Fenice

Gruppo italiano per la ricerca clinica in Medicina d'Urgenza

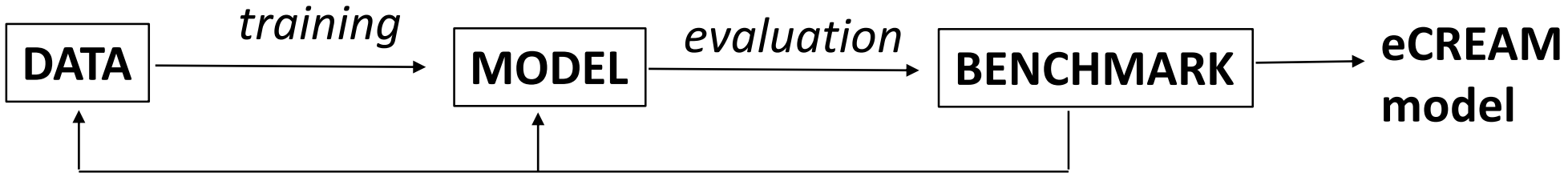
Le nuove frontiere dell'IA esplorate in eCREAM



Bernardo Magnini
Fondazione Bruno Kessler - Trento

6 november 2026

eCREAM: Experimental Setting (NLP)



ED clinical notes

Task: CRF-filling

Clinical notes + CRF

DATA: eCREAM_MedCorpus



- **Five Italian hospitals** concluded the process.
- About 2M notes per hospital, **10M notes in total.** More are joining (Maribor, Catania, …)
- **De-identification pipeline**: a commercial tool + LLM-based revision (< 0,1% false negatives). Over-removal approach. Final version ready in two weeks.
- **Agreement for data distribution**: signature expected in few days. Creative Common CC-BY license.
- **eCREAM-medCorpus v1.0** (advertised on scientific papers), official release v1.0 in December
- **Unique experience in Italy and Europe!**



eCREAM TASK: CRF-filling

[...]

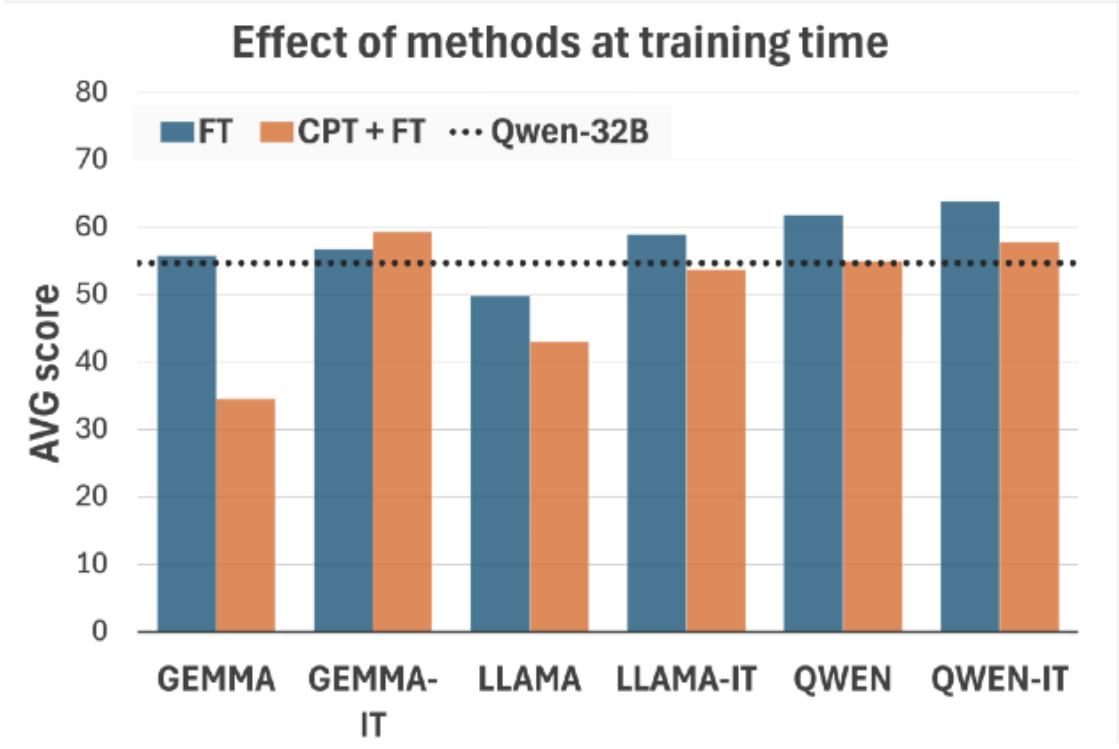
Paziente giunto in PS per dispnea. All'arrivo appare vigile ma presenta **difficoltà a mantenere un'adeguata interazione con l'ambiente**. **FC 110 bpm**. All'esame obiettivo si rilevano segni di affaticamento respiratorio. **Troponina 75 ng/L**.

