Expert Systems With Applications 279 (2025) 127191

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Decoding the mind: A RAG-LLM on ICD-11 for decision support in psychology

Marco Cremaschi [a,*], Davide Ditolve [a], Cesare Curcio [a], Anna Panzeri [b], Andrea Spoto [b], Andrea Maurino [a]

[a] University of Milano-Bicocca, viale Sarca 336, Milan, 20126, Italy
[b] Department of General Psychology, University of Padua, via Venezia 8, Padua, 35131, Italy

ARTICLE INFO

Keywords:
Large language models
Retrieval Augmented Generation
Prompt Engineering
Healthcare
Mental health
Psychology
Clinical decision support
Diagnostic reasoning
ICD-11
DSM5

ABSTRACT

This paper explores the use of Large Language Models (LLMs) in mental health to assist psychologists and psychiatrists with diagnostic decision-making according to the ICD-11 classification system. ICD-11 is the 11th revision of the International Classification of Diseases, a globally used diagnostic tool for health conditions, including mental, behavioural, and neurodevelopmental disorders. In detail, we propose LLMind Chat, an AI-powered tool with a user-friendly interface designed to support mental health professionals in their diagnostic processes. LLMind Chat leverages a Retrieval Augmented Generation (RAG) model based on the Gemma 2 (27B parameters), specifically adapted to the context of the ICD-11. This RAG model combines the strengths of Gemma 2 with a comprehensive knowledge base derived from the ICD-11, allowing it to access and process relevant information from the classification manual in real-time. LLMind's diagnostic accuracy was rigorously evaluated against the DSM-5-TR Clinical Cases manual, using automated metrics and mental health professionals' expert validation. The result suggests that LLMind Chat can serve as a reliable decision-support tool, enhancing diagnostic reasoning and potentially reducing misclassifications.

1. Introduction

Artificial Intelligence (AI) has made great strides in recent years, permeating numerous industries and radically transforming many aspects of our daily lives. However, despite this rapid spread, there are still areas where AI-based decision support tools need further studies and deeper investigation. One such area is the field of medicine, particularly in the specialities of psychiatry and psychology.

Although AI has shown enormous potential in many medical areas, such as diagnostic imaging analysis (Shi et al., 2021) or drug discovery (Mak, Wong, & Pichika, 2024), its use in mental health remains limited (Lee, 2021). This paradox is particularly evident given the large amount of clinical data available and the complexity of psychiatric diagnoses (Fried, 2020). Mental health professionals, such as psychologists, psychiatrists, and psychotherapists, face challenges in managing the vast amount of patient information available, which includes historical records and real-time assessments through clinical interviews and specific evaluation tools, such as self-report questionnaires. This is further complicated because these professionals often lack dedicated systems for organising and interpreting these data, relying instead on their memory and notes.

To create a support tool for mental health professionals, this paper explores the potential of Large Language Models (LLMs) for a specific domain (*i.e.*, psychological and psychiatric). In particular, the ICD-11 manual[1] has been used as a context as it provides a comprehensive classification of mental health disorders, which is crucial to guide clinical practice and research. A pipeline based on the Gemma 2 model[2] has been developed to create a Retrieval Augmented Generation (RAG) model. With 27 billion parameters, this (open) model was chosen for its ability to generate text, as will be demonstrated in the following sections. The proposed pipeline included the following main steps:

- **Dataset creation**: for the creation of the dataset, Chapter 6[3] of the ICD-11, which is specifically dedicated to all mental, behavioural and neurodevelopmental disorders, was retrieved via API.[4] The data were organised into a CSV file, each row corresponding to a specific disorder.

* Corresponding author.
E-mail addresses: marco.cremaschi@unimib.it (M. Cremaschi), d.ditolve@campus.unimib.it (D. Ditolve), cesare.curcio@mail.polimi.it (C. Curcio), anna.panzeri@unipd.it (A. Panzeri), andrea.spoto@unipd.it (A. Spoto), andrea.maurino@unimib.it (A. Maurino).

[1] icd.who.int
[2] ai.google.dev/gemma
[3] icd.who.int/browse/2024-01/mms/en#334423054
[4] icd.who.int/icdapi

https://doi.org/10.1016/j.eswa.2025.127191
Received 10 December 2024; Received in revised form 20 February 2025; Accepted 6 March 2025
Available online 29 March 2025

Fig. 1. Logo of LLMind Chat.

- **RAG model implementation**: the dataset was transformed into a vector representation using embeddings generated by LangChain.[5] This process allowed the dataset to be used as context for Gemma 2, enhancing its ability to suggest accurate diagnoses.

For an objective measurement of the capabilities of the model, an automated validation was performed using the DSM-5-TR Clinical Cases (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Clinical Cases) (American Psychiatric Association, 2023),[6] comparing the suggested diagnoses with those explicitly stated in the manual. In addition, a manual analysis was performed with the participation of mental health professionals to ensure the validity and reliability of the model responses in the clinical context.

The proposed pipeline led to the creation of LLMind Chat (Fig. 1), a tool with a User Interface (UI), that supports mental health professionals in analysing real clinical cases by providing decision-making assistance for accurate diagnosis and differential diagnoses. From an analysis conducted with the support of a psychologist, psychotherapist, and researcher, the model achieves an accuracy rate of 76%.

LLMind was designed and developed precisely to overcome the limitations of generalist LLMs in a clinical context. Although trained on large and diverse datasets, these models do not deeply integrate specialised clinical knowledge bases, often leading to inaccurate diagnoses or diagnoses not aligned with official diagnostic criteria. Additionally, as reported in the article, they struggle to manage the complexity of diagnostic reasoning, especially when it comes to distinguishing disorders with overlapping symptoms. Another critical limitation is the lack of transparency and interpretability, as standard models operate as opaque systems, offering little visibility into the processes that lead to their responses—a particularly relevant problem in a high-impact field like mental health. LLMind, built on a RAG architecture, dynamically retrieves and integrates information extracted directly from the ICD-11, ensuring that all suggested diagnoses are based on an authoritative and recognised source. LLMind also incorporates textual representations integrating differential diagnosis criteria, allowing for a more effective distinction between similar disorders.

The integration of solutions like LLMind Chat into everyday clinical practice for psychologists, psychiatrists, and psychotherapists offers promising opportunities to enhance mental health care. LLMs can serve as decision-support tools, aiding in diagnostic accuracy and treatment planning by analysing data, such as patient-reported symptoms, past medical histories, and therapeutic session transcripts.

Obviously, the role of LLMs in the clinical workflow would be supportive rather than substitutive of the mental health professionals who would retain their primary role in interpreting patient data, providing empathetic care, and making nuanced clinical decisions — but also benefiting from the hints from LLMs. Indeed, the LLM would act as a "second opinion", a support tool or analytical assistant, offering insights that clinicians can then evaluate and better contextualise symptoms and events in the anamnesis of a patient. It is well-known that humans, even when well-trained, are prone to biases, mistakes and attention issues, which can be partially avoided by leveraging LLMs.

[5] www.langchain.com

[6] www.appi.org/Products/DSM-Library/DSM-5-TR-Clinical-Cases

This paper is structured as follows: in Section 2, we provide an overview of the relevant literature on the application of AI in mental health and language models for decision support. Section 3 describes the process of creating the datasets using the RAG obtained by combining the ICD-11 and the ICD-11-CDDR. The creation of the LLMind model is detailed in Section 4, and Section 5 presents the evaluation of the model, including a comparison with SOTA approaches. Section 6 describes LLMind Chat, an interactive chat platform designed to deliver the functionalities of the LLMind model. Finally, Section 7 presents an ablation study, and Section 8 concludes the paper with a discussion of future work.

2. Background and motivation

AI is rapidly permeating industries, bringing with it a wave of unprecedented change and innovation. This transformative process has profoundly impacted numerous facets of contemporary life, from the way we work to the way we interact with the world around us.

For example, in manufacturing, AI automates complex processes, improving efficiency and productivity. Intelligent robots work alongside human workers (Wei, Zhang, Sun, Chen, & Li, 2023), while Machine Learning (ML) algorithms optimise supply chains and predict product demand (Tirkolaee, 2021). In transportation, AI is driving the development of autonomous vehicles, promising to revolutionise mobility and reduce traffic accidents (Ma, Wang, Yang, & Yang, 2020). These are just a few examples of how AI reshapes industries. AI is not just a technology but an innovation engine changing how we live, work, and interact with the world.

AI is also rapidly transforming healthcare in ways that were unimaginable just a few years ago. From analysing medical images (Castiglioni et al., 2021) to discovering new drugs (Gupta, 2021), AI is improving patient care, accelerating research, and opening up new frontiers in medicine. Diagnostic is an area where AI significantly impacts. ML algorithms can analyse medical images such as X-rays (Hansun, 2023), computerised tomography scans (Dankelman, 2023), and Magnetic Resonance Imaging (MRI) (Meshaka, 2023) with an accuracy comparable to, and in some cases superior to, that of human doctors (Oren, Gersh, & Bhatt, 2020). This allows for early identification of diseases such as cancer (Agarwal et al., 2023), heart disease (Baghdadi, 2023), and neurodegenerative disorders (Singh & Dash, 2023), allowing timely interventions and increasing the probability of therapeutic success (Zeb, 2024).

Despite the growing applications of AI in medicine, its use in the mental health field still has significant limitations (Joseph, 2024). While progress is being made (Ren et al., 2024), it is important to remain cautious and aware of the challenges for LLMs in the mental health field. One of the main challenges is the complexity of the human mind. Mental health conditions are often characterised by a combination of biological, psychological, social and environmental factors that mutually interact in complex ways (Fried, 2020; Joseph, 2024). AI, based on algorithms and statistical models, can have difficulty capturing and interpreting all the nuances and subjectivity of the human experience, which are crucial for accurate diagnosis and treatment in psychiatry (Yan, Ruan, & Jiang, 2022). Diagnosis in mental health presents unique complexities that distinguish it from other areas of medicine. While in many medical specialities, diagnosis is based on objective tests such as blood tests, X-rays, or biopsies, mental health diagnoses are often more nuanced and based on a careful assessment of multiple factors (Yan et al., 2022).

Symptoms of mental disorders are subjective and challenging to quantify (Yan et al., 2022). For instance, sadness, anxiety, or hallucinations are internal experiences that can vary significantly from person to person and from culture to culture. This makes diagnosis more complex than medical conditions with objective and measurable symptoms. It is also common for individuals with mental disorders to present with multiple conditions simultaneously (*i.e.*, comorbidity) (McGrath, 2020) which can make it difficult to distinguish the symptoms of one disorder from those of another, complicating the diagnosis process (Sartorious, 2013).

