InfoLab-FEUP at MultiClinSum-2: Agentic Multilingual Clinical Case Summarization with Biomedical Concept Enrichment

Artur Oliveira[1,2], Tiago Rodrigues[1,3] and Carla Teixeira Lopes[1,2]

[1]Faculty of Engineering, University of Porto, Porto, Portugal
[2]INESC TEC, Porto, Portugal
[3]Opnova

Abstract

Clinical case reports contain detailed patient information, but their length and complexity make automatic summarization challenging, particularly in multilingual clinical settings. We present a multilingual clinical case summarization system that uses an agentic pipeline to generate summaries for English and Portuguese clinical case reports. Each report is first decomposed into four clinical sections: patient characteristics, diagnostics, interventions, and outcomes. These structured sections are then reviewed through an iterative agent-critic loop and used as context for final summary generation. To improve clinical grounding, the pipeline incorporates biomedical keyword extraction. Three enrichment configurations were evaluated: keywords only, keywords with concept definitions, and keywords with definitions assigned to the corresponding clinical sections. For Portuguese, two pipelines were compared: a direct Portuguese pipeline and a translation-assisted pipeline in which the source text was translated into English, summarized using the English pipeline, and translated back into Portuguese. The system was evaluated using lexical-overlap, semantic-similarity, and qualitative criteria including faithfulness, completeness, fluency, and consistency. Results show that concept definition enrichment produced the strongest English configuration, while direct Portuguese generation with keyword and definition enrichment achieved the best Portuguese automatic scores. The translation-assisted workflow performed worse than direct Portuguese generation, suggesting that the added translation steps introduced more noise than benefit in this setting.

Keywords

Clinical text summarization, Clinical case reports, Large Language Models, Multilingual summarization, Agentic workflow, QuickUMLS, Biomedical enrichment

1. Introduction

Clinical documentation continues to grow in both volume and complexity, creating significant challenges for healthcare professionals who must synthesize large amounts of medical information into concise and clinically relevant summaries [1]. Recent advances in Large Language Models have demonstrated strong potential for biomedical summarization tasks, though factual inconsistencies, and hallucinations remain major issues [2, 3]. The MultiClinSum-2 shared task, organized as part of the fourteenth BioASQ challenge, provides an opportunity to investigate solutions to this problem by evaluating multilingual clinical summarization systems across clinical case reports in multiple languages [4, 5].

In this work, we present a multilingual agentic summarization framework that combines the extraction of clinical information into structured sections, keyword extraction and enrichment using QuickUMLS [6] and the UMLS Metathesaurus [7], and iterative critic-feedback refinement. The system decomposes case reports into clinically relevant sections, including patient characteristics, diagnostics, interventions, and outcomes, before generating the final summary through local LLM inference. The pipeline was also adapted to Portuguese, where we explored both direct target-language summarization, using Portuguese prompts and Portuguese UMLS-based resources for keyword extraction and definition enrichment, and a translation-assisted workflow, where Portuguese case reports were translated into

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
artur.alb@outlook.com (A. Oliveira); trodrigues@opnova.ai (T. Rodrigues); ctl@fe.up.pt (C. Teixeira Lopes)
https://github.com/ArturAlb (A. Oliveira)

English and processed using English prompts and English UMLS resources. This comparison was intended to evaluate the trade-off between native-language processing and an English-centric strategy supported by broader English UMLS coverage.

Our main goal is to evaluate whether a structured, keyword enriched, and locally deployable LLM pipeline can improve multilingual clinical case summarization while preserving clinically relevant information. In particular, we compare different system configurations to analyze how each affects summary quality.

2. Related Work

Clinical text summarization is a long-standing task in biomedical natural language processing, motivated by the need to condense complex medical information into concise and clinically meaningful summaries. This challenge is made harder by the need to preserve factual accuracy, relevant medical terminology, and important diagnostic or therapeutic details. Existing approaches are commonly divided into extractive and abstractive methods: extractive systems select relevant sentences or passages directly from the source document, while abstractive systems generate new text that synthesizes and reorganizes the source information [1, 2, 8].

A further development is the use of modular, agentic, and multi-path orchestration to divide complex information processing into more controlled stages. Rather than treating context as a single undifferentiated block, these systems use query planners, specialized agents, or multi-turn retrieval strategies to identify which information should be retrieved, and passed forward. Volkova et al. [9] illustrates this idea in a cross-disciplinary setting, using a central query-planning component and specialized retrieval, translation, and reasoning agents to select relevant knowledge sources and synthesize information across multiple domains. Choi et al. [10] applied a multi-agent design to pharmacovigilance, where specialized agents retrieve and classify drug, adverse event associations, and critic agents validate intermediate outputs before they are passed to later stages. Similar agent-based architectures were applied to clinical report management, academic research support, and online mental health assistance [11, 12, 13].

These works suggest that context division can support the handling of long and heterogeneous medical information by separating into task-specific steps.

3. Methodology

This work implements a modular clinical case summarization pipeline based on large language models. The pipeline was designed to decompose clinical case reports into smaller clinically meaningful sections before final summary generation. Instead of prompting the model with the full unstructured case report, the system first extracts relevant clinical sections, optionally enriches the extracted context with biomedical concepts and UMLS definitions, and then generates the final summary from the structured representation.

