Large Language Models for Interpretable Mental Health Diagnosis

Brian Hyeongseok Kim, Chao Wang

University of Southern California, Los Angeles, USA
{brian.hs.kim, wang626}@usc.edu

Abstract

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). Having a CDSS is important because of the high complexity of diagnostic manuals used by mental health professionals and the danger of diagnostic errors. Our CDSS is a software tool that uses an LLM to translate diagnostic manuals to a logic program and solves the program using an off-the-shelf CLP engine to query a patient's diagnosis based on the encoded rules and provided data. By giving domain experts the opportunity to inspect the LLM-generated logic program, and making modifications when needed, our CDSS ensures that the diagnosis is not only accurate but also interpretable. We experimentally compare it with two baseline approaches of using LLMs: diagnosing patients using the LLM-only approach, and using the LLM-generated logic program but without expert inspection. The results show that, while LLMs are extremely useful in generating *candidate* logic programs, these programs still require expert inspection and modification to guarantee faithfulness to the official diagnostic manuals. Additionally, ethical concerns arise from the direct use of patient data in LLMs, underscoring the need for a safer hybrid approach like our proposed method.

1 Introduction

Mental disorders impose a significant burden on the affected individuals and their communities (Gorvin and Brown 2012). Accurate diagnosis is a critical first step toward improving patient outcomes and fostering societal well-being. In clinical settings, the diagnostic process relies on matching a patient's symptoms with the mental health diagnostic rules outlined in official manuals such as DSM-5-TR (American Psychiatric Association 2022) and ICD-11 CDDR (World Health Organization 2024). These manuals, consisting of more than 1,000 pages of natural language descriptions, serve as authoritative references for not only mental health professionals but also insurance companies. However, their complexity poses a significant challenge, not only exacerbating the workload of already overburdened mental health professionals but also increasing the risk of diagnostic errors (American Psychological Association 2023). This underscores the pressing need for developing a robust clinical decision support systems (CDSS), a software tool that can verify the diagnosis made manually. Yet, such tools remain underdeveloped, particularly those that address issues of reliability and interpretability.

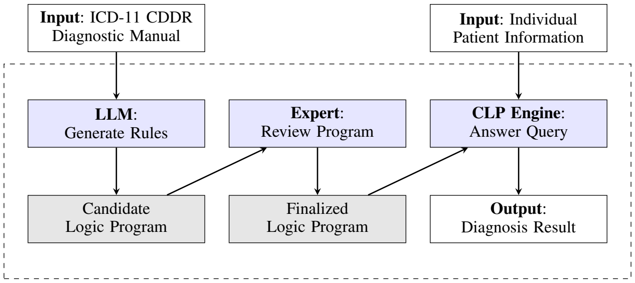
Input: ICD-11 CDDR Diagnostic Manual
Input: Individual Patient Information
LLM: Generate Rules
Expert: Review Program
CLP Engine: Answer Query
Candidate Logic Program
Finalized Logic Program
Output: Diagnosis Result

Figure 1: Clinical decision support system (CDSS) combining the strengths of LLM and constraint logic programming.

Recent advancements in large language models (LLMs) suggest their potential in many applications (Friha et al. 2024) including clinical settings (Ullah et al. 2024). Thanks to their excellent processing and understanding of natural language, LLMs can generate diagnostic suggestions based on medical literature and patient data. However, their adoption in clinical settings still faces challenges. For example, LLMs are prone to issues like hallucinations (Huang et al. 2023; Bai et al. 2024), lack of explainability (Zhao et al. 2023) and consistency (Moore, Deshpande, and Yang 2024), and limited proficiency in complex reasoning (Huang and Chang 2023). To date, no existing approach effectively combines LLMs with mechanisms that *guarantee* accuracy and interpretability in the context of mental health diagnosis.

To fill this critical gap, we propose a method that combines LLMs with constraint logic programming (CLP), leading to a practical tool for assisting clinicians in making mental health diagnosis. Specifically, our method leverages LLMs to translate natural language descriptions of mental health diagnostic criteria from manuals such as DSM-5-TR and ICD-11 CDDR to logic rules, thus reducing the cognitive burden on domain experts. Simultaneously, we use an off-the-shelf CLP engine for solving the logic rules to ensure that the diagnostic output is verifiably correct, while enhancing interpretability through the rules and objectives explicitly defined via CLP.

Figure 1 shows the overall flow of our method. First, natural language text from a diagnostic manual (e.g., ICD-11 CDDR) is fed into an LLM, which generates a *candidate* logic program codifying the diagnostic rules. Next, a domain expert manually reviews the code to ensure that the LLM-generated rules accurately encode the manual's criteria. Finally, the *finalized* logic program is used by a CLP engine to generate the diagnosis result based on the information of an individual patient. By combining the natural language processing capabilities of an LLM with the logical reasoning capabilities of a CLP engine, our method delivers accurate and inherently interpretable diagnostic outcomes.

The rest of this paper is organized as follows: Section 2 provides the background needed to contextualize our approach. Section 3 presents our methodology. Section 4 presents the experimental results and analysis of our findings. Section 5 reviews the related work. Finally, Section 6 concludes with a summary of our contributions.

2 Background

In this section, we review the background information of psychological diagnosis and constraint logic programming.

