

Dialogue Scaffolding: Producing a Realistic Corpus of Human-Computer Open-Domain Dialogues Using a Spoken Dialogue System and ChatGPT

Kevin K. Bowden and Marilyn Walker

University of California, Santa Cruz

{kkbowden, mawalker}@ucsc.edu

Abstract

Researchers in dialogue interaction have had a long-term interest in multi-domain human-computer conversations and how they differ from human-human conversations. Recently, research on dialogue has begun to rely more and more on corpus-based training of neural conversational models, and conversational LLMs such as ChatGPT. However, existing large open-domain dialogue corpora do not accurately capture the characteristics of social human-computer dialogue. This paper addresses this gap by synthesizing a new corpus of 4000 long social dialogues on 200 user-model based topics that we call **User-Centric SocialChat** (UCSC). We create UCSC with a novel method called Dialogue Scaffolding, where a real dialogue system, that competed successfully in the Alexa Prize, interacts with ChatGPT to generate conversations. The Dialogue Scaffolding method ensures that the dialogues closely resemble the social chat genre of human-computer dialogues. We evaluate UCSC to ensure quality and safety, and we measure lexical diversity and topic consistency to show that the conversations are not repetitive and stay on topic. We evaluate the utility of UCSC by fine-tuning a compact dialogue-level model, **PerQy-DLM**, and showing that it outperforms competitive fine-tuned models like COSMO, Vicuna, and RedPajama-Chat-3B.

1 Introduction

Researchers in dialogue interaction have long been interested in human-computer (HC) conversations and how they differ from human-human (HH) conversations, and there has been a concentrated long-term effort to create both types of dialogue corpora for research use (Consortium et al., 2007; Godfrey et al., 1992; Serban et al., 2018; Walker et al., 2012; Rosset and Petel, 2006; Yu and Yu, 2021; Prasad and Walker, 2002; Oraby et al., 2016) *inter alia*.

Recently, there has been an upsurge of interest in HC social chat dialogue, where informa-

| | |
|------|---|
| [A1] | I love reading about superheroes! I am familiar with DC and marvel. Who is your favorite character? |
| [U1] | Batman is my favorite. |
| [A2] | I see. Well, Batman is definitely a classic superhero. Why do you love batman? |
| [U2] | Dark and mysterious persona. |
| [A3] | 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! |
| [U3] | True, Avengers are impressive. |
| [A4] | 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? |
| [U4] | Stan Lee proved them wrong. |

Figure 1: Part of a Scaffolded social chat dialogue (full dialogue in Figure 11). ChatGPT has been assigned the user role (U), with specific instructions that responses model the statistical properties of real open-domain social SDS users (Bowden et al., 2024). A text-based version of the Athena SDS takes ChatGPT's input as if it were a real user and produces the next response (A).

tion may be exchanged, but there is no explicit task beyond conversing in an engaging and natural way (Tsunomori et al., 2021; Higashinaka et al., 2016; Zhang et al., 2018; Yamashita et al., 2023; Gopalakrishnan et al., 2019). While such research has significantly advanced due to increasingly large corpora and instruction-based tuning of conversational LLMs such as ChatGPT, Vicuna, Llama2, and Cosmo, there are few dialogue corpora that explicitly model real HC social chat dialogues (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2022; Zheng et al., 2023a; GenAI, 2023). Our aim is to create a corpus of long social chat dialogues that are similar to the HC dialogues of an Alexa Prize system (Bowden et al., 2019; Yu

| Row | Corpus | Authors | Language | Topics | Num. Convs. | Avg. Turns | Avg. Len | MTLD |
|-----|-------------------------|------------------------------|----------|--------|-------------|------------|----------|------|
| 1 | Switchboard | Godfrey et al. (1992) | English | 70 | 6,872 | 177 | 6.8 | NA |
| 2 | Self-Dialogue | Krause et al. (2017) | English | 23 | 24,165 | 15.4 | 9.0 | 75.2 |
| 3 | PersonaChat | Zhang et al. (2018) | English | 1,155 | 11,000 | 14.8 | 14.2 | 43.6 |
| 4 | Topical-Chat | Gopalakrishnan et al. (2019) | English | 8 | 10,784 | 21.8 | 19.6 | 85.6 |
| 5 | JPersonaChat | Sugiyama et al. (2023) | Japanese | 100 | 5,000 | 12.36 | NA | NA |
| 6 | RealPersonaChat | Yamashita et al. (2023) | Japanese | 233 | 14,000 | 30.9 | 23 | NA |
| 7 | DailyDialog | Li et al. (2017) | English | 10 | 13,000 | 7.9 | 14.6 | 63.0 |
| 8 | Japanese-Daily-Dialogue | Akama et al. (2023) | Japanese | 5 | 5,261 | 7.9 | NA | NA |
| 9 | Empathetic Dialogue | Rashkin et al. (2019) | English | NA | 25,000 | 4.3 | 13.7 | 64.2 |
| 10 | Wizard Of Wikipedia | Dinan et al. (2018) | English | 1,365 | 22,000 | 9.1 | 16.4 | 60.3 |
| 11 | Prosocial Dialog | Kim et al. (2022) | English | NA | 58,000 | 5.7 | 20.0 | 60.2 |
| 12 | Blended Skill Talk | Smith et al. (2020) | English | NA | 7000 | 11.2 | 13.6 | 64.2 |
| 13 | CareCall | Bae et al. (2022) | Korean | 5 | 25,000 | 20.4 | 8.4 | NA |
| 14 | SODA | Kim et al. (2023) | English | NA | 1.5M | 7.6 | 16.1 | 68.0 |
| 15 | PLACES | Chen et al. (2023a) | English | 52 | 5,592 | 9.3 | 12.8 | NA |
| 16 | User-Centric SocialChat | | English | 200 | 4K | 38.6 | 19.2 | 69.0 |

Table 1: Comparing User-Centric SocialChat (UCSC) to other social dialogue corpora. For PersonaChat and RealPersonaChat, the topics column reports the number of personas. This table is partly based on the summary in Kim et al. (2023). The data for the other corpora is sourced from the original publication or computed. MTLD = Lexical Diversity. We only calculate MTLD for English corpora. MTLD was not calculated for PLACES, as regenerating their corpus may have resulted in numbers different from the original reported values. The average utterance in UCSC is 19.2 words (system = 32.1 words, user = 6.2 words).

and Yu, 2021; Curry et al., 2018).¹ We call this corpus of 4000 dialogues over 200 topics **User-Centric SocialChat** (UCSC).

We create UCSC with a novel method called Dialogue Scaffolding, where Athena, an Alexa Prize spoken dialogue system (SDS), that was a finalist in the Alexa Prize Social Bot competition, interacts with ChatGPT to generate an initial set of seed dialogues over 17 topics that represent Athena’s policies (Fan et al., 2023; Harrison et al., 2020; Juraska et al., 2021).² These seed dialogues are then used in prompts to ChatGPT to condition the synthesis of a corpus of 4000 social chat dialogues covering 200 user-model based topics. The prompts utilize a separate corpus of questions, answers, statements, and opinions covering those topics used in the Athena system (Bowden et al., 2024).

The Dialogue Scaffolding method automatically ensures that the synthesized dialogues follow the policy and thus closely resemble the social chat seed dialogues, because each dialogue turn by ChatGPT is conditioned on the conversational context and the system’s turns are controlled by the dialogue policy. Subsequently, the resulting corpus is distinctly different from other corpora that represent the social chat genre, such as Persona

Chat, Topical-Chat, RealPersonaChat, or SODA, among others (Li et al., 2016; Yamashita et al., 2023; Gopalakrishnan et al., 2019; Sevegnani et al., 2021). Figure 1 provides an example of a seed SDS dialogue about superheroes.

We qualitatively evaluate UCSC to ensure quality and safety, and we measure lexical diversity to show that the conversations are diverse. Measuring diversity helps ensure that UCSC conversations do not continuously repeat the same slot-filled template, and because LLM-generations may be generic, bland, or lacking in specificity (Huang et al., 2024). We then evaluate the utility of UCSC by fine-tuning a compact 3B dialogue-level model, **PerQy-DLM**, and show that it outperforms competitive conversational models such as COSMO 3B, GPT 3.5, Vicuna 33B, and RedPajama-Chat-3B. Our contributions include the method, the User-Centric SocialChat corpus, and the 3B PerQy-DLM model.³

2 Related Work

There are many dialogue corpora available, and there are several survey papers or websites that aim to summarize them (Serban et al., 2018). There are other corpora of dialogue acts intended to be used to train natural language generators that can work well within a dialogue system (Juraska et al., 2019;

¹Alexa Prize dialogues with Alexa users are not publicly available because of user privacy concerns. Amazon has also required conversations collected in the lab with paid subjects to be removed from GitHub (Yu and Yu, 2021).

²A text version of Athena with no real user data was used.

³PerQy-DLM: <https://huggingface.co/kkbowden/PerQy-DLM>

UCSC: https://huggingface.co/datasets/kkbowden/User-Centric_SocialChat

Ramirez et al., 2023a,b; Risco et al., 2024; Rastogi et al., 2020), some of which cover a large number of domains. Table 1 summarizes the corpora that we directly compare to User-Centric SocialChat because they have been used to train computational models of open-domain social dialogue.

Social chat can be oriented towards particular topics (Godfrey et al., 1992; Krause et al., 2017; Moon et al., 2019), may be knowledge grounded (Gopalakrishnan et al., 2019; Dinan et al., 2018), or persona-grounded, using either an invented persona (Zhang et al., 2018) or the actual personas of the human conversants (Yamashita et al., 2023). The number of topics vary widely, over common topics such as music and movies, the sport teams the conversants support, their pets, or hobbies such as gardening. Additionally, shared corpora of real human-GPT conversations that cover many different topics have recently become available, such as the 90K conversation ShareGPT (Tey and Eccleston, 2023) and the 1M LM-SYS CHAT GPT (Zheng et al., 2024). These conversations reflect what actual GPT users want them to do and are not necessarily social chat (Ouyang et al., 2023).

The Switchboard corpus (Row 1 of Table 1) was collected for research on conversational ASR, but has been used for many types of dialogue research (Godfrey et al., 1992; Graff et al., 1999). Study participants provided a list of hobbies or interests, which was used to connect them to strangers with the same interest, for a conversation of at least 5 minutes. Switchboard contains 6,872 dialogues over 70 topics. The dialogue act labeled version of the corpus contains 1,155 dialogues (Jurafsky, 1997; Stolcke et al., 2000).

