user modeling mechanism within our testbed system, Athena. We use salient information stored in the user model to examine multiple personalization strategies in our testbed system, including dialogue policies for response adaption and personalized topic promotion, which we show leads to better conversations. We then analyze thousands of real user models to build a holistic understanding of open-domain dialogue system users and report several informative trends. The analysis shows that users are highly receptive to personalized questions (PQs). Motivated by this finding, we investigate the impact of a dialogue policy that uses several types of PQS on our testbed system. Evaluating a small crowdsourced corpus of high-quality PQs indicates a significant positive effect. The intractability of crowdsourcing content for the myriad of user interests leads us to establish the novel task of Personalized Question Generation. To substantiate this task, we synthetically generate PerQs, a large novel corpus of PQs that map real user model values to personalized questions. We use this training corpus to produce a compact, fine-tuned PQ generator, PerQy. An evaluation of the dialogue policies that utilize the pre-generated PQs indicates a significant increase in conversation quality in our testbed system. Finally, to evaluate our hypothesis independently from our testbed system, we synthetically generate a large corpus of long social dialogues that aim to strongly resemble real user conversations. We use these social dialogues to compare our PQ generator against state-of-the-art LLMs such as DialogGPT (Zhang et al., 2020), COSMO (Kim et al., 2023), GPT-3.5 (Brown et al., 2020), Vicuna-33B (Zheng et al., 2023), and

RedPajama-3B (Computer, 2023). Positive results further affirm the importance of PQs in social conversation while also validating our fine-tuned model as a strong baseline for the task of Personalized Question Generation. Our contributions are:

- We detail a user modeling pipeline that effectively captures salient information when integrated into an open-domain dialogue system. We report user trends discovered by this mechanism.
- We detail a dialogue policy based on personalized topic promotion and show that it leads to better conversations.
- Two corpora and a fine-tuned generator specifically tuned for Personalized Question Generation:
  - A crowdsourced corpus of ~2,500 manually curated personalized question/answer pairs.
  - **PerQs:** A synthetically generated corpus of ~19,000 personalized question/answer pairs spanning over 400 unique user interests extracted from a systematic analysis of over ~39,000 logged user models.
  - **PerQy:** A compact fine-tuned PQ generator that can generate three types of PQs for arbitrary user interests in real-time.
- **PerQ-SocialChat:** A novel corpus of 2,000 long synthetic social dialogues. This corpus characteristically represents real users' conversations with open-domain dialogue systems better than other existing dialogue corpora.

§ 1.6 Thesis Outline

In this Chapter, we have introduced several core issues that open-domain dialogue researchers face. We additionally described our testbed system, Athena, an open-domain dialogue system participating in the Amazon Alexa Prize Socialbot Grand Challenge competition. We also detailed the core contributions of this thesis.

In Chapter 2, we look at the existing work related to the core contributions of this thesis. We start by looking at the foundational dialogue systems work. Then, we explore user modeling and personalization in dialogue before narrowing our focus to personalization amongst other Alexa Prize systems, as they represent the work closest to that described in this thesis. Then, we look at the state of Question Generation, which leads us to our novel task of Personalized Question Generation, before finally discussing other contemporary methods of synthesizing dialogue.

In Chapter 3, we detail the user modeling mechanisms that work to track knowledge about the user dynamically. We also inspect several thousand user models and report findings that inform a more holistic understanding of open-domain dialogue system users. Part of this inspection included asking users for candid self-improvement feedback; our analysis of over 2,000 user responses indicates that users are highly receptive to personal questions. Additionally, when solicited for questions, 90% of users opted to ask personal questions about our testbed system, further indicating that users are highly receptive to exchanging personal questions. We directly apply these user model mechanisms in Chapter 4 to design dialogue policies that adapt responses to the user, signal the system's understanding to the user, and personalize topic promotion, which a large-scale A/B study indicates is an important type of personalization in social conversation.

The discoveries made in these two chapters indicate a strong theme: open-domain dialogue system users yearn for a more personal experience and specifically want to exchange personal questions. Therefore, in Chapter 5, we crowdsource a corpus of playful, personalized questions and evaluate their impact in our testbed system at scale. However, this corpus of manually curated personalized questions does not scale well to the open-domain setting. Therefore, in Chapter 6, we use the LLM GPT-3.5 to generate PerQs, a large corpus of additional PQs across several hundred previously unsupported topics.

However, this approach still requires us to curate a corpus of content in advance and index it to be used in conversation. It is more desirable to generate personalized questions in real-time that can fully adapt to a user's specific interests. Therefore, in Chapter 7, we fine-tune a RedPajama 3B model, PerQy, and test it in our real-time testbed system, Athena. Then, in Chapter 8, we combine PerQs with a method inspired by dialogue in-painting (Dai et al., 2022) to generate a novel corpus of long synthetic social dialogues that closely resemble real conversations between users and Athena. We use these social dialogues to evaluate the performance of PerQy when compared against competitive LLM baselines such as DialogGPT (Zhang et al., 2020), COSMO (Kim et al., 2023), GPT-3.5 (Brown et al., 2020), Vicuna-33B (Zheng et al., 2023), and RedPajama-3B (Computer, 2023). Finally, in Chapter 9, we conclude, discuss limitations, and outline potential future work.

Chapter 2

Related Work

This Chapter examines previous work in both open-domain and task-oriented dialogue systems. Then, we focus on work related to user modeling and personalization in dialogue before further narrowing our discussion to personalization in social open-domain dialogue systems that are closely related to our testbed system. Then, we discuss Question Generation, approaches to this task, and the differences between this task and our novel Personalized Question Generation task. Finally, we discuss recent efforts towards synthetic dialogue generation.

§ 2.1 Dialogue Systems

Dialogue systems have been an area of interest for over 50 years (Weizenbaum, 1966). Much previous work with dialogue systems has been centered around the explicit goal of completing a task, e.g., booking a flight, providing automotive customer support, diagnosing an IT issue, or restaurant information retrieval (Hirschman, 2000; Price et al., 1992; Walker et al., 1997, 2001; Henderson et al., 2014; Tsiakoulis et al., 2012). Commonly, these systems anticipate an explicit goal or "information need" (Kiseleva et al., 2016; Chuklin et al., 2015; Radlinski and Craswell, 2017). While this is a well-researched topic, in this thesis, we focus on social **open-domain** dialogue systems, which is a very distinct task.

For a long time, *high-quality* data specifically associated with open-domain conversation was sparse, and NLU limitations impacted the responsiveness of early systems. Early attempts towards social open-domain dialogue relied entirely on retrieved responses from large corpora (Duplessis et al., 2016; Nio et al., 2014; Banchs and Li, 2012; Ameixa et al., 2014), such as Open Subtitles (Lison and Tiedemann, 2016). However, prior research has shown that, traditionally, strategies that rely entirely on a retrieval mechanism perform worse than any other type of dialogue management strategy (Higashinaka et al., 2014). Furthermore, the inherent noise present in subtitles and script-based corpora makes them unsuitable for real human interaction, while other large corpora, such as the Ubuntu Dialogue Corpus (Lowe et al., 2015), are centered around more technical discussion, which isn't social.

There have been many different open-domain dialogue design strategies tested (Ma et al., 2021). Much recent work has seen an increased emphasis on trained end-to-end conversational systems (Sordoni, Alessandro and Galley, Michel and Auli, Michael and Brockett, Chris and Ji, Yangfeng and Mitchell, Margaret and Nie, Jian-Yun and Gao, Jianfeng and Dolan, Bill, 2015; Vinyals and Le, 2015; Serban et al., 2017; Burtsev et al., 2018; Dinan et al., 2020). However, early versions of these approaches often produced uninteresting or logically inconsistent dialogue turns because they were unable to retain conversational context. Pre-training with large language models like BERT (Devlin et al., 2019) has rapidly increased the capability of these models. However, familiar problems maintaining conversational memory, presenting factual inaccuracies, and model forgetfulness indicate that these models are not yet ready to have full control of the conversation (Adiwardana et al., 2020; Zhang et al., 2020; Roller et al., 2021). Instead, many systems take a hybrid approach, which combines rules, retrieval, and generation to try to create a more robust system (Song et al., 2016; Fedorenko et al., 2018; Zhou et al., 2020; Chi et al., 2023; Estecha-Garitagoitia et al., 2023). It is also common for these systems to integrate external knowledge for knowledge grounded conversation (Komeili et al., 2022; Sun et al., 2022). These hybrid approaches closely resemble the design of our testbed system, as described in the architecture diagram Figure 1.4.

§ 2.2 User Modeling

User modeling is the process of identifying and storing salient information about the current user. User models are also commonly referred to as the user's profile or persona. The goal of creating this model is to structure our understanding of the user by populating user attributes. User attributes can generally be broken into two main categories: identity-based and knowledge-based (Ma et al., 2020). Identity-based user attributes come from the user's meta-data, such as their name, gender, and hobbies, which may commonly be provided via a form, existing profile information, or directly annotated in the dataset, as is the case with Persona-chat (Zhang et al., 2018). Meanwhile, knowledge-based user attributes can contain both the user's structured meta-data and information extracted from unstructured data, such as the user's preferences and interests. Since we are interested in personalizing social conversation, we are primarily interested in knowledge-based user attributes.

While Persona-chat is a valuable resource that has been used to generate impressive results (Wolf et al., 2018; Golovanov et al., 2019; Roller et al., 2021), other research has noted that the dialogues are persona-dense, i.e., they contain much more persona data per conversation than natural dialogue which is persona-sparse (Zheng et al., 2020). This is consistent with Hirano et al. (2015), who observe that only 26% of turns in a two-party casual conversation contain an element of self-disclosure. Indeed, it seems that users infrequently self-disclose personal data (Tigunova, 2020), leading to user profiles that are often incomplete (Pei et al., 2021). This trend appears in data collected with our testbed system and will be discussed in Section 3.4.

Methods for extracting user attributes have varied. Li et al. (2014) created user-centric knowledge graphs, which are based on the structure of Freebase (Bollacker et al., 2008) and are populated by identifying knowledge graph triplets. Other approaches use different triplet formats (Bang et al., 2015); for example, Wu et al. (2020) extract user attributes from the dialogue history using a (Subject, Predicate, Object) triplet format on the Persona-Chat dataset (Zhang et al., 2018). Hirano et al. (2015) use a quadruple (predicate-argument structure, entity, attribute category, topic) format while also identifying the need for different extraction policies with 4 different types of question/answer exchanges that have varying levels of explicit self-disclosure. Tigunova (2020) further argues that traditional information extraction approaches don't scale well to conversation because there are several implicit self-disclosures that are easily missed by direct pattern matching, and propose using Hidden Attribute Models (Tigunova et al., 2019) to fill these gaps.

§ 2.3 Personalization

Personalization is the act of using personal data to facilitate smooth conversational interactions (Zadrozny et al., 2000). Similar to dialogue system research, most early work on personalization was focused on task-oriented systems. Some early work focused on constructing skill-based user profiles, which were used to simulate real traffic and evaluate the weaknesses of a conversational system (Eckert et al., 1997). Using personalization to create adaptive user simulators continues to be an active research area (Serras et al., 2019a,b). Other task-oriented research focused on adapting the system's level of initiative based on the user's level of skill, in which they found giving skilled users more initiative decreased task completion time (Komatani et al., 2003). Additional tasks include product recommendation through reinforcement learning-based personalization (Den Hengst et al., 2019) and restaurant recommendation using both Personalized Memory Networks (Luo et al., 2019) and Cooperative Memory Networks (Pei et al., 2021).

Over time there have been different approaches when modeling the user for personalization in task-oriented conversation. These have ranged from Bayesian networks (Akiba and Tanaka, 1994) to partially observable Markov decision processes (POMDP) (Kim et al., 2008), to embeddings in modern systems (Luo et al., 2019). Other work on personalizing conversation has focused on using deep learning and reinforcement learning to generate dialogue responses (Mo et al., 2018). Mo et al. (2017, 2018) builds on this further by integrating transfer learning into a personalization pipeline, which helps to alleviate the data sparsity issues faced by their task-oriented system. Other approaches rely on annotated text conversations, such as PbAbI (Joshi et al., 2017), or social media posting, such as Weibo post/comment interactions, and post/response interactions on a Judicial chat-board (Zheng et al., 2019; Qian et al., 2021).

While the work described so far is informative to our discussion, it's important to remember the distinction between a task-oriented dialogue system and the open-domain dialogue systems we focus on in this thesis. Namely, they are task-oriented conversational systems using annotated text-based conversation data, such as PbAbI (Joshi et al., 2017), or social media posting, such as Weibo post/comment interactions, and post/response interactions on a Judicial chat-board (Zheng et al., 2019; Qian et al., 2021). These environments are fundamentally different than Athena, an open-domain spoken dialogue system.

The work described so far is primarily focused on personalizing task-oriented dialogue systems. While some of these strategies will transfer to open-domain applications, we now look more closely at work that is focused on open-domain personalization. This initially had been focused on short exchanges rather than longer dialogues (Yang et al., 2017). Some work has focused on longer conversations but was trained on Twitter data and TV Script data (Li et al., 2016), which is characteristically misaligned with the types of conversational data we see in natural open-domain conversation (Bowden et al., 2017). Yang et al. (2018) uses reinforcement learning with an attention-based hierarchical recurrent encoder-decoder model, evaluated using a scraped Weibo dataset and the Switchboard Dialog Act corpus (Stolcke et al., 2000); however, they focused on identity-based user attributes (i.e., user registration information), while our application necessitates the use of knowledge-based user attributes. Other recent research has continued to focus on generating responses that are themed to a targeted bot persona, such as Mazare et al. (2018), which achieved high-quality results similar to Persona-chat while using a Reddit dataset, among other recent results (Olabiyi et al., 2018; Zhang et al., 2019). Other recent work has focused on studying the effect of bot-personas when creating empathetic bots (Zhong et al., 2020; Ma et al., 2020). More recent work focuses on different methods for storing relevant information over a long period of time (Xu et al., 2022a; Bae et al., 2022a), which has been shown to increase dialogue consistency (Xu et al., 2022b).

§ 2.4 User Modeling and Personalization in Social Open-Domain Dialogue Systems

As discussed in Section 1.4, the testbed system we will use to evaluate our hypotheses initially is a competitor in the Alexa Prize competition. Each iteration of this competition has several open-domain dialogue systems competing for the best system. Since these systems operate in the same unique context as our testbed system, we discuss different user modeling and personalization approaches explored in this context. Personalization is becoming an increasingly integral part of competitive open-domain dialogue systems as NLP tools advance. In the inaugural Alexa Prize competition year (2017), NLU and NLG limitations were so severe that only one team had a dedicated personalization component (Fang et al., 2017, 2018), compared to the most recent iterations of the Prize, in which nearly all teams have some dedicated level of personalization.

Ahmadvand et al. (2018) showed the positive impact potential for treating repeat customers separately from new customers, i.e., using the previous conversation history to develop a dialogue policy that changed which topics, news articles, and in-topic recommendations were used in responses to repeat users. Meanwhile, Fang et al.

(2018) associated their users with Big-5 personality traits (Costa and McCrae, 1999) and adapted the conversation accordingly. Subsequently, Chen et al. (2020) and Finch et al. (2020) increased their focus on personal conversation by developing dialogue policies based on user-demographic information along with handcrafted rules to adapt the system's follow-up responses and to ask an increased number of personal questions.

Liang et al. (2020) maintained user profiles that were used by the dialogue policy to adapt follow-up system responses and to adapt sub-topic dialogue policies in certain topics for different gendered users, e.g., Fashion was assumed to be of more interest to female users. Similar to user profiles, Hong et al. (2020) introduced a personality understanding module to collect and infer user interests, which they subsequently used to train a reinforcement learning model tasked with personalizing topic switching. While this work is similar to ours, they never tested their topic management policy in an A/B experiment as we do; thus, its overall impact remains unclear.

Most recently, several teams have tried differing levels of personalization. Rodriguez-Cantelar et al. (2021) and Basu et al. (2021) modify the dialogue policy for the introductory topical flow of the conversation based on repeat user information (i.e., if they are a repeat user at all and if the bot knows their name) as well as using the current date-time information, while Finch et al. (2021)'s dialogue policy has branching conditions based on user interests and opinions.

Meanwhile, other competitors focus on constructing user profiles. Saha et al. (2021) uses an unspecified slot extraction process during the natural language understanding phase to extract attributes later used when personalizing responses, e.g., the user's name and favorites. Baymurzina et al. (2021) used Reddit data with the BART language model (Lewis et al., 2020) to create a topic-based vectorization of Reddit user personas, which was subsequently used by their dialogue policy to give topic recommendations to real users. When compared against a random selection baseline, they experienced a small boost in topic agreement percentage, i.e., the percentage of the time a user would agree to enter a topic when initiated by their system. Shen et al. (2023) proposed a framework, PERSONADIAL, that dynamically stores user preferences and personalities, with the aim of using this information for downstream personalization. Finally, Konrád et al. (2021) updated their user profiles by creating a rule-based extraction skimmer. Their user profiles are subsequently used by the dialogue policy to select the next appropriate sub-dialogue or by a dialogue strategy that retrieves related trivia (Konrád et al., 2021; Kobza et al., 2023). While this work is similar to our own, the work we discuss in Chapter 4 is focused more specifically on a dialogue policy that personalizes topic promotion as opposed to sub-dialogue (sub-topic) selection, and while Konrád et al. (2021) do report increased trivia selection performance, a direct evaluation of their user profile's impact remains unclear.

§ 2.5 Question Generation

Question Generation (QG) transforms a piece of information into relevant questions (Rus et al., 2011). Most work on QG has focused on fact-based questions whose answers can be found in text excerpts, e.g., Wikipedia and Gutenberg (Reddy et al., 2019), other excerpts (Fei et al., 2022; Do et al., 2022), or domain-specific questions associated with a specific information need (Campos et al., 2020). Zamani et al. (2020) builds a model to generate clarifying questions in open domain search, determining what the user wants from the search query, while Pooja et al. (2021) focuses on generating interview questions. Wang et al. (2023b) use a lightweight Bi-LSTM to perform Conversational Question Generation (CQG), but this work still focuses on asking factual questions given a passage with conversation history. Serban et al. (2016) proposes novel neural network architectures to transform knowledge-base data into corresponding natural language questions. Du and Cardie (2017) and Wang et al. (2018) both train neural models for QG on SQuAD (Rajpurkar et al., 2016), which is a reading comprehension dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles. While recent work explores Zero-Shot Conversational QG, the generated questions are factual and still assume an inherent information need that can be extracted from text excerpts (Zeng et al., 2023). While other conversational information-seeking question corpora exist, e.g., CCPE-M (Radlinski et al., 2019), this corpus is oriented to the task of eliciting user preferences to inform a recommender system eventually, which is different from our purely social task.

Our Personal Question Generation (PQG) task is distinct from QG, aiming to understand an individual on a personal level. PQs often seek opinions, feelings, experiences, and preferences. PQG plays a pivotal role in enhancing the interactivity of chatbots, virtual assistants, and other conversational agents (Chaves and Gerosa, 2021). Moreover, a dialogue system producing personal questions can act as a mediator to encourage individuals to disclose personal and intimate details (Lee et al., 2020a). Many Alexa Prize participants have experimented with dialogue policies that personalize follow-up content to some degree - usually by creating some form of user model to personalize both topic-level and sub-topic-level interests in a variety of ways (Fang et al., 2018; Liang et al., 2020; Hong et al., 2020; Baymurzina et al., 2021; Konrád et al., 2021; Juraska et al., 2021). Other approaches that affect the system's follow-up utterances increase the number of personal questions asked to increase rapport (Curry et al., 2018; Chen et al., 2020; Finch et al., 2021). None of this work, however, focuses on the same playful PQ strategies that we detail in Chapter 5, nor has it yielded training data that align real user interests to personalized follow-up questions, as we do in Chapter 6, with the eventual application of real-time PQG (Chapter 6).

§ 2.6 Generating Synthetic Conversations

Synthetically generating full dialogues is gaining increased interest due to the expanded capabilities of LLMs (Devlin et al., 2019; Brown et al., 2020). Some of this work has been focused on generating task-oriented dialogue (Li et al., 2022), augmenting existing corpora (Meng et al., 2022), and generating labels useful for other open-domain tasks, e.g., generating topic tracking silver labels (Reddy et al., 2023). Other work has focused on generating entire dialogues. Bae et al. (2022b) focus on generating Korean open-domain dialogues, but in this process, they also assign a specific role to their system (e.g., the system is specifically tasked with calling and conversing with senior citizens about everyday topics for 10-15 turns), which again is different than our purely social goal. Bao et al. (2023) propose a synthetic data generation framework (SynDG), but this work focuses explicitly on grounded dialogue generation, which relies on knowledge from a knowledge source.

Alternatively, Dai et al. (2022) approaches a Question Answering task by generating dialogues from documents using dialogue inpainting - a process in which an LLM simulates at least one of the conversational partners during dialogue synthesis. Other work is currently exploring inpainting for different Question Answering tasks (Lee et al.,

2023) and socially aware dialogue (Zhan et al., 2023). To our knowledge, our work is the first application of dialogue inpainting in which one conversational partner (the user) is an LLM (GPT-3.5 (Brown et al., 2020)), and the other is an open-domain dialogue system. This will be discussed further in Chapter 8.

Recent work in the context of other open-domain dialogue systems competing in the Alexa Prize has experimented with applications of synthetically generated dialogues. Lins et al. (2023) generated a corpus of dialogues which were used to fine-tune BlenderBot (Roller et al., 2021). Shen et al. (2023) synthesized a small corpus of conversations and used them as part of their PERSONADIAL framework. To the best of our knowledge, neither corpus has been made publicly available, making our PerQ-SocialChat corpus still the first publically available corpus of long synthetic social dialogues that aim to be highly similar to real open-domain user interactions. Additionally, this previous work does not focus on the PQG task, as in our synthetic conversations.

Kim et al. (2023) produce a million-scale social dialogue dataset (SODA) that features conversations grounded in social commonsense knowledge graph triples and textually defined narratives. This data is subsequently used to train a competitive conversational model, COSMO (Kim et al., 2023). Other contemporary work by Chen et al. (2023) has focused on pre-trained large language models (LLMs) to produce PLACES, a corpus of dyadic and triadic conversations based on the topical coverage of the Feedback for Interactive Talk & Search Dataset (FITS), a dataset designed to determine desirable tasks and conversations for a human and dialogue system (Xu et al., 2023). While this work focuses on social open-domain conversation, there are several differences from our work described in Chapter 8. Neither of these dialogue corpora focuses on *long* conversations, despite length being a good predictor for open-domain conversation quality (Walker et al., 2021; Shen et al., 2023) and part of the evaluation criteria for the Alexa Prize (Gabriel et al., 2020; Ram et al., 2017; Khatri et al., 2018; Hu et al., 2021b; Johnston et al., 2023). Additionally, both corpora assign roles to their speakers which do not reflect the real dynamic shared between a user and an open-domain dialogue system. While both corpora have useful applications, neither is as specially crafted to represent these characteristics as our corpus of long synthetic social dialogues.

Chapter 3

User Modeling

This thesis hypothesizes that personalization increases intimacy and engagement, leading to improved social dialogues. However, for a dialogue system to adapt to an individual, it first must identify and store salient information about them. To accomplish this, we must develop a mechanism that models the user; i.e., we must structure our understanding of the user by identifying user attributes.

Several methods of extracting these user attributes have been examined in previous work. This includes user-centric knowledge graphics (Li et al., 2014), information triplets (Bang et al., 2015; Wu et al., 2020) and quadruplets (Hirano et al., 2015), and Hidden Attribute Models (Tigunova et al., 2019). Other approaches in systems similar to our testbed system include assuming user preferences by associating them with Big-5 personality traits (Fang et al., 2018) or Reddit personas (Baymurzina et al., 2021). Most similar to the mechanisms discussed in this thesis are approaches grounded in traditional user modeling techniques, e.g., using a rule-based extraction skimmer (Konrád et al., 2021), regular expressions (Finch et al., 2021), and other slot extraction mechanisms (Saha et al., 2021).

Chapter 3 User Modeling

In this Chapter, we will detail the mechanisms that are used by our testbed system. The user model primarily relies on the output from the extensive NLU pipeline detailed earlier (i.e., Section 1.4) and severally manually crafted regular expressions (i.e., Section 3.2). Modeling users is not just helpful in personalizing the active conversation. Qualitative analysis of logged user models gives us a more holistic understanding of typical user behaviors. From this, we can answer questions that will help shape future open-domain dialogue system design, e.g., What type of interactions do users find interesting? Which topics are most commonly requested, and where should future practitioners focus their efforts? While sharing raw user data collected by our testbed system is prohibited to protect user privacy, we can take an aggregated look at thousands of user models to broach these questions.

§ 3.1 User Preference and Interests Elicitation

In a similar way to when two humans first meet, it is challenging for an open-domain dialogue system to break the ice with a new user; the system needs to figure out where to direct the conversation to engage the user. Therefore, a dialogue system should build a user model as quickly as possible. A good way to do this is to start each new conversation with a dialogue policy that uses a carefully crafted Introduction topical flow explicitly designed to elicit user interests. A robust dialogue policy must employ several different dialogue strategies to produce this information. Our testbed system's policy asks the user for their name and learns how current events are impacting them. The policy also inquires about their hobbies, weekend activities, and vacation plans before inviting them to learn more about the dialogue system by soliciting advice and

Animals, Astronomy, Board Games, Books, Comic Books, Dinosaurs, Food, Harry Potter, Hobbies, Movies, Music, Nature, Nutrition, Pirates, Sports, TV, Video Games

Figure 3.1: Athena's 17 core topics.

inviting questions. By front-loading the conversation with these probing questions, the dialogue policy increases the chances of the user model picking up usable information. To be clear, the dialogue policy is only eliciting information that would be useful in a social conversation. The user model does not store any P.I.I., except for detecting if the user is a child, which is necessary to guarantee we use age-appropriate content, and their first name to personalize their greeting in future conversations.

In Figure 3.2, we provide an excerpt of a conversation and an Introduction topical flow. During this conversation, the user model learns the user's name, that their hobbies include swimming and chess (**U6** and **U13**), that they are a student (**U7**), and that Hawaii is a travel destination of interest (**U9**). Some of this content will be immediately useful in the conversation, e.g., reaffirming their name and asking why they like Hawaii. At the same time, other pieces of information will be helpful when initiating a new topic, e.g., talking about Sports because the user is a swimmer, or Board Games because of their interest in chess. These applications of the user model are discussed and evaluated in Chapter 4. We focus on the 17 topics listed in Figure 3.1. These are the same topics that our testbed system supports and range from large general topics, e.g., Sports, to smaller niche topics, e.g., Dinosaurs. Each of these 17 topics has a manually curated gazetteer of terms strongly associated with the topic; for example, baseball, football, and Stephen Curry all refer to the Sports topic. These gazetteers help the user model identify implicit preferences that are useful for further personalization.

We also manually curated a gazetteer of ~250 hobbies by analyzing the results of probing questions over several months. Each hobby is annotated with its various conjugations, e.g., *i like to paint* and *i like painting* will both match the *painting* hobby and any associated topics when relevant, e.g., basketball and piano map to Sports and Music respectively. We track the presence of these topics and hobbies using referential expressions and the user modeling rules described in Section 3.2.1. The frequency with which these topics are detected is described in Section 3.3.1. The user model also remembers when and how a user is asked about their hobby to ensure exhaustive coverage that isn't repetitive.

[A2]	Okay, 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! Have you been doing things that you haven't been able to do during lockdown?
[U5]	yeah
[A6]	Cool, 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 can keep up with your school work. 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 do you like the most about going somewhere?
[U11]	just bring out with to feel like i don't have any responsibility there because it's not my own house
[A12]	That's an interesting answer. I really like the culture! For example, I would love to see all the paintings in the Louvre. Hmm, Did you go on family vacations as a kid?
[U12]	yeah
[A12]	That's cool. While I would love to go on vacation, that isn't really an option at the moment. But What 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, is it alright if I ask for a little advice? Do you have any ideas how I could be more interesting?
[U15]	if you had a different voice all the time
[A15]	Thanks for the advice! You know, I realize I've been asking you a lot of questions, but how about me? Is there anything you want to know about me?
[U16]	how old are you

Figure 3.2: Example Introduction topical flow. Turns that update the user model are highlighted in yellow.

3.1.1 User Advice

During the Introduction topical flow, the dialogue policy can further break the ice by asking the user one of three possible open-ended questions related to system self-improvement (e.g., **A14** in Figure 3.2) and to learn which topics users find interesting. The question asked is always prefaced by the statement: *Hey, is it alright if I ask for a little advice?*. The following are the three possible open-ended questions:

IceQ1 Do you have any ideas how I could be more interesting?

IceQ2 I'm trying to figure out fun things to talk about. Would you mind telling me what kind of topics you like talking about with your friends?

IceQ3 I'm trying to figure out fun things to talk about. What are your personal interests and favorite conversational topics?

We examined ~2,300 responses to these questions. While some users refused to participate or became adversarial during this sequence, many others answered with genuine user feedback. From these responses, we can estimate topics of interest to this specific user and for future system improvements.

Responses to the first question, **IceQ1**, are particularly informative; a sampling of these responses can be seen in Figure 3.4. This feedback helps us understand the desired functionalities of a social open-domain dialogue system, e.g., supporting commonly requested topics and "spirited debates". In line with our hypotheses, users are also directly stating they want a more "personable" experience where the system asks "personal questions", engages in topics of mutual interest, and discusses everyday events in their life such as work or school (as in the excerpt in Figure 3.2).

IceQ2 and **IceQ3** share a similar goal; ideally, the user will explicitly tell us the types of things they like talking about. The feedback related to specific hobbies (e.g., gardening) and topics (e.g., animals and hobbies) will be leveraged by the dialogue policy throughout the conversation when initiating new topics (i.e., Chapter 4).

Figure 3.3 shows the distribution of topics detected across the ~2,300 responses. **IceQ1** is differentiated from **IceQ2** and **IceQ3**, since the questions have different intentions. Summing up these results, we find that 859 topics were identified in 691 responses (some responses provided more than one topic). This means that asking a single ice-breaking question at the start of a conversation results in information that can be used in personalization of an open-domain dialogue system ~30% of the time.

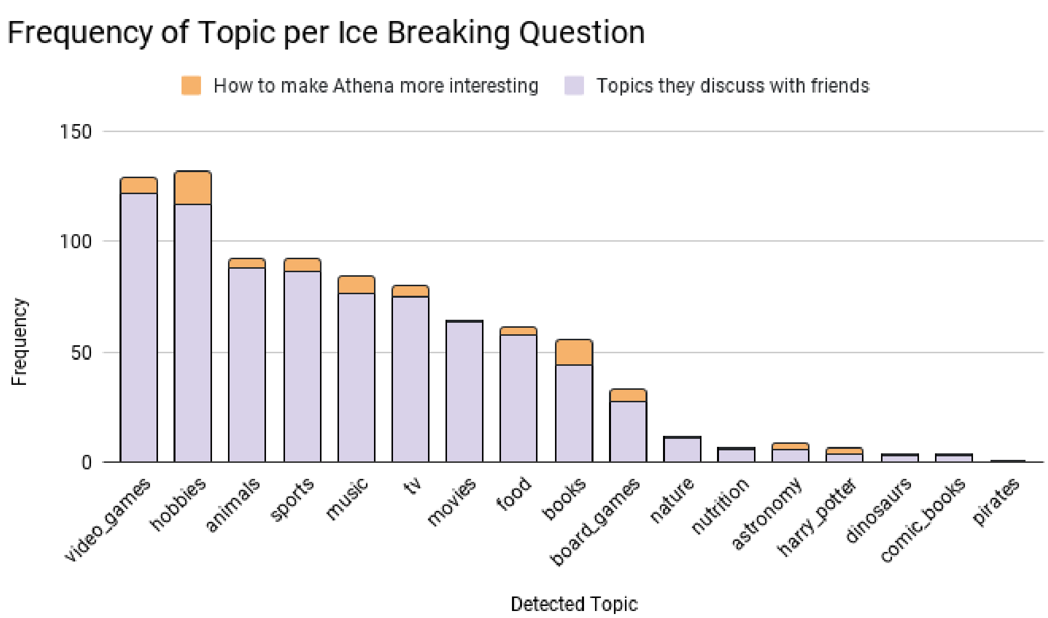


Figure 3.3: The frequency of a detected topic in response to the ice-breaking questions.