The proposed approach combines LLM-based and non-LLM-based modules. LLMs are used in stages that require interpretation, rewriting, or classification, such as clinical section extraction, critic revision, section enrichment, and final summary generation. In contrast, biomedical concept extraction and definition lookup rely on external biomedical resources, namely QuickUMLS and the UMLS Metathesaurus. This separation was intended to make the pipeline more controllable, since clinical information extraction, biomedical grounding, and summary generation could be modified or evaluated independently.

The pipeline was also designed to support multilingual adaptation. The same general architecture can be reused across languages, but some modules require language-specific adaptation. In particular, prompts must be translated or rewritten for the target language, and biomedical concept extraction depends on the availability and coverage of the corresponding UMLS language index. For Portuguese, we therefore evaluated both a direct Portuguese pipeline and a translation-assisted workflow, where the English pipeline and English UMLS resources could be reused after translating the source text.

The complete pipeline and its individual stages are described in Section 4, while the submitted run configurations are detailed in Section 5.2 and 5.3.

3.1. Model selection and deployment

The system was deployed locally using Ollama [14]. This choice was motivated by the objective of running the full summarization pipeline on limited consumer hardware while avoiding dependence on external APIs. Local deployment also supports data privacy, which is particularly relevant in clinical settings, even when working with de-identified texts.

During development, we evaluated two lightweight open-weight models: Llama 3 8B [15] and Mistral 7B [16] by generating summaries using the sample dataset provided for MultiClinSum-2 with the objective to compare lexical and semantic metric scores. As shown in Table 1, the difference between the two models was modest. However, Llama 3 8B obtained consistently higher scores across all evaluated metrics, including ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1. These metrics were selected because they were the ones used to evaluate competition submissions in MulticlinSum-2.

Table 1
Comparison between Llama 3 8B and Mistral 7B.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
Llama 3 8B	0.384	0.131	0.240	0.862
Mistral 7B	0.381	0.126	0.235	0.859

Due to the results observed, Llama 3 8B was selected for the final submitted runs. We also selected Llama 3 8B:instruct over the base model because the proposed pipeline requires the model to follow explicit task instructions across multiple stages, including extraction, critique, and final summary generation. The model was used with a temperature of 0.1 for minimal variance between results.

Since fine-tuning was not part of this work, the models were used as is. This reduced the memory footprint compared with a full-precision model and made it possible to execute batch generation on a consumer-grade machine with limited GPU memory.

All experiments and batch generations were executed on a local system equipped with an AMD Ryzen 7 4800H CPU, an NVIDIA GeForce RTX 2060 GPU with 6GB of VRAM, 16GB of RAM operating at 3200 MT/s, and a TEAM TM8FPK001T SSD.

4. Summarization Pipeline

Figures 1, 2, and 3 show the pipeline configurations used in the submitted runs. All configurations follow the same general structure, starting from the input clinical case report and applying clinical section extraction before the final agentic summary generation stage. The differences between the runs are related to the biomedical enrichment components included in the pipeline.

Figure 1 represents the baseline submitted configuration, where biomedical enrichment is limited to the extraction of clinical keywords. These are later appended as a list to the extracted sections, creating a context which has 2 parts: the extracted sections, and the list of keywords.

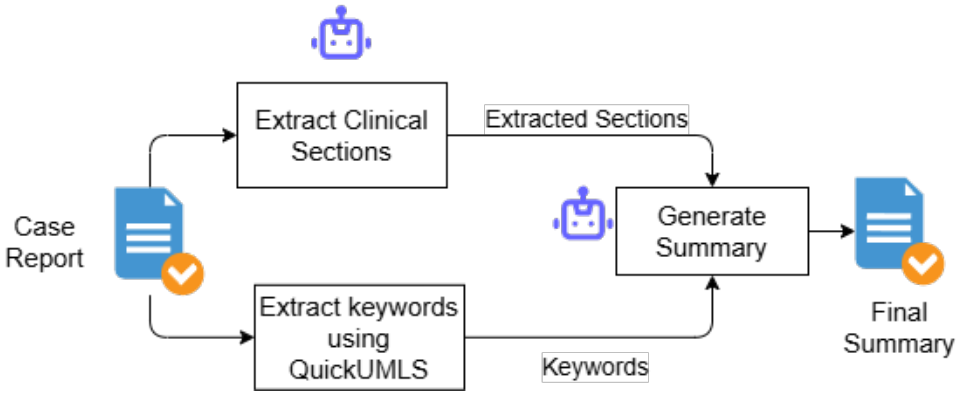

Figure 1: Pipeline configuration used for the baseline submitted run, without UMLS definition enrichment or section-aware keyword assignment. The robot icon indicates stages that require the use of an LLM.

Figure 2 extends this configuration by adding UMLS definition enrichment, allowing extracted biomedical terms to be expanded with definitions when a corresponding UMLS match is found and then provided as context in the same way as the previous configuration.

