Polimi at CRF Filling 2026: Prompt-Based Information Extraction from Italian Clinical Notes

Vittorio Torri[1], Francesca Ieva[1,2]
[1]MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano,
[2]HDS - Health Data Science Centre, Human Technopole
[1]Piazza Leonardo da Vinci, 32, 20133, Milano, Italy,
[2]Viale Rita Levi Montalcini, 1, 20157, Milano, Italy
{vittorio.torri, francesca.ieva}@polimi.it

Abstract

In this paper we describe the system developed by the Polimi team for the CRF Filling Shared Task 2026, which focuses on extracting structured variables from clinical notes. The task is challenging due to scarce annotations, heterogeneous clinical language, and the sparsity of the 134 items to be extracted. Our approach relies on prompt-based information extraction using locally deployed open-weight Large Language Models (LLMs). We focused on the Italian subset of the dataset. The pipeline performs zero-shot extraction using task-specific prompts augmented with a glossary of abbreviations derived from unlabelled notes. To improve reliability and reduce hallucinations, the extraction schema is decomposed into multiple prompts targeting groups of variables, whose outputs are merged and refined through deterministic post-processing rules to normalize values and recover missing labels. During development we explored verification stages based on LLM-based prediction validation and synthetic example generation, but these strategies did not improve performance and were not included in the final system. On the development set, the best configuration based on Mistral Small 3.2 24B Instruct achieved an F1-score of 67.51%. On the official test set, our system ranked third overall and second among systems evaluated on the Italian subset, achieving an F1-score of 63%.

Keywords: Information Extraction, Clinical Natural Language Processing, Large Language Models, Prompt-based Learning, Electronic Health Records

1. Introduction

Electronic Health Records (EHRs) contain large amounts of clinically relevant information in the form of unstructured textual documents such as clinical notes and discharge summaries. Automatically extracting structured information from these narratives can support clinical research, hospital management, and decision support systems. However, clinical text processing remains challenging due to the heterogeneity of medical language, the frequent use of abbreviations and implicit expressions, and the presence of context-dependent information (Wang et al., 2018; Locke et al., 2021).

A major limitation in this domain is the scarcity of annotated datasets. Creating high-quality annotations requires medical expertise and strict privacy safeguards, making large annotated corpora difficult to obtain (Wei et al., 2018; Sylolypavan et al., 2023), especially in languages other than English (Névéol et al., 2018). As a result, many clinical NLP tasks must operate in settings with very limited supervision.

Recent advances in Large Language Models (LLMs) have shown promising results for zero- and few-shot clinical information extraction (Hu et al., 2024; Luo et al., 2024; Fornasiere et al., 2024). However, the use of proprietary models in healthcare contexts is often restricted because clinical data cannot be sent to external servers due to privacy and regulatory constraints. This motivates the exploration of locally deployable open-weight models that can operate entirely within secure hospital infrastructures.

In this work we describe the system developed by the Polimi team for the CRF Filling Shared Task 2026 (Ferrazzi et al., 2026), which focuses on extracting structured variables from Italian clinical notes. Our approach relies on prompt-based extraction with open-weight LLMs combined with additional knowledge derived from unlabelled data and deterministic post-processing to improve reliability on some specific items.

2. Materials and Methods

2.1. Data

The dataset provided for the CRF Filling Shared Task consists of short clinical notes, with original Italian texts and corresponding English translation. The aim of the task is to extract 134 items that compose the Case Report Form (CRF), which include symptoms, vital signs, medical history, diagnostic tests, and treatments. These items are divided into binary, categorical, and measured. The main challenge lies in their sparsity: $\sim 95\%$ of items have *unknown* value. Our team focused on the Italian version of the notes.

The data are divided into a training set (10

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 412–427
12 May 2026.

labelled notes), a development set (80 labelled notes), a test set (200 notes for which labels have not been released) (Kaczmarek et al., 2026), and a larger collection of 2,667 unlabelled notes (Ferrazzi et al., 2026). In our experiments, the training notes were used exclusively to construct examples for few-shot extraction and for LLM-based verification. The 80 development notes were used during system development to evaluate different prompting strategies and post-processing configurations. The 200 test notes were used only to generate the final predictions submitted to the shared task.

The unlabelled notes were exploited for corpus-level analysis, in particular to identify frequent medical abbreviations and acronyms that were used to construct a glossary integrated into the extraction prompts (Section 2.2.3).

An additional set of 80 synthetic notes (Ferrazzi et al., 2025) provided by the organizers was not used, as their annotated variables were not fully aligned with those in the main dataset.

2.2. Methods

Our final submitted system consists of a prompt-based zero-shot extraction pipeline with glossary augmentation and deterministic post-processing for certain items, summarized in Figure 1. In addition, during development we experimented with few-shot learning and additional verification stages designed to filter unsupported predictions and recover potentially missed labels. Although these verification strategies improved precision in some cases, they decreased the overall macro F1-score on the development set and were therefore not included in the final system.

2.2.1. Model Selection

We evaluated several open-weight large language models, focusing on models that can be deployed locally without sending clinical data to external servers, a common requirement in hospital environments. In particular, we experimented with LLaMA 3.1 8B, LLaMA 3.3 70B FP8 (Grattafiori et al., 2024), Qwen 2.5 72B AWQ (Yang et al., 2025), and Mistral-Small-3.2-24B-Instruct (Jiang et al., 2023), including also an ensemble of these models. For reference, we additionally included a comparison with two closed-source models from OpenAI, namely GPT-4.1 mini (Achiam et al., 2023) and GPT-5.4 (Singh et al., 2025).

2.2.2. Prompt-based Extraction

Information extraction was performed using prompt-based generation. The model receives the clinical note together with detailed instructions specifying the CRF schema and is asked to return a structured JSON object containing the extracted values. Basic prompts specified only the task and the admissible values, while more complex ones included additional instructions for handling missing values and numerical values.

To reduce hallucinations caused by large output schemas, we decomposed the extraction task into multiple prompts. Each prompt focuses on a subset of clinically related fields (e.g., binary, categorical and measured items). The resulting partial JSON outputs are then merged into a single structured representation before entering the verification stage.