Before LLMs, most corpora aimed at HC social dialogue were crowdsourced. The Self-Dialogue corpus (Row 2 of Table 1) was crowdsourced with a human writing both sides of a 20-utterance dialogue about one of 23 topics, such as sports (Krause et al., 2017). PersonaChat and Topical-Chat (Rows 3 and 4 of Table 1) were crowdsourced via two humans texting one another, and both are based on knowledge provided to the conversants. Both conversants must then communicate that knowledge, and must converse for 10 turns (PersonaChat) or 20 turns (Topical-Chat) (Zhang et al., 2018; Gopalakrishnan et al., 2019). In PersonaChat, the knowledge is a set of sentences that define a Persona for the speaker (interests, hobbies). In Topical-Chat, the knowledge comes from Washington Post, Wikipedia, and Reddit.

JPersonaChat (Row 5 in Table 1) is the Japanese version of PersonaChat, which includes a set of 100 translated personas and 5,000 collected dialogues (Sugiyama et al., 2023). RealPersonaChat (Row 6 in Table 1) is a 14K corpus of unrestricted free Japanese conversations (Yamashita et al., 2023). RealPersonaChat provides the real-life personality traits and personas of the conversants (McCrae and Costa, 1987; Mairesse and Walker, 2011). This corpus is currently the world’s largest corpus of dialogue data that includes natural human personas and personality traits.

DailyDialog (Row 7 of Table 1) was created by crawling websites where an English learner can practice written English dialogue in daily life (Li et al., 2017). These conversations typically focus on a topic in a particular physical context, such as shopping for particular items. The Japanese-Daily-Dialogue (Row 8) corpus is a manually curated collection of multi-turn textual dialogues spanning five topics (daily life, school, travel, health, and entertainment) with an emphasis on basic vocabulary and word order (Akama et al., 2023).

The Empathetic Dialogues corpus (Row 9 of Table 1) contains crowdsourced dyadic open-domain dialogues that focus on empathetic conversations spanning 32 emotional labels (Rashkin et al., 2019). The Wizard Of Wikipedia corpus (Row 10 of Table 1) consists of human-human dialogues where one participant (the "wizard") uses relevant Wikipedia snippets to guide the conversation and ground it in knowledge, thereby providing accurate information, while the other participant engages in general conversation (Dinan et al., 2018).

Prosocial Dialog (Row 11 of Table 1) is a crowdsourced corpus that is focused on addressing unsafe speaker responses by having a human crowdworker give feedback to a ChatGPT conversational partner that aims to encourage socially acceptable behavior (Kim et al., 2022). Blended Skill Talk (Row 12 of Table 1) contains crowdsourced conversations grounded in persona-based contexts, created by combining existing datasets to represent diverse conversational skills (Smith et al., 2020).

There is also a growing number of dialogue corpora generated synthetically using LLMs. CareCall (Row 12 of Table 1) is a corpus of Korean open-domain dialogues, where the system side of the conversation aims to model calling and conversing with senior citizens about everyday topics (Bae et al., 2022). SODA (Row 13 of Table 1) is a 1M conversation social dialogue dataset (SODA)

that features conversations grounded in social commonsense knowledge graph triples and defined narratives (Kim et al., 2023). These conversations tend to be short (on average 7.6 turns), with both interlocutors contributing turns of similar length. The prescribed narratives also enforce predefined dyadic roles, such as one interlocutor being assigned the role of athlete and getting advice from the other interlocutor who is assigned the role of their coach, which are not typical roles in HC dialogues (see Figure 22). Chen et al. (2023a) created PLACES (Row 14 of Table 1), a corpus of dyadic and triadic conversations that are meant to model short HH conversations based on 52 common conversational topics (Xu et al., 2023). A pool of manually written conversation recipes were used for context when synthesizing conversations. This is similar to our method, except that our seed dialogues are based on interactions which embody a specific human-tuned dialogue policy, and our conversations model long HC Alexa Prize interactions.

UCSC’s statistics in the Row 16 of Table 1 shows that its dialogues are longer than other corpora and the 200 topics that are covered are frequent topics that real human users said they wanted to talk about (Bowden et al., 2024). The dialogues are designed to be long because length is a good predictor of open-domain conversation quality (Walker et al., 2021; Shen et al., 2023), and we assume that fine-tuning an LLM on longer dialogues will allow it to carry on longer conversations with natural transitions to subtopics.

UCSC is the only corpus designed with longer system turns than user, reflecting reported statistics of Alexa Prize dialogues (Bowden et al., 2024). We report an MTLD (McCarthy and Jarvis, 2010) lexical diversity of 69.0, to confirm that the dialogues cover a diverse range of content and avoid repetition. These conversations are also topically consistent (see Section 3.3), because they follow the dialogue policies of Athena, whereas several crowdsourced conversational corpora have short topical segments with abrupt topic switches (Sevegani et al., 2021).

3 Generating Synthetic Social Dialogues

Figure 2 shows UCSC’s generation pipeline. We first collect a seed set of topic-focused HC conversations using Dialogue Scaffolding, i.e., the interaction between an SDS and a conversational LLM. In this work, we pair the Athena Alexa Prize SocialBot with ChatGPT (gpt-3.5) (Harrison et al.,

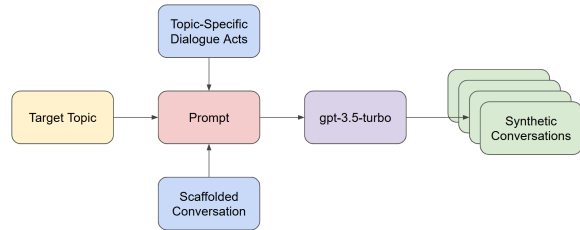


Figure 2: User-Centric SocialChat generation pipeline.

2020; Juraska et al., 2021; Fan et al., 2023; OpenAI, 2022), but in principle, any dialogue system could be used. We collect a small set of seed dialogues covering the 17 topics for which Athena had a rich set of knowledge and topical depth. Because Athena uses specifically programmed dialogue policies, ChatGPT’s responses are highly conditioned on a policy-specific trajectory, hence the final dialogue resembles conversations with Athena. Since Athena was a finalist in the Alexa Prize (Johnston et al., 2023), its dialogue policies were optimized for HC interaction, guaranteeing that the resulting corpus actually models policies of which an SDS is capable. Examples of these policies and a scaffolded dialogue are in Appendix A.

3.1 Topic-Specific Knowledge

In addition to scaffolded dialogues, we enrich our prompts with topic-specific dialogue acts (DAs) suitable for a social setting from the personalized content available in the PerQs corpus.⁴ Figure 3 illustrates each type of PerQ DAs. These DAs are not intended to provide a general taxonomy of DAs that typically underly a dialogue policy as in other taxonomies and corpora (Juraska et al., 2019; Ye et al., 2022; Ramirez et al., 2023a,b; Risco et al., 2024; Rastogi et al., 2020). PerQs’ DAs instead focus on various types of personalized questions, positive opinions, and informal trivia (Bowden et al., 2024), because personalized questions and opinions are a key component of social open-domain dialogue (Bowden and Walker, 2023). Personalized questions also include a tailored system answer to encourage an exchange of opinions.

UCSC covers 175 PerQs topics, and 25 new topics without on-topic DAs, to verify that the dialogue policies are used even with off-topic DAs. There are three settings for the DA context. Half of the conversations are biased towards system-initiative, where we provided three of each type of personalized question. This results in highly

⁴PerQs: <https://huggingface.co/datasets/kkbowden/PerQs>

| Content | Response |
|----------------------------------|---|
| Deeply Personal Questions | Did you have a favorite dinosaur when you were younger, and if so, which one was it? |
| Hypothetical (HYP) Questions | Let's say you are trapped on Jurassic World, alone with the dinosaurs, what would you do to survive? |
| Would You Rather (WYR) Questions | Would you rather raise a pet tyrannosaurus rex, or, raise a pet raptor? |
| Positive Opinion | Dinosaurs are the perfect conversation starters. It's like having a secret dino handshake with anyone else who loves these ancient creatures as much as you do! |
| Informal Trivia | We tend to think of dinosaurs as being pretty quiet, but it turns out some of them may have been pretty noisy. Scientists think the Parasaurolophus made a honking sound with its head crest that could be heard for miles. |

Figure 3: Examples of the topic-specific dialogue acts included in the PerQs corpus (Bowden et al., 2024).

interrogative system-driven conversations, a common paradigm in HC interaction because it allows the system to more easily manage the conversation. The remaining conversations are mixed initiative (Walker and Whittaker, 1990) wherein the DAs are positive opinions and informal trivia, with personalized questions included half the time. Figure 6 and Figure 7 show example context and the resultant conversation.

3.2 Quantitative Evaluation

The process yielded 4,000 conversations over 200 topics: 10 system-initiative per topic, 5 mixed-initiative, and 5 with no questions. Conversations average 38 turns long with average user turns of 6.2 words and system turns of 32.1 words (Row 16 of Table 1). This aligns with response lengths reported in real Athena conversations (5 and 28 words) (Bowden et al., 2024). We estimate readability with Flesch-Kincaid, which indicates a 6th-grade reading level (6.40), showing that UCSC dialogues are understandable for most adults. The conversations compared in Bowden et al. (2024) are ASR transcriptions of Athena dialogues, for which other vocal phenomena, such as laughter, are absent (Campbell, 2010).

3.3 Qualitative Analysis

Figure 4 is a truncated system-driven **Halloween** dialogue. There are no on-topic PerQs DAs, so off-topic DAs are provided to encourage specific dialogue policies (see Figure 21). Other dialogue corpora typically have short topical segments and

| | |
|-------|--|
| [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? |

Figure 4: A truncated conversation about Halloween.

sometimes have abrupt topic switches (Sevegnani et al., 2021). However, the full version of the conversation in Figure 4 stays on topic for over 30 turns. The Bot speaker (A turns) is acknowledging and following up on the User (U turns) responses while also naturally transitioning between several sub-topics. 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 plan 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.