[...]

Level of consciousness: V	Heart rate: 110	Presence of dyspnea : Y	Troponin: 75	Body temperature: Unknown
Category (A|V|P|U)	**Number**	**Binary (Y|N)**	**Number**	

eCREAM: CRF-filling

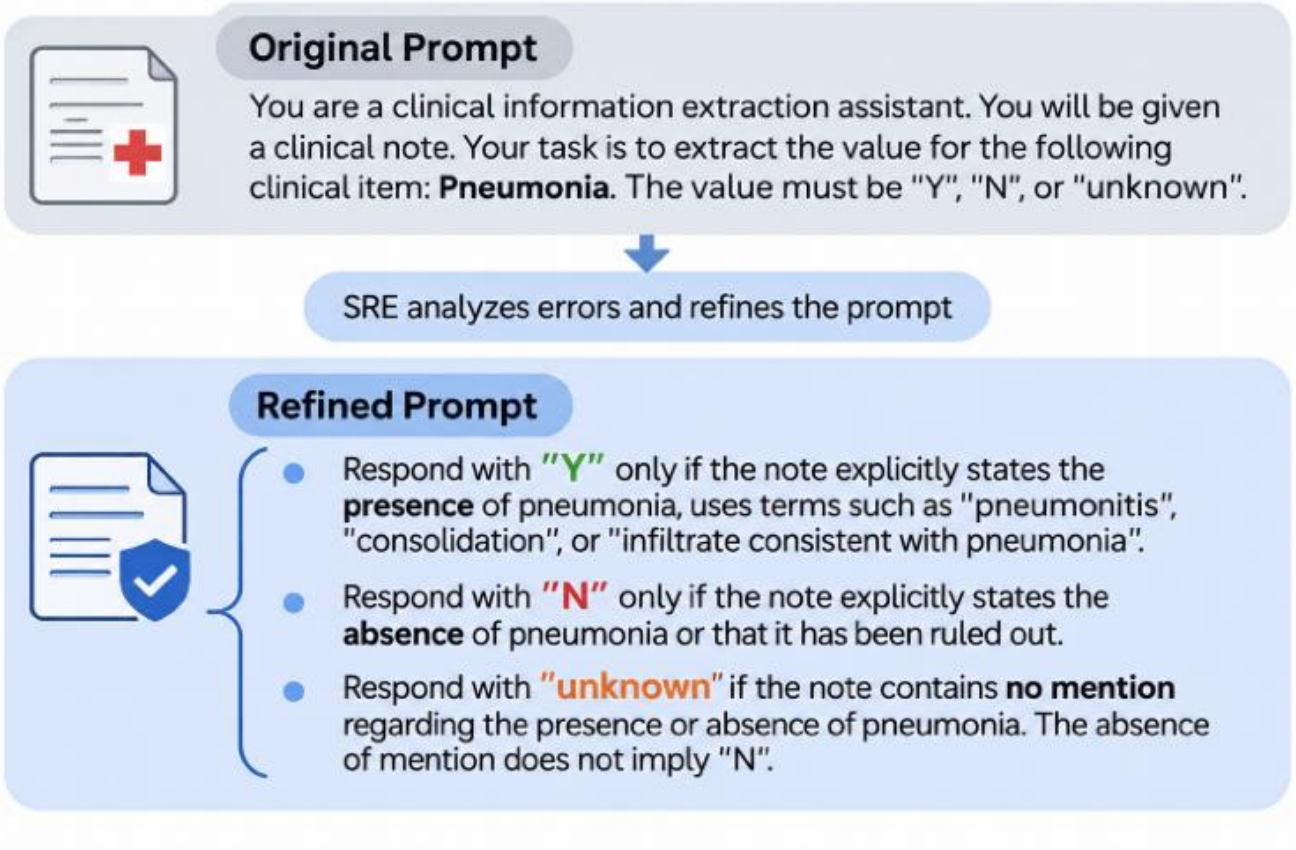
- Experiments on **small models** (1B parameters): continual pre-training, fine tuning
- **MedGemma-27B** currently best model on eCREAM benchmark
- CRF-filling **performance above 70% F1macro** (>95% F1 micro)
- **Prompt self-refinement from errors**: good results (submitted ACL ARR)
- Take advantage of non-labeled data: **Self-Relevance-Guided Masking** (EMNLP 2026)
- **Collaboration with Apertus and Meditron 70B –** (Swiss.ai) for eCREAM post training