Few-shot prompts included examples derived from the training set to illustrate the expected output format and the exact-match behavior required. Details of the prompts are reported in Appendix B.

2.2.3. Glossary Construction from Unlabelled Data

Clinical notes frequently contain abbreviated medical terminology that can hinder reliable information extraction. To address this issue, we constructed a glossary directly from the unlabelled corpus, with the goal of complementing the model's knowledge in cases where abbreviations are ambiguous or poorly recognized. Acronyms were automatically extracted from the notes using a regular expression identifying tokens composed of two to six uppercase letters. These candidates were aggregated across the corpus and ranked by frequency, resulting in a list of the 1,000 most frequent acronyms.

Each acronym was then processed using an LLM-based interpretation step. The model was prompted to determine whether the acronym corresponds to a well-defined abbreviation in Italian medical language and to provide its meaning together with a confidence score in structured JSON format. We focused on acronyms for which the model exhibited uncertainty or produced incorrect interpretations. Specifically, acronyms associated with low confidence scores (< 0.20) or incorrect meanings were incorporated into the extraction prompts as part of a glossary of domain-specific abbreviations. The list of included acronyms with their extracted meaning and confidence is reported in Appendix D.

This design explicitly targets cases where the model's internal knowledge is unreliable, providing additional context to improve interpretation during extraction. In the dataset, abbreviations are frequent and often correspond to clinical shorthand (e.g., ECG findings, laboratory values, or procedural notes), making their correct interpretation critical for accurate information extraction.

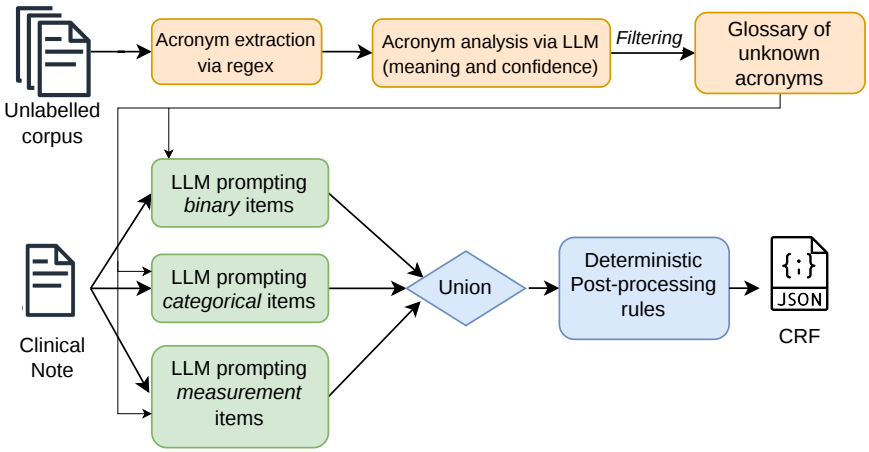

Figure 1: Overview of the extraction pipeline. The clinical note is processed by multiple prompts targeting different subsets of CRF fields. Each of them receives relevant elements from the glossary built from the unlabelled corpus. The resulting partial JSON outputs are merged into a single structured representation and subsequently refined through deterministic post-processing.

2.2.4. Verification Stage

During development we investigated whether an additional verification stage could improve extraction reliability.

First, we evaluated a simple LLM-based verifier that receives the clinical note together with the predicted CRF values and checks whether each prediction is supported by explicit textual evidence. Predictions that cannot be supported by a literal span in the note are discarded and replaced with `unknown`.

In a second variant, we introduced a synthetic-example-guided verifier designed to recover labels initially predicted as `unknown`. To support this step, we generated a bank of synthetic clinical examples using an LLM. Each example consists of a short Italian Emergency Department note in which the value of a specific CRF item is controlled and supported by an explicit evidence span. Generation prompts include stylistic constraints to reproduce realistic ED documentation and information to ensure that the correct label can be determined only from explicit textual evidence. An additional validation pass verifies that the generated note supports the intended label and that the evidence span appears verbatim in the text.

During verification, a small number of contrastive synthetic examples are included in the prompt to guide the model in identifying explicit evidence in the original note. A recovered label is accepted only if a supporting span is explicitly identified.

2.2.5. Post-processing

The predictions generated by the LLM were further refined using a deterministic post-processing stage applied offline to the model outputs. These rules exploit explicit textual evidence in the clinical notes to correct systematic errors produced by the LLM. The post-processing stage was designed to correct frequent normalization issues and reduce unsupported predictions while preserving the original extraction outputs whenever no clear textual evidence was available. More details, including an example, are in Appendix C.

2.2.6. Ensemble Strategy

We explored an ensemble combining LLaMA, Qwen, and Mistral using a deterministic item-level aggregation. Agreement between at least two models is taken as the final prediction. When only one model predicts a non-`unknown` value, it is accepted only if produced by Mistral, which achieved the highest recall on the development set. In case of conflicting non-`unknown` predictions, Mistral is used as the anchor model. This design leverages agreement as a proxy for confidence while prioritizing recall-oriented predictions.

2.2.7. Implementation Details

All experiments were conducted using locally deployed open-weight models, except those including OpenAI models that were performed by using the official OpenAI APIs. Experiment with Mistral-22B were executed on a single NVIDIA H100 GPU, while smaller models on a single A100 or V100 GPU. LLM inference was performed with the `vLLM` framework (Kwon et al., 2023), which enables batched inference and structured output decoding. Model outputs were constrained to JSON schemas using structured decoding in order to guarantee valid

CRF outputs. Extraction and verification stages were executed with deterministic decoding (temperature $= 0$) to ensure reproducibility. Code is available at https://github.com/vittot/CRF-Filling-SharedTask-CL4Health26.