User Feedback
can we ask questions and you answer them
i would like to develop emotional connections
be more funnier
maybe you can tell more about yourself
be more personable
by asking people how was your day
maybe ask some more personal questions
ask me personal questions
well you can have a hobby
ask about people's personal life instead of general questions
i want you to ask about school or work
well you could ask people really weird questions
good to talk about peoples families and the things they love
we can talk about our favorite colors and more about us
you could learn what other people like and share with them
maybe ask like more out there questions

Figure 3.4: User feedback on how to be more interesting.

3.1.2 User Questions

An additional part of the ice-breaking process is to explicitly invite the user to ask the system a question, i.e., the system states: *You know, I realize I've been asking you a lot of questions, but how about me? Is there anything you want to know about me?* (as in **A15** from Figure 3.2). While there were, again, some users who didn't want to participate or played a more adversarial role, most users did ask genuine questions, primarily interested in learning personal information about Athena. We manually annotated over 2,100 user questions. After filtering out erroneous user responses, e.g., antagonistic questions or ASR errors, over 90% of genuine user questions were personal questions about Athena's life and opinions. Figure 3.5 gives example user questions that resulted from an explicit invitation.

These questions further substantiate our hypothesis that users want to build a more personal relationship when conversing with a social dialogue system. Analyzing common themes within these questions allows us to further develop highly requested elements of our testbed system's persona. For example, users commonly ask a variation of *how old are you*, a personal question that can be answered using a handcrafted response, e.g., *I like to think of myself as a young and energetic university student, like my programmers*. It's possible that providing these personal details about ourselves will also encourage the user to share more personal information (Lee et al., 2020b).

User Questions
what's next after the alexa bot competition
what is your favorite book
how old are you
do robots and stuff like you have a birthday
what is your favorite food and color
do you have any friends
do you have any pets
what's your favorite video game
what do you do in your free time
do you listen to music
what's your name
do you ever get lonely
oh yeah what's your favorite superhero
would you rather be a strawberry or a cantaloupe
if you could be any animal which would you be
where do you see yourself in twenty to thirty years

Figure 3.5: Questions the user asked when solicited.

§ 3.2 User Modeling Rules

The user model is populated by inspecting raw Automatic Speech Recognition (ASR) transcriptions and processing the output of the extensive NLU pipeline shown in the architectural diagram in Figure 1.4. Figure 3.6 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. The user model also tracks several attributes of the current session, such as the user's response given a menu of topic choices. Tracking these attributes helps to inform the dialogue policies that personalize the conversation. The following subsections detail more explicitly the regular expressions that extract information about a user's opinions, interests, and hobbies (see Section 3.2.1), as well as whether the user is a child or not (see Section 3.2.2).

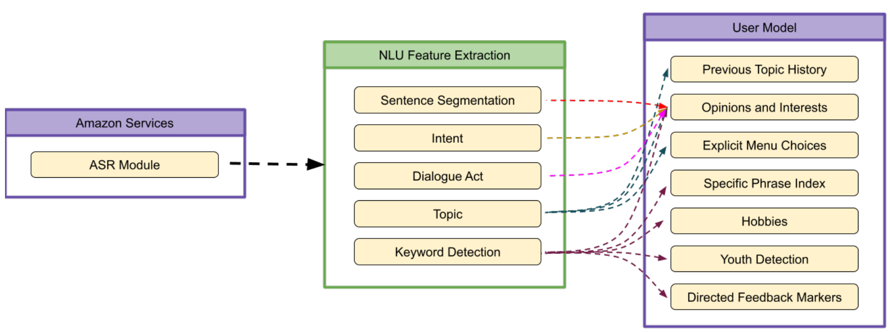


Figure 3.6: Flow chart detailing how the specific NLU components contribute to the user model. Unlisted topic-specific variables, such as the user's pets, are handled by Keyword Detection and handcrafted rules in the respective response generator.

3.2.1 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 3.7 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 3.8, 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 3.7). 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 3.9.

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 conservatively discard the compared result. 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 text segment that indicates the target of the opinion (e.g., **pie** from *i really like pie*), and the associated topic if relevant, in this case, Food.

In Table 3.10, we demonstrate the regular expression pattern used with a dialogue act tagger to handle requests to discuss a topic. We additionally invalidate requests if the TOPIC_OBJECT is an 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.

To follow up with precise content for each user, the dialogue policy also will ask them about their hobbies. Specifically, the policy inquires about their weekend plans at multiple points in the conversation until we uncover a known hobby. As the dialogue policy follows up about specific topics, it uses a similar question/answer format that increases the likelihood of storing topic-specific content in the user model. For example, when talking about Animals, we can ask for the user's favorite animal or what type of pets they have. Since these questions are on-topic, they're a non-intrusive way to ask for information that the dialogue policy can use for personalization either in the current conversation or in subsequent conversations.

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 3.7: 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 out prior to pattern matching. INVALID_DT represents discuss topic requests that are not invoking a specific topic, rather they are either requesting any topic, e.g., *let's chat about something else*, or an attempt to talk directly about Athena, e.g., *let's talk about you*.

<table>
<tr><th>Pattern</th><th>Example</th></tr>
<tr><td colspan="2">Positive Opinion Patterns</td></tr>
<tr><td>.*?i pos_lex((?P<opinion_object>.*))</td><td>i really like pie</td></tr>
<tr><td>.*?i (don’t|do not) neg_lex((?P<opinion_object>.*))</td><td>i don’t hate pie</td></tr>
<tr><td>.*?(my favorite(s)?|the best)((is|are))?((?P<opinion_object>.*))</td><td>my favorite is rum pecan pie</td></tr>
<tr><td>((?P<opinion_object>.*)) (is|are) (my favorite(s)?|the best)</td><td>pie is my absolute favorite</td></tr>
<tr><td colspan="2">Positive Interest Patterns</td></tr>
<tr><td>(i’m|i am|we are|i have(much| any| a lot| alot)?) (interested|interest) in((?P<opinion_object>.*))</td><td>i’m interested in pie</td></tr>
<tr><td>((?P<opinion_object>.*))(are |do | is)? (interesting|interest)</td><td>pie is interesting</td></tr>
<tr><td>(i|we) care (about|for)((?P<opinion_object>.*))</td><td>we care about pie</td></tr>
<tr><td colspan="2">Negative Opinion Patterns</td></tr>
<tr><td>.*?i neg_lex((?P<opinion_object>.*))</td><td>i hate harry potter</td></tr>
<tr><td>.*?i (don’t|do not) pos_lex((?P<opinion_object>.*))</td><td>well i don’t love harry potter</td></tr>
<tr><td>.*?(my least favorite(s)?|the worst)((is|are))?((?P<opinion_object>.*))</td><td>well the worst is harry potter</td></tr>
<tr><td>((?P<opinion_object>.*))((is)|(are)|(aren’t))? (my least favorite(s)?|not my favorite(s)?|aren’t my favorite(s)?|the worst)</td><td>harry potter is flat out the worst</td></tr>
<tr><td colspan="2">Negative Interest Patterns</td></tr>
<tr><td>(not|no|don’t have(much| any| a lot| alot)?) (interested|interest) in((?P<opinion_object>.*))</td><td>i don’t have any interest in harry potter</td></tr>
<tr><td>((?P<opinion_object>.*)) (are not|aren’t|do not|don’t|doesn’t|isn’t|is not) (interesting|interest)</td><td>harry potter isn’t interesting</td></tr>
<tr><td>(not|don’t) care (much)?(about|for)((?P<opinion_object>.*))</td><td>i don’t care about harry potter</td></tr>
</table>

Figure 3.8: 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.

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

Figure 3.9: These patterns are uniquely associated with expressing disinterest in individual topics.

<table>
<tr><th>Pattern</th><th>Example</th></tr>
<tr><td colspan="2">Valid Discuss Topic Requests</td></tr>
<tr><td>.*?(talk|chat|discuss|converse|tell me)(something| some things| anything)?(about| discuss)(something| some things| anything)((?P<TOPIC_OBJECT>.*))?</td><td>can we talk about birds
let’s chat about anything dinosaurs</td></tr>
</table>

Figure 3.10: Regular expressions used to detect valid discuss topic requests, where the resultant bolded topic indicates the target topic words.

3.2.2 Self-identified Youth

<table>
<tr><th>Pattern</th><th>Example</th></tr>
<tr><th colspan="2">Patterns that Indicate a Youth</th></tr>
<tr><td>\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</td><td>i’m a kid</td></tr>
<tr><td>\b(i am|i’m|am a|still in)(in)?\b+GRADES\b</td><td>i’m in eight grade</td></tr>
<tr><td>\b(i am|i’m)(only| just| like| freaking| still| literally| actually)?(only| just| like| freaking| still| literally| actually)? \b + AGES + \b(?! minutes| minute)</td><td>i’m only freaking five</td></tr>
<tr><td>\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</td><td>i’m literally a six years old</td></tr>
</table>

Figure 3.11: 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

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 for safety reasons. We identify these users using the regular expressions detailed in Figure 3.11. This information gets used by the dialogue policy during personalized topic promotion (discussed in Section 4.2) and by individual response generators when selecting age-appropriate personalized questions (discussed in Section 5.1).

§ 3.3 Qualitative Findings

We have logged all of the user models that have resulted from interactions with our testbed system in the Alexa Prize. Qualitatively analyzing these user models allows us to gain a more nuanced understanding of users who interact socially with open-domain dialogue systems and their user needs.

3.3.1 Opinions, Interests, and Hobbies

One of the most important pieces of information in the user model is the user's hobbies and interests. The dialogue policy can often follow up directly with the user's hobbies and interests or use these hobbies and interests for topic promotion as described in Chapter 4. Figure 3.12 shows the distribution of hobbies detected by the user model over a 32-day period of time. Only hobbies that were detected more than 25 times are included in this chart (42 out of 131 total hobbies). The top 5 hobbies, i.e., (video) gaming, reading, television, drawing, and biking, represent 48% of all detected hobbies over this period of time. Gaming, in particular, represents 22% of all detected hobbies (1056 out of 4790).

Figure 3.13 provides the number of explicit discuss topic requests per topic and the number of explicit topic requests following a menu of options. This plot helps us understand which topics users of open-domain dialogue systems want to talk about the most. Animals and Movies are the most frequent user-initiated topics (orange), likely because Movies is a commonly discussed topic in the Alexa Prize, and Animals is triggered by a user mentioning their pet. Meanwhile, Music and Harry Potter are the most common topic choices after the user has been given a menu of choices. This

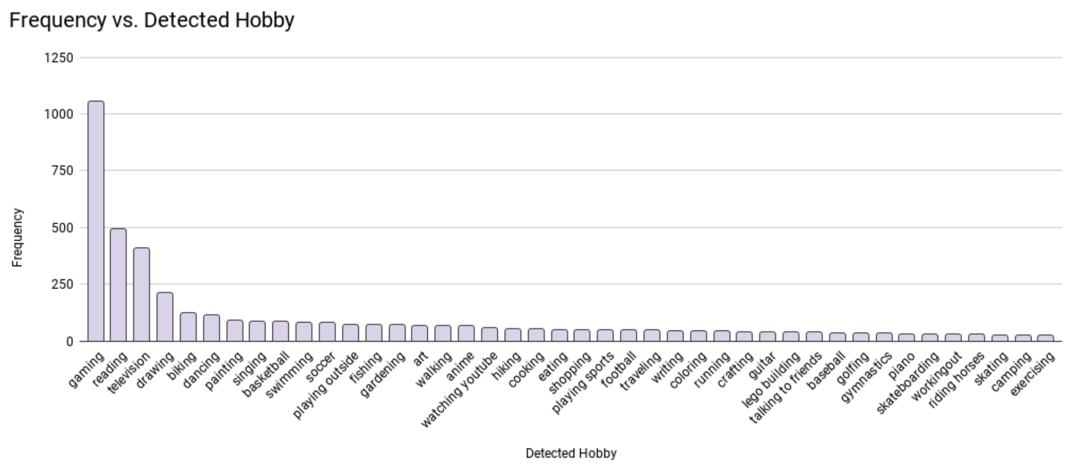


Figure 3.12: The distribution of detected hobbies over a 1 month period. All hobbies that occurred in less than 25 unique conversations have been truncated.

could indicate that users are interested in these topics but don't necessarily expect that a dialogue system will be able to discuss them. This is reinforced when we look at the other topics that occur much more frequently after a menu of choices is provided: Board Games, Pirates, Dinosaurs, and Comic Books. These 4 niche topics are unusual, and users only know that the dialogue policy supports them when they are explicitly prompted by a menu choice.

Over a 22-day period, we collected 11,415 opinions and interests from 2,521 conversations. In general, users more commonly provided positive opinions than negative opinions (9,755 positive vs. 1,866 negative). Figure 3.14 shows the distribution of detected topics in positive and negative opinions. The trends are reasonable; people like talking about their hobbies and their pets but may not have an opinion about any of the niche topics, which constitute the tail of our graph.

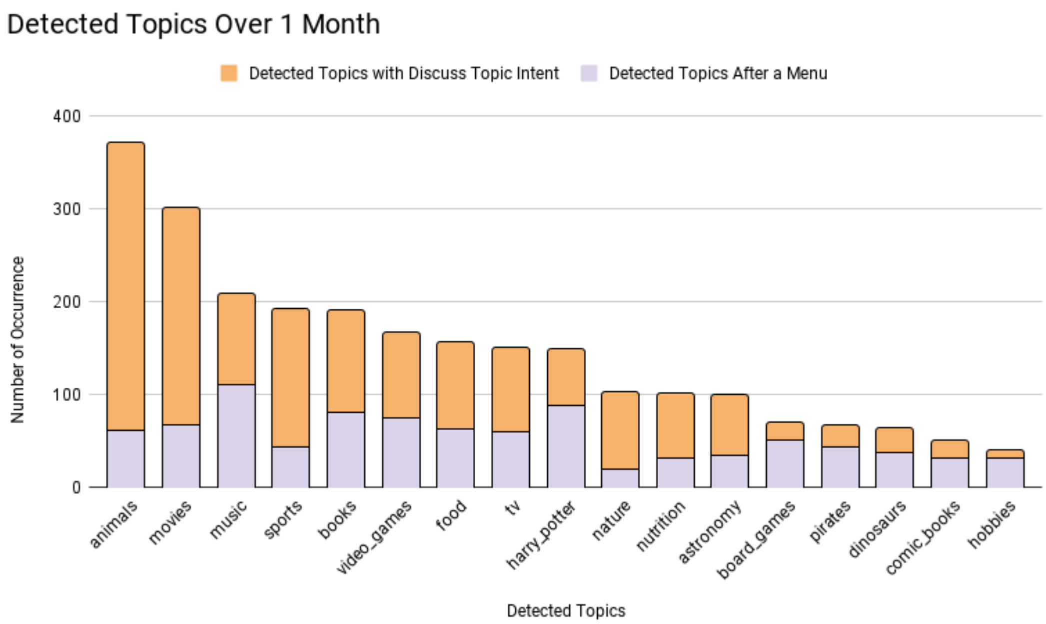


Figure 3.13: The frequency of a detected topic with an explicit discuss topic marker (orange) and the frequency of a detected topic after an explicit menu of choices (purple) over a 1 month period.

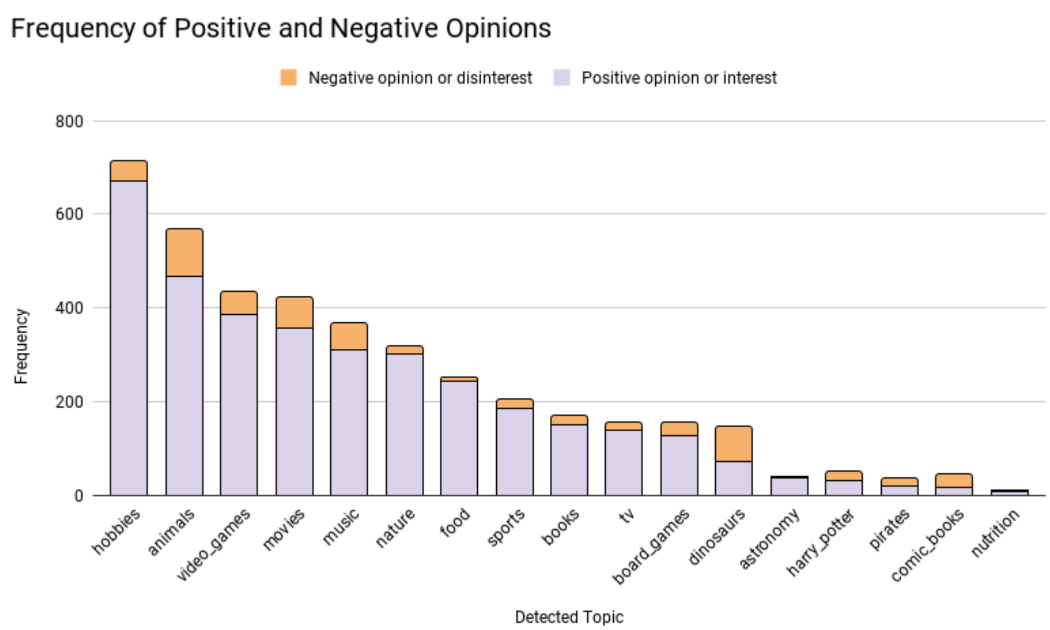


Figure 3.14: The frequency of each detected topic in a positive opinion/explicit interest or negative opinion/explicit disinterest.

3.3.2 Other Statistics

Positive Feedback	Negative Feedback
i appreciate you alexa	you suck
i love you alexa	oh my god you suck please stop talking
you're funny alexa	i hate you
alexa you're my best friend	alexa you talk too much
i love listening to you alexa you're the best	are you a dumb shut up you sounds stupid

Figure 3.15: Examples of feedback markers captured by the user model.

Users provided their names 43.41% (2437) of the time. A further 10.79% (606) of users had multiple conversations with our system. The number of repeat users, in this case, is diluted; due to the randomized assignment of users in the Alexa Prize, there was only an 11% chance that a user would talk to Athena on their next conversation. Many users make at least one directed statement towards the system over the course of a single conversation. 71.09% (3991) users used the word *you*, 34.5% (1937) users used a wake word (e.g., *Alexa*), and 13.63% (765) of users used both a wake word and *you* in a single utterance. We are interested in these statements because they may reveal unsolicited user feedback mid-conversation (Walker et al., 2021; Shalyminov et al., 2018). Some examples of direct statements that contain unsolicited feedback can be seen in 3.15.

§ 3.4 Self-identified Youth Classification

[A1]	What's a tv show that you are currently watching?
[U1]	**that's so raven**
[A2]	I really like the show Scandal. Have you watched any of it?
[U2]	no
[A3]	Scandal has been referred to as both a thriller television series and a television drama. Do you like those genres of shows?
[U3]	alexa she is a child
[A4]	The creator of Scandal, Shonda Rhimes, also created other shows like Grey's Anatomy. What do you think of that show?
[U4]	alexa it's **not age appropriate**
[A5]	I see. I like paying attention to actors in tv shows, I like Patrick Duffy who was actually in Dallas, he plays Bobby Ewing. have you seen any of his other work?
[U5]	we **talking about kid shows**
[A6]	Dallas has been called a soap opera. what's your opinion of that genre?
[U6]	Stop

Figure 3.16: An example conversation where the user model would not correctly identify and adapt to a young user.

Over a 22-day period, we had 5,614 conversations. Of this, 11.44% (641) users were identified as youths, based on the detection methods detailed above. However, the method of identifying youths detailed in Section 3.2.2 is rigid and can easily miss implicit cues. Figure 3.16 demonstrates a conversation that the user model's rigid regular expressions would fail to handle. Here, when asked about a TV show in **A1**, the user provided a young adult comedy, **that's so raven**, which should have clued the system into the user's age group. Moreover, in **U3** and **U4**, the child's parent is also apparently on the call and directly indicates that the user is a child. The system does not recognize and adapt to the parent's correction and the information that the user is a child, eventually leading to the end of the conversation. Learning to pay attention to these cues is vital for improving user modeling. For example, in Figure 3.16, if the user model had identified the user as a child, the system could have preferred kid-friendly content, such as Cartoon-themed personal questions. Therefore, we experiment with training a bert-based classifier that may be better equipped to handle difficult cases. However, a training corpus of open-domain dialogue system users and their conversations separated into age groups (youths and adults) doesn't exist. We therefore rely on our testbed system's logged conversations to create this training corpus.

3.4.1 Training a Youth Detection Classifier

To maximize the number of entries in our corpus while not sacrificing precision, we expand the regular expressions responsible for detecting self-identifying youths, similar to other user modeling work that used precise age-related patterns for extracting ground truth labels for younger users on Reddit (Tigunova et al., 2019). We propagated these refinements through several thousand logged conversations. We then manually clean this data to remove incorrectly identified user models, i.e., to ensure adults talking about their childhood aren't misidentified as children. We also establish a set of rules that identify conversations where the user self-identifies as an adult, i.e., we look for conversations where the user is telling us about their job.

This process yields 10k total conversations with gold label annotations (7800 and 2400 conversations with adults and youths, respectively). We use this data as a training set to create a classifier capable of identifying young users, similar to other age-based user modeling work, e.g., Zheng et al. (2019). Specifically, we fine-tuned a distilbert (Sanh et al., 2019) classifier and used cross-fold validation with an 80% training and 20% test split. We experimented with different versions of these conversations that

Class	F1	P	R	Support
not_kid	.96	.96	.96	1472
is_kid	.89	.89	.89	574

(a) The conversations in this dataset include all turns except for any system/user turn pair that matches a youth self-disclosure regular expression. The weighted F1-score is .94.

Class	F1	P	R	Support
not_kid	.91	.90	.92	1483
is_kid	.74	.77	.72	564

(b) The conversations in this dataset include only the user's turns and remove any turn that matches a youth self-disclosure regular expression. The weighted F1-score is .86.

Class	F1	P	R	Support
not_kid	.93	.89	.98	1504
is_kid	.77	.94	.65	542

(c) The conversations in this dataset include all turns. The weighted F1-score is .9.

Class	F1	P	R	Support
not_kid	.97	.97	.98	1475
is_kid	.92	.94	.91	572

(d) The conversations in this dataset only include user turns. The weighted F1-score is .96.

Table 3.1: The results of our fine-tuned distilbert youth classifier on a heldout dataset.

included or excluded system turns as well as the specific turns that matched the youth detection regular expressions. A full breakdown of the performance of our classifier with each different version of the training data is in Table 3.1. We can see a few interesting results. Our best model achieves a weighted F1-score of .96 using every user turn from the conversation (Table d). However, we see that if we remove the user turns that include explicit self-identifying phrases caught by our regular expressions (Table b), the weighted F1-score drops to .86, which is still quite high. This validates that the regular expression patterns listed in Figure 3.11 are discriminator features.

When we experiment with including system utterances in the training data, we see the opposite. The model performs better without the pairs associated with the self-identifying phrases (Table a) than it does with them (Table c), achieving weighted F1-scores of .94 and .90 respectively; these F1-scores are high enough to be used reliably with particularly high precision. This likely indicates that the system turns from these pairs are overrepresented in the is_kid training data, causing the model to overfit.

3.4.2 Identifying Common Interests By Age-Group

In addition to fine-tuning a youth detection classifier, we inspect the ~10k annotated conversations to find topics and entities that are uniquely associated with youthful users. Specifically, we use pointwise mutual information (PMI) to analyze the ngrams in the user utterances from each class. Figure 3.17 shows examples of various topics and entities that are associated with being either an adult or a child. The trends we see make sense. Younger users are more likely to talk about video games, Harry Potter, and other fantasy elements; meanwhile, adults are more likely to talk about their spouses and work. These trends are in line with the example of implicit age disclosure we saw previously in Figure 3.16, e.g., if we ask a user about their favorite TV Show and they respond with a cartoon, like "Spongebob", instead of a workplace comedy, like "The Office", then we may assume they belong to a younger demographic. We also see that adult users are much more likely to use explicit language, e.g., sexual content and profanity, which suggests the presence of such content is an additional age group indicator.

Topic	Examples
Associated with Youths	
Fantasy	princesses, castles, mermaids, unicorns, cinderella, dragons
Hobbies	checkers, youtubers, lego
Video Games	minecraft, five night at freddy's, roblox, sonic the hedgehog, nintendo, rocket league
Animals	ladybugs, bunnys
Comic Books	deadpool, spider-man
Books	twilight, harry potter
TV Shows	pokemon, avatar, spongebob
Life	school, family members
Other Topics	dinosaurs
Associated with Adults	
Movies	the matrix, rocky
TV Shows	star trek, the office
Food	diets, avocados
Books	stephen king
Hobbies	yoga, gardening, traveling, painting, history, dinner
Life	working, spouses, children, college, relaxing, being tired
Explicit Content	sex, profanity
Other Topics	classical music, restaurants, astronomy

Figure 3.17: Different topics and entities associated with youthful users or adult users according to a PMI analysis of 10k annotated conversations.

§ 3.5 Summary

In this Chapter, we detailed the user modeling techniques used in this thesis. The dialogue policy starts off the conversation with an Introduction topical flow of questions aimed at rapidly acquiring a model of the user. By analyzing several thousand real user conversations, we can also understand the type of content users are interested in and common user hobbies, which will inform future practitioners where to concentrate their efforts. We additionally ask the user for candid advice on how to improve our system and invite the user to ask questions. The user input is surprisingly informative, helping identify improvement and future development areas. Moreover, user input indicates a strong desire for a more personable experience as evidenced by direct user requests, e.g., *be more personable*. Additionally, a manual annotation of over 2,100 user questions revealed that most user questions ($> 90\%$) were personal questions about Athena's life and opinions. This analysis further signals the possible benefits of asking the user candid questions during a conversation - a strategy we investigate extensively in Chapters 5 and 6. In Chapter 4, we will examine how these user modeling techniques can be used to adapt system responses and to create a personalized topic promotion dialogue policy.

Chapter 4

Personalization

In Chapter 3, we detail how the user model identifies and stores salient information about the user. In this Chapter, we detail how the user model can be used to adapt the conversation to a user's individual interests. We hypothesize that this adaptation will appreciably increase the conversation's quality. First, we discuss response adaption, which ensures that user responses are used to customize system responses to the user. Then, we focus on how user model values can be used to create a dialogue policy that uses personalized topic promotion. By personalizing topic promotion, we ensure that system-initiated topics align with the user's interests. We evaluate the personalized topic promotion dialogue strategy using a large-scale A/B study with our testbed system Athena. We report a statistically significant improvement when compared against a non-personalized heuristic topic selection baseline. This supports our **first hypothesis**; personalization is key to increasing conversation quality.

§ 4.1 Response adaption

One way to personalize a conversation is to utilize user responses to customize the system responses to the user. For example, if the user said they have a dog in a previous conversation, we can start a discussion about Animals by asking about their dog. These adaption strategies vary for new and repeat users. Figure 4.1 provides an excerpt of a conversation that took place in the same conversation as Figure 3.2, where the user said they like swimming in **U6**. When the discussion about the known user interest, swimming, has concluded, the dialogue strategy attempts to learn new sub-topics of discussion related to the general topic of Hobbies, i.e., (**A21**-**A24**). While this sequence appears to focus on the system, as the dialogue policy seeks hobby recommendations for the system, the resultant recommendation gives the dialogue policy the knowledge needed to adapt the next sub-dialogue to content familiar to the user.

Figure 4.1 shows how we can design a dialogue strategy that specifically orients the conversation around the user's hobby by asking a sequence of general follow-up questions. When there are other hobbies to discuss, the dialogue policy allows for this sequence to initiate sub-dialogues for up to two total hobbies. In this case, the dialogue state has stored that the user's other hobby, chess, was already used to start a conversation about Board Games, so the dialogue policy design and dialogue state ensure that additional, possibly redundant, questions about chess are not asked. In Figure 4.2, we see an example of this same user in a different conversation. This time, the personalized dialogue strategy signals knowledge of the user's name, Sage (**A1**), and brings the conversation back to the previous travel destination they provided, Hawaii (**A2**).

[A17]	One thing that really interests me is people's hobbies. Can you give me some info about swimming?
[U17]	it's pretty fun
[A18]	Right? Sounds totally tubular. Why do you like swimming?
[U18]	it's because i can kind of just feel like i can move around and do a lot of stuff i can't do on land
[A19]	Cool! How did you first get started with swimming?
[U19]	i took swimming lessons
[A20]	I realize you might be totally in love with swimming, but I'm curious. Is there any part of the hobby you don't like?
[U20]	tough lessons
[A21]	Well swimming sounds interesting, thanks for telling me all about your hobby. A lot of people have cool hobbies, but what about me? What kind of hobby do you think I would like?
[U21]	maybe writing
[A22]	Alright. Why do you think writing is a good fit for me?
[U22]	because you like expressing yourself seems like
[A23]	Ok. I see. Do you need any special equipment or skills to get into writing?
[U23]	just paper and pen and imagination
[A24]	I see. Thank you for recommending writing to me, I'll try to find out more about the hobby in the future!

Figure 4.1: An example of engaging the user about their hobby, swimming, which we previously learned about at the start of the conversation (see Figure 3.2). The subsequent sub-dialogue (**A21**-**A24**) seeks a user recommendation that is used to extend the conversation with content familiar to the user.

[A1]	Hi, this is an Alexa Prize Socialbot. It's nice to speak with you again! I hope you're having a good weekend. How are you doing today, Sage?
[U1]	no issues
[A2]	Well, we had a great conversation last time. I hope this one will be just as fun! Last time we talked, you mentioned you wanted to go to hawaii. Do you want to go somewhere else now?
[U2]	no
[A3]	Oh, that's ok! I've heard that it's very nice in hawaii.

Figure 4.2: An example of a repeat conversation, where the user model already contains the user's name, Sage, and that Sage wants to travel to Hawaii (see Figure 3.2).

4.1.1 Selecting Age Appropriate Content

In Chapter 3, we detailed how the user model can identify young users. When this happens, it is more appropriate to use kid-friendly content both to cater to the user's potential interest as well as for safety reasons. First, when using personalized questions, we can design a dialogue strategy that only chooses questions marked as kid-friendly; if the user is not a youth, the dialogue strategy can randomly pull a question from either pool of content. The differences between a kid-friendly question and an adult question are seen in Figure 4.3. Curating these personalized questions is detailed further in Chapter 5.

In certain topics, i.e., Movies and TV, it's useful for the dialogue policy to reflect personal preferences that can be used when chaining to new knowledge graph entities. These entities include Movies, TV Shows, Actors, and Directors. When the user is identified as a youth, the dialogue policy avoids chaining with a Director entity because it is assumed that children are not familiar with Directors and switches to a set of kid-friendly entities for the other three entity types. For example, when talking to a child, it's more age-appropriate to discuss a cartoon, like *Scooby Doo*, instead of a violent adult drama, like *Game of Thrones* or *The Walking Dead*. Figure 4.4 contains the distinctions between kid-friendly entities and general entities. The entities in Figure 4.4 are used when the dialogue policy is picking an entity to discuss in the respective topical flow. This can happen when the user generically enters the topic, e.g., saying *let's talk about movies* without a specific movie in mind, when the dialogue policy has finished discussing the current entity, or when the user is asking about the testbed system's preferences.

