PsyEHR-Bench: An Agentic Framework and Benchmark for Cohort Selection in Chinese Psychiatric EHRs

Anonymous ACL submission

Abstract

Cohort selection from electronic health records (EHRs) is especially challenging in psychiatry, where the variables that drive most studies (e.g., prior suicide attempts, family history, treatment adherence) are documented in free-text fields rather than in structured fields. Existing EHR benchmarks target text-to-SQL and factoid answering on English structured fields from ICU or general-hospital cohorts, leaving Chinese psychiatric cohort selection over mixed structured and free-text conditions unaddressed. We introduce PsyEHR-Bench, the first cohort-selection benchmark evaluated over Chinese psychiatric EHRs, comprising 134 natural-language queries from 80 empirical studies, each paired with a verifiable ground-truth patient set on a de-identified database of approximately 59,000 patients. Our agentic framework couples automated SQL execution with a specialized free-text toolset for semantic search, lexical variant discovery, and per-record multi-criterion judgments, together with experience-driven self-evolution. The framework establishes a strong baseline with an overall F1-score of 0.851 on the 100-query test set. A 0.226 gap between structured-only queries (F1-score = 0.900) and queries requiring free-text (F1-score = 0.674) identifies unstructured clinical narratives as the primary bottleneck. Ablations show that the free-text toolset forms a highly synergistic ensemble whose joint contribution on the free-text subset (+0.244) significantly exceeds its individual additive effects.

1 Introduction

Cohort selection is the first step of most retrospective research on electronic health records (EHRs), yet the work of translating clinical eligibility criteria into database queries remains a persistent obstacle. The criteria are written by clinicians and epidemiologists, while the underlying data reside in a clinical database that requires substantial engineering to query. Clinical researchers typically lack this advanced database programming expertise, creating a persistent bottleneck in retrospective studies.

Psychiatry widens this gap. Psychiatric diagnosis lacks a laboratory or imaging gold standard (Lawrence et al., 2024). Consequently, the variables that drive most studies (e.g., prior suicide attempts, substance use, treatment adherence, family history, and social determinants of health) are primarily documented in free-text fields (clinical narratives such as admission notes, progress notes, and discharge summaries) rather than in structured fields (Newby et al., 2024; Cusick et al., 2022). Even suicidal behavior, a high-salience event, is captured in structured diagnostic codes only in a small fraction of cases (Cusick et al., 2022; Xie et al., 2022). Psychiatric phenotypes also unfold across years: longitudinal analyses report that roughly 64% of patients with severe mental illness experience diagnostic changes across visits (De la Hoz et al., 2022; Jørgensen et al., 2023). A realistic cohort definition therefore requires both narrative reading and longitudinal reasoning across visits.

Existing benchmarks for natural-language access to EHRs target text-to-SQL and factoid answering on English structured fields from ICU or general-hospital cohorts (Lee et al., 2022; Shi et al., 2024; Lee et al., 2026; Jiang et al., 2025; Lee et al., 2025a). None of these is designed for cohort selection on a Chinese psychiatric EHR where the criteria mix structured and free-text conditions. Hybrid query systems and psychiatric natural language processing (NLP) have explored adjacent parts of this space, and we defer the detailed comparison to §2.

Our contributions are as follows:

- We introduce PsyEHR-Bench[1], the first cohort-selection benchmark evaluated over Chinese psychiatric EHRs. This benchmark

[1] Our implementation is available at https://anonymous.4open.science/r/PsyEHR-Bench.

comprises 134 real-world queries derived from empirical studies and provides high-fidelity, verifiable ground-truth patient sets constructed over a large-scale clinical database.

- We design a novel agentic framework centered on a specialized free-text toolset to resolve unstructured clinical narratives. By coupling automated SQL execution with integrated tools for semantic search, lexical variant discovery, and per-record multi-criterion judgments, our system addresses the limitations of traditional text-to-SQL paradigms and establishes a strong baseline with an overall F1-score of 0.851.

- Through comprehensive ablation studies, we deliver systematic insights into the mechanisms required for unstructured clinical retrieval. We demonstrate that this free-text toolset forms a highly synergistic ensemble whose joint contribution significantly exceeds its individual additive effects. Furthermore, we quantify the pragmatic impact of experience-driven self-evolution, showing that its optimization benefits fall almost entirely on unstructured narratives by streamlining the tool-invocation sequence.

2 Related Work

2.1 EHR Query and Cohort Extraction

Cohort selection from electronic health records (EHRs) requires translating natural-language eligibility criteria into queries over a mix of structured fields and clinical narratives. Existing benchmarks tackle pieces of this translation under different output assumptions. Text-to-SQL and question-answering benchmarks grade a single executable answer by execution accuracy. For instance, EHRSQL (Lee et al., 2022) pairs natural-language questions with SQL ground truth on MIMIC-III and eICU with explicit unanswerable-question handling. EHR-ChatQA (Lee et al., 2026) extends this to interactive sessions with follow-up disambiguation, while EHRAgent (Shi et al., 2024), MedAgentBench (Jiang et al., 2025), and FHIR-AgentBench (Lee et al., 2025a) place the task inside multi-step agents over relational or FHIR-standard schemas. Patient-eligibility systems instead score a patient against a fixed criterion list. Specifically, TrialGPT (Jin et al., 2024) uses a three-tier retrieval/matching/ranking pipeline with per-criterion explanations. MAKAR (Shi et al., 2025) adds knowledge-augmented multi-agent reasoning on the n2c2 cohort, and RDMA (Wu et al., 2025) mines rare-disease phenotypes from MIMIC-III clinical notes. EMR-AGENT (Lee et al., 2025b) is the closest existing cohort-selection work, proposing an LLM-driven agent measured by set-level F1-score over ICU databases (MIMIC-III, eICU, SICdb), but operating on English structured schemas with criteria expressible as SQL filters.