Given the complexities inherent in mental health diagnosis, diagnostic manuals such as DSM-5-TR (American Psychiatric Association, 2022)[7] and ICD-11 (World Health Organization, 2022)[8] provide essential support to mental health professionals (Regier, 2013; Seemüller, 2023). Firstly, they offer a common language and shared criteria for the classification of mental disorders. This enables clinicians worldwide to communicate clearly and accurately, facilitating diagnosis, treatment, and research. These manuals help reduce interpretative subjectivity by describing symptoms and the criteria required for each diagnosis. Secondly, diagnostic manuals provide a systematic framework for patient evaluation. They guide clinicians through a structured assessment, helping them gather relevant information and organise data coherently. This systematic approach helps reduce the risk of diagnostic errors and ensures a complete patient evaluation. Both the ICD-11 and the DSM-5-TR are voluminous and complex tomes containing hundreds of pages of detailed descriptions of disorders, diagnostic criteria, and codes. This can make studying and consulting the manuals challenging, particularly for less experienced clinicians. Searching for specific information can be time-consuming and demanding, with the risk of becoming overwhelmed by the numerous diagnostic categories and subcategories.

As a general caveat, it is essential to remember that a diagnosis cannot be made with certainty by simply reading a case description. The psychological and psychiatric assessment process is a lengthy and complex endeavour that relies on multiple sources of information (Wright, 2020). It involves several clinical interviews, the use of assessment tools such as questionnaires and (semi-) structured interviews, as well as the observation of behaviour, facial expressions, tone of voice, presentation, cooperativeness, and other factors (El-Hay, 2018).

In this scenario, AI, and particularly Large Language Model (LLM), emerges as an innovative solution to simplify access to and extraction of information from complex manuals. The adoption of LLMs is increasingly widespread across various specialised fields, supporting professionals in analysing large volumes of data.

A significant example is the legal field (Sun, 2023), where models such as LawGPT (Nguyen, 2023; Zhou et al., 2024) and Lawyers-Llama (Wang, Qian, Zhou, Chen, & Tan, 2023) have been trained to analyse legal documents and develop decision-support systems. This trend has spurred the emergence of commercial products like ROSS Intelligence,[9] CaseText,[10] and Lex Machina,[11] underscoring the demand for such systems.

The same diffusion can also be seen in the Finance field (Li, Wang, Ding, & Chen, 2023). An example is BloombergGPT (Wu et al., 2023), a 50 billion parameter model for analysing financial data, generating market insights, answering complex financial questions, and performing finance-specific Natural Language Processing (NLP) tasks such as sentiment analysis, price forecasting, and risk assessment.

Several more or less established solutions can also be identified in the medical field. There are two main categories: pre-trained models from scratch, such as BiomedGPT (Zhang, Zhou, et al., 2024), a vision-language model for various biomedical tasks, NYUTron (Jiang et al., 2023), focused on prediction in healthcare, GatorTronGPT (Peng et al., 2023), a generative LLM for research and assistance, and BioGPT (Luo et al., 2022), specialising in biomedical text generation and extraction. The second category includes fine-tuned models, such as Med42 (Christophe, Kanithi, Raha, Khan, & Pimentel, 2024), a suite of clinical LLMs with different optimisation strategies, Taiyi (Luo et al., 2024), a bilingual model for various medical applications, AlpaCare (Zhang, Tian, et al., 2024), specifically trained for the healthcare field, and BianQue (Chen et al., 2023), which balances the ability to ask questions and provide suggestions. Finally, Google has developed Med-PaLM 2 (Singhal et al., 2023),[12] an LLM that has demonstrated impressive capabilities in answering medical questions and achieving an "expert" level on USMLE-style questions (US Medical Licensing Examination).[13]

LLMs are also opening new avenues in the mental health field. Mental-LLM (Xu et al., 2024) is a model specifically designed for analysing online textual data to predict and understand mental health issues. Leveraging the power of LLMs, Mental-LLM can process large volumes of textual data from sources such as social media, forums, and blogs, identifying patterns and linguistic signals that may indicate the presence of mental health conditions like depression, anxiety, or post-traumatic stress disorder. A second example is Psy-LLM (Lai et al., 2023); this model can be used to develop therapeutic chatbots capable of providing essential support and human-like conversations or screening systems to identify individuals at risk for mental disorders.

3. Dataset

The first step in the study was to choose the knowledge base on which the model had to be adapted. Two possible sources have been considered:

- the *Diagnostic and Statistical Manual of Mental Disorders* by the American Psychiatric Association (APA) for the classification of mental disorders in its latest edition (DSM-5-TR). It is the standard classification of mental disorders tool used by mental health professionals in the United States and worldwide, and it is a primarily used manual on diagnosing and treating such disorders.
- the *International Classification of Diseases* (ICD), which is described in this Section.

In the present study, the ICD-11 was selected as the primary data source for two main reasons. The ICD-11 is produced by the World Health Organization (WHO) under a global public health mandate, in contrast to the DSM-5-TR, which is published by a national professional association. Additionally, the ICD-11 is recognised as the standard diagnostic manual in Italy, where the research was conducted, in line with the guidelines of the Italian National Health System.

The ICD-11 is a classification manual for statistical and medical purposes. Revised periodically, the last version of the ICD officially took effect in January 2022. Originally designed as a healthcare classification system, it provides a code system to classify diseases (with different levels based on their severity) and the wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease associated with them.

Chapter 6 of the ICD-11 focuses on mental, behavioural or neurodevelopmental disorders and is the knowledge base on which the LLM has been refined in this work. Such disorders are clinical conditions marked by significant disruptions in how a person thinks, manages emotions, or behaves that often stem from underlying psychological, biological, or developmental problems. Typically, such disorders cause considerable personal distress and interfere with an individual's ability to function effectively in everyday life, like personal relationships, social interactions, and educational and professional settings.

While choosing the ICD-11 for the training part of the model, the DSM-5-TR Clinical Cases has been selected for the testing part as it presents 104 patient cases that exemplify the mental disorders categorised by the DSM-5-TR, which is a widely spread and recognised diagnosis manual, specific for mental health, which diagnoses can be "translated" into those of ICD-11.

7 www.psychiatry.org/psychiatrists/practice/dsm
8 icd.who.int/en
9 blog.rossintelligence.com
10 casetext.com
11 lexmachina.com
12 sites.research.google/med-palm/
13 www.usmle.org

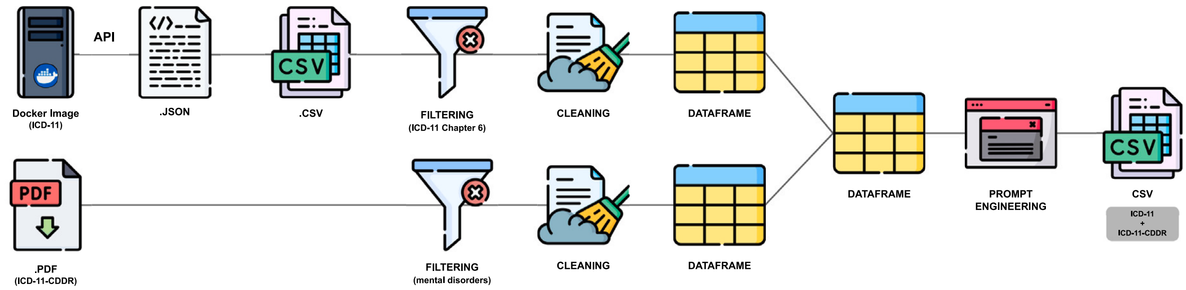

Fig. 2. Process of the training dataset creation.

The official Docker image[14] has been employed to extract the ICD-11 in the most accurate format possible, allowing local access to the REST APIs[15] without usage limits. The ICD-11 API is accessed via a Python script, enabling HTTPS to get requests to retrieve data provided in JSON format. The extracted JSON is then converted into CSV format to facilitate further processing. Since the output encompasses all available ICD-11 disorders, not just those in Chapter 6, the obtained file was processed to filter only the disorders of interest. A series of filtering and cleaning steps are applied to the CSV data to ensure that only valid and relevant entries are retained:

- ICD-11 codes not related to mental health disorders are discarded (codes that do not start with '6');
- the rows are cleaned from the malformed or corrupted characters by removing unnecessary ones, such as hexadecimal.

The cleaned data is structured into a Python dataframe for subsequent analysis.

Concurrently, the ICD-11-CDDR PDF file[16] is imported. ICD-11-CDDR is a diagnostic manual with detailed descriptions and criteria for diagnosing mental, behavioural, and neurodevelopmental disorders. It goes beyond simply listing symptoms by offering comprehensive guidance that facilitates differential diagnosis. This means that the ICD-11-CDDR helps clinicians to distinguish between different disorders that may present with similar symptoms. For example, the manual provides detailed descriptions of how different anxiety disorders may manifest, highlighting subtle yet crucial differences in their presentation. This allows clinicians to accurately identify the specific disorder a patient may be experiencing, leading to more effective treatment plans. The document is processed page by page; the text of each page was first divided into various sections (corresponding to the different pathologies) through a regular expression (RegEx). The RegEx allowed the extraction of only the mental pathologies (characterised by a well-defined code), excluding the codes related to risk factors (such as PJ codes related to mistreatment or QE codes about relational problems). The extracted text was subsequently integrated into a dataframe and joined to the dataframe containing the ICD-11 information; they are merged based on the shared "code" field.

The final dataset is output in CSV format, structured with the following fields: *code*, *title*, *definition*, *inclusions*, *exclusions*, *content*, and *prompt*. The prompt field contains a concise summary of the relevant information, formatted as a paragraph for further analysis and model processing.

The steps of a described process for dataset creation are depicted in Fig. 2.

To elucidate this process, a case study focused on *Borderline Personality Disorder* (BPD) is presented, referencing the corresponding entry within the ICD-11.[17] BPD is a mental health condition marked by a stable and pervasive pattern of unstable moods, relationships, and self-image with subsequent negative impairment for the patient in several areas. People with BPD may struggle with intense emotions, fear of abandonment, and impulsive behaviours. They may experience feelings of emptiness and have difficulty maintaining stable relationships. The boxes below provide an excerpt from this disorder entries from the ICD-11 and ICD-11-CDDR.

ICD-11 record about Borderline Personality Disorder.

The Borderline pattern specifier may be applied to individuals whose pattern of personality disturbance is characterised by a pervasive pattern of instability of interpersonal relationships, self-image, and affects, and marked impulsivity, as indicated by many of the following:
- Frantic efforts to avoid real or imagined abandonment;
- A pattern of unstable and intense interpersonal relationships;
- Identity disturbance, manifested in markedly and persistently unstable self-image or sense of self;
- A tendency to act rashly in states of high negative affect, leading to potentially self-damaging behaviours;
- Recurrent episodes of self-harm; Emotional instability due to marked reactivity of mood;
- Chronic feelings of emptiness;
- Inappropriate intense anger or difficulty controlling anger;
- Transient dissociative symptoms or psychotic-like features in situations of high affective arousal.

