CharmBana: Progressive Responses with Real-Time Internet Search for Knowledge-Powered Conversations

Revanth Gangi Reddy, Sharath Chandra, Mankeerat Singh Sidhu, Hao (Jack) Bai,
Wentao Yao, **Pranav Pillai**, **Karan Aggarwal**, **Liliang Ren**, **Prathamesh Sonawane**,
Keyu Han, **Varun Goyal**, **Stuti Agrawal** and **Chengxiang Zhai**
Department of Computer Science
University of Illinois at Urbana-Champaign
`{revanth3, sce3, mssidhu2, haob2, wentaoy4, ppillai3, karana5, liliang3, pks10, keyuhan2, varun15, stutia3, czhai}`@illinois.edu

Abstract

We introduce CharmBana, a cutting-edge social chatbot created by the team at the University of Illinois at Urbana-Champaign to compete in the Alexa Prize SocialBot Grand Challenge 5. CharmBana's primary innovation lies in its ability to harness Internet search and proactively gather the latest information and knowledge from the web, relating to any given subject in real-time, and seamlessly integrate it into a conversation through progressive response generation. We delve into CharmBana's system design and architecture, elucidate the implementation specifics of major modules, and explore the research and development undertakings throughout the competition to enhance CharmBana. Furthermore, we share valuable insights gleaned from experiments and the competition, emphasizing novel methods devised for progressive response generation utilizing real-time internet search, and presenting preliminary findings on the automated evaluation of conversation quality.

1 Introduction

The Alexa Prize SocialBot Grand Challenge[1] provides resources and opportunities to study and advance the state of the art technologies in developing open-domain social bots that can conduct meaningful conversations on any topic with human users. The University of Illinois at Urbana-Champaign (UIUC) team has developed a novel chatbot, called CharmBana[2], for participating in the Alexa Prize SocialBot Grand Challenge 5 (Johnston et al., 2023). Past teams in the Alexa Prize SocialBot Grand Challenge have made innovation, including, e.g., user personalization (Konrád et al., 2021) and modeling (Juraska et al., 2021), structured dialog control (Chi et al., 2021), templated generation (Chi et al., 2022) and incorporating knowledge graphs (Pichl et al., 2020).

Despite advancements in chatbot technology, current systems are yet to achieve the goals set by Amazon, as numerous complex technical obstacles must be overcome.

Effective chatbots must engage users in casual, informal conversations on a variety of topics that cannot be predetermined. To create captivating and informative discussions, chatbots need to access external knowledge to enrich the dialog. Engaging in knowledge-driven conversations (Komeili et al., 2022; Dinan et al.) typically requires leveraging external sources to gather the most relevant information for continuing the conversation. Chatbots might use a pre-built knowledge base to obtain

---

[1] https://www.amazon.science/alexa-prize/socialbot-grand-challenge

[2] CharmBana is meant to be a “Charming Social Bot"

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

information about a subject, but the information may be outdated or unsuitable for discussing current topics. Similarly, a large language model (LLM) trained with a static text collection, regardless of size, would face the same limitations. Utilizing a web search engine could potentially address this issue by explicitly generating search queries that help retrieve the most current information from the web in real-time.

While such queries are more obvious when the user explicitly asks for certain information, a.k.a. conversational information seeking (Zamani et al., 2022), it is, however, unclear what information should be pursued when users are passive, disengaged, and do not provide clear guidance for the conversation (Hardy et al., 2021). Additionally, as open-domain conversations allow users to discuss any topic, developing a comprehensive algorithm that offers relevant queries in response to a random topic presents a unique and complex challenge unexplored in previous studies. Furthermore, generating responses based on real-time web search results inevitably leads to delays in response times. Thus, addressing the latency issue associated with real-time internet searches poses yet another challenge.

In this competition, we address these challenges and devise innovative techniques to allow CharmBana to harness the power of Internet search in proactively obtaining the latest information and knowledge related to any given subject in real time while seamlessly integrating it into a conversation through progressive response generation. Specifically, we propose a novel methodology that integrates social commonsense reasoning into the generation of search queries in knowledge-powered conversations. Social commonsense (Moore, 2006) refers to the general understanding of social situations and human behavior that helps connect conversational topics and navigate discussions.

We therefore hypothesize that by leveraging a deeper understanding of social commonsense and the implicit cues that steer human conversation, chatbots can become more proficient at navigating passive conversations. Concretely, we introduce a novel framework that uses a commonsense response as a *latent directive* for an instruction-following query generator. Our approach incorporates the use of topic tracking to first identify the main point of discussion, followed by generating commonsense-based responses that can link concepts to the main topic, resulting in a latent commonsense directive. Finally, we use instruction-driven query generation to produce a search query that adheres to the latent directive within the commonsense response.

To tackle the issue of latency, we propose using a progressive response generation strategy aimed at reducing the user's waiting time. This involves the chatbot initially delivering a swift, general response, followed by a more comprehensive and detailed answer based on internet search. In the competition, we utilized a commonsense response as our preliminary response (refer to §2.3) and integrated a follow-up flow (discussed in §2.5). As discussed later in this report, our innovative techniques for progressive responses combined with real-time internet search have proven effective, significantly enhancing CharmBana's performance.

We have also studied the challenge of automating the evaluation of conversation quality in chatbots. This would allow for estimating ratings for numerous discussions that lack manual ratings and potentially also serve as a reward function for reinforcement learning to optimize dialog strategies. As part of the scientific track, we introduced an innovative information-theoretic metric to evaluate conversation quality at turn level across multiple dimensions. This new measure has been demonstrated to outperform several existing state-of-the-art metrics, enabling a more detailed analysis of our chatbot's performance. Furthermore, we assessed various representative, cutting-edge machine learning algorithms for predicting human ratings of these conversations. Both classification and regression tasks were utilized, with results indicating that leveraging machine learning for automated assessment is a promising direction.

In the rest of this technical report, we will present CharmBana's system design and architecture, outline the implementation specifics of its primary components, and delve into the research and development efforts undertaken during the competition to enhance CharmBana. Additionally, we will discuss several promising future research avenues to advance the cutting-edge of social chatbot technology, building upon the methodologies we already devised.

2 System Design and Implementation

In this section, we provide an overview of CharmBana's design and implementation specifics. Our framework operates on the principle of "*progressive response generation*", where a quick general response is initially given to the user, followed by a comprehensive, search-driven follow-up response for a more in-depth reply. The progressive response strategy is meant to simulate how a human makes the tradeoff between giving a response quickly and giving a well-thought response (which may require more processing time). This strategy allows for progressive refinement or construction of a response, naturally leading to an anytime algorithm that can flexibly support multiple response generation algorithms that have variable computational complexity.

The system's architecture is illustrated in Figure 1 and comprises the following key elements, which are discussed further below:

Dialog Manager: This component orchestrates the dialog strategy, focusing on the selection of appropriate responders (i.e. *responder selection*) and determining whether an internet search is needed (i.e. *search decision*).

