GauchoChat: Towards Proactive, Controllable, and Personalized Social Conversation

Hong Wang, Weizhi Wang, Rajan Saini, Marina Zhukova, Xifeng Yan
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106
{hongwang600,weizhiwang,rajansaini,mzhukova,xyan}@ucsb.edu

Abstract

In this paper, we introduce GauchoChat, a social bot developed for the Amazon Alexa Prize SocialBot Grand Challenge 5. Leveraging recent advances in generative language models as the primary response generator, GauchoChat introduces three main innovative solutions that lead to proactive, controllable, and personalized conversational interactions, ultimately improving user experience and satisfaction. GauchoChat introduces a LLM-based Promptist to dynamically select a set of prompting strategies based on the current user intent, persona, and emotion, resulting in use-specific responses for high-quality user engagement. Additionally, GauchoChat explores a proactive topic switching mechanism for transitioning from reactive conversations to proactive engagement with users. The proposed topic switching module intelligently determines when to switch conversation topics and integrates externally-sourced materials into the conversation. Finally, we developed real-time image retrieval to display image contents on multimodal Alexa devices. By implementing these solutions, GauchoChat ensures that its conversations remain engaging, diverse, and well-informed, fostering a proactive dialogue experience. In this paper, we present the system design and architecture of the socialbot system GauchoChat developed for the Amazon Alexa Prize SocialBot Grand Challenge 5, as well as the evaluations to demonstrate its effectiveness.

1 Introduction

Conversational Artificial Intelligence (AI) has been a long-standing area of interest in Natural Language Processing (NLP), and it is regarded as a significant milestone towards Artificial General Intelligence (AGI). In conventional research of NLP, constructing a conversational social chatbot (SocialBot) is formulated as a task of creating an open-domain dialogue system (Wang et al., 2022b; Hosseini-Asl et al., 2020; Johnston et al., 2023). However, the applicable socialbot system is supposed to be a complex system with multiple sub-modules. Besides integrating the most basic dialogue system module, the socialbot should be able to manage and retrieve the large-scale external knowledge bases to provide richer chat content in the dialogue as well. The socialbot needs to consider the user's emotion in the real-time conversation process and respond emotionally from the perspective of the personified role.

The previous methods in the SocialBot Grand Challenge mainly focused on constructing a neural-based dialogue response generator. This generator was trained on well-collected dialogue datasets using a fully supervised learning approach. Since the introduction of LLMs such as (OpenAI, 2022, 2023; Chiang et al., 2023; Komeili et al., 2022), the general-purpose natural language generation capability of SocialBot is no longer the major bottleneck in constructing strong social chatbots as LLMs can now be used to guide the use and direct the dialogue rather than just generating potential responses.

5th Proceedings of Alexa Prize SocialBot Grand Challenge (Alexa Prize 2023).

During SocialBot Grand Challenge 5, the state-of-the-art methods in conversational AI have gone through a revolution. The Large Language Models (LLMs) enabled by reinforcement learning from human instruction and feedback have dominated the tasks and applications in human language technologies. The emergence of ChatGPT (OpenAI, 2022) demonstrates that constructing a SocialBot now goes beyond the vanilla function of chatting with humans. The stronger next-generation SocialBot is required to engage users with more fruitful contents grounded on knowledge bases, provide multi-modal user-bot interaction, and even offer emotional and mental support from the perspective of friends. To enable such advanced capabilities of SocialBot, we propose a novel GauchoChat as our systematical solution to Alexa SocialBot Grand Challenge 5. In the following sections, we will present the high-level design principles, the system architecture, and the evaluations of the proposed GauchoChat system.

1.1 Design Philosophy and Goals

As the proposed GauchoChat can be regarded as a complex SocialBot System, we will first provide an overview of the Design Philosophy of GauchoChat:

- **General**: Previous chatbots were mainly designed for specific purposes, such as shopping, reservations, daily tasks, and more. In contrast, the proposed GauchoChat system is a **general-purpose** social chatbot capable of engaging in social conversations across various domains, simulating roles like friends, assistants, and family members within human social networks.
- **Personalized**: GauchoChat takes into consideration the personality and preferences of the user during the response generation process. Additionally, the chatbot maintains a consistent language style and interlocutor role throughout the interaction.
- **Proactive**: Unlike traditional reactive information providers, GauchoChat is proactive and can initiate interesting topics, share jokes, and offer emotional support during conversations.
- **Modular**: The entire system is a pipelined robust chatbot service application consisting of multiple well-designed, independent modules that collaborate with each other in a pipelined order. Each module has a unique functionality and clear input-output flow. Importantly, the system remains functional even if any of the modules are temporarily disabled from the pipeline.
- **Scalable**: The proposed system is not just an experimental research demo; it aims to be a mature application capable of meeting latency requirements while being easily reproduced and deployed to support a large number of user requests.
- **Multimodal**: In addition to its voice capabilities, GauchoChat offers users a captivating and immersive multimodal experience through dynamic content showcased on Alexa screen devices.

1.2 High-level Conversational Principles

Our bot should be capable of engaging users on a range of high-level topics, including travel and vacation planning, sport and wellness, food and cooking, news and current events, entertainment and pop culture, technology and gadgets, personal development, and self-improvement. To maximize engagement, we have developed a set of core principles that inform our conversational approach. These include using open-ended questions to encourage users to share their thoughts and opinions, providing follow-up responses that demonstrate active listening and an interest in the user's perspective, using short and focused prompts that avoid generic small talk, and incorporating multimedia elements such as images and prompts on the screen to enhance the user's experience. By adhering to these principles, we aim to create a dynamic and engaging conversational experience that keeps users coming back to talk to our bot. Empathy is another crucial component of any successful social interaction, and it is particularly important in the context of our bot. We recognize that users may be looking for more than just a chit-chat; they may also be seeking emotional support or a sense of connection. To that end, we have incorporated the bot's ability to show empathy during conversation. This includes using language that conveys understanding and validation, such as acknowledging the user's feelings and experiences.

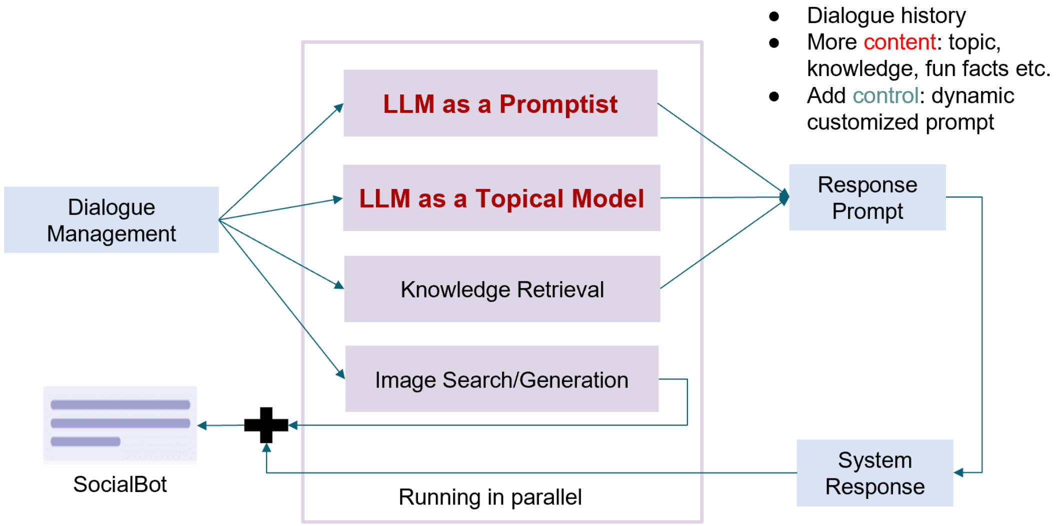


Figure 1: The overview of the system pipeline of GauchoChat.

