JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2024, VOL 7, NO S9

Chatbot in the E-Service of Mental Health Using the Reprogramming of the GPT-2 Model

Hiba Malik Mohssen[1]*Dr. Hayder H. Safi[2]

[1] Informatics Institute for Postgraduate Studies / Iraqi commission for computers & informatics/ Baghdad, Iraq
[2] Informatics Institute for Postgraduate Studies / Iraqi commission for computers & informatics / Baghdad, Iraq
[1]* Corresponding author's Email: ms202220729@iips.edu.iq
[2]* Corresponding author's Email: hayder.h.safi@uomustansiriyah.edu.iq

ABSTRACT

Large language models LLMs, have revolutionized the field of natural language generation NLG model by exhibiting human-like text generation capabilities. This paper explores the development of an advanced the NLG model for mental health support using the GPT-2 language model. The paper aims to overcome the limitations of pre-trained models in domain-specific applications by implementing fine-tuning strategies and memory augmentation techniques. The methodology involves reprogramming the GPT-2 model using a dataset of mental health conversations with a focus on Alzheimer's disease. The model incorporates an (archive technique) array-based storage method for maintaining context across interactions, enabling more coherent and personalized responses. Evaluation metrics include BLEU scores, cosine similarity, and training loss to assess the quality and relevance of generated responses. Results demonstrate the model's ability to generate contextually appropriate and informative text on complex medical and social topics related to mental health. The paper highlights the potential of combining large language models with specialized memory approaches to enhance e-service of mental health interventions. Future work suggestions include expanding the training dataset, implementing the proposed system in other and different e- services fields. This paper contributes to the ongoing efforts to improve the accessibility and quality of mental health support through AI-driven conversational agents.

Keywords: Artificial Intelligence in Healthcare, Chatbot, GPT2, Mental Health, Natural Language Processing, NLG model.

1. Introduction

Natural Language Processing (NLP) has undoubtedly enabled machine learning systems to understand and communicate with users in natural languages by increasing the number of applications such as mental health support. More recently, the work of Large Language Models (LLMs), typically in the form of transformers such as Generative Pre-trained Transformer 2 (GPT-2), has shown that these can write text almost at the human level and be engaged in conversations [1]. These models have proved incredibly flexible despite being trained with much more extensive and varied data, and they can handle a range of language tasks without needing task-specific

training. For the past few years, significant interest has been acknowledged in using NLP and LLMs in the mental health domain reported [2]. This technology has contributed to portable all-time chatbots or virtual support for mental health. They can help with Triage coping skills and act as the first line of defence to get someone linked with mental health resources.

Based on LLMs ability to grasp the context, formulate empathetic replies and provide academic information, this application is well suited for these highly sensitive fields [3] Yet, while they have potential, there are challenges in using LLMs for mental health [4] The challenge is that pre-trained models, e.g., GPT-2, are not well-suited to a specific domain, such as mental health conditions. They require substantive fine-tuning or customization, even though they perform excellent language tasks [5]. This paper aims to overcome these challenges by implementing a more advanced the NLG model architecture for mental health support based on a modified GPT-2 model. This paper seeks to create an open the NLG model that independently provides conversational explanations relevant to mental health conversations. Therefore, we aim to refine the GPT-2 language model by developing a fine-tuning strategy specifically for mental health and feeding output knowledge generation of GPT2 and input prompts to retraining. We are also working on functionality to allow the chatbot to remember conversations and recall necessary details during prolonged engagements. The ultimate goal of this paper is to help develop a more efficient, contextually grounded and niche-focused the NLG model for mental health support. As part of the broader objective to use advances in NLP technologies for improving access and quality resources generated in real-time.

This research addresses two interconnected challenges: a scientific-technical problem in natural language processing and an application issue in mental health support within healthcare. The scientific-technical problem concerns the limitations of pre-trained language models like GPT-2 in maintaining context and personalization in domain-specific and determining the efficiency and effectiveness of GPT-2. We propose an innovative approach to improve the GPT-2 model by fine-tuning it on mental health data, specifically Alzheimer's disease conversations [6]. As for an application issue, it lies in the extent to which pre-trained language models such as GPT-2 can be applied while maintaining context and customization in order to generate a coherent text capable of providing electronic services in multiple fields. Our method implements an array-based storage for short- term context retention and a mechanism for periodic retraining by seed reprogramming technique, effectively creating a long-term memory for the NLG model. This approach combines transfer learning with continuous learning, enhancing the model's performance in mental health-related dialogues over time. The application problem addresses the need for accessible, personalized, and context-aware e-service of mental health interventions. We apply our enhanced model to create a chatbot system providing support related to Alzheimer's disease. The system engages in contextually coherent conversations, offering information and support to patients, caregivers, and family members. By continuously learning from interactions, the chatbot aims to improve the quality of mental health support over time [5],[6].