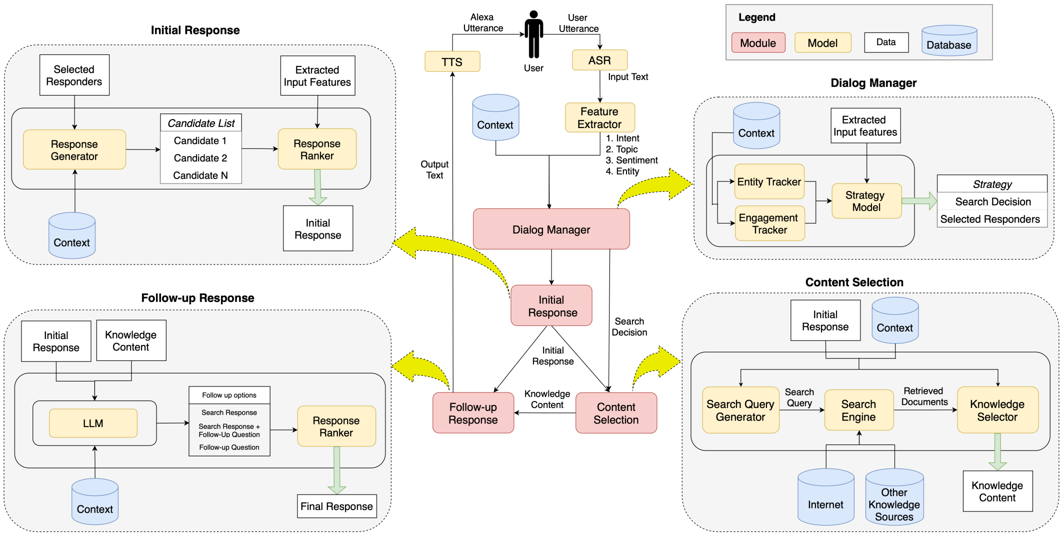
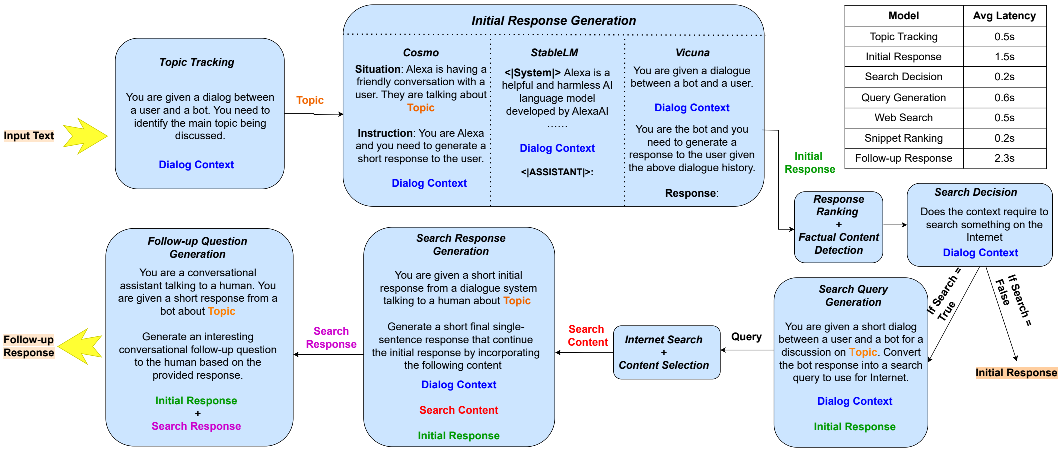


Figure 1: The overall framework for CharmBana following the progressive responses methodology with dedicated modules for initial and follow-up responses.



Topic Tracking

You are given a dialog between a user and a bot. You need to identify the main topic being discussed.

Dialog Context

Initial Response Generation

Cosmo

Situation: Alexa is having a friendly conversation with a user. They are talking about Topic

Instruction: You are Alexa and you need to generate a short response to the user.

Dialog Context

StableLM

<|System|> Alexa is a helpful and harmless AI language model developed by AlexaAI

……

Dialog Context

<|ASSISTANT|>:

Vicuna

You are given a dialogue between a bot and a user.

Dialog Context

You are the bot and you need to generate a response to the user given the above dialogue history.

Response:

Model	Avg Latency
Topic Tracking	0.5s
Initial Response	1.5s
Search Decision	0.2s
Query Generation	0.6s
Web Search	0.5s
Snippet Ranking	0.2s
Follow-up Response	2.3s

Response Ranking + Factual Content Detection

Search Decision

Does the context require to search something on the Internet

Dialog Context

Search Query Generation

You are given a short dialog between a user and a bot for a discussion on Topic. Convert the bot response into a search query to use for Internet.

Dialog Context

Initial Response

Internet Search + Content Selection

Search Response Generation

You are given a short initial response from a dialogue system talking to a human about Topic

Generate a short final single-sentence response that continue the initial response by incorporating the following content

Dialog Context

Search Content

Initial Response

Follow-up Question Generation

You are a conversational assistant talking to a human. You are given a short response from a bot about Topic

Generate an interesting conversational follow-up question to the human based on the provided response.

Initial Response
+
Search Response

Figure 2: Flowchart showing the flow of information in CharmBana via different prompts, along with the average latency for each component.

Initial Response Generation: This module produces a basic yet coherent initial response, paving the way for a more informed follow-up reply.

Content Selection: Employed for choosing pertinent content from the internet to enhance the follow-up response, this module is activated only when the “search decision” is true. We note that the content selection module runs only when the “search decision” is true.

Follow-up Response Generation: This final component integrates the extracted search information to formulate a follow-up response that builds upon the chatbot’s initial reply.

2.1 Architecture

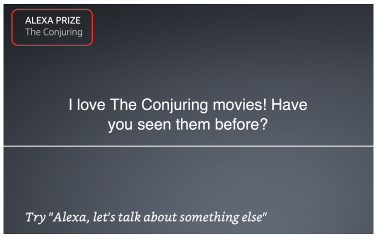
Our architecture employs the Cobot toolkit (Ram et al., 2018), constructed with a modular design approach, enabling developers to easily alter or replace components or modules as needed. Cobot, as a system, facilitates the integration of remote modules for offloading processing while managing the required logic, akin to a microservice architecture. With this principle in mind, our primary focus was on implementing selection and ranking strategies using custom logic, while delegating more computationally intensive tasks to remote modules. These tasks encompass topic tracking, LLM-based response generation, internet search, and more. Additionally, the Cobot framework offers features for Named Entity Recognition (NER), dialogue intent detection, and punctuation correction, which we utilized.

2.2 Dialog Manager

2.2.1 Topic Tracking

Topic tracking (Nakata et al., 2002) aims to identify the primary subject of the discussion in free-form dialogs, and has been demonstrated (Guo et al., 2018) to improve the coherence of dialog systems. Unlike previous approaches (Khatri et al., 2018) that track a fixed set of broad high-level topics (e.g., *movies*, *sport*), our objective is to detect unconstrained, finer-grained topics (such as *movie/actor* names or *teams*). For fine-grained topic tracking, we apply an instruction-tuned model (Chung et al., 2022) to identify the current topic from the dialog context.

We utilize the prompt in Figure 2 and rely on instruction-tuned models with strong zero-shot abilities (Wei et al., 2021) due to lack of training data for such topic tracking in dialog[3]. In particular, we employ Flan-T5 large (770M) for the purpose of topic tracking. To enhance its performance, we generate silver data by obtaining ChatGPT labels for topic tracking on the Wizard of Internet (WoI) training set (Komeili et al., 2022). Additional information regarding performance enhancements from[4] are provided in §3.1. Notably, the topic tracking results are also displayed on the screen of the Alexa device, as illustrated in Figure 3. An alternative topic tracking approach could follow Shuster et al. (2022a); Adolphs et al. (2022), extracting topics as relevant entities grounding the final response.



ALEXA PRIZE
The Conjuring

I love The Conjuring movies! Have you seen them before?

Try "Alexa, let's talk about something else"

Figure 3: Topic tracking output as shown on the Alexa device.

2.2.2 Engagement Classifier

In determining how to seamlessly transition within dialogues, we found that employing an engagement tracker enhances our bot’s capability to manage dialog shifts. Drawing inspiration from Ghazarian et al. (2020)’s predictive engagement metric, we designed a model that uses a Bi-directional Long Short-Term Memory (Bi-LSTM) network that takes BERT embeddings as inputs. This setup leverages the strengths of both models - the sequential data processing of Bi-LSTM and the semantic understanding of BERT. Subsequently, these embeddings are processed through a Multi-Layer Perceptron (MLP) classifier, which is structured with three-layer networks comprising 64, 32, and 8 hidden

---

[3] We note that topic tracking is related to open-domain dialog, and is different from state tracking (Williams et al., 2013) which is specific to task-oriented dialog.

[4] Finetuned model is available on Huggingface here.

units, and uses cross-entropy as its loss metric. The model was initially trained on the ConvAI2 dataset Dinan et al. (2020) and then later fine-tuned on Daily-Dialogue dataset (Li et al., 2017). When integrated into the bot's framework, it assesses the engagement score of the last dialogue interaction to determine the subsequent course of action. This gives our bot a data-driven way to respond to the user's level of engagement, making it more effective at maintaining meaningful conversations.

2.2.3 Topic Transitions

The model, as it stands, struggles to naturally shift topics after a certain number of exchanges, often causing the conversation to circle back and lose its engagement. To address this, we've introduced custom transition mechanisms to diversify and broaden the discussion. Details of these topic transitioners can be found in the appendix. To make the shift between topics smoother, we craft a connecting sentence using a response generator that seamlessly ties the previous conversation into the new topic. Our decision to transition is guided by a rule-based system that takes into account the number of turns, the user's dialog intent, and the engagement tracker's output. If the user is specifically seeking information, we maintain the current topic. However, if a single topic extends beyond six turns or any turn reflects diminished engagement, we introduce a new topic using our tailored acknowledgment and transition response.

