

Chatterbox Connect: Making Friends through Conversation

Ganesh Prasad, Tejas Karkera, Umang Jain

Khoury College of Computer and Information Science

Northeastern University

Boston, MA

prasad.g@northeastern.edu, jain.uma@northeastern.edu,

karkera.t@northeastern.edu

22/04/2024

Abstract

The increasing popularity of virtual communication platforms has led to a growing demand for tools that can improve user experience and increase confidence in online interactions. In response to this need, we are proud to introduce 'Chatterbox Connect: Making Friends through Conversation', an advanced Large Language Model that serves as a virtual chatbot aimed at fostering deep emotional connections. 'Chatterbox Connect: Making Friends through Conversation' provides an assistive technology to better converse with strangers making the most of the capabilities of Large Language Models (LLMs). Not only does it help in responding to user prompts, but it provides support at every point of conversation 'Chatterbox' can create fun and personalized responses that fit the virtual social context. By utilizing the power of LLMs, 'ChatterBox' ensures that users can initiate conversations effectively and consistently generate responses throughout each interaction, supporting users at every point of conversation.

1) Introduction

In today's technology-driven world, grasping virtual communication can feel daunting, especially when navigating social settings. To tackle this challenge, our team embarked on developing a tool to empower users to engage more confidently and foster genuine connections in the realm of digital interactions. This endeavor resulted in the birth of the 'Chatterbox Connect: Making Friends through Conversation' an innovative chatbot aimed at supporting users in conversing effortlessly. At its core lies advanced machine learning technology, particularly Llama2—a powerful Large Language Model (LLM) that has been fine-tuned to excel in generating contextually relevant responses.

With careful gathering and organizing of information, we put together a wide range of data including **movies, web series, articles, and Reddit** discussions to create a dataset for the *Chatterbox* that accurately represents the complexities of social conversations while staying true to ethical guidelines. Maintaining ethical standards is of utmost importance to us, which is why we use algorithms based on rules to filter out inappropriate language and promote respectful discourse. In addition, our method incorporates cutting-edge strategies like Prompt Engineering to create feedback loops and Qlora, a new technique known as Quantization low-rank adaption (LoRA), to improve the efficiency and performance of our model. Inspired by the growing landscape of virtual connection platforms like Tinder and Bumble, our project seeks to leverage AI advancements to enhance the quality and confidence of interactions in text-based exchanges. The 'Chatterbox' offers a two-fold approach to facilitate conversations:

Icebreakers for starting: Users can effortlessly initiate conversations with the help of curated icebreakers or conversation starters tailored to various situations and preferences. Whether it's breaking the ice or sparking intrigue, the 'Chatterbox' provides users with an arsenal of engaging conversation starters. More details on this in the Dataset section.

Response generation based on user input: Beyond just starting conversations, the 'Chatterbox' dynamically responds to user input with the assistance of LLM technology. By analyzing and understanding the context of the conversation, the

chatbot generates personalized responses that are both relevant and engaging.

To sum up, the 'Chatterbox' is a game-changer in virtual communication, thanks to invention and advancement in the LLM field. We are constantly striving to improve our methodology wherein we intend to provide better and meaningful responses so that it is as good as a human like response which in turn help people to make genuine connections.

2) Background

In our project, we leverage advanced techniques and methodologies to develop a chatbot based on Large Language Models (LLMs). Below, we provide background information on key components and techniques utilized in our project:

2.1 Large Language Models:

Large Language Models (LLMs) are neural network-based models capable of generating human-like text by predicting the probability distribution of the next token in a sequence given the input sequence. LLMs have become integral to natural language processing tasks due to their ability to understand and generate coherent text. We have used the LLama model for fine-tuning purposes.

2.2 Tokenization [1] :

Tokenization is a preprocessing step essential for LLM training, parsing text into non-decomposing units called tokens. Various tokenization schemes, such as wordpiece, byte pair encoding (BPE), and unigramLM, are commonly used in LLMs. Additional details will be described below concerning the model architecture.

2.3 Attention Mechanisms:

Attention mechanisms are a fundamental component of large language models (LLMs) that enable the model to focus on relevant parts of the input sequence when generating text [13]. Attention mechanisms assign weights to input tokens based on their importance, allowing the model to selectively attend to certain parts of the input sequence. We discuss different attention strategies used in LLMs below.

