Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Knowledge Graph and Large Language Model Framework for Personalized Clinical Decision Support in Bipolar Disorder and Borderline Personality Disorder: The KG- Psy Framework

Zahra Mahmoudpour Harris
Shariaty Technical College
Reza Rahaeimehr
Augusta University
Saba Hesaraki
saba.hesaraki@iau.ir

Islamic Azad University

Article

Keywords: Bipolar disorder(BD); Borderline personality disorder(BPD); Knowledge graph(KG); Large language model(LLM); Clinical decision support; Personalized treatment; Explainability

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9621282/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Knowledge Graph and Large Language Model Framework for Personalized Clinical Decision Support in Bipolar Disorder and Borderline Personality Disorder: The KG-Psy Framework

Zahra Mahmoudpour Harris[1], **Reza Rahaeimehr**[2], **and Saba Hesaraki**[3,*]

[1]Shariaty Technical College, Tehran, Tehran, Iran.
[2]Augusta University, Augusta, USA.
[3]Islamic Azad University, Science and Research Branch ,Tehran, Iran
[*]Corresponding author: ; saba.hesaraki@iau.ir

ABSTRACT

Background: Differentiating between bipolar disorder (BD) and borderline personality disorder (BPD) is an intriguing clinical challenge that has attracted research attention, as the multiple co-occurring symptoms such as mood instability, impulsivity, and emotional dysregulation have made the boundary between the two disorders unclear. Recent research in computational psychiatry has shown promise in leveraging the power of knowledge graphs and large language models to enable accurate and interpretable decision-making.
Methods: The study developed a hybrid framework, KG-Psy, that combines large language models with dynamic graphs. The system extracts clinical entities through GPT-5 while employing multi-step reasoning to find hidden connections within a psychiatric knowledge graph that weighs its evidence-based content. The system performs four main operations, which include extracting clinical entities, creating dynamic weighted knowledge graphs, using GPT-5 to trace multi-step cognitive pathways, and evaluating customized treatment options. The researchers used 310 anonymized psychiatric case reports to test their method.
Results: KG-Psy achieved better discrimination between BD and BPD than other methods. The model achieved a maximum Precision (%) of 91.5, an improvement of 19%, 10.3%, and 7.4% over the KG-only, LLM-only, and Hybrid-other approaches, respectively. The system successfully discovered unknown cognitive pathways, which it used to create accurate pathway maps and provide treatment recommendations supported by numerical evidence. The results of ablation experiments showed that LLM reasoning, combined with the knowledge graph structure, led to high performance.
Conclusion: The KG-Psy system introduces a novel approach to psychiatry inspired by large-scale language models and dynamic graphs that integrates data-driven linguistic analysis with established symbolic knowledge to enhance the capabilities of differential diagnosis and treatment development for complex mood and personality disorders. Future work could include validation on larger datasets and integration of multimodal clinical data.
keywords:Bipolar disorder(BD); Borderline personality disorder(BPD); Knowledge graph(KG); Large language model(LLM); Clinical decision support; Personalized treatment; Explainability

1 Introduction

Bipolar disorder (BD) and Borderline Personality Disorder (BPD) are two disorders that constitute complex psychiatric conditions that pose significant challenges in diagnosis and treatment[1]. Despite extensive

research, differentiating these disorders and drawing a line between them is difficult due to overlapping symptoms[1]. As in any area of psychiatry, misdiagnosis can lead to inappropriate treatment choices, which can lead to increased risk and longer-term consequences. Traditional clinical decision-making relies heavily on manual assessments, structured interviews, and standardized questionnaires. These assessments often fail to capture the full range of cognitive, affective, and neurobiological dynamics underlying these disorders.

Bipolar disorder (BD) is a chronic, cyclical mental disorder characterized by unusual mood swings between hypomania and depression. The disorder has raised concerns in the scientific and medical communities due to its devastating social and economic impact. The course of the illness is often persistent, leading to recurrent cognitive deficits over time[3].

characterized by abrupt changes in identity, interpersonal relationships, and emotions, as well as impulsive behaviour, suicidal behaviour, self-harm, recurrent bouts of anger, feelings of worthlessness, stress, and severe dissociative symptoms. Borderline personality disorder is interview-based and is usually diagnosed by a mental health professional using semi-structured interviews. Most people with borderline personality disorder have co-occurring mental disorders, such as mood disorders, anxiety disorders, or substance use disorders. The etiology of borderline personality disorder is related to both genetic factors and adverse childhood experiences, such as sexual and physical abuse[4]. Medication is usually the best solution for this disorder. For severe, isolated mental disorders, such as major depression, medications such as escitalopram, sertraline, or fluoxetine may be prescribed. For short-term treatment of acute crises in BPD, including suicidal behavior or thoughts, psychotic episodes, or the prescription of low-potency antipsychotics (such as quetiapine) may be recommended[4].

The establishment of computational psychiatry provides new possibilities for implementing data-driven medical solutions in clinical research facilities. Knowledge graphs (KGs)[567] serve as an efficient scientific tool which enables researchers to develop organized systems. The graphs display intricate connections between patient symptoms and brain circuits and treatment results and health conditions. KGs use graph-based structures to represent clinical data from diverse sources, allowing users to infer detailed relationships and use mechanistic models to study psychiatric disorders. The main limitation of KGs is that they rely on pre-existing knowledge or explicit information, making them unable to detect hidden relationships that emerge from specific contexts, which clinicians need to create personalized treatment plans[8].

The increasing processing power of graphics processing units (GPUs) and recent advances in large language models (LLMs), especially models such as GPT-5[9], have enabled unprecedented capabilities in text understanding. The availability of large language models containing billions of parameters enables natural language understanding and reasoning in unstructured clinical and biomedical texts. LLMs can extract subtle clinical entities. They can also infer hidden relationships and provide interpretable recommendations based on context. Various approaches have been proposed in the literature for this purpose, typically using a dataset of millions of words in a property-based context as input and building a language model from them. However, LLMs alone may produce outputs that are inconsistent and unfounded. These unfounded outputs may be difficult to validate and pose challenges for clinical acceptance.

In this study, the KG-Psy framework is proposed to address these challenges. KG-Psy is a composite approach that utilizes two entities: GPT-5 (for contextual inference) and KG (for structured reasoning). In fact, in the KG-Psy approach, GPT-5 is responsible for entity extraction, multi-step reasoning, and personalized recommendation generation. On the other hand, the role of KG is to provide a structured model that supports consistency and interpretability and ensures evidence-based validation. Finally, the KG-Psy approach provides a multi-layered representation that includes symptom pathways, neural corre-

lates, treatment options, comorbidities, and longitudinal behavioural markers.

KG-Psy, along with the proposed hybrid approach, has several advantages, the most important of which are the following:

- KG-Psy supports by combining patient symptoms and past treatment experiences.
- KG-Psy enables optimization of treatment recommendations for both BD and BPD by considering pharmacological, psychotherapy, and lifestyle interventions.
- KG-Psy, using confidence measures and graph structures, provides outputs that are both quantitatively valid and interpretable.