2. Related Work

2.1 Overview of the NLG Model in Mental Health

Mental health interventions are now utilizing the NLG model to extend the reach of mental health care within existing populations, leveraging NLP for efficient and effective tech-enabled support. Work such as G. Rosario and D. Noever in 2023 [7] an evaluation and analysis of the responses generated by large language models chatgpt3, which presented 60 questions and recorded the responses based on several metrics, including BLEU and its comparison with human texts. It obtained excellent results in machine generation but doesn't mimc text appears human. More recently, Researchers S. Priccilia and A. Girsang (2024) [8] presented a service that provides academic information resources in the field of education for Indonesian university students through the process of digitizing the answer to questions between students and student service staff. They implemented a generative dialogue system to answer their inquiries using the GPT2 model with three sizes and showed good performance quality but complexity in understanding the context.

A. M. Hasani et al. 2024 [9] presented a comparative study of the quality of content of clinical diagnosis and decision-making reports for radiology reports generated from the GPT4 model and compared with radiology specialists, where the reports generated by artificial intelligence for similarity quality were high, but they suffer from clear brevity and variation in sentence length. Also study by X. Zhang and Z. Luo 2024 [10] addressed the limitations of many traditional dialogue system s in maintaining context across multiple interactions, leading to repetitive responses. In addition, dialogue systems rely on text-based dialogues and lack the ability to refer to previous conversations, which reduces their efficiency and often exposes sensitive user information to the risk of disclosure. The researchers presented the dialogue system SOULSPEAK, a new conversational agent that has a dual memory system that distinguishes between short-term and long-term memory, works to retain key information from previous interactions and use it in subsequent interactions and contains a privacy module to hide sensitive user information. The responses were contextually similar to human responses but less readable. Additionally, Guntuku et al. [11] study has shown that it can capture user experience based on prior inputs of users and identify their mood with machine learning algorithms for mental health monitoring.

2.3 Recent Advancements in LLM Fine-tuning and Memory Augmentation

During the past few years, there has been considerable success in developing mental health the NLG model, and new ideas spring up every year. In 2020, L. Athota et al. [12] proposed AI-driven chatbots in healthcare, focusing on patient engagement and satisfaction. Through their solution, they built an NLP and machine-learning-enabled AI chatbot that provided customized support to users for improved healthcare processes. This work signposted the path for the future and showed that AI chatbots could be a great opportunity in mental health care. Two important works in the field were published in 2022; the first was conducted by Noble et al. [13], who and his team responded to the challenges resulting from COVID-19. Where his work developed and tested a mental health chatbot for medical staff and their families. The proposed Mental Health Intelligent Information Resource Assistant aimed to navigate the mental health system and provide personalized support during crises.

The second work was also published in the same year (2020) by Rathnayaka et al. [2]. He and his team implemented a conversational agent combining Cognitive behavioral Therapy (CBT) with personalized Behavioral Activation (BA), advancing this concept further. The chatbot was provided in terms of emotional support, personalized or customized assistance, and remote monitoring, which is very effective, showing that AI integrates with BA therapy within a chatbot setup. In 2023, Arriba-Pérez et al. [14] designed an intelligent conversational system for elder cognitive impairment monitoring, utilizing NLP techniques to entertain and conduct cognitive assessments. Field tests demonstrated that the technology was able to identify cognitive impairments. Researchers N. Calderon et al. 2024 [15] presented a general description of compressing NLG models by distilling knowledge for specific NLG models to address the bias problem with the challenge of maintaining performance efficiency by proposing a co-teaching method by distilling knowledge at the word level implemented in different tasks. They evaluated the computational performance and the performance results were average as it depends on the task, model and model setup which in turn depends on the batch size and does not depend on the dataset also It didn't take into account the computational costs of training time memory mechanisms add computational overhead. In 2024, Kulkarni et al. [16] created a mental health support chatbot using the LSTM and BERT for analyzing guided open-ended conversation data, concluding how promising AI tools could be in transforming the landscape of providing help with mental health while suggesting their application as an ad hoc tool to accompany conventional clinical practices. In 2023, researcher A. Ouyang [17] studied the challenges associated with the growth of large language models in size and complexity and their increasing resources, which leads to high computational power costs and difficulties in deployment. He proposed performance engineering efforts aimed at improving access time, deployment efficiency, and cost effectiveness. The results indicate that larger models perform better. The results indicate that while larger models achieve advanced performance in various natural language processing tasks, the associated computational costs pose significant challenges. The research emphasizes the need for improved performance engineering to mitigate these problems and enhance the feasibility of deploying large models. One of the newest works in 2024, conducted by Jain et al. [18], reviewed the use of AI in developing chatbots for mental health, focusing on Machine Learning (ML) and Deep Learning (DL) methods such as NLP and sentiment analysis that relieved symptoms like anxiety or depression.

In the same year (2024**),** Jo et al. [19] evaluated the effects of Long-Term Memory (LTM) information in affecting user engagement and self-disclosure concerning a chatbot powered by LLM, examining perceptions towards CareCall, concluding that provision of appropriate health domain-contextualized long-term memory can enhance quality interactions. Still, it also poses challenges associated with personal health data privacy concerns . Zhong et al. [20] MemoryBank proposed a novel memory mechanism aimed at LLMs to provide long-term correspondence in chatbots for maintained interaction scenarios. This demonstrates its ability to provide long-term companionship and mental health fulfilment but limiting real-time application and challenges in balancing memory retention and forgetting for dynamic tasks. D. Gu in 2024 [21] proposed developing a GPTHF model based on the GPT transformer to improve the performance efficiency of LLMs in terms of response speed and computational power resources, as traditional generation models require time and large floating point operations (FLOPs). The aim is to reduce the number of FLOPS while maintaining and improving performance compared to the standard. The results

indicate achieving good performance in faster generation time but but the computational costs of the algorithm are large, may limit the actual speedup achieved.