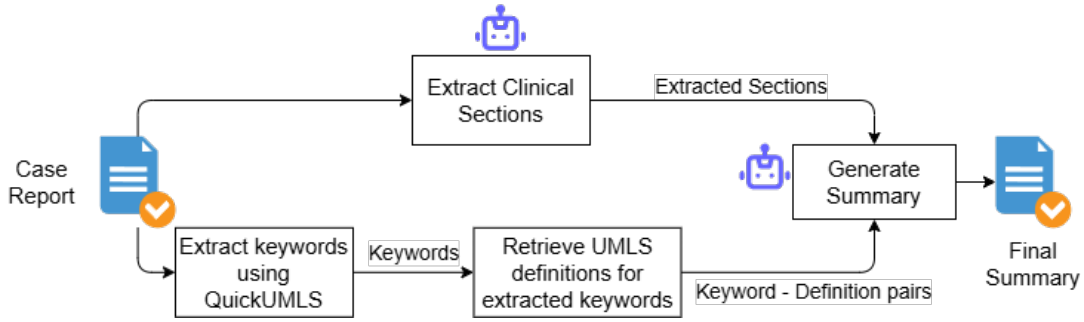

Figure 2: Pipeline configuration with biomedical enrichment through UMLS definitions, but without section-aware keyword assignment. The robot icon indicates stages that require the use of an LLM.

Finally, Figure 3 represents the complete version of the pipeline. In this configuration, the extracted keywords and their corresponding UMLS definitions are not provided to the summary generation agent as a single undifferentiated list. Instead, they are assigned to the relevant clinical sections, allowing each section to receive biomedical context that is more directly related to its content.

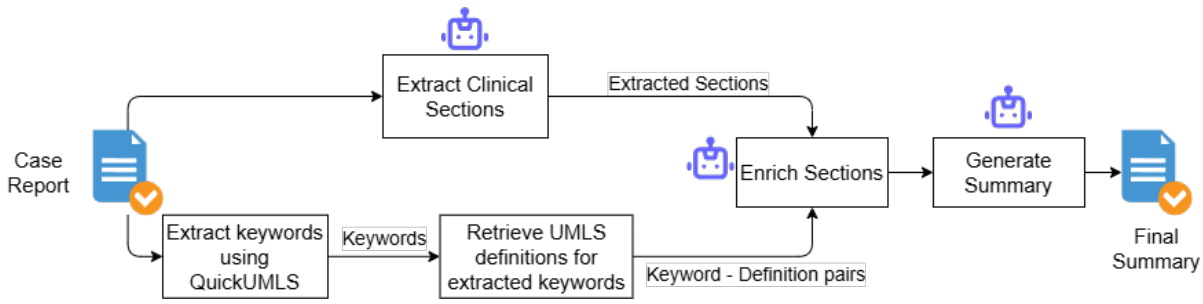

Figure 3: Complete architecture of the proposed summarization pipeline. The robot icon indicates stages that require the use of an LLM.

4.1. Clinical Section Extraction

The first step consists of a series of extraction agents, each responsible for extracting information related to a specific case report section: patient characteristics, diagnosis, interventions/treatments, and

outcomes and follow-ups: patient characteristics, diagnosis, interventions/treatments, and outcomes and follow-ups. These sections were selected based on the 13-item CARE guideline checklist for case reports, which includes *title*, *key words*, *abstract*, *introduction*, *patient information*, *clinical findings*, *timeline*, *diagnostic assessment*, *therapeutic interventions*, *follow-up and outcomes*, *discussion*, *patient perspective*, and *informed consent* [17].

Of those 13 elements, only 5 were available in the fulltexts provided for the competition: patient characteristics, clinical findings, diagnostic assessment, therapeutic interventions, and outcomes and follow-ups. Clinical findings and diagnostic assessment were merged into a single diagnostics section, since these elements are closely related and often provide complementary information about each other.

Following this, four extraction agents were implemented to isolate these specific sections. This decomposition produces a structured representation of the clinical case, allowing the final generator to rely on a structured context, with information divided into sections, rather than the full unstructured case report.

The extraction prompts followed a role-based prompting strategy, where each prompt assigned the LLM the role of a clinical extraction assistant specialized in the corresponding section. Rather than using a single generic extraction instruction, the system used a shared prompt template instantiated for patient characteristics, diagnosis, treatment/interventions, and outcomes and follow-up. This template made the expected behavior of each extraction step explicit and instructed the model to extract only information supported by the source case report.

Section extraction prompt template

```
You are a clinical information extraction assistant.

Extract ONLY the {section_name} from the clinical case report below.
Be concise and accurate.

CASE REPORT:
{full_text}
```

Following the agent-critic design pattern developed by Choi et al. [10], each section extraction was followed by a critic-review step, shown in Figure 4. In this design, reviewer feedback is returned to the generating agent, which revises its output before the pipeline continues. In the proposed pipeline, the critic agent evaluated whether the extracted content contained all the necessary information for the corresponding section. If relevant information was missing, the extraction agent was prompted to revise its output using the critic feedback; otherwise, the extracted section was accepted and added to the structured clinical context.

