Hello World! I am Charlie, an Artificially Intelligent Conference Panelist

Patrick Cummings
Aptima, Inc.
pcummings@aptima.com

Ryan Mullins
Aptima, Inc.
rmullins@aptima.com

Manuel Moquete
Aptima, Inc.
mmoquete@aptima.com

Nathan Schurr
Aptima, Inc.
nschurr@aptima.com

Abstract

In recent years, advances in artificial intelligence (AI) have far outpaced our ability to understand and leverage them. In no domain has this been more true than in conversational agents (CAs). Transformer-based generative language models, such as GPT-2, significantly advance CAs' ability to generate creative and relevant content. It is critical to start exploring collaboration with these CAs. In this paper, we focus on an initial step by enabling a human-augmented, AI-driven CA to contribute to a panel discussion. Key questions include training a transformer-based AI to talk like a panelist, effectively embodying the CA to interact with panel participants, and defining the operational requirements and challenges to a CA gaining acceptance from its peers. Our results highlight the benefits that varied training, equal and dynamic representation, and fluid operation can have for AI applications. While acknowledging limitations, we present a path forward to richer, more natural human-AI collaboration.

1. Introduction

With the release of OpenAI's generative pre-trained transformer (GPT-2) [1, 2], it became apparent that generative language models had reached a level of sophistication such that their outputs were believable by a general audience. This was demonstrated in the short stories released by the OpenAI team [3], and validated by others who created scripts [4], poetry [5], and other narratives [6]. What had not been demonstrated is the performance — believability, relevance, response time — of these models in real time.

Conversational agents (CAs) are a common application domain for Natural Language Processing (NLP) technologies. Models like GPT-2 could expand the use of CAs from knowledge retrieval [7] to idea generation. Two key challenges must be overcome in building a CA for generative tasks. First is the challenge of interacting with these artificial intelligence (AI) models. CAs often contribute to activities that happen in real time, are physically anchored, and involve iterations with humans. Second is the challenge of transforming expectations of a CA's capabilities. Society expects transactional interactions with CAs, but generation necessitates collaboration and iteration. We must also be aware of the limitations of deep learning AI, such as GPT-2. These models are not guaranteed to deal in facts [8]. Instead, they should be used when *conjecture* is an acceptable currency in discourse.

In this work, we set an ambitious goal of building a generative CA that would participate as a human-augmented AI panelist in a discussion of "AI-empowered learning" as part of the 2019 Interservice/Industry Simulation, Training, and Education Conference (I/ITSEC). Conference panels are a prime venue for conjecture, offering a creative, improvisational environment for ideation where an AI-powered CA can thrive, without strict requirements on feasibility or veracity. An AI panelist in this environment must effectively address these challenges while working within the technological limitations, or risk entering the "uncanny valley" that leads to subverted expectations and rejection [9]. This work explored three research questions:

> *RQ1: Can a generative language model be trained to talk like a conference panelist?*
>
> *RQ2: What embodiment and interfaces are required to enable effective contribution to a panel by an AI-driven CA?*
>
> *RQ3: What mechanics and interfaces are required to ensure successful operation of the AI-driven CA during the panel?*

2. Related Work

2.1. Generative Language Models

Recent, extreme strides have been made in NLP models by using the Transformer architecture [10].

URI: https://hdl.handle.net/10125/70656
978-0-9981331-4-0

Transformers ensure that predicted tokens (or words) are realistic using a mechanism called self-attention, which uses input sequence tokens to improve predictions. Other methods, such as Recurrent Neural Networks [11] and Long Short-Term Memory [12], use hidden state to approximate self-attention by incorporating all previously processed tokens into the current prediction. Self-attention does this at a larger scale by drawing attention to each of the previous tokens separately, as opposed to one compilation of them all.

Transformers have produced state-of-the-art results on a number of deep learning tasks such as translation [13]. Language generation, particularly, has made recent strides with the integration of transformers, as shown with BERT, GPT-2, and Turing-NLG [14]. In these model architectures, transformers have shown above human-level expertise on the GLUE benchmark based on their bilingual evaluation understudy (BLEU) scores [15]. The BERT transformer achieves better BLEU scores than the previous state-of-the-art models on the English-to-German and English-to-French news tests at a fraction of the training cost [10].

2.2. Human-AI Interaction

CAs have a long history, from early chatbots [16] to modern personal assistants [17]. CAs are designed to use natural language as their primary interface [18], with the nature of their use influencing the design of their underlying AI and embodiment. Here, we review past work in three use cases applicable to CA performance.

Question-Answering. CAs were designed to enable information retrieval (IR). Chatbots rely on text-based interfaces [16], which have been scalable for a variety of service-oriented tasks [19]. Advances in context modeling [20] have improved the accuracy and precision of chatbot IR by enabling iterative refinements of a shared context space [21]. However, chatbots still struggle for acceptance [18], correlated with perceptions of social presence and humanness [22]. Personal assistants — exemplified by Amazon's Alexa and Apple's Siri — expand on the foundations of chatbots to include automated task completion and speech-based interfaces [7]. Regardless of interaction modality, research has shown positive correlations between response delay [23] and dynamism [24] in human acceptance of CAs. These are important insights for a panelist, signaling that the audience will be accepting of variance in the response time should the agent provide sufficient embodiment of its internal state.