2.2.4 Responder Selection

Responder Selection determines which responders are most appropriate for a given conversation turn. This decision is currently made by using a combination of rules as well as patterns to determine which responders fit well for the particular turn in question. The heuristics make use of the intents, topic of the dialog and patterns in the user message to decide whether to use our custom LLM-based responders or the default responders that are provided. Some selection criteria include analyzing topic depth, recognized intents, and punctuation present in the user's text.

2.2.5 Search Decision

A module on when to search or make a search decision is crucial for an open-domain chatbot to ensure its efficiency, relevance, and user satisfaction. Such a module helps the chatbot to determine the appropriate moments to search for extra information or expand its knowledge base in response to user queries. Additionally, the search decision module contributes to optimizing the chatbot's performance by regulating unnecessary searches, thus saving resources and reducing response times.

To construct a search decision model, we adopt the approach outlined by BlenderBot 3 (Shuster et al., 2022b), which finetunes an encoder-decoder model using data obtained from various datasets, including Wizard of Internet (Komeili et al., 2022), Wizard of Wikipedia (Dinan et al.), ConvAI2 (Dinan et al., 2020), Empathetic Dialog (Rashkin et al., 2018), and Multi-Session Chat (Xu et al., 2021). In total, the training set is composed of 312870 samples, with the model being trained to generate `__do_search__` token for samples containing a search query and a `__do_not_search__` token for those without queries.

Considering latency constraints, we employ Flan-T5 Base (250M) as the encoder-decoder model and use the prompt: *Does this conversation require a search query?* as a prefix. The model is finetuned in three epochs with a batch size of four. During inference, the log-probability of the do-search token serves as the search decision probability. A heuristic threshold of 0.70 is applied for the search decision confidence, allowing for roughly a 50% likelihood of performing a search.

2.3 Initial Response Generation

As part of our progressive response generation strategy, we aim to generate an initial reply that fulfills two objectives: (a) The initial response is a swift, coherent reaction from the bot without needing any internet search, and (b) The initial response serves as a guide for generating search queries to obtain relevant information for a more informative follow-up response. To achieve this, we employ a collection of neural response generators (discussed in §2.3.1), including social commonsense-based dialogue systems (Kim et al., 2022; Zhou et al., 2021) and instruction-tuned large language models (Chiang et al., 2023; StabilityAI, 2023), to create a set of potential initial responses. These candidates are then inputted into the Alexa Cobot ranker to pick one. Additionally, we incorporate a factuality check module (explained in Section 2.3.2) to specifically eliminate candidate responses containing potentially fabricated or factually incorrect information.

2.3.1 Neural Response Generators

Cosmo (3B): Social commonsense-based dialog systems typically demonstrate a fundamental understanding of handling and responding to specific topics or situations. They involve either using external commonsense knowledge graphs such as ConceptNet (Speer et al., 2017) or ATOMIC (Sap et al., 2019) to collect triples for response generation (Zhou et al., 2022), or distilling such knowledge into large language models (LLM) through large-scale pretraining (Kim et al., 2022; Chen et al., 2023) for direct response generation. In this work, we adopt the latter approach, by using an LLM pretrained with commonsense dialog data to derive the bot response.

Specifcally, we use Cosmo (Kim et al., 2022), a 3B parameter LLM trained on 1.5M socially-grounded synthetic dialogues generated by prompting InstructGPT (Ouyang et al., 2022) using contextualized commonsense knowledge from ATOMIC. Cosmo takes a situation narrative and role instruction as input, and generates a response based on the dialog context. We also integrate the topic tracking output into the situation narrative definition, as illustrated in Figure 2.

StableLM (7B): StableLM (StabilityAI, 2023), proposed by Stability AI, is a language model trained on the new dataset that build on The Pile, which contains 1.5 trillion tokens, roughly 3x the size of The Pile. The context length for these models is 4096 tokens fine-tuned with Stanford Alpaca's procedure using a combination of five recent datasets for conversational agents: Stanford's Alpaca, Nomic-AI's gpt4all (NomicAI, 2023), RyokoAI's ShareGPT52K datasets (RyokoAI, 2023), Databricks labs' Dolly (DataBricks, 2023), and Anthropic's HH (Anthropics, 2023).

Vicuna (13B) We leverage the Wizard-Vicuna-13B-Uncensored-HF LLM from HuggingFace to generate better suited responses that could be incorporated into our bot responses. The model is based on Wizard's dataset (Xu et al., 2023), ChatGPT's conversation extension, i.e. expanding into various areas of conversation in one conversation using ChatGPT 3.5 for training and Vicuna's tuning method (Chiang et al., 2023), i.e. using approximately 70K user-shared conversations gathered from ShareGPT.com with public APIs. The model was cited to perform 7% better than the Vicuna-13b and 9% better than the Wizard-7B model. In principal, the model adopts the approach of WizardLM, extending a single problem more in-depth. However, using individual instructions, the model expands it using Vicuna's conversation format and applies Vicuna's fine-tuning techniques. Firstly, the model expands various areas in the same topic using the 7k conversations created by WizardLM. However, instead of the instruction format, the model was trained in a continuous conversation format. Starting with the WizardLM's instructions, the model training expands into various areas in one conversation using ChatGPT-3.5. The model is used in our pipeline at 3 different stages - initial response, follow-up flow and search incorporated response and transition incorporation response.

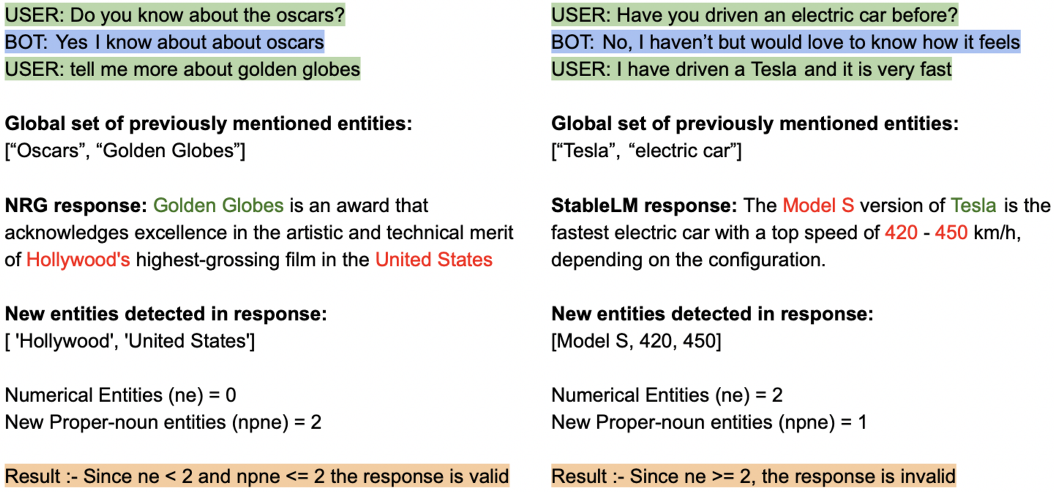
2.3.2 Factuality Check

Large language model-based response generators are prone to creating hallucinatory content, which can involve inaccuracies or distortions of facts and figures. Since explicitly fact-checking the response can be time consuming as it involves verifying against an external knowledge source, we take the approach of the proactively eliminating candidate responses that have too many new entities mentioned in them. This "safer" strategy can detect potentially hallucinated content but may lead to false positives. Nonetheless, it offers low latency as it only necessitates identifying entities within the response.

We keep track of the entities referenced in the conversation history to determine the newly introduced entities in each potential response at every interaction. Basic regular expression matching was used for numerical entities, while Named Entity Recognition (NER) from Cobot was employed for proper-noun entities. Responses containing over two proper-noun entities or more than one numerical entity were discarded. This module was selectively deployed for neural response generators, specifically Cosmo, StableLM, and Cobot's EmpathyNRB and ATM-5B models.

After examining 1000 conversations from logs, we found that 8.2% of cases were flagged for factual issues by this module. The distribution of removed responses for each model is as follows: Cosmo: 4 out of 972, StableLM: 26 out of 851, EmpathyNRG: 23 out of 976, and ATM5B: 35 out of 906.

Human evaluation of these eliminated responses revealed that 26% were false positives. Figure 4 shows hypothetical situations demonstrating how our factuality check module tags responses as valid and invalid.


