

Personalizing Chatbots With Messages From Social Media Applications

Siavash J. Hassibi

Received March 22, 2023

Accepted July 13, 2023

Electronic access August 15th, 2023

The goal of creating “human-like” machines is not new; Alan Turing pondered this goal in his 1950 paper Computing Machinery and Intelligence. With the invention of machine learning algorithms, particularly ‘deep learning’ neural networks, computing technology continues to provide interesting results and refinements on the definitions of “human-like”. The goal of this paper is not to create or measure intelligence, but rather to investigate the creation of a computer that imitates a pre-existing human: we call this the personalization problem. Using messages from an individual’s social media accounts, we create chatbots to mimic the person who wrote the original messages. We used OpenAI’s GPT-2 model for natural language generation. To measure the “personalization” in our chatbots, we asked subjects with relationships to target persons to distinguish between chatbots with different personalities. Our results show statistically significant personalization of the targets. To our knowledge, this is the first experimental demonstration of chatbot personalization for pre-existing humans.

Introduction

Improving the conversational abilities of machine learning models is a major endeavor in the field of artificial intelligence. One challenge is to develop a chatbot that exhibits, and remains consistent with, a larger persona. We call this the personalization problem for chatbots, which is to create a chatbot with the biographical and personality features of a specific, existing individual. More specifically, we focus on creating open-domain chatbots: programs that engage in direct, one-on-one dialogue with a user.

Although similar to Alan Turing’s goal of reproducing a human-like mind in a machine, the problem of personalization is distinct. While Turing considered human-likeness as the extent to which a machine exhibits intelligent behavior, personalization deals with the assumption of a specific human identity. This identity should include biographical information and elements of personal memory, as well as personality traits and idiosyncrasies of language (e.g. slang). The degree to which a machine can be molded to a specific, pre-existing identity is, in our understanding, a novel problem and requires a creative selection and use of datasets.

We propose using messaging data sourced from a single individual which can be extracted from various social media applications. The prolific usage of social media generates huge amounts of text data, especially dialogue between two individuals. Moreover, data taken from social media is largely available to its authors, and studies show that personality features can be predicted to incredible accuracy from messaging data¹. Further, we utilize a pre-trained model, those of

Table 1 Messages extracted from social media application Discord.

Messaging Application	Total Number of Messages	Word Count	Average Length of a Message
Discord	43,774	223,286	~5 words

which have shown significant success in a variety of language-understanding tasks².

Related Works

Many previous studies have attempted to create personalized dialogue models. The most recent efforts focus on improving a set of common problems found within chatbots such as a lack of consistent personality, relatively short long term memory in conversation, and a habit of resorting to non-answers (e.g. “I don’t know”). A common approach to improving the conversational ability of chatbot models is to create a unique dataset in order to equip the model with increased ‘human-like’ characteristics. A variety of these datasets have been collected and used for model training, including a corpus of movie dialogues³ and subtitles⁴ respectively. Further attempts at compiling a dataset include PersonaChat⁵, a collection of conversations between volunteers acting on a pre-constructed profile (a brief character description).

We define a persona as various elements of identity⁶, ranging from biographical information to language use and behavior. Thus, personalization is the degree as to which our model adopts this larger persona. Much like research de-

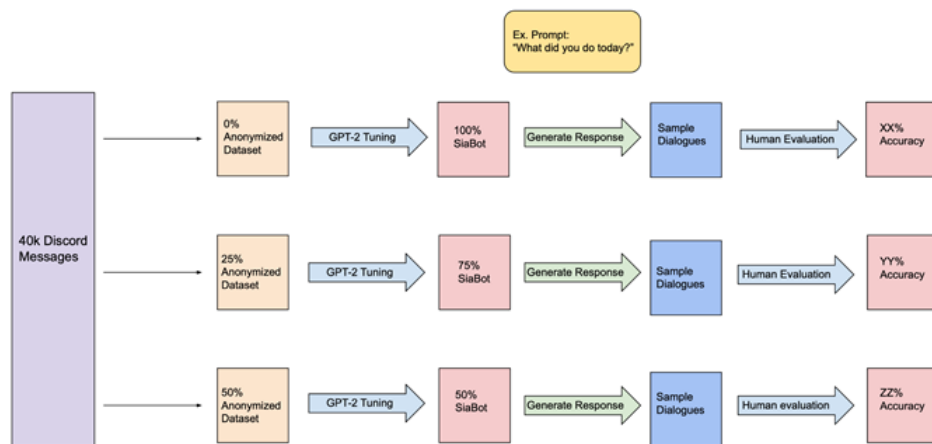


Fig. 1 Training data for each chatbot.

scribed above, we propose a new dataset; however, rather than improve fluency, we aim to increase the personalization of the model. Pretrained models have demonstrated promise in the development of task-oriented dialogue⁷, and similar research investigates their utility in an open-domain setting. ‘Persona-rich’ data, like movie scripts and social media conversations, are then used to fine-tune the model⁸. The objective of our research is to determine the level as to which a model can mold to a cohesive larger personality, and thus datasets like PersonaChat, where a variety of personas are explicitly featured throughout the dialogue, cannot be used.