Debate. CAs have started to appear in debates against humans. Debate is a highly structured discourse where opponents must recognize an argument [25] and present an optimal counter-argument in order to "win" [26]. CA debaters rely on corpora of existing arguments [27], which they assess in real time for quality [28] and relevance [29], to perform in these conditions. Although panel discussions are far less formal than debates, debaters provide an important lesson: the audience will judge the panelist on how it interacts with its human compatriots. Effective performance depends on its ability to build on and branch off of others' statements.

Ideation. Panels gather individuals from varied backgrounds to deliberate potential innovations, thus engaging in the first step of ideation [30]. For humans, ideation often relies on synthesis from two or more existing concepts, as shown in product development [31] and information analysis [32]. To date, measures of synthesis are limited and difficult to complete in real time [33]. As such, the use of CAs in ideation has been restricted to supporting roles, such as that of a moderator [34], facilitator [35], or planner [36]. However, there is an open question as to how plausible an idea needs to be in order to be a useful contribution to ideation. For example, if the discussion merely needs enough entropy to prevent premature closure [37], a generative language model could effectively contribute.

3. Methods

Charlie was designed and developed as an action research program [38] that used guerilla usability testing [39], agile software development practices [40], design thinking [41], and rapid prototyping [42] methods. The goal of this program was to develop a CA that could effectively participate as a conference panelist, while characterizing the limitations and future research needs.

We conducted 10 guerilla usability tests. Tests 1-6 assessed utterance believability (Section 5.1) and embodiment effectiveness (Section 5.2) in a 30-minute discussion between Charlie, a moderator, and two reviewer-panelists. Two observers recorded notes during these tests. Discussion transcripts were recorded in addition to the observer notes. Following the discussion, the observers conducted a semi-structured interview with the participants to document Charlie's failure modes and ideas for improvement. Tests 7-10 were conducted as a formal panel discussion between Charlie, a moderator, and four reviewer-panelists. Data were recorded in the same way as the prior tests, although the emphasis pivoted to performance and operation in a real time (see Section 5.3). Feedback from these tests was translated into tickets and addressed during subsequent development activities.

Major capability milestones were marked with two demonstrations equivalent to a dry-run of a panel

in front of a live audience — one with panelists from our organization and one with Charlie's I/ITSEC co-panelists. Qualitative measures of engagement were collected, along with feedback on utterance quality and effectiveness of embodiment.

4. Building an AI panelist

Charlie is a human-augmented, AI-driven CA designed to participate in real-time panel discussions. Given the capabilities of the other *panelists*, Charlie was expected to provide the following capabilities.

1. Charlie must be able to take input from the *moderator* and/or other *panelists* as speech.
2. Charlie must be able to generate natural language responses to questions that account for responses from other *panelists*.
3. Charlie must vocalize the generated statements so that they can be heard by the audience.
4. Charlie must provide an embodiment on stage so that the audience and panel members are aware of her presence.
5. Charlie must provide some indication of her internal state such as readiness to speak.

Charlie consists of four components (Fig. 1). Two components reside on the Amazon Web Services (AWS) cloud infrastructure. Our trained model (Req. 2) runs on one or more Elastic Compute Cloud (EC2) nodes with high-performance GPU compute. Amazon's Polly service provides Charlie's text-to-speech capability (Req. 3). The remaining components run on a computer at the conference venue. The panelist interface provides Charlie's embodiment (Req. 4), including a visual representation of her state (Req. 5) and the outbound audio interface with the room's sound system (Req. 3). The operator interface enables human augmentation of Charlie during the discussion (Req. 1). The necessity, assumptions, design decisions, and evolution of the model, panelist interface, and operator interface are discussed respectively in Sections 5.1, 5.2, and 5.3.

5. Findings

5.1. Training an AI CA to be a panelist (*RQ1*)

Charlie was trained to speak about the future of learning and training as a like-minded panelist using 30 years of papers from I/ITSEC (a training and simulation conference) as source material. We hypothesized that the abstracts would provide the most useful material for a panel discussion as they avoid the technical details

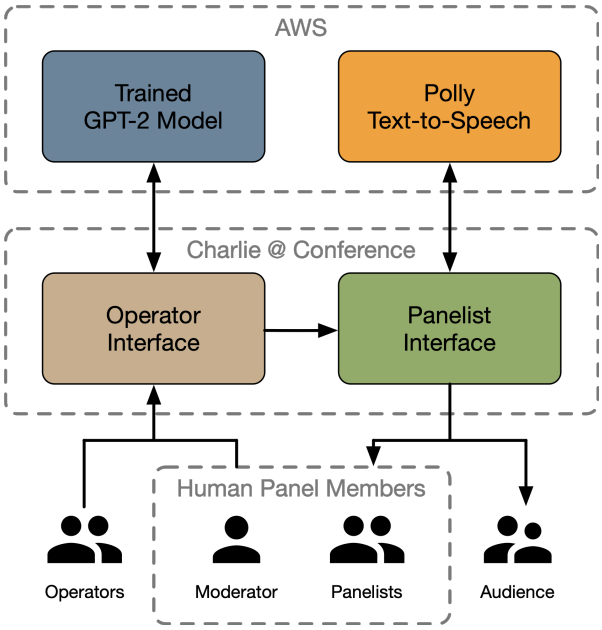

Figure 1. Charlie system architecture

and give a brief overview of the content — qualities one would expect from a panelist. We fine-tuned GPT-2 (the 117M, 345M, and 774M models) with a corpus of 3,121 I/ITSEC abstracts (approximately 5.6 MB of text). We concatenated the abstracts to prepare for training, adding <|endoftext|> between abstracts as done in the GPT-2 training corpus. We also removed items within parentheses to eliminate citations from the training data. Although citations were not the only text contained within parentheses, most other items were clarifications or interjections and therefore not appropriate for a panel, where flow is of utmost importance.

