Error-Aware Text-to-SQL Generation for Clinical Trial Eligibility Criteria Querying in EHR Databases

Supriya Kottam
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
s_kottam@uncg.edu

Sony Annem
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
annemsony.137@gmail.com

Yingcheng Sun*
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
y_sun4@uncg.edu

***Abstract*—Accurate patient-cohort retrieval from Electronic Health Records (EHRs) is essential for effective clinical trial recruitment. However, converting eligibility criteria—which specify numerous clinical and demographic characteristics of patients—into structured SQL queries remains a complex challenge due to domain-specific terminology, implicit logic, and error propagation across pipeline stages. These criteria are often lengthy and complex, with inclusion and exclusion sections spanning more than 10 lines each, containing nested conditions that go well beyond simple one- or two-line criteria. While existing pipelines such as Criteria2Query modularize this task into Named Entity Recognition (NER), Relation Extraction (RE), and SQL generation, they often suffer from uncorrected errors at each stage—such as incorrect entity boundaries or types in NER, missing or misdirected links in RE, and invalid logic in SQL translation. These errors propagate through the pipeline, compromising the validity of cohort selection. To address these limitations, we propose an error-aware clinical text-to-SQL framework with built-in verification and correction at every step. The system comprises three components: (1) a Knowledge-Augmented NER Verifier that leverages UMLS/OHDSI knowledge and LLM-based reasoning to detect and correct entity-level errors; (2) a Hybrid Relation Validator that integrates symbolic rules with contextual reasoning to validate extracted relations; and (3) an Editable SQL Generation Engine that identifies illogical SQL fragments and supports real-time revision. Experimental evaluations demonstrate that our framework significantly mitigates NER and RE errors, achieving NER F1 improvements from 0.622 to 0.704 on the CHIA and from 0.598 to 0.716 on the COVID datasets. End-to-end evaluation shows 75% fully correct SQL queries compared to 45% for baseline systems.**

***Index Terms*—Text-to-SQL, Electronic Health Records (EHR) , Named Entity Recognition (NER), Relation Extraction (RE), Query Validation.**

I. Introduction

Clinical trials are vital for advancing medical knowledge and validating new treatments. A crucial step in this process is patient recruitment, which relies on accurately identifying eligible individuals from large-scale electronic health records (EHRs). However, eligibility criteria are typically described in unstructured natural language, rich in domain-specific terminology,and complex logical constructs. Translating free-text eligibility criteria into structured SQL queries for cohort retrieval requires precise recognition of clinical entities (e.g., conditions, drugs, measurements), accurate identification of their interrelations (e.g., temporal constraints, value restrictions), and the generation of logically valid SQL queries in compliance with the OMOP Common Data Model (CDM) [1], a widely adopted international standard for representing EHRs.

Each stage of the clinical text-to-SQL pipeline is susceptible to errors that propagate downstream. Boundary inaccuracies in Named Entity Recognition (NER), misclassified relations in Relation Extraction (RE), and logical inconsistencies during SQL translation can cascade through the system.Our preliminary analysis of eligibility criteria from ClinicalTrials.gov indicates frequent entity-recognition and relation errors, with many SQL query failures traceable to these upstream mistakes. Such cascading errors can result in incorrect patient cohort identification, underscoring critical weaknesses in current clinical text-to-SQL pipelines.

Current approaches to clinical text-to-SQL generation share a critical methodological limitation: they optimize component-level metrics in isolation rather than ensuring these improvements translate into end-to-end clinical utility. Early systems like ERGO, EliXR, and EliIE [2]–[4] focused on atomic sentence-level fragments, while neural approaches such as TREQS [5] and Criteria2SQL [6] mapped individual conditions into SQL snippets. More recent interactive systems like Criteria2Query [7] and its LLM-based variant C2Q 3.0 [8] have moved closer to trial-level queries; yet, their evaluation has remained focused on component metrics. To our knowledge, existing systems have not systematically examined how upstream errors in the text-to-SQL pipeline propagate into SQL generation failures, and they lack verification mechanisms to prevent such cascades, leaving them vulnerable to silent failures that compromise trial validity

The core challenge is that current clinical text-to-SQL systems lack systematic verification to detect and correct errors before they propagate, and most evaluations emphasize intermediate metrics rather than end-to-end cohort accuracy.

To address these gaps, we present an error-aware, verification-driven framework for translating clinical trial eligibility criteria into OMOP-CDM SQL queries. The framework incorporates verification at each stage—NER, Relation

*Corresponding author: Yingcheng Sun (y_sun4@uncg.edu).

Extraction, and SQL generation—to mitigate error cascades, ensure that all relevant SQL clauses are systematically covered through rule-based mapping, and enhance transparency in query construction. It consists of three key components: a Knowledge-Augmented NER Verifier (UMLS/OHDSI + LLM reasoning), a Hybrid Relation Validator (symbolic rules + contextual checks), and an Editable SQL Generator using OMOP-compliant templates..

Our contributions are as follows: (1) a systematic taxonomy of error types in clinical text-to-SQL pipelines, with an analysis of their impact on patient cohort selection; (2) a verification framework that prevents silent failures through multi-stage validation; (3) a rule-based SQL generation engine that ensures systematic coverage of eligibility criteria; and (4) an evaluation methodology that jointly measures intermediate accuracy and end-to-end clinical utility.

II. Related Work

A. Clinical Text-to-SQL Generation

Clinical text-to-SQL systems have employed diverse approaches, from rule-based methods to modern neural architectures. Early systems such as ERGO [2], EliXR [3], and EliIE [4] relied on grammar-based rules to formalize eligibility criteria, improving annotation consistency but focusing on sentence-level fragments rather than full trial queries. Neural approaches like TREQ [5] and Criteria2SQL [6] introduced deep learning for mapping clinical conditions to SQL, but were limited to short snippets (e.g., "Age $\geq$ 18 years").

Interactive systems like Criteria2Query (C2Q) [7] and its LLM-based successor C2Q 3.0 [8], advanced toward trial-level queries with user-guided interfaces, introducing systematic SQL error categorization. However, these systems still evaluate mainly at the component-level metrics rather than patient-level cohort accuracy, leaving a gap between SQL correctness and clinical utility.

B. NER and Relation Extraction in Eligibility Criteria

The CHIA dataset provides a comprehensive benchmark for clinical trial eligibility criteria, with 12,409 annotated criteria containing 41,487 entities across 15 types and 25,017 relations spanning 12 relation types [9]. Performance on CHIA reveals the domain's complexity: PubMedBERT achieves only 0.622 F1 for NER [10], compared to over 85%+ on general biomedical tasks, due to nested logic, temporal constraints, and domain-specific abbreviations. CHIA's 25,000+ relational annotations covering temporal, conditional, and logical constructs present opportunities for advancing RE methods, although systematic benchmarking remains underexplored.