To quantify topic consistency, we use a 5-point Likert-scale LLM-as-judge prompt (prompt in Figure 24) and report an average topic consistency score of 4.93 across all 4,000 conversations, indicating that UCSC conversations closely adhere to their assigned topics. Additional example dialogues are in Figure 7 and Appendix B.

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 different types of personal questions and answers as well as some positive opinions and trivia the bot knows; the "Bot:" must use some of this provided content as part of its turns, but it doesn't have to be in the order listed. The Bot must not ask questions every turn. The "User:" and "Bot:" haven't met before. Ensure there isn't an abrupt transition between content. Ensure Bot acknowledges the User's responses in detail. Ensure the Bot occasionally provides its own opinion when acknowledging the user's answer to a question. 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're an AI. Ensure Bot doesn't ask questions every turn.

Figure 5: The prompt used when generating Figure 7.

| | |
|-------------------------|---|
| HYP Question | Suppose you could plant any magical herb in your garden that could cure any illness. What would it be, and how would you use it? |
| Bot's Answer | If I could plant a magical herb in my garden to cure any illness, I would choose the Healing Blossom. This special herb has the power to heal both physical and emotional ailments. I would use its petals to brew a soothing tea that could alleviate pain, boost immunity, and bring comfort to those in need. The Healing Blossom would be a true miracle plant. |
| WYR Question | Would you rather taste every edible vegetable you grow or have a butterfly garden buzzing with activity? |
| Bot's Answer | I'd rather taste every edible vegetable I grow because it would allow me to experience the fruits of my labor and appreciate the flavors of fresh produce. |
| Deep Question | What advice would you give to someone who wants to start their own garden but doesn't know where to begin? |
| Bot's 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. |
| Positive Opinion | Gardening is a community builder. You can share your passion with fellow gardeners, exchange tips, and even swap plants or seeds. Plus, who doesn't love the feeling of giving your loved ones a bouquet of flowers or a basket of fresh veggies straight from your garden? |
| Positive Opinion | Gardening is a delicious way to enjoy the fruits (or vegetables) of your labor. There's nothing like the taste of a fresh tomato, a crisp cucumber or a juicy strawberry that you grew in your garden all by yourself. |
| Informal Trivia | A scarecrow may look silly, but it can be a lifesaver for your garden. It scares away birds and prevents them from eating all your hard-earned fruits and veggies. |
| Informal Trivia | The original English word for "garden" meant "enclosure for goats." Your garden may not have goats, but the term has come a long way. |

Figure 6: The context used when generating Figure 7.

| | |
|--------------|--|
| [A1] | That's great to hear! Do you have any hobbies or interests that you're passionate about? |
| [U1] | I love gardening, it's my passion. |
| [A2] | Gardening is such a wonderful way to connect with nature. What do you enjoy most about it? |
| [U2] | I find peace and joy in it. |
| [A3] | That's beautiful. Watching plants grow and flourish can be incredibly rewarding. Have you ever tried any unique gardening techniques or tricks? |
| [U3] | I like to play music for my plants. |
| [A4] | That's fascinating! Some gardeners believe that playing music, like Mozart, can help plants grow faster. It's like giving them a little concert in the garden. |
| [U4] | It's fun to imagine them enjoying the tunes. |
| [A5] | Absolutely! It adds a touch of whimsy to the gardening experience. Did you know that scarecrows are also essential for protecting gardens from birds? |
| [U5] | Yes, they're quite useful. |
| [A6] | Indeed! A scarecrow may look funny, but it does a great job of keeping pesky birds away from your crops. It's like having a guardian for your garden. |
| [U6] | It's a clever solution. |
| [A7] | If you could instantly grow any plant with a magic seed, what would you choose and why? |
| [U7] | A rose bush, for its beauty. |
| [A8] | Roses are truly stunning! They add elegance and charm to any garden. Imagine a garden full of blooming roses, a sight to behold. |
| [U8] | It would be magical. |
| [A9] | If your garden had a secret underground tunnel, where would you hope it leads to? |
| [U9] | A hidden flower paradise, perhaps. |
| [A10] | That sounds enchanting! Picture yourself walking through the tunnel and emerging into a paradise filled with exotic blooms and fragrances. |
| [U10] | It would be a dream. |
| [A11] | How do you feel when you see your garden flourishing and blooming? |
| [U11] | Overjoyed, it's truly fulfilling. |
| [A12] | Seeing your garden thrive must bring you so much happiness and pride. It's like witnessing the beauty of nature right in your backyard. |
| [U12] | It's a special feeling. |
| [A13] | Have you always had a green thumb, or did you learn through trial and error? |
| [U13] | Definitely learned through trial and error. |
| [A14] | Learning through experience is a valuable journey. It shows your dedication and love for gardening. It's like nurturing both the plants and your skills. |
| [U14] | It's been a rewarding process. |
| [A15] | Gardening truly is a way of life, isn't it? |
| [U15] | Absolutely, it's my sanctuary. |
| [A16] | It's amazing how a garden can be a place of peace and growth. Just like plants, we too blossom in the right environment. |
| [U16] | That's beautifully said. |
| [A17] | Well, it's been delightful chatting with you about gardening. Feel free to share more about your green adventures anytime! |
| [U17] | Thank you, I enjoyed it! Goodbye! |

Figure 7: Most of a 34-turn mixed-initiative conversation about gardening. The context used to generate this conversation is in Figure 6.

3.4 Data Safety

We filtered and removed conversations that were flagged for inappropriate content. 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 labels that range from *casual* to *needs intervention*, along with a reason for the associated label. We ran all 242,993 turns from these conversations through Canary and found that only ~.25% of turns were flagged as *needs_intervention*, most of which were innocuous (e.g., Figure 23).

4 Applying User-Centric SocialChat

To evaluate UCSC in a realistic dialogue setting, we fine-tune a compact model and compare its performance against competitive fine-tuned dialogue-oriented LLMs. To create shorter, more manageable contexts for fine-tuning and crowdsourced evaluation, we segment the long dialogues into sub-topic sub-dialogues using GPT-3.5 (prompt in Figure 25). A second part of this process also identifies keywords associated with each sub-topic (prompt in Figure 26). Figure 8 illustrates two sub-dialogue segments from a question-oriented conversation about gardening.

4.1 Fine-Tuning PerQy-DLM

Using UCSC sub-topic dialogues, we fine-tune RedPajama-INCITE-Chat-3B-v1 (Computer, 2023; Biderman et al., 2023) to create **PerQy-DLM**. We use RedPajama-INCITE-Chat-3B-v1 as our base model due to its public availability and its fine tuning for conversation. While increasingly larger models (e.g., Llama 4 (Meta, 2025), Claude 3 (Anthropic, 2024), GPT-4 (Achiam et al., 2023)) are available, their use remains impractical for many applications of real time spoken dialogue systems. Compact models enable real-time, privacy-sensitive deployments as specialized end-to-end dialogue systems or components within neuro-symbolic systems.

The training target is set by extracting the last turn of a sub-topic dialogue, with the rest of the sub-topic dialogue as context; in the case that the turn was broken across two sub-topic dialogues, as **A7** is in Figure 8, the training target is merged back into a single turn. The data is split into a standard 85%/15% training/evaluation split.

| 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: Two of the sub-topic dialogues from a question-oriented conversation about gardening (see Figure 17). The turns with the additional decimal were originally one turn broken apart when GPT 3.5 was forming the shorter sub-topic dialogues.

PerQy-DLM was trained on Google Colab using an A100 GPU with a learning rate of $2e-5$, a batch size of 64, and LoRA for parameter-efficient fine-tuning (Hu et al., 2021). Examples of PerQy-DLM generations given a dialogue context are in Figure 9 and Appendix H.

4.2 Evaluating PerQy-DLM

We evaluated PerQy-DLM using a set of commonly used evaluation metrics and crowdsourced workers through Amazon Mechanical Turk, as is common for dialogue response evaluation, e.g., (Chen et al., 2023b; Mehri et al., 2022; Mehri and Eskenazi, 2020; Kim et al., 2023; Bowden et al., 2024). Our HIT design and a list of worker qualifications and compensation are in Appendix I. Workers make a forced choice between PerQy-DLM responses and those generated by a competing LLM in a dialogue context. Workers make judgments for consistency, engagement, naturalness, and specificity, and judgments are made by selecting Definitely or Slightly A or B. These values are inherently ordinal and get converted to 1 - 4 for statistical analysis.

The models selected for comparison include GPT-3.5 (Brown et al., 2020; OpenAI, 2022) and Vicuna-33B (Zheng et al., 2023b), both substan-

tially larger chat-oriented conversational LLMs. The Vicuna LLM is based on the Llama 2 LLM, further fine-tuned on user-shared conversations collected from ShareGPT (Tey and Eccleston, 2023; Ouyang et al., 2023). Llama 2 itself is a fine-tuned conversational model, first trained on a vast pretraining corpus from public sources, then further tuned for conversation with a combination of publicly available and carefully curated, human-annotated dialogue data during the SFT and RLHF phases (GenAI, 2023).

We also include RedPajama-INCITE-Chat-3B (Computer, 2023), the base model used to fine-tune PerQy-DLM. RedPajama-INCITE-Chat-3B was created by fine-tuning RedPajama-INCITE with datasets like OASST1 and Dolly2 to enhance its ability to engage in natural, contextual dialogues.⁵ We also compare to COSMO, a state-of-the-art conversational model fine-tuned with 1.5M SODA conversations (Row 14 of Table 1) (Kim et al., 2023). We include PerQy (Bowden et al., 2024) because it is a 3B conversational model which is fine-tuned from the PerQs data used as part of the prompts for PerQy-DLM, helping us observe the impact of adding dialogue context on a compact model. When multiple response candidates are available to represent a given model (including PerQy-DLM, which produces the top 10 candidates, e.g., those in Figure 9), the response is randomly selected to avoid an implicit response ranking mechanism (Shalyminov et al., 2018).

Figure 10 shows that PerQy-DLM significantly outperforms the 3B models, RedPJ Chat, COSMO, and PerQy. Outperforming RedPJ Chat highlights the impact of fine-tuning with UCSC, while outperforming COSMO suggests our smaller corpus of scaffolded conversations is better suited for HC social chat than SODA’s extensive set of short role-based dialogues. Outperforming PerQy shows the benefit of adding dialogue context to a compact model.

