Large Language Models for Interpretable Mental Health Diagnosis

Brian Hyeongseok Kim and Chao Wang
Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, USA

1. Introduction

We propose a clinical decision support system (CDSS) for mental health diagnosis that combines the strengths of large language models (LLMs) and constraint logic programming (CLP). Our method leverages LLMs to translate natural language descriptions from diagnostic manuals into logic programs with interpretable rules and objectives and then solves them using a CLP engine to ensure that the diagnostic output is verifiably correct for the given patient data. We present our results here and discuss the applicability of LLMs for mental health diagnosis at large.

2. Background

Psychological Diagnosis

- Process by which clinicians assess if a patient's symptoms meet the criteria for mental disorders
- Specifications outlined in diagnostic manuals (e.g., DSM-5-TR by APA [1], ICD-11 CDDR by WHO [2])
- 1,000+ page manuals → complexity increases the risk of diagnostic errors [3]

At least two of the following symptoms must be present most of the time for a period of 1 month or more. At least one of the qualifying symptoms should be from items (a) to (d) below: [List of symptoms, omitted for brevity].

ICD-11 CDDR diagnostic criteria for schizophrenia.

Constraint Logic Programming (CLP)

- Programming paradigm using logical rules of desired computations, rather than the actual implementation
- *What* should be computed, rather than *how*
- Enables verifying correctness & logical soundness
- In our work: Datalog with Soufflé solver engine [4]

```
.decl Edge(x:number, y:number)           Input:    Edge(1, 2).
.decl Path(x:number, y:number)                     Edge(2, 3).
.input Edge
.output Path                             Output:   Path(1, 2).
Path(x, y) :- Edge(x, y).                          Path(2, 3).
Path(x, y) :- Path(x, z), Edge(z, y).              Path(1, 3).
```

Listing 1: Example logic program expressed in Datalog (left) and its corresponding input-output example (right).

[1] American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. American Psychiatric Association Publishing. ISBN 9780890425763.

[2] World Health Organization. 2024. *Clinical descriptions and diagnostic requirements for ICD-11 mental, behavioural and neurodevelopmental disorders*. World Health Organization. ISBN 9789240077263.

[3] American Psychological Association. 2023. Psychologists reaching their limits as patients present with worsening symptoms year after year: 2023 Practitioner Pulse Survey.

[4] Jordan, H.; Scholz, B.; and Subotić, P. 2016. Soufflé: On Synthesis of Program Analyzers. In *Computer Aided Verification*, 422–430. Springer International Publishing.

3. Methodology

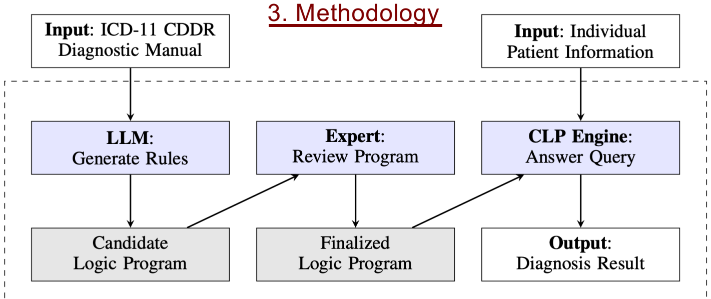

Figure 1: Our proposed CDSS pipeline that combines LLM and CLP.

Patient Information

```
.decl Observed(Patient:symbol,
               Symptom:symbol, Week:float)
.decl History(Patient:symbol,
               Condition:symbol, Count:number)
.decl Diagnosis(Patient:symbol, Disorder:symbol)
.input Observed
.input History
.output Diagnosis
```

Diagnostic Rules

```
AllPatients(P)   :- Observed(P, _, _).
Core(P, S, W)    :- Observed(P, S, W), (S = "SymptomA"; S = "SymptomB"), Week>=2.
Qual(P, S, W)    :- Observed(P, S, W), (S = "SymptomC"; S = "SymptomD"), Week>=2.
CoreCount(P, count:Core(P, _, _)) :- Core(P, _, _).
CoreCount(P, 0)                   :- !Core(P, _, _), AllPatients(P).
QualCount(P, count:Qual(P, _, _)) :- Qual(P, _, _).
QualCount(P, 0)                   :- !Qual(P, _, _), AllPatients(P).
TotalCount(P, CC + QC)            :- CoreCount(P, CC), QualCount(P, QC).
Diagnosis(P, "DisorderD")         :- CoreCount(P, CC), TotalCount(P, TC),  History(P, "ConditionC", HC),
                                     CC>=1, TC>=2, HC>=1.
```

Listing 2: Example logic program for encoding diagnosis in Datalog. Soufflé takes this program and ***Observed / History*** *patient data as input and return* ***Diagnosis*** *as output.*

Core symptom = must be present
Qualifying (Qual) symptom = can be present

System: *You are an expert at translating mental health diagnostic criteria into a Datalog program in Soufflé.*
Prompt: *The patient data is given as input to the program as Observed and History relations. The patient diagnosis is returned as output from the program as Diagnosis relation. [Explain the relations.]*
Example: *[Include an ICD-11 CDDR diagnostic criteria for a disorder and its corresponding Datalog program.]*
Task: *Translate the given criteria into a Datalog program using Soufflé syntax. [Include relevant Observed symptom names, History condition names, and the ICD-11 CDDR diagnostic criteria for each disorder.]*

Our LLM prompt template for translating diagnostic criteria into logic programs.