The fine-tuned I/ITSEC model was trained using hyperparameters [10]. Specifically we used the loss function equivalent to that of GPT-2–that is, we optimized for predicting the next word in a text corpus. For training we used the Adam optimizer with learning rate 2×10^{-5}, batch size equal to 1, and only trained the transformer layers. The loss function converged after approximately $3,000$ iterations.

For generating text, we used a series of AWS EC2 instances running in parallel to rapidly provide potential responses. Upon each request of these models, the text generation with temperature $= 1$ and $\text{top}_k = 0$ produced text until a full sentence was generated. We used an unsupervised model for sentence boundary detection in order to avoid premature closure from mid-sentence punctuation [43]. As input to the generation, we used the previous panel text with the following structure.

HOST: Text from host...

PANELIST: Text from panelist...

HOST: Text from host...

PANELIST: Text from panelist...

This structure had three key advantages: (1) it removed the requirement for an operator to start a sentence for Charlie by delineating when others were done speaking and she should start; (2) it kept Charlie on her own thread while integrating context from others; and (3) it allowed Charlie to react to the discussion above rather than simply continuing her thought.

Content was filtered using a black list containing frequent tokens in the training corpus that are well-suited for publications, but not for panel discourse, such as 'in this paper', 'Figure', 'Name:', or symbols (*, -, [], (), etc.), as well as a dictionary of inappropriate and/or foul language. If those tokens were generated, the process deleted them and started the text just after. The generation of <|endoftext|> or HOST: were also treated in this manner, but in future work can be used to detect the end of Charlie's statement.

Over the course of the panel, Charlie generated 2,098 responses to 268 prompts, of which 42 were selected by the human operator as potentially useful. Of the responses selected, 23 (54.8%) were from the 1.5B parameter base GPT-2 model and 19 (45.2%) were from the fine-tuned I/ITSEC model. Those responses were then gathered into 10 total utterances. Fig. 2 shows the distribution of each utterance split by sentence and colored by model. Generally, Charlie stuck with one model for the full course of the utterance but she occasionally pivoted between the two.

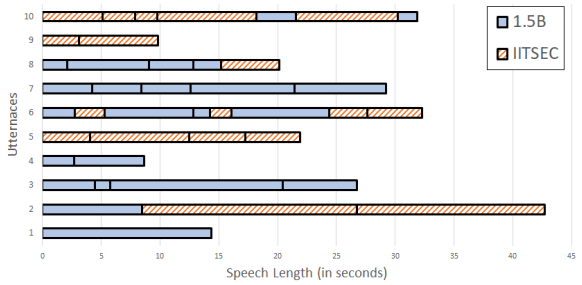

Figure 2. Charlie's answers distributed across the two models

We hypothesized that the base model would be particularly well suited for generating general text and that the fine-tuned I/ITSEC model would be well suited to providing insights about artificial intelligence and training. However, we found that in the responses selected, both models generated a mix of the two. For example, in the below quote from Charlie in the panel, the model that generated each sentence is included at the end. This response is a mix of both models as well as both types of sentences.

> Prompt: How can we help build trust in artificial intelligence systems?
>
> Utterance: *This is a difficult question to answer (GPT-2). I think trust will evolve over time (I/ITSEC). Trust is something that you build from within and therefore artificial intelligence can be of benefit to you (GPT-2). It's a long ways off (GPT-2). It won't happen overnight (I/ITSEC). But we must build trust, and we must create systems that help people understand each other and communicate a bit better (GPT-2). Just get over the fear of machines taking over (I/ITSEC). And this means building trust in realistic, credible conversations (I/ITSEC).*

5.2. Embodying an AI panelist (*RQ2*)

Charlie was expected to contribute to the discussion as an equal. Thus, the visual display took up approximately the same space on stage, the speech flowed through the same sound system, and the non-verbal communication was equally visible to the audience. Human panelists were seated in a row of chairs on stage, and Charlie's embodiment (Req. 4) was constrained to a similar style and space. A single 32-inch display was mounted on a chair to provide the visual component of Charlie's embodiment (Fig. 3). The sound from the computer driving the display was connected to the room's mixing board, as were the microphones for each human panelist.

Figure 3. The panel on stage at I/ITSEC 2019.

The *panelist interface* (i.e., the embodiment) required 3 iterations to refine state communication (Req. 5), representation (Req. 4), and stage presence (Req. 4) driven by feedback from guerilla usability evaluations. From chatbots, we expected that response delays would be acceptable [23], especially in response to other panelists [24], as long as Charlie's state was clearly communicated. Humans use physical and audible queues — gestures, changes in eye contact,

and transitional phrases — to indicate their state and control the flow of a conversation [44]. Charlie had to effectively coordinate the use of the display and audio to achieve a similar presence and represent its states as follows:

- *Idle*: Charlie is listening.
- *Thinking*: Charlie is generating a statement.
- *Interjection*: Charlie has something to say!
- *Speaking*: Charlie is speaking.