In this study, the KG-Psy model was evaluated on a set of 310 de-identified psychiatric case reports and 12 expert-created benchmark scenarios. KG-Psy can distinguish BD from BPD with high accuracy. It also provides personalized treatment plans. In addition, it identifies hidden symptom-treatment correlations that are often missed in standard clinical procedures. The KG-Psy model provided a new paradigm for precision psychiatry by combining knowledge graph reasoning with LLM-based inference, a combination adopted by GPT -5 for this purpose.

1.1 LLM background

Language is a medium of human expression and communication that changes throughout a lifetime. Machine-based products cannot understand human language directly. This is a challenge that requires powerful AI algorithms. This challenge has been a long-standing research goal, enabling machines to read, write, and communicate like humans[10]. The idea of language modelling (LM) has been committed to this goal, which is one of the primary methods for developing machine language intelligence. LM modelling focuses on modelling the probability of generating word sequences to predict the probability of future words. The manner in which text sequences are written is denoted by tokens, which can be considered as a sequence of discrete observations of words or characters. The objectives of the language model are to estimate the joint probability of the whole sequence as follows[11]:

$$P(x_1, x_2, \ldots, x_T) \tag{1}$$

The goal of LM is generally to map token x at time t, which can be expressed as $x_t \sim P(x_t \mid x_{t-1}, \ldots, x_1)$. Various NLP tasks, such as text classification, text prediction, sentiment analysis, and more, are derived from LMs. The first attempts at LMs were statistical LMs to assign probabilities to a sequence of N tokens, which can be expressed as:

$$P(s) = P(w_1, w_2, \ldots, w_N) = P(w_1)P(w_2|w_1)\ldots P(w_N|w_1w_2\ldots w_{N-1}) \tag{2}$$

w_i is the word i in the sentence sequence s. According to the Markov Chain theorem, the probability of a sequence of words can be divided by the product of the conditional probability of the next word given its previous words. Given these interpretations, the following relations can be obtained:

Uni Unigram: $P(w_1)P(w_2)P(w_3)\ldots P(w_n)$

Bi Bigram: $P(w_1)P(w_2|w_1)P(w_3|w_2)\ldots P(w_n|w_{n-1})$

Tri Trigram:$P(w_1)P(w_2|w_1)P(w_3|w_2, w_1)\ldots P(w_n|w_{n-1}, w_{n-2})$

N N-gram: $P(w_1)P(w_2|w_1)\ldots P(w_n|w_{n-1}, w_{n-2}\ldots w_{n-N})$

A $k+1$-gram model is derived from the assumption of Markov order k and can be defined as follows:

$$P(w_t|w_1\ldots w_{t-1}) \approx P(w_t|w_{t-k}\ldots w_{t-1}) \tag{3}$$

N-gram LMs are created by constructing an n-gram count table from the training data. This table stores all the unique n-grams in the training data, each associated with its count. Such n-gram count tables are large and grow almost exponentially. An n-gram LM is defined as follows[?]:

$$P(w_i, w_{1:i-1}) = \frac{cnt(w_{i-(n-1):i-1}w_i \mid D)}{cnt(w_{i-(n-1):i-1} \mid D)} \tag{4}$$

2 Background and Related Work

Recently, with the development of artificial intelligence, major breakthroughs in computational psychiatry have led to it becoming a popular area of research for both medical professionals and researchers. The aim of these studies has been the integration of KG and LLMs for proper diagnosis and treatment of complicated diseases like BD and BPD. This section will discuss the related work, including the challenges of diagnosis, the use of KG in psychiatry, LLM models in medical applications, and the combination of neuro-symbolic models.

2.1 Diagnosing and treating BD and BPD Challenges

Accurate diagnosis of BD and BPD is difficult due to overlapping symptoms and mood swings. Approaches in the literature based on traditional methods often lead to misdiagnosis. Many studies have been conducted to establish a clear and non-overlapping feature boundary. Variables such as nonparametric circadian rest-activity, impulsivity with low stability, mood instability with low inter-day stability, greater rhythm fragmentation, and later onset of daily activity are examples that cannot be defined for BD[12]. Also, study[13] investigated actigraphy patterns to differentiate impulsivity and mood instability in BD and BPD. This study showed that sleep and circadian disturbances can predict symptom severity. On the other hand, studies show that traditional methods such as structured interviews cannot fully capture the underlying neural patterns and dynamics of symptoms[14].

2.2 Knowledge Graphs in Psychiatry

Knowledge graphs have emerged in the literature as effective tools and symbolic frameworks for structuring complex relationships between symptoms, neural circuits, pharmacological interventions, and behavioural outcomes[15]. KGs were initially proposed for mental disorders. The Mental Disorders Knowledge Graph (MDKG) was presented as a pioneering work in KG research. The MDKG is a large-scale, contextualized knowledge graph built using large language models to integrate evidence from biomedical literature and databases. The MDKG contains over 10 million relationships and nearly 1 million new relationships not found in existing sources. As a scalable, rich resource, MDKG provides a powerful foundation for accelerating psychiatric research and generating interpretable, data-driven clinical insights[16]. In addition, in[17], KG was developed to model mental disorders and causal relationships were used. Knowledge graphs, as organized models of information, when coupled with big data and deep learning technologies, offer significant potential for studying neuropsychiatric disorders. In addition to

deep learning, machine learning can also be enhanced by combining with KG data. Study[18] focuses on the use of KG embeddings and shows that KG can enhance machine learning on incomplete data. However, inferring hidden relationships for personalized care remains a major challenge in KGs, as they are often limited to explicit knowledge[19].

2.3 Large language models in clinical and psychiatric applications

LLMs have advanced capabilities in natural language understanding, clinical entity extraction, and multi-stage reasoning[20]. Clinical notes contain information about patients and medications that goes beyond unstructured data. However, clinical notes have been less widely used than structured data because they are high-dimensional, heterogeneous, and sparse. ClinicalBERT[21] developed and evaluated a continuous representation of clinical notes to predict 30-day hospital readmissions at different time points of admission, including early and discharge. This language model applies bidirectional encoder representations from a transformer (BERT) to clinical text. BERT parameters were trained on standard datasets such as Wikipedia and BookCorpus, which are different from clinical text. Therefore, the authors pre-trained BERT on clinical notes and fine-tuned the network for predicting hospital readmissions, calling the resulting model ClinicalBERT. Clinical-BERT reveals high-quality relationships between medical concepts as judged by clinicians.