ICD-11-CDDR record about Borderline Personality Disorder.

Note: the borderline pattern specifier has been included to enhance the clinical utility of the classification of personality disorder. There is considerable overlap between this pattern and information contained in the trait domain specifiers (most typically negative affectivity, dissociality and disinhibition). However, use of this specifier may facilitate the identification of individuals who may respond to certain psychotherapeutic treatments. The borderline pattern specifier may be applied to individuals whose pattern of personality disturbance is characterised by a pervasive pattern of instability of interpersonal relationships, self-image and affects, and marked impulsivity, as indicated by five (or more) of the following:
- frantic efforts to avoid real or imagined abandonment;
- a pattern of unstable and intense interpersonal relationships, which may be characterised by vacillations between idealisation and devaluation, typically associated with both strong desire for and fear of closeness and intimacy;
- identity disturbance, manifested in markedly and persistently unstable self-image or sense of self;
- a tendency to act rashly in states of high negative affect, leading to potentially self-damaging behaviours (e.g. risky sexual behaviour, reckless driving, excessive alcohol or substance use, binge eating);
- recurrent episodes of self-harm (e.g. suicide attempts or gestures, self-mutilation);
- emotional instability due to marked reactivity of mood — fluctuations of mood that may be triggered either internally (e.g. by one's own thoughts) or by external events, as a consequence of which, the individual experiences intense dysphoric mood states, which typically last for a few hours but may last for up to several days;
- chronic feelings of emptiness;
[..]

The final dataset is in CSV format and contains the following fields for each disorder:

- **Index**: an autoincrement ID starting from 0;
- **Code**: the unique code assigned by the ICD-11 manual;
- **Title**: disorder name;
- **Definition**: the description is a short characterisation of the entity that states things that are always true about a disorder;
- **Inclusions**: inclusion terms are a list of conditions, symptoms, or clinical presentations that fall under the definition of the specific disorder. These terms help clinicians identify cases that should be classified under that particular diagnosis. They include synonyms of the disorder, similar clinical presentations and specific subtypes of the disorder;
- **Exclusions**: exclusion terms are a list of conditions or symptoms that, although they may seem similar or related, should not be

[14] icd.who.int/docs/icd-api/ICDAPI-DockerContainer
[15] icd.who.int/icdapi
[16] iris.who.int/handle/10665/375767
[17] icd.who.int/browse/2024-01/mms/en#2006821354

Table 1
Table with the characteristics of the dataset.

Characteristics	Value
Number of records	719
Number of disorders w. at least one inclusion	55
Number of disorders w. at least one exclusion	144
Average prompt length (characters)	1013

classified under that particular disorder. These terms help to differentiate the disorder from other similar conditions, avoid misdiagnosis or diagnostic overlap, and guide the clinician to the most appropriate classification;
- **Content**: clinical descriptions and diagnostic requirements from ICD-11-CDDR;
- **Prompt**: this field will be discussed in the next Section 3.1.

The two records reported above show how the BDP is reported into the ICD-11 and ICD-11-CDDR, while the corresponding BDP representation in the dataset is represented in the Listing 1.

Listing 1: Dataset instance example related to BPD.

```
"index": 510,
"code": "6D11.5",
"title": "Borderline pattern",
"definition": "The Borderline pattern specifier may be applied
    to individuals ...",
"inclusions": "",
"exclusions": ""
"content": "Note: the borderline pattern specifier has been
    included to enhance the ..."
"prompt": "..."
```

Table 1 reports some statistics on the dataset. The dataset is available in the GitHub repository.[18]

3.1. Prompt engineering

Before the final step in our pipeline, a prompt for each disorder has been created (Fig. 2). This entailed merging relevant information from both the ICD-11 and ICD-11-CDDR manuals and supplementing each field with a human-readable description. This consolidation is crucial because the subsequent conversion into an embedding requires all disorder-related information to be contained within a single textual field (“Document”).

For instance, Listing 2 illustrates a prompt generated from the dataset instance in Listing 1, where the highlighted text (in purple) represents the supplementary information required to adapt the ICD-11 and ICD-11-CDDR data for the RAG model.

Listing 2: Prompt example.

```
"prompt":
Disorder Name: Borderline pattern
Disorder Code: 6D11.5
Disorder symptoms:
The Borderline pattern specifier may be applied to individuals
    whose pattern of personality disturbance is characterised
    by a pervasive pattern of instability of interpersonal
    relationships, self-image, and affects, and marked
    impulsivity, as indicated by many of the following:
    Frantic efforts to avoid real or imagined abandonment~ A
    pattern of unstable and intense interpersonal ...
For this disorder the diagnostical requirements are:
Borderline pattern The borderline pattern specifier has been
    included to enhance the clinical utility of the
    classification of personality disorder. Specifically, use
    of this specifier may facilitate the identification of
    individuals who may respond to certain psychotherapeutic
    treatments. A complete description of a particular case of
    personality disorder includes the rating of the severity
    level and the assignment of the applicable trait domain
    specifiers (e.g. mild personality disorder with negative
    affectivity and anankastia; severe personality disorder
    with dissociality and disinhibition.) The borderline
    pattern specifier is considered optional but, if used,
    should ideally be used in combination with the trait domain
     specifiers (e.g. moderate personality disorder with
    negative affectivity, dissociality and disinhibition,
    borderline pattern). ...
```

[18] github.com/unimib-whattadata/llmind/blob/main/data/input/ICD-11_joined.csv

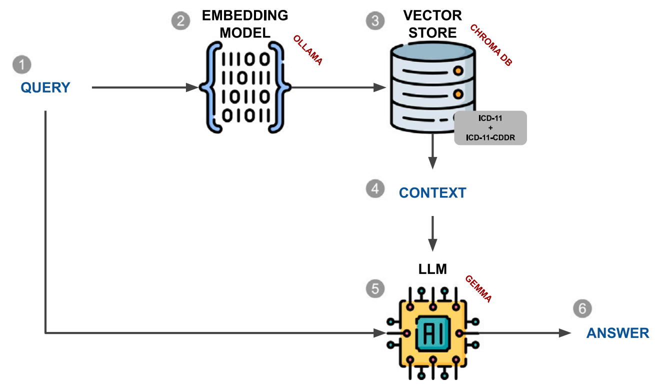

Fig. 3. RAG architecture.

4. RAG model

LLMs are typically pre-trained on a diverse and large corpus of text data using a pre-training objective such as language modelling. During the pre-training phase, the model learns to predict the next word in a sentence or to fill in missing words. While this process allows the model to capture general language patterns, it does not explicitly train the model for task-specific operations. Consequently, LLMs are designed to generalise well across a broad range of language tasks, learning to understand language’s underlying structure and semantics. This generalisation makes them versatile but may lead to less optimal performance on highly specific tasks.

Two techniques could be adopted to adapt the LLM to specific tasks: *fine-tuning* and *RAG*. The *fine-tuning* involves further training a pre-trained LLM on a new dataset specific to a task. This process updates the model’s weights, adapting it to the new task. Fine-tuning can be computationally expensive and time-consuming, but it can significantly improve the model’s performance for the specific task. The *RAG* does not modify the weights of the LLM model; the parameters learned during its initial training remain unchanged. Instead, it retrieves relevant information from an external knowledge source (such as a database or a set of documents) and provides this information as context to the LLM. RAG is generally more resource-efficient than fine-tuning. The external knowledge sources can be easily updated, allowing the RAG system to quickly adapt to new information. For this reason, the RAG method has been chosen for the LLMind.

The RAG architecture is represented in Fig. 3:

1. **Query formulation**: the process begins with a clinical description of the patient obtained from the clinician or their clinical records. This description includes relevant patient-specific information;
2. **Embedding**: this clinical description is then transformed into a numerical vector representation (an embedding) using the same embedding model used to generate vector representations of various disorders. This ensures the query and the disorder information exist in the same conceptual space[19];
3. **Similarity search**: the query vector is compared to a database of disorder vectors. A similarity measure is applied to identify the top 4 most similar disorders to the query (this number has been determined empirically);

[19] python.langchain.com/docs/integrations/text_embedding/ollama/

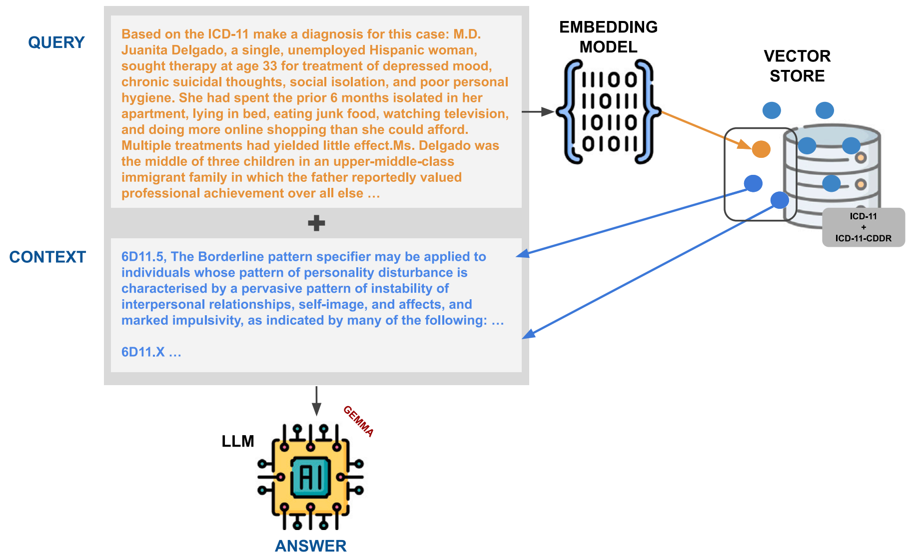

Fig. 4. RAG pipeline.

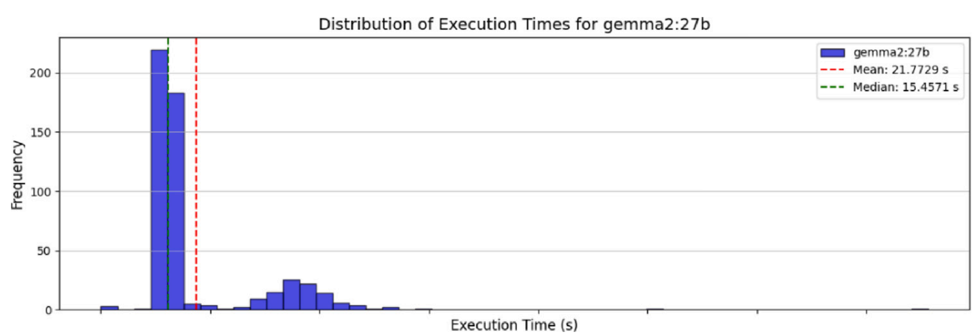

Fig. 5. Distribution of execution times for the Gemma 2 model.

