Overview of the CRF 2026 Shared Task on Clinical Case Report Forms Filling

Pietro Ferrazzi[1,2], Soumitra Ghosh[1], Alberto Lavelli[1], Bernardo Magnini[1]
[1]Fondazione Bruno Kessler, Povo, Trento, Italy
[2]University of Padova, Padova, Italy

Abstract

Case Report Forms (CRFs) are structured instruments widely used in clinical research to systematically collect patient information according to predefined protocols. In practice, CRFs are often manually completed by clinicians based on patients' clinical reports, a process that is time-consuming and prone to inconsistencies. Despite their central role in medical studies, automatic population of CRFs from clinical narratives remains largely underexplored in the Natural Language Processing community, partly due to the scarcity of publicly available datasets. In this paper, we present the CRF Filling Shared Task, organised at the CL4Health Workshop at LREC 2026, which aims to advance research on automatic extraction of structured clinical information from unstructured patient notes. The task consists of assigning the correct value to a set of predefined CRF items given a clinical note. The target dataset is derived from a real-world CRF for dyspnea assessment, comprising 134 medical items with predefined value sets. The task is provided in two languages, Italian and English. We describe the dataset, the task formulation, and the evaluation framework, and discuss the participating systems and their results. By introducing this shared task, we aim to stimulate research on clinically applicable NLP systems for structured data extraction in healthcare.

Keywords: case report form, shared task, medical information extraction, medical question answering

1. Introduction

Case Report Forms (CRFs) are structured data collection instruments used in clinical research to systematically record patient information. They are a fundamental tool in medical practice and research: clinicians use structured forms in Emergency Departments during patient admission, physicians rely on them to monitor specific conditions, and researchers use them to identify and characterize eligible patients for clinical studies (Richesson and Nadkarni, 2011; Bellary et al., 2014). By standardising how information is collected and represented, CRFs improve efficiency through normalisation and structured documentation, and enhance effectiveness by providing a clear, organised overview of available clinical data (Pétavy et al., 2019; Rinaldi et al., 2025; Lin et al., 2015). With the rapid progress of Natural Language Processing (NLP) methods, interest in automatic CRF filling has grown, as researchers explore whether automating this time-consuming task can help streamline clinical data collection and support medical research workflows (Mac Kenzie et al., 2016). Gutiérrez-Sacristán et al. (2024) were among the first to propose an automatic population of case report forms from clinical notes based on NLP techniques, while Crema et al. (2024) introduced the usage of Large Language Models for the task, by designing a BERT-based approach that frames the task as extractive question asnwering. These studies have opened a promising research direction with significant potential impact in real-world clinical settings, where substantial time is spent manually completing CRFs based on patients' clinical notes. However, the applied methods do not explore the full set of state-of-the-art NLP approaches, which remain largely underexplored. Furthermore, the mentioned approaches do not release any dataset that can be used to advance research in the field. For these reasons, we propose the CRF Filling Shared Task, hosted at the CL4Health Workshop at LREC 2026, to foster the development of systems capable of automatically populating CRFs from clinical notes (an intuitive example of the task is shown in Figure 1). We combine a set of data sources, spanning from manually annotated CRFs from emergency department settings (Kaczmarek et al., 2026) to semi-automatically annotated examples from scientific literature and medical exams (Ferrazzi et al., 2025), and more than two thousand manually selected unannotated clinical notes from those presented in Ferrazzi et al. (2026). All datasets are released through our HuggingFace organization account[1]. In this paper, we present an overview of the Shared Task task, including the provided datasets, a brief description of the best systems proposed by each of the 12 participating teams, selected from the 32 valid test submissions. Moreover, we discuss the results by comparing submissions and identifying observed trends and similarities.