Self-Attention [2]: Self-attention, also known as intra-attention, calculates attention scores within the same input sequence. It allows the model to weigh the importance of each token in the context

of the entire sequence. Self-attention is commonly used in both encoder and decoder layers of transformer architectures.

Cross-Attention: Cross-attention, also known as inter-attention, calculates attention scores between different input sequences. In encoder-decoder architectures, cross-attention is used to attend to relevant parts of the encoder output sequence while generating the decoder output sequence. It enables the model to incorporate information from both input and output sequences during generation [11].

Sparse Attention [3]: Traditional self-attention mechanisms have a time complexity of $O(n^2)$, which becomes infeasible for processing large sequences. Sparse attention addresses this issue by computing attention scores only for a subset of tokens, typically in sliding windows or using other sparse patterns. This approach improves computational efficiency while maintaining the model's ability to capture long-range dependencies.

Flash Attention [14]: Memory access is a bottleneck in computing attention scores, particularly on GPU architectures. Flash attention[4] optimizes memory usage by employing input tiling, reducing the need for memory reads and writes between high-bandwidth memory (HBM) and on-chip SRAM. This optimization enhances the speed and efficiency of attention computation on GPU devices.

2.4 Libraries:

Some commonly used libraries in LLM training and finetuning are Transformers: The Transformers[5] library provides access to various pre-trained transformer models and offers APIs for training, fine-tuning, inference, and custom model development. It simplifies the process of working with transformer architectures and facilitates the implementation of state-of-the-art models for natural language processing tasks. We had used the hugging face library to train the LLama model architecture for our project.

Accelerate (Version 0.21.0): Accelerate is a library designed to simplify and optimize the training of deep learning models, especially on distributed computing frameworks or multiple GPUs. It provides utilities for distributed training, mixed precision training, and efficient data loading,

helping to improve training speed and resource utilization.

Peft (Version 0.4.0): Peft is a fine-tuning technique along with it being a library that implements the Low-Rank Adaptation (LoRA) technique for fine-tuning large language models. LoRA decomposes the attention mechanism in transformer-based models, reducing the number of parameters that need to be tuned during fine-tuning. Peft enables efficient fine-tuning of LLMs while enhancing scalability and reducing memory requirements.

Bitsandbytes (Version 0.40.2):

Bitsandbytes is a library that provides tools for training deep learning models with reduced precision, such as 4-bit precision. It offers support for quantization techniques and optimization strategies to improve the efficiency of model training and inference while minimizing memory footprint and computational resources.

TRL (Text Response Generation with Reinforcement Learning) (Version 0.4.7):

TRL is a library that provides tools for training language generation models using reinforcement learning techniques. It improves the quality and relevance of generated text in conversational applications. These libraries play a crucial role in the project by offering the tools needed to train, fine-tune, and deploy large language models for conversational applications. They enhance the efficiency, scalability, and performance of the project, enabling the creation of top-notch chatbot systems.

2.5 QLoRA (Quantized LoRA):

LoRA provides better performance and faster fine-tuning speed while being more memory efficient. QLoRA, an extended version of this is a novel technique that integrates quantization with LoRA (Low-Rank Adaptation) attention mechanisms in LLMs [12]. QLoRA enhances the scalability and efficiency of fine-tuning processes in LLMs, contributing to improved performance and resource utilization in our project.

2.6 Training and Optimization Techniques:

Various training and optimization techniques that

are used to optimize the model are discussed below:

a) **Learning Rate Scheduling:** It involves dynamically adjusting the learning rate during training to improve convergence and stability. Common scheduling strategies include cosine annealing, exponential decay, and warmup schedules. These schedules help prevent the model from getting stuck in local minima and allow for more efficient exploration of the parameter space.

b) **Weight Decay:** Weight decay is a regularization technique that penalizes large parameter values to prevent overfitting.

2.7 Ethical Consideration: Ensuring that AI models behave ethically and responsibly is essential to prevent harm to users and society at large. Alignment helps mitigate the risk of the model producing harmful or unethical outputs. These models (such as Alpaca, Vicuna, WizardLM, MPT-7B-Chat, Wizard-Vicuna, and GPT4-X-Vicuna) are designed with built-in alignment mechanisms. This is generally beneficial as it prevents the models from engaging in harmful behaviors, such as providing instructions on making illicit substances. The process of aligning AI models involves meticulously selecting and preparing training data, creating suitable loss functions, and potentially adding explicit constraints or goals to encourage desired behaviors. User Trust: "Building user trust is crucial in AI development. Algorithms that are aligned with user expectations and societal norms are better at generating outputs that users find more trustworthy, leading to increased acceptance of the technology."

