# Project Midterm Report

**Harshil Bhullar**
hbhullar18@g.ucla.edu

**Krish Patel**
knpatel@g.ucla.edu

**Siddarth Chalasani**
siddarth@cs.ucla.edu

## 1 Problem Introduction

In our project, we tackle the task of creating "persona chatbots", i.e., language models that are capable of carrying a fluent, continued conversation while basing the content of their utterances on a predefined persona. Although the problem we attempt to solve is not novel, previous attempts usually either failed to maintain a consistent personality [1] or suffered from long-term memory degeneration [3]. In our project, we survey current literature on persona chatbots, evaluate baseline models currently being employed, and suggest possible improvements to yield improved results.

## 2 Literature Review and Baseline Model Selection

After reviewing the 3 papers we initially planned on building our project with, we decided to drop the Facebook dataset and Cornell Movie Dialogue, which will be discussed further below. We are going to continue with Google's Synthetic Persona Chat Dataset. We now do a short review of each of the papers that we considered for our project.

### 2.1 Facebook AI's Persona-Chat Dataset and Baseline Model

The paper "Personalizing Dialogue Agents: I have a dog, do you have pets too?" reintroduces general chit-chat models as a valuable end-application [4]. To aid in the task of creating such models, the authors created the PERSONA-CHAT dataset, which consists of 162,064 utterances between randomly paired crowdworkers who were each assigned one of 1155 personas to base their conversation on. These personas, consisting of at least 5 sentences, were also generated by crowdworkers and assigned randomly. The paired crowdworkers were then asked to chat in a natural, engaging manner. To avoid trivial word overlap in the dataset, each of the 1155 personas was rewritten multiple times by crowdworkers, with explicit word overlap being forbidden.

The paper then evaluates two classes of models: ranking models, which the paper uses as its baseline; and generative models, which the paper uses as its novel approach. The ranking models employed are a greedy IR model and Starship, and the generative model used is based on an LSTM architecture with the hidden encoder state initialized to the embedding of the provided persona. The paper found that, although their generative model did not necessarily outperform the ranking models on their evaluation metrics (perplexity and hits@1), both classes of models were significantly better at the task of persona-based chit-chat when conditioning prediction on their own persona.

Therefore, we consider their ranking models and LSTM-based architecture as our baseline. The paper also presents a platform for training and evaluating dialogue models, ParlAI [2], which could aid our experiments. We decide to discard the PERSONA-CHAT dataset in favor of Google's Synthetic Persona Chat dataset (described in the following section) because of its superior quantity of data at similar quality.

### 2.2 Google's Synthetic Persona Chat Dataset

The paper "Faithful Persona-based Conversational Dataset Generation with Large Language Models" introduces a novel method for creating high-quality conversational datasets that are faithful to user personas. The authors propose a structured approach involving three key components: User Generation, User Pairing, and Conversation Generation.

User Generation involves two sub-steps: Persona Expansion and User Profile Construction. The process starts by bootstrapping seed persona

attributes using prompts to generate a broader set of persona attributes. These attributes are then used to construct user profiles with the aid of a Natural Language Inference (NLI) model, ensuring the consistency and non-redundancy of the profiles.

User Pairing identifies potential pairs of user profiles for conversation based on their semantic similarity, which is assessed using the BERT model. The goal is to pair profiles that share common persona attributes to potentially enhance the engagement level of the conversations.

Conversation Generation uses a Generator-Critic architecture. The Generator creates initial conversational exchanges between paired profiles, which are then evaluated by the Critic based on predefined quality metrics. This iterative process aims to refine the conversations further, ensuring they meet the desired standards of relevance and persona consistency.

This framework not only allows for the generation of personalized dialogues but also ensures that the conversations are aligned with the constructed personas, thereby avoiding the generation of dialogues that contradict the persona attributes. This means that the dataset created with this paper's approach is perfect for our use case for creating persona chat bots.

### 2.3 Cornell Movie Dialogue Dataset Issues

Initially, our team considered using the Cornell Movie Dialogue Dataset, which contains around 9,000 personas or different users and over 304,000 conversations. However, while delving deeper into the dataset, several issues emerged that led us to reconsider its usability for our project.

Firstly, the dataset lacked fundamental information about the speakers involved in the conversations. While the gender of the speakers was available, other relevant details such as their interests, backgrounds, or roles in the conversations were absent, and fields such as the movie name and speaker name were irrelevant in this scenario. This lack of speaker information limited our ability to analyze their dialogue and understand the context of the conversation.

Additionally, the dataset only contained dialogues extracted from movies, which presented more challenges. Unlike real-life conversations, movie dialogues are scripted and may not accurately reflect the nuances and dynamics of natural language interactions. Along with that, the context surrounding these dialogues, such as the setting, character relationships, and plot developments, was often missing or ambiguous.

## 3 Description of Approach

For our evaluation metrics, we decided to use Perplexity, BLEU, and ROUGE to evaluate our fine-tuned model's performance

- Perplexity: It measures how well a probability distribution or probability model predicts a sample. In this context of language models, perplexity quantifies how well the model predicts a given sequence of words. A lower perplexity indicates better performance, as it suggests that the model is more confident and accurate in its predictions, and doesn't deviate from the actual topic in consideration.

- BLEU: It measures the similarity between the generated text and one or more reference translations(in this case, the output of Facebook's LSTM persona chatbot. BLEU operates by comparing n-grams (sequences of n consecutive words) between the generated text and the reference text. It calculates a precision score for each n-gram size (usually up to 4-grams), and then combines these scores using a modified brevity penalty to calculate the final BLEU score. A BLEU score of 1.0 indicated perfect match.

- ROUGE or Recall-Oriented Understudy for Gisting Evaluation is a set of metrics used for evaluating text generation tasks. It measures the overlap between the generated summary or text and one or more reference summaries. ROUGE calculates various measures such as ROUGE-N (which focuses on n-gram overlap), ROUGE-L (which measures the longest common subsequence between the generated and reference texts), and ROUGE-W (which considers weighted versions of ROUGE-N). Similar to the above metric, a score of 1.0 indicates a perfect match between the two comparison dialogues.

## 4 Project Progress Until Now

- Analyzed other datasets and their respective literatures. Rejected the Facebook's and the Cornell Movie Dialogue datasets due to limitations such as limited data quantity or lack of a context.
- Evaluated baseline models for the persona dataset, and decided to move forward with Facebook's approach with LTSMs.
- Set Up our access to OpenAI's GPT-3.5 Turbo Model using their API to query with simple personal injection.

## 5 Subsequent Weeks

- Week 5-7: We'll integrate persona information into our model via APIs, including details like interests and hobbies, to personalize responses. Additionally, we'll fine-tune OpenAI's model to better suit our needs, optimizing its ability to generate relevant responses based on provided personas.
- Week 8: Our aim is to consolidate our evaluations from the previous weeks, generate visualizations or demos to showcase our models' capabilities, and work on preparing a presentation summarizing our findings and progress.
- Weeks 9-10: During these final weeks, we'll finalize our experimental results and conduct a comprehensive qualitative and quantitative analysis of our proposed method compared to baseline approaches. Additionally, we'll dedicate time to completing our final report, including methodology, results, and insights gained throughout the project.

## References


[1] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In Lucia Specia, Matt Post, and Michael Paul, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[3] Oriol Vinyals and Quoc V. Le. A neural conversational model. *ArXiv*, abs/1506.05869, 2015.

[4] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.