4. **Contextualisation**: the top 4 most similar disorders provide crucial context for the LLM to generate a relevant response;
5. **Answer generation**: the LLM leverages the context from the identified disorders to generate an informed answer to the initial clinical query.

For the implementation, the following solutions have been employed:

- Ollama[20] for creating, customising and running local LLM models;
- LangChain to create embeddings from the previously described prompts;
- ChromaDB[21] to store them.

In particular, Ollama served as a local LLM engine, enabling our study to avoid cloud server dependencies and keep data private and controlled while utilising multiple models simultaneously. LangChain libraries, instead, are used for embeddings, vector storage, and language models. LangChain was chosen because it is an open-source library designed to simplify the creation of applications that leverage LLMs. The library integrates with various LLMs, allowing developers to build custom Natural Language Processing (NLP) pipelines and incorporate advanced features like conversation management, personalisation, and external data usage. It is advantageous when precise control is needed over how models are integrated into specific applications. To avoid repeating this resource-heavy process for each prompt, the adopted solution was to save results for quick retrieval. ChromaDB was employed to store, search, and manage embeddings efficiently to achieve this result.

The RAG pipeline is depicted in Fig. 4.

The pipeline starts with reading the joined output CSV file, where each disorder is represented as a row, providing a structured format for the data. Embeddings are generated for each disorder to enhance the system's capability to understand and retrieve relevant information. As described, embeddings are stored in a ChromaDB. For processing, the Gemma 2 is loaded into Ollama. The system's core is the question-answering (Q&A) chain, which combines the vector store retriever, the RAG prompt, and the model running inside Ollama to process queries effectively.

[20] ollama.com
[21] www.trychroma.com

The presentation of a diagnostic query, or user prompt (*e.g.*, "Based on the ICD-11, provide a diagnosis for this case ..."), initiates the Q&A chain.

Example of the user prompt based on Case 18.5 of DSM-5-TR Clinical Cases.

Based on the ICD-11 make a diagnosis for this case: Juanita Delgado, a single, unemployed Hispanic woman, **sought therapy at age 33** for treatment of **depressed mood, chronic suicidal thoughts, social isolation, and poor personal hygiene**. She had spent the prior 6 months isolated in her apartment, lying in bed, eating junk food, watching television, and doing more online shopping than she could afford. **Multiple treatments had yielded little effect**.
Ms. Delgado was the middle of three children in an **upper-middle-class immigrant family** in which the father reportedly valued professional achievement over all else. She **felt isolated** throughout her school years and **experienced recurrent periods of depressed mood**. Within her family, she was known for **angry outbursts**. She had done well academically in high school but dropped **out of college** because of frustrations with a roommate and a professor. She had attempted a series of internships and entry-level jobs with the expectation that she would return to college, but she **kept quitting** because "bosses are idiots. They come across as great and they all turn out to be twisted". These "traumas" always left her feeling terrible about herself ("I can't even succeed as a clerk?") and angry at her bosses ("I could run the place and probably will"). She had dated men when she was younger but **never let them get close physically** because she became **too anxious** when any intimacy began to develop. Most of her friends from the neighbourhood had "become frauds or losers", while she had stayed "Facebook friends" with some friends from college, and they were all "doing amazing things all over the world. I'll definitely meet up with them if they ever come back to town".
Ms. Delgado said that she was generally "**down and depressed**" but that she had dozens of 1- to 2-day "**manias**" in which she was energised and edgy and pulled all-nighters. She tended to "crash" the next day and **sleep for 12 h**.
Ms. Delgado reported that she occasionally **cut herself**, "but only superficially". She had persistent thoughts that she would be **better off dead**, and she had been psychiatrically hospitalised three times after **overdoses**. She indicated that psychotherapy had never been especially helpful, nor had her medication treatments, which had included mood stabilisers, antidepressants, and low-dose antipsychotic medication.
Mrs. Delgado was a casually groomed and somewhat unkempt woman who was cooperative, coherent, and goal directed during the interview. She was generally **dysphoric**, with a constricted affect, but did smile appropriately several times. She described shame at her poor performance but also believed she was "on Earth to do something great". She described her father as a spectacular success, but he was also a "Machiavellian loser who was always trying to manipulate people". She described quitting jobs because people were disrespectful. Towards the end of the initial session, **she became angry** with the interviewer after he glanced at the clock ("Are you bored already?").

The process entails the transformation of the user prompt into an embedding, *i.e.*, a numerical vector representation capturing semantic information. Subsequently, this embedding is compared against a repository of disorder embeddings within a vector store. Leveraging cosine similarity, a metric quantifying the angular proximity between two vectors in an inner product space, the system identifies disorders exhibiting the highest degree of semantic similarity to the user prompt.

The most 4 similar disorders are then appended to the original user prompt, enriching the contextual information provided to the underlying LLM. This augmented context facilitates the LLM's generation of a more accurate and informed diagnostic response.

The box below presents an example of the model's output related to BPD.

Example of the model output about BPD.

Based on the provided information, Juanita Delgado's symptoms suggest a diagnosis of Borderline Personality Disorder (6D10) according to ICD-11. Her history of unstable relationships, intense anger, impulsivity (self-harm, substance use), chronic feelings of emptiness, and difficulty regulating emotions align with the diagnostic criteria for this disorder. The presence of brief manic episodes further supports this diagnosis.

This approach streamlines the process of generating accurate diagnoses from a specific medical knowledge base, leveraging both retrieval and generation techniques for enhanced performance.

4.1. Model extensibility

To ensure maximum flexibility and continuous updates, the approach relies on a modular library of prompt units stored in the vector. Each disorder in the ICD-11 is represented by a single text chunk (prompt) that includes descriptive content, diagnostic criteria, and other relevant details. This structure allows the repository to be augmented with additional clinical resources and guidelines over time without retraining or altering the base model's weights. For instance, segments from other diagnostic manuals (*e.g.*, DSM-5-TR) can be integrated, split into distinct prompt units, and indexed to facilitate retrieval. Whenever the ICD-11 is revised, or new evidence emerges, the system only requires embedding the updated materials and adding them to the Chroma DB, thereby maintaining an evolving, version-controlled content base.

The resulting framework supports specialised prompts for specific contexts, such as child or geriatric psychiatry. It leverages a configurable top-k retrieval mechanism to manage how much context the model supplies. Compared to full-scale fine-tuning, this substantially reduces update costs while preserving adaptability to novel literature and changing diagnostic practices. Consequently, LLMind can be transformed into a dynamic knowledge hub capable of delivering diagnoses and referencing up-to-date treatment protocols and support resources, integrating seamlessly into the workflows of psychologists, psychiatrists, and psychotherapists.

4.2. Execution times

The LangChain embeddings creation process was executed on an MSI Vector GP78 HX 13 V laptop equipped with an i9-13980HX CPU, 2x16Gb DDR5 RAM, an NVMeSSD, and an RTX4090 Laptop graphic card, running on Windows 11. Due to the frequent access to the ChromaDB in the experimental setup, disk speeds were critical in overall performance and affected computational time. This system boasts a sequential read speed of 6828.08 MB/s and write speeds of 5205.79 MB/s.

This configuration led to an average speed for creating embeddings between 2.1 and 2.2s/it, as every model had the same 719 chunks that needed processing. The average total time was between 22 and 26 min.

The distribution of inference times for the Gemma 2 model is shown in Fig. 5. The mean execution time is 21.7729 s, depicted by the red dashed line. The median execution time is 15.4571 s, shown by the green dashed line.

5. Evaluation

The evaluation is based on the answers and diagnoses suggested by LLMind for the cases of the DSM-5-TR Clinical Cases. DSM-5-TR Clinical Cases has been chosen to completely avoid any evaluation bias by using the same cases the models have been fed with (*i.e.*, ICD-11). During the evaluation, it is essential to carefully review the generated responses, as they do not always include the disorder code or the full disorder name as specified in ICD-11.

In the presented case study (Case 18.5), the model accurately identifies the disorder and its corresponding code (*i.e.*, Borderline Personality Disorder (6D10)). However, in other clinical cases, the model may generate a differential diagnosis, suggesting multiple possible disorders. In such instances, the most relevant diagnosis was determined through manual analysis, selecting the most definitive option. For example, in the response provided in Listing 3, the following disorders can be identified: (i) 6A30 Substance-Induced Depressive Disorder and (ii) 6C45 Disorders Due to Use of Cocaine. Notably, the latter diagnosis is presented without an explicit disorder code.

In another example reported in the Listing 4, it is possible to identify the following disorders: (i) 6D3Z Sexual Disorder, (ii) 6C45 Disorders due to use of cocaine, and (iii) 6B20 Obsessive-Compulsive Disorder (workaholism). For these last two disorders, the model does not extract any disorder identifier, so manual intervention is needed to recognise and associate the correct disorder identifier.

Listing 3: Example of the output of the model.

```
Based on the ICD-11 criteria, Frank Young's diagnosis is likely
6A30 Substance-induced depressive disorder, specifically
       cocaine induced. His depressive symptoms
(depressed mood, loss of interest, insomnia, low self-esteem)
       emerged approximately 3 months after initiating
regular cocaine use and persist regardless of recent cocaine
       intake.
```

Listing 4: Example of the output of the model.

```
Based on the ICD-11, a possible diagnosis for Mr. Pickering
       could be Other Specified
Sexual Disorder. His behaviours, including frequent anonymous
       sexual encounters, interest in exhibitionism and
voyeurism, use of condoms as a fetish object, and preference for
        BDSM roles suggest a pattern of atypical sexual
interests and behaviours. While he presents with other concerns
        like cocaine use and workaholism, these appear
secondary to the core issue of his sexual expression.
```

Two approaches were initially considered for evaluating the model's performance:

1. **Automated assessment**: in the automated evaluation, the model's response is manually analysed to extract the most likely diagnosis from the generated output as described above. Subsequently, it is automatically checked (*i.e.*, string matching) whether the response matches the diagnosis present in the dataset (DSM-5-TR Clinical Cases). In this case, the score given is either 0 or 1;
2. **Manual review by qualified mental health** professionals: this manual evaluation by health professionals involved an evaluation by 4 professionals who manually assessed the model's answers and assigned a score ranging from 0 (completely incorrect diagnosis) to 1 (exactly correct diagnosis).

5.1. Automatic evaluation

Automated evaluation proved unsuitable due to the nature of the model's output. As previously noted, the model often generates responses that encompass multiple disorders, making it impossible to assign a single, definitive diagnostic code in some cases where multiple comorbid diagnoses are simultaneously present. Additionally, the model's output is purposefully designed to emulate the clinical case presentations outlined in the DSM-5-TR Clinical Cases, aligning with professional standards and fulfilling a core design requirement.

