

POLITECNICO DI TORINO

Master's Degree in ICT for Smart Societies



**Politecnico
di Torino**

Master's Degree Thesis

An AI based privacy oriented approach to administer Psychological Projective Test

Supervisors

Prof. Gabriella OLMO

Dott. Andrea COLA

CANDIDATE

Alessandro CLEMENTE

DECEMBER 2023

Abstract

In recent years, there has been a sharp rise in the demand for mental health services, with an increasing number of individuals seeking assistance for psychological and emotional challenges. This escalating need for mental health support has placed a significant strain on healthcare professionals, necessitating innovative solutions to manage the growing workload effectively.

This research introduces a pioneering approach that leverages Large Language Models (LLMs) to automate psychological triage, offering a potential solution to reduce the burden on mental health professionals. The primary objective of this study is to streamline the psychological assessment process while ensuring patient anonymity and data privacy.

The proposed framework centers on the administration of the Thematic Apperception Test (TAT) on edge devices, followed by facial emotion recognition and text transcription performed by existing state-of-the-art Neural Networks. The collected data is securely encrypted and transmitted to an LLM for comprehensive analysis. The result is the generation of an extensive PDF report, which can be shared with specialists to gain an initial understanding of the patient's profile.

In a world where the demand for mental health services continues to grow, this research presents a novel perspective on psychological triage, aiming to strike a balance between efficiency, patient privacy, and in-depth analysis. This work includes an introduction to the problem statement, a theoretical background of the tools used to realize the framework, a description of the methodology adopted, the results obtained as well as some challenges and possible future development.

Keywords— Mental Health, Psychological Triage, Large Language Models (LLMs), Thematic Apperception Test (TAT), Emotion Recognition, Data Privacy, Automation, Telepsychology, Privacy, Edge devices.

Table of Contents

List of Tables	IV
List of Figures	V
Acronyms	VIII
1 Introduction	1
1.1 General Context	1
1.2 Proposed Framework and Rationale	1
2 Literature Review	3
2.1 Telemedicine	3
2.1.1 Definition of telemedicine	3
2.1.2 Telemental health and Telepsychology	4
2.1.3 Impact of COVID-19 on telepsychology	4
2.1.4 Regulatory Aspects: Privacy and Informed Consent.	6
2.1.5 Risks related to digital platforms for telepsychology	7
2.2 Thematic Apperception Test (TAT): Origins and Principles	7
2.2.1 Foundations of the TAT	8
2.2.2 Methodology and Administration	8
2.2.3 Interpretation and Psychodynamic Insights	8
2.2.4 Evolution and Critique	9
2.3 Large Language Models - LLMs	9
2.3.1 Transformers	10
2.3.2 Transformer architecture	10
2.3.3 RNNs and LTSMs challenges	13
2.3.4 Positional Encoding	14
2.3.5 Self-Attention	16
2.4 LLMs in the field of Mental Health	18
2.4.1 LLMs' potentialities	19

2.4.2	Existing framework	21
3	Proposed Framework	24
3.1	High level design	24
3.2	Implementation	25
3.2.1	Thematic Apperception Test on edge device	25
3.2.2	FER and Speech to Text	26
3.2.3	Story Encryption	29
3.2.4	LLM analysis	31
3.2.5	Report creation	37
4	Testing and Results	39
4.0.1	Sentiment Analysis Models	39
4.0.2	Pre-Diagnosis and Analysis	40
4.0.3	Conclusion	41
5	Challenges and Future works	42
5.1	Quantization	42
5.2	Audio analysis integration	44
5.3	Specialized LLM	45
	Bibliography	46

List of Tables

3.1	Emotion Model Structure. Table from "HyperExtended LightFace: A Facial Attribute Analysis Framework"	28
-----	---	----

List of Figures

2.1	Transformer Architecture	11
2.2	Positional Embedding Effects: Curves Vary with i . The image demonstrates how curves change at different i values, indicating various frequencies. Distant positions show early deviations in y -coordinates, emphasizing the role of position and embedding dimension in capturing word order within the Transformer model [22]	16
2.3	(left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.	17
2.4	Example of attention matrix after training with dimension 7, not normalized [23]	18
2.5	Mental health professionals report increased workloads since the start of the pandemic as demand for therapy soars. Image: American Psychological Association	19
3.1	High level workflow of proposed framework	25
3.2	Sample of TAT images used	27
3.3	A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets [31]	29
3.4	32
3.5	33
3.6	Example of prompt and good response of Falcon-7B	33
3.7	GPT 3.5 response to same prompt as Falcon 7b	34

3.8	GPT 3.5 response when incorporating list of emotions displayed during the test administration. Image taken from the pdf report created with the framework developed.	35
3.9	GPT 3.5 emotion analysis of a story.	36
3.10	GPT 3.5 common theme analysis among all the stories. Image taken from the final pdf report.	36

Acronyms

APA

American Psychological Association

ASR

Automatic Speech Recognition

BERT

Bidirectional Encoder Representations from Transformers

GPT

Generative Pre-trained Transformer

LLM

Large Language Model

NLP

Natural Language Processing

TAT

Thematic Apperception Test

Chapter 1

Introduction

1.1 General Context

Access to mental health services has become an issue of increasing relevance in our rapidly evolving society. The number of people in the modern world who are seeking help for mental health issues is constantly rising. This trend reflects a variety of factors, including escalating levels of stress, anxiety, depression, and other psychological conditions related to the complexities of modern life. However, the fundamental issue emerging from this situation is the limited accessibility to mental health services. Many people in need of psychological support face barriers preventing them from obtaining timely assistance.

Restricted access to mental health services can have significant impacts on the quality of life for those affected, delaying the diagnosis and treatment of psychological conditions. Moreover, the growing demand for help places considerable strain on available resources, resulting in lengthy waiting lists and placing mental health professionals under an increasingly demanding workload.

This challenge underscores the need to develop new approaches and innovative solutions to make mental health services more accessible and manageable. This thesis delves into the exploration of an Artificial Intelligence-based framework, focusing on its potential to streamline and automate the work of mental health professionals. By providing clear data and initial patient analysis following psychological assessments, this research contributes to improving the accessibility and manageability of mental health services in our increasingly digitalized world.

1.2 Proposed Framework and Rationale

In response to the challenges outlined in the preceding sections, this thesis presents a comprehensive framework that leverages the capabilities of Artificial Intelligence

(AI) and Large Language Models (LLMs) to automate and streamline the psychological triage process. The proposed framework aims to address the increasing demand for mental health services and the growing workload of mental health professionals. It seeks to improve accessibility and offer a preliminary patient analysis by collecting, analyzing, and summarizing data obtained during psychological tests conducted on edge devices.

The rationale behind the development of this framework is rooted in the need for more efficient and user-friendly mental health support systems. With a substantial portion of the population seeking psychological assistance, there is an urgent requirement to reduce the burden on mental health professionals and expedite the patient assessment process. By harnessing AI, speech-to-text, and emotion recognition technologies, the framework intends to provide a preliminary analysis of patients' emotional and psychological states, enabling quicker interventions and improved resource allocation.

In the following sections, this thesis will delve into the intricacies of the research. Chapter 2 is dedicated to an extensive examination of key areas, including telemedicine, specifically telepsychology, the Thematic Apperception Test (TAT), and Large Language Models (LLMs) and their role in mental health. Each of these subjects will be explored in-depth to provide a comprehensive understanding of their relevance and significance to the study. Chapter 3 will detail the methodology employed in the research and how the proposed framework was implemented. Chapter 4 will present the results and findings obtained during the development and testing phases, shedding light on the framework's potential and the challenges faced. Chapter 5 will address the obstacles confronted during the framework's implementation. It will also shed light on promising directions for future research and improvements that could enhance the developed solution.

Chapter 2

Literature Review

2.1 Telemedicine

2.1.1 Definition of telemedicine

The definition of telemedicine has evolved over time, primarily due to its wide-ranging application within the healthcare sector. Various formulations have been proposed in the literature, all sharing the core idea that information should travel instead of the patient.

In 1975, K.T. Bird described telemedicine as «the practice of medicine through audio-video telecommunications, without the need for physical encounters between the physician and the patient.» In 1983, D.W. Conrath defined it as «the use of communication technologies to enhance healthcare services.» According to the World Health Organization (WHO), telemedicine is «the delivery of care and assistance in situations where distance is a critical factor, by healthcare providers using information and communication technologies to exchange information useful for diagnosis, treatment, prevention, research, evaluation, and continuous training of medical personnel, for the benefit of individual and community health.»

In 1990, the European Union agreed that telemedicine includes «the monitoring, management, and education of patients and staff through the use of systems that enable timely access to consultations with experts and patient information, regardless of their location.» Telemedicine is applied in various medical domains, assuming different names depending on the context, such as teleradiology, home telehealth, telecardiology, telepathology, teledermatology, telerehabilitation, teleconsultation, telepsychology and tele-education.

2.1.2 Telemental health and Telepsychology

In the field of telemedicine, There are various branches that address various medical needs. One important branch is "Telemental Health" or "eMental Health". Telemental health involves providing mental health services using digital tools and communication technologies. This approach allows people to access therapy, counseling, and support for their mental well-being remotely [1]. A related aspect within telemental health is telepsychology. This refers to the practice of delivering psychology services from a distance. Psychologists can use video calls and digital platforms to assess patients, diagnose conditions, offer therapy, and prescribe medications, even when they are far apart. This use of technology is transforming how we approach mental healthcare and is addressing challenges like distance and limited access to care. By embracing telemental health and telepsychiatry, healthcare providers are making mental health support more available and inclusive for everyone. [2].

2.1.3 Impact of COVID-19 on telepsychology

The COVID-19 pandemic has had a profound impact on the world of telemedicine and telepsychology, as it has created an urgent need for remote and accessible healthcare services. As pandemic unfolded, telemedicine gradually became a considerably valuable service to avoid contagion between healthcare professionals and patients and has encompassed a growing number of medical disciplines [3]. The coronavirus (COVID-19) pandemic has not only affected the physical health of millions of people worldwide, but also their mental health. The pandemic has caused various psychological problems, such as anxiety, depression, stress, and trauma, among different populations, such as children, elderly, people with disabilities, and health workers [4]. Telepsychology can be effective for many people, providing an alternative option to traditional in-person psychiatric services. That could prove even more beneficial in light of the fact that the COVID-19 pandemic has caused enormous mental and psychological suffering affecting millions of people all over the world, hence such new methods are probably even more pertinent and useful for helping as many people as possible who require assistance..

Telepsychology has been shown to be effective for a variety of mental health problems and populations, such as depression, anxiety, post-traumatic stress disorder, substance use, and eating disorders. A systematic review found that telepsychology delivered by video and phone is effective for depression, anxiety, and adjustment disorder. Research has also indicated that telepsychology can achieve similar outcomes and satisfaction levels as face-to-face psychotherapy. For example, [5] found that in-person supervision and telesupervision had equivalent outcomes; there was no significant decrease in participant-rated supervision satisfaction or supervisory working alliance after participants switched from the initial in-person supervision

format to telesupervision. Prior to the COVID-19 pandemic, in the USA, psychologists performed 7.07% of their clinical work with telepsychology, which increased 12-fold to 85.53% during the pandemic, with 67.32% of psychologists conducting all of their clinical work with telepsychology [6].

However, as Nancy McWilliams [7] has explained by sharing her experience, numerous psychotherapists face different challenges when spending extended periods in front of a computer screen. Although the ability to gain insight into patients' living environments is advantageous, this shift to digital platforms can be demanding.

To offer guidance and support to psychologists and psychotherapists, the "Society for Psychoanalysis and Psychoanalytic Psychology" has provided certain recommendations [8]:

- The pandemic experience is uniquely perceived by each individual. The role of psychotherapists entails making patients cognizant of the situation while avoiding the assumption of being the sole problem solver that led patients to seek therapeutic assistance.
- Ensuring the safety of patients and alleviating uncertainties associated with remote sessions holds paramount importance. A prudent selection of suitable devices and platforms is crucial to maintain the privacy and confidentiality integral to these therapeutic methodologies.
- The elevated empathetic engagement demanded in virtual consultations can lead to fatigue and exhaustion among healthcare providers. Identifying symptoms and respecting intervals for breaks and rest are crucial. Additionally, therapists must sustain focus throughout sessions, which can be challenging after prolonged exposure to digital screens. Attaining an optimal level of empathy with the patient is intricate, and even if achieved, a physical disconnect may persist, exacerbating the challenge of bridging this gap.

These insights represent only a portion of the recommendations presented by the APA, serving as guiding principles for those seeking to integrate psychotherapy sessions with telemedicine beyond the emergency phase. This approach also encourages therapists to embrace innovative digital tools. The crisis has presented an occasion for both patients and practitioners to recognize the potential for new modes of action, particularly in times of heightened difficulty within the field of mental health.

