Investigating Multimodal Empathetic Conversational AI Utilizing Ensemble Learning and Humor

[1]Darshan Bhavesh Mehta

[1]Independent AI Researcher, Mumbai, Maharashtra, India

[1]darshanbmehta@hotmail.com

***Abstract*—Engaging in conversations to enhance mindfulness and combat loneliness can play a significant role in reducing and preventing mental health issues. Anxiety, depression, and stress are increasingly prevalent in today's fast-paced world, impacting people of all ages and backgrounds. To address these challenges, a virtual companion is being designed to tackle problems with an inclusive approach, offering a holistic understanding by considering multiple perspectives. This virtual companion leverages advanced natural language processing (NLP) models, such as DialoGPT and T5 Transformers, along with techniques like prompt engineering, Retrieval Augmented Generation (RAG), and fine-tuning on an augmented dataset.**

The system also incorporates text-to-speech engines, including Speech T5 and Meta's Massively Multilingual Speech (MMS), which can generate naturalistic speech in multiple languages, capturing the nuances of human communication. The goal is to provide an empathetic and friendly solution that meets the diverse mental health needs of users. By combining AI with empathy, this research aims to develop an innovative tool for empathic care, helping people build resilience in the face of life's challenges. The tool generates real-world solutions and responses, including humor tailored to the user's receptivity, to foster engaging conversations and journaling activities. This approach not only helps users manage their mental health but also offers a safe space for emotional expression and support.

***Index Terms*— Therapeutic Conversational Agent, DialoGPT, Meta's Massively Multilingual Speech, Whisper large v2, Speech T5**

I. Introduction

In today's fast-paced world, mental health disorders such as stress, anxiety, depression, procrastination, and overthinking are becoming increasingly prevalent. These issues can have severe health consequences, including infertility, heart attacks, stroke, reduced stamina, hypertension, increased susceptibility to cancer, obesity due to stress eating, and irritable bowel disease (IBD). Such problems can significantly impact a person's motivation, productivity, and overall quality of life. Anxiety disorders alone affect over 284 million people worldwide, highlighting the need for effective counseling services. However, implementing counseling is challenging, especially in developing countries where up to 75% of patients do not receive any assistance or treatment. This lack of emotional support systems is alarming and underscores the need for innovative solutions. Traditional systems often respond in a robotic manner, lacking empathy and personalization, which can lead to generalized answers that fail to meet users' specific needs. To address these challenges, researchers are developing conversational AI assistants for mental health support. These systems integrate advanced text generation and summarization models to provide empathetic outputs. They use sophisticated text-to-speech (TTS) models to produce natural responses and allow users to customize the synthesized voice. The inclusion of humor, when appropriate, helps lighten conversations and enhance user interaction[4]. The proposed system employs a hybrid approach using DialoGPT and T5 Transformers for text generation. It combines Whisper large v2 for Automatic Speech Recognition (ASR) with state-of-the-art TTS engines like Speech T5 and Meta's Massively Multilingual Speech to generate synthetic speech that retains human interaction nuances. This approach aims to offer personalized support, addressing the limitations of existing chatbots, which often lack emotional intelligence, cultural understanding, and integration with professional care systems. Ultimately, the goal is to create a system that not only provides empathetic and helpful responses but also fosters long-term engagement and addresses ethical concerns, ensuring that mental health chatbots can effectively support those in need.

II. Literature Survey

Hsu et al. (2023) propose a transformative approach to empathetic conversational AI systems by integrating a transformer-based language model with attribute models for both affective and cognitive empathy[12]. Their research combines the Empathetic Dialogues dataset with a modular architecture, allowing for dynamic language generation adjustments with minimal retraining. This enhances the model's empathetic communication capabilities, enabling it to adapt to diverse user needs and emotional states. The integration of DialoGPT and T5, along with advanced NLP techniques and data augmentation methods, forms a robust system for empathetic conversational AI [10]. This system is designed to provide personalized support, addressing the emotional and psychological needs of users in a more human-like manner.

