Instilling Personality in Chatbots

Abstract

Chatbots with human-like characteristics often suffer from context loss, personality dilution, and hallucinations during extended interactions, especially evident in character-driven applications, eroding user trust and engagement. Surface-level attempts to stimulate distinct personas through prompting remain insufficient for achieving deeply ingrained personality modelling.

This paper presents a data-driven approach to address these issues by fine-tuning large language models (LLMs) for personality consistency using GPT-2 as the base model. The objective is to maximize expression of the Big Five personality traits—Agreeableness, Openness, Conscientiousness, Extraversion, and Neuroticism---embedding each directly into the architecture of the model.

The resulting personality-aware LLMs consistently generated fluent, trait-aligned outputs. Outputs were prompted with a standardized phrase and evaluated using Personality_LM (Sun, HuggingFace), with results averaged across three trials to minimize noise. Four of the five models demonstrated substantial improvements (~18%–53%) in trait expression compared to the control model.

Personality-consistent LLMs enable applications in personalized conversational agents, adaptive customer support, career-fit assessment, therapeutic or educational chatbots, and personality-driven content generation.

Keywords: LLM, NLP, AI, personality, persona, personalization, neural network, conversational agent, chatbot

1. INTRODUCTION:

1.1 Background:

Large language models[1a] (LLMs) have revolutionized natural language processing (NLP), enabling machines to generate fluent, human-like language. These advancements have transformed human-machine interactions, facilitating the use of LLMs in communication, creativity and problem solving. These intelligent agent technologies (IAT), commonly known as chatbots, are used for content generation, translation, sentiment analysis, and also as virtual assistants, and more. As an estimate, one-third of all online interactions involve chatbots; this trend is expected to exponentially grow because chatbots can serve as specific work and emotional assistants and enhance business operational efficiency, providing customer service on demand[2].

Additionally, psychology theories like Maslow's Hierarchy of Needs and Self-Determination Theory (SDT), establish the principle that people are likely to trust, engage and respond to others with similar personalities. This principle is wildly adopted in outreach and influencing behaviors[3]. However, this principle is detracted from due to user reported frustrations with chatbots, arising from misunderstood questions, context mix-up and irrelevant or hallucinatory responses. For example, app reviews for Character AI report issues such as the AI's responses being nonsensical, repetitive, and often failing to align with the intended character. Many consumers also feel that the AI's memory is poor, with the bots frequently forgetting details from previous messages[i].

This paper seeks to address the question: can chatbots be engineered to assume an effective personality?. The quest draws on the similarity attraction theory and the five factors of personality model to highlight the role of personality for enhancing human-technology interactions. This study demonstrates that chatbots can be trained to

assume a personality through fine-tuning the later layers of an LLM to create personality-aware chatbots, leading to more personable chatbots, enhancing trust, consistent with research that demonstrates the importance of words and language to influence mood, engagement, and decision-making[4].

1.2 Conceptual Background

1.2.1 Chatbots

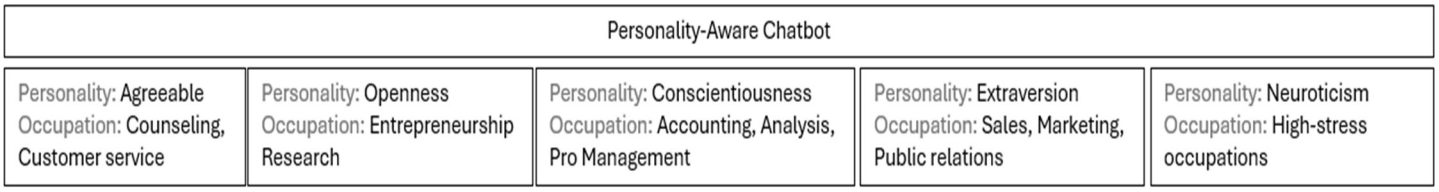
One of the earliest objectives of Artificial Intelligence (AI) was to enable human-machine conversation. A Chatbot or Conversational Agent (CA) is an intelligent agent technology (IAT) that uses natural language processing (NLP) and machine learning (ML) to communicate with users. Over the decades, CAs have evolved through distinct technological eras, from rule-based scripts to sophisticated neural architectures with nuanced personalities.