The American Psychological Association emphasizes the importance of mental health therapy professionals becoming familiar with the dangers associated with handling sensitive data and managing confidential information. It is crucial not only to ensure the confidentiality of information discussed during consultations but

also to effectively communicate precise details to patients about the mechanics of teleconsultations, including potential risks [9].

The importance of meticulous health data management will be covered in more detail in the paragraph that follows.

2.1.4 Regulatory Aspects: Privacy and Informed Consent.

The integration of telepsychiatry into mental health services offers unprecedented convenience and accessibility for patients and practitioners. However, along with these advantages come potential risks and challenges, particularly concerning the privacy of sensitive patient information. As telepsychiatry involves the transmission of personal and confidential data over digital channels, it is crucial to address privacy concerns adequately.

One primary concern is the security of electronic communication and data storage. Patient medical records, diagnoses, treatment plans, and other sensitive information are susceptible to unauthorized access, hacking, or data breaches. In Italy, the health data is subject to the regulations outlined in Title V of Legislative Decree No. 196 of 2003, which distinguishes it from sensitive data only when processed by healthcare personnel or healthcare organizations for the purpose of safeguarding the health of the individual or the community. Should these data be used for other purposes, the classification as health data ceases, and they are considered sensitive data, thus falling under a different set of regulations. Safeguarding this information against cyber threats is paramount to maintain patient confidentiality and trust in telepsychiatry services and telemedicine in general.

It is crucial to emphasize that when health data leave the sphere of healthcare and is processed by entities other than healthcare organizations, robust measures such as encryption and anonymization become imperative. These measures are vital in ensuring that the data cannot be traced back to the individual owner, even indirectly. This proactive approach not only upholds the privacy rights of individuals but also aligns with the broader principles of data protection and confidentiality. Another crucial aspect to consider is that of informed consent [10]. This process empowers individuals with comprehensive knowledge, enabling them to make informed decisions regarding their personal healthcare. Within the ethical responsibility of healthcare providers, it becomes essential to transparently communicate all potential risks, including those stemming from the absence of physical contact or the potential inability to provide immediate assistance during emergencies. Furthermore, it is the duty of healthcare professionals to elucidate the various technological tools in use, along with the protocols ensuring data protection. This effort should also involve identifying and addressing any concerns or anxieties the patient may hold. Hence, a comprehensive grasp of, and adherence to, the relevant guidelines, regulations, and standards is mandatory. This commitment to

informed consent serves as the foundational step in establishing a constructive and positive relationship with the patient.

2.1.5 Risks related to digital platforms for telepsychology

The use of digital platforms for telepsychology introduces the risk of unintended exposure. Patients participating in remote sessions may inadvertently disclose sensitive details of their personal lives if not provided with a private and secure environment. It is essential for both patients and practitioners to ensure that they are in a confidential space during sessions to prevent unauthorized individuals from overhearing or witnessing confidential conversations.

Another consideration is the potential for inadequate encryption or weak security measures in the chosen telecommunication platforms. Insecure platforms could compromise patient data and undermine the confidentiality of therapeutic interactions. Mental health professionals must carefully select and utilize platforms with robust encryption and security features to mitigate such risks effectively.

To address these challenges, regulatory bodies and professional organizations have issued guidelines and recommendations for secure telepsychology practices. Mental health practitioners engaging in telepsychology must adhere to these guidelines to ensure the protection of patient privacy. Additionally, staying informed about evolving cybersecurity threats and regularly updating security protocols is vital to maintain the integrity and confidentiality of telepsychology services.

2.2 Thematic Apperception Test (TAT): Origins and Principles

The Thematic Apperception Test (TAT) is a projective test that, after a period of limited use in clinical practice, has been regaining a fundamental role in psychodiagnostic investigation and the initial phase of psychological treatment in recent years. The tool is extremely useful in understanding the psychological functioning of the individual, whether they are an adult or an adolescent. The TAT, along with the famous Rorschach Inkblot Test, belongs to the sphere of projective tests. These are a type of personality test in which the patient is asked to give responses to ambiguous scene, words or images [11]. A person's response are thought bring to light the hidden conflicts or emotions, that are later to be addressed through psychotherapy or other treatment. The TAT is thought to provide valuable information regarding various aspects of the patient's functioning, such as the level of patient's dependency/emancipation or the management of emotion.

2.2.1 Foundations of the TAT

The TAT was developed as a measure of Henry Murray's need theory. Murray proposed a set of psychological needs that determined personality. He also defined common environmental forces—presses—which acted on personality and behavior. Murray believed that the projective responses to the ambiguous TAT cards would reveal an individual's needs and presses. Currently, the TAT is used in clinical as well as research settings to measure personality constructs. [12]

2.2.2 Methodology and Administration

The TAT consists of 31 black and white enigmatic pictures designed for use as follows –11 were designed for all sexes and ages (MF cards), 7 for only girls and women (GF cards), 7 for only boys and men (BM cards), 1 for only women (F card), 1 for only men (M card), 2 for only boys (B cards), 2 for only girls (G cards), and 1 blank card for both men and women [13]. The examiner begins by selecting 20 relevant cards and divides them into two sets of 10 cards, to be administered a week apart. Each set is presented based on a predetermined sequence, with the first set reflecting everyday situations and the second presenting more unusual and dramatic situations [14]. Participants are presented with these visual stimuli and asked to tell a story for each picture including the following: what has led up to the event shown, what is happening at the moment, what the characters are feeling and thinking, what the outcome of the story may be.

2.2.3 Interpretation and Psychodynamic Insights

The Thematic Apperception Test (TAT) provides valuable insights into an individual's personality and how they interact with the world around them. When people respond to the TAT cards, which show unclear situations, they reveal aspects of their own personality and how they see things. [15]. The stories people tell give insight into their inner world, showing their beliefs, fantasies, and how they see things. The way they structure their stories reflects their emotional state and their underlying outlook on life, whether it's optimistic or pessimistic. So, by looking at these different aspects, examiners can gain a deeper understanding of the participant's thoughts, feelings, and perspectives.

When examiners analyze responses to the Thematic Apperception Test (TAT), they usually focus on three main aspects: the content of the stories that the participant shares, the emotional tone of the stories and how the participant expresses their feelings and the participant's behaviors beyond their verbal responses.

2.2.4 Evolution and Critique

The TAT is often criticized for not being standardized, meaning there are no rules of administration or formal scoring system. Clinicians often vary in how they administer the test. The absence of stringent standardization protocols has led to concerns regarding the consistency and comparability of results across different administrations and interpreters. This variability in administration and scoring procedures may introduce subjectivity and undermine the test's reliability. Several common evaluation approaches have emerged within the realm of TAT analysis, offering valuable insights into different facets of personality and cognition. Some of these widely-used methods include:

- Defence Mechanisms Manual
- Social Cognition and Object Relations (SCOR) Scale
- Personal Problem-Solving System—Revised (PPSS-R)

2.3 Large Language Models - LLMs

Large language models (LLMs) are deep learning algorithms that can recognize, summarize, translate, predict, and generate content using very large datasets [16]. They can perform a variety of Natural Language Processing (NLP) tasks such as generating and classifying text, answering questions in a conversational manner, and translating text from one language to another. These models can create text based on given starting points called "prompts," which help them perform tasks in natural language or code.

LLMs are trained by data scientists and researchers on extensive amounts of unstructured data using self-supervised learning. In this process, the model is presented with sequences of words, some of which are missing. The model's objective is to predict these missing words, leading to the acquisition of valuable weight parameters for the model, along with embeddings for each input word.

During the inference phase, users input prompts to the LLMs—short text excerpts that guide the model's response. Each token in the prompt is transformed into an embedding which are used to predict the likelihood of various tokens that could come next. The model then selects the next token in a somewhat random manner and continues this process until a STOP token is chosen. LLMs are built on neural network architectures, particularly the *transformer* architecture, which allows them to capture complex language patterns.

There are three main types of LLMs based on the transformer architecture [17]:

1. *Autoregressive Language Models* (e.g., GPT): These models generate text by predicting the subsequent word in a sequence based on the words that came

before it. Training involves maximizing the likelihood of each word within the training dataset, considering its contextual surroundings. An illustrative example is the series of Generative Pre-trained Transformers (GPT) developed by OpenAI. Among these, GPT-4 stands as the latest and most advanced iteration.

2. *Autoencoding Language Models* (e.g., BERT): On the contrary, autoencoding models create fixed-size vector representations (also called embeddings) of input text. This is achieved by reconstructing the original input from a masked or altered version of itself. The models are designed to predict missing or masked words in the input text by leveraging the contextual context surrounding them. Notably, Bidirectional Encoder Representations from Transformers (BERT), devised by Google, exemplifies a well-known autoencoding language model. BERT's versatility extends to fine-tuning for a range of Natural Language Processing (NLP) tasks, including sentiment analysis, named entity recognition, and question answering.
3. The third category involves a *fusion* of both autoencoding and autoregressive techniques, as seen in models like T5. These models have shown promising capabilities by integrating features from both ends of the spectrum, further enriching their performance and applications.

2.3.1 Transformers

Transformers represent a significant advancement in natural language processing and other fields of machine learning. This architecture was introduced in the paper "Attention Is All You Need" [18], which marked a turning point in the neural network approach to language processing. Transformers were designed to address challenges such as handling long-range dependencies and computational efficiency in language models. By enabling parallel processing of data sequences and capturing relationships between distant elements within a sequence, Transformers have revolutionized the field of Language Models (LMs). Their widespread usage has been observed in natural language processing, leading to substantial improvements in automatic translation, text processing, and the generation of coherent and fluent text.

2.3.2 Transformer architecture

A Transformer, Fig.2.1, is composed of two main components: an encoder module, which encodes input data into an intermediate representation, and a decoder module, which converts it back into readable text. For instance, when training a model to translate English text into French, the encoder initially translates English

into an imaginary language, while the decoder—equipped with knowledge of this language—proceeds to perform the translation into French.

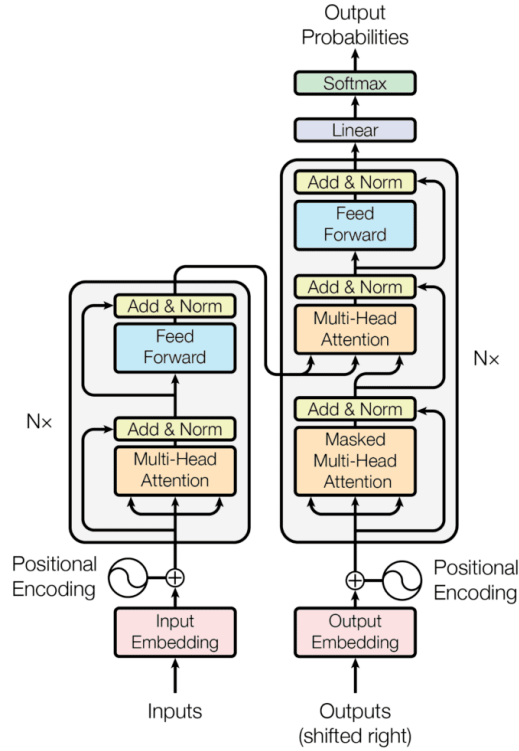


Figure 2.1: Transformer Architecture

The encoder module comprises the following components:

1. A learned embedding layer that facilitates the conversion of textual input into a machine-processable format.
2. Positional encoding, which supplements the input embedding with positional information, aiding the model in comprehending the placement of various input elements within the entirety of the input. Notably, Transformers lack recurrence, necessitating an alternative approach for incorporating positional data, achieved through positional encoding.
3. The encoder segment, repeatable for N iterations, progressively refining the encoding's accuracy. It encompasses a multi-head attention component and a feedforward component. Integrating residual connections, which involve forwarding the original input to the output, enhances training efficacy by facilitating gradient flow during backpropagation.

4. The multi-head attention element allows for the segmentation of inputs into queries, keys, and values. Analogously, consider searching for a specific Netflix title; a query is matched against series characteristics (keys), yielding relevant matches (values). By employing self-attention in the keys-values mapping, the encoder is guided to focus on specific segments of phrases when encoding particular words. This mechanism is multi-headed, meaning that queries, keys, and values are divided into N parallel components, enabling diverse perspectives on the mapping process. The outputs of these multi-head attention blocks are then summed and normalized using the residual input.
5. The feedforward component generates the encoding for each individual input, producing a high-dimensional representation. Once again, a residual connection ensures smooth gradient flow.

