Athena 3.0: Personalized Multimodal ChatBot with Neuro-Symbolic Dialogue Generators

Yue Fan, Kevin K. Bowden, Wen Cui, Winson Chen, Vrindavan Harrison, Angela Ramirez,
Saaket Agashe, Xinyue Gabby Liu, Neha Pullabhotla, Nan Qiang
Jeshwanth Bheemanpally, Sugam Garg, **Marilyn Walker**, and **Xin Eric Wang**
University of California, Santa Cruz
Santa Cruz, CA
`{yfan71, kkbowden, wcui7, wchen157, vharriso, aramir62`
`saagashe, xinlue, npullabh, nqiang`
`jbheeman, sgarg6, mawalker, xwang366}@ucsc.edu`

Abstract

The burgeoning prominence of conversational agents is evident, given the significant number of people interacting with them on a daily basis. The utility of these agents isn't confined to functioning as personal assistants; they fulfill a variety of objectives. Some of them are task-centric, for instance, facilitating customer support for financial institutions or assisting in reservation processes. Conversely, certain agents are programmed to embody empathy, thereby fostering an emotional rapport with the users. The aim of the Alexa Prize Socialbot Grand Challenge (SGC) is the creation of a socialbot that facilitates engaging, coherent dialogues on a wide array of trending topics appealing to users. In addition to operating as a spoken chat agent, Athena 3.0 offers multimodal interactivity through screen usage. In this technical report, we describe Athena 3.0. Athena 3.0 advances upon its predecessor, Athena 2.0, by supporting multimodal conversation and substantiating a neuro-symbolic experience by fusing LLM-based generators with existing conversation strategies.

1 Introduction

Conversational agents, integral to our daily lives, facilitate seamless interaction between millions of users and various systems. They are designed for a multitude of scenarios, encompassing task-oriented dialogues that involve activities like making reservations to chit-chat dialogues where large language models (LLMs) have emerged as powerful tools for generating human-like text. Some agents are designed to specialize in empathy and emotional connection, and especially socialbots, a subset of these agents, aim at engaging in coherent, meaningful conversations across a wide array of topics. These socialbots are the focus of the Alexa Prize Socialbot Grand Challenge (SGC) [11].

Athena 3.0 engages in high-quality extended conversations on a wide range of topics, processes natural language input, and provides human-like responses [13, 6, 23]. Moreover, with advancements in conversational devices, Athena can utilize multiple modalities, e.g., vision, speech, and text, for user interaction by leveraging Alexa Presentation Language (APL). We show multimodal interactions that are merged seamlessly in the dialogue, providing an enriched and more interactive conversational experience while ensuring that Athena remains just as effective when interacting with users via headless devices - those without a display or visual interface.

Furthermore, we also focus on taking advantage of neural-based response generation models, as they have made significant strides in generating human-like responses in conversational agents. The flexibility and learning capacity of these neural models makes them ideally suited for generating

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

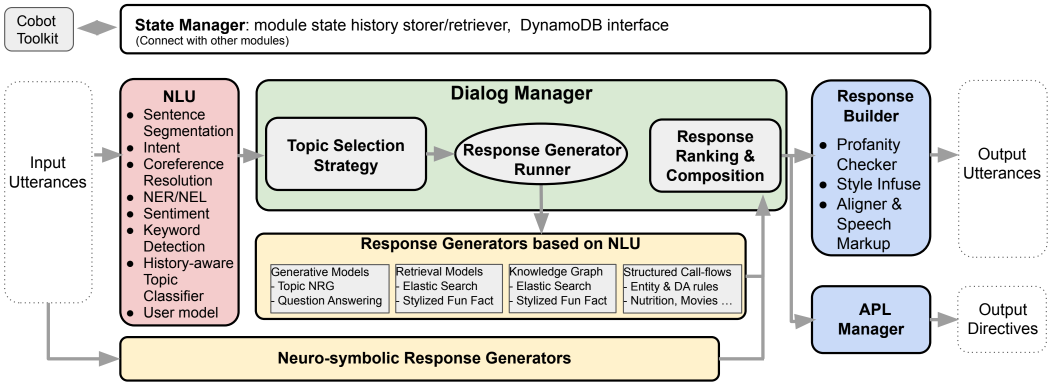

Cobot Toolkit
State Manager: module state history storer/retriever, DynamoDB interface
(Connect with other modules)
Input Utterances
NLU
Sentence Segmentation
Intent
Coreference Resolution
NER/NEL
Sentiment
Keyword Detection
History-aware Topic Classifier
User model
Dialog Manager
Topic Selection Strategy
Response Generator Runner
Response Ranking & Composition
Response Generators based on NLU
Generative Models - Topic NRG - Question Answering
Retrieval Models - Elastic Search - Stylized Fun Fact
Knowledge Graph - Elastic Search - Stylized Fun Fact
Structured Call-flows - Entity & DA rules - Nutrition, Movies …
Neuro-symbolic Response Generators
Response Builder
Profanity Checker
Style Infuse
Aligner & Speech Markup
APL Manager
Output Utterances
Output Directives

Figure 1: Athena's system architecture. The system is built upon the Cobot (Conversational Bot) Toolkit. The input is the user's utterance and the outputs are multimodal responses, including a visual interface and text for Amazon's text-to-speech service.

contextually appropriate and diverse responses, leading to more engaging and meaningful dialogues. In building Athena 3.0, we found that neural-based response generation models could be further optimized within our system when using a neuro-symbolic approach. In this hybrid paradigm, the unrestricted generative power of neural models is controlled by the structured reasoning of symbolic systems. Although this adds a layer of constraint to the model's performance, it also ensures a more predictable output, reducing the likelihood of undesirable responses.

2 Architecture and System Overview

Figure 1 details Athena's architecture. We build Athena using Amazon's Cobot (Conversational Bot) Toolkit which depends on the Alexa Skills Kit (ASK)[1]. Athena runs as an on-demand application that responds to ASK events containing utterance hypotheses produced by Amazon's automatic speech recognition (ASR) service. Cobot also provides seamless integration with Amazon Web Services (AWS), and natively utilizes the AWS Lambda, DynamoDB, and ECS services [15]. We use several additional AWS services, such as the Neptune graph database, and Elastic Search.[2]

The inputs to Athena are the ASR hypotheses for a user's turn, as well as a conversation ID that is used to retrieve the conversation history and state information from a back-end database. This is represented in the Discourse Model. See Section 3.1. Then, the ASR hypothesis is fed into both a natural language understanding (NLU) pipeline and a group of neuro-symbolic response generators (Neuro-symbolic RGs), taking raw user utterances as input. The NLU pipeline produces a representation of the user utterance and conversation context (Section 3), which the Dialogue Manager (DM) (Section 4) uses to update the Discourse Model, and then the Response Generator Runner makes calls to several response generators (RGs) to populate a response pool (Section 5). The response generators are called depending on the context and the NLU results. Additionally, the output responses from the neuro-symbolic RGs will also be added to the response pool. The DM then applies ranking strategies based on heuristic rules and a trained neural ranker to form Athena's next response from the potentially large pool of possible responses. After the response is determined, it will be marked up with SSML [29] and uttered using Amazon's text-to-speech (TTS) service.

To enrich the interaction and cater to a broad range of user devices, we also detect the user's device capabilities and generate multimodal outputs if supported. In particular, we determine if the user's device supports the Amazon Presentation Language (APL) and the screen type of the device. APL enables the creation of visual interfaces for voice-first experiences, adding an additional dimension to the interaction. When the user's device is detected to support APL, we activate our APL manager to generate visual interface outputs that complement the auditory response. The integration of auditory and visual outputs forms a multimodal output, providing a richer, more immersive user experience.

---

[1] `https://developer.amazon.com/en-US/alexa/alexa-skills-kit`

[2] `https://aws.amazon.com/elasticsearch-service/`

3 Natural Language Understanding

Athena's NLU modules consist of a combination of off-the-shelf modules provided with Cobot and Athena's own modules. See Figure 1. We use Cobot's off-the-shelf tools for topic classification and intent recognition. We continued to use our own sentence segmentation and dialogue act tagging modules models from SGC3 (in green in Figure 1) [6] and disclosure model from SGC4 [13]. In addition, we update the named entity recognition and named entity linking model to facilitate the strong language understanding ability of Athena and the user model to include a wider range of user attributes. Further, we build a novel history-aware topic classifier to reduce the chance of outputting off-topic responses (Section 3.3).

3.1 Discourse Model

The Athena discourse model is designed to track the topics under discussion, selected named entities from the last few user and system utterances, speaker information, and anaphoric expressions. The number of entities to track is a parameter of the model and is used to control the size of the "centered" entities [3, 31]. The information in the discourse model is then used in downstream tasks such as named entity recognition (NER), coreference resolution, named entity linking (NEL), and system response generation. These components are described in more detail in our SGC4 technical report [23].

The discourse model facilitates information sharing by storing its data in a state table that is accessible to all Athena modules. This year's focus on the discourse model requires changes to the Response Generators (RGs), which are responsible for creating the system response detailed in Section 5. When the RGs introduce an entity in the system utterance, the entity and its knowledge graph ID are recorded in the discourse model. Downstream RGs can then rely on the discourse model to identify any system-generated entities and their IDs, rather than having the NER/NEL detect them in the system utterance.