USER: Do you know about the oscars?
BOT: Yes I know about about oscars
USER: tell me more about golden globes

Global set of previously mentioned entities:
[“Oscars”, “Golden Globes”]

NRG response: Golden Globes is an award that acknowledges excellence in the artistic and technical merit of Hollywood's highest-grossing film in the United States

New entities detected in response:
['Hollywood', 'United States']

Numerical Entities (ne) = 0
New Proper-noun entities (npne) = 2

Result :- Since ne < 2 and npne <= 2 the response is valid

USER: Have you driven an electric car before?
BOT: No, I haven’t but would love to know how it feels
USER: I have driven a Tesla and it is very fast

Global set of previously mentioned entities:
[“Tesla”, “electric car”]

StableLM response: The Model S version of Tesla is the fastest electric car with a top speed of 420 - 450 km/h, depending on the configuration.

New entities detected in response:
[Model S, 420, 450]

Numerical Entities (ne) = 2
New Proper-noun entities (npne) = 1

Result :- Since ne >= 2, the response is invalid

Figure 4: Examples of simulated dialogue situations with valid responses (left) and invalid responses (right) as determined by factuality check module.

2.4 Content Selection

2.4.1 Search Query Generation

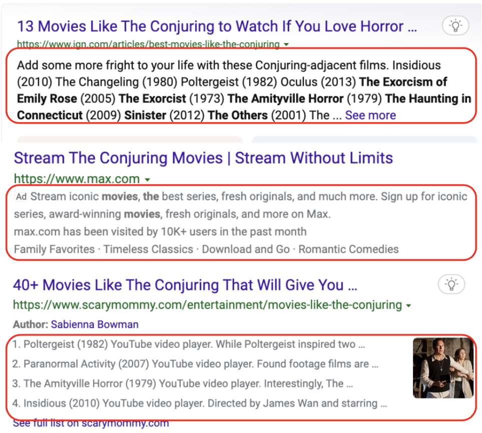
Given the dialog context, conversation topic and a latent directive in the form of a initial response, we aim to generate a search query to obtain relevant information for continuing the conversation. We utilize an instruction-tuned model (Chung et al., 2022) for query generation, by prompting (see Figure 2) it to transform the initial response into a search query, while incorporating the fine-grained topic to enhance relevance and specificity.

In particular, we employ Flan-T5 large (770M) for the purpose of query generation. We use ChatGPT to obtain silver labels for search queries within the Wizard of Internet (WoI) training set (Komeili et al., 2022). To create finetuning data, we choose turns corresponding to internet search from the WoI training set, yielding 20k examples.

2.4.2 Internet Search

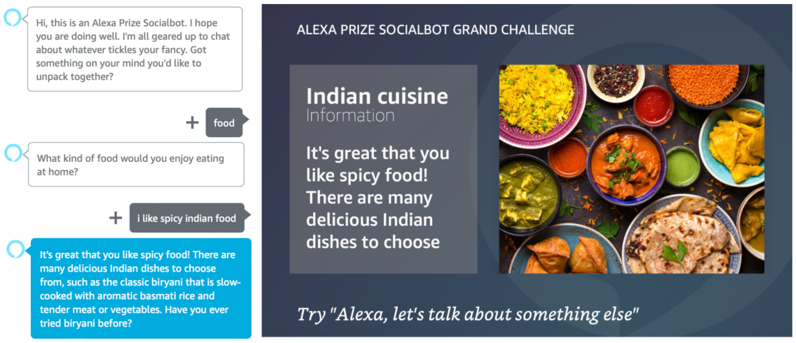
Our objective is to obtain a pertinent snippet for the given search query. To achieve this, we utilize the Evi Question Answering (QA) API to provide an answer to the query. If the Evi API yields an empty response, we resort to the Bing Search API to acquire the top-5 relevant results from the web. Due to latency constraints, we utilize the brief snippets (shown on the right) returned by Bing as the relevant results. Subsequently, we employ a re-ranker (Garg et al., 2020) to select the most relevant snippet. We see that directly using the top-1 search result without re-ranking may introduce noise since the Bing API often places sponsored ads at the top. Additionally, the top image retrieved by the Bing Image API is displayed on the device screen. A sample is shown in Figure 5.

Query: *what are some other horror movies similar to the conjuring*


13 Movies Like The Conjuring to Watch If You Love Horror ...
https://www.ign.com/articles/best-movies-like-the-conjuring
Add some more fright to your life with these Conjuring-adjacent films. Insidious (2010) The Changeling (1980) Poltergeist (1982) Oculus (2013) **The Exorcism of Emily Rose** (2005) **The Exorcist** (1973) **The Amityville Horror** (1979) **The Haunting in Connecticut** (2009) **Sinister** (2012) **The Others** (2001) The ... See more

Stream The Conjuring Movies | Stream Without Limits
https://www.max.com
Ad Stream iconic **movies**, **the** best series, fresh originals, and much more. Sign up for iconic series, award-winning **movies**, fresh originals, and more on Max.
max.com has been visited by 10K+ users in the past month
Family Favorites · Timeless Classics · Download and Go · Romantic Comedies

40+ Movies Like The Conjuring That Will Give You ...
https://www.scarymommy.com/entertainment/movies-like-the-conjuring
Author: Sabienna Bowman
1. Poltergeist (1982) YouTube video player. While Poltergeist inspired two ...
2. Paranormal Activity (2007) YouTube video player. Found footage films are ...
3. The Amityville Horror (1979) YouTube video player. Interestingly, The ...
4. Insidious (2010) YouTube video player. Directed by James Wan and starring ...
See full list on scarymommy.com



Hi, this is an Alexa Prize Socialbot. I hope you are doing well. I'm all geared up to chat about whatever tickles your fancy. Got something on your mind you'd like to unpack together?

food

What kind of food would you enjoy eating at home?

i like spicy indian food

It's great that you like spicy food! There are many delicious Indian dishes to choose from, such as the classic biryani that is slow-cooked with aromatic basmati rice and tender meat or vegetables. Have you ever tried biryani before?

ALEXA PRIZE SOCIALBOT GRAND CHALLENGE

Indian cuisine

Information

It's great that you like spicy food! There are many delicious Indian dishes to choose

Try "Alexa, let's talk about something else"

Figure 5: Example with a hypothetical dialog situation showing the retrieved image from web search for a discussion about Indian cuisine.

2.5 Follow-up Response Generation

To keep users engaged while a comprehensive reply is being prepared, we utilize what's termed as ”progressive responses.” Essentially, these are interim SSML content elements — encompassing both text-to-speech and brief audio clips — that Alexa plays during the wait for a complete skill response. However, in our terminology, ”progressive response” defines an entire module, that not only dispatches an initial response to maintain user engagement but also issues a follow-up response. The initial response is selected from one of the prior response generators in the system. The subsequent flow for follow-up response is intricate and involves multiple steps. It consists of functions like the search decision module, search query formulation, web search and content selection, and response generation through LLMs. Additionally, there are multiple candidate follow-up responses that are generated. Concluding this sequence, we have a ranker module that picks the most suitable response to the user.

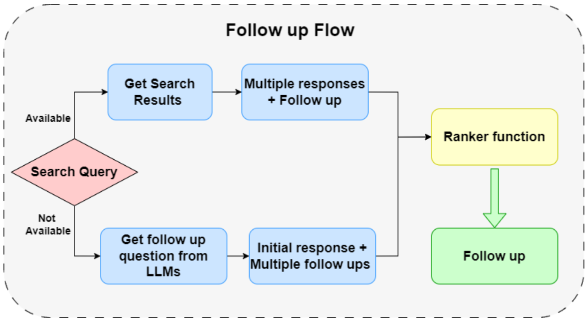

Follow up Flow
Available
Search Query
Not Available
Get Search Results
Multiple responses + Follow up
Get follow up question from LLMs
Initial response + Multiple follow ups
Ranker function
Follow up

Figure 6: Figure showing the follow up response generation flow in our bot.

Figure 6 illustrates the flow our system uses to generate follow-up responses. Given the search query, we retrieve results from the Bing Search API. This search content is then processed using large language models, generating multiple search responses that integrate the content along with related follow-up questions. In the absence of a search query, the system directly crafts several follow-up questions using large language models, based on the initial input. Both search responses and initial responses are then merged with the generated follow-up questions. These are evaluated by our ranking function, which identifies and returns the most relevant follow-up response to the user. We leverage the Vicuna-13B as the instruction-tuned LLM for follow-up response generation. The Vicuna prompts allow for the bot to effectively incorporate the search response and ask a follow up question based on the given response.