The decoder module endeavors to translate the imaginary language back into written language or perform other tasks, based on its training. It comprises the following components:

1. A learned embedding layer, akin to the encoder, facilitates the transformation of textual input into machine-readable format. In Vaswani et al. (2017), weight matrices are shared between input and output embedding layers. Similar to the encoder, positional encoding is applied to the embedding.
2. The decoder segment, also repeatable for N iterations, includes:
 - A masked multi-head attention segment, producing self-attention for desired outputs by focusing on relevant input words. This attention is masked to hide future words in a phrase, given that future information is unknown. This segment also features the multi-head split, residual addition, and layer normalization.
 - Another multi-head attention segment, where encoder output keys and values are merged with values from the masked multi-head attention block. This combination facilitates the mapping of "key-value" relationships to optimal outcomes, allowing for best-matching results between input and encoded phrases. Information from the masked segment is also incorporated via a residual flow.
 - A feedforward network, accompanied by a residual connection, processes the input-output combination to yield a W -dimensional vector, with W representing the word count. Employing a Softmax function, the most probable word is generated.
3. The predicted outputs are fed back into the decoder segment, enhancing the prediction of subsequent output tokens based on the encoded input token

and previous predictions. This approach has yielded remarkable outcomes in Natural Language Processing, addressing challenges of long-term memory through attention and computation speed by omitting recurrent segments. Consequently, recurrent neural networks may no longer be the primary choice for language model generation.

2.3.3 RNNs and LTSMs challenges

In the last years, in the field of NLP, there has been a significant shift from utilizing Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to adopting transformer-based models. RNNs, while groundbreaking in their ability to handle sequential data, have been associated with several challenges that have limited their performance and efficiency in natural language processing tasks.

One major issue is that of *long-range dependencies* [19]. RNNs face difficulty in capturing meaningful relationships between distant words in a text document. When processing lengthy sequences, like long text documents, RNNs struggle to retain context over extended distances. This limitation makes it challenging to maintain the coherence and understanding of complex language structures, hampering their ability to comprehend and generate longer sequences of text. Moreover, RNNs are susceptible to *gradient vanishing and explosion* during training [20]. As information is propagated back through time during the training process, gradients can either become extremely small (vanishing gradients) or excessively large (exploding gradients). This can destabilize the training procedure and lead to slow convergence or divergence during optimization, causing difficulties in effectively updating the model's parameters.

Another concern is the computational inefficiency associated with their architecture. RNNs are essentially unrolled over time, resulting in a deep network structure where each time step corresponds to a layer. Consequently, this leads to a substantial number of interlinked parameters, requiring extensive computational resources and a longer training time. This extended training period can hinder the development and deployment of RNN-based models in practical applications.

Furthermore, RNNs suffer from a *lack of parallelism* in computation [21]. Due to their inherent sequential nature, they process input data one step at a time, preventing the utilization of parallel computing architectures such as GPUs to their full potential. This inefficiency in parallel processing limits their ability to efficiently process and analyze large amounts of data, reducing overall training and inference speed.

In order to overcome some of these limitation, Long Short Term Memory networks (LSTMs) were developed [21]. LSTMs were introduced as an improvement over traditional RNNs, aiming to alleviate some of the inherent problems. One notable

enhancement is their ability to capture long-range dependencies more effectively. Through the integration of memory cells and various gating mechanisms, LSTMs can selectively retain and propagate information across multiple time steps. This allows them to better capture contextual information from distant words in a text, mitigating the issue of long-range dependencies that RNNs struggle with.

Additionally, LSTMs help mitigate the problem of gradient vanishing and explosion. By incorporating mechanisms like forget gates, input gates, and output gates, LSTMs regulate the flow of information and gradients during backpropagation, preventing them from becoming too small or too large. This stability in gradient flow enhances the training process and contributes to more efficient convergence.

Despite these improvements, LSTMs still retain some limitations and the same problem that happens to RNNs generally, happen with them, i.e. when sentences are too long, LSTMs still don't do too well. The reason for that is that the probability of keeping the context from a word that is far away from the current word being processed decreases exponentially with the distance from it.

The advent of Transformer models has revolutionized the landscape of sequence-to-sequence learning and addressed the previously discussed challenges. Transformers possess two fundamental innovations that grant them remarkable proficiency in handling extensive language models: positional encodings and self-attention [21][17][18].

2.3.4 Positional Encoding

The inherent structure of language relies heavily on the positioning and sequence of words, which in turn dictate the grammatical structure and semantic meaning of sentences. Recurrent Neural Networks and Long Short Term Memory networks, naturally incorporate word order by processing sentences sequentially, word by word. This sequential approach embeds the word order within them. Transformers on the other hand take all the hidden state embeddings all at once, this allows transformers to be much faster but as a side effect they lose the critical information related to word ordering. Consequently, they lack awareness of the specific arrangement of words within the sequence, disregarding the distinction between the first and last word. The order of words holds profound significance as it can profoundly alter the meaning of a sentence. For instance, consider the sentence "*The cat chased the dog*" versus "*The dog chased the cat.*" The simple rearrangement of words completely transforms the intended message, highlighting the pivotal role of word order in conveying accurate semantics. Consequently, a method is required to infuse this positional information into the model.

A potential solution involves imparting each word with an additional piece of information that reflects its position within the sentence. This augmentation is called "positional encoding."

One initial approach might involve assigning numerical values within the $[0, 1]$ range to represent the sequence's progression, where 0 corresponds to the first word and 1 corresponds to the final time-step. However, this approach presents challenges, notably the inability to discern the quantity of words within a specific range, rendering time-step delta inconsistent across sentences.

Alternatively, a linear assignment of numerical values to time-steps is conceivable, where in the first word is assigned "1", the second "2", and so forth. Yet, this approach poses issues, as values might become unwieldy and exceed training data parameters. Furthermore, accommodating longer sentences during inference could strain the model's applicability, potentially hindering its generalization.

The authors of [18] introduce an ingeniously simple approach to encoding. Unlike a single numeric value, the encoding is represented as a vector in a dimensional space, capturing details about a particular location within a sentence. This encoding isn't directly embedded within the model structure; instead, this vector is employed to augment each word with contextual information regarding its position in the sentence. In essence, the model's input is enriched to incorporate the sequential order of words. One challenge is that positional embedding values at a given position should remain the same independently from the text total length or any other factor. The author of the paper decided then to use wave frequencies to capture position information:

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d}}\right) \quad (2.1)$$

$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \quad (2.2)$$

where pos is the position, i represent the indices of each of the position embedding dimensions and d is the size of the position embedding and must be the same as the word embedding. That is, each dimension of the positional encoding corresponds to a sinusoid. When the curve is plotted at various values of i , a series of curves with distinct frequencies is obtained. The rationale behind these curves having varying frequencies is as follows: if two points are in proximity along the curve, their correspondence remains unchanged even at higher frequencies. It is only at significantly higher frequencies that discrepancies in y -coordinates between points on the curve emerge, allowing for differentiation. On the other hand, when points are more distant from each other, one can observe them deviating in terms of curve heights at an earlier stage. Consequently, both the positional information and the embedding dimension contribute to conveying the word order. In Fig. 2.2 this concept is illustrated.

In summary, instead of feeding words within a sentence sequentially into the neural network, thanks to positional encoding, the words can be fed in non-sequentially.

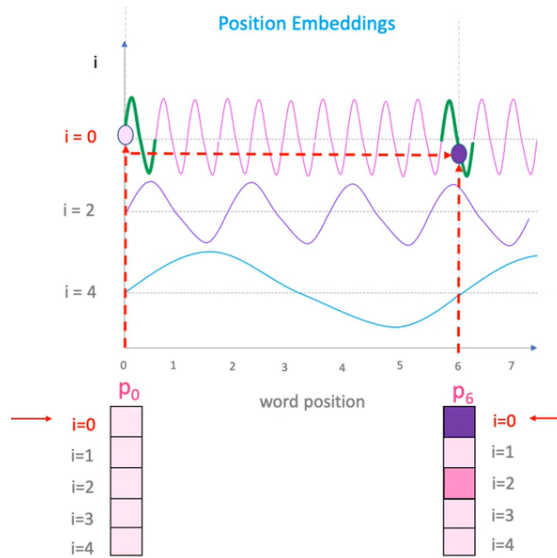


Figure 2.2: Positional Embedding Effects: Curves Vary with i . The image demonstrates how curves change at different i values, indicating various frequencies. Distant positions show early deviations in y -coordinates, emphasizing the role of position and embedding dimension in capturing word order within the Transformer model [22]

2.3.5 Self-Attention

Another key innovation that sets Transformers apart is the incorporation of the attention mechanism. Attention plays a crucial role in revolutionizing the way models process and comprehend sequential data, such as language. Attention in Transformers introduces a fundamentally different approach to information processing, allowing the model to dynamically weigh the significance of different parts of the input sequence while generating output. This ability to selectively focus on relevant segments not only enhances the model’s understanding of context but also enables the capturing of intricate patterns and dependencies that span across considerable distances. Transformers in particular use *self-attention mechanism*. The difference between simple attention and self-attention is that simple attention selectively focuses on words with respect to some external query. The more important a word is in determining the answer to that query, the more focus it is given. Self-attention on the other hand also takes the relationship among words within the same sentence into account.

In Fig. 2.3 we can see a schematized representation of the multi-head attention block.

The first component in this block are three linear layers. A linear layer is simply

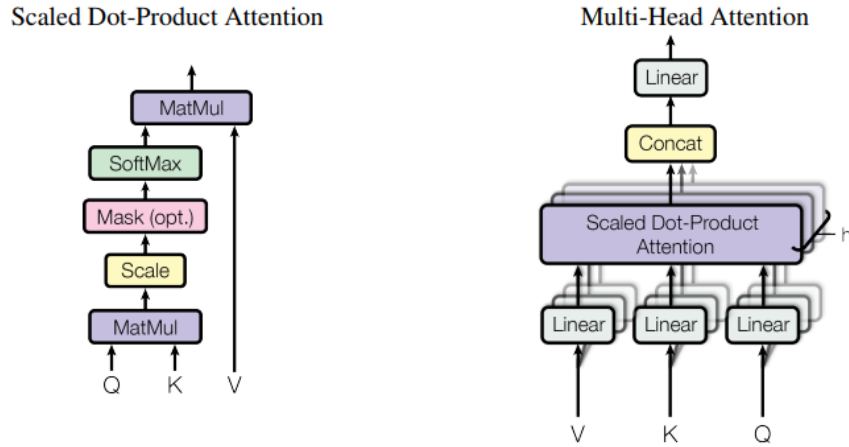


Figure 2.3: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

composed of a bunch of fully connected neurons without the activation function. They may serve two main purposes. Number one, mapping inputs onto the outputs. Number two, changing the dimensions of the inputs themselves. After embedding, the inputs are transformed using separate linear layers. This means the embedding layers are multiplied by different sets of weights for each layer. These weights are unique to each layer. After the multiplication, we get new matrices from each layer: the *query*, *key*, and *value matrices*.

The query, key, and value are fundamental components within the self-attention mechanism of the multi-head attention block in Transformers. They play distinctive roles in capturing relationships between words within the same sentence and enabling the model to comprehend contextual information.

As illustrated in Fig. 2.3 (left), a basic multiplication between the query and key matrices is performed, obtaining an entity referred to as an attention filter. Initially, the attention filter contains somewhat random numbers, but as the training process advances, these numbers start to hold more meaningful significance, representing attention scores. In the example below we can see how the highest attention is typically directed towards the word itself, and the next highest attention score is assigned to the word that closely follows. The attention scores are then scaled by a factor $\sqrt{d_k}$ and finally softmax function is used to keep the values between 0 and 1.

Subsequently, the attention filter is multiplied by the value matrix, it generates in this way a filtered value matrix which assigns high focus to the features that are more important and this is the final output of one of the Scaled Dot-Product

	When	you	play	the	game	of	thrones
When	89	20	41	10	55	10	59
you	20	90	81	22	70	15	72
play	41	81	95	10	90	30	92
the	10	22	10	92	88	40	89
game	55	70	90	88	98	44	87
of	10	15	30	40	44	85	59
thrones	59	72	92	90	95	59	99

Figure 2.4: Example of attention matrix after training with dimension 7, not normalized [23]

Attention block.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (2.3)$$

Since the authors talk about Multi-Head Attention, in the transformer structure there are multiple attention filter, each focusing on a different linguistic phenomenon. Each attention head, 8 in the original paper, therefore outputs its own attention filter that are then concatenated and passed through a linear layer to shrink back its size.

2.4 LLMs in the field of Mental Health

In recent years, the field of mental health care has been experiencing a significant surge in demand for services, accompanied by a growing need for personalized and effective therapeutic interventions. The challenge lies in providing tailored care to meet individual patient needs while also managing the increasing demand for mental health services. As the complexities of mental health issues continue to evolve, innovative solutions that leverage cutting-edge technology are emerging to bridge this gap and revolutionize the landscape of mental health care. According to a survey by the American Psychological Association, psychologist claim that there was a significant increase of demand for treatments after the pandemic, Fig. 2.5.