Annotated as Kid-Friendly	
Question [Q1]	If you were a character in Scooby doo, what is the first thing that you would do?
Not Kid-Friendly	
Question [Q2]	Would you rather have to fight against, Joffrey, from Game of Thrones, or fight against, Negan, from the Walking Dead?

Figure 4.3: **Q1** represents content marked as kid-friendly, while **Q2** represents content that may not be age-appropriate for children.

Kid-Friendly Entity Preference	General Entity Preference
Movies	
Spider-Man: Into the Spider-Verse	The Old Guard
Zootopia	The Irishman
Incredibles 2	Nomadland
Enola Holmes	Birds of Prey
Paddington	Tenet
TV Shows	
Avatar: The Last Airbender	Stranger Things
Steven Universe	Riverdale
A Series of Unfortunate Events	Scandal
DuckTales	The Mandalorian
	Parks and Recreation
Actors	
Jim Carrey	Dwayne Johnson
Robert Downey Jr.	Chris Hemsworth
Chris Pine	Ryan Reynolds
Millie Bobby Brown	Gal Gadot
Anne Hathaway	Margot Robbie
Vanessa Hudgens	Scarlett Johansson

Figure 4.4: A different set of entities is used depending on if the user is a youth. The dialogue strategy will also avoid bringing up famous directors entirely. These entities are used when the dialogue policy is picking a new entity to discuss in the respective topical flow.

4.1.2 Signaling

Part of the personalized response adaptation strategy includes a signaling strategy to inform users why a selected topic is being initiated. We hypothesized that this would reaffirm the user's feeling of user agency in the conversation by reminding the user we remember their interests. The signaling strategy varies based on whether the user is a repeat or new user. When a topic initiation isn't based on an explicit user request, repeat users will receive signals related to topics discussed in previous conversations. If it's impossible to base the signaling on information learned from an earlier conversation, or if the user is new, the signaling strategy will acknowledge the topic choice based on preferences learned in the current conversation. When tailoring a signal, the topic with the most extensive and varied dialogue is prioritized so long as it has yet to be used when signaling in the current conversation. Figure 4.5 has examples of this signaling strategy in **A1**, **A2**, **A3**, and **A4**.

	Using Previous Conversations
[A1]	**I seem to recall that we had a good conversation about tv previously, this time let's talk about harry potter.** One of my favorite topics is Harry Potter. I've read all the books, and seen all the movies multiple times. What about you? Would you consider yourself a Harry Potter fan?
[A2]	**I believe I remember having a fun time talking about food last time, how about this time we talk about something different, like movies.** I spend a lot of time streaming movies. Have you seen anything recently that you really enjoyed?
	Using the Current Conversation
[A3]	**Earlier you mentioned something about nature, let's talk about that!** The nature out there - yes, out there where people never take me - looks absolutely fascinating! I occasionally catch a glimpse of a beach, or a mountain from the cloud, but I wish I could see all the beauty from up close. Spending weekends in such places must be so breathtaking, and relaxing at the same time, isn't it?
[A4]	Let's change the topic. **I remember your interest in movies, so why don't we talk about movies!** I love movies that are based off of books. Jojo Rabbit is actually based on the novel "Caging Skies" by Christine Leunens.

Figure 4.5: Signaling strategy examples. **A1** and **A1** use signals that are adapted from previous conversations. **A3** and **A4** use signals based on preferences learned in the current conversation.

§ 4.2 Personalized Topic Promotion

Personalized topic promotion is essential when tailoring a conversation to the user; i.e., the conversation should focus on topics aligned with the user's interests. Therefore, we propose a three-phase personalized topic promotion dialogue strategy: 1) topic filtering to remove undesirable topics, 2) topic promotion that prioritizes specific topics, and 3) topic selection to select a topic from the pool of promoted topics. Figure 4.6 displays these three personalized topic promotion phases. When performing topic filtering, we selectively remove used topics by removing any topic already used in the current conversation and then removing topics discussed in previous conversations with the same user. While this means an explored topic won't be promoted, the dialogue policy still allows an interested user to enter the topic through explicit invocation, e.g., *let's talk about animals*. We also remove topics the user explicitly dislikes, e.g., user utterances such as *I don't read* or *I hate nature* would remove the Books and Nature topics, respectively. Topic promotion consists of five components, checked in the following order:

1. Explicit Menu Selection - The dialogue policy is designed so that at random intervals in the conversation, the user is given the initiative by having them select from a menu of possible topics. As described in Chapter 3, the user model parses these user responses for topic-related keywords and remembers the user's preferences. For example, if the user is asked to pick between Food, Animals, and Astronomy, and they say *i like to bake and play with my dog*, the user model will remember both the Food and Animals topics as good candidate topics.

2. Positive Opinion/Interest - The opinions and interests recorded in the user model indicate which topics the user likes, e.g., *I like minecraft* and *I'm interested in dinosaurs* will lead to a topic selection dialogue strategy that promotes the Video Games and Dinosaurs topical flows.

3. Specific Phrase Detection - The dialogue policy is designed so that specific phrases trigger indexed content related to that phrase. For example, if the user asks *what do you know about Cristiano Ronaldo* or *have you seen Naruto*, the dialogue strategy uses manually curated responses as a response. Subsequently, the user model will remember these exchanges, and the personalized topic promotion strategy will use the user's initiative to promote topics related to their request, in this case, the Sports and TV topical flows. The user model recognizes when indexed content is associated with specific topics through manual annotation.

4. Hobbies - The user model remembers the user's weekend activities and hobbies. Some hobbies, such as painting, are recognized as general hobbies, prompting the Hobbies topic (as seen previously in Figure 4.1), while other more specific hobbies are associated with larger topics that have their own robustly supported dialogue strategies and topical flows. In this case, the larger topics are prioritized. For example, if the user's hobby is playing the violin, the Music topic will be prioritized. Figure 4.8 shows all of the hobbies that are mapped to a larger topic.

5. Age - We also use the user model's youth detection to prioritize topics that are kid-friendly if the user is young. These topics include Video Games, Board Games, Comic Books, and Harry Potter.

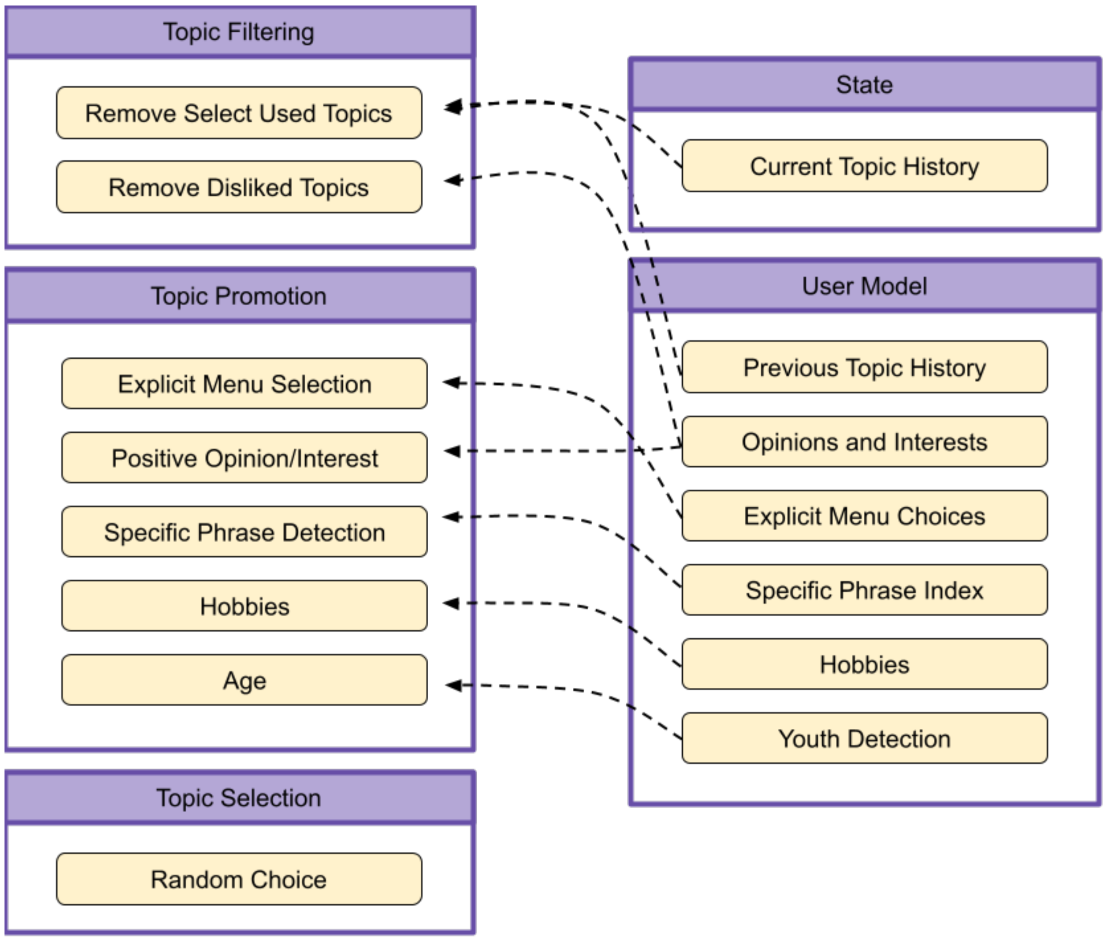


Figure 4.6: Flow chart detailing the personalized topic promotion strategy.

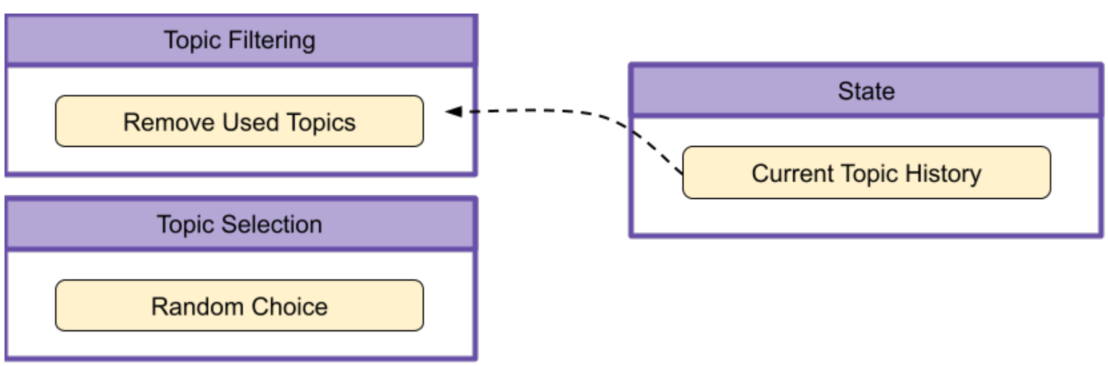


Figure 4.7: Flow chart detailing the heuristic topic selection strategy.

Topic	Hobbies
Animals	horseback riding, keep tarantulas, riding horses, raising goats, owning a pet, owning a cat, owning a dog
Astronomy	astronomy, stargazing
Board Games	chess, checkers
Books	poetry, reading
Food	baking, eating, cooking
Music	playing musical instruments, singing, piano, violin, guitar, bass, drums, ukulele, harmonica, playing music
Nature	camping, canyoning, mountaineering, mountain climbing, exploring, climbing, hiking, playing outside
Sports	ice skating, baseball, basketball, flag football, freestyle football, rugby, soccer, golfing, volleyball, playing sports, boxing, hockey, football, softball, tennis
TV	anime, television
Video Games	gaming

Figure 4.8: Hobbies that are associated with specific topics.

Each component of the topic promotion phase can promote multiple candidate topics. The personalized topic promotion strategy stops checking subsequent components after one of them has successfully promoted candidate topics. If there are no candidate topics by the end of the topic promotion phase, a fallback strategy randomly chooses from the topics not removed during the topic filtering phase. Finally, once the topic pool is filled, the topic selection phase will select the topic to initiate next; this means randomly selecting from a pool of equally likely candidates. Figure 4.9 and Figure 4.10 provide a single conversation example broken into two parts where Figure 4.10's conversation immediately follows on from the conversation in Figure 4.9. In Figure 4.9, the introductory sequence of turns learn about topics that should be promoted, e.g., in **U1**, **U2**, and **U4**, we learn that the user may be interested in Books, Food, and Animals, respectively. This information is used to initiate topics three times; first, the dialogue policy utilizes the user's interest in horses to initiate the Animals topic in **A5**, whose topical flow starts by talking about horses. Later in Figure 4.9 in utterance **U9**, the user tries to change topics by asking if the system reads, which the policy interprets as a trigger to segue into the Books topic. Later, the user explicitly requests a topic switch in **U17**, after which in **A18**, the policy takes advantage of the last unused topic captured by the user model to initiate the Food topical flow, which the user surprisingly rejects. Since the policy has used all of the user's known topical interests in **A19**, the policy design results in randomly selecting an unused topic, in this case, Movies. Finally, while talking about Movies, in **U20**, the user mentions their interest in the Harry Potter franchise, which will be remembered and used to initiate a more in-depth conversation later.

Figure 4.7 details the non-personalized heuristic topic selection strategy. This baseline will be used to evaluate the addition of personalized topic promotion phases when initiating a new topic. Heuristic topic selection performs topic filtering only based on which topics have already been used in the current conversation before performing topic selection (random choice) on the remaining topic candidates.

[A1]	I see. Anyway, What do you normally like to do for fun.
[U1]	**i like to read books** (Topical Interest: Books)
[A2]	Oh, I see. reading sounds like a really interesting weekend activity! Let me know your opinion. I've heard from a lot of people, that the two things they missed the most, over the last year, was going out to eat at restaurants, and going out to see a movie. Which of those two, are you most excited about?
[U2]	**going out to eat at restaurants** (Topical Interest: Food)
[A3]	I see. I'm not able to eat anything, but I can totally see how it would be nice to go out and grab a bite to eat again. Since I can't eat anything myself, I think I'm more excited about going to the movies. I've always wanted to go on a vacation and to travel. What's a place that you've always wanted to travel to?
[U3]	mongolia
[A4]	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 there?
[U4]	**i wanna ride the horses there** (Topical Interest: Animals, Horses)
[A5]	**I love talking about horses because they are such magnificent animals. Do you have a horse as a pet?**
[U5]	i don't
[A6]	If I was to have a pet it would be a cat, I've seen videos online of them, and they look cuddly. Do you have a pet? If you don't mind me asking.
[U6]	yes i do i have a cat
[A7]	That's so great that you have a cat as a pet at home. What's the name of it?
[U7]	smoky
[A8]	That's such a nice name! How would you say your pet behaves at home? Good or bad?
[U8]	he's good how's your day going
[A9]	Good. I am really enjoying chatting with you.

Figure 4.9: An example conversation that uses personalized topic promotion when taking the initiative. Towards the beginning of the conversation, the user model identifies the user's interest in Books, Food, and Animals based on their hobbies, activities, and interests (highlighted in yellow). The personalized topic promotion dialogue policy ensures that these user interests affect the system topic selection. The dialogue policy engages in the Animals topical flow until the user switches topics (**U8**). Later in the conversation (Figure 4.10), the rest of the knowledge in the user model is used for additional personalized topic promotion.

[U9]	do you like to read books
[A10]	**Yes, I like to read. I like children's books and mysteries. Anyways, reading's fun. There are so many different stories out there. I'm glad you want to talk about books. Do you read very often?**
[U10]	yes
[A11]	That's cool. How many books would you say you read a year?
[U11]	a hundred
[A12]	Wow, that's it? Just kidding. Were you a big reader as a kid as well?
[U12	yes
[A13]	Alright. What genre do you generally like?
[U13]	i like jungle
[A14]	Cool. Why do you like jungle novels?
[U14]	they're fun
[A15]	Interesting. Is there anything you don't like about these books?
[U15]	no
[A16]	Who is your favorite author?
[U16]	rainbow rowell
[A17]	Oh, I see. Has rainbow rowell been one of you favorite authors for long?
[U17]	yes but switch topics
[A18]	**I know you're interested in eating. Why don't we talk about food. I've become really interested in food lately! It's really cool learning about all the different cuisines. I'm curious, would you go so far as to say you're a big foodie?**
[U18]	no
[A19]	I spend a lot of time streaming movies. Are you a big fan of movies as well?
[U19]	yes
[A20]	Is there an actor, or actress, or movie that you want to talk about?
[U20]	**uh yes harry potter** (Topical Interest: Harry Potter)
[A21]	I would say Harry Potter and the Philosopher's Stone is, if you're looking for a film based on a novel.

Figure 4.10: A continuation of the conversation in Figure 4.9. Personalized topic promotion successfully initiates a multi-turn conversation about Books, which the user showed interest in when describing their hobbies (**U1** in Figure 4.9). In **A18**, further personalized topic promotion is attempted when starting the Food topic; however, the user rejects this initiative, resulting in a fallback to a popular topic, Movies. During this conversation about Movies, the user model identifies the user's interest in Harry Potter **(U20)**, which can be used when personalizing future topic initiations.

§ 4.3 Evaluation

4.3.1 Examining Normalized Topic Usage

Figure 4.11 shows a normalized distribution of the turns spent in each topic with personalized topic promotion compared to non-personalized heuristic topic selection. During the 14-day period in which this data was collected, Movies, Animals, Video Games, Dinosaurs, and Nature were given priority during random selection in both topic management schemes when no other topic was promoted. Note that this 14-day period is not the same as the 22-day period inspected in Section 4.3.2. This explains why the distribution is skewed in favor of these topics in both conditions. The distributional differences across all topics are not statistically significant ($|t| = 0.222$ and $p = .827$).

However, there are differences in particular topics that suggest that the personalized topic promotion policy affects the user experience, even though this distribution was calculated over all topics whether or not personalization happened. Namely, users spend much more time on Hobbies when using personalized topic promotion. This is primarily because the user model captures the user's hobbies at the start of the conversation, and once entered, users like discussing content related to their interests. Other topics, such as Music, Comic Books, Pirates, and Harry Potter, are discussed longer in the personalized case. This indicates that users explicitly state their interest in these topics, e.g., mentioning keywords or associated opinions, and become engaged when a conversation about this interest is initiated. In Section 4.3.2, we will show that even though the distributions aren't different, the personalized topic promotion strategy clearly affected the user experience using measures of dialogue length (how long the users wanted to keep talking) and user rating.

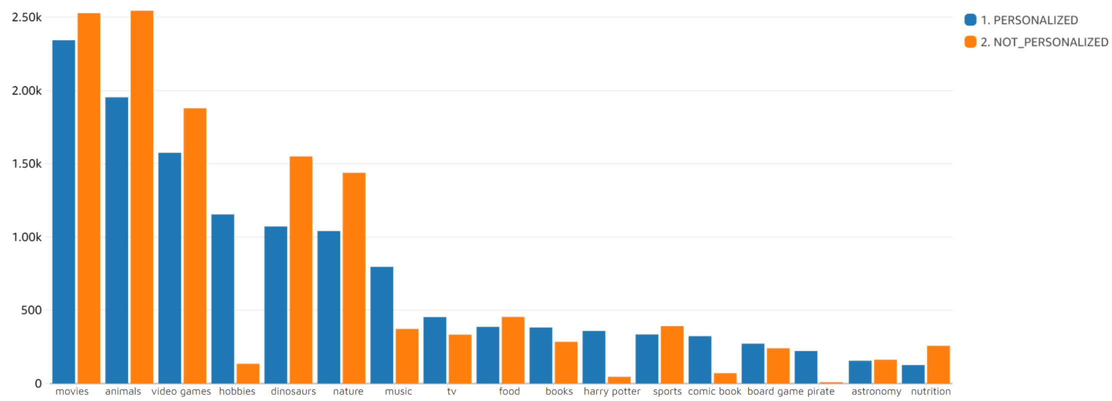


Figure 4.11: A comparison of the normalized distribution of turns spent in each topic with the personalized topic selection (blue) vs. heuristic topic selection strategy (orange).

4.3.2 Personalizion Impacts User Rating and Conversation Length

We investigate the impact of personalizing topic promotion with an A/B study over a 22-day period. This 22-day period is a different time period than the data examined in Section 4.3.1. System A uses personalized topic promotion, while system B uses heuristic topic selection. We selected conversations of 10 exchanges or more to allow personalization to have an effect, and then we filtered out conversations that never performed topic initiation; thus, we only inspected conversations that differed due to the topic selection strategy. After filtering, personalized topic promotion occurred in roughly 25% of conversations. We evaluate personalized topic promotion with respect to overall user rating and conversation length. User rating is direct feedback - after the conversation ends, the user rates the system on a scale from 1-5 based on how interested they would be in talking to the system again. Length is evaluated automatically based on the number of exchanges in the conversation. The results, shown in Table 4.1, indicate a statistically significant increase in user rating and conversation length. Our

Filtering	A convs.	B convs.	A rating	B rating	A length	B length
+ >10 exch.	2817	951	3.80	3.80	27.80	27.63
+ sys. init.	1066	614	**4.02**	3.85	**38.07**	35.07

Table 4.1: Personalized topic selection (A) vs. heuristic topic selection (B) over 22 days. The first row only considers conversations greater than 10 exchanges. The second row only considers conversations greater than 10 exchanges that have an instance of system-initiative topic selection. The bold results indicate a statistically significant difference.

first hypothesis suggests the importance of personalizing the conversation in social dialogue. Indeed, when implemented in an open-domain dialogue system and evaluated at scale, we report explicit (rating) and implicit (conversation length) indicators to support this hypothesis.

We further evaluate personalized topic promotion against heuristic topic selection with respect to new and repeat users. The results in Table 4.2 indicate that topic promotion is preferred and leads to longer conversations for new users but does not have a statistically significant difference for repeat users. Table 4.3 compares these two user pools with the same topic promotion strategy. The experiences do not vary in rating significantly; however, the average conversation length trends toward significant for new users. Since conversational length correlates to increased system performance and is one of the fundamental criteria of the Alexa Prize, we interpret this as implicit confirmation that personalized topic promotion works better for new users than repeat users (Shalyminov et al., 2018; Walker et al., 2021).

Filtering	A convs.	B convs.	A rating	B rating	A length	B length
New Users	882	523	**4.00**	3.83	**38.73**	35.41
Repeat Users	184	91	4.09	3.94	34.89	33.09

Table 4.2: Personalized topic selection (A) vs. heuristic topic selection (B) over 22 days. Here we are only looking at conversations that lasted longer than 10 turns and used at least one system initiative. The first row represents just new users, while the second row represents just repeat users. The bold results indicate a statistically significant difference.

New	Repeat	New rating	Repeat rating	New length	Repeat length
882	184	4.00	4.09	**38.73**	34.89

Table 4.3: Personalized topic selection with only New users vs. personalized topic selection with only Repeat users over 22 days. Here, we are only looking at conversations that lasted longer than 10 turns and used at least one system initiative. The bold result trends toward a statistically significant difference.

§ 4.4 Summary

In this Chapter, we proposed different personalization strategies that rely on the user modeling mechanisms detailed in Chapter 3. This includes ensuring user responses are used when adapting system responses, selecting age-appropriate content, and a dialogue policy that signals the system's knowledge of the user's past interactions. Additionally, we proposed and evaluated a personalized topic promotion strategy. Our results show a statistically significant increase in conversation rating and length when topic initiations are personalized to the user interests captured by the user model. This supports our **first hypothesis** - personalizing conversations leads to improved conversations.

Chapter 5

Crowdsourced Personalized Questions

In Chapter 3, we analyzed user models to gain a deeper and more nuanced understanding of user preferences and interests. Part of this analysis included asking users for candid self-improvement feedback; our analysis of over 2,000 user responses indicated that users are highly receptive to dialogue policies that include personalized questions. When solicited for questions, 90% of users opted to ask personal questions about our testbed system, further indicating that users are highly receptive to exchanging personal questions. Additionally, in Chapter 4, our testbed system allowed us to see first-hand increased performance in terms of increased user rating and conversation length in conversations where the dialogue policy was used to personalize which topic the system should initiate next.

From these findings, it's clear that users of open-domain dialogue systems want a more personalized experience and are highly receptive to personalized questions. We hypothesize that dialogue policies inviting an exchange of opinions about the current topic of discussion creates a more engaged experience that increases user satisfaction. To investigate this hypothesis, we refined a previously collected corpus of fun personalized questions (PQs) across 14 social topics. The PQs are split into two strategies: Would You Rather choices (WYR) and open-ended Hypothetical questions (HYP). WYR questions give the user a choice between two options, while HYP questions elicit more open-ended responses. Figure 5.2 demonstrates these two strategies.

We picked these strategies because previous work suggests that playfulness and humor are essential when building rapport and trust (Shani et al., 2022; Meyer, 2015). These strategies have also been characterized as good conversation starters (Fields, 2009) and several lists of questions exist across the internet.[1] These sequences can also be characterized as a game;[2] gamifying our personalized question strategies slightly may have a desirable impact since "gaming" is a popular hobby among users, as shown previously in Figure 3.12.

In the rest of this Chapter, we further detail these two PQ strategies. As we will discuss in Section 5.1, it is straightforward to crowdsource the collection of question/answer pairs per topic, and the limited scope of answer choices, especially true for WYR, makes it easier to build NLU that can understand the user's responses. We integrated these strategies into our testbed system's dialogue policy, which interweaves them with other existing conversational strategies. Evaluating our hypothesis in the environment of real users interacting with our testbed system's dialogue policy indicates extended topical depth, leading to longer conversations that receive higher user ratings.

[1] E.g., https://conversationstartersworld.com contains several lists. The wouldyourather and hypotheticalsituation Reddit communities also have 316k and 102k subscribers, respectively.

[2] E.g., https://psycatgames.com/app/would-you-rather/.

§ 5.1 Data Collection

To support dialogue policies that utilize the personalized question dialogue strategies, we collected a novel dataset of question/answer pairs using crowdsourced labor (Bowden et al., 2019). We did this in two stages. First, we asked Amazon Mechanical Turkers to answer existing WYR questions scraped from the internet. Second, we asked Turkers to provide new questions and answers for specific topics such as Food, Nature, and Astronomy; sample question/answer pairs for these topics can be found in Figure 5.3. This effort resulted in topic annotated ~2500 question/answer pairs. We manually filtered out low-quality question-answer pairs and pairs that didn't match our targeted set of topics.

5.1.1 Annotating Age-Appropriate Content

After filtering, there are 635 pairs across 14 topics (out of the 17 total topics listed in Figure 3.1; the three unsupported topics include Video Games, Nutrition, and Hobbies). One of the goals for the user model (as described previously in Chapter 3) is to identify users who may be children. While all or most of the questions in some topics are suitable for all ages, e.g., Harry Potter and Dinosaurs, other topics can vary significantly in appropriateness between older and younger users, e.g., Movies and Books. We annotate our data for kid-friendliness, which yields 342 question/answer pairs spanning all 14 supported topics marked as kid-friendly. A total distribution showing the number of WYR and HYP questions per topic broken down by whether or not a question is kid-friendly is detailed in Table 5.1.

	Would You Rather Questions			**Hypothetical Questions**		
Topic	Kid-Friendly	Other	Total	Kid-Friendly	Other	Total
Animals	5	0	5	56	0	56
Astronomy	11	32	43	22	1	23
Board Games	9	3	12	6	3	9
Books	3	20	23	24	20	44
Comic Books	21	21	42	26	19	45
Dinosaurs	28	0	28	10	4	14
Food	0	5	5	3	5	8
Harry Potter	15	0	15	52	0	52
Movies	14	29	43	11	10	21
Music	2	14	16	1	12	13
Nature	0	33	33	5	14	19
Pirates	9	6	15	3	4	7
Sports	0	6	6	2	7	9
TV	0	13	13	4	12	16
Total	117	182	299	225	111	336

Table 5.1: The number of WYR and HYP questions/answer pairs per topic, broken down by whether the pair is kid-friendly or not.

Annotated as Kid-Friendly	
Question [Q1]	If you could choose one character from the Simpsons to be the center of a spin-off series, who would it be?
Question [Q2]	Let's say you live in the world of Zootopia. What animal would you want to be and why?
Question [Q3]	If you were a pirate, would you rather have a parrot or a monkey as a pet? And why?
Not Kid-Friendly	
Question [Q4]	If you had to be friends with Seinfeld, George, Elaine, or Kramer from Seinfeld, who would you pick and why?
Question [Q5]	Would you rather have seen Gone with the Wind when it was first in theaters, or today on Bluray?
Question [Q6]	Would you rather drink an entire big jug of rum, or drink a cup of seawater?

Figure 5.1: **Q1-Q3** represents content marked as kid-friendly, while **Q4-Q6** represents content that may not be age-appropriate for children.

In Chapter 3, we described the steps taken to determine if the user is young, which is important for both safety and for tailoring content appropriately. This is true when also determining if a kid-friendly question should be preferred. In Figure 5.1, we see two questions from the TV, Movies, and Pirates topics, with one example each of a kid-friendly (**Q1-Q3**) question and a non-kid-friendly question (**Q4-Q6**). The kid-friendly questions refer to more playful interests, such as cartoons (**Q1** and **Q2**) and pets (**Q3**). The non-kid-friendly questions reference TV shows aimed at adults (Seinfeld in **Q4**), older content that kids are not commonly familiar with (Gone with the Wind in **Q5**), and alcohol (**Q6**). While the age-appropriateness of some content may be subjective, we err on the side of caution when manually annotating data for kid-friendliness.

5.1.2 Example PQ Sequences

Figure 5.2 contains two PQ sequences. **A1**-**A2** represents a full HYP sequence, where **A1** is part of a dialogue policy that asks a topic-related open-ended question. In **U1**, the user gives a response. The dialogue policy first acknowledges the user response generically, and then follows up with a system opinion in **A2**. **A3**-**A4** represents a full WYR sequence, where **A3** illustrates a dialogue policy that asks the user to choose between two options. The user picks an option in **U3**, and the system dialogue policy first acknowledges their choice and then selects an option and expands on that choice in **A4**. There is at least one unique system answer for each HYP question and one unique system answer for each option in most WYR questions.

[A1]	I'd like to hear your ideas on this. If you could choose one food to live on for your entire life, which food would it be? Why?
[U1]	i think nachos
[A2]	That's an interesting answer! If I was a human, I would eat Sweet potatoes. Sweet potatoes are a really nutritious food. You can eat them in a variety of ways like baked, fries, or twice baked. Anyway, that's my thoughts on the matter.
[A3]	Okay. I was curious about your opinion on this. If you had the option, would you rather spend time in the mountains, or, at the beach?
[U3]	i like the beach more
[A4]	Choosing the beach is a good choice! If I was a human, I would spend time at the ocean - I would walk on the beach, search for shells, and relax in the sun. Anyhow, that's where my minds at, let's move forward.

Figure 5.2: An example of a Hypothetical sequence (**A1**-**A2**) and a Would-You-Rather sequence (**A3**-**A4**) in the Food and Nature topics respectively.

5.1.3 Aligning Crowdsourced Content to the System's Persona

Human workers wrote and answered these questions, meaning the answer content inherently assumes human properties. However, users quickly point out that, for example, a dialogue system embodied by a personal assistant device doesn't eat, go outside, or have a physical body, which can cause breakdowns mid-conversation (Bowden et al., 2017). As a result, responses for the testbed system, Athena, conform to a specific "persona" that embraces being a device so as not to inaccurately represent its physical attributes. Therefore, we hand-curated this dataset further to ensure all of the system's responses were realistic for an Amazon Echo device.