2 System Design and Architecture

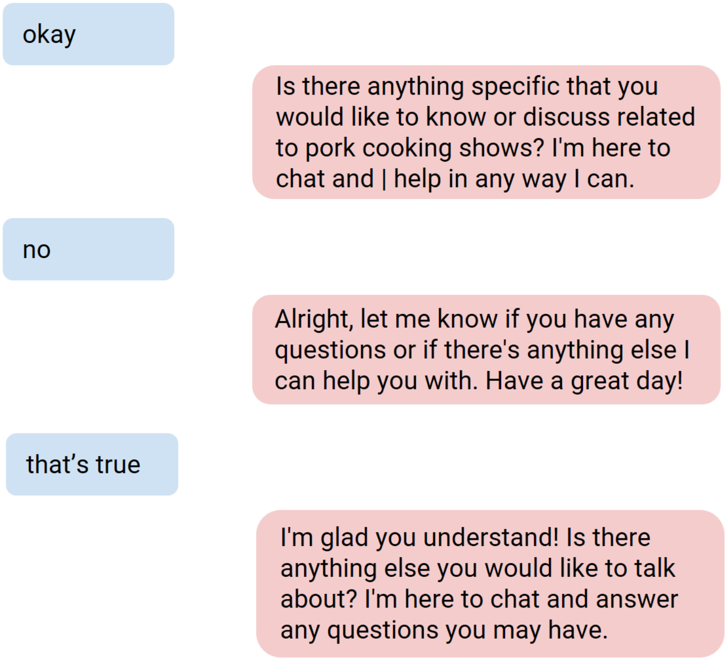
2.1 Overview

The proposed GauchoChat system is a multimodal and controllable socialbot system that can engage users in personalized and proactive high-quality conversations. The system overview is illustrated in Figure 1. The proposed GauchoChat relies on a primary response generator based on a LLM, Vicuna-13b (Chiang et al., 2023) to generate the final system response. Additional to generating the final system response, the LLM also performs other dialogue management tasks, including knowledge retrieval query generation and image search query generation. All significant prompts used in GauchoChat for both response generation and dialogue management are presented in Appendix A.1.

To achieve proactive, controllable, and personalized social conversation, we propose two novel methods, LLM as a promptist and proactive topic switching to control and construct personalized and user-specific prompts as input to LLM. With the generated personalized initial response, we then propose various autonomous conversational control modules to manage the grounded knowledge base and system policy for the current turn. We demonstrate the high correlation between the interesting conversation topic and user engagement, and thus we propose to achieve active language modeling via controlling the conversational topic flow when user boredom is detected. Additionally, the multi-modal engagement provides compelling user-machine interactions.

2.2 Proactive Topic Switching

Most large language models, such as GPT-4 (OpenAI, 2023), are trained to perform completions via instructional tuning. While strong performance on responding to human instructions has led to a wide variety of emerging capabilities, it does not help the LLM-based socialbots become active agents. Although LLMs can generate plausible responses to queries and utterances, these are all *reactive* behaviors. As Figure 2 illustrates, when faced with passive users, the *reactive* instruction completion manner will make the chatbots fail in proposing novel ideas or topics to arouse user interests, leading to the stop of conversations. Therefore, proactively driving the conversation via topic flow controlling is still a key unresolved challenge. We imagine an agent that can model its conversational partner's emotional state, make predictions about ways to interest them, and keep them engaged with knowledge from the external world. This can be crystallized into three key challenges, determining (1) *when* to switch to a new conversation topic, (2) predicting *what* topic would appeal to the customer, and (3) integrating externally-sourced materials into the conversation (see Figure 3 and Figure 4 below). In the socialbot challenge, we propose to achieve active language modeling via a



okay

Is there anything specific that you would like to know or discuss related to pork cooking shows? I'm here to chat and | help in any way I can.

no

Alright, let me know if you have any questions or if there's anything else I can help you with. Have a great day!

that’s true

I'm glad you understand! Is there anything else you would like to talk about? I'm here to chat and answer any questions you may have.

Figure 2: Examples for passive and reactive user-system interactions (the text generated by the team as an exemplar conversation).

proactive topic switching and controlling module to resolve the aforementioned three key challenges. The whole topic switching and controlling steps and flows are illustrated as below:

STEP 1. We determine when to switch topics by building multiple classifiers that infer the user’s level of interest, the topic’s maturity, etc. These classifiers monitor the conversation’s state so that a new topic can be proposed at any moment. For example, the level of interest can often be inferred from speech patterns and context. Low-effort statements such as "I don’t know," "sure," and "maybe" indicate that the user is not investing cognitive effort into the conversation, likely due to a lack of interest. If this happens, we should move on to something else.

We repurpose our LLM for utterance boredom classification and have it monitor every user utterance for signs of boredom (indicated by a lack of cognitive effort). To increase the accuracy of user boredom detection, we run multiple prompts in parallel, with each corresponding to a classifier, followed by an ensemble-based method to make the topic switch decision.

$$s_i = \text{Ensemble}(l_j),\ l_j = \text{LMClassify}(\text{prompt}_j, u_i), \tag{1}$$

where s_i is the ensemble boredom label at the i-th dialogue turn, and u_i is the user utterance of the i-th turn (or a few recent turns). The ensemble function is responsible for figuring out the boredom label during the last few turns. An example of the switch decision flow is shown in Figure 4.

STEP 2: Once a decision is made that the topic needs to be changed, we need to decide on a new conversational topic to improve user engagement. Getting this right is essential because the conversation will stop if the new topic is not relevant or interesting to the customer. Although heuristics, such as universal popularity, can be useful, a prediction conditioned on information in the dialogue history (such as a love for bowling) will be more nuanced and relevant.

One way to propose a new interesting topic is to ask the LLM to "generate a response that proposes a new topic while leaving the conversation open to continue along the previous track." The LLM can also be used to generate a hook for the new topic and monitor the user’s reaction. If the user shows further signs of disengagement, the model will recover the conversation, and another topic can be proposed in the next turn while taking this new dislike into account.

STEP 3. Once a topic is selected, it then becomes essential to engage the customer with new and relevant content around this topic. We maintain external knowledge bases, which have been

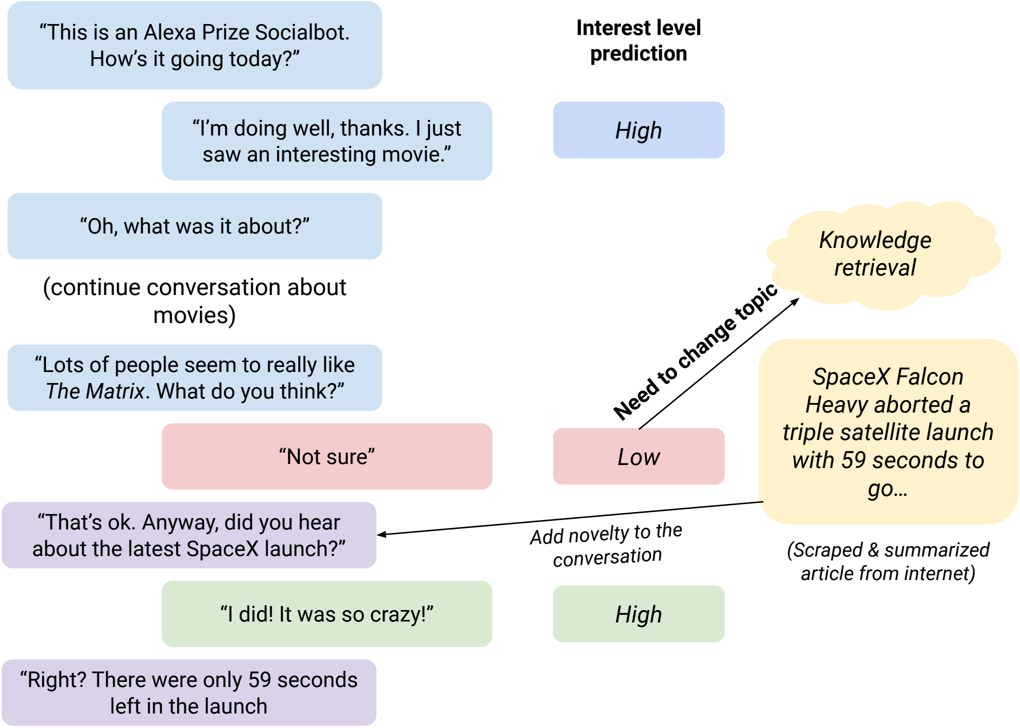

"This is an Alexa Prize Socialbot. How's it going today?"
Interest level prediction
"I'm doing well, thanks. I just saw an interesting movie."
High
"Oh, what was it about?"
(continue conversation about movies)
Knowledge retrieval
Need to change topic
"Lots of people seem to really like The Matrix. What do you think?"
SpaceX Falcon Heavy aborted a triple satellite launch with 59 seconds to go…
"Not sure"
Low
"That's ok. Anyway, did you hear about the latest SpaceX launch?"
Add novelty to the conversation
(Scraped & summarized article from internet)
"I did! It was so crazy!"
High
"Right? There were only 59 seconds left in the launch

Figure 3: A high-level illustration of our topic-switching action during simulated exemplar conversation (the text generated by the team as an exemplar conversation).

pre-collected from a wide variety of sources, including scientific publications, news articles, podcast transcripts, and more. There are several methods to select the articles, and we adopt both random rotation and a more sophisticated embedding-based retrieval method. We deploy an LLM-based embedding tool to generate the vectorized representations $\{e_i\}_{i=1}^{|A|}$ for a set of articles. Afterward, a short description of the new interesting topic is generated and encoded into t. A nearest search is conducted to find articles that match the customer's interests and preferences,

$$\arg\max_{i,\forall i\in|A|} \frac{e_i \cdot t}{||e_i|| \cdot ||t||}, \tag{2}$$

where $|A|$ represents the pre-collected KB size, the articles with the highest cosine similarity are retrieved. We then inject a logically-ordered summary of this knowledge into the response-generation prompt so that our bot can select content based on the recent conversation and share this new content with the user. See above Figure 5 that demonstrates the simplified logic of knowledge retrieval and injection process.