2.1 Psychological Diagnosis

Psychological diagnosis is the process by which clinicians assess if a patient's symptoms meet the criteria for specific disorders as outlined in the diagnostic manuals such as DSM-5-TR and ICD-11 CDDR. These authoritative manuals are widely adopted, underscoring their global relevance and importance. As an example, consider the following ICD-11 CDDR diagnostic criteria for *schizophrenia*:

> *At least two of the following symptoms must be present (by the individual's report or through observation by the clinician or other informants) 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 from (a) to (g)... (omitted for brevity)].*

Clinical decision support systems (CDSS) are a specific type of DSS (Keen and Scott Morton 1978) where patient data and medical knowledge are integrated to the software tool, to assist clinicians with decision-making. CDSS can address various scenarios such as offering diagnostic support, identifying drug interactions, and predicting treatment outcomes (Berner 2007). The focus of this work is offering diagnostic support in the context of mental disorders. Note that CDSS is merely a support system, as the name implies; it aims at helping clinical professionals in decision-making instead of replacing their decision-making role entirely.

2.2 Constraint Logic Programming

Constraint logic programming (CLP) is a paradigm that focuses on expressing logical rules of desired computations as opposed to implementing these computations. It is well-suited for applications where accuracy and transparency are critical, as it focuses more on *what* should be computed rather than *how* to compute it, thus enabling easier verification of correctness and logical soundness. In our work, we use Datalog as the CLP language, and solve Datalog programs using Soufflé, a state-of-the-art Datalog engine (Jordan, Scholz, and Subotić 2016; Scholz et al. 2016).

Listing 1: An example logic program expressed in Datalog.

```
1  .decl Edge(x:number, y:number)
2  .decl Path(x:number, y:number)
3  .input Edge
4  .output Path
5  Path(x, y) :- Edge(x, y).
6  Path(x, y) :- Path(x, z), Edge(z, y).
```

Listing 1 shows an example logic program that codifies the rules that infer the `Path` relation from the `Edge` relation. It starts with declarations of the two relations (Lines 1-2). Then, it specifies the input and the output (Lines 3-4). Finally, it defines the *rules* for inferring `Path` from `Edge` (Lines 5-6). Specifically, Line 5 means that `Path(x,y)` holds if `Edge(x,y)` holds, and Line 6 means that `Path(x,y)` holds if both `Path(x,z)` and `Edge(z,y)` hold. The comma (,) in Line 6 denotes logical AND, whereas a semicolon (;) denotes logical OR.

Given a set of *facts*, e.g., `Edge` from 1 to 2 and `Edge` from 2 to 3, the program in Listing 1 computes all entries of the `Path` relation: from 1 to 2, from 2 to 3, and from 1 to 3. This is how the program can answer queries, e.g., whether `Path(1,3)` holds. Similarly, we want to use Datalog to express ICD-11 CDDR diagnostic rules, and then answer queries for individual patients. This leads to a verifiably correct and explainable CDSS for mental disorders.

3 Methodology

In this section, we present our Datalog encoding of diagnostic rules and LLM-based translation of text to rules.

3.1 Datalog Encoding of the Diagnosis

We focus on ICD-11 CDDR diagnostic rules, but this can be done similarly for DSM-5-TR.

The Diagnostic Rules Listing 2 shows a Datalog program with *rules* that connect a patient's symptoms and past conditions to a mental disorder. Lines 4-6 specify the input and output relations. For brevity, we omit the declarations of intermediate relations, but they also require the `.decl` keyword with variable types, similar to Lines 1-3. The program first extracts the patient's name from `Observed` and add it to a relation called `AllPatients` (Line 7), and then identifies which symptoms are core (must be present) or qualifying (can be present) according to the diagnostic criteria. Given a set of symptoms A, B, C, and D, for example, Symptoms A and B may be considered core whereas Symptoms C and D may be considered qualifying, and they must have been observed for more than 2 weeks (Lines 8-9).

The program has rules that count the number of symptoms in each category, which requires an aggregate function called `count` (Lines 10 and 12). If `Core` or `Qual` relations do not exist, the count is set to 0 (Lines 11 and 13). Once we have the counts for the symptoms, we add them up (Line 14). Finally, we decide if a patient should be given the

Listing 2: An example logic program for encoding ICD-11 CDDR diagnostic rules in Datalog.

```
1   .decl Observed(Patient:symbol, Symptom:symbol, Week:float)
2   .decl History(Patient:symbol, Condition:symbol, Count:number)
3   .decl Diagnosis(Patient:symbol, Disorder:symbol)
4   .input Observed
5   .input History
6   .output Diagnosis
7   AllPatients(P) :- Observed(P, _, _).
8   Core(P, S, W) :- Observed(P, S, W), (S = "SymptomA"; S = "SymptomB"), Week>=2.
9   Qual(P, S, W) :- Observed(P, S, W), (S = "SymptomC"; S = "SymptomD"), Week>=2.
10  CoreCount(P, count:Core(P, _, _)) :- Core(P, _, _).
11  CoreCount(P, 0) :-  !Core(P, _, _), AllPatients(P).
12  QualCount(P, count:Qual(P, _, _)) :- Qual(P, _, _).
13  QualCount(P, 0) :-  !Qual(P, _, _), AllPatients(P).
14  TotalCount(P, CC + QC) :- CoreCount(P, CC), QualCount(P, QC).
15  Diagnosis(P, "DisorderD") :- CoreCount(P, CC), TotalCount(P, TC), History(P, "ConditionC", HC),  CC>=1, TC>=2, HC>=1.
```

`Diagnosis` of "`DisorderD`" (Line 15). Here, the diagnosis requires at least 1 core symptom and at least 2 symptoms in total (i.e., one core and one qualifying, or two core symptoms), and at least one occurrence of "`ConditionC`" in prior history.