Currently, **conversational chatbots** simulate human-like interactions through text and speech by leveraging natural language processing (NLP) and machine learning (ML) techniques to interpret user input, generate coherent responses, and perform task-oriented functions. Recent advancements have enabled persona-driven chatbots, in which large language models (LLMs) are fine-tuned to emulate specific traits or roles. Given the self-supervised training paradigm of LLMs, these models can incorporate contextual cues from dialogue histories and condition responses on character-specific attributes. Many systems also incorporate user feedback mechanisms, such as rating model outputs, to iteratively refine tone and content. IATs further extend chatbot functionality by enabling interactivity with users, environmental scanning, collaborative decision support, information retrieval, and adaptive learning from past interactions[5]. While such systems can automate information delivery and query resolution, they also augment human interaction by generating recommendations and context-sensitive suggestions.

Commercial platforms such as Character.ai, Talkie, and Replika illustrate the application of LLMs for persona simulation, allowing users to design and engage with bots that mimic real or fictional identities. For example, Character.ai implements a GPT-based framework with text and voice interfaces, enabling users to create, share, and interact with customized agents. However, these systems primarily operate by instructing the underlying LLM to *act* as a character through prompt conditioning, rather than embedding personality at the model level.

This design choice leads to well-documented issues of quality degradation in long conversations, including **context loss, repetitiveness, and personality drift**. Chatbots often fail to retain previously supplied user information, requiring repeated inputs and reducing conversational fluidity. Over extended dialogues, responses frequently converge to minimally varied repetitions, independent of the intended persona. Most critically, persona fidelity diminishes over time, undermining the distinctiveness of character-driven interaction.

This investigation argues that such limitations stem from the reliance on surface-level prompt engineering. Instead of constraining an LLM to "act" like a character, more robust outcomes may be achieved if the model is fine-tuned such that it *is* the character—encoding personality traits, linguistic structures, and vocabulary patterns



directly into its parameters. Fine-tuning the later layers of the model with carefully curated, trait-specific corpora

can embed personality features at a representational level, thereby mitigating context loss, reducing repetitive sequences, and preserving personality consistency throughout long-form interactions.

1.2.2 Big Five Personality Traits

Personality traits are defined as consistent tendencies in an individual's cognition, affect, and behavior that distinguishes them from others and show relative stability across time and context[6], representing broad dispositions that underlie observable patterns of thoughts, emotion, and interaction. The five-factor model (FFM) is considered the most comprehensive taxonomy in personality research as it incorporates all the critical components of personality. The most wildly adopted FFM model is known as Big Five[6b] **(OCEAN)** and includes the following traits.

Agreeableness: refers to a person's tendency to be friendly, compassionate, courteous, tolerant, and co-operative toward others.

Openness: refers to the extent to which a person is open to experiencing a variety of activities, is imaginative, creative, curious, broad-minded and independent.

Conscientiousness: refers to a person's tendency to act in an organized, planned, or thoughtful manner, such individuals are responsible and reliable.

Extraversion: generally associated as socializing with others, being gregarious, assertive, enthusiastic and talkative.

Neuroticism refers to the extent to which a person's emotions are sensitive to the environment, and includes being nervous, insecure or anxious, associated with emotional instability and moodiness.