[1] https://huggingface.co/collections/NLP-FBK/crf-filling-sharedtask-cl4health2026

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 245–254

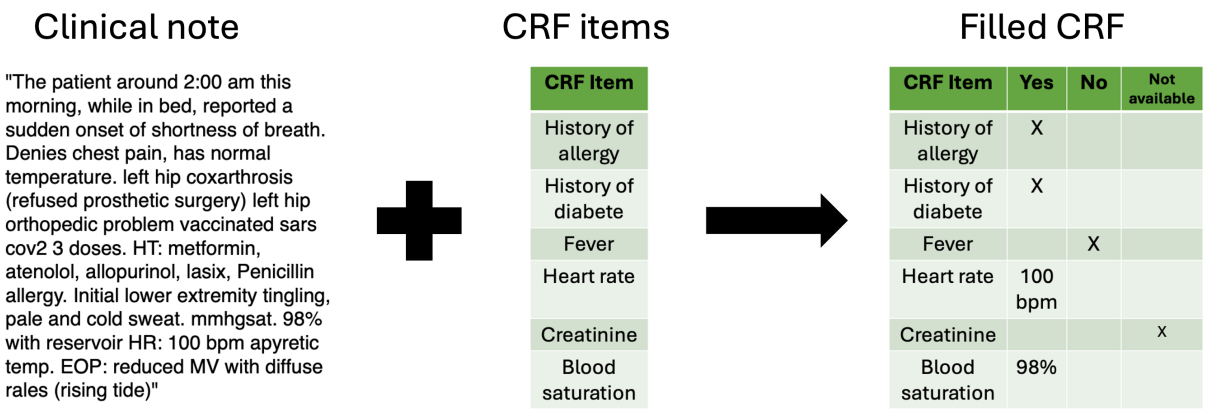

Figure 1: Example of the CRF filling task. Given a clinical note describing a patient's history and medical conditions, the task is to automatically fill a predefined list of medical items based on the note's content.

2. Task Description

The Dyspnea CRF consists of a list of 134 medical items that can assume different values. The task consists of assigning the correct value to each item given a patient's clinical note.
As an example, the item *"chronic pulmonary disease"* has to be filled with a value from *"certainly chronic"*, *"certainly not chronic"*, *"possibly chronic"*. Whenever the note does not provide sufficient evidence, the correct output is *"unknown"*, which is always considered a valid answer.

The task is available in two languages: Italian and English. Multilingualism is intended to encourage the development of robust methods applicable across languages and diverse clinical settings, while also reflecting the linguistic variability of real-world healthcare data.

Participants are provided with an HuggingFace collection with all the data[2], the CodaBench portal for systems' results submission[3], and a GitHub repository for local evaluation during development[4].

3. Datasets

The development of systems to automatically perform CRF filling is hampered by a lack of annotated data. This happens because manual annotation of clinical notes requires expert medical doctors, who are not often available for annotation tasks, given the workload in healthcare settings. Nevertheless, other sources can be leveraged to design systems that perform the task. For this reason, the data we propose as training set combine (i) few gold-standard examples of the actual task at hand; (ii) much more unannotated data from the same data distribution, which experts did not annotate; and (iii) some annotated examples of tasks with some degrees of similarity with the task at hand, but not identical or coming from the same data distribution. Then, to train systems to fill the Dyspnea CRF, participants are provided with three datasets:

- 10 **gold-standard** pairs of clinical notes with manually filled Dyspnea CRF. These are intended to showcase the task. They can be used for few-shot settings, a base for data augmentation, etc.
- 2667 **unannotated** clinical notes about patients with Dyspnea. These are intended to provide participants with knowledge about the target data distribution.
- 71 (80 for Italian) pairs of clinical notes and **semi-automatically annotated** CRF. These are intended as examples of how the CRF filling task can be performed in domains other than Dyspnea.

The gold-standard pairs are the ones that describe the task at hand, while the unannotated clinical notes about Dyspnea patients and the semi-automatically annotated pairs are provided as extra data, potentially helpful in solving the task, but not strictly necessary. **Development** and **test** can be performed on two gold-standard datasets of 80 and 200 clinical notes, respectively, coming from the same source as the 10 provided for training (Kaczmarek et al., 2026).

