# Augmenting Training Data for a Virtual Character Using GPT-3.5

**Elizabeth Chen*** and **Ron Artstein**
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista CA 90094-2536, USA
ec105@wellesley.edu , artstein@ict.usc.edu

## Abstract

This paper compares different methods of using a large language model (GPT-3.5) for creating synthetic training data for a retrieval-based conversational character. The training data are in the form of linked questions and answers, which allow a classifier to retrieve a pre-recorded answer to an unseen question; the intuition is that a large language model could predict what human users might ask, thus saving the effort of collecting real user questions as training data. Results show small improvements in test performance for all synthetic datasets. However, a classifier trained on only small amounts of collected user data resulted in a higher F-score than the classifiers trained on much larger amounts of synthetic data generated using GPT-3.5. Based on these results, we see a potential in using large language models for generating training data, but at this point it is not as valuable as collecting actual user data for training.

## Introduction

The Digital Survivor of Sexual Assault (DS2A) is a system which allows users to interact with a digital representation of a survivor of sexual assault in a manner that mimics face-to-face conversation, in order to learn the survivor's story, establish a connection, and build empathy through direct interaction (Artstein et al. 2019). The system uses a large library of pre-recorded video statements by the survivor; a classifier trained on linked questions and answers allows the system to retrieve an appropriate response to new, unseen questions (Leuski and Traum 2011). The first character, Jarett Wright, was trained on a set of 7759 questions, 1926 responses, and 18279 links, following a months-long effort of data collection from real users. A second survivor of sexual assault, Samantha Downey, was recorded but was not subject to substantial user data collection. This paper explores the possibility of using a Large Language Model (LLM) to create synthetic training data in lieu of actual user data.

Previous works have shown that LLMs encode a great deal of knowledge about language use (Wiedemann et al. 2019; Tenney, Das, and Pavlick 2019), and can generate training data for classification (Piedboeuf and Langlais 2023; Møller et al. 2023). We use the GPT-3.5 (also, ChatGPT) model from OpenAI to generate training data for the Samantha Downey character. Our success depends on prompting the LLM to predict what people are likely to say in interaction, and to identify which responses would be appropriate for such utterances. We explore the use of zero-shot and one-shot prompting to create synthetic training data in place of real user data, and test the performance of classifiers trained on such synthetic data. Our results show that GPT-3.5 prediction can be useful to some extent, but is still not as useful as the collection of live interaction data.

## Method

All the experiments follow the same general structure: we train a classifier using NPCEditor (Leuski and Traum 2011) with a set of linked questions and responses, and test the classifier with a set of unseen questions. The responses are the same for all classifiers; the classifiers differ only in the training questions and links between questions and responses. We use the same test set for all the classifiers: 358 questions, manually linked to appropriate responses.

**The baseline classifier** is trained using only the data from the recording interviews: 1861 questions, 1857 responses, and 1862 links between questions and responses. Additional classifiers use several augmentation methods; LLM prompting follows the guidelines of the Prompt Engineering Guide (https://www.promptingguide.ai).

**Method 1: New Questions** We ask GPT-3.5 to generate new questions for each of Samantha's responses. We prompt the model with some context, an example, the response, and the corresponding baseline question, following a one-shot prompting strategy. The temperature and frequency penalty parameters for the model were both set to 0.5 to allow for some creativity and randomness. This added 6993 new questions and 7758 new links to the baseline system.

**Method 2: New Links** We use GPT-3.5 to find additional links between the questions and responses of the baseline dataset. Due to time and cost constraints, we run each question through the baseline classifier to find the 15 top-ranked responses. Then, we present GPT-3.5 with each of these responses and the question, and ask it whether the response is appropriate for the question. We set the temperature and frequency penalty parameters to 0 for a more deterministic output. This added 3078 new links to the baseline system.

*Now at Wellesley College

**Method 3: Questions and Links** We combine the first two methods by simply linking each question generated by Method 1 to all the responses linked by Method 2 to the corresponding baseline question. The resulting dataset contained 6988 new questions and 22,102 new links, which were added to the baseline system.

**Method 4: Jarett Questions** We link questions collected from live user interaction for a previously completed character, Jarett Wright (Artstein et al. 2019), to responses of Samantha, using the same prompting as in Method 2. Since both characters are survivors of sexual assault, some of the questions that users ask Jarett are also likely to be asked of Samantha. From Jarett's training data, GPT-3.5 identified 2386 new usable questions and 7067 new links, which were added to the baseline system.

**Method 5: Collected Questions** A small amount of interaction data collected for Samantha was not included in the test set. We manually cleaned this data by removing any invalid questions and correcting the automatic transcriptions, then annotated each usable question with links to appropriate responses. This resulted in 196 new questions and 320 new links, which were added to the baseline system.