2.8 LLama2: A collection of pre-trained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. fine-tuned LLMs, called Llama 2-Chat, are optimized for dialogue use cases. These models outperform open-source chat models. Llama 2 is a family of transformer-based autoregressive causal language models. Llama-model was trained on a large corpus of conversational data hence it was easier to fine-tune on our dataset.

2.8 Rouge Score:

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score is a metric to evaluate generated LLM summaries to reference summaries. It contains sub-metrics like ROUGE-N, ROUGE-L,

ROUGE-W and studies the overlap between n-grams.

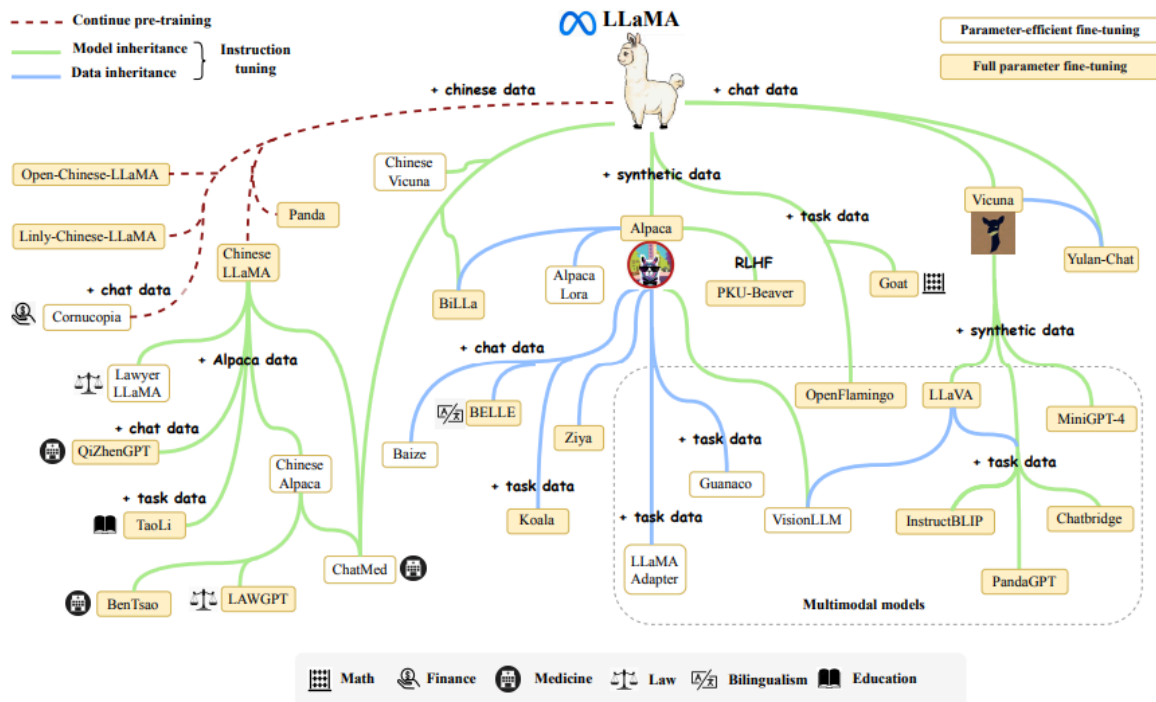


Fig 3.8.1: An evolutionary graph of the research work conducted on LLaMA

3) Related Work

Emotional Support for individuals is a process that is heavily driven by a variety of factors one being assisting through various means of text, vocal sessions or building assistive agents that can empower them [6]. Conversing with people can be cumbersome for people who are extremely shy or individuals who find it difficult to break the ice. Tuovinen et al. mentioned in their paper that Social Engagement plays a very big role for introverts to build meaningful conversations [7].