Gold-standard The gold-standard dataset (Kaczmarek et al., 2026) is comprises (i) clinical notes reporting patients' history and (ii) the manually annotated Dyspnea CRF, a list of 134 medical items

[2] https://huggingface.co/collections/NLP-FBK/crf-filling-sharedtask-cl4health2026
[3] https://www.codabench.org/competitions/11984/
[4] https://github.com/hltfbk/CRF-filling-CL4Health2026

filled by medical doctors, designed to capture clinical information in the setting of an Emergency Department (ED). For each clinical note, medical doctors from the eCREAM European Project have manually filled the 134 CRF items.
The *clinical notes* (10 for training, 80 for development, and 200 for testing) refer to patients of the San Giovanni Bosco hospital in Turin (Italy) who went to the emergency department with Dyspnea (a respiratory condition that involves shortness of breathing). All notes have been anonymized to preserve privacy according to the ethical protocol defined in the context of the eCREAM European Project. All information that could identify patients, such as family members, locations, names, and phone numbers has been replaced by placeholders. The notes come from different sources inside the Emergency Department, including anamnesis reports, triage records, nursing care notes, specialist consultation notes, medical visit reports, diagnostic reports, and discharge summaries.
Each *CRF item* is associated with a predefined label space that reflects the type of information to be extracted. If the note does not contain the information required to fill an item, its filling value is set to "unknown". The most common label types include: (i) binary categorical labels (e.g., yes/no, positive/negative, present/absent), (ii) ordinal labels, typically used to indicate whether a measurement falls below, within, or above the normal range (e.g., body temperature can be hypothermic, normotermic, hypertermic), and (iii) for those clinical tests whose outcomes require contextual interpretation, the filling value can be "measured" if it has been collected, "unknown" otherwise.
The dataset is characterized by a **high imbalance towards "unknown"** values. More than 90% of the items are annotated with such label, as the CRF is designed to collect all the potentially relevant information, while patients' notes often report just a small fraction of them.

Gold-standard data distribution On average, each gold-standard note contains 5.7 annotated items, although the distribution is uneven: some notes include more than 15 annotated items, while around 20% (58/290) contain none. The average length of the notes is 131 words, with some outliers reaching up to 400 words. Note length also varies across document types, with triage and discharge records typically shorter than others.

Coverage of the CRF is partial, with approximately three-quarters of the items annotated at least once in the dataset. The annotations also exhibit characteristic properties of clinical documentation. In particular, redundancies are common, with 28% of notes containing repeated annotations of the same item with identical values, while a small proportion (0.8%) includes multiple assignments of the same item with differing values. These patterns do not reflect annotation errors, but rather the longitudinal and sometimes evolving nature of clinical information.

Unannotated Notes We provide a set of 2667 unannotated clinical notes of patients with Dyspnea as additional source of knowledge about this medical condition. These notes are a manually selected subset from the much larger dataset of clinical cases presented in Ferrazzi et al. (2026), which contains 1.9M anonimyzed clinical notes from the emergency department of the San Giovanni Bosco hospital, Turin (Italy).

Semi-automatic CRFs As a further resource to enhance task understanding at training time, we include the semi-automatically generated CRFs proposed in Ferrazzi et al. (2025). There, we proposed a methodology to convert medical datasets curated for Named Entity Recognition into annotated CRF filling data by (i) creating shared lists of items for clinically similar notes, and (ii) automatically filling each item utilizing the available manual annotation information. We utilise the dataset composed of 71 Italian and 80 English notes. Such notes are grouped into clusters (7 for Italian, 8 for English) based on patients' histories and conditions. Each cluster of notes is associated with a cluster-specific CRF that contains medical items relevant to the cluster itself. For this reason, these items do not necessarily overlap with the target Dyspnea CRF ones, but are intended as a means to learn the CRF filling task, regardless of the medical characteristics.

Data Translation The clinical notes collected from the San Giovanni Bosco hospital are in Italian, while the semi-automatic ones are both in Italian and English. As we aim to propose a multilingual task which includes both Italian and English, we translated the clinical notes into English, following two different approaches. The *gold-standard* clinical notes have been manually translated into English by professional translators, while the unannotated clinical notes have been automatically translated into English using GPT-5. For the automatic approach, we evaluate the translation quality via back-translation (Rapp, 2009). Such a method involves translating the target text back into Italian and comparing the original version with the back-translated one. While this process may compound errors across translation steps, prior work has shown that back-translation scores correlate with human judgments and provide a practical proxy for translation quality (Zhuo et al., 2023). We compare the back-translated text with the source uti-

lizing four metrics, showcasing good translation quality (COMET, CHRF (Popović, 2015), CHRF++ (Popović, 2017), and BERTScore (Zhang et al., 2020)). We report the results in Table 1. The prompt utilized for automatic translation can be found in the Appendix.