All six classifiers (baseline and five augmentations) were tested on the same test set, using NPCEditor. After training on the appropriate dataset, NPCEditor returns zero or more responses for each of the test questions, and calculates the precision, recall, and F1 of the retrieved responses relative to the annotated correct responses.

## Results

The overall scores are generally very low, because for most of the test questions, the classifiers do not return any response, resulting in scores of zero on all three measures. The baseline system is the lowest, with F1=0.036. Each of the augmentation methods manages to nearly double performance over the baseline, likely due to the extreme poverty of the baseline classifier (Method 1: 0.067, Method 2: 0.075, Method 3: 0.059, Method 4: 0.102, Method 5: 0.103).

Although Method 1 attempts to encourage more creative outputs, many of the generated questions were still paraphrases of the original question. Some of the generated questions also contained awkward phrasing and/or the use of semi-formal language, not reflective of typical conversational language use. Additionally, many questions generated by GPT-3.5 addressed very specific sections of the response rather than the response as a whole.

The classifier trained on GPT-found links (Method 2) performed marginally better than the classifier trained on GPT-generated questions. Overall, the new question-response links were reasonable, with the response usually at least partially answering the question. However, some possible question-response links were incorrectly discarded.

Method 3 leads to worse performance than either Method 1 or Method 2. This is probably because GPT-generated questions are not always equivalent or even similar in meaning to the example; it therefore does not follow that responses appropriate to one question would be appropriate for a new question generated using the first question as an example.

Finally, we note that the highest performing classifiers, and the only ones to score above 10% on all three measures, are the ones which augment the baseline with questions from the Jarett system (Method 4) or questions collected directly for Samantha (Method 5).

## Discussion

This work has shown that synthetic data generated through prompting GPT-3.5 can be used to improve the performance of a classifier with severely impoverished training data. However, the synthetic data has not been shown to match the quality of user data collected in interaction. Adding just under 200 questions and 320 links of collected user data is equivalent or superior to thousands of questions and links generated or identified by GPT-3.5. In other words, we have not demonstrated the ability of a LLM to predict what users might say in an interaction to the same extent that collected data from some users can predict what other users might say. This is not to say that such prediction is not possible: it may be that better prompt design or better language models (such as the most recent model of GPT-4) could improve prediction. However, given the results of our experiment, it appears that collecting actual user data in interaction is still invaluable for training interactive systems.

Our results suggest several directions for future research. To better evaluate the limitations observed from these experiments, it may be worth replicating the experiment with more recent models and exploring different prompt design or fine-tuning strategies. Furthermore, the relative success of the method that uses GPT-3.5 to link questions from a related but slightly different domain (Jarett questions) suggests that a fruitful use of LLMs may be not so much in predicting what users might say, but rather in helping transfer the predictions from one domain to another.

The topic of sexual assault can raise issues that are very sensitive and personal. To mitigate against potential leakage of sensitive information, all the queries to GPT were sent through the Enterprise API, which does not store the data long-term and does not use it to train new GPT models. The text generated by GPT-3.5 is used purely for training the classifier, and is thus not visible on the user-facing side of the system; users only experience recorded content, and the system complies with the requirements of consent, fair representation, veracity, and informedness for dialogue agents that represent real persons (Artstein and Silver 2016). It is still possible that biases encoded in the language model could affect *which* recorded statements are played to the user, but the adverse effects of such bias should be minimal.

## Acknowledgments

The first author was supported by NSF award 2150187 "REU Site: The Face and Mind of Artificial Intelligence" (PI: Ron Artstein). The second author was sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

## References


Artstein, R., and Silver, K. 2016. Ethics for a combined human-machine dialogue agent. In *Ethical and Moral Considerations in Non-Human Agents: Papers from the AAAI Spring Symposium*, 184–189. Stanford, California: AAAI Press.

Artstein, R.; Gordon, C.; Sohail, U.; Merchant, C.; Jones, A.; Campbell, J.; Trimmer, M.; Bevington, J.; Engen, C. C.; and Traum, D. 2019. Digital survivor of sexual assault. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, 417–425. Marina del Rey, California: ACM.

Leuski, A., and Traum, D. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine* 32(2):42–56.

Møller, A. G.; Dalsgaard, J. A.; Pera, A.; and Aiello, L. M. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. ArXiV preprint 2304.13861.

Piedboeuf, F., and Langlais, P. 2023. Is ChatGPT the ultimate data augmentation algorithm? In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 15606–15615. Singapore: Association for Computational Linguistics.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. Florence, Italy: Association for Computational Linguistics.

Wiedemann, G.; Remus, S.; Chawla, A.; and Biemann, C. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.