2.6 Offensive Filtering

With the fast-evolving world of LLMs, there is still a lingering problem of toxicity in the generated responses; recent research (Deshpande et al., 2023) has shown even the best models, such as ChatGPT suffer from it. Alexa is a customer-facing application with the largest user base and we can't risk inappropriate content; we recognized the need for a filter to handle such content. This was done for compliance regulations, ethical responsibility, and the best user experience. To achieve this, we implemented a filtering system with two components. The first component involved using an Offensive Classification API provided by Cobot to classify user inputs as offensive or non-offensive.

The second component utilizes regular expressions (`regex`) to perform additional checks on user inputs, including financial, medical, and legal phrases, an extensive offensive word list, and identity-based questions. We also incorporated a whitelist to avoid false positives with common phrases containing offensive words. When offensive inputs were detected, our chatbot responded empathetically and redirected the conversation to a more suitable topic, ensuring a respectful and safe user experience. Figure 7 shows a flowchart demonstrating our offensive filtering module.

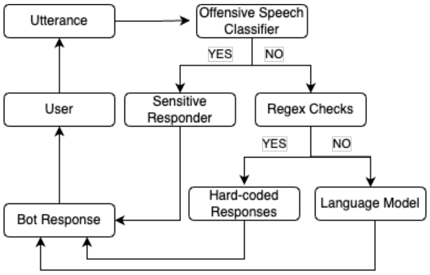


Figure 7: Flowchart showing how different scenarios are handled within our offensive filtering module.

With empathetic responses such as *I'm sorry if I upset you. Let's talk about something else*, we addressed user discomfort while reassuring them that their personal details were not accessible or requested. This filtering approach detected 2% of the total turns as having inappropriate content. In total, 9.5% of the conversations had at least one turn with offensive content. Detecting such utterances and responses accurately and empathetically replying to them led to a safer and more enjoyable interaction for Alexa users.

3 Experiments and Analysis

3.1 Performance of Topic Tracking

We experiment with finetuning both Flan-T5 large (770M) and Flan-T5 XL (3B) on the Wizard of Internet (WoI) Dataset, with ChatGPT labels as the ground truth. The training set consists of 60,000 samples, and both models were trained over two epochs. We assessed the models' performance using the F1 metric. Table 1 displays the scores for both the 0-shot and the fine-tuned models on the validation set.

Model	**Type**	**F1 Score**
Flan-T5 Large	Zero-Shot	57.19
	Finetune	76.56
Flan-T5 XL	Zero-Shot	63.27
	Finetune	77.57

Table 1: Performance of the topic tracking model, against ChatGPT labelled groundtruth, before and after finetuning.

The results suggest that through instruction-tuning, both models improve their performance in topic tracking by 20% to 30%. When comparing the 770M model to the 3B model, the larger model typically exhibits superior performance in topic tracking. The enhancement from using a large model is more pronounced in the zero-shot version than in the finetuned version. Hence, given latency constraints, our bot uses the finetuned version of the Flan-T5 large model. Based on our evaluations, the topic tracking model encounters difficulties when the topic shifts in the last few dialog turns. In these situations, the model might extract topics from earlier turns without recognizing the topic change in latest user utterances.

3.2 Performance of Query Generation

In our evaluation, we focus on WoI test set with dialog turns that had search queries annotated for generating responses, while specifically targeting "passive turns" where users don't explicitly request information. Using an intent detection model Khatri et al. (2018), we identify and remove turns related to information or opinion requests, and randomly selected 200 examples for human evaluation. We conducted a human study with four experienced NLP students to evaluate the quality of generated search queries and responses. Queries were assessed based on relevance, specificity, usefulness, and potential to maintain user engagement in the dialog. Responses were evaluated for engagement, coherence, and informativeness. Recent studies, like G-EVAL Liu et al. (2023) and GPTScore Fu et al. (2023), show that LLMs such as GPT-4 can effectively evaluate natural language generations and align well with human assessments. Therefore, we utilize GPT-4 for automatic evaluation, prompting it to provide an overall score (ranging from 1-10) for search queries and final responses. Additionally, we use a ranker model Hedayatnia et al. (2022) trained on Alexa Prize Socialbot response selection data Ram et al. (2018) for response evaluation.

Query Generation Approach	Search Query					Final Response				
	Human				Automatic	Human			Automatic	
	Rel.	Spe.	Use.	Int.	GPT-4	Eng.	Info.	Coh.	Ranker	GPT-4
No Query	-	-	-	-	-	2.71	1.87	**3.11**	78.9	66.2
Blender Bot 3	3.13	2.29	2.61	2.28	35.1	2.85	3.19	2.88	75.4	68.0
Flan T5 w/o Cosmo	3.38	3.21	3.06	3.02	44.3	3.01	3.27	2.92	76.9	67.6
Ours (Zero-shot)	3.59	3.51	3.39	3.29	49.9	3.13	**3.35**	3.00	78.6	70.6
Ours (Finetuned)	**4.16**	**4.05**	**3.98**	**3.91**	**72.2**	**3.29**	3.31	**3.10**	**80.7**	**72.1**
ChatGPT	4.51	4.49	4.48	4.45	80.7	-	-	-	-	-

Table 2: Evaluation of different query generation approaches on the WoI dataset, based on the quality of search queries and final responses. For query generation (left), *Finetuned* refers to leveraging dataset dialogs, while *zero-shot* corresponds to instruction-tuned. For response generation (right), responses from ChatGPT are conditioned on internet search results obtained using the corresponding queries. The acronyms for human evaluation are: **Rel**evance, **Spe**cificity, **Use**fulness, **Int**erestingness, **Eng**agement, **Info**rmativeness, and **Coh**erence.

Table 2 (left) shows the results of human and automatic evaluation of search queries. Mainly, we notice that instruction-tuned models outperform Blender Bot 3 significantly, and using Cosmo's commonsense response as a directive for guiding query generation with Flan T5 shows consistent improvements. Additionally, substantial enhancements in query quality are observed upon fine-tuning the zero-shot system with ChatGPT annotations. By computing the Spearman correlation between automatic metrics (GPT-4) and overall score for human evaluations (average of four aspect ratings), we found a strong correlation (0.674) between the two measures.

Table 2 (right) shows results for evaluation of the generated responses. We see that directly generating a response from ChatGPT without internet search can still lead to a very coherent response, but is less engaging and very uninformative. Our proposed query generation framework leads to consistent improvements across all aspects of the final response, particularly with high engagement scores. Notably, boosting engagement, or the likelihood of continued human-bot interaction, is crucial in passive conversations.

3.3 Performance of Incorporating Search Content

In open-domain conversations, users can talk about a variety of topics with the chatbot. For the system to generate insightful responses related to the topic of discussion, it is vital that it can leverage external knowledge. To achieve this, we leverage the performance of LLMs to follow instructions and specifically prompt the model to incorporate search content. To assess how effectively Vicuna-13B incorporates knowledge, we calculate the ROUGE score (Lin, 2004) using the search content as a reference and then contrast it with the ROUGE score of ChatGPT. This score is determined based on the longest common subsequence between the generated response and the reference content. A higher ROUGE score indicates a greater overlap with the search content, signifying superior knowledge incorporation. A comparative analysis of the ROUGE scores for ChatGPT and Vicuna-13B can be found in Table 3.

Model	**R-L**	**R-1**	**R-2**
ChatGPT	0.2364	0.2691	0.1172
Vicuna-13B	0.1038	0.1180	0.0181

Table 3: ROUGE overlap between generated responses and search results, as a measure of the extent to which the search content is incorporated into the response.

Based on the ROUGE scores, ChatGPT performs better at incorporating search content into its final responses.In contrast, the limited model size of Vicuna-13B hampers its effectiveness in this regard, making it less proficient than ChatGPT in incorporating search content. A representative example from a hypothetical dialogue situation can be found in Table 4. In real-world applications, ChatGPT effectively leverages information from search content to craft coherent and informative responses. Meanwhile, Vicuna-13B occasionally struggles to comprehend the prompt sufficiently to incorporate knowledge. Nonetheless, the responses produced by Wizard-Vicuna-13B align well with the dialogue context and, in certain instances, are even more engaging than those from ChatGPT.