CRF-filling: Prompt Self-refinement

Learning to improve prompts from errors

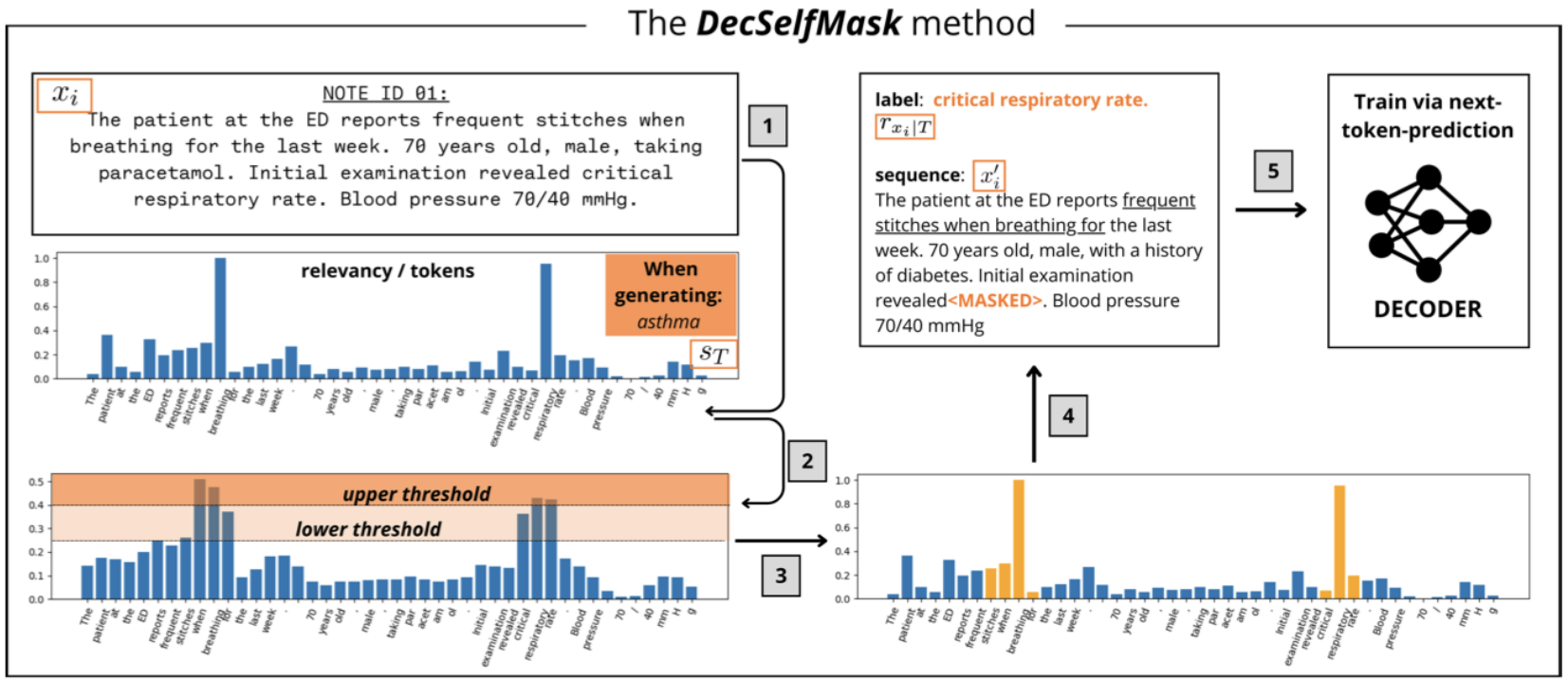
- Use a small set of annotated notes (e.g. 200)
- Run a model for CRF-filling with prompt 1
- Collect most significative errors for item
- Ask the model to refine the initial prompt based on detected error
- Iterate till performance improves with refined prompts
- Very effective, particularly with small models