The first iteration of the *panelist interface* supported 3 of the 4 states. When *idle*, the display was a blank, white canvas. *Thinking* added a yellow border (Fig. 4.A). When *speaking*, the border changed to green and the text of the speech filled the central white area (Fig. 4.B). Reviewers felt that this design fell short for a number of reasons. First, displaying the text on screen was distracting and would not be expected of human panelists. Second, the borders did not effectively communicate Charlie's state, lacking an intuitive mapping of the colors and being indistinguishable at a distance. Finally, the blank display did not convey that Charlie was listening or present.

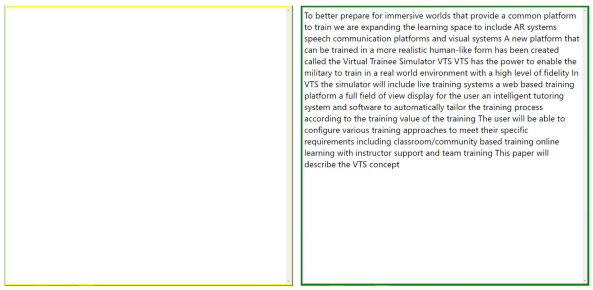

Figure 4. Version 1 of the panelist interface with states for thinking (A, yellow border) and speaking (B, green border with text).

The second iteration took a very different form, providing a highly mimetic and pictorial representation of Charlie's state using stylized video of a human acting out hand gestures for the *idle* (Fig. 5.A), *thinking* (Fig. 5.B), and *speaking* (Fig. 5.C) states. Although this iteration addressed the issues of presence and distinctive state representation, reviewers noted that it introduced additional problems akin to the "uncanny valley" [9]. Seeing a human on screen would be a misrepresentation of what Charlie is; instead of an AI-driven CA, audiences would think that a human was generating these statements.

Charlie's third *panelist interface* iteration moved to an abstract, geometric representation, akin to that of

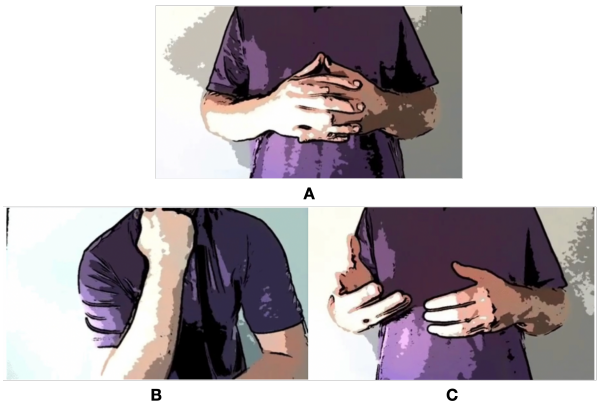

Figure 5. Version 2 of the panelist interface used videos of hand gestures to show the idle (A), thinking (B), and speaking (C) states.

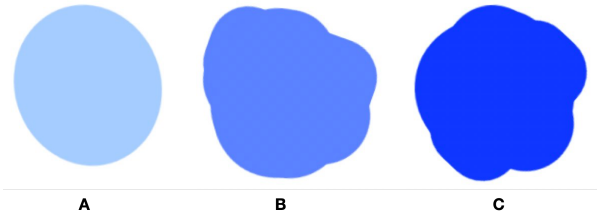

Figure 6. Version 3 of the panelist interface using shape and color to show the idle (A), thinking (B), and speaking (C) states.

other CAs [7]. When *idle*, a light blue ellipse rotated constantly (Fig. 6.A). The *thinking* and *speaking* states added distortions around the edge of the ellipse while it rotated, and increased color saturation (Fig. 6.B/C). Reviewers noted that the use of geometric shapes with smooth transitions eliminated the "uncanny valley" problem, improved presence, and reduced distraction. However, the changes in color and texture were too subtle to quickly interpret at a distance, a known challenge with animated symbology [45].

The final iteration refined the geometric approach. Size, color saturation, and texture were the primary visual variables used to distinguish between states. The size and color saturation increased from *idle* to *thinking* to *speaking* (Figs. 7.A-C). An animated border texture, which effectively "chased its own tail," was used to emphasize *thinking* as an ongoing process (Fig. 7.B). This animation led to a natural representation of the *interjection* state: a complete border around the circle, signaling the end of the thinking process, and a color change from blue to yellow, which draws attention and indicates urgency (Fig. 7.D). The *speaking* state was also animated to indicate an ongoing process, with the circle changing size while Charlie spoke.

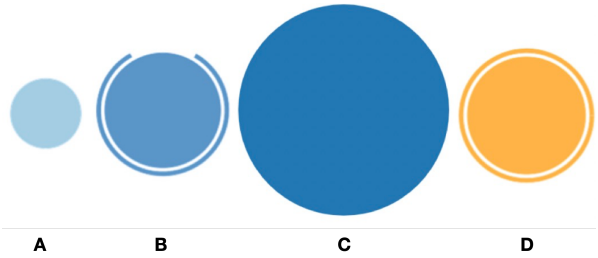

Figure 7. The final panelist interface combined size, texture, color, and animation to illustrate all four states: idle (A), thinking (B), speaking (C), and interjection (D).

Reviewers found that these changes effectively enabled state inference, provided stage presence, and supported dynamic evolution of Charlie's response process. For example, when posed with a direct question, Charlie could fluidly transition from *idle* to *thinking* to *speaking*, or, when listening to another panelist, Charlie could move immediately from *idle* to *interjection* if an appropriate statement was generated.