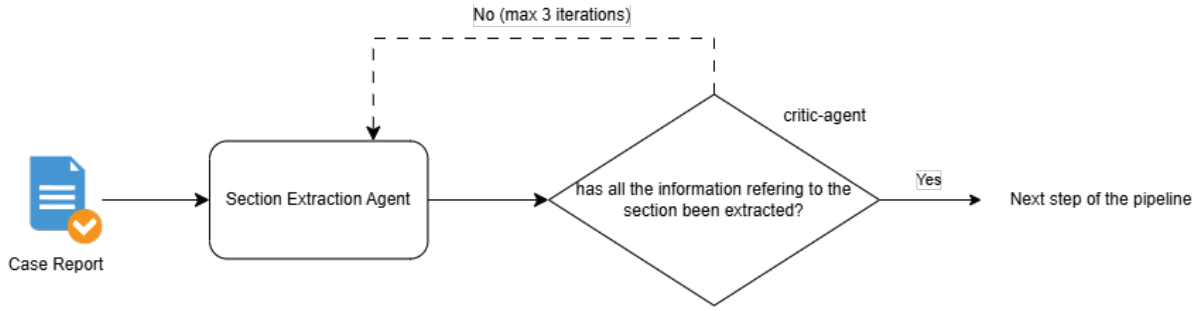

Figure 4: Generic Section Extraction with critic review loop

To avoid infinite agent-critic interactions, the maximum number of refinement loops was set to three iterations. When issues are detected, the feedback produced by the critic is appended to the end of the original extraction prompt, and the agent is asked to regenerate the corresponding section. The full critic prompt template is provided in Appendix A.1.

4.2. Keyword Extraction

A specific part of the experimentation focused on evaluating whether domain-specific keyword extraction could improve the summarization pipeline. The goal was not only to identify relevant medical terms from each case report, but also to determine whether these terms could provide useful contextual support during summary generation. For this purpose, we compared three keyword extraction approaches: spaCy [18], ClinicalBERT [19], and QuickUMLS [6]. These methods were selected to represent different levels of linguistic and domain-specific grounding.

SpaCy was included as a lightweight general-purpose baseline, relying on linguistic processing and named entity recognition to identify relevant terms with low computational cost. ClinicalBERT was tested as a neural domain-adapted alternative, motivated by previous BioASQ work showing that fine-tuned BERT-based biomedical models can achieve strong performance in biomedical named entity recognition tasks. In particular, Pamio and Di Nunzio [20] compared CRF and BERT-based approaches for biomedical NER and relation extraction, reporting that fine-tuned BERT models outperformed their CRF baselines for the NER subtask, with biosyn-sapbert-bc2gn obtaining the best submitted NER performance in their system. Such a finding suggested that specialized models could be useful for identifying clinically relevant terms. Although no model was found that was specifically trained for keyword extraction, we considered that ClinicalBert, being trained on a vast array of medical knowledge, would still be relevant to our research.

Finally, QuickUMLS allowed mapping text spans directly to concepts in the Unified Medical Language System (UMLS). Its inclusion was also motivated by previous MultiClinSum work, where QuickUMLS was used in a concept-based extractive summarization method to identify clinical concepts from source sentences [21].

After applying the three extraction methods, their outputs were compared in terms of both the number of extracted keywords and their suitability for prompt enrichment. Since no gold-standard keyword annotations were available for the clinical case reports, we performed a qualitative assessment of the extracted terms, considering four main criteria: clinical relevance, case specificity, normalization, and usefulness for section-level summarization. Clinical relevance refers to whether the extracted term corresponds to a meaningful biomedical concept. Case specificity refers to whether the term captures information that is central to the specific patient case rather than generic medical language. Normalization refers to whether the term can be linked to a standardized biomedical concept or terminology resource. Finally, prompt usefulness refers to whether the extracted term could reasonably help the language model identify or preserve relevant clinical information during summary generation.

This assessment showed that the number of extracted keywords alone was not a reliable indicator of quality. Although ClinicalBERT extracted the highest average number of keywords, as can be seen in Figure 5, manual inspection of the generated keyword lists showed that many of these terms were not directly useful for guiding the summarization process. In several cases, the BERT-based method returned broad or weakly contextualized entities that were medically related but not central to the clinical case, as can be seen in Table 2. It also produced fragmented subword units and expressions that were difficult to interpret as standalone clinical concepts. Similarly, spaCy frequently extracted general linguistic phrases, isolated adjectives, or non-specific noun chunks that reflected the surface structure of the text rather than clinically meaningful concepts. As a result, both methods introduced additional noise into the prompt, increasing the risk of distracting the language model with terms that were not essential for producing the final summary.

For this reason, QuickUMLS was selected for the submitted runs. Although it did not extract the largest number of terms, its outputs were more often recognizable biomedical concepts, more interpretable as clinical terms, and more suitable for normalization through UMLS resources. This choice was also motivated by previous MultiClinSum work, where a concept-based extractive method using QuickUMLS achieved competitive results and showed that clinical concepts extracted from UMLS could help align selected content with the medical terms present in reference summaries [21].