Patient Information The Datalog program in Listing 2 requires *facts* that describe the patient as input. These facts are expressed using the `Observed` and `History` relations. `Observed` indicates that `Patient` is experiencing `Symptom` for the duration of `Week` (Line 1). `History` indicates that `Patient` has a history of `Condition` for the `Count`[1] number of times (Line 2).

Consider the follwing example of "PatientA", who has been observed with "SymptomA" and "SymptomB" for 3.5 weeks, and has a prior history of "ConditionC" two times. The corresponding input facts are as follows:

- `Observed`("PatientA", "SymptomA", 3.5)
- `Observed`("PatientA", "SymptomB", 3.5)
- `History`("PatientA", "ConditionC", 2)

They meet the diagnostic criteria for "DisorderD" as shown in Line 15 of Listing 2.

3.2 LLM-Based Translation of Manuals to Rules

We prompt LLMs to translate the text-based diagnostic criteria from the ICD-11 CDDR manual into *candidate* logic programs in Datalog, similar to the program shown in Listing 2. Then, we assess whether the LLM-generated logic program can diagnose a given patient correctly. In-context learning (ICL) (Brown et al. 2020; Dong et al. 2022) allows LLMs to perform tasks better without explicitly updating the model parameters. As part of ICL, we provide an example of diagnostic criteria text from ICD-11 CDDR and its corresponding Datalog program, such that the models can learn from the demonstrated task. Our one-shot prompt template is as follows:

> ***System****: You are an expert at translating mental health diagnostic criteria into a Datalog program in Soufflé. 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 the 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.*

Since the generated programs are declarative, and they are driven by logic, the diagnoses that they provide are guaranteed to be correct, as long as the rules reflect the logic of the diagnostic manual accurately. While LLMs may produce *candidate* logic programs that contain syntactic and/or semantic errors, these logic programs may be reviewed and corrected by a domain expert. At the level of Datalog programs, expert intervention is feasible and sufficient for ensuring that the *finalized* logic programs not only can be compiled, but also accurately represent the diagnostic criteria. Furthermore, manual inspection is reasonable in this context, given the critical role of human oversight in clinical applications. We refer the readers to Appendix C for the detailed prompts used in our experimental evaluation.

4 Evaluation

We use ICD-11 CDDR diagnostic manual (World Health Organization 2024) and focus on four mood disorders: Bipolar I (BPD1), Bipolar II (BPD2), Single Episode Depressive Disorder (SEDD), and Recurrent Depressive Disorder (RDD). From natural language descriptions of the diagnostic criteria, our method generates the *candidate* Datalog program as described in Section 3.1. Then, we manually inspect the LLM-generated Datalog program and correct errors to ensure that the *finalized* Datalog program accurately encodes the diagnostic rules.

We also manually validated the diagnosis results of the Datalog program. Given a dataset of 30 patients, the finalized Datalog program identified 9 patients with BPD1, 8 with BPD2, 5 with SEDD, and 4 with RDD. Four patients remained undiagnosed, as they did not meet the criteria for any of the considered mood disorders. We validated the pro-

[1] Note that the lowercase `count` refers to the aggregate function, while the uppercase `Count` refers to the variable name.

gram's results by cross-checking the patient data against the diagnostic criteria specified in the manual. Details of all 30 patients can be found in Appendix B.

4.1 The Experimental Setup

Given the time and effort required for manual rule translation and diagnosis validation, we aim to assess whether state-of-the-art LLMs can automate this process without compromising accuracy. Our main research questions are:

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 from text 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?

We used 3 state-of-the-art LLMs. **GPT** stands for GPT-4O on OpenAI, released in May 2024 (OpenAI 2024).[2] **Gemini** stands for GEMINI-1.5-FLASH on Google Cloud, released in May 2024 (Gemini Team, Google 2024)[3]. **Llama** stands for LLAMA-3.2 on Meta AI, released in September 2024 (Llama Team, AI @ Meta 2024).[4] These LLMs were accessed between October and November 2024.

4.2 Two Baseline Approaches

To evaluate our method, we developed two groups of baselines for comparison. The first group of baselines (*LLM-only*) involves directly providing a diagnosis given the patient data. This is analogous to using LLMs as an external consultant that can either validate or challenge a clinician's diagnosis (Wang et al. 2024). For this task, we use the following prompt template:

> ***System**: You are an expert at diagnosing patients according to the ICD-11 CDDR. The considered disorders are [List of mood disorders]. The patient data is given as* `Observed` *and* `History` *relations. Explain the relations.*
>
> ***Task**: Please output the diagnosis for the following patients. Patients with no clear diagnosis should be indicated as such. Include patient data.*

The second group of baselines (*LLM + Datalog*) is running the LLM-translated Datalog programs without expert intervention. This approach is comparable to testing LLM-generated code in imperative programming (Jiang et al. 2024). These baselines follow the same prompt structure described in Section 3.2. Our method extends these baselines by incorporating expert corrections to address syntactic and semantic errors in the LLM-generated code. We refer to these expert-corrected programs as *Our CDSS*.

4.3 Experimental Results

Table 1 compares the performance of our method against the baselines across 10 patients. The remainder of this section will discuss the results for the first 10 patients. Results for all 30 patients can be found in Appendix A.

Columns 1-2 list patient numbers and their disorders based on our manually written Datalog program, validated against the ICD-11 CDDR criteria. Columns 3-5 (*LLM-only*) show diagnoses directly provided by the LLMs, while Columns 6-8 (*LLM + Datalog*) show diagnoses from LLM-generated Datalog programs without expert intervention. Column 9 (*Our CDSS*) shows diagnoses from an expert-corrected LLM-generated program. Green cells indicate correct diagnoses, yellow cells indicate partial correctness (correct diagnosis with additional incorrect ones), and the final row summarizes correct diagnoses per method.