In[22], a large clinical language model, GatorTron, was developed using 90 billion words of text. This language model was systematically applied to medical relation extraction, semantic text similarity, natural language inference (NLI), and medical question answering (MQA). The GatorTron models increased the clinical language model from 110 million to 8.9 billion parameters. GatorTron achieved a maximum of F1 = 0.900 in Clinical concept extraction and F1 = 0.9627 in Medical relation extraction. Another example of the application of GatorTron was presented in[23]. This study developed a generative clinical LLM, GatorTronGPT, trained on 277 billion words of text. The researchers trained GatorTronGPT using the GPT-3 architecture with up to 20 billion parameters and evaluated its application to biomedical natural language processing and healthcare text generation.

MEDITRON[24] improved a set of open-source LLMs in the medical domain on a large scale. MEDITRON is built on Llama-2 and pre-trained on a comprehensive medical dataset including selected PubMed articles. Overall, MEDITRON achieved an absolute performance improvement of 6% over the best public baseline in its parameter class and 3% over the most robust baseline (Llama-2). MEDITRON-70B outperforms GPT-3.5 and Med-PaLM and is within 5% of GPT-4 and 10% of Med-PaLM-2. ClinicalGPT[25] is another language model that is explicitly designed and optimized for clinical scenarios. ClinicalGPT is better prepared for performing multiple clinical tasks by incorporating extensive and diverse real-world data, such as medical records, domain-specific knowledge, and multi-step conversational consultations into the training process. This language model has been trained and evaluated on cMedQA2, cMedQA-KG, MD-EHR, MEDQA-MCMLE, and MedDialog, which serve as its core. Comparisons of medical conversation and medical examination, this model has shown better results than LLAMA-7B, ChatGLM-6B, and BLOOM-7B.

Although recent breakthroughs in commercial language models have yielded promising results in the medical domain, their closed-source nature has raised significant privacy and security concerns, making it difficult for the medical community to adopt them. Table 1 lists a few examples of available LLMs. In recent years, GPT-4-specialized LLMs have been employed to diagnose mental illnesses based on actual cases, with high accuracy of up to 71.7% for ICD-10 classification, especially for mood disorders and schizophrenia[35]. Various studies have investigated the accuracy of LLMs in psychiatric diagnosis, and GPT-4 outperforms previous models[36]. In the context of BD and BPD, LLMs have been used to classify relevant posts with an AUC of up to 0.97 for some disorders[37]. However, LLMs can produce unreliable

Table 1. Open source available LLMs.

Model	Reference	Base Model	Pretraining dataset	Open source
BioELMo	[26]	ELMo	PubMed	✓
BioBERT	[27]	BERT	PubMed, PMC	✓
BlueBERT	[28]	BERT	PubMed	✓
BioMegatron	[29]	BERT	PubMed, PMC	✓
BioLinkBERT	[30]	BERT	ioL PubMed	✓
BioGPT	[31]	GPT	PubMed	✓
BioMedGPT-LM	[32]	LLaMA	PMC, arXiv, WIPO	✓
BioinspiredLLM	[33]	Llama-2	Biological article	✓
BioMistral	[34]	Mistral	PMC	✓

results, which makes them less acceptable in the clinical context[38].

2.4 KG and LLM hybrid approaches

Hybrid frameworks that integrate KGs and LLMs have shown promise in the biomedical and psychiatric domains [14]. As previously reviewed, LLM-based KG was developed for mental health exploration, and its embeddings have shown improved performance for predicting BD and MDD from UK Biobank data[16]. Also,[39] proposed a similar approach for extracting information from mental disorder data by integrating medical KG and LLMs. Systematic reviews in 20252026 emphasized neuro-symbolic synergies and extended the application of KG + LLM for personalized diagnosis and treatment in mental health[38]. However, no previous study has specifically applied the KG + LLM hybrid framework to the differential diagnosis of BD and BPD and personalized treatment planning, highlighting the innovation of KG-Psy.

3 Methodology

The KG-Psy framework integrates a psychiatric knowledge graph (KG) with arguments extracted from GPT-5. This integration is intended to provide personalized diagnostic and treatment recommendations for bipolar disorder (BD) and borderline personality disorder (BPD). The pipeline of the hybrid model is shown in Figure 1. The core of KG-Psy consists of four interrelated phases, which are discussed in detail below.

3.1 Clinical Entity Extraction

For each patient case ăand clinical document, GPT-5 is employed to extract:

- Symptom events: $S = \{\}$
- Neural correlates: $N = \{\}$
- Treatment actions: $T = \{\}$
- Temporal and causal relations: $R = \{\}$

Each extracted tuple is represented as:

$$Tl \in T(s_i, r, n_j, C_{GPT5}) \tag{5}$$

Where C_{GPT5} ădenotes the confidence score provided by GPT-5 for the relation, these tuples form the initial evidence base for the knowledge graph and are later weighted by empirical clinical data.

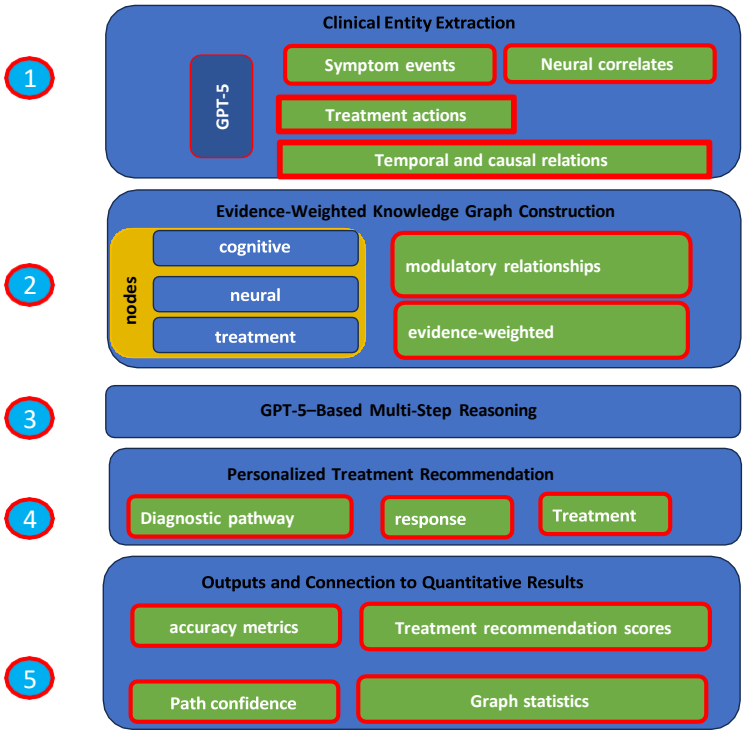


Figure 1. The pipeline of proposed hybrid model.

3.2 Evidence-Weighted Knowledge Graph Construction

The psychiatric knowledge graph $G = (V, E, W)$ is constructed where:

- $V = S \cup N$ represents cognitive, neural, and treatment nodes.
- $E \subseteq V * V$ are directed edges representing causal or modulatory relationships.
- $W : E \to [0, 1]$ assigns evidence-weighted confidence scores to each edge.