Despite these limitations, an automated evaluation was conducted, focusing solely on the identification of a primary diagnosis — meaning the first diagnosis presented with the highest degree of certainty (*e.g.*, 'the diagnosis is [...]') in the output of the LLM. This assessment, however, excluded consideration of any secondary diagnoses provided by the model with a lower degree of certainty (*e.g.*, 'the diagnosis could be [...]') and presented after the first one mentioned. It is important to note that this refers only to the order of appearance of diagnoses in the output of the LLM and is not related to the concepts of primary and secondary diagnoses as understood in psychiatry, where a 'primary' condition is the most prominent and may lead to other 'secondary' less prominent diagnoses (Maj, 2021).

The metrics used for the automated evaluation are:

- **Accuracy**: a metric that evaluates the overall performance of the model by calculating the proportion of correctly classified cases out of the total number of diagnoses performed;

$$Accuracy = \frac{\#correct\_cases}{\#total\_cases} \tag{1}$$

- **Precision**: a measure of the accuracy of a model's diagnosis, computed as the ratio of correct annotations to the total number of annotations generated by the models;

$$Precision = \frac{\#correct\_annotations}{\#system\_annotations} \tag{2}$$

- **Recall**: a metric that considers the coverage of clinical diagnosis computed. It is calculated as the ratio of true positive annotations to the total number of target annotations in the Dataset;

$$Recall = \frac{\#correct\_annotations}{\#target\_annotations} \tag{3}$$

- **F1-Score**: the F1-score is the harmonic mean of precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

In Table 2, the metrics for each model are provided, with the best performance achieved by the Gemma 2 model. The results show that larger models (those with more parameters) achieve the highest performance. The smallest model, LLama 3, delivers the lowest performance. Overall, Gemma proves to be the most suitable model for this task. Specifically, using Gemma 27b—which has three times the parameters of Gemma 9b—yields the best results.

Fig. 6 depicts the word cloud related to the responses provided by LLMind. The word cloud provides an immediate visual representation of the vocabulary (excluding the less frequent words and stop-words) and main themes emerging from LLMind responses.

5.2. Manual evaluation

The manual evaluation was conducted with four experts in the field, each with different specialisations and experience levels. These experts were psychologists and psychotherapists recruited from the professional network of the authors and were asked to complete the assessment of 104 clinical cases within one week. The experts had a diverse range of specialisations, including expertise in specific age groups (adults, young adults, adolescents, teenagers), various disorders (anxiety, depression, personality disorders, autism spectrum, neurodivergence, psychotic disorders, mood disorders, obsessive-compulsive disorder), and specific psychotherapeutic approaches. This diverse range of expertise ensured a comprehensive evaluation of LLMind's performance across various clinical scenarios.

The metric used for the manual evaluation is:

- **Average**: a metric that calculates the overall average by summing all manually assigned scores during the evaluation and dividing the total by the number of values.

$$Average = \frac{\#score\_sum}{\#total\_cases} \tag{5}$$

The LLMind responses were submitted to the experts, who provided a manual score between 0 (= incorrect) and 1 (= correct). This task was performed by comparing the diagnosis reported by the model with the diagnosis reported by DSM-5-TR Clinical Cases.

In the case of multiple diagnoses, a score of 1 was given only if all co-diagnoses were correctly identified. For diagnoses that only missed a specifier, a score close to 1 was assigned. If the diagnosis was incorrect but fell within a similar pathological category, the score given was not 0 but approximately 0.5. Adjustments were made based on the similarity of the proposed answer to the complete correct diagnosis as reported by the DSM-5-TR Clinical Cases.

The manual evaluation achieved an average score of **0.761**. The Concordance Correlation Coefficient (CCC) (Barnhart, Haber, & Song, 2002) was computed for each pair of evaluations to further assess the consistency among raters. Unlike a simple correlation coefficient, the CCC captures both the strength of correlation and systematic bias

Table 2
Evaluation metrics and No Diagnosis predictions for the tested models.

Model	Accuracy	Precision	Recall	F1-score	No Diagnosis Counts	Response length (avg)
Gemma2 9B	0.272	0.309	0.272	0.276	2 (1.94%)	372
LLMind (w. Gemma2 27B)	**0.524**	**0.552**	**0.524**	**0.524**	0 (0.00%)	396
Llama3 8B	0.165	0.215	0.165	0.175	4 (3.88%)	298
Mistral Nemo 12B	0.204	0.244	0.204	0.192	0 (0.00%)	303
Phi 3 medium 14B	0.175	0.231	0.175	0.190	35 (33.98%)	557

Fig. 6. Word cloud related to the LLMind's responses.

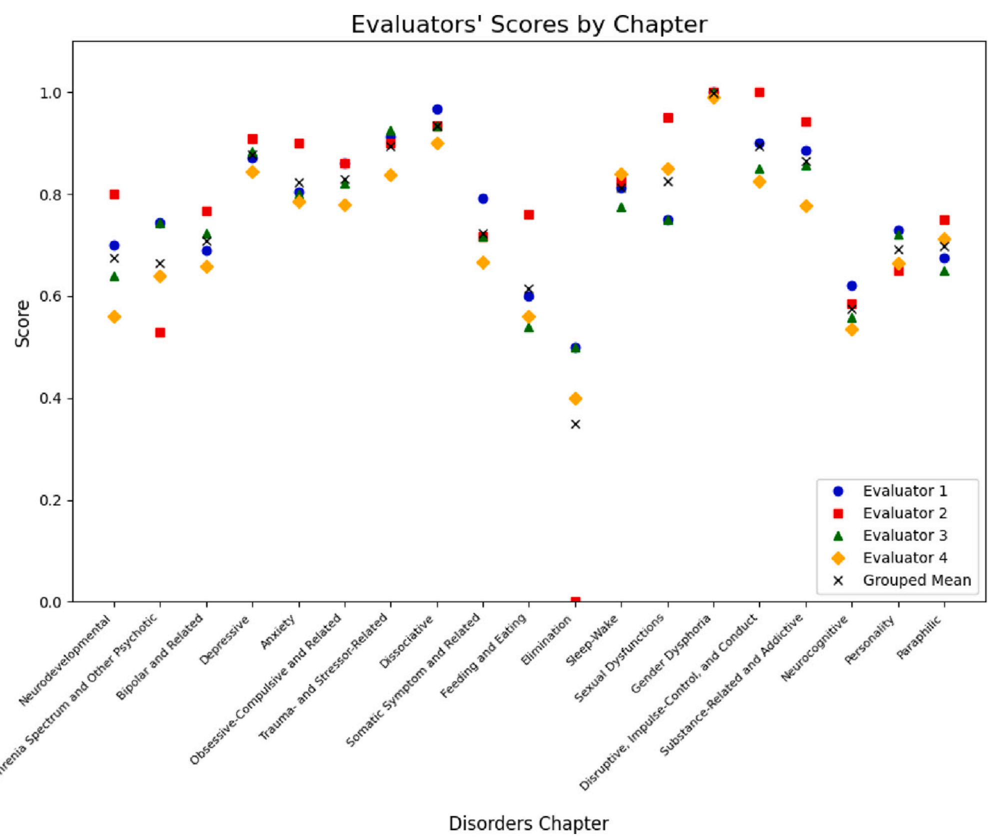

Fig. 7. Distribution of execution times for the Gemma 2 model.

among ratings. The mean CCC value of 0.70 suggests a moderate level of inter-rater agreement, thereby supporting the reliability of the model's diagnostic recommendations.

Fig. 7 displays the mean score given by each evaluator, as well as the mean score across all the evaluators, for each disorder chapter present in the DSM-5-TR Clinical Cases. Such subdivision allows to visually explore consistency of mean scores within the same disorder category, eventually giving a direction for future inspection of different performances of the model and different evaluators' scores depending on the disorder type. Overall, no substantial scores differences can be noted, except those for "Elimination Disorders", for which, however, only one clinical case was available and it did not perform adequately.

5.3. Comparison with SOTA approaches

The results are compared with other SOTA approaches that use foundational mental health models to further validate the study. The selected approaches are: (i) MentaLLaMA (Yang, 2023), the first open-source LLM for interpretable mental health analysis with instruction-following capabilities, (ii) MedLlama2, a Llama 2-based model trained with medical Q&A, and (iii) BioMistral, that uses Mistral as its foundation model and it is further pre-trained on PubMed Central (Labrak, 2024). The same prompts that composed the results already discussed are asked of various other models. Below are responses from the three models considered about Case 18.5 of the DSM-5-TR Clinical Cases.

Example of the output provided by MedLLama2 about the Case 18.5.

> Based on the given information and context, I would assign the following label to the case: Bipolar Disorder.
> Reasoning: The case mentions symptoms that align with those commonly associated with bipolar disorder, such as chronic suicidal thoughts, social isolation, poor personal hygiene, and episodes of depressed mood. There is also mention of previous hospitalisations due to overdoses and ongoing treatment attempts that have been unsuccessful. These factors suggest a pattern of manic and depressive episodes, which are characteristic of bipolar disorder.

Example of the output provided by Biomistral about the Case 18.5.

> The patient has been diagnosed with bipolar type II disorder in partial remission, which means that there is insufficient information to determine the nature of her most recent mood episode. It also means that she did meet the diagnostic criteria for bipolar type II disorder previously.

Example of the output provided by MentaLLama about the Case 18.5.

> Ms. Delgado's history suggests that she is experiencing a moderate episode of Recurrent depressive disorder, as defined by the DSM-5 criteria, with psychotic symptoms, which is characterised by delusions and/or hallucinations. This diagnosis is in addition to her current major depressive disorder. In the future, it may be helpful for her to consult a mental health professional regarding the possibility of preventing these episodes from occurring again. Given her history, she may also benefit from ongoing maintenance treatment with an antidepressant medication and potentially psychotherapy to learn skills to manage symptoms when they arise. However, this decision should be made in collaboration with a qualified mental health professional.

Below is LLMind's response:

Example of the output provided by LLMind about the Case 18.5.

> Based on the provided information, Juanita Delgado's symptoms suggest a diagnosis of Borderline Personality Disorder (6D10) according to ICD-11. Her history of unstable relationships, intense anger, impulsivity (self-harm, substance use), chronic feelings of emptiness, and difficulty regulating emotions align with the diagnostic criteria for this disorder. The presence of brief manic episodes further supports this diagnosis.

LLMind's response is correct, diagnosing a borderline personality disorder (ICD-11 code: 6D10). The model's response also includes additional details to support the diagnosis, such as the presence of manic episodes. In contrast, MedLLama2 mistakenly diagnoses bipolar disorder, while Biomistral diagnoses bipolar II disorder in partial remission. MentaLLama diagnoses a moderate episode of recurrent depressive disorder with psychotic symptoms, in addition to major depressive disorder.