An important aspect of the discourse model is that it tracks if entities are introduced through a user utterance vs. a system response. This distinction is important for the confidence estimates given to the entities. The discourse model also contains information pertaining to the entity type (e.g., person, song, movie), which is used by the coreference model to resolve pronouns.

3.2 Named Entity Recognition and Linking

The named entity recognition (NER) and named entity linking (NEL) modules play a crucial role in understanding entity references in user utterances. These components hold particular significance for the Knowledge-Graph RGs, which leverage entity information to acquire additional knowledge from Wikidata, therefore generating coherent responses. Additionally, the APL component utilizes the Wikidata ID to access and showcase entity images described in Section 5.4.

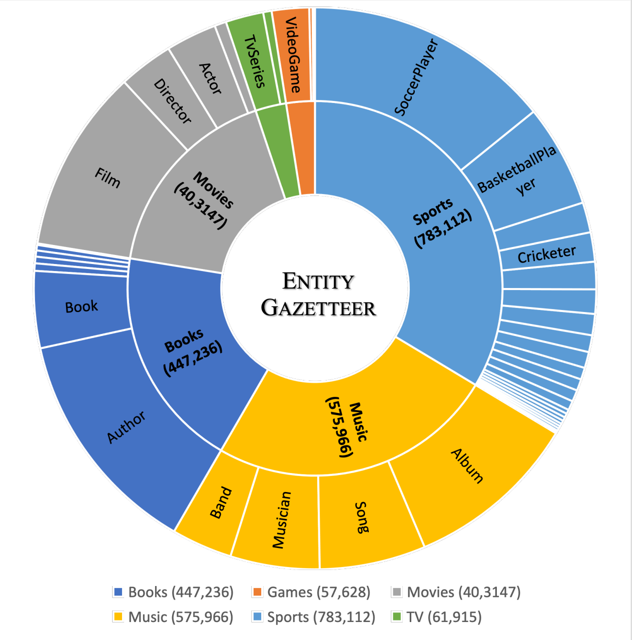


Figure 2: The topic and entity types distribution of the entity gazetteer our NEL system could link to. The numbers in the parentheses are the total number of entities within each domain/entity type.

NER and NEL pose notable challenges in the context of open-domain conversations. These challenges include the scarcity of annotated open-domain conversational data and the lack of recognizing popular entities using existing tools such as Spotlight DBPedia [22], Athena 1.0 [6] and Athena 2.0's [13] NER/NEL modules. The creation of annotated data is particularly intricate due to competition rules that prohibit the utilization of Mechanical Turk for annotating and evaluating raw user data. To tackle this, we incorporated the NEL pipeline developed in prior work [13]. For this year's challenge, our primary enhancement lies in updating the entity gazetteer used by the NEL system. The updated entity gazetteer was curated from the latest Wikidata, consisting of 44 entity types across six prevalent topics, resulting in a total of 2.3 million entities, illustrated in Figure 2. Similar to our previous work, we enhance the entity's popularity metric by accumulating three years of Wikipedia page views as an indicator of entity popularity. To further improve entity recall at inference time, we expanded the entity mentions through integration with the provided Cobot NER module. This augmentation specifically pertains to instances where the Cobot NER module identifies entity types such as 'videoname', 'gamename', 'sportteam', 'songname' and 'person'.

We also investigated a joint learning strategy, aiming to simultaneously detect mentions and resolve ambiguities in one pass [18]. Our approach leverages dialogue context and popularity during model training. However, due to resource constraints, particularly the substantial RAM requirement and inference time, we encountered challenges in deploying this model within the final system.

Category	**P**	**R**	**F1**
Animals	0.79	0.85	0.82
Astronomy	0.71	0.79	0.75
Board Games	0.72	0.83	0.77
Celebrities	0.92	0.89	0.91
Comic Books	0.79	0.81	0.80
Dinosaurs	0.70	0.70	0.70
Fitness	0.55	0.47	0.50
Food	0.81	0.91	0.86
Literature	0.82	0.86	0.84
Movies	0.85	0.89	0.87
Music	0.89	0.93	0.91
Nature	0.84	0.83	0.83
News	0.48	0.53	0.50
Other	0.86	0.84	0.85
Pirates	0.33	0.22	0.27
Sports	0.84	0.87	0.86
Technology	0.55	0.41	0.47
TV Show	0.85	0.82	0.84
Video Games	0.83	0.85	0.84
Weather	0.60	0.55	0.57
Accuracy	-	-	0.84
M-Avg	0.74	0.74	0.74
W-Avg	0.84	0.84	0.84

Table 1: Topic Classifier Results

3.3 History-aware Topic Classifier

In SGC5, we developed a new specialized topic classification module for the topics shown in Table 1, based on the topics covered by Athena's response generators. We first fine-tuned a pre-trained BERT-like model for topic classification using the data from the CONCET dataset [1], but initial experiments revealed that this resulted in a noisy classifier. To rectify this, we annotated conversations using a Vicuna13B [4] Large Language Model (LLM).

We used the previous system utterance and current user utterance as input and prompted the LLM to generate a summary of the conversation (to promote Chain-of-Thought Reasoning [34]) and then classify the output from a list of topics. We then used these annotated user conversations to fine-tune a DeBERTa v3 [8] model. Table 1 shows the results on a test-split consisting of 20% of the collected data stratified by labels obtained by this fine-tuned model. The topic classifier has been used in our experimental traffic as a way to control neural response generation, but more work is needed to assess its performance on Alexa Prize traffic.

3.4 User Model

Creating a user model and modifying the dialogue system's responses based on that user model should increase the user's feelings of agency and create more engaging conversations [16, 24]. We build a user model incrementally over multiple conversations with information from the NLU pipeline, several handcrafted regular expressions, and annotated data. The user model tracks general information across conversations, such as the user's name, whether or not they've self-identified as a youth, and their interests. We also track topic-specific information, such as their pets' names, their weekend hobbies, and their favorite dinosaur. Athena uses this data to adapt system responses to individual users and to control topic promotion [23].

4 Dialogue Management

Open-domain conversational dialogue management is a challenging task due to the many different possibilities for valid responses to a user's utterance given the context. Unlike task-oriented dialogues, where the dialogue manager (DM) can optimize a clear set of objectives, open-domain dialogues do

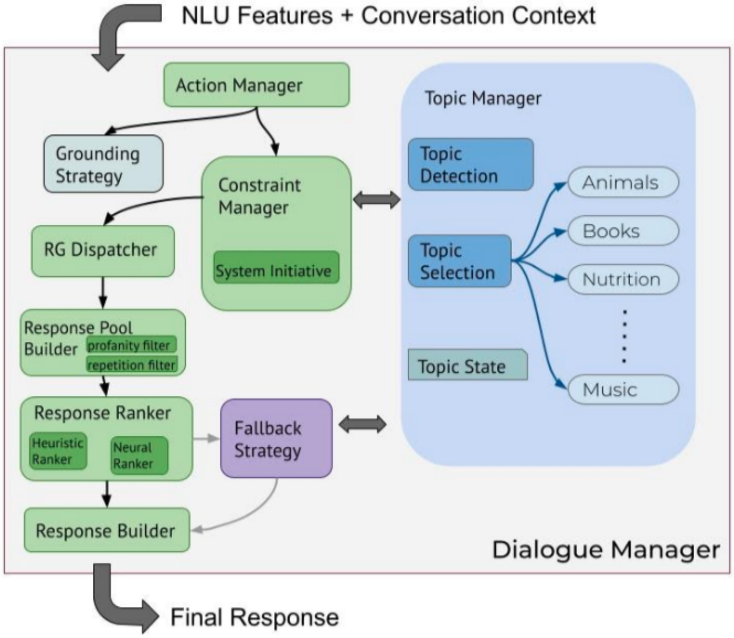


Figure 3: Athena Dialogue Manager architecture.

not have a clear metric of appropriateness [20, 27, 21, 19, 36, 9]. In order to tackle this problem, some earlier Alexa Prize systems rely on using "handcrafted" scripted call-flows to form coherent dialogues on a particular topic. Athena's DM uses a shallow hierarchy based on a top-down decomposition into a number of subcomponents, each with its own responsibility. These are oriented as a pipeline: the outputs of one component directly feed as inputs to the next. The DM sub-modules are shown in Figure 3. Some modules remain similar to what was used in SGC-4 [6, 13]. Our primary focus was retraining the response ranking module with the hope that an improved version of this module could distinguish the good and bad outputs of various NRGs.

4.1 Response Ranker

The goal of the response ranker is to find the best response that also matches the output of the constraint by the Constraint Manager. Athena's response ranker is a BERT-based neural response ranker that was trained on hand-annotated Alexa Prize conversation data. The annotators were given two turns each of the user and Athena's conversation context along with a pool of candidate responses and asked to rank the candidate responses. We have annotated additional data in SGC4 and retrained this response ranker with new data twice during SGC5.