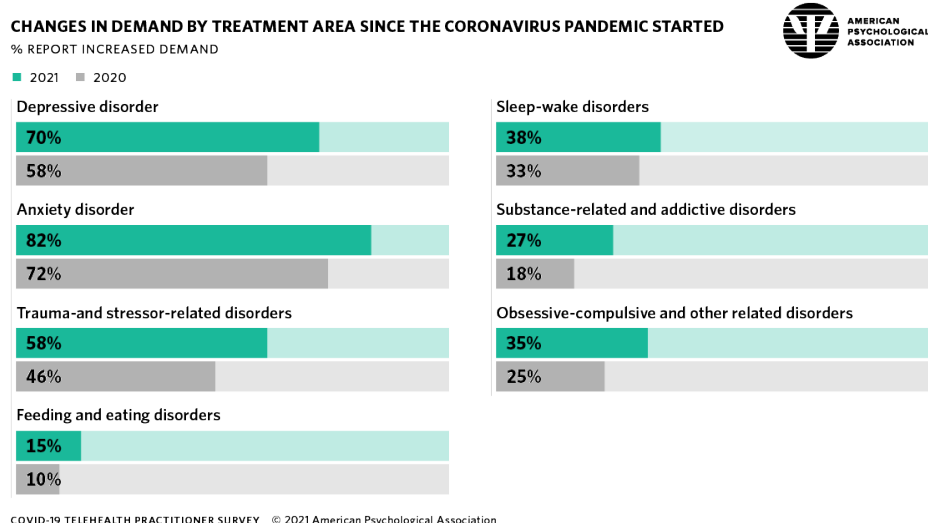


Figure 2.5: Mental health professionals report increased workloads since the start of the pandemic as demand for therapy soars. Image: American Psychological Association

To cope with the exponential increase in patient load, LLMs can triage patients based on the severity and urgency of their conditions, provide 24/7 support in case of crisis and refer the more serious cases to emergency services. LLMs can also be leveraged to generate concise summaries of each patient’s medical background, which may bring to the physician’s attention relevant patient information that may be essential in disease diagnosis. This facilitates a more efficient and comprehensive consultation, allowing the physician to dedicate more time for patient interaction and lead to greater patient satisfaction.

Large language models (LLMs) such as ChatGPT and GPT-3/4, built on artificial intelligence, hold immense potential to support, augment, or even replace psychotherapy. Enthusiasm about such applications is mounting in the field as well as industry. These developments promise to address insufficient mental healthcare system capacity and scale individual access to personalized treatments [24].

2.4.1 LLMs’ potentialities

Gaining insights into individuals’ mental states based on their language expression is a crucial aspect of mental health analysis and in this context, Large Language Models might reveal themselves as promising tools. These instruction-finetuned and general-purpose models, exhibit the remarkable ability to comprehend diverse inputs, thereby eliminating the necessity to develop multiple models for distinct

tasks. This versatility provides a strategic advantage within the field of mental health care, enabling LLMs to effectively address a multitude of challenges without the burden of training specialized models for each scenario. Research studies have probed the potential applications of LLMs in understanding and supporting various aspects of mental health.

In [25], a comprehensive assessment was conducted to evaluate the performance of both ChatGPT and GPT-3/4 in the field of mental health analysis. This evaluation encompassed a rigorous examination of eleven diverse datasets across five distinct tasks related to mental health analysis, including emotion recognition within conversations, causal entailment of emotions, detection of mental health conditions, and their underlying causes. Furthermore, diverse prompting strategies were explored to incorporate emotional information and augment the capabilities of ChatGPT. The findings underscore the robust proficiency of ChatGPT in mental health analysis and emotional reasoning within intricate conversations. In the emotion recognition task, ChatGPT with zero-shot prompting outperforms traditional supervised methods including CNN and cLSTM on IEMOCAP, MELD, and EmoryNLP datasets, indicating that its generalizability can make up for the lack of task-specific model architectures. In the Mental Health Analysis task, ChatGPT zero-shot performs significantly better than traditional light-weighted neural network-based methods such as CNN, GRU, BiLSTM_{Att} on binary mental health condition detection. The research [25], however, concludes by saying that is imperative to underscore that the performance of ChatGPT in the domain of mental health analysis exhibits notable variability with minor alterations in prompt wording. For instance, the inclusion of adjectives describing the degree of mental health conditions can significantly impact ChatGPT's performance.

Another study [26] conducts a comparative analysis of the efficacy of various LLMs, including Alpaca, FLAN-T5, GPT-3.5, and GPT-4, in mental health tasks such as detecting stress, depression, and suicide detection. The study also investigates strategies to enhance LLM performance, such as zero-shot prompting, few-shot prompting, and instruction-based finetuning. The paper's conclusions are as follows:

- Zero-shot prompting exhibits promising albeit limited LLM performance in mental health tasks. Performance can be enhanced by incorporating more contextual or domain-specific information in the prompts.
- Few-shot prompting similarly improves LLM performance by supplying a few examples per class. However, the extent of improvement varies based on the model and task.
- Instruction-based finetuning across multiple datasets significantly enhances LLM performance across a range of mental health tasks. The fine-tuned models,

Mental-Alpaca and Mental-FLAN-T5, surpass larger models like GPT-3.5 and GPT-4, and achieve performance on par with the leading task-specific model, Mental-RoBERTa.

LLMs possess distinct capacities and limitations for mental health reasoning tasks. While models like GPT-4 can generate accurate and insightful reasoning, others may generate incorrect or irrelevant reasoning. Ethical concerns and challenges associated with LLM use for mental health applications are also highlighted. The paper indicates that GPT-4 performs favorably among the evaluated LLMs. However, it still has room for improvement and faces challenges within the mental health domain. Although a direct comparison of prompting strategies is not explicit, the paper implies that instruction-based finetuning proves to be the most effective approach to enhance LLM capabilities for mental health tasks.

2.4.2 Existing framework

The American Psychiatric Association estimates there are more than 10,000 mental health apps circulating on app stores [27]. Many of these application leverage Large Language Models or artificial intelligence take the form of chatbots. One of the most popular and FDA approved application is Wysa.

Wysa is an AI-enabled Coach for mental and emotional wellness. that offers immediate support as a first step of care through its clinically validated artificial intelligence. The application is completely anonymous, without stigma and without limits. Wysa has held over half a billion mental health conversations with more than five million people in 95 countries. The application also offers structured programs and on-demand self-care through evidence-based exercises, including resources for anxiety, sleep, managing difficult conversations, and improving productivity [28].

Launched in 2017, the service provides early intervention to high-risk groups through 3 methods:

- An AI chatbot
- A library of evidence-based self-help tools
- Messaging-based support from human therapists

Wysa uses evidence-based techniques like Cognitive Behavioral Therapy (CBT), which are suggested and approved by professional psychologists for use in a self-help context, and by their scientific advisory board. As specified in the site of the company, the AI Coach is restricted in the means of response and the intended usage is for providing evidence-based tools and techniques to manage emotions and encourage mental well-being in a self-help context. It is *not intended for providing diagnosis, treatment, or cure of a condition or disease or disorders*. From the

privacy point of view, the usage of the application does not require any personal data information and all conversations with Wysa's AI Coach and Wysa Emotional Well-being Professional are private and not shared with anyone. The data provided by the users is encrypted using strong TLS encryption protocols and transmitted securely from the mobile device to the company's servers. The data is also encrypted at-rest in the storage servers using strong industry compliant AES-256 encryption protocols.

Woebot represents another existing application in the realm of mental health. As the application previously presented, *Woebot* is also an AI-powered chatbot that employs the principles of Cognitive Behavioral Therapy (CBT) to assist users in effectively managing distressing thoughts and emotions. Upon registration, users engage with *Woebot* through the app's messaging platform. The interaction involves periodic check-ins where users report their mood and provide contextual details about their emotional state. In response, *Woebot* offers tailored suggestions, tools, skills, and strategies to alleviate the reported distress. Alongside messaging, the platform offers users the ability to visualize a timeline of their mood entries and access psychoeducational content. Guided exercises, such as "challenge negativity" and "challenge stress," are also available. After each conversation, users have the opportunity to rate the quality of their interaction with *Woebot*. It's important to note that *Woebot* is not intended for use during emergencies or to manage psychiatric crises.

Koko is a peer-support platform based on the concept of crowdsourced cognitive therapy. It was designed to connect teens and adults who need mental health help to volunteers through messaging apps like Telegram and Discord. On Discord, users sign in to the *Koko Cares* server and send direct messages to a *Koko* bot that asks several multiple-choice questions (e.g., "What's the darkest thought you have about this?"). It then shares a person's concerns—written as a few sentences of text—anononymously with someone else on the server who can reply anonymously with a short message of their own. However, a few months ago, *Koko* co-founder Rob Morris announced on Twitter that his company ran an experiment to provide AI-written mental health counseling for 4,000 people: volunteers providing assistance to others had the option to use a response automatically generated by OpenAI's GPT-3 large language model instead of writing one themselves and as a result the «messages composed by AI (and supervised by humans) were rated significantly higher than those written by humans on their own ($p < .001$).» The experiment however, raised ethical concerns about the lack of informed consent, the potential risks and harms, and the regulation and oversight of AI in health-related fields.

Established in 2012, *NovoPsych* [29] was founded with the core mission of empowering mental health professionals with the latest advancements in psychometric science to enhance clinical practice. *NovoPsych* collaborates with a diverse range of stakeholders, including practitioners, private clinics, government agencies, and

non-governmental mental health services, to implement cutting-edge measurement tools. These tools are designed to track symptoms over time, support diagnosis, and facilitate case formulation, all while adhering to scientifically validated psychometric scales. The platform's commitment to continuous innovation is evident through the incorporation of normative data for the automated calculation of percentiles. Leveraging machine learning, NovoPsych offers an array of decision support tools that aid mental health professionals in their work. One significant feature of NovoPsych is its capacity for automated and remote symptom monitoring, as well as the evaluation of treatment efficacy post-discharge. Clinicians can effectively monitor patient outcomes between in-person visits or after treatment has concluded, paving the way for automated treatment program assessments.

NovoPsych's test-delivery system employs graphical representations to illustrate symptom progression over time, providing clinicians with a historical perspective of a client's journey. These easy-to-read graphs showcase the progress clients have achieved. Additionally, NovoPsych automatically computes percentile ranks, offering clear insights into a client's symptoms and their comparison to normative data. It boasts an extensive collection of over 50 assessments tailored for mental health settings. These assessments cover a wide spectrum of conditions, including depression, anxiety, PTSD, OCD, ADHD, and more. NovoPsych prides itself on delivering tests with proven psychometric properties and clinical utility. It places a strong emphasis on providing detailed information on the data used for norming and other metrics, enabling clinicians to effectively interpret the results for improved patient care. It becomes apparent that the current products available in the market, that does not require the intervention of a specialist while using them, are primarily centered around providing emotional support to individuals. These offerings are not designed to offer diagnoses or manage complex mental health issues, nor do they serve as substitutes for the administration of projective tests, which inherently require more extensive time for proper assessment. Given this context, it is noteworthy that existing solutions predominantly utilize AI technology in the form of chatbots, fulfilling the role of emotional support tools rather than providing diagnostic or pre-diagnostic capabilities intended for submission to medical professionals. While applications that assist psychologists in their assessment work do exist, they primarily pertain to psychological tests featuring closed-ended questions and well-defined evaluation standards.

Chapter 3

Proposed Framework

3.1 High level design

In this study, our objective is to establish a framework for conducting psychological triage, using the capabilities of LLMs, in an automated and anonymous manner, thereby minimizing the psychologist's involvement and reducing their workload. At a high level we can divide the functioning of the framework in five main steps:

1. *Psychological test on edge device:* The process begins by conducting a verbose psychological test on a local edge device. Clear instructions guide the user through the steps. By following user commands, the device records audio and video, allowing the user to start and stop recording between successive images.
2. *Emotion recognition and transcription :* On the edge device, the next phase involves Face Emotion Recognition and Speech to Text tasks, facilitated by the deepFace and Whisper libraries, respectively. Audio and video data are analyzed locally, ensuring sensitive information remains on the device.
3. *Encryption and Anonymization:* Subsequently, the recorded user stories undergo anonymization and are sent, along with the associated emotions displayed during the test, to the OpenAI API for analysis. This step ensures that no personally identifiable information leaves the user's device.
4. *LLM Analysis:* The encrypted data is then sent by a large language model that will analyze emotions, narratives, and recurring themes across the user's stories. It provides summaries, initial patient analysis, emotion insights, and identifies any common threads.
5. *PDF Report Generation:* Utilizing the responses and analyses generated by the LLM, a comprehensive PDF report is created. This report can be shared

with a specialist, offering an early understanding of the patient’s profile. This preparatory report enables the specialist to expedite and organize their assessment more effectively.

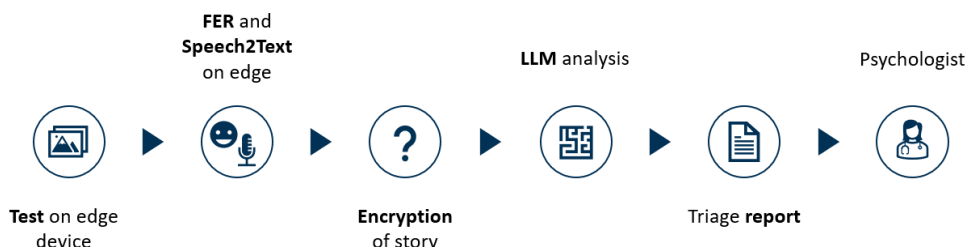


Figure 3.1: High level workflow of proposed framework

3.2 Implementation

In this sessions we are going to dive deep into the implementation of the framework, analysing which components are used, how the data is handled and how the final report is created thanks to tailored prompts for the LLM.