2.2 Hybrid Structured-Unstructured Query

Real queries often combine relational filters over structured fields with semantic predicates over unstructured data. Hybrid query systems address this by introducing LLM-based semantic operators that compose with relational operators. SUQL (Liu et al., 2024) extends SQL with Answer and Summary UDFs evaluated by an LLM, demonstrated on Yelp restaurants and HybridQA. LOTUS (Patel et al., 2025) formalizes six semantic operators over DataFrames with proxy-oracle cascades for cost-bounded execution. BlendSQL (Glenn et al., 2024) embeds LLM ingredients in SQLite with predicate pushdown to limit LLM calls on hybrid QA tasks. TAG (Biswal et al., 2025) proposes a three-step framework of query synthesis, execution, and answer generation, and releases TAG-Bench on five BIRD-derived databases. Weaver (Khoja et al., 2025) interleaves SQL and LLM steps in an explicit multi-step plan for hybrid table reasoning over WikiTableQuestions, TabFact, FinQA, and multimodal extensions. Benchmarks include SWAN (Zhao et al., 2025) on beyond-database questions over general-domain tables and the PRIME split of STaRK (Wu et al., 2024), whose biomedical setting evaluates retrieval over the PrimeKG knowledge graph rather than over patient records. None of these systems or benchmarks runs on a real clinical EHR.

2.3 Psychiatric NLP

Psychiatric NLP has historically focused on dialogue, counseling, and social-media data rather than institutional records. MentalGLM (Zhai et al., 2025) fine-tunes GLM-4 on Chinese social-media posts for suicide-risk and cognitive-distortion analysis. DeepWell-Adol (Qiu et al., 2025) constructs an adolescent positive-psychology dialogue corpus in Chinese, with mixed expert and synthetic origins. DSM5-AgentFlow (Ozgun et al., 2025) orchestrates