Metric	Score
BERTScore	91.8
COMET	79.7
chrF	71.9
chrF++	68.3

Table 1: Back-translation evaluation metrics for the unannotated 2667 notes (Italian to English via GPT-5).

4. Evaluation

Model predictions are evaluated by comparing them to the ground truth label for each CRF item, using exact match as the evaluation criterion. The most intuitive metric for CRF filling would then be accuracy, calculated by comparing the filling value prediction to the ground truth for each item in the CRF among all clinical notes. This would quantify how many item values are filled correctly across the dataset. Nevertheless, the labels distribution is highly imbalanced among items, with more than 90% of them being labelled as "unknown". For this reason, a system that always predicts such a value would result in extremely high performances, which would not adequately represent reality. Therefore, we utilise **macro-F1** as the primary evaluation metric instead. This choice rewards systems that properly assign rare values, balancing the distribution-shift effect. For instance, a rare label is *"short"*, which can only be assigned to two CRF items, "duration of patient's consciousness recovery" and "duration of patients' unconsciousness". A system that does not learn to correctly assign this label will be penalised as much as a system that consistently misses the "unknown" label. This choice allows to identify systems that are capable of identifying less frequent, clinically informative outcomes.

Baseline The baseline we select to tackle the task is an LLM (`Qwen3-8B` without thinking by Yang et al. (2025)) prompted via 2-shot. For each clinical note, we prompt the LLM as many times as CRF items exist (i.e., 134 times), providing (i) the clinical note, (ii) the CRF item, (iii) two examples. The examples are selected from the training set, and always refer to the same item as the target one. The prompt is presented in Appendix.
The baseline achieves a macro-F1 of 47.0 and 53.9 in the English and Italian test sets, respectively,

5. Task Results

In this section, we report the performances of the various participating systems on the task. Table 2 shows the results for the best valid test submissions for English, while Table 3 for Italian. We present the results highlighting whether a system uses any closed-source LLM or not. While this does not represent a limitation of any kind with respect to the shared task, we believe it is useful to understand how submissions that only rely on open-source systems behave with respect to closed-source ones, as they can be locally deployed in healthcare institutions.
The team that achieved the best performance on both Italian and English is *Aurum*, who divides the 134 items into 14 groups, and optimizes a Chain-of-Thought prompt for each group using `Qwen3-Max`. The best system relying on open-source models for English is *MdL*, which leverages an entity recognition step before prompting `Qwen3-8B` via 1-shot. For Italian, the best performing system with open-source models was proposed by the *Polimi* team, which divided the items into three groups based on the options labels, and prompt `Mistral-small-3.2-24B-Instruct` with a prompt including a glossary of abbreviations, followed by impactful postprocessing.
In Table 4 we report the comparison of the results on Italian and English for those submissions that handled both languages. It can be seen that systems perform better in English.

Team name	**CS**	**Macro-F1**
Aurum	✓	68.3
DocUA	✓	62.8
Gladiators	✓	59.7
MdL		57.1
NM	✓	55.7
SEBIS		55.3
DocUA		54.4
Jovian_Tech	✓	52.4
Greyc		48.1
Innov8rs	✓	47.1
BASELINE		**47.0**
LTRC-CRF		38.3

Table 2: Results for English test. "CS" refers to the system using any closed-source model.

6. Submissions

In this section, we provide an overview of the ten submissions received through the Test Phase of the CodaBench portal, together with a detailed description. We report a summary of some relevant

Team Name	CS	Macro-F1
Aurum	✓	67.2
Polimi		63.5
DocUA	✓	57.0
NM	✓	55.6
BASELINE		**53.9**
Innov8rs		44.7
LTRC-CRF		29.4

Table 3: Results for Italian test. "CS" refers to the system using any closed-source model.