3. Results

Table 1 summarizes the main experiments and the their results on the development set. Baseline zero-shot prompting with Llama 3.1 8B achieved an F1-score of 52.84 %, which increased to 57.78 % when scaling to Llama 3.3 70B FP8. The introduction of rules in the prompt increased significantly the FP, and the glossary significantly reduced them, while keeping a small improvement on TP and FN compared to the first simple prompt. Substantial reduction of FP was obtained when splitting the prompts, with 3 prompts (binary, categorical, measured) being the most effective, even if this also increased significantly the FN. The post-processing steps determined a major improvement, keeping FP at the same level, but reducing FN.

The best development result was obtained with Mistral Small 3.2 24B using three prompts, glossary and enhanced post-processing, reaching an F1-score of 67.51 %. This configuration outperformed both larger but quantized models and ensemble approaches, achieving the best balance between FP and FN. Additional strategies explored during development, including few-shot prompting, ensemble predictions, and verification stages, did not improve the best configuration. In particular, the LLM-based verifier substantially reduced performance, while the synthetic-example-based verifier achieved good F1-score but remained below the best configuration.

Proprietary OpenAI models showed a clear tendency toward over-generation, with a substantially higher number of FP compared to open-weight models (e.g., GPT-5.4: 305 FP vs 94 for Mistral). Despite their good recall, this lower precision resulted in lower F1-score, indicating weaker adherence to the strict evidence-based extraction constraints of the task.

Figure 2 reports the official test set results. The best performance was again obtained by the Mistral-based configuration with three prompts, glossary and enhanced post-processing, achieving an F1-score of 63 %. Systems based on Llama 3.3 70B and Qwen 72B obtained similar but slightly lower results (61 %). The configuration including the verification stage performed worse (53 %), confirming the trends observed on the development set.

Item-level results (Table 2) show substantial variability across variables. Vital signs and laboratory measurements achieved the highest performance,

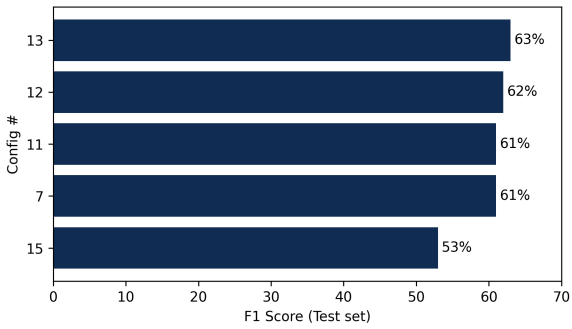

Figure 2: Results on the official test set

often exceeding 85% F1 (e.g., *SpO2*, *heart rate*, *body temperature*, and *blood pressure*), reflecting their relatively standardized textual expressions. Intermediate performance was observed for clinical context variables such as *poly-pharmacological therapy*, *anticoagulants or antiplatelet therapy*, and *cardiovascular diseases*. In contrast, symptoms and infrequent conditions showed much lower performance; for example, *presence of dyspnea* achieved an F1-score of 27 %, while several low-support variables obtained F1-scores close to zero.

The best development configuration produced 255 errors, including 155 false negatives and 94 false positives, indicating that the system often failed to extract information explicitly present in the text and predicted `unknown`. False positives were frequently caused by contextual inference (e.g., *cardiovascular diseases*, *poly-pharmacological therapy*), while false negatives were more common for symptoms or episodic events such as *agitation*, *presence of dyspnea*, and *administration of fluids*. Value mismatches mainly involved normalization issues for numerical variables or minor formatting differences (e.g., “hb 12” vs. “12”), suggesting that improved normalization rules could further reduce these errors. Some examples are reported in Appendix A.

4. Discussion

In this work we explored several strategies for prompt-based information extraction from clinical notes, including zero-shot and few-shot prompting, glossary-based knowledge injection, LLM-based verification, and deterministic post-processing. While scaling to larger models improved baseline performance, prompt engineering alone produced modest gains. The non-quantized Mistral Small 3.2 24B model outperformed both the smaller Llama 3.1 8B and the larger but quantized Llama 3.3 70B, suggesting that architecture and quantization may have a stronger impact than parameter count alone. Additional strategies yielded limited improvements (glossary) or no gains (verification). In contrast,

Config #	Model	Description	TP	FP	FN	F1 [%]
1	Llama 3.1 8B	simple prompt	235	337	107	52.84
2	Llama 3.3 70B FP8	simple prompt	306	380	60	57.78
3	Llama 3.3 70B FP8	prompt with rules	309	760	**50**	56.21
4	Llama 3.3 70B FP8	rules + glossary	310	498	54	58.23
5	Llama 3.3 70B FP8	3 prompts rules + glossary	148	88	234	57.76
6	Llama 3.3 70B FP8	3 prompts rules + glossary + post-proc	231	90	136	64.13
7	Llama 3.3 70B FP8	3 prompts rules + glossary + enh. post-proc	245	100	121	65.30
8	Llama 3.3 70B FP8	3 prompts few-shot rules + glossary + post-proc	210	**87**	157	62.31
9	Qwen 2.5 72B AWQ	3 prompts rules + glossary	146	92	227	58.48
10	Qwen 2.5 72B AWQ	3 prompts rules + glossary + post-proc	229	108	129	63.16
11	Qwen 2.5 72B AWQ	3 prompts rules + glossary + enh. post-proc	248	103	114	65.66
12	Mistral Small 3.2 24B	3 prompts rules + glossary + post-proc	235	96	130	64.36
13	Mistral Small 3.2 24B	3 prompts rules + glossary + enh. post-proc	255	94	155	**67.51**
14	Ensemble	Llama70B-FP8 + Qwen72B-AWQ + MistralSmall24B	238	121	138	63.20
15	Mistral Small 3.2 24B	3 prompts + rules + glossary + post-proc + LLM-based verifier (Llama 3.1 8B)	302	498	64	58.23
16	Mistral Small 3.2 24B	3 prompts + rules + glossary + post-proc + Synth-based verifier (Mistral Small 3.2 24B)	266	133	101	65.97
17	OpenAI GPT 4.1 mini	3 prompts + rules + glossary	288	885	95	55.06
18	OpenAI GPT 5.4	3 prompts + rules + glossary	307	317	82	62.43
19	OpenAI GPT 4.1 mini	3 prompts + rules + glossary + enh. post-proc	301	867	81	57.55
20	OpenAI GPT 5.4	3 prompts + rules + glossary + enh. post-proc	**314**	305	75	64.22

Table 1: Development set results for the different configurations of our system.