Raamkumar et al. (2023), under the IEEE umbrella, explore various empathic conversational AI forms by utilizing the Empathetic Dialogues dataset. This research emphasizes the importance of speech-based interactions and highlights the need for more progressive empathetic approaches[7]. The proposed method leverages DialoGPT and T5, driven by a strong data augmentation strategy and advanced NLP techniques. By focusing on speech-based unities, this research underscores the significance of

naturalistic communication in AI systems, enhancing user comfort and engagement. The integration of these technologies aims to bridge the gap between human-like empathy and AI-driven support systems[23].

Karna et al. (2023) observed that transformer-based models like BERT and RoBERTa are effective in detecting depression on social media platforms[5]. By experimenting with sophisticated GPUs and various AI frameworks, including LSTM, CNN, and BiLSTM, these models were enhanced with BERT and RoBERTa layers. The methodology improved by integrating advanced technologies like DialoGPT and T5, creating a system that aims for precision and efficiency in depression detection[8]. This approach not only enhances the accuracy of mental health analysis but also provides a framework for early intervention and support.

Bird and Lotfi (2023) investigated the efficiency of high-level model chatbots for treating mental health issues such as anxiety and depression. Their chatbots achieved high accuracy, with predictions reaching 96.49% and 97.88%. This hybrid study combined DialoGPT and T5 technologies with advanced NLP methods and data augmentation, enhancing chatbot capabilities and providing personalized support to users. The goal was to create a system that could predict user needs accurately and respond empathetically, fostering a supportive environment for mental health care.

The investigation by S. Alharbi et al. (2021) highlights the crucial role of transformers in speech patterns, emphasizing the importance of large datasets for resilience and accuracy. Transformers have become the backbone of modern speech recognition systems due to their ability to handle complex linguistic structures and nuances. Trabelsi et al. (2023) conducted a comparative study on ASR tools, focusing on accuracy and inference time, and suggested that Kaldi is highly successful due to its efficiency and robustness. Kaldi's performance in various environments makes it a preferred choice for applications requiring high accuracy and fast processing times.

Ao et al. (2022) developed the SpeechT5 framework, which specializes in spoken language processing. It uses a shared encoder-decoder network supported by pre-nets and post-nets, trained on a large volume of unlabeled data. SpeechT5 excels in tasks such as speech recognition, speech translation, and speaker identification, making it a versatile tool for speech-based applications. The framework's ability to handle diverse speech patterns and languages enhances its utility in global communication systems.

Liu et al. (2023) addressed the limitations of contemporary speech technology by introducing the Massively Multilingual Speech (MMS) project. MMS leverages self-supervised learning on religious texts from over 1,000 languages to overcome linguistic barriers, improving multilingual speech recognition without requiring extensive labeled data. This approach not only expands the reach of speech recognition systems but also provides a cost-effective solution for developing countries where linguistic diversity is high[11].

Kaur (2023) emphasized the potential of machine learning (ML) in mental health analysis, particularly in handling unstructured data from social media and medical records. However, Kaur noted challenges such as small datasets and limited accuracy[19]. The author advocated for multimodal approaches combining text, image, video, and audio to improve ML performance in detecting depression, highlighting the potential for more effective mental health analysis. By integrating diverse data types, ML models can better capture the complexities of mental health issues, leading to more accurate diagnoses and personalized interventions.

In conclusion, the integration of advanced AI technologies like DialoGPT, T5, and MMS offers a promising path forward for empathetic conversational AI systems. These systems have the potential to revolutionize mental health support by providing personalized, empathetic, and accessible care to individuals worldwide. As research continues to evolve, the focus on multimodal approaches and advanced NLP techniques will be crucial in enhancing the effectiveness and reach of these systems.

