

A hybrid AI + rules system detects more patient data than any single approach alone, ensuring strict GDPR compliance in bilingual hospitals

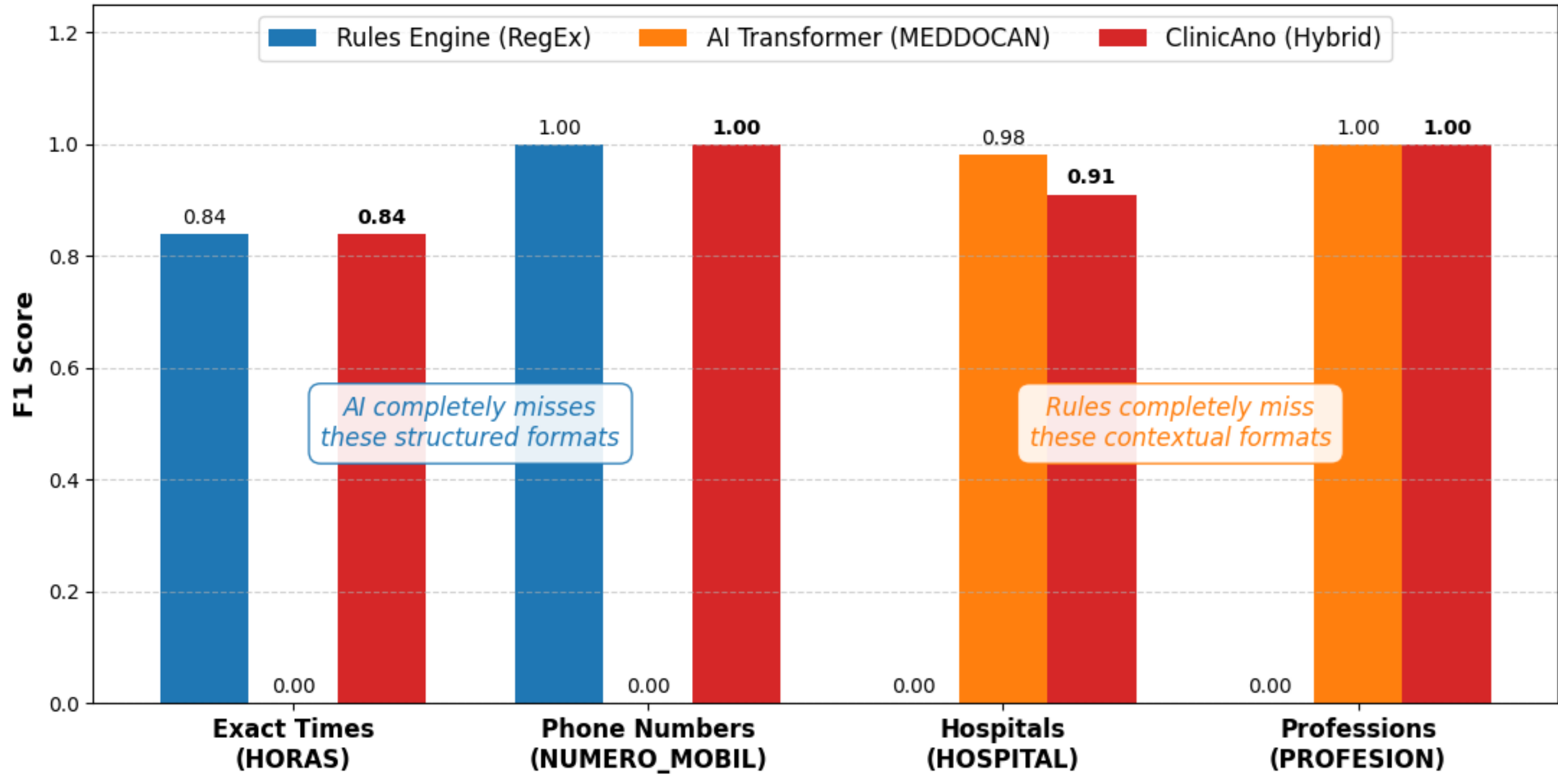
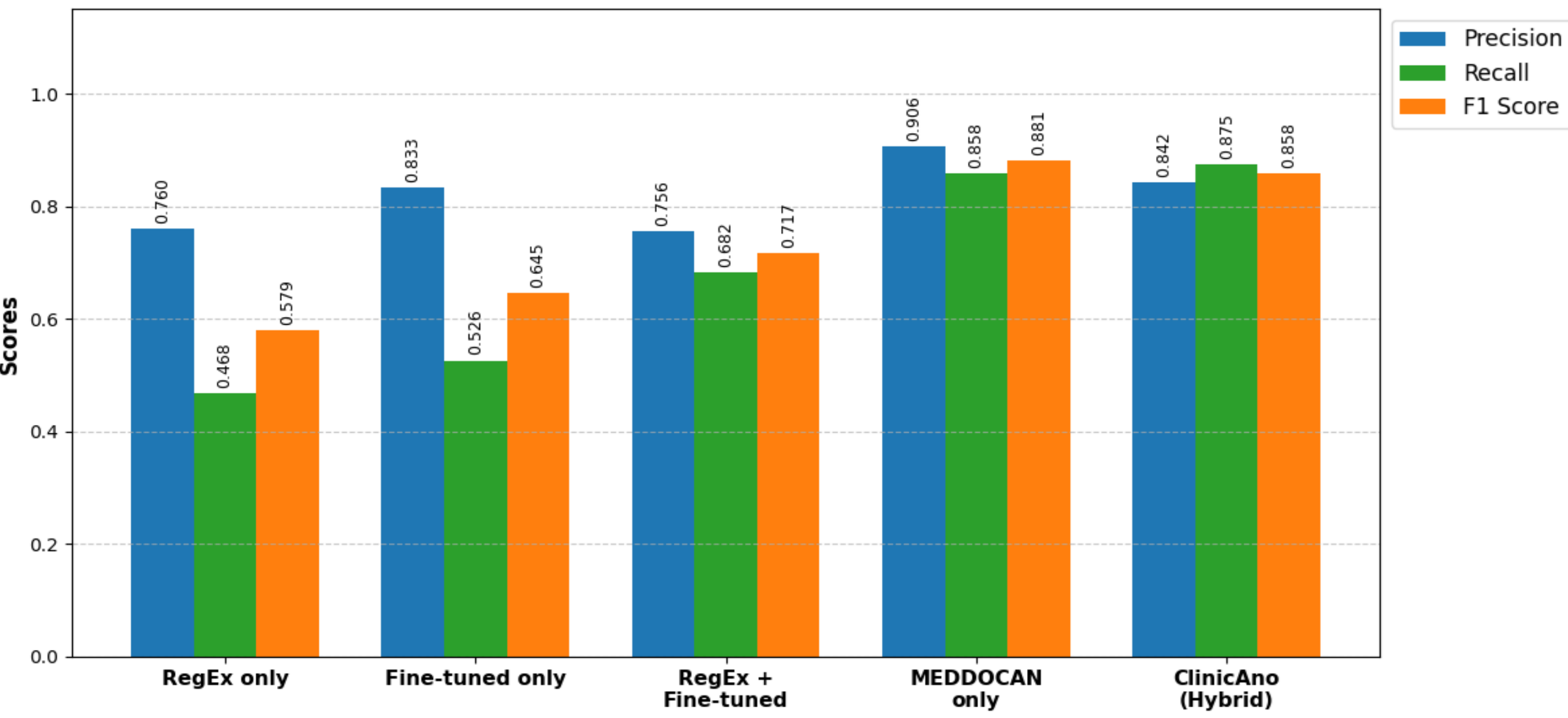
Bilingual Clinical De-Identification in Free-Text Annotations: A Hybrid Approach Combining Biomedical Transformers and heuristic Rules