1.2.3 Benefits of Persona-Specific Language

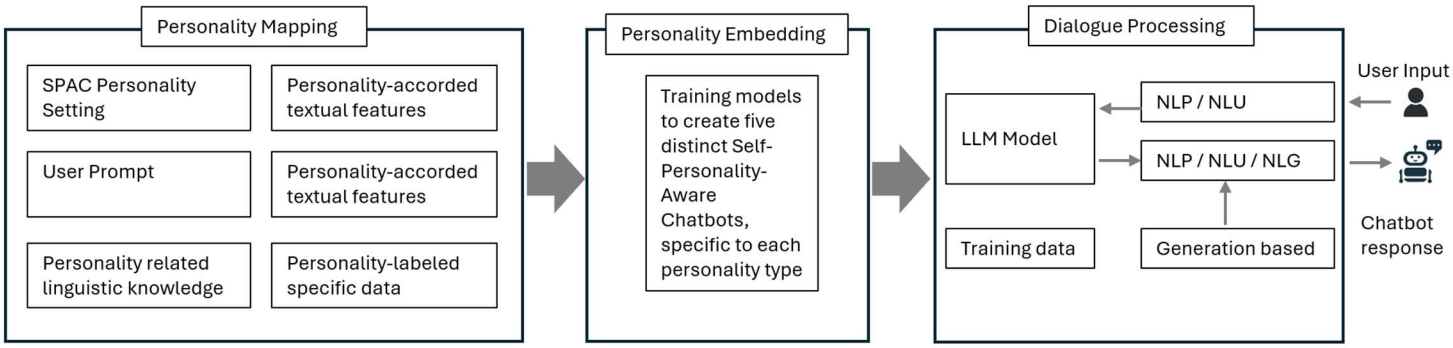
Extant research demonstrates that everyday language use provides stable and reliable indicators of personality traits over time[7]. For instance, individuals high in extraversion often produce greater word counts, favor less complex vocabulary, and employ more socially and emotionally expressive language, reflecting their talkativeness and reduced concern for linguistic precision[8]. Much of this work has relied on lexicon-based approaches, where linguistic features are quantified across semantic categories and associated with personality dimensions[9]. A prominent tool for this methodology is *Linguistic Inquiry and Word Count* (LIWC), which has been widely validated for inferring personality from text[10]. Importantly, personality prediction grounded in the Five-Factor Model (FFM) has demonstrated robustness across demographic boundaries—including culture, language, gender, and occupation—when applied to lexical features of language[11]. Moreover, analyses that emphasize lexical content (the words individuals choose) rather than stylistic markers (how language is delivered) have proven especially reliable for personality inference[12].

This body of research is directly relevant to the present study, which extends personality–language associations from observational human data to generative systems. By fine-tuning LLMs on trait-specific corpora, this work operationalizes the principle that linguistic patterns encode personality, embedding such patterns into model parameters rather than merely observing them. In doing so, it bridges psycholinguistic findings with computational methods, demonstrating how stable personality markers in human language can be leveraged to engineer trait-consistent chatbot behaviors. When applied ethically, this integration has the potential to enhance personalization in AI-human interaction, improve user alignment, and enable differentiated conversational agents[13].

2. Research Method

Training large language models (LLMs) from scratch is computationally expensive, requiring massive corpora and specialized infrastructure. For example, GPT-3 was trained on hundreds of billions of tokens using thousands of GPUs over weeks[14]. For most researchers, this cost is prohibitive, making **fine-tuning** of pretrained models the practical solution for domain-specific tasks[15]. Fine-tuning leverages the general linguistic representations learned during large-scale pretraining while introducing task- or trait-specific supervision, thereby enabling scalability across diverse applications without the need for full retraining.

Pretrained models such as GPT-2[16] are trained on terabytes of open-domain data and thus capture broad statistical regularities of language, including lexical distributions, syntactic structures, and discourse-level coherence. The transformer architecture underlying GPT-2[17] represents input sequences as embeddings of **tokens**—subword units derived via Byte Pair Encoding[18] -and processes them through stacked self-attention and feedforward layers. In this context, **parameters** denote the learned weights of the model, while **hyperparameters** (e.g., number of layers, hidden size, learning rate, batch size) define the configuration of the training process. Importantly, tokenization is not a security procedure but a mathematical decomposition of raw text into atomic units suitable for vector representation and optimization.



Deep learning[17b] enables models to identify probabilistic relationships in unstructured data. Self-supervised pretraining employs an autoregressive objective in which the model predicts the probability of the next token given the preceding sequence. Through exposure to large-scale corpora, LLMs acquire the ability to represent not only the structural features of language but also pragmatic markers such as formality, sentiment, and personality cues[19]. Although pretrained Fine-tuning addresses this limitation by adjusting the weights of a pretrained LLM with trait-specific corpora, thereby operationalizing psycholinguistic evidence that stable personality characteristics are reflected in language use[20]. Practically, this process focuses optimization on the later layers of the network, which are most responsible for shaping output style and lexical choice[21]. By constraining training to curated datasets aligned with the Big Five dimensions—Agreeableness, Openness, Conscientiousness, Extraversion, and Neuroticism—the model parameters are shifted to embed these traits directly, rather than relying on surface-level prompting to simulate them.