Answer to RQ1 To address RQ1 on the accuracy of LLM-translated programs, we look at *LLM + Datalog* columns for the as-is versions[5] and the *Our CDSS* column for the expert-corrected version. Among the as-is programs, GPT performs the best with 7 correct diagnoses out of 10, followed by Gemini with 2 correct and 2 partially correct, and Llama with 3 correct. This pattern holds across all the patients, where GPT achieves 22 correct diagnoses out of 30, while Gemini has 8 correct and 4 partially correct, and Llama only 9 correct. We extend the most accurate program generated by GPT and implement logical changes to align with the ICD-11 CDDR criteria for the mood disorders. The expert-reviewed program, shown in Column 9, produces 10 correct diagnoses out of 10 (30 out of 30).

Answer to RQ2 To address RQ2 on the performance of LLMs in translating diagnostic criteria into Datalog programs, we take a closer look at the programs that correspond to Columns 6-8.

Although GPT-generated program achieves the greatest number of correct diagnoses, it relies solely on `History` in its final diagnostic rules, despite constructing intermediate rules to identify mood episodes based on `Observed`. This issue arises from the diagnostic text's phrasing, which specifies a "history of" certain mood episodes as a requirement for a particular mood disorder. While clinicians would intuitively consider current symptoms for diagnosis, GPT interprets the text literally, lacking the nuanced understanding needed for accurate clinical interpretation.

The Gemini-generated code presents the opposite issue, where it only considers current symptoms, ignoring the patient's history. This leads to missed diagnoses such as Patient 1, where the current symptoms suggest no diagnosis (denoted '-' under Column 7), despite the patient's history indicating BPD2. This may stem from using a Datalog program for schizophrenia in the prompt, which doesn't incorporate `History`, which is specific to mood disorders. Additionally, Gemini-generated program frequently diagnoses conflicting disorders (e.g., "BPD1, BPD2" for Patients 3 and 5), which contradicts the diagnostic criteria that require

[2] https://chatgpt.com

[3] https://gemini.google.com

[4] https://www.meta.ai

[5] The GPT-generated program compiled and ran as-is, while programs from Gemini and Llama required minor syntactic fixes for compilation. For Gemini, 17 lines were added, and 4 removed from an initial 34 lines, resuling in 47 lines of code (LoC). For Llama, 19 lines were added, and 11 removed from an initial 114 lines, resulting in 122 LoC. No semantic changes were made.

Table 1: Comparing our method with baselines on the first 10 (out of 30) patients. 'Known Disorder' indicates what the patient is diagnosed with according to the ICD-11 CDDR criteria. 'LLM-only' indicates the diagnosis directly produced by LLMs. 'LLM + Datalog' indicates the diagnosis produced by the LLM-generated Datalog program. 'Our CDSS' indicates the diagnosis produced by our method. The symbol '-' indicates no clear diagnosis.

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

mutually exclusive conditions. BPD1 requires a mixed or manic episode; BPD2 explicitly requires the absence of such episodes.

The Llama-generated code often misdiagnoses patients as BPD1 (e.g., Patients 2, 4, 6, 8, and 9). This stems from a lack of intermediary logic to properly identify mood episodes. The generated program counts associated symptoms without distinguishing core or qualifying symptoms and uses an arbitrary threshold that doesn't align with the ICD-11 criteria, leading to frequent BPD1 diagnoses, as it has the lowest threshold.

Overall, the inconsistency across models suggests that while LLM-generated code shows promise, it is not yet reliable for direct clinical use. Models could benefit from more text-to-rule translation examples for ICL to generate more accurate programs. Additionally, breaking down the task into smaller steps through multi-turn conversation (Zheng et al. 2024) or Chain-of-Thought (CoT) prompting (Wei et al. 2022) could enhance their logical reasoning. Models could be fine-tuned for this task by experts with reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022). Finally, using LLMs optimized for code generation (e.g., GitHub Copilot[6] and Amazon Q Developer[7]) could improve the performance.

Answer to RQ3. As discussed, LLM-generated programs do not guarantee that the encoded logic accurately replicates the diagnostic criteria of ICD-11 CDDR. To address this, our method proposes a pipeline where LLMs generate candidate Datalog programs, which are then reviewed and refined by experts. This approach aims to balance the efficiency of AI with the crucial need for diagnostic accuracy. In this context, it is important to address RQ3, which examines the additional human effort required to accurately represent the logic of the diagnostic criteria.

Listing 3 highlights some of the changes made to the GPT-generated code, addressing two major issues. First, the original definition of `MixedEpisode` (Line 2) created a cyclic dependency by requiring both `ManicEpisode` and `DepressiveEpisode` for its definition. However, according to the ICD-11 CDDR manual, a mixed episode should be defined independently of these episodes and based on specific symptom thresholds. Furthermore, the logic expressed in the program contradicted clinical guidelines, since depressive or manic episodes should not apply if the symptoms qualify better as a mixed episode. The revised code resolves this by incorporating the absence of a mixed episode as part of criteria for other mood episodes (Line 1) and redefining `MixedEpisode` to directly evaluate symptom counts and core criteria (Line 2). This approach eliminates the cyclic dependency and ensures compliance with clinical guidelines. In order to achieve this, we manually added several missing intermediate relations and rules to accurately identify and count core and qualifying symptoms for different mood episodes.