3.2.1 Thematic Apperception Test on edge device

The selection of the Thematic Apperception Test (TAT) as our psychological assessment tool is a result of a meticulous evaluation of various options. While considering different psychological tests, such as the Rorschach, we recognized the distinct advantages that the TAT offers in line with our framework’s goals. A key element was the TAT’s distinctive verbal format, which encourages participants to speak freely and extensively. In contrast, the administration and interpretation of tests like the Rorschach require a higher level of psychological participation. Furthermore, the TAT’s narrative-driven approach provides an ideal platform for our framework’s integrated technology components, including speech-to-text analysis and face emotion recognition. The TAT offers an open-ended style in contrast to many other assessment tests now in use for illnesses like depression, ADHD, and anxiety, which rely on closed-ended responses and deterministic interpretative norms. This distinctiveness enables our framework to capture not only textual content but also subtle emotional expressions for analysis by our Large Language Model (LLM). As a result, the TAT was found to be the best option for an automated, anonymous psychological triage system that takes advantage of the intrinsic richness and emotional depth of human expression.

To ensure the utmost privacy and confidentiality of user data, the Thematic Apperception Test is conducted on an edge device. This strategic approach ensures that sensitive information such as facial features and voice recordings never leaves the confines of the user’s device. By employing this localized execution, users can engage with the assessment confidently, knowing that their personal data remains under their control.

Audio and Video recordings

Since to our knowledge there is no Python library or solution for simultaneous video and audio recording, we had to implement a solution ourselves. For managing audio data, we leveraged the capabilities of the *pyaudio* and *wave* libraries, while for handling video data, we utilized the *cv2* library. For the Audio and Video recording process, two internal Python classes, *AudioRecorder* and *VideoRecorder*, were created within the *Recorder* class. These classes utilize the Python *'thread'* library, enabling simultaneous audio and video recording, stopping and saving. This approach leverages multithreading, a programming technique that enables different parts of a program to execute concurrently, enhancing efficiency. The TAT process commences with a user-friendly welcome message, elucidating the purpose of the test and providing clear instructions. Upon initiation, the initial trial image is presented to the user. At the user’s discretion, recordings can be initiated by pressing a designated key. When the user completes their narration, pressing another key concludes the recording. The captured data is stored locally, and the subsequent trial image is displayed.

In this study, a small set of trial images were utilized, Figs. 3.2. These images were administered to a diverse pool of 49 participants, encompassing women and men, of different ages. Verbal consent was obtained from each participant before commencing the test, ensuring adherence to ethical protocols. The images used were chosen due to their popularity and frequent usage among psychologists during test administration. Nevertheless, it’s essential to emphasize that in the framework, psychologists have the flexibility to change the selection of images as they see fit.

3.2.2 FER and Speech to Text

Once the video and audio recordings per each image are obtained and saved locally as a .avi and .wav files, they are processed by two models to extract emotions from videos and a transcription of the stories from audios. The *deepFace* and *whisper* libraries were chosen to tackle the Face Emotion Recognition (FER) and trascription tasks.



Figure 3.2: Sample of TAT images used

DeepFace

Deepface is a lightweight face recognition and facial attribute analysis (age, gender, emotion and race) framework for python. It is a hybrid face recognition framework wrapping state-of-the-art models: VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib and SFace. Age, gender, emotion, and race/ethnicity predictions are included in the module on face attribute analysis [30]. The VGG-Face model, which is the cornerstone and is renowned for being straightforward and precise, serves as the framework and is merged with pre-existing weights. The output layer of this base model was modified to fit the intended task. To encourage transfer learning during model development and training, the first layers were kept unchanged. Regarding the emotion prediction, the FER-2013 dataset was used to train the emotion model. Because this dataset has 48x48 shaped grayscale images, the emotion model is not built on VGG-Face structure. As shown in Table 3.1, the architecture built, presents five convolution layers and three fully connected layers. Its output layer has seven nodes corresponding: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral [30]

To implement the Face Emotion Recognition method within our framework, we adopted a sequential frame-by-frame analysis approach. We processed each recorded video by splitting it into individual frames using the OpenCV (cv2) library. These frames were then passed as input to the default model provided by the deepFace library, which is based on the VGG-16 architecture as described previously. For each frame, the deepFace model produced an output probability distribution across seven recognizable emotions: angry, fear, sadness, neutral, surprise, happy, and disgust. To determine the predominant emotion over a specific time interval, we aggregated the predictions from a number of frames corresponding to a 10-second segment of the video. This aggregation involved calculating the average probability

values for each emotion category within the selected timeframe. The choice of considering 10-second intervals was deliberate. It allows for a sufficiently extended time window to capture diverse emotional expressions as users speak, mitigating the bias introduced by excessively high values assigned to a particular emotion within a single frame.

	1 conv	2 mpool	3 conv	4 conv	5 apool	6 conv
Filters	64	-	64	64	-	128
Kernel	5	-	3	3	-	3
Pool	-	5	-	-	3	-
Strides	-	2	-	-	2	-
Units	-	-	-	-	-	-
	7 conv	8 apool	9 fc	10 fc	11 fc	12 softmax
Filtera	128	-	-	-	-	1
Kernel	3	-	-	-	-	-
Pool	-	3	-	-	-	-
Strides	-	2	-	-	-	1
Units	-	-	1024	1024	7	0

Table 3.1: Emotion Model Structure. Table from "HyperExtended LightFace: A Facial Attribute Analysis Framework"

Whisper

Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web [31]. The authors decided for an encoder-decoder Transformer [18] as this architecture has been well validated to scale reliably. All audio is re-sampled to 16,000 Hz, input audio is split into 30-second chunks, converted into a log-Mel spectrogram, and then passed into an encoder. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation [31]. Fig. 3.3 summarizes the model architecture.

Other than being able to handle many different type of languages, others advantages of using whisper is its robustness and its ability to handle long audio. Since Whisper models are trained on 30-second audio chunks and cannot consume longer audio inputs at once, the authors developed a strategy to perform buffered transcription of long audio by consecutively transcribing 30-second segments of audio and shifting the window according to the timestamps predicted by the model.

The library offers five different models of varying sizes, four with English-only versions, each striking an optimal balance between speed and accuracy. As one of the framework’s primary aim was to operate the ASR model on an edge device, the choice was made to select the *base* model with 74 million parameters, which is recognized for its superior performance among the available English models according to the authors.

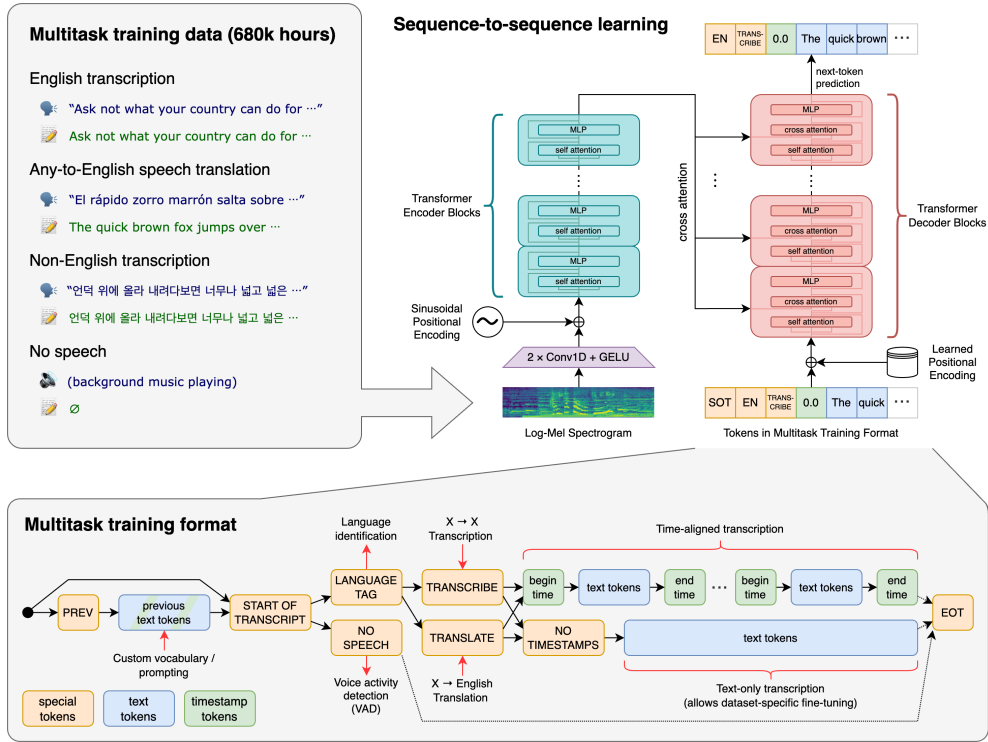


Figure 3.3: A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets [31]

3.2.3 Story Encryption

Until now, all information gathering and recording processes have taken place on the edge devices. However, in order to fully harness the capabilities of the selected Large Language Model it became necessary to engage the LLM API and transmit

the patient narratives for comprehensive analysis. Given that the data is going to leave the edge device, the need for an encryption method became apparent to ensure the complete privacy of the patients' information. The encryption process in this context employs the RSA (Rivest-Shamir-Adleman) algorithm, which offers a high level of security and privacy for patient data during the transfer from edge devices to the LLM.

RSA encryption

The RSA algorithm is categorized as an asymmetric cryptographic system. The term 'asymmetric' conveys that it operates using two distinct keys: the Public Key and the Private Key. As the nomenclature suggests, the Public Key is made available to all, while the Private Key remains confidential and is not disclosed to others [32].

In Asymmetric Encryption algorithms, you use two different keys, one for encryption and the other for decryption. The key used for encryption is the public key, and the key used for decryption is the private key. But, of course, both the keys must belong to the receiver.

RSA's security is based on the challenge of factoring large integers that result from multiplying two prime numbers. While multiplication is straightforward, determining the original prime factors—or factoring—is considered impractical due to the significant time required, even with today's supercomputers.

The most intricate aspect of RSA cryptography is the process of generating public and private keys. Two substantial prime numbers, denoted as 'p' and 'q,' are generated employing the Rabin-Miller primality test algorithm. These primes are used to compute a modulus, 'n,' by multiplying 'p' and 'q.' 'n' plays a dual role, serving as the foundation for both the public and private keys and is often measured in bits, referred to as the 'key length.'

The public key comprises 'n' and a public exponent, 'e', which is typically set at 65537, chosen due to its being a prime number of reasonable size. 'e' does not need to be a secretly chosen prime since the public key is openly shared with all.

The private key is composed of 'n' and a private exponent, 'd.' The determination of 'd' involves the Extended Euclidean algorithm, identifying its multiplicative inverse concerning 'n's totient.

Each patient's narrative is encrypted using their unique public key, ensuring that only authorized personnel with access to the corresponding private key can decrypt and access the original content. This robust encryption mechanism serves as a critical component of the framework, assuring the confidentiality of patient narratives as they traverse from edge devices to the LLM for analysis.

3.2.4 LLM analysis

Once the encrypted and anonymized stories are obtained, they are sent to a LLM to be analysed. Several open-source Large Language Models (LLMs) were assessed during the framework's implementation, among them there was Falcon-7b-instruct [33], which is considered to date one of the best 7b open source model, outperforming comparable open-source models (e.g., MPT-7B, StableLM, RedPajama etc.) [33].

With *Falcon-7b-instruct*, a variety of prompts and temperature settings were methodically tested during the experimentation phase. At first, a series of meticulous and accurate instructions were used. Following there are some example of prompts used:

- *I am going to give you a story told by a psychological patient while looking at the image number Image number of the TAT test. "{STORY}". The patient is doing the Thematic Apperception Test. The Thematic Apperception Test (TAT) is a projective psychological assessment tool used to evaluate a person's patterns of thoughts, attitudes and emotional responses. Analyse the story told by the patient from a psychological point of view, what can you tell me about this patient? Can you identify the protagonist of the story? Can you tell me the relationship among the characters in the story? What does the character do in this story and how do they feel?*
- *I am going to give you a story told by a psychological patient while looking at the image number 3BM of the TAT test. You are going to analyse the story from a psychological point of view by responding to the questions. If you do not know how to respond about something do not say anything - BEGINNING {STORY} - END The patient is doing the Thematic Apperception Test. The Thematic Apperception Test (TAT) is a projective psychological assessment tool used to evaluate a person's patterns of thought, attitudes, and emotional responses. Analyse the story told by the patient from a psychological point of view. Who is the main character in the story told? Are there any other characters in the story and what is their relationship with the boy? What are the boy's feelings? What is the outcome of the story? What can you tell me about the patient who told the story from a psychological point of view?*
- *You are going to be provided with a story invented by a psychological patient about the image number {image number} of the Thematic Apperception Test. Analyse the story from a psychological point of view and make insight about the inner emotions of the patient. {STORY}*

These questions were created to direct the model's analysis toward a thorough evaluation of significant narrative components found in Thematic Apperception Test (TAT) stories. Unlike the testing phase, to assess the quality of responses,

we employed a set of sample stories obtained from the internet. In the different attempts that were made, it was assessed that the Falcon 7b did not work well with a highly detailed prompt, showing repetition and sometimes difficulty in identifying the protagonist of the story; Fig. 3.4 and Fig. 3.5 show some attempts of prompt engineering with Falcon7b and the responses given.