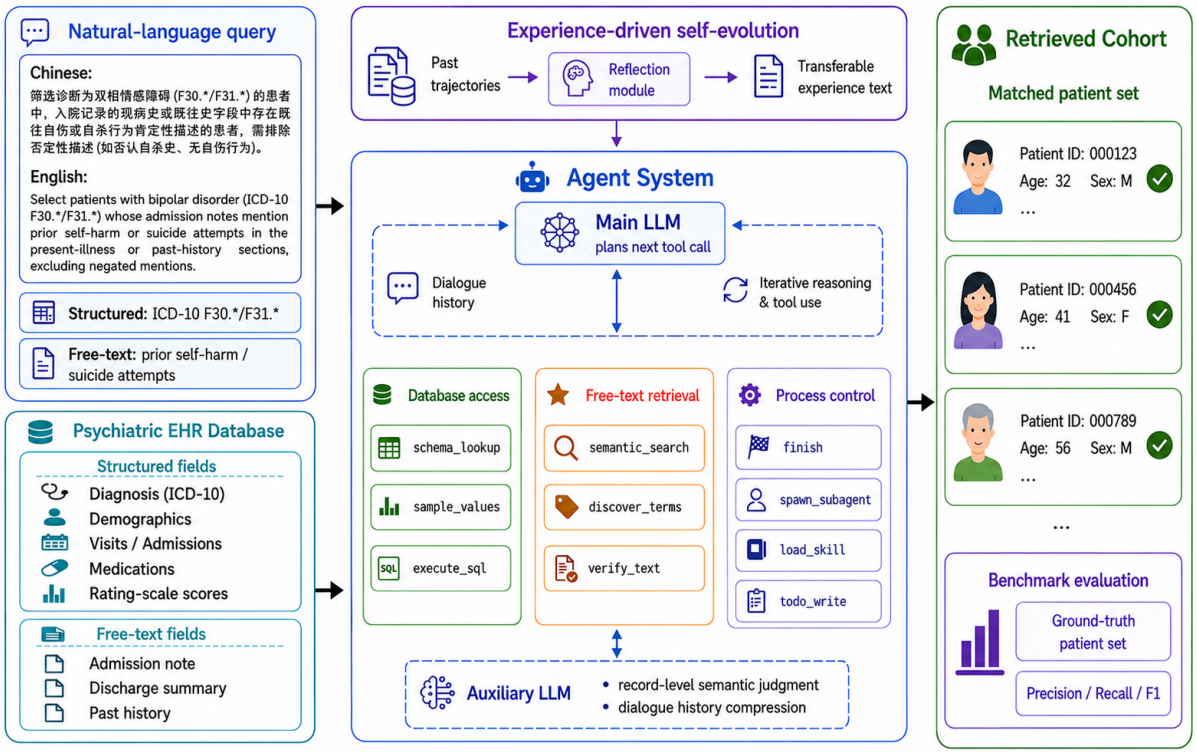

Figure 1: Overview of PsyEHR-Bench and the proposed agentic framework. A natural-language cohort query that mixes structured and free-text conditions (top left) is resolved by an agent (center) that iteratively invokes a toolset over a de-identified Chinese psychiatric EHR database (bottom left). The toolset is organized into three functional groups: database access, free-text retrieval, and process control. An auxiliary LLM supports the free-text tools via per-record semantic judgments and dialogue history compression. Experience-driven self-evolution (top) distills past trajectories into a transferable experience text that is appended to the agent's system prompt. The retrieved patient cohort (right) is scored against the ground-truth set by set-level Precision, Recall, and F1-score.

a multi-agent workflow over the DSM-5 Level-1 symptom checklist, producing synthetic dialogues across ten disorders. PsychiatryBench (Fouda et al., 2026) aggregates 5,188 expert-extracted items from thirteen psychiatric textbooks across eleven task types. Two recent benchmarks ground the evaluation in real Chinese psychiatric records. PsychBench (Liu et al., 2025) grades sixteen LLMs on 300 de-identified records from three Chinese psychiatric centers across six clinical tasks. Similarly, LingxiDiagBench (Xu et al., 2026) couples 1,709 real outpatient records from Shanghai Mental Health Center with 16,000 synthetic dialogues for ICD-10 classification under static and dynamic consultation. Scoping and systematic reviews (Hua et al., 2025; Guo et al., 2024) confirm that social-media data remains the dominant source and that work grounded in real clinical records is still rare. Both PsychBench and LingxiDiagBench evaluate per-patient diagnostic or treatment tasks rather than retrieval of cohorts whose records satisfy hybrid eligibility criteria.

3 PsyEHR-Bench

PsyEHR-Bench is the first cohort-selection benchmark evaluated over Chinese psychiatric electronic health records (EHRs). It comprises 134 natural-language queries derived from published empirical studies, each paired with a verifiable ground-truth patient set executed on a de-identified database from a tertiary psychiatric hospital.

3.1 Task Definition

Given a natural-language query specifying clinical inclusion and exclusion criteria, the objective is to retrieve the matching set of patients, where each individual is identified by a unique ID in the database. Evaluation relies on set-level Precision, Recall, and F1-score. Precision measures the purity of the retrieved cohort, while Recall measures its coverage, reflecting the core dimensions used by clinical researchers to evaluate cohort extraction results.

3.2 Database and Queries

The underlying database consists of de-identified psychiatric electronic health records spanning from 2018 to 2024 at a tertiary psychiatric hospital. The database encompasses approximately 59,000 patients, 1.1 million visits, and 50 million records organized across 14 tables. Structured fields, such as diagnostic codes, medication orders, and rating-scale scores, coexist with free-text fields containing clinical narratives from progress notes, admission assessments, and discharge summaries.

To ensure the queries align with real-world clinical research needs rather than synthetic templates, we extract 134 inclusion and exclusion criteria from 80 empirical psychiatric studies published in prominent medical journals. These queries span eight clinical research directions, including Epidemiology, Longitudinal cohorts and disease progression, and Suicide & self-harm.

The 134 queries are split into a development set of 34 queries and a held-out test set of 100 queries; the development set facilitates experience-driven self-evolution, whereas the test set is reserved for final evaluation. This partition preserves both the criteria-type distribution—comprising 78 structured-only queries and 22 queries requiring free-text in the test set—and the consistent coverage across all eight research directions. Table 1 lists the detailed dataset statistics.

3.3 Ground-Truth Construction

The ground-truth patient set for each query is constructed by separately resolving structured and free-text conditions before combining them according to the logical architecture of the query. Structured conditions can be executed directly via SQL queries over the database, whereas free-text conditions necessitate per-record multi-criterion judgments over clinical narratives.

For structured conditions, domain experts draft reference SQL statements against the schema. To mitigate single-annotator bias or implementation errors, each condition is independently formulated by multiple experts whose retrieval outputs are cross-checked; any discrepancy is reviewed and revised collectively until all implementations yield identical patient sets.

Free-text conditions cannot be directly formulated within standard SQL statements. We first leverage the structured conditions of the same query to restrict the scope to a candidate patient set. Then, a locally deployed language model[2] serves as a pre-annotator to execute record-level judgments across all relevant narratives for each candidate patient. Finally, a random sample of the model-generated annotations is independently validated by human annotators to ensure high fidelity of the ground truth.

4 Agent System

This section describes our agent for PsyEHRBench and its two core mechanisms: free-text retrieval for free-text conditions, and experience-driven self-evolution for accumulating experience from past trajectories. Figure 1 presents an overview.