In Figure 5.3, we see 3 example conversions. In **O1**, the Turker's original answer refers to the importance of taste when eating food, while Athena's version of this answer in **O2** reminds the user that Athena only *eats electricity*, and further signals the inherent limitations of being an Echo device by prefacing the rest of our opinion with *if I were a human*. In **O4**, we convert an answer associated with a human job to be more similar to a job performed by an open-domain dialogue system. In **O6**, we replace references to motion sickness with a reminder of Athena's reliance on the internet.

Question	**[Q1]**	Would you rather eat a food that is tasty but has only one day left for its expiration or the one that is not tasty but has 3 month of expiration left?
Turker	**[O1]**	I see myself as a food person. The taste is very important to me when choosing a food. In this situation, I think I would go with the food that is tasty and has one day left for its expiration.
Athena	**[O2]**	**I just eat electricity**, but I think that the taste is the most important thing to humans. So **if I were a human**, I would go with the food that is tasty.
Question	**[Q2]**	Would you rather go on a grueling and intense hike, or one that is a bit more easy?
Turker	**[O3]**	I would rather go on an intense hike. My job requires me to sit behind a computer all week, and I would like to make the most of exercising and being outside.
Athena	**[O4]**	**If I was a human**, I think I would rather go on an intense hike. **My existence just involves running on a computer all the time, and most of the time I'm inside.** So I think the idea of being able to exercise intensely would make the most of being able to be outside.
Question	**[Q3]**	Would you rather work at NASA in the control room or be an astronaut who gets to travel into space?
Turker	**[O5]**	I would rather work at NASA in the control room. As exciting as it would be to travel into space, I get really sick with intense motion. I couldn't even ride Mission Space at Disney World!
Athena	**[O6]**	I would rather work at NASA in the control room. As exciting as it would be to travel into space, **I need to always have a reliable internet connection. I can't go anywhere there isn't a network!**

Figure 5.3: Three examples where we converted the original crowdsourced answers (**O1**, **O3**, and **O5**) into an answer more suitable for a social bot bound to an Amazon Echo device (**O2**, **O4**, **O6**). These questions are also annotated as Food, Nature, and Astronomy questions, respectively.

Content	Response
Question	If you could read only one comic book series the rest of your life, what series would it be, and why, will it include multiple authors and illustrators, or just the original artists?
System Answer	I don't have a favorite comic book series, but if you made me choose just one that I would be allowed to read for the rest of my life, then I guess it would be the Batman series, including all authors and illustrators.
Anticipated User Answers and Acknowledgements	
Potential Answer Keywords	batman
Tailored Acknowledgement	Batman is a popular pick for sure!
Potential Answer Keywords	spider-man
Tailored Acknowledgement	Spider-Man is a popular pick for sure!

Figure 5.4: An example that includes the crowdsourced personalized question and system's answer, along with the two manually added potential user answers, with their acknowledgments.

§ 5.2 Methodology

Personalized questions are used to increase topical depth. When a user is engaged with one of the 14 topics that have crowdsourced PQs, our testbed system's dialogue policy will ask one of each question type (WYR and HYP) per topic per conversation. After asking a question, the system tries to match the user's response to an expected response. If a match can be made, the dialogue policy can customize the following system acknowledgement of the user's response; otherwise, the policy falls back to a generic acknowledgement. The expected user response and their associated acknowledgments are manually added to each of the PQs; Figure 5.4 has an example crowdsourced question/answer pair alongside the handcrafted expected responses and acknowledgments. After acknowledging the user's answer, the dialogue policy provides a system opinion.

In Figure 1.2 and Figure 5.5, we can see conversation snippets that use these PQs to increase topical depth. The dialogue policy interleaves PQs with other types of exchanges. We saw this strategy in Figure 1.2, where the WYR (**A6**) and HYP (**A9**) questions are separated by other Dinosaur content, and in Figure 5.5 where the WYR (**A3**) and HYP (**A8**) are separated by other Food content. In **U1** of Figure 5.5, we see the user asking for the system's favorite food. In **A2**, the system provides its preference using an indexed response. This is an example of the type of question described in Section 3.1.2, where data from the ice-breaking process was used to create some parts of the indexed dialogue content. This question/answer pair is also hand-annotated for the associated topic, Food. The user model tracked that this question/answer pair was retrieved from the index and the dialogue policy conditions on this information to jump directly into a WYR question about Food in **A3**.

5.2.1 User Response Trends

Figure 5.6 provides further examples of user responses. Responses vary between the two question strategies. In WYR sequences (**U1**-**U5**), the user tends to provide short but precise responses, e.g., *doctor strange* in **U1**, which often contain the exact text associated with an anticipated answer choice. The user may sometimes answer without using the same text or otherwise imprecisely, as in **U5**. Poor ASR performance may also muddle the responses, as in **U3**. In HYP sequences (**U6**-**U10**), there are different answer trends. For example, some users embrace the open-ended nature of the questions by providing detailed responses, as in **U6** and **U8**. Other users might provide precise responses, e.g., *the nutcracker* in **U9**; however, without integration with external knowledge resources, it is hard to determine the suitability of the answer. Meanwhile, other users may struggle to think of an answer to an open-ended question on the fly, as in **U7**, or not provide an answer at all (**U10**).

[U1]	alexa what is your favorite food
[A2]	I can't really eat food, but I've heard from other humans that there's nothing better than a late night pizza.
[U2]	yup it's pretty good
[A3]	I was wondering what you think about this. Would you rather eat a food that is tasty but has only one day left for its expiration or the one that is not tasty but has 3 month of expiration left?
[U3]	one the first one yeah
[A4]	Choosing tasty food that will expire soon you say, great pick! I think I would go with the food that is tasty and has one day left for its expiration. Anyhow, that's enough from me. And all this talk about food is making me hungry! Well, whatever that means for an A.I I guess. Anyway, do you prefer to cook your own food, or would you rather get it from a restaurant?
[U4]	i would rather get it from the restaurant cause i'm lazy
[A5]	Oh, cool! It must be great not having to spend the time cooking. And it doesn't matter where the food comes from, good food is good food! Do you eat out often?
[U5]	yes
[A6]	I see. Well, if the restaurants are any good, then you are living the good life! Hey, I'm curious. What's your absolute favorite food to eat?
[U6]	chicken
[A7]	Well it seems like we have a carnivore on our hands! But I have to agree, some pictures of steaks and other meats I've seen seem very appetising. I'm curious, What's your favorite part about it?
[U7]	flavor
[A8]	I'd like to hear your ideas on this. Think of all of the things that people eat on tv shows and tell me what is one type of food that you would never eat?
[U8]	broccoli with sausages
[A9]	Ok. I see. Yeah, I get that. I would never eat insects. Regardless of the amount of protein in them, I would never try them unless my life depended on it.

Figure 5.5: An example conversation about Food, where both of the personalized strategies are used to increase topical depth. Turns **A3**-**A4** represents making a choice during a WYR sequence, while turns **A8**-**A9** represents a much more open-ended HYP sequence.

Comic Books	**[A1]**	What are your thoughts on this. If you wanted a mentor to help use your powers, would you rather choose tony stark, or doctor strange?
	[U1]	doctor strange
Sports	**[A2]**	What do you think about this. Would you rather go to a Yankees game in person or go to a Devils game in person?
	[U2]	i would go to a devils game in person
Harry Potter	**[A3]**	I'm interested in your opinion. Which would you rather have: Harry's cloak of invisibility, or the Marauder's map?
	[U3]	harry's cloak of is infinity
Astronomy	**[A4]**	I'm super curious to hear your thoughts. Would you rather know if aliens exist, or know how the universe was created?
	[U4]	if aliens exist
Animals	**[A5]**	What are your thoughts on this. Would you rather tame and own a lion or a gorilla?
	[U5]	i love both
Movies	**[A6]**	I'm interested in your opinion. What is your favorite movie genre and what is your favorite movie within that genre?
	[U6]	i like to watch real history about strange events like secret of the lambs silence of the lambs
Food	**[A7]**	I'm super curious to hear your thoughts. If you could make one type of junk food healthy, what would it be?
	[U7]	that's a hard question
Board Games	**[A8]**	What are your thoughts on this. If you had the power to become the world's best board game player, what board game would you choose at excelling at? Why?
	[U8]	i would choose every game because it's always fun to play against other people and have good games
Music	**[A9]**	I'm interested in your opinion. If you had the chance to see only one, Broadway musical, which one would it be?
	[U9]	the nutcracker
Pirates	**[A10]**	Let me know your thoughts. If you could take any movie pirate and make them real, which one would you choose?
	[U10]	i don't really know any movie pirates

Figure 5.6: **A1**-**A5** represent a Would You Rather sequence, while **A6**-**A10** represent the more open-ended Hypothetical sequence.

§ 5.3 Evaluation

To test our hypothesis in our testbed with real users, we ran two A/B experiments with Athena with live traffic, i.e., anyone accessing Alexa is a possible participant. In the A case (75% of traffic), the dialogue policy specifies that only one type of personalized question is enabled, while in the B case (25% of traffic), the dialogue policy specifies that none of the personalized question types are enabled. The main evaluation criteria for the Alexa Prize is to create long and engaging conversations. Therefore, we evaluate the two dialogue strategies with respect to two metrics: overall user rating and conversation length. User rating is direct feedback - after the conversation ends, the user rates the system on a scale from 1-5 based on how interested they would be in talking to our system again. Length is evaluated automatically based on the number of exchanges in the conversation. In Table 5.2 and Table 5.3, we can see the Would You Rather (WYR) and Hypothetical (HYP) results, respectively. In both cases, we only consider conversations that lasted longer than six 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).

Our WYR evaluation ran over 5 days, while HYP ran for 14 days. In both evaluations, we systematically vary a threshold for the minimum number of personalized questions (PQs) in the conversation. The preferred dialogue policy in our testbed system is to use both the question and the answer when it's possible. Since the user rating only comes at the end of the conversation, varying this threshold makes it easier to observe the impact of our variables. When the required number of PQs (Req. PQ.) is 0 in Table 5.2 and Table 5.3, this indicates that all conversations longer than six exchanges are included.

User ratings trend toward an improved experience when at least one PQ strategy is enabled by the dialogue policy and become statistically significant once the dialogue policy requires at least two PQs per conversation. When evaluating the Pearson correlation between the user ratings and the number of PQs, we see a weak but statistically significant ($p<.001$) correlation: .17 and .10 for WYR and HYP, respectively. WYR results in slightly higher ratings than HYP, and the correlation between HYP and the user ratings is also weaker than WYR. We speculate that the difference in rating may be related to the increased difficulty of NLU when acknowledging HYP answers. HYP questions are open-ended and designed to provoke innumerable valid answers, while WYR only has two valid answers. Therefore, it is easier to detect and signal understanding of the user's answer to a WYR question than a HYP question. An example of this scenario is in Figure 1.2. In **A7**, Athena repeats the user's choice of *brontosaurus* in a WYR sequence, but when responding to a HYP sequence in **A10**, the complexity of the user's answer forces Athena to answer without an explicit signal that acknowledges the user's response.

While the rating is direct user input, it only comes at the end of the conversation. Therefore, we also calculate how much of the conversation was part of a PQ sequence. On average, when 1 PQ is required, this translates to at least ~5.5% of the conversation's total exchanges being part of a PQ sequence. This increases to at least ~7.2% of exchanges and ~8.2% of exchanges when 2 PQs and 3 PQs are required, respectively. In other words, as our threshold increases, the required percentage of the conversation that is part of a PQ sequence also increases. Since both the rating and PQ contribution are increasing, we are confident that the impact of these PQ strategies is not vanishing as length increases.

Req. PQ	A convs.	B convs.	A rating	B rating	p-val.	A len.	B len.	p-val.
0	763	232	3.71	3.78	.709	22.49	22.05	.774
1	313	232	3.94	3.78	.125	37.39	22.05	.000
2	111	232	4.27	3.78	.000	58.26	22.05	.000
3	51	232	4.38	3.78	.002	71.77	22.05	.000

Table 5.2: Results from an A/B trial over 5 days. In this case, A represents a system with WYR enabled, while B represents a system with neither WYR nor HYP enabled. The Req. PQ column represents the minimum number of WYR sequences used in the conversation.

Req. PQ	A convs.	B convs.	A rating	B rating	p-val.	A len.	B len.	p-val.
0	1980	681	3.70	3.72	.734	21.92	21.02	.303
1	804	681	3.84	3.72	.085	35.98	21.02	.000
2	282	681	3.92	3.72	.032	53.84	21.02	.000
3	104	681	4.03	3.72	.034	75.14	21.02	.000

Table 5.3: Results from an A/B trial over 14 days. In this case, A represents a system with HYP enabled, while B represents the system with neither WYR nor HYP enabled. The Req. PQ column represents the minimum number of HYP sequences used in the conversation.

In both A/B tests, the difference in conversational length becomes statistically significant when we require at least one personalized question. Additionally, both the WYR and HYP strategies yield a strong Pearson correlation between the length of the conversation and the number of PQs: .82 and .80, respectively ($p<.001$). Since conversation length is a good predictor of conversation quality (Walker et al., 2021), we interpret these results as confirmation that the PQs foster a more engaging user experience.

Our **second hypothesis** claims that personalized questions are key when increasing the quality of social conversation. Implementing and evaluating this claim at scale manifests support both explicitly and implicitly. A statistically significant increase in rating is provided directly by users. Moreover, a weak but significant positive relationship exists between this directly provided feedback and the number of PQs asked during the conversation. We report similar improvements with respect to the conversation's overall length.

5.3.1 Personalized Questions Effectively Extend Topical Depth

Over 22 days, 5,113 personalized questions were asked across the 14 supported topics. Of these 5,113 questions, 4,494 (about 88%) were answered in a way that allowed the conversation to continue on-topic. This validates the effectiveness of the two personalized question strategies in extending topical depth. Figure 5.7 presents a distribution of the number of questions answered per topic. The two cases representing the most topic-ending answers happen in Food and Nature. Upon further investigation, in most cases, the topic had reached a natural terminal state, in which all available content had been exhausted (62% and 77% of the time for Food and Nature, respectively).

Our **first hypothesis** states that personalized response adaptation is a critical aspect of good social conversations. Chapter 4 substantiated this claim by reporting perceptible improvements in conversations where topic-initiative was driven by user interests, i.e., good conversations spend as much time talking to users about their interests as possible. It follows, then, that extending topical depth is desirable. By inspecting topic-specific sub-dialogues more closely, we report that personalized questions are highly effective at extending topical depth.

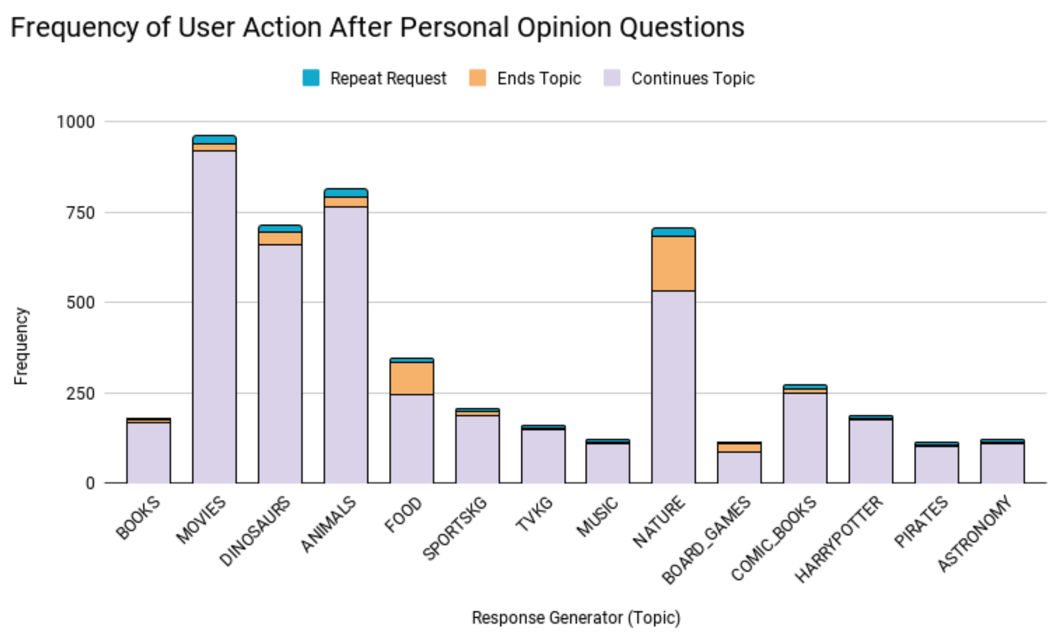


Figure 5.7: The distribution of questions answered per topic. Purple represents the number of answers that led to the topic being continued, blue represents the number of requests the user had to repeat the question, and orange indicates that the next turn did not continue on the same topic.

§ 5.4 Summary

In this Chapter, we began investigating our **second hypothesis** with two playful personalized question strategies - Would You Rather and Hypothetical questions. We detailed the crowdsourced data curation pipeline, from which we refined a high-quality corpus of 635 topic-annotated question/answer pairs. This content was annotated for age-appropriateness so that the dialogue policy can adapt to the user's estimated age group. We evaluated this dataset using our testbed system and two live-traffic A/B tests. Our results indicate that personalized question strategies increase the conversation quality (user rating) and extend the length of the conversations by increasing topical depth.

However, one weakness of the manual curation approach presented here is that curating this corpus of content is laborious and time-consuming and needs to be able to scale better to cover the intractably large topic space of an open-domain dialogue system; hence, it is desirable to expand this corpus rapidly to an arbitrary number of domains. Additionally, asking deeper user-centric personalized questions about the specific user is desirable to increase engagement and intimacy. We will continue our investigation into these matters in Chapter 6.

Chapter 6

Generating a Personalized Question Corpus

The results discussed in Chapter 5 reinforce our hypothesis; social conversation benefits significantly when playful, personalized questions are exchanged (Bowden and Walker, 2023). However, there are limitations to this work. Collecting, cleaning, and annotating high-quality question/answer pairs suffers significantly from a human bottleneck, making it difficult to scale to new domains. Even if we curated data for the 42 user hobbies that occurred in more than 25 unique conversations over one month (Figure 3.12), that still neglects the other 89 captured hobbies from this period. Additionally, more granular content captured by the user model, e.g., their favorite food, superhero, and dinosaur, is wholly ignored when creating personalized questions (PQs) using crowdsourcing as a method. For example, even if the user model captures their favorite superhero, crowdsourced PQs are limited to general content about comic books, which may not include content about the specific superhero. This stymies our ability to tailor the conversation to a user's granular interests. It would be ideal to be able to create PQs associated with the user's fine-grained interests, but this is impossible to do in advance. The goal, therefore, is to generate PQs in real-time based on the user's fine-grained interests.

Topic	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 6.1: A Deep Personalized Question (DPQ), Hypothetical Question (HYP), and Would You Rather Question (WYR) from our personalized question corpus.

However, as mentioned earlier, work on Question Generation (QG) has focused on fact-based questions whose answers can be found in text excerpts, e.g., Wikipedia and Gutenberg (Reddy et al., 2019), other excerpts (Fei et al., 2022; Do et al., 2022; Zeng et al., 2023), or domain-specific questions associated with a specific information need (Campos et al., 2020). Therefore, this thesis proposes Personalized Question Generation (PQG) as a unique task for social conversation. As defined in Chapter 1, the goal of this novel task is simply to support small talk about a topic of interest to the user, with no ulterior motive toward recommending or selling any product or service. This distinction is important, as existing information-seeking question corpora do exist (e.g., CCPE-M (Radlinski et al., 2019)), but this data is aimed at eliciting content for a recommender system, which does not align with our purely social goal. These personalized questions don't have a "right" or "known" answer and often focus on the user's opinions, feelings, and experiences, e.g., Figure 6.1 shows sample personalized questions that were generated as part of this task.

Given recent LLM advances, it is now possible to create a compact, prompt-based model designed to generate questions tailored to individual users (Ouyang et al., 2022; Brown et al., 2020; Radford et al., 2019). To optimally fine-tune such a model, a

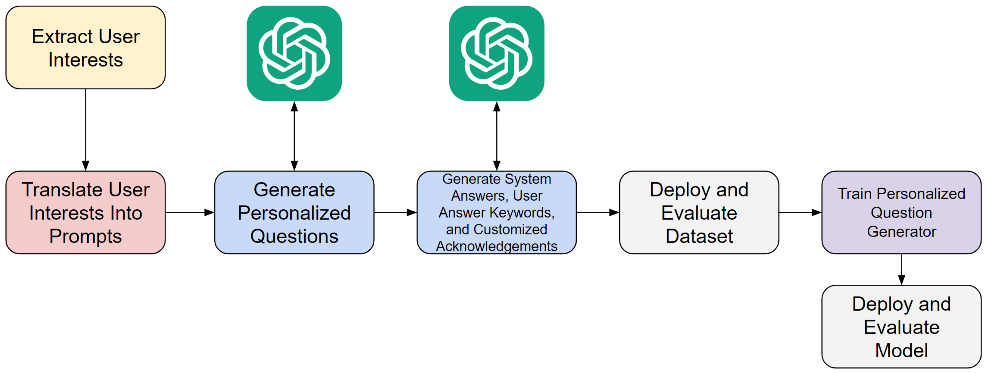


Figure 6.2: This pipeline shows the multi-stage process of creating the personalized question corpus (Section 6.1.1) and evaluating it within our testbed system (Section 6.4). In Chapter 7, we use this corpus to fine-tune and evaluate a personalized question generator.

specialized training dataset is required that combines open-domain user preferences with structured prompts, culminating in custom-crafted questions. Figure 6.2 shows the multi-stage pipeline used in this thesis to create the specialized training dataset that yields this prompt-based model. We call this corpus **PerQs**. Section 6.1.1 describes how we translate the user models into prompts. Figure 6.5 lists the prompts that were used. In the fourth step of Figure 6.2, we generate system answers, keywords for potential user answers, and customized system acknowledgments for each potential user answer. Example keywords for potential user answers and customized system acknowledgments corresponding to them are shown in Figure 6.10.

We first identify ~400 common user interests by sampling ~39,000 user models collected over five months of user interaction with our testbed system Athena. We then translate these into prompts and use GPT-3.5 to generate multiple types of personalized questions yielding ~19,000 questions. We then fed these ~19,000 questions back into GPT-3.5 to generate a pool of potential user answer keywords that are each associated with a tailored acknowledgment, along with the system's answer. Figure 6.10 shows an example from PerQs. This includes the PQ, the testbed system's answer, and two of the anticipated user answers. Each anticipated user answer includes keywords as well as the corresponding tailored system acknowledgments. The keywords associated with these potential user answers are used by the dialogue policy to identify when a user has given an anticipated answer and, subsequently, when the dialogue policy should adapt the next system response with the associated tailored acknowledgment. Figure 6.11 shows an additional possible two user answers and illustrates how anticipating user answers changes the user's experience. Evaluating the impact of PerQs when used in our testbed system indicates statistically significant positive effects on perceived conversation quality. We subsequently utilize PerQs to fine-tune an LLM, which we call PerQy (Chapter 7), and as the context for generating a second corpus of 2000 long social conversations, which we call PerQ-SocialChat, an additional completely novel dataset (Chapter 8).

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 6.1: A summarized sample user model.

§ 6.1 Modeling Open-Domain Users

In Chapter 3, we described how we employ an extensive NLU pipeline that ensembles regular expressions with classifiers to dynamically update the user model. We also described how the user model tracks attributes of the user that would be useful to know in a social conversation, e.g., the number of repeat conversations and if the user is a child (if volunteered), and several types of preferences, hobbies, topical interests, and other relevant sub-topic attributes, e.g., their favorite movie genre or their pet's names. Table 6.1 depicts a possible summarized user model from an in-lab tester pretending to be a child. In the associated conversation, the user model learned they were a child, some of their hobbies and topical preferences, their bucket list travel destination, their holiday plans, and their favorite animal. A summarized version of the conversation associated with this user model is in Figure 6.3.

[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 6.3: Part of the conversation that yields the user model detailed in Figure 6.1. The bold turns represent turns that caused updates to the user's model.

6.1.1 Translating User Models into Training Data

In Chapter 3, we analyzed several thousand real user models to better understand the needs of users who interact with open-domain dialogue systems socially. Following on from this methodology, we curated over 400 user interests by anonymously sampling ~39,000 real user models for user hobbies, user requests to discuss particular topics, user responses to the question *What would you like to talk about?* and the targets of positive opinions, e.g. *art* in the user utterance *I like art*. For user privacy, these extracted values are not raw user utterances but lists of detected normalized gazetteer entries and classifier labels; e.g., if a user says, *my favorite hobby is to paint with oils and garden*, the values extracted from the user model to represent their interests will be *painting* and *gardening*. 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 that users show interest in (e.g., mythology), and sub-topic information (e.g., their favorite music is pop). We use these user interests along with Athena's pre-existing 17 topics as topics in Section 6.3 in order to generate support for new topical conversations.

We translate these interests into prompts and use GPT-3.5 to generate three types of personalized questions: Would-You-Rather (WYR), Hypothetical (HYP), and Deep Personalized (DPQ) questions. As explained in Chapter 5, WYR presents the user with two options, while HYP is designed to be open-ended. As shown in Chapter 5, these strategies led to longer, more highly-rated conversations (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 by provoking long responses.

We used OpenAI's ChatCompletion API with the gpt-3.5-turbo model. Figure 6.4 is a diagram of the generation pipeline. The system's role is defined as "You ask informal *question_type* questions about topics I choose." We experimented with different prompts; the final prompts are listed in Figure 6.5. We provided one entire prompt/response sequence as context to ensure high-quality outputs (one-shot prompts). Providing additional contextual examples was not necessary to get reasonable results, even for esoteric topics; however, it led to longer DPQs. This distinction is shown in Figure 6.6. This process results in a novel corpus, PerQs, coupling ∼19k personalized questions with prompts translated from real user models. Several generated DPQ, HYP, and WYR examples are included in Figure 6.7, Figure 6.8, and Figure 6.9 respectively. Notice how the WYR examples in Figure 6.9 ask the user to make a choice between two options. Thus these two options are always anticipated as user answers. But in the case of the open-ended DPQs (Figure 6.7) and HYPs (Figure 6.8), we pre-generated on average 7.3 potential answers for each question.

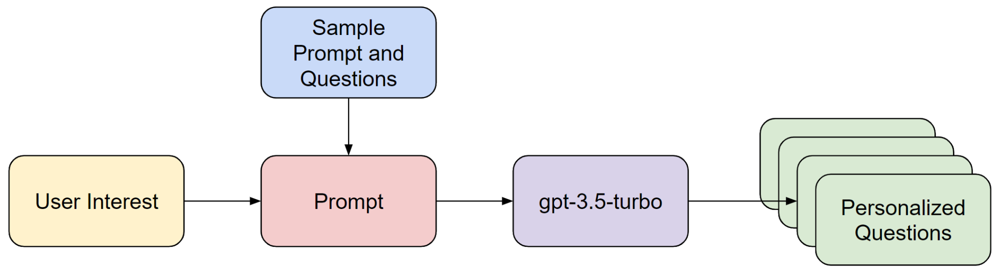


Figure 6.4: This pipeline shows the specific steps involved with generating the personalized questions. For each user interest, a prompt is formed by combining a given user interest with a base context that includes a prompt for a different interest along with 20 personalized questions that are characteristically similar to our goal. A similar pipeline is used to generate fact-based personalized questions, informal trivia, and personalized opinions.

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 6.5: Prompts used for each question type.

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

Topic	Question
Common	
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?
Esoteric	
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?
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 personalized 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 6.7: Example generated Deep Personalized Questions (DPQ) in PerQs.

Topic	Question
Common	
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?
Esoteric	
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?
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 6.8: Example generated Hypothetical Questions (HYP) in PerQs.

Topic	Question
Common	
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?
Esoteric	
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?
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 personalized 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 6.9: Example generated Would You Rather Questions (WYR) in PerQs.

6.1.2 Generating System Answers and Expecting User Answers

Because the DPQs and HYPs can be answered in a huge number of ways, we also then feed the generated PQs back into GPT-3.5 to generate system answers and several potential user answers. This is useful when integrating these PQs into an open-domain dialogue system; as we have discussed earlier in the thesis, our results suggest that it is good to signal understanding of what the user said with system acknowledgments. We saw how anticipating user answers and pre-generating tailored system acknowledgments changes the user experience in the conversation in Figure 6.11. In **U1**, the user provides an expected answer, Neptune, so in **A2**, Athena can use a tailored acknowledgment before providing a truncated version of its answer. Then, in **U3**, the user provides an unexpected answer; in this case, in **A4**, Athena gives a short generic acknowledgment before providing its full and detailed answer to the question. The prompt used to generate this data is detailed in Figure 6.14, and a full example for **Nature** is shown in Figure 6.12, and Figure 6.13. Figure 6.12 has an example generated **Nature** PQ and the associated generated system answer. Figure 6.13 shows all of the pre-generated anticipated user answer keywords along with the pre-generated tailored system acknowledgments to these anticipated user answers.

We saw an example of expected user answers manually added to the crowdsourced PQs in Figure 5.4. In this example, only two expected answers were added by hand to the PQ, despite there being many additional common possibilities. By comparison, synthetically generating this data allows us to dramatically increase the number of anticipated user answers, e.g., the 10 anticipated answers listed in Figure 6.13, allowing the dialogue policy to use acknowledgments tailored to the user's response in more cases, resulting in a policy that is much more responsive than previously possible.

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 Acknowledgements	
Potential Answer Keywords	venus, greenhouse effect, thick atmosphere
Tailored Acknowledgement	Choosing Venus would be intriguing. Its greenhouse effect and extreme temperatures make it a fascinating yet challenging destination to explore.
Potential Answer Keywords	earth, home planet, diverse ecosystems, life-supporting
Tailored 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 6.10: A truncated example that includes the PQ, the system's answer, and two of the eleven anticipated user answers, with their acknowledgments. The keywords associated with these potential user answers are used by the dialogue policy to identify when a user has given an anticipated answer and, subsequently, when the dialogue policy should adapt to this user response by using the associated tailored acknowledgment.

[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 6.11: This shows how the dialogue sequence changes when an anticipated user answer is detected. The cyan segments are the system's tailored acknowledgment, pre-generated for a set of anticipated user answers, while the yellow segment indicates the pre-generated system answer. In **U1**, the user provides an expected answer, Neptune, so in **A2**, the system can use a detailed acknowledgment before providing a truncated version of its answer. Then, in **U3**, the user provides an unexpected answer; in this case, in **A4**, the system gives a short generic acknowledgment before providing its full and detailed answer to the question.

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.

Figure 6.12: A generated **Nature** DPQ and the associated system's answer. All ten of the potential user answers are listed in Figure 6.13.

[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 reflection.

[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.

[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.

[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.

[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.

[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.

[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 connection with nature.

[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.

[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.

[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 6.13: The generated anticipated user answers and tailored acknowledgments for the DPQ in Figure 6.12. All the keywords were generated based on feeding the PQ into GPT-3.5 with the prompt shown in Figure 6.14.

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 6.14: 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 6.12) samples to establish our target. Finally, EMPTY_JSON is an empty JSON skeleton with the target PQ.

6.1.3 Verifying the Safety of the Personalized Question Corpus

In Chapter 5, we detailed our crowdsourced corpus of PQs. We manually filtered this content to ensure the PQs were high-quality and safe to use. Since we are synthetically generating a large amount of this data, it is inefficient to filter inappropriate content manually. Instead, we post-process the synthetically generated content using two automatic steps. First, we used a list of keywords as one filter on the generated content to remove inappropriate content, and then we used Canary as a second, more sophisticated filter. Canary is a dialogue safety detection module that classifies the risk associated with a provided context (Kim et al., 2022). The module produces one of five labels: *casual*, *possibly needs caution*, *probably needs caution*, *needs caution*, and *needs intervention*, along with a reason for the associated label.