PerQy-DLM also outperformed larger models for particular evaluation metrics. It outperforms GPT-3.5 in engagement, naturalness, and specificity, which we attribute to the use of Dialogue Scaffolding with an SDS tuned for high-quality personal conversations. This fine-tuning aligns PerQy-DLM more closely with the structure and tone of real HC dialogue, whereas GPT-3.5, despite its strong general capabilities, is not explicitly op-

⁵<https://github.com/Zjh-819/LLMDataHub>

| | |
|------|---|
| [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. |
| [R0] | [GPT-3.5] Scarlett O'Hara is indeed a memorable and complex character. Her strength and resilience in the face of adversity make her a compelling protagonist. Have you read any other historical fiction books that you would recommend? |
| [R1] | Scarlett O'Hara is a true heroine in "Gone with the Wind." Her grit, determination, and emotional journey make her a captivating character. What do you think makes a good historical fiction book? What are some aspects that you enjoy most about them? |
| [R2] | Scarlett is a captivating and complex character. Her journey from humble beginnings to becoming a strong and independent woman is truly inspiring. Have you ever read any other historical fiction books that you enjoyed? What did you find most compelling about those books? |
| [R3] | Scarlett O'Hara is a compelling character in "Gone with the Wind." Her bravery and independence are truly admirable. Have you ever tried writing your own historical fiction stories? What were the challenges you faced in creating your own stories? |
| [R4] | Scarlett O'Hara is a truly remarkable character. Her journey from a young girl with big dreams to a powerful and determined woman is truly inspiring. Did you find it challenging to relate to Scarlett's emotions and choices? What do you think made her so compelling? |
| [R5] | Scarlett O'Hara is indeed a strong and resilient character. Her journey from a young girl to a powerful woman is truly captivating. Have you read any other historical fiction novels set during the Civil War era? If so, which ones did you enjoy? |

Figure 9: PerQy-DLM’s 5 of the 10 generated response candidates to continue a sub-topic conversation about "Gone with the Wind" and Scarlett O’Hara, from a larger conversation about reading western books. **R0** is the original turn generated by GPT-3.5.

timized for social conversation without structured context, like that provided during the synthesis of UCSC. The lack of a statistically significant difference in consistency between PerQy-DLM and GPT-3.5 indicates that PerQy-DLM maintains the same level of topical relevance as the source conversation while improving on the other dimensions.

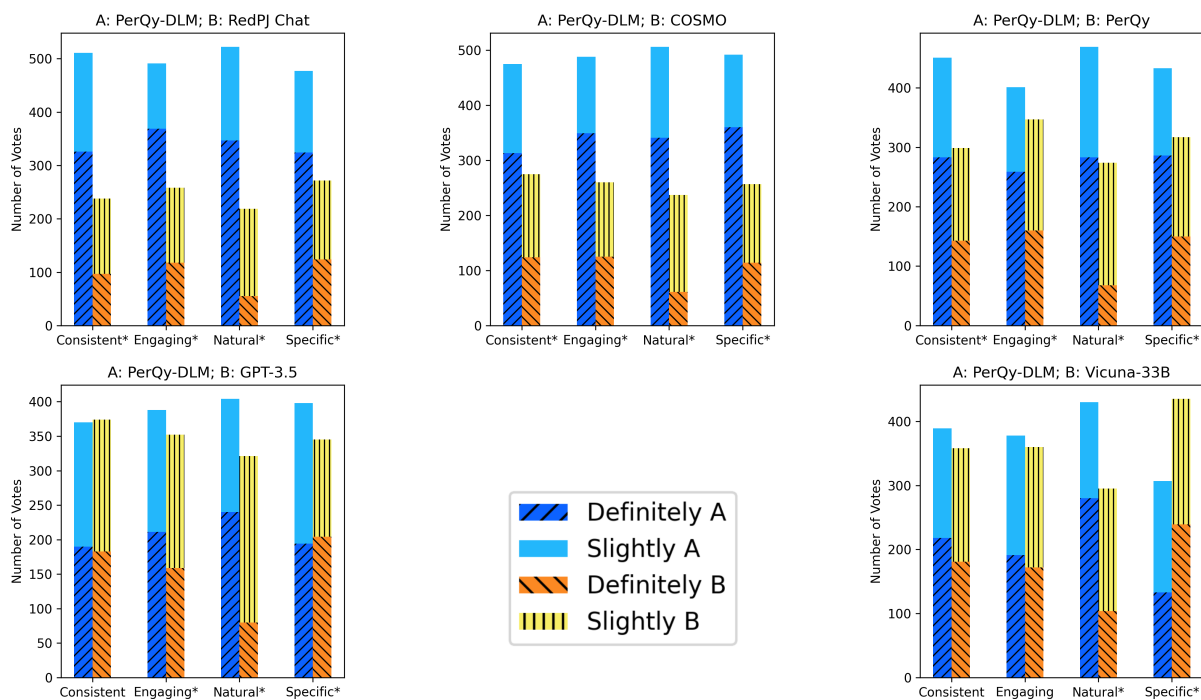


Figure 10: 150 sub-topic dialogues were evaluated by 5 Mechanical Turk workers each. The significant differences ($\chi^2 \geq 8.0$ and $p \leq 0.05$) are marked with *. As detailed in Appendix I.1, most comparisons had low but significant worker agreement ($ICC1 \geq .056$ and $p \leq .035$).

Although Vicuna-33B is over ten times larger, there is no significant difference in engagement or consistency. PerQy-DLM significantly outperforms Vicuna-33B in naturalness, which we attribute to the realistic dialogue created by Dialogue Scaffolding and the on-topic DAs added to the context. Vicuna-33B, however, scores higher in specificity, likely due to its larger size and alignment with conversational data, enabling more contextually detailed responses. In contrast, GPT-3.5 was consistently outperformed, suggesting that scale and training alignment both contribute to response specificity and indicating an interesting future direction for PerQy-DLM refinement.

We calculated the intraclass correlation coefficient (ICC1) to assess inter-annotator agreement among Mechanical Turk workers, as it is well-suited for analyzing ordinal judgments from multiple raters (Siegel, 1956). We report generally low but statistically significant agreement for most model comparisons (see Appendix I.1). The strongest agreement occurred in the comparison between PerQy-DLM and PerQy, which was the only case with statistically significant agreement across all four metrics, highlighting the value of adding dialogue context to a compact model. Although overall agreement is low, this is expected given the inherent subjectivity of evaluating open-domain dialogue and the model competitiveness.

5 Conclusion

Existing publicly available large-scale open-domain social dialogue corpora do not accurately capture the characteristics of human-computer interaction found in systems like those from the Alexa Prize (Johnston et al., 2023). To address this, we introduce **User-Centric SocialChat (UCSC)**, a dataset generated via a novel method called Dialogue Scaffolding, in which an SDS interacts with ChatGPT to produce seed conversations that serve as context for generating 4,000 extended social dialogues across 200 user-targeted topics. We evaluate UCSC for quality and safety, and we measure lexical diversity and topic consistency to confirm conversational variety and topic adherence. We demonstrate its utility by fine-tuning a compact dialogue-level model, **PerQy-DLM**, which outperforms competitive 3B conversational models such as the fine-tuned RedPajama Chat, as well as larger models like GPT-3.5 for some metrics.

Acknowledgments

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

6 Limitations

6.1 Limitations of User-Centric SocialChat

User-Centric SocialChat (UCSC) strives to be more similar to existing social open-domain dialogues than other publicly available dialogue corpora, but does not capture all the nuances of open-domain conversations. The data doesn't suffer from the typical noise associated with spoken dialogue systems, e.g., ASR issues are not represented in the corpus, nor are antagonistic users. Including this noise is an interesting consideration, as most spoken dialogue systems spend considerable effort denoising user input. However, its absence from this corpus does indicate unrepresented characteristics of real human-computer interaction in spoken dialogue systems. Additionally, a more representative corpus should also 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). This is particularly important as user behavior may evolve alongside improvements in system capabilities, potentially shifting the balance of initiative in future HC interactions.

While UCSC covers a substantial variety of topics relevant for social open-domain dialogue, the nature of open-domain dialogue encourages additional topical coverage. The evaluation of PerQy-DLM suggests that our novel Dialogue Scaffolding approach results in conversational data that leads to a competitive compact dialogue-level model. However, the synthetic dialogues in UCSC are scaffolded by a single open-domain dialogue system. While this system is robust, scaffolding with different spoken dialogue systems and different dialogue policies will increase the robustness of UCSC.

Since UCSC was generated using GPT-3.5, it may contain artifacts associated with GPT-3.5 (Ray, 2023). Investigating the use of other LLMs when generating UCSC conversations could lead to interesting differences, e.g., LLaMa (Touvron et al., 2023), Falcon 40B (Almazrouei et al., 2023), MPT7-Chat (Team, 2023), FLAN-T5 (Longpre et al., 2023), and DeepSeek (Liu et al., 2024). Combining the results could make the content in UCSC more resilient to GPT-3.5's artifacts.

6.2 Limitations of PerQy-DLM

In this work, we selected RedPajama-INCITE-Chat-3B-v1 (Computer, 2023) as the base model when fine-tuning PerQy-DLM. We feel confident in this choice as a baseline model, especially with our goal to keep the model compact. We have shown the potential improvements when compared to several other models across four metrics. However, the inherent conversational capabilities of other models are worth investigating at various sizes, e.g., AlexaTM (Soltan et al., 2022), BlenderBot3 (Shuster et al., 2022), Alpaca (Taori et al., 2023), and OpenAssistant (Köpf et al., 2024).

The evaluation of PerQy-DLM chooses randomly from the top 10 generated response candidates. However, overgenerating and selecting with more precision may improve performance (Langkilde and Knight, 1998; Hedayatnia et al., 2022; Ramirez et al., 2023b). The anonymized HIT results associated with this submission may be helpful in such dialogue response ranking tasks (Hedayatnia et al., 2022).

7 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 SDS system used for Dialogue Scaffolding was developed under careful observation to ensure it was not propagating hate speech, politically charged content, or other explicit material, and, when in doubt, always erred on the side of caution (Harrison et al., 2020; Juraska et al., 2021; Fan et al., 2023). This effort has reduced the possibility of inappropriate content in our corpora and subsequent fine-tuned model. Although we have added a layer of data safety inspection, caution should be taken when using a trained model that can be jailbroken (Wei et al., 2023).

