

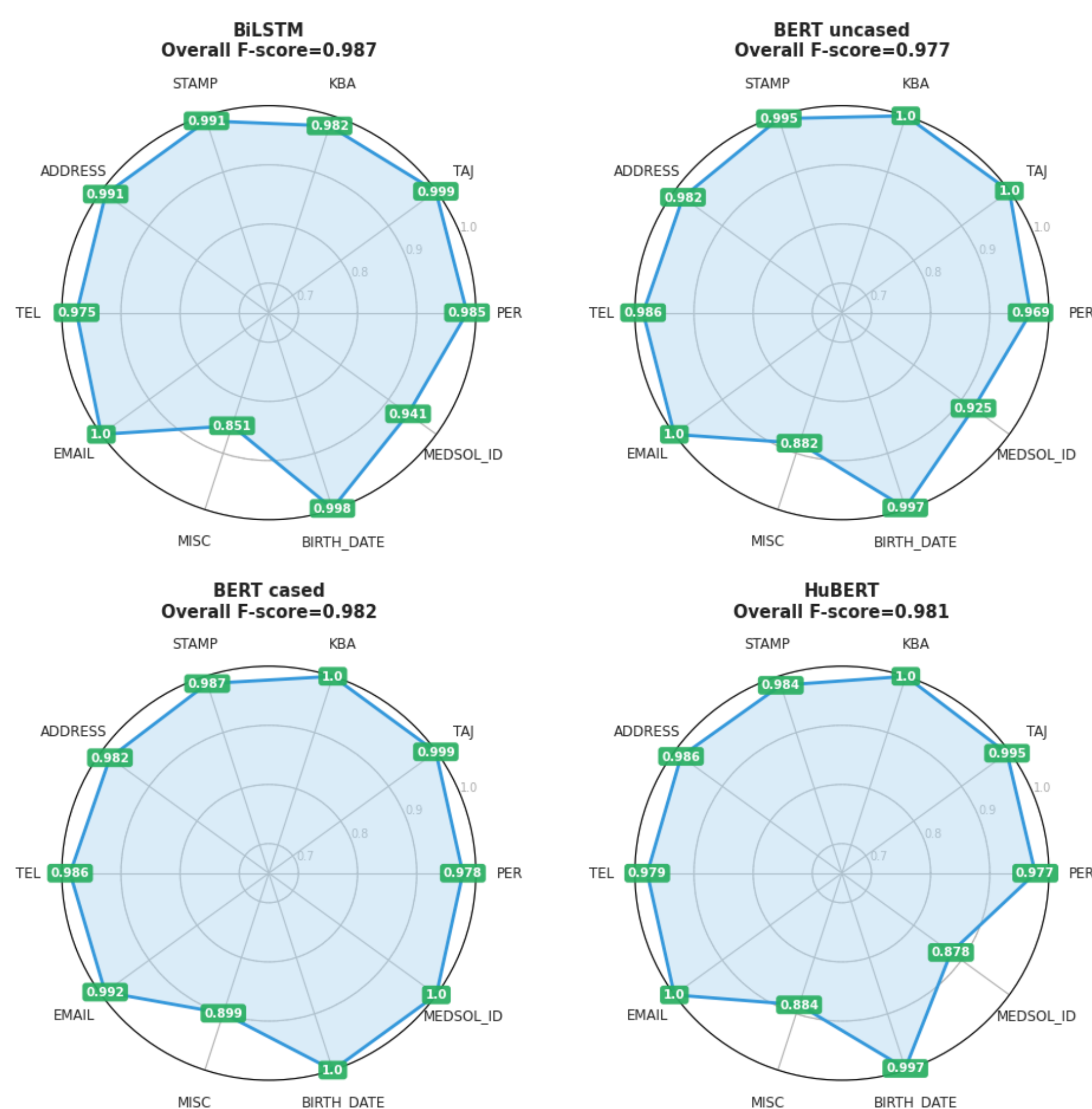
High-Performance NLP Models for De-identifying Hungarian Clinical Texts to Enable Observational Research

Production-Ready De-identification Models for Low-Resource Hungarian Clinical Data

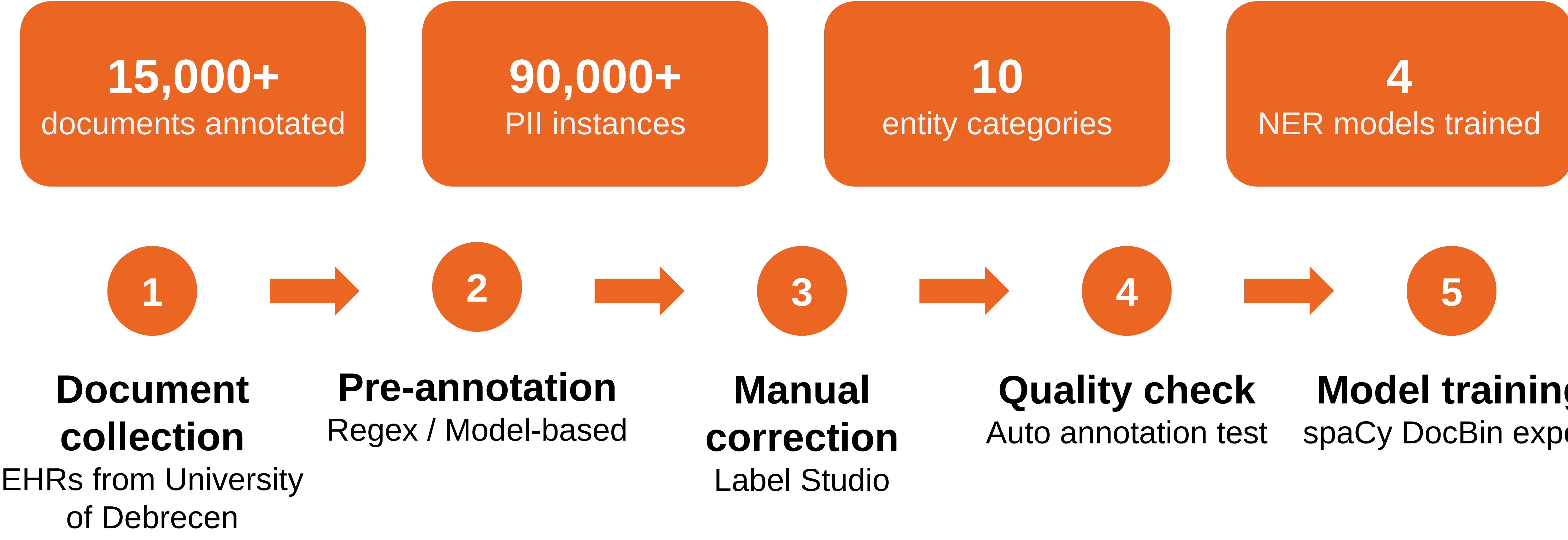
Background: The automated removal of Personally Identifiable Information (PII) is a critical preprocessing step, yet robust solutions are often lacking for non-English, low-resource languages. This study addresses this gap for Hungarian, developing and validating natural language processing (NLP) models for the de-identification of free-text clinical documents.

Result 1: Validation F-score results over 0.977.

Result 2: Model performance in terms of cost and efficiency.



Methods



Conclusion: We present highly effective, production-ready NLP models for de-identifying Hungarian clinical text. This tool enables the secure and GDPR-compliant secondary use of EHR data for observational health research, facilitating participation in collaborative networks like OHDSI. The model and methodology can serve as a blueprint for adapting de-identification pipelines to other low-resource language contexts.