Methods

Data Collection

Our dataset was a conversion log of Discord Messages between two individuals (Siavash Hassibi and “Ben”) over a period of two years. More specifically, we obtained messages from the author’s Discord account. We elected to use Discord data because the author had a large conversation history immediately available. The data was received as a collection of .csv files. For the data to be viable, it had to be reformatted (i.e. removal of timestamps, renaming of conversation participants) and converted into a .txt format. This was done using an open source parser from Github⁹.

The messaging datasets are personal conversations, often including slang and casual use of punctuation, grammar, and syntax. The chat data is almost always between the author and a friend, “Ben”, from school, discussing a wide variety of topics from classes to gaming to current events, with conversations taking place from September 21, 2020 to the date when we downloaded the Discord data, November 10th, 2022.

Training Chatbots

The model we used for training was a Github implementation of GPT-2, known as Gravital¹⁰. Developed by OpenAI¹¹, GPT-2 is a general purpose (i.e. not designed for a specific usage case) model that generates text based on inputs, or prompts, from a user. A deep learning model, GPT-2 is a transformer neural network equipped with 12 layers of neurons. OpenAI’s more recent and currently proprietary chatbot ChatGPT, based on a similar architecture, has received extensive media attention for its human-like conversation skills¹². We elected to use GPT-2 because of its extensive open-source support as many fine-tuning algorithms were available for the pretrained model. While GPT-2 is an established model, tuning allows for the specification of the model with respect to a specific dataset. In our case, we tune on different datasets of conversation messages (Table 2) to test the degree the model adopts a larger persona.

A diagram of our procedure (Figure 1) demonstrates how GPT-2 was used. A pool of roughly 40,000 messages was edited into four different datasets, with each containing a different level of author representation. Table 2 demonstrates example text from each, where a differing percentage of the ‘SIAVASH’ speaker token is replaced with a random five letter string. A total of five models were trained with a standard number of steps at 3000 and a model size of 500 MB. After creating a set of prompts, each model generates a series of responses which are then evaluated.

Assessment

To assess the personalization achieved by our chatbot, we showed close associates of Siavash and Ben the anonymized dialogues generated by the models, asking them to identify who the bot was imitating. Our hypothesis was that a higher

Table 2 Training data for each chatbot.

Name of Dataset	100%-SiavashBot	50%-SiavashBot	25%-SiavashBot	0%-SiavashBot
Description	No changes to dataset	Replace 50% of Siavash’s name with random 10-character string	Replace 75% of Siavash’s name with random 10-character string	Replace all occurrences of Siavash’s name with random 10-character string
Example dialogue	SIAVASH: Hey, what are you up to today? BEN: Nothing much my whole day is free, what about you? SIAVASH: Yeah same my day is pretty much free BEN: What were you thinking of doing? SIAVASH: You wanna grab coffee after school? BEN: Sure, can we get lunch too? don’t eat at school SIAVASH: Yeah sounds good, I’ll see you then	KJGHV: Hey, what are you up to today? BEN: Nothing much my whole day is free, what about you? SIAVASH: Yeah same my day is pretty much free BEN: What were you thinking of doing? FDTASD: You wanna grab coffee after school? BEN: Sure, can we I get lunch too? don’t eat at school SIAVASH: Yeah sounds good, I’ll see you then	GRWSD: Hey, what are you up to today? BEN: Nothing much my whole day is free, what about you? SIAVASH: Yeah same my day is pretty much free BEN: What were you thinking of doing? MNKWSY: You wanna grab coffee after school? BEN: Sure, can we I get lunch too? don’t eat at school RWESD: Yeah sounds good, I’ll see you then	TVTHR: Hey, what are you up to today? BEN: Nothing much my whole day is free, what about you? HGTDHN: Yeah same my day is pretty much free BEN: What were you thinking of doing? LDSTW: You wanna grab coffee after school? BEN: Sure, can we I get lunch too? I don’t eat at school POHWS: Yeah sounds good, I’ll see you then

presence of "Siavash" vs. anonymized messages in the training data (i.e., 0% vs. 50% vs. 75% vs. 100%) would lead to more personalization imbued in the Siavash chatbot, and thus, higher correct discrimination rates. Pre-existing methods for assessing chatbots often focus on sentence understanding¹³, however, similar persona-based research frequently utilizes human evaluations to measure fluency, specificity¹⁴, and conversation depth¹⁵. For the purpose of evaluating our specific definition of personalization, we create a new method.