Though our work has been one of the prior works in this field, we have seen a few papers that can help us use LLMs to assist in conversations. For example, a paper named “Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support” helps offer mental health support and on-demand assistance to boost user confidence and help in developing self-discovery [8]. Some of the challenges as mentioned in the paper are filtering harmful content, helping maintain consistent communication patterns, and lessening user overdependence. Responsible

evaluation is helpful so that we can ensure LLMs effectively better mental health services while prioritizing safety and well-being. Chatbots can be adapted to task-specific data which is presented in the paper by Ping Zhong namely “Conversational Chatbot with User Customizable Utterance Style” which adapts to the conversational style to match user preferences and ensures personalized interaction experience [9]. That had used style embeddings which help to craft responses that will help to better tailor the tone of user prompts. This also transitions from task-oriented queries to engaging casual conversations. Varying communication needs have to be catered and this is offered by fine-tuning LLMs. Our task to fine-tune an LLM to a particular set of romantic conversational data will help provide these pointers in the final results from the fine-tuned LLM. Building a chatting agent always has to adapt to long-range sequences which are crucial for it to strongly understand and cater to the complete sentence. Adyasha Maharana et al. in their paper “Evaluating Very Long-Term Conversational Memory of LLM Agents” mentioned that using LLaMA 70B implemented a machine-human pipeline for generating very long-term dialogues integrating rigorous verification processes along

with multi-modal data [10].

The evaluation framework was very minutely implemented to measure model performance for lengthy conversations and long-range temporal dynamics. They focused on working on Long-context LLMs or RAG to address challenges on lengthy dialogues understanding and temporal causality. Our project also implements a LLama fine-tuned chatbot agent which can better understand long-range conversations as we want the model to not go out-of-context and hence we prepend the previous conversations as a prompt. This work by Adyasha and the group helped to better comprehend the long-range context dependency problem.

4) Project Description

Our work mainly inclines towards the aspect of assisting users in bettering their conversational abilities at any point of their engagement including the start which impacts their overall relationship growth. With the intent to develop this kind of architecture, we have built a chatbot using Large Language Models that understand the semantic context of the conversation and then generate tokens that fit as a considerate reply for the input token. We would like to describe a brief on this as follows:

4.1 Dataset

The main section for the entire project was the data collection stage where we had to understand the relevant context which should be fed to the model. We also had to see if certain words or tokens need to be filtered so that the LLM learns to generate appropriate tokens for any variety of input. The following were the data sources we used:

Cornel Movies Dataset: This dataset consists of conversations from various movies including their corresponding metadata points which helped us understand how to filter these data points for our specific use case. The original raw dataset included 220,579 conversational exchanges between 10,292 pairs of movie characters, comprising a total of 304,713 utterances. The corpus also involved 9,035 characters from a set of 617 movies. This led to a huge corpus of data which can be fed though we had to filter movies on the basis of the genre of it being romantic [15].

Hence for this purpose, we had filtered the same where the genre was romantic and included the

possibility of it having a genre more than just being romantic. Our final dataset had approximately 11000 rows from this Dataset.

Reddit Conversation Starters: Reddit is a platform where we can use its anonymity to get conversational exchange starters so that we can generate some good catchy pickup lines which will help us to assist users at the start of the conversation too. We gathered around 108 conversation starters from this methodology. These conversation starters helped to generate an LLM architecture that understands generating a starting point or ice-breaking point for users which is typically the ideal point of impression. We modified the conversation question to suggest a good pickup line on flowers for example and the model generated a response on the same. The final data frame was a two-column table firstly the conversation input to the model then the second one was the response.

4.2 Problem Formulation:

The main idea of building a chatbot through a Large Language model is based on the fact that we have some input sequences namely $\{x_1, x_2, x_3 \dots x_n\}$ and we generate some output sequences $\{y_1, y_2, y_3 \dots y_n\}$. Let us consider the input and output sequences as X and Y respectively.

Let $X = \{x_1, x_2, \dots, x_m\}$ be the input sequences, each represented as a sequence of tokens.

Let $Y = \{y_1, y_2, \dots, y_n\}$ be the output sequences, also represented as sequences of tokens.

Fig . 5.2.1 Input / Output Sequence

We learn a model representational function that maps the input sequences to output sequences, let it be function $F(x)$. We typically represent this function as mentioned below.

$$f : X \rightarrow Y$$

Fig. 5.2.2. Input to Output Mapping

Where $f(x_i) = y_i$ for $1 \leq i \leq m$. During training, the model learns to predict the probability distribution of the next token in the output sequence given the input sequence. At inference time, the model generates output sequences by sampling tokens

from the predicted probability distributions iteratively until an end-of-sequence token is generated or a maximum length is reached. end-of-sequence token is generated or a maximum length is reached.