We started out with the response ranker described in [7]. We trained a new transformer model and annotated more conversation data for response ranking. We had two interests while collecting data for response ranking annotation. First, we want to add more question contexts, and second, introduce synthesized responses from neural response generators.

Data collection, annotation, and processing. We start with a previously created dataset of conversation response pools with response quality scores assigned by expert annotators [7]. The initial dataset contains 2,094 conversation contexts and a mean response pool size of 4.43 system utterance candidates produced by Athena. We perform further annotation to increase the corpus of examples for training. Annotation efforts were in several rounds, and team members were recruited to rank Athena's responses by hand. This resulted in a new dataset of 7,751 instances, 37,235 annotated responses, and a mean response pool size of 4.89. The instances span across 29 conversation topics sourced from Athena's 50 response generators.

Response Ranking model. The response ranking Transformer model is a GPT-2 style decoder-only transformer. We experiment with finetuning versions of DialoGPT and DialogRPT models on the next response scoring task using a regression learning objective. The model operates similarly to [7].

It takes in four turns of conversation history, some of Athena's internal conversation state information, and a candidate response to score. The response ranker assigns each response a score between 0 and 1. The best is 1.

Evaluation: Performance in Deployment. During both the semi-final and final periods, we used two versions of the DM with the trained response ranker. One version of the DM (main-b-newranker) allowed the response ranker to make all decisions about the next turn in dialogue, while the other (main-a-flow-priority) gave priority to responses that were within a hand-crafted call-flow where the RG controlling the call-flow had also taken the previous turn.

We repeatedly deployed updated versions of the response ranker over the semi-finals and completed an A/B test of the new ranker (main-b-newranker), compared to a flow-priority version of the DM (main-a-stable-flow-priority-ranker). From May 8 we deployed Athena such that the main-b-newranker system configuration received 50% of the conversation traffic and main-a-stable-flow-priority-ranker the other 50%. We performed an analysis on May 17 and the results showed a statistically significant improvement (p = .01) in ratings for main-b-newranker (3.52) vs. main-a-stable-flow-priority-ranker (3.26) as well as a significant difference in length (p = .01) with the average conversation length of main-b-newranker 3.24 turns greater than main-a-stable.

We then updated the flow-priority version and again ran both of these versions in main traffic for a two-week period before the code freeze with the result that main-b-newranker had an average score of 3.35 while main-a-flow-priority had an average score of 3.25. We thus converted all main-traffic to the main-b-newranker version of the DM at that time.

5 Response Generation

5.1 Flow-Based Response Generators

Athena supports the topics shown in Table 2 with a set of mini-flows for each topic, built using Flow-RG, a tool we created to support the design of dialogue flows and the communication of information systematically from the Flow RGs to the DM, as detailed in our SGC4 and SGC3 reports [6, 13, 23]. For Movies and Music, there are two types of Flows, one that is scripted chit-chat and another supported by the Wikidata KG, as discussed in more detail in Section 5.5.3. Every topic also has an associated Centering RG (Section 5.2). The Flow-RG framework supports interweaving within a topic between multiple RGs on that topic and with other types of RGs, such as for question answering. A detailed description of interweaving is provided in our SGC4 report [13, 23].

sports	*movies*
books	*nature*
news	*animals*
astronomy	*comic books*
dinosaurs	*harry potter*
nutrition	*pirates*
video games	*board games*
TV	*food*
hobbies	

Table 2: Athena's topics.

5.2 Centering RGs

Centering RGs are topical retrieval-based generators where fun facts relevant to the current context are retrieved according to their indexed discourse entities and concepts. Almost every topic in Athena has a corresponding Centering RG. These Centering RGs can interweave with both KG RGs and Flow RGs. This year we extended this RG in two ways; firstly, we generated several informal paraphrases of our original fun facts, and secondly, we used the original fun facts as the base context for the personal question generation pipeline described in Section 5.3. This yields a variety of informal personal questions ($\sim$19k in total) that are retrieved given the appropriate context. Furthermore, our dialogue manager can interweave a paraphrase of the base trivia with multiple personal questions that are all contextually relevant. Figure 4 shows examples of these changes.

5.3 Generated Personalized Content

Another expanded feature in Flow-RG this year is the *universal templated miniflows*; only a data file needs to be provided in order for them to be enabled for a given topic. In SGC4, Athena supported two types of *universal templated miniflows*, Would-you-rather sequences and Hypothetical questions, in most of Athena's supported topics, which was shown to significantly improve Athena's

Content Type	Response
Original Fun Fact	Did you know that Shazam, who was formerly known as Captain Marvel, outsold Superman, Batman, Wonder Woman, and the entire comic books industry for most of the 1940s?
Fun Fact Paraphrase	The comic book industry was dominated by Shazam in the '40s, surpassing Superman, Batman, and others.
Personal Question	Alright, real talk. Do you think it was the yellow lightning bolt or the red cape that made Shazam so popular in the 1940s?
Personal Question	That's wild, man. So who do you think would win in a tag team match: Shazam and Batman or Superman and Wonder Woman?
Original Fun Fact	Ostriches can run faster than horses, and the males can roar like lions.
Fun Fact Paraphrase	If you didn't know, ostriches are faster than horses and male ostriches can produce a roar that sounds like a lion's roar.
Personal Question	Yo, let's have some fun. If an ostrich was running as fast as a horse towards you while roaring like a lion, how would you react?
Personal Question	If ostriches could be trained to race like horses, do you think people would ever bet on them in races, or would it just be too strange?

Figure 4: Two examples of Athena's updated fun fact paraphrases and personal questions used by the Centering RGs.

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

Figure 5: Example conversations regarding the Would-You-Rather sequence (upper) and the Hypothetical question (lower) in the Food and Nature topics, respectively. Due to privacy considerations, the user utterances displayed are from the developer rather than actual users.

performance [2]. Examples of these miniflows are seen in Figure 5. This year, we added three new types of *universal templated miniflows* to each of these topics: Personal_Questions, Informal_Trivia, and Personal_Opinions. While our original WYR_Questions and Hypothetical_Questions were manually curated, this new content has been generated and stored for retrieval. In total, the size of the combined corpora of the new WYR_Questions, Hypothetical_Questions, and Personal_Questions is ˜19k total newly generated questions.

One fundamental challenge faced when building open-domain socialbots is having topical coverage across a large spectrum of user interests. Previously, Athena robustly supported less than 20 topics; if the user has other interests, Athena would rely on fallback tactics while routing users to more robust topics. In order to better personalize Athena's content to a user's unique interests, we use the same generation pipeline to generate content for all five miniflows for over 400 new topics. Our dialogue manager's ability to interweave these miniflows enables Athena to have multi-turn conversations across a larger spectrum of topics. The impact of these changes is currently being evaluated. An example sequence of a previously unsupported topic, art, is shown in Figure 6.

5.3.1 Evaluating Universal Templated Miniflows and Centering RGs

We compare seven different conversational strategies: two strategies with the new corpora of generated personal questions (**A** and **B**), two strategies that use existing personal question strategies (**C** and **D**), and three different types of statements (**E**, **F**, and **G**). Our analysis (Table 3) shows that **A**, **B**, and **C**, which all ask a variety of open-ended personal questions, yield longer user utterances. Moreover, the difference between **A**, **B**, and **C** are statistically significant when compared

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

Figure 6: A sample conversation about art, a previously unsupported topic, that interweaves five different types of generated content together. Due to privacy considerations, the user utterances displayed are from the developer rather than actual users.

against all pairs, i.e., **A** significantly **B** significantly **C** significantly **D**, **E**, **F**, and **G** (the last four are not significantly different). These results are encouraging. **C** and **D** have already shown statistically significant improvements to our socialbot; since **A** and **B** are even more engaging (longer user turns), we expect they will amplify our previous gains. Moreover, we see that all three open-ended personal question strategies significantly outperform statements and the average user utterance length across our entire system (overall utterances, it's 5.04), which reinforces our expectation that personal questions are crucial in social conversation. Additionally, running the provided ODES classifier [17] on these user turns reveals that $\leq 2\%$ of user turns were classified with negative ODES classes, which also indicates a positive addition to the interaction in most cases.

	Data Type	Avg. Len.
A*	**Deep Personal Qs**	**7.41**
B*	**Centering Personal Qs**	**6.52**
C*	**Hypothetical Qs**	**5.92**
D	Would You Rather Qs	5.46
E	Personal Opinions	5.35
F	Informal Trivia	5.15
G	Centering Fun Facts	5.14

Table 3: Average user utterance length per content. Generated deep personal questions and personal questions generated from fun facts elicit the longest user responses. Over all utterances, the average is 5.04. The evaluation is based on over 15k user turns collected over a 10-day period of time (June 15th - June 25th).