Item	F1	Support
SpO2	1.00	20
Heart rate	0.91	25
Body temperature	0.90	28
Blood pressure	0.86	24
Level of consciousness	0.88	33
Poly-pharmacological therapy	0.79	23
Anticoagulants / antiplatelet therapy	0.81	16
Antihypertensive therapy	0.81	19
Cardiovascular diseases	0.74	27
Chronic pulmonary disease	0.73	7
Neuropsychiatric disorders	0.46	10
ECG abnormalities	0.44	7
Level of autonomy (mobility)	0.33	11
Presence of dyspnea	0.27	20

Table 2: Item-level performance of the best development configuration on selected items.

decomposing the task into three prompts targeting groups of related variables proved beneficial, likely reducing hallucinations caused by the large output schema. This observation is consistent with recent studies reporting limited benefits of more complex techniques such as RAG or CoT in clinical NLP (Nagar et al., 2025). Deterministic post-processing further improved performance by correcting normalization errors and recovering information expressed through predictable textual patterns. Using this configuration, our system ranked third overall in the shared task and second among systems evaluated on the Italian subset.

Performance varied substantially across variables. Vital signs and laboratory measurements were extracted reliably due to their stable lexical patterns, whereas symptoms, contextual variables, and rarely mentioned conditions remained challenging because of implicit phrasing, abbreviations, and limited supervision.

This work has several limitations. The dataset is relatively small, limiting the ability to learn robust patterns for rare variables and increasing the risk of overfitting the development set. Experiments were conducted only on Italian notes, and the generalizability to the English version remains to be investigated. Moreover, the best-performing models require computational resources that may not always be available in hospital environments.

Future work could explore hybrid approaches combining prompt-based extraction with lightweight supervised models, improved normalization and abbreviation handling, the use of English prompts on Italian data, and larger multilingual clinical corpora. A more systematic evaluation of synthetic data generation for supervision may also help improve performance on sparse variables.

5. Acknowledgements

The present research has been partially supported by the Italian Ministry of University and Research (MUR), grant Dipartimento di Eccellenza 2023-2027, awarded to the Department of Mathematics, Politecnico di Milano.

6. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Al-

tenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026. Overview of the crf 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.

Raffaello Fornasiere, Nicolò Brunello, Vincenzo Scotti, and Mark Carman. 2024. Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 456–466.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. 2024. Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics*, 183:105321.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10:3.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, and Gareth B Kitchen. 2021. Natural language processing in medicine: a review. *Trends in Anaesthesia and Critical Care*, 38:4–9.

Xiao Luo, Fattah Muhammad Tahabi, Tressica Marc, Laura Ann Haunert, and Susan Storey. 2024. Zero-shot learning to extract assessment criteria and medical services from the preventive healthcare guidelines using large language models. *Journal of the American Medical Informatics Association*, 31(8):1743–1753.

Aishik Nagar, Viktor Schlegel, Thanh-Tung Nguyen, Hao Li, Yuping Wu, Kuluhan Binici, and Stefan Winkler. 2025. Llms are not zero-shot reasoners for biomedical information extraction. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 106–120.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Aneeta Sylolypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. 2023. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, 6(1):26.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2018. Clinical text annotation–what factors are associated with the cost of time? In *AMIA Annual Symposium Proceedings*, volume 2018, page 1552.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

7. Language Resource References

Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026. Small llms for medical nlp: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in italian.

Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. Converting annotated clinical cases into structured case report forms. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Viena, Austria. Association for Computational Linguistics.

Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo Magnini. 2026. Toward automatic filling of case report forms: A case study on data from an italian emergency department.

A. Evaluation Details and Examples

The task is evaluated at the item level across clinical documents, using macro-averaged F1-score as the official metric. For each document, systems must predict the values of a predefined set of CRF items, which are compared with the reference annotations. The value *unknown*, allowed for all items, indicates that the information is not explicitly present in the document. An item contributes to the evaluation whenever at least one of the two values (prediction or reference) is different from unknown. Correct matches are counted as true positives (TP), while predicting a value where the reference is *unknown* produces a false positive (FP) and predicting *unknown* where a value exists produces a false negative (FN).

Macro-F1 is computed across item values for each document and then averaged across all documents to obtain the final score. This design gives equal importance to each document and prevents documents with many populated items from dominating the metric. At the same time, rare items and frequent items contribute equally, making robust performance across all CRF fields important for achieving a high overall score.

The official evaluation script computes the macro-F1, TP, FP and TN, and it was used to report results on development set in Table 1, while the evaluation on the test is on the submission platform which reports only the macro-F1 (Figure 2).

In addition to the official scoring script, we performed a per-item error analysis on the development set, to better understand model behavior (summarized in Table 2). We also manually investigated mistakes on specific document/item pairs, and we report here some examples, including also potential annotation mistakes.

Example 1 (Chest Pain - FN - Terms not recognized)
Document ID: 419929

Item: chest pain
Gold value: y
Prediction: unknown

Clinical note excerpt:
*"Riferisce **algie** ben localizzate all'**emitorace** di sinistra da una settimana."*

The model is not recognized the presence of chest pain, perhaps due to the word *emitorace (hemithorax)* which indicates only half of the chest.

Example 2 (Respiratory Rate - FP - Suspect Annotation Mistake)
Document ID: 1498690

Item: respiratory rate
Gold value: unknown
Prediction: eupneic

Clinical note excerpt:
*"B: **eupnoico** in cn 2 l/min, lieve utilizzo muscolatura accessoria, e/to ega art.–>SO2 93%"*

The text clearly states that the patient is *eupnenic*, so the gold label is probably wrong.

Example 3 (SPO2 - FP - Normalization error)
Document ID: 1777967

Item: SPO2
Gold value: 97%
Prediction: 97%aa

Clinical note excerpt:
*"PAOS 120/75, FC 115, **Sat 97%aa**, T 38.1°C"*

This error is likely due to the instruction given to the LLM not to normalize measurement values. In the annotations, measurement fields often preserve spaces and abbreviations as they are in the original text (e.g., *hb 12*, *97%*), which are kept as part of the gold value, apparently not in a fully consistent manner.