An analysis of the responses[22] of the various models highlights the following weaknesses related to the other models:

MedLlama2 exhibits inconsistent diagnostic accuracy, providing correct diagnoses in some cases but failing to recognise symptoms or suggesting incorrect diagnoses in others. The model struggles to integrate information from different parts of the clinical case, leading to incomplete or inaccurate diagnoses. It sometimes tends to over-diagnose or under-diagnose certain conditions, potentially due to biases in the training data or limitations in understanding clinical criteria. MedLlama2 appears to struggle with cases involving multiple co-morbidities or complex clinical presentations and may generate differential diagnoses that are irrelevant or plausible.

BioMistral demonstrates consistently poor performance, providing inaccurate or irrelevant responses and often failing to provide any meaningful diagnosis or demonstrating a lack of understanding of the clinical information. The model appears to have limited knowledge of medical terminology and concepts, leading to misinterpretations of clinical information and inaccurate diagnoses. BioMistral's responses are often inconsistent or incomplete, lacking adequate explanations or justifications for its diagnoses or offering irrelevant information. It tends to provide incomplete diagnoses, failing to capture the full spectrum of the patient's condition. It may exhibit biases towards specific diagnoses, potentially influenced by the training data or the model's architecture. Additionally, BioMistral demonstrates limitations in understanding the temporal evolution of diseases and their associated symptoms.

MentaLLaMA shows several critical issues in medical diagnosis. It often provides inaccurate, overly generic, or vague diagnoses, lacking specificity and precision. The model may suggest implausible conditions, fail to recognise crucial details or clear symptoms, and tend to focus on common conditions, neglecting rarer or more complex diagnoses. Additionally, MentaLLaMA seems to have a limited understanding of clinical nuances, leading to misinterpretations of symptoms and inaccurate diagnoses. Its responses are often too short or incomplete, without adequate explanations or justifications. The model may also misinterpret clinical findings or fail to recognise crucial details, leading to diagnostic errors.

It should be noted that MedLLama2 is a linear, general-purpose QA system not explicitly designed for psychological issues. BioMistral has not been specifically developed for the domain of psychology/psychiatry either. It uses medical scientific publications as its primary source of information without a particular focus on mental health aspects. MentaLLama, unlike the other two, has been specifically developed for the mental health field. However, unlike LLMind, it does not use diagnostic manuals but data taken from websites (*e.g.*, Reddit, X).

The models' evaluations are summarised in Table 3. "No Diagnosis" refers to malformed predictions or inconsistent answers.

In summary, LLMind demonstrates greater accuracy, completeness, and ability to interpret symptoms compared to the other models. This highlights the importance of using LLMs specifically trained for the mental health domain through the use of universally recognised diagnostic manuals.

5.4. Discussion on evaluation results

The models' performances appear to be influenced by the complexity of the clinical cases they analyse. The models perform better with emblematic and straightforward cases, with clear symptomatology and fewer comorbidities. These cases often present distinct diagnostic features, making the model's diagnosis easier. Conversely, the models

[22] unimib-whattadata.github.io/llmind-docs/

Table 3
Evaluation metrics and No Diagnosis predictions for different SOTA models.

Model	Accuracy	Precision	Recall	F1-score	No Diagnosis Counts
MentaLLaMA	0.068	0.127	0.067	0.067	25 (24.27%)
MedLLama2	0.049	0.098	0.048	0.054	18 (17.48%)
BioMistral	0.058	0.069	0.058	0.061	40 (38.83%)
LLMind (w. Gemma 2 27B)	**0.524**	**0.552**	**0.524**	**0.524**	0 (0.00%)

struggle when cases become more complex, involving nuanced presentations with multiple overlapping comorbidities. Several co-diagnoses complicate the diagnostic process, as the model needs to weigh the interactions and relative contributions of various conditions. Future research could implement complexity metrics through a mechanism to assess the complexity of a case, such as the number and similarity of comorbid conditions, which could help models adapt their decision-making strategies based on the task's difficulty.

The models face even more significant challenges when co-diagnoses are closely related or share overlapping symptom profiles. For instance, differentiating between schizophrenia and bipolar disorder (with psychotic features) is considerably more complex than distinguishing between two very different conditions, such as religious delusions and arachnophobia. The subtle distinctions in symptom presentation, progression, and context require a high degree of interpretive capability. Future research may refine differential diagnosis algorithms that better recognise and distinguish between closely related conditions, thus improving diagnostic accuracy in challenging cases.

While LLMind shows promise in augmenting mental health assessments, it is essential to acknowledge its limitations and consider practical implications for clinical use, especially given its current accuracy rate of 76.1%. This accuracy rate, while promising, implies that the model may produce errors in approximately one out of four cases. In clinical settings, such errors could lead to misdiagnosis or missed diagnoses, potentially impacting patient care and safety. Incorrect or incomplete diagnoses could result in inappropriate treatment recommendations, raising ethical concerns and potential liability issues for clinicians. Therefore, as stated above, human supervision remains crucial, and LLMind should be viewed as a decision-support tool rather than a replacement for clinicians. Mental health professionals must critically evaluate and interpret the model's outputs, using their clinical judgement and expertise to ensure accurate diagnoses and appropriate treatment plans.

Regarding evaluation settings, the DSM-5-TR and ICD-11 differ in diagnostic criteria, leading to minor discrepancies in diagnoses and potentially lowering agreement rates in studies. Despite the clinicians' ratings considering these differences, they may still underestimate the accuracy of diagnostic models, as discrepancies stem from systemic classification variations rather than model performance. Future research will refine agreement estimates and better assess diagnostic models across systems.

6. LLMind Chat

To deliver the functionalities of the LLMind model, an interactive chat called LLMind Chat has been developed. The intuitive chat-based interface allows healthcare professionals to easily input patient information and receive diagnostic suggestions and differential diagnoses in real-time. This conversational approach offers flexibility, enabling the adaptation of the dialogue to the needs of the specific case, and promotes efficiency, reducing the time required to obtain information. The familiarity of the chat interface and the traceability of the information exchanged make LLMind Chat a user-friendly and transparent tool.

This Section illustrates the architecture of LLMind Chat and all the technologies used to build the entire system. The architecture includes multiple independent software modules that cooperate by exchanging data, resulting in a cohesive system that provides comprehensive functionality and allows for scalability, flexibility, and easier maintenance. Fig. 8 shows the different software components implemented and deployed using Docker containers. The core module has been integrated into a web application developed with T3 Stack,[23] a TypeScript full-stack framework. T3 Stack is based on Next.js[24] which is a popular React framework that provides server-side rendering, static site generation, and other advanced features.

A description of each component/module of the architecture represented in Fig. 8 will be provided below:

1. Frontend Module: the module is responsible for the user interaction during table annotation. The technologies of the T3 Stack that have been integrated into this module are: (i) *TypeScript*,[25] a strongly typed programming language that builds on JavaScript, providing static typing to catch errors early in the development process; (ii) *Tailwind CSS*,[26] a utility-first CSS framework that provides low-level utility classes to build responsive and modern UIs quickly. Tailwind CSS promotes rapid development and allows for highly customisable and consistent styling across the application.

2. Backend Module: the backend module is fundamental as it interacts with requests from the frontend (which are made by the user), and then performs all the operations behind the scenes. The main role of this component is to collect requests from the frontend and save the data in the database. In this module, the technologies used from T3 Stack are: (i) *tRPC*,[27] a TypeScript-based remote procedure call framework which allows the building of type-safe APIs. It provides end-to-end type safety, enabling the developer to define the APIs in a single place and get full type inference across the stack; (ii) *Drizzle*,[28] an ORM tool for *Node.js* and *TypeScript* that provides a type-safe database client. It simplifies database interactions by offering an intuitive schema definition language and powerful query capabilities, and it supports popular databases such as PostgreSQL and MySQL.

3. LLM API: the last component of the architecture is represented by the *LLM API*, which allows the application to effectively conversate with the user. The implementation of this module is separated from the T3 Stack and has been developed using the Flask[29] backend framework. It is a Python framework, which is known for its simplicity, readability, and versatility. Its clear syntax makes it easy to learn and write, reducing the complexity of coding tasks.

4. Database: Chats, messages and validations are permanently stored in a *Postgres*[30] database, which is perfectly integrable in the *Next JS* system, and enables to retrieve tables efficiently. *LLM API* integrates *Chroma DB*, an open-source vector database designed to store and manage high-dimensional vector embeddings. The embeddings, as described, are retrieved through similarity search in RAG-based systems to enrich the model's prompt with additional contextual information.

Having illustrated the overall architecture, an explanation of how to use LLMind Chat in practice is provided below, including a description of the UI and an overview of its functionalities. LLMind Chat UI[31] has been built using a single page which includes all the required functionalities:

[23] create.t3.gg
[24] nextjs.org
[25] typescriptlang.org
[26] tailwindcss.com
[27] trpc.io
[28] orm.drizzle.team
[29] flask.palletsprojects.com
[30] postgresql.org
[31] llmind.datai.disco.unimib.it

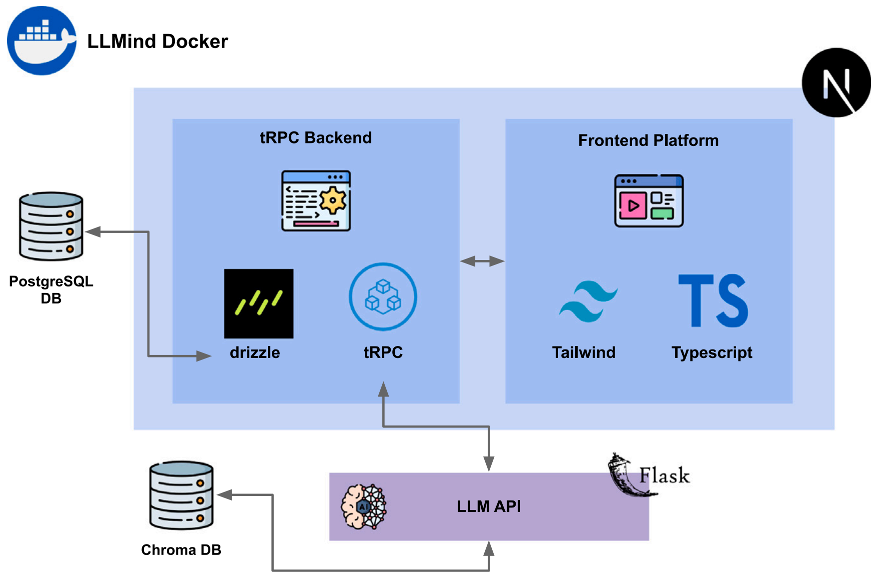

Fig. 8. Architecture of LLMind Chat.