Another work in 2024 was done by Guo et al. [22] introduced Low-rank Prompt Tuning (LoPT), a non-traditional type of so-called prompt engineering for language generation models, such as GPT-2. Guo et al. aim to optimize low-rank prompts, critically cutting down the trainable parameters compared with full prompt tuning. This method has competitive results and is more efficient than current methods; it could change how we transfer large language models to various tasks but Prompt tuning methods are limited in domain-specific customization reduced effectiveness in contextually dynamic dialogues.

Y. Wu., Z. Wang, et al. in 2024 [23], introduced PIM-GPT, a process-in-memory architecture for efficient GPT inference. It uses DRAM and memory to carry out multiply-accumulate operations directly inside the memory chips, which can dramatically reduce data transfers. Optimal mapping disperses matrix operations to different DRAM channels and banks to parallelize the data. This allows for vastly improved performance and energy efficiency when running GPT-2 or other large language models compared to traditional CPU and GPU implementations but challenges of performance efficiency with limits on coherence or contextual relevance in the output.

Finally, Berrezueta-Guzman et al. [24] published their work in 2024, tested ChatGPT for supporting Attention Deficit Hyperactivity Disorder (ADHD) therapy, and demonstrated that the empathy-driven adaptation of this system could have a substantial impact in enhancing ADHD care while simultaneously recommending more significant improvements regarding privacy and cultural sensitivity to enable deployment into healthcare. The journey of AI support for mental health takes place from 2020 to 2024, and the reviewed works show how fast the field of AI-driven solutions has evolved by year with more understanding and technologies that can help change the approach towards access barriers.

3. Development Aspect Highlighted in a Scientific Manner

The NLG model's development centers on fine-tuning the GPT-2 model for mental health conversations, particularly those related to Alzheimer's disease. This process involves optimizing the model parameters θ to minimize cross-entropy loss L over a domain-specific dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ (1), where x_i is the input prompt and y_i is the corresponding response:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}(\theta; D) \quad (1)$$

The loss function for each pair (x_i, y_i) is defined as:

$$\mathcal{L}(\theta; x_i, y_i) = -\sum_{t=1}^{T} y_{i,t} \log P\left(y_{i,t} \mid x_i, \theta\right) \quad (2)$$

To maintain context in conversations, the system uses an array $A_t = \left\{\left(p_j, r_j\right)\right\}_{j=1}^{t-1}$ that stores recent prompts and responses. This array is updated at each interaction:

$$A_t = A_{t-1} \cup \{(p_t, r_t)\} \quad (3)$$

This approach ensures the model retains context, improving the coherence and relevance of its responses. Additionally, the NLG model's continuous learning mechanism allows it to adapt over time by periodically reprogramming to retraining on new user interactions, further enhancing its ability to provide contextually appropriate mental health support.

4. Algorithms Used in the System - Mathematical Modeling

4.1 Fine-Tuning of GPT-2 Model

The fine-tuning of the GPT-2 model is formulated as minimizing cross-entropy loss $\mathcal{L}$ over a specialized dataset $D = \{ (x_i, y_{ii=1}) \}_{i=1}^{N}$ where x_i represents the input prompt, and y_i is the corresponding response [32]:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}(\theta; D) \quad (4)$$

The loss (2) for each example is defined as [33]:

$$\mathcal{L}(\theta; x_i, y_i) = -\sum_{t=1}^{T} y_{i,t} \log P\left(y_{i,t} \mid x_i, \theta\right) \quad (5)$$

where:

- T is the length of the target response sequence.
- $P\left(y_{i,t} \mid x_i, \theta\right)$ is the predicted probability of word $y_{i,t}$ given the input x_i and model parameters.
- Parameters θ are iteratively updated using gradient descent until convergence [31].

4.2 Array-Based Context Storage

To maintain context after several rounds of dialogue, the NLG model uses an array $A_t = \left\{\left(p_j, r_j\right)\right\}_{j=1}^{t-1}$ that stores recent prompts and responses, where p_j is a prompt, and r_j is the corresponding response. The array is updated at each interaction [32]:

$$A_t = A_{t-1} \cup \{(p_t, r_t)\} \quad (6)$$

This ensures the model generates contextually relevant responses by referencing recent interactions.

