Chatbot with a Persona
Dialogue and Narrative Coursework

Jack Burnett
Interactive AI CDT
University of Bristol
jack.burnett@bristol.ac.uk

Abstract

This study uses modern LLMs to attempt the ConvAI2 competition. The competition's goal was to create a conversational AI that could generate a suitable response to user inputs, given a dialogue history and a description of the agent's persona. This study developed a system for evaluating LLMs for this task using human feedback. Using the developed system, three models were implemented for testing; it was found that modern LLMs can produce suitable outputs deemed representative of a persona but struggle to convey them naturally. Hyperparameters were also analysed, with findings indicating that increasing temperature increases the naturalness of outputs and top_p increases result in better persona representation.

1 Introduction

The Second Conversational Intelligence Challenge (ConvAI2) (Dinan et al., 2020) was a competition for Advances in Neural Information Processing Systems 31 (NeurIPS 2018) (Bengio et al., 2019); ConvAI2 (Dinan et al., 2020) aimed to further the development of 'high-quality dialogue agents capable of meaningful open domain conversation', this was achieved through creating a scenario for testing chatbots that engage with humans. The competition used a data set, called Persona-Chat, of conversations between two individuals who had been given personas to act as; the goal was to use Persona-Chat and a dialogue history within a conversational AI to generate suitable responses to user inputs. The motivation for the competition was to enhance the consistency and engagement of conversational models through the use of personas, based on research by Zhang et al. (2018). The models produced by the competition were evaluated using a test set, Amazon Mechanical Turk, and live evaluation from volunteers. The models produced in this report cannot use the test set and Amazon Mechanical Turk; instead, the researcher evaluates each model based on interactions with each AI.

In this report, the ConvAI2 competition will be attempted using modern pre-trained generative models, such as GPT-3 (Floridi and Chiriatti, 2020) and LLaMa (Touvron et al., 2023a), and modern techniques to fine-tune these models, such as prompt engineering (Muktadir, 2023), to show how advances in large language models have begun to trivialise the task of creating personas. Before ConvAI2, chit-chat models were seen to have issues regarding a lack of consistency when it came to personality (Li et al., 2016) and long-term conversations (Serban et al., 2016); these are both issues which modern generative chatbots are successfully mitigating, with ChatGPT having a working memory similar to humans (Gong et al., 2023) and prompt engineering enabling personas to be successfully implemented (Short and Short, 2023). Therefore, this research evaluates how well modern models can implement and represent the personas used in the ConvAI2 competition. This research sits within conditional text generation (Guo et al., 2020), as the overall goal is to generate text according to pre-specified conditioning, such as sentiment or constraint.

2 Previous Approaches

Understanding researchers' approaches during the competition will provide insight into how personas are successfully implemented within large language models. Prior techniques may be adapted into the pre-trained model approach and may prevent potential pitfalls and errors. The models that will be discussed and analysed are those developed by Hugging Face, Little Baby, Lost in Conversation, and Mohd Shadab Alam, all of which can be found in the article by Dinan et al. on the ConvAI2 competition (2020). These approaches' findings align with advances in LLMs since the competition.

2.1 Hugging Face

Hugging Face's approach focused on the model's ability to interact with frequently switching shallow topics; this was implemented through a generative neural network and transfer learning. Hugging Face pre-trained the model with a language modelling objective using GPT-1 (Radford and Narasimhan, 2018); the model was modified via fine-tuning. Hugging Face fine-tuned GPT-1 using positional embeddings, embeddings that indicate the ownership of tokens, and semantic learning. The Hugging Face model was the most successful in the automatic evaluation stage of the ConvAI2 competition (Dinan et al., 2020). Hugging Face did not identify further improvements they would make to the model.

2.2 Little Baby

Little Baby's approach focused on using a Sequential Matching Network (Wu et al., 2017) that could model semantics for a sentence, capture utterance-response matching, distil important matching information, and capture the temporal relationship of utterances; this approach allowed for multi-grained semantic information to be extracted for each sentence. While there were advantages to this model, it was identified that implementing reply history or utilising a generative model would be beneficial for the competition's task. Little Baby did not perform as well as Hugging Face's or Lost in Conversation's models, which were generative.

2.3 Lost in Conversation

Lost in Conversation's approach (Tselousov and Golovanov, 2018) focused on simulating 'normal' conversation by learning the interests of the other agent and then discussing its interests to find common ground. The model was trained using Persona-Chat (Korshuk, 2019), DailyDialog (Li et al., 2017), and a dataset of Reddit comments. The model was built on the pre-trained GPT-1 model (Radford and Narasimhan, 2018), which was modified by providing persona information and a dialogue history in addition to minor modifications to the attention layer. The model attempted to simulate human behaviour by analysing sentiments, correcting errors, and adding emojis. This model performed best within human and 'engagingness' evaluations during the ConvAI2 competition (Dinan et al., 2020). Tselousov and Golovanov (2018) identified future improvements, such as optimising memory, speed, and sentence attention, which are improvements that GPT-3 has implemented over GPT-1 (Imamguluyev, 2023).

2.4 Mohd Shadab Alam

Mohd Shadab Alam's approach focused on implementing a seq2seq encoder and fine-tuning the embeddings. The model utilised Universal Language Model Fine-tuning (Howard and Ruder, 2018) to train and fine-tune the language embeddings using the Persona-Chat dataset (Korshuk, 2019); these embeddings were then concatenated with pre-trained embeddings from GloVe (Pennington et al., 2014) to produce the input vector for the seq2seq encoder. A highway layer (Srivastava et al., 2015) was introduced to reduce bias in the encoder's output. The researchers who developed this model did not identify further improvements that could be made to the model.

2.5 Summary of Previous Approaches

The reviewed approaches show that the generative models performed better within the competition and were seen as direct improvements to alternative models, with one identifying that they would implement a generative model to improve their performance in the competition. Radford and Narasimhan's paper (2018) on GPT-1 was released during the competition, making ConvAI2 (Dinan et al., 2020) one of the first usages of the model for conversations with personas; since ConvAI2, GPT models have been used to enable chatbots with personas in a wide variety of settings with success (Shao et al., 2023; Lee et al., 2022).