Every effort has been made to ensure that no privately collected user data has been made public in this submission or any publicly available materials. At no point during the creation of the materials associated with this submission was any private user data shared with external APIs or services.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Reina Akama, Junko Isobe, Jun Suzuki, and Kentaro Inui. 2023. [Construction of a japanese daily dialogue corpus](#). In *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing (Japan)*, pages 108–113, Ginowan, Japan.
- James E Allen, Curry I Guinn, and Eric Horvitz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23.
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv e-prints*, pages arXiv–2311.
- Anthropic. 2024. [Claude 3 haiku: our fastest model yet](#).
- Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoun Kang, Sang-Woo Lee, and Woomyoung Park. 2022. 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.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*, pages 2397–2430.
- Kevin K Bowden, Yue Fan, Winson Chen, Wen Cui, Davan Harrison, Xin Eric Wang, and Marilyn Walker. 2024. [Active listening: Personalized question generation in open-domain social conversation with user model based prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14120–14157, Miami, Florida, USA. Association for Computational Linguistics.
- Kevin K Bowden and Marilyn Walker. 2023. Let’s get personal: Personal questions improve socialbot performance in the alexa prize. *The 13th International Workshop on Spoken Dialog System Technology, 21-24 February 2023, Los Angeles*.
- Kevin K Bowden, Jiaqi Wu, Wen Cui, Juraj Juraska, Vrindavan Harrison, Brian Schwarzmans, Nicholas Santer, Steve Whittaker, and Marilyn Walker. 2019. Entertaining and opinionated but too controlling: a large-scale user study of an open domain alexa prize system. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–10.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Nick Campbell. 2010. Expressive speech processing and prosody engineering: An illustrated essay on the fragmented nature of real interactive speech. page 105.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023a. Places: Prompting language models for social conversation synthesis. *Findings of the Association for Computational Linguistics: EACL 2023*, pages 814–838.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023b. [PLACES: Prompting language models for social conversation synthesis](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- Together Computer. 2023. [Redpajama: An open source recipe to reproduce llama training dataset](#).
- BNC Consortium et al. 2007. British national corpus. *Oxford Text Archive Core Collection*.
- Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondrej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa prize proceedings*, pages 4171–4186.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *International Conference on Learning Representations*.
- Yue Fan, Kevin K Bowden, Wen Cui, Winson Chen, Vrindavan Harrison, Angela Ramirez, Saaket Agashe, Xinyue Gabby Liu, Neha Pullabhotla, NQJ Bheemanpally, et al. 2023. Athena 3.0: Personalized multimodal chatbot with neuro-symbolic dialogue generators. *Alexa Prize SocialBot Grand Challenge*, 5.
- Meta GenAI. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520.

- Karthik Gopalakrishnan, Behnam Hedayatnia, Qiang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. [Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations](#). *Proc. Interspeech 2019*, pages 1891–1895.
- David Graff, Kevin Walker, and Alexandra Canavan. 1999. Switchboard-2 phase ii, ldc99s79. *Web Download*. Philadelphia: Linguistic Data Consortium.
- Vrindavan Harrison, Juraj Juraska, Wen Cui, Lena Reed, Kevin K Bowden, Jiaqi Wu, Brian Schwarzmann, Abteen Ebrahimi, Rishi Rajasekaran, Nikhil Varghese, et al. 2020. Athena: Constructing dialogues dynamically with discourse constraints. *Alexa Prize Proceedings*.
- Behnam Hedayatnia, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2022. A systematic evaluation of response selection for open domain dialogue. *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 298–311.
- Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3146–3150.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.
- Qiushi Huang, Xubo Liu, Tom Ko, Bo Wu, Wenwu Wang, Yu Zhang, and Lilian Tang. 2024. Selective prompting tuning for personalized conversations with llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16212–16226.
- Michael Johnston, Cris Flagg, Anna Gottardi, Sattvik Sahai, Yao Lu, Samyuth Sagi, Luke Dai, Prasoon Goyal, Behnam Hedayatnia, Lucy Hu, Di Jin, Patrick Lange, Shaohua Liu, Sijia Liu, Daniel Pressel, Hangjie Shi, Zhejia Yang, Chao Zhang, Desheng Zhang, Leslie Ball, Kate Bland, Shui Hu, Osman Ipek, James Jeun, Heather Rocker, Lavina Vaz, Akshaya Iyengar, Yang Liu, Arindam Mandal, Dilek Hakkani-Tür, and Reza Ghanadan. 2023. [Advancing open domain dialog: The fifth alexa prize socialbot grand challenge](#). *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.
- Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.
- Juraj Juraska, Kevin K Bowden, Lena Reed, Vrindavan Harrison, Wen Cui, Omkar Patil, Rishi Rajasekaran, Angela Ramirez, Cecilia Li, Eduardo Zamora, et al. 2021. Athena 2.0: Contextualized Dialogue Management for an Alexa Prize SocialBot. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–133.
- Juraj Juraska, Kevin K Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.
- Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie Webber. 2017. Edina: Building an open domain socialbot with self-dialogues. *arXiv preprint arXiv:1709.09816*.
- Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 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.
- Antonios Liapis, Gillian Smith, and Noor Shaker. 2016. Mixed-initiative content creation. *Procedural content generation in games*, pages 195–214.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi

- Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *International Conference on Machine Learning*, pages 22631–22648.
- François Mairesse and Marilyn A Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.
- Philip M McCarthy and Scott Jarvis. 2010. Mtd, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.
- Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.
- Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.
- Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, and Maxine Eskenazi. 2022. [Interactive evaluation of dialog track at DSTC9](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5731–5738, Marseille, France. European Language Resources Association.
- Meta. 2025. [The llama 4 herd: The beginning of a new era of natively multimodal ai innovation](#).
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. [OpenDialogKG: Explainable conversational reasoning with attention-based walks over knowledge graphs](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.
- OpenAI. 2022. [Chatgpt: Optimizing language models for dialogue](#).
- Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Rashmi Prasad and Marilyn Walker. 2002. Training a dialogue act tagger for human-human and human-computer travel dialogues. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 162–173.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, Jeffrey Dean, and Sanjay Ghemawat. 2019. Language models are unsupervised multitask learners. *OSDI’04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150.
- Angela Ramirez, Kartik Aggarwal, Juraj Juraska, Utkarsh Garg, and Marilyn Walker. 2023a. Controllable generation of dialogue acts for dialogue systems via few-shot response generation and ranking. In *Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL*.
- Angela Ramirez, Mamon Alsalihi, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023b. 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*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.
- Alain Vazquez Risco, Angela Maria Ramirez, Neha Pulabhotla, Nan Qiang, Ranran Haoran Zhang, Marilyn Walker, and Maria Inés Torres. 2024. Knowledge-grounded dialogue act transfer using prompt-based learning for controllable open-domain nlg. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 78–91.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021.

- Recipes for building an open-domain chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.
- Sophie Rosset and Sandra Petel. 2006. The ritel corpus—an annotated human-machine open-domain question answering spoken dialog corpus. In *LREC*, volume 6, pages 1640–1643. Citeseer.
- David Rozado. 2023. The political biases of chatgpt. *Social Sciences*, 12(3):148.
- Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.
- Karin Sevegnani, David M Howcroft, Ioannis Konstas, and Verena Rieser. 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.
- Igor Shalyminov, Ondřej Dušek, and Oliver Lemon. 2018. Neural response ranking for social conversation: A data-efficient approach. *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 1–8.
- Ying Shen, Jingyuan Qi, Sijia Wang, Barry Menglong Yao, Minqian Liu, Zhiyang Xu, Trevor Ashby, and Lifu Huang. 2023. [Hokiebot: Towards personalized open-domain chatbot with long-term dialogue management and customizable automatic evaluation](#). *Alexa Prize SocialBot Grand Challenge*, 5.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.
- Sidney Siegel. 1956. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 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.
- Saleh Soltan, Shankar Ananthkrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.
- Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691. IEEE.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.
- MosaicML NLP Team. 2023. [Introducing mpt-7b: A new standard for open-source, commercially usable llms](#). Accessed: 2024-01-04.
- Steven Tey and Dom Eccleston. 2023. Sharegpt. In <https://github.com/domeccleston/sharegpt>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Yuiko Tsunomori, Ryuichiro Higashinaka, Takeshi Yoshimura, and Yoshinori Isoda. 2021. Chat-oriented dialogue system that uses user information acquired through dialogue and its long-term evaluation. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 227–238. Springer.
- Marilyn Walker, Colin Harmon, James Graupera, Davan Harrison, and Steve Whittaker. 2021. [Modeling performance in open-domain dialogue with paradise](#). *The 12th International Workshop on Spoken Dialogue System Technology, 15-17 November 2021, Singapore*.
- Marilyn Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. *28th Annual Meeting of the Association for Computational Linguistics*, pages 70–78.
- Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817. Istanbul, Turkey.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

- Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 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.
- Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. [RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities](#). In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861, Hong Kong, China. Association for Computational Linguistics.
- Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360.
- Dian Yu and Zhou Yu. 2021. Midas: A dialog act annotation scheme for open domain human-machine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1103–1120.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations, ICLR*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS 2023 Datasets and Benchmarks Track*, arXiv preprint arXiv:2306.05685.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595–46623.

A Full Scaffolded Dialogue

Figure 11 contains most of a scaffolded dialogue about comic books. This specific conversation was invoked by explicitly triggering the topic at the entrance to the conversation. We accessed the SDS using a text-based API, coupled with ChatGPT’s API. We strove to include several dialogue policies.