4.3 Response Generation and Ranking

- **Response generation by Top-K and Top-P Sampling**: The model applies sampling techniques to choose the most suitable response after generating potential ones. It is essential to balance diversity and relevance in the response. Two primary sampling techniques are used: Top-K and Top-P (nucleus). As in the equation is among the top-k (7):

$$S_K = \{v_j : (y_t = v_j \mid y_{<t}, x; \theta) \quad (7)$$

Mathematically, for each time step t, the model sorts the vocabulary of possible next tokens V= $\{v_1, v_2, \ldots, v_n\}$ by their conditional probabilities $P(y_t = v_j \mid y_{<t}, x; \theta)$ and selects the top-K tokens. The model then samples from this restricted set, where K is chosen to balance coherence and diversity (8).

$$S_P = \left\{v_j : \sum_{i=1}^{j} P(y_t = v_i | y_{<t}, x; \theta) \leq P\right\} \quad (8)$$

Mathematically, Top-p sampling defines a set of possible next tokens S_P. Additionally, the BLEU score is calculated to assess how closely the response r_t^k matches a reference response y_t (9):

$$S_{\text{BLEU}}(r_t^k, y_t) = \text{BLEU}\ (r_t^k, y_t) \quad (9)$$

- **Ranking Prompt:** The model ranking prompt $P_t = \{p_t^1, p_t^2, \ldots\}$ for a given input prompt. Each prompt p_t^k is evaluated using cosine similarity (10):

$$S_{\cos}(p_t, q_t^k) = \frac{v(p_t) \cdot v\left(rq_t^k\right)}{\|v(p_t)\| \|v(q_t^k)\|} \quad (10)$$

where $v(p_t)$ and $v(q_t^k)$ are the vector representations of the prompt and the prompt in dataset, respectively.

The highest-ranked prompt is selected from the exact sine similarity after comparing it with the ideal prompt which is determined using mean Last Hidden State technique then determined the best vector of embedding that used to find the highest-ranking prompt as in equation (11).

$$P_{highest} = arg \max_{p_i \in P} S(V_i, V_{mean}) \quad (11)$$

- Final Response Selection: The response corresponding to the highest rank prompt is selected from among the candidate responses through the mathematical equation (12).

$$R_{select} = arg \max_{r_i \in R} S\ (P_{highest}, c_i) \quad (12)$$

Where $P_{highest}$ represents the prompt with the highest rank and c_i represents the responses that selected from candidate responses which they can be represented by the new hybrid prompt technique then stored in archive technique (array).

5. Methodology

In this section, the details of the methodology are discussed. Figure 1. This General framework represents a comprehensive and iterative methodology designed to fine-tune a GPT-2 model, interact with users, and continuously improve based on ongoing data and feedback. Python programming language and its libraries have been used to conduct the work in this paper.

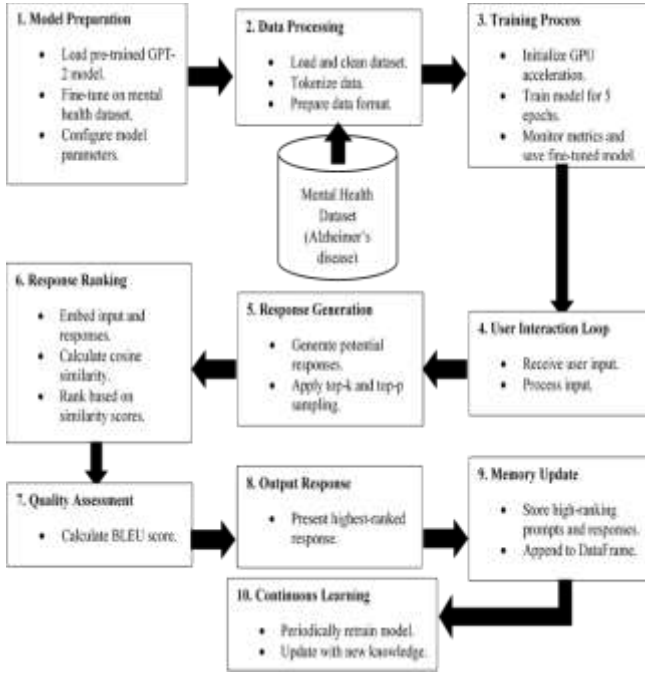

Figure 1: General Framework of Methodology Flowchart Steps.

5.1 Model Architecture

The basis of this paper is built on GPT-2, the pre-trained large language model by transformer architecture. This model is a strong baseline for understanding and generating various NLG tasks. Fine-tuning is the process that adapts GPT- 2 to a specific domain (mental health in our case). For instance, transfer learning is used here when we use that pre-trained GPT-2 model and fine-tune it as training using data tailored to mental health field requirements. Different hyperparameters, such as learning rate, batch size, and number of epochs, can be tuned for an optimized training process. The project addresses the challenge of maintaining information across interactions by implementing an array-based storage method. This includes specifying an array to accumulate candidate prompts and generated responses. This array serves as a storage context (Archive), enabling the model to reference recent interactions for more coherent conversations. Instead of a complex long-term memory system, this method balances maintaining relevant context and computational efficiency in mental health support scenarios. This includes specifying a list to accumulate the candidate prompts, generated responses, and concatenation in DataFrame. This data frame will train the model again by seed reprogramming technique, enabling it to effectively (remember) and apply past-built knowledge for future interactions. Instead, this creative method can be considered as long memory for a NLG model. Table (1) Hyperparameter of the GPT-2 language model. table (2) represents an example of configuring the tokenizer tool used in the proposed work.

Table 1: Hyperparameter of The GPT-2.