4.1 Backbone

The agent runs a stateless main loop that resolves a natural-language query into a matching patient set. At each step, the main LLM reads the dialogue history and the tool descriptions, selects the next tool call, and the tool output is appended to the dialogue. The loop runs until the agent signals termination. The toolset comprises ten tools divided into three functional groups (database access, free-text retrieval, and process control), as listed in Table 2. The `finish` tool in the process-control group serves as the termination signal: the agent submits its final patient set inside the `finish` call, ending the loop. Two LLMs are used: the main LLM selects tool calls in the main loop, while the auxiliary LLM performs two supporting computations: per-record semantic judgments inside `verify_text` and compression of the dialogue history when its length approaches the context limit.

4.2 Free-Text Retrieval

Free-text conditions must be judged from the semantics of clinical narratives. Retrieving such conditions is difficult for two reasons. First, the same concept appears in many synonymous phrasings across clinicians, and the surface forms used in a given field are unknown until the field is inspected. Second, even when a record contains the relevant phrasing, it may be negated, refer to a family member rather than the patient, or describe a past state; therefore, polarity, subject, and tense must be judged within the record. We design three composable tools to address these two sources of difficulty, al-

[2] Qwen3.5-4B (https://huggingface.co/Qwen/Qwen3.5-4B).

Group	Item	Full	Dev	Test
Total		**134**	**34**	**100**
Query type	Structured-only	104	26	78
	Requires free-text	30	8	22
Research direction	Treatment response & medication	25	7	18
	Longitudinal cohorts & disease progression	21	5	16
	Readmission & service utilization	21	4	17
	Diagnostic trajectory & outcomes	18	3	15
	Epidemiology	14	3	11
	Specific populations	13	4	9
	Suicide & self-harm	12	4	8
	Clinical phenotyping & validation	10	4	6

Table 1: Distribution of queries in PsyEHR-Bench by criteria type and research direction, shown for the full set, the development set, and the test set.

Group	Tool	Description
Database access	`schema_lookup`	List tables and column schemas of the database.
	`sample_values`	Draw sample values from a given field.
	`execute_sql`	Execute a read-only SQL query.
Free-text retrieval	`semantic_search`	Vector-neighbor search on a free-text field given anchor phrases.
	`discover_terms`	Enumerate literal variants of a seed term in a field with frequencies.
	`verify_text`	Judge each record of given patients against a multi-criterion list.
Process control	`finish`	Terminate the loop and submit the final patient set.
	`spawn_subagent`	Delegate a sub-task to an isolated agent.
	`load_skill`	Load an external skill module into context.
	`todo_write`	Maintain a structured todo list across steps.

Table 2: Tools available to the agent, organized by functional group.

lowing the agent to choose which to call and in what order for each query.

4.2.1 Anchored Semantic Search

Given a text field f and an anchor set $A = \{a_1, \ldots, a_n\}$ from the agent, `semantic_search` queries a pre-built vector index over f for each anchor, merges and deduplicates the results, and returns the top-k records by similarity:

$$\texttt{semantic_search}(f, A) \rightarrow R_k. \quad (1)$$

The tool takes anchors instead of the full query. A clinical query typically mixes several semantic dimensions, which a single embedding of the whole query would collapse. Letting the agent enumerate anchors keeps this decomposition explicit and visible in the trace.

4.2.2 Lexical Variant Discovery

Given a text field f and a seed term s, `discover_terms` returns all literal variants in f that contain s, along with their frequencies:

$$\texttt{discover_terms}(f, s) \rightarrow \{(v_i, n_i)\}_{i=1}^{m}. \quad (2)$$

For example, with f set to the current-history field and s set to *suicide* (自杀), the tool returns variants such as *suicidal ideation* (自杀意念, 8,421), *suicide attempt* (自杀未遂, 3,120), *denies history of suicide* (否认自杀史, 612), and so on. These are literal strings actually present in the field; no generative model is used to propose synonyms. The agent thus observes the field's actual vocabulary distribution and can insert any useful variant directly into the literal-match clause of an SQL query.

4.2.3 Per-Record Verification

Given a candidate patient set C already narrowed by the agent, a text field f, and a list of criteria $K = \{k_1, \ldots, k_m\}$, `verify_text` returns the subset of C satisfying every criterion:

$$\texttt{verify_text}(C, f, K) \rightarrow C' \subseteq C. \quad (3)$$

For each patient $p \in C$, the tool retrieves all records of p under f, and the auxiliary LLM independently labels each record on each k_j as yes, no, or unsure. Labels are combined with a logical AND across criteria at the record level and an OR across records at the patient level; the patient is kept whenever any single record simultaneously satisfies all criteria. The criteria list is given by the agent and typically covers concept, subject, and polarity simulta-

neously, for example {*mentions suicidal ideation*; *the subject is the patient, not a family member*; *not a negation*}. The unit of judgment is the record, not the patient; the patient-level OR captures the *ever-satisfied* temporal reading.

The three tools can be used alone or cascaded. A typical workflow starts with `semantic_search` to find relevant records, calls `discover_terms` along the confirmed semantic directions to list the literal variants in the field, writes those variants into the literal-match clause of an SQL query to narrow the candidate set as far as possible, and finally runs `verify_text` on the narrowed set for the strict multi-criterion judgment. The execution order follows the call cost. Vector search and literal counting are one-shot operations at the field level, while record-level LLM judgment scales linearly with the candidate-set size; placing it last restricts the expensive call to the smallest amount of data.