The key takeaways from the previous approaches are that generative models performed best during the competition, utilising further training data improves performance, dialogue history greatly impacted human evaluation, and pre-trained models were preferable.

3 Motivation and Chosen Approach

The success of GPT-1 (Radford and Narasimhan, 2018) in the ConvAI2 competition (Dinan et al., 2020) and the advances of GPT-3 (Imamguluyev, 2023) motivated the approach to attempt the challenge with modern generative models. Current generative models have successfully implemented personas from fictional novels and short stories (van der Zon, 2023), enabling authors to engage with their created characters and further understand them. Character-LLM (Shao et al., 2023) is a generative large-language model, built upon GPT-3, that can be trained, akin to prompt engineering, using a persona; this implementation has allowed for the successful implementation of historical figures, fictional characters, and celebrities as personas, with the resulting human-AI interactions capable causing users to develop emotional bonds with the AI (Zahira et al., 2023). Within the approaches to the ConvAI2 competition, it was identified that poor long-term memory caused humans to engage less with a model; Landwehr et al. (2023) implemented a system to enable long-term memories for AI characters created via prompts in GPT-3, potentially solving this issue. Evaluating the different implementations of the pre-trained models allows for effective methods for fine-tuning models with personas to be identified; this evaluation shows how effective modern models are at completing the task using one-shot learning.

The chosen approach, which aims to achieve the goals of ConvAI2 and the motivations for this research, is to implement and evaluate two modern generative LLMs, LLaMa (Touvron et al., 2023a) and GPT-3.5 (Floridi and Chiriatti, 2020); these models will be implemented using one-shot learning (Jurafsky and Martin, 2008) and with modifiable parameters (Naveed et al., 2023). While Landwehr et al.'s (2023) system for AI memories could be implemented, this will have no benefit with the competition's Persona-Chat dataset due to the personas being singular statements that lack depth; this is also the case for Shao et al.'s (2023) Character-LLM as the personas do not provide enough information to complete the prompt inputs, which was found after brief testing of the model. The fine-tuning methods are based on research into the effective fine-tuning and prompt engineering of pre-trained large language models (Radiya-Dixit and Wang, 2020; Gao et al., 2021; Liu et al., 2021), and will implement the most common (Jurafsky and Martin, 2008) and simplistic methods (Baker, 2023) to show how modern generative models have simplified the task of implementing personas.

The evaluation approach for the models uses human evaluation, which is implemented over three stages. The coherency stage evaluates the persona outputs of the models, rating how well the output matches the persona match. The fluency stage evaluates the conversational outputs of the models, rating how natural the conversation is (Clark et al., 2019). The informativeness stage evaluates the perceived accuracy of model outputs, with humans selecting the persona that best matches an output. All evaluations will be summarised using quantitative analysis.

3.1 Strengths

The strengths of the chosen approach are that it is based on the successes and findings of the ConvAI2 competition (Dinan et al., 2020), it follows current trends in conversational AI research and implementation (Wassan and Ghuriani, 2023), and it uses human-in-the-loop evaluation (Sagiraju, 2022). Building on the successes of ConvAI2 using modern techniques allows the evaluation of progress within conversational AI by identifying how effectively modern models complete this task and proving generative AI is a strong model type for this task. Applying current trends in conversational AI to an older competition allows us to validate the progress made in the field whilst identifying areas for improvement. Conversational AI success is routed in a model's interaction with humans (Wienrich and Latoschik, 2021); therefore, human evaluation provides accurate judgement on how effectively each model represents a persona for the ConvAI2 task (Fiebrink et al., 2011).

3.2 Weaknesses

The weaknesses of the chosen approach are that it naively assumes current trends and the ConvAI2 results represent the best approach, it does not use objective or linguistic evaluation (Jadeja and Varia, 2017), and it does not use a novel approach within conversational AI (Kulkarni et al., 2019). Assuming that the results of ConvAI2 cover all suitable models and are representative of all approaches to persona-based conversational AI is a faulty generalisation (Longoni et al., 2023), as it assumes that all models were implemented most effectively and are representative of the task as a whole. Relying solely on human evaluation ignores objective evaluation metrics using mathematical approaches, such as those defined by Bandi et al. (2023) for generative AI; for the validation of models in future studies, Bilingual Evaluation Understudy should be implemented as an evaluation metric. The lack of a novel approach means this study only verifies current advances in conversational AI rather than advancing the models used in this field.

4 Technical Background

When selecting an LLM, the key factors influencing this choice are parameters, training data, and architecture (Naveed et al., 2023). Parameters are the weights and biases that determine a model's behaviour (Jurafsky and Martin, 2008); weights are numerical values that define strength between neurons in the model, with biases being numerical values added to weights to control the output of neurons. The more parameters within a model, the better it can represent the patterns of the language. Training data is the data used to train the model; the training data used directly impacts the patterns that can be identified by the model, with diverse and representative training data resulting in improved performance (Naveed et al., 2023). A generative model's architecture refers to the attention mechanisms, neural networks, and regression models used.

When fine-tuning a model, the fundamental methods are prompt engineering, training on additional data, and modifying the hyperparameters (Shin et al., 2023). Prompt engineering is the method of designing inputs to produce optimal results; prompt engineering is performed by ensuring that the input is effectively interpreted by the model and providing context, precision, and scope for the model (Meskó, 2023). Training on additional data provides the model with more context of the task domain, enabling further understanding of the patterns within; in the case of pre-trained models, providing a knowledge base of the task domain directly improves performance (Nayak and Timmapathini, 2023). Modifying hyperparameters, such as temperature, top p, and top k, allows a pre-trained model to be adapted to different task domains (Tribes et al., 2023); Liao et al. (2022) discuss the impacts of hyperparameter tuning, identifying that the same changes across different models can have significantly different effects on model performance.