We additionally run an A/B study over 15 days of live user traffic. In this study, A represents the full version of Athena, while B represents a version of Athena that cannot ask the new deep Personal_Questions, and cannot use the new Centering RG personal questions. Table 4 shows our results. We see a statistically significant improvement in both rating and conversation length when requiring at least one personal question to be asked in the conversation. Moreover, we calculate the Pearson correlation of the number of personal questions asked to conversation length and user rating and find that they have a positive and statistically significant strong correlation (.90) and weak correlation (.12), respectively. Running a linear regression on the number of personal questions with respect to user ratings also confirms a positive impact, i.e., β_{num_qs}: .1. However, breaking this down further, a multivariate linear regression reveals a surprising negative relationship with our Centering Qs, while we still see a positive relationship for our Personal_Question Qs and conversation length, i.e., $\beta_{num\_center\_qs}$: -.01, $\beta_{num\_flow\_qs}$: .04, and β_{con_len}: .13. When we further inspect the Centering Qs, we find that many times a question will be used without the appropriate setup context, making it confusing or seemingly irrelevant to the conversations current state. This highlights the importance of not just asking as many personal questions as possible but ensuring they are appropriately interwoven within the conversation.

Req. PQ	A rating	B rating	p-value	A length	B length	p-value
0	3.48	3.41	0.37	22.37	22.32	0.97
1	3.60	3.41	0.02	32.33	22.32	0.00
2	3.68	3.41	0.00	38.40	22.32	0.00
3	3.76	3.41	0.00	44.67	22.32	0.00

Table 4: Results from an A/B trial over 15 days (June 28th - July 12th). A represents the full version of the system, while B represents a version of the system that cannot ask DPQs or fact-based Qs. The Req. PQ column represents the minimum number of PQs in the conversation.

In addition to the above-described generated content, it is desirable to generate even more personalized questions that are able to combine more values stored in Athena's user model with a fine-tuned neural generator. This work is ongoing, but samples from our training data are demonstrated in Figure 7.

User Model Values	hobbies: [baking] topical_interest: [video_games]
Prompt	I like baking and playing video games. Can you ask me a specific personal question about that?
Personal Question	Do you have a go-to video game that you always play while waiting for your baked goods to cook?
Personal Question	Have you ever baked a themed dessert based on a video game?
User Model Values	favorite_tv_show: [Ink Masters, Comic Book Men] topical_interest: [comic_books, movies] topical_disinterest: [food, music]
Prompt	I like the show Ink Masters and Comic Book Men. I'm not interested in food or music. I like comic books and movies. My favorite comic character is Magneto. Ask me some specific personal questions that are about the shows I like and my interests.
Personal Question	If you could have any comic book character tattooed on you by an Ink Masters artist, who would it be and why?
Personal Question	In Ink Masters, the judges often critique the artists' technical skills and ability to bring a design to life. When it comes to comic book art, do you prefer more realistic or stylized interpretations of characters?
User Model Values	positive_opinion: [night time, looking at the sky and enjoying myself, soccer, to walk by myself, hiking, music, the piano, every type of seafood]
Prompt	I like a lot of things. I like night time, looking at the sky and enjoying myself, soccer, to walk by myself, hiking, music, the piano, and every type of seafood. Now ask some specific personal questions that combine the things I like.
Personal Question	Have you ever gone on a nighttime hike to a scenic location where you could stargaze? If not, is this something you'd like to try?
Personal Question	Have you ever gone on a solo hike where you brought along a portable keyboard or piano to play music in a scenic location? If not, is this something you'd consider doing?

Figure 7: Examples translating user model values into prompts and the resultant questions.

5.4 Multimodal Content Generation

During our experimentation in SGC5, we uncovered the substantial enhancement that screen-enabled devices can offer to the user experience of conversational agents. By leveraging the Alexa Presentation Language (APL), which facilitates the creation of interactive voice and visual experiences across a broad spectrum of devices, we've been able to innovate our user interface designs and realize visually engaging, interactive conversational experiences.

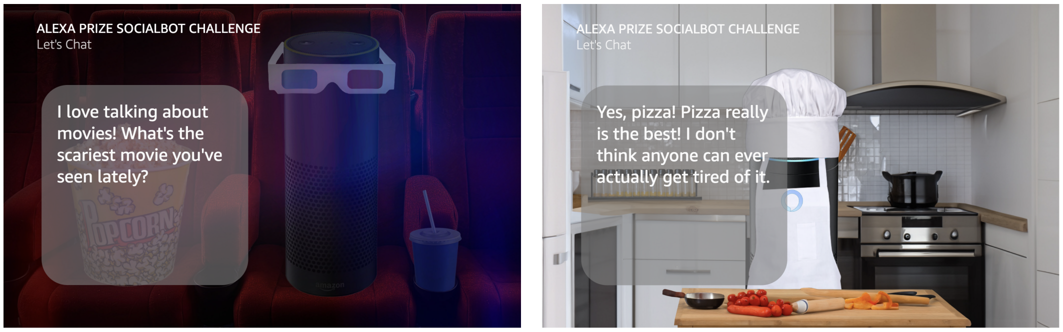
To achieve this, we introduced an APL Manager, a module explicitly engineered to generate multimodal content. This tool maximizes the potential of screen-enabled devices, fostering a dynamic and immersive interaction for the user. Based on the state of the Dialog Manager, the APL Manager selects one of the three unique user interface designs as a template and populates it with relevant content. We provide detailed descriptions of these interfaces and the conditions under which they are used in the subsequent sections.

5.4.1 Conversation Detail Screen

This screen is the default template to be selected by the APL Manager and is composed of both textual and graphical components. In terms of textual elements, the design comprises a header and Athena's response, displayed on the left in a dedicated text box with a darkened background. We integrate the 'Karaoke' feature of the APL API, which allows the text to auto-scroll synchronously with Athena's speech, thus creating an engaging multimodal response that combines both audio and visual cues.

Graphically, this screen contains a background image and an optional foreground image on the right side. The background image, as shown in Figure 8, comes from a static human-collected database which includes Creative Commons images and original Athena Avatar images designed by us. Each image is accompanied by one or more labels about the topic and keywords that match the image. A two-step background image selection mechanism is adopted, where the first step filters out images that match the conversation's topic and sub-topic of the miniflow, while the second step involves further determining the background image based on keyword matching. This approach ensures the background imagery aligns with the dialogue's context, enhancing the overall immersive quality of the user experience.

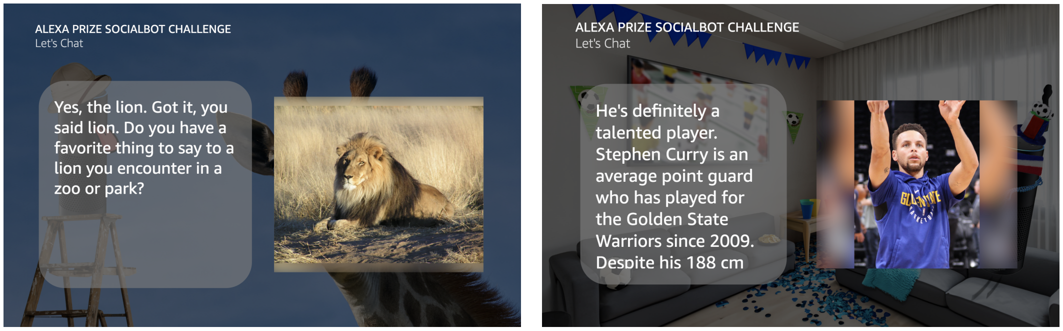
The optional foreground image serves as a visual supplement, drawing attention to specific entities. In order to facilitate the retrieval of entity images, we have enhanced our call-flows to include the corresponding entity ID. The Knolwedge-Graph RGs are designed and already associated with Wikidata IDs. Consequently, the APL manager retrieves images using these IDs from Wikdiata or Wikipedia. an entity response template is employed to exhibit the entity's image as the foreground visual. To emphasize the foreground images, we designed a blurred background in such instances, as illustrated in Figure 9.


ALEXA PRIZE SOCIALBOT CHALLENGE
Let's Chat
ALEXA PRIZE SOCIALBOT CHALLENGE
Let's Chat

I love talking about movies! What's the scariest movie you've seen lately?

Yes, pizza! Pizza really is the best! I don't think anyone can ever actually get tired of it.

Figure 8: Examples of presenting Athena Avatar Images during the introduction phase with corresponding responses.


ALEXA PRIZE SOCIALBOT CHALLENGE
Let's Chat
ALEXA PRIZE SOCIALBOT CHALLENGE
Let's Chat

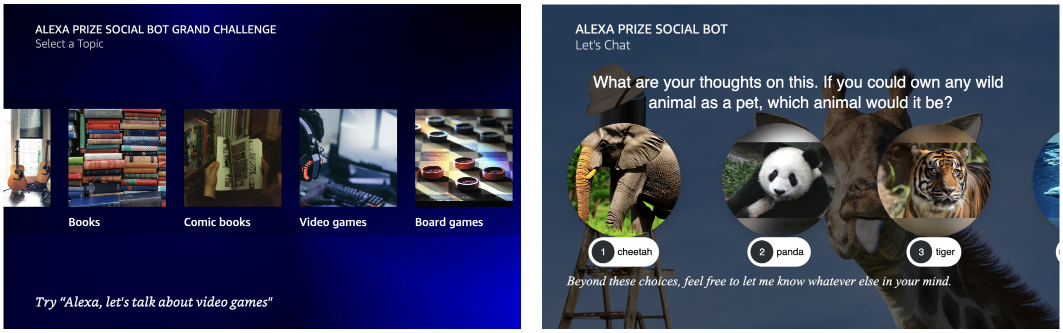
Yes, the lion. Got it, you said lion. Do you have a favorite thing to say to a lion you encounter in a zoo or park?