5. Results

Table 1: Our method compared against two baselines on the first 10 (out of 30) patients.

Patient ID	Known Disorder	*Diagnosis by LLM-only Approach*			*Diagnosis using LLM + Datalog*			*Diagnosis by Our CDSS*
		Llama	**Gemini**	**GPT**	**Llama**	**Gemini**	**GPT**	**GPT**
No. 1	BPD2	BPD2	BPD1	BPD2	-	-	BPD2	BPD2
No. 2	RDD	SEDD	SEDD	SEDD	BPD1	SEDD	SEDD	RDD
No. 3	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1, BPD2	BPD1	BPD1
No. 4	BPD2	SEDD	BPD2	BPD2	BPD1	SEDD	-	BPD2
No. 5	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1, BPD2	-	BPD1
No. 6	BPD2	BPD2	BPD2	BPD2	BPD1	SEDD	BPD2	BPD2
No. 7	BPD1	-	BPD1	BPD1	-	BPD1	BPD1	BPD1
No. 8	SEDD	SEDD	SEDD	SEDD	BPD1	-	SEDD	SEDD
No. 9	SEDD	SEDD	SEDD	SEDD	BPD1	-	SEDD	SEDD
No. 10	-	-	-	-	-	-	-	-
Correct Diagnosis (Total):		7/10	8/10	9/10	3/10	(2+2)/10	7/10	10/10

RQ1: Accuracy for *LLM + Datalog*
- GPT (best): 7/10 → 22/30 correct

RQ2: Quality of LLM-generated Datalog programs
- GPT interprets the text literally and relies only on History
- Gemini ignores History, outputs conflicting diagnoses
- Llama doesn't distinguish Core and Qual symptoms

RQ3: LoC changes from *LLM + Datalog* to *Our CDSS*
- 57 added (+), 10 removed (-) from the initial 107 LoC
- Fix cyclic dependencies and clinical inconsistencies

RQ4: Accuracy for *LLM-only*
- GPT (best): 9/10 → 22/30 correct
- No guarantee / transparency in probabilistic predictions

Ethical Statement: The proposed CDSS aims at helping clinical professionals in decision-making. It is not meant to replace or refute the diagnoses provided by qualified clinicians.

4. Evaluation

Research Questions

- **RQ1**. How accurate are the diagnostic outputs generated by the LLM-translated programs?
- **RQ2**. To what extent can LLMs accurately interpret and translate diagnostic criteria into Datalog?
- **RQ3**. How much additional human effort is required to correct errors in the LLM-translated programs?
- **RQ4**. How effective are LLMs in diagnosing a patient when given their data directly?

Experimental Setup

- 4 mood disorders: 1) Bipolar I (**BPD1**), 2) Bipolar II (**BPD2**), 3) Single Episode Depressive Disorder (**SEDD**), 4) Recurrent Depressive Disorder (**RDD**).
- 30 patients: 9 with BPD1, 8 with BPD2, 5 with SEDD, 4 with RDD, 4 undiagnosed (do not meet the criteria)
- 3 LLMs: 1) Meta's **Llama**-3.2, 2) Google's **Gemini**-1.5-Flash, 3) OpenAI's **GPT**-4o.

Approaches

1. *LLM-only*: directly provide a diagnosis
2. *LLM + Datalog*: translate criteria into Datalog programs
3. *Our CDSS*: experts correct errors in LLM-translated Datalog programs

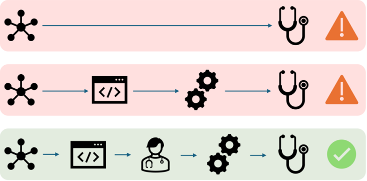

Figure 2: Comparison of the two baseline approaches (LLM-only and LLM + Datalog in Rows 1 and 2) with Our CDSS (Row 3).

6. Conclusion

Our method utilizes LLMs to generate logic programs that encode psychological diagnostic rules, and CLP engines to produce diagnostic results based on patient data. We propose that this hybrid approach, combined with expert validation, ensures that diagnostic reasoning is aligned with clinical criteria, enhancing reliability and safety in clinical decision-making for mental health diagnosis.

Paper

Contact

Page: https://briankim113.github.io/
Email: brian.hs.kim@usc.edu