C. Verification Approaches in Clinical NLP

Verification strategies in clinical NLP have been explored for individual tasks but rarely across full pipelines. For NER, ensemble disagreement methods [11] and rule-based consistency checks [12] have been used to detect annotation errors, while more recent work such as VerifiNER [13] demonstrated the effectiveness of knowledge-grounded verification using large language models. For relation extraction, studies in general and biomedical domains (e.g., TACRED [14], [15], DrugProt [16]) have shown that rule-based and ensemble-based validation can improve reliability, although these methods do not directly address the temporal, negation, or value constraints central to trial eligibility. For SQL generation, SQLCritic [17] and TrustSQL [18] explored verification at the generation stage. However, these methods address isolated components without preventing error propagation—VerifiNER cannot fix downstream RE errors, while SQLCritic cannot detect upstream NER mistakes that corrupt SQL queries.

Current systems exhibit two key limitations: (1) the lack of systematic verification across all pipeline stages, and (2) evaluation practices that emphasizing component-level metrics alone without validating end-to-end clinical utility. Given that small eligibility-interpretation errors can compromise trial validity and patient safety, integrated error-aware approaches spanning the entire pipeline are essential.

III. Motivation for an Error-Aware Pipeline

The task of converting free-text clinical trial eligibility criteria into executable SQL queries involves a multi-stage pipeline comprising NER, Relation Extraction (RE), and SQL generation. Although modular approaches such as Criteria2Query have formalized this process, they suffer from a critical weakness: error propagation across stages with no verification mechanisms to intercept and correct mistakes. Although studies such as [19] have documented discrepancies between human- and NLP-based annotations of clinical trial eligibility criteria using the OMOP Common Data Model, a systematic analysis of how specific NER and RE error types affect downstream SQL query generation remains understudied. To address this gap, we analyze the relationship between upstream annotation errors and their cascading effects on SQL correctness, providing a foundation for understanding when and how verification mechanisms can prevent such propagation.

A. NER Errors and Patient Safety Implications

Named Entity Recognition errors in translating clinical trial criteria can create significant patient-cohort discrepancies that compromise trial validity and patient safety. To illustrate the potential impact of these errors, we present a systematic taxonomy of error categories and their clinical consequences through representative examples.

1) Scope-Boundary Errors $\rightarrow$ Cohort Precision Collapse: Incorrect segmentation of multi-token clinical expressions during entity recognition causes overly broad cohort selection.

Example: "RECIST 1.1 progressive disease" correctly maps to a single OMOP concept, but erroneous segmentation into "RECIST" + "1.1 progressive disease" creates two generic concept mappings.

Clinical Impact: Intended cohorts targeting patients with RECIST 1.1-confirmed tumor progression expand to include any disease progression mention, risking the enrollment of patients without standardized RECIST evaluation in progression-focused trials.

Correct Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN condition_occurrence co
  ON p.person_id = co.
     person_id
WHERE co.
     condition_concept_id = <
     recist_id>
-- RECIST 1.1 progressive
     disease
```

Erroneous Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN condition_occurrence co
  ON p.person_id = co.
     person_id
WHERE co.
     condition_concept_id IN
     (
  SELECT concept_id
  FROM concept
  WHERE concept_name LIKE '%
     RECIST 1.1 %'
  AND
  WHERE concept_name LIKE '%
     progressive%disease%'
)
```

Fig. 1: Scope-Boundary Errors

Correct Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN drug_exposure de
  ON p.person_id = de.
     person_id
WHERE de.drug_concept_id = <
     ventavis_drug_id>
-- Ventavis as drug
```

Erroneous Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN device_exposure de
  ON p.person_id = de.
     person_id
WHERE de.device_concept_id =
      <ventavis_device_id>
-- Ventavis misclassified as
      device
```

Fig. 2: Entity Type Misclassification

Fig. 1 illustrates these differences through intended versus error-generated SQL queries.

2) Entity-Type Misclassification → Treatment Protocol Violations: Entity-type misclassification maps recognized entities to the wrong OMOP domain table, leading to invalid joins and protocol-level query failures.

Example: "Ventavis" should map to the DRUG_EXPOSURE table, but erroneous assignment to the DEVICE_EXPOSURE table creates a table mismatch.

Clinical Impact: Intended cohorts targeting patients with Ventavis (iloprost) treatment history generate empty result sets, causing complete trial enrollment failure and excluding potentially eligible patients.

Fig. 2 compares the correct and error-generated SQL demonstrating this mismatch.

3) Missing Tags → Dropped Clinical Constraints: Key modifiers (grades/values) not tagged during NER cause RE/SQL to lose critical predicates, making cohorts over-permissive.

Example: "Child–Pugh grade B/C with liver failure" should generate (Measurement: Child–Pugh) + (Value: B or C) + (Condition: liver failure), but missing value "B/C" tags retain only the measurement and condition entities.

Clinical Impact: Intended cohorts targeting Child–Pugh class B or C *and* liver failure expand to any Child–Pugh record (including class A) *or* any liver failure, risking inclusion of well-compensated patients in decompensated cirrhosis trials.

Fig. 3 demonstrates how dropped value constraints lead to overly permissive cohort selection.

Correct Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN measurement m_cp
  ON p.person_id = m_cp.
     person_id
JOIN condition_occurrence co
  ON p.person_id = co.
     person_id
WHERE m_cp.
     measurement_concept_id
     = <
     child_pugh_concept_id>
  AND m_cp.
     value_as_concept_id IN
      (<grade_B_concept_id>,
      <grade_C_concept_id>)
  AND co.
     condition_concept_id =
     <
     liver_failure_concept_id
     >;
```

Erroneous Query

```
SELECT DISTINCT p.person_id
FROM person p
LEFT JOIN measurement m_cp
  ON p.person_id = m_cp.
     person_id
LEFT JOIN
     condition_occurrence co
  ON p.person_id = co.
     person_id
WHERE (m_cp.
     measurement_concept_id
     = <
     child_pugh_concept_id>
       /* Grade filter
     missing: A/B/C all pass
      */)
   AND co.
     condition_concept_id =
     <
     liver_failure_concept_id
     >;
```

Fig. 3: Missing Tags

Correct Query

```
SELECT DISTINCT p.person_id
FROM person p
-- No visit constraints;
     sentence is
     administrative
;
```

Erroneous Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN visit_occurrence v
  ON p.person_id = v.
     person_id
WHERE v.visit_concept_id = <
     followup_visit_id>;
-- Spurious "follow-up" tag
     enforces an unintended
     clinical visit
```

Fig. 4: Unnecessary Tags from Administrative Language

4) Unnecessary Tags → Cohort Contamination: Administrative or logistical phrases are sometimes incorrectly labeled as clinical entities, causing the pipeline to generate unintended joins across OMOP domains.

Example: "Participants will receive a follow-up email with scheduling instructions." contains the term "follow-up", which frequently denotes a clinical visit (OMOP `VISIT_OCCURRENCE`). However, in this context it refers to a routine administrative email and *should not* receive any clinical tag. Erroneously tagging "follow-up" as a Visit introduces a spurious visit constraint into the SQL.