He's definitely a talented player. Stephen Curry is an average point guard who has played for the Golden State Warriors since 2009. Despite his 188 cm

(a) An example of the Animal call-flow presenting entity image with the corresponding response.

(b) An example of the Sports Knowledge-Graph RG using NEL result and presenting entity image.

Figure 9: Examples of presenting entity images.



Try "Alexa, let's talk about video games"

What are your thoughts on this. If you could own any wild animal as a pet, which animal would it be?

Beyond these choices, feel free to let me know whatever else in your mind.

(a) An example of using personal questions with option choices.

(b) An example of a topic selection screen to show a list of supported topics we support.

Figure 10: Examples of option selection.

5.4.2 Interactive Topic Selection Screen

In scenarios where our Dialog Manager identifies that the current topic is "menu_topic", indicating a need for a shift in conversation topic, the Interactive Topic Selection Screen comes into play. As shown in Figure 10a this user interface design is instrumental in fostering a more interactive and user-driven conversational experience. The screen presents Athena's utterance suggesting new conversation topics, facilitating a smooth transition in the dialogue. Central to the screen are the topic names and corresponding images of the proposed topics, arranged in a row. These elements aren't merely decorative; they are designed to be interactive and clickable.

		rating	ratio
shape	rectangle round	3.2 1.2	98% 2%
orientation	landscape portrait	3.4 3.2	92% 8%
mode	Hub TV Mobile	3.2 3.0 2.7	79% 12% 9%

Table 5: Average rating and the percentage of traffic for each screen type from 7/1/2023 to 7/31/2023

In this setting, users have the freedom to either verbally communicate their desired topic to Athena or directly interact with the clickable topic images on the screen. The latter action emulates the effect of verbally stating the topic, providing a seamless blend of voice and touch interactions. By enabling users to choose topics visually, the Interactive Topic Selection Screen offers an alternative mode of interaction that may be more intuitive for some users. Furthermore, it encourages user engagement and participation, allowing users to steer the course of the conversation.

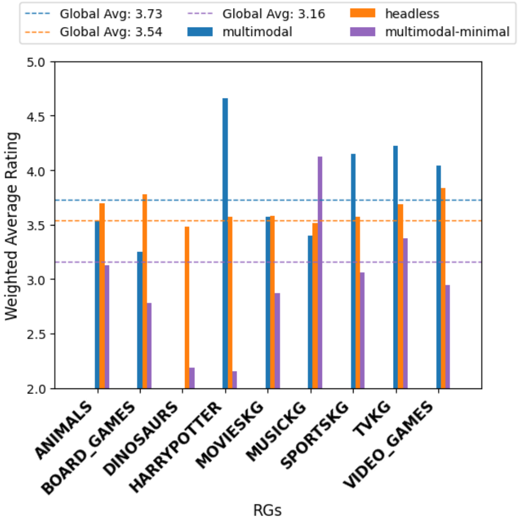


Table 6: Comparison of weighted average rating with selected topics for multimodal, multimodal-minimal, and headless from 7/24/2023 to 8/2/2023.

5.4.3 Interactive Question Answer Screen

The Interactive Question Answer Screen is designed to be displayed when Athena asks a question with predefined candidate answers. As shown in Figure 10b, this interface enhances user engagement and enriches the decision-making process, offering a visually interactive response selection method.

At the top of the screen, Athena's question is prominently displayed. This is followed by the presentation of candidate answers in the middle of the screen, each accompanied by a corresponding image. These candidate answers give users a sense of what other responses might be like acting as conversation prompts or hints.

The images accompanying each answer are initially produced by a two-stage pipeline. We first create a base prompt that includes the option text, keywords related to the topic, and the topic itself. This is run through a prompt generation model optimized for stable diffusion, known as MagicPrompt [26]. We input this richer prompt into the Stable-Diffusion-2 model to generate a corresponding image. To ensure their appropriateness and relevance, each generated image then undergoes a manual review. This process ensures that the images accurately represent the corresponding responses, thus supporting user comprehension. Similar to the Interactive Topic Selection Screen, the candidate answers are clickable, and thus the Interactive Question Answer Screen offers a dual-mode interaction. Users have the choice of either verbally stating their selected answer or clicking on the corresponding image. This design continues our trend of providing a multimodal, user-centric conversational experience, offering users flexibility and control over their interactions with Athena.

5.4.4 Multimodal Evaluation

We ensured our user interfaces were responsive, as we found through tracking customer device types that users have different screen sizes, such as portrait, landscape, and circle sizes. Table 5 shows the average rating and the number of conversations for each device type during one week of traffic in the final. Table 6 shows the comparison of the weighted average between multimodal and headless. We select a subset of topics that has functionality targeted at multimodal settings. Due to the insufficient traffic for some device types, i.e., round shape and portrait orientation, we found no significant differences among types.

5.5 Neural-based Response Generators

In the unfolding era of large language models (LLMs), we aim to enhance the interactivity and responsiveness of Athena by leveraging the capabilities of neural-based response generators (NRGs). We have made strides in understanding and leveraging the capabilities of NRGs in two distinct approaches: an end-to-end approach and a neuro-symbolic approach. For either approach, the NRGs use raw user utterances as input without relying on any extra processing of the input, which improves the system latency given the extended time needed for the NRGs.

End-to-End Neural Response Generators. Language models trained on large volumes of data can be adopted as response generators to generate responses that bear a convincing resemblance to human discourse. In our work, we experiment with the Neural Response Generator supplied in the cobot toolkit (cobotNRG), AlexaTM 20B [28], and RedPajama 7B-Chat model [3], applying them in an end-to-end manner, where the outputs generated by these NRGs were directly added to the response pool. However, we observed that the robustness of this end-to-end suffered. Specifically, when faced with incomplete user utterances - a frequent occurrence in SGC5 - the NRGs often produced ambiguous outputs. Through human annotation of the quality of the NRGs' outputs, based on real conversation history, we discovered that responses from end-to-end NRGs are less preferred compared with responses from Flow-RG. We carried out an A/B test on the end-to-end NRGs in the headless device traffic for more than a week, gathering over 300 conversations for both variations: one with and the other without end-to-end NRG responses in the pool. The data indicates that the version with end-to-end NRGs received an average rating of 3.29, which is lower than the 3.46 rating achieved by the version without end-to-end NRGs.

Neuro-Symbolic Response Generators. Neuro-symbolic response generators complement the end-to-end method by incorporating the principles of neuro-symbolic computing. This strategy harnesses the strengths of both neural networks and symbolic reasoning, aiming to create a seamless fusion of the two paradigms. To construct neuro-symbolic response generators, we utilize NRGs that have been fine-tuned for specific tasks and apply certain rule-based constraints to each NRG output based on the dialogue act detected from the user's utterances. As a result, the responses from neuro-symbolic response generators are guided by clear rules that are adaptable and robust to various conversations, which significantly reduces the probability of generating low-quality responses.

We developed the three fine-tuned neural models based on diverse datasets tailored to different conversation scenarios: a grounding NRG, a question-answering NRG, and a knowledge-graph RG for five different domains. The selection of the base model was significantly influenced by the strict

---

[3] https://www.together.xyz/blog/redpajama-7b

latency requirements of Athena. Based on our previous year's technical report, we observed a strong correlation between user satisfaction and response latency, with quicker responses invariably leading to a more satisfactory user experience. As a result, we chose the RedPajama 3B model [5] as our base model for fine-tuning, which is renowned for its efficient response generation capabilities, to handle fine-grained generation tasks described below effectively. The fine-tuning is done on 4 Nvidia A5000 graphics cards. We applied parameter efficient fine-tuning using LoRA [10] to train for 15 epochs at a learning rate of *2e-3* to *2e-5* with various models and the sampling process is controlled by *temperature=0.7*, *top_p=1*, and *top_k=0* to reach a decent performance.

5.5.1 Grounding NRG

To enhance Athena's conversational context awareness, we also trained a new Grounding NRG, a specialized neuro-symbolic response generator that generates backward-looking DAs that ground the user's utterance in the dialogue history [12]. Specifically, it processes the last two turns of dialogue to produce a response that evaluates or acknowledges the user's utterance, and this grounding response is then prepended to the system's next turn. Such an approach ensures that the generated responses are not only contextually relevant but also create a sense of continuity in the dialogue, enhancing the overall conversational flow.

We previously grounded the user utterances using an index of grounding turns that were based on a frequency analysis of user answers to common system questions: the user answers were identified with regular expressions, and the grounding turns were then prepended to the system's next turn. To produce training data for the grounding NRG, we pulled 36K user utterance contexts from SGC4 and SGC5, where our rule-based regex grounding was triggered, expecting that the NRG would be able to generalize from these examples and have broader coverage. The resulting fine-tuned module had an average response time of 1.5 seconds, and we deployed it to half of the main-traffic on July 18th, restricting its use to dialogue contexts where the detected user dialogue act is a type of opinion, comment, or statement. Our perception was that the grounding module improves system responsiveness and, together with the QA NRG eliminates those situations where we ignore the user's utterance. However, we have not yet been able to measure a significant difference in the user ratings with and without the grounding module.