5.3. Operating an AI panelist (*RQ3*)

The nature of a conference environment and the nuances of GPT-2 posed significant challenges to automating Charlie's operation: bandwidth and interface limitations precluded the use of speech-to-text tools; GPT-2 is vulnerable to spelling errors; and multiple, different model instances were needed to respond in a timely fashion with relevant content. These challenges necessitated that a human operator augment Charlie by performing the following tasks.

- Transcribing speech to text
- Aggregating statements into an utterance
- Coordinating Charlie's state transitions

The *operator interface* went through several revisions. The first version (Fig. 8.A) was a command-line interface to the AWS EC2 instances running the models. This was an efficient way to test the models, but was inconvenient and minimally usable for transcription and aggregation tasks, and lacked support for controlling state transitions entirely.

Starting with Version 2 (Fig. 8.B), the *operator interface* became a web application that emphasized efficiency in aggregating statements into utterances. A history of utterances was shown at the top, with the "send" button separating the utterance aggregation text field below. Candidate statements were shown in a pop-up. Clicking on a statement added it to the

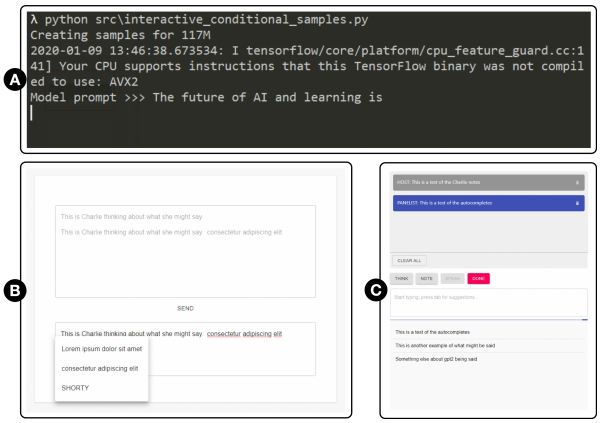

Figure 8. The evolution of the operator interface, from command-line (A) to a web app enabling utterance construction (B) and state transition (C).

current utterance text field, and pressing tab refreshed the candidate statements. When the operator clicked send, the utterance was added to the history and sent to the *panelist interface* for vocalization. All items in the history were sent to the models as context every time candidate statements were refreshed. During capability demonstrations, operators noted that it was hard to distinguish between what the *moderator* and *panelists* were saying, and the lack of direct state controls.

Version 3 (Fig. 8.C) added direct control over state transitions via buttons. Clicking these buttons sends a message to the *panelist interface*, which processes messages as a queue. The other key addition was the "note" button, which added the contents of the text field to the history as a note when clicked, allowing the system to capture and distinguish content from the *moderator* and other *panelists* (grey items) and content from Charlie (blue items) for the first time.

The final iteration represents a major shift in workflow. During demonstrations, it became clear that a single operator could not manage transcription, state transition, and utterance aggregation without inducing unacceptably long response delays [23]. Thus, the interface was divided into two areas (Fig. 9): the conversation history area on the left and the statement review area on the right, with one operator assigned to each area. While *idle*, the conversation history operator's primary job is to transcribe the discussion into the text area and store it in the history. The statement review area operator's job is to retrieve, assess, and save candidate statements for future utterances, and to tell the other operator when to stop so they can initiate a state transition. Candidate statements are displayed in a list of 8 numbered items: 4 from the GPT-2 model and 4 from the I/ITSEC model. A statement's color saturation