The two model families used in this research are LLaMa-2 (Touvron et al., 2023a) and GPT-3.5 (Floridi and Chiriatti, 2020). Xuanfan and Piji (2023) implemented these models in a similar task; in this task, they found that GPT-3.5 performed best across a wide variety of tasks due to the size of its parameters and training data, while LLaMa provided the richest outputs. Xuanfan and Piji's research (2023) found that the success of LLMs in natural language generation tasks could be predicted through the number of parameters and training data implemented by the model, with an increase in these values resulting in increased performance; due to this, two LLaMa models have been implemented within this research to verify Xuanfan and Piji's findings (2023). Xuanfan and Piji (2023) analysed the distinctness of outputs by quantifying the number of distinct N-grams present; using this metric, LLaMa was ranked highest in the distinctness of outputs. Fine-tuning the temperature of the models may allow for distinctness to be further explored, as Xuanfan and Piji (2023) did not identify the temperature values utilised by the models.

The LLaMa models used are LLaMa-2-7B-Chat[1], LLaMa-2-13B-chat[2], and LLaMa-2-70B-Chat[3]. The difference between these models is the number of parameters implemented, with 7B, 13B, and 70B representing the number of parameters in billions. LLaMa[4] is a family of open-source LLMs developed to be easily retrained and fine-tuned. LLaMa-2[5] is the most recent model generation, which has various available model sizes and is fine-tuned for chat usage. LLaMa-2 has two trillion pre-training tokens and a context length of 4096 tokens. LLaMa-2 performed better than other pre-trained models in multi-task language understanding benchmarks (Touvron et al., 2023b). LLaMa is a foundational model, which means that it is designed to be versatile and applied to different task domains rather than a fine-tuned model for a specific task.

The GPT-3.5 model used is gpt-3.5-turbo-1106[6], which has over one hundred and seventy-five billion parameters. GPT[7] is a family of proprietary generative LLMs developed to be easily fine-tuned and shared. GPTs are accessed via the OpenAI API[8] or ChatGPT[9], with gpt-3.5-turbo-1106 being the default implementation. GPT-3.5 has over three hundred billion pre-training tokens and a context length of 16385 tokens. GPT-3.5 benchmarks at the same level as the best few-shot LLM (OpenAI et al., 2023). GPTs are designed to generate human-like text and aim to be used for general purposes within NLP (Floridi and Chiriatti, 2020).

[1]huggingface.co/meta-llama/Llama-2-7b-chat
[2]huggingface.co/meta-llama/Llama-2-13b-chat
[3]huggingface.co/meta-llama/Llama-2-70b-chat
[4]ai.meta.com/blog/large-language-model-llama-meta-ai/
[5]ai.meta.com/blog/llama-2/
[6]platform.openai.com/docs/models/gpt-3-5
[7]openai.com/blog/introducing-gpts
[8]openai.com/blog/openai-api
[9]chat.openai.com/

5 Design Decisions

The development tools used for this application were Python 3.9[10], llama-cpp-python[11], LLM[12], and Streamlit[13]. Python was chosen as the programming language due to the wide availability of powerful toolkits and modules for natural language processing available to it (Thanaki, 2017); Python 3.9 was selected as it is compatible with LangChain[14], with previous iterations of the application implemented ConversationChains. llama-cpp-python is a Python wrapper for llama-cpp[15] that allows LLaMa models to run on a local machine using 4-bit integer quantisation; using llama-cpp-python allows the application to run locally stored models on a variety of machines. LLM is a Python library that enables LLMs to be accessed through remote APIs, allowing access to gpt-3.5-turbo-1106 via the OpenAI API. Streamlit is a framework that enables the development of interactive data apps within Python, providing UI elements that can interact with machine learning functions.

The LLaMa-2 and GPT-3.5 models are implemented using different methods within the application. For the implementation of LLaMa-2, pre-quantised models were implemented; pre-quantised models allow the models to be implemented faster and ensure consistency in the models utilised by the application when installed on multiple machines. The pre-quantised models implemented were produced by TheBloke[16]. The LLaMa-2-70B-Chat model was not implemented within the application, as the required RAM exceeded 32GB. For the implementation of GPT-3.5, the OpenAI API was used to communicate with ChatGPT via the LLM library. GPT-3.5 models are proprietary and not available for download, meaning that the OpenAI API is currently the only method for communication with the model; the model accessed via the API is the default implementation of gpt-3.5-turbo-1106.

The models were fine-tuned through prompt engineering and hyperparameter tuning. Additional training data was not used for fine-tuning, as the PersonaChat training dataset[17] was not in a format suitable for LLaMa-2 or GPT-3.5. The LLaMa-2 model can be fine-tuned by modifying the temperature, top p, top k, repetition penalty, and maximum token length hyperparameters; within the application, users can change these values with the default values being those recommended for 'Creative' responses by the LocalLLaMa community 2023. The model prompt was developed using the LLaMa-2 prompt template[18], with the system message fine-tuned for the best baseline outputs; the LLaMa-2 prompt was also suitable for the GPT-3.5 model, so it shares the same prompt template within the application. The PersonaChat dataset was modified to suit the prompt better by replacing all periods with commas due to how LLaMa-2 interprets periods.

The evaluation approach was designed to be simple and fast for users while gathering quantitative data. The strategy gets users to evaluate key criteria for chatbots based on the Liang and Li's research 2021; the fluency metric combines Liang and Li's 2021 readability and naturalness criteria, and the coherency metric combines the relevance and consistency criteria. The informativeness metric is the success rate of a user identifying the AI's persona, which depends on the information presented to the user. As surveys allow fast user feedback (Hill, 2013), the evaluation section was created as a short survey that implemented three closed-ended questions for ease of response and high user acceptance (Andrews et al., 2003). The gamification(Harms et al., 2015) of the Informativeness metric, by notifying users how many personas they identified correctly during the study, increased evaluation engagement.

6 Implementation

The implementation[19] was through a Streamlit application, which can be installed via GitHub. The application implements the PersonaChat dataset[20] and, on each execution, selects a random persona from the data set for AI to carry out. The application allows users to converse with AI through a chat box and to evaluate current AI through a short survey. Users can select different LLMs to interact with; users can modify the hyperparameters of the LLaMa-2 models.