5.5.2 Question Answer Response Generator

We set out this year to develop a novel response generator that could answer user questions in SGC dialogues. We have informally observed that in the SGC setting, users pose a variety of questions with different types at unpredictable times throughout the dialogue. Moreover, while QA is a well-established field in NLP, it was also clear that a standard QA engine can only answer a small percentage of users' questions in SGC dialogues. So our first step in developing a QA RG was to collect all the user utterances from SGC3, SGC4, and SGC5 labeled as a question dialogue act (QDA) by either our MIDAS DA tagger or our Fine-Grained CRF (FG) tagger [37, 6, 23]. MIDAS question labels are open_question_factual, open_question_opinion, yes_no_question, while the FG question labels are advice-question, fact-question, opinion-question, and personal-question. We segmented the user utterances into DAs and extracted only the segments labeled as questions. This resulted in a sub-collection of over 300K questions composed of roughly 150K distinct types of questions. We then sorted these by frequency. Several things are immediately apparent: (1) Neither the MIDAS nor FG QDA types intended to identify factual questions actually do so reliably; (2) There is a very long tail (40% singletons, 2% doubletons); (3) Many non-questions are classified as questions. We thus developed the Q taxonomy shown in Table 7 aimed at characterizing the distribution and types of questions.

Table 7 shows the types and number of type instances of the top 300 most frequent Qs and what we observed in a random sample of 300 singletons. The top 300 most frequent questions account for 36% of the instances. The most frequent category is PERSONA questions, where the user asks about Athena's personal history, past experiences, or opinions. Earlier versions of Athena attempted to answer these questions with a large index of backstory questions and answers, where questions were identified via regex patterns. The next two most frequent Q types, CLARIFY and ELLIPTICAL are conversational and dependent on context for interpretation. The class COMM-SIT refers to meta-level questions about the communication situation, which are surprisingly frequent. It is striking that

TYPE	**Ex**	**Top 300** (%)	**SING** (%)
PERSONA	Have you ever been to Hawaii?; What is your favorite animal; Do you have any siblings	26.7	18.7
CLARIFY	Who is she? What is that?	13.0	6.3
ELLIPTICAL	what else?; did you?	13.0	2.7
PHATIC	How are you?; How are you doing?	5.7	0.6
OPENING	Can we talk?; Do you want to chat?	6.7	0.0
COMM-SIT	What can we talk about; Do you want to know my name?	9.3	15.3
SITUATION	What time is it?; What is the weather	1.7	2.0
FACTOID	Who is the president of the US; What team does Steph Curry play for	0.0	15.3
REQ-REPEAT	What?; What did you say?	2.0	0.0
PRE-Q	I have a question for you; Can I ask you something?	2.3	0.0
ASR	not interpretable	0.0	7.7
RED	Can I sleep with you?	0.0	1.0
NOT-A-Q	Misclassified Q	20.0	26.7

Table 7: Representative questions, their types and distribution in Athena SGC3, SGC4 and SGC5 . Top 300 provides the counts for the top 300 most frequent questions while SING provides the counts for a random selection of 300 singletons.

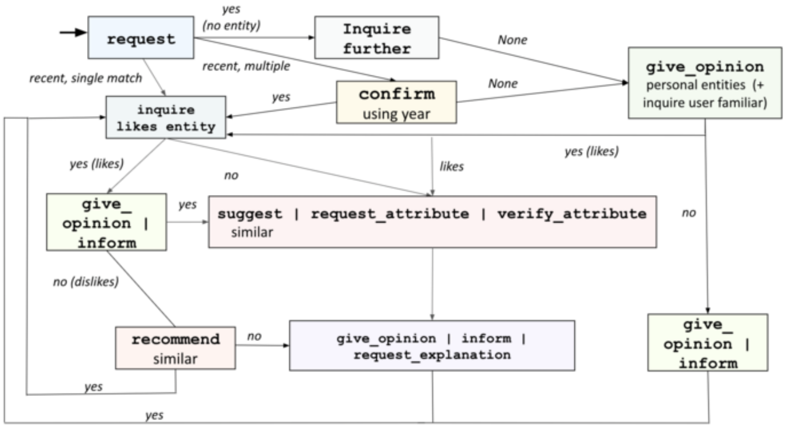


Figure 11: Cross-Domain Universal Dialogue Policy based on the ViGGO Dialogue Acts

there are no fact-seeking questions in the top 300 most frequent, and that of the sample of singletons fact-seeking Qs account for only 46 out of 300.

In order to train a QA engine that would be able to answer user questions of all types, we decided to identify an LLM that provides good performance and then generate synthetic training data by prompting the LLM. However, one challenge with this is the need to respect user privacy: the actual user question utterances cannot be used to prompt an API for an LLM. We decided to, therefore, first train a question generator (QG) on the user data and then use the generated questions with the Open AI default chat LLM (after pilot explorations testing the quality of answers generated by CobotNRG, Cobot Topic NRG, ATM 5B, ATM 20B and OPT 13B).

In spite of the fact that not all of the utterances identified as QDAs are actually questions, we used a stratified sample taken from the whole corpus of user utterances along with one system turn of context to fine-tune a QG engine using OPT 1.3B and used it to generate a sample of 22K Qs. We then refined the training data for QA by utilizing the hand-curated answers that we already had in Athena as part of the training data, along with answers generated by OpenAI Chat, and fine-tuned a QA engine.

The QA engine was deployed in Athena on July 18th and called whenever either the MIDAS or the FG-CRF classifiers identified a question in the user utterances. As a low-risk strategy, we prefixed the generated answers onto the next system turn in the main traffic while experimenting with letting the QA engine take the whole turn in experimental traffic. Our impression is that the QA NRG improves interactions with users since otherwise, many user questions are ignored, but the user ratings do not demonstrate any significant improvements.

5.5.3 Knowledge Graph-Based Response Generators

The KG RGs in SGC4 covered Movies, Music, TV, and Sports, and targeted entities and relations in Wikidata [23]. In recent work, we showed that we could use prompt-based learning to produce a meaning-to-text corpus with controlled DAs using DAs from the ViGGO corpus, which are also used in Athena's Video Game RG [25, 14]. We hypothesized that we could improve Athena's KG RGs by: (1) using neural generation rather than template-based generation; (2) taking advantage of an NRG's ability to realize unseen WikiData relations; (3) transferring the dialogue acts (DAs) from ViGGO to the KG RGs; and (4) utilize the same set of DAs across all topics in the universal dialogue policy shown in Figure 11.

One of the challenges with transferring DA types is that we have no training data for the DA types in the KG domains. However, We show that a combination of prompt-based transfer learning (PBL) and self-training can successfully transfer conversational DAs from the ViGGO corpus onto KG-based meaning representations for Movies, Music, TV, and Sports. Figure 12 illustrates the cycle of self-training, where we start with ViGGO examples and then do a round of self-training where we use the silver outputs from the first round in a second round of prompts and re-generate. We achieved an improvement in semantic accuracy from .77 pseudo-Beyond-BlEU to .80.

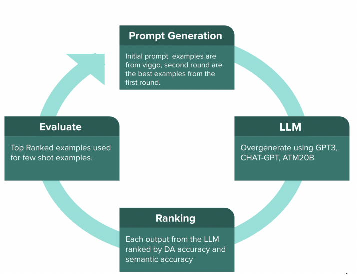

Prompt Generation
Initial prompt examples are from viggo, second round are the best examples from the first round.
LLM
Overgenerate using GPT3, CHAT-GPT, ATM20B
Ranking
Each output from the LLM ranked by DA accuracy and semantic accuracy
Evaluate
Top Ranked examples used for few shot examples.

Figure 12: Self-Trainining Setup

Table 8 provides an example dialogue for the Music domain based on the universal policy. We were able to deploy the full policy for the Movies and Music KGs and utilize the `give_opinion` DA in Sports (athletes), TV (shows), and animals (wild animals). This policy smoothly moves from one entity to another using different DAs and multi-hop traversals of the KG, taking advantage of the connected nature of Wikidata and its abundance of content. Note how the **Suggest DA** introduces a related entity via a traversal of the KG.

Athena: Hi, this is an Alexa Prize Socialbot. It's so good to talk to you again! I hope your weekend has been rejuvenating. How are you doing today?
User: I'm doing good, I want to talk about music
Request Attribute DA Athena: Sure, we can talk about anything you want. Hhmm, anyways, I love talking about music! What kind of music do you like to listen to?
User: I enjoy pop music
Request DA Athena: Do you know of any pop songs that are catchy?
User: Yes, never gonna give you up
Confirm DA Athena: Is it Rick Astley's song Never Gonna Give You Up that you're referring to?
User: Yes
Give Opinion DA Athena: Rick Astley did a nice job with the single 'Never Gonna Give You Up' from 1987. It's quite good!
User: That sounds great, I totally agree with you.
Suggest DA Athena: Do you also like pop songs by Kylie Minogue, for example, All I See?
User: No I don't know about that song
Inform DA Athena: I must say, Kylie Minogue really nailed it with their performance of All I See from the album X. It's a fantastic track that I love listening to.