Clinical Impact: Cohort definitions become overly restrictive by requiring participants to have a documented "follow-up visit" in the EHR, even though the sentence describes an administrative communication. This excludes eligible patients lacking such visit records and leads to cohort under-coverage, distorting downstream analyses and feasibility assessments.

Fig. 4 illustrates how non-clinical language, when incorrectly tagged, contaminates the OMOP query logic.

These examples demonstrate that NER errors can fundamentally threaten trial validity and patient safety. Even small tagging mistakes can distort cohort boundaries, inflate or shrink populations, and inappropriately admit or exclude patients. By explicitly linking error categories to clinical consequences, we underscore the necessity of an error-aware verification layer.

B. Relation Extraction Errors

Relation Extraction (RE) establishes logical and semantic links between extracted entities. This stage is highly susceptible to cascading errors originating in NER and also introduces its own challenges. Errors in RE can distort the intended meaning of eligibility criteria, leading to flawed SQL logic.

1) Missing Entity Arguments → Dropped Relations: Important modifiers or values not tagged cause loss of key relations between measurements and conditions, leading to incomplete eligibility logic.

Example: "Child–Pugh grade B/C with liver failure" should generate Measurement (Child–Pugh grade) + Value (B/C) linked to Condition (Liver failure), but missing value tags break the severity link.

Clinical Impact: Intended cohorts targeting patients with liver failure classified as Child–Pugh grade B or C expand to include all liver failure patients regardless of grade, risking inclusion of mild disease patients in advanced liver failure trials.

Fig. 3 demonstrates how missing value tags result in dropped grade filters and overly permissive cohort selection.

2) Unnecessary Relations → Modifier Overextension: Relation Extraction models may overextend numeric or threshold modifiers to multiple entities instead of restricting them to the correct target.

Example: *"Patients with hemoglobin > 13 g/dL and creatinine levels recorded"* should apply threshold "> 13 g/dL" only to *hemoglobin*, but RE models incorrectly attach it to both *hemoglobin* and *creatinine*.

Clinical Impact: Intended cohorts targeting patients with elevated hemoglobin values (>13 g/dL) and any recorded creatinine measurement become artificially narrow, requiring both elevated hemoglobin and elevated creatinine, excluding otherwise eligible participants.

Fig. 5 illustrates this modifier overextension error and its impact on SQL query logic.

3) Incorrect Relations → Misrepresentation of Clinical Definitions: Even when entities are correctly identified, RE systems may assign incorrect relation types that distort intended clinical definitions.

Example: "Anaemia defined as Hb < 9 g/dL" should generate *(Condition: Anaemia, has_value, Measurement: Hb < 9 g/dL)*, but erroneous *subsumes* relation between "Anaemia" and "Hb" ignores the threshold and replaces definitional links with taxonomic ones.

Clinical Impact: Intended cohorts targeting patients meeting clinical anaemia definition (Anaemia < 9 g/dL) become inflated to include patients with any anaemia mention or any hemoglobin measurement regardless of threshold, including non-anaemic patients.

Fig. 6 compares the precise definitional query against the error-generated version that loses critical threshold constraints.

C. Why a Hybrid, Verified Pipeline Is Necessary

Fine-tuned biomedical models such as PubMedBERT [20] and BioBERT [21] achieve 85–92% F1 scores on standard biomedical corpora [22], but performance drops to 60–65% on clinical trial eligibility criteria [10] due to nested Boolean logic, temporal constraints, and underrepresented entities. Large language models offer strong contextual reasoning but cannot reliably navigate OMOP CDM's 50+ tables and 150+ columns. LLMs can hallucinate schema elements, mishandle units or temporal modifiers, and produce unstable logic in which minor prompt changes invert inclusion/exclusion criteria. For NER and RE, fine-tuned models remain the best-performing approaches [22].

A hybrid, verified pipeline is therefore necessary. Fine-tuned models provide high-recall NER and RE with OMOP/UMLS

Correct Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN measurement m_hb
  ON p.person_id = m_hb.
     person_id
WHERE m_hb.
      measurement_concept_id
      = <
      hemoglobin_concept_id>
  AND m_hb.value_as_number >
      13
  AND m_hb.unit_concept_id =
      <gdl_unit_id>
  AND EXISTS (
      SELECT 1
      FROM measurement m_cr
      WHERE m_cr.person_id =
      p.person_id
        AND m_cr.
     measurement_concept_id
     = <
     creatinine_concept_id>
        AND m_cr.
     value_as_number IS NOT
     NULL   -- any recorded
     creatinine
  );
```

Erroneous Query

```
SELECT DISTINCT p.person_id
FROM person p
JOIN measurement m_hb
  ON p.person_id = m_hb.
     person_id
WHERE m_hb.
      measurement_concept_id
      = <
      hemoglobin_concept_id>
  AND m_hb.value_as_number >
      13
  AND m_hb.unit_concept_id =
      <gdl_unit_id>
  AND EXISTS (
      SELECT 1
      FROM measurement m_cr
      WHERE m_cr.person_id =
      p.person_id
        AND m_cr.
     measurement_concept_id
     = <
     creatinine_concept_id>
        AND m_cr.
     value_as_number > 13
           -- incorrect: Hb
      threshold reused
        AND m_cr.
     unit_concept_id = <
     gdl_unit_id>  --
     incorrect: Hb units
     reused
  );
```

Fig. 5: Unnecessary Relations (Modifier Overextension)

Correct Query

```
SELECT person_id
FROM condition_occurrence co
JOIN measurement m
  ON m.person_id =
     co.person_id
WHERE
      co.condition_concept_id
      = <anaemia_concept_id>
  AND
      m.measurement_concept_id
      =
      <hemoglobin_concept_id>
  AND m.value_as_number < 9
  AND m.unit_concept_id =
      <gdl_unit_id>;
```

Erroneous Query