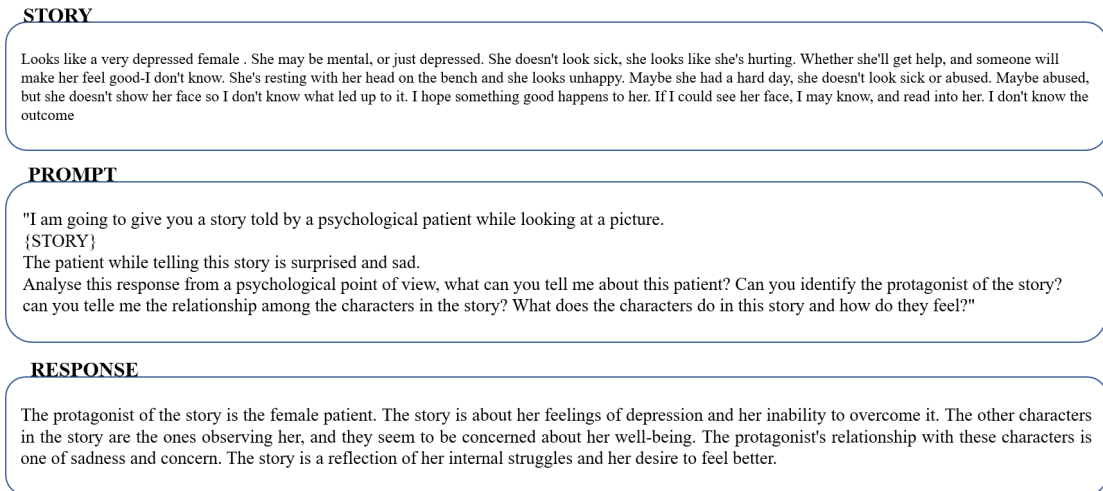


Figure 3.4

Instead did a much better job with more generic instruction as shown in Fig. 3.6.

These attempts were conducted using a sample story as a basis for analysis. The image sheds important light on the model's interpretive abilities when exposed to various stimuli and setups, giving useful insights into the variances in model replies, for instance we registered an improvement in the quality of the responses when highlighting the beginning and the end of the stories with quotes or when not giving too much information about the TAT test or the physical situation.

Even if with its promising results, it ultimately lagged behind the well-known GPT-3.5 Turbo model created by OpenAI, which is known for powering the well-liked ChatGPT, in terms of analysis and adaptability.

In comparison to the Falcon 7b model, the response generated by GPT 3.5 exhibits a notable distinction, fig. 3.7. GPT 3.5's reply stands out for its comprehensive nature, offering a more elaborate and intricate stream of information. This added complexity not only enriches the response but also provides valuable insights that can be particularly beneficial for medical professionals.

In the subsequent stages of our framework development, we enriched our prompts further by incorporating results from emotion analysis, the emotions where passed as a list of dictionaries where each dictionary represents the mean of the emotions

STORY

'Matt is bored out of his mind and would rather just stare at it and waste the time away. It's hard to see if one eye is open and the other is shut. Before? The teacher/parents just placed it on the surface in front of him and he's like WTF. Feeling? Hard to tell. I went with bored at first, but now I don't know...he could be thinking WTF do I do with this. End? He smashes it over his mother's head and stabs his mom with it. I was never into wood instruments. I prefer the drums to be it. Looking back? Just be getting out of jail and wondering why he killed his mom with a \$3000 gift. I don't know. I never really think of situations in detail. My frustrations coming out.'

PROMPT

I am going to give you a story told by a psychological patient while looking at the image number 1 of the TAT test.

You are going to analyze the story from a psychological point of view by responding to the questions.

If you do not know how to respond about something do not say anything

- % BEGINNING STORY

{STORY}

- % END STORY

The patient is doing the Thematic Apperception Test. The Thematic Apperception Test (TAT) is a projective psychological assessment tool used to evaluate a person's patterns of thought, attitudes, and emotional responses.

Analyze the story told by the patient from a psychological point of view.

Who is the main character in the story told?

Are there any other characters in the story and what is their relationship with the boy?

What are the boy's feelings?

What is the outcome of the story?

What can you tell me about the patient who told the story from a psychological point of view?

RESPONSE

The main character in the story is the boy. The boy is bored out of his mind and would rather just stare at the gift he received from his mother. He is feeling bored and frustrated. The relationship between the boy and his mother is strained, as he is frustrated with her for giving him a gift that he does not like. The outcome of the story is that the boy smashes the gift over his mother's head and stabs her with it. The boy is feeling angry and frustrated with his mother and the situation. The patient is likely experiencing feelings of anger and frustration towards his mother. The main character in the story is the boy. The boy is bored out of his mind and would rather just stare at the gift he received from his mother. He is feeling bored and frustrated. The relationship between the boy and his mother is strained, as he is frustrated with her for giving him a gift that he does not like. The outcome of the story is that the boy smashes the gift over his mother's head and stabs her with it. The boy is feeling angry and frustrated with his mother and the situation. The patient is likely experiencing feelings of anger and frustration towards his mother.

Figure 3.5

STORY

"The boy is contemplating a violin, which he probably just finished playing, forced by his mother. He doesn't enjoy playing violin, and he was hoping to play with his friends, and instead he's stuck at home, and he's forced to learn how to play this instrument."

PROMPT

You are going to be provided with a story invented by a psychological patient about the image number $\{image\}$ of the Thematic Apperception Test. Analyse the story from a psychological point of view and make insight about the inner emotions of the patient. {STORY}

RESPONSE

The patient's story indicates feelings of frustration and resentment towards the violin, likely stemming from a perceived obligation to practice and perform despite personal disinterest. The choice of the instrument as the focal point of the story may reflect the patient's resistance to the task imposed upon them. The inner emotions portrayed in the story could suggest a struggle with autonomy and a desire for freedom from the expectations placed upon them.

Figure 3.6: Example of prompt and good response of Falcon-7B

displayed in 10 seconds, as described in the section "*FER and Speech to Text*"; to each emotion is assigned a probability score, here interpreted as an emotion score. This allowed us to obtain even more intricate and contextually relevant analyses, a notable advantage for our medical applications. It is worth highlighting that GPT

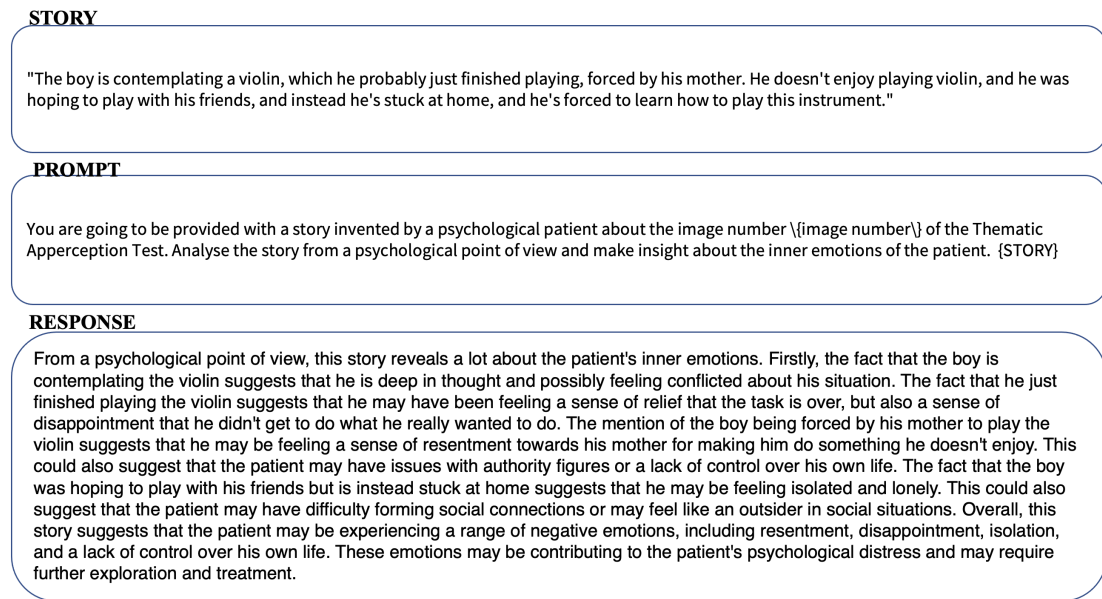


Figure 3.7: GPT 3.5 response to same prompt as Falcon 7b

3.5 Turbo continued to excel in handling these augmented prompts, demonstrating its competence in comprehending and responding to complex queries Fig. 3.8.

We also ventured into creating simpler prompts to extract summaries and perform sentiment analysis on the patient narratives. Sentiment analysis of the story was performed via the following prompt:

- *Given the provided story, please identify the most relevant phrases and analyse them providing a sentiment. Please follow the example structure:*
 - *'sentence of interest': analysis of the sentiment plus breath explanation*
 - *'sentence of interest': analysis of the sentiment plus breath explanation*

To obtain the desired output it was required to give GPT the structure of the response wanted directly by giving an example in the input prompt. The analysis conducted by GPT 3.5 and taken directly from the final report created is shown in Fig. 3.9.

In these tasks, once again, GPT 3.5 Turbo consistently outperformed the competition; an example of response by Falcon-7b is provided below:

- *The most relevant phrases are:*
 - *"forced by his mother" (negative sentiment)*
 - *"playing with his friends" (positive sentiment)*
 - *"stuck at home" (negative sentiment)*
 - *"learning how to play this instrument" (positive sentiment)*

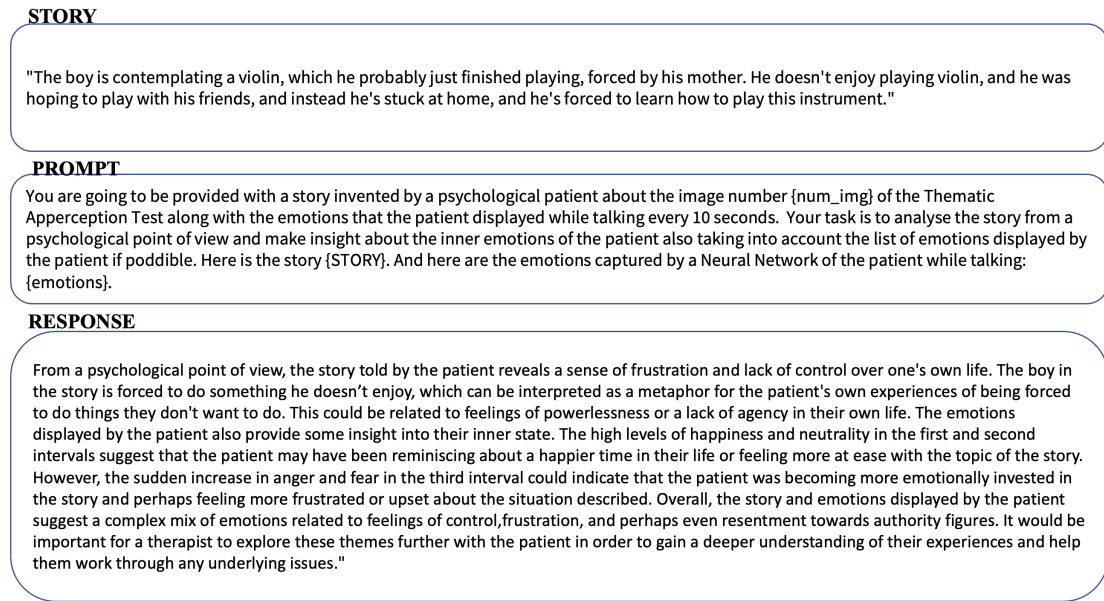


Figure 3.8: GPT 3.5 response when incorporating list of emotions displayed during the test administration. Image taken from the pdf report created with the framework developed.

The sentiment of the story is predominantly negative, as the boy is being forced to play the violin and is not enjoying it.