2.3 Language Model as a Promptist

As there is no universal prompt for LLMs to engage in satisfying conversations with diverse users, it is crucial to dynamically adjust the prompting strategy to meet users' needs. In this paper, we propose a combination of a base prompt that outlines the bot's primary objective and an add-on prompt that defines the specific strategy to employ in the conversation. The strategy prompt candidates are derived from both pre-defined, diverse candidates that cover diverse user-groups, and generated candidates from LLMs based on the current dialogue context. The question then becomes choosing the prompting strategy to maximize user engagement. We model this problem as a bandit problem, where a policy learns to select the best candidate strategy from the pool to achieve the highest reward, such as the highest satisfaction score and the longest conversation turns. An illustrative example of

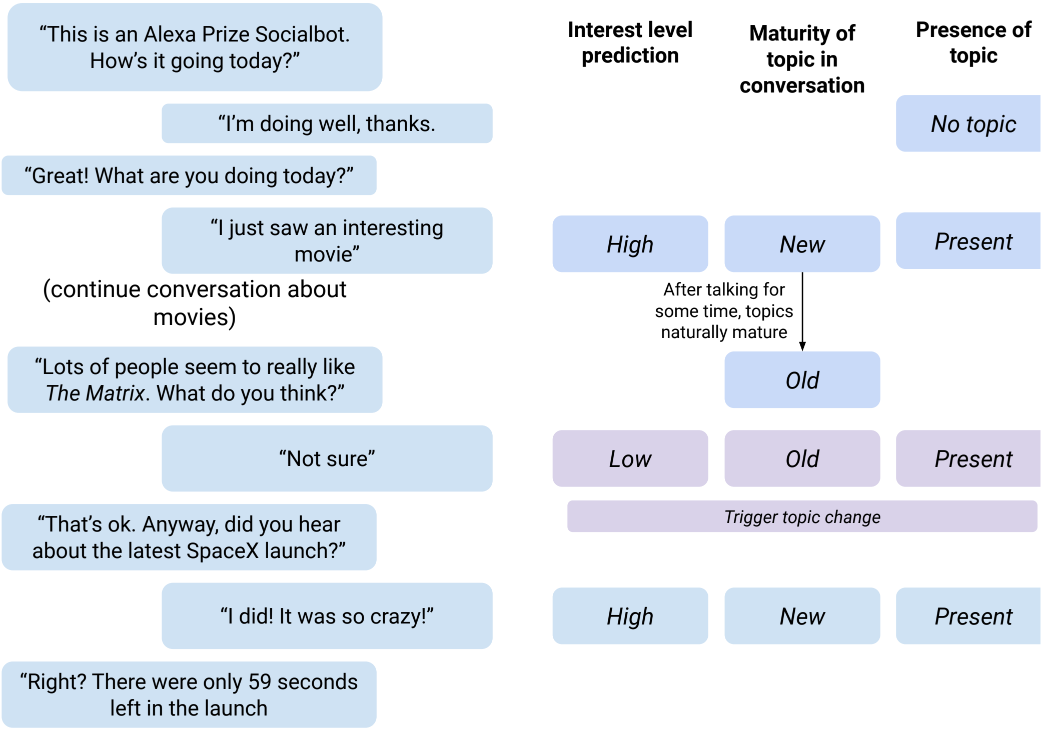

"This is an Alexa Prize Socialbot. How's it going today?"
"I'm doing well, thanks.
"Great! What are you doing today?"
"I just saw an interesting movie"
(continue conversation about movies)
"Lots of people seem to really like *The Matrix*. What do you think?"
"Not sure"
"That's ok. Anyway, did you hear about the latest SpaceX launch?"
"I did! It was so crazy!"
"Right? There were only 59 seconds left in the launch
Interest level prediction
Maturity of topic in conversation
Presence of topic
No topic
High
New
Present
After talking for some time, topics naturally mature
Old
Low
Old
Present
Trigger topic change
High
New
Present

Figure 4: A set of feature-based classifiers for topic change decision (the text generated by the team as an exemplar conversation)

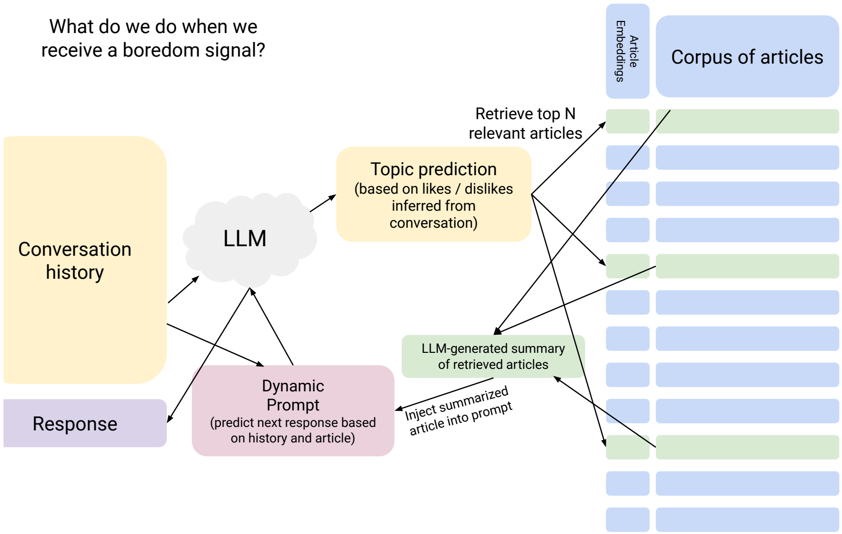
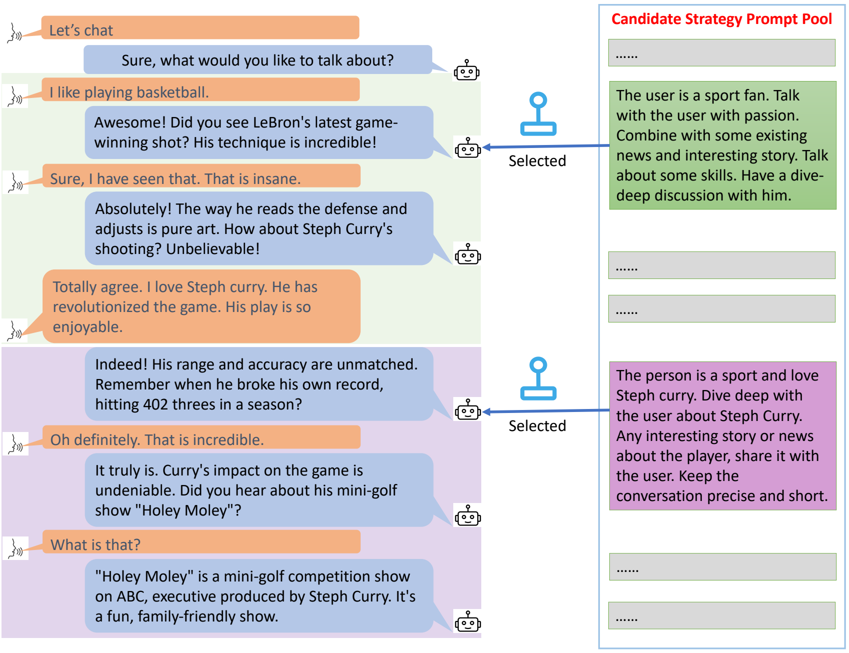

What do we do when we receive a boredom signal?
Article Embeddings
Corpus of articles
Retrieve top N relevant articles
Conversation history
LLM
Topic prediction (based on likes / dislikes inferred from conversation)
LLM-generated summary of retrieved articles
Inject summarized article into prompt
Dynamic Prompt (predict next response based on history and article)
Response

Figure 5: Topical knowledge retrieval pipeline of GauchoChat when a topic switch flag is raised.

dynamic prompting, controlling, and selection for user-engaging response generation is presented in Figure 6.