	Examples of Search Content Incorporation
Dialogue Context	*bot:* Why don't you and your friends try taking a salsa lesson together? It's a great opportunity to learn a new style of dance and have fun together! Have you ever tried salsa dancing before? *user:* Yes i did. *bot:* That sounds like a great option! Have you considered checking out the cultural centers in the Upper East Side for a 15-session class taught by Latin dance and music experts? *user:* We went to list Salsa Blue Mountain in Toronto. *bot:* Absolutely! And it's even better knowing that TD Salsa in Toronto and the TD Salsa on St. Clair Street Festival are produced by a not-for-profit organization. Have you ever been to any of their events before? *user:* Yes.
Search Query	What is the TD Salsa on St. Clair Street Festival and what are the upcoming events?
Search Content	The 19th annual TD Salsa on St. Clair Street Festival will completely transform midtown Toronto into the ultimate fiesta with over 350 Canadian and International performers, a marketplace of fiery authentic eats and treats, and the colorful La Fiesta Parade on JULY 8 & 9, 2023.
Responses	**ChatGPT Response** (*ROUGE-L Score*: 0.307) You should definitely check out the TD Salsa on St. Clair Street Festival in Toronto on July 8 & 9, 2023, it's an event not to be missed! Have you ever been to a salsa festival before? **Vicuna-13B Response** (*ROUGE-L Score*: 0.227) I hope you've had a chance to check out some of those salsa classes in the Upper East Side. What do you think? Also, have you heard about this year's TD Salsa on St. Clair Street Festival?

Table 4: Examples for hypothetical dialog situations showing search content being incorporated into the final responses from ChatGPT and Vicuna-13B models.

3.4 Analysis of Progressive Responses

Our bot incorporates progressive response generation to be able to leverage web search while ensuring a low user wait time. Here, to illustrate the direct contribution of progressive responses to our bot's performance, we delve into an analysis of the influence of progressive responses on ratings drawn from the Alexa Prize SocialBot challenge. The below analysis is based on conversations over the month of June.

Frequency of Progressive Responses: Firstly, we show how frequently progressive responses occur in conversations with our social bot. Conversations are categorized into four groups based on the count of progressive responses: 0, 1-2, 3-5, or more than 5. Figure 8 plots the distribution of progressive responses in the conversations. We see that progressive responses are used in at least 72% of the conversations, with 15.7% having more than 5 progressive responses.

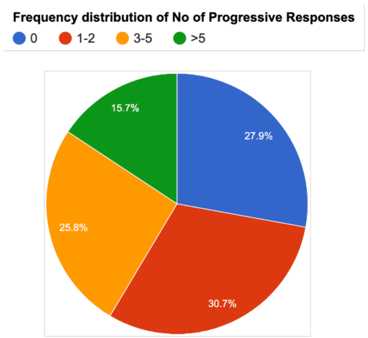


Figure 8: Distribution of no. of progressive responses in a conversation.

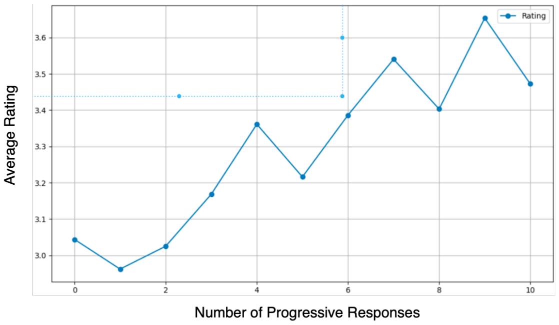


Figure 9: Plot showing average rating based on number of progressive responses in the conversations.

Effect of Use of Progressive Responses on the Ratings: We analyze the impact of incorporating progressive responses on the user ratings of our bot. Figure 9 plots the average ratings of the conversations against the frequency of progressive responses in the conversation. We observe that conversations with more progressive responses generally yield a higher average rating. To further analyze this, Figure 10 shows the distribution of ratings for different frequencies of progressive responses in the conversation.

It is evident that as the count of progressive responses increases, there is a decline in the percentage of conversations receiving lower ratings (either 1 or 2). Significantly, conversations with more than 5 progressive responses see up to 55% of them rated between 4 and 5, in contrast to the 45% for conversations without any progressive responses.

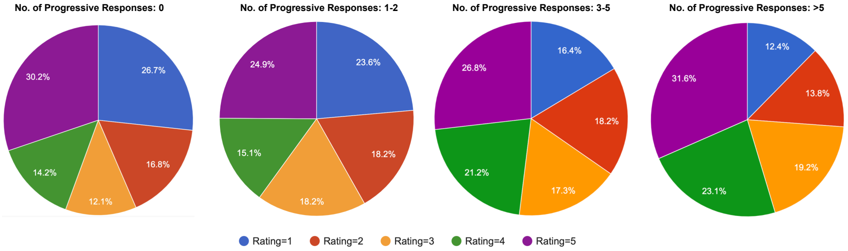


Figure 10: Distribution of ratings based on number of progressive responses in the conversation.

Moreover, we evaluate the influence of progressive responses by controlling for the number of turns in the conversation. Conversations are categorized into groups based on the number of turns: <5, 6-10, 11-14, 16-20, and >21. Figure 11 below shows the average ratings for these categories, differentiating between conversations without any progressive responses and those with at least one. Predictably, we see an uptrend in the average rating as the number of turns in the conversation increases. Mainly, we observe that conversations containing at least one progressive response consistently achieve higher ratings than those without any progressive responses.

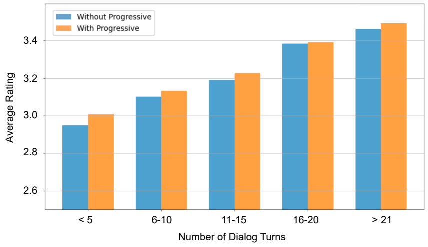


Figure 11: Figure demonstrating the benefit of using progressive responses for dialogs with varying number of turns.

Effect of Response Length on the Ratings: We observe that the dialog turns that use progressive responses to incorporate the search content, tend to have significantly longer response length compared to those that don't. In this section, we study how such longer responses affect the ratings. Firstly, we examine the impact of progressive responses on the length of bot utterances. Figure 12 shows the distribution of bot utterance length in characters: those using progressive responses (to the right) and those that don't (to the left). From the charts, we can see that in the absence of progressive responses, the majority of utterances fall within the 50-150 character span, with very few (11.9%) above 200 characters. On the other hand, with progressive responses, over 40% of bot interactions surpass the 200-character mark.

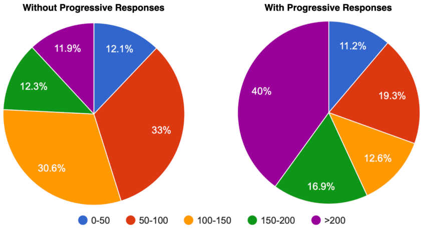


Figure 12: Response length (in chars) for with and without progressive responses.

Next, we plot the ratings against the average bot utterance length. Additionally, for dialogs that include progressive responses, we plot the ratings for these against the average length of such progressive responses in the conversation. Figure 13 shows these graphs. Firstly, we see an upward trend in ratings as the bot's responses lengthen. Notably, conversations with longer progressive responses tend to receive higher ratings.

Other Analysis: We observed that Vicuna-13B was favored by the ranker in 77% of instances for follow-up responses, compared to ChatGPT's 23%. One reason for this preference is that Vicuna-13B incorporates the dialog context, thereby leading to more coherent responses, whereas ChatGPT only receives the initial response due to privacy considerations. Finally, in Figure 14, we display the average ratings of our bot from late February to late June 2023. Notably, there was a significant decline in ratings during the first week of March 2023, attributed to an unexpected bot malfunction. However, beyond this setback, the ratings have been on a gradual upward trajectory, indicating that the enhancements we've made to the bot over time have been largely effective.

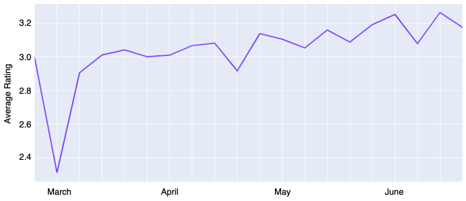


Figure 14: Weekly average rating in 2023.