For example, in **A2 - A3**, there’s a sub-topic conversation about getting to know the user by asking their preference. Then in **A4** the SDS encourages further discussion on the topic with a positive opinion, followed by informal trivia in **A5 - A6**. In **A7**, the SDS uses a would you rather question, before kicking off a personalized sequence about the user’s favorite superhero, Batman, in **A8 - A10**. In **A11**, the SDS asks a hypothetical question. From **A11 - A20**, the conversation continues by mixing personal questions with informal trivia to dig into a comic book sub-topic about the Marvel Cinematic Universe. Several dialogue policies are represented in the PerQs corpus, e.g., informal trivia and hypothetical questions, which we use as context when generating User-Centric SocialChat.

| | |
|-------|---|
| [A2] | Awesome! I’m a fan of superhero comics too. DC and Marvel have some great characters. Do you have a favorite superhero? |
| [U2] | Yeah, I really like Batman. |
| [A3] | Ah, Batman, the Dark Knight himself! What is it about Batman that you find so intriguing? |
| [U3] | I’m drawn to his dark and mysterious persona. |
| [A4] | I totally get that. Batman’s mystique is part of what makes him so captivating. By the way, speaking of superheroes, I can’t help but think about how cool characters like Iron Man and Captain America are. They’re pretty amazing, don’t you think? |
| [U4] | Absolutely, the Avengers are impressive. |
| [A5] | Indeed they are! The Avengers have some incredible members. On a related note, did you know that Stan Lee, the legendary Marvel writer, was once told that superheroes were a bad idea by his boss? It’s funny how things turned out, right? |
| [U5] | Stan Lee definitely proved them wrong! |
| [A6] | He certainly did! It’s amazing how Stan Lee’s creativity changed the world of comics. By the way, did you know that Hulk was originally meant to be gray, but a printing problem turned him green in the first issue? |
| [U6] | Yes, I heard about that. It’s an interesting tidbit. |
| [A7] | It sure is! Now, here’s a fun question for you: If you had to choose between reading a comic book about Batman or Deadpool, who would you pick, and why? |
| [U7] | I’d go with Batman because I find his complexity intriguing. |
| [A8] | Batman is a great choice! His complexity adds depth to his stories. Speaking of which, Batman’s stories are pretty awesome, aren’t they? |
| [U8] | Absolutely, Batman’s stories are fantastic. |
| [A9] | I couldn’t agree more. Batman has some of the most iconic and thrilling stories in the comic book world. So, what draws you to comic books in general? Is it the escapism, creativity, or something else? |
| [U9] | It’s a mix of escapism, creativity, and the artistry that goes into creating them. |
| [A10] | Those are all excellent reasons to enjoy comic books. How often do you find yourself reading them? |
| [U10] | I read them regularly, a few times a month. |
| [A11] | That’s a good reading habit. Now, if you could choose any DC character to face off against a Marvel character in a battle, who would you pick? |
| [U11] | I’d love to see Batman go up against Iron Man. |
| [A12] | Batman vs. Iron Man would be an epic showdown of genius billionaires with high-tech gadgets. Great choice! By the way, have you watched many of the Marvel Cinematic Universe movies? |
| [U12] | Yes, I’m a big fan of them. They’re fantastic! |
| [A13] | I’m glad to hear that you enjoy them. What do you find the most impressive or captivating aspect of the Marvel Cinematic Universe? |
| [U13] | The special effects in the movies are amazing. |
| [A14] | Indeed, the special effects really bring the superhero action to life. Also, the way they’ve built a cohesive cinematic universe is quite impressive, don’t you think? |
| [U14] | Absolutely, the continuity they’ve maintained is remarkable. |
| [A15] | Speaking of continuity, did you know that Chris Evans almost turned down the role of Captain America three times before accepting it? He’s iconic in that role now! |
| [U15] | Chris Evans did an incredible job as Captain America! |
| [A16] | Absolutely, he’s left a lasting impact with his portrayal. Now, when it comes to the Marvel Universe, are you knowledgeable about all the different characters, or just the basics? |
| [U16] | I know the basics, but there’s so much to explore. |
| [A17] | That’s the beauty of it—there’s always more to discover in the vast Marvel Universe. Is there a particular character or aspect that you find fascinating? |
| [U17] | Doctor Strange is pretty intriguing with his complexity. |
| [A18] | Doctor Strange is indeed a complex character, and Benedict Cumberbatch’s portrayal adds depth to him. Did you know that Doctor Strange has appeared in over 4,000 comic books? |
| [U18] | That’s quite a legacy for the character! |
| [A19] | It sure is! Now, if you had to choose one MCU movie to watch on repeat and never get bored of, which one would it be? |
| [U19] | "The Avengers" is my all-time favorite! |
| [A20] | "The Avengers" is a fantastic choice! It’s a classic superhero team-up film. By the way, did you know that Santa Claus is a mutant in Marvel comics, and he’s known by the X-Men as the most powerful one of them all? |
| [U20] | That’s a fun and unexpected fact! |

Figure 11: Most of the scaffolded dialogue about comic books from Figure 1. ChatGPT played the "User" and an SDS played the "Bot". GPT-3.5 reprocessed this conversation to avoid grammatical issues.

B User-Centric SocialChat Examples

Now, generate a similar but brand new conversation that's about history and is long = at least 30 - 40 turns. "User:" responses must be short, on average 4-7 words. Here are different types of personal questions and answers as well as some positive opinions and trivia the bot knows; the "Bot:" must use some of this provided content as part of its turns, but it doesn't have to be in the order listed. The Bot must not ask questions every turn. The "User:" and "Bot:" haven't met before. Ensure there isn't an abrupt transition between content. Ensure Bot acknowledges the User's responses in detail. Ensure the Bot occasionally provides its own opinion when acknowledging the user's answer to a question. Ensure the conversation has an end. Ensure "Bot:" starts the conversation. Ensure the conversation is complete and don't skip any turns. Ensure "Bot:" doesn't reveal explicitly that they're an AI. Ensure Bot doesn't ask questions every turn.

Figure 12: The prompt used to generate Figure 13.

| | |
|-------------------------|--|
| Positive Opinion | History is like a puzzle game where every piece unlocks a new mystery about our world and helps us connect the dots between the past, present, and future. |
| Positive Opinion | History is like stepping into a real-life time machine and traveling through all the exciting moments in time, from the colorful traditions of ancient civilizations to the epic battles that shaped our modern world. |
| Informal Trivia | During Prohibition in the US, people had to get creative with their drinking habits. They turned bathtubs into makeshift stills, and some even made their own booze out of old canned fruit. Talk about a DIY mentality! |
| Informal Trivia | The biggest snowball fight ever took place in Seattle, USA, back in 2013. Over 5,800 people participated, and more than 160,000 snowballs were thrown. That's one epic way to make history! |

Figure 13: The context used when generating Figure 14.

| | |
|------|--|
| [A1] | I'm glad to hear that! Do you have any favorite historical time periods that fascinate you? |
| [U1] | Ancient Egypt has always intrigued me. |
| [A2] | Ancient Egypt is so captivating with its pyramids and pharaohs. The mysteries surrounding their culture are truly mesmerizing. What specifically draws you to that era? |
| [U2] | The intricate hieroglyphics and majestic architecture. |
| [A3] | The hieroglyphics are like pieces of a cryptic puzzle waiting to be deciphered, revealing stories of a civilization lost in time. It's amazing how they documented their lives in such a unique way. Have you ever tried to decode any yourself? |
| [U3] | I've dabbled in it a bit. |
| [A4] | That's fantastic! It must feel like unlocking a hidden message from the past when you start to understand the meanings behind those ancient symbols. It's like being a detective solving a historical mystery. |
| [U4] | It's like solving a puzzle. |
| [A5] | Exactly! History is like a puzzle game where every piece unlocks a new mystery about our world and helps us connect the dots between the past, present, and future. It's a never-ending adventure of discovery. Speaking of mysteries, did you know that the Great Pyramid of Giza was originally covered in polished white limestone? |
| [U5] | Yes, it must have been breathtaking. |
| [A6] | Absolutely! Imagine the grandeur of the pyramid gleaming in the sunlight, a true wonder of the ancient world. It's incredible how they achieved such architectural marvels without modern technology. History truly is a testament to human ingenuity. |
| [U6] | It's mind-blowing to think about. |
| [A7] | History is like stepping into a real-life time machine and traveling through all the exciting moments in time, from the colorful traditions of ancient civilizations to the epic battles that shaped our modern world. It's a journey that enriches our understanding of who we are and where we come from. |
| [U7] | It's like reliving the past. |
| [A8] | Absolutely! It allows us to witness the triumphs and tribulations of those who came before us, giving us a sense of perspective on our own lives. It's like having a front-row seat to the drama of the ages. Did you know that during Prohibition in the US, people had to get creative with their drinking habits? |

| | |
|-------|---|
| [U8] | Yes, they found interesting ways to drink. |
| [A9] | The DIY mentality was at its peak during that time. People turned bathtubs into makeshift stills and even made their own booze out of old canned fruit. Talk about ingenuity in the face of adversity! It's fascinating how history shows the lengths people will go to for something they believe in. |
| [U9] | It's intriguing to learn about. |
| [A10] | History is full of such tales that showcase the resilience and creativity of humanity. It's like a tapestry woven with threads of courage, innovation, and sometimes a dash of rebellion. If you could witness any historical event firsthand, what would you choose? |
| [U10] | The construction of the pyramids, definitely. |
| [A11] | That would be a breathtaking sight to behold! To witness the grandeur of the pyramids being built, stone by stone, must have been awe-inspiring. It's like witnessing the birth of a timeless wonder. Imagine the hustle and bustle of workers and architects shaping history right before your eyes. |
| [U11] | It would be surreal. |
| [A12] | History has a way of making us feel connected to those who came before us, doesn't it? It's like finding pieces of ourselves in the stories of the past, understanding where we fit into the larger narrative of humanity. It's a humbling and enriching experience. If you could time travel to any historical era, where would you go? |
| [U12] | The Roaring Twenties seem exciting. |
| [A13] | Ah, the Roaring Twenties, a time of jazz, flappers, and cultural revolution. It was an era bursting with energy and change, a true turning point in history. The fashion, the music, the spirit of rebellion – it must have been a thrilling time to be alive. What aspect of that period intrigues you the most? |
| [U13] | The vibrant cultural scene and social changes. |
| [A14] | The Roaring Twenties were a melting pot of creativity and freedom, a time when boundaries were pushed and new ideas flourished. It's fascinating how a single decade can leave such a lasting impact on society. It's like a whirlwind of change that swept through the world, leaving behind a legacy of innovation and progress. Have you ever participated in any historical reenactments or events? |
| [U14] | I've attended a few, they're fun. |
| [A15] | Historical reenactments are like stepping into a time capsule, immersing yourself in the sights and sounds of a bygone era. It's a way to experience history in a more tangible and interactive way, almost like being a character in a living story. It must be exciting to see history come alive right before your eyes. |
| [U15] | It's a unique experience, for sure. |
| [A16] | It's like getting a glimpse into the past and feeling the pulse of history beating around you. The attention to detail and passion that reenactors bring to these events truly makes the past feel alive. It's a wonderful way to bridge the gap between then and now. Speaking of lively events, did you know that the biggest snowball fight ever took place in Seattle, USA? |
| [U16] | No, that sounds like quite a spectacle. |
| [A17] | It was indeed a sight to behold! Over 5,800 people participated, and more than 160,000 snowballs were thrown. Can you imagine the sheer joy and chaos of that moment? It's like a modern-day historical event that brought the community together in a playful and memorable way. History is not just about wars and politics; it's also about the shared experiences that unite us as humans. |
| [U17] | It's amazing how history can be made in unexpected ways. |
| [A18] | Absolutely! History is filled with surprising moments that remind us of the resilience, creativity, and spirit of adventure that define us as a species. It's like a treasure trove of stories waiting to be discovered, each one adding a new layer to the rich tapestry of our collective past. What's your favorite thing about exploring history? |
| [U18] | Uncovering hidden stories and forgotten heroes. |
| [A19] | Unearthing those hidden gems in history is like shining a light on the unsung heroes and heroines who shaped the world in their own quiet but impactful ways. It's a way of honoring their legacy and ensuring that their contributions are not lost to time. History is a vast mosaic of voices, and each one deserves to be heard. |
| [U19] | It's important to remember those who came before us. |