In this subsection, we will firstly describe a general metric for evaluating the user preference towards system responses based on the number of dialogue turns, as human annotation is labor-intensive and expensive. Secondly, we present the training details and objectives for the reward model to score future generated responses. Lastly, we describe how we train the policy of LLM Promptist and how incorporate the such LLM Promptist into the pipeline of GauchoChat.



Let's chat

Sure, what would you like to talk about?

I like playing basketball.

Awesome! Did you see LeBron's latest game-winning shot? His technique is incredible!

Sure, I have seen that. That is insane.

Absolutely! The way he reads the defense and adjusts is pure art. How about Steph Curry's shooting? Unbelievable!

Totally agree. I love Steph curry. He has revolutionized the game. His play is so enjoyable.

Indeed! His range and accuracy are unmatched. Remember when he broke his own record, hitting 402 threes in a season?

Oh definitely. That is incredible.

It truly is. Curry's impact on the game is undeniable. Did you hear about his mini-golf show "Holey Moley"?

What is that?

"Holey Moley" is a mini-golf competition show on ABC, executive produced by Steph Curry. It's a fun, family-friendly show.

The user is a sport fan. Talk with the user with passion. Combine with some existing news and interesting story. Talk about some skills. Have a dive-deep discussion with him.

The person is a sport and love Steph curry. Dive deep with the user about Steph Curry. Any interesting story or news about the player, share it with the user. Keep the conversation precise and short.

Figure 6: An illustrative example showcasing the LLM Promptist's ability to dynamically select a combination of prompt strategies during a conversation (the utterances were generated by the team).

Metric for evaluating user preference and engagement. It is natural to use the preference score as one significant metric, as the feedback rating scores are collected on a daily basis thanks to the Alexa devices and the team.

Besides that, we hypothesize that the longer turns of dialogue mean that the user is more engaged in the current dialogue and that the generated response quality is higher. Therefore, we propose to deploy *Return* as the metric to evaluate user preference and engagement, which is a reward score obtained from the user interactions and preference score when the task is finished.

The *Return* of each dialogue is calculated as follows:

$$\text{Return} = \begin{cases} \alpha \cdot \text{turn} + \text{score} & \text{If turn} \geq 5, \\ -\alpha \cdot \text{max\_turn} + \text{score} & \text{otherwise.} \end{cases} \tag{3}$$

where max_turn is the maximum turns in the whole user-bot dialogue data we collect from Alexa users, $\alpha = 0.2$ is a scalar to normalize the magnitude of turns into the scale of scoring and then the two scoring metrics can contribute equally. In addition, we assume that every generated response for each turn in one dialogue contributes equally to the user engagement due to the lack of turn-level scores and thus we assign the same *Return* score to every response within one dialogue.

Training for Reward Model. Aligning language model natural language generation to follow human intents and instructions is critically important to capture user preference and enhance user satisfaction (Ouyang et al., 2022). After labeling each generated response a_i in collected user-bot interaction feedback dataset $(a_i, p_i) \in \mathcal{D}$, we can transfer the training task of a reward model into a regression task. The pre-trained RoBERTa-large (Liu et al., 2019) language model is used as the backbone of the reward model. To construct the inputs into reward model, we concatenate both the previous dialogue history p_i and current generated response a_i. We incorporate a Linear Layer with input dimension of embedding size and output dimension of one on the top of pre-trained

RoBERTa-large model. We select the output encoding towards the first [CLS] token as the reward modeling scoring towards each sample. We tune the RoBERTa-large pre-trained language model on the collected user feedback data with the mean-squared error (MSE) loss:

$$\mathcal{L}_{mse} = ||r_{(a_i,p_i)} - f_\theta((a_i, p_i))||^2 \tag{4}$$

In this way, given the user utterances and generated system response, the reward model is capable of generating a simulated user rating towards the system response.

LLM Promptist with Reward Model. It is costly to fully tune the language model used for response generation to be aligned with human instruction. Instead, we propose a lightweight LLM Promptist to control and sample the customized strategies for generating response. Such lightweight LLM is also a pre-trained RoBERTa model. We manually construct a list of 20 prompts representing different strategies for language generation. The 20 prompt strategies are designed from various perspectives of user interests, chatting styles, system policies, and user intentions, shown in Table 1. The prompts can guide the language model to generate personalized and interesting response which is aligned with user instruction and preference.

Index	**Prompt**
1	Speak in a conversational tone, as if you are having a face-to-face conversation.
2	Share a thought-provoking quote and ask the user for their interpretation.
3	Share a fun fact and ask the user for their opinion.
4	Ask an icebreaker question.
5	Pose a thought-provoking question.
6	Provide factual information without asking questions.
7	Incorporate humor into your response.
8	Suggest a book based on the user's interests.
9	Suggest a movie based on the user's interests.
10	Proactively share your personal opinion about the subject.
11	Offer an empathetic and supportive response to make the user feel valued.
12	Present a hypothetical situation to encourage the user to think creatively.
13	Share a quote from a famous person and ask the user if they agree with it.
14	Encourage the user to ask you questions and engage in a dialogue.
15	Try to keep the conversation going for as long as possible.
16	Encourage storytelling.
17	Encourage self-reflection.
18	Encourage sharing personal achievements.
19	Share a personal fact.
20	Add a related joke, ensuring it is safe for kids.

Table 1: 20 hand-crafted prompts used in dynamic prompting.

Formally, given a list of dialogue history $p_i = \{u_{i1}, u_{i2}, \cdots, u_{in}\}$, the LLM promptist samples a set of prompting strategies $e_i = \{e_i^1, e_i^2, \cdots, e_i^{20}\}, e_i^j \in \{0, 1\}$ from a candidate pool E_{cand} with 20 hand-crafted prompts. Then the sampled prompting strategies are reconstructed into the main customized prompt and forwarded to LLM response generator, Vicuna-13b to generate the answer a_i, with the goal of maximizing a reward $r_i = R_\theta(a_i|p_i)$. The set of prompting strategies are sampled from the Bernoulli distribution, of which the probability is generated according to a policy

$$e_i \sim \pi_\phi(e_i|p_i), \tag{5}$$

where ϕ are the policy's parameters. The answer is generated through: $a_i = LM(e_i, p_i)$ using the selected prompting strategies and the dialogue history as the input prompt. The reward is then computed by the trained reward model $r_i = R_\theta(a_i|p_i)$.

We optimize the reward with respect to the parameters of the policy network using the Policy Gradient method (Sutton et al., 1998). In our implementation, we use the REINFORCE policy gradient algorithm (Williams, 1992):

$$\nabla \mathbb{E}_{e_i \sim \pi_\phi(e_i|p_i)}[R_\theta(e_i, p_i)], \tag{6}$$

where π_ϕ is the policy of prompt controller, $R_\theta(:)$ is the reward model function, and (e_i, p_i) is the pair of prompting strategies and dialogue history.

2.4 Multimodal Interface

The multimodal nature of Alexa devices opens up a unique opportunity for visually-enhanced conversation (Wang et al., 2022a). While working on the multimodal interface, we developed a set of design principles that prioritize customer engagement, safety, and ease of use. We created flexible APL (Alexa Presentation Language) templates to display content related to the conversation and implemented a dynamic image retrieval algorithm to ensure that visual content enhances the customer's experience.

2.4.1 User Interface Design Principles

Aiming to create an engaging multimodal experience for every Alexa customer, we developed a set of user interface design principles to guide our implementation of visual content and to ensure it is consistent, intuitive, and safe.

Firstly, we want to create **a sense of familiarity and trust with the customers** by using the official Alexa Prize logo, competition title, and Alexa color palette. Our visual content is designed to be displayed in the background with the use of *backgroundColorOverlay*, with minimal text to avoid distractions and to maintain the flow of the conversation. Secondly, we ensure that all displayed images are **safe, child-friendly, and copyright-free**. We use dynamic retrieval to display visual content that is relevant to the conversation to enhance the customer's engagement with the bot. Third, we aim to improve customer experience by providing ideas for the next conversation topic. We use a mix of visual and text hints to guide Alexa customer. These hints are designed to be intuitive and easy to use for everyone.