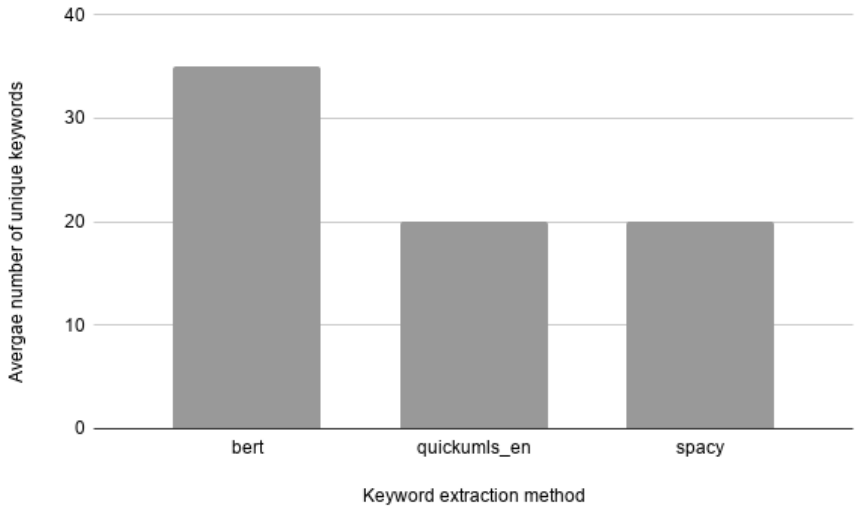

Figure 5: Average number of extracted keywords per clinical case for the three evaluated keyword extraction methods.

Table 2
Example keyword extraction outputs for one clinical case.

Method	Extracted keywords
spaCy	she; thyroid; normal; patient; neck; treatment; rash; examination; sign; week; test; serum; daily; which; difficulty; chest; time; suggestive; anterior; wolkiet
QuickUMLS-en	3 times; abdominal; activities; air; alert; anemia; anterior chest; anterior neck mass; appearance; aspirate; atrophy; auscultation; avoidance; back; blood; blood pressure; blood pressures; body; breathing; cellular
ClinicalBERT	52-year-old; violace; thighs; surgical treatment; ir; goiter; ##rita; climbing; healthy; hair loss; bilateral; two years ago; one week preceding; three; squat; po; 100 mg; body fatigue; anterior; heat intolerance; multi; generalized; oral; rash; toxic; diagnosis; stiff; female; 30 years; prop; ##thiouracil; ##yl; symptoms; four months prior; anterior neck; swelling; six months following

4.3. UMLS definition enrichment

The definition enrichment stage was designed to provide the LLM with additional context for clinically relevant terms. After QuickUMLS extracted candidate biomedical keywords from the case report, a separate UMLS lookup module attempted to associate each extracted keyword with a UMLS Concept Unique Identifier and retrieve a corresponding definition. This module used the UMLS Metathesaurus files directly, relying on `MRCONSO.RRF` to map English concept names to Concept Unique Identifiers and on `MRDEF.RRF` to retrieve definitions associated with those CUIs. During this process, each keyword was lowercased and searched in the UMLS term index. When an exact match was not found, a simple singular fallback was attempted for terms ending in `s`. If no matching CUI or definition was available, the keyword was kept unchanged.

As a result, the enrichment stage produced a mixed list of plain keywords and keyword-definition pairs. This allowed the prompt to include not only clinically relevant terms, but also short explanatory information for terms that could otherwise be ambiguous, abbreviated, or highly specialized. This was particularly relevant for diagnoses, procedures, symptoms, and domain-specific terminology that may be difficult for the model to interpret from the source text alone.

4.4. Section Enrichment

Following the extraction strategy described above, the extracted keywords and their corresponding definitions were also assigned to their respective clinical sections, namely patient characteristics, diagnostics, interventions, and outcomes. This allowed for the model to receive keywords directly associated to each section rather then a simple list of concepts.

This stage corresponds to the section enrichment component introduced in Run 3. The assignment was performed using a role-based prompt that instructed the LLM to act as a clinical terminology specialist. The model received the structured clinical context and the list of keyword-definition pairs, and was asked to group each pair according to the most appropriate clinical section. This design aimed to make the biomedical enrichment more useful during summary generation by aligning biomedical enrichment with the section of the clinical case where it was most relevant.

Section enrichment prompt

```
You are a clinical terminology specialist. Classify each keyword below into
one or more sections of a structured clinical case.

SECTIONS AND WHAT BELONGS IN EACH:
- "patient characteristics": demographics, comorbidities, risk factors,
baseline health status, medical history
- "diagnosis": confirmed or suspected conditions, clinical findings,
lab/imaging results, diagnostic criteria
- "treatment/interventions": medications, procedures, interventions, dosages,
therapeutic strategies
- "outcomes and follow-ups": clinical response, adverse events, disease
progression, follow-up findings, survival or discharge status

CLASSIFICATION RULES:
- Assign each keyword to every section it is relevant to (a keyword may appear
in multiple sections).
- Preserve the full keyword string including any definition after a colon or
dash.
- Omit keywords that are not clinically relevant to any section.
- Do not invent or infer information not present in the keyword list.
- Do not alter information already present in the structured context.

KEYWORDS:
{kw_list}

STRUCTURED CONTEXT:
{structured}

Return ONLY a valid JSON object with exactly these four keys:
{{
  "patient characteristics": [],
  "diagnosis": [],
  "treatment/interventions": [],
  "outcomes and follow-ups": []
}}

Your response must begin with {{ and end with }}. Do not include any text,
explanation, or formatting outside the JSON object.
ONLY return the JSON object.
```

4.5. Agentic Summary Generation

After the structured sections are combined and enriched with section-specific keywords and definitions, the final summary generator produces a concise clinical case summary from this structured context. The summary generation prompt followed the same role-based prompting strategy used in the previous LLM-based stages. In this case, the prompt instructed the model to act as a clinical summarization expert and to generate a medically accurate, fluent, and concise single-paragraph summary based on the enriched structured context.