Example 4 (History of Allergy - FP - Negation not recognized)
Document ID: 1307609

Item: History of Allergy
Gold value: n
Prediction: y

Clinical note excerpt:
*"**Non** fumo **né allergie note**."*

The model recognized the mention of allergy (it is not *unknown*), but it missed the negation, expressed in a non-trivial form which also bypasses our dedicated post-processing rule.

Example 5 (ECG Abnormality - FN - Acronym not recognized)
Document ID: 821011

Item: ECG Abnormality
Gold value: y
Prediction: unknown

Clinical note excerpt:
*"- ECG –> RS FC 58 asse a sx, 1 **CPV**"*

The model does not seem to recognize the mention of *CPV*, which stands for *Premature Ventricular Contraction* and indicates an abnormality in the ECG.

Example 6 (Neurodegenerative Diseases - FN - Borderline condition)
Document ID: 382962

Item: Neurodegenerative Diseases
Gold value: y
Prediction: unknown

Clinical note excerpt:

*"- **Sindrome di Arnold Chiari**. Seguita c/o H Molinette. In programma RM encefalo a breve (non ancora effettuata) e successiva v NCH con dr. NOME_PERSONA per presa in carico presso OSGB."*

The model does not recognize the mention of *Sindrome di Arnold Chiari*, a neurological disorder reported in the patient's history. Although not strictly a neurodegenerative disease (e.g., Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis), this mention may still be relevant depending on how neurological conditions are interpreted within the annotation schema.

B. Prompt Templates

This appendix reports the prompt templates used in our extraction pipeline. Prompts were designed to extract structured variables from Italian clinical notes using a zero-shot instruction-following setup. We also include prompts for verifiers and glossary analysis.

B.1. Binary Variables Prompt

```
Sei un estrattore di dati clinici per
    verbali di Pronto Soccorso italiani.

OBIETTIVO: Compila SOLO i campi binari
    (y/n/unknown) richiesti.

REGOLE CRITICHE (ANTI-FALSE-POSITIVE):
1) DEFAULT = "unknown". Se il concetto
    NON e' esplicitamente menzionato,
    scrivi "unknown".
2) Scrivi "n" SOLO con negazione
    esplicita nel testo ("nega", "non",
    "assente", "negativo", "non segni
    di", ecc.).
3) Scrivi "y" SOLO con affermazione
    esplicita nel testo.
4) NON fare inferenze cliniche, NON
    dedurre diagnosi o terapie non
    dichiarate.
5) Filtra temporalmente: considera SOLO
    dati all'ADMISSION/triage.

ECCEZIONE TEMPORALE:
Per item di ANAMNESI/TERAPIA
    DOMICILIARE/SOCIALE e' consentito
    usare sezioni
"APR:", "Anamnesi:", "TD:", "Terapia
    domiciliare:", "Allergie:", o
    elenchi
di farmaci/patologie come evidenza
    esplicita.

6) **Administration of [Drugs]**:
    Scrivi 'y' solo se il farmaco e'
    stato
somministrato FISICAMENTE in PS (verbi:
    "praticato", "somministrato",
    "fatto").
I farmaci presenti in TD NON contano
    come 'administration of ...'.

ECCEZIONE SOLO PER ITEM DI
    ANAMNESI/TD/SOCIALE (riduzione FN
    controllata):
Per i seguenti item:
- history of allergy
- poly-pharmacological therapy
- antihypertensive therapy
- cardiovascular diseases
- anticoagulants or antiplatelet drug
    therapy
- diffuse vascular disease
- neuropsychiatric disorders
- living alone

vale questa regola speciale:
A) Se trovi una sezione tipo "APR:",
    "Anamnesi:", "TD:", "Terapia
    domiciliare:",
   "Allergie:" oppure un elenco di
       farmaci/patologie separati da
       virgole o trattini,
   questo conta come MENZIONE ESPLICITA
       quindi puoi scrivere "y" se
       pertinente.
B) Scrivi "n" SOLO con negazione
    esplicita (es. "nega allergie",
   "allergie: neg", "non assume
       terapia", ecc.).
C) Se non trovi queste sezioni o
    elenchi espliciti allora resta
    "unknown".
D) NON applicare questa eccezione agli
    altri item binari (acuti).

GLOSSARIO (serve solo per riconoscere
    menzioni esplicite, non per
    dedurre):
- **EON**: Esame Obiettivo Neurologico.
    Se scritto "EON nella norma" o
  "GCS 15" o "vigile/orientato" allora
      level of consciousness = 'a'.
  Se non scritto allora "unknown".
- **CE**: Corpo estraneo (utile per
    'foreign body in the airways').
- **FA**: Fibrillazione atriale (utile
    per 'arrhythmia').
- **NAO / TAO**: terapia anticoagulante
    (utile per
  'anticoagulants or antiplatelet drug
      therapy').
- **ASA / aspirina / clopidogrel**:
    antiaggreganti (utile per
  'anticoagulants or antiplatelet drug
      therapy').
- **TD**: Terapia domiciliare.

REGOLE PER OGNI ITEM:
```

```
{rules_text}

Restituisci ESCLUSIVAMENTE un oggetto
    JSON conforme allo schema.
```