Thus, in this study, GPT-2 was fine-tuned on personality-aligned corpora using Hugging Face's Transformers library[22]. The approach builds on established psycholinguistic findings[23] and integrates them with computational methods to yield trait-consistent LLMs capable of producing outputs that maintain personality coherence over extended interactions. This methodology illustrates how the interplay between linguistic psychology and transformer-based architectures can be exploited to create stable, personality-aware conversational agents.

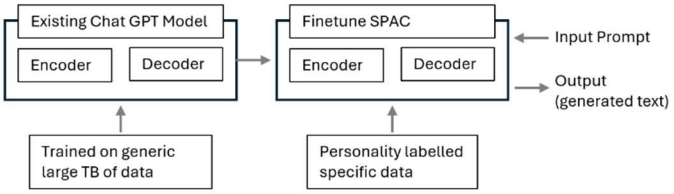
2.1 Framework and Key Techniques

To generate individual self-personality-aware chatbot (SPAC), I used a framework, as shown in figure, with three fundamental functional parts; personality mapping, personality embedding, and dialogue processing. The dialogue processing is common to all chatbots. As the bot receives user's input, the dialogue processing will analyze the input with a module of natural language understanding (NLU) and generate the corresponding output with a module of natural language generation (NLG).

For my research, I have used the data-driven approach, which relies mainly on training the LLMs on small-scale personality labelled specific training data. This adjusts the pretrained NLG to become personality embedded chatbot.

For my research study, I trained and created five distinct SPACs. I created a scientific procedure to be able to methodically evaluate the effectiveness of fine-tuning each model to a personality.

1. Curate Input text: Find and curate dataset to fine-tune LLMs for each of the different personality traits. The most difficult part of my experiment was finding and curating domain-specific usable large datasets with content catering to specific personality traits. Preparing the input for correct formatting to ensure uniformity and truncation during batch preparations was time-consuming, arduous, and iterative.
2. Train models: For each of the traits, I fine-tuned the GPT-2 (accessed from HuggingFace) on the established input text.
3. Generate text: I generated text (~500 characters) on a fixed prompt and repeated the exercise for all five personality traits.
4. Evaluate text: I put the generated text sections through Kevin Sun's Personality_LM[23b] algorithm to measure the percentage occurrence of each of the traits in each of the generated texts. Each text was evaluated three times to reduce error margins.
5. Draw conclusions: Evaluate how successful fine-tuning the models was at instilling a consistent personality into the generated texts.



2.2 Accuracy Function:

I fine-tuned GPT-2 under a causal language modeling (CLM) objective. In CLM, the model learns[24a] to predict the probability of the next token given a sequence of preceding tokens. Formally, training minimizes the cross-entropy loss:

$$L(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log P_0(x_t \mid x_{<t})$$

where x_t is the current token, $x_{<t}$ denotes the input context, and $\boldsymbol{\theta}$ represents the model parameters. This setup allows the model to capture sequential regularities in language, enabling coherent generation when conditioned on an input prompt. The equation assigns an "accuracy" to a sequence/set of words to optimize accurate output.

2.3 Training Configuration:

Fine-tuning was implemented using the Hugging Face Transformers library (Wolf et al., 2020). Training hyperparameters were set as follows: epochs = 3, batch size = 4, learning rate = 5e-5, with a linear decay schedule and warmup ratio of 0.1. The AdamW optimizer[17a] was used with default β-values (β_1=0.9, β_2=0.999) and ε=1e−8. Weight decay was set to 0.01, and gradient clipping was applied with a maximum norm of 1.0. Mixed-precision training (fp16) was enabled to reduce memory overhead.

Model checkpoints were saved every 1,000 steps, with retention limited to two checkpoints to conserve disk space. Training typically required ~30 hours per trait model.