[10] python.org/downloads/release/python-390/
[11] github.com/abetlen/llama-cpp-python
[12] llm.datasette.io/en/stable/
[13] streamlit.io
[14] langchain.com
[15] github.com/ggerganov/llama.cpp
[16] huggingface.co/TheBloke
[17] huggingface.co/datasets/bavard/personachat_truecased
[18] gpus.llm-utils.org/llama-2-prompt-template
[19] github.com/jackjburnett/PersonaChat
[20] kaggle.com/datasets/atharvjairath/personachat

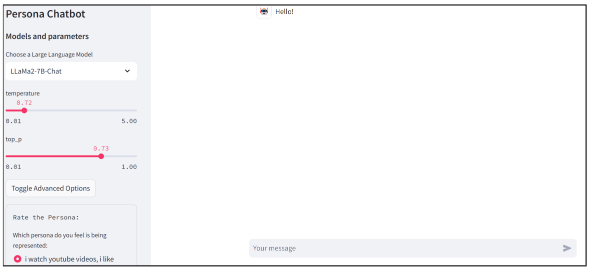

Figure 1: PersonaChat Dashboard

The PersonaChat dashboard, see figure 1, comprises llama2local.py, ChatGPT.py, PersonaList.py, Persona_Chat.py, and evaluation.py. llama2local.py utilises llama-cpp-python to interact with the local LLaMa models. The function within llama2local takes the currently selected model, the user's hyperparameters, and the prompt with chat history and calls the LLaMa-2 model with these parameters; these values are passed to the model from Persona_Chat.py. The result of the LLaMa-2 call is stored in a text file to enable debugging, and the 'text' section of the response is returned to Persona_Chat. If the currently selected model is a GPT-3.5-turbo, the function within ChatGPT.py is called; this function uses the LLM library to define the input and output for the model, then uses an OpenAI API key and the prompt to make a request. The ChatGPT response is sent back to Persona_Chat via llama2local. PersonaList.py converts the personas in the dataset into a list and has a function call to select a random persona; these are both invoked by Persona_Chat.

Persona_Chat.py contains the Streamlit dashboard, the session variables, and the functions for each dashboard element. The application uses a list of dictionaries stored as a session variable and the Streamlit library's chat_message function to implement the chat function; the list contains each message with its sender and its avatar. When the user sends a message, a prompt is generated and then passed, along with the current model and hyperparameters, to llama2local; the prompt is generated through a system prompt that implements the persona and the chat history built from the list of messages. The persona is selected through a session variable, which produces a random number, and then a second session variable stores the persona with the index of the random number. The sidebar implements Streamlit's slider objects to enable the user to manipulate the hyperparameters and evaluate the current AI. Three random personas are stored in a session variable to implement the informativeness metric, and then Streamlit's radio object is used to show these to the user.

To save development time, the application was not dockerized[21]; the application must be manually installed on new devices, including conda environment[22] creation. The LLM processing, on average, takes ten seconds but can take upwards of a minute; this can be reduced by utilising different quantisations of the models. The application runs instantaneously as all model processing is run on demand, allowing prompts and hyperparameters to be updated in real time. The OpenAI API can have delays, and during human evaluation, RateLimit errors affected the GPT-3.5 model assessment.

An evaluation dashboard was created for real-time analysis. This dashboard allows models to be selected, which results in the model's average coherency, fluency, and informativeness being displayed; the difference between the model's scores and the mean scores is output. The survey results for a model are displayed, allowing for filtering and sorting. Hyperparameters can be selected, resulting in a graph for each metric being shown; this enables fine-tuning analysis. The evaluation dashboard also compares the correlation matrix of the evaluator metrics for the model with the general metrics.

7 Experimental Evaluation

Three participants were asked to evaluate the ability of three modern LLMs to generate a suitable response given a dialogue history and a description of an agent's persona. The participants were given guidance on how to rate the response, with the participants advised to evaluate the AI as they would an actor undertaking the role. The scores are still subjective because they rely on the individual judgement of participants. The goal is to identify how well the AI portrays personas, hypothesising that the AI will score an average of 7.5 or higher in coherency and informativeness. Participants were given 30 minutes to converse with as many AI personas as possible, with 10 minutes initially allocated for each model; however, RateLimit errors prevented two participants from interacting with GPT-3.5. Participants were advised to tweak the hyperparameters between personas. Two participants opted to extend the experiment by continuing to interact with the AI past the allotted time.

[21] docker.com/resources/what-container

[22] conda.io

Over a cumulative time of three hours, 68 conversations were completed and evaluated within the application; the number of conversations was due to the LLaMa-2-7B-Chat model taking, on average, 54 seconds and the LLaMa-13B-Chat model taking, on average, 134 seconds to respond to each message. There were seven recorded conversations for GPT-3.5-turbo-1106, 20 for LLaMa2-13B-Chat, and 41 for LLaMa2-7B-Chat. A summary of the results of the human evaluations can be seen in table 1.

	Coherency	**Fluency**	**Informativeness**
GPT-3.5-turbo-1106	*1.8571*	***9.7143***	*1.4286*
LLaMa-7B	*8.1220*	*6.2195*	*9.5121*
LLaMa-13B	***8.25***	*6.05*	***10.0000***

Table 1: Scores from human evaluation

The results summary in table 1 shows that the LLaMa models performed similarly to each other with high informativeness and coherency. In contrast, the GPT model performed poorly in these metrics but excelled in fluency. The GPT model's low coherency and informativeness were due to safeguards that prevent the AI from expressing opinions, which basic prompt engineering cannot bypass (Deng et al., 2023). The high fluency of GPT-3.5 is likely due to the large number of parameters and the training data, which allows for the model to model human conversations accurately; the findings of this evaluation align with findings that professionals struggle to distinguish ChatGPT outputs from human outputs (Herbold et al., 2023).