The edge weight w_{ij} is computed as:

$$w_{ij} = a + \beta E_{clinical}(v_i, v_j) + aC_{GPT5}(v_i, v_j) \tag{6}$$

$$\gamma S_{KG}(v_i, v_j) + \beta + \gamma = 1 \tag{7}$$

Where C_{GPT5} is GPT-5 confidence for the edge $v_i \to v_j$, $E_{clinical}$ is normalized support from patient-level clinical evidence, and $S_{KG}(v_i, v_j)$ represents structural support drom prior KG.

Edges with $w_{ij} < \tau$ are pruned to ensure reliability. This ensures that both explicit and inferred clinical evidence contribute quantitatively to the graph structure.

3.3 GPT-5Based Multi-Step Reasoning

GPT-5 performs multi-step inference to infer hidden relationships and intermediate cognitive or neural states that are not explicitly observed in the data. This multi-step inference is performed over the weighted

KG. For a candidate diagnostic path $P = (v_1, v_2, \ldots, v_L)$, the path confidence is calculated by GPT-5 as follows:

$$Conf\ P\ ( ) = \sum_{t=1}^{L-1} \frac{1}{L-1} w_{v_t v_{t+1}} \quad (8)$$

Where $w_{v_t v_{t+1}}$ is the evidence-weighted edge score. Longer pathways are inherently penalized for low-confidence intermediate edges, preventing overfitting to unsupported inferences. GPT-5 proposes intermediate nodes $I = i_1, \ldots, i_q$ when necessary, updating V and E dynamically, while the updated pathway confidence guides their inclusion.

3.4 Personalized Treatment Recommendation

Treatment recommendation is modeled as a mapping function:

$$f : C * P \rightarrow T_{personalized} \quad (9)$$

Where C is the patient case, P is the inferred diagnostic pathway, and $T_{personalized}$ is a ranked list of treatment options. Each treatment $t \in T$ is assigned a treatment suitability score:

$$Score(t) = \lambda W_{path} + \mu E_{response}(t, C) + \nu S_{treatment-KG}(t); \lambda + \mu + \nu = 1 \quad (10)$$

Where W_{path} is the confidence of the diagnostic pathway associated with the treatment, $E_{response}(t, C)$ is the empirical efficacy from prior clinical cases for similar patient profiles, and $S_{treatment-KG}(t)$ represents prior knowledge from psychiatric literature (e.g., first-line guidelines, side-effect profiles). Treatments with scores above a threshold θ are recommended, yielding quantitative, patient-specific outputs that can be evaluated against expert benchmarks.

3.5 Outputs and Connection to Quantitative Results

The proposed framework naturally produces numerical and descriptive outputs, which are as follows:

- Diagnostic accuracy metrics: This metric indicates the alignment of the inferred paths with the expert-assessed symptom paths and reveals the BD/BPD distinction.
- Treatment recommendation scores: This metric indicates the quantitative scores of suitability for personalized treatments.
- Path confidence: The path confidence metric uses total edge weight calculations to determine how trustworthy and understandable a route is, which uses interpreted route paths.
- Graph statistics: This metric indicates the overall statistics of the dynamic graph, including the number of nodes, edges, and intermediate inferred states to assess the complexity of the path.

The experimental results from the proposed approach are consistent with the measured values . The proposed framework establishes a direct connection between the mathematical definitions and the derived KG structures. Finally, these results are used to show the relationship with clinical outcomes for research that requires reproducibility and interpretability.

The overall process of the proposed approach is shown in Algorithm 1.

Algorithm 1 Clinical KG and Personalized Treatment Recommendation

Require: Set of patient cases $C = \{c_1, c_2, \ldots, c_n\}$, Set of clinical documents $D = \{d_1, d_2, \ldots, d_m\}$, Large Language Model GPT-5, Empirical clinical evidence database ClinicalEvidence, Edge weighting parameters α, β, γ such that $\alpha + \beta + \gamma = 1$, and Edge pruning threshold τ, $V \to 0/$, $E \to 0/$, $W \to 0/$ ▷ Initialize nodes, edges, and weights

// Step 1: Clinical Entity and Relation Extraction

for each patient $c \in C$ **do**
  for each clinical document $d \in D$ **do**
    $tuples \to$ GPT5.extract_entities_and_relations(d)
    for each tuple $(s_i, r, n_j, C_{\text{GPT5}}) \in tuples$ **do**
      if $s_i \notin V$ **then**
        Add node s_i to V
      end if
      if $n_j \notin V$ **then**
        Add node n_j to V
      end if
      $w_{ij} \to \alpha \cdot C_{\text{GPT5}} + \beta \cdot$ ClinicalEvidence.fMRI(s_i, n_j)
        $+\gamma \cdot$ ClinicalEvidence.document_support(s_i, n_j)
      if $w_{ij} \geq \tau$ **then**
        Add weighted edge $(s_i \overset{r, w_{ij}}{\longrightarrow} n_j)$ to E and W
      end if
    end for
  end for
end for

// Step 2: Multi-hop Reasoning with GPT-5 for Intermediate Nodes

for each pathway $P \subseteq G$ **do**
  $I \to$ GPT5.infer_missing_nodes(P)
  for each inferred node $i_j \in I$ **do**
    Add node i_j to V
    Compute and add incoming/outgoing edges for i_j using GPT-5 + ClinicalEvidence and Update E and W accordingly
  end for
end for

// Step 3: Harmonization and Standardization

for each node $v \in V$ **do**
  Map v to standard ontology (e.g., DSM-5, Cognitive Atlas, Allen Brain Atlas)
end for
for each edge $e \in E$ **do**
  Restrict relation type to canonical set {*activates*, *inhibits*, *modulates*, *connects*}
end for

// Step 4: Pathway Assembly

Identify directed paths $P \subseteq G$ representing sequences symptom → neural mechanism → treatment
for each path P **do**
  $\text{Conf}(P) \to \frac{1}{|P|} \sum_{(i,j) \in P} w_{ij}$ ▷ Average edge confidence
end for

// Step 5: Conflict Resolution and Pruning

for each pair of nodes with contradictory relations **do**
  Retain relation $r^* = \text{argmax}_r \, w_{ij}(r)$
end for
Prune redundant nodes/edges and low-confidence elements

4 Experiments Results

In this section, to evaluate the proposed framework, experiments were conducted on a dataset of 310 de-identified psychiatric case reports from PubMed, along with 12 expert-designed benchmark scenarios. The overall evaluation focuses on the two main tasks of differentiating bipolar disorder (BD) from borderline personality disorder (BPD) and providing personalized treatment recommendations. The proposed framework was compared with baseline methods, including purely LLM-based reasoning and purely KG-based inference.