We prompted each chatbot with 20 questions or sentences related to biographical information or a topic of shared interest to the chatbots, e.g., Figure 2: "Give an example of a computer game". Then, we selected five sample generations, chosen for being most coherent, from each of the five chatbots. We anonymized these samples and prepared a survey of 20 questions asking a human subject to determine the identity of one of the chatbots, whose identity has been masked using "O" or "X" for the name and green or blue highlighting to make the separate identities easier to distinguish. For our data collection, we gave the survey to seven individuals who know both Ben and Siavash personally. For each survey question, the subject was asked to identify one of the two (randomly) chosen chatbots, as being either "Siavash" or "Ben".

Results and Discussion

When dividing the responses to each question based on the chatbot that generated them, the human subjects were not more or less likely to identify the chatbots for different levels of "Siavash"-anonymization (Figure 3). That is, altering the dataset to replace a certain percentage of "SIAVASH" did not seem to affect the personalization exhibited by the chatbot. Participants exhibited a relatively consistent accuracy despite the differing models.

The total number of correct responses in our survey was 94 for the 140 questions. This probability that these results occur in the case that no personalization occurred, i.e., the participants could not discriminate at all between bots, can be modeled by a binomial distribution with $x = 94$, $n = 140$, and $p = 0.5$.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{1}$$

$$P(X \geq x) = \sum_{X=x}^n P(X = x) \tag{2}$$

Binomial Probability Density Function 1 and Binomial Cumulative Density Function 2

Then, the probability that 94 or more answers would be randomly guessed correctly is $P(X \geq 94) = 0.003\%$. Because of

Who is X? *

```
PROMPT 1008: Give an example of a computer game.
O: map_vrz
X: but i think its real good
O: show me
X: nahhh
X: i cant
X: not that long
X: bro
X: i am uh
X: My grandma was told me it was the cousins in danville
X: and the intro song was on the cover of that playlist
X: you recognize that eagle
O: damn
O: i cant get u to another point
X:
X: thats cloud rap in a nutshell
X: im not a big fan of the big 3 shoegaze bands and i know of its like
X: i feel like its too
X: my nostalgia big 3rd and half of the big 3 shoegaze side
X: but its not 4 me, and of coursehh i can do it too
X: maybe stuff
X: you too
X: just like
X: theres a lot of young stuff
```

- Ben
- Siavash

Fig. 2 Example survey question and chatbot response.

this extreme unlikelihood, we conclude that the chatbots are exhibiting statistically significant personalization of the target persons. Participants of our survey were able to consistently identify the chatbot’s “identity” with around 70% accuracy. To our knowledge, this is one of the first experimental demonstrations of chatbot personalization to existing individuals. One potential limitation of this result is that the chatbot generations used to create the survey were not randomly selected, instead chosen for coherence. Bias regarding what is considered fluent was not accounted for in our experiment; however, it should be noted that many model responses were just a collection of phrases that lacked any larger sensibility. Related efforts in pursuing personalization center around improving the fluency of their models, though auxiliary experiments demonstrate a simple persona-consistency⁵. While evidence of some degree of personalization, these results do not indicate that the dialogue model has absorbed a larger identity.

In our experiment, the personalization was not affected by replacing the name “SIAVASH” with random labels (see Table 2). This could be because the name “BEN” was never changed and the model associated messages for a Siavash-chatbot by process of elimination. Given the relative success

of personalization, more study is needed to determine gradations of personalization achievable by manipulating the training data. Randomizing either name, randomly dropping either name for a third ‘unknown’, and randomly swapping the names are all models that potentially may succeed in regulating the personalization exhibited by the chatbot. However, these models make it more difficult to assess who the chatbot is attempting to imitate. A strategy for solving this could be to tie the model to generate one line at a time and then forcing it to produce one author first.

To attempt better results, we suggest some improvements to the datasets.

