

INTRO

- The Problem:** Structured diagnostic codes suffer from low sensitivity and high underreporting in complex behaviors like suicidality.
- The Opportunity:** Large Language Models (LLMs) offer transformative potential for extracting and normalizing unstructured clinical narratives into the **OMOP Common Data Model (CDM)**, but require validation to mitigate model errors and ensure evidentiary reliability.
- The Scale:** This study utilizes the **BIFAP database**, containing over **2 billion unstructured health records** from 30 million patients in Spain.

METHODS

The study implemented a six-step process to transform raw narratives into standardized evidence.

I. Use case definition:

Outcomes were categorized into completed suicide, suicide attempt, suicidal ideation, and not related.

II. Corpus generation:

Large-scale filtering of 2 billion records using 70+ regular expressions to identify event-related text.

- LLM pre-annotation:** Auto-labeling: Qwen-2.5 (14B) for initial classification.
- Human validation:** Expert review of 5,040 texts (>90% inter-annotator agreement).

III. Model training:

Fine-tuning was performed on both encoder (XLM-RoBERTa) and decoder (Qwen family) models.

IV. The Cascaded System:

To balance speed and accuracy, an encoder model acts as the first filter, any record with a confidence score below specific thresholds is delegated to a Qwen-2.5 (7B) decoder or human review.

V. Large-Scale Execution:

The best model was applied across the entire 2-billion-record dataset.

VI. OMOP integration:

Validated events were mapped to **SNOMED CT** concepts and stored in the **Note_nlp** table for global interoperability.

RESULTS

- High-Fidelity performance:** Preliminary results show robust classification, achieving an **F1-score of 0.968** for "completed suicide" and **0.931** for "suicidal ideation".
- Efficiency:** The system maintained high throughput, with a delegation rate of only **9%** for uncertain cases
- Uncertainty:** The "not related" category accounted for the highest level of model uncertainty

Clinician-Driven AI

How 5,000 expert annotations drive LLM precision across BIFAP database (>2 Billion health records)

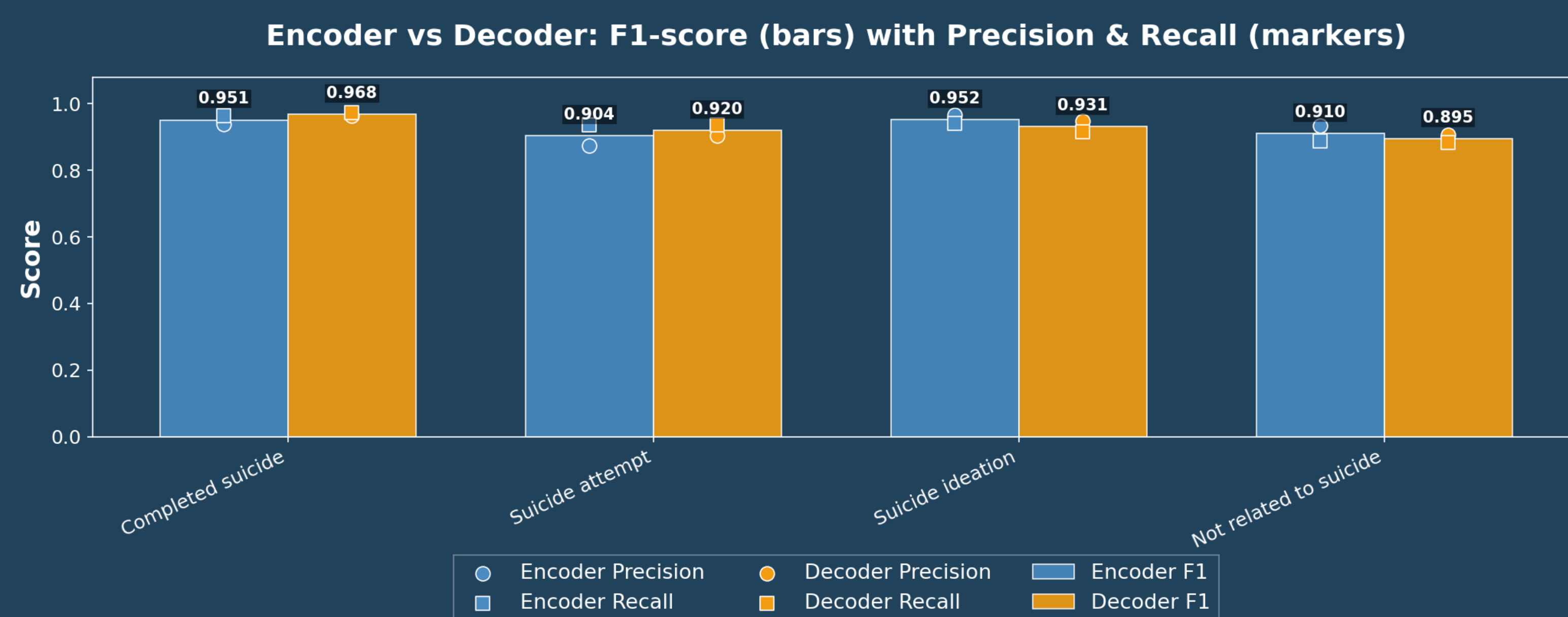
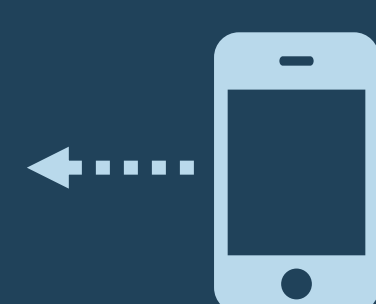


Fig 2. Benchmark evaluation of fine-tuned transformer models: precision, recall, and F1-score across suicidal behavior categories.



Take a picture to visit BIFAP



Fig 1. Workflow of the process.

CONCLUSIONS

This hybrid architecture successfully balances **computational scalability** for multi-billion-record datasets with **high-performance accuracy**. By integrating these outputs into the **OMOP CDM**, we enable more precise cohort definitions and robust epidemiological analyses for public health challenges.

LIMITATIONS & FUTURE WORK

- Population-Level validation**
 - While preliminary results are robust, future efforts will prioritize cross-validation against external, heterogeneous data sources to establish the generalizability of the hybrid approach and its superiority over traditional diagnostic code-based strategies
- Scalability of human review**
 - Although the cascaded architecture successfully filters the vast majority of records, the **9% delegation rate** for uncertain cases presents a significant operational challenge. At a scale of **2 billion records**, manual clinical review of this volume is currently unfeasible, necessitating the development of more advanced automated uncertainty resolution or secondary LLM-based verification layers.

H. Martínez-Alcalá, A. Peñaranda-Navazo, C. Justo-Astorgano, A. Llorente, V. Arroyo, MA. Macía, M. Martín, E. Martín-Merino, M. Gil, J. Santamaría