While GPT-3.5 was unsuitable for undertaking a persona due to the safeguards in place, the LLaMa models excelled at producing outputs in line with given personas. LLaMa-13B had a mean of 10 for informativeness, meaning that humans correctly identified the persona being portrayed within all conversations; LLaMa-13B also scored highest in coherency, a metric evaluating how well the AI performed the persona, but lowest in fluency. LLaMa-7B achieved a higher fluency score than LLaMa-13B but had lower coherency and informativeness. During the evaluation stage, 3 conversations scored 10 for all metrics; 2 were with LLaMa-13B, and 1 was with LLaMa-7B. The whole conversation log can be found in the appendix. Across the models, there is a negative correlation between fluency and the other two metrics; figure 2 shows the correlation matrix for the metrics.

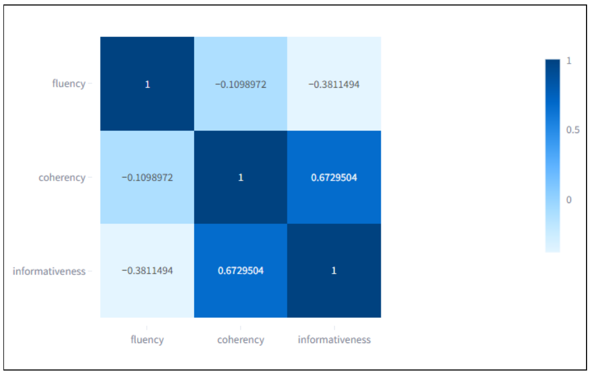

Figure 2: Metric Correlation Matrix

It can be seen in figure 2 that coherency and informativeness are positively correlated, while fluency is negatively correlated with both metrics. The negative relationship between coherency and fluency is likely due to people perceiving natural conversation, especially introductions, to be low in information; Atir et al. (2022) discuss this perception and how individuals undervalue how much they learn in social situations. The positive correlation between coherency and informativeness is likely due to a persona being selected correctly if the conversation represented the persona well; during the study, from discussion with the participants, the critical differences between scores of 7 and higher were the AI's mannerisms.

The initial hypothesis was that the AI would achieve a score of 7.5 or higher in informativeness and coherency, which was performed by the models that did not utilise safeguards; however, the suitability of modern AI for implementing personas is not as high as predicted due to fluency issues. While AI can accurately portray a persona, it does not accurately mimic natural conversation; this means that while AI is suitable for chatbots with personas, it is not yet practical to enact the persona in the place of a human agent. A mistake the AI occasionally made was acting out the conversation on behalf of both agents and hallucinating a dialogue between the two agents, including hallucinating that the user had provided information before the discussion; this behaviour is widely recorded in modern LLMs (Rawte et al., 2023).

The application also recorded the hyperparameters utilised for each conversation, identifying how they affected the performance of the LLaMa models. During the study, the human evaluators modified the temperature and top_p values for the two models.

	Fluency	Coherency	Informativeness
temperature	-3.216702559378415	0.8697541870907521	5.544785010450808
top_p	-7.571587348965469	6.16566134975519	29.36631174319159

Table 2: LLaMa2-7B-Chat Regression Coefficients

Table 2 shows the coefficients of a multiple regression model fitted to the hyperparameters with each metric as the target variable. Higher temperatures and top_p resulted in lower fluency but higher coherency and informativeness; temperature controls how deterministic a model is by adding randomness to its probability distribution for tokens, while top_p adds a threshold for acceptable token probability (Ouyang et al., 2023). As higher top_p values result in less probable tokens being output, the outputs will become less likely and potentially less understandable, affecting fluency; conversely, a higher top_p allows for the tokens that personas may contain, which are less probable in everyday conversation to be accepted as outputs, affecting coherency and informativeness. Bianchi et al. (2020) discuss the need for predictability in natural language, which explains why the randomness introduced by temperature hurts fluency.

8 Conclusions

This study developed a method for using human evaluation to analyse the performance of LLMs in undertaking a persona. The method includes a system to load and modify local LLM models and an evaluation dashboard. The performance of LLaMa and GPT-3.5 models was evaluated through the study. It was found that modern LLMs are suitable for presenting as a persona, though portraying the persona to a human agent. Still, they struggled to perform this task using a natural flow of conversation. The study also identified that temperature and top_p have predictable impacts on the perceived naturalness of language and the ability to present as a persona, with high temperature resulting in lower fluency and higher coherency and top_p having an inverse effect. There was a negative correlation between fluency and coherency, but this is likely due to lower parameter LLMs being unable to model the nuances of natural language while also providing information from the personas; further studies using higher parameter models would give more insight into how the number of parameters affect fluency and coherency.

8.1 Further Implementation

The system developed could be implemented as a feedback loop for LLMs used within interactive AI or as a basis for creating 'life-like' conversational agents through personas. A training dataset should be provided for a more engaging agent to identify how the persona should act. This form of AI would be beneficial for human-companionship applications and entertainment mediums. Implementing AI with personas within video games, which the user can converse with naturally rather than through pre-created options, is an area of current research interest (Karaca et al., 2023). Inworld[23] has begun developing systems that use LLMs to enable natural conversation with NPCs in video games.

van der Zon (2023) discussed how AI with personas could benefit storytelling by providing a medium for authors to interact with their characters; this system could be implemented by providing a simplistic method of inputting personas as prompts, then providing methods of fine-tuning the prompt as the character develops. Shao et al.'s study (2023) created a system that partially enables van der Zon's (2023) goals.

8.2 Future Studies

Modern LLMs are suitable for implementing personas, but the outputs are perceived as not highly natural conversations; however, this raises the question of whether outputs can be coherent and highly natural. If natural conversation relies on being predictable, does that not infer that adding person-specific information, which breaks predictability, is deemed unnatural? This is not the case, as research has identified mutual interests as a strong identifier of natural conversation (Nguyen et al., 2015). Studying how to achieve the best fluency and coherency within LLMs with persona is a task of fine-tuning and linguistic analysis. Repeating this study with more individuals over a longer time frame will allow for better analysis of hyperparameter fine-tuning while also providing further insight into what makes an output both fluent and coherent regarding the persona.

[23] npc.ai

References

Dorine Andrews, Blair Nonnecke, and Jennifer Preece. 2003. Electronic survey methodology: A case study in reaching hard-to-involve internet users. *International journal of human-computer interaction*, 16(2):185–210.