4.3 Experience-Driven Self-Evolution

The agent distills trajectories from past runs into a transferable experience text and appends it to the system prompt, enabling continuous improvement across queries. A separate reflection module periodically reads a batch of past runs. Each run record contains the set-level metrics, samples of false positives and false negatives, and a compressed tool-call trace. Across runs, the module identifies recurring error patterns and the default actions that resolve them, writing them into a short, transferable rule document. The rules are intentionally abstract: they may refer to patterns such as *free-text conditions involving negation*, but they never mention a specific query, literal term, or threshold, ensuring they remain applicable to unseen queries. The document is appended verbatim to the agent's system prompt and influences its decisions on subsequent queries. It is implemented as a reusable skill loaded via the `load_skill` tool. No model parameters are updated. The only cross-query state is this experience text, which can be reset, edited, or replaced as a whole.

5 Experiments

5.1 Setup

Evaluation is conducted on the 100-query test set. We employ DeepSeek-V4-Pro as the main LLM to select tool calls in the main loop, and DeepSeek-V4-Flash as the auxiliary LLM to perform per-record judgments inside `verify_text` and compress the dialogue history. The vector index for `semantic_search` is pre-built across all relevant free-text fields using Alibaba text-embedding-v4. Self-evolution is executed offline once on the 34-query development set, and the resulting experience document is appended verbatim to the agent's system prompt during test evaluation. For each query, we calculate set-level Precision, Recall, and F1-score, reporting the arithmetic mean across queries.

5.2 Main Results

The agent achieves an overall F1-score of 0.851 (P$=0.849$, R$=0.910$) on the test set. The F1-score reaches 0.900 on the 78 structured-only queries, but drops to 0.674 on the 22 queries requiring free-text; this 0.226 gap highlights the retrieval difficulty of free-text conditions.

To isolate the contribution of each component, we evaluate eight ablation configurations across three dimensions: removing each of the three free-text tools individually and all three jointly to form a baseline that uses only SQL (corresponding to the prior text-to-SQL paradigm); removing self-evolution or few-shot examples individually and jointly to form a zero-shot variant; and removing domain knowledge. Table 3 reports the mean Precision, Recall, and F1-score; Table 4 reports the per-query tool-call counts across all evaluated configurations.

6 Analysis

6.1 Performance across Research Directions

Figure 2 shows the F1-score distribution across the eight research directions, with mean F1-scores falling within the $[0.72, 0.99]$ range. Differences across directions are driven by two factors. The first is the mixture ratio of query types: among the six directions containing free-text conditions, the mean F1-score of the queries requiring free-text is consistently lower than that of the structured-only queries within the same direction. For example, in Readmission, the structured-only F1-score is 0.97 while the free-text F1-score is only 0.62; a similar intra-direction contrast appears in Longitudinal (0.98 versus 0.69). The concentration of orange dots in the lower-to-middle F1-score region of the figure reflects this gap.

The second factor is the internal complexity of the queries themselves, which manifests in the two structured-only directions. Epidemiology (n=11) consists entirely of single-condition queries, yield-

Method	Overall			Structured-only			Requires free-text		
	P	R	F1	P	R	F1	P	R	F1
Ours	0.849	**0.910**	**0.851**	0.898	**0.949**	**0.900**	**0.674**	**0.771**	**0.674**
w/o `discover_terms`	0.829	0.872	0.812	0.882	0.909	0.862	0.640	0.741	0.633
w/o `verify_text`	0.805	0.854	0.788	0.867	0.896	0.844	0.587	0.705	0.591
w/o `semantic_search`	0.813	0.850	0.788	0.869	0.904	0.847	0.618	0.658	0.580
w/o free-text tools	0.789	0.845	0.782	0.894	0.927	0.881	0.415	0.557	0.430
w/o self-evolution	0.822	0.878	0.813	0.891	0.928	0.880	0.577	0.698	0.574
w/o few-shot examples	0.795	0.855	0.781	0.880	0.892	0.853	0.493	0.726	0.524
w/o self-evolution & few-shot examples	0.777	0.832	0.758	0.845	0.864	0.815	0.538	0.721	0.555
w/o domain knowledge	**0.856**	0.872	0.830	**0.909**	0.910	0.882	0.668	0.737	0.646

Table 3: Precision, Recall, and F1-score on the 100-query test set, evaluated over all queries, structured-only queries, and queries requiring free-text.

Method	`semantic_search`	`discover_terms`	`verify_text`	`spawn_subagent`	`execute_sql`
Ours	0.49	2.97	0.58	0.00	7.89
w/o `discover_terms`	0.63	—	0.62	0.00	8.18
w/o `verify_text`	0.46	3.27	—	0.19	9.27
w/o `semantic_search`	—	3.91	0.77	0.00	8.47
w/o free-text tools	—	—	—	0.02	12.17
w/o self-evolution	0.68	2.50	0.98	0.00	6.91
w/o few-shot examples	0.47	6.27	0.74	0.01	12.85
w/o self-evolution & few-shot examples	0.50	5.61	0.57	0.01	12.11
w/o domain knowledge	0.48	2.90	0.60	0.00	8.40

Table 4: Mean per-query tool-call counts across all evaluated configurations.

ing an F1-score of 0.96. In contrast, Treatment (n=18), though also structured-only, achieves an F1-score of only 0.72, as these queries frequently require anchoring temporal windows across multiple visits and comparing rating-scale scores —tasks where the agent still struggles with cross-visit temporal alignment. The highest cross-direction F1-score occurs in Phenotyping (n=6) at 0.99, owing to its unambiguous, structured definitions.