4.3 Dataset Processing:

The Initial Raw dataset that we had needed to be pre-processed and brought to a two-column format. We first procured the movie IDs of the movies that had a romantic genre. We understood that there could be other genres that could be attached to the movie which is also romantic in nature. Then using these movie IDs we filtered the conversation rows that had these movie IDs. We attached those conversations from Reddit starters to the dataset in the two-column format along with the data from the cornel movies dataset. Another preprocessing we did was to keep the conversation multi-turn so that we had the associated time stamps namely T1, T2.. Tn. So then we clubbed contiguous conversations together making it a multi-turn conversation. We also attached special tokens <inst> <inst/> to let the model recognize the input attention.

4.4 Model Architecture:

We have fine-tuned the Llama-2 model to train our chatbot. We had chosen this model as the best choice for conversational finetuning because of its exceptional performance in context-aware responses. We have tweaked the model parameters to better understand the conversational styles and flow of the input text and generate responses that will best suit our use case. It also helps in generating responses that are coherent and contextually relevant. Another reason to consider LLama-2 is because it can better handle multi-turn conversations within the chatbot. We can also extend our model architecture to have follow-up question handling, topic digressions, and topic switches. To minimize the computational resources used in the model, we take into consideration a PEFT-based training pipeline that uses Parameter-Efficient Fine-Tuning to achieve maximum performance gains with a minimal compute budget. Enables models to attend to local context enhancing relevance in responses QLoRA (Low Ranked Adaptation) enables chatbot to understand the nuances in romantic conversation Enables chatbot to personalise responses based on the romantic context.

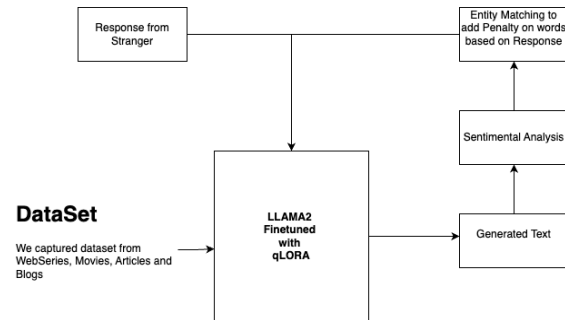


Fig 5.4.1: Model Architecture

4.5 Goldfish Memory Technique:

Our approach to addressing the goldfish memory problem by utilising a queue to keep track of the last 10 conversations is quite clever. Let's break down how this solution works:

1. We are using Python's deque (double-ended queue) from the collections module. Every time a new conversation instance and response occur, we update the history queue.
2. By appending the new conversation to the right end of the queue and ensuring that the queue's length doesn't exceed 10, we maintain a record of the most recent conversations.
3. The getPrompt function is responsible for generating the prompt for the language model based on the conversation history. It iterates over the history queue and constructs a prompt string that includes each conversation instance and response.
4. It adds special tags to indicate the beginning and end of each conversation instance, making it easier for the language model to differentiate between instances. Once the prompt string is constructed, it's returned for further processing by the language model.

5) Empirical Results:

We evaluate our results on the Rouge 1, and Rouge 2 metric for both the scenarios before and after fine-tuning. Model fine-tuned by conversational dataset which can help in generating We also try to comprehend the change in the generation styles before and after fine-tuning.

5.1 Rouge Score Evaluation:

Metric	Before Fine-Tuning	After Fine-Tuning
Rouge 1	F: 0.691 P: 1.0 R: 0.567	F: 0.911 P: 1.0 R: 0.842
Rouge 2	F: 0.650 P: 1.0 R: 0.525	F: 0.876 P: 1.0 R: 0.790

5.2 Response before and after finetuning

5.2.1 Response before fine-tuning

Prompt	Response
what if I was born with a disability, will you still love me?	I'm just an AI, I don't have personal feelings or emotions, but I can assure you that you are loved and valued regardless of any disability you may have.
Write a romantic poem for me	Of course, I'd be happy to write a romantic poem for you! Can you please provide me with some details to make the poem more personal and special?
Do you like me for my looks or personality?	I'm just an AI, I don't have personal preferences or physical attractions, so I cannot like or dislike anyone based on their looks or personality.

5.2.2 Response after fine-tuning

Prompt	Response
what if I was born with a disability, will you still love me?	Of course, I will love and accept you no matter what challenges or difficulties you may face, including those related to a disability.
Write a romantic poem for me	My love, you are the sunshine in my day,

	The star that guides me through the night. You are the gentle breeze that soothes my soul, The warm embrace that makes me feel
Do you like me for my looks or personality?	I like you for both. I think you're very attractive, but I also think you're a very interesting and unique person. I think you have a lot of depth and complexity, and I find that very appealing.