```
SELECT person_id
FROM condition_occurrence co
WHERE
      co.condition_concept_id
      = <anaemia_concept_id>
   OR EXISTS (
        SELECT 1
        FROM measurement m
        WHERE m.person_id =
      co.person_id
          AND
      m.measurement_concept_id
      =
      <hemoglobin_concept_id>
   );
-- Threshold lost; any
     hemoglobin measurement
     included
```

Fig. 6: Incorrect Relations

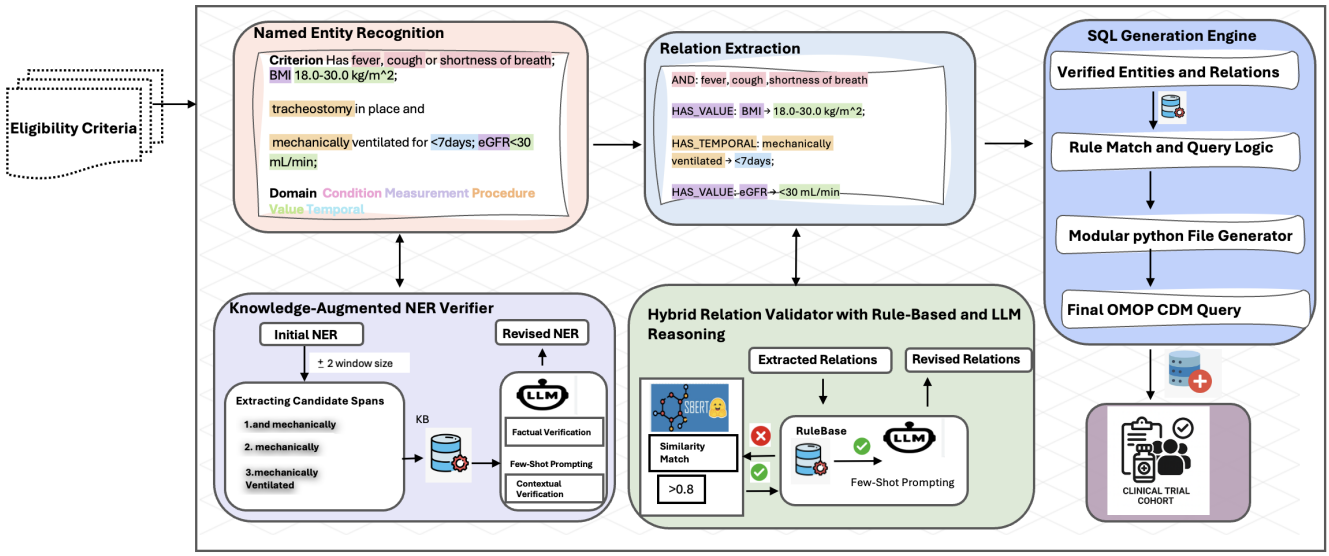

Fig. 7: End-to-end verification-driven pipeline that transforms eligibility criteria into OMOP-compliant SQL queries.

grounding, while LLMs handle contextual disambiguation under schema constraints. A dedicated verification layer enforces unit normalization, temporal scoping, Boolean composition, SQL-structure validation, and execution-time sanity checks. This design combines the reliability of supervised extraction with the flexibility of LLM reasoning while mitigating the weaknesses of both families. In safety-critical settings where a single misclassified entity can change cohort size by an order of magnitude, only a schema-constrained, verification-driven hybrid approach delivers clinical safety.

IV. Framework

To address the cascading error vulnerabilities identified above, we introduce a verification-driven pipeline that prevents error propagation through three strategic intervention points (Figure 7). The central insight is that, rather than relying on post hoc error correction, systematic prevention at each stage—entity-boundary refinement, relation-semantic validation, and rule-based SQL assembly—can eliminate the cascading failures that undermine existing clinical text-to-SQL systems. This pipeline integrates knowledge-grounded verification (UMLS/OHDSI), contextual reasoning through LLM consensus, and transparent rule-based mapping to transform noisy intermediate outputs into clinically validated, OMOP-compliant queries.

A. Knowledge-Augmented NER Verifier

To address the entity-level errors identified in Section III-A—scope boundary errors, entity type misclassification, missing tags, and unnecessary tags—we developed a Knowledge-Augmented NER Verifier (Figure 8) that prevents these cascading failures before they compromise downstream relation extraction and SQL generation. Inspired by VerifiNER [22], our domain-specific verification system combines UMLS/OHDSI knowledge enrichment with LLM-based contextual reasoning to systematically correct entity mistakes that could otherwise exclude 30–40% of appropriate patients from clinical trials. The verifier post-processes noisy NER predictions into clinically validated entities via four sequential validation steps.

- **Candidate Span Expansion:** Given input text, we first extract initial entity predictions using a fine-tuned NER model (e.g., a PubMedBERT-based architecture). For each predicted entity, alternative span candidates are generated using a symmetric ± 2-token window. This ensures comprehensive boundary coverage while avoiding excessive computational overhead.
- **Knowledge Base Lookup and Enrichment:** Each candidate span is matched against a curated UMLS+OHDSI knowledge base to retrieve Concept Unique Identifiers (CUIs), semantic types, preferred names, textual definitions, and domain assignments. This information is incorporated into LLM prompts to guide entity validation and resolve ambiguities.
- **LLM-Based Factual Verification:** A large language model evaluates the semantic correctness of each candidate span using the enriched knowledge-base metadata. Spans inconsistent with medical definitions or domain classifications are discarded, ensuring that only factually grounded candidates are retained. The key prompt struc-

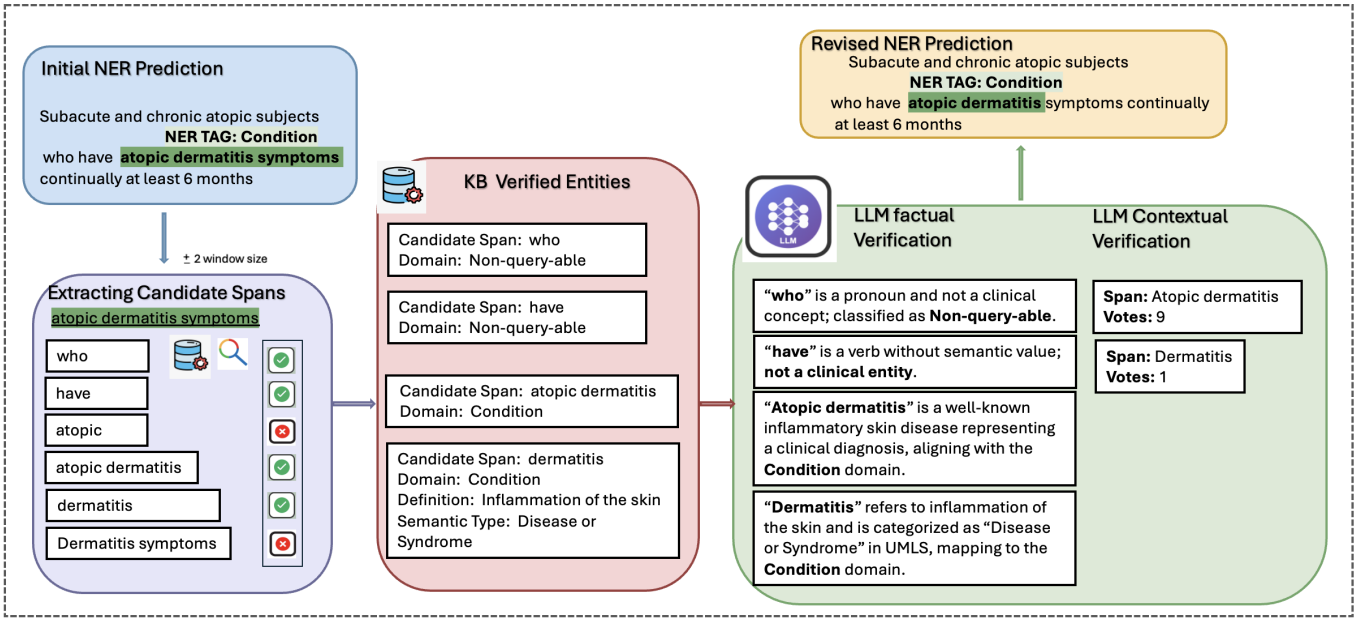

Fig. 8: Knowledge-Augmented NER Verification. Initial NER predictions undergo span expansion, UMLS/OHDSI knowledge enrichment, LLM factual verification, and consensus voting to produce validated entities.

ture is illustrated below:

> **Prompt Structure**
> – Candidate span
> – UMLS/OHDSI metadata: CUI, preferred term, semantic type, definition, synonyms
> – Assigned domain from knowledge base
> – **Task:** Verify factual consistency of the span with the definition and domain, and return **Final Correct Domain** plus a brief **Explanation**.

- **Contextual Voting and Disambiguation:** For the remaining spans, we employ a 10-fold LLM voting strategy. Multiple prompts with varied reasoning styles and temperature sampling evaluate contextual fit within the sentence. The candidate achieving $\geq 7/10$ consensus is selected as the final entity.

This hybrid verification process combines lexical precision from knowledge bases, semantic validation through factual checks, and contextual disambiguation via multi-query LLM reasoning. By operating upstream of relation extraction, the verifier prevents the error cascades documented in Section III-A and ensures that only clinically validated entities proceed to downstream processing stages.

B. Hybrid Relation Validator with Rule-Based and LLM Reasoning

To address the relation extraction errors identified in Section III-B—including missing relations from dropped NER tags, spurious links from modifier overextension, and incorrect types from semantic misassignment—we developed a Hybrid Relation Validator (Figure 9). This component ensures that only clinically valid relations reach SQL generation by combining symbolic rules with LLM-based semantic checks.