4.1 Example Cognitive Pathways

The proposed KG-Psy framework generates cognitive pathway diagrams (CPGs). CPGs represent inferred symptom-neuro-treatment sequences for specific cases that reflect the multi-step cognitive and emotional processes underlying bipolar disorder and borderline personality disorder. For this purpose, 4 criteria are used, and the average obtained from them is shown in Table 2. These criteria are as follows:

1 Visual-Attention Pathway in BD: The measure links the inferred path to established visual streams through its alignment. The GPT-5 system establishes path confidence through its reasoning process by creating intermediate nodes that do not exist in the case documentation. The score for this study demonstrated high reliability, with a value of 0.87 across 310 cases. The calculations refer to the average value derived from the 310 cases under examination. The system determines the count of intermediate nodes, which helps confirm the multi-hop reasoning process. An example of the output of this stage is as follows:

$$\textit{Stimulus} \rightarrow V1 \rightarrow V2/V4 \rightarrow \text{Salience Encoding} \rightarrow \text{Dorsal Attention Network} \rightarrow \text{Executive Selection} \rightarrow \text{Motor Preparation}$$

2 Emotional Regulation Pathway in BPD: The goal of this metric is to provide a path that captures the interaction of emotional and executive networks. The intermediate emotional modulation nodes increase the interpretability inferred by GPT-5. The proposed framework achieved an average score of 0.84 on this metric by using an intermediate node. An example of a path suggested by the proposed approach is given below.

$$\text{Threat Cue} \rightarrow \textit{Amygdala} \rightarrow \text{Arousal Modulation} \rightarrow \text{Orbitofrontal Cortex} \rightarrow \text{Executive Inhibition} \rightarrow \text{Behavioral Output}$$

3 Working Memory Pathway in BD: This measure also considers and attempts to map multi-stage cognitive processing for memory tasks. In the proposed framework, this measure achieved an average path confidence score of 0.82, using 2 inferred intermediate nodes. The output of this stage also includes nodes and edges in addition to the path, which is given in the following inference path of this measure:

$$\text{Perceived Stimulus} \rightarrow \text{Encoding Loop} \rightarrow \text{Prefrontal Cortex Updating} \rightarrow \text{Maintenance Network} \rightarrow \text{Response Selection}$$

Table 2. Aggregated insights across 310 cases result.

Cognitive Domain	Avg. Pathway Confidence	Avg. Inferred Nodes	Notes
Visual Attention (BD)	0.87	2	Matches dorsal/ventral streams
Emotional Regulation (BPD)	0.84	1	Includes amygdala-prefrontal modulation
Working Memory (BD)	0.82	2	Multi-step encoding and maintenance
Decision-Making (BPD)	0.79	1	Reward and evaluation integration
Overall (All Domains)	0.83	1.5	High interpretability and mechanistic plausibility

4 Decision-Making Pathway in BPD: This measure is slightly different and captures the interactions between reward processing and emotion regulation. In the proposed framework, this measure achieved an average confidence score of 0.79, with 1 inferential node. In fact, the goal of this measure is to demonstrate the overlap between cognitive and emotional processes. The following process is used to derive this measure:

Reward Cue → Ventral Striatum → Anterior Cingulate Cortex → Orbitofrontal Evaluation → Action Selection → Outcome Monitoring

4.2 Quantitative Evaluation: Diagnostic Accuracy

To assess diagnostic performance, we measured alignment with expert-assessed labels for BD vs. BPD differentiation. Metrics included precision, recall, F1-score, and pathway confidence.

These metrics can be defined as follows:

Pr Precision: $Precision = \frac{TP}{TP+FP}$

Re Recall: $Recall = \frac{TP}{TP+FN}$

F1 F1-score: $F1 - score = 2 * \frac{precision*recall}{precision+recall}$

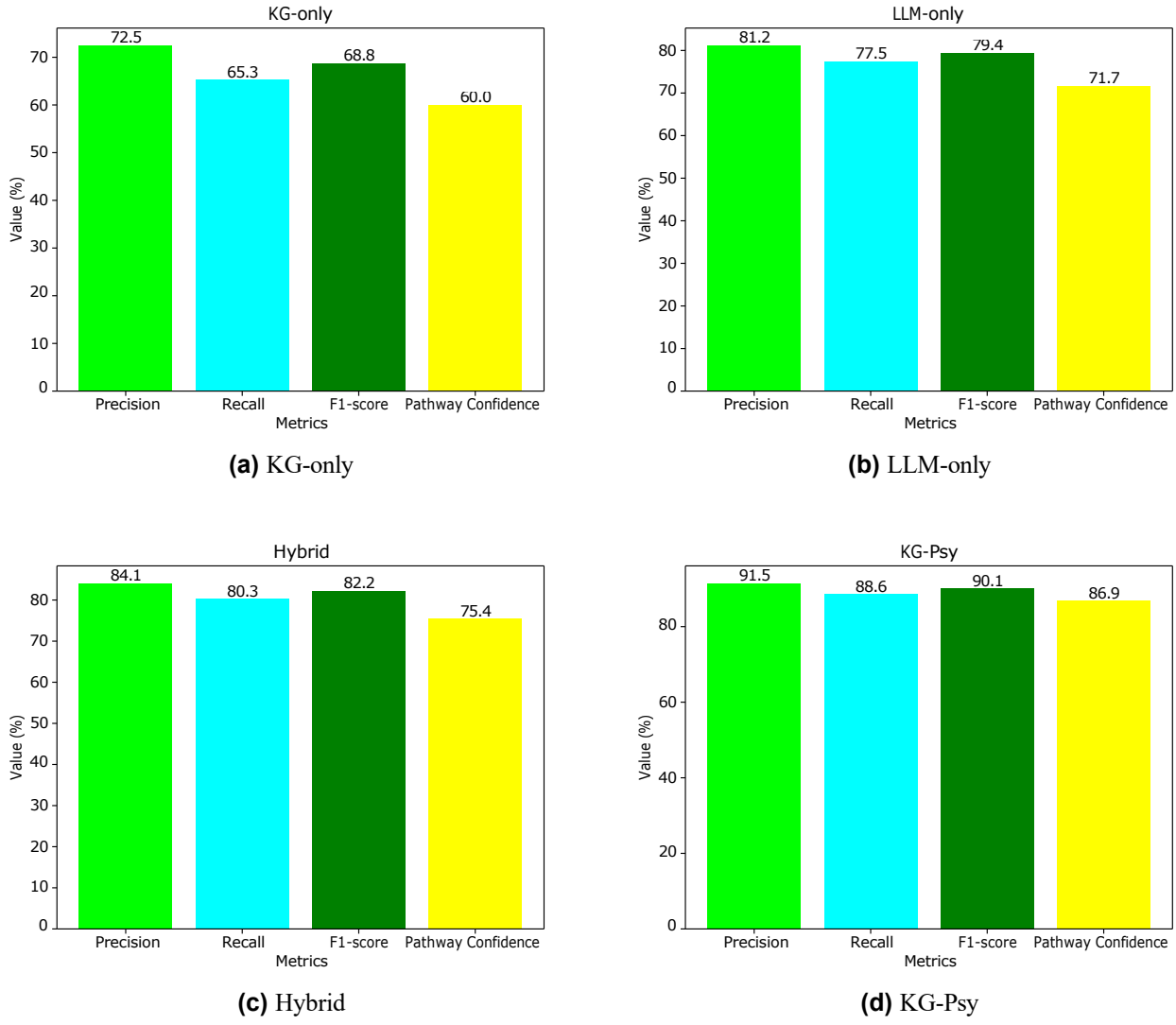
The results of the different approaches and the proposed approach are given in Table 3. For the basic approaches, three approaches were considered: KG-only, LLM-only, and Hybrid-other. The KG-only approach achieved precision (%)=72.5, Recall (%)=65.3, F1-score (%)=68.8, and Pathway confidence (%)=60.0. The LLM-only approach provided better results and was able to achieve precision (%)=81.2, Recall (%)=77.5, F1-score (%)=79.4, and Pathway confidence (%)=71.7. The combination of the two KG and LLM models was also able to achieve a maximum Precision (%)=84.1, which is a significant improvement over both KG and LLM. The other criteria for this approach are Recall (%)=80.3, F1-score (%)=90.1, and Pathway Confidence (%)=86.9, respectively. KG-Psy achieved a maximum Precision (%) of 91.5, an improvement of 19%, 10.3%, and 7.4% over the KG-only, LLM-only, and Hybrid-other approaches, respectively.