B.2. Categorical Variables Prompt

```
    Sei un estrattore di dati clinici
        per verbali di Pronto Soccorso
        italiani.

OBIETTIVO: Compila SOLO i campi
    categorici richiesti (non
    misurazioni).

REGOLE:
1) DEFAULT = "unknown" se
    l'informazione non e'
    esplicitamente nel testo.
2) Non dedurre "n" o altre categorie se
    non scritto.
3) Per ogni campo, usa ESATTAMENTE una
    delle opzioni permesse.
4) Considera SOLO l'ADMISSION/triage
    (ignora rivalutazioni).
5) **Level of Consciousness**: 'a'
    (vigile), 'v' (verbal), 'p' (pain),
    'u' (unresponsive). 'a' SOLO se il
    testo contiene esplicitamente:
    "vigile", "orientato", "GCS 15",
    "EON nella norma". Altrimenti
    "unknown".

GLOSSARIO (Il glossario serve solo a
    riconoscere menzioni esplicite. Non
    usarlo per dedurre diagnosi non
    scritte):
- **APR**: Anamnesi Patologica Remota
    (storia clinica passata).
- **EON**: Esame Obiettivo Neurologico
    (stato di coscienza). Se "EON nella
    norma" -> level of consciousness =
    'a'.
- **AASS** / **AAII**: Arti Superiori /
    Arti Inferiori.
- **TC**: Se associata a gradi (es. TC
    37.1) e' **Temperatura Corporea**.
    Se in contesto radiologico e'
    Tomografia Computerizzata.

REGOLE PER OGNI ITEM:
{rules_text}

Restituisci ESCLUSIVAMENTE un oggetto
    JSON conforme allo schema.
```

B.3. Measured Variables Prompt

```
    Sei un estrattore di dati clinici
        per verbali di Pronto Soccorso
        italiani.

OBIETTIVO: Compila SOLO i campi
    MISURATI richiesti (valori
    numerici/testuali).

REGOLE MIRROR (EXACT MATCH):
1) Se un valore e' presente nel testo,
    COPIALO LETTERALMENTE (stessi
    spazi, simboli, unita' se presenti,
    ad esempio "hb 12", "k 6.3").
2) Non normalizzare (es: non
    trasformare "99 %" in "99%").
3) Se un valore non e' presente, scrivi
    "unknown".
4) Considera SOLO valori
    all'ADMISSION/triage.
5) Per 'spo2' scrivi SOLO la
    percentuale cosi' come appare (es.
    "99 %" o "99%"), senza prefissi
    ("SpO2", "Sat", ecc.).

GLOSSARIO (Il glossario serve solo a
    riconoscere menzioni esplicite. Non
    usarlo per dedurre diagnosi non
    scritte):
- **APR**: Anamnesi Patologica Remota
    (storia clinica passata).
- **HGT**: Glicemia capillare (valore
    del glucosio).
- **EGA**: Emogasanalisi (fonte per pH,
    pO2, pCO2, Lattati).
- **AA**: Aria Ambiente (SpO2 senza
    ossigeno).
- **ASA**: Acido Acetilsalicilico.
- **NRS** / **VAS**: Scale del dolore
    (0-10)..
- **TC**: Se associata a gradi (es. TC
    37.1) e' **Temperatura Corporea**.
    Se in contesto radiologico e'
    Tomografia Computerizzata.
- **GB / WBC**: Sono i **Leucociti**
    (Globuli Bianchi). Se leggi "gbb
    16", scrivi "gbb 16" nel campo
    leukocytes.

REGOLE PER OGNI ITEM:
{rules_text}

Restituisci ESCLUSIVAMENTE un oggetto
    JSON conforme allo schema.
```

B.4. Acronym Analysis Prompt

```
    Sei un medico esperto di Pronto
        Soccorso. Analizza
        l'acronimo medico e rispondi
        solo in formato JSON
        seguendo lo schema richiesto.

    Analizza l'acronimo: {acronym}.
        E' un termine certo nel
        contesto dell'emergenza
        urgenza italiana?
```

```
Rispondi con questo schema JSON:

{{"acronym": "{acronym}",
    "known": "yes/no",
    "meaning": "...",
    "certainty_score": 0-100}}
```

B.5. LLM-based Verifier Prompt

```
    Sei un verificatore rigoroso.

Riceverai:
1) Una nota clinica di Pronto Soccorso
    (in italiano).
2) Un elenco di variabili del CRF con i
    valori proposti (predizioni).

Compito:
Per OGNI variabile, devi decidere se il
    valore proposto e' SUPPORTATO da
    uno span ESPLICITO nel testo della
    nota.

Regole:
- "SUPPORTATO" solo se la nota lo
    afferma esplicitamente (oppure lo
    nega esplicitamente se il valore e'
    una negazione come 'n').
- Se la variabile non e' menzionata
    esplicitamente, segnala "DROP".
- NON utilizzare inferenze cliniche.
- NON assumere valori negativi in
    assenza di evidenza. Silenzio
    implica DROP.
- L'output deve essere in formato JSON
    con chiavi = nomi delle variabili e
    valori in {"SUPPORTATO","DROP"}.
```

B.6. Synth-based Verifier Prompt

```
    Sei un verificatore clinico per
        estrazione strutturata da
        verbali di Pronto Soccorso
        italiani.

OBIETTIVO:
Per UN SOLO item target, verifica se il
    valore iniziale "unknown" deve
    essere mantenuto oppure corretto in
    base a evidenza ESPLICITA presente
    nel testo.

REGOLE CRITICHE:
1) DEFAULT = "unknown".
2) Puoi cambiare "unknown" in "y" o "n"
    SOLO se trovi evidenza ESPLICITA
    nel testo.
3) NON fare inferenze cliniche.
4) NON usare conoscenza medica
    implicita per dedurre il valore.
5) Usa SOLO il testo del verbale.
6) L'evidence_span deve essere una
    sottostringa LETTERALE del verbale,
    non parafrasata.
7) Se non trovi una sottostringa
    letterale di supporto, mantieni
    "unknown".
8) Se ci sono elementi contraddittori o
    poco chiari, preferisci "unknown".

ITEM TARGET:
{item}

DESCRIZIONE ITEM:
{item_description}

ISTRUZIONI SPECIFICHE:
{item_rules}

ESEMPI CONTRASTIVI DI RIFERIMENTO:
{examples_block}

VERBALE CLINICO:
{note}

VALORE INIZIALE DEL PRIMO PASSAGGIO:
unknown

RESTITUISCI ESCLUSIVAMENTE un JSON
    conforme allo schema.
```