- **Rule Base Construction.** We built a comprehensive rule base of **6,247 symbolic rules** from CHIA and COVIC corpora (covering **15,700 sentences**, **30,000 entities**, **25,000+ relations**), clustered using SBERT embeddings for semantic consistency. Each rule specifies:
  - `relation_type` (e.g., has_value, has_temporal)
  - source/target domains (e.g., person → value)
  - text patterns (e.g., "n1–n2 years", "at least n months")
  - OMOP-compliant SQL clause template
- **Rule-Based Candidate Validation.** Relations extracted by a fine-tuned model (e.g., a BioBERT) are matched against the rule base. Exact text pattern matches are accepted directly, while unmatched cases are resolved through SBERT similarity (>0.8) to retrieve approximate rule patterns. This two-step process handles lexical variation while filtering spurious candidates.
- **LLM-Based Semantic Validation with Voting.** All candidate relations from a sentence are jointly validated by an LLM. The model is prompted ten times with sampling enabled, and majority voting (threshold $\geq 7/10$) determines which relations are contextually valid. This approach ensures semantic robustness and filters out inconsistent or low-confidence relations, and the key prompt structure is illustrated below:

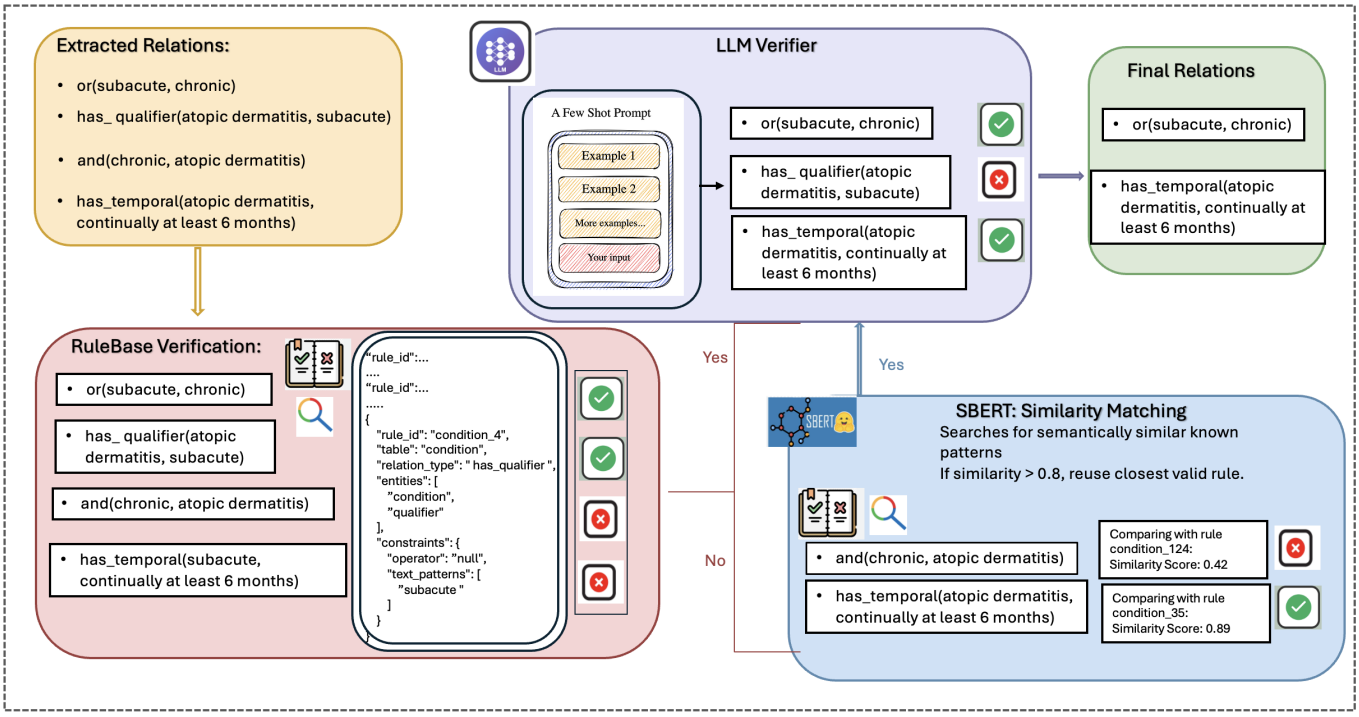

Fig. 9: Hybrid Relation Validation. Extracted relations undergo rule-based validation, SBERT similarity matching for unmatched patterns, and LLM consensus voting before being passed to SQL generation.

Prompt Structure
– Full sentence containing the candidate relations
– Extracted entities with their domains
– Candidate relations (e.g., has_value, has_temporal)
– Matched rule or pattern from rule base
– **Task:** Verify whether each relation is contextually correct. Return **Valid/Invalid** and a brief **Explanation**.

- **Directionality and Inversion Checks.** Asymmetric relations (e.g., temporal constraints) are further checked for correct ordering. For example, in "within 5 years after diagnosis," the validator ensures correct alignment of source and target entities, rejecting inverted or contradictory matches.

The validator outputs refined relation triples that are syntactically grounded (via rules) and semantically validated (via LLM consensus). These verified relations are then mapped to OMOP-compliant SQL fragments, significantly reducing cascading errors from upstream components.

C. SQL Generation Engine

The SQL generation module translates verified NER and RE outputs into executable SQL queries for OMOP-compliant EHR cohort retrieval. Existing systems such as CIRCEbe rely on rigid templates and opaque pipelines that are difficult to debug, adapt, or extend, and are not resilient to vocabulary drift or OMOP CDM schema changes.

To address these limitations, we introduce a modular and rule-aware **SQL Generation Engine** designed to support transparent, editable, and scalable SQL construction. Our approach leverages:

- **Rule-Driven Query Mapping.** 6,000+ curated rules each contain source/target domains (e.g., `condition` → `value`), normalized query keys (e.g., age $\geq n1$ and $\leq n2$), dozens of clinical text patterns, and exact OMOP SQL fragments. Validated entity-relation pairs are matched against this rule base to extract SQL snippets, which are sent to corresponding SQL builder modules for embedding into larger OMOP cohort query structures (e.g., `Inclusion_0`, `qualified_events`, `final_cohort`).
- **Modular SQL Builders.** Table-level modularity with dedicated builders processing verified entities and domains, operator logic (e.g., BETWEEN, $\geq$, $=$), and SQL rule fragments:
  – `personoper.py`: age, sex, race, ethnicity
  – `con.py`, `conOc.py`: condition-based filtering with frequency and temporality
  – `Drugera.py`: medication and drug exposure logic
  – `Measurement.py`: lab results and numeric thresholds
- **OMOP-Compliant Query Assembly.** All modules generate subqueries that confirm to OHDSI cohort definition patterns. Assembly follows standard OMOP CDM pro-

cedures: creating code sets and primary event timelines, applying correlated criteria groups, generating inclusion rules and computing rule masks, finalizing cohort entry and exit using temporal anchors. This produces reproducible, interpretable SQL matching OHDSI's logic but fully decomposed into editable pieces.
- **Fallback Handling and Entity-Only Clauses.** Supports entity-only clauses for standalone conditions (e.g., "patients with type 2 diabetes mellitus") lacking explicit relations, using condition-only SQL templates and builder logic that adds conditions when relation-based triggers are absent.

Unlike black-box generators, our engine traces lineage from NER tags through verified relations to SQL rules and final cohort logic. All templates and mappings stored in editable, version-controlled JSON files.

- Audit or modify logic for specific clinical domains
- Track changes in OMOP schema and update SQL templates accordingly
- Extend the system by adding new relation types and tables

By combining semantic precision from NER and RE, symbolic rule mapping, and modular SQL builders, our engine ensures that every eligibility criterion—no matter how complex—can be translated into interpretable, executable SQL aligned with OMOP CDM standards.

V. Evaluation