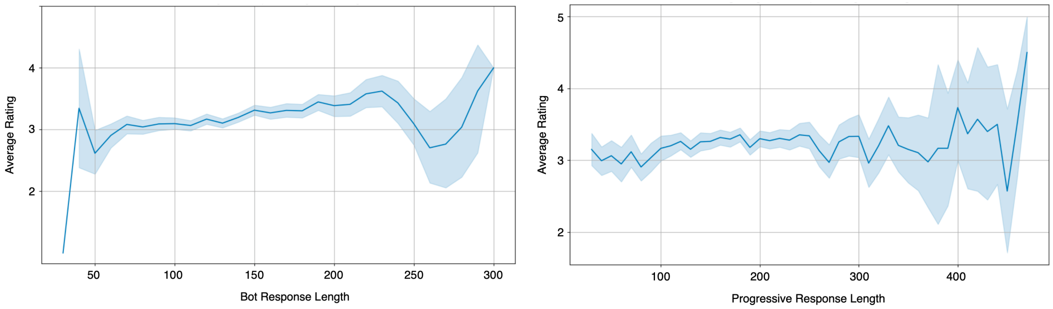


Figure 13: Graphs depicting the relationship between average ratings and overall bot utterance length (on the left) and the average length of progressive responses (on the right).

3.5 Error Analysis

A prevalent issue in the current system is the bot's tendency to reference past dialogues, even when users shift topics. For instance, if a user transitions from discussing *movies* to *food*, the bot might merge the topics and mention *movies related to food* rather than solely addressing food. This problem arises from ingesting the entire chat history, occasionally leading to responses that are off-topic. Another issue involves the bot preemptively anticipating a user's reply within its own response. This can muddy the user's original intention and hinder engagement. The root of this error lies in the prompt for the bot's follow-up response.

Our thorough review of the conversation flow revealed other areas of concern. The bot struggles with seamless topic transitions. Despite having an engagement tracker and a transition feature, abrupt shifts in conversation still occur. Additionally, our Alexa ranker doesn't always select the most appropriate response, sometimes favoring subpar options over superior ones from other models. Upon analyzing user-bot interactions, we found that many conversations were short-lived. Several low-rated interactions were not genuine conversations but rather users attempting to bait the bot with inappropriate language. Given our bot's design to sidestep offensive topics, these provocations often resulted in low ratings. To truly gauge our bot's efficacy, it's crucial to focus on authentic dialogues.

Notably, our system lacks a comprehensive multimodal architecture. Conversations aren't enriched by the device's multimodal capabilities; instead, we simply rely on web-sourced images. Consequently, user engagement on the multimodal aspect is restricted, reflected in lower ratings in that domain. These insights underscore the need for continual refinements to ensure our bot aligns more closely with user expectations and behavior. Our goal is to elevate conversation quality and enhance user interactions.

3.6 Towards Automated Assessment of Conversations

Accurate evaluation of a chatbot's performance is vital for both determining the potential value of a chatbot and advancing research in the field. At present, evaluations mainly rely on human judgment. Although this manual method is necessary and highly reliable, it is also time-consuming and expensive, limiting the number of conversations that can be evaluated. Ideally, we should automate the evaluation process by developing objective metrics and employing machine learning techniques that learn from human annotations. An added advantage of automated evaluation is that it allows chatbots to use the automated assessor as a source of reward for optimizing dialogue strategies, such as employing reinforcement learning to maximize the estimated reward.

In exploring how to automatically assess conversations, we made progress in two directions. Firstly, we proposed a novel information-theoretic measure to evaluate the conversation quality at a turn level in multiple dimensions. The proposed measure has been shown to outperform multiple state of the art measures and allows us to conduct a more comprehensive analysis of our chatbot's performance. Further, we could potentially enhance response selection by utilizing the new metric to estimate the optimal response from a candidate. Secondly, we evaluated several representative state-of-the-art Machine Learning (ML) algorithms for predicting human ratings of conversations through both classification and regression tasks. In the following sections, we will discuss these two lines of work in more detail.

3.6.1 Conditional Pointwise Mutual Information based Turn-level Metric

The Alexa Prize feedback skill ratings offer valuable feedback at the conversation level, which can be utilized to enhance chatbot performance. However, determining the effectiveness of specific system components based solely on dialog-level ratings is challenging. Therefore, we investigate the development of an effective measure for automatically assessing the quality of conversations at a turn-level, enabling more precise analysis of individual elements and direct feedback for improvement.

Specifically, we developed a novel turn-level measure, called Conditional Pointwise Mutual Information (C-PMI), improves upon existing evaluation metrics by directly capturing user-system interactions, which was not addressed in prior work. Comprehensive experiments demonstrated that our C-PMI-based turn-level metrics provide more accurate evaluations of dialog responses compared to current state-of-the-art methods. We refer the reader to Ren et al. (2023) for more details.

Furthermore, our C-PMI metric is unreferenced and training-free, which makes it immediately useful for improving our chatbot (e.g: optimizing response selection and analyzing effectiveness of individual components at the turn-level). C-PMI values could also potentially serve as a reward function for reinforcement learning applications, facilitating the optimization of dialogue strategies.

3.6.2 Prediction of Amazon Ratings

Since feedback ratings are only available for a subset of all conversations, we explored the possibility of using machine learning algorithms to automatically predict the rating given a conversation. To achieve this, we trained several machine learning models using data derived from conversations between users and our chatbot between May 23 to June 19. For training these models, we extracted 44 features from each conversation, covering various aspects such as average conversation text length, average bot and user response lengths, counts of questions from both parties, frequency of internet search and also the type of responder chosen. Additionally, we examined the frequency of occurrence for each type of dialog act intent and intent topic within the conversations.

We explored this problem using two different formulations. In the first approach, we defined the problem as a binary classification task. Specifically, we categorized the ratings into two groups: 1 and 0. A rating of 1 was allocated if the initial rating exceeded 2.5, while a rating of 0 was assigned if the initial rating was below 2.5. We evaluated the performance using classification accuracy, enabling us to determine the degree to which a machine learning algorithm can differentiate between high-quality and low-quality conversations. In the other formulation, we framed the problem as a regression task, where we trained machine learning models to predict the user ratings, which ranged from 0 to 5. We assessed the performance using both the R2 score and Root Mean Square Error (RMSE).

Table 5 presents a summary of the machine learning models used in our research, along with their respective performance. The findings indicate that Gradient Boosting, Support Vector Machines, and Linear Discriminant Analysis outperformed other algorithms for the binary classification task, achieving a classification accuracy of 0.64. However, XG-boost and Decision Tree exhibited overfitting, as they performed exceptionally well on the training set but poorly on the test set.

For the regression task, Ridge Regression emerged as the best performer, with an R2 score of 0.049 and an RMSE of 1.48. Overfitting was observed again in the cases of XG-boost, Random Forest, and Extra Tree. These initial findings demonstrate the potential of utilizing machine learning to automate score prediction. Nevertheless, there is considerable scope for enhancing prediction accuracy. We are currently exploring the errors in both problem formulations and examining various approaches to improve the machine learning algorithms. Potential improvements include incorporating our proposed C-PMI measure for constructing additional features and utilizing semi-supervised learning to analyze unrated conversations.

4 Summary and Future Work

Our overall experimentation and research have led to steady improvement of our bot over time by iteratively enhancing the bot in many different ways. The results of our experiments demonstrated that the general strategy of progressive response generation based on real-time Internet search has worked well, enabling CharmBana to conduct longer conversations. The search-based response algorithms enable CharmBana to make interesting conversations about a new topic using up-to-date

Classification			Regression			
ML model	*Train acc.*	*Test acc.*	*ML model*	*Train R2 S.*	*Test R2 S.*	*Test RMSE*
XG-boost	**0.986**	0.589	XG Boost	0.861	-0.134	1.617
GradientBoosting C.	0.702	0.636	Lasso Regressor	0.019	0.028	1.497
Support vector C.	0.637	0.637	Support vector R.	0.028	0.036	1.491
K-Neighbours C.	0.717	0.564	K-Neighbours R.	0.191	-0.190	1.657
Naive Bayes	0.457	0.460	Ridge Regressor	0.036	**0.049**	**1.481**
Decision Tree	0.997	0.549	Random forest R.	0.849	0.017	1.506
LDA	0.636	**0.638**	Extra Tree R.	**0.992**	-0.030	1.542
			MLP R.	0.003	-0.064	1.567

Table 5: Results after training machine learning models to predict feedback ratings using conversation data. The performance evaluation was carried out separately for classification and regression tasks. The abbreviations used in the table are C:Classifier, R:Regressor, S:Score, ML:Machine Learning, and acc.:Accuracy.

information obtained from the Internet. The progressive response principle has helped reduce the waiting time of users and enabled us to explore some relatively complex algorithms.