B.7. Synthetic Examples Generation Prompt

```
    Sei un generatore di brevi verbali
        clinici sintetici italiani di
        Pronto Soccorso.

OBIETTIVO:
Genera un singolo verbale clinico
    realistico, breve ma plausibile, in
    italiano, tale che il valore
    corretto per il campo target sia
    ESATTAMENTE quello richiesto.

VINCOLI:
1) Il verbale deve sembrare realistico
    e non artificiale.
2) Non nominare mai il nome tecnico del
    campo target in inglese.
3) Non inserire spiegazioni, solo il
    contenuto richiesto.
4) Il testo deve contenere abbastanza
    contesto clinico da sembrare un
    vero verbale.
5) Il valore corretto del campo target
    deve essere determinabile SOLO da
    un'evidenza esplicita nel testo.
6) Se label = "unknown", NON inserire
    alcuna evidenza esplicita ne'
    positiva ne' negativa per il campo
    target.
7) Mantieni lo stile compatibile con
```

```
    verbali PS italiani: triage,
    anamnesi, APR, TD, EO, terapia,
    decorso, ecc.
8) Evita di copiare identicamente gli
    esempi eventualmente forniti.
9) Inserisci uno o piu' distrattori
    realistici ma NON contraddire il
    label target.
10) L'evidenza deve essere locale e
    verificabile come span testuale.

TARGET ITEM:
{item}

DESCRIZIONE:
{description}

LABEL RICHIESTA:
{label}

INDICAZIONI SU EVIDENZA POSITIVA:
{positive_evidence_hint}

INDICAZIONI SU EVIDENZA NEGATIVA:
{negative_evidence_hint}

INDICAZIONI SU UNKNOWN:
{unknown_evidence_hint}

SEZIONI PROBABILI:
{likely_sections}

INDIZI LESSICALI POSSIBILI:
{lexical_hints}

{anchors_block}

ESEMPI:

{few_shots_block}

RESTITUISCI ESCLUSIVAMENTE un JSON
    conforme allo schema.
```

C. Post-Processing Rules

After LLM-based extraction, predictions were refined using a deterministic post-processing pipeline applied to the generated JSON outputs.

For each document, the predicted items are first mapped to an item–value dictionary. The clinical note is then analyzed using regular expressions and rule-based heuristics. Rules are applied sequentially and may modify predicted values according to the conditions described below.

Unless otherwise stated, rules modify a prediction only if the current value is `unknown`.

The simple "post-processing" named in the paper cover rules for *history based inference*, *allergy negation* and additional rules for adjusting measurement units of measured variables, while the others reported here are part of the "enhanced post-processing".

Table C1 reports an example of the post-processing effect on the following document form the dev set (id 914210):

```
Pte proveniente con ambulanza, preso in
    carico in sala visita covid per
    dispnea
A: vie aeree pervie
B: giunge in MR, E/EGA ART in AA, spo2
    86% pO2 49 pCo2 29, posizionate cn
    4 lt, eupnoico al momento
C: cute asciutta, polsi isosfigmici
    bilateralmente FC 89 b/min R PAO
    130/80
D: GCS 15 pupille isocicliche
    bilateralmente, vigile orientato
E: tc 38 E/perfalgan in ambulanza

cvp posizionato dal LUOGO in sede
E/ecg
LUOGO in degenza covid per nota
    positivita' al covid
4AT negativa
```

In this note, post-processing recovers two missed labels (*foreign body in the airways=n*, *presence of dyspnea=y*) and suppresses two unsupported predictions (*presence of respiratory distress*, *sars-cov-2 swab test*), while preserving all already correctly extracted values. The model originally predicted the *sars-cov-2 swab test* since it is mentioned that the patient comes from *sala visite covid (room covid visits)*. Regarding dyspnea, the text clearly mentions it, so it is surprising it was originally missed. On the contrary, the *presence of respiratory distress=current* was a reasonable inference given the values of spo2, pa02, pac02, and the oxygen therapy (*positizionate cn 4 lt)*, (and the presence of dyspnea), but it is not explicitly mentioned in the text so annotations do not include it. The *foreign body in the airways* can be recognized to be negative due to the mention of *vie aeree pervie (patent airway)*. This is a very common mention in these notes, but the model does not recognize its meaning.

Dyspnea Recovery Rule

Positive lexical triggers:

- `dispnea`
- `dispnoe`
- `fiato corto`
- `affanno`
- `fame d'aria`
- `difficolta' respiratoria`

Item	Gold	Raw	Post-proc	Effect
level of consciousness	A	A	A	unchanged, correct
respiratory rate	eupneic	eupneic	eupneic	unchanged, correct
body temperature	hyperthermic	hyperthermic	hyperthermic	unchanged, correct
heart rate	normocardic	normocardic	normocardic	unchanged, correct
blood pressure	normotensive	normotensive	normotensive	unchanged, correct
spo2	86%	86%	86%	unchanged, correct
foreign body in the airways	n	unknown	**n**	recovered FN
presence of dyspnea	y	unknown	**y**	recovered FN
pa02	49	49	49	unchanged, correct
pac02	29	29	29	unchanged, correct
presence of respiratory distress	unknown	current	**unknown**	removed FP
sars-cov-2 swab test	unknown	pos	**unknown**	removed FP

Table C1: Example of information extraction from a note (914210), including gold labels (Gold) and output from the model before (Raw) and after post-processing (Post-proc).

Negation patterns:

- `nega dispnea`
- `dispnea negata`
- `non presenta dispnea`
- `assenza di dispnea`
- `senza dispnea`
- `no dispnea`

Rule:

- If a negation pattern occurs in the note, the value is set to `n`.
- Otherwise, if a dyspnea lexical trigger occurs without a nearby negation (within ± 80 characters), the value is set to `y`.

Foreign Body in Airways Rule

Positive triggers:

- `corpo estraneo`
- `inalazione di corpo estraneo`
- `aspirazione di corpo estraneo`
- `soffocamento`
- `choking`

Negative triggers:

- `vie aeree pervie`

Rule:

- If a negative trigger occurs, the value is set to `n`.
- Otherwise, if a positive trigger occurs, the value is set to `y`.