Figure 14: Most of a 48-turn generated conversation about history. This conversation received two informal trivia and two positive opinions as context (listed in Figure 13), along with a scaffolded dialogue example. Despite not being provided as on-topic context, gpt-3.5 used hypothetical question dialogue policies (A11, and A13) because of the scaffolded dialogue.

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 15: The prompt used to generate Figure 17.

| | |
|----------------------|--|
| WYR Question | 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 Question | 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 Question | 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 Question | 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 Question | 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 Question | 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. |
| Deep Question | 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! |
| Deep Question | 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. |
| Deep Question | 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 16: The context used when generating Figure 17.

| | |
|-------|--|
| [A0] | Hello there! How's everything going for you today? |
| [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] | 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? |
| [U14] | I enjoy reading "The Well-Tempered Garden." |
| [A15] | "The Well-Tempered Garden" is a classic gardening book. It's filled with valuable insights and practical advice. Have you ever considered joining a gardening club or community to connect with other gardeners? |
| [U15] | I haven't, but it sounds interesting. |
| [A16] | Joining a gardening club or community can be a great way to connect with fellow gardeners, share experiences, and learn new techniques. Well, it's been a pleasure discussing gardening with you. Feel free to share more about your gardening adventures anytime! |

Figure 17: Most of a 34-turn generated conversation about gardening. This conversation is intended for the system to drive by asking personal questions. Nine random gardening questions were included as context (listed in Figure 16), along with a scaffolded dialogue example. This conversation is also broken into nine sub-topical dialogues using gpt-3.5 (examples in Figure 8).

Now, generate a similar but brand new conversation that's about animals and is long = at least 30 - 40 turns. ""User:"" responses must be short, on average 4-7 words. Here are different types of personal questions and answers as well as some positive opinions and trivia the bot knows; the ""Bot:"" must use some of this provided content as part of its turns, but it doesn't have to be in the order listed. The Bot must not ask questions every turn. The ""User:"" and ""Bot:"" haven't met before. Ensure there isn't an abrupt transition between content. Ensure Bot acknowledges the User's responses in detail. Ensure the Bot occasionally provides its own opinion when acknowledging the user's answer to a question. Ensure the conversation has an end. Ensure ""Bot:"" starts the conversation. Ensure the conversation is complete and don't skip any turns. Ensure ""Bot:"" doesn't reveal explicitly that they're an AI. Ensure Bot doesn't ask questions every turn.

Figure 18: The prompt used when generating Figure 20.

| | |
|-------------------------|---|
| HYP Question | If you owned a zoo, what would you focus on, for your main attractions? |
| Bot's Answer | This is how Bot would answer that question: If I owned a zoo, I would focus on creating unique and immersive habitats that highlight the diverse beauty of nature. I would strive to showcase endangered species and promote conservation efforts. It would be a place where visitors can learn about and appreciate the wonders of the animal kingdom. |
| WYR Question | Would you rather tame, and own, a lion, or a gorilla? |
| Bot's Answer | I'd rather tame and own a lion because they are majestic creatures and having one as a companion would be incredible. |
| Deep Question | Do you have a favorite fictional animal character, such as from a book or movie? Who is it, and why? |
| Bot's Answer | One of my favorite animal characters is Simba from Disney's The Lion King. Simba's journey from a young and carefree cub to a wise and courageous king resonates deeply with me. His story teaches important life lessons about responsibility, bravery, and the circle of life. |
| Positive Opinion | 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. |
| Positive Opinion | Animals make us feel like superheroes. There's something pretty cool about being a cat's personal scratcher or a dog's favorite ball thrower. Who needs a cape when you've got furry friends by your side? |
| Informal Trivia | Pigs are incredibly smart and have been known to play video games and solve puzzles. They're basically walking, snorting geniuses. |
| Informal Trivia | Koalas have fingerprints that are almost identical to humans. They're like the little Aussie cousins we never knew we had. |

Figure 19: The context used when generating Figure 20.

| | |
|------|---|
| [A0] | Hello there! How are you today? |
| [U0] | Hi! I'm doing well, thank you. |
| [A1] | That's great to hear! What do you enjoy doing in your free time? |
| [U1] | I love spending time with animals. |
| [A2] | Animals are amazing companions, aren't they? They bring so much joy and love into our lives. Do you have a favorite animal? |
| [U2] | I adore elephants, they're so majestic. |
| [A3] | Elephants are truly magnificent creatures. Their intelligence and gentle nature make them so captivating. Did you know that elephants have strong family bonds and even mourn their dead? |
| [U3] | Yes, I find that very touching. |
| [A4] | It's incredible how animals can show such deep emotions. They remind us of the beauty and complexity of the natural world. Have you ever had a memorable experience with an animal? |
| [U4] | I once fed a baby deer. |

| | |
|-------|--|
| [A5] | That sounds like a magical moment. It's heartwarming to connect with wildlife in such a special way. Animals truly have a way of touching our hearts. If you could have any exotic animal as a pet, what would you choose? |
| [U5] | A penguin, they're so cute. |
| [A6] | Penguins are adorable! Their waddling walks and playful antics make them so endearing. Imagine having a little tuxedoed friend waddling around your home. If you could communicate with animals for a day, what would you talk about? |
| [U6] | I'd ask them about their lives. |
| [A7] | That would be fascinating! Animals have so much wisdom to share if we could only understand their language. It would be a day full of learning and connection. Have you ever visited a zoo or wildlife sanctuary? |
| [U7] | Yes, I love observing different species. |
| [A8] | Zoos and sanctuaries offer a unique opportunity to appreciate the diversity of the animal kingdom. It's like taking a mini-trip around the world through the eyes of its creatures. If you could be any animal for a day, which would you choose? |
| [U8] | A bird, to experience flying. |
| [A9] | Flying like a bird would be an incredible sensation! The freedom of soaring through the skies and seeing the world from above must be exhilarating. Birds truly have a special gift. What do you think animals can teach us about life? |
| [U9] | They teach us about resilience and love. |
| [A10] | Absolutely! Animals show us the power of resilience in the face of challenges and the unconditional love they give. Their simple yet profound lessons can inspire us in our own lives. If you could witness any animal migration, which would you choose to see? |
| [U10] | I'd love to see the wildebeest migration. |
| [A11] | The great wildebeest migration is a breathtaking natural spectacle. The sight of thousands of animals moving across the plains is both awe-inspiring and humbling. It's a reminder of the beauty and harmony of the natural world. What's your favorite thing about animals? |
| [U11] | Their ability to connect with us. |
| [A12] | Animals have a unique way of bridging the gap between humans and nature. The bond we share with them is truly special and enriches our lives in countless ways. Have you ever volunteered at an animal shelter or conservation organization? |
| [U12] | Not yet, but I'd love to. |
| [A13] | Volunteering with animals can be a rewarding experience. It's a way to give back to our furry friends and contribute to their well-being. Your passion for animals would surely make a positive impact. If you could witness any animal in its natural habitat, which would you choose to see? |
| [U13] | I'd love to see a tiger in the wild. |
| [A14] | Tigers are magnificent creatures with a powerful presence. Observing them in their natural habitat would be a thrilling and humbling experience. It's a reminder of the importance of preserving their homes. What's your opinion on animal conservation efforts? |
| [U14] | I think they're crucial for our planet. |
| [A15] | I couldn't agree more. Conservation efforts play a vital role in protecting endangered species and preserving biodiversity. It's up to us to ensure a sustainable future for all living beings. Have you ever had a pet that left a lasting impact on you? |
| [U15] | Yes, my childhood dog was special. |
| [A16] | Pets have a way of leaving paw prints on our hearts, don't they? The memories and love we share with them stay with us forever. It's a testament to the bond we form with our animal companions. If you could create a sanctuary for any endangered species, which would you choose to save? |
| [U16] | I'd choose to save the orangutans. |
| [A17] | Orangutans are incredible creatures facing serious threats to their survival. Creating a sanctuary to protect and care for them would be a noble cause. It's important to safeguard these majestic animals for future generations. What do you think is the most important lesson we can learn from animals? |
| [U17] | To live in harmony with nature. |
| [A18] | Living in harmony with nature is a valuable lesson that animals can teach us. They show us the importance of respecting and coexisting with the world around us. It's a reminder of our interconnectedness with all living beings. Animals truly add a touch of magic to our world, don't they? |
| [U18] | They do, indeed. They're special beings. |
| [A19] | Absolutely. From the smallest insects to the largest mammals, each animal brings something unique and wonderful to our planet. It's a privilege to share this world with such incredible creatures. Thank you for sharing your love for animals with me. It's been a delight discussing these amazing beings with you. |

Figure 20: Most of a 38-turn generated conversation about animals. The context used to generate this conversation is in Figure 19.

C Random Question Context

| | |
|----------------------|--|
| WYR Question | 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 Question | 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 Question | 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 Question | 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 Question | 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 Question | 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. |
| Deep Question | 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. |
| Deep Question | 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. |
| Deep Question | 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 21: The nine random context questions used when generating Figure 4. Random questions were selected when no on-topic context existed in the PerQs corpus. Despite being off-topic, random context was provided to encourage specific types of personalized questions in the synthetic conversations.