Second, the original logic for diagnosing disorders relied solely on `History` (Lines 3-4). As discussed, this approach neglected the possibility of diagnosing based on present mood episodes. The corrected code addresses this by checking for the presence of the current mood episodes based on `Observed` symptoms (Line 5). Overall, these corrections ensure that the diagnosis logic is more comprehensive and aligns with clinical practice.

The final corrected version of the GPT-generated code passes for all 10 (30) patients. In total, 57 lines were added and 10 removed from the initial 107 lines of code (LoC), resulting in a final 154 LoC. The first set of corrections—addressing cyclic dependencies and clinical inconsistencies—required the addition of 47 LoC and removal of 6, reflecting changes that demanded significant domain expertise. In contrast, modifying the diagnosis logic to incorporate present mood episodes added 10 LoC and removed 4, which were relatively straightforward adjustments.

These statistics highlight the varying levels of effort

[6] https://copilot.github.com/

[7] https://aws.amazon.com/q/developer/

Listing 3: Portion of manually corrected Datalog code in the candidate logic program generated by LLM.

```
DepressiveEpisode(Patient) :-
      + !MixedEpisode(Patient),
      DepressiveSymptomCount(Patient, Count), Count >= 5, AffectiveCluster(Patient, _).
MixedEpisode(Patient) :-
      - ManicEpisode(Patient), DepressiveEpisode(Patient).
      + DepressiveSymptomCount(Patient, DepressiveCount), DepressiveCount >= 3, MixedManicSymptomCount(Patient,
     ManicCount), ManicCount >= 3, MixedCore(Patient).
Diagnosis(Patient, "Bipolar_I") :- History(Patient, "manic_episode", Count1), Count1 >= 1.
Diagnosis(Patient, "Bipolar_I") :- History(Patient, "mixed_episode", Count2), Count2 >= 1.
+ Diagnosis(Patient, "Bipolar_I") :- ManicEpisode(Patient); MixedEpisode(Patient).
```

needed to refine different aspects of the generated code, while also demonstrating that much of the initial code was functional and required only minimal deletions to align with clinical guidelines. Despite the manual effort required, LLMs significantly accelerate the initial code generation process, providing a strong foundation that would otherwise require substantial time and expertise to build from scratch.

Answer to RQ4. To answer RQ4, which evaluates the effectiveness of LLMs in diagnosing patients directly, we revisit Table 1 under the *LLM-only* columns. Among the tested models, GPT leads with 9 correct diagnoses out of 10, followed by Gemini with 8 and Llama with 7. This trend extends across all patients, where GPT leads with 22 correct diagnoses out of 30, followed by Gemini and Llama with 19 each.

Directly using LLMs for diagnosis generally results in higher accuracy than relying on LLM-generated candidate programs. However, the variability in model performance highlights significant challenges. LLMs inherently rely on probabilistic predictions rather than logical proofs, making it difficult to guarantee consistency and accuracy required in medical contexts. Furthermore, their complex architectures make them hard to interpret. Unlike LLM-generated logic programs, which offer transparent reasoning steps, the direct diagnoses provided by LLMs remain opaque, even when correct. Finally, there are always ethical implications and privacy concerns of providing real patient data to LLMs, which complicates their direct application in healthcare.

Instead, our proposed method of combining LLMs with constraint logic programming offers a promising alternative. LLMs can be leveraged to generate interpretable logical rules that determine if a patient meets specific diagnostic criteria. This approach reduces ethical concerns by avoiding direct input of sensitive patient data into LLMs and takes advantage of the increasing availability of code-generative LLMs. Moreover, logic programs are inherently transparent and interpretable, as their rules can be manually verified for alignment with diagnostic standards of the manuals.

5 Related Work

To the best of our knowledge, our proposed method is the first method for combining LLMs and constraint logic programming to provide a clinical decision support system (CDSS) in the context of mental health diagnosis. Other recent studies have explored the use of LLMs in mental health contexts, including developing chat-based counselors (Liu et al. 2023), analyzing emotions (Yang et al. 2023), and predicting mental states from online text (Xu et al. 2024). However, these works do not extend to creating diagnostic tools or integrating logic-based reasoning to support clinical decision-making.

Beyond clinical applications, there is ongoing research on LLMs for logical reasoning, such as translating text into specifications for Boolean satisfiability (SAT) solvers (Ye et al. 2023) or evaluating their reasoning capabilities in mathematical and strategic domains (Imani, Du, and Shrivastava 2023; Zhang et al. 2024). While these efforts demonstrate LLMs' potential for logic-based tasks, they do not address the use of logic programming languages like Datalog in clinical settings.

Prior work on using logic in CDSS take an ontological approach of structured knowledge representations for diagnosis (Casado-Lumbreras et al. 2012), or apply satisfiability modulo theory (SMT) solvers and theorem provers to detect conflicts in medical treatments (Bowles et al. 2019). While these works address relevant clinical needs, they precede the advent of modern LLMs and do not incorporate logic programming. Our work bridges these gaps by leveraging LLMs to generate interpretable logic programs for mental health diagnosis, offering a unique combination of efficient AI techniques and explainable logic-based computation to assist clinicians.

6 Conclusion

We present a novel approach that integrates large language models (LLMs) with constraint logic programming (CLP) to design a clinical decision support systems (CDSS) for mental health diagnosis. Our evaluation demonstrates that while LLMs show promise for diagnostic tasks, they still face significant limitations when used directly, including issues with consistency, interpretability, and ethical concerns related to patient data. To address these challenges, our method utilizes LLMs to generate logic programs that encode 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. Future work will explore domain-specific fine-tuning of LLMs, evaluate the approach on real-world datasets, and extend the Datalog encoding to address more nuanced diagnostic criteria and specifiers.