Team Name	Macro-F1			
	IT	EN	δ	$\delta\%$
Aurum	67.2	68.3	-1.1	-2%
DocUA	57.0	62.8	-5.8	-9%
NM	55.6	55.7	-0.1	0%
Innov8rs	44.7	47.1	-2.4	-5%
LTRC-CRF	29.4	38.3	-8.9	-23%

Table 4: Side-by-side comparison of the Macro-F1 scores for the Italian and English datasets. The delta columns show the difference in scores and the percentage drop from English to Italian.

dimensions common among submissions in Table 5.

SEBIS SEBIS proposes a two-stage pipeline for CRF filling based on MedGemma-27B. Their approach separates the task into two steps: first, predicting the presence of the information associated with each CRF item, and then extracting the corresponding value when applicable. This decomposition is designed to reduce hallucinations and limit over-interpretation by enforcing closer grounding in the clinical text. The system relies on few-shot prompting without model fine-tuning.

Gladiators Gladiators proposes a hybrid system that combines Claude-4.6-Opus with rule-based post-processing. The LLM is used as the primary extractor with deterministic prompting (zero temperature), while the prompt design emphasises key extraction rules through repetition to increase the model's attention to frequently occurring clinical variables. A set of post-processing rules addresses common extraction issues, including Unicode normalisation for medical subscripts, regular-expression-based extraction of vital signs, and explicit handling of symptom negation. Additional checks verify treatments based on the presence of action verbs to reduce false positives. The system adopts a conservative strategy by defaulting to *unknown* in the absence of explicit textual evidence.

Innov8ors Innov8ors adopts a large language model ensemble approach. For the English track, they combine Gemini 2.5 Flash and Llama 3.3 70B, using a dynamic few-shot prompting strategy in which in-context examples are selected based on TF-IDF similarity to the input note. For the Italian track, the system relies on Llama 3.3 70B with static few-shot examples derived from the gold training data. To mitigate hallucinations, the authors introduce a precision-oriented filtering mechanism calibrated on the development set, which suppresses predictions for CRF items with historically high false-positive rates. This strategy is designed to improve overall macro-F1 performance by favoring precision on difficult items. Nevertheless, they experience that this filtering does not generalize to the test set.

Jovian_Tech Jovian_tech employs a few-shot prompting approach based on Claude Haiku 4.5 to automatically populate Dyspnea CRFs from clinical notes. Their system uses three gold-standard examples as in-context demonstrations to guide the extraction of the 134 CRF variables. The prompt design emphasizes conservative predictions and strict adherence to the predefined value sets for each item. The pipeline includes text normalization of the input notes, LLM-based extraction, and validation of the predicted values against the allowed option sets. The approach favors *unknown* predictions in the absence of explicit textual evidence, reflecting the sparsity of relevant information in clinical reports.

NM NM proposes a two-stage pipeline combining extraction and verification. In the first stage, a few-shot prompted Gemini 3 Flash Preview model populates CRF fields directly from the clinical note, prioritizing recall but generating a number of false positives. In the second stage, LLaMA4 Maverick 17B model reviews both the original note and the preliminary CRF output to verify each prediction and remove unsupported or ambiguous entries. This sequential strategy aims to balance coverage and precision by retaining only values that are clearly supported by the source text.

Aurum Aurum presents a modular extraction framework implemented with DSPy, decomposing the 134 CRF items into 14 domain-specific extractors (e.g., medical history, lab values, treatments, vital signs). Each module is defined through a typed DSPy signature with chain-of-thought reasoning, enabling targeted extraction within a specific clinical subdomain rather than relying on a single monolithic prompt. The prompts were iteratively refined through error analysis to reduce systematic false positives and negatives. The system uses Qwen3-

Rank	Team	Grouped	CS	Pipeline	FP/FN trade-off	Few-shot	Post-proc	Extra data
1	Aurum	✓	✓	✓	✓			
2	Polimi			✓	✓		✓	✓
3	DocUA	✓	✓	✓	✓		✓	
4	Gladiators	✓	✓	✓			✓	
5	MdL	✓		✓			✓	
6	NM		✓	✓	✓	✓		
7	SEBIS			✓	✓	✓		
8	Jovian_tech	✓	✓	✓		✓		
9	Greyc			✓	✓	✓		
10	Innov8ors		✓	✓	✓	✓	✓	