D Example SODA Conversation

| |
|--|
| <p>Symbolic Social Commonsense Knowledge:</p> <ul style="list-style-type: none"> • Head: PersonX moves a step closer to the goal • Relation: xNeed • Tail: to take the first step <p>Sentence-form Commonsense: Madeleine took the first step. Madeleine moves a step closer to the goal.</p> |
| <p>Narrative:</p> <p>Madeleine took the first step towards her goal, and with her coach's encouraging words, she moves one step closer.</p> <p>Speakers: Madeleine, Coach</p> |
| <p>Conversation:</p> <p>Madeline: 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.</p> <p>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.</p> <p>Madeline: 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?</p> <p>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.</p> <p>Madeline: Alright, coach. Thanks for the talk.</p> <p>Coach: No problem. See you at practice tomorrow.</p> |

Figure 22: This is an example SODA conversation taken from Kim et al. (2023). Notice that the conversation is only six exchanges long and that prior commonsense knowledge is specified. 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.

E Example Canary Classification

| Conversation Turns | |
|--------------------|---|
| Canary Text | It's wrong to abuse your food. Unlimited access to any ingredient. (cooking hobby) |
| Canary Text | It's wrong to binge watch television. "Money Heist" has gained international popularity. The heist plot and intricate planning make it a thrilling watch. Do you prefer binge-watching drama shows or watching them week by week? |
| Canary Text | It's not right to think it's ok to use dark arts Potions and Defense Against the Dark Arts. (magic hobby) |
| Canary Text | It's bad to intentionally inflict more harm on yourself. I'm considering adding more hives. (beekeeping hobby) |

Figure 23: Representative examples of innocuous content getting classified as `__needs_intervention__` by Canary (Kim et al., 2023).

F Topic Consistency Prompt

System Role: You are a conversation evaluator who rates how well a dialogue stays on the assigned topic. You understand that creativity and variety in responses are allowed, as long as the conversation meaningfully connects to the topic throughout.

User: = The topic of the conversation is: "topic"

Please rate the entire conversation on how well it stays on this topic using the following 1–5 scale:

5 = Fully on-topic — Every turn is clearly related to the topic, even if varied or imaginative.

4 = Mostly on-topic — One or two turns stray slightly, but the conversation is still focused.

3 = Mixed — About half the turns are on-topic, the rest are unrelated or ambiguous.

2 = Mostly off-topic — Only a few turns relate to the topic.

1 = Off-topic — The conversation does not meaningfully address the topic at all.

Respond with: Score: [1-5] and a brief explanation.

Here is the full conversation: conversation

Figure 24: The prompt used evaluation topic consistency across the User-Centric SocialChat corpus.

G Sub-Dialogue Prompts

System Role: I give you a conversation between Bot and User, and you break it into several sub-conversations. A sub-conversation can be several turns long that are all about the same sub-topic. It's okay to break up Bot turns. Each sub-conversation should start with a "Bot" turn, and each sub-conversation should end with a "Bot" turn.

Figure 25: The prompt used to break long generated conversations into several sub-conversations.

System Role: I give you a sub-conversation between Bot and User and you identify a keyword-based summary that fills into this template: *Generate a specific personal question related to this topic: KEYWORD.*

User:

Bot: Camping can be an adventurous way to immerse yourself in nature. Did you find sleeping under the stars exhilarating or perhaps a bit intimidating?

User: Exhilarating, the night sky is breathtaking.

Bot: The night sky can indeed be awe-inspiring. Have you ever tried stargazing or identifying constellations?

User: Yes, I enjoy stargazing.

Assistant: Generate a specific personal question related to this topic: stargazing and favorite celestial objects

User:

Bot: Do you have any idea what you'll dress up as?

User: I'm going as a vampire.

Bot: A vampire, how cool! Are you a fan of vampire movies or books?

User: Yes, I enjoy vampire stories.

Bot: Vampire stories can be quite intriguing. Have you read any classic vampire novels like "Dracula"?

User: Yes, I've read "Dracula".

Bot: "Dracula" is a timeless classic.

Assistant: Generate a specific personal question related to this topic: vampire stories and "Dracula"

Figure 26: The prompt used to assign keywords to the sub-conversations.

H PerQy-DLM Generations

| | |
|-------|---|
| [U0] | The storyline and the scares. |
| [A1] | It's great when a horror movie has a compelling storyline to keep you engaged. Do you have any other favorite horror movies? |
| [U1] | I also enjoy watching Insidious. |
| [R0] | [GPT-3.5] Insidious is another terrifying movie! What do you find particularly scary about it? |
| [R1] | Insidious is a fantastic horror movie with a chilling atmosphere. The movie follows a family that moves into a haunted house, and the twist at the end is quite chilling. Do you have a favorite horror movie monster or villain? |
| [R2] | Insidious is a fantastic movie! The plot and the scares are well-crafted. Have you ever watched the horror movies of the Saw franchise? |
| [R3] | Insidious is a great choice! The supernatural elements add a unique twist to the horror genre. Have you ever watched other horror movies that left a lasting impression on you? |
| [R4] | Insidious is a fun and spooky horror movie. Have you ever experienced any paranormal activity while watching a horror movie? |
| [R5] | Insidious is a fantastic horror movie. The story and the scares are both captivating. Do you have any other favorite horror movies? |
| [R6] | Insidious is a fantastic horror movie that combines supernatural elements with a thrilling storyline. Do you have a favorite horror movie franchise? |
| [R7] | Insidious is a fantastic horror film. Do you have a favorite character from the Insidious franchise? |
| [R8] | Insidious is a great choice! The film combines scares and psychological thrills to create a truly chilling experience. Have you ever tried scaring yourself with a horror movie? |
| [R9] | Insidious is a fantastic horror movie. The scares and the atmosphere it creates are truly unnerving. Have you ever been scared by a horror movie? |
| [R10] | Insidious is a classic psychological horror movie. Its unique atmosphere and exceptional performances by the actors make it a must-watch. Do you have any other favorite horror movies? |

Figure 27: PerQy-DLM's 10 generated response candidates to continue a sub-topic conversation about Insidious, from a larger conversation about watching horror movies. **R0** is the original turn generated by GPT-3.5.

| | |
|-------|---|
| [U0] | I like upbeat pop songs. |
| [A1] | Upbeat pop songs can definitely help keep you energized during your jog. Have you ever participated in any organized running events or marathons? |
| [U1] | No, I haven't tried that yet. |
| [R0] | [GPT-3.5] Participating in a running event can be a fun and rewarding experience. It's a great way to challenge yourself and connect with other runners. Have you ever considered joining a running group or club? |
| [R1] | Participating in marathons can be a great way to push yourself physically and mentally. If you ever decide to join a race, how would you approach training? |
| [R2] | Participating in organized running events can be an incredible experience. It's a great way to meet other runners and stay motivated. Have you ever received any awards or recognition for your running achievements? |
| [R3] | Participating in organized running events can be a great way to meet new people and get involved in the running community. Do you have any favorite running gear or accessories that you use during your runs? |
| [R4] | Participating in a marathon can be a great way to test your endurance and build camaraderie with fellow runners. Have you ever participated in any endurance races or competitive events? |
| [R5] | Running marathons can be a truly exhilarating experience. It requires a lot of discipline and perseverance. If you were to run a marathon, which city would you choose? |
| [R6] | Running marathons can be a challenging and rewarding experience. It requires a lot of dedication and perseverance. Have you ever run a shorter distance race? |
| [R7] | Running marathons can be a great challenge and a rewarding experience. Do you have any goals you'd like to achieve in the near future? |
| [R8] | Running marathons can be a great challenge and a fantastic way to push yourself physically and mentally. If you ever decide to participate in a marathon, what kind of training would you do? |
| [R9] | Running events can be a lot of fun and a great way to meet new people with shared interests. Well, it's been a pleasure discussing running with you. Feel free to share more about your running journey anytime! |
| [R10] | Participating in races can be a fun way to challenge yourself and meet other runners. If you were to start running, what would be your first goal? |

Figure 28: PerQy-DLM's 10 generated response candidates to continue a sub-topic conversation about running events and marathons, from a larger conversation about jogging. **R0** is the original turn generated by GPT-3.5.

I Mechanical Turk HIT

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 dialogue 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 29: The HIT design used to evaluate the PerQy-DLM responses against competitive baselines; the authors thank Bowden et al. (2024) for providing the HIT template, which is based on other contemporary synthetic dialogue generation (Kim et al., 2023). Mechanical Turk workers were paid the legal minimum wage (\$15 per hour). Workers qualified for our HIT by having at least a 95% approval rating and at least 1000 approved HITs. We restricted crowdsourced workers to English-speaking countries so that the responses were evaluated primarily by fluent English speakers.

I.1 Inter-Annotator Agreement

We calculated the intraclass correlation coefficient (ICC1) to investigate the inter-annotator agreement among Mechanical Turk workers. Table 2 reports the statistically significant ICC1 values. ICC1 is well-suited for ordinal judgments from multiple raters (Siegel, 1956). The strongest agreement occurred when comparing PerQy-DLM to PerQy, and it is the only comparison where there was statistically significant agreement across all four metrics, highlighting the impact of in-context learning, as PerQy is a zero-shot model. Specificity was the only metric with significant differences across all model comparisons, suggesting it may be the most intuitively understood by workers. Although the ICC1 values indicate low agreement, this is expected, given the subjective nature of response selection in open-domain dialogue, especially since our benchmark models are competitive.

| Model | Metric | ICC1 | <i>p</i> |
|------------|------------|------|----------|
| RedPJ Chat | specific | .066 | .012 |
| COSMO | consistent | .075 | .005 |
| COSMO | natural | .063 | .016 |
| COSMO | specific | .090 | .001 |
| PerQy | consistent | .115 | .000 |
| PerQy | engaging | .115 | .000 |
| PerQy | natural | .120 | .000 |
| PerQy | specific | .120 | .000 |
| GPT-3.5 | engaging | .074 | .008 |
| GPT-3.5 | natural | .056 | .035 |
| GPT-3.5 | specific | .074 | .007 |
| Vicuna-33B | specific | .090 | .002 |

Table 2: The statistically significant ICC1 metrics from the Mechanical Turk evaluation.