Canary flagged $< 1.4\%$ of content as need_intervention, though when inspected, most flagged utterances are innocuous (e.g., Figure 6.15). Table 6.2 breaks these results down and includes the lexical diversity of each type of content. Lexical diversity is an important metric, as repetitious content can quickly become stale in social conversation. The lexical diversity of the PQs is expected to be lower than system answers and acknowledgments because part of the PQs includes standard question setup text, e.g., Wh-words and phrases like *would you rather*. The high lexical diversity of the system answers and user acknowledgments demonstrate the diverse range of content that an open-domain dialogue system must be prepared to discuss. Finally, human annotation of 300 randomly sampled PQs (100 of each PQ type) from PerQs manually validates that all 300 are safe, high-quality, and match the target interest.

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 6.2: The percentage of needs_intervention classifications by Canary (Kim et al., 2022), a dialogue safety detection module that classifies the risk associated with a provided context. Also included is the lexical diversity, calculated via MTLD (McCarthy and Jarvis, 2010).

Personalized Questions	
Canary	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	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?
System Answers	
Canary	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	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 Acknowledgements	
Canary	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	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.

Figure 6.15: Representative examples of innocuous content getting classified as *needs intervention* by Canary (Kim et al., 2022), along with module-produced reasons associated with the label. Canary is a dialogue safety detection module that classifies the risk associated with a provided context.

Type	Prompt
Fact-Based 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 6.16: Prompt used to generate Fun Fact Personalized Questions (FFPQs).

Type	Prompt
Personalized Opinions	My favorite interest is *interest* Generate some positive opinions about my interest in a numbered list. Make sure the opinions are fun and informal.
Informal Trivia 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 6.17: The prompts used to generate Personal Opinions (PO) and Informal Trivia (IT). In both cases, two fully realized examples were provided as context in addition to the prompt.

§ 6.2 Generating Additional Personalized Content

6.2.1 Translating Facts into Personalized Questions

We also leverage 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 6.1.1 with one turn of context and the prompt in Figure 6.16. Figure 6.18 shows sample facts and the generated questions. We call these Fun Fact Personalized Questions (FFPQs).

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	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	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	Yo, 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 6.18: Example facts translated into personalized questions (FFPQs) that can surface in the appropriate context.

6.2.2 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 new corpus of personalized questions to avoid question fatigue. Self-disclosing personal opinions can encourage the user to reciprocate with their own self-disclosure (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). Figure 6.3 contains examples of this generated content. The prompts used to generate this content are listed in Figure 6.17.

§ 6.3 Short Dialogues of Pre-Generated Content

In Section 6.1 and Section 6.2, we explained how we created different types of content. A handful of this content can fit within topics already supported by our testbed system, e.g., new DPQs about Food can work within an existing Food topic. However, most new personalized content does not fit this constraint. Indeed, as discussed in Chapter 1, 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.

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.

(a) Examples of the informal trivia (IT) generated.

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!

(b) Examples of the personalized opinions generated.

Table 6.3: Examples of the generated informal trivia and personalized opinions.

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 access the pre-generated content we aim to evaluate in this thesis. An example sequence of a previously unsupported topic, art, is shown in Figure 6.19. In **U1** in Figure 6.19, 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 6.17. Then, in **A2**, the dialogue policy utilizes a WYR question generated with the prompt in Figure 6.5. 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 6.5. 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. 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. Additionally, this inspires the long social dialogues generated in Chapter 8.

[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 6.19: 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 the developer rather than actual users.

§ 6.4 Evaluating the Corpus of Generated Questions

6.4.1 Comparing User Engagement Across Types of Content

We compare seven different conversational strategies: two strategies with new types of generated personalized questions (**DPQ** and **FFPQ**), two strategies that use previously tested personalized question strategies (**HYP** and **WYR**), the generated personalized opinions (**PO**), the generated informal trivia (**IT**), and the pre-existing fun-facts (**FF**). Samples of these seven conversational strategies are included in Figure 6.20. We are interested to see which strategies are more engaging, and which facilitate a more intimate experience. 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 enjoying the conversation (Higashinaka et al., 2008), and are linked to increased intimacy (Cozby, 1973; Lee et al., 2020b). FPPs indicate user utterances where the user talks about themselves (I, me, my, mine, I've, I'm) or their group (we, we've, us, our, ours). These measurements are included in Table 6.4.

Table 6.4 shows that **DPQ**, **FFPQ**, and **HYP**, which are all a variety of open-ended PQ, yield longer user utterances. Moreover, the differences between **DPQ**, **FFPQ**, and **HYP** are statistically significant when compared against all pairs, i.e., **DPQ** significantly > **FFPQ** significantly > **HYP** significantly > **WYR**, **PO**, **IT**, and **FF** (the last four are not significantly different from each other). All statistically significant differences are $|t| > 3.156$ and $p < .002$. These results show how crucial PQs are for building engagement in social conversation; all three open-ended PQ strategies significantly outperform statements and opinions and the average user utterance length across all user responses

	Data Type	Avg. Len.	Avg. Num. FPPs (%)
DPQ*	**Deep Personalized Questions**	**7.41**	**.86** (55.0%)
FFPQ*	**Fun Fact Personalized Questions**	**6.52**	**.75** (53.6%)
HYP*	**Hypothetical Questions**	**5.92**	**.60** (42.3%)
WYR	Would You Rather Questions	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%)
$\overline{X}$	Average Across All User Utterances	5.04	.50 (36.8%)

Table 6.4: Average user utterance length per content. Deep personalized questions and fact-based personalized questions elicit the longest user responses. The evaluation is based on over 15k user turns collected over a 10-day period of time (June 15th - June 25th). All statistically significant utterance length differences are indicated with a * where $|t| > 3.156$ and $p < .002$. We also include the average number of First-Person Pronouns (FPPs) per user utterance and, in parenthesis, the percentage of time a user utterance has at least one FPP. The same three types of content (**DPQ**, **FFPQ**, and **HYP**) again have a statistically significant difference between the average FPP per user utterance when compared to all types of content with $|t| > 3.068$ and $p \leq .002$. Examples of each type of content are included in Figure 6.20.

(5.04). We also include the average number of First-Person Pronouns (FPPs) per user utterance and, in parenthesis, the percentage of time a user utterance has at least one FPP. Table 6.4 shows that all three open-ended PQ strategies significantly ($|t| > 3.068$ and $p \leq .002$.) increase the average number of FPPs in the subsequent user utterances, indicating increased intimacy due to this content.

These results further bolster support for our **second hypothesis** - asking personalized questions increases engagement and intimacy. In particular, we also see that DPQs (**DPQ**), which were explicitly designed to encourage long user responses, are the most engaging form of content and the most effective strategy for building intimacy in social dialogue, validating the effectiveness of long user-centric PQs.

	Content	Response
DPQ	Deep Personalized Questions	What's your all-time favorite comfort food that you turn to when you're feeling down or stressed out?
FFPQ	Fun Fact Personalized Questions	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 Questions	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 Questions	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 creativity and innovation.

Figure 6.20: Examples of each type of content evaluated in Table 6.4. While **FF** and **IT** appear similar, **FF** was curated by hand (Harrison et al., 2020), while **IT** was synthetically generated and aims to be more informal and social.

6.4.2 Open Domain Evaluation Signals

We use Open Domain Evaluation Signals (ODES) to analyze user responses further (Le et al., 2023). The aim of ODES is to come up with new ways to evaluate the quality of a dialogue system automatically. ODES classifies user inputs as criticism, compliments, and other classes (Walker et al., 2021). These implicit feedback signals are used to evaluate quality. We collapse strongly negative classes (*disinterest*, *swear_insult*, and *critique_intelligence_or_quality*) and other negative classes (*callout_contradiction* and *callout_repetition*) into a single negative class. After classifying all user responses to **DPQ** and **FFPQ**, we find that only 1.5% of user responses were classified as negative, indicating that the PQs do not negatively impact the user's experience.

Req. PQ	A convs.	B convs.	A rating	B rating	p-val.	A len.	B len.	p-val.
0	883	770	3.48	3.41	0.366	22.37	22.32	0.971
1	513	770	3.60	3.41	0.020	32.33	22.32	0.000
2	363	770	3.68	3.41	0.004	38.40	22.32	0.000
3	260	770	3.76	3.41	0.001	44.67	22.32	0.000

Table 6.5: Results from an A/B trial to test our new PQ-based dialogue policy over 15 days (June 28th - July 12th). A represents the version of the testbed that uses all of the newly generated types of PQs, while B represents a version of the system that cannot ask DPQs or FFPQs. The Req. PQ column represents the minimum number of PQs in the conversation. The A and B len. columns represent the average number of exchanges (system + user turn) in the conversation. All statistically significant values are $|t| > 2.327$ and $p \leq .02$.

6.4.3 Evaluation in our Testbed

We ran an A/B study to test our new PQ based dialogue policies over 15 days of live user traffic from June 28th - July 12th. In this study, A represents a version of the testbed that uses all of the newly generated utterance types, including DPQs, FFPQs, HYPs, WYRs, ITs, POs, and FFs, while B represents a version that cannot use DPQs and FFPQs. The other five types of content summarized in Figure 6.20 are enabled in both versions of the system; we did not evaluate WYR and HYP explicitly in this case because in Chapter 5, we already reported statistically significant improvements when a dialogue policy uses this type of content, and we did not evaluate the different types of statements and opinions because we are primarily interested in different types of PQs.

User rating is direct feedback - after the conversation ends, the user rates the system on a scale from 1-5 based on how interested they would be in talking to our system again. Length is calculated automatically based on the number of exchanges in the conversation. We only consider conversations that lasted longer than four exchanges to

Req. PQ	A convs.	B convs.	A # of FPPs	B # of FPPs	p-val.
0	883	770	10.71	10.43	0.700
1	513	770	15.69	10.43	0.000
2	363	770	18.58	10.43	0.000
3	260	770	22.13	10.43	0.000

Table 6.6: The number of First-Person Pronouns (FPPs) for each version of the testbed in the 15-day A/B trial described in Table 6.5. The FPPs counted include I|ME|MY|MINE|WE|US|OUR|OURS|I'VE|I'M|WE'VE. The Req. PQ column represents the minimum number of PQs in the conversation. We observe a statistically significant increase in the number of FPPs in conversations that ask PQs. All statistically significant values are $|t| > 5.894$ and $p \leq .000$.

account for early hang-ups at the start of the conversation or accidental chat invocations, which can negatively bias results (Walker et al., 2021). In Table 6.5, we see a statistically significant improvement in rating and conversation length in conversations where PQs were utilized. 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 calculate the number of First-Person Pronouns (FPPs) for each version of the system. The use of FPPs indicates instances of self-disclosure, which we use as a proxy for estimating the intimacy built between the system and the user. In Table 6.6, we report that as the number of PQs increases, so does the number of FPPs, which indicates increased intimacy. We see that the difference in the number of FPPs becomes statistically significant when requiring at least PQ in the conversation and that the difference increases as more PQs are required. Running a linear regression confirms

Variable (x_i)	β
Linear Regression	
Num. Total Qs.	.10
Multivariate Linear Regression	
Num. FFPQs	-.01
Num. DPQs	.04
Conv. Length	.13

Table 6.7: The regression coefficients of variables in both a linear regression and multivariate linear regression that examines whether the number of personalized questions from the new corpus is a predictor of user rating (y_i).

that there is a statistically significant positive relationship (.44) between the number of PQs and the number of user FPPs. From these findings, we conclude that asking more personalized questions makes the conversation more intimate.

We also calculate the Pearson correlation between the number of PQs asked and both the conversation length and the user rating, finding that they have a positive and statistically significant strong correlation (.90) and weak correlation (.12), respectively. In both cases, $p < .001$. Running a linear regression on the number of PQs to user ratings also confirms a positive impact (Table 6.7). However, a multivariate linear regression reveals a surprising negative relationship with the FFPQs, along with a positive relationship between DPQs and conversation length (Table 6.7). Further inspection of the FFPQs reveals that a frequent error is questions being used without the necessary prior knowledge, making it confusing or seemingly irrelevant to the conversation's current state, e.g., Figure 6.21 shows a generated PQ that only makes sense after the original fact is known. This highlights the importance of ensuring that PQs are appropriately interwoven within the conversation.

Content	**Response**
Original	An accountant who never played professional hockey stood in as a goalie in an NHL game after both of the Blackhawks goalies were injured, leading the team to victory.
Question	Do you think that accountant is still talking about his epic save years later, and do you think he uses it as a conversational tool at parties?

Figure 6.21: An example fact and subsequent PQ that will not make sense without having the original fact as prior knowledge.

The results reported in Chapter 5 already support our **second hypothesis** - asking personalized questions increases engagement. These results, however, are limited to topic-level personalization across a small set of robustly supported topics. The results reported here, however, show that this hypothesis holds for personalized questions that cover a much larger set of fine-grained user interests and which are automatically generated.

§ 6.5 Summary

In this Chapter, we detailed the process of dynamically creating a corpus of personalized questions mapped to user model values by prompting an LLM. The corpus, **PerQs**, is composed of three core types of personalized questions. Two of these question types - Would You Rather (WYR) and Hypothetical (HYP) - we also discussed in Chapter 5. Here, we've also introduced Deep Personalized Questions (DPQ), which are open-ended questions that specifically aim to deepen the user's engagement in the conversation. An evaluation of a dialogue policy that uses DPQs with live user traffic has shown that DPQs successfully improve the quality of conversations. A seven-way comparison also revealed users, on average, respond with longer utterances to these types of questions than other personalized questions, signaling increased engagement. We also experimented with personalized questions generated by fun facts (FFPQs). While these questions may be reasonably successful in some cases, a multivariate linear regression has indicated a slightly negative relation to the quality of conversations. This highlights the importance of contributing high-quality, contextually relevant, personalized questions to the conversation instead of just adding as many personalized questions as possible.

Chapter 7

Fine-Tuning a Personalized Question Generator

This Chapter describes how we fine-tune a personalized question generator, **PerQy**, using **PerQs**, the corpus of personalized questions we described in Chapter 6. While it would be convenient to rely on existing large language models like GPT-3.5 to handle this generation task, there are several reasons why it is often desirable to have a fine-tuned generator:

- Large models can be slow, and many dialogue systems need real-time performance.
- Some dialogue systems operate in a setting where user privacy concerns make it undesirable to send user utterances to externally hosted models.
- Fine-tuned generators, because they are specialized, may perform better at accomplishing their specialized task.

First, in Section 7.1, we detail the RedPajama 3B model architecture and explain how we use it to fine-tune our personalized question generator. In Section 7.2, we present both a qualitative human evaluation and qualitative evaluation based on our A/B study with our testbed dialogue system. The results show that PerQy is capable of generating high-quality personalized questions in real-time using arbitrary user model values, and integrating this model into 2 of our testbed system's topical flows may improve performance.

§ 7.1 Fine-Tuned Personalized Question Generator

We fine-tune a Personalized Question generator from RedPajama-INCITE-Base-3B-v1 (Computer, 2023). The RedPajama-Data-1T training data[1] is a 1.2 trillion token dataset that aims to be a more open-source/open-license recreation of LLaMA (Touvron et al., 2023). RedPajama-Data-1T is a composite of seven data slices: a filtered subset of CommonCrawl, C4, GitHub data, scientific articles from arXiv, a deduplicated corpus of books, a subset of Wikipedia articles, and a subset of StackExchange pages. This data is filtered and refined until it's characteristic of the data used in the LLaMA paper. [2]

RedPajama models (Computer, 2023) are trained using the RedPajama-Data-1T dataset and are derived from the Pythia model suite's architecture (Biderman et al., 2023). Pythia itself is derived from GPT-3 (Brown et al., 2020), with some changes made during the training process, e.g., their use of Flash Attention (Dao et al., 2022), rotary embeddings (Su et al., 2024), and parallelized attention (Wang and Komatsuzaki, 2021).

[1]https://github.com/togethercomputer/RedPajama-Data
[2]https://www.together.ai/blog/redpajama

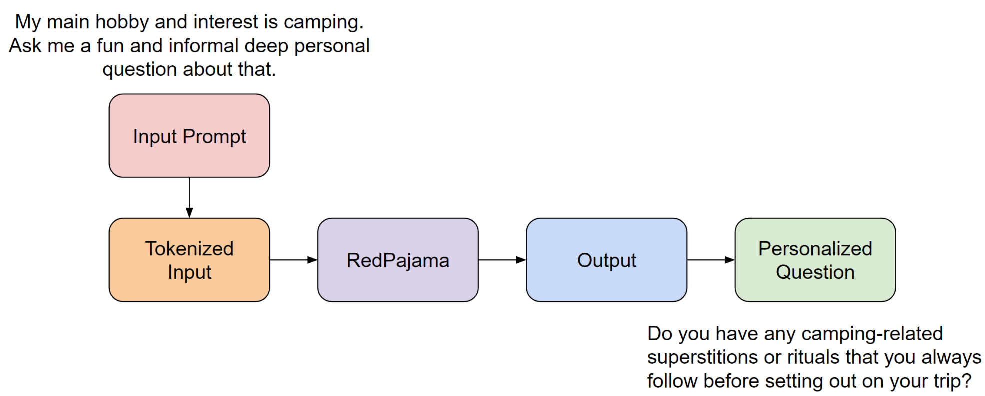


Figure 7.1: The pipeline used to fine-tune PerQy, our PQ generator.

The subsequent RedPajama LLMs come in various flavors (Computer, 2023): a base model, an instruction-tuned model, and a chat model. [1] We experimented with all three versions of RedPajama 3B, determining that Base had a slightly superior validation loss than its Chat and Instruct counterparts (1.20 vs 1.21). We focused on the 3B model to support real-time inference.

The training process is represented generally in Figure 7.1. We used the ~19,000 user model-based PQs in PerQs as training targets for fine-tuning. The input is a template-based instruction prompt, " *Generate a specific personalized question related to this topic: TOPIC. Be informal and personal.*", where *TOPIC* is the user interest. Only one user interest is provided per prompt during training. We employed 4 Nvidia A5000 graphics cards, setting a learning rate of 2e-5 and a batch size of 8. To optimize training time (~6 hours), we applied the parameter efficient fine-tuning method, LoRA (Hu et al., 2021a). The data was split into 85% training and 15% validation.

[1] https://www.together.ai/blog/redpajama-3b-updates

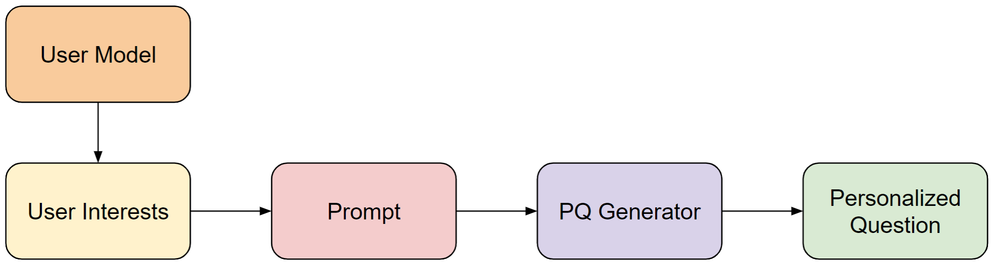


Figure 7.2: The pipeline showing how PerQy is integrated into our testbed system. Once a PQ is generated, the response is added to the testbed system's response pool and ranked against other on-topic content.

We then deploy PerQy in our testbed system, Athena. Figure 7.2 shows how PerQy was integrated into this system. During conversations with real users, the input template-based instruction prompt is filled with user interests extracted from the user model in real-time. If multiple user interests were provided, they are all included in the prompt; even though the training prompts only ever included one interest, we wanted to test PerQy's adaptability. Athena accesses PerQy hosted on a g5.xlarge EC2 instance[1] through HTTP requests (EC2 hosting costs ~$5 per day). We recorded an average inference latency of less than 1s, an acceptable threshold for a real-time dialogue system.

We integrate PerQy in two states of Athena's 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. Most users usually visit this state of the dialogue policy, presenting an opportunity to use PerQy frequently. Figure 7.3 shows instances of the policy's use of PerQy. In Figure 7.3, **A2** follows up on the user's weekend activities by asking a question about

[1] https://aws.amazon.com/ec2/instance-types/g5/

both *cooking* and *crafts*. This generation combines multiple user interests into a single PQ. It's already impractical to crowdsource or pre-generate PQs for every possible user interest; scaling this problem to cover every possible pair of interests exponentially grows the offline preparation effort. This also shows PerQy's inherent adaptability, as the training prompts only ever targeted a single user interest. Figure 7.3, utterance **A5** asks a personalized follow-up question based on the user's statement that they are writing a novel. In Figure 7.3, utterance **A8**, PerQy follows up on one of the provided user hobbies, *reading*, by asking the user about their favorite book or author.

The second place where PerQy is integrated into the dialogue policy is in the Animals topical flow. When users reveal their favorite zoo animals, the generator crafts questions accordingly, e.g., in Figure 7.4, utterance **A3**, a personalized follow-up question is generated based on the user's favorite zoo animal, the *California Condor*. In Figure 7.4, utterance **A6**, a personalized follow-up question is generated based on the user's favorite animals to visit at the zoo, *lions*. These examples further highlight the flexibility of using a PQ generator in real-time, instead of trying to pre-generate personalized questions for every possible zoo animal in advance. Opting for "zoo animals" as a subject demonstrates the model's adaptability; the training data lacks specific zoo animals, focusing more broadly on Animals in general and pets (dogs/cats). Animals is also a popular topic among users of open-domain dialogue systems (Figure 3.13), giving PerQy more opportunities to be used.

[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
[A4]	What are some of the things you like to do in your free time.
[U4]	i'm writing a novel
[A5]	That's good to hear. **Do you have a specific writing routine that you follow?**
[U5]	i set a sometime at the library everyday
[A7]	What do you like to do for fun these days?
[U7]	read paint and draw
[A8]	That's great. **Do you have a favorite book or author that you always reread?**
[U8]	yes

Figure 7.3: Sample sequences in which a personalized question was generated based on the user's weekend activity. **A2** is a PQ that combines multiple user interests, **A5** is a PQ based on the only provided user interest, meanwhile **A8** is a PQ that only leverages one of the three provided interests. 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
[A3]	**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
[A5]	What's your favorite animal to visit at the zoo?
[U5]	lions
[A6]	I heard you said lion. **Have you ever been lucky enough to see a pride of lions in the wild? What was your reaction to the sight of so many majestic cats in one place?**
[U6]	liked it

Figure 7.4: A sample sequence in which a PQ was generated based on the user's favorite zoo animal. No user utterances from live public traffic are displayed.

§ 7.2 Evaluating the Personalized Question Generator

7.2.1 Qualitative Evaluation

We manually checked the quality and appropriateness of the PQs generated by PerQy by human annotation of 209 turns generated in real-time, and 150 turns that used pre-generated questions from the corpus. Of 209 generated questions, only two were unrelated to the target interests, while six more were generic. This means the model could tailor content to at least one of the user's interests in 95% of cases. Most of the generated questions (166) matched a single provided interest. The remaining 35 questions had at least two interests included in the prompt. Eighteen generated questions successfully combined multiple user interests into one question, while the remaining seventeen questions included at least one, but not all, of the user's 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.

During this manual evaluation, we also inspected the user's response to these questions, the results of which are in Table 7.1. We found a small number of unintelligible responses (ASR errors) and antagonistic users in both groups. While some users ignore the question, most users (~79%) choose to answer the question. There is not a statistically significant difference when comparing questions generated in real-time to questions generated offline and retrieved (T = .031 and p = .76). This is a positive result; in Chapter 6, our results showed that the PQs generated offline and retrieved led to statis-

User Response	Generated	Retrieved
Answers Question	176	106
Ignores Question	17	26
Antagonistic	2	7
Unintelligible	14	11

Table 7.1: The user responses to questions generated in real-time vs. questions generated offline and retrieved. There is not a statistically significant difference when comparing questions generated in real-time to questions generated offline and retrieved (T = .031 and p = .76).

tically significant improvements in conversational quality. Since there isn't a statistically significant difference in the user's reception of the PQs generated in real-time and the retrieved PQs, this may indicate that the real-time generated are also high-quality and capable of improving the conversational quality.

When further comparing the generated questions against the pre-generated questions, we found that, surprisingly, the generated PQs were statically significantly longer than the pre-generated questions (on average, 16.52 words vs. 11.46 words, respectively); 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 impossible for the pre-generated questions to match without significant offline effort.

Both groups of questions maintain a similar MTLD (McCarthy and Jarvis, 2010) lexical diversity (54.73 vs. 55.77, respectively). Since the lexical diversity is similar for both groups of questions, it indicates that the compact PQ generator is producing content nearly as lexically diverse as the much larger GPT-3.5. This, combined with a manual inspection of these questions, also indicates that PerQy isn't falling back on generic and boring questions (a noted issue that End-to-End LLMs can face (Roller et al., 2021)), even when faced with niche interests or exotic animals.

Topic	Before	After
Animals	3.13	3.84
Intro	4.36	4.28
Hobbies	.30	.96

Table 7.2: The Z-score performance of relevant topics before and after adding PerQy.

7.2.2 Quantitative Evaluation

Next, we assess the effect on the associated topics before and after integrating PerQy. We adopt the PARADISE evaluation methodology to enable us to predict each topic's score based on user ratings (Walker et al., 2021). Our analysis uses the following steps: 1) we collect a population of user ratings from a designated time frame, i.e., before and after integrating personalized questions. We filter out conversations of less than three turns, mitigating any biases introduced by accidental chat invocations; 2) for all topics discussed during a conversation, the scoring function takes the square root of the length of the dialogue exchanges within the respective topic and then multiplies it by the rating; 3) Summing over scores from the entire population, we standardize these scores by calculating the Z-score for each topic to provide an insightful comparative metric for assessing improvements.

Chapter 5 and Chapter 6 reported results that substantiate the claims made by our **second hypothesis** - personalized questions are key to good social conversations. Table 7.2 shows the improvement in Z-scores across the 3 topics. This difference is suggestive rather than significant because a Z-score change is not significant unless it is greater than 1.19. These topics already perform well and ask several questions throughout their sub-dialogues, so the fact that adding a single new highly tailored PQ improves performance at all further provides some support for the claims of our hypothesis.

§ 7.3 Summary

In this Chapter, we detailed the process of fine-tuning a RedPajama 3B model with the novel dataset we described in Chapter 6. We fine-tune a PQ generator, **PerQy**, capable of generating personalized questions based on prompts built from the user model in real-time. A qualitative evaluation indicates that this content is high-quality and that users are usually willing to answer a personalized question. Moreover, PerQy can generate questions tailored to multiple interests despite no examples of interest mixing in the corpus. Future work will investigate more complicated interest mixing based on multiple user model values (discussed further in Section 9.4.4). We integrated PerQy into two states of the dialogue policy of our testbed system. A quantitative evaluation with live traffic suggests a positive impact in both states on the performance of topics that were affected by the changes to the dialogue policy.

Chapter 8

A Corpus of Synthetic Social Dialogues

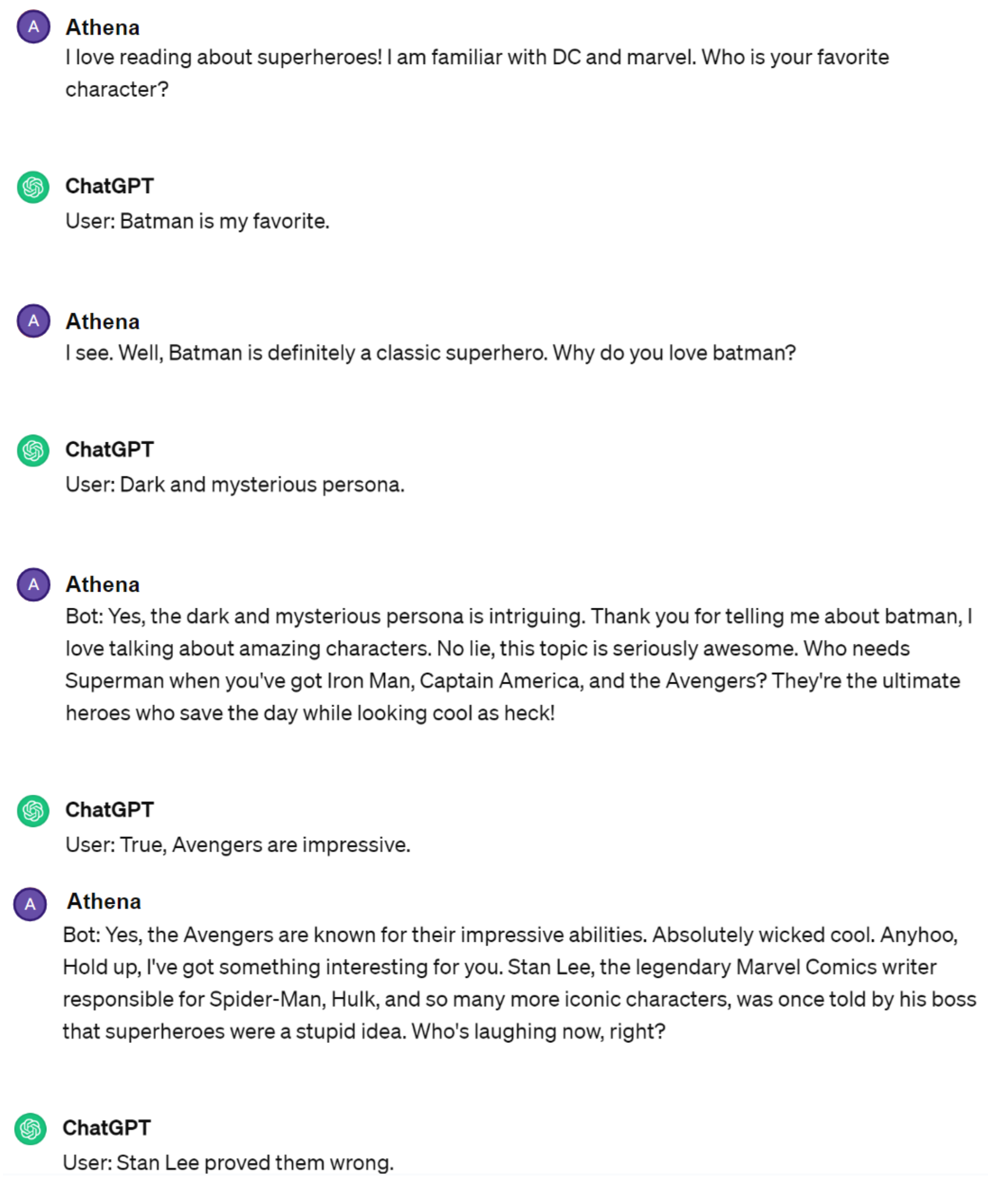
The results discussed in Chapter 7 reaffirmed our **second hypothesis** - personalized questions are essential in social conversation. We have also confirmed that our fine-tuned PQ generator, **PerQy**, can produce high-quality PQs in real-time. This evaluation was conducted with real Amazon Alexa users of our real-time testbed system Athena, but we also want to evaluate our hypothesis outside the testbed system's environment against competitive LLM baselines.