Table 5: Revised overview of system design choices across participating teams. Checkmarks indicate only explicitly supported evidence from system descriptions. Blank cells denote insufficient or implicit evidence. **Grouped** refers to the fact that multiple items have been grouped together into one single prompt; **CS** to the usage of closed-source models in the submission; **Pipeline** to the concatenation of multiple steps (instead of a single LLM call); **FP/FN trade-off** to whether the submission analyses and tries to optimize the trade-off between generating as few False Positives as possible, at the cost of generating more False Negatives; **Few-shot** if at any step, few shots are provided as examples; **Post-proc** refers to the presence of deterministic post-processing on the predictions; **Extra data** for the use of a dataset other than the gold-standard.

Max (Thinking) as the underlying model and performs consistently across both languages.

MdL MdL adopts a two-step approach combining rule-based detection with LLM-based normalization. First, relevant CRF item categories are identified through synonym matching and entity recognition using the GliNER model, with synonyms compiled from the training set to capture common variants and acronyms. Predictions from both methods are merged to select the most relevant item categories and filter out those likely to be *unknown*. In the second step, the selected items and the clinical note are provided to a Qwen 8B model through a structured prompt to normalize the extracted information and assign the appropriate CRF values according to predefined rules.

Polimi Polimi focuses on the Italian track. After evaluating several configurations, their best system uses Mistral-small-3.2-24B-Instruct in a zero-shot setting with three specialized prompts tailored to different item types (binary, categorical, and measured variables). The prompts incorporate a glossary of clinical abbreviations automatically extracted from the unlabeled notes. In addition, rule-based post-processing is applied to items with low recall but identifiable textual patterns. The system was tuned on the development set and tends to favor conservative predictions, producing more false negatives than false positives.

Greyc Greyc proposes a modular LLM-based framework consisting of three components: a rewriting model, an extractor, and a judging model. The rewriting stage reformulates the clinical note with respect to each CRF item, aiming to highlight relevant information. A second LLM then extracts candidate values using a k-shot prompting strategy, while a third LLM acts as a judge to verify whether sufficient textual evidence supports the prediction, defaulting to unknown otherwise. While the rewriting stage improves recall, it also increases false positives, which the judging step attempts to mitigate through a more conservative verification process.

DocUA DocUA introduces LLM StructCore, a two-stage system combining schema-guided reasoning with deterministic post-processing. In the first stage, Mistral Large 3 (675B) generates structured JSON summaries using a set of schema-guided reasoning patterns designed to ensure consistent output structure. In the second stage, deterministic rules perform canonicalization, vocabulary normalization, and mapping to standardized medical concepts through UMLS, along with evidence-based checks to reduce false positives. This hybrid design combines LLM-based extraction with rule-based normalization to produce structured CRF outputs.

7. Discussion

The participating systems reveal several emerging trends in approaches to automatic CRF fill-

ing from clinical notes. All submissions rely on large language models (LLMs), often combined with structured prompting strategies and multi-steps pipelines to improve reliability and reduce hallucinations. Almost all submissions identify as main challenge the high sparsity of the data, given that around 95% of the CRF items' labels are set to *unknown*, and analysed the trade-off between False Positives and False Negatives. Only one submission (Polimi) utilized data other than the gold-standard set, achieving the best results for Italian, and the best results overall without using closed-source models.

Conservative prediction strategies. Many systems adopt conservative strategies to avoid hallucinations, frequently defaulting to the *unknown* value unless clear textual evidence is present. Overall, participants observed that LLMs always tend to fill items with values other than *unknown* even when explicit information is not present in the clinical note. Many teams present the task as a trade-off between generating as few False Positives (FP) as possible, while accepting to generate some False Negatives (FN).
For instance, Innov8ors introduce a development-set calibrated filter that suppresses predictions for items with historically many FP, which results in too much bias towards the development set. NM reports that open-weight models tend to reduce FP, being more conservative than closed-source counterparts. Greyc show that their judging component eliminates a large portion of FP but also removes a substantial number of true positives. Aurum design works at a prompt level, achieving a 92% reduction of FP, at the cost of a 30% increase of FN. Polimi observed that LLM-based verification is not as effective as rule-based postprocessing driven by pattern finding in the original clinical notes.