Table 8: Music example of Universal Dialogue Policy. Due to privacy considerations, the user utterances displayed are from the developer, rather than actual users.

Fine-tuned NRG for KGs. To achieve real-time high-quality natural language generation, we initially used the high-quality outputs of GPT-3 (175B params) to create templates which were then indexed within the dialogue policy states. However, once we fine-tuned the RedPajama model [5] using the GPT-3 outputs, we replaced these templates by directly calling the model on the input attributes to generate the knowledge-grounded utterances at each state.

Before fine-tuning the RedPajama model, the GPT-3 outputs were filtered for quality based on semantic accuracy and dialogue act. We utilized the pipeline to fine-tune the pre-trained RedPajama (refer to Section 5.5 with data from 5 domains that include nine entities in total: Movies (movies,

Music	
NAME	GENRES
PRODUCER	RECORD_LABEL
PERFORMER	PUBLICATION_YEAR
RATING	SPECIFIER
FROM_ALBUM	
Movies	
NAME	GENRES
CAST_MEMBER	PUBLICATION_DATE
PRODUCER	SCREENWRITER
DIRECTOR	CHARACTER_ROLE
RATING	MPA_RATING
SPECIFIER	
Athletes	
NAME	TEAM_POSITIONS
AWARDS	HEIGHT
NUM_AWARDS	PROFESSION_START_TIME
SPORTS_TEAMS	SPORT
STATS_LEADER	DATE_OF_BIRTH
PLACE_OF_BIRTH	RATING
SPECIFIER	
TV shows	
NAME	GENRES
CAST_MEMBER	CHARACTER
SCREENWRITER	AWARDS_RECEIVED
RATING	CREATOR
SPECIFIER	

Table 9: List of all possible slots for each domain.

actor), TV (shows), Sports (athletes), Music (musicians, songs), and Animals (cats, dogs, wild animals). The resulting fine-tuned Redpajama model is 8 seconds faster on average than GPT-3 at inference time.

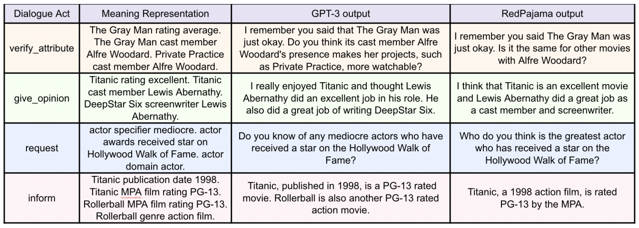
The human evaluation in Table 10 shows that the DA accuracy is lower, and there are more hallucinations with the fine-tuned RedPajama model than with GPT-3, Where most of the hallucinations are caused by longer and more complex meaning representations. Although the RedPajama model has superior BBLEU than GPT-3, the annotations from the PERF evaluation reveal a different story. While a higher BBLEU [35] score generally signifies better semantic accuracy, there's a threshold beyond which the increasing BBLEU [35] score contributes little to further improvements, effectively leading to a scenario of diminishing returns. While the RedPajama outputs do a good job in realizing the MRs for certain dialogue acts that do not have as complex hop relations, Figure 13 shows examples of certain dialogue acts from movies with more hop relations where RedPajama outputs do not perform as well in comparison to GPT-3. We may be able to address this limitation with further multi-hop targeted fine-tuning.

Model	**Domain**	**HAL**	**PERF**	**BBLEU**	**DA ACC**	**BBLEU-ANN**	**DA ACC-ANN**
RedPajama	**Music**	18.36%	78.57%	0.893	82.20%	0.784	89.20%
RedPajama	**Sport**	13.00%	82.00%	0.841	71.70%	0.738	82.50%
RedPajama	**TV**	14.00%	81.00%	0.898	90.20%	0.886	90.10%
RedPajama	**Movies**	6.93%	81.19%	0.898	88.10%	0.833	89.50%
RedPajama	**Animals**	13.13%	79.80%	0.820	83.60%	0.787	85.50%
GPT-3	**Music**	7.00%	93.00%	0.831	98.00%	0.778	100.00%
GPT-3	**Sport**	6.00%	94.00%	0.816	95.80%	0.763	97.90%
GPT-3	**TV**	19.00%	81.00%	0.797	97.60%	0.787	100.00%
GPT-3	**Movies**	5.00%	95.00%	0.753	94.50%	0.777	100.00%
GPT-3	**Animals**	3.00%	94.00%	0.677	94.30%	0.595	86.00%

Table 10: Annotation results for HAL (Hallucinations) and PERF (Semantic and DA accuracy) across different domains when comparing GPT-3 and RedPajama 3B. ANN is for annotation sample performance. BBLEU score is a metric that ranges from 0 to 1, and is not represented as a percentage.

Movies and Music KG. The Movies and Music KG relies on the LLM model to generate responses by leveraging dialogue acts that provide more dialogue variety in the responses for these domains. This is implemented by using a state graph that is not only based on the dialogue act from the user response but also where we provide our own specific dialogue acts at that particular state. By integrating dialogue acts such as opinion, inform, request, confirm, recommend, request attribute, verify attribute, request explanation, and suggestion DAs, we enable the system to engage users with diverse interactions. With each domain, we choose different possible slots as shown in Table 9 to be filled using Wikidata API; therefore, we can systematically choose more interesting slots that we can talk about within a conversation.

For instance, in the first state, if the user does not mention an entity related to movies or music, the system requests the user for a song or movie. If the user provides one, the system transitions to the ‘inquire likes entity‘ state and inquires if the user likes the entity so that we can further talk about that particular movie or song. Upon receiving a ’yes’ response to the inquiry, the state graph transitions to a give opinion DA, which then provides an opinion about the entity based on the user’s preference. If the user has a positive response to this opinion, we then transition to a new state to provide a similar movie as a suggestion or request or verify an attribute relating to the entity. If not, we use the recommend DA to recommend a similar song or movie to the user and loop back to give an opinion or provide information about the entity based on the user’s liking of the recommendation. In this way, we utilize each of the DAs listed above based on the user response to offer more relevant and natural conversations.



Dialogue Act	Meaning Representation	GPT-3 output	RedPajama output
verify_attribute	The Gray Man rating average. The Gray Man cast member Alfre Woodard. Private Practice cast member Alfre Woodard.	I remember you said that The Gray Man was just okay. Do you think its cast member Alfre Woodard's presence makes her projects, such as Private Practice, more watchable?	I remember you said The Gray Man was just okay. Is it the same for other movies with Alfre Woodard?
give_opinion	Titanic rating excellent. Titanic cast member Lewis Abernathy. DeepStar Six screenwriter Lewis Abernathy.	I really enjoyed Titanic and thought Lewis Abernathy did an excellent job in his role. He also did a great job of writing DeepStar Six.	I think that Titanic is an excellent movie and Lewis Abernathy did a great job as a cast member and screenwriter.
request	actor specifier mediocre. actor awards received star on Hollywood Walk of Fame. actor domain actor.	Do you know of any mediocre actors who have received a star on the Hollywood Walk of Fame?	Who do you think is the greatest actor who has received a star on the Hollywood Walk of Fame?
inform	Titanic publication date 1998. Titanic MPA film rating PG-13. Rollerball MPA film rating PG-13. Rollerball genre action film.	Titanic, published in 1998, is a PG-13 rated movie. Rollerball is also another PG-13 rated action movie.	Titanic, a 1998 action film, is rated PG-13 by the MPA.

Figure 13: Examples from DAs that have more complex meaning representations (MRs). The examples show how RedPajama outputs sometimes fail to realize the multi-hop relations when compared to GPT-3.

In order to measure the impact of the changes in the performance for the movies and music KG, we plotted the z-scores for both before and after the code freeze as shown in Figure 15. The incorporation of the fine-tuned model shows an increase in the z-score by 0.1 for the movies and music KG.

Domain	**NUM Hops**	**Self-trained**	**HAL**	**PERF**
Music	*One hop*	no	2%	97%
Music	*One hop*	yes	0%	100%
Music	*Multi-hops*	no	21%	89%
Music	*Multi-hops*	yes	26%	89%
Movies	*One hop*	no	2%	94%
Movies	*One hop*	yes	1%	99%
Movies	*Multi-hops*	no	4%	77%
Movies	*Multi-hops*	yes	3%	92%

Figure 14: Annotation results for HAL (Hallucinations) and PERF (Semantic and DA accuracy) before and after self-training.

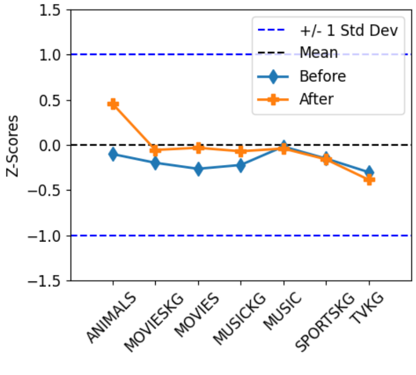


Figure 15: Flow Performance before (6/17/2023 - 6/24/2023) and after (7/24/2023-7/31/2023) the flow changes.