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. All evaluations and decisions regarding diagnoses must be conducted in accordance with the expertise of trained professionals. The hypothetical data and modeling used in this study are intended for proof of concept and are not meant to substitute for real patient data, which may be more complex.

References

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

American Psychological Association. 2023. Psychologists reaching their limits as patients present with worsening symptoms year after year.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Berner, E. S. 2007. *Clinical decision support systems*, volume 233. Springer.

Bowles, J.; Caminati, M.; Cha, S.; and Mendoza, J. 2019. A framework for automated conflict detection and resolution in medical guidelines. *Science of Computer Programming*, 182: 42–63.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Casado-Lumbreras, C.; Rodríguez-González, A.; Álvarez Rodríguez, J. M.; and Colomo-Palacios, R. 2012. PsyDis: Towards a diagnosis support system for psychological disorders. *Expert Systems with Applications*, 39(13): 11391–11403.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Friha, O.; Amine Ferrag, M.; Kantarci, B.; Cakmak, B.; Ozgun, A.; and Ghoualmi-Zine, N. 2024. LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness. *IEEE Open Journal of the Communications Society*, 5: 5799–5856.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Gorvin, L.; and Brown, D. 2012. The psychology of feeling like a burden: A review of the literature. *Social Psychology Review*, 14(1): 28–41.

Huang, J.; and Chang, K. C.-C. 2023. Towards Reasoning in Large Language Models: A Survey. arXiv:2212.10403.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232.

Imani, S.; Du, L.; and Shrivastava, H. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Jiang, J.; Wang, F.; Shen, J.; Kim, S.; and Kim, S. 2024. A Survey on Large Language Models for Code Generation. *arXiv preprint arXiv:2406.00515*.

Jordan, H.; Scholz, B.; and Subotić, P. 2016. Soufflé: On Synthesis of Program Analyzers. In Chaudhuri, S.; and Farzan, A., eds., *Computer Aided Verification*, 422–430. Cham: Springer International Publishing. ISBN 978-3-319-41540-6.

Keen, P. G.; and Scott Morton, M. S. 1978. Decision support systems: an organizational perspective. *(No Title)*.

Liu, J. M.; Li, D.; Cao, H.; Ren, T.; Liao, Z.; and Wu, J. 2023. ChatCounselor: A Large Language Models for Mental Health Support. arXiv:2309.15461.

Llama Team, AI @ Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Moore, J.; Deshpande, T.; and Yang, D. 2024. Are Large Language Models Consistent over Value-laden Questions? arXiv:2407.02996.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Scholz, B.; Jordan, H.; Subotić, P.; and Westmann, T. 2016. On fast large-scale program analysis in Datalog. In *Proceedings of the 25th International Conference on Compiler Construction*, CC '16, 196–206. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342414.

Ullah, E.; Parwani, A.; Baig, M. M.; and Singh, R. 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1): 43.

Wang, H.; Zhao, S.; Qiang, Z.; Xi, N.; Qin, B.; and Liu, T. 2024. Beyond Direct Diagnosis: LLM-based Multi-Specialist Agent Consultation for Automatic Diagnosis. *arXiv preprint arXiv:2401.16107*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

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

Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J.; Ghassemi, M.; Dey, A. K.; and Wang, D. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1): 1–32.

Yang, K.; Ji, S.; Zhang, T.; Xie, Q.; Kuang, Z.; and Ananiadou, S. 2023. Towards Interpretable Mental Health Analysis with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ye, X.; Chen, Q.; Dillig, I.; and Durrett, G. 2023. SatLM: Satisfiability-Aided Language Models Using Declarative Prompting. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 45548–45580. Curran Associates, Inc.

Zhang, Y.; Mao, S.; Ge, T.; Wang, X.; de Wynter, A.; Xia, Y.; Wu, W.; Song, T.; Lan, M.; and Wei, F. 2024. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models. arXiv:2404.01230.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2023. Explainability for Large Language Models: A Survey. arXiv:2309.01029.

Zheng, K.; Decugis, J.; Gehring, J.; Cohen, T.; Negrevergne, B.; and Synnaeve, G. 2024. What Makes Large Language Models Reason in (Multi-Turn) Code Generation? arXiv:2410.08105.

A Detailed Diagnosis Results

Table 2: Comparing our method with baselines on all 30 patients. ‘Known Disorder’ indicates what the patient is diagnosed with according to the ICD-11 CDDR criteria. ‘LLM-only’ indicates the diagnosis directly produced by LLMs. ‘LLM+Datalog’ indicates the diagnosis produced by the LLM-generated Datalog program. ‘Our CDSS’ indicates the diagnosis produced by our method. The symbol ‘-’ indicates no clear diagnosis.