Hyperparameter	GPT-2
Epochs	5
Batch size	32
Learning rate	0.0001
Epsilon	1e-05
Optimizer	Adam

Table 2: Configuration of The Tokenizer Tool Using The GPT-2 Model

Tokenizer Tool of GPT-2	Configuration
name_or_path	gpt2
vocab_size	50257
model_max_length	1024
fast-False, padding side	right
truncation_side	right
special_tokens	['bos_token": "<(endoftext", "eos_token': 'clendofte xt/>', 'unk_token': '</endoftext]>'}
added tokens decoder	50256
Added Token	("<endoftext/>", rstrip-False, stripalse, single arabe, ormalized-true, special=True),

5.2 Data Processing and Training

First, loading and cleaning our initial dataset kicks us off into data processing. Then, it is taken in, tokenized with the GPT-2 tokenizer, and formatted correctly for

model use (block size of 1024, batch size of 32). The model training is accelerated using GPUs via Unified Device Architecture (GPU T4×2) for faster computation and CPU (RAM max 29 GiB). A training loop is created here for a specific number of epochs. For every five questions, we reprogram or retrain the model, in this case, to monitor and log specific metrics at regular intervals for performance evaluation during the training phase by the loss function or renderer. One of the highlights of this system is its continuous learning capabilities. This is done by adding functionality to add new questions and responses generated in the test set back into the training data. Over time, this updated dataset is fed back to the model for retraining it periodically, allowing the system to learn and include new knowledge. This way, with every interaction, the knowledge base of the NLG model will be constantly updated and growing over time. Table 3 illustrates the structure information of the metadata.

Table 3: DataFrame Metadata Structure Information

Range Index: 25177 entries, 0 to 25176 Data columns: total 2 columns Data types: total 2 objects Memory usage: 393.5+ KB		
Column	**Non-null count**	**Dtype**
Questions	25173	object
Answers	25154	object

5.3 Response Generation and Ranking

The response generator leverages the GPT-2 model fine-tuned around typing an answer on an input prompt. Techniques such as top-k and top-p sampling are used here to generate better quality, more diverse text. The system then uses an advanced sampling methods to decide the best response. This requires us to encode a given input prompt and available responses by the GPT-2 model for further calculating coherence scores between them through BLEU score. Then, used the exact sine similarity to sort the prompts in descending order and choose the highest ranked prompt and store it in an archive. The Bilingual Evaluation Understudy (BLEU) score evaluates the generated text and ensures its quality. A comprehensive evaluation is performed using both the sentence-level BLEU score and corpus-level BLEU scores. This provides deep insights into how well NLG model accurate responses in a linguistic and context-guaranteed manner. The last part of this methodology is seed reprogramming technique that maintains domain specificity and the interactive system. This enables users to input questions via direct text or choose from a list of randomly generated questions. The question pool is automatically updated. For example, it features a different number of questions each time to keep users in touch with the app and offers an array-based context storage (archive) from which they can choose one - creating diverse conversations throughout. In conclusion, mixing advanced NLP features with bespoke implementations produced a mental health NLG model that can provide context-specific responses and remember them for much longer, making the NLG models suitable to support or inform in the Mental Health domain.

6. Experimental Setup

This paper tests the NLG model specific to mental health using a reprogramed GPT-2 model. The setup consists of data preparation and definition of evaluation measures to obtain a solid analysis of the system's capabilities.

6.1 Dataset Description

The conversational dataset is published by the American Mental Health Association [25] and is associated with government agencies. The dataset consists of FAQ about Mental Health that are conversations between users and experts in the field of psychology about mental health and its relationship to Alzheimer's disease (Alzheimer's chat dataset, Alzhimer_chat_Leader, Mental_Health_FAQ.csv, NLP Mental Health Conversations). It was carefully collected and anonymized and included two columns, a question column and an answer column, to train the transformer, thus providing users or patients with appropriate guidance in answering their questions. The dataset is essential as it will be the base to fine-tune the pre-trained GPT-2 model, making responses auto-context relevant and personalized for mental health conversations.

6.2 Evaluation Metrics

To evaluate the performance of the fine-tuned GPT-2 model, several key metrics are used:

- Training metrics: are the metrics that give an idea about the progress and performance during training and provide an understanding of how well the model learns, scientific analysis, identifying potential problems, and improving the model, such as the global step, training run time, training samples per second, training steps per second, and floating point operations per second (FLOPS).
- BLEU Score: This metric is used to assess the quality of the generated text by comparing it with reference sentences. It evaluates how well the model-ranked prompts align with expected answers.
- Cosine Similarity: This metric measures the similarity between the input prompt and the prompts in dataset, ensuring the NLG model's responses are customized and contextually relevant.
- Training Loss: It is the meaning of the entropy-cross loss. Monitoring the training loss over epochs provides insights into how well the model learns from the data and whether it converges towards an optimal solution.

7. Results and Discussion

In this section, we present and analyze the results of our GPT-2 based NLG model for mental health and its relationship to Alzheimer's disease support, focusing on model performance, quality of generated responses, the effectiveness of our array storage approach and seed reprogramming.