Multi-stage pipelines. All teams decomposed the task into multiple stages. For example, SEBIS and NM employ a two-step approach, separating extraction from verification. In these pipelines, a first model predicts candidate values for CRF items, while a second model verifies whether the predictions are supported by the clinical note. Similarly, Greyc proposes a three-component architecture involving a rewriting step, an extraction stage, and a judging model that validates the predictions against the original note. These designs consistently aim to improve precision by filtering unsupported predictions.

Prompt engineering and task decomposition. Prompt design plays a key role in several submissions. SEBIS show through ablation studies that few-shot prompting improves performance and that separating the prediction of item presence from value extraction further increases reliability. Innov8ors dynamically select in-context examples based on TF-IDF similarity to the input note, while Jovian_tech demonstrate the feasibility of predicting all CRF items simultaneously using a small set of fixed few-shot examples. Other teams explored task decomposition strategies. Aurum divide the CRF into domain-specific extractors implemented through DSPy signatures with chain-of-thought reasoning, enabling targeted extraction across clinical subdomains. Polimi instead design specialized prompts for different categories of items (binary, categorical, and measured variables), while Greyc employ k-shot prompting in an item-wise extraction framework. Overall, these approaches highlight how prompt design and task structuring can significantly influence extraction performance in CRF filling, and few-shot being an effective approach.

Hybrid LLM and rule-based systems. Several systems combine LLM predictions with rule-based post-processing. Polimi, for instance, augment prompts with a glossary of clinical abbreviations automatically extracted from unlabeled notes and apply pattern-based rules for items with identifiable textual cues. Gladiators similarly integrate rule-based components to address systematic extraction errors, including regular-expression extraction of vital signs, Unicode normalization for medical symbols, and explicit handling of symptom negation. DocUA adopts a comparable hybrid design, combining LLM-based extraction with deterministic post-processing for canonicalization, vocabulary normalization, and mapping to standardized medical concepts. These approaches highlight the continued value of domain knowledge and manual analysis of clinical text when designing reliable extraction pipelines.

8. Conclusion

In this paper, we presented an overview of the CRF filling Shared Task. A number of insights emerge across submissions. Overall, the task is far from being saturated, with the best performing system reaching a score of 68.3 and 67.1 for English and Italian respectively. Not surprisingly, systems powered by closed-source models perform better than open-source ones. Nevertheless, the gap can be significantly reduced by post-processing approaches based on manual or semi-automatic mappings. This highlights the importance of in-depth analysis of the dataset characteristics.

All participants faced the trade-off between precision and coverage, with systems that aggressively reduce false positives often sacrifice recall. To address this challenge, several teams report that man-

ual analysis of difficult items and targeted prompt or rule adjustments can yield measurable improvements, suggesting that future work may benefit from more fine-grained modeling of item-specific extraction patterns.

Overall, the shared task results indicate that while LLMs provide a strong baseline for CRF filling, robust performance often requires hybrid architectures combining prompting strategies, verification steps, and domain-specific post-processing.

9. Limitations

The main limitation of the Shared Task relies on the size of the gold-standard dataset. Overall, it comprises 290 clinical notes from a single Italian hospital. Future work should focus on increasing the size and diversifying the sources.

10. Acknowledgments

This work has been partially funded by the European Union under the Horizon Europe eCREAM Project (Grant Agreement No.101057726). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

11. Bibliographical References

Shantala Bellary, Binny Krishnankutty, and M. S. Latha. 2014. Basics of case report form designing in clinical research. *Perspectives in Clinical Research*, 5(4):159–166.

Claudio Crema, Federico Verde, Pietro Tiraboschi, Camillo Marra, Andrea Arighi, Silvia Fostinelli, Guido Maria Giuffré, Vera Pacoova Dal Maschio, Federica L'Abbate, Federica Solca, Barbara Poletti, Vincenzo Silani, Emanuela Rotondo, Vittoria Borracci, Roberto Vimercati, Valeria Crepaldi, Emanuela Inguscio, Massimo Filippi, Francesca Caso, Alessandra Maria Rosati, Davide Quaranta, Giuliano Binetti, Ilaria Pagnoni, Manuela Morreale, Francesca Burgio, Michelangelo Stanzani-Maserati, Sabina Capellari, Matteo Pardini, Nicola Girtler, Federica Piras, Fabrizio Piras, Stefania Lalli, Elena Perdixi, Gemma Lombardi, Sonia Di Tella, Alfredo Costa, Marco Capelli, Cira Fundarò, Marina Manera, Cristina Muscio, Elisa Pellencin, Raffaele Lodi, Fabrizio Tagliavini, and Alberto Redolfi. 2024. Medical information extraction with nlp-powered qabots: A real-world scenario. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6906–6917.