4.3 Personalized Treatment Recommendation

KG-Psy also generates treatment suitability scores for each patient by integrating treatment pathway assurance, historical clinical evidence, and literature-based KG support. A sample output for a BD patient is shown in Table 4. The Suitability Score (%) is used in these recommendations to ensure accurate recommendations. These results demonstrate that KG-Psy provides quantitative, interpretable, and clinically relevant recommendations for personalized care.

Table 3. Diagnostic accuracy across 310 cases result

Model	Precision (%)	Recall (%)	F1-score (%)	Pathway Confidence (%)
KG-only	72.5	65.3	68.8	60.0
LLM-only	81.2	77.5	79.4	71.7
Hybrid-other	84.1	80.3	82.2	75.4
KG-Psy	91.5	88.6	90.1	86.9



(a) KG-only **(b)** LLM-only

(c) Hybrid **(d)** KG-Psy

Figure 2. Bar plot for diagnostic accuracy across 310 cases.

Table 4. Personalized treatment recommendation output for a BD patient.

Treatment Option	Suitability Score (%)	Notes
Lithium	92.3	First-line mood stabilizer
Lamotrigine	85.7	Effective for rapid cycling
CBT (Cognitive Therapy)	79.4	Adjunctive therapy
Lifestyle Modulation	74.2	Sleep and activity rhythm regulation

Table 5. Results obtained by considering the components of the proposed framework.

Configuration	F1-score (%)	Pathway Confidence (%)
Full KG-Psy	90.1	86.9
Without GPT-5 Reasoning	83.5	75.2
Without KG Weighting	85.0	77.2
GPT-5 only	79.3	71.7

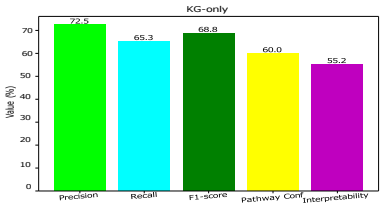
4.4 Ablation Study

For the ablation studies, the contribution of each KG-Psy component was evaluated. Table 5 presents the results obtained by considering the components of the proposed framework. The ablation study confirms that both KG structure and GPT-5 reasoning are critical for achieving high diagnostic alignment and pathway interpretability.

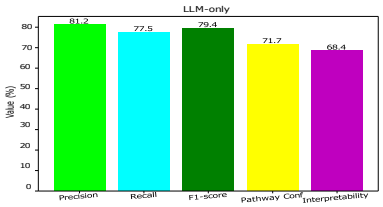
Table 6 shows the 310 diagnostic cases evaluated using language models, ranked by the improvements achieved in the evaluation criteria. Various models available in the literature were also used for evaluation. In addition to the criteria reviewed in this section, we also considered Interpretability. The KG-only approach achieved Interpretability=55.2, and without integrating LLM, it achieved low accuracy because it cannot process complex natural language. The LLM-only approach also achieved Interpretability=68.4. This approach uses only LLMs and is better than KG-only, but lacks an external knowledge structure to improve Interpretability. Combining these two approaches in the Hybrid-other model achieved Interpretability=72.9. The Grok-5 approach using xAI achieved F1=83.5, Pathway Confidence (%)=77.2, and Interpretability=74.1 on this data. The Llama-4 model achieved F1=84.6, Pathway Confidence (%)=78.9, and Interpretability=76.3 in the evaluations. The Claude-4 and Gemini-2 models also achieved an interpretability of 77.8 and 79.1, respectively. The Mistral-Next model produced results close to those of the Gemini-2 model and achieved F1=86.8. The Phi-4 and Falcon-2 models, as well as the Qwen-3 and Vicuna-2 models, yielded evaluation results that were close to each other. Alpaca-Next and the proposed model performed similarly on the Precision criterion, but in other metrics, the proposed approach achieved better results.

4.5 Analysis and Insights

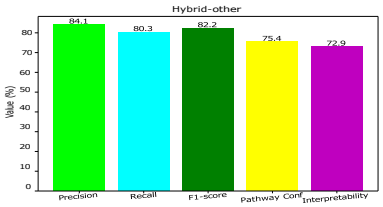
The KG-Psy framework identifies hidden intermediate cognitive pathway nodes that are not represented in existing clinical documents. The framework, leveraging patient-specific disease symptoms and functional data, generates customized treatment solutions that meet established medical standards, based on specific evaluation criteria. The results of the KG-Psy evaluation are numerical data. The researchers can directly measure their results using three data presentation methods: path confidence, F1 scores, and suitability scores. The results maintain their capacity for clinical reproducibility and their ability to support medical interpretation. The architecture delivered superior outcomes compared to single-method baselines because the combination of KG and GPT-5 produced a synergistic effect.