Summary generation prompt

```
You are a clinical summarization expert.

Using the STRUCTURED CONTEXT below, generate a concise, fluent clinical case
report summary.
Only provide the summary as output with no extra information.

STRUCTURED CONTEXT:
{structured}{kw_block}

Requirements:
- Include patient characteristics, diagnosis, treatment/interventions, and
outcomes and follow-ups
- Be medically accurate
- Write as a single paragraph
- Prioritize factual accuracy over keyword usage
- When a keyword has a definition, use that definition to guide interpretation
```

The same critic-review mechanism described in Section 4.1 was also applied to the final summary generation, but with the evaluation focused on the complete summary rather than on an individual extracted section. The critic prompt summary template is provided in Appendix A.2.

5. Evaluation

This section describes the evaluation setup used to assess the submitted systems. It first presents the datasets used for official and complementary evaluation, then summarizes the English and Portuguese run configurations, and finally analyzes the results across automatic metrics and qualitative evaluation dimensions.

5.1. Datasets

Two datasets were used to evaluate the submitted systems. The primary evaluation was performed on the official MultiClinSum-2 test dataset, which consists of 1,000 full clinical case reports per language. The test cases were derived from the same biomedical sources as the training data, namely the PMC-Patients subset of PubMed Central and manually curated clinical case reports from PubMed [5]. This dataset was used for the official competition submissions in both English and Portuguese.

In addition to the official test set, the English runs were also evaluated on the gold standard dataset from the previous MultiClinSum edition [1] [22]. This dataset was manually selected by clinical professionals and contains curated clinical case reports paired with reference summaries. It was used as an additional evaluation set to analyze whether the behavior observed in the official competition results was consistent on a smaller and more curated collection of clinical cases.

[1] https://temu.bsc.es/multiclinsum/

5.2. English Runs Configurations

- **Run 1:** Used QuickUMLS keyword extraction and the agent-critic loop. This was the simplest pipeline and served as the baseline against which the following runs were compared.
- **Run 2:** Extended the baseline approach by incorporating UMLS concept definitions associated with the extracted keywords, providing additional grounding to the LLM during generation.
- **Run 3:** Further extended the previous configuration through section-aware keyword assignment, where extracted keywords and their corresponding definitions were grouped according to their clinical context, specifically patient characteristics, diagnostics, interventions, and outcomes.

5.3. Portuguese Runs Configurations

- **Run 1:** Implemented a direct Portuguese summarization pipeline using Portuguese prompts, QuickUMLS keyword extraction based on the Portuguese UMLS index, and the agent-critic loop.
- **Run 2:** Implemented a translation-assisted workflow where Portuguese source documents were translated into English, summarized using the English pipeline with the English QuickUMLS index, English prompts, and the agent-critic loop, and subsequently translated back into Portuguese.
- **Run 3:** Followed a direct Portuguese summarization workflow combining Portuguese prompts, QuickUMLS extraction with the Portuguese index, the agent-critic loop, and UMLS concept definitions also from the portuguese UMLS index.

5.4. Results

Table 3 presents the official evaluation results for the submitted English and Portuguese runs. The systems were evaluated using lexical overlap metrics, namely ROUGE-1, ROUGE-2, and ROUGE-LSum, as well as semantic similarity through BERTScore. In addition, the outputs were assessed using qualitative dimensions related to Faithfulness, Completeness, Fluency, and Consistency.

Table 3
Evaluation results for the English and Portuguese submitted runs. Best results for each language are shown in bold.

Lang.	Run	R-1	R-2	R-LSum	BERTScore	Faith.	Compl.	Fluency	Cons.
en	Run 1	0.3534	0.1290	0.2266	**0.8616**	0.7583	0.8854	0.7000	0.6936
en	Run 2	**0.3538**	**0.1328**	**0.2271**	**0.8616**	**0.7603**	**0.8902**	0.7000	0.6928
en	Run 3	0.3469	0.1267	0.2221	0.8605	0.7553	0.8599	**0.7074**	**0.6964**
pt	Run 1	0.3578	0.1320	0.2274	0.8620	**0.7453**	**0.8421**	0.6996	**0.5849**
pt	Run 2	0.3364	0.1037	0.2121	0.8581	0.7269	0.7967	0.6992	0.5457
pt	Run 3	**0.3610**	**0.1352**	**0.2294**	**0.8630**	0.7443	0.8273	**0.6998**	0.5752

5.4.1. English Results

The English runs showed minimal variation across all metrics, indicating that the different QuickUMLS enrichment strategies produced relatively stable outputs. Run 2 achieved the best lexical overlap among the English submissions, with the highest ROUGE-1, ROUGE-2, and ROUGE-LSum scores, reaching 0.3538, 0.1328, and 0.2271 respectively. This suggests that adding UMLS concept definitions to the extracted QuickUMLS keywords provided a small improvement in overlap with the reference summaries when compared with the baseline keyword-only configuration.