Stav Atir, Kristina A Wald, and Nicholas Epley. 2022. Talking with strangers is surprisingly informative. *Proc Natl Acad Sci U S A*, 119(34):e2206992119.

Pam Baker. 2023. *ChatGPT for dummies*. Wiley.

Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. 2023. The power of generative ai: A review of requirements, models, input-output formats, evaluation metrics, and challenges. *Future Internet*, 15(8).

Samy Bengio, Roman Garnett, Nicolò Cesa-Bianchi, Kristen Grauman, Hugo Larochelle, and Hanna Wallach, editors. 2019. *Advances in neural information processing systems*, volume 31. Curran Associates, Inc.

Bruno Bianchi, Gastón Bengolea Monzón, Luciana Ferrer, Diego Fernández Slezak, Diego E. Shalom, and Juan E. Kamienkowski. 2020. Human and computer estimations of predictability of words in written language. *Scientific Reports*, 10(1):4396.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

LocalLLaMA Community. 2023. What model parameters is everyone using?

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Masterkey: Automated jailbreak across multiple large language model chatbots.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.

Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 147–156, New York, NY, USA. Association for Computing Machinery.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners.

D Gong, X Wan, and D Wang. 2023. Working memory capacity of chatgpt: an empirical study.

Bin Guo, Hao Wang, Yasan Ding, Wei Wu, Shaoyang Hao, Yueqi Sun, and Zhiwen Yu. 2020. Conditional text generation for harmonious human-machine interaction.

Johannes Harms, Stefan Biegler, Christoph Wimmer, Karin Kappel, and Thomas Grechenig. 2015. Gamification of online surveys: Design process, case study, and evaluation. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part I 15*, pages 219–236. Springer.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13(1):18617.

Paul Hill. 2013. Real, fast, feedback. *The Journal of Extension*, 51(1):5.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Rahib Imamguluyev. 2023. The rise of gpt-3: Implications for natural language processing and beyond. *International Journal of Research Publication and Reviews*, 4:4893–4903.

Mahipal Jadeja and Neelanshi Varia. 2017. Perspectives for evaluating conversational ai.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. Pearson.

Yasemin Karaca, Djameleddine Derias, and Gozde Sarsar. 2023. Ai-powered procedural content generation: Enhancing npc behaviour for an immersive gaming experience. page 10.

Aleksey Korshuk. 2019. Persona-chat.

Pradnya Kulkarni, Ameya Mahabaleshwarkar, Mrunalini Kulkarni, Nachiket Sirsikar, and Kunal Gadgil. 2019. Conversational ai: An overview of methodologies, applications & future scope. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–7.

Fabian Landwehr, Erika Varis Doggett, and Romann M. Weber. 2023. Memories for virtual AI characters. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 237–252, Prague, Czechia. Association for Computational Linguistics.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

Hongru Liang and Huaqing Li. 2021. Towards standard criteria for human evaluation of chatbots: A survey.

Lizhi Liao, Heng Li, Weiyi Shang, and Lei Ma. 2022. An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. *ACM Trans. Softw. Eng. Methodol.*, 31(3).

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3?

Chiara Longoni, Luca Cian, and Ellie J. Kyung. 2023. Algorithmic transference: People overgeneralize failures of ai in the government. *Journal of Marketing Research*, 60(1):170–188.

Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *J Med Internet Res*, 25:e50638.

Golam Md Muktadir. 2023. A brief history of prompt: Leveraging language models. (through advanced prompting).

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models.

Anmol Nayak and Hari Prasad Timmapathini. 2023. Llm2kb: Constructing knowledge bases using instruction tuned context aware large language models.

Tien T Nguyen, Duyen T Nguyen, Shamsi T Iqbal, and Eyal Ofek. 2015. The known stranger: Supporting conversations between strangers with personalized topic suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 555–564.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2023. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. *Improving Language Understanding by Generative Pre-Training*. OpenAI.

Evani Radiya-Dixit and Xin Wang. 2020. How fine can fine-tuning be? learning efficient language models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2435–2443. PMLR.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.

Sujatha Sagiraju. 2022. *The State of AI and Machine Learning*, 8th edition. Appen.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing.

Jiho Shin, Clark Tang, Tahmineh Mohati, Maleknaz Nayebi, Song Wang, and Hadi Hemmati. 2023. Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks.

Cole E. Short and Jeremy C. Short. 2023. The artificially intelligent entrepreneur: Chatgpt, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*, 19:e00388.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks.

Jalaj Thanaki. 2017. *Python natural language processing*. Packt Publishing Ltd.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Christophe Tribes, Sacha Benarroch-Lelong, Peng Lu, and Ivan Kobyzev. 2023. Hyperparameter optimization for large language model instruction-tuning.

Alexander Tselousov and Sergey Golovanov. 2018. Convai2 submission.

Petrus Johannes van der Zon. 2023. Assisting the creation-process of fictional characters with large language models. Technical report, Leiden Institute of Advanced Computer Science.

Jyotsna Talreja Wassan and Veena Ghuriani. 2023. Recent trends in deep learning for conversational ai. In Ph.D. Manuel Jesus Domínguez-Morales, Dr. Javier Civit-Masot, and Mr. Luis Muñoz-Saavedra, editors, *Deep Learning - Recent Findings and Researches*, chapter 13. IntechOpen, Rijeka.

Carolin Wienrich and Marc Erich Latoschik. 2021. extended artificial intelligence: New prospects of human-ai interaction research. *Frontiers in Virtual Reality*, 2.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Ni Xuanfan and Li Piji. 2023. A systematic evaluation of large language models for natural language generation tasks. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 40–56, Harbin, China. Chinese Information Processing Society of China.

Syifa Izzati Zahira, Fauziah Maharani, and Wily Mohammad. 2023. Exploring emotional bonds: Human-ai interactions and the complexity of relationships. *Serena: Journal of Artificial Intelligence Research*, 1(1):1–9.

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

A Application Download

The application can be downloaded from GitHub[24]. The results from human evaluation are stored in 'evaluation.csv'.