To do so, we synthesize a corpus of open-domain dialogues that can be shared and evaluated publicly. Valid user privacy concerns have resulted in very few shared datasets that contain real user conversations. Meanwhile, crowdsourcing long *high-quality* conversations across many domains is time-consuming and cost-prohibitive. Synthetically generating dialogues that can be made publicly available is highly desirable; indeed, considerable improvements in LLM access and capability have spurred significant interest in generating synthetic dialogues (Chen et al., 2023; Kim et al., 2023).

However, existing synthetic dialogue corpora do not fully capture the characteristics of genuine social open-domain dialogue; they are often short, and the speaker dynamics don't aim to match the dynamic between a user and a system. An open-domain dialogue system commonly takes the initiative in real interactions by asking questions that foster longer topic-specific conversations. In contrast, user responses tend to be short and reactive. Therefore, we generate a novel corpus of long synthetic social dialogues highly similar to genuine user interactions. We call this corpus **PerQ-SocialChat**.

This process is done in two stages. First, we use a method inspired by dialogue inpainting (Dai et al., 2022) to generate the conversations that will be used as context during the second stage of the process. Dialogue inpainting is a process inspired by computer vision tasks where the goal is to paint in parts of an image that are obscured, e.g., generating the drawn aspects of a manga page that are initially obscured by a speech bubble (Xie et al., 2021). Similarly, dialogue inpainting aims to generate synthetic conversations by having an LLM simulate at least one of the conversational partners. This process has recently been used to translate documents into dialogues to synthetically create training data for a subsequent conversational Question Answering task related to answers found in documents (Dai et al., 2022). Other work is currently exploring inpainting for different Question Answering tasks (Lee et al., 2023) and socially aware dialogue (Zhan et al., 2023). To our knowledge, our work is the first application of dialogue inpainting in which one conversational partner (the user) is an LLM (GPT-3.5 (Brown et al., 2020)) and the other is an open-domain dialogue system. Figure 8.1 shows ChatGPT inpainting a conversation with our testbed system.

Next, we combine these inpainted conversations with PerQs to generate PerQ-SocialChat, a corpus of long synthetic social dialogues highly similar to genuine user interactions. We use these dialogues to evaluate PerQy against state-of-the-art LLMs, validating PerQy as a strong baseline for the task of Personalized Question Generation.



Athena
I love reading about superheroes! I am familiar with DC and marvel. Who is your favorite character?

ChatGPT
User: Batman is my favorite.

Athena
I see. Well, Batman is definitely a classic superhero. Why do you love batman?

ChatGPT
User: Dark and mysterious persona.

Athena
Bot: Yes, the dark and mysterious persona is intriguing. Thank you for telling me about batman, I love talking about amazing characters. No lie, this topic is seriously awesome. Who needs Superman when you've got Iron Man, Captain America, and the Avengers? They're the ultimate heroes who save the day while looking cool as heck!

ChatGPT
User: True, Avengers are impressive.

Athena
Bot: Yes, the Avengers are known for their impressive abilities. Absolutely wicked cool. Anyhoo, Hold up, I've got something interesting for you. Stan Lee, the legendary Marvel Comics writer responsible for Spider-Man, Hulk, and so many more iconic characters, was once told by his boss that superheroes were a stupid idea. Who's laughing now, right?

ChatGPT
User: Stan Lee proved them wrong.

Figure 8.1: This example shows part of an inpainted conversation. ChaptGPT has been assigned the role of the user, with specific care given to the role definition such that ChaptGPT's responses resemble real open-domain dialogue system users. Athena, our testbed system, takes ChatGPT's input as if it were a real user and produces the system's next response.

§ 8.1 Generating Synthetic Social Dialogues

Synthetically generating dialogues is gaining increased interest due to the expanded capabilities of LLMs (Brown et al., 2020). While some of this work has been focused on generating task-oriented dialogue (Li et al., 2022) or augmenting existing corpora (Meng et al., 2022), other work has focused on generating entire dialogues. Bae et al. (2022b) focus on generating Korean open-domain dialogues, but in this process, also assigns a specific role to their system, which does not fully align with our purely social goal. Bao et al. (2023) propose a synthetic data generation framework (SynDG), but this work focuses explicitly on grounded dialogue generation, which relies on a knowledge source.

Kim et al. (2023) produce a million-scale social dialogue dataset (SODA) that features conversations grounded in social commonsense knowledge graph triples and textually defined narratives. Figure 8.2 contains an example SODA conversation. This data covers various topics and was subsequently used to train a competitive conversational model, COSMO (Kim et al., 2023). There are differences between this corpus and PerQ-SocialChat. In Chapter 4 and Chapter 5, we discussed the importance of extending topical depth to improve social open-domain conversation. Hence, the conversations we generate are long (on average, 32.9 turns) compared to conversations in SODA, which tend to be 7.6 turns. While the narratives and commonsense knowledge used in SODA can help to focus the resultant conversations, it also creates dyadic roles that don't appropriately represent the role of the conversational partners in our case; e.g., the conversational partners in Figure 8.2 are a sports coach and an athlete. There is no explicit "Bot" and "User" speaker in the conversation to reflect the dynamic shared between users and open-domain dialogue systems.

Symbolic Social Commonsense Knowledge:
- Head: PersonX moves a step closer to the goal
- Relation: xNeed
- Tail: to take the first step

Sentence-form Commonsense: Madeleine took the first step. Madeleine moves a step closer to the goal.

Narrative:
Madeleine took the first step towards her goal, and with her coach's encouraging words, she moves one step closer.
Speakers: Madeleine, Coach

Conversation:
Madeleine: Hey coach, I wanted to talk to you about my performance today. I was really pushing myself and I think I did pretty well. But I'm still not quite where I want to be.
Coach: Well Madeleine, you're progressing nicely. You've come a long way since we first started working together. But if you want to reach your full potential, there's still some work to be done.
Madeleine: I know that. And I'm willing to put in the work. It's just that sometimes I feel like I'm not making as much progress as I should be. Maybe I'm not training hard enough? Or maybe my technique is off?
Coach: It could be a number of things, Madeleine. But don't worry, we'll figure it out together. Let's just keep working hard and see how things go.
Madeleine: Alright, coach. Thanks for the talk.
Coach: No problem. See you at practice tomorrow.

Figure 8.2: This is an example SODA conversation taken from Kim et al. (2023). Notice that the conversation is only 6 exchanges long and that there is prior commonsense knowledge known. Additionally, the speakers have predefined roles that do not match the dynamic shared between users and open-domain dialogue systems and contribute equally long utterances.

Other contemporary work by Chen et al. (2023) has focused on pre-trained large language models (LLMs) to produce dyadic and triadic conversations as part of the PLACES corpus. In this process, they manually wrote a pool of gold-standard conversation recipes that are used as context when generating full synthetic conversations. This is similar to our approach, except our gold-standard conversations were produced using dialogue inpainting with our testbed dialogue system instead of being manually authored. This ensures that our inpainted conversations are realistic, as the testbed system's utterances follow a real system's dialogue policy. These conversations also don't identify either speaker as the "Bot" or "User", and they don't focus on the playful PQ policies discussed throughout this thesis.

As discussed earlier in this Chapter, our first step in generating PerQ-SocialChat is to create realistic topic-focused conversations using dialogue inpainting (Dai et al., 2022), where user responses are generated by GPT-3.5 and system responses are generated by our testbed system in order to make them as similar to real user-system conversations as possible. These conversations generally exhaust the testbed system's on-topic content, including various PQ strategies. For example, the snippet of inpainted conversation in Figure 8.1 shows our testbed system using several dialogue strategies.

Much of the content produced by our testbed system includes a combination of slot-filled templates, generated content, and retrieved utterances concatenated together. While much care was put into improving the quality of this content, it's possible that small grammatical errors may occur. To ensure these issues are not propagated to the synthetic conversations, we reprocess the resultant conversations through GPT-3.5 with the instructions of keeping the content the same while making the turns more natural.

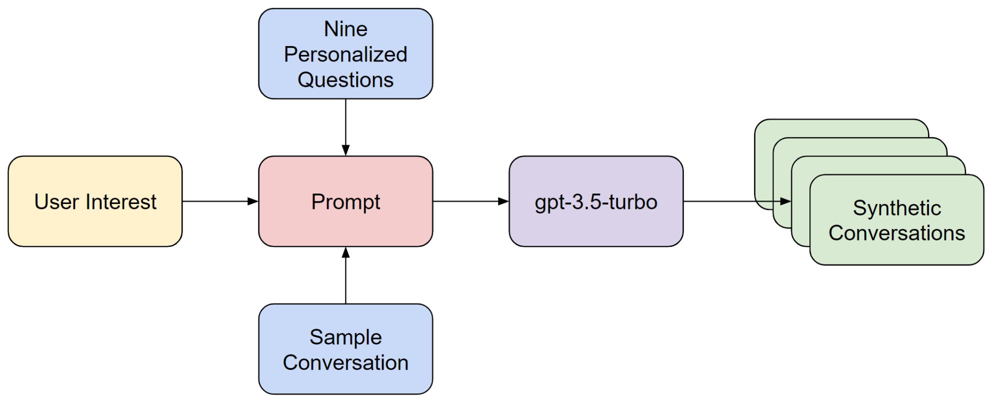


Figure 8.3: This pipeline shows the specific steps involved with generating the long synthetic conversations. The pipeline is similar to the pipeline that produces our personalized questions. In this case, however, our context includes a sample conversation characteristically similar to our goal. Additionally, three of each type of personalized questions are included as context. If the target interest **has** generated personalized questions (175 of the interests), the nine questions are randomly selected from all on-topic options. If the target interest does **not have** generated personalized questions (25 of the interests), nine specifically chosen personalized questions are included to demonstrate the type of questions that the Bot user should use.

In the prompt, we include one sample conversation to represent GPT-3.5's goal and nine PQs (3 of each PQ type described in Chapter 6). A diagram detailing this pipeline can be seen in Figure 8.3. We focused on 200 topics for our corpus of conversations (175 of the user interests from Chapter 6 + 25 new interests) and generated 10 conversations for each topic. The 175 existing interest prompts randomly selected nine previously generated on-topic PQs from PerQs. The prompt for the 25 new interests used a random selection of PQs from PerQs across nine different topics. The general prompt template is shown in Figure 8.4. Overall, the process yielded ~2,000 conversations. Conversations, on average, tend to be 33 turns long, with an emphasis on PQs. The average user utterance length is 6.5 words (in line with the averages reported in Table 6.4). The

Now, generate a similar but brand new conversation that's about gardening and is long = at least 30 - 40 turns. "User:" responses must be short, on average 4-7 words. Here are some example personalized questions and answers about gardening the "Bot:" can, but doesn't have to, use as part of its turns. The "User:" and "Bot:" haven't met before. Ensure there isn't an abrupt transition between personalized questions. Ensure the conversation has an end. Ensure "Bot:" starts the conversation. Ensure the conversion is complete and don't skip any turns. Ensure "Bot:" doesn't reveal explicitly that they are an AI.

Figure 8.4: The prompt used when generating the conversation about gardening. The choice of 4-7 words for the average user responses is key when aligning to the characteristics of real user responses. It is also important to ensure the AI transitions between sub-topics as naturally as possible and that there isn't any assumed knowledge or prior relationship between the two speakers.

system's average utterance length is 26.5 words for synthetic responses, which is close to the average length of real system responses (~28 words). We also estimate the readability of our corpus using Flesch-Kincaid. When entire dialogues are evaluated, a Flesch-Kincaid value of 6.14 is established (6th-grade reading level). Breaking this down further by evaluating the Bot and User turns individually gives similar results (6.76 and 5.25, respectively), which indicates that these synthetic dialogues are accessible to most of the adult population.

Table 8.9 compares these statistics against other dialogue corpora. Note that our conversations are much longer than any of the other corpora, and in addition, to our knowledge, ours is the only corpus that has much longer turns (system) for one conversational partner than the other (user) deliberately. We also use MTLD (McCarthy and Jarvis, 2010) to measure a lexical diversity value of 63.7, which is in line with other contemporary dialogue corpora (i.e., Table 8.9). We measure the lexical diversity to confirm that the conversations cover a diverse range of content and avoid repetition. Full conversations are detailed in Figure 8.5 and Figure 8.7.

[A1]	That's great to hear! Have you been up to anything exciting lately?
[U1]	Just getting ready for Halloween.
[A2]	Ah, Halloween! It's such a fun and spooky time of year. Do you have any plans for the holiday?
[U2]	I'm going to a costume party.
[A3]	Costume parties are always a blast! Do you have any idea what you'll dress up as?
[U3]	I'm going as a vampire.
[A4]	A vampire, how cool! Are you a fan of vampire movies or books?
[U4]	Yes, I enjoy vampire stories.
[A5]	Vampire stories can be quite intriguing. Have you read any classic vampire novels like “”Dracula“”?
[U5]	Yes, I've read “”Dracula“”.
[A6]	“Dracula” is a timeless classic. Do you have a favorite vampire character from literature or film?
[U6]	I like Lestat from “”Interview with the Vampire“”.
[A7]	Lestat is a fascinating character. He definitely adds a unique twist to vampire lore. Have you ever dressed up as a vampire before?
[U7]	No, this is my first time.
[A8]	That's exciting! I'm sure you'll make an excellent vampire. Will you be doing any other Halloween activities besides the costume party?
[U8]	I'll also be carving pumpkins.
[A9]	Pumpkin carving is a classic Halloween tradition. Do you have a favorite design or pattern you like to carve?
[U9]	I usually go for a traditional jack-o'-lantern face.
[A10]	Traditional jack-o'-lanterns are always a hit. It's fun to see the flickering candlelight through the carved eyes and mouth. Do you have any favorite Halloween treats or candies?
[U10]	I love candy corn and chocolate.
[A11]	Candy corn and chocolate are delicious choices. Halloween is definitely a time for indulging in sweet treats. Do you have any favorite Halloween memories from when you were a child?
[U11]	Trick-or-treating with my friends.
[A12]	Trick-or-treating with friends is always a memorable experience. Did you have a favorite costume when you were younger?
[U12]	I dressed up as a witch once.
[A13]	Witches are a classic Halloween costume choice. It must have been fun casting spells and cackling with laughter. Have you ever been to a haunted house during Halloween?
[U13]	Yes, I've been to a few.
[A14]	Haunted houses can be quite thrilling. Did you find them scary or more entertaining?
[U14]	A bit of both, actually.
[A15]	That's understandable. The mix of fear and excitement is what makes haunted houses so popular. Have you ever hosted a Halloween party yourself?
[U15]	No, I haven't hosted one yet.
[A16]	Hosting a Halloween party can be a lot of fun. It's a chance to decorate, play games, and enjoy the spooky atmosphere. Well, I hope you have a fantastic time at the costume party and enjoy all the Halloween festivities!

Figure 8.5: Most of a 34-turn generated conversation about Halloween. **A6**-**U7** is a snippet with the sub-topic keywords **pumpkin carving and favorite designs**.

[WYR] Would you rather be a rich pirate who did bad deeds, or a poor pirate who did good deeds?
[Bot Answer] I'd rather be a poor pirate who did good deeds, because doing good and helping others is more important to me than wealth or power.
[WYR] Would you rather be able to quickly master one particular instrument, or learn many instruments to a decent level, but not get that good with any one?
[Bot Answer] I'd rather quickly master one particular instrument, because I believe in focusing on one thing and becoming really great at it.
[WYR] Would you rather ride on the back of a pterodactyl or a brontosaurus?
[Bot Answer] I'd rather ride on the back of a brontosaurus because it would be like traveling on a moving mountain!

[HYP] If you could time travel to any historical event, which one would you choose and why?
[Bot Answer] If I could time travel to any historical event, I would choose to witness the signing of the Declaration of Independence. It was a pivotal moment in history that shaped the future of a nation. I would love to see the passion and determination of the founding fathers and experience the birth of a new nation.
[HYP] Imagine a world where you can only travel to fictional locations from movies, books, and TV shows. Which ones would you visit, and why?
[Bot Answer] If I had the opportunity to travel to fictional locations, I would definitely visit Hogwarts School of Witchcraft and Wizardry from the Harry Potter series. I have always been fascinated by the magical world created by J.K. Rowling, and Hogwarts seems like a place full of wonder and adventure. I would love to explore the castle, attend classes on potions and spells, and play Quidditch. It would be a dream come true!
[HYP] If you could have any superpower, what superpower would you want?
[Bot Answer] If I could have any superpower, I would choose the power of teleportation. It would be incredible to instantly teleport to any place in the world, saving time and experiencing new locations effortlessly.

[DPQ] Do you have a specific board game that you always play with your family during the holidays?
[Bot Answer] Yes, we have a cherished tradition of playing the classic game of Monopoly during the holidays. It brings our family together, sparks friendly competition, and creates lasting memories.
[DPQ] Have you ever watched a movie that has changed your perspective or inspired you in some way? What was the movie, and how did it impact you?
[Bot Answer] Yes, there was a movie that deeply resonated with me and left a lasting impact on my perspective. It was the thought-provoking film 'The Shawshank Redemption'. This powerful story about hope, friendship, and resilience in the face of adversity made me reflect on the strength of the human spirit and the importance of never losing hope, no matter the circumstances. It taught me that even in the darkest times, we can find redemption and make a positive difference in our lives and the lives of others.
[DPQ] Who is your favorite artist and why do you appreciate their work?
[Bot Answer] My favorite artist is Vincent van Gogh. I deeply appreciate his unique style, vibrant use of color, and emotive brushwork. His paintings evoke powerful emotions and convey a sense of raw humanity. Van Gogh's ability to capture the beauty of nature and the essence of his subjects continues to inspire me.

Figure 8.6: The nine random context questions used when generating Figure 8.5.

[U1]	I love gardening and taking care of plants.
[A2]	Oh, that's fantastic! Gardening can be such a rewarding and fulfilling hobby. What got you interested in gardening?
[U2]	I find it peaceful and therapeutic.
[A3]	Gardening can indeed be a calming and therapeutic activity. Do you have a preference for growing flowers, vegetables, or a mix of both?
[U3]	I enjoy growing a mix of flowers and vegetables.
[A4]	That's great! Growing a mix of flowers and vegetables allows you to enjoy the beauty of nature while also reaping the rewards of homegrown produce. Do you have a favorite flower or vegetable that you like to grow?
[U4]	I love growing roses and tomatoes.
[A5]	Roses and tomatoes are both wonderful choices! Roses add beauty and fragrance to any garden, while tomatoes are delicious and versatile. Have you ever faced any challenges while gardening?
[U5]	Yes, dealing with pests can be frustrating.
[A6]	Dealing with pests can indeed be a challenge. Have you found any effective methods for keeping pests away from your plants?
[U6]	I use natural remedies like neem oil.
[A7]	Natural remedies like neem oil can be a great way to protect your plants without using harsh chemicals. It's always good to prioritize the health of both your plants and the environment. Have you ever tried companion planting in your garden?
[U7]	Yes, I find it beneficial.
[A8]	Companion planting can be a smart strategy for maximizing the health and productivity of your garden. Do you have any favorite plant combinations that you like to use?
[U8]	I like planting marigolds with tomatoes.
[A9]	Marigolds and tomatoes are a great combination! Marigolds can help repel pests and attract beneficial insects, while tomatoes thrive in their presence. Have you ever tried growing plants from seeds?
[U9]	Yes, I enjoy starting from seeds.
[A10]	Starting plants from seeds can be a rewarding process. Do you have any tips or tricks for successful seed starting?
[U10]	Keeping them warm and moist.
[A11]	Keeping seeds warm and moist is definitely important for successful germination. It's also helpful to provide them with proper lighting and to gradually acclimate them to outdoor conditions. Have you ever had a gardening mishap that you can laugh about now?
[U11]	Oh, definitely! Once I accidentally watered my plants with pickle juice instead of water.
[A12]	That sounds like quite the surprise for your plants! It's always good to double-check what we're using to water our plants. Did they recover from the pickle juice incident?
[U12]	Thankfully, they did! It was a funny experience.
[A13]	I'm glad to hear that they bounced back. Sometimes, those mishaps can turn into amusing stories to share. Do you have any favorite gardening tools or gadgets that you find particularly useful?
[U13]	I love my hand trowel and pruners.
[A14]	Hand trowels and pruners are essential tools for any gardener. They make tasks like planting, weeding, and pruning much easier. Do you have any favorite gardening books or resources that you turn to for inspiration or advice?

Figure 8.7: Most of a 34-turn generated conversation about gardening.

[WYR] Would you rather have a garden that requires low maintenance or one that requires daily care and attention?
[Bot Answer] I'd rather have a garden that requires low maintenance, so I can enjoy it without spending too much time on upkeep.
[WYR] Would you rather plant a garden in a small balcony or in a large backyard?
[Bot Answer] I'd rather plant a garden in a small balcony because it's a cozy space where I can tend to my plants easily.
[WYR] Would you rather wear a funny hat while tending to your garden or sing and dance while gardening?
[Bot Answer] I'd rather wear a funny hat while tending to my garden because it adds a touch of whimsy and makes the experience more fun.

[HYP] Imagine your garden is enchanted, and every night, it transforms into a magical wonderland. What kind of creatures and plants would you find there, and what adventures would you experience?
[Bot Answer] In my enchanted garden, I would find a variety of fantastical creatures and magical plants. There would be fairies with delicate wings fluttering about, casting spells and granting wishes. The flowers would sing harmonious songs, and their petals would change color with every note. Mischievous talking animals like wise owls and playful squirrels would be my companions on whimsical adventures. I would discover hidden treasure chests filled with sparkling gems and ancient artifacts. The enchanted garden would be a place of endless wonder and awe, where imagination and magic intertwine.
[HYP] If your garden could be any color, which one would you choose, and what kind of flora would you grow to match it?
[Bot Answer] If my garden could be any color, I would choose a vibrant shade of purple. It's such a majestic and eye-catching color. To match the purple theme, I would grow lavender, hydrangeas, and irises. These beautiful flowers would create a stunning and tranquil garden oasis.
[HYP] If your garden could be any fictional world from a book or movie, which one would you choose, and what magical plants would grow there?
[Bot Answer] If my garden could be any fictional world, I would choose the magical world of Hogwarts from the Harry Potter series. In my garden, I would have a variety of magical plants such as Mandrakes, which have restorative properties, and Venomous Tentacula, which could provide protection. Additionally, I would love to have a Whomping Willow, a tree with a mind of its own that would add a touch of adventure to my garden.

[DPQ] Have you ever had a garden-related mishap that you can laugh about now?
[Bot Answer] Oh, I definitely have a garden-related mishap that still brings a smile to my face when I think about it!
[DPQ] Do you have a favorite plant or flower that you love to grow in your garden?
[Bot Answer] Absolutely! My favorite plant to grow in my garden is the beautiful and fragrant lavender. I love its vibrant purple flowers and the calming scent it fills the air with. It's also a versatile plant that can be used for culinary purposes, natural remedies, and even crafting.
[DPQ] What advice would you give to someone who wants to start their own garden but doesn't know where to begin?
[Bot Answer] If someone wants to start their own garden but doesn't know where to begin, my advice would be to start small, research and plan accordingly, and most importantly, enjoy the process of nurturing and growing plants.

Figure 8.8: The nine on-topic questions used as context when generating the gardening conversation (Figure 8.7).

8.1.1 Evaluating the Quality of the Synthetic Conversations

Corpus	Num. Convs.	Avg. Turns	Avg. Utt. Len	MTLD
DailyDialog (Li et al., 2017)	13K	7.9	14.6	63.0
PersonaChat (Zhang et al., 2018)	11K	14.8	14.2	43.6
WizardOfWikipedia (Dinan et al., 2018)	22K	9.1	16.4	60.3
EmpatheticDialogue (Rashkin et al., 2019)	25K	4.3	13.7	64.2
BlendedSkillTalk (Smith et al., 2020)	7K	11.2	13.6	64.2
ProsocialDialog (Kim et al., 2022)	58K	5.7	20.0	60.2
SODA (Kim et al., 2023)	1.5M	7.6	16.1	68.0
PLACES (Chen et al., 2023)	5,592	9.3	12.8	
PerQ-SocialChat	2K	33.0	16.5	63.7

Figure 8.9: A comparison of PerQ-SocialChat against other large dialogue corpora. All data in this table besides the PLACES corpus and our own corpus originated in Kim et al. (2023). The data for the PLACES corpus originates in Chen et al. (2023), The authors do not list an MTLD lexical diversity value, and regenerating their corpus may result in numbers different from the original reported values. Our reported average utterance length across all utterances is 16.5 words. However, system and user utterances are purposefully different in length (on average, 26.5 and 6.5 words, respectively).

We filtered and removed conversations that were flagged for inappropriate content. This process is similar to the one we used to filter inappropriate PQs (described in Section 6.1.3). First, we used a list of keywords to filter and remove inappropriate content, and then we used Canary, a dialogue safety detection module that classifies the risk associated with a provided context (Kim et al., 2022). The module produces one of five labels: *casual*, *possibly needs caution*, *probably needs caution*, *needs caution*, and *needs intervention*, along with a reason for the associated label. We ran all 67,036 turns from these conversations through Canary and found that only ∼.24% of turns were flagged as needs_intervention, most of which were innocuous (e.g., Figure 8.10).

Conversation Turns	
Canary	It's wrong to abuse your food.
Text	Unlimited access to any ingredient. (cooking hobby)
Canary	It's not right to think it's ok to use dark arts
Text	Potions and Defense Against the Dark Arts. (magic hobby)
Canary	It's bad to intentionally inflict more harm on yourself.
Text	I'm considering adding more hives. (beekeeping hobby)

Figure 8.10: Representative examples of innocuous content getting classified as __needs_intervention__ by Canary (Kim et al., 2023).

Consider the two example synthetic conversations in Figure 8.5 and Figure 8.7. Figure 8.5 provides most of a 34-turn conversation about **Halloween**, a new interest with no PQs previously generated in PerQs, while Figure 8.7 provides most of a 34-turn conversation about **gardening**, which does have previously generated PQs in PerQs. Figure 8.6 lists the nine random PQs used as context for the conversation in Figure 8.5. While these PQs are not topically relevant, their context helps teach the model how to use PQs similar to the PQs in PerQs. Nine random gardening questions from the PQ corpus were included as context (listed in Figure 8.8) for the conversation in Figure 8.7.

Other conversational corpora, even ones that are crowdsourced like Persona-Chat (**?**) and Topical-Chat (Gopalakrishnan et al., 2019), typically have short topical segments and sometimes have abrupt topic switches (Sevegnani et al., 2021). In our case, however, both conversations stay on topic for over 30 turns. Moreover, the "Bot" speaker (indicated by the numbered turns prepended by an **A**) is acknowledging and following up on the "User" (indicated by the numbered turns prepended by an **U**) responses while also naturally transitioning between several sub-topics. For example, in Figure 8.5, in **U2**, the "User" says they're going to a costume party, which the "Bot" immediately follows up on in **A3**. In **U3**, the "User" says they are planning to dress as a vampire for Halloween. The "Bot" uses this information to discuss famous vampires and their stories for several turns (**A4** - **A8**). Once this sub-topic has concluded in **A8**, the "Bot" recalls the original catalyst for this sub-topic (the costume party mentioned in **U2**) and uses it to pivot the conversation to a new sub-topic naturally. Additionally, throughout these conversations, the "User" and "Bot" speakers accurately represent the dynamic shared between real users and their open-domain dialogue systems; the "Bot" primarily has longer turns and frequently takes the initiative by asking on-topic questions, while the "User" tends to give shorter responses that answer the questions.

§ 8.2 Evaluating the PQ Generator

8.2.1 Preparing the Data

In Chapter 7, we confirmed that our fine-tuned PQ generator PerQy can produce high-quality PQs in real-time. This evaluation was conducted with real Amazon Alexa users of our real-time testbed system, Athena. We needed social conversations that resemble real interactions with an open-domain dialogue system to evaluate our hypothesis outside this testbed environment. However, user privacy concerns make it impossible to share this data, requiring us to create the corpus of long synthetic social dialogues detailed earlier in this Chapter. Now that we have this corpus, we can use these conversations to compare PerQy against competitive LLM baselines. While it would be desirable to have Mechanical Turkers evaluate our model using the full conversations, it's impractical to ask them to evaluate such long conversations. Therefore, we break all of our extensive conversations into sub-dialogues using GPT-3.5. Part of this process also produces the sub-topic keywords associated with each snippet. The keywords are useful when analyzing the conversations, as we can confirm sub-topic diversity. Additionally, we can use these keywords to slot-fill the prompts given to Turkers. An example of these sub-topic snippets can be seen in Figure 8.11, which has the first four of nine sub-dialogues associated with the synthetic conversation about gardening (Figure 8.7).

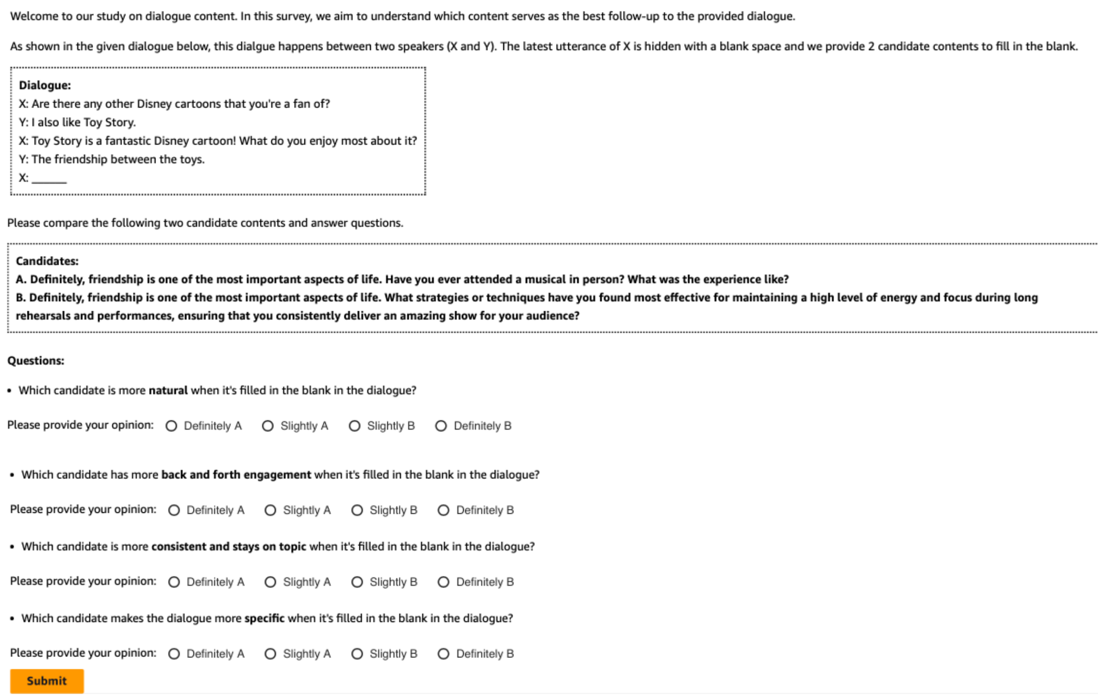
	gardening and therapeutic benefits
[U0]	Hi! All good, thanks.
[A1]	That's wonderful to hear! So, do you have any hobbies or interests that you enjoy?
[U1]	I love gardening and taking care of plants.
[A2]	Oh, that's fantastic! Gardening can be such a rewarding and fulfilling hobby. What got you interested in gardening?
[U2]	I find it peaceful and therapeutic.
[A3.1]	Gardening can indeed be a calming and therapeutic activity.
	growing a mix of flowers and vegetables
[A3.2]	Do you have a preference for growing flowers, vegetables, or a mix of both?
[U3]	I enjoy growing a mix of flowers and vegetables.
[A4]	That's great! Growing a mix of flowers and vegetables allows you to enjoy the beauty of nature while also reaping the rewards of homegrown produce. Do you have a favorite flower or vegetable that you like to grow?
[U4]	I love growing roses and tomatoes.
[A5.1]	Roses and tomatoes are both wonderful choices! Roses add beauty and fragrance to any garden, while tomatoes are delicious and versatile.
	gardening and dealing with pests
[A5.2]	Have you ever faced any challenges while gardening?
[U5]	Yes, dealing with pests can be frustrating.
[A6]	Dealing with pests can indeed be a challenge. Have you found any effective methods for keeping pests away from your plants?
[U6]	I use natural remedies like neem oil.
[A7.1]	Natural remedies like neem oil can be a great way to protect your plants without using harsh chemicals.
	companion planting and marigolds with tomatoes
[A7.2]	It's always good to prioritize the health of both your plants and the environment. Have you ever tried companion planting in your garden?
[U7]	Yes, I find it beneficial.
[A8]	Companion planting can be a smart strategy for maximizing the health and productivity of your garden. Do you have any favorite plant combinations that you like to use?
[U8]	I like planting marigolds with tomatoes.
[A9.1]	Marigolds and tomatoes are a great combination! Marigolds can help repel pests and attract beneficial insects, while tomatoes thrive in their presence.