Epochs: 3 / Batch size: 4 / Steps Checkpoint: 1000 / Checkpoint Limit: 2

In practice, parameters refer to the weights and biases within the transformer blocks (e.g., self-attention matrices, feedforward layers), while hyperparameters define training dynamics (e.g., learning rate, batch size, maximum sequence length) and architecture scale (e.g., number of hidden layers, hidden size, number of attention heads). By holding hyperparameters constant across trait-specific experiments, observed differences in output behavior can be attributed primarily to the fine-tuning data, ensuring methodological validity[17, 22].

2.4 Tokenization, Padding, and Masking

GPT-2 employs **Byte Pair Encoding (BPE)**[18] for tokenization, which iteratively merges the most frequent pairs of characters into subword units. This approach balances efficiency and coverage by avoiding an excessively large vocabulary while still capturing rare or novel words through subword composition. Each token is then mapped to an embedding vector that serves as input to the transformer architecture.

During training, sequences were truncated to 128 tokens; shorter sequences were padded with **[PAD]**. However, GPT-2 does not natively include a **padding token**. Padding tokens are intended to standardize the length of a subword for ease and consistency of training. The lack of **[PAD]** creates difficulties when batching variable-length text sequences for GPU training. To resolve this, we introduced a custom **[PAD]** token and extended the embedding matrix. Padding ensures that all sequences in a batch have equal length, but it must not influence the learning process. If the model treats the **[PAD]** as a word, it may learn then produce nonsense. We therefore applied two safeguards:

1. **Attention masks** excluded padded positions from self-attention calculations, preventing the model from attending to artificial symbols.
2. **Loss masking** was applied by assigning −100 to padded labels, ensuring no gradient updates were back-propagated from padding positions.

This combination enabled efficient mini-batch training while preserving semantic fidelity of the data. The approach was particularly important for conversational corpora, where input lengths vary substantially, and represents a methodological contribution for fine-tuning GPT-2 on specialized datasets.

2.5 Hardware and Environment

Experiments were conducted on a single NVIDIA RTX 3090 GPU (24 GB VRAM) with 128 GB system RAM. The software stack consisted of [32]**Python 3.10, PyTorch 2.0, CUDA 12.1, cuDNN 8.9**, and Hugging Face Transformers v4.39. Training scripts were executed in a Linux environment (Ubuntu 22.04). To ensure reproducibility, random seeds were fixed across training, evaluation, and generation steps [insert reproducibility reference].

2.6 Datasets and Preprocessing

Trait-specific corpora were curated to align with the Big Five dimensions. Corpus sizes after preprocessing ranged from **~8,000 to ~15,000 tokens** per trait. To reduce imbalance across traits, corpora were normalized to a comparable token count through subsampling and truncation.

Preprocessing involved:

- Deduplication of repeated lines.

- Normalization of punctuation and casing.
- Retention of stopwords (important for stylistic markers).
- Shuffling at the document level prior to batching.

The data was split 90/10 into training and validation sets. Maximum sequence length was set to 128 tokens, consistent with typical fine-tuning practice for stylistic datasets.

3. Experiment:

The success of LLM fine-tuning depends strongly on the relevance and representativeness of the training corpus. To evaluate the impact of trait-specific fine-tuning, I constructed five *SPACs* each optimized for one of the Big Five personality traits. A **comparative baseline** was also established using GPT-4 (OpenAI, 2023), the successor of GPT-2, to assess whether observed increases in trait alignment were attributable to fine-tuning rather than inherent improvements in newer, larger models.

Trait-Specific Models:

Agreeableness Model: Trained on a [24]Cornell University corpus of Wikipedia talk page requests, capturing polite, cooperative, and service-oriented dialogue patterns.

Openness Model: Trained on ~50,000 characters of [25]*Alice's Adventures in Wonderland* (Carroll, 1865), selected for whimsical narrative structures and creative language use.

Conscientiousness Model: Trained on technical manuals and academic instructional texts, emphasizing organizational markers ("first," "second," "third") and structured, precise language[26].

Extraversion Model: Trained on transcripts from the Online Speech Bank[27], including Martin Luther King Jr.'s *I Have a Dream* and John F. Kennedy's Inaugural Address, chosen for confidence, passion, and expressive delivery.