III. METHODS

Text Generation for Conversions

To develop an empathetic mental health chatbot, the team explored the world of open-source language models, including BERT, RoBERTa, DistilBERT, and others. These models are akin to enormous AI minds, pre-trained on extensive datasets, which significantly reduces the time and effort needed to create a new model from scratch. The team trained and fine-tuned DialoGPT and T5 Transformers on a custom dataset to generate empathetic responses that offer solutions and view problems from multiple perspectives, such as addiction, psychological, and social lenses. The approach of combining retrieval and generative methods has been demonstrated in models like those created by Beredo et al.. The novelty in this approach lies in using a custom-generated dataset combined with other publicly available datasets, converted into a dialogue pair format suitable for the system[16]. The model was trained to generate empathetic responses with humor, which can be disabled by the user if not receptive. BERT, an older text generation model, performed well with a perplexity of 18.4 and excelled at 81% of the tasks. However, its improved version, RoBERTa, was more comprehensible (perplexity 15.9), completed more tasks correctly (85%), and provided faster responses. DistilBERT, a distilled version, was smaller and faster but slightly less precise (perplexity 19.2, tasks 78%). DialoGPT surprised the team with its understanding capabilities, showing the least perplexity (12.7) and excelling at most tasks (91% completion). It uses local attention in every other layer with a window size of 256 tokens, similar to GPT2 but with improvements. DialoGPT is a conversational model trained on a unique dataset of Reddit conversations collected over several years, enabling it to generate responses that are informative, engaging, and closely reflect human dialogue dynamics.

Speech to Text/Automatic Speech Recognition

The team tested the Whisper large v2 speech recognizer developed by OpenAI, which excelled at navigating accents and providing near-accurate transcriptions, making it perfect for the task[18]. Whisper is a state-of-the-art speech recognition model trained on a massive dataset of around 680,000 hours of audio and text. It can transcribe speech with high accuracy, even in noisy environments. Whisper large V2 is used for humanizing the mental health app due to its impressive multilingual capabilities and

audio characteristics, which are encouraging for further research in mental health data processing. Comparing the transcript of each model against the original audio recording was crucial. The team observed aspects such as transcription accuracy, emotional tone management, and prevention of hesitation, which can affect mental health conversations. This work focuses on a qualitative approach, revealing the strengths and weaknesses of Whisper large V2, the baseline model in the framework of the application [22]. The performance of the ASR system was evaluated using Word Error Rate (WER) and accuracy. WER provides a detailed measure of errors made by the system, while accuracy shows the overall percentage of correctly transcribed words. WER is calculated as:

$$WER = \frac{Substitutions + Insertions + Deletions}{(No.of\ words\ (Reference)}$$

For instance, if the reference transcription is "The feline took a seat on the desk" and the ASR transcription is "The female took a seat on desk," the WER would be calculated as follows:

- Substitutions = 1 ('feline' as 'female')
- Insertions = 0 (No new word inserted)
- Deletions = 1 ('the' missing in ASR Transcription)
- No. of words in reference transcription = 8

Hence, the WER would be 2/8, i.e., 25%. Accuracy is calculated using a predefined formula similar to WER:

$$Accuracy = \frac{No.of\ words - Substitutions - Insertions - Deletions}{No.of\ words\ in\ reference} * 100$$

For supporting multilingual conversations, the team employed Meta's Massively Multilingual Speech, which is developed upon VITS by training it on religious text reader recordings in over 1,000 languages. This model is used to convert translated text (from English to the required language) into speech.

Table 1: Comparison of ASR Models

Model	*Word Error Rate (WER)*	ACCURACY (%)
Vosk API	9.7%	90.3
Whisper v2 (Large)	8.5%	91.5

Text to Speech

Adding multi-modality to conversations can provide a personalized touch and help users build trust, enhancing usability. The team rigorously tested popular Text-to-Speech models like Tacotron 2, known for generating naturalistic human-like speech with natural inflections and pauses. Tacotron 2 uses a neural network architecture with an attention mechanism and a post-processing vocoder to create realistic speech directly from text. Another model, VITS, is efficient and generates high-quality speech using fewer parameters than similar models. It combines the strengths of autoregressive and GAN-based text-to-speech models[27]. The Speech T5 model is flexible and handles various speech tasks such as speech recognition, translation, and text-to-speech synthesis. Based on the T5 Transformer architecture, it is adept at comprehending and manipulating spoken language. Speech T5 produces audio waveforms from text without intermediary representations, leading to better quality and simplified pipelines. It also allows manipulation of speech features like pitch, speed, and emphasis, enabling the generation of sophisticated, subtle, and natural-sounding artificial voices. It is quick and effective, making it an ideal choice for the system.