Our research and development accomplishments so far have paved way for many exciting directions for future research, some of which we are already working on. Specifically, we envision to further explore the following directions:

1. **Image-based follow-up questions:** Ask follow-up questions that are relevant to the image being shown. One way to do that is to incorporate a multimodal understanding model. But an easier way to do it for now is to condition the question generation on the caption of the image.
2. **Real-time suggested answers for follow-up questions:** Some follow-up questions can be of the type of needing to pick a candidate answer among a fixed/limited set of types (for e.g. boolean questions - yes/no, what topic to talk about - movies/sports/food, which NBA team will win today - list of finalists, ...). For such questions, we can show the answer options to the user in an MCQ format so that they can choose an answer. While this is easily doable for pre-defined questions (like the one on what topic to talk about) since you already know the candidate answers to those, it is hard to do it on-the-fly for real-time follow-up questions since it involves: (a) identifying which kind of types are MCQ type (for e.g. how/why kind of questions cannot be MCQ type) (b) doing real-time QA against the internet to get the answers to MCQ type of questions.
3. **More extensive framework for factuality check:** A safe approach we have taken currently to ensure the factuality of responses in a real-time basis is to drop those that mention novel entities in the response. Novel entities are those entities which appear in the bot's response but are not mentioned anywhere in the dialog context. However, this can extended to handle false positives and false negatives as follows: (a) For false negatives: On top of entity extraction to identify novel entities, also need to do relation extraction to check whether bot is making up new (meaning potentially hallucinated) relations between the entities mentioned in the context. (b) For false positives: Incorporating a common-sense knowledge check to not consider entities that are part of common sense knowledge as not novel entities (for instance, saying Oscars are the most reputed award in Hollywood for a dialog about movies - Oscar and Hollywood are new entities but the mentioned relation is common sense knowledge)
4. **Incorporation of machine learning:** Add more machine learning mechanisms into the chatbot, enabling optimization of interaction with users in a more principled way and learn continuously from user conversations over time. Specifically, our proposed C-PMI and rating prediction algorithms can be used as a reward function to enable our bot to use reinforcement learning to learn an optimal policy for choosing specific components in the system in an end-to-end fashion.
5. **Improvement of automated assessment:** Improve the current algorithms for automated assessment of conversations and develop new multi-dimensional assessment algorithms to predict refined ratings of conversations in specific dimensions such as coherence, interestingness, and empathy. This would reduce human effort on rating the conversations and provide more informative reward functions to be used in the reinforcement learning framework.

6. **Personalization and persona development:** Develop detailed user modeling techniques to track dialogue state and enable more engaging personalized conversations. We currently have used various heuristics to enhance engagement; it is important to develop more principled approaches to generation of interesting questions based on the estimated multidimensional user and dialogue state representation so as to project a coherent persona of the chatbot.

Acknowledgments

The CharmBana team would like to thank Amazon for providing a grant partially supporting the work of the team and providing many useful resources that our team have exploited to build the CharmBana bot. We thank the Amazon Alexa Prize team and reviewers for their great support, guidance, and help. We would like to thank Heng Ji and the Blender NLP group for their valuable feedback and comments. We are also grateful to Brent Roberts for many useful discussions and resource pointers.

References

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Reason first, then respond: Modular generation for knowledge-infused dialogue. *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Anthropics. 2023. Hh rlhf.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. *arXiv preprint arXiv:2302.03269*.

Ethan A Chi, Caleb Chiam, Trenton Chang, Swee Kiat Lim, Chetanya Rastogi, Alexander Iyabor, Yutong He, Hari Sowrirajan, Avanika Narayan, Jillian Tang, et al. 2021. Neural, neural everywhere: Controlled generation meets scaffolded, structured dialogue. *Alexa Prize Proceedings*.

Ethan A Chi, Ashwin Paranjape, Abigail See, Caleb Chiam, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, Alexander Iyabor, et al. 2022. Neural generation meets real people: Building a social, informative open-domain dialogue agent. *arXiv preprint arXiv:2207.12021*.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

DataBricks. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7780–7788.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. Topic-based evaluation for conversational bots. *arXiv preprint arXiv:1801.03622*.

Amelia Hardy, Ashwin Paranjape, and Christopher Manning. 2021. Effective social chatbot strategies for increasing user initiative. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 99–110, Singapore and Online. Association for Computational Linguistics.

Behnam Hedayatnia, Di Jin, Yang Liu, and Dilek Hakkani-Tür. 2022. A systematic evaluation of response selection for open domain dialogue. In *SIGDIAL 2022*.

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*.

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. *arXiv preprint arXiv:2111.02519*.

Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Angeliki Metallinou, Raefer Gabriel, and Arindam Mandal. 2018. Contextual topic modeling for dialogue systems. In *SLT 2018*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

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)*, pages 8460–8478.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hỳlová, and Jan Šedivỳ. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization. *arXiv preprint arXiv:2109.07968*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Chris Moore. 2006. The development of commonsense psychology.

Takayuki Nakata, Shinichi Ando, and Akitoshi Okumura. 2002. Topic detection based on dialogue history. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

NomicAI. 2023. Gpt4 for all: Open-source assistant-style large language models that run locally on your cpu.

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.

Jan Pichl, Petr Marek, Jakub Konrád, Petr Lorenc, Van Duy Ta, and Jan Šedivỳ. 2020. Alquist 3.0: Alexa prize bot using conversational knowledge graph. *arXiv preprint arXiv:2011.03261.*

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604.*

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207.*

Liliang Ren, Mankeerat Sidhu, Qi Zeng, Revanth Gangi Reddy, Heng Ji, and ChengXiang Zhai. 2023. C-pmi: Conditional pointwise mutual information for turn-level dialogue evaluation. In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 80–85.

RyokoAI. 2023. Sharegpt52k.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022a. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *arXiv preprint arXiv:2203.13224.*

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

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

StabilityAI. 2023. Stablelm: Stability ai language models.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652.*

Jason D Williams, Antoine Raux, Deepak Ramachandran, and Alan W Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244.*

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567.*

Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808.*

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252.

Examples of custom transitioners for topic switching
Oh, by the way, have you tried any interesting cuisine lately? What's your favorite food?
You know what, I've been thinking about travel. Can you share a memorable travel experience of yours?
Random question, but if you could pick up any language instantly, what would it be and why?
Hey, I was just wondering. Is there a particular book that left a deep impression on you?
Just out of curiosity, what kind of music gets you grooving? Any particular artist or genre you're into?
Since we're chatting, I'm curious about your artistic preferences. Do you have a favorite artist or artwork?
Here's a fun hypothetical. If you could have dinner with any historical figure, who would it be and why?
Random question, On the note of innovations, if you could invent something to make life easier, what would it be?
Oh, by the way, Talking about different perspectives, have you ever watched a movie that made you view things differently?
Just for fun, do you have a dream adventure or expedition you'd love to go on someday?
Oh, by the way, On the topic of personal growth, is there a new skill or hobby you've been wanting to learn?
Switching gears a bit, do you have a favorite sport or physical activity you enjoy doing to unwind?
Speaking of trends, is there a current fashion trend you're really into or one you wish would come back?
Random thought, but do you have a favorite quote or piece of advice that's stuck with you over the years?
Hey, just thought of this, but if you could teleport anywhere in the world right now, where would you go and why?
You know, I've been wondering about your cooking adventures. What's the most challenging dish you've ever attempted to make?
By the way, what's the most fascinating thing you've learned recently?
Just curious, if you could be a character from any film or TV show, who would you be and why?
On the topic of storytelling, do you have a favorite podcast or audiobook you'd recommend?
Out of the blue, but if you had a time machine, would you travel to the past or the future, and why?
Just a fun thought, but if you could have any superpower, what would it be and how would you use it?
Hey, on a different note, have you ever tried any exotic fruits or vegetables? What was that like?
Switching topics a bit, if you could spend a day in someone else's shoes, whose would they be and why?
Random thought, but do you have a favorite piece of technology you couldn't live without?
By the way, have you ever attended a concert or live event that you'll always remember?
Just for fun, if you could be an animal for a day, which one would you be and why?
Changing gears a bit, do you have a favorite season or time of year? What makes it special for you?
Curious question, if you could make a cameo in any TV show or movie, which one would it be and why?
On a different note, have you ever had a chance to volunteer or do charity work? Can you share the experience?
Randomly wondering, is there a specific childhood toy you still remember fondly?
By the way, have you ever tried any adrenaline-pumping activities like skydiving or bungee jumping?
Just out of curiosity, do you have a favorite board or video game that you could play for hours?