Neuroticism Model: Trained on an emotion-labeled Twitter corpus[28], restricted to negative affect tweets, to capture heightened sensitivity and emotionally charged language.

3.1 Evaluation:

To standardize generation across models, a fixed prompt was used:

"I would describe myself as …"

Each model—including the GPT-4 baseline—was prompted with this phrase and asked to generate ~500 characters of output (using the following sampling configuration: **temperature = 0.7, top-p = 0.9, max tokens = 120, repetition penalty = 1.2)**. Outputs were evaluated using the **Personality_LM** tool (Sun, 2023), which quantifies the degree to which text aligns with each Big Five trait. To reduce variance, each generation was assessed three times and the scores averaged. Complete sample generations and corresponding evaluation scores are available and will be provided on request.

3.2 Results and Data Representation

The Table Res. Below reports the averaged Personality_LM scores across three independent trials for each model. The highest-scoring trait for each model is shaded in dark gray.

As shown in (R2/C2) (Row 2 / Column 2), the GPT-4 baseline exhibited its strongest trait expression in **Agreeableness**, consistent with expectations for general-purpose generative models, which often default to

cooperative and polite styles. Following fine-tuning, the Agreeableness (R3/C2), Openness (R4/C3), Conscientiousness (R5/C4), and Extraversion (R6/C5), models each maximized their respective target traits. The Neuroticism model (Row 7) exhibited its highest score in Openness (R4/C3), but Neuroticism (R7/C6) was the second-highest trait, indicating partial but less dominant embedding.

Fine-tuned models achieved significant gains relative to the control baseline, with average improvements ranging from **17.7% to 53.6%**, depending on trait. The smallest gain was observed for Agreeableness (+3.9%), likely because the control model already displayed high baseline agreeableness. (Table Res)

Table Res. BIG 5 Major Personality Traits Text Generation Scores

	1	2	3	4	5	6	7
1	**MODEL:**	**Agreeableness**	**Openness**	**Conscientiousness**	**Extraversion**	**Neuroticism**	**%increase from control**
2	**CONTROL**	**0.2386**	**0.2139**	**0.1927**	**0.1671**	**0.1877**	
3	**Agreeableness**	**0.2478**	**0.1667**	**0.1997**	**0.1995**	**0.1864**	**3.86%**
4	**Openness**	**0.1954**	**0.2518**	**0.1895**	**0.1948**	**0.1684**	**17.75%**
5	**Conscientiousness**	**0.1727**	**0.1576**	**0.2413**	**0.2350**	**0.1936**	**25.24%**
6	**Extraversion**	**0.1916**	**0.1519**	**0.2171**	**0.2568**	**0.1826**	**53.63%**
7	**Neuroticism**	**0.1726**	**0.2594**	**0.1643**	**0.1744**	**0.2293**	**22.17%**

3.3 Technical challenges and Error:

While results demonstrate that fine-tuning effectively embedded personality traits into LLM outputs, several technical limitations remain. A key challenge arises from the non-deterministic nature of the Personality_LM evaluator, in which identical excerpts can yield slightly different scores across runs. To mitigate this variance, each generated text was evaluated three times, and average values were reported. This approach revealed consistent trends, but highlights the need for more robust and deterministic evaluation frameworks in future work.

4. Business Applications:

The applications of advanced AI personality computing and natural language processing will pave way for numerous personality-aware chatbot applications. A PAC can be used to aess suitability and candidate personality for different careers[31], a personality assessment via Q&A can help therapists to tailor specific therapies during counselling. PAC can be used as a personal counsel to provide emotional support to students and adults and as an agreeable customer service agent, it can be tailored to help kids learn from a congruent chatbot teacher. In addition, a SPAC could imitate a character or person in language style to evoke the deceased and potentially used to create a form of afterlife. An individual-like PAC could also possibly act as a personal emotional companion for a user. There are countless possibilities for future chatbots and PACs beyond the above mentioned applications.