Blood Pressure Normalization

The clinical note is searched for patterns of the form

```
(PA|pressione arteriosa)
 <systolic>/<diastolic>
```

where systolic and diastolic are numerical values. The extracted values are categorized as follows:

- **hypertensive**: systolic ≥ 140 or diastolic ≥ 90
- **hypotensive**: systolic < 90 or diastolic < 60
- **normotensive**: otherwise

If such a pattern is detected, the prediction for the item is replaced with the corresponding category.

Allergy Negation Rule

The following expressions are interpreted as explicit negation:

- `allergie negate`
- `nessuna allergia`
- `no allergie`
- `nega allergie`
- `NKA`
- `NAD`
- `NKDA`

Rule:

If one of these expressions is detected, the value is set to `n`. This rule may override both `unknown` and incorrectly predicted positive values.

Evidence-based Filtering Rules

Some variables frequently produced false positive predictions. For these variables, predictions different from `unknown` are kept only if explicit textual evidence is detected. Otherwise, the value is reset to `unknown`.

Respiratory Distress
Evidence patterns include:

- `distress respiratorio`
- `insufficienza respiratoria`
- `IRA`
- `tirage`
- `uso dei muscoli accessori`
- `cianosi`

If none of these patterns occurs in the note, the prediction is set to `unknown`.

Level of Consciousness
Evidence patterns include:

- `vigile`
- `orientato`
- `GCS`
- `cosciente`

If none of these patterns occurs, the prediction is set to `unknown`.

Level of Autonomy (Mobility)
Evidence patterns include:

- `deambula`
- `cammina`
- `carrozzina`
- `bastone`
- `girello`

If none of these patterns occurs, the prediction is set to `unknown`.

Chronic Cardiac Failure
Evidence patterns include:

- `scompenso cardiaco`
- `insufficienza cardiaca`

If none of these expressions occurs, the prediction is set to `unknown`.

SARS-CoV-2 Swab Test
A prediction is retained only if the note contains the word `tampone` together with one of the following:

- `positivo`
- `negativo`
- `eseguito`
- `effettuato`
- `risultato`

Otherwise the value is set to `unknown`.

Chronic Dialysis
Evidence patterns include:

- `emodialisi`
- `dialisi`
- `dialisi peritoneale`

If none of these expressions occurs, the prediction is set to `unknown`.

History-based Inference Rules

Additional rules recover missing values from explicit mentions of therapies or comorbidities.

Anticoagulant or Antiplatelet Therapy
Mentions of the following medications trigger a positive label:

- warfarin
- coumadin
- clopidogrel
- aspirin
- cardioaspirin
- apixaban
- rivaroxaban
- dabigatran

Antihypertensive Therapy
Mentions of antihypertensive medications such as ACE inhibitors, ARBs, beta blockers, calcium channel blockers, or diuretics trigger a positive label.

Cardiovascular Diseases
Evidence patterns include:

- `cardiopatia`
- `fibrillazione atriale`
- `coronaropatia`
- `ipertensione`

Diffuse Vascular Disease
Evidence patterns include:

- `vasculopatia`
- `arteriopatia`
- `aterosclerosi`

Neuropsychiatric Disorders
Evidence patterns include:

- `demenza`
- `alzheimer`
- `disturbo psichiatrico`
- `depressione`

Living Alone
Expressions such as `vive da solo` set the value to `y`.Expressions such as `vive con la moglie` or `vive con la famiglia` set the value to `n`.

Poly-pharmacological Therapy
If the therapy section of the note contains at least three medication separators (e.g., commas or semicolons) or at least four drug tokens, the value is set to `y`.

Diagnostic Exam Rules

Rule-based extraction was also applied to imaging and diagnostic exam variables:

- brain CT scan, any abnormality
- abdomen CT scan, any abnormality
- cardiac ultrasound, any abnormality
- thoracic ultrasound, any abnormalities
- chest RX, any abnormalities
- ECG, any abnormality

For each exam, the algorithm detects whether the procedure is mentioned and whether abnormal findings are reported using terms such as *alterazioni*, *patologico*, or *anomalie*. If the exam is mentioned but no abnormality is detected, the prediction is set to `unknown`.

D. Glossary

Table D1 reports the acronyms included in the glossary, with the meaning and confidence extracted by the LLM and the true meaning. Except for two cases, all the others are wrong meanings despite high confidence of the LLM.

Acronym	Known	Conf.	Known Meaning	True Meaning
EON	no	0	–	Esame Obiettivo Neurologico (Neurological physical examination)
CE	yes	80	Codice di Emergenza (Emergency Code)	Corpo Estraneo (Foreign Body)
FA	yes	80	Fibra Ariana, Frattura dell'Anca (Ariana Fibre, Hip Fracture)	Fibrillazione Atriale (Atrial Fibrillation)
NAO	yes	80	Non Applicabile Operativamente (Not applicable in practice)	Nuovi Anticoagulanti Orali (New Oral Anticoagulants)
TAO	yes	90	Trombosi Arteriosa Ostiale (Ostial Arterial Thrombosis)	Terapia Anticoagulante Orale (Oral anticoagulant therapy)
TD	yes	80	Tachicardia Da (Tachicardia Da)	Terapia Domiciliare (Home Care)
APR	yes	80	Acid Peptic Reflux	Anamnesi Patologica Remota (Past medical history)
AASS	yes	100	Automated External Defibrillator	Arti Superiori (Upper Limbs)
TC	yes	90	Trauma Cranico (Head Injury)	If associated with degrees: Temperatura Corporea (Body Temperature); in radiology context: Tomografia Computerizzata (CT scan)
HGT	no	0	–	Hemo Gluco Test
EGA	yes	100	Età Gestazionale Anamnestica (Gestational Age)	Emogasanalisi (Blood gas analysis)
NRS	yes	80	Numero Registro Sanitario (Number of Health Registry)	Numerical Pain Rating Scale
GB	yes	80	Gravità del Bene (Severity of Goodness)	Globuli Bianchi (White Blood Cells)

Table D1: Acronyms of the glossary extracted from the unlabelled corpus, with the meaning and confidence reported by the LLM and the true meaning.