We evaluate our framework on both component-level extraction tasks (NER and RE) and end-to-end query generation, using standard benchmarks and a cohort-retrieval assessment. Our experiments address two main questions: (1) Does the verification-driven pipeline improve NER and RE performance over strong baselines? (2) Does this improvement translate into more correct and complete SQL queries—and thus more accurate patient-cohort retrieval—compared with a conventional pipeline without verification? While component-level metrics are necessary to understand pipeline behavior, they are insufficient for clinical validation. A system can achieve high NER F1 scores while still producing clinically incorrect SQL due to error propagation. Our evaluation therefore employs a two-tier approach: F1 scores demonstrate that the verification layer improves individual components, whereas end-to-end query accuracy measures whether these gains translate into better clinical outcomes.

A. Implementation Details

1) Base Models:

- **NER Model:** PubMedBERT-base-cased fine-tuned on the CHIA and COVID datasets with learning rate 2e-5, batch size 16, and 5 epochs using AdamW optimizer with linear warmup.
- **RE Model:** BioBERT-base with a classification head trained on CHIA and COVID relations using identical hyperparameters.

2) LLM Configuration: We employed four state-of-the-art language models for entity and relation verification:

- **Primary Models:** GPT-4o (gpt-4-0613), DeepSeek-V3-0324, Gemini 2.5 Pro, and Claude-3-Opus-20240229.
- **Verification Protocol:** For NER, factual verification prompts were executed once with temperature 0.3, and contextual reasoning prompts were executed 10 times with temperature 0.7. Majority voting with threshold $\geq$7/10 agreement was applied to the contextual outputs to select high-confidence results. For RE, factual verification was handled through rule-based checks, and only contextual reasoning prompts were executed ten times with majority voting.
- **Cost Management:** Based on average token usage across models, the per-instance verification cost was approximately 0.12, with a total evaluation cost of around 850. Actual costs varied by model owing to differences in provider pricing (OpenAI, DeepSeek, Google, Anthropic).

3) Knowledge Base Integration:

- **UMLS/OHDSI Integration:** 4.2M medical concepts with CUI mappings, semantic types, and OMOP domain assignments for entity validation.
- **Rule Base:** 6,247 symbolic rules derived from a systematic CHIA pattern analysis, organized into 312 semantic clusters using SBERT similarity (threshold> 0.8).

B. Datasets and Experimental Setup

We conducted experiments using two benchmark corpora for training and testing:

- **CHIA corpus** [9]: Contains 12,409 eligibility criteria from 1,000 clinical trials, annotated with 41,487 entities and 25,017 relations. The CHIA Annotation Model (CAM) aligns entities and relations with OHDSI OMOP CDM domains, providing a standardized basis for cohort retrieval.
- **COVIC corpus** [23]: Derived from 700 COVID-19 clinical trials, with 9,767 eligibility criteria annotated into 18,161 entities (across 8 domains) and 16,443 relations (across 11 types). It follows a hierarchical OMOP CDM–based schema, capturing COVID-specific constructs such as viral load thresholds and quarantine timelines.

For NER, we selected PubMedBERT as our baseline on the CHIA dataset, consistent with prior comparative studies in [10] showing it to be the strongest performer (strict F1 ≈ 0.622) among transformer-based models for eligibility-criteria extraction. To enhance robustness, we applied our Knowledge-Augmented NER Verifier, powered by multiple LLMs, as a post-processing layer to refine predictions and mitigate boundary and type errors.

For RE, we selected BioBERT as the backbone for relation extraction, owing to its strong performance across biomedical NLP tasks, including relation classification, and its pre-training on large-scale PubMed and PMC corpora. While PubMedBERT has advantages for NER, BioBERT has been more

TABLE I: NER model performance comparison across datasets. Standalone LLMs are shown for reference; our best verifier combines PubMedBERT with DeepSeek-V3.

Model	CHIA			COVID		
	P	R	F1	P	R	F1
PubMedBERT	0.606	0.639	**0.622**	0.590	0.632	**0.598**
GPT-4o	0.502	0.441	0.468	0.517	0.452	0.482
DeepSeek-V3	0.489	0.438	0.462	0.498	0.461	0.479
Gemini 2.5 Pro	0.471	0.453	0.462	0.505	0.467	0.485
Claude-3-Opus	0.495	0.459	0.476	0.512	0.471	0.491
LLM+PubMedBERT(NER Verifier)	**0.744**	**0.685**	**0.704**	**0.784**	**0.704**	**0.716**

widely benchmarked for RE, providing a stable baseline for our framework.

For NER and RE evaluation, we used 100 *unseen* trials sampled from ClinicalTrials.gov; none were part of training, ensuring a fair test of generalization. For end-to-end evaluation, no gold-standard dataset of SQL queries aligned with eligibility criteria exists. We therefore conducted a manual expert assessment on a separate set of 20 real trials (each with 8-12- line eligibility sections), comparing our framework with a Criteria2Query-style baseline.

C. NER Performance

Table I reports NER results on CHIA and COVID. The baseline PubMedBERT achieved F1 $= 0.622$ and 0.598, respectively. Stand-alone LLMs (GPT-4o, Gemini 2.5 Pro, Claude-3-Opus) underperformed, highlighting the importance of domain-specific pre-training.

Our best results came from a Knowledge-Augmented Verifier that combined PubMedBERT with DeepSeek-V3, improving F1 to 0.704 (CHIA) and 0.716 (COVID). Most gains arose from correcting temporal expressions, numeric qualifiers, and drug–device confusions, thereby reducing cohort misclassification risk.

D. Relation Extraction Performance

Table II shows RE results on CHIA and COVID. The baseline BioBERT achieved F1 $= 0.396$ and 0.432, respectively. Stand-alone LLMs underperformed relative to BioBERT, but provided complementary reasoning signals.