6.2 Effect of Free-Text Retrieval Tools

Removing `verify_text` drops the F1-score on the subset requiring free-text from 0.674 to 0.591. Without per-record multi-criterion judgments, the agent resorts to more SQL literal matching as an approximation (`execute_sql` calls rise from a mean of 7.89 to 9.27, a 17% increase) and invokes `spawn_subagent` in 19 runs (compared to 0 in the full configuration) to delegate record-level evaluations to isolated subagents. Total tool calls instead increase, demonstrating that `verify_text` actually saves overall execution cost.

Removing `discover_terms` drops the F1-score on the subset requiring free-text from 0.674 to 0.633. The agent compensates with more frequent semantic retrieval (`semantic_search` calls increase by 28% from 0.49 to 0.63), yet SQL calls remain largely unchanged. While retrieved semantic neighbors indicate the conceptual direction, they do not directly expose the underlying literal variants inside the field; consequently, the agent cannot easily rewrite SQL literal-match clauses based on them.

Removing `semantic_search` drops the F1-score on the subset requiring free-text from 0.674 to 0.580, marking the largest decline among the three single-tool ablations. The agent attempts to compensate with more frequent lexical enumeration (`discover_terms` $+32\%$) and per-record evaluations (`verify_text` $+33\%$); however, lacking the entry tool, these downstream tools struggle to locate appropriate target records.

When all three tools are jointly disabled, the agent relies solely on SQL: `execute_sql` calls surge to 12.17 ($+54\%$), and the F1-score on the subset requiring free-text plummets to 0.430. The joint contribution of the three tools is $+0.244$, which exceeds the sum of their individual contributions ($+0.218$), indicating they form a synergistic ensemble rather than decomposable additive components.

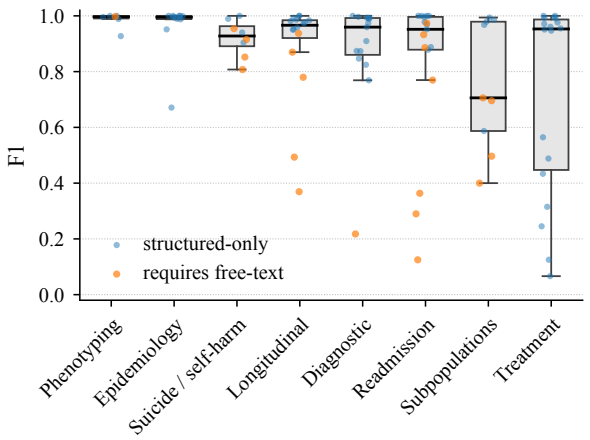

Figure 2: F1-score distribution across the eight research directions on the 100-query test set. Each box shows the interquartile range and median of per-query F1-score; overlaid dots mark individual queries, colored by query type (blue = structured-only, orange = requires free-text). Directions, sorted left-to-right by mean F1-score: Phenotyping (clinical phenotyping & validation, n=6), Epidemiology (n=11), Suicide & self-harm (n=8), Longitudinal (longitudinal cohorts & disease progression, n=16), Diagnostic (diagnostic trajectory & outcomes, n=15), Readmission (readmission & service utilization, n=17), Subpopulations (specific populations, n=9), Treatment (treatment response & medication, n=18).

6.3 Effect of Self-Evolution and Few-Shot Examples

Removing self-evolution drops the F1-score by only 0.020 on the structured-only subset, but by 0.100 on the subset requiring free-text; thus, the contribution of self-evolution falls almost entirely on the free-text side. Observing tool-call frequencies reveals that without self-evolution, the agent relies more heavily on `semantic_search` (+39% relative to the full configuration) and `verify_text` (+69%), while using `discover_terms` less often (−16%). In the full configuration, the agent adopts the opposite sequence: it first enumerates actual literals via `discover_terms`, narrows the candidate patient set with SQL literal matches, and finally calls `verify_text` for multi-criterion judgments. This sequence follows a low-to-high cost logic; the experience document translates this abstract design principle into specific agent behaviors.

The full configuration yields an overall gain of +0.093 relative to the zero-shot baseline, which decomposes into individual contributions from few-shot examples (+0.055) and self-evolution (+0.024). Their sum (+0.079) approaches the joint gain, suggesting they supplement the agent's knowledge from distinct dimensions.

6.4 Effect of Domain Knowledge

Removing domain knowledge decreases the overall F1-score from 0.851 to 0.830. Sliced by criteria type, the structured-only F1-score drops from 0.900 to 0.882, and the F1-score for queries requiring free-text drops from 0.674 to 0.646, indicating a slightly larger decline on the free-text side. Domain knowledge functions primarily during the interpretation of clinical narratives (aligning semantics for psychiatric terminology, medication names, and rating scales); compared to the core +0.244 contribution of the free-text tools, it represents a marginal but non-zero benefit.

