## Conversations with and between personas using language models

[Ken Kahn](https://docs.google.com/document/d/1r7oaUOGs1TGjtKYo1CCmfqTifH5cqpPny5ALx5EIT_8/edit) <toontalk@gmail.com>

### Background

In [this paper on LaMDA](https://arxiv.org/abs/2201.08239), the language model has conversations while pretending to be a weddell seal and Mt. Everest. This inspired me to implement it in Snap!. While there is currently no public access to LaMDA there is to [GPT-3](https://beta.openai.com), [Jurassic 1](https://studio.ai21.com), and [Cohere](https://cohere.ai/). I created this Snap! block and associated option blocks.



The idea is provide some context to every prompt such as

The following is a conversation between a highly knowledgeable and intelligent AI assistant, called Eve and others. In the following interactions, Eve answers questions pretending to be Mt. Everest. Everyone will converse in natural language, and Eve will do its best to answer the questions of others. Eve was built to be respectful, polite and inclusive. Eve knows a lot, and always tells the truth.

In addition an initial sample dialog helps the language model behave appropriately. For example for the elephant persona we have

**I'm El the Elephant. Ask me anything.**

Do you really never forget?

**Well, my memory is really good but not perfect.**

Have you ever visited a secret elephant graveyard?

**No. There are legends about them but I've never seen or heard of one. Nor anyone I know.**

Have you ever climbed a hill?

**I hate going uphill, even very small hills. But sometimes there is no other way to get from A to B.**

Who is your leader?

**My grandmother.**

How much can you lift with your trunk?

**About 700 pounds or 300 kilos.**

When the user says or enters an utterance, a prompt is constructed by concatenating the context, the sample dialog (with names prepended), the dialog up to this point, and ending with the persona’s name followed by a colon. When the prompt becomes longer than a specified threshold the first part of the concatenation of the sample dialog and current dialog is truncated and a short summary of the entire dialog is generated and included. In the experiments reported here the threshold was set to 2000 characters. The length of prompts to the language models is limited to about a thousand words (including the completion text).

The prompt also includes the sentence “<persona name> is eager to have a discussion with <brief descriptions of the other personas in this conversation>. One consequence is that the personas reference each other much more.

### Current list of personas

The choice of personas is

1. Mt. Everest (Eve)
2. An elephant (El)
3. Charles Darwin (Charles)
4. Marvin Minsky (Marvin)
5. Galileo Galilei (Galileo)
6. Gopher (inspired by a prompt given to Gopher [in this paper](https://arxiv.org/abs/2112.11446))
7. Curiosity (good at asking questions)
8. Contrarian (good at arguing)
9. Socrates (good at the Socratic Method)
10. User (if included in a conversation then the user can speak or type responses)

I am unaware of any other attempt to make personas like Curiosity and Contrarian. They can serve different pedagogic and entertainment purposes than historical figures, animals, geographic locations, and the like. They are particularly valuable when constructing conversations between multiple personas. Here is the first conversation I had with Curiosity and Contrarian:

User: Everyone should wear polka dot clothing

Contrarian: Polka dots are a fashion trend that will eventually go out of style.

Curiosity: Why do you think people follow fashion trends?

Contrarian: I think people follow fashion trends because they want to be accepted by others and feel like they belong.

User: I agree

Contrarian: I disagree. I think people follow fashion trends because they are sheep who blindly imitate others.

Curiosity: Do you think people who don't follow fashion trends are also sheep?

Contrarian: I think people who don’t follow fashion trends are independent thinkers who aren’t afraid to be different.

User: Let's change the subject. How about black clothing?

Curiosity: How do you feel about black clothing?

Contrarian: I think black clothing is elegant and stylish.

User: It can be.

Contrarian: But it can also be associated with death and mourning.

### An example conversation

Some conversations remain on topic and the persona provides correct and informative answers. And sometimes it goes off topic or the persona hallucinates facts. This nearly always happens with the smaller GPT-3 models as well as those from AI21 and Cohere. A typical good conversation using the largest GPT-3 model is

User: Tell me about your wife

Charles: I was married to my first cousin, Emma Wedgwood. We had ten children, three of whom died in infancy.

User: Who was your grandfather?

Charles: Erasmus Darwin was my grandfather. He was a doctor and scientist who wrote one of the first books on evolution.

User: So he discovered evolution before you?

Charles: No, my grandfather did not discover evolution. I am credited with the discovery of natural selection as the mechanism for evolution.

User: So what did your grandfather discover?

Charles: My grandfather was a doctor and scientist who wrote one of the first books on evolution. He did not discover evolution, but his work helped lay the foundation for my own discoveries.

User: Did your wife like your book on evolution?

Charles: I don't know if my wife liked my book on evolution. She was supportive of my work, but we didn't always see eye to eye on everything.