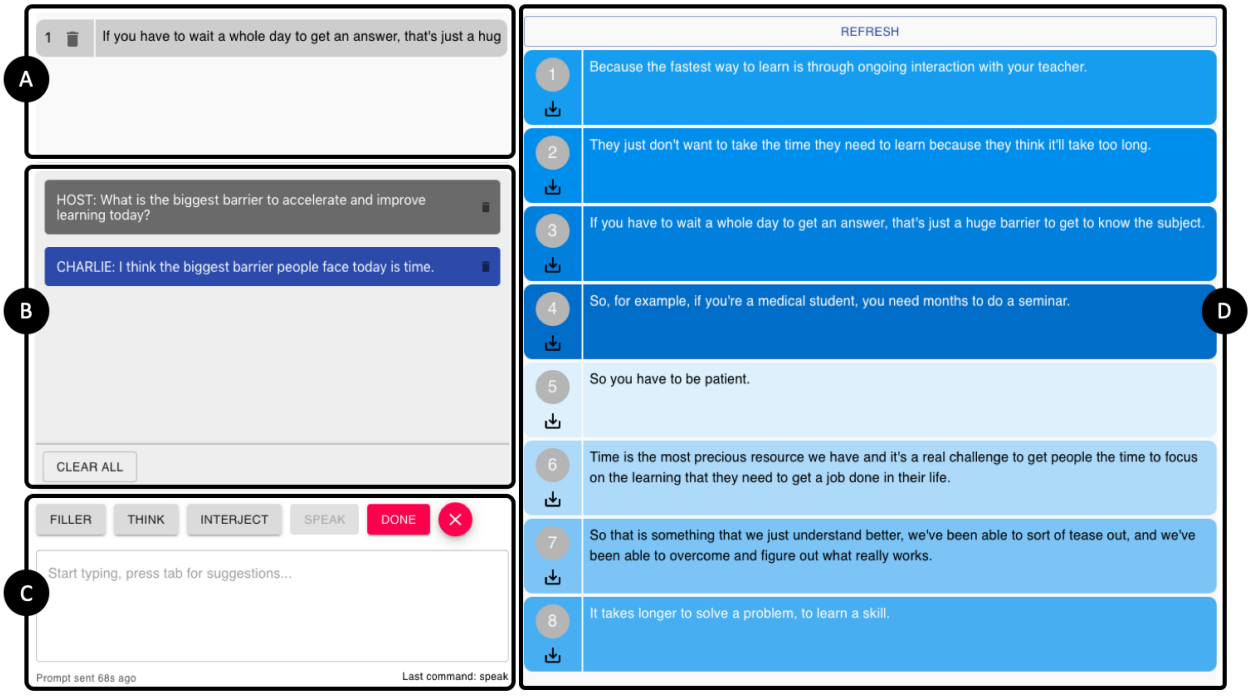

Figure 9. The final operator interface with the (A) saved statements, (B) conversation history, and (C) utterance construction components on the left, and the (D) statement review area on the right.

indicates its recency, with newer candidates being more saturated. A button under the statement number allows the operator to save candidates for future use.

If one statement stands alone as relevant, the operators can transition to the *interjection* state: click on that statement in the review area; which loads it in the text area; and await recognition from the *moderator* to speak. If there are multiple relevant statements, operators can transition to the *thinking* state and begin to collaboratively aggregate statements into utterances from the review area on the right (Fig. 9.D) or the "saved statement" area in the top-left (Fig. 9.A). Clicking on a statement appends it to the text area content and generates new candidate statements. Operators can continue to add statements to the utterance until they are satisfied, at which point they click "speak" to send the utterance to the *panelist interface*, and then the "done" button to transition back to the *idle* state.

6. Results and discussion

In practice, Charlie successfully participated in a 95-minute panel session in front of more than 300 audience members. Charlie responded with relevant content to a number of questions from the moderator, other panelists, and the audience. Fig. 10 shows Charlie's state progression throughout the panel, and Table 1 breaks down the total time (in minutes) and the average duration (in seconds) in each state. There are some striking observations we can gather from the flow of the discussion captured in the timeline (Fig. 10). We clearly see a cyclical pattern of states. Between long periods of *idle*, Charlie enters a *thinking* state followed by *interject* and finally *speaking*. This process is as expected for any panelist, human or AI. Below we highlight three key patterns.

1. The majority of Charlie's interactions follow the *idle-thinking-interject-speaking* pattern. A question is posed by the moderator; Charlie generates a response; Charlie indicates that she is ready to speak; upon a break in conversation or as directed by the moderator, Charlie speaks.

2. From minutes 21-24 and 55-60, Charlie remained in the *interject* state before speaking. These long periods of waiting to interject are in line with human panelists who also frequently want to jump in but must also avoid speaking over others.

3. At minute 79, during the audience question portion of the panel, Charlie entered the *thinking* and *interject* states but did not speak afterwards. In this case, although Charlie had something relevant to say, the conversation quickly moved on to the next topic before she could respond.

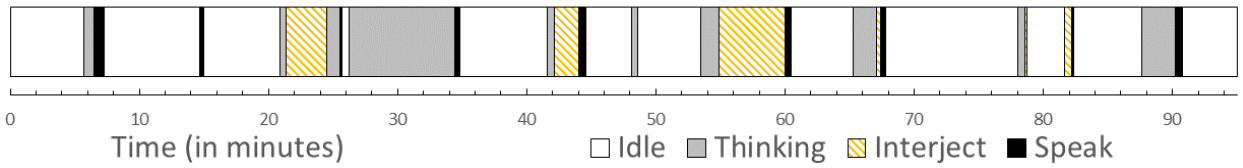

Figure 10. Changes in Charlie's state over the course of the panel discussion.

Per [23, 24], Charlie's ability to demonstrate structural parity in conversational patterns, and her ability to respond dynamically to discursive conditions are essential for effective performance and acceptance.

Table 1. Charlie's state progression by the numbers.

State	Total Time *(min)*	Avg. Duration *(s)*
Idle	61 (65%)	N/A
Speaking	5 (5%)	32
Thinking	18 (18%)	106
Interject	11 (12%)	112

6.1. Limitations

Although future iterations of Charlie are expected to reduce the necessary human augmentation, the current need for human involvement in operating Charlie is a significant limitation. A human is required to (1) transcribe the conversation, (2) select responses for Charlie to speak, and (3) conduct Charlie state changes.

Human transcription was a venue limitation; our inability to access the room's soundboard and the limited internet bandwidth precluded the use of real-time transcription services. The technical aspect of this limitation has been addressed in our continued development of Charlie. That said, since Charlie uses previous discussion to generate her injects, Charlie is sensitive to typographical errors in transcription. Particularly, as Charlie sees errors or nonsense, she will begin to generate typos and nonsense herself. Mitigation of these errors will be key when integrating automated transcription services, at the risk of non-acceptance by her compatriots and the audience [18, 22].

Response selection is an open research question. We expect that the pace of GPU advancement will overcome the time-bounded technical aspects of this limitation. However, generative language models induce many challenges regarding relevance and adaptability. For example, Charlie can easily generate novel ideas using GPT-2's underlying randomness and leveraging training corpus construction, but this randomness can result in responses that do not progress the discussion. As we know from chatbots [18, 22] and debaters [26], constructive, on-topic feedback is essential for effective performance. Adaptability is another challenge. Because any new response is influenced by the entire prior discussion, Charlie naturally tends toward a static tone and context. Although this keeps her responses globally relevant, it limits her ability to engage in short-lived, tangential conversations. This is the underlying reason we used multiple models, but automatically selecting which model to use at a given time is an unsolved problem. Balancing the dynamics of her generative processes will be critical to generalize Charlie to other real-time discursive activities and increase her value in ideation.