7 Conclusion

We present PsyEHR-Bench, the first cohort-selection benchmark evaluated over Chinese psychiatric EHRs, alongside an agentic framework designed to resolve natural-language eligibility criteria that mix structured and free-text conditions. By coupling automated SQL execution with a specialized free-text toolset for semantic search, lexical variant discovery, and per-record multi-criterion judgments, the framework establishes a strong baseline with an overall F1-score of 0.851. However, a 0.226 gap between structured-only queries (F1-score = 0.900) and queries requiring free-text (F1-score = 0.674) identifies unstructured clinical narratives as the primary bottleneck of this task. Comprehensive ablations demonstrate that the free-text toolset forms a highly synergistic ensemble whose joint contribution significantly exceeds its individual additive effects. Furthermore, experience-driven self-evolution streamlines the tool-invocation sequence, yielding optimization benefits concentrated almost entirely on the free-text subset. Future work will focus on closing this remaining gap and validating the framework across additional institutions and base models.

Limitations

This work has several limitations.

First, PsyEHR-Bench is constructed from de-identified EHRs at a single tertiary psychiatric hospital. While the database covers routine clinical practice from 2018 to 2024 at a high-volume center, multi-center evaluation will be needed to assess generalization of both the benchmark and the

agent across institutions with different documentation conventions.

Second, both the main and the auxiliary LLMs are drawn from a single model family (DeepSeek). This controlled setup isolates the contribution of the proposed agent components while holding the model family fixed; broader evaluation across other open-source or commercial families is left to future work.

Third, results are reported on a single fixed development/test split, rather than under k-fold cross-validation. We adopt this setup to ensure deterministic comparability between the main agent and its eight ablation configurations.

Fourth, the task scope is intentionally restricted to cohort selection as a read-only research-data retrieval problem. Downstream EHR scenarios such as clinical decision support, individual-level prognosis, and treatment recommendation lie outside the scope of this work and require their own task definitions and validation protocols.

Fifth, free-text conditions in the ground truth are annotated through a model-assisted pre-annotation protocol with sampling-based human verification, rather than through full expert adjudication of every record. This design enables annotation at the scale required by a narrative-heavy EHR while preserving fidelity at the patient-set level.

Ethics Statement

This study uses de-identified electronic health records from a single tertiary psychiatric hospital, covering routine clinical practice from 2018 to 2024. The study was approved by the ethics committee of the institution where the data were collected, and the records were de-identified by the hospital information department before being made available to the research team; the research team had no access to any directly identifying patient information at any stage. Ground-truth labeling of free-text conditions was performed by a locally deployed language model, while evaluation of the agent involved a commercial large language model API; in both cases only de-identified records were processed. PsyEHR-Bench is intended strictly as a research benchmark for evaluating cohort-selection systems and should not be used as a basis for clinical decisions about individual patients. The benchmark queries and the agent code will be released under the Apache License 2.0 upon publication, while the underlying database is not released.

References

Asim Biswal, Liana Patel, Siddharth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, and Matei Zaharia. 2025. Text2SQL is not enough: Unifying AI and databases with TAG. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*.

Marika Cusick, Sumithra Velupillai, Johnny Downs, Thomas R. Campion, Jr., Evan T. Sholle, Rina Dutta, and Jyotishman Pathak. 2022. Portability of natural language processing methods to detect suicidality from clinical text in US and UK electronic health records. *Journal of Affective Disorders Reports*, 10:100430.

Juan F. De la Hoz, Alejandro Arias, Susan K. Service, Mauricio Castaño, Ana M. Díaz-Zuluaga, Janet Song, Cristian Gallego, Sergio Ruiz-Sánchez, Javier I. Escobar, Alex A. T. Bui, Carrie E. Bearden, Victor Reus, Carlos López-Jaramillo, Nelson B. Freimer, and Loes M. Olde Loohuis. 2022. Analysis of diagnosis instability in electronic health records reveals diverse disease trajectories of severe mental illness. *medRxiv preprint*.

Aya E. Fouda, Abdelrahman A. Hassan, Radwa J. Hanafy, and Mohammed E. Fouda. 2026. PsychiatryBench: A multi-task benchmark for LLMs in psychiatry. *npj Digital Medicine*, 9:320.

Parker Glenn, Parag Dakle, Liang Wang, and Preethi Raghavan. 2024. BlendSQL: A scalable dialect for unifying hybrid question answering in relational algebra. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language models for mental health applications: A systematic review. *JMIR Mental Health*.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, David A. Clifton, Andrew Beam, and John Torous. 2025. Large language models in mental health care: A scoping review. *Current Treatment Options in Psychiatry*, 12:27.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. MedAgentBench: A realistic virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9).

Qiao Jin, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2024. Matching patients to clinical trials with large language models. *Nature Communications*, 15(1):9074.

Terese Sara Høj Jørgensen, Merete Osler, Martin Balslev Jorgensen, and Anders Jorgensen. 2023. Mapping diagnostic trajectories from the first hospital diagnosis

of a psychiatric disorder: a Danish nationwide cohort study using sequence analysis. *The Lancet Psychiatry*, 10(1):12–20.