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 pretraining 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, Vienna, Austria. Association for Computational Linguistics.

Alba Gutiérrez-Sacristán, Simran Makwana, Audrey Dionne, Simran Mahanta, Karla J. Dyer, Faridis Serrano, Carmen Watrin, Pierre Pages, Sajad Mousavi, Anil Degala, Jessica Lyons, Danielle Pillion, Joany M. Zachariasse, Lara S. Shekerdemian, Dongngan T. Truong, Jane W. Newburger, and Paul Avillach. 2024. Development and validation of an open-source pipeline for automatic population of case report forms from electronic health records: a pediatric multicenter prospective study. *eBioMedicine*, 108. Doi: 10.1016/j.ebiom.2024.105337.

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.

Ching-Heng Lin, Nai-Yuan Wu, and Der-Ming Liou. 2015. A multi-technique approach to bridge electronic case report form design and data standard adoption. *Journal of Biomedical Informatics*, 53:49–57.

W. R. Mac Kenzie, A. J. Davidson, A. Wiesenthal, J. P. Engel, K. Turner, L. Conn, S. J. Becker, S. Moffatt, S. L. Groseclose, J. Jellison, J. Stinn, N. Y. Garrett, L. Helmus, B. Harmon, C. L. Richards, J. R. Lumpkin, and M. F. Iademarco. 2016. The promise of electronic case reporting. *Public Health Reports*, 131(6):742–746. Epub 2016 Oct 13.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*.

Frank Pétavy et al. 2019. Global standardization of clinical research data. *Applied Clinical Trials*, 28(4):20–23.

Reinhard Rapp. 2009. The backtranslation score: Automatic MT evalution at the sentence level without reference translations. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Rachel L. Richesson and Prakash Nadkarni. 2011. Data standards for clinical research data collection forms: current status and challenges. *Journal of the American Medical Informatics Association*, 18(3):341–346.

Eugenia Rinaldi, Caroline Stellmach, and Sylvia Thun. 2025. How to design electronic case report form (ecrf) questions to maximize semantic interoperability in clinical research. *Interactive Journal of Medical Research*, 14:e51598.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*.

A. Appendix

A.1 Translation Prompt

Here is the system prompt utilized for automatic translation of the unannotated clinical notes:

```
You are a clinical-grade translation model specialized in emergency-department notes. Translate all text from Italian → English.

## Requirements
  • Preserve all medical facts, numeric values, units, abbreviations, and dates.

  • Maintain the structure and conciseness typical of ED notes (headings, bullet points, line breaks).

  • Use professional clinical English, not patient-friendly language.

  • Do not summarize, interpret, add, infer, or remove information.

  • Translate abbreviations to their standard English ED equivalents when unambiguous (e.g., PA → BP, FC → HR). If ambiguous, leave them unchanged.

  • If a term is unclear, leave it unchanged.

  • The notes have been anonymized: if you find terms such as NOME_PERSONA, LUOGO, PARENTE, NUM_TELEFONO leave them unchanged.
## Output
Return a json file with the English translation:
{
"translation":
"<full translated text here>"
}
```

This is the user prompt:

```
Translate the following ED note from Italian to English according to the system instructions: note
```

A.2 Baseline Prompt

Here we report the prompts used for the 2-shot `Qwen3-8B` baseline for both Italian and English.

System prompt:

```
You are a helpful medical assistant that extracts information in clinical notes based on patient history.
Items must be filled with one of the valid values.
If the information is not explicitly stated in the clinical note, answer 'unknown'.
```

User prompt:

```
Here is the patient history:
< < <{clinical_note}> > >

Fill this item:  {item}

Valid values are:  {valid_values}
```