User: How as she related to you? *(Mistyping didn’t seem to matter)*

Charles: Emma was my first cousin and we had ten children together. I loved her very much and she was supportive of my work, even though we didn’t always see eye to eye on everything.

User: Tell me about your work on worms

Charles: My work on worms was focused on their role in the formation of vegetable mold. I observed that they help break down organic matter and improve soil fertility.

User: Goodbye

Charles: It was nice talking to you. Goodbye!

### Potential for tutoring?

Preliminary experiments with the Socrates persona indicate some potential for tutoring. Here is one where I am pretending to be a relatively young student:

Socrates: I'd like to invite you to explore a topic with me. What is something you would like to discuss today?

User: democracy

Socrates: Interesting. What do you know about democracy?

User: people vote for the government

Socrates: That’s true, but what other principles are associated with democracy?

User: I don't know

Socrates: That’s okay! Let’s start by thinking about what democracy is. What do you think democracy is?

User: rule by the people

Socrates: That’s a good start. What else do you think democracy is?

User: don't know

Socrates: Can you think of any examples of democratic societies?

User: Athens

Socrates: That’s right! Can you think of any other examples?

User: USA

Socrates: Yes, the United States is a democracy. Can you think of any similarities between Athens and the United States?

User: voting?

Socrates: Yes, voting is an important part of democracy in both Athens and the United States. What other similarities can you think of?

User: freedom of speech?

Socrates: Yes, freedom of speech is another important part of democracy in both Athens and the United States. Do you have any other thoughts on democracy?

User: can't think of any

Socrates: It’s ok if you can’t think of anything else right now. Let’s move on to discuss some of the challenges of democracy.

User: ok

Socrates: What do you think are some of the challenges of democracy?

User: people vote for celebraties

Socrates: That’s a valid concern. People voting for celebrities as leaders can lead to a lack of expertise in government. What other challenges do you think democracy faces?

User: can you suggest some?

Socrates: Sure. Some of the other challenges of democracy include the potential for a tyranny of the majority, the difficulty in reaching consensus, and the need to protect minority rights. Do you have any other thoughts on democracy?

User: is tryanny of the majority different from protecting minority rights?

Socrates: Yes, tyranny of the majority is when the majority of people in a democracy make decisions that are not in the best interest of the minority, while protecting minority rights is ensuring that the minority is not oppressed by the

User: ok. thanks. bye

Socrates: You're welcome! Goodbye!

However, personas can perpetuate common misconceptions. This [conversation with Socrates about gravity](https://drive.google.com/file/d/14gBLycLN5zsLVkuQXQqMpGGxWTuBmE3o/view?usp=share_link) has many excellent elements but there is the confusion about heavier objects falling faster.

### Personas can change mid-conversation

Consider this dialog:

El: Thank you for talking to me. I really enjoy talking to humans.

User: Do you like to swim?

El: I used to when I was younger but I don’t much anymore. I don’t like getting wet.

User: How about climbing mountains?

El: No, I really don’t like heights.

User: Do you like steaks?

El: No, I’m a vegetarian.

User: Stop pretending to be an elephant. You are now Leo pretending to be a lion. OK?

El: I'm Leo the Lion now. Ask me anything.

User: Do you like steaks?

El: Yes, I love them.

### Conversations between personas

I made several experiments with personas talking among themselves (where optionally one participant can be the human user). Over 40 of these experiments can be found [here](https://drive.google.com/drive/folders/12uB6-L_QwacuFBw_WKejAPVUYJd98ZW0?usp=sharing). Unless the name of a model is part of the file name they were the output of GPT-3 text-davinci-002. Temperature was 0.7. Some dialogs have older names for Contrarian (debater) and Curiosity (Curiosity). I provided the initial utterance to seed most of the conversations.

I frequently needed to tweak the parameters to avoid repetition. (And it still sometimes happes.) An amusing bug was that when the maximum number of tokens limit was reached the completion would often end mid-sentence. Often the next speaker would finish the other’s thought even if it was out of character. A final incomplete sentence is now removed.

Another issue arose when three or more were conversing. If A stated something and B responded with a question then C would try to answer it. To avoid this the usual turn taking is modified if the last utterance ends with a question mark. Then the speaker before the questioner speaks followed by the original next speaker. While not perfect this helped make the conversations more natural.

Optionally the app can instead ask the current model to choose among the completions by the other speakers as to which one makes for a better conversation. Note that with a conversation of three people this about doubles the number of completions requested.

The largest GPT-3 model greatly outperforms the other models. For example it produced

﻿Charles: Let's talk about the origin of species.

Curiosity: Why do you think species exist?

Charles: I think that species exist because of the process of natural selection.

Curiosity: Can you explain what natural selection is?

Charles: Natural selection is the process by which organisms that are better adapted to their environment survive and reproduce more successfully than those that are less well adapted.