7.1 Data Preprocessing and Training Results

During preprocessing and training, some analyses were done to understand how data was structured within our dataset and what had to be prepared for on the model side. Figure 2 shows the successful installation of essential libraries such as pandas, numpy, Torch and transformers, indicating a complete setup for data manipulation and model training. Table 3 depicts the examination showing that this dataset has 25177 entries and covers two columns, namely prompt (questions) and answers. The dataset's memory usage was measured at around 393.5 KB, suggesting a compact yet sufficient corpus for fine-tuning. These findings demonstrate the successful preparation of a specialized dataset for mental health and Alzheimer's disease conversations, laying the groundwork for subsequent model training and evaluation processes.

Questions	Answers
1430 Can exposure to electromagnetic fields increa...	The evidence regarding the association betwee...
8250 Are all head injuries equally associated with...	Head injuries that result in loss of conscious...
6540 How does family history impact Alzheimer's risk?	Having a family history of Alzheimer's increas...
8302 Can head injuries lead to other neurodegenerat...	Yes, head injuries may increase the risk of ot...
3043 How does sugar intake during pregnancy influen.....	Limited research suggests that maternal sugar

Figure 2: Calling Necessary Libraries and Setting up The Environment.

7.2 GPT-2 Model Fine-tuning and Metrics Results

Table 4 Training Metrics for the proposed fine-tuned GPT-2 model. The global step count of 4105 and epoch number of 5.0 indicated the extent of model training. A train runtime 2123.9143 was recorded, with both train samples and steps per second measured at 3.863. The total Floating Point Operations (FLOPs) were calculated at 42,878 TFLOPS, demonstrating the efficient performance and high speed in generating responses. Notably, the training loss was observed at 1.0506, this losses is good for the text generation task., suggesting room for further optimization in the model's performance on the mental health conversation dataset. These metrics provided valuable insights into fine-tuning and the model's adaptation to the specialized domain.

Table 4: Training Metrics for Fine-Tuned GPT-2 Model

Metrics Training	Result
Epoch	5
Global step	4105
Train runtime	2123.9143
Train samples per second	3.863
Train steps per second	1.933
Train loss	1.0506
Total FLOPS	4287802245120000.0

Table (5) presented BLEU score calculations for the model's output at both sentence and corpus levels. The BLEU values corresponding to the reference sentences A and G represent the BLEU value of reference sentence A = 0.176 which is bad and the BLEU value of sentence G = 0.549 which is very good for the text generation task. The BLEU scores achieved 0.534 for both single sentence and corpus scores perform uniformly across the dataset its very good.

Table 5: BLEU Score Calculations at Sentence and Corpus Levels

BLEU score	Result
BLEU score for A	0.1768
BLEU score for G	0.5495
BLEU score for a single sentence	0.5345
BLEU score for a corpus	0. 5345

7.3 Evaluation Prompt of Similarity Scores

Table 6 shows a data frame containing the cosine similarity scores for the prompt order. By calculating the cosine similarity between the input prompt and the prompt in the training dataset, they are arranged in descending order. The prompt with the highest rank is selected and the corresponding response is selected according to a specific mechanism to be stored in the archiving technique that helps to preserve and recall texts in previous rounds, especially complex rounds. Then, they are returned to the training dataset using the seed reprogramming technique to customize the model and keep the responses from hallucination.

Table 6: Ranking of Prompts Based on Similarity Scores

	Answers	Rinked
0	Could you explain what causes Alzheimer's dis.....	1.000000
1	What are the early brain changes associated wi.....	0.999714
2	What causes Alzheimer's disease, and how much	0.999695
3	Can you explain the difference between Alzheim...	0.999687
…..	……	…..
195	How can dementia manifest, and are there diffe...	0.998509
196	Is there a distinction between Alzheimer's dis...	0.998286
197	Are there non-medication approaches before con ...	0.998252
198	Is there an increase in the diagnosis of Alzhe..	0.997960
199	What is dementia, and how does it vary in seve...	0.997849

7.4 Evaluation of Model Performance

The text coherence quality values of the Blue scale that we obtained from tuning large language models GPT-2 to perform NLG tasks on a mental health dataset (Alzheimer's chat dataset, Alzhimer_chat_Leader, Mental_Health_FAQ.csv, NLP Mental Health Conversations) , we did not find any previous studies that applied their study to it, so we made the comparison on the results of generating pre-trained LLMs only and some of which used the fine-tuning technique, and our study achieved satisfactory results for the health dataset (BLEU = 0.5345) compared to the results achieved by G. Rosario and D. Noever, 2023, which used ChatGPT for generation and obtained results ranging between (0.50 -0.65), as our model is considered to have achieved good quality because this study was not subjected to the fine-tuning technique. While A. M. Hasani et al. who achieved a score of (BLEU = 0.5008) and relied on the GPT-4 model and Prompt engineering also considered our text quality slightly higher, which is a good result, in addition to S. Priccilia and A. Girsang, 2024 who achieved a good score (BLEU = 0.565) and used the fine-tuning method on data similar to our data consisting of a pair of questions and answers to adapt the GPT-2 model, but the application in the field of education and with several sizes of the GPT-2 model achieved a result very weak according to our results.