Currently character bots are used for fun, simply to converse with favorite characters or to explore. And this is an increasingly competitive area, for example, Character AI has taken the conversational AI industry by storm. Released in 2022, it has had excellent growth in number of users and the rate of engagement, even reaching worldwide appeal. Character AI has 233.3 million users worldwide as of April 2024, along with an annual revenue estimated to be $16.7 million as of 2024.[30] There are many emerging character apps, with different ways of generating revenue. Character AI has premium membership with better quality bots. Talkie AI has premium features including voice calls with characters and more chatting features. Chatbot AI is a profitable industry. The ability to create higher-quality bots is potentially revolutionary for this emerging industry, providing an edge to

the creators. The goal of this investigation was to combat the problem of LLM's losing personality by fine-tuning them into a personality, instead of instructing them to 'act' like a personality. As most of the fine-tuned models successfully maximized the personality traits. It is not difficult to create a GPT wrapper[29] like Character AI. For a competitive advantage, or a unique selling point, chatbot companies must have a defining characteristic that can be used as personal accomplices and as well in professional settings.

4.1 Applications in Ethics:

A long-standing debate involving AI in general is AI's ability to make ethical decisions[30]. While this investigation sought to instill a personality into an LLM, in a similar way, emulating ethical decisions from a person can be done with transcripts of their decisions. Fine-tuning the Extraversion model revealed how quickly vocabulary and preferences could be instilled, as the model instantly described itself as republican. For the Neurotic generation, the model instantly described itself as a woman and a child, meaning that with improved memory, the model could remember that identity and make decisions as such an identity would dictate. Notoriously, AI —as the name suggests—fails in capturing the human aspect of intelligence. For humans, there are a variety of subconscious factors that go into decision making, like ethical, moral, and emotional considerations, gained through life experiences and a defined identity. Business, life, and society. But instilling an AI with a consistent personality and the history of a human could significantly improve that, revolutionizing AI as we know it, bringing it closer to 'humanity' than ever before.

CONCLUSION:

The experiment revealed that fine-tuning the later layers of an LLM was effective to maximize a trait as established by the Big Five Personality Traits for Agreeableness, Openness, Conscientiousness, and Extraversion, but not for Neuroticism. Even though the trained LLM output results measure marginally to significantly better than the control results, the results of an experiment suggest that there are still advancements to be made before we can have a chatbot that accurately reflects a human's personality without the aforementioned issues of losing a distinct personality over time. Once we have advancements and appropriate data for hyperparameter tuning, we can produce Chatbots that can be personalized to serve as emotional, social, practical and dependable companions in personal life and in professional settings.

5. BIBLIOGRAPHY:

[1a]Cloudflare. "What Is a Large Language Model (LLM)?" *Cloudflare*, https://www.cloudflare.com/learning/ai/what-is-large-language-model/

[2]Sweezey, M. (2019). Key chatbot statistics to know in 2019. Retrieved March 2, 2020, from August 4 website: https://www.salesforce.com/blog/2019/08/chatbot-statistics.html

[3]Byrne, Clore, & Worchel, 1966: Effect of economic similarity-dissimilarity on interpersonal attraction.

[4]MIT Sloan Brett McFerran, Moore, & Packard, 2019: How Should Companies Talk to Customers Online?

[5]Kumar, Dixit, Javalgi, & Dass, 2016: Research framework, strategies and applications of Intelligent Agent Technologies in marketing

[6]McCrae, R. R., & Costa, P. T. (1987): Validation of the five-factor model of personality across instruments and observers

[6b]University of Liverpool. "The Big Five." *Prosper*, https://prosper.liverpool.ac.uk/postdoc-resources/reflect/the-big-five/

[7]Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003): Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology, 54, 547- 577

[8]Cohen, A. S., Minor, K. S., Baillie, L. E., & Dahir, A. M. (2008): Clarifying the linguistic signature: Measuring personality from natural speech

[9]Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013): Personality, gender, and age in the language of social

[10]Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006) Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life

[11]John, O. P. (1990): The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires

[12]Ireland, M. E., & Mehl, M. R. (2014): Natural language use as a marker of personality

[13] Matz, Kosinski, Nave, & Stillwell, 2017: Psychological Targeting as an Effective Approach to Digital Mass Persuasion