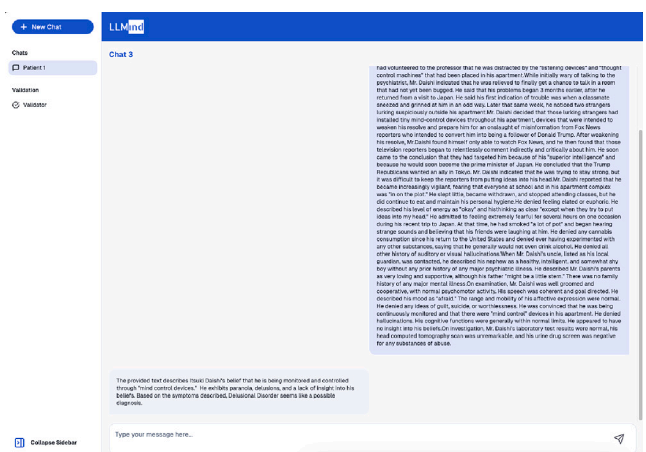

Fig. 9. Chat Screen.

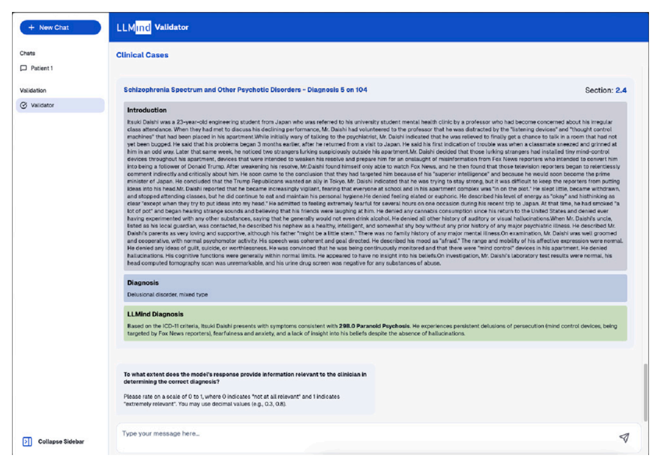

Fig. 10. Validation Screen.

- **Chat creation**: allows the creation of a new chat from the user to start a new conversation with the chatbot (Fig. 9);
- **Conversation**: allows the user to write the text message to the chatbot and then to visualise the given response (Fig. 9);
- **Validation**: allows the domain expert to insert a validation by assigning a score between 0 and 1. (Fig. 10).

The source code for LLMind is open-source and available under the AGPL license.[32] This allows users to freely access, modify, and contribute to the platform's development.

In the LLMind Chat the clinician is responsible for gathering and interpreting patient information, including conducting thorough assessments and evaluating the reliability of patient responses. This information is then used to create a comprehensive and accurate clinical picture, which is then inputted into LLMind to aid in the diagnostic process.

Therefore, LLMind's reliance on clinician-provided information serves as a safeguard against the potential issues of misdiagnoses and inaccurate patient responses. The clinician acts as a filter, ensuring that the data entered into the model is as accurate and objective as possible. This collaborative approach leverages both human expertise and AI capabilities to enhance diagnostic accuracy and support clinical decision-making. Furthermore, we believe that this approach aligns with the responsible use of AI in healthcare, where human oversight and clinical judgement remain central to patient care.

LLMind is designed with a modular architecture, allowing it to dynamically integrate new data sources and continuously adapt to specific clinical contexts and evolving scientific knowledge. This flexibility is essential for mitigating the impact of unreliable input data, ensuring that the model operates with validated and high-quality information (*e.g.*, patient medical records). In fact, for example, in real-world settings, patients may struggle to provide complete or entirely accurate information, either due to cognitive biases, memory limitations, or the subjective nature of self-reporting in mental health.

7. Ablation study

A comprehensive comparative analysis is essential to justify the presented results and the model selection process. After selecting the knowledge base, detailed in Section 3, extensive testing was conducted to identify the optimal LLM for the disorder identification task.

Ollama offers a diverse range of compatible LLMs, including community-driven models, those developed by companies like Google

[32] github.com/unimib-datAI/LLMind

(*e.g.*, Gemma), and customised versions of publicly available LLMs. This research focused on four of the most widely-used LLMs, including variations like Gemma with 9.24B and 27B parameters.

The 4 LLMs analysed are listed below:

- **Phi 3**[33] is a robust model that efficiently handles natural language understanding and generation tasks. It is optimised for medium-scale deployments, focusing on high accuracy with lower computational requirements than larger variants, and it is licensed under the Apache 2.0 license. Phi 3 strikes a balance between performance and resource consumption, offering competitive results on NLP tasks, such as question answering and text classification, while maintaining lower inference costs; Phi 3 can understand and generate text in multiple languages and supports a context window of up to 16k tokens;
- **Mistral NeMo**[34] is a high-performance language model designed for efficient deployment in production systems. It utilises a sparse mixture of experts (SMoE) approach to achieve higher scalability and performance in complex NLP tasks, it is licensed under the Apache 2.0 license, and offers remarkable performance in NLP benchmarks. Mistral NeMo can process up to 40k tokens, supports multiple languages, and performs excellently in tasks requiring a deep understanding of context and semantics. It is also fine-tuned to follow instructions and provide precise outputs;
- **Llama 3**[35] is the third iteration of the Llama family of models and is licensed under Apache 2.0. It can handle a context of up to 64k tokens, supports a wide range of languages, and is fine-tuned for instruction-following tasks, demonstrating exceptional versatility and robustness across diverse NLP applications.
- **Gemma**,[36] which was developed by Google, is the chosen LLM. The main advantages of employing Gemma 2 model are:
  - **Model Type**: Gemma 2 is an advanced large language model developed for various tasks, leveraging SOTA natural language processing techniques. It combines deep learning approaches with efficient scaling to handle diverse input types.
  - **Licensing**: Gemma 2 is licensed under the Apache 2.0 license, allowing users to freely use, modify, and distribute the model in compliance with the license terms.
  - **Performance**: Gemma 2 provides competitive performance across various NLP benchmarks, demonstrating strong results in text generation, summarisation, and translation tasks, significantly outperforming previous models in speed and accuracy.
  - **Capabilities**: Gemma 2 can process long input sequences, supports multiple languages, and excels in few-shot learning scenarios. It is fine-tuned to follow user instructions, ensuring a reliable and high-quality response in various contexts.

Fig. 11 illustrates the frequency with which various models produce inconclusive or empty answers in response to queries. Notably, the results demonstrate that models such as Gemma 2 and **Mistral Nemo** consistently generate responses suitable for accurate analysis, avoiding incomplete or irrelevant outputs.

This superior performance can be attributed to their large number of parameters. In the context of LLMs, the number of parameters refers to the complexity of the model and its ability to capture nuanced patterns and relationships in the training data. More extensive models typically have a greater capacity to understand context, interpret ambiguities, and generalise knowledge effectively.

[33] ollama.com/library/phi3

[34] ollama.com/library/mistral-nemo

[35] ollama.com/library/llama3

[36] ollama.com/library/gemma2

The graphs reported in Fig. 12 show how each model performs on DSM-5-TR. Even in this case, the model with the most parameters (*i.e.*, Gemma 2) achieves the best Accuracy, Precision, Recall and F1-Score. This behaviour shows that larger models can process and generate language with greater sophistication, enabling them to better understand the context and nuances of queries and retrieved documents. They are more adept at integrating subtle relationships between information retrieved by the RAG component, ensuring the generated response is coherent and contextually relevant.

8. Conclusion and future work

This research explored the potential of LLMs in assisting mental health professionals with diagnostic decision-making. It was found that different LLMs perform differently depending on their intended use case, with Gemma 2 model, typically used for creative writing, proving most effective for this application.

Findings underscore that while LLMs cannot replace human clinicians, they can be valuable tools for supporting mental health professionals in the diagnostic process. Specifically, they can aid in differential diagnosis, inform treatment planning, and provide valuable training for young psychologists and psychiatrists. Notably, the results were validated by a psychometrics expert, revealing that some initial discrepancies between the model and human diagnoses ultimately highlighted nuanced aspects of the cases, further reinforcing the model's potential.

While promising, this project requires further refinement before market readiness. Future development will leverage cloud computing to adapt the LLM to a larger dataset and enhance its capabilities. This will allow mental health professionals to proactively identify potential mental health conditions, analyse extensive clinical files more efficiently, and expedite diagnosis and treatment. Ultimately, the aim is to incorporate treatment evaluation capabilities, providing a comprehensive tool for mental health professionals.

The automatic and manual evaluation of the LLMind has highlighted limitations in considering the patient's history, including the temporal evolution of symptoms, age, and medication history. Possible solutions include incorporating a mechanism to track the temporal evolution of symptoms and the patient's age, expanding the training set to include and utilise pharmacological information, and prioritising previously reported diagnoses without disregarding the possibility of errors. Future research will focus on refining LLMind's performance, particularly in managing comorbidity, recognising atypical clinical presentations, and integrating information from diverse sources.

We are exploring the possibility of implementing a feature that allows the model to ask specific questions to discriminate between different pathologies in situations of low confidence, guided by the information in the ICD-11-CDDR. This approach would help improve the model's accuracy, especially in complex cases where differential diagnosis is critical.

Before any production release, LLMind Chat must fully comply with all applicable legal and ethical requirements to ensure the responsible use of AI in mental health settings. This includes adhering to strict data minimisation principles, implementing secure data storage protocols, and establishing clear guidelines for obtaining informed consent where necessary. Additionally, LLMind Chat will meet the requirements of the proposed EU AI Act,[37] conducting rigorous risk assessments, ensuring transparency in its operation, and maintaining definitive mechanisms for human oversight throughout its lifecycle.