Table (7) shows that the proposed work shown as for the FLOPS measure of model performance and computational requirements for training neural networks and response generation speed, which is comparable to the trillion TFLOPS level, the

results of our proposed model on health data with 12 baselines (layers), batch size 32, and generated sequence length = 100 for the text generation task, the FLOPS value is (4287.8 trillion), which represents excellent performance, When comparing our proposed model with A. Ouyang 2023 who used an input sequence length of 201 and no batch size for the transformer-based OPT model and for the text generation task, the TFLOPS value was (49.0993), indicating average performance compared to our results. As for D. Gu's study, he used a baseline of 12 (layers), batch size of 32, and generated sequence length = 100 for the text generation task of the GPT2 model, and the TFLOPS value was (2.46), which represents a below-average performance rate. Nitai Caldero obtained TFLOPS metrics for a batch size of 32, a baseline of 12 layers, and a GPT2 model for three sizes GPT2 = 0.0136, GPT2-M = 0.0388, and GPT2-L = 0.084, which are very weak values compared to our work and obtained BLEU metric values for the three models GPT2 = 0.078, GPT2-M = 0.097, and GPT2-L = 0.119, which is a very weak constraint.

Table 7: Performance Comparison of Proposed Work with Existing the NLG Model

Metrics	BLEU Score	TFLOPS	Dataset	Response Generation Technique
Proposed work	**0.5345**	**4287.8**	American Mental Health Association	GPT-2 Top-k and top-p sampling
G. Rosario and D. Noever, 2023	0.50 - 0.65	Not reported	Quora	ChatGPT
A. M. Hasani et al., 2024	0.5008	Not reported	Set of Medical radiology reports	Prompt engineering + GPT-4
S. Priccilia & A. Girsang, 2024	0.565	Not reported	Pair of question-answer collected online	GPT-2
Nitay C. & s. m. 2023	0.119	0.084	XSUM40K	GPT2-L
	0.097	0.038		GPT2-M
	0.078	0.136		GPT-2
D. Gu, 2024	Not reported	2.46	Wikipedia, ArXiv, OpenWebText	GPTHF based-GPT
A. Ouyang, 2023	Not reported	49.09	Not specified	GPT-style decoder-only

Figure 3 shows an example of generating a text response for the NLG model based on a large language model (LLM) pre-trained and fine-tuned on mental health data. The GPT-2 model was retrained on the resulting knowledge and the input prompt and ranked in the highest order to maintain specificity in the mental health domain.

Q: How does Alzheimer's disease affect the family?
A: Alzheimer's disease is often said to be a "family" disease. At some point, individuals who are diagnosed cannot live alone, resulting in a minimum of two people being affected with each case of the disease. As the disease progresses, the person affected will increasingly require more care as Alzheimer's robs the person not only of function, but also of language and communication skills, thinking, reasoning and decision-making skills, and will eventually require total care. Caring for a person with Alzheimer's is a job that no one "applies" for, so family members are not trained, and are unexperienced or unprepared for this role making them susceptible to physical and emotional exhaustion, depression and feeling a sense of hopeless and helplessness in the face of the growing demands of care.
G: How does Alzheimer's disease affect the family? "Alzheimer's disease can significantly impact the family. A love done with Alzheimer's has an increased need for support, and caregivers may need to provide as much as possible of their own physical and emotional

Figure 3: Example of Generated Text

8. Conclusion and Future Work

The NLG model contributes to contextually coherent generative text, providing information and support to patients, caregivers, and family members. By continuously learning from interactions, the NLG model aims to improve the quality of mental health support over time. Our research demonstrates the potential of combining the language understanding capabilities of the GPT-2 model pre-tuned to mental health data with a novel long-term memory-like storage array approach (archive technique) to support mental health. The developed system showed promising results in maintaining context, generating empathetic responses, and personalizing interactions across extended conversations by selecting the highest-order prompt. These improvements address key limitations of traditional The NLG model in mental health applications, enhancing the quality and effectiveness of e-service for mental health interventions.

Future work should focus on several areas to further develop this technology:

1- Work on applying the proposed the NLG model to a local dataset and testing it instead of a global dataset.

2- Work on applying the proposed system to other areas of e-services in addition to the e-health service area.

3- Investigating the potential of combining the NLG model and human therapists in a hybrid support model.

These developments could lead to more effective, accessible, and personalized mental health support tools, potentially revolutionizing e-services or digital interventions for mental health.

References

X. Zhang and Z. Luo, "Advancing Conversational Psychotherapy: Integrating Privacy, Dual-Memory, and Domain Expertise with Large Language Models," arXiv preprint arXiv:2412.02987, 2024.

A. Jain, G. Srivastava, S. Singh, and V. Dubey, “Application of Artificial Intelligence (AI) Technologies in Employing Chatbots to Access Mental Health,” Computer Vision and AI-Integrated IoT Technologies in the Medical Ecosystem, 2024. doi: 10.1201/978100342960919.

A. P. B. Hafver, “Beyond words?,” DNV. Accessed: Aug. 01, 2024. [Online]. Available: https://www.dnv.com/research/future-of-digital-assurance/beyond-words-large-language-models/

A. S. D. Oliveira and R. D. S. Fernandes," Exploring the impact of intermediate languages on machine translation", Federal University of Do Rio De Janeiro Computer Institute Bachelor of Computer

Science, 2023.