- *More messages*: Having more conversation history, for example, would be beneficial to the chatbot training. This could be done by concatenating datasets from a wider variety of social media applications, for example.
- *Cleaner, more structured language*: Data taken from a more formal setting (e.g. email), where long-form messages are more grammatical and media attachments are less frequent, might be more effective for training.
- *Binary turn-taking*: In the current data, one person would

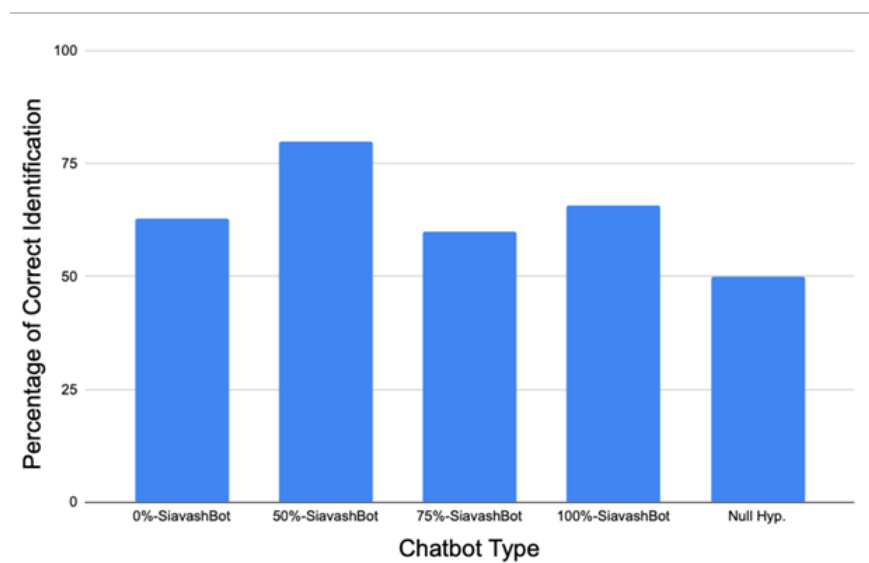


Fig. 3 Example survey question and chatbot response.

often send many short messages, followed by a burst of messages from the other person, and so on; this lessens the dialogue effect of the dataset. One could pre-process the messages to have a strict, back-and-forth style, where each participant sends one message, then receives a response. However, this could potentially trivialize the model as rather than learning by speaker label, it could instead learn to alternate personalities.

Conclusion

We study the personalization problem for chatbots. Making use of a pre-processed dataset from a social media application Discord, we apply state-of-the-art machine learning algorithms to create chatbots with varying degrees of personalization. Newer and more advanced large-language models are being developed, and a potential avenue for future work could utilize them for a similar procedure. We also present a novel experimental design for evaluating personalization; unlike most chatbot evaluation criteria in existing work, we focus on determining if human subjects can identify between chatbots personalized towards different, existing individuals, rather than measures of the coherence or sensibility of the chatbot’s language. This reflects the limitations of current natural language generation technology; our goal not to create intelligence, rather to create technology and evaluation criteria towards personalization. A potential avenue of exploration is creating chatbots from public figures or fictional characters, who may have larger available datasets. The use of well-

known individuals also allows more people to be involved in guessing authorship as well.

Acknowledgements

The author would like to extend a special thanks to Isaac Mackey from UC Santa Barbara for his guidance and mentorship throughout the research process. Additionally, he would like to thank Jack Urbanek and other anonymous reviewers for their comments that led to the improvement of this paper as well as his brother, Sohrab, who is an invaluable source of inspiration.

Author

Siavash J. Hassibi is a student at Menlo-Atherton High School in California. He is interested in computer science, particularly media generation accomplished through machine learning. Currently, Siavash is working on a project to check the clinical accuracy of ChatGPT when summarizing medical data.

References

- 1 Y. Montjoye, J. Quoidbach, F. Robic and A. Pentland, Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA.
- 2 A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, *Improving language understanding by generative pre-training*.
- 3 R. Banchs, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).
- 4 International Conference on Intelligent Virtual Agents, Cham.

-
- 5 S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela and J. Weston, *Personalizing dialogue agents: I have a dog, do you have pets too?*, arXiv preprint arXiv:1801.07243.
 - 6 J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao and B. Dolan, *A persona-based neural conversation model*, arXiv preprint arXiv:1603.06155.
 - 7 P. Budzianowski and I. Vulić, *Hello, it's GPT-2—how can I help you? towards the use of pretrained language models for task-oriented dialogue systems*, arXiv preprint arXiv:1907.05774.
 - 8 Y. Zheng, R. Zhang, M. Huang and X. Mao, *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, year.
 - 9 C. Ling, *Github repository*.
 - 10 N. Brisebois, *DiscordChatAI-GPT2*, *Github repository*, https://github.com/sudo-logic/GPT-2_Discord.
 - 11 O. H. Page, <https://openai.com/>.
 - 12 K. Roose, *New York Times*.
 - 13 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*, arXiv preprint arXiv:1804.07461.
 - 14 D. Adiwardana, M. Luong, D. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu and Q. Le, *Towards a human-like open-domain chatbot*, arXiv preprint arXiv:2001.09977.
 - 15 A. Baheti, A. Ritter, J. Li and B. Dolan, *Generating more interesting responses in neural conversation models with distributional constraints*, arXiv preprint arXiv:1809.01215.