BERTScore remained almost unchanged across the three English runs, with Runs 1 and 2 both achieving 0.8616 and Run 3 obtaining 0.8605. This indicates that all English configurations preserved a similar level of semantic similarity to the reference summaries. The addition of section-aware keyword assignment in Run 3 did not improve ROUGE or BERTScore, although it achieved the highest Fluency and Consistency scores among the English runs, with 0.7074 and 0.6964 respectively. This may indicate

that organizing keywords by clinical section helped produce slightly more coherent outputs, even if it did not improve lexical overlap or completeness.

In terms of qualitative evaluation, Run 2 obtained the highest Faithfulness and Completeness scores, with 0.7603 and 0.8902 respectively. This supports the hypothesis that adding UMLS definitions can provide useful semantic grounding for the model, helping it preserve clinically relevant information without substantially changing the overall generation behavior.

Run 1, which used QuickUMLS keyword extraction together with the critic agent, also compares favourably with the extractive QuickUMLS-based concept method reported by ExtraSum in the previous MultiClinSum working lab. Their concept-based English system used QuickUMLS to extract clinical concepts and rank sentences according to concept-set similarity, obtaining a BERTScore F1 of 0.8446 and a ROUGE F1 of 0.2313 [21]. In comparison, our Run 1 reached a BERTScore of 0.8616, indicating stronger semantic similarity to the reference summaries, while achieving a ROUGE-LSum score of 0.2266. This suggests that using QuickUMLS concepts as prompt-level guidance within an abstractive LLM pipeline, combined with critic-based refinement, can preserve semantic content more effectively than relying on concept-based extractive sentence selection alone, although the lexical overlap remains comparable.

5.4.2. Portuguese Results

The Portuguese submissions showed clearer differences between configurations. Run 3 achieved the best overall lexical and semantic performance, with the highest ROUGE-1, ROUGE-2, ROUGE-LSum, and BERTScore values among the Portuguese runs. This suggests that, in this evaluation setting, the direct Portuguese pipeline benefited from UMLS definition enrichment more than from the translation-assisted workflow.

Run 1, which used direct Portuguese generation with the Portuguese QuickUMLS index but without definitions, also had good performance but on different metrics. It achieved the highest Portuguese Faithfulness and Completeness scores, with 0.7453 and 0.8421 respectively. This indicates that direct Portuguese generation was effective at preserving clinically relevant information and remaining grounded in the source text.

Run 2, the translation-assisted pipeline, obtained the lowest scores across all Portuguese metrics. Although the workflow was designed to take advantage of the larger English UMLS index and English prompting, the results suggest that the added translation steps may have introduced information loss or inconsistencies. This is reflected in its lower ROUGE-LSum score of 0.2121, BERTScore of 0.8581, and Consistency score of 0.5457. These results indicate that, for this setup, the potential benefits of English-domain processing did not compensate for the risks introduced by translating the input and output.

5.4.3. Comparison with the Gold Standard Dataset

As stated in Section 5.1, the three English configurations were also evaluated on the gold standard dataset described in Section 5.1. This comparison was intended to assess whether the behavior of the submitted English runs remained consistent in a smaller and more curated evaluation setting.

Across all three runs, the results obtained on the manually selected gold standard dataset show higher ROUGE scores than the official competition results. This improvement is most visible in ROUGE-1 and ROUGE-LSum, suggesting that the generated summaries had stronger lexical overlap with the references in the curated dataset. In contrast, BERTScore F1 remained very similar across both evaluations, with only small decreases on the gold standard dataset. This indicates that the semantic similarity of the generated summaries remained stable, while the lexical alignment varied more noticeably depending on the dataset.

The higher ROUGE scores on the gold standard dataset may be partly explained by differences in dataset composition and reference-summary characteristics. ROUGE is sensitive to exact wording, summary length, and the level of detail included in the reference, so changes in these properties can

Table 4
Comparison between the official English results and the results obtained on the manually selected gold standard dataset. Best result for each metric is shown in bold.

Dataset	Run	R-1	R-2	R-LSum	BERTScore
Official test	Run 1	0.3534	0.1290	0.2266	**0.8616**
Official test	Run 2	0.3538	0.1328	0.2271	**0.8616**
Official test	Run 3	0.3469	0.1267	0.2221	0.8605
Gold standard	Run 1	0.3819	0.1326	0.2399	0.8594
Gold standard	Run 2	0.3858	0.1345	**0.2412**	0.8601
Gold standard	Run 3	**0.3864**	**0.1350**	0.2404	0.8584

affect lexical overlap even when semantic similarity remains stable. This is consistent with the relatively small variation in BERTScore across the two evaluations.