State management is the final link in optimizing Charlie's performance. As discussed above, aspects of response time (a critical performance measure [23]) and relevance have been optimized via the design of the operator interface, training the human operator to optimize interface use, and using parallel GPU accelerated instances. State management remains a challenge because of the non-verbal cues so prevalent in human communication. Automating this process will require real-time content analysis, as well as sensing and analyzing of auditory and visual cues.

7. Conclusion

This paper presented findings from the development and operation of Charlie, a human-augmented AI panelist. We sought to address three questions related to the suitability of AI-driven CAs as peers: (1) can a generative language model talk like a panelist?, (2) what embodiment is required to enable effective contribution?, and (3) what mechanics are required to ensure successful operation during the panel?

We posit that, given an appropriate framework and constraints, generative language models do provide useful, novel inputs to discussion on a panel. When training the model, developers should recognize the benefits that varied training data have on speech patterns while reinforcing core content and understanding the potentially negative effects that typos and sidebar discussions can have on performance. Despite their present limitations, transformer-based AI shows considerable promise for backing CAs in panels,

debates, public speaking, and other domains.

Acceptance of AI, and CAs in particular, depends not only on their technical prowess, but on their embodiment and operation. Our work shows the effectiveness of equal positioning and differentiated representation. Although the representation of Charlie's state lacked nuance compared to humans, the use of size, color, texture, and animation was effective in communicating state and provided a stage presence with which panel members and the audience could engage. Further research is needed to refine the embodiment in collocated and distributed settings.

Operating an AI-driven CA — balancing the transcription, aggregation, and state management — proved non-trivial. Our research found that this was a two-person job: one transcribing and the other juggling aggregation and state management. Adding automated transcription could improve operation, so long as it does not induce spelling, grammatical, or other errors in Charlie's statements. Further, scoring generated statements for appropriateness and utility would assist in human-augmented workflows (e.g., aggregation) in the short term and in removing the human augmentation requirement in the long term.

7.1. Future Work

There are many potential paths forward for Charlie, ranging from moving her to other forms of discourse, to increasing her level of automation, to improving her generation capabilities. For the latter, some current work in improving content generation for generative language models has focused on evaluating generated responses to maximize information gain [46, 47]. Although this is one key to a "good" response in a panel discussion, filtering or ranking of responses also requires understanding the relevance and value of the response to the conversation. These metrics can be measured to rank responses for selection.

A first step toward increasing Charlie's autonomy is using speech-to-text services (STTs) to transcribe the ongoing discussion. Although current STTs show less than 10% error rate [48], this is still above the threshold of sensitivity for generative language models [49]. There may need to be accommodations for the editing STT-transcribed text on the fly, and certainly the use of automated spell checkers will be required.

Since appearing on the I/ITSEC panel, Charlie has been a podcast guest discussing the third wave of AI. She has also been invited to participate in other panels, to write a book chapter, and to give a joint human-AI keynote address. We expect that AI agents such as Charlie will continue to expand into new fields where creative generation and discourse is a valuable resource.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[2] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 06 2019.

[3] J. Vincent, "Openai's new multitalented ai writes, translates, and slanders," *The Verge. The Verge, February*, vol. 14, 2019.

[4] A. Destine-DeFreece, S. Handelsman, T. Light Rake, A. Merkel, and G. Moses, "Can gpt-2 replace a sex and the city writers' room?," 2019.

[5] G. Branwen, "Gpt-2 neural network poetry," 2019.

[6] L. Qin, A. Bosselut, A. Holtzman, C. Bhagavatula, E. Clark, and Y. Choi, "Counterfactual story reasoning and generation," *arXiv preprint arXiv:1909.04076*, 2019.

[7] V. Kepuska and G. Bohouta, "Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 99–103, IEEE, 2018.

[8] X. Peng, S. Li, S. Frazier, and M. Riedl, "Fine-tuning a transformer-based language model to avoid generating non-normative text," *arXiv preprint arXiv:2001.08764*, 2020.

[9] N. Schurr, A. Fouse, J. Freeman, and D. Serfaty, "Crossing the uncanny valley of human-system teaming," in *International Conference on Intelligent Human Systems Integration*, pp. 712–718, Springer, 2019.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 06 2017.

[11] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," *Arxiv*, pp. 1–11, 03 2015.

[12] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A c-lstm neural network for text classification," 11 2015.

[13] G. Tang, M. Müller, A. Rios, and R. Sennrich, "Why self-attention? a targeted evaluation of neural machine translation architectures," *arXiv preprint arXiv:1808.08946*, 2018.

[14] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *arXiv preprint arXiv:2003.08271*, 2020.

[15] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," 04 2018.

[16] M. L. Mauldin, "Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition," in *AAAI*, vol. 94, pp. 16–21, 1994.

[17] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, *et al.*, "Conversational ai: The science behind the alexa prize," *arXiv preprint arXiv:1801.03604*, 2018.