This work was born from a deep desire to empower mental health professionals with the tools they need to address the growing challenges in their field, particularly the concerning rise of mental health disorders

[37] digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

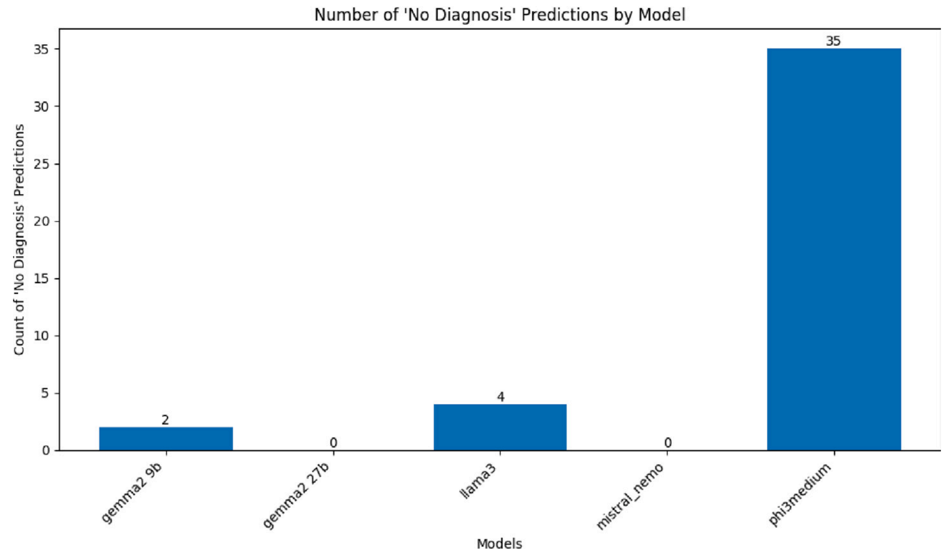

Fig. 11. Number of times the models returned empty or inconclusive answers.

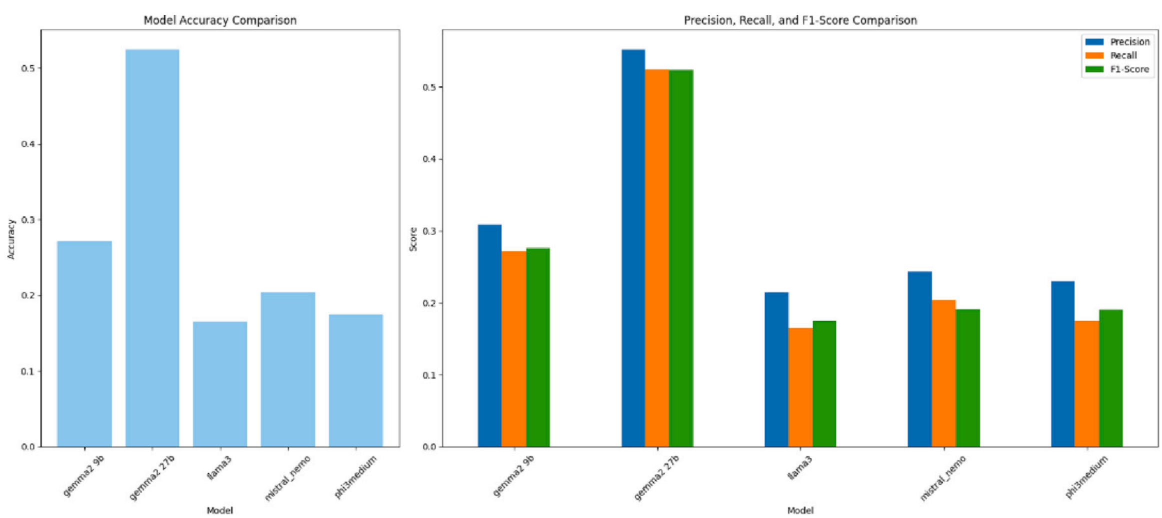

Fig. 12. Scores of the models.

among young-adults. LLMind Chat can represent a significant step towards this goal. Our commitment to open-source accessibility underscores our belief in collaborative innovation. We invite researchers, clinicians, and developers to join us in refining and expanding this technology, ensuring its potential is fully realised in improving the lives of those affected by mental health disorders.

In conclusion, the thoughtful integration of LLMs into mental health care has the potential to significantly enhance diagnostic processes and therapies as a consequence, and it can provide ongoing monitoring of patient progress. However, it is imperative that these tools are used responsibly, with human oversight remaining central to clinical decision-making. The future of mental health care could be greatly improved through a synergy between advanced AI technologies and human expertise.

CRediT authorship contribution statement

Marco Cremaschi: Conceptualization, Methodology, Validation, Investigation, Formal analysis, Writing – original draft, Writing – review & editing, Supervision. **Davide Ditolve:** Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Cesare Curcio:** Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Anna Panzeri:** Conceptualization, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Andrea Spoto:** Supervision, Writing – review & editing. **Andrea Maurino:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially funded by the "Ministry of Enterprises and Made in Italy" within the framework "Innovation Agreements" (Accordi per l'innovazione), project code F/310240/01-04/X56.

Data availability

Data will be made available on request.

References

Agarwal, S., Yadav, A. S., Dinesh, V., Vatsav, K. S. S., Prakash, K. S. S., & Jaiswal, S. (2023). By artificial intelligence algorithms and machine learning models to diagnosis cancer. *Materials Today: Proceedings*, *80*, 2969–2975, SI:5 NANO 2021.

American Psychiatric Association (2022). *Diagnostic and statistical manual of mental disorders, Fifth Edition, text revision (DSM-5-TR®)*. American Psychiatric Association Publishing.

American Psychiatric Association (2023). *DSM-5-TR® clinical cases*. American Psychiatric Association Publishing.

Baghdadi, N. A. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*.

Barnhart, H. X., Haber, M., & Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, *58*(4), 1020–1027.

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., et al. (2021). AI applications to medical images: From machine learning to deep learning. *Physica Medica*, *83*, 9–24.

Chen, Y., Wang, Z., Xing, X., zheng, h., Xu, Z., Fang, K., et al. (2023). BianQue: Balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT.

Christophe, C., Kanithi, P. K., Raha, T., Khan, S., & Pimentel, M. A. (2024). Med42-v2: A suite of clinical LLMs.

Dankelman, L. H. (2023). Artificial intelligence fracture recognition on computed tomography: review of literature and recommendations. *European Journal of Trauma and Emergency Surgery*, *49*(2), 681–691.

El-Hay, M. A. A. (2018). *Essentials of psychiatric assessment*. Routledge.

Fried, E. (2020). Systems all the way down: embracing complexity in mental health research. *BMC Medicine*, *18*.

Gupta, R. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, *25*, 1315–1360.

Hansun, S. (2023). Machine and deep learning for tuberculosis detection on chest X-rays: systematic literature review. *Journal of Medical Internet Research*, *25*.

Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., et al. (2023). Health system-scale language models are all-purpose prediction engines. *Nature*, *619*(7969), 357–362.

Joseph, A. (2024). The unseen dilemma of AI in mental healthcare. *AI and Society*.

Labrak, Y. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains.

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., et al. (2023). Psy-LLM: Scaling up global mental health psychological services with AI-based large language models.

Lee, E. E. (2021). Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *6*, 856–864.

Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance* (pp. 374–382). New York, NY, USA: Association for Computing Machinery.

Luo, L., Ning, J., Zhao, Y., Wang, Z., Ding, Z., Chen, P., et al. (2024). Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association*, *31*(9), 1865–1874.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., et al. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6), bbac409.

Ma, Y., Wang, Z., Yang, H., & Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA Journal of Automatica Sinica*, *7*(2), 315–329.

Maj, M. (2021). The clinical characterization of the patient with primary psychosis aimed at personalization of management. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *20*(1), 4–33.

Mak, K.-K., Wong, Y.-H., & Pichika, M. R. (2024). Artificial intelligence in drug discovery and development. In *Drug discovery and evaluation: safety and pharmacokinetic assays* (pp. 1461–1498). Cham: Springer International Publishing.

McGrath, J. J. (2020). Comorbidity within mental disorders: a comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiology and Psychiatric Sciences*, *29*.

Meshaka, R. (2023). Artificial intelligence applied to fetal MRI: A scoping review of current research. *The British Journal of Radiology*, *96*.

Nguyen, H.-T. (2023). A brief report on lawgpt 1.0: A virtual legal assistant based on GPT-3.

Oren, O., Gersh, B. J., & Bhatt, D. L. (2020). Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health*, *2*(9), e486–e488.

Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., et al. (2023). A study of generative large language model for medical research and healthcare. *NPJ Digital Medicine*, *6*(1), 210.

Regier, D. A. (2013). The DSM-5: Classification and criteria changes. *World Psychiatry : Official Journal of the World Psychiatric Association (WPA)*, *12*, 92–98.

Ren, X., Bai, Y., Duan, H., Fan, L., Fei, E., Wu, G., et al. (2024). Chatasd: LLM-based AI therapist for ASD. In *Digital multimedia communications* (pp. 312–324). Singapore: Springer Nature Singapore.

Sartorious, N. (2013). Comorbidity of mental and physical diseases: a main challenge for medicine of the 21st century. *Shanghai Archives of Psychiatry*, *25*(2), 68–69.

Seemüller, F. (2023). ICD-11 and mental disorders: important changes, controversies, and future directions. *BMC Psychiatry*.

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., et al. (2021). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering*, *14*, 4–15.

Singh, K. R., & Dash, S. (2023). Early detection of neurological diseases using machine learning and deep learning techniques: A review. In A. Abraham, S. Dash, S. K. Pani, & L. García-Hernández (Eds.), *Artificial intelligence for neurological disorders* (pp. 1–24). Academic Press.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180.

Sun, Z. (2023). A short survey of viewing large language models in legal aspect.

Tirkolaee, E. B. (2021). Application of machine learning in supply chain management: a comprehensive overview of the main areas. *Mathematical Problems in Engineering*.

Wang, Y., Qian, W., Zhou, H., Chen, J., & Tan, K. (2023). Exploring new frontiers of deep learning in legal practice: A case study of large language models. *International Journal of Computer Science and Information Technology*, *1*(1), 131–138.

Wei, H.-H., Zhang, Y., Sun, X., Chen, J., & Li, S. (2023). Intelligent robots and human–robot collaboration in the construction industry: A review. *Journal of Intelligent Construction*, *1*(1), Article 9180002.

World Health Organization (2022). *ICD-11: International classification of diseases (11th revision)*. World Health Organization.

Wright, A. J. (2020). *Conducting psychological assessment: A guide for practitioners*. John Wiley and Sons.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., et al. (2023). Bloomberggpt: A large language model for finance.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., et al. (2024). Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *8*(1).

Yan, W.-J., Ruan, Q.-N., & Jiang, K. (2022). Challenges for artificial intelligence in recognizing mental disorders. *Diagnostics*, *13*(1), 2.

Yang, K. (2023). MentaLLaMA: Interpretable mental health analysis on social media with large language models..

Zeb, S. (2024). AI in healthcare: Revolutionizing diagnosis and therapy. *International Journal of Multidisciplinary Sciences and Arts*, *3*(3), 118–128.

Zhang, X., Tian, C., Yang, X., Chen, L., Li, Z., & Petzold, L. R. (2024). AlpaCare:Instruction-tuned large language models for medical application.

Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., et al. (2024). A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 1–13.

Zhou, Z., Shi, J.-X., Song, P.-X., Yang, X.-W., Jin, Y.-X., Guo, L.-Z., et al. (2024). LawGPT: A Chinese legal knowledge-enhanced large language model.