Background: Clinical notes at Hospital del Mar mix Spanish and Catalan languages, and patient identifiers are sparse. This study presents ClinicaAno, a custom de-identification pipeline designed for Hospital del Mar. We propose a methodology that moves beyond simple fine-tuning, employing data-centric AI techniques to adapt a state-of-the-art Spanish model to the specific dialect and sparsity constraints of our hospital. ClinicaAno combines deterministic rules with a biomedical transformer to safely de-identify free-text.

Result 1 (The Privacy Win): The hybrid system achieves the highest recall (0.875), leaking the least patient data. The Transformer alone scores higher F1 (0.881), but in privacy, missing a name is worse than over-redacting.

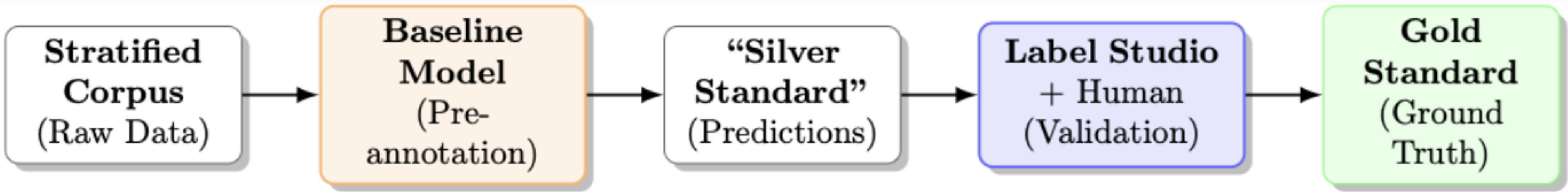
Result 2 (Complementarity):

- Transformer detects context (Hospitals, Professions: F1=1.00)
- Rules Engine catches formats the AI misses (Phone numbers, Exact times: F1=1.00).
- Neither works alone.



Methods

- 1 Stratified Active Sampling:** Scanned 5,000 notes; extracted the top 100 Spanish & 100 Catalan notes with the highest PHI density.
- 2 Synthetic Header Injection:** Added fake patient headers to training data so the AI wouldn't ignore sparse names.
- 3 Silver Standard Annotation:** AI pre-labeled the 200 notes; human experts corrected errors in Label Studio. Faster with gold-standard quality.
- 4 Hybrid Fusion Engine:** Rules and AI run in parallel. A conflict-resolution engine prioritizes maximum PHI detection.



ClinicaAno successfully de-identifies bilingual clinical notes by fusing rule-based precision with AI recall to meet strict GDPR standards. Future work includes expanding the training corpus and deploying the system as a REST API for OMOP CDM integration.



Ricard Garcia-Isern, Antonio Arcas, Angela Leis, Juan Manuel Ramírez-Angueta, Carla Arxé Jan Carreras, Miguel-Angel Mayer