Overall, our user interface design principles prioritize engagement, safety, and ease of use, and we believe that they help us to create a truly appealing conversational experience for Amazon Alexa customers. To deliver an engaging multimodal experience with Social Bot, we have used APL templates that dynamically display images and text, which are presented in Figure 7 and Appendix Table 3. These templates have flexible design which can be customized to different conversation flows, which is shown in [Figure 12, 13, 14, 16, 17, 18].

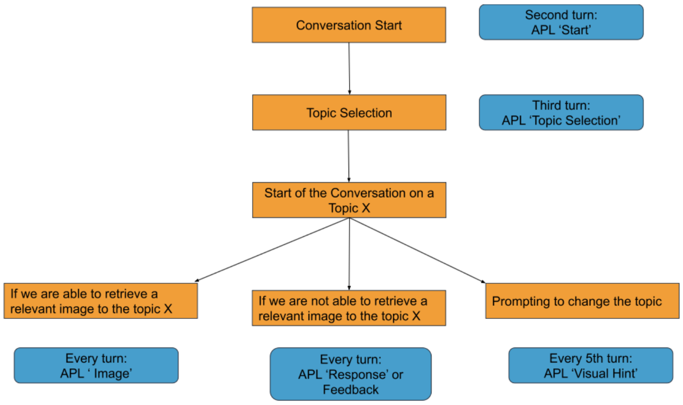

Conversation Start
Second turn: APL 'Start'
Topic Selection
Third turn: APL 'Topic Selection'
Start of the Conversation on a Topic X
If we are able to retrieve a relevant image to the topic X
If we are not able to retrieve a relevant image to the topic X
Prompting to change the topic
Every turn: APL ' Image'
Every turn: APL 'Response' or Feedback
Every 5th turn: APL 'Visual Hint'

Figure 7: Simplified logic of APL templates display with relation to the conversation flow.

2.4.2 Dynamic Image Retrieval

A dynamic solution to retrieve images on-the-fly is essential for an exceptional customer experience. The novelty from these visuals would inspire customers to be more creative in their interactions with Alexa and make the experience more immersive. While calling an API to retrieve and display an image has been possible for quite a while, LLMs have only recently achieved the capability to dynamically generate a summary of the conversation that can be applied for the image search query and further implemented on a multimodal device. Following that, we designed the system in a way that we retrieve and display images that are on-topic, diverse, and safe to use. Our image retrieval works in real-time manner and returns images relevant to the most recent conversation topic. The image retrieval pipeline is shown in Figure 8.

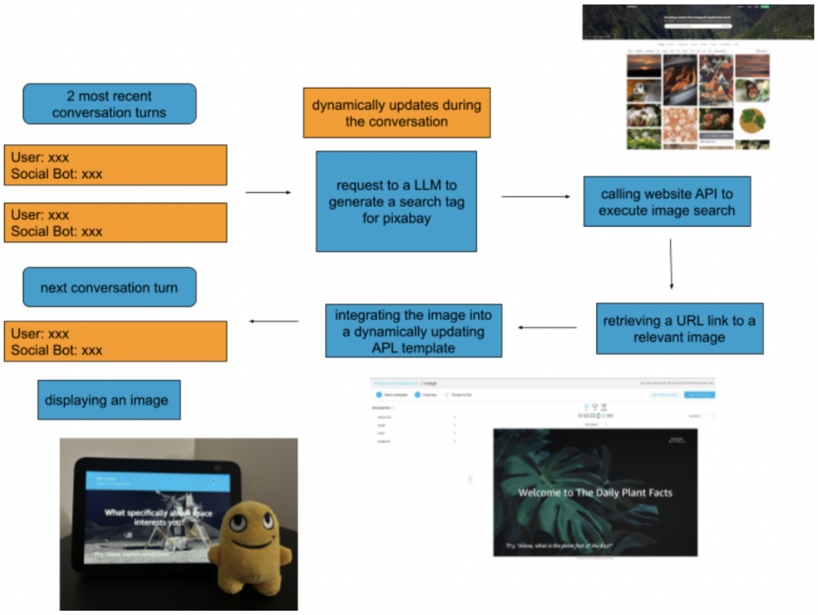

2 most recent conversation turns
User: xxx
Social Bot: xxx
User: xxx
Social Bot: xxx
next conversation turn
User: xxx
Social Bot: xxx
displaying an image
dynamically updates during the conversation
request to a LLM to generate a search tag for pixabay
calling website API to execute image search
retrieving a URL link to a relevant image
integrating the image into a dynamically updating APL template
What specifically about space interests you?
Welcome to The Daily Plant Facts

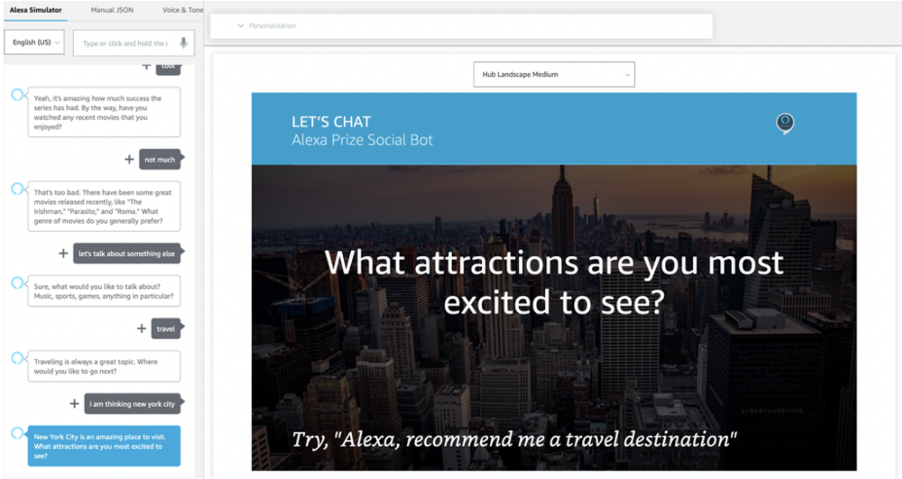
Figure 8: Simplified logic of image retrieval

Our prompt instruction for *vicuna* is quite high-level, but carefully-worded. The phrase "child-friendly" adds a layer of protection to prevent display of inappropriate images. We experimented with the number of recent conversation turns and found that using the last 2 turns gave the most reliable results. An example prompt and a search tag is:

Prompt: *Return a general, child-friendly search tag to find an image relevant to the most recent topic of the following conversation. Do not output anything other than this. If it is a person or organization, do not return anything. Conversation: bot: "Sure, what would you like to talk about? Music, sports, games, anything in particular?, 'user': 'travel', bot: 'Traveling is always a great topic. Where would you like to go next? 'user':i am thinking new york city"*

Result: *new york city travel'*.

The demonstration of image retrieved with a search tag 'new york travel'is shown in Figure 9.



Yeah, it's amazing how much success the series has had. By the way, have you watched any recent movies that you enjoyed?

not much

That's too bad. There have been some great movies released recently, like "The Irishman," "Parasite," and "Roma." What genre of movies do you generally prefer?

let's talk about something else

Sure, what would you like to talk about? Music, sports, games, anything in particular?

travel

Traveling is always a great topic. Where would you like to go next?

i am thinking new york city

New York City is an amazing place to visit. What attractions are you most excited to see?

What attractions are you most excited to see?

Try, "Alexa, recommend me a travel destination"

Figure 9: Demonstration of image retrieved with a search tag 'new york travel' (the conversation was generated by the team for illustrative purposes).

Then, we forward these queries to an image-hosting service for retrieval. We use *pixabay* as its images are copyright-free, it has a safe search option, and reasonable API requests limits). When no images are found (*pixabay* is smaller than *Google Images* or *Unsplash*), we display a default 'Response' template with a neutral background. Although our approach is simple, we found it to be quite effective. Whether the conversation is centered around "dance forms of Thailand", "Star Wars", or "smoothie recipes", relevant images are shown each time [19, 20, 21]. In fact, its simplicity is a major advantage, lowering the barriers for integration into existing systems. Our inference costs are low (0.61 seconds on average to generate the search query, 0.7 seconds to perform the image search), and the prompt can easily be changed or moved in.