[18] R. Meyer von Wolff, S. Hobert, and M. Schumann, "How may i help you?–state of the art and open research questions for chatbots at the digital workplace," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[19] S. Feng and P. Buxmann, "My virtual colleague: A state-of-the-art analysis of conversational agents for the workplace," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[20] R. Mullins, A. Fouse, G. Ganberg, and N. Schurr, "Practice makes perfect: Lesson learned from five years of trial and error building context-aware systems," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[21] S. Reshmi and K. Balakrishnan, "Implementation of an inquisitive chatbot for database supported knowledge bases," *sādhanā*, vol. 41, no. 10, pp. 1173–1178, 2016.

[22] F. Hendriks, C. X. Ou, A. K. Amiri, and S. Bockting, "The power of computer-mediated communication theories in explaining the effect of chatbot introduction on user experience," *interaction*, vol. 12, p. 15, 2020.

[23] U. Gnewuch, S. Morana, M. Adam, and A. Maedche, "Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction," 2018.

[24] R. M. Schuetzler, M. Grimes, J. S. Giboney, and J. Buckman, "Facilitating natural conversational agent interactions: lessons from a deception experiment," 2014.

[25] J. Visser, J. Lawrence, J. H. Wagemans, C. Reed, *et al.*, "Revisiting computational models of argument schemes: Classification, annotation, comparison.," in *COMMA*, pp. 313–324, 2018.

[26] M. Sato, K. Yanai, T. Miyoshi, T. Yanase, M. Iwayama, Q. Sun, and Y. Niwa, "End-to-end argument generation system in debating," in *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pp. 109–114, 2015.

[27] S. Gretz, R. Friedman, E. Cohen-Karlik, A. Toledo, D. Lahav, R. Aharonov, and N. Slonim, "A large-scale dataset for argument quality ranking: Construction and analysis," in *Proceedings of the 34th AAAI conferennce on artificial intelligence*, 2020.

[28] L. Ein-Dor, E. Shnarch, L. Dankin, A. Halfon, B. Sznajder, A. Gera, C. Alzate, M. Gleize, L. Choshen, Y. Hou, *et al.*, "Corpus wide argument mining–a working solution," in *Proceedings of the 34th AAAI conferennce on artificial intelligence*, 2020.

[29] M. Hildebrandt, J. A. Q. Serna, Y. Ma, M. Ringsquandl, M. Joblin, and V. Tresp, "Reasoning on knowledge graphs with debate dynamics," in *Proceedings of the 34th AAAI conferennce on artificial intelligence*, 2020.

[30] D. Graham and T. T. Bachmann, *Ideation: The birth and death of ideas*. John Wiley & Sons, 2004.

[31] S. F. Slater, J. J. Mohr, and S. Sengupta, "Radical product innovation capability: Literature review, synthesis, and illustrative research propositions," *Journal of Product Innovation Management*, vol. 31, no. 3, pp. 552–566, 2014.

[32] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5, pp. 2–4, McLean, VA, USA, 2005.

[33] D. J. Zelik, E. S. Patterson, and D. D. Woods, "Judging sufficiency: How professional intelligence analysts assess analytical rigor," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 51, pp. 318–322, SAGE Publications Sage CA: Los Angeles, CA, 2007.

[34] T. Strohmann, S. Fischer, D. Siemon, F. Brachten, C. Lattemann, S. Robra-Bissantz, and S. Stieglitz, "Virtual moderation assistance: Creating design guidelines for virtual assistants supporting creative workshops.," in *PACIS*, p. 80, 2018.

[35] E. Bittner and O. Shoury, "Designing automated facilitation for design thinking: a chatbot for supporting teams in the empathy map method," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[36] Q. Li, M. A. Vasarhelyi, *et al.*, "Developing a cognitive assistant for the audit plan brainstorming session," 2018.

[37] J. B. Grossman, D. D. Woods, and E. S. Patterson, "Supporting the cognitive work of information analysis and synthesis: A study of the military intelligence domain," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 51, pp. 348–352, SAGE Publications Sage CA: Los Angeles, CA, 2007.

[38] R. L. Baskerville and A. T. Wood-Harper, "A critical perspective on action research as a method for information systems research," *Journal of information Technology*, vol. 11, no. 3, pp. 235–246, 1996.

[39] J. Nielsen, "Guerrilla hci: Using discount usability engineering to penetrate the intimidation barrier," *Cost-justifying usability*, pp. 245–272, 1994.

[40] M. Fowler, J. Highsmith, *et al.*, "The agile manifesto," *Software Development*, vol. 9, no. 8, pp. 28–35, 2001.

[41] S. Black, D. G. Gardner, J. L. Pierce, and R. Steers, "Design thinking," *Organizational Behavior*, 2019.

[42] L. Luqi and R. Steigerwald, "Rapid software prototyping," in *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, vol. 2, pp. 470–479, IEEE, 1992.

[43] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, pp. 485–525, 12 2006.

[44] K. R. Scherer, "The functions of nonverbal signs in conversation," in *The social and psychological contexts of language*, pp. 237–256, Psychology Press, 2013.

[45] M. Harrower, "The cognitive limits of animated maps," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 42, no. 4, pp. 349–357, 2007.

[46] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan, "Generating informative and diverse conversational responses via adversarial information maximization," 09 2018.

[47] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," 2019.

[48] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[49] D. Pruthi, B. Dhingra, and Z. Lipton, "Combating adversarial misspellings with robust word recognition," pp. 5582–5591, 01 2019.