As the final step, we sought to identify recurring themes across the various patient narratives. To accomplish this, we processed all the patient stories with GPT 3.5 Turbo and designed a prompt to conduct a thematic analysis of the emerging patterns within the narratives:

- *You are going to be provided with [number_of_stories] different stories invented by a psychological patient during the Thematic Apperception Test. Analyse the stories and find any common theme among them.*

Figure 3.10 shows a sample taken from a report created with our framework of the analysis over all the stories to search for a common theme.

This analytical approach served as a powerful tool in uncovering hidden insights that might have eluded traditional manual analysis. Through its cognitive prowess, GPT 3.5 Turbo not only revealed the thematic links but also offered a nuanced interpretation of the significance of these recurring elements. The insights obtained in this manner carry significant implications for the field of psychology and mental health, as they can potentially aid clinicians in understanding and addressing recurrent psychological themes among their patients.

Story emotion analysis:

Sentence	emotion_analysis
"forced by his mother	negative sentiment. This phrase suggests that the boy is not playing the violin out of his own free will, but rather because he is being coerced by his mother. This can create a feeling of resentment or frustration towards the activity.
"He doesn't enjoy playing violin	negative sentiment. This phrase indicates that the boy does not find pleasure in playing the violin, which can lead to a lack of motivation and interest in the instrument.
"he was hoping to play with his friends	positive sentiment. This phrase suggests that the boy would rather be spending time with his friends, which can create a feeling of excitement and anticipation towards socializing.
"he's stuck at home	negative sentiment. This phrase implies that the boy feels trapped or confined to his home, which can create a feeling of boredom or restlessness.
"forced to learn how to play this instrument	negative sentiment. This phrase reinforces the idea that the boy is being forced to do something he does not want to do, which can create a feeling of resentment or resistance towards the activity.

Figure 3.9: GPT 3.5 emotion analysis of a story.

Common themes:

The common theme among these stories is a sense of feeling trapped or forced into a situation that the characters do not want to be in. In the first story, the boy is forced to play the violin by his mother, and in the second story, the woman is drinking to cope with an abusive husband. In the third story, the son is being pressured by his father to pursue a path that he does not want to take. All three stories depict characters who are struggling with a lack of control over their lives and are being pushed into situations that they do not want to be in.

Figure 3.10: GPT 3.5 common theme analysis among all the stories. Image taken from the final pdf report.

OpenAI GPT-3.5 turbo

Generative Pre-trained Transformer (GPT) is a class of large language models that utilizes the transformer architecture. Developed by OpenAI, GPT stands out as

one of the pioneering advancements in natural language processing. Its primary function is to generate coherent and contextually relevant text, making it highly useful for a wide range of applications, including text completion, summarization, question answering, and even creative writing.

GPT's creation involves two main stages: pre-training and fine-tuning. During the pre-training phase, the model learns from vast amounts of text data to predict the next word in a given sequence of words. This process enables GPT to grasp intricate patterns and semantic relationships within the text. The training data can include a diverse range of text sources, making GPT proficient in understanding and emulating various writing styles, tones, and topics.

The heart of GPT's success lies in its transformer architecture, which employs multi-head self-attention mechanisms to capture dependencies and relationships across words in a sequence. This self-attention mechanism enables the model to contextualize each word with respect to its surrounding words, effectively encoding rich semantic information.

GPT's uniqueness also lies in its generative ability. After pre-training, GPT can generate text by predicting the next word given a prompt or seed text. The model leverages its learned knowledge to produce coherent and contextually fitting text that aligns with the input prompt.

Fine-tuning is the subsequent step, where GPT is fine-tuned on specific tasks or domains to enhance its performance in those areas. This phase involves training the model on labeled datasets for tasks like sentiment analysis, translation, or even domain-specific tasks.

One of GPT's remarkable features is its size and scale. Models like GPT-4 boast billions of parameters, enabling them to capture intricate nuances of language. This extensive capacity contributes to GPT's success in various natural language understanding and generation tasks.

3.2.5 Report creation

After a thorough analysis using Large Language Models, the next crucial stage of this study was producing comprehensive reports. These PDF-formatted reports were created to summarize the findings from the analyses performed by GPT 3.5 Turbo, offering an in-depth overview of the psychological details and narrative insights of every patient.

This goal was achieved by implementing a custom PDF report generation process that made use of the Python ReportLab library. During this process, a unique PDFReport class was created and carefully constructed to accept the wide range of analyses produced by the LLM. This class manages the process of creating customized patient reports that include multiple sections and takeaways from the patient narratives.

The essential elements of the PDF reports—patient identification, initial diagnoses, narrative summaries, sentiment analysis findings, and themes that recur in the patient stories, will be covered in detail in the following subsections.

Patient Identification

The initial segment of each generated report features patient identification information. This includes patient names, unique identifiers, and relevant background data that is essential for medical records and subsequent analysis.

Preliminary Diagnoses

Building upon the insights obtained from the patient narratives, preliminary diagnoses are introduced in the report. These diagnoses are generated by GPT 3.5 Turbo and represent the model's interpretation of the underlying psychological factors and conditions exhibited in the narratives.

Narrative Summaries

Incorporating the narrative summaries section, the reports encapsulate the key elements of each patient's story.

Sentiment Analysis

To offer a comprehensive psychological assessment, sentiment analysis results are presented in form of a table. These results reveal the emotional undercurrents within each story, ranging from positive and negative sentiments to neutral observations.

Identification of Common Themes

The reports culminate with the identification of common themes prevalent across the patient narratives.

Upon the successful creation of these detailed PDF reports, they are securely stored on the edge device of the patient. This design allows for a highly personalized and patient-centric approach to the process. The patient, now in possession of these insightful reports, holds the autonomy to decide whether they wish to share this valuable information with their psychologist. This level of control not only respects the patient's privacy and consent but also empowers them to actively participate in their mental health journey.

Chapter 4

Testing and Results

In this chapter, we dive into the testing phase of our proposed framework. The framework was put to the test with a diverse pool of 49 participants, encompassing a wide range of ages and genders.

4.0.1 Sentiment Analysis Models

We firstly evaluated the effectiveness of the sentiment analysis models. Understanding sentiment analysis's performance was essential, as it played a crucial role in providing accurate information to GPT-3.5, as the prompt were also enhanced with sentiment information. Participants were asked to express specific emotions falling into either the negative or positive emotional spectrum.

We initially employed a video-based sentiment analysis model (FER) to assess the participants' emotional expressions, utilising the DeepFace library. It was decided to aggregate and take the average of emotions in a number of frames corresponding to 10 seconds of videos so to give the possibility to the participants to elaborate a thought and a story and to not have only humming. In this way we ended up with a list of dictionaries with key-value pair represented by emotion and its probability score for each 10 seconds of video. The highest probability score in each dictionary, represented the most dominant emotion for the 10 seconds of video under exam. This list was meant to be given alongside the narrative of the story so to enrich the analysis provided by the LLM. The results from this approach alone were not entirely satisfactory. While the emotion recognition on a single frame taken from a popular dataset such as FER-2013 [34] worked really well, reaching accuracy scores grater than 92% on emotions like anger and happiness, the scenario changes when considering a 10-second video clip from a real recording. In this case, the accuracy scores drop drastically, failing to surpass 76% for various emotions. In our opinion, this decline in performance can be attributed to the fact that, during the storytelling process, many participants remained silent for

extended periods or engaged in humming while contemplating how to articulate their thoughts. During these contemplative moments, participants exhibited facial expressions that were not directly correlated or caused by the events narrated in the story. These unrelated facial expressions could have added complexity to the model’s interpretation, contributing to the observed decrease in accuracy. In conclusion, if the emotion displayed is not ‘strong’ enough DeepFace model falls short, instead if the participant is strongly invested in his/her story then the results are still acceptable.

In the early diagnostic phase, the GPT model did well in keeping the initial responses unaffected by any incorrect emotions from the earlier approach. However, to further ensure accuracy, a decision was made to eliminate this list of emotions from the prompt. Instead, a concise paragraph was introduced, highlighting the scores for the seven emotions identified throughout the story, so to give as much information as possible to the psychologist. To obtain the scores for each emotions, we integrated results from a text emotion analysis, which may be able to better catch ‘mild’ emotions and so to improve the emotion score for the story. This improved scoring system was strengthened by utilising GPT-3.5’s capabilities. We incorporated an additional step by instructing the model to conduct a comprehensive emotion analysis on the entire narrative, thereby fortifying the reliability of the final emotion scores.

It’s important to highlight that, out of the 49 recorded videos, only 43 were deemed suitable for further analysis due to their appropriate duration.

4.0.2 Pre-Diagnosis and Analysis

From our pool of participants, 31 agreed to proceed with the analysis of their responses by the LLM. First, we carefully examined how accurate the narrative summaries produced by the models were. Out of the 31 reports created, 25 effectively captured the essential elements of the narratives.

Subsequently, we engaged in discussions with the participants regarding the pre-diagnoses generated by the LLM. This process allowed us to determine which pre-diagnoses correctly identified relevant concepts and which ones did not. Among the 31 reports, 22 were approved.

Furthermore, the LLM exhibited the capability to accurately detect instances of anxiety, emotionally impactful events, aggressive traits, and concerns related to body weight in specific participants. In particular:

- 2 cases of concerns related to body weights (man and woman)
- 6 aggressive traits
- anxiety in 1 participant after identifying a particular event in his/her life

- in 3 testers it was able to identify more than 3 psychological traits

Our results demonstrate the LLM's ability to identify marked traits in participants with reasonable accuracy. However, the model faces challenges when dealing with more general and less distinct traits.

4.0.3 Conclusion

In conclusion, it is noteworthy that LLMs exhibit a commendable proficiency in addressing these themes even without fine-tuning, relying on techniques like few-shot learning or extracting information from knowledge bases. The successful definition of our framework, validated by test results and its adoption by the company of the internship for further projects, indicates its feasibility. While the framework has inherent limitations, such as the current inability of LLMs to run on mobile devices, due to power and computational demand, it provides a robust foundation for private diagnoses on mobile devices. While the framework's performance, coupled with its potential for aiding in pre-diagnosis, holds promise for medical and psychological applications, it is crucial to recognize that, at its current state, LLMs can serve only as an adjunct for psychologists. They can streamline their workflow, attempt to highlight certain patient traits, but the ultimate diagnostic decisions and therapeutic planning must remain in the hands of the qualified psychological professional, who can comprehensively assess and interpret the nuances of the patient's condition. Nevertheless these findings pave the way for further exploration of specialized LLMs, as we will mention in the next chapter.

Chapter 5

Challenges and Future works

In this chapter, we delve into the challenges encountered during the development of our proposed framework and explore potential future directions for improvement.

5.1 Quantization

In the initial stages of this project, our intention was to develop a framework that could operate entirely on edge devices. As part of this objective, we explored various solutions to bring a Large Language Model (LLM) onto edge devices, considering one of the most popular approaches: model quantization. Quantization, in the context of machine learning, involves reducing the memory footprint and computational requirements of a model by representing numerical values with lower precision data types [35]. Such an approach was seen as a promising solution to address the challenges posed by LLMs, including their response times and significant model sizes, which often necessitate deployment on multiple GPUs.

There are two primary approaches to obtaining quantized models [36]:

1. **Post-Training Quantization (PTQ)**: PTQ involves the conversion of the weights of a pre-trained model to a lower precision without requiring retraining. While it is a straightforward and easily implementable method, PTQ may lead to a slight degradation in the model's performance due to the reduction in weight precision.
2. **Quantization-Aware Training (QAT)**: In contrast to PTQ, QAT incorporates the weight conversion process during the model training phase. This approach often results in improved model performance but demands more computational resources. An example of a widely used QAT technique is the QLoRA (Quantization with Loss-optimized Retraining Approach).

Model quantization involves the process of representing numerical values in a model with lower-precision data types, typically transitioning from floating-point numbers (32-bit or 16-bit) to fixed-point numbers (8-bit or lower). By reducing precision, quantized models require fewer bits to store and process numerical values, resulting in significant memory and computational savings. As reported in [35], this approach offers several advantages, including:

- **Reduced Memory Footprint:** Quantized models consume less memory, making them suitable for deployment on resource-constrained devices like edge devices.
- **Faster Inference:** Reduced precision and memory requirements enable faster inference, which can enhance the efficiency of processing.
- **Energy Efficiency:** Quantized models lead to improved energy efficiency, a valuable feature for applications where power consumption is a concern.
- **Deployment Flexibility:** Smaller model sizes and reduced computational demands make quantized models easier to deploy over networks, even in scenarios with limited bandwidth or storage capacity.
- **Compatibility with Hardware Acceleration:** Quantized models often work well with hardware acceleration techniques like GPUs or TPUs, further enhancing performance.
- **Cost Savings:** Smaller memory requirements and faster inference times translate into cost savings, particularly in cloud-based or server-based environments where resources directly impact costs.

The benefits of model quantization come with inherent challenges and trade-offs, making it a less desirable choice for our project.