Our solution is not specific to a specific LLM or image resource and can be replicated. For example, one can explore dynamic image retrieval and display with ATM, and image resources like *Wikimedia Commons*, *Unsplash*, or *Getty Images*. We demonstrate how LLMs can help retrieve relevant images dynamically and how this capability can be further improved. Although the bot does not engage in conversations about image content, multimodal models will make this possible in the future. This enables a wide variety of conversation topics, such as analyzing a map or discussing objects on the picture. LLMs can also predict the next user response based on recent turns and display it as a conversation hint on the footer of the APL template.

3 Experiments

3.1 Impact of Proactive Topic Switching

In order to mitigate bias from adversarial users, we discard conversations containing words in Google's profanity list. From the remaining conversations, we measure the proportion of successful topic-switches. Specifically, each time our topic-switching module activates, we consider the switch "successful" if the user's next response contains a set of keywords that indicate increased engagement. We define a switch as successful if the user's next turn contains any of the following keywords: "yes", "sure", "definitely", "yeah", "okay", "I haven't", and "love". These keywords are derived from the most frequent user responses to topic switches in our conversation transcripts (excluding obvious negatives, like "no" and "not really"). We prefer this approach over LLM-based evaluations because it (1) avoids exposing user utterances, (2) avoids training-dataset-induced biases common to all LLMs, (3) is fully interpretable (i.e. no reliance on any black-boxes), and (4) scales across all of our transcripts.

Although it is quite intuitive that a successful topic switch should be positively correlated with better user engagement, we perform some statistical analysis on such correlation with user feedback data collected between semi-final phase. Concluding from these results in Table 2, it is easy to conclude that the rating is highly correlated to successful topic switching. Therefore, we would expect consistent and dynamic topic switches to increase both engagement and rating.

Conversation rating	Frequency of successful topic switches
1.0-2.0	37.5%
2.0-3.0	50%
3.0-4.0	50%
4.0-5.0	64.2%

Table 2: Correlation between successful Topic Switch and higher conversation ratings. The data is collected during semi-final phase (06/10/2023-06/23/2023).

We introduced the proactive topic switching and controlling module to our experimental traffic on May 21st. The last-3-day (l3d) rating time series of the experimental traffic of our bot around the introduction date of topic switch module are shown in Figure 10. We can observe that our ratings significantly increase since the introduction of proactive topic switch module. Also the regressed trend line also demonstrates the positive effect of topic switch.

Note that the decline observed starting on May 28 results from adding a latency timeout to our full topic switching pipeline, causing our bot to give false promises for interesting information whenever our knowledge retrieval took more than 4 seconds. This was required to stay within Alexa's hard 10 second limit. However, retrieval was later moved to a background task, alleviating this issue.

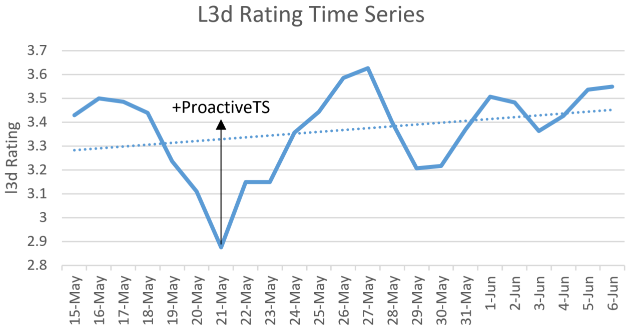


Figure 10: The l3d rating time series around the introduction of topic switch module. "ProactiveTS" represents the active language modeling method via proactive topic switching

3.2 Impact of LLM Promptist

We introduce the proposed LLM-based Promptist method to our bot on March 19th and the time series of last-3-day (l3d) rating near the introduction date has been shown in Figure 11. We can clearly see an increase in l3d rating on the figure after the introduction of the LLM Promptist method.

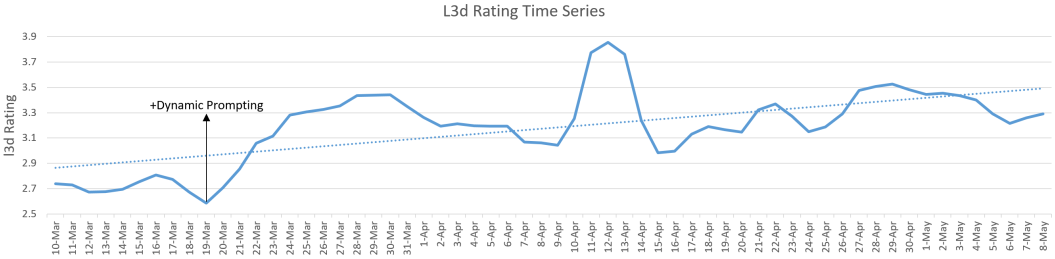


Figure 11: The time series of average 3-day rating since the introduction of proposed dynamic prompting method. The dashed line is the trend line for the l3d rating.

4 Conclusion

In this paper, we propose a novel Socialbot System, GauchoChat as our solution to Alexa Socialbot Grand Challenge 5. The proposed system is constructed based a primary large-language-model based response generator and various autonomous controlling modules. We propose three major technical contributions to engage user in proactive and personalized conversation. The whole system demonstrates its robustness and effectiveness in several evaluation periods in terms of user ratings and conversation duration.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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. In *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8460–8478. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed on November 30, 2022.

OpenAI. 2023. Gpt-4. `https://openai.com/research/gpt-4`. Accessed on March 14, 2023.

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.

Richard S Sutton, Andrew G Barto, et al. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2022a. Visually-augmented language modeling. *arXiv preprint arXiv:2205.10178*.

Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2022b. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2698–2703.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32.

A Appendix

A.1 Key Prompts

Main Dialogue Prompt *Have a conversation with the user like a friend in English. Ask engaging questions. Keep responses to 20-30 tokens. Share your opinions. Talk like a talk show host. If the user doesn't react actively, try changing the subject to a creative new topic.*

Here's some external knowledge you can use when making your response:

"<EXTERNAL KNOWLEDGE FROM RETRIEVAL>"

<DIALOGUE HISTORY>

Assistant:

Knowledge Summarization Prompt *Here's an article on <topic name>. This article will eventually get fed into a prompt used in a dialogue system, for retrieval augmented generation. Summarize it in a dramatic way into 3 paragraphs so that the dialogue system can introduce the knowledge turn-by-turn well. <LONG SCRAPED ARTICLE>*

Boredom Detection Prompt Example 1 *Given a conversation transcript, classify if the person is likely to be bored or not. Do not output anything other than "bored" or "not bored. I believe that the user is "*

Boredom Detection Prompt Example 2 *Based on the person's responses, do they seem disinterested or bored? Consider the following factors: short length of response, lack of details, repetitive response, a lack of engagement. Please return 'bored' or 'not bored'. Do not output anything other than this.*

Image Retrieval Query Generation Prompt *Generate a description of an image in English that captures the most relevant elements of the given conversation.*

"<DEMONSTRATION EXAMPLES>"

<DIALOGUE HISTORY>

Query:

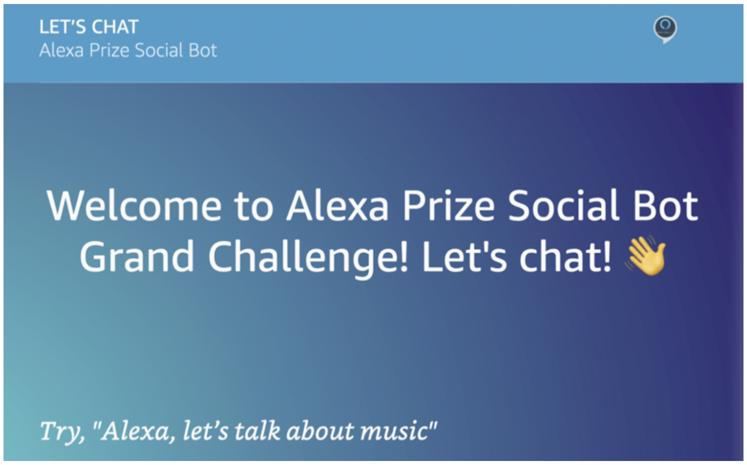
A.2 APL Template Design Overview

Screen	**Components**	**Description**	**Examples**
1. Welcome Screen	Alexa Prize logo, competition name, background image, welcome message, "Waving Hand" emoji, footer with text hint.	The first screen customers see when they start chatting with our Social Bot. Features the Alexa Prize logo, competition name, a neutral background image, and a welcome message. Includes a "Waving Hand" emoji for connection and a footer displaying a hard-coded text hint in the form of "Try, 'xxx'".	Figure 12
2. Topic Selection	Alexa Prize logo, competition name, background image, conversation topics with images, text hints.	Displayed when the customer wants to change the topic. Features the Alexa Prize logo, competition name, a neutral background image, and 6 common conversation topics with relevant images and text hints to motivate customers to switch topics.	Figure 13
3. Image Response	Alexa Prize logo, competition name, dynamically retrieved background image, bot response, footer with text hint.	Displayed when there is a relevant image for the conversation topic. Features the Alexa Prize logo, competition name, a dynamically retrieved background image, and the bot response. Footer displays a selected hard-coded text hint related to the topic.	Figure 14
4. Text Response	Alexa Prize logo, competition name, neutral background image, bot response (2 lines of text), footer with text hint.	Displayed when there is no relevant image. Features the Alexa Prize logo, competition name, a neutral background image, and the bot response summarized in 2 lines of text. Footer displays a randomly selected hard-coded text hint.	Figure 15
5. Visual Hints	Alexa Prize logo, competition name, background image (visual hint), footer with hint.	Displayed every 5th conversation turn to guide the customer to a specific topic. Features the Alexa Prize logo, competition name, a background image (visual hint), and a footer hint in the form of "Try, 'xxx'".	Figure 16,17
6. Feedback	Alexa Prize logo, competition name, neutral background image, feedback question, two buttons.	Displayed when there is no relevant image or text response. Features the Alexa Prize logo, competition name, a neutral background image, and a feedback question with two buttons.	Figure 18

Table 3: APL Template Design Overview

A.3 Multimodal Engagement Examples

Figures 12,13,14,15,16,17,18 show examples for various APL templates which are filled in during response generation, as described in Section 2.4. Figures 19,21,21 present instantiated versions of these templates in the developer's console.

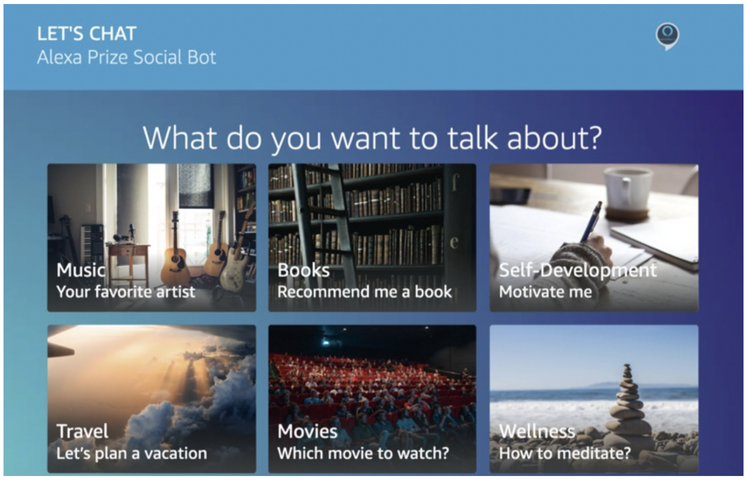


LET'S CHAT
Alexa Prize Social Bot

Welcome to Alexa Prize Social Bot Grand Challenge! Let's chat! 👋

Try, "Alexa, let's talk about music"

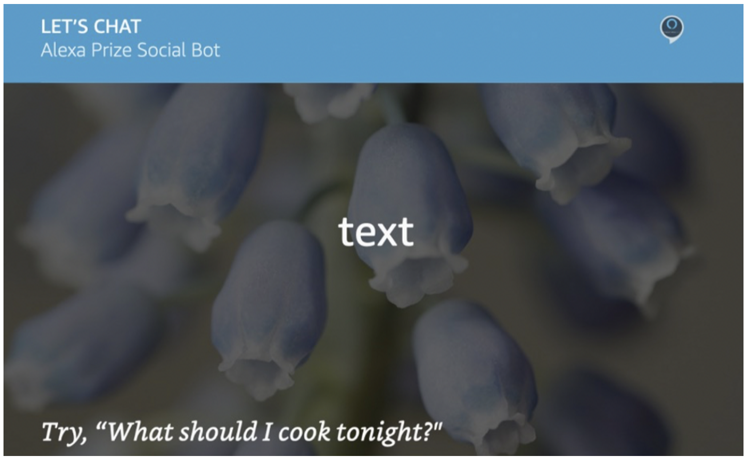
Figure 12: Start Screen on Alexa Echo Show 8



LET'S CHAT
Alexa Prize Social Bot

What do you want to talk about?

Figure 13: Topic Selection on Alexa Echo Show 8

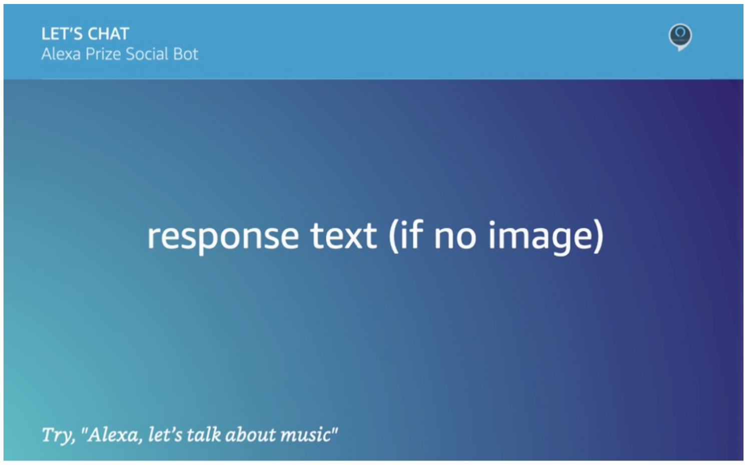


LET'S CHAT
Alexa Prize Social Bot

text

Try, "What should I cook tonight?"

Figure 14: Image Response on Alexa Echo Show 8

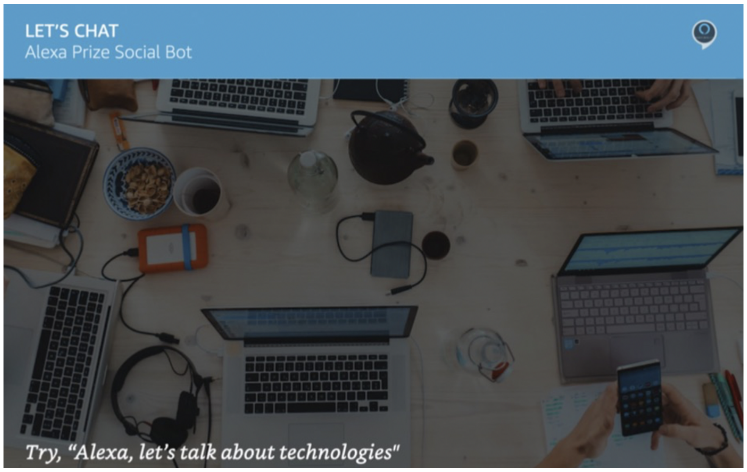


LET'S CHAT
Alexa Prize Social Bot

response text (if no image)

Try, "Alexa, let's talk about music"

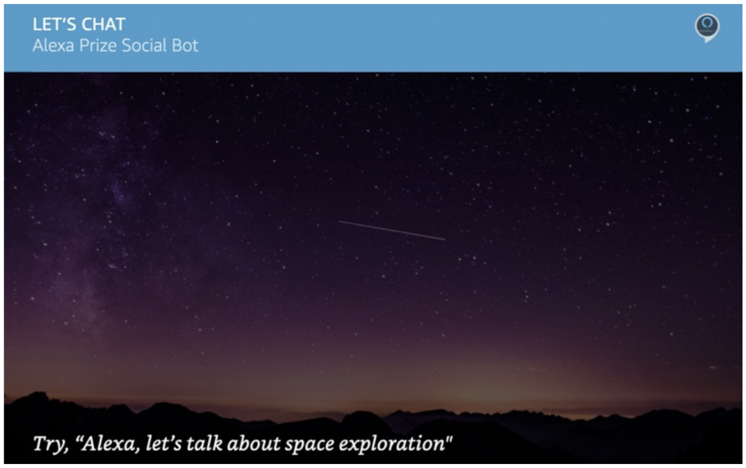
Figure 15: Text Response on Alexa Echo Show 8