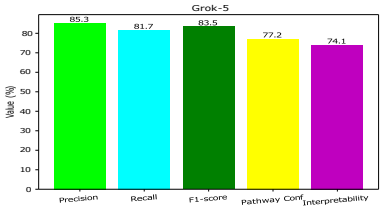
(a) KG-only



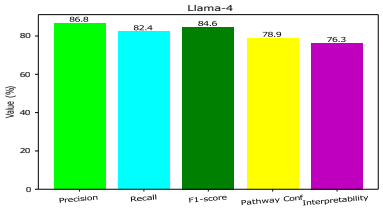
(b) LLM-only



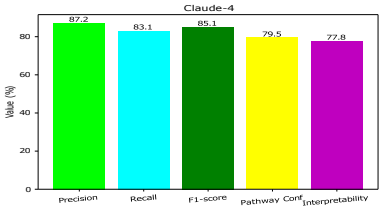
(c) Hybrid



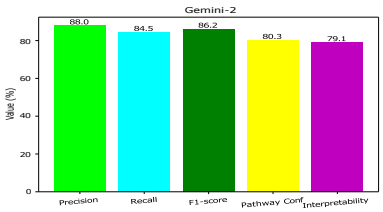
(d) Grok-5



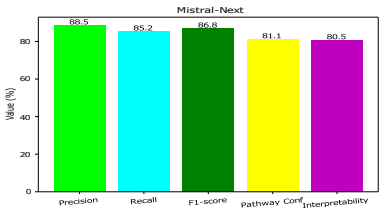
(e) Llama-4



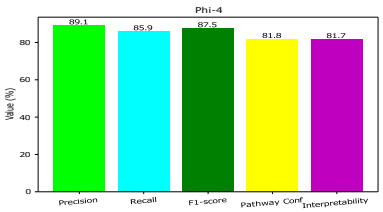
(f) Claude-4



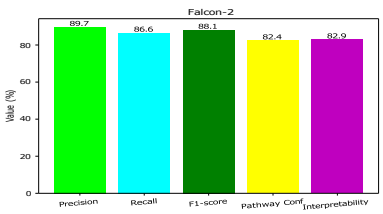
(g) Gemini-2



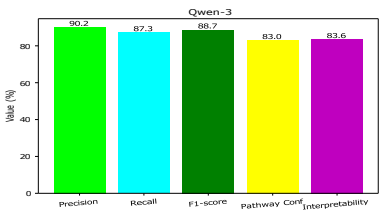
(h) Mistral-Next



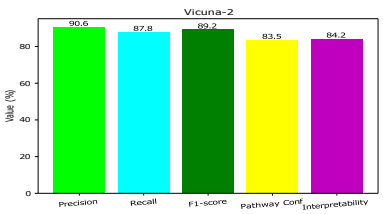
(i) Phi-4



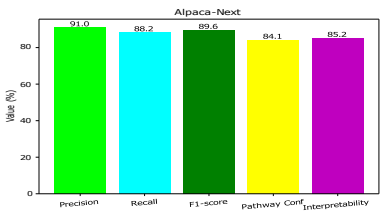
(j) Falcon-2



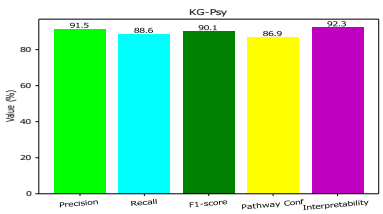
(k) Qwen-3



(l) Vicuna-2



(m) Alpaca-Next



(n) KG-Psy

Figure 3. Bar plot for diagnostic accuracy across 310 cases.

Table 6. Ability study on Interpretability and base line models.

Model	Precision (%)	Recall (%)	F1(%)	Pathway Confidence (%)	Interpretability
KG-only	72.5	65.3	68.8	60.0	55.2
LLM-only	81.2	77.5	79.4	71.7	68.4
Hybrid-other[40]	84.1	80.3	82.2	75.4	72.9
Grok-5	85.3	81.7	83.5	77.2	74.1
Llama-4[41]	86.8	82.4	84.6	78.9	76.3
Claude-4[42]	87.2	83.1	85.1	79.5	77.8
Gemini-2[43]	88.0	84.5	86.2	80.3	79.1
Mistral-Next[44]	88.5	85.2	86.8	81.1	80.5
Phi-4[45]	89.1	85.9	87.5	81.8	81.7
Falcon-2[46]	89.7	86.6	88.1	82.4	82.9
Qwen-3[47]	90.2	87.3	88.7	83.0	83.6
Vicuna-2[48]	90.6	87.8	89.2	83.5	84.2
Alpaca-Next[49]	91.0	88.2	89.6	84.1	85.2
KG-Psy	91.5	88.6	90.1	86.9	92.3

5 Conclusion

The idea of using hybrid approaches of language models and knowledge graphs in the literature has yielded remarkable results. In this study, KG-Psy was introduced. KG-Psy is a hybrid framework that integrates a psychiatric knowledge graph with arguments extracted from the large language model GPT-5 to provide personalized clinical decision support for patients with BD and BPD. In fact, the framework systematically extracts clinical events, neural correlates, and treatment interventions from unstructured case reports in the first two phases, builds evidence-weighted knowledge graphs in the second phase, and finally infers multi-step cognitive pathways using the large language model GPT-5. Different outputs were considered as the final result of this framework. In fact, the resulting cognitive path graphs (CPGs), which include documented and hidden intermediate nodes, provide interpretable and logically plausible representations of cognitive and emotional processes. The evaluations showed that KG-Psy achieved high diagnostic accuracy with an F1 score=90.1% and an average path confidence = 86.9%, outperforming both the LLM-only and KG-only baselines. The ability studies confirmed that combining structured knowledge with GPT-5 reasoning systems establishes essential components for accurate path inference and interpretability. KG-Psy generates numerical treatment recommendations which are tailored to individual patients. The scores demonstrate path confidence, along with clinical evidence and prior understanding, that underpin personalized, evidence-based treatment. KG-Psy bridges the gap between data-driven linguistic model reasoning and dynamic graphs. The framework has the following advantages: accurate differentiation between BD and BPD, reconstruction of cognitive pathways with intermediate nodes, provision of evidence-based, patient-specific treatment recommendations, and a transparent, replicable approach to computational psychiatric research. Like any other computational model, it relies on rich data, which is considered a weakness. Future work could focus on richer datasets. Language models like Grok could also provide better recommendations.