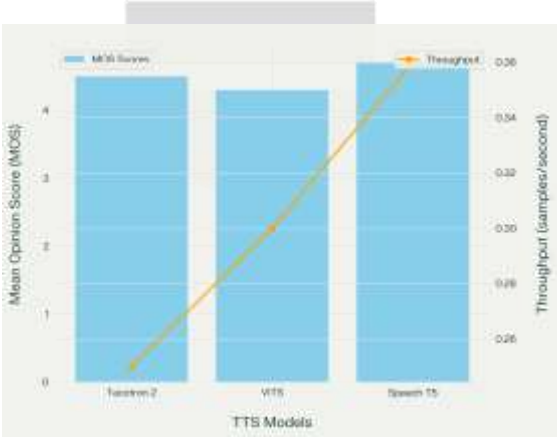

Figure 1: Comparison of TTS Models by MOS and Throughput

Dataset

The dataset was created using PersonaChat data by Meta, a cleaned dataset from Reddit containing 7,650 records, the Mental Health Corpus with text and records related to anxiety, depression, and mental health issues (27,972 records), and a dataset named Mental Health Chatbot created by Mark Daniel Lampa with around 100 dialogue pairs. This diverse dataset ensures that the model can handle a wide range of mental health topics and user interactions, providing empathetic and personalized support.

Integration of Models

The system integrates various cutting-edge technologies to ensure coherent performance. The codes for the Automatic Speech Recognition (ASR) model, specifically Whisper Large v2, are combined with those for text/dialogue generation and summarization using T5 Transformers and DialoGPT[21]. Additionally, the system incorporates text-to-speech functionality using Speech T5, along with translation modules. An Application Programming Interface (API) was developed to integrate all these components seamlessly into the backend of a web application.

This integration allows the system to handle multiple tasks simultaneously, such as speech recognition, text generation, and speech synthesis, providing a comprehensive conversational experience. The use of an API ensures that these functionalities can be easily accessed and managed through a unified interface, enhancing the overall user experience.

Testing

Rigorous manual testing rounds were conducted to evaluate the system's usability and performance. Feedback from users and mentors was incorporated to refine the model further. ChatGPT was utilized to create human-like test cases, simulating real-world interactions to assess the system's conversational capabilities. The model was also tested for its multilingual performance to ensure it could support users across different languages[20]. These testing phases were crucial in identifying areas for improvement and ensuring that the system could adapt to diverse user needs and preferences. By employing both manual and automated testing methods, the team aimed to create a robust and reliable conversational AI system.

Results

This research focuses on key natural language processing tasks, including text generation, summarization, translation, automatic speech recognition (ASR), and text-to-speech (TTS) synthesis. The system leverages a powerful ensemble of cutting-edge methods to address these challenges effectively. For text generation, DialoGPT and T5 Transformers were chosen due to their ability to produce contextually appropriate and diverse outputs. Google Translate was used for translation tasks, proving efficient in converting text across various languages. Whisper Large-v2 was the primary ASR model, offering precise speech-to-text translation even in challenging acoustic environments[14]. Speech T5 excelled in TTS synthesis, generating high-quality synthetic speech in English, while Meta's Massively Multilingual Speech model supported numerous languages, though with slightly less natural results.

IV. Conclusion

This work proposes a novel multimodal conversational agent designed to understand user queries and generate empathetic, human-like responses for mental health support. The system incorporates humor in conversations when users are receptive, aiming to create a natural and light-hearted interaction. By leveraging large language models like T5 and DialoGPT, along with Whisper Large v2 for ASR and Speech T5 for TTS, the model engages in empathetic two-way conversations via text and audio. The system operates primarily through Retrieval Augmented Generation (RAG) of responses from the dataset, generating new responses when necessary. It employs prompt engineering and fine-tuning on a custom dataset created through web scraping and data augmentation. Experiments showed that the agent could understand user intentions, engage in empathetic discussions, and lighten moods with humorous responses[29]. While still a work-in-progress, this conversational AI system has the potential to make mental healthcare more accessible globally, reducing stress and positively impacting anger. Future work will focus on expanding the agent's knowledge base, improving personalization, and conducting user studies to evaluate real-world impact.