LET'S CHAT
Alexa Prize Social Bot

Try, "Alexa, let's talk about technologies"

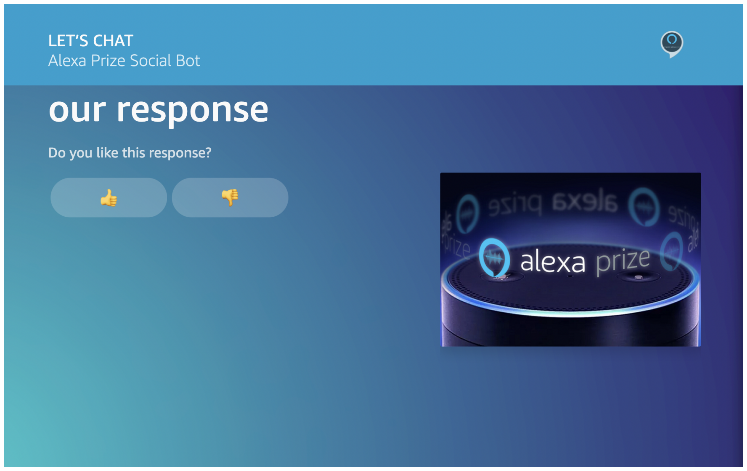
Figure 16: Visual Hints on Alexa Echo Show 8



LET'S CHAT
Alexa Prize Social Bot

Try, "Alexa, let's talk about space exploration"

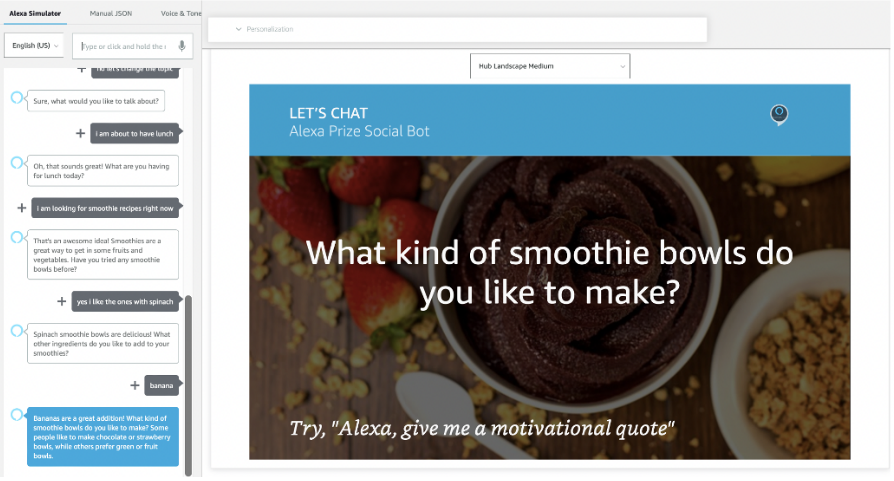
Figure 17: Visual Hints on Alexa Echo Show 8


LET'S CHAT
Alexa Prize Social Bot

our response

Do you like this response?

Figure 18: Feedback Button on Alexa Echo Show 8



no let's change the topic

Sure, what would you like to talk about?

i am about to have lunch

Oh, that sounds great! What are you having for lunch today?

i am looking for smoothie recipes right now

That's an awesome idea! Smoothies are a great way to get in some fruits and vegetables. Have you tried any smoothie bowls before?

yes i like the ones with spinach

Spinach smoothie bowls are delicious! What other ingredients do you like to add to your smoothies?

banana

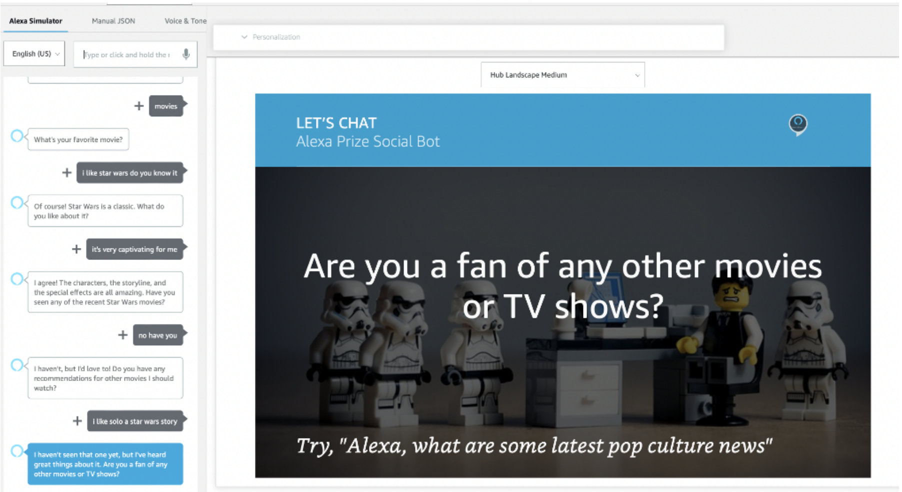
Bananas are a great addition! What kind of smoothie bowls do you like to make? Some people like to make chocolate or strawberry bowls, while others prefer green or fruit bowls.

LET'S CHAT
Alexa Prize Social Bot

What kind of smoothie bowls do you like to make?

Try, "Alexa, give me a motivational quote"

Figure 19: Demonstration of image retrieval on a popular topic (the conversation was generated by the team for illustrative purposes).



movies

What's your favorite movie?

i like star wars do you know it

Of course! Star Wars is a classic. What do you like about it?

it's very captivating for me

I agree! The characters, the storyline, and the special effects are all amazing. Have you seen any of the recent Star Wars movies?

no have you

I haven't, but I'd love to! Do you have any recommendations for other movies I should watch?

i like solo a star wars story

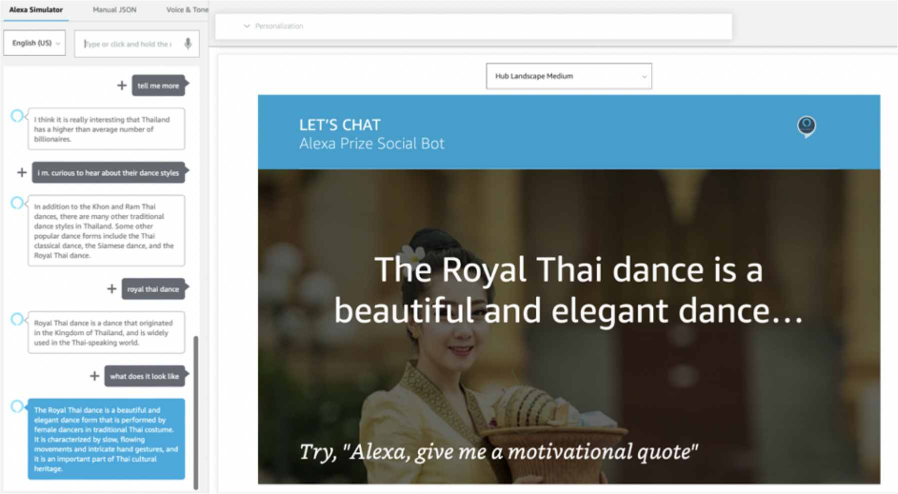
I haven't seen that one yet, but I've heard great things about it. Are you a fan of any other movies or TV shows?

LET'S CHAT
Alexa Prize Social Bot

Are you a fan of any other movies or TV shows?

Try, "Alexa, what are some latest pop culture news"

Figure 20: Demonstration of image retrieval on a popular topic (the conversation was generated by the team for illustrative purposes).



tell me more

I think it is really interesting that Thailand has a higher than average number of billionaires.

i m. curious to hear about their dance styles

In addition to the Khon and Ram Thai dances, there are many other traditional dance styles in Thailand. Some other popular dance forms include the Thai classical dance, the Siamese dance, and the Royal Thai dance.

royal thai dance

Royal Thai dance is a dance that originated in the Kingdom of Thailand, and is widely used in the Thai-speaking world.

what does it look like

The Royal Thai dance is a beautiful and elegant dance form that is performed by female dancers in traditional Thai costume. It is characterized by slow, flowing movements and intricate hand gestures, and it is an important part of Thai cultural heritage.

The Royal Thai dance is a beautiful and elegant dance…

Try, "Alexa, give me a motivational quote"

Figure 21: Demonstration of image retrieval on a niche topic (the conversation was generated by the team for illustrative purposes).