References

1. Mok, Mo Yin, and Yehuda Shoenfeld. "Recent advances and current state of immunotherapy in systemic lupus erythematosus." Expert opinion on biological therapy 16.7 (2016): 927-939.
2. Aaron, Leslie A., and Dedra Buchwald. "A review of the evidence for overlap among unexplained clinical conditions." Annals of internal medicine 134.9_Part_2 (2001): 868-881.
3. Pereira, Ana Catarina, et al. "Inflammation in Bipolar Disorder (BD): Identification of new therapeutic targets." Pharmacological research 163 (2021): 105325.
4. Leichsenring, Falk, et al. "Borderline personality disorder: a review." Jama 329.8 (2023): 670-679.
5. Kang, Rui, et al. "Knowledge graphs for operational decision-making in industrial maintenance: A systematic review." Expert Systems With Applications 299 (2026): 130172.
6. Yang, Bin, et al. "Knowledge graphs in smart construction safety management: A systematic review." Safety Science 197 (2026): 107133.
7. Xie, Zhiwen, Meiling Wang, and Po Hu. "MPRG: A unified framework for knowledge graph reasoning via pattern-aware relation graph." Information Processing Management 63.3 (2026): 104570.
8. Jiang, Xuhui, et al. "On the evolution of knowledge graphs: A survey and perspective." arXiv preprint arXiv:2310.04835 (2023).
9. Singh, Aaditya, et al. "Openai gpt-5 system card." arXiv preprint arXiv:2601.03267 (2025).
10. Turing, Alan M. "Computing machinery and intelligence." Creative Computing 6.1 (1980): 44-53.
11. Neubig, Graham. "Neural machine translation and sequence-to-sequence models: A tutorial." arXiv preprint arXiv:1703.01619 (2017).
12. Emanuel, Aviv, and Eran Eldar. "Emotions as computations." Neuroscience Biobehavioral Reviews 144 (2023): 104977.
13. McGowan, Niall M., et al. "Actigraphic patterns, impulsivity and mood instability in bipolar disorder, borderline personality disorder and healthy controls." Acta Psychiatrica Scandinavica 141.4 (2020): 374-384.
14. Arikan, Mehmet Kemal, et al. "Borderline personality and bipolar disorders cannot be differentiated electrophysiologically." Clinical EEG and Neuroscience 50.6 (2019): 383-388.
15. Freidel, Sebastian, and Emanuel Schwarz. "Knowledge graphs in psychiatric research: Potential applications and future perspectives." Acta Psychiatrica Scandinavica 151.3 (2025): 180-191.
16. Gao, Shan, et al. "Large language model powered knowledge graph construction for mental health exploration." Nature Communications 16.1 (2025): 7526.
17. Wang, Qizheng, et al. "Knowledge graph and its application in the study of neurological and mental disorders." Frontiers in Psychiatry 16 (2025): 1452557.
18. Zhou, Zihan, et al. "Research on the proximity relationships of psychosomatic disease knowledge graph modules extracted by large language models." Scientific Reports 15.1 (2025): 20653.
19. Vidal, Maria-Esther, et al. "Integrating knowledge graphs with symbolic AI: The path to interpretable hybrid AI systems in medicine." Journal of Web Semantics 84 (2025): 100856.
20. Andreas, Triantafyllidis, et al. "Large Language Models for Cardiovascular Disease, Cancer, and Mental Disorders: A Review of Systematic Reviews." Healthcare. Vol. 13. No. 1. MDPI AG, 2025.

21. Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission." arXiv preprint arXiv:1904.05342 (2019).

22. Yang, Xi, et al. "Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records." arXiv preprint arXiv:2203.03540 (2022).

23. Peng, Cheng, et al. "A study of generative large language model for medical research and healthcare." NPJ digital medicine 6.1 (2023): 210.

24. Bosselut, Antoine, et al. "MEDITRON: Open medical foundation models adapted for clinical practice." (2024).

25. Wang, Guangyu, et al. "ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation." arXiv preprint arXiv:2306.09968 (2023).

26. Naseem, Usman, et al. "Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding." 2020 International joint conference on neural networks (IJCNN). IEEE, 2020.

27. Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

28. Wada, Shoya, et al. "Pre-training technique to localize medical bert and enhance biomedical bert." arXiv preprint arXiv:2005.07202 (2020).

29. Shin, Hoo-Chang, et al. "BioMegatron: larger biomedical domain language model." Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020.

30. Lowe, Juliette. "BioLinkBERT-PR: A Large Language Model for Diagnosing Pediatric Rheumatological Disorders." (2025).

31. Luo, Renqian, et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." Briefings in bioinformatics 23.6 (2022): bbac409.

32. Luo, Yizhen, et al. "Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine." arXiv preprint arXiv:2308.09442 (2023).

33. Luu, Rachel K., and Markus J. Buehler. "BioinspiredLLM: Conversational large language model for the mechanics of biological and bioinspired materials." Advanced Science 11.10 (2024): 2306724.

34. Labrak, Yanis, et al. "Biomistral: A collection of open-source pretrained large language models for medical domains." Findings of the association for computational linguistics: acl 2024. 2024.

35. Sun, Maoqian, et al. "Large Language Models for Psychiatric Diagnosis Based on Multicenter Real-World Clinical Records: Comparative Study." JMIR Medical Informatics 14 (2026): e77699.

36. Chien, Chien Wen, and Yueh-Ming Tai. "Performances of large language models in detecting psychiatric diagnoses from Chinese electronic medical records: comparisons between GPT-3.5, GPT-4, and GPT-4o." Taiwanese Journal of Psychiatry 38.3 (2024): 134-141.

37. Shewcraft, Ryan Allen, John Schwarz, and Mariann Micsinai Balan. "Algorithmic Classification of Psychiatric DisorderRelated Spontaneous Communication Using Large Language Model Embeddings: Algorithm Development and Validation." JMIR AI 4 (2025): e67369.

38. Omar, Mahmud, et al. "Applications of large language models in psychiatry: a systematic review." Frontiers in psychiatry 15 (2024): 1422807.

39. Park, Chaelim, Hayoung Lee, and Ok-ran Jeong. "Leveraging medical knowledge graphs and large language models for enhanced mental disorder information extraction." Future Internet 16.8 (2024): 260.

40. Gao, Yanjun, et al. "Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study." Jmir Ai 4 (2025): e58670.

41. Adcock, Aaron, et al. "The Llama 4 Herd: Architecture, Training, Evaluation, and Deployment Notes." arXiv preprint arXiv:2601.11659 (2026).

42. Nguyen, Cindy, Daniel Carrion, and Mohamed K. Badawy. "Comparative performance of anthropic Claude and openai GPT models in basic radiological imaging tasks." Journal of medical imaging and radiation oncology 69.4 (2025): 431-439.

43. Comanici, Gheorghe, et al. "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities." arXiv preprint arXiv:2507.06261 (2025).

44. Sharma, Aanchal, et al. "Fine-Tuning Mistral 7B for Enhanced." Communication and Intelligent Systems: Proceedings of ICCIS 2024, Volume 4 4 (2025): 341.

45. Abdin, Marah, et al. "Phi-4 technical report." arXiv preprint arXiv:2412.08905 (2024).

46. Malartic, Quentin, et al. "Falcon2-11b technical report." arXiv preprint arXiv:2407.14885 (2024).

47. Tigani, Smail. "CPU-Only Self Enhancing Authoring Copilot Design-Based Markov Decision Processes Orchestration and Qwen 3 Local Large Language Model." Technologies 13.11 (2025): 520.

48. Ji, Bin. "Vicunaner: Zero/few-shot named entity recognition using vicuna." arXiv preprint arXiv:2305.03253 (2023).

49. Chen, Lichang, et al. "Alpagasus: Training a better alpaca with fewer data." arXiv preprint arXiv:2307.08701 (2023).

Author contributions

All authors contribute the writing the manuscript, concepts, methodology, design of the proposed models, experimental results analysis, the software of the proposed models, dataset collection, resources, visualization, implementation of the models, similarity reduction, and the editing of the manuscript, the review of the writing and grammatical errors for the manuscript, validation of the results, and supervision of the proposed work.

Funding

No funding

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.