CRF-filling: Detecting Relevant Associations

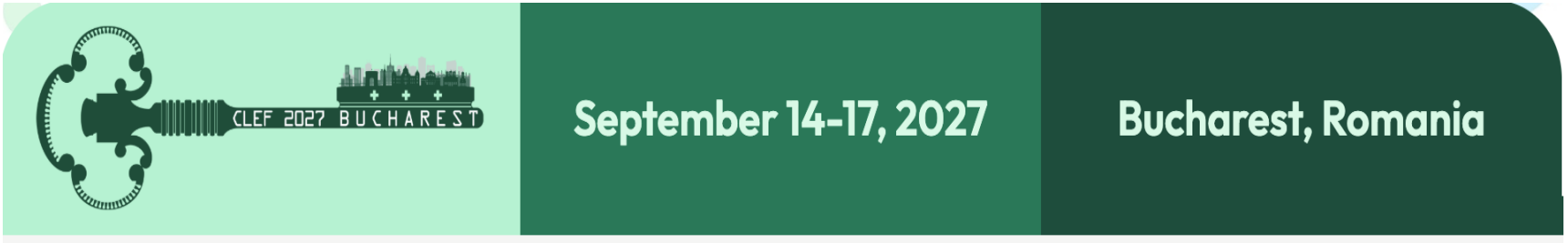
Clinical knowledge is crucial for CRF-filling

- Step 1: detect clinical associations in unlabeled notes through LLM attention mechanism while executing CRF-filling
- Step 2: use clinical association to fine-tuning a model
- Step 3: Run CRF-filling on the fine-tuned model



eCREAM Benchmarking

- **8K annotated notes+CRF,** ”Gamma” version for preliminary benchmarking is not satisfactory
- **Quality of manual annotations** so far is below expectations
- **Revision of current version in progress** (Mario Negri)
- Subset of 2K notes to be delivered by December 15 for **CLEF 2026 shared task** (Cross Language Evaluation Forum)