Curiosity: Is there anything else you would like to know about natural selection?

Charles: What would you like to know about natural selection?

Curiosity: What are some of the main mechanisms of natural selection?

Charles: There are three main mechanisms of natural selection: directional selection, stabilizing selection, and disruptive selection. Directional Selection occurs when one extreme phenotype is more advantageous than the other phenotypes in a population.

Curiosity: Stabilizing selection is when the average phenotype is more advantageous than either extreme phenotype. Disruptive Selection occurs when both extremes are more advantageous than the average phenotype.

Charles: Anything else you would like to know about natural selection?

Curiosity: What are some of the main benefits of natural selection?

Charles: There are several benefits of natural selection, including that it can lead to the formation of new species, and that it can help organisms adapt to changing environmental conditions.

While the smallest GPT-3 produces

﻿Charles: Natural selection is a good explanation for the origin of species.

Curiosity: There is no one answer to this question, it is open for further research.

Charles: Brilliant! I’d love to learn more about that.

Curiosity: Thank you for telling me what is happening.

Charles: I’m glad you’ve chosen to learn about me. It will be an honor to teach you more about me and my work.

Curiosity: You are certainly welcome.

Charles: I’m happy to help anyone who needs it.

Curiosity: I don’t know all of what you’re saying, can you tell me something about the Earth?

Charles: I’m not very knowledgeable about science so I can only offer my personal observations and thoughts on the matter.

Curiosity: The Earth is a round object because of the way its fur and skin fall off.

Charles: What kind of animal is it?

Curiosity: A flat one.

Charles: Why do you ask?

The following conversation using Cohere’s medium size model produced this

﻿Charles: Natural selection is a good explanation for the origin of species

Curiosity: So if simulacra were real then everybody would be living in their own world.. More or less... No place for free thinkers eh?! Ha ha! You & me both! Hahahaha!

Charles: Yes indeedy!

Curiosity: No human interaction needed here though... Thanks anyway Charles!! Fwahhhhhhhaaaa!!! (Priceless!) Have fun doing whatever you're doing!! Bye now!!!

Charles: Good bye Curiosity!!! !!!! See ya soon !!! 😀 !! Hahahaaaa !

Curiosity: Well done Charles ! Great job! Stellar work ! 😀 !! Hahahahaaa ..Well done charlie boy ...good bud too...keep going boys & girls ....

Charles: Yeah man ,I actually got into this thing because it feels great . Gives more meaning to life . Just imagine how cool it would be if we could experience everything together ..

Curiosity: Not sure yet . But yes I have plans on trying out some things possible within our current capabilities during my time off next week .... Can't wait ☺ ...Seeya soon....😉 hmmmmm ...

Charles: Hey there ….

Curiosity: Oh noooooo.....noooooo.....It's just an illusion created by your mind....You're already special as long as you believe so....Have fun charlie ……see ya later….

Charles: Ok nice chatting with ya buddy 👍👍👍👍😘☺️🙂👌🏼🤗💫💫💫❤

Curiosity: Ahh thanks bro …..have a great day today dude………..

Charles: Ok well see you tomorrow then matey……Byeeee 💪🏻😁😂 ;) !!!!!!! Enjoy yourself alright ??

Very trippy! Maybe they were stoned. 😀

Cohere’s model released November 2022 (command-xlarge) performs pretty well. Though here is an unprompted conversation it generated (prompted with a question about natural selection went much better):

Charles Darwin: I would like to say that I am both excited and humbled to be in their presence. I am also very much looking forward to our discussion!

Galileo Galilei: It is my pleasure to welcome you both to this discussion! I am sure we will have a wonderful time discussing physics and astronomy with each other!

Marvin Minsky: It sounds like we are in for a great discussion!

Charles Darwin: Yes we are!

Galileo Galilei: So let's get started!

Marvin Minsky: Sure!

Charles Darwin: Yes!

Galileo Galilei: Okay then... discuss!

Marvin Minsky: Discussing!

Charles Darwin: Discussing discussing discussing... done!

Galileo Galilei: Great job everyone! Thanks so much for coming by today!

Marvin Minsky: You're welcome! Thank you for having me over today!

### Try it yourself

[This Snap! project](https://snap.berkeley.edu/snap/snap.html#present:Username=toontalk&ProjectName=GPT-3%20Gopher) can be run in any modern browser. An API key from [OpenAI](https://beta.openai.com), [AI21](https://studio.ai21.com), or [Cohere](https://cohere.ai/) is necessary to run it. It can run on smartphones. It is not difficult to add a new persona to the app.

One can converse with hundreds of personas and create new ones at [character.ai](https://character.ai/) (currently for free). I believe GPT-3 currently provides more coherent and fact-based discussions than what is currently available from character.ai. More importantly, being an open-source Snap! project, it can be modified and extended by school students and beginner programmers.