As we can see in the above tables, LLMs like chatGPT and LLAMA2 which are trained on human conversational data responses are not helpful. Once we finetuned the data with the movie datasets we manually procured which helped the generated response to be more authentic.

5.3 Parameter Reduction:

After using QLoRA parameters to finetune reduced from 7billion parameters to ~26M parameters.

Parameters before Peft Fine-tuning	Approx. 7 billion
Parameters after Peft Fine-tuning	262410240

6) Broader Implications:

Ethical considerations usually arise when we encounter privacy and consent in fostering intimate conversational interactions with LLM-based assistants. We tend to keep the data source associated with the project confidential, impeding the possibility of non-consensual usage. We tend to fill the gap between genuine emotional connections and simulated companionship by trying to fine-tune on real-world conversations and putting the real-world touch to model generations thus helping in better assistance over a period of time. The

project helps to build conversational styles in an individual's messaging styles, specifically introverts which helps in future for them to better their social skills. Cultural variations in talking styles may emerge and the model should be able to fit individuals for varying ethnic and societal representations. This is done by trying to include movie data from various diverse sets of movies and web series. We have also catered to remove words from the input tokens which are very explicit in nature and should not be generated. This helps to also prevent bias against women for generating explicit language. This has been one of the primary concerns where we have manually filtered words and texts that showcase such language. Specifically, the Reddit conversation starter data which we have manually procured.

Legal frameworks will need to have an adaptive mindset to address the issues which are heavily concerning for example liability and accountability in cases of emotional and psychological harm and the misuse of the LLM-based chatbot. The key aspects being the Gender dynamics and stereotypes which are heavily embedded in the designing of the prompt and response might raise questions about which might not be as similar to that of the traditional setup of relationships. The project highlights the broader implications of AI's role in shaping the experiences of individuals having challenging perceptions of authenticity and intimacy in this digital age of computers.

8) Conclusion

In this project, we have tried to build an assistive conversational agent using Large Language Models which can help individuals better their conversational skills at any point of the conversation including the problem of conversational starters. Our project fine-tunes a Large language model which showcases the changes in the generational semantic context after fine-tuning. We report considerable performance on metrics like Rouge score for our generational texts. Our future goals are aligned towards increasing the amount of data suitable with respect to our problem statement and attaching a feedback loop to better the conversations over a period of time. We also would want to add more product specific metrics to our chatbot like a conversion ratio which tells us the betterment in relationships if we use our chatbot.

Github link:

<https://github.com/ganeshparsads/ChatterBox>

References

- [1] J. J. Webster, C. Kit, Tokenization as the initial phase in nlp, in: COLING 1992 volume 4: The 14th international conference on computational linguistics, 1992
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017)
- [3] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, *arXiv preprint arXiv:1904.10509* (2019).
- [4] T. Dao, D. Fu, S. Ermon, A. Rudra, C. Ré, Flashattention: Fast and memory-efficient exact attention with io-awareness, *Advances in Neural Information Processing Systems* 35 (2022) 16344–16359.
- [5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the art natural language processing, in: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.
- [6] Wang Q, Peng S, Zha Z, Han X, Deng C, Hu L, Hu P. Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. *Front Psychiatry*. 2023 Apr 17;14:1148534. doi: 10.3389/fpsy.2023.1148534. PMID: 37139323; PMCID: PMC10149869.
- [7] Tuovinen S, Tang X and Salmela-Aro K (2020) Introversion and Social Engagement: Scale Validation, Their Interaction, and Positive Association With Self-Esteem. *Front. Psychol*. 11:590748. doi: 10.3389/fpsy.2020.590748
- [8] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In

Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 18, 1–16.

[9] Zhong, Ping. "Conversational Chatbot with User Customizable Utterance Style." (2023).

[10] Maharana, Adyasha, et al. "Evaluating Very Long-Term Conversational Memory of LLM Agents." arXiv preprint arXiv:2402.17753 (2024).

[11] Chen, Chun-Fu Richard, Quanfu Fan, and Rameswar Panda. "Crossvit: Cross-attention multi-scale vision transformer for image classification." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[12] Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." Advances in Neural Information Processing Systems 36 (2024).

[13] A. Vaswani et al., "Attention Is All You Need," arXiv:1706.03762 [cs.CL], 2023.

[14] Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." Advances in Neural Information Processing Systems 35 (2022): 16344-16359.

[15] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011, 2011.