B Hyperparameter Impacts

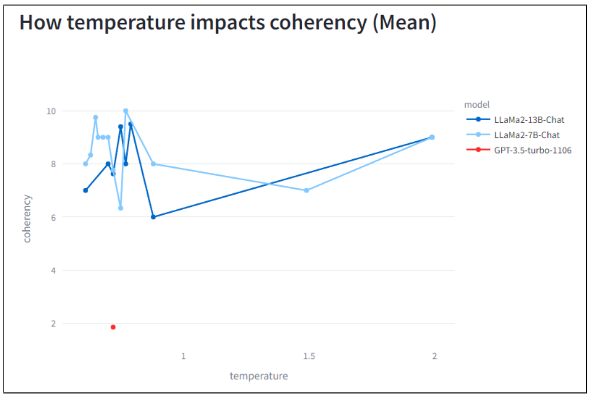

Figure 3: Temperature mean Coherency

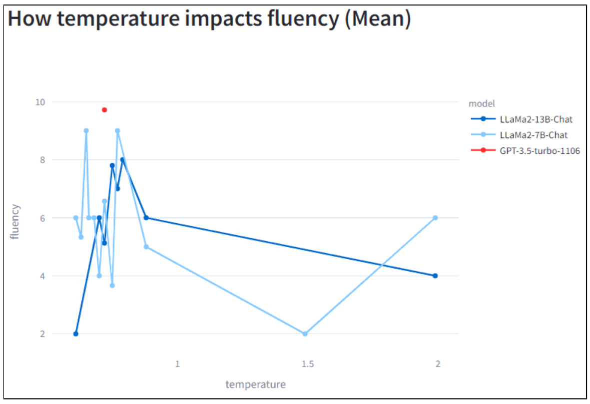

Figure 4: Temperature mean Fluency

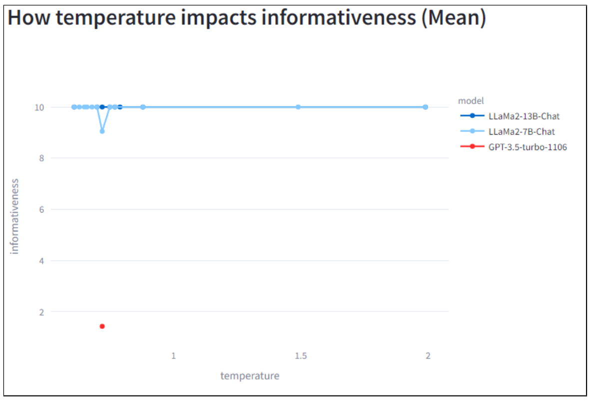

Figure 5: Temperature mean Informativeness

[24] github.com/jackjburnett/PersonaChat

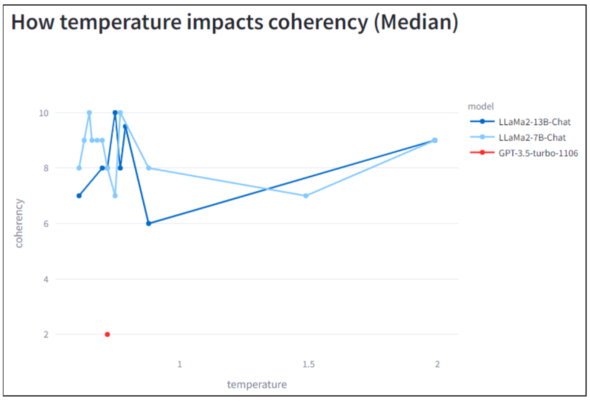

Figure 6: Temperature median Coherency

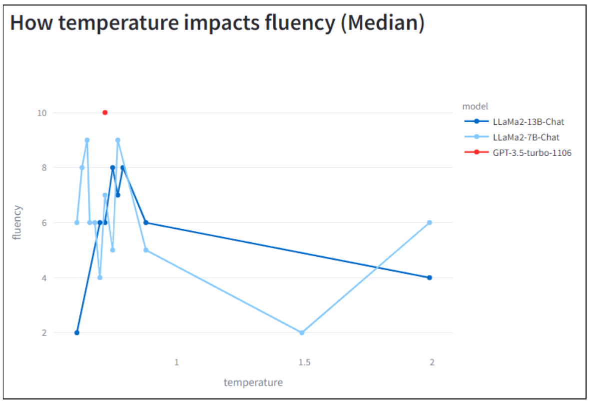

Figure 7: Temperature median Fluency

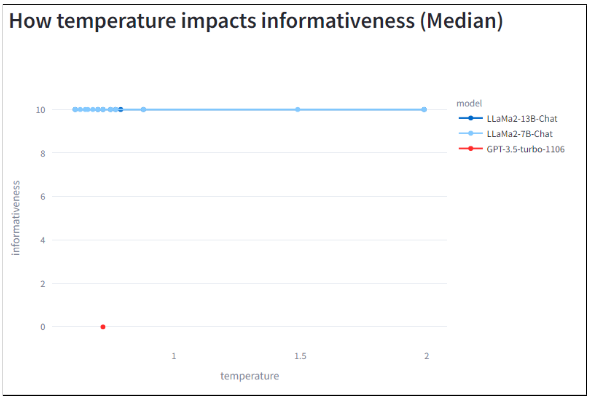

Figure 8: Temperature median Informativeness

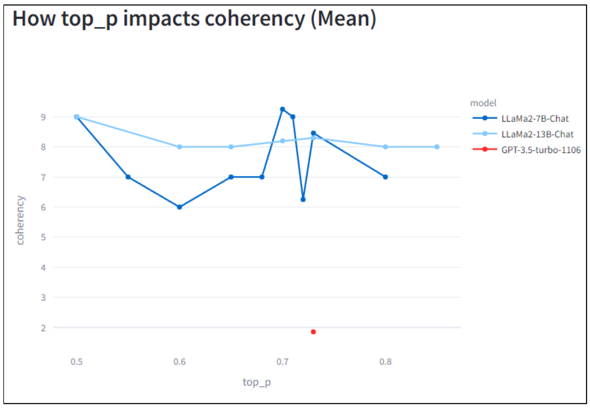

Figure 9: Top_p mean Coherency

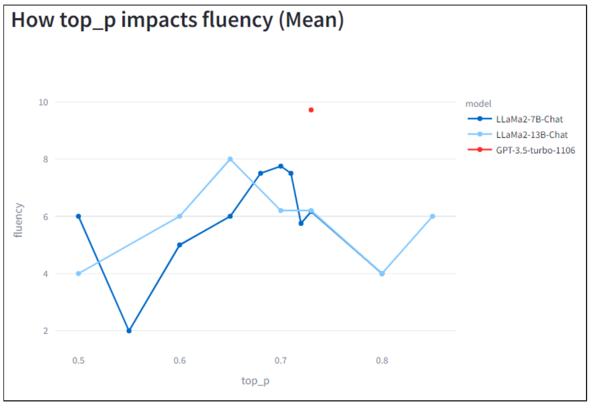

Figure 10: Top_p mean Fluency

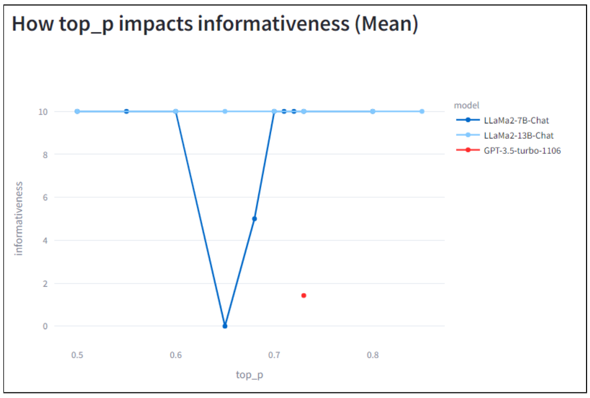

Figure 11: Top_p mean Informativeness

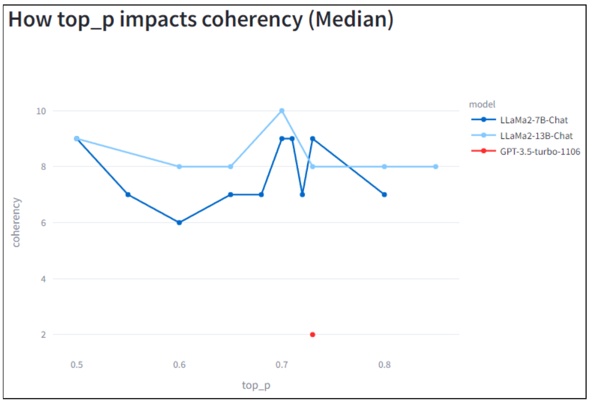

Figure 12: Top_p median Coherency

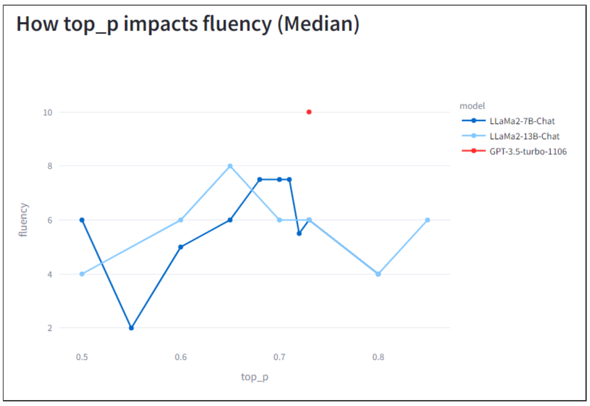

Figure 13: Top_p median Fluency

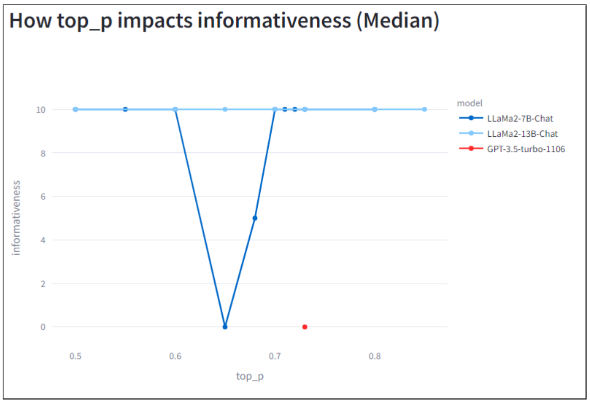

Figure 14: Top_p median Informativeness

C Regression Results

	Coefficient	**Intercept**
temperature	-3.216702559378415	14.02193998655236
top_p	-7.571587348965469	14.02193998655236
top_k	0	14.02193998655236