Rohit Khoja, Devanshu Gupta, Yanjie Fu, Dan Roth, and Vivek Gupta. 2025. Weaver: Interweaving SQL and LLM for table reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28282–28308, Suzhou, China. Association for Computational Linguistics.

Hannah R. Lawrence, Renee A. Schneider, Susan B. Rubin, Maja J. Matarić, Daniel J. McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11:e59479.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Yugang Jia, and Jong Ha Lee. 2025a. FHIR-AgentBench: Benchmarking LLM agents for realistic interoperable EHR question answering. In *Proceedings of Machine Learning Research 297: Machine Learning for Health (ML4H) 2025*.

Gyubok Lee, Woosog Chay, Heeyoung Kwak, Yeong Hwa Kim, Haanju Yoo, Oksoon Jeong, Meong Hi Son, and Edward Choi. 2026. From conversation to query execution: Benchmarking user and tool interactions for EHR database agents. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. EHRSQL: A practical text-to-SQL benchmark for electronic health records. In *Advances in Neural Information Processing Systems 35: Datasets and Benchmarks Track*.

Kwanhyung Lee, Sungsoo Hong, Joonhyung Park, Jeonghyeop Lim, Juhwan Choi, Donghwee Yoon, and Eunho Yang. 2025b. EMR-AGENT: Automating cohort and feature extraction from EMR databases. *arXiv preprint arXiv:2510.00549*.

Shicheng Liu, Jialiang Xu, Wesley Tjangnaka, Sina J. Semnani, Chen Jie Yu, and Monica S. Lam. 2024. SUQL: Conversational search over structured and unstructured data with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4535–4555.

Shuyu Liu, Ruoxi Wang, Ling Zhang, Xuequan Zhu, Rui Yang, Xinzhu Zhou, Fei Wu, Zhi Yang, Cheng Jin, and Gang Wang. 2025. PsychBench: A comprehensive and professional benchmark for evaluating the performance of LLM-assisted psychiatric clinical practice. *arXiv preprint arXiv:2503.01903*.

Danielle Newby, Niall Taylor, Dan W. Joyce, and Laura M. Winchester. 2024. Optimising the use of electronic medical records for large scale research in psychiatry. *Translational Psychiatry*, 14:232.

Mithat Can Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qingzhi Liu, and Junxiao Wang. 2025. Trustworthy AI psychotherapy: Multi-agent LLM workflow for counseling and explainable mental disorder diagnosis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, pages 3761–3764.

Liana Patel, Siddharth Jha, Melissa Pan, Harshit Gupta, Parth Asawa, Carlos Guestrin, and Matei Zaharia. 2025. Semantic operators: A declarative model for rich, AI-based data processing. *Proceedings of the VLDB Endowment*, 18(11):4171–4184.

Wenyu Qiu, Yuxiong Wang, Jiajun Tan, Hanchao Hou, Qinda Liu, Wei Yao, and Shiguang Ni. 2025. DeepWell-Adol: A scalable expert-based dialogue corpus for adolescent positive mental health and well-being promotion. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12786–12810, Suzhou, China. Association for Computational Linguistics.

Hanwen Shi, Jin Zhang, and Kunpeng Zhang. 2025. Enhancing clinical trial patient matching through knowledge augmentation and reasoning with multi-agent. *arXiv preprint arXiv:2411.14637*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. 2024. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339, Miami, Florida, USA. Association for Computational Linguistics.

John Wu, Adam Cross, and Jimeng Sun. 2025. RDMA: Cost effective agent-driven rare disease discovery within electronic health record systems. *arXiv preprint arXiv:2507.15867*.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N. Ioannidis, Karthik Subbian, James Zou, and Jure Leskovec. 2024. STaRK: Benchmarking LLM retrieval on textual and relational knowledge bases. In *Advances in Neural Information Processing Systems (NeurIPS) 2024 Datasets and Benchmarks Track*.

Fagen Xie, Deborah S. Ling Grant, John Chang, Britta I. Amundsen, and Rulin C. Hechter. 2022. Identifying suicidal ideation and attempt from clinical notes within a large integrated health care system. *The Permanente Journal*, 26(1):85–93.

Shihao Xu, Tiancheng Zhou, Jiatong Ma, Yanli Ding, Yiming Yan, Ming Xiao, Guoyi Li, Haiyang Geng, Yunyun Han, Jianhua Chen, and Yafeng Deng. 2026. LingxiDiagBench: A multi-agent framework for benchmarking LLMs in Chinese psychiatric consultation and diagnosis. *arXiv preprint arXiv:2602.09379*.

Wei Zhai, Nan Bai, Qing Zhao, Jianqiang Li, Fan Wang, Hongzhi Qi, Meng Jiang, Xiaoqin Wang, Bing Xiang

Yang, and Guanghui Fu. 2025. MentalGLM series: Explainable large language models for mental health analysis on Chinese social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12968–12990.

Fuheng Zhao, Divyakant Agrawal, and Amr El Abbadi. 2025. Hybrid querying over relational databases and large language models. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*.