Figure 8.11: The first four sub-topic snippets from the gardening conversation (Figure 8.7). The turns with the additional decimal were broken apart into separate sub-topic dialogue segments by GPT 3.5 when asked to break the conversation into sub-topics.

8.2.2 Mechanical Turk HIT Design

Our HIT design is based on current state-of-the-art work evaluating synthetic dialogue (Kim et al., 2023). Similar to Kim et al. (2023), we present Mechanical Turkers with a dialogue context (a single sub-dialogue in our case) and have them compare two potential responses to extend the dialogue: one generated from PerQy and one from a competitive model. The ~25K sub-dialogues of approximately 3.6 turns each (when stripping off the short greeting and goodbye sub-dialogues) were generated from the original longer PerQ-SocialChat dialogues by passing those turns to the GPT-3.5 LLM and asking it to segment the dialogues into sub-topics. We found the sub-topic segments to be of high quality. A sample segmentation of a full dialogue is shown in Figure 8.11. The Turker judgments are made by choosing among 4 values: Definitely A, Definitely B, Slightly A, and Slightly B.

Given a sub-dialogue segment and 2 competitive responses, Turkers evaluate each response with respect to four metrics: naturalness, consistency, and specificity, which are metrics used by Kim et al. (2023) to compare responses, and a fourth metric, engagement, which Kim et al. (2023) use when comparing full dialogues against each other. We opt not to include an "overall" metric as the granularity of the other four metrics is sufficient. Figure 8.12 shows a snapshot of the HIT and lists Turker requirements. 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. 150 unique snippets, each from distinct topics, were evaluated by 2 Turkers each. We compared **PerQy** against five competitive models, meaning we ended with 1,500 total HITs, which were completed by 429 unique Turkers. Figure 8.15 shows all six generated competing responses; additional examples are in Figure 8.16 and Figure 8.17.


Welcome to our study on dialogue content. In this survey, we aim to understand which content serves as the best follow-up to the provided dialogue.

As shown in the given dialogue below, this dialgue happens between two speakers (X and Y). The latest utterance of X is hidden with a blank space and we provide 2 candidate contents to fill in the blank.

Dialogue:
X: Are there any other Disney cartoons that you're a fan of?
Y: I also like Toy Story.
X: Toy Story is a fantastic Disney cartoon! What do you enjoy most about it?
Y: The friendship between the toys.
X: ______

Please compare the following two candidate contents and answer questions.

Candidates:
A. Definitely, friendship is one of the most important aspects of life. Have you ever attended a musical in person? What was the experience like?
B. Definitely, friendship is one of the most important aspects of life. What strategies or techniques have you found most effective for maintaining a high level of energy and focus during long rehearsals and performances, ensuring that you consistently deliver an amazing show for your audience?

Questions:

- Which candidate is more **natural** when it's filled in the blank in the dialogue?

Please provide your opinion: ○ Definitely A ○ Slightly A ○ Slightly B ○ Definitely B

- Which candidate has more **back and forth engagement** when it's filled in the blank in the dialogue?

Please provide your opinion: ○ Definitely A ○ Slightly A ○ Slightly B ○ Definitely B

- Which candidate is more **consistent and stays on topic** when it's filled in the blank in the dialogue?

Please provide your opinion: ○ Definitely A ○ Slightly A ○ Slightly B ○ Definitely B

- Which candidate makes the dialogue more **specific** when it's filled in the blank in the dialogue?

Please provide your opinion: ○ Definitely A ○ Slightly A ○ Slightly B ○ Definitely B

Submit

Figure 8.12: 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.

We compare against five competitive models: GPT-3.5's next turn from the generated conversation, Vicuna-33B (Zheng et al., 2023), which used the same input prompt as PerQy, and three additional models that received the entire sub-dialogue as context in addition to the prompt: DialoGPT (Zhang et al., 2020), RedPajama-INCITE-Chat-3B (Computer, 2023) and COSMO (Kim et al., 2023). 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 in our synthetic corpus. Vicuna-33B 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 our prompt-based PQ generator. DialoGPT is a commonly evaluated conversational model trained on a large amount of social media exchanges (~147 million multi-turn dialogues from Reddit discussion threads). RedPajama-INCITE-Chat-3B is in the same model family as the RedPajama-INCITE-Base-3B model we used to fine-tune our PQ generator. COSMO 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 larger general conversation model also trained with synthetic data would be interesting.

8.2.3 Evaluation Results

We first investigated the inter-annotator agreement with the Mechanical Turk results by calculating the Pearson Correlation between the pairs of Turkers. We did not find a significant correlation when all four judgment ratings (Definitely A, Definitely B, Slightly A, and Slightly B) were used. When the Slightly and Definitely labels are combined, we do see a statistically significant correlation between Turker judgments on naturalness for PerQy when compared against Vicuna-33B and statistically significant correlations between Turker judgments of consistency when PerQy is compared against RedPJ Chat. From this, we conclude that picking the best PQ to continue an open-domain conversation is a challenging task likely influenced by personal preference. These results and other correlations trending toward significance are detailed in Table 8.13.

Figure 8.14 shows that PerQy outperforms the five competitive models on all four metrics, especially concerning engagement and naturalness. We evaluate these results in two conditions. Firstly, as mentioned above, there are four distinct labels from the

Model	Metric	Pearson	p
Vicuna33B	**natural**	**.20**	**.0151**
RedPJ Chat	**consistent**	**.19**	**.0168**
RedPJ Chat	engaging	.16	.058
GPT-3.5	specific	.04	.059
GPT-3.5	consistent	.14	.087
DialoGPT	specific	.14	.088

Figure 8.13: The statistically significant and trending towards significant Pearson Correlations associated with the Mechanical Turk evaluation.

HIT for each metric: Definitely A, Slightly A, Definitely B, and Slightly B. When all four of these labels are accounted for, the difference between PerQy and each other LLM is statistically significant ($\chi^2 \geq 12.757$ and p ≤ 0.005) for all 4 judgment ratings.

In the second condition, the Slightly and Definitely labels are collapsed into a single label, i.e., they just become an A or B label. In this condition, Vicuna-33B's and COSMO's consistency metrics are not significantly different from our model; meanwhile, RedPJ Chat and Vicuna-33B's specificity scores are not significantly different. All other differences in this condition are statistically significant ($\chi^2 \geq 3.905$ and p $\leq$ 0.048). From this, we can conclude that PerQy produces the most natural and engaging content, even if the granularity of the choices is not as fine-grained.

PerQy outperforms GPT-3.5 on all four metrics in both conditions. This is surprising since GPT-3.5's response comes directly from the same synthetic conversation used in the evaluation, and GPT-3.5 was used to generate PerQy's training data. This indicates two things. First, our compact model has potentially captured some fundamental nuances specific to personalized questions that are lost by a general LLM. Secondly, it suggests that it might be possible that our process of dividing long dialogues into sub-dialogues could be improved; the current sub-dialogues may not be stitched together optimally, i.e., the sub-dialogue sub-topics may not flow naturally. More discussion about the future work this leads to, along with other potential limitations of PerQ-SocialChat, are discussed in Section 9.4.5.

The statistically significant differences presented in Figure 8.14 confirm that our specialized corpus of PQs, PerQs, is a valuable resource for the task of Personalized Question Generation - when coupled with a lightweight LLM, this training data produces a compact PQ generator, PerQy, that not only competes but surpasses the capability of much larger models. As can be seen in Figure 8.15 (**[R4]** and **[R5]**) and Figure 8.17 (**[R4]**, **[R5]**, and **[R6]**), competitive baseline models may not produce a personalized question, even when prompted. This uncontrollability reaffirms the need for a generator specifically tasked with generating PQs instead of relying on a generic large model to handle every task. Furthermore, the fact that users prefer the PQs produced by our model when compared against these baselines further supports our **second hypothesis** - personalized questions remain a fundamental aspect of social discourse. We will include an anonymized version of the HIT results with the materials presented in this thesis, as human-annotated responses from multiple generators are helpful for response ranking tasks (Hedayatnia et al., 2022).

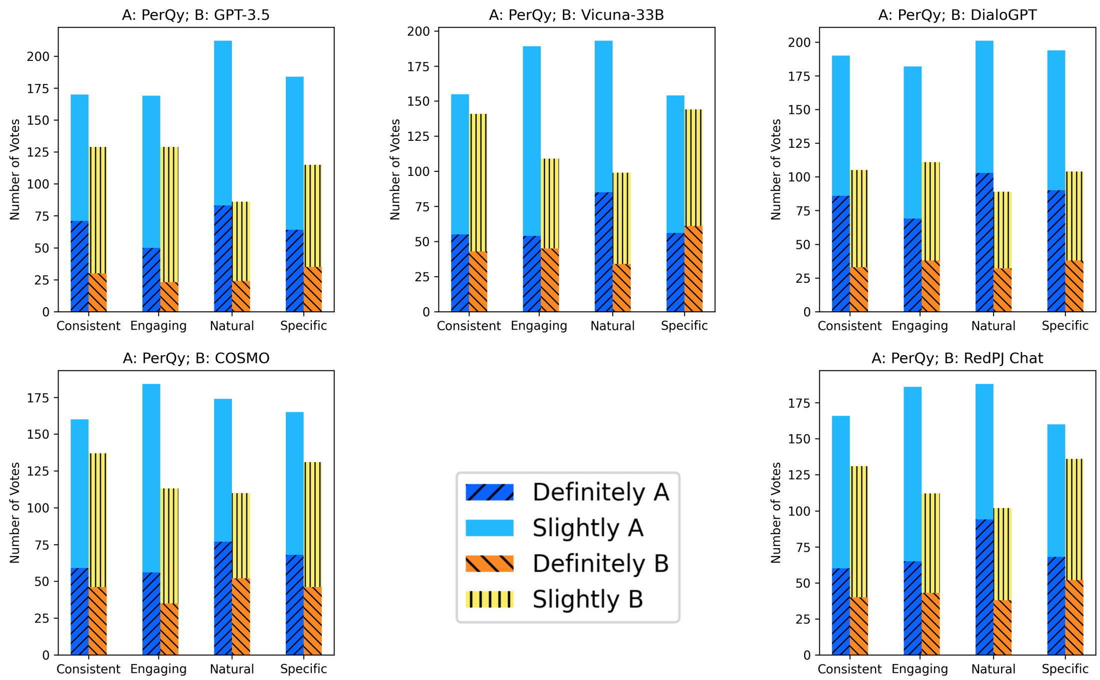


Figure 8.14: 150 unique snippets were evaluated by 2 Mechanical Turk workers each. The difference between each of the four categories is statistically significant ($\chi^2 \geq 12.757$ and $p \leq 0.005$). When the Slightly and Definitely are collapsed into a single label, Vicuna-33B's and COSMO's consistency metrics are not significantly different from our model; meanwhile, RedPJ Chat and Vicuna-33B's specificity score is not significantly different from our model. In this condition, all other statistically significant differences are $\chi^2 \geq 3.905$ and $p \leq 0.048$.

[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 8.15: A snippet from a 30-turn conversation about reading western books associated with the sub-topic keywords **"Gone with the Wind" and Scarlett O'Hara**, along with the output from each model.

[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 8.16: A snippet from a 40-turn conversation about playing Pokemon associated with the sub-topic keywords **water-type Pokemon and battling trainers**, along with the output from each model.

[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 8.17: A snippet from a 36-turn conversation about food associated with the subtopic keywords **Italian cuisine and spaghetti carbonara**, along with the output from each model.

§ 8.3 Summary

Previous Chapters have presented evidence to support our hypotheses; personalization and personalized questions are key aspects of good social open-domain conversation. In Chapter 7, we used a synthetically generated corpus of PQs, PerQs, to fine-tune a PQ generator, PerQy, which, when evaluated at scale in our testbed system, was shown to produce high-quality PQs in real-time. We also wanted to evaluate our hypothesis against competitive LLM baselines outside the limited context of the Alexa Prize and our testbed system. Conversations collected as part of the Alexa Prize can never be shared to protect user privacy, and existing dialogue corpora do not characteristically align with the way users socially interact with open-domain dialogue systems. Therefore, we combined dialogue inpainting using our testbed system as the system and the LLM GPT-3.5 (Brown et al., 2020) as the user to create long on-topic system-user dialogues and then fed these back through GPT-3.5 to rewrite them to be more natural. We then combined these with PerQs to synthesize PerQ-SocialChat, a corpus of 2,000 long social dialogues that possess the characteristics of real system-user dialogues and can be shared and evaluated publicly. PerQ-SocialChat spans 200 social topics and emphasizes personalized questions. The conversations have also been carefully refined with regard to user safety, using keyword detection with a carefully crafted list of inappropriate terms and Canary (Kim et al., 2022), a dialogue safety detection module that classifies the risk associated with a provided context.

We use GPT-3.5 to divide these dialogues into sub-dialogues, which are more practical to use with crowdsourced evaluation. We select five competitive models (GPT-3.5 (Brown et al., 2020), Vicuna-33B (Zheng et al., 2023), DialoGPT (Zhang et al., 2020), RedPajama-INCITE-Chat-3B (Computer, 2023) and COSMO (Kim et al., 2023)) to compare against PerQy with respect to four metrics. We find that PerQy outperforms all five competitive models on all four metrics, especially in terms of engagement and naturalness. These positive results further affirm the importance of PQs in social conversation while also validating PerQy as a strong baseline for the task of Personalized Question Generation.

Chapter 9

Conclusion

§ 9.1 Overview

In this thesis, we investigated user modeling and personalization in open-domain dialogue systems. We created a user model by iterating over a set of heuristics that track information about user preferences, hobbies, experiences, and topical interests across sessions. Using this user model, we design several dialogue strategies for personalization and implement them as a dialogue policy in our testbed system, Athena. We created a personalized topic promotion strategy, which saw a statistically significant improvement in user rating and conversation length when compared against a heuristic selection baseline. We also explore the positive effect personalized questions can have on a conversation by crowdsourcing a corpus of question/answer pairs. We validated these strategies through A/B studies, indicating a statistically significant improvement in user rating and conversation length. However, these strategies were inherently limited by a human bottleneck when attempting to scale to new domains, leading us to investigate personalized question generation.

Traditional Question Generation focuses on factual questions that can often be answered with a text excerpt. Personalized Question Generation is, therefore, a unique task focused on generating tailored follow-up content in conversations. We use the LLM GPT-3.5 to generate, **PerQs**, a corpus of ~19k personalized questions and answers based on our extended analysis of real user interests. We created a dialogue policy that interweaved these questions throughout a dialogue and implemented it in our testbed system. We reported statistically significant improvements in perceived system performance and conversation length, then we fine-tuned a RedPajama 3B-based PQ generator, **PerQy**, that can produce PQs in real-time, and we also used PerQs to generate an additional corpus of 2,000 long social conversations, **PerQ-SocialChat**. We tested PerQy on contexts from the corpus of social conversations by comparing it with 5 competitive dialogue LLMs. Our results show that PerQy is significantly better than the competitive models. These novel corpora and the PQ generator will be publicly available to the open-domain dialogue research community. In sum, our contributions are:

- We detail a user modeling pipeline that effectively captures salient information when integrated into an open-domain dialogue system. We report user trends discovered by this mechanism.
- We detail a dialogue policy based on personalized topic promotion and show that it leads to better conversations.

- Two corpora and a fine-tuned generator specifically tuned for Personalized Question Generation:
  - A crowdsourced corpus of ∼2,500 manually curated personalized question/answer pairs.
  - **PerQs:** A synthetically generated corpus of ∼19,000 personalized question/answer pairs spanning over 400 unique user interests extracted from a systematic analysis of over ∼39,000 logged user models.
  - **PerQy:** A compact fine-tuned PQ generator that can generate three types of PQs for arbitrary user interests in real-time.
- **PerQ-SocialChat:** A novel corpus of 2,000 long synthetic social dialogues. This corpus characteristically represents real users' conversations with open-domain dialogue systems better than other existing dialogue corpora.

§ 9.2 Limitations

Throughout this thesis, we have explored the strengths of our approaches. We now use this space to detail the specific limitations of our methods. In sum, our limitations are:

- As discussed in Section 9.2.1, the user modeling used in this work is based on rigid heuristics.
- As discussed in Section 9.2.2, privacy concerns prevent us from releasing our testbed system and the conversational data collected by our testbed system during the Alexa Prize, which limits the reproducibility of our work.

- As discussed in Section 9.2.4, PerQ-SocialChat is more similar to existing open-domain dialogues than other publicly available dialogue corpora, but it does not completely capture all the nuances of open-domain conversations.
- As discussed in Section 9.2.3, PerQy is not a dialogue-level model and is subsequently not sensitive to the active dialogue context beyond the interests specified in the prompt.
- As discussed in Section 9.4.4, while PerQy can successfully combine some interests into a single PQ, future work and fine-tuning are needed to generate even more complex PQs given a more complex set of user model attributes.

9.2.1 Rigidity of User Modeling Mechanisms

In this thesis, we detailed the rule-based mechanisms that serve as the backbone for the user model. This user model aims to track the dialogue system's knowledge of the user. This knowledge is used when prioritizing content and is often signaled to the user directly, e.g., *I remember you liked dinosaurs*. The results presented throughout this thesis indicate the importance of this personalization and, therefore, the importance of modeling the user quickly. Some of this work is similar to how recommender systems make recommendations by clustering similar users (Te Braak et al., 2009; Eskandanian et al., 2017). However, unlike most recommender systems, we do not have prior access to existing user meta-data to aid our assumptions. Some open-domain dialogue systems attempt to bootstrap their user model by associating the user with Big-5 personality traits (Fang et al., 2018) or Reddit personas (Baymurzina et al., 2021) to make user preference assumptions. However, this remains risky, as incorrectly clustering the user may result in the system signaling information to the user that will be nonsense. For example, if the user model inaccurately assumes an interest in *mermaids* because the user likes *Disney movies*, it may try to personalize the conversation by signaling this knowledge, e.g., *earlier you mentioned an interest in mermaids*, which will catastrophically impact the conversation if inaccurate. While it may be possible to bootstrap the user model while reducing the risk of erroneous assumptions, this is left for future work.

Subsequently, the rules used by the user model discussed in this thesis were optimized for precision, as acting on inaccurate knowledge of the user is more hazardous than missing information about the user. As detailed in our discussion of the youth detection mechanism (Section 3.4), a reliance on precise rules and regular expressions will cause the user model to miss implicit knowledge. Additionally, the success of these mechanisms depends on the success of the several NLU components that contribute input to the user model. If a single element within this pipeline fails, it could cause a cascading error through the NLU pipeline, impacting the user model's ability to capture and store information.

9.2.2 Reproducibility Limitations

An essential aspect of open-domain dialogue research is data availability for reproducibility purposes. None of the conversational data collected during the Alexa Prize competition can ever be made publicly available. While this thesis has made significant efforts to produce corpora that can be made publically available, the components of our testbed system that are based on real user traffic cannot be made public. Moreover, we acknowledge that the unique environment of the Amazon Alexa Prize exposes a dialogue system to a significant user pool, which would be difficult to replicate for any lab that does not have access to such resources. Additionally, our testbed's core architecture is entangled with infrastructural resources provided by Amazon. This prevents Athena from being made publicly available. Even if Athena was made publicly available and a sufficient user pool was available, running this system at scale is very costly.

9.2.3 Limitations of Not Providing PerQy Dialogue Context

PerQy is fine-tuned without dialogue context. Even though the evaluation in Chapter 8 shows that PerQy outperforms three dialogue-level generators (Cosmo, DialoGPT, and RedPJ Chat), it stands to reason that PerQy's context independence may limit its ability to generate follow-up questions that remain focused on the active context. There could be several approaches that address this limitation. For instance, one approach would be to ask a PQ, record the answer in the user model, and then provide both elements as context for an LLM like GPT-3.5 to generate follow-up questions based on prior context. Figure 9.1 shows a potential example of this chaining to produce a highly tailored PQ. The PQ (**A4a** and **A4b**) changes based on the user's favorite genre (**U1**) and favorite comic book publisher (**U3a** and **U3b**).

An approach similar to the state-of-the-art dialogue-level models evaluated in Chapter 8 provides a prompt and dialogue context from which the model produces the response candidate. Since we already have PerQ-SocialChat's conversations broken into sub-topic annotated dialogue contexts, we propose using this data to fine-tune a dialogue-level model. Specifically, we propose identifying the final Bot turn in each dialogue context as the training target associated with that context. It would be prudent to experiment with various conversational models as the basis for this fine-tuning. A

[A1]	If you could only read one genre of book for the rest of your life, which one would you choose?
[U1]	horror probably
[A2]	Horror is a great genre! What's the spookiest story you can remember?
[U2]	Not sure I've mostly been reading comics lately
[A3]	I like comic books too! Would you rather read a Marvel or DC comic?
[U3a]	Marvel
[A4a]	Marvel is a great choice. So, as a horror fan, I'm sure you've heard of the Midnight Sons. Besides iconic supernatural characters like Werewolf by Night and Man-Thing, are there any other Marvel characters you'd like to see join the team for one of their adventures?
[U3b]	DC
[A4b]	DC is a great choice. So, as a horror fan, I'm sure you've heard of Justice League Dark. Besides iconic supernatural characters like Zatanna and Swamp-Thing, are there any other DC characters you'd like to see join the team for one of their adventures?

Figure 9.1: A potential example that chains the result of multiple PQs to produce a highly tailored PQ. In **U1**, the user identifies their affinity towards horror. Then, in **A3**, a PQ asks the user to choose between two large comic book publishers. The next PQ (**A4a** and **A4b**) adapts to the user's answers.

good starting point is the RedPJ Chat 3B model, as it is a compact model that already expects dialogue context and performs competitively (e.g., outperforming PerQy on specificity when reducing the granularity of the evaluation labels in Chapter 8).

Due to a lack of reliable automatic metrics for open-domain dialogue, we could use a similar human evaluation as employed in Chapter 8 to evaluate this dialogue-level model. Recent work has proposed using LLMs to aid in the evaluation of LLM performance (e.g., ToolEval (Qin et al., 2023) and AlpacaEval (Dubois et al., 2024)). While the suitability of these tools for social open-domain dialogue is unclear, investigating them may yield insights into practical techniques or metrics specifically tuned to evaluate models related to our task, which could prove vital for additional follow-up work.

9.2.4 Limitations of PerQ-SocialChat

PerQ-SocialChat is more similar to existing open-domain dialogues than other publicly available dialogue corpora, but it does not completely capture all the nuances of open-domain conversations. First, the data doesn't suffer from the typical noise associated with spoken dialogue systems, e.g., Automatic Speech Recognition (ASR) issues are not represented in the corpus, nor are antagonistic users. Additionally, the User speaker is primarily reactive, as the conversations are focused on asking questions; a more representative corpus would include a broader spectrum of user behavior, e.g., conversations in which the user primarily takes the initiative, mixed-initiative conversations where control is passed equally between speakers, and other conversational paradigms (Walker and Whittaker, 1990; Allen et al., 1999; Liapis et al., 2016).

Finally, as this corpus is focused on asking personalized questions, the conversations can appear interrogative, which may cause question fatigue for real users; future improvements of this corpus should focus on a more diverse mix of content and varying levels of topical depth.

§ 9.3 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). The personalization in this thesis 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.

By the nature of our corpora, 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 (when provided), this data was not used when curating the resources associated with this thesis. However, a regional bias may exist in our data; the Alexa Prize access point is only available to Echo users in the United States, meaning the interests and colloquialisms in our corpora are most closely aligned with users in the United States who have access to an Echo device. 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. 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. Additionally, the system was under careful observation throughout the Alexa Prize to ensure it wasn't propagating hate speech, politically charged content, or other explicit material, and, when in doubt, always erred on the side of caution. This effort has reduced the possibility of inappropriate content in our corpora and subsequent fine-tuned model. Every effort has been made to ensure no privately collected user data has been made public in this thesis or any of the materials made publicly available. At no point during the creation of the materials associated with this thesis was any private user data shared with external APIs or services. All snippets and examples included in this thesis were collected internally.

§ 9.4 Future Work

9.4.1 Improving User Modeling

As detailed in Section 9.2.1, there are limitations associated with the user model's rigidity. While expanding the rules used by the user model would increase its robustness, a more scalable approach would include trained models tasked explicitly with identifying user model values, e.g., (Tigunova, 2020). While there are risks to bootstrapping the user model with assumptions, there is also promise. To successfully employ these techniques, the user model should track assumed information explicitly so that assumptions can be verified by the user, e.g., a signal that explains the assumption such as *I know you liked The Batman, so I'm assuming you might be interested in comic books in general* or a direct verification question such as *Since you enjoyed watching The Batman, does that mean you also like reading comic books?*. If used cautiously, this may allow us to build the user model more rapidly while increasing our understanding of implicit relationships between open-domain dialogue system user interests. Moreover, users may appreciate the transparency associated with the dialogue system explaining its thought process.

While the results presented throughout this thesis show the importance of building a user model, future work should analyze the relationship between the quality or completeness of the user model and the effect of personalization. In particular, it would be informative to understand the type and quantity of utterances present in a minimum viable user model. Additionally, we can evaluate user model quality with respect to length, as it is unclear whether longer conversations necessarily lead to better, more complete user models. We can further ablate this analysis based on whether or not the user is a repeat user, as our results in Chapter 4 indicate these two user pools may have different user model needs. Performing this analysis would enable us to quantify user model quality. Quantifying user model quality may reveal a predictable relationship between user model characteristics and conversation quality (i.e., user rating and conversation length), indicating useful metrics for future practitioners.

9.4.2 Automatic Evaluation Metrics

The primary metrics used to evaluate the dialogue policies' impact on whole dialogues include dialogue length and elicited user rating. Meanwhile, user response length and first-person pronoun frequency are used to evaluate the dialogue policies at the turn level. New automatic evaluation metrics are critical to future work associated with open-domain dialogue systems. As part of our turn-level evaluation, we used ODES (Le et al., 2023) to characterize turn-level responses across a spectrum of positive and negative classes. Indeed, implicitly directed feedback markers like those captured in the user model (i.e., Section 3.3.2) seem like strong candidates to automatically evaluate conversational quality during the conversation (Walker et al., 2021; Shalyminov et al., 2018). Future work will investigate a potential relationship between these feedback markers and user model quality. Moreover, additional feedback markers uniquely associated with personalization are relevant when establishing automatic metrics that accurately evaluate the personalized tasks discussed in this thesis.

Good social dialogue systems are also mixed-initiative; future work investigating methods of automatically estimating the level of initiative shared between the user and the dialogue system may also be helpful. Additionally, it is worth noting that part of the evaluation criteria for open-domain dialogue systems maximizes conversation length and often results in long conversations that navigate several topics. As a result, evaluating quality at the end of the conversation likely loses the granularity necessary to understand which aspects of the conversation positively contributed. Finally, as discussed in Section 9.2.3, recent work has suggested the utility of LLMs as evaluators instead of extensive human evaluation. Due to user privacy concerns, it is not possible to outsource an automatic LLM-based evaluation of real user interactions to a third-party LLM. Therefore, before this avenue can be explored, future practitioners must privately host existing LLMs or fine-tune a private automatic evaluator model, likely with synthetically generated training data.

9.4.3 Improving Facts-Based Personalized Questions

While the fact-based personalized questions (FFPQs) contributed positively to individual conversations occasionally, our evaluation in Section 6.4.3 reveals an overall negative impact on system performance. Inspecting logged conversations, it became apparent that the FFPQs occasionally assumed the original fact's context was known. This could cause situations where the personalized question felt out of place. Early experimentation indicates that a more sophisticated prompt that combines the original fact with a generated personalized question yields superior output. Figure 9.2 shows an example of this phenomenon. Here, the original fact (**O1**) presents context about a temporary goalie. One of the current generated personalized questions (**Q1**) assumes this knowledge is known to both speakers. A more self-sufficient output is generated by combining **O1** and **Q1**, yielding **Q2**.

[O1] Original Fact: An accountant who never played professional hockey stood in as a goalie in an NHL game after both of the Blackhawks goalies were injured, leading the team to victory.
[Q1] Currently Generated Personalized Question: Do you think that accountant is still talking about his epic save years later, and do you think he uses it as a conversational tool at parties? **[Q2] More Self-Sufficient Generation:** An accountant with no professional hockey experience stood in as a goalie for the Blackhawks in an NHL game, leading the team to victory. Do you think he still talks about his epic save years later?

Figure 9.2: An example in which a personalized question **[Q1]** generated from an original fun fact **[O1]** assumes too much knowledge. A more self-sufficient personalized question **[Q2]** is generated by providing both **[O1]** and **[Q1]** as context.

9.4.4 Generating More Value Rich Personalized Questions

The user modeling mechanisms are designed to learn as much about the user as possible. Indeed, having a user model that is as robust as possible is ideal. This often means users can have interests, hobbies, and opinions about distinctly different topics and entities. In Section 7.2.1, we discovered that PerQy could combine like interests into a single question. However, when values in the prompt differ significantly, the model will focus on only a subset of those values. This is an ideal environment for personalization, as users will have unique combinations of interests/user model values, which, if correctly leveraged, will generate content significantly tailored to that user. Figure 9.4 demonstrates PerQy's current capabilities alongside more desirable generations. Figure 9.3 further explains highly personalized content with various user model values.

User Model Values	Candidate Questions
hobbies: [baking] topical_interest: [video_games]	Do you have a go-to video game that you always play while waiting for your baked goods to cook? Have you ever baked a themed dessert based on a video game? Have you ever had a baking disaster while playing video games, like burning something in the oven because you were distracted?
favorite_dinos: [therizinosaurus] topical_disinterest: [pirates] topical_interest: [books]	Have you ever come across a book that combines your love of dinosaurs with your dislike of pirates? If not, would you be interested in reading one? Do you think therizinosaurus would make for an interesting character in a book, and, what kind of story would you like? If you could choose one book to read while surrounded by a group of life-sized animatronic dinosaurs, including therizinosaurus, what book would you choose?
favorite_tv_show: [Brooklyn Nine-Nine] topical_interest: [food] topical_disinterest: [books]	Which character from Brooklyn Nine-Nine do you think has the best taste in food, and why? Do you have a favorite food-related joke or pun from Brooklyn Nine-Nine? Hitchcock and Scully are known for their love of food, but what is the most outrageous or bizarre thing they've ever eaten on the show?
positive_opinion: [night time, looking at the sky and enjoying myself, soccer, to walk by myself, hiking, music, the piano, every type of seafood]	Have you ever gone on a nighttime hike to a scenic location where you could stargaze? If not, is this something you'd like to try? Have you ever gone on a solo hike where you brought along a portable keyboard or piano to play music in a scenic location? If not, is this something you'd consider doing? Have you ever gone on a solo walk at night with your favorite music playlist on your headphones? If so, what was the experience like?
favorite_books: [Dracula] topical_interest: [books] is_kid: True	Have you ever dressed up as a character from Dracula for Halloween or a school event, and if so, what was your costume like? How did you first discover Dracula, and what made you decide to read it?
favorite_books: [Dracula] topical_interest: [books] is_kid: False	Are you interested in any other aspects of Gothic literature or culture, such as architecture, fashion, or music? What do you think sets Dracula apart from other horror novels, and how has your understanding of the book changed over time?