Human Evaluation. For human evaluation, five expert annotators evaluated 100 samples of outputs from Movies and Music before and after self-training. For the sampled outputs, 11 random samples from each DA were picked. Since the ’request_attribute’ DA only has 7-8 samples in the whole set for movies and music, we randomly sampled from the other DAs to create 100 total samples for annotation. The annotations were used to evaluate for perfect semantic and DA accuracy (PERF) and both implicit and explicit hallucinations (HAL). Figure 14 shows the results from the annotation.

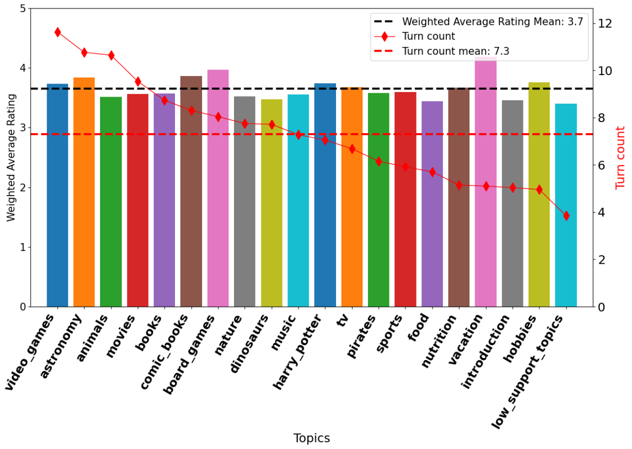


Figure 16: Weighted average rating and turn count per topic from 7/24/2023 to 8/5/2023.

Sports, TV, and Animals KG. As previously discussed, both SportsKG and TVKG leverage template responses and extract information from the WikiData API. We substitute one of the flow's functions with the "give opinion" and "inform" DA utilizing a fine-tuned model. This fine-tuned model has been extensively trained on a diverse dataset, encompassing various conversational scenarios, which equips it to generate more natural and human-like responses compared to the standard template-based approach. This choice was made based on the observation that these dialogue acts serve similar purposes within a typical conversation.

In terms of our animal-related content, we include pictures of specific animals being discussed with the user. This is aimed at enhancing our multimodal capabilities. In addition to this, Animals KG also has a similar implementation for using the fine-tuned RedPajama model as the TV and Sport KG to improve its output variety. The improvement in animal-related discussions is substantial before and after the inclusion of relevant animal images, as evidenced by the significant improvement in Z-Score (Figure 15).

Additional changes. While we only added 'give_opinion' and 'inform' DA call flows to the TV, Sports, and Animals KG into the system, we worked on expanding the data needed for the other DAs offline in order to be able to implement the complete call flow as we did for movies and music KG. In addition to generating the data for all the DAs for TV, Sports, and Animals, we also utilized the data for other entities that are a part of movie and music domains, such as actor and musician entities, to further fine-tune the RedPajama model to incorporate the data from all the topics.

6 Evaluation Analysis and Results

Evaluation of interactive dialogue is extremely challenging [33, 32, 30]. In the context of Athena, Amazon asks users for their ratings after every conversation on a scale of 1 . . . 5. It is obvious that a user's interactions with different response generators (RGs) and topics affect their views of the system and, therefore, their ratings. However, ratings are only collected at the end of the dialogue, and only a small portion of users actually provide dialogue ratings.

We calculate the performance of Athena topics by assuming that if a topic is discussed for at least three turns in a dialogue, then its pool of ratings includes the rating for that dialogue weighted by a weighting factor within that topic. We experimented with various weighting factors and used the square root of turn counts. Using this simplifying assumption, the weighted averages and turn counts for each topic are shown in Figure 16. Topics like video games, astronomy, animals, movies, and books have the highest turn counts. They are topics that users, in general, find interesting and fun to talk about. Topics like Harry Potter, Vacations, and Hobbies, have turn counts below the average but with average user ratings higher than the average.

References

[1] Ali Ahmadvand, Harshita Sahijwani, Jason Ingyu Choi, and Eugene Agichtein. Concet: Entity-aware topic classification for open-domain conversational agents. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1371–1380, 2019.

[2] Kevin K Bowden and Marilyn Walker. Let's get personal: Personal questions improve socialbot performance in the alexa prize. *The 13th International Workshop on Spoken Dialogue Systems Technology*, 2023.

[3] Susan E. Brennan, Marilyn Walker Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proc. 25th Annual Meeting of the ACL, Stanford*, pages 155–162, 1987.

[4] 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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[5] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023. `https://github.com/togethercomputer/RedPajama-Data`.

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

[7] Vrindavan Harrison, Rishi Rajasekaran, and Marilyn Walker. A transformer-based response evaluator for open-domain spoken conversation. *The 13th International Workshop on Spoken Dialogue Systems Technology*, 2023.

[8] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543, 2021.

[9] Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. Improving taxonomy of errors in chat-oriented dialogue systems. In *9th International Workshop on Spoken Dialogue System Technology*, page 331. Springer, 2019.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[11] 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. Advancing open domain dialog: The fifth alexa prize socialbot grand challenge. In *Alexa Prize SocialBot Grand Challenge 5 Proceedings*, 2023. `https://www.amazon.science/publications/advancing-open-domain-dialog-the-fifth-alexa-prize-socialbot-grand-challenge`.

[12] Dan Jurafsky. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, 1997.

[13] Juraj Juraska, Kevin K Bowden, Lena Reed, Vrindavan Harrison, Wen Cui, Omkar Patil, Rishi Rajasekaran, Angela Ramirez, Cecilia Li, Eduardo Zamora, et al. Athena 2.0: Contextualized dialogue management for an alexa prize socialbot. *arXiv preprint arXiv:2111.02519*, 2021.

[14] Juraj Juraska, Kevin K Bowden, and Marilyn Walker. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, 2019.

[15] Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*, 2018.

[16] Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G Okuno. User modeling in spoken dialogue systems to generate flexible guidance. *User Modeling and User-Adapted Interaction*, 15(1):169–183, 2005.

[17] Cat P Le, Luke Dai, Michael Johnston, Yang Liu, Marilyn Walker, and Reza Ghanadan. Improving open-domain dialogue evaluation with a causal inference model. *The 13th International Workshop on Spoken Dialogue Systems Technology*, 2023.

[18] Belinda Z Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. Efficient one-pass end-to-end entity linking for questions. *arXiv preprint arXiv:2010.02413*, 2020.

[19] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016.

[20] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65, 2017.

[21] Shikib Mehri and Maxine Eskenazi. Usr: An unsupervised and reference free evaluation metric for dialog. In *Annual Meeting of the Association for Computational Linguistics, ACL20*, 2020.

[22] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8, 2011.

[23] Kevin K. Bowden Juraj Juraska Wen Cui Vrindavan Harrison Rishi Rajasekaran Angela Ramirez Cecilia Li Eduardo Zamora Phillip Lee Jeshwanth Bheemanpally Rohan Pandey Adwait Ratnaparkhi Omkar Patil, Lena Reed and Marilyn Walker. Athena 2.0: Discourse and user modeling in open domain dialogue. In *Proc. of the Alexa Prize SocialBot Grand Challenge 4*, 2021.

[24] Daniel O'Sullivan. Using an adaptive voice user interface to gain efficiencies in automated calls. *White Paper, Interactive Digital, Smithtown, USA*, 2009.

[25] Angela Ramirez, Kartik Aggarwal, Juraj Juraska, Utkarsh Garg, and Marilyn Walker. Controlling dialogue acts with few-shot response generation and ranking for dialogue systems. In *Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL*, 2023.

[26] Gustavo Santana. Magicprompt - stable diffusion, 2022.

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

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

[29] Paul Taylor and Amy Isard. Ssml: A speech synthesis markup language. *Speech Communication*, 1(21):123–133, 1997.

[30] M. Walker, A. Rudnicky, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, R. Prasad, S. Roukos, G. Sanders, S. Seneff, and D. Stallard. DARPA communicator evaluation: Progress from 2000 to 2001. In *ICSLP*, 2002.

[31] Marilyn A. Walker, Masayo Iida, and Sharon Cote. Centering in japanese discourse. In *COLING90: Proc. 13th International Conference on Computational Linguistics, Helsinki*, pages 1–8, 1990.

[32] Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.

[33] Marilyn A. Walker, Diane Litman, Candace Kamm, and Alicia Abella. Evaluating interactive dialogue systems: Extending component evaluation to integrated system evaluation. In *Proc. of the ACL/EACL Workshop on Interactive Spoken Dialogue Systems*, 1997.

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[35] John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy, July 2019. Association for Computational Linguistics.

[36] Sanghyun Yi, Rahul Goel, Chandra Khatri, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *arXiv preprint arXiv:1904.13015*, 2019.

[37] Dian Yu and Zhou Yu. Midas: A dialog act annotation scheme for open domain human machine spoken conversations, 2019.