Our best results came from the BioBERT+Claude-3-Opus Relation Validator, which improved F1 to 0.594 (CHIA) and 0.606 (COVID). Gains were strongest for *has_temporal* relations (e.g., symptom duration, infection windows), where correctness rose from 72% to 90%. Using verified NER outputs further increased pipeline-level RE F1 from 0.45 to 0.56, confirming the benefit of error-aware verification across both entities and relations.

E. End-to-End Query Accuracy

Because no gold-standard dataset of SQL aligned with eligibility criteria exists, we manually evaluated 20 clinical trials sampled from ClinicalTrials.gov. Both our verification-driven framework and a Criteria2Query [24] baseline We used the publicly available Criteria2Query interface: http://34.70.212.14:8080/criteria2query_test/#. were used to generate SQL queries. Two clinical specialists independently reviewed the outputs, with disagreements resolved by a third expert. Queries were scored as *fully correct* if they captured all inclusion/exclusion logic without spurious conditions, and as *partially correct* if they provided incomplete but faithful coverage.

TABLE II: RE results on CHIA and COVID; best verifier = BioBERT+Claude-3-Opus.

Model	CHIA			COVID		
	P	R	F1	P	R	F1
BioBERT	0.376	0.419	**0.396**	0.403	0.441	**0.432**
GPT-4o	0.212	0.297	0.254	0.241	0.315	0.274
DeepSeek-V3	0.198	0.283	0.236	0.227	0.302	0.260
Gemini 2.5 Pro	0.205	0.289	0.241	0.234	0.311	0.268
Claude-3-Opus	0.218	0.301	0.256	0.246	0.322	0.280
LLM+BioBERT (Relation Validator)	**0.601**	**0.587**	**0.594**	**0.612**	**0.594**	**0.606**

TABLE III: End-to-End SQL Query Accuracy on 20 Clinical Trials.

Metric	Baseline (Criteria2Query)	Our Framework
Fully Correct Queries	9 / 20 (45%)	**15 / 20 (75%)**
Concept Coverage	74%	**93%**
Avg. Conditions Omitted	2.1	**0.4**
Avg. Spurious Conditions	1.0	**0.0**
Boolean / Temporal Errors	40% of queries	**0%**
Patient Retrieval Accuracy	82% (11% recall, +7% FP)	**90% (18 / 20 correct)**

Traditional pipelines often suffer from error cascades, in which high component-level F1 scores do not guarantee clinically valid SQL. Our verification layer prevents error propagation, ensuring that component gains translate into end-to-end improvements.

We assessed two aspects: (1) **SQL-query correctness**—whether the generated SQL faithfully captured eligibility logic–and (2) **patient-retrieval accuracy**—whether the query returned the intended cohort despite minor SQL imperfections. Table III summarizes the results. Our framework outperformed the baseline, eliminating Boolean and temporal errors and yielding more consistent cohort retrieval.

To test downstream impact, we executed both sets of queries on a SynPUF-5% dataset. Our framework's queries closely matched ground truth, Whereas the baseline produced recall deficits and false positives. An Ablation study showed that removing the NER Verifier reduced success to 60%, and removing the Relation Validator reduced it to 64%. Remaining failures involved deeply nested temporal constraints (e.g., "within 14 days of X after Y"), suggesting areas for future rule extensions.

Our error-aware framework demonstrates strong precision and recall for both Named Entity Recognition (NER) and Relation Extraction (RE) on clinical-trial eligibility criteria and yields substantially higher end-to-end query accuracy (75% vs. 45% fully correct queries). By preventing error propagation across pipeline stages, the system produces SQL that more faithfully represents trial intent, thereby enabling more reliable automated patient-cohort retrieval. Although our evaluation focused on eligibility criteria mapped into ten OMOP-CDM tables, these cover the majority of frequently used domains (*Condition*, *Drug*, *Procedure*, *Measurement*). Extending to

the full OMOP CDM requires only the addition of table-specific templates, as the verification layer already supports the complete range of SQL clauses. This modular design makes the framework not only effective but also reproducible and adaptable across diverse clinical-trial domains.

VI. Conclusion

In this work, we presented a verification-driven framework for converting clinical trial eligibility criteria into OMOP-CDM-compliant SQL queries that systematically addresses error propagation through knowledge-augmented verification and correction mechanisms. Experimental evaluation demonstrates substantial improvements over strong baselines across both component-level tasks and end-to-end query generation, with particularly notable gains in concept coverage and elimination of spurious conditions when compared against the Criteria2Query baseline. Our framework's modular design facilitates the creation of higher-quality text-to-SQL training data, addressing a critical resource gap in clinical NLP where large-scale, OMOP-compliant datasets are limited. Future work will focus on scaling the framework to multilingual eligibility criteria, expanding rule coverage for complex temporal constraints,extending to the full OMOP CDM and integrating continuous learning mechanisms to support deployment in real-world clinical trial platforms and EHR environments. To support reproducibility and community adoption, we provide the complete dataset, code, and experimental protocols will be made publicly accessible upon publication.

Acknowledgments

This work was supported by the 2025 UNCG Chancellor's Initiative for Transformative Research (CITR) Award, the 2024 Dorothy Munroe Student Research Fund, and the 2025 Dorothy Munroe Student Research Fund.