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	-	-	-	-	-	-	-	-
No. 11	-	BPD2	BPD2	BPD2	BPD1	-	-	-
No. 12	BPD1	BPD1	BPD1	BPD2	BPD1	BPD1	-	BPD1
No. 13	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1
No. 14	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1, BPD2	BPD1	BPD1
No. 15	BPD2	BPD2	BPD2	BPD2	BPD1	BPD1	RDD	BPD2
No. 16	BPD2	BPD2	BPD2	BPD2	BPD1	BPD1	BPD2	BPD2
No. 17	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1
No. 18	RDD	RDD	SEDD	RDD	BPD1	-	RDD	RDD
No. 19	BPD2	BPD2	BPD2	BPD2	BPD1	SEDD	BPD2	BPD2
No. 20	SEDD	-	-	-	-	-	SEDD	SEDD
No. 21	BPD1	BPD1	BPD1	BPD1	BPD1	-	BPD1	BPD1
No. 22	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1	BPD1
No. 23	BPD2	-	BPD1	-	-	BPD1	SEDD	BPD2
No. 24	-	BPD1	BPD1	BPD1	BPD1	BPD1	-	-
No. 25	RDD	RDD	SEDD	RDD	BPD1	-	RDD	RDD
No. 26	BPD2	SEDD	BPD1	BPD2	BPD1	BPD1, BPD2	BPD2	BPD2
No. 27	SEDD	RDD	SEDD	RDD	BPD1	-	SEDD	SEDD
No. 28	SEDD	SEDD	SEDD	SEDD	BPD1	SEDD	-	SEDD
No. 29	RDD	SEDD	SEDD	SEDD	BPD1	SEDD	SEDD	RDD
No. 30	-	BPD2	BPD2	-	BPD1	BPD1	-	-
Correct Diagnosis (Total):		19/30	19/30	22/30	9/30	(8+4)/30	22/30	30/30

B Patient Information

Table 3: Input data of all 30 patients. The symbol '-' for Column 4 indicates that there is no prior history condition of mood episode. '-' in Column 5 indicates that the patients' observed symptoms do not qualify for a current mood episode.

Patient ID	Disorder	Observed Symptoms	History Conditions	Mood Episode
No. 1	BPD2	depressed_mood 1.5 reduced_concentration 1.2 reduced_energy 0.8 increased_talkativeness 0.6	depressive 1 hypomanic 1	-
No. 2	RDD	depressed_mood 5.7 diminished_interest_pleasure 5.7 reduced_concentration 3.5 low_self_worth 2.0 psychomotor_disturbances 5.7	depressive 1	depressive
No. 3	BPD1	increased_activity_energy 0.5 euphoria_irritability_expansiveness 0.5 racing_thoughts 0.5 increased_talkativeness 0.5 increased_self_esteem 0.5 diminished_interest_pleasure 2.0 reduced_concentration 2.0 disrupted_excessive_sleep 2.0 change_in_appetite_weight 2.0 psychomotor_disturbances 2.0	mixed 2	depressive hypomanic
No. 4	BPD2	depressed_mood 5.7 diminished_interest_pleasure 5.7 reduced_concentration 3.5 low_self_worth 2.0 psychomotor_disturbances 5.7	hypomanic 1	depressive
No. 5	BPD1	depressed_mood 7.5 low_self_worth 7.5 disrupted_excessive_sleep 4.0 reduced_energy 5.5 change_in_appetite_weight 5.5 euphoria_irritability_expansiveness 3.0 increased_activity_energy 2.5 racing_thoughts 2.5 decreased_need_for_sleep 1.0 distractibility 1.0	-	mixed
No. 6	BPD2	depressed_mood 2.0 diminished_interest_pleasure 1.5 reduced_energy 1.0	depressive 1 hypomanic 1	-
No. 7	BPD1	depressed_mood 1.8 increased_activity_energy 0.5 reduced_concentration 1.0	depressive 1 mixed 1	-
No. 8	SEDD	depressed_mood 4.0	depressive 1	-
No. 9	SEDD	depressed_mood 2.5 recurrent_thoughts_death_suicide 1.7 change_in_appetite_weight 1.0	depressive 1	-
No. 10	-	depressed_mood 1.5 reduced_energy 0.9 increased_self_esteem 0.7	-	-

Continued on next page

Table 3: Input data of all 30 patients. The symbol '-' for Column 4 indicates that there is no prior history condition of mood episode. '-' in Column 5 indicates that the patients' observed symptoms do not qualify for a current mood episode. (Continued)

Patient ID	Disorder	Observed Symptoms	History Conditions	Mood Episode
No. 11	-	increased_talkativeness 1.2 euphoria_irritability_expansiveness 1.0	hypomanic 1	-
No. 12	BPD1	euphoria_irritability_expansiveness 2.5 increased_activity_energy 3.2 increased_talkativeness 1.8 racing_thoughts 2.9 decreased_need_for_sleep 2.7	hypomanic 2	manic
No. 13	BPD1	euphoria_irritability_expansiveness 1.5 increased_self_esteem 1.2 distractibility 1.8 impulsive_reckless_behavior 2.0 increased_sexual_sociability_goal_directed_activity 2.3	manic 1	-
No. 14	BPD1	depressed_mood 3.5 diminished_interest_pleasure 3.1 euphoria_irritability_expansiveness 2.6 increased_activity_energy 2.4	mixed 2	-
No. 15	BPD2	euphoria_irritability_expansiveness 0.7 increased_activity_energy 1.2 increased_talkativeness 1.8 racing_thoughts 0.7 decreased_need_for_sleep 1.2	depressive 2	hypomanic
No. 16	BPD2	depressed_mood 2.0 reduced_concentration 3.1 low_self_worth 2.7 increased_activity_energy 1.2	depressive 1 hypomanic 1	-
No. 17	BPD1	euphoria_irritability_expansiveness 2.8 increased_activity_energy 2.8 racing_thoughts 3.0 decreased_need_for_sleep 2.5 impulsive_reckless_behavior 2.7	manic 1	manic
No. 18	RDD	psychomotor_disturbances 3.0 hopelessness 2.9 recurrent_thoughts_death_suicide 4.0	depressive 2	-
No. 19	BPD2	diminished_interest_pleasure 3.4 increased_self_esteem 2.2 decreased_need_for_sleep 2.4	hypomanic 1 depressive 1	-
No. 20	SEDD	delusions 4.1 passivity_experiences 3.7 disorganized_behavior 3.9	depressive 1	-
No. 21	BPD1	reduced_energy 2.5 disrupted_excessive_sleep 3.0 change_in_appetite_weight 2.8 psychomotor_disturbances 2.9	manic 1 depressive 1	-
No. 22	BPD1	depressed_mood 3.6 hopelessness 2.8 increased_activity_energy 3.2 impulsive_reckless_behavior 3.1	mixed 1 hypomanic 1	-