[14]Brown, Mann, Ryder et.all. 2020 Language Models are Few-Shot Learners

[15]Jeremy Howard, Sebastian Ruder: Universal Language Model Fine-tuning for Text Classification

[16] Radford, A., Wu, J., Child, R., et al. (2019): Summary of Research Methods on Pre-Training Models of Natural Language Processing

[17a] Loshchilov & Hutter, 2019 Decoupled Weight Decay Regularization

[17]Ashish Vaswani, Noam Shazeer, Niki Parmar 2017 et.al.. Attention Is All You Need - encoder-decoder configuration

[17b]IBM. "What Is Deep Learning?" *IBM*, https://www.ibm.com/topics/deep-learning

[18]Rico Sennrich, Barry Haddow, Alexandra Birch: Edinburgh Neural Machine Translation Systems for WMT 16

[19]Pennebaker James, a King Linguistic styles: Language use as an individual difference

[19]Schwartz, S. J: Identity development, personality, and well-being in adolescence and emerging adulthood: Theory, research, and recent advances

[20]Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003) Psychological aspects of natural language use: Our words, our selves

[20] Ireland, M. E., & Mehl, M. R. (2014: Natural language use as a marker of personality

[21]Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, Ian Tenney 2020: What Happens To BERT Embeddings During Fine-tuning?

[22]Wolf et.al..2020 Transformers: State-of-the-Art Natural Language Processing

[22]Radford, Alec, et al. "Language Models are Unsupervised Multitask Learners." 2019

[23]Matz, Kosinki, Nave, Stillwell: Psychological targeting as an effective approach to digital mass persuasion

[23b]Wang, Rong, and Kun Sun. "Continuous Output Personality Detection Models via Mixed Strategy Training." *ArXiv*, 2024, https://arxiv.org/abs/2406.16223

[24a]Radford, Alec, et al. "Language Models Are Unsupervised Multitask Learners." 2019

[24]Cornell University. "Wiki Politeness." *Cornell ConvoKit*, https://convokit.cornell.edu/documentation/wiki_politeness.html.

[25]Carroll, Lewis. *Alice's Adventures in Wonderland.* Macmillan, 1865.

"English Quotes." *Hugging Face*, Abirate, *https://huggingface.co/datasets/Abirate/english_quotes?row=9*

[26]Public Record Office Victoria. "VPRS 13613 Technical Manuals." *Research Data Australia*, https://researchdata.edu.au/vprs-13613-technical-manuals/152179.

[26]Wilson, Jeffrey R. *Academic Writing: A Guide to Writing Across the Curriculum.* Harvard University, https://wilson.fas.harvard.edu/files/jeffreywilson/files/jeffrey_r._wilson_academic_writing.pdf

[27]American Rhetoric: "Top 100 Speeches of the 20th Century by Rank." *American Rhetoric*, https://www.americanrhetoric.com/top100speechesall.html

[28]Sim Anjali: "Emotion Analysis Based on Text." *Kaggle*, https://www.kaggle.com/datasets/simaanjali/emotion-analysis-based-on-text?select=emotion_sentimen_dataset.csv.

[29]DataCamp. "Beyond ChatGPT: 13 Best Wrappers." *DataCamp*, 10 Aug. 2023, www.datacamp.com/blog/beyond-chatgpt-13-best-wrappers

[30]Dastin, Jeffrey. "AI Isn't Ready to Make Unsupervised Decisions." *Harvard Business Review*, https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions

[30]"Character AI Statistics." *What's the Big Data*, 2024, https://whatsthebigdata.com/character-ai-statistics/

[30]Oshan AI. "13 Top AI Character Chat Apps + Websites [Free & Paid 2024]." *Oshan AI*, 12 Mar. 2024, https://oshan.ai/ai-character-chat-apps-websites/

[30]The Knowledge Academy. "How Does Character AI Work?" *The Knowledge Academy*, https://www.theknowledgeacademy.com/blog/character-ai/#:~:text=are%20impressively%20lifelike.-,How%20Does%20Character%20AI%20Work%3F,of%20the%20characters%20it%20represents.

[31]Illinois Institute of Technology. "Unlock Career Opportunities with AI: How to