V. Future Scope

The future scope of AI-driven mental health systems offers numerous opportunities for innovation, particularly in enhancing cultural sensitivity, predictive capabilities, and personalized care. Future research should focus on optimizing these systems by integrating art-based interventions to reduce anxiety and improve emotional well-being. For instance, AI systems can incorporate culturally sensitive art therapies, such as digital art generation or music therapy, which align therapeutic practices with users' cultural identities. Additionally, linguistic nuances should be addressed by training models to understand dialect-specific idioms, metaphors, and social norms. Collaborative efforts with anthropologists and linguists could improve contextual awareness, ensuring that responses resonate with regional values and cultural experiences. Expanding language support is another critical area of development. Leveraging advancements in massively multilingual models like Meta's Massively Multilingual Speech (MMS) can extend support to underrepresented languages. Although current systems cover over 1,000 languages, fine-tuning for low-resource dialects remains essential for equitable access. Community-centric design should also be prioritized by partnering with local mental health organizations to tailor systems for marginalized groups, including LGBTQ+ individuals and indigenous communities. This approach would address unique stressors such as discrimination or acculturation challenges. Predictive behavioral modeling presents a promising direction for proactive mental health care[1][14][3]. Real-time risk detection can be achieved by integrating data from wearable devices, such as sleep patterns and heart rate variability, along with social media activity to predict emotional states. Machine learning frameworks like XGBoost could enable early intervention by identifying behavioral shifts indicative of crises. Developing in-depth user profiles using federated learning would allow systems to analyze historical interactions while preserving privacy. These profiles could adjust support strategies dynamically, such as recommending mindfulness exercises during detected stress spikes or suggesting tailored interventions based on individual needs. Hybrid human-AI support systems represent another critical area for future development. Seamless expert referrals can be implemented through API-based pathways that connect users to licensed therapists during high-risk scenarios[16]. For example, chatbots could triage severe cases to video consultations while providing clinicians with interaction summaries to reduce diagnostic delays. Continuous learning mechanisms should also be deployed to refine responses based on user feedback. Such systems could retain context from prior conversations to avoid re-traumatization and improve rapport with users over time. Institutional applications of these technologies hold significant potential for improving mental well-being in workplaces and educational settings[5]. AI tools embedded into employee wellness platforms can address burnout and productivity challenges by mediating conflict resolution or suggesting micro-breaks during high-

stress periods. Similarly, in universities, AI systems can support students navigating academic pressures by using sentiment analysis of journal entries to identify at-risk individuals and provide timely assistance.

Ethical and technical considerations must also be addressed in the future development of these systems. Bias mitigation is essential to ensure that care quality remains consistent across demographic groups[3]. Regular audits should be conducted to identify disparities in depression detection accuracy across gender and ethnic cohorts. Furthermore, localized automatic speech recognition (ASR) models should be developed to comply with regional data regulations like GDPR, particularly in the European Union and Global South regions. These advancements have the potential to transform mental healthcare into a proactive, culturally attuned, and universally accessible ecosystem. By bridging technological innovation with human-centered design principles, future systems could significantly reduce global treatment gaps while fostering resilience across diverse populations.

REFERENCES

[1] Ayisire OE, Babalola F, Aladum B, Oyeleye-Adegbite OC, Urhi A, Kilanko A, et al. "A Comprehensive Review on the Effects of Humor in Patients With Depression." Cureus. 2022 Sep 17;14(9):e29263. doi: 10.7759/cureus.29263.

[2] Hsu J-H, Chang J, Kuo M-H, Wu C-H. "Empathetic Response Generation Based on Plug-and-Play Mechanism with Empathy Perturbation." IEEE/ACM Trans Audio Speech Lang Process. 2023;31:2032-2042. doi: 10.1109/TASLP.2023.3277274.

[3] Raamkumar AS, Yang Y. "Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities." IEEE Trans Affect Comput. 2023 Oct-Dec;14(4):2722-2739. doi: 10.1109/TAFFC.2022.3226693.