A. M. Hasani et al., "Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports," European Radiology, vol. 34, no. 6, pp. 3566-3574, 2024.

Abdelhay, M., Mohammed, A., & Hefny, H. A. (2023). " Deep learning for Arabic healthcare: MedicalBot" . Social Network Analysis and Mining, 13(1). https://doi.org/10.1007/s13278-023-01077-w

S. Priccilia and A. Girsang, "Indonesian generative chatbot model for student services using GPT," International Journal of Informatics and Communication Technology (IJ-ICT), vol. 13, p. 50, 04/04 2024, doi: 10.11591/ijict.v13i1.pp50-56.

D. Gu, "Text Compression for Efficient Language Generation," 2024.

E. Jo, Y. Jeong, S. Park, D. A. Epstein, and Y.-H. Kim, "Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention," in Proceedings of the CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM, May 2024, pp. 1–21. doi: 10.1145/3613904.3642420.

F. de Arriba-Pérez, S. García-Méndez, F. J. González-Castaño, and E. Costa-Montenegro, "Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities," J Ambient Intell Humaniz Comput, vol. 14, no. 12, 2023, doi: 10.1007/s12652-022-03849-2.

G. Chalvatzaki, A. Younes, D. Nandha, A. T. Le, L. F. R. Ribeiro, and I. Gurevych, "Learning to reason over scene graphs: a case study of finetuning GPT-2 into a robot language model for grounded task planning," Front Robot AI, vol. 10, 2023, doi: 10.3389/frobt.2023.1221739.

https://www. thekimfoundation.org/faqs/ , https://www. mhanational.org/frequently-asked-questions , https://www. wellnessinmind.org/frequently-asked-questions/ and https://www. heretohelp.bc.ca/questions-and-answers.

https://www2.it.uu.se/edu/course/homepage/systemid/vt12/ch2.pdf .

J. M. **Noble** et al., "Developing, Implementing, and Evaluating an Artificial Intelligence-Guided Mental Health Resource Navigation Chatbot for Health Care Workers and Their Families During and Following the COVID-19 Pandemic: Protocol for a Cross-sectional Study," JMIR Res Protoc, vol. 11, no. 7, 2022, doi: 10.2196/33717.

G. Rosario and D. Noever, "Grading conversational responses of chatbots," arXiv preprint arXiv:2303.12038, 2023.

Lekha **Athot**a; Vinod Kumar Shukla; Nitin Pandey; Ajay Rana, "Chatbot for Healthcare System Using Artificial Intelligence," International Research Journal of Modernization in Engineering Technology and Science, 2023, doi: 10.56726/irjmets34164.

N. C. Chung, G. Dyer, and L. Brocki, "Challenges of Large Language Models for Mental Health Counseling," Nov. 2023.

P. **Rathnayak**a, N. Mills, D. Burnett, D. De Silva, D. Alahakoon, and R. Gray, "A Mental Health Chatbot with Cognitive Skills for Personalised Behavioural Activation and Remote Health Monitoring," Sensors, vol. 22, no. 10, 2022, doi: 10.3390/s22103653.

A. Ouyang, "Understanding the Performance of Transformer Inference," Massachusetts Institute of Technology, 2023.

S. Berrezueta-Guzman, M. Kandil, M. L. Martín-Ruiz, I. Pau de la Cruz, and S. Krusche, "Future of ADHD Care: Evaluating the Efficacy of ChatGPT in Therapy Enhancement," Healthcare (Switzerland), vol. 12, no. 6, 2024, doi: 10.3390/healthcare12060683.

S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," 2017. doi: 10.1016/j.cobeha.2017.07.005.

S. D. K. C. **Shouchang Guo**, "LoPT: Low-Rank Prompt Tuning for Parameter Efficient Language Models," arXiv:2406.19486 [cs.CL], 2024.

S. **Kulkar**ni, E. Parkar, R. Lonkar, P. Pareek, S. Patil, and S. Kusal, "Conversational AI for Mental Health Support,"in 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), IEEE, Apr. 2024, pp. 1–7. doi:

10.1109/MITADTSoCiCon60330.2024.10575117.

N. Calderon, S. Mukherjee, R. Reichart, and A. Kantor, "A systematic study of knowledge distillation for natural language generation with pseudo-target training," arXiv preprint arXiv:2305.02031, 2023.

W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "MemoryBank: Enhancing Large Language Models with Long-Term Memory," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, pp. 19724–19731, Mar. 2024, doi: 10.1609/aaai.v38i17.29946.

Y. Wu, Z. Wang and W. D. Lu "PIM GPT a hybrid process in memory accelerator for autoregressive transformers," npj | unconventional computing , 2024. https://doi.org/10.1038/s44335-024-00004-2.

Z. A. Ahmed, and M. Raafat, " An Extensive Analysis and Fine-Tuning of Gmapping's Initialization Parameters", International Journal of Intelligent Engineering and Systems, Vol.16, No.3, 2023, doi: 10.22266/ijies2023.0630.10.