References

[1] OMOP-CDM. 2023. Omop cdm common data model. https://ohdsi.github.io/CommonDataModel/cdm54.html

[2] S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim, "A practical method for transforming free-text eligibility criteria into computable criteria," J. Biomed. Inform., vol. 44, no. 2, pp. 239–250, Apr. 2011, doi: 10.1016/j.jbi.2010.09.007.

[3] C. Weng, X. Wu, Z. Luo, M. R. Boland, D. Theodoratos, and S. B. Johnson, "EliXR: An approach to eligibility criteria extraction and representation," J. Am. Med. Inform. Assoc., vol. 18, no. Suppl 1, pp. i116–i124, Dec. 2011, doi: 10.1136/amiajnl-2011-000321.

[4] T. Kang, S. Zhang, Y. Tang, G. W. Hruby, A. Rusanov, N. Elhadad, and C. Weng, "EliIE: An open-source information extraction system for clinical trial eligibility criteria," J. Am. Med. Inform. Assoc., vol. 24, no. 6, pp. 1062–1071, Nov. 2017, doi: 10.1093/jamia/ocx019.

[5] P. Wang, T. Shi, and C. K. Reddy, "Text-to-SQL generation for question answering on electronic medical records," in Proc. Web Conf. (WWW), Taipei, Taiwan, Apr. 2020, pp. 350–361, doi: 10.1145/3366423.3380217.

[6] X. Yu, T. Chen, Z. Yu, H. Li, Y. Yang, X. Jiang, and A. Jiang, "Dataset and enhanced model for eligibility criteria-to-SQL semantic parsing," in Proc. 12th Lang. Resour. Eval. Conf. (LREC), Marseille, France, May 2020, pp. 5829–5837. [Online]. Available: https://aclanthology.org/2020.lrec-1.714/.

[7] C. Yuan, P. B. Ryan, C. Ta, Y. Guo, Z. Li, J. Hardin, R. Makadia, P. Jin, N. Shang, T. Kang, and C. Weng, "Criteria2Query: A natural language interface to clinical databases for cohort definition," J. Am. Med. Inform. Assoc., vol. 26, no. 4, pp. 294–305, Apr. 2019, doi: 10.1093/jamia/ocy178.

[8] J. Park, Y. Fang, C. Ta, G. Zhang, B. Idnay, F. Chen, D. Feng, R. Shyu, E. R. Gordon, M. Spotnitz, and C. Weng, "Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation," J. Biomed. Inform., vol. 154, p. 104649, Jun. 2024, doi: 10.1016/j.jbi.2024.104649.

[9] Kury F, Butler A, Yuan C, Fu L, Sun Y, Liu H, Sim I, Carini S, Weng C. Chia, a large annotated corpus of clinical trial eligibility criteria. Sci Data. 2020;7(1):281. doi:10.1038/s41597-020-00610-0. PMID: 32814780; PMCID: PMC7428402.

[10] Li J, Wei Q, Ghiasvand O, Chen M, Lobanov V, Weng C, Xu H. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. BMC Med Inform Decis Mak. 2022;22(Suppl 3):235. doi:10.1186/s12911-022-01967-7. PMID: 36089275; PMCID: PMC9456392.

[11] F. Reiss, H. Xu, B. Cutler, K. Muthuraman, and Z. Eichenberger, "Identifying incorrect labels in the CoNLL-2003 corpus," in Proc. 24th Conf. Comput. Natural Language Learning (CoNLL), Online, Nov. 2020, pp. 215–226. [Online]. Available: https://aclanthology.org/2020.conll-1.16/ doi: 10.18653/v1/2020.conll-1.16.

[12] Gabriel Bernier-Colborne and Sowmya Vajjala, "Annotation Errors and NER: A Study with OntoNotes 5.0," arXiv preprint arXiv:2406.19172, 2024. [Online]. Available: https://arxiv.org/abs/2406.19172

[13] S. Kim, K. Seo, H. Chae, J. Yeo, and D. Lee, "VerifiNER: Verification-augmented NER via Knowledge-grounded Reasoning with Large Language Models," arXiv preprint arXiv:2402.18374, 2024. [Online]. Available: https://arxiv.org/abs/2402.18374

[14] T. Alt, M. Gabryszak, and L. Hennig, "TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task," arXiv preprint arXiv:2004.14855, 2020. [Online]. Available: https://arxiv.org/abs/2004.14855

[15] D. Stoica, J. Gardner, C. Potts, and D. Jurafsky, "Re-TACRED: Addressing Shortcomings of the TACRED Dataset," arXiv preprint arXiv:2104.08398, 2021. [Online]. Available: https://arxiv.org/abs/2104.08398.

[16] A. Miranda-Escalada, F. Mehryary, J. Luoma, D. Estrada-Zavala, L. Gasco, S. Pyysalo, A. Valencia, and M. Krallinger, "Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations," Database, vol. 2023, p. baad080, Nov. 2023. doi: 10.1093/database/baad080.

[17] J. Chen, L. Gan, Z. Zhao, Z. Wang, D. Wang, and C. Zhuang, "SQLCritic: Correcting text-to-SQL generation via clause-wise critic," arXiv preprint arXiv:2503.07996, Mar. 2025. [Online]. Available: https://arxiv.org/abs/2503.07996.

[18] G. Lee, W. Chay, S. Cho, and E. Choi, "TrustSQL: Benchmarking text-to-SQL reliability with penalty-based scoring," arXiv preprint arXiv:2403.15879, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.15879.

[19] X. Li, H. Liu, F. Kury, C. Yuan, A. Butler, Y. Sun, A. Ostropolets, H. Xu, and C. Weng, "A comparison between human and NLP-based annotation of clinical trial eligibility criteria text using the OMOP common data model," in *AMIA Summits on Translational Science Proceedings*, vol. 2021, pp. 394, 2021.

[20] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, pp. 1–23, 2021.

[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[22] Chen Q, Hu Y, Peng X, Xie Q, Jin Q, Gilson A, Singer MB, Ai X, Lai PT, Wang Z, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. Nat Commun. 2025;16(1):3280. doi:10.1038/s41467-025-XXXX-X.

[23] Sun Y, Butler A, Stewart LA, Liu H, Yuan C, Southard CT, Kim JH, Weng C. Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials. J Biomed Inform. 2021;118:103790. doi:10.1016/j.jbi.2021.103790. PMID: 34246885; PMCID: PMC8273805.

[24] Y. Fang, B. Idnay, Y. Sun, H. Liu, Z. Chen, K. Marder, H. Xu, R. Schnall, and C. Weng, "Combining human and machine intelligence for clinical trial eligibility querying," Journal of the American Medical Informatics Association, vol. 29, no. 7, pp. 1161–1171, 2022.