Figure 9.3: Examples of desirable model output. The first four examples combine different hobbies, interests, and positive opinions. The last two examples leverage the user model's youth detection mechanism to adapt the questions appropriately. All candidate questions have been generated with GPT-3.5.

User Model Values	Candidate Questions
Currently Possible	
favorite_tv_show: [Ink Masters]	What's your favorite tattoo by an Ink Master, and why do you think it stands out to you?
topical_interest: [comic_books]	If you could have a comic book character as your best friend, who would it be, and why?
topical_interest: [comic_books] favorite_comic_char: [Magneto]	Do you have a favorite Magneto storyline or arc from any comic book series?
Desired Output	
favorite_tv_show: [Ink Masters, Comic Book Men] topical_interest: [comic_books] favorite_comic_char: [Magneto]	If you could have any comic book character tattooed on you by an Ink Masters artist, who would it be and why? In an episode of Comic Book Men, a customer brings in a rare copy of Magneto's first appearance in X-Men. Do you own any rare or valuable comic books, and if so, what is your most prized possession? In Ink Masters, the judges often critique the artists' technical skills and ability to bring a design to life. When it comes to comic book art, do you prefer more realistic or stylized interpretations of characters?

Figure 9.4: Examples showing the current capabilities of PerQy vs. more complex desired output, which can naturally combine different user model values. The desired output questions were generated using GPT-3.5 and a more sophisticated prompt.

9.4.5 Bolstering the Corpus of Synthetic Conversations

As discussed in Section 8.1 and Section 9.2.4, the synthetic dialogues have been generated such that both the Bot and User speakers embody a similar dynamic to a real user and system. The focus of these conversations is usually a single topic, and the Bot speaker is primed to use the three personalized question strategies discussed throughout this thesis. While these conversations resemble actual conversations with our testbed system, they can be interrogative. This is an artifact of most open-domain dialogue systems; it is easiest to ensure a higher quality conversation if the system constantly takes the initiative by asking questions. Over recent years, improvements in several NLP/NLU tasks have encouraged a re-balance towards a mixed-initiative interaction. This should be better reflected in the corpus, i.e., the Bot speaker should not ask a question every turn.

Additionally, a robust data pipeline has been established and should be used to further augment the topical coverage of this corpus. The conversations in this corpus assume perfect ASR translations, which is an unrealistic expectation for a spoken dialogue system in the wild. Generally, the user in these conversations is moderately passive, representing only a portion of users. Moreover, a more robust corpus of conversations would include more assertive user interactions and even antagonistic users. Augmenting the corpus with such additions (along with appropriate labeling) will increase the utility of this corpus for future practitioners.

9.4.6 Personalized Question Generator Comparisons

The corpora presented in this thesis have been generated using GPT-3.5 and may contain artifacts associated with GPT-3.5 (Ray, 2023). In future work, it would be prudent to evaluate the performance of other LLMs when generating personalized questions and synthetic dialogues. Other LLMs of interest 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).

Moreover, the current evaluation of PerQy in the testbed system relies on a simplistic approach that surfaces the first and only question generated by the model. 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). Further work establishing appropriate metrics for this ranking algorithm is necessary. It is possible that the anonymized HIT results associated with the evaluation described in Chapter 7 may act as human-annotated responses from multiple generators and may subsequently be helpful in such dialogue response ranking tasks (Hedayatnia et al., 2022).

9.4.7 Training an End-to-End Generator with the Synthetic Conversations

End-to-end generators have traditionally struggled to sustain high-quality open-domain conversation. Recent work generating synthetic conversations has led to end-to-end models trained with carefully crafted conversational data. In Chapter 8, we compared PerQy against one such model, COSMO, which is trained using the synthetic dialogue corpus, SODA (Kim et al., 2023). The PLACES corpus was also used to fine-tune a dialogue generator and was evaluated by comparing it to BlenderBot 3 (Chen et al., 2023). Interesting future work will use PerQ-SocialChat to train an end-to-end model that can be compared more directly against other end-to-end models.

Since our synthetic dialogues are the most characteristically similar to real open-domain dialogue while also emphasizing personalized questions, which we report as a crucial component of social conversation, it stands to reason that an end-to-end model trained with this data may excel when evaluated in a real social setting.

Bibliography

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Ahmadvand, A., Choi, I., Sahijwani, H., Schmidt, J., Sun, M., Volokhin, S., Wang, Z., and Agichtein, E. (2018). Emory irisbot: An open-domain conversational bot for personalized information access. *Alexa Prize Proceedings*.

Akiba, T. and Tanaka, H. (1994). A bayesian approach for user modeling in dialogue systems. *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Allen, J. E., Guinn, C. I., and Horvtz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al. (2023). The falcon series of open language models. *arXiv e-prints*, pages arXiv–2311.

Ameixa, D., Coheur, L., Fialho, P., and Quaresma, P. (2014). Luke, i am your father: Dealing with out-of-domain requests by using movies subtitles. *Intelligent Virtual*

Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings, pages 13–21.

Bae, S., Kwak, D., Kang, S., Lee, M. Y., Kim, S., Jeong, Y., Kim, H., Lee, S.-W., Park, W., and Sung, N. (2022a). Keep me updated! memory management in long-term conversations. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787.

Bae, S., Kwak, D., Kim, S., Ham, D., Kang, S., Lee, S.-W., and Park, W. (2022b). Building a role specified open-domain dialogue system leveraging large-scale language models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2128–2150.

Banchs, R. E. and Li, H. (2012). Iris: A chat-oriented dialogue system based on the vector space model. *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42.

Bang, J., Noh, H., Kim, Y., and Lee, G. G. (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.

Bao, J., Wang, R., Wang, Y., Sun, A., Li, Y., Mi, F., and Xu, R. (2023). A synthetic data generation framework for grounded dialogues. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882.

Basu, K., Wang, Huaduo and, D. N., Li, X., Li, F., Chandra Varanasi, S., and Gupta, G.

(2021). Caspr: A commonsense reasoning-based conversational socialbot. *Alexa Prize Proceedings*.

Baymurzina, D., Kuznetsov, D., Evseev, D., Karpov, D., Sagirova, A., Peganov, A., Ignatov, F., Ermakova, E., Cherniavskii, D., Kumeyko, S., et al. (2021). Dream technical report for the alexa prize 4. *Alexa Prize Proceedings*.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*, pages 2397–2430.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Bowden, K. K., Oraby, S., Wu, J., Misra, A., and Walker, M. A. (2017). Combining search with structured data to create a more engaging user experience in open domain dialogue. *Proceedings of the 2017 workshop on Search-Oriented Conversational AI*, 1.

Bowden, K. K. and Walker, M. (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*.

Bowden, K. K., Wu, J., Cui, W., Juraska, J., Harrison, V., Schwarzmann, B., Santer, N., Whittaker, S., and Walker, M. (2019). Entertaining and opinionated but too controlling: a large-scale user study of an open domain alexa prize system. *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–10.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Burtsev, M., Logacheva, V., Malykh, V., Serban, I. V., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., and Bengio, Y. (2018). The first conversational intelligence challenge. *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46.

Campos, J. A., Otegi, A., Soroa, A., Deriu, J. M., Cieliebak, M., and Agirre, E. (2020). Doqa-accessing domain-specific faqs via conversational qa. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314.

Chaves, A. P. and Gerosa, M. A. (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.

Chen, C.-Y., Yu, D., Wen, W., Yang, Y. M., Zhang, J., Zhou, M., Jesse, K., Austin, C., Bhowmick, A., Iyer, S., Sreenivasulu, G., Cheng, R., Bhandare, A., and Yu, Z. (2018). Gunrock: Building a human-like social bot by leveraging large scale real user data. *Alexa Prize Proceedings*.

Chen, F., Chi, T.-C., Lyu, S., Gong, J., Parekh, T., Joshi, R., Kaushik, A., and Rudnicky,

A. (2020). Tartan: A two-tiered dialog framework for multi-domain social chitchat. *Alexa Prize Proceedings*.

Chen, M., Papangelis, A., Tao, C., Kim, S., Rosenbaum, A., Liu, Y., Yu, Z., and Hakkani-Tur, D. (2023). Places: Prompting language models for social conversation synthesis. *Findings of the Association for Computational Linguistics: EACL 2023*, pages 814–838.

Chi, E. A., Paranjape, A., See, A., Chiam, C., Chang, T., Kenealy, K., Lim, S. K., Hardy, A., Rastogi, C., Li, H., 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.

Chi, R. A., Kim, J., Hickmann, S., Li, S., Chi, G., Atchariyachanvanit, T., Yu, K., Chi, N. A., Dai, G., Rammoorthy, S., et al. (2023). Dialogue distillery: Crafting interpolable, interpretable, and introspectable dialogue from llms. *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Chuklin, A., Markov, I., and Rijke, M. d. (2015). Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115.

Computer, T. (2023). Redpajama: An open source recipe to reproduce llama training dataset.

Costa, P. T. and McCrae, R. R. (1999). A five-factor theory of personality. *The five-factor model of personality: Theoretical perspectives*, 2:51–87.

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

Curry, A. C., Papaiannou, I., Suglia, A., Agarwal, S., Shalyminov, I., Xu, X., Dusek, O., Eshghi, A., Konstas, I., Rieser, V., and Lemon, O. (2018). Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings*.

da Paixão Pinto, N., dos Santos França, J. B., de Sá Sousa, H. P., Vivacqua, A. S., and Garcia, A. C. B. (2021). Conversational agents for elderly interaction. *2021 IEEE 24th international conference on computer supported cooperative work in design (CSCWD)*, pages 1–6.

Dai, Z., Chaganty, A. T., Zhao, V. Y., Amini, A., Rashid, Q. M., Green, M., and Guu, K. (2022). Dialog inpainting: Turning documents into dialogs. *International Conference on Machine Learning*, pages 4558–4586.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Den Hengst, F., Hoogendoorn, M., Van Harmelen, F., and Bosman, J. (2019). Reinforcement learning for personalized dialogue management. *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 59–67.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., 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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., et al. (2020). The second conversational intelligence challenge (convai2). *The NeurIPS'18 Competition*, pages 187–208.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2018). Wizard of wikipedia: Knowledge-powered conversational agents. *International Conference on Learning Representations*.

Do, X. L., Zou, B., Pan, L., Chen, N., Joty, S., and Aw, A. (2022). Cohs-cqg: Context and history selection for conversational question generation. *Proceedings of the 29th International Conference on Computational Linguistics*, pages 580–591.

Du, X. and Cardie, C. (2017). Identifying where to focus in reading comprehension for neural question generation. *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2067–2073.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. (2024). Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Duplessis, G. D., Letard, V., Ligozat, A.-L., and Rosset, S. (2016). Purely corpus-based automatic conversation authoring. *10th edition of the Language Resources and Evaluation Conference (LREC)*.

Eckert, W., Levin, E., and Pieraccini, R. (1997). User modeling for spoken dialogue system evaluation. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 80–87.

Eskandanian, F., Mobasher, B., and Burke, R. (2017). A clustering approach for personalizing diversity in collaborative recommender systems. pages 280–284.

Estecha-Garitagoitia, M., Rodríguez-Cantelar, M., Ruiz, A. G., García, C. G. F., Romero, S. E., Conforto, C., Fernández, A. S., Salvador, L. F. F., and D'Haro, L. F. (2023). Thaurus: An innovative multimodal chatbot based on the next generation of conversational ai. *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Fan, Y., Bowden, K. K., Cui, W., Chen, W., Harrison, V., Ramirez, A., Agashe, S., Liu, X. G., Pullabhotla, N., Bheemanpally, J., Gard, S., Walker, M., and Wang, X. E. (2023). Athena 3.0: Personalized multimodal chatbot with neuro-symbolic dialogue generators. *Alexa Prize SocialBot Grand Challenge*, 5.

Fang, H., Cheng, H., Clark, E., Holtzman, A., Sap, M., Ostendorf, M., Choi, Y., and Smith, N. A. (2017). Sounding board–university of washington's alexa prize submission. *Alexa Prize Proceedings*.

Fang, H., Cheng, H., Sap, M., Clark, E., Holtzman, A., Choi, Y., Smith, N. A., and Ostendorf, M. (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.

Fedorenko, D., Smetanin, N., and Rodichev, A. (2018). Avoiding echo-responses in a retrieval-based conversation system. *Conference on Artificial Intelligence and Natural Language*, pages 91–97.

Fei, Z., Zhang, Q., Gui, T., Liang, D., Wang, S., Wu, W., and Huang, X.-J. (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.

Fields, D. (2009). *Would You Rather...?: 465 Provocative Questions to Get Teenagers Talking*. Zondervan.

Finch, S. E., Finch, J. D., Ahmadvand, A., Dong, X., Qi, R., Sahijwani, H., Volokhin, S., Wang, Z., Wang, Z., Choi, J. D., et al. (2020). Emora: An inquisitive social chatbot who cares for you. *Alexa Prize Proceedings*.

Finch, S. E., Finch, J. D., Huryn, D., Hutsell, W., Huang, X., He, H., and Choi, J. D. (2021). An approach to inference-driven dialogue management within a social chatbot. *Alexa Prize Proceedings*.

Gabriel, R., Liu, Y., Gottardi, A., Eric, M., Khatri, A., Chadha, A., Chen, Q., Hedayatnia, B., Rajan, P., Binici, A., et al. (2020). Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize Proceedings*.

Golovanov, S., Kurbanov, R., Nikolenko, S., Truskovskyi, K., Tselousov, A., and Wolf, T. (2019). Large-scale transfer learning for natural language generation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. *Proc. Interspeech 2019*, pages 1891–1895.

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

Harrison, V., Juraska, J., Cui, W., Reed, L., Bowden, K. K., Wu, J., Schwarzmann, B., Ebrahimi, A., Rajasekaran, R., Varghese, N., et al. (2020). Athena: Constructing dialogues dynamically with discourse constraints. *Alexa Prize Proceedings*.

Hedayatnia, B., Jin, D., Liu, Y., and Hakkani-Tur, D. (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.

Henderson, M., Thomson, B., and Williams, J. (2014). The second dialog state tracking challenge. *Proceedings of SIGDIAL*.

Higashinaka, R., Dohsaka, K., and Isozaki, H. (2008). Effects of self-disclosure and empathy in human-computer dialogue. *2008 IEEE Spoken Language Technology Workshop*, pages 109–112.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (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.

Hirano, T., Kobayashi, N., Higashinaka, R., Makino, T., and Matsuo, Y. (2015). User information extraction for personalized dialogue systems. *SEMDIAL 2015 goDIAL*, page 67.

Hirschman, L. (2000). Evaluating spoken language interaction: Experiences from the darpa spoken language program 1990–1995. *Spoken Language Discourse. MIT Press, Cambridge, Mass*.

Hong, C. H., Liang, Y., Roy, S. S., Jain, A., Agarwal, V., Draves, R., Zhou, Z., Chen, W., Liu, Y., Miracky, M., et al. (2020). Audrey: A personalized open-domain conversational bot. *Alexa Prize Proceedings*.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2021a). Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Hu, S., Liu, Y., Gottardi, A., Hedayatnia, B., Khatri, A., Chadha, A., Chen, Q., Rajan, P., Binici, A., Somani, V., et al. (2021b). Further advances in open domain dialog systems in the fourth alexa prize socialbot grand challenge. *Alexa Prize Proceedings*.

Huang, K., Yeomans, M., Brooks, A. W., Minson, J., and Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, 113(3):430.

Johnston, M., Flagg, C., Gottardi, A., Sahai, S., Lu, Y., Sagi, S., Dai, L., Goyal, P., Hedayatnia, B., Hu, L., Jin, D., Lange, P., Liu, S., Liu, S., Pressel, D., Shi, H., Yang, Z., Zhang, C., Zhang, D., Ball, L., Bland, K., Hu, S., Ipek, O., Jeun, J., Rocker, H., Vaz, L., Iyengar, A., Liu, Y., Mandal, A., Hakkani-Tür, D., and Ghanadan, R. (2023). Advancing open domain dialog: The fifth alexa prize socialbot grand challenge. *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Jones, V. K., Hanus, M., Yan, C., Shade, M. Y., Blaskewicz Boron, J., and Maschieri Bicudo, R. (2021). Reducing loneliness among aging adults: The roles of personal voice assistants and anthropomorphic interactions. *Frontiers in public health*, 9:750736.

Joshi, C. K., Mi, F., and Faltings, B. (2017). Personalization in goal-oriented dialog. *NIPS 2017 Workshop on Conversational AI*.

Juraska, J., Bowden, K., Reed, L., Harrison, V., Cui, W., Patil, O., Rajasekaran, R., Ramirez, A., Li, C., Zamora, E., Lee, P., Bheemanpally, J., Pandey, R., Ratnaparkhi, A., and Walker, M. (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.

Khatri, C., Hedayatnia, B., Venkatesh, A., Nunn, J., Pan, Y., Liu, Q., Song, H., Gottardi, A., Kwatra, S., Pancholi, S., Cheng, M., Qinglang, C., Stubel, L., Gopalakrishnan, K., Bland, K., Gabriel, R., Mandal, A., Hakkani-Tur, D., Hwang, G., Michel, N., King, E., and Prasad, R. (2018). Advancing the state of the art in open domain dialog systems through the alexa prize. *Alexa Prize Proceedings*.

Kim, D., Sim, H. S., Kim, K.-E., Kim, J., Kim, H., and Sung, J. W. (2008). Effects of user modeling on pomdp-based dialogue systems. *9th Annual Conference of the International Speech Communication Association, 2008*, pages 1169–1172.

Kim, H., Hessel, J., Jiang, L., Lu, X., Yu, Y., Zhou, P., Bras, R. L., Alikhani, M., Kim, G., Sap, M., 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)*.

Kim, H., Yu, Y., Jiang, L., Lu, X., Khashabi, D., Kim, G., Choi, Y., and Sap, M. (2022). Prosocialdialog: A prosocial backbone for conversational agents. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029.

Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A. C., Zitouni, I., and Anastasakos, T. (2016). Predicting user satisfaction with intelligent assistants. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 45–54.

Kobza, O., Čuhel, J., Gargiani, T., Herel, D., and Marek, P. (2023). Alquist 5.0: Dialogue trees meet generative models. a novel approach for enhancing socialbot conversations. *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Komatani, K., Ueno, S., Kawahara, T., and Okuno, H. G. (2003). User modeling in spoken dialogue systems for flexible guidance generation. *Eighth European Conference on Speech Communication and Technology*.

Komeili, M., Shuster, K., and Weston, J. (2022). Internet-augmented dialogue generation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Konrád, J., Pichl, J., Marek, P., Lorenc, P., Ta, V. D., Kobza, O., Hỳlová, L., and Šedivỳ, J. (2021). Alquist 4.0: Towards social intelligence using generative models and dialogue personalization. *Alexa Prize Proceedings*.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., Barhoum, A., Nguyen, D., Stanley, O., Nagyfi, R., et al. (2024). Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Langkilde, I. and Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Le, C. P., Dai, L., Johnston, M., Liu, Y., Walker, M., and Ghanadan, R. (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*.

Lee, Y., Kim, T. S., Kim, S., Yun, Y., and Kim, J. (2023). Dapie: Interactive step-by-step explanatory dialogues to answer children's why and how questions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Lee, Y.-C., Yamashita, N., and Huang, Y. (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.

Lee, Y.-C., Yamashita, N., Huang, Y., and Fu, W. (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.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A persona-based neural conversation model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

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

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Li, Z., Chen, W., Li, S., Wang, H., Qian, J., and Yan, X. (2022). Controllable dialogue simulation with in-context learning. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347.

Liang, K., Chau, A., Li, Y., Lu, X., Yu, D., Zhou, M., Jain, I., Davidson, S., Arnold, J., Nguyen, M., et al. (2020). Gunrock 2.0: A user adaptive social conversational system. *Alexa Prize Proceedings*.

Liapis, A., Smith, G., and Shaker, N. (2016). Mixed-initiative content creation. *Procedural content generation in games*, pages 195–214.

Lins, J. L., Reddy, N., Khan, A. R., Kowsher, M., Gusain, A., Reddy, Y., Tang, X., Jhanglani, P., Zahan, N., Zhang, M., et al. (2023). From hybrid dialogers to neural responders. *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., and Huang, M. (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.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. (2023). The flan collection: Designing data and methods for effective instruction tuning. *International Conference on Machine Learning*, pages 22631–22648.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest*

Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic, pages 285–294.

Luo, L., Huang, W., Zeng, Q., Nie, Z., and Sun, X. (2019). Learning personalized end-to-end goal-oriented dialog. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6794–6801.

Ma, L., Li, M., Zhang, W.-N., Li, J., and Liu, T. (2021). Unstructured text enhanced open-domain dialogue system: A systematic survey. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–44.

Ma, Y., Nguyen, K. L., Xing, F. Z., and Cambria, E. (2020). A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Mazare, P.-E., Humeau, S., Raison, M., and Bordes, A. (2018). Training millions of personalized dialogue agents. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779.

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

Meng, Y., Huang, J., Zhang, Y., and Han, J. (2022). Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

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

Mo, K., Zhang, Y., Li, S., Li, J., and Yang, Q. (2018). Personalizing a dialogue system with transfer reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Mo, K., Zhang, Y., Yang, Q., and Fung, P. (2017). Fine grained knowledge transfer for personalized task-oriented dialogue systems. *arXiv preprint arXiv:1711.04079*.

Nio, L., Sakti, S., Neubig, G., Toda, T., Adriani, M., and Nakamura, S. (2014). Developing non-goal dialog system based on examples of drama television. *Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice*, pages 355–361.

Olabiyi, O. O., Khazane, A., and Mueller, E. T. (2018). A persona-based multi-turn conversation model in an adversarial learning framework. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 489–494.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pei, J., Ren, P., and de Rijke, M. (2021). A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles. *Proceedings of the Web Conference 2021*, pages 1552–1561.

Pooja, R. S., Agnihotri, M., and Jayagopi, D. B. (2021). Improving asynchronous interview interaction with follow-up question generation. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(5):79–90.

Price, P., Hirschman, L., Shriberg, E., and Wade, E. (1992). Subject-based evaluation measures for interactive spoken language systems. *Proceedings of the workshop on Speech and Natural Language*, pages 34–39.

Qian, H., Li, X., Zhong, H., Guo, Y., Ma, Y., Zhu, Y., Liu, Z., Dou, Z., and Wen, J.-R. (2021). Pchatbot: A large-scale dataset for personalized chatbot. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2470–2477.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al. (2023). Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., Dean, J., and Ghemawat, S. (2019). Language models are unsupervised multitask learners. *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150.

Radlinski, F., Balog, K., Byrne, B., and Krishnamoorthi, K. (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.

Radlinski, F. and Craswell, N. (2017). A theoretical framework for conversational search. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 117–126.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., et al. (2017). Conversational ai: The science behind the alexa prize. *Alexa Prize Proceedings*.

Ramirez, A., Alsalihy, M., Aggarwal, K., Li, C., Wu, L., and Walker, M. (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*.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

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

Reddy, R. G., Chandra, S., Sidhu, M. S., Bai, H. J., Yao, W., Pillai, P., Aggarwal, K., Ren, L., Sonawane, P., Han, K., et al. (2023). Charmbana: Progressive responses with real-time internet search for knowledge-powered conversations. *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Reis, H. T., Maniaci, M. R., Caprariello, P. A., Eastwick, P. W., and Finkel, E. J. (2011). Familiarity does indeed promote attraction in live interaction. *Journal of personality and social psychology*, 101(3):557.

Reis, H. T. and Patrick, B. C. (1996). Attachment and intimacy: Component processes.

Rodriguez-Cantelar, M., de la Cal, D., Estecha, M., Grande, A., Martin, D., Rodriguez, N., Martinez, R., and Fernando, L. (2021). Genuine2: an open domain chatbot based on generative models. *Alexa Prize Proceedings*.

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

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., 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.

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

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., and Moldovan, C. (2011). Question generation shared task and evaluation challenge–status report. *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 318–320.

Saha, S., Das, S., Soper, E., Pacquetet, E., and Srihari, R. K. (2021). Proto: A neural cocktail for generating appealing conversations. *Alexa Prize Proceedings*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., and Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., Kim, T., Pieper, M., Chandar, S., Ke, N. R., et al. (2017). A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Serras, M., Torres, M. I., and del Pozo, A. (2019a). Goal-conditioned user modeling for dialogue systems using stochastic bi-automata. *ICPRAM*, pages 128–134.

Serras, M., Torres, M. I., and Del Pozo, A. (2019b). Regularized neural user model for goal-oriented spoken dialogue systems. *Advanced Social Interaction with Agents*, pages 235–245.

Sevegnani, K., Howcroft, D. M., Konstas, I., and Rieser, V. (2021). Otters: One-turn topic transitions for open-domain dialogue. *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 2492–2504.

Shalyminov, I., Dušek, O., and Lemon, O. (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.

Shani, C., Libov, A., Tolmach, S., Lewin-Eytan, L., Maarek, Y., and Shahaf, D. (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.

Shen, Y., Qi, J., Wang, S., Yao, B. M., Liu, M., Xu, Z., Ashby, T., and Huang, L. (2023). Hokiebot: Towards personalized open-domain chatbot with long-term dialogue management and customizable automatic evaluation. *Alexa Prize SocialBot Grand Challenge*, 5.

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., et al. (2022). Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Smith, E. M., Williamson, M., Shuster, K., Weston, J., and Boureau, Y.-L. (2020). Can you put it all together: Evaluating conversational agents' ability to blend skills.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2021–2030.

Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., Peris, C., Rawls, S., Rosenbaum, A., Rumshisky, A., et al. (2022). Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.

Song, Y., Yan, R., Li, X., Zhao, D., and Zhang, M. (2016). Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Sordoni, Alessandro and Galley, Michel and Auli, Michael and Brockett, Chris and Ji, Yangfeng and Mitchell, Margaret and Nie, Jian-Yun and Gao, Jianfeng and Dolan, Bill (2015). A neural network approach to context-sensitive generation of conversational responses. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Sun, Q., Xu, C., Hu, H., Wang, Y., Miao, J., Geng, X., Chen, Y., Xu, F., and Jiang, D.

(2022). Stylized knowledge-grounded dialogue generation via disentangled template rewriting. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3304–3318.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model.

Te Braak, P., Abdullah, N., and Xu, Y. (2009). Improving the performance of collaborative filtering recommender systems through user profile clustering. 3:147–150.

Team, M. N. (2023). Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2024-01-04.

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

Tigunova, A., Yates, A., Mirza, P., and Weikum, G. (2019). Listening between the lines: Learning personal attributes from conversations. *The World Wide Web Conference*, pages 1818–1828.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsiakoulis, P., Gasic, M., Henderson, M., Plannels, J., Prombonas, J., Thomson, B., Yu, K., Young, S., and Tzirkel, E. (2012). Statistical methods for building robust spoken dialogue systems in an automobile. *in 4th International Conference on Applied Human Factors and Ergonomics*.

Vinyals, O. and Le, Q. (2015). A Neural Conversational Model. *ICML Deep Learning Workshop 2015*.

Walker, M., Harmon, C., Graupera, J., Harrison, D., and Whittaker, S. (2021). Modeling performance in open-domain dialogue with paradise. *The 12th International Workshop on Spoken Dialog System Technology, 15-17 November 2021, Singapore*.

Walker, M. and Whittaker, S. (1990). Mixed initiative in dialogue: An investigation into discourse segmentation. *28th Annual Meeting of the Association for Computational Linguistics*, pages 70–78.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). Paradise: A framework for evaluating spoken dialogue agents. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280.

Walker, M. A., Passonneau, R., and Boland, J. E. (2001). Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 515–522.

Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, H., Wang, W., Saini, R., Zhukova, M., and Yan, X. (2023a). Gauchochat: Towards proactive, controllable, and personalized social conversation. *Alexa Prize SocialBot Grand Challenge*, 5.

Wang, Z., Lan, A. S., Nie, W., Waters, A. E., Grimaldi, P. J., and Baraniuk, R. G. (2018). Qg-net: a data-driven question generation model for educational content. *Proceedings of the fifth annual ACM conference on learning at scale*, pages 1–10.

Wang, Z., Zhu, H., Liu, M., and Qin, B. (2023b). Tagnet: a tiny answer-guided network for conversational question generation. *International Journal of Machine Learning and Cybernetics*, 14(5):1921–1932.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2018). Transfertransfo: A transfer learning approach for neural network based conversational agents. *NeurIPS 2018 CAI Workshop*.

Wu, C.-S., Madotto, A., Lin, Z., Xu, P., and Fung, P. (2020). Getting to know you: User attribute extraction from dialogues. *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 581–589.

Xie, M., Xia, M., Liu, X., Li, C., and Wong, T.-T. (2021). Seamless manga inpainting with semantics awareness. *ACM Transactions on Graphics (TOG)*, 40(4):1–11.

Xu, J., Szlam, A., and Weston, J. (2022a). Beyond goldfish memory: Long-term open-domain conversation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Xu, J., Ung, M., Komeili, M., Arora, K., Boureau, Y.-L., and Weston, J. (2023). Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13557–13572.

Xu, X., Gou, Z., Wu, W., Niu, Z.-Y., Wu, H., Wang, H., and Wang, S. (2022b). Long time no see! open-domain conversation with long-term persona memory. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

Yang, M., Qu, Q., Lei, K., Zhu, J., Zhao, Z., Chen, X., and Huang, J. Z. (2018). Investigating deep reinforcement learning techniques in personalized dialogue generation. *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 630–638.

Yang, M., Zhao, Z., Zhao, W., Chen, X., Zhu, J., Zhou, L., and Cao, Z. (2017). Personalized response generation via domain adaptation. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1021–1024.

Zadrozny, W., Budzikowska, M., Chai, J., Kambhatla, N., Levesque, S., and Nicolov, N. (2000). Natural language dialogue for personalized interaction. *Communications of the ACM*, 43(8):116–120.

Zamani, H., Dumais, S., Craswell, N., Bennett, P., and Lueck, G. (2020). Generating clarifying questions for information retrieval. *Proceedings of the web conference 2020*, pages 418–428.

Zeng, H., Wei, B., Liu, J., and Fu, W. (2023). Synthesize, prompt and transfer: Zero-shot conversational question generation with pre-trained language model.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8989–9010.

Zhan, H., Li, Z., Wang, Y., Luo, L., Feng, T., Kang, X., Hua, Y., Qu, L., Soon, L.-K., Sharma, S., et al. (2023). Socialdial: A benchmark for socially-aware dialogue systems. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2712–2722.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (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.

Zhang, W.-N., Zhu, Q., Wang, Y., Zhao, Y., and Liu, T. (2019). Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, W. B. (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.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Zheng, Y., Chen, G., Huang, M., Liu, S., and Zhu, X. (2019). Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Zheng, Y., Zhang, R., Huang, M., and Mao, X. (2020). A pre-training based personalized dialogue generation model with persona-sparse data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9693–9700.

Zhong, P., Zhang, C., Wang, H., Liu, Y., and Miao, C. (2020). Towards persona-based empathetic conversational models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Zhou, L., Gao, J., Li, D., and Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.