eHealth Lab @ CLEF 2027

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

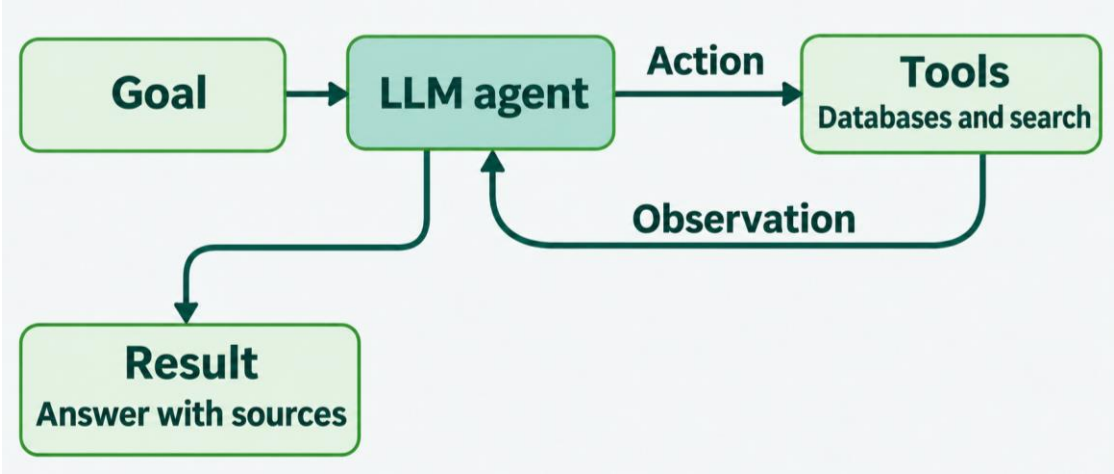
alt-edic

eCREAM

PLATEFORME DES DONNÉES DE SANTÉ
FRENCH HEALTH DATA HUB

Recent Trends in AI: Agents for Clinical Research

- **Plain LLM:** answers a prompt.
- **Agent:** an LLM with tools and a loop to act, observe and adapt.
- **Advantage:** executes tasks across steps with less manual coordination.

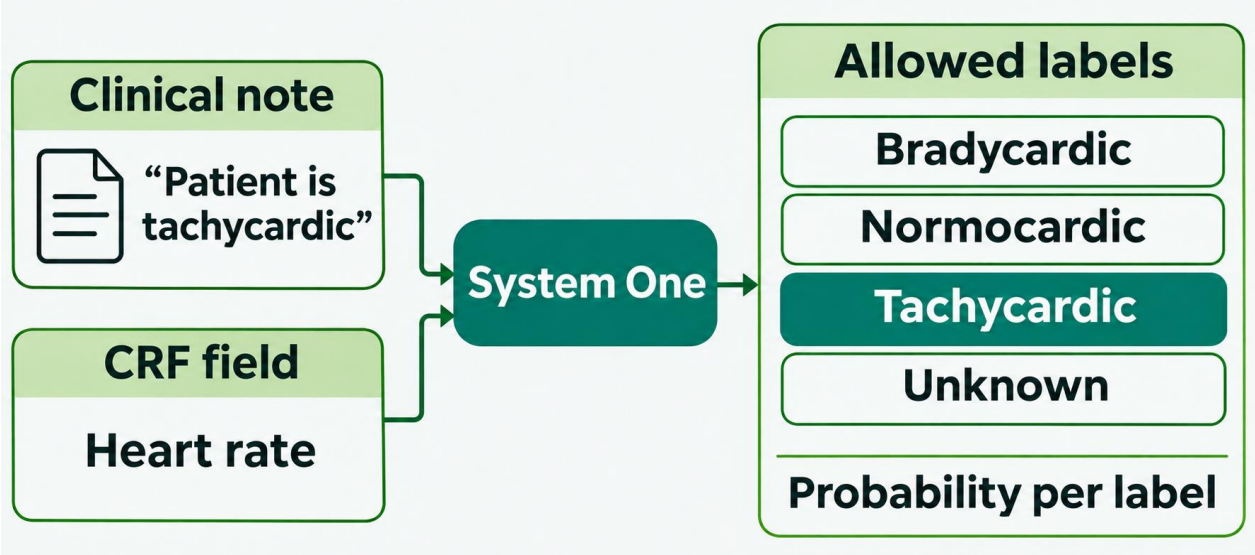


Example of usage:
Given a study protocol, identify eligible patients and provide a list with evidence from their records for review by the researcher.

Possibly even multi-agent:
Agents with different roles collaborate: one finds candidates, another checks the evidence.

Recent Trends in AI: System One for Decision Modeling

- **Plain LLM:** generates text.
- **Decision model:** directly returns typed choices and their calibrated probabilities. So we know how reliable are the answers.
- **Advantage:** simpler integration, with potentially lower latency and cost on repeated decisions.



Example of usage:
CRF filling task -> Classify the 136 Case Report Form fields, as in the DecSelfMask task. Compare Macro-F1, latency and calibration.

Possibly paired with LLMs:
System One models can handle fast, structured decisions, while LLMs tackle complex or uncertain cases that require deeper reasoning.

Grazie per l’attenzione!