- **Decreased Model Accuracy:** Model quantization leads to reduced precision of numerical values, which can compromise model accuracy, especially for complex tasks requiring high precision calculations.
- **Limited Dynamic Range:** Quantized models have a limited dynamic range due to fewer available values to represent numbers, which can affect their performance on certain types of data.
- **Increased Quantization Error:** The quantization process introduces quantization errors, which are discrepancies between the original floating-point values and their quantized representations. These errors can accumulate during computations and affect the final model output.

- **Training Challenges:** Training quantized LLM models is more challenging than full-precision models, requiring specialized techniques and training strategies to achieve optimal performance while mitigating the impact of quantization on accuracy.
- **Compatibility and Deployment Issues:** Some hardware and software platforms may have limitations or compatibility issues with running quantized models. Ensuring the target deployment platform supports the specific quantization scheme and data types used by the model is essential.
- **Increased Development Complexity:** Developing and optimizing quantized LLM models can be complex and time-consuming compared to full-precision models. Calibration, fine-tuning, and optimization processes are necessary to minimize the impact of quantization on model performance.

For such reasons and for simplicity sake, we opted to use the model via API and implement an encryption phase to keep sensitive data secure even when leaving the edge device.

5.2 Audio analysis integration

In our pursuit of a comprehensive psychological analysis framework, we explored the integration of not only facial emotion recognition but also a component for speech emotion recognition. This endeavor aimed to provide a more holistic view of a patient’s emotional state by considering both facial and vocal cues.

The initial approach to speech emotion recognition involved the creation of a neural network using Conv1D layers and Bidirectional LSTM layers. The training data comprised audio samples from various datasets, including Crema-D [37], RAVDESS [38], SAVEE [39], and TESS [40], encompassing a range of emotions such as surprise, happiness, sadness, anger, and neutrality. These datasets provided a diverse collection of audio samples, each with a maximum duration of 7 seconds and a mode of 2 or 3 seconds depending on the dataset.

During the model development and testing phase, we observed promising results, with model accuracies consistently exceeding 80% when evaluated on test datasets. However, the transition to real-world audio data, such as those collected during the TAT test sessions, posed new challenges.

Real-world audio data often extends beyond the 7-second window considered during training and introduces background noise, including humming sounds, which can confound the accuracy of emotion recognition models. These factors led to suboptimal performance when processing real patient audio data in the TAT test scenario.

As part of future work and research extensions, we picture the integration of a more robust and adaptable speech emotion recognition model. This model could be designed to work in conjunction with the facial emotion recognition component, allowing for a more complete analysis of a patient's emotions. Such an integration would contribute to the continued evolution of our psychological analysis framework and open new avenues for understanding and assisting patients in their mental health journeys.

5.3 Specialized LLM

One important direction for future research is to investigate customised Large Language Models (LLMs) for the domains of psychological and medical analysis. The concept revolves around fine-tuning LLMs using medical and psychological texts and data to create models equipped with the specific knowledge and capabilities necessary for in-depth analysis in these domains.

Fine-tuning, in the context of LLMs, involves adapting a pre-trained model to perform specific tasks or address particular domains [41]. This process customizes the model's behavior by exposing it to domain-specific data, thereby enhancing its capacity to generate contextually relevant and accurate responses. The benefits of specialized LLMs include the ability to interpret medical or psychological content, generate insights, and provide tailored information to healthcare practitioners, psychologists, and even patients.

Even though fine-tuning is a less demanding task than training completely from scratch a new LLM, the development and fine-tuning of specialized LLMs involves computational challenges and considerations. Fine-tuning a model of the scale and complexity of LLMs often demands substantial computational resources. The process involves running extensive training sessions on high-performance hardware, such as Graphics Processing Units (GPUs) or specialized AI accelerators like TPUs. These powerful computational resources are required to manage the vast amount of data and parameters involved in fine-tuning, ensuring optimal model performance. The acquisition and allocation of such computational power are critical prerequisites for effective fine-tuning.

In conclusion, we envision future iterations of the framework implementing a stacked strategy. This involves an initial non-fine-tuned model performing a preliminary diagnosis, guiding users towards a more specialized, fine-tuned model tailored to specific psychological conditions. This layered approach holds the potential to enhance the accuracy and effectiveness of the diagnostic process, providing a nuanced and targeted analysis for improved mental health outcomes.

Bibliography

- [1] National Institute of Mental Health. *Telemental Health*. Sito Web. URL: <https://www.nimh.nih.gov/health/publications/what-is-telemental-health> (cit. on p. 4).
- [2] American Psychological Association. «Guidelines for the Practice of Telepsychology». In: (2013). Articolo, Ultimo aggiornamento nell'agosto 2023. URL: <https://www.apa.org/practice/guidelines/telepsychology> (cit. on p. 4).
- [3] Solimini R., Busardò F.P., Gibelli F., Sirignano A., Ricci G. «Ethical and legal challenges of telemedicine in the era of the COVID-19 pandemic». In: *Medicina* (2021) (cit. on p. 4).
- [4] Bilal Javed, Abdullah Sarwer, Erik B. Soto, and Zia-ur-Rehman Mashwani. «The coronavirus (COVID-19) pandemic's impact on mental health». In: (2020). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7361582/> (cit. on p. 4).
- [5] Kevin R. Tarlow, Carly E. McCord, Jordan L. Nelson, and Paula A. Bernhard. «Comparing in-person supervision and telesupervision: A multiple baseline single-case study». In: *Journal of Psychotherapy Integration* 30.2 (2020), pp. 383–393. DOI: 10.1037/int0000210. URL: <https://doi.org/10.1037/int0000210> (cit. on p. 4).
- [6] Bradford S. Pierce, Paul B. Perrin, Carmen M. Tyler, Grace B. McKee, Jack D. Watson. «The COVID-19 Telepsychology Revolution: A National Study of Pandemic-Based Changes in U.S. Mental Health Care Delivery». In: (2021). URL: <https://psycnet.apa.org/fulltext/2020-61592-001.html> (cit. on p. 5).
- [7] Nancy McWilliams. «Psychotherapy in a Pandemic,» in: (2020). URL: <https://www.linkedin.com/pulse/psychotherapy-pandemic-written-nancy-mcwilliams-nicky-abdinor> (cit. on p. 5).

- [8] Essig, t., Russel, G. I., Mcwilliams, N. «Providing psychodynamic care during COVID-19. How to deepen the treatment with telehealth during the pandemic.» in: (2020). URL: www.apaservices.org. (cit. on p. 5).
- [9] *Guidelines for the Practice of Telepsychology*. American Psychologist, 1959, pp. 791–800 (cit. on p. 6).
- [10] Rajendra A. Kadam. «Informed consent process: A step further towards making it meaningful!» In: *Perspectives in Clinical Research* 8.3 (2017), pp. 107–112. DOI: 10.4103/picr.PICR_147_16. URL: https://doi.org/10.4103/picr.PICR_147_16 (cit. on p. 6).
- [11] Encyclopaedia Britannica. *Projective Test*. Sito Web. URL: <https://www.britannica.com/science/projective-test> (cit. on p. 7).
- [12] M. Jain, A. Singh, S.K. Kar, I.B. Weiner, and K. Kuehnle. «Projective Assessment of Children and Adolescents». In: *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier, 2017. ISBN: 978-0-12-809324-5. DOI: <https://doi.org/10.1016/B978-0-12-809324-5.05058-6>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128093245050586> (cit. on p. 8).
- [13] Praveen Shrestha. «Detailed Procedure of Thematic Apperception test». In: (Nov. 2017) (cit. on p. 8).
- [14] Edward Aronow, Kim Altman Weiss, and Marvin Reznikoff. *A PRACTICAL GUIDE TO THE THEMATIC APPERCEPTION TEST: The TAT in Clinical Practice*. Brunner-Routledge, 2001 (cit. on p. 8).
- [15] B. I Murstein. *A conceptual model of projective techniques applied to stimulus variations with thematic techniques*. Journal of Consulting Psychology, 1959, pp. 23, 3–14 (cit. on p. 8).
- [16] NVIDIA. *Large Language Models Explained*. 2023 (cit. on p. 9).
- [17] Ajitesh Kumar. *Large Language Models (LLMs): Concepts & Examples*. July 2023. URL: <https://vitalflux.com/large-language-models-concepts-examples/> (cit. on pp. 9, 14).
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. «Attention Is All You Need». In: Sept. 2017. URL: <https://arxiv.org/abs/1706.03762> (cit. on pp. 10, 14, 15, 28).
- [19] Community Members. *Why does the transformer do better than RNN and LSTM in long-range context dependencies?* Stack Exchange - Artificial Intelligence. 2021. URL: <https://ai.stackexchange.com/questions/20075/why-does-the-transformer-do-better-than-rnn-and-lstm-in-long-range-context-depen> (cit. on p. 13).

- [20] Baeldung. *RNNs vs. Transformers in NLP*. Baeldung. 2023. URL: <https://www.baeldung.com/cs/rnns-transformers-nlp> (cit. on p. 13).
- [21] Giuliano Giacaglia. «How Transformers Work: The Neural Network used by OpenAI and DeepMind». In: *Medium, Towards Data Science* (Mar. 2019). URL: <https://towardsdatascience.com/transformers-141e32e69591> (cit. on pp. 13, 14).
- [22] Hedu Ai by Batool Haider. *Visual Guide to Transformer Neural Networks - (Episode 1) Position Embeddings*. YouTube video. 2021. URL: <https://www.youtube.com/watch?v=dichIcUZf0w&t=612s> (cit. on p. 16).
- [23] Hedu Ai by Batool Haider. *Visual Guide to Transformer Neural Networks - (Episode 2) Multi-Head Self-Attention*. YouTube video. 2021. URL: <https://www.youtube.com/watch?v=mMa2PmYJlCo> (cit. on p. 18).
- [24] Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle Ungar, H Andrew Schwartz, David B Yaden, Joao Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. «Artificial Intelligence Will Change the Future of Psychotherapy: A Proposal for Responsible, Psychologist-Led Development». In: 1 (2023), p. 2. URL: https://www.researchgate.net/publication/370401072_Artificial_intelligence_will_change_the_future_of_pschotherapy_A_proposal_for_responsible_psychologist-led_development (cit. on p. 19).
- [25] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. «Towards Interpretable Mental Health Analysis with ChatGPT». In: *arXiv preprint arXiv:2304.03347* (2023) (cit. on p. 20).
- [26] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. «Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data». In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.1 (2023), pp. 1–27 (cit. on p. 20).
- [27] American Psychiatric Association. «Thousands of Mental Health Apps. . . » In: *Psychiatric News* (July 2019). URL: <https://phys.org/news/2023-07-korean-team-room-temperature-ambient-pressure-superconductor.html> (cit. on p. 21).
- [28] Wysa. *Wysa: Mental Health Application*. URL: <https://www.wysa.com/> (cit. on p. 21).
- [29] NovoPsych. *NovoPsych Official Website*. 2023. URL: <https://novopsych.com.au/> (visited on 08/25/2023) (cit. on p. 22).

- [30] Sefik Ilkin Serengil and Alper Ozpinar. «HyperExtended LightFace: A Facial Attribute Analysis Framework». In: *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE. 2021, pp. 1–4. DOI: 10.1109/ICEET53442.2021.9659697. URL: <https://doi.org/10.1109/ICEET53442.2021.9659697> (cit. on p. 27).
- [31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. «Robust Speech Recognition via Large-Scale Weak Supervision». In: *arXiv* (Dec. 2022). URL: <https://arxiv.org/abs/2212.04356> (cit. on pp. 28, 29).
- [32] GeeksforGeeks. *RSA Algorithm in Cryptography*. 2023. URL: <https://www.geeksforgeeks.org/rsa-algorithm-cryptography/> (cit. on p. 30).
- [33] Ebtesam Almazrouei et al. «Falcon-40B: an open large language model with state-of-the-art performance». In: (2023) (cit. on p. 31).
- [34] *FER2013*. Kaggle dataset (cit. on p. 39).
- [35] Explore With Yasir. *Quantized Large Language Model*. Jul 6. URL: <https://medium.com/@developer.yasir.pk/quantized-large-language-model-e80bdc81a52#:~:text=While%20quantized%20LLM%20models%20offer,to%20a%20loss%20of%20information.> (cit. on pp. 42, 43).
- [36] Miguel Carreira Neves. *What are Quantized LLMs?* Updated on Oct 1. Sep 19. URL: <https://www.tensorops.ai/post/what-are-quantized-llms> (cit. on p. 42).
- [37] *CREMA-D*. Dataset for Machine Learning (cit. on p. 44).
- [38] *RAVDESS Emotional Speech Audio*. Kaggle dataset (cit. on p. 44).
- [39] *Surrey Audiovisual Expressed Emotion (SAVEE)*. Kaggle dataset (cit. on p. 44).
- [40] *Toronto Emotional Speech Set (TESS)*. Kaggle dataset (cit. on p. 44).
- [41] Gaurav Garg. *Understanding LLM Fine Tuning — A Complete Guide for Everyone*. GoPenAI. Sept. 2023 (cit. on p. 45).