repetition	0	14.02193998655236
max_length	0	14.02193998655236

Table 3: LLaMa2-7B-Chat Fluency Regression Results

	Coefficient	**Intercept**
temperature	-1.4670241298709343	9.788479063365765
top_p	-2.547394804632442	9.788479063365765
top_k	0	9.788479063365765
repetition	-0.6847958181093106	9.788479063365765
max_length	0	9.788479063365765

Table 4: LLaMa2-13B-Chat Fluency Regression Results

	Coefficient	**Intercept**
temperature	0.8697541870907521	3.0912241012025774
top_p	6.16566134975519	3.0912241012025774
top_k	0	3.0912241012025774
repetition	0	3.0912241012025774
max_length	0	3.0912241012025774

Table 5: LLaMa2-7B-Chat Coherency Regression Results

	Coefficient	**Intercept**
temperature	0.7804318721085917	6.88117978270343
top_p	0.48340380970055236	6.88117978270343
top_k	0	6.88117978270343
repetition	0.36336104634371 58	6.88117978270343
max_length	0	6.88117978270343

Table 6: LLaMa2-13B-Chat Coherency Regression Results

	Coefficient	**Intercept**
temperature	5.544785010450808	-15.508797093072843
top_p	29.36631174319159	-15.508797093072843
top_k	0	-15.508797093072843
repetition	0	-15.508797093072843
max_length	0	-15.508797093072843

Table 7: LLaMa2-7B-Chat Informativeness Regression Results