[4] Karna P, Keshari SK, Kumar Mandal A, Chakraborty B. "BERT-Driven Deep Learning Approach for Depression Detection in Social Media Posts." Int Conf Optim Tech Learn. 2023:1-6. doi: 10.1109/ICOTL59758.2023.10435285.

[5] Bird JJ, Lotfi A. "Generative Transformer Chatbots for Mental Health Support: A Study on Depression and Anxiety." Proc 16th Int Conf Pervasive Tech Assist Environ. 2023.

[6] Alharbi S, et al. "Automatic Speech Recognition: Systematic Literature Review." IEEE Access. 2021;9:131858-131876. doi: 10.1109/ACCESS.2021.3112535.

[7] Trabelsi A, Warichet S, Aajaoun Y, Soussilane S. "Evaluation of the Efficiency of State-of-the-Art Speech Recognition Engines." Procedia Comput Sci. 2022. doi: 10.1016/j.procs.2022.09.534.

[8] Ao J, Wang R, Zhou L, Wang C, Ren S, Wu Y, et al. "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing." ArXiv preprint. 2021 Oct. arXiv:2110.07205.

[9] Kim J, Kong J, Son J. "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech." ArXiv preprint. 2021 Jun. arXiv:2106.06103.

[10] Pratap V, Tjandra A, Shi B, Tomasello P, Babu A, Kundu S, et al. "Scaling Speech Technology to 1,000+ Languages." ArXiv preprint. 2023 May. arXiv:2305.13516.

[11] Kaur V, Gupta K. "A Brief Review of Machine Learning Methods Used in Mental Health Research." Int Conf Artif Intell Appl. 2023:1-6. doi: 10.1109/ICAIA57370.2023.10169520.

[12] Beredo JL, Ong EC. "A Hybrid Response Generation Model for an Empathetic Conversational Agent." Int Conf Asian Lang Process. 2022 Nov:300-305. doi: 10.1109/IALP57159.2022.9961311.

[13] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." ArXiv preprint. 2019 Jul. arXiv:1907.11692.

[14] Sanh V, Debut L, Chaumond J, Wolf T. "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter." ArXiv preprint. 2019 Oct. arXiv:1910.01108.

[15] Zhang Y, Sun S, Galley M, Chen Y-C, Brockett C, Gao X, et al. "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation." ArXiv preprint. 2019 Nov. arXiv:1911.00536.

[16] Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, et al. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions." ArXiv preprint. 2017 Dec. arXiv:1712.05884.

[17] Mehta S, Tu R, Beskow J, Sz√©kely √â, Henter GE. "Matcha-TTS: A Fast TTS Architecture with Conditional Flow Matching." ICASSP. 2024:11341-11345. doi: 10.1109/ICASSP48485.2024.10448291.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. "Attention is All You Need." Adv Neural Inf Process Syst. 2017 Dec;30.

[19] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. "Language Models are Unsupervised Multitask Learners." OpenAI Blog. 2019;1(8):9.

[20] Devlin J, Chang MW, Lee K, Toutanova K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT. 2019:4171-4186.

[21] Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, et al. "Recipes for Building an Open-Domain Chatbot." EACL. 2021:300-325.

[22] Rashkin H, Smith EM, Li M, Boureau YL. "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset." ACL. 2019:5370-5381. doi: 10.18653/v1/P19-1534.

[23] Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. "LaMDA: Language Models for Dialog Applications." ArXiv preprint. 2022 Jan. arXiv:2201.08239.

[24] Bahdanau D, Cho K, Bengio Y. "Neural Machine Translation by Jointly Learning to Align and Translate." ICLR. 2015.

[25] Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models." ArXiv preprint. 2021 Dec. arXiv:2112.10741.

[26] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." ACL. 2020:7871-7880.

[27] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." J Mach Learn Res. 2020;21(140):1-67.

[28] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. "Language Models are Few-Shot Learners." Adv Neural Inf Process Syst. 2020;33:1877-1901.

[29] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. "Learning Transferable Visual Models From Natural Language Supervision." ICML. 2021:8748-8763.