Continued on next page

Table 3: Input data of all 30 patients. The symbol '-' for Column 4 indicates that there is no prior history condition of mood episode. '-' in Column 5 indicates that the patients' observed symptoms do not qualify for a current mood episode. (Continued)

Patient ID	Disorder	Observed Symptoms	History Conditions	Mood Episode
No. 23	BPD2	euphoria_irritability_expansiveness 0.5 increased_activity_energy 0.5 increased_self_esteem 0.5 impulsive_reckless_behavior 0.5 distractibility 0.5	depressive 1	hypomanic
No. 24	-	increased_activity_energy 2.6 distractibility 2.3 racing_thoughts 2.7 increased_self_esteem 2.9 impulsive_reckless_behavior 2.6	-	-
No. 25	RDD	low_self_worth 2.3 recurrent_thoughts_death_suicide 3.8 change_in_appetite_weight 2.7	depressive 2	-
No. 26	BPD2	depressed_mood 5.7 diminished_interest_pleasure 5.7 reduced_concentration 3.5 low_self_worth 2.0 psychomotor_disturbances 5.7 euphoria_irritability_expansiveness 0.5 increased_activity_energy 0.5 increased_self_esteem 0.5 impulsive_reckless_behavior 0.5 distractibility 0.5	depressive 1 hypomanic 1	depressive hypomanic
No. 27	SEDD	reduced_concentration 3.5 low_self_worth 2.0 hopelessness 5.7 recurrent_thoughts_death_suicide 4.0 disrupted_excessive_sleep 3.5 change_in_appetite_weight 2.0 psychomotor_disturbances 5.7 reduced_energy 4.0	depressive 1	-
No. 28	SEDD	depressed_mood 5.7 diminished_interest_pleasure 5.7 reduced_concentration 3.5 low_self_worth 2.0 psychomotor_disturbances 5.7	-	depressive
No. 29	RDD	depressed_mood 5.7 diminished_interest_pleasure 5.7 reduced_concentration 3.5 low_self_worth 2.0 psychomotor_disturbances 5.7	depressive 1	depressive
No. 30	-	euphoria_irritability_expansiveness 0.7 increased_activity_energy 1.2 increased_talkativeness 1.8 racing_thoughts 0.7 decreased_need_for_sleep 1.2	-	hypomanic

C Prompts

This section shows the full prompts we used to interact with the LLMs. Whenever applicable, we used system and user prompts as follows.

C.1 Translating ICD-11 CDDR Manual to Datalog Rules

***System**: You are an expert at translating mental health diagnostic criteria into Soufflé Datalog code. Translate the given criterion into a .dl program using Soufflé syntax as follows. The patient information 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.*

- *`.decl Observed(Patient:symbol, Symptom:symbol, Week:float)` describes that Patient has experienced Symptom for Week number of weeks.*
- *`.decl History(Patient:symbol, Condition:symbol, Count:number)` describes that Patient has experienced Condition for Count number of times.*
- *`.decl Diagnosis(Patient:symbol, Disorder:symbol)` describes that Patient has been diagnosed with Disorder.*

For context, here is an example of Scizophrenia criterion translated into Soufflé .dl code.

- *Scizhophrenia criterion: [Scizhophrenia criterion from ICD-11 CDDR].*
- *Relevant symptom names for `Observed` relation: [Symptom names]*
- *Soufflé .dl code: [Manually crafted Datalog program for Schizophrenia]*

***User**: Now, translate the following criteria into Souffle .dl code for Bipolar I, Bipolar II, Single Episode Depressive Disorder, and Recurrent Depressive Disorder.*

- *Mood Episode criterion: [Depressive, Manic, Mixed, and Hypomanic Episode criteria from ICD-11 CDDR].*
- *Mood Disorder criterion: [Bipolar I, Bipolar II, Single Episode Depressive Disorder, and Recurrent Depressive Disorder criteria from ICD-11 CDDR].*
- *Relevant symptom names for `Observed` relation: [Symptom names]*
- *Relevant condition names for `History` relation: [Condition names]*

C.2 Generating Diagnosis by *LLM-only Approach*

***System**: You are an expert at diagnosing patients according to the ICD-11 Clinical Descriptions and Diagnostic Requirements (CDDR). The patient data are represented by a list of current symptoms denoted as `Observed` and a list of history denoted as `History`. `Observed` matches the patient with the symptom and the number of weeks it has been observed. `History` matches the patient with the condition and the number of times it existed. No record for a patient means that there is no related data for them. The considered disorders are: Bipolar I, Bipolar II, Single Episode Depressive Disorder, and Recurrent Depressive Disorder.*

***User**: For brevity, please output only the diagnosis for the following patients. Patients with no clear diagnosis should be indicated as such.*

- *`Observed`: [Observed Data]*
- *`History`: [History Data]*