The behavior of the runs generated using the gold standard was also consistent with the official evaluation. Run 2 remained one of the strongest configurations, achieving the highest ROUGE-LSum and BERTScore F1 on the gold standard dataset. Run 3 obtained the highest ROUGE-1 and ROUGE-2 scores, suggesting that section-aware keyword and definition assignment may improve lexical overlap in a more curated evaluation setting. However, the differences between Runs 2 and 3 were small, indicating that the additional section-aware organization did not lead to a large performance gain. Overall, this comparison suggests that the proposed pipeline generalizes well to manually selected clinical cases, with UMLS definition enrichment remaining the most stable configuration across both evaluation settings.

5.4.4. Overall Analysis

Across both languages, the results show that UMLS-based enrichment had a generally positive but modest impact. In English, the addition of definitions improved ROUGE and qualitative completeness, while in Portuguese the combination of direct generation and definitions produced the best overall automatic scores. However, section-aware keyword grouping did not lead to a measurable improvement in lexical or semantic metrics for English, despite slightly improving fluency and consistency.

The critic-review mechanism was triggered in a relatively small proportion of cases. Considering both critic-review stages, namely section extraction and final summary generation, the critic requested a revision 45 times across the generation of 1,000 official test cases summaries. This indicates that most outputs passed through the critic-review stage without the need for revision, so its impact on the final results was likely limited to a small subset of cases.

Regarding multilingual summarization, direct Portuguese generation outperformed the translation-assisted strategy, suggesting that maintaining the original language throughout the pipeline may be preferable. The translation-based approach introduced additional processing complexity and produced lower performance, likely due to cascading errors during translation or mismatches between the translated clinical text and the final Portuguese output.

A brief qualitative inspection of the translation-assisted workflow helped contextualize this result. In one inspected case, two issues were observed. First, a Portuguese clinical term referring to excessive salivation was translated as a different clinical finding related to eyelid drooping. Second, the English output was substantially shorter than the Portuguese source text, suggesting that the model compressed parts of the case report despite being used for translation rather than summarization. Since Run 2 performs section extraction, keyword matching, and summary generation over the translated English text, both terminology shifts and unintended compression may have affected the information passed through the pipeline. However, this analysis was limited in scope and should be interpreted as illustrative rather than conclusive.

Overall, the best English configuration was Run 2, based on its stronger ROUGE, Faithfulness, and Completeness scores, while the best Portuguese configuration was Run 3, based on its highest ROUGE

and BERTScore values. These findings suggest that the enrichment through QuickUMLS and UMLS definitions can improve clinical summarization performance, particularly when applied directly in the target language.

6. Conclusion

In this work, we presented a multilingual clinical case summarization pipeline for MultiClinSum-2, combining structured clinical-domain extraction, iterative agent-critic refinement, and keyword enrichment using QuickUMLS and UMLS definitions. The results showed that QuickUMLS-based grounding provided a stable baseline across both English and Portuguese, while the addition of UMLS definitions led to the strongest English configuration and the best automatic scores in Portuguese. The Portuguese experiments also indicated that direct target-language generation was more effective than the translation-assisted workflow, suggesting that the additional translation steps introduced more noise than benefit in this setup.

Future work will focus on improving the critic-feedback mechanism and expanding it to the multilingual components of the system. In particular, critic prompts could be developed for translation validation. Further experiments should also assess the effect of larger or medically adapted open-weight models and improved Portuguese clinical terminology resources. Finally, a broader evaluation of section-aware keyword assignment and UMLS definition enrichment through an ablation study would help determine when keyword and definition grounding improves factual consistency, completeness, and summary quality.

Declaration on Generative AI

During the preparation of this work, the author(s) used Generative AI to check grammar, wording, and rephrase text for readability. After using these tools, the author(s) reviewed and edited the content and take full responsibility for the publication's content.

References

[1] G. Adams, E. Alsentzer, M. Ketenci, J. Zucker, N. Elhadad, What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4794–4811. doi:10.18653/v1/2021.naacl-main.382.

[2] D. V. Veen, C. V. Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerova, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, A. S. Chaudhari, Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization, Nature Medicine 30 (2024) 1134–1142. doi:10.1038/s41591-024-02855-5. arXiv:2309.07430.

[3] Y. Gao, D. Dligach, T. Miller, D. Xu, M. M. Churpek, M. Afshar, Summarizing Patients Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models, 2022. URL: http://arxiv.org/abs/2208.08408. doi:10.48550/arXiv.2208.08408. arXiv:2208.08408.

[4] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International

Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[5] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, A. Mash, M. Krallinger, Overview of multiclinsum-2 at clef 2026: Data, evaluation, and results for multilingual clinical case report summarization across 15 languages, in: CLEF, 2026.

[6] GitHub - Georgetown-IR-Lab/QuickUMLS: System for Medical Concept Extraction and Linking · GitHub, https://github.com/Georgetown-IR-Lab/QuickUMLS, 2016.

[7] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.

[8] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, K. Kochut, Text Summarization Techniques: A Brief Survey, 2017. doi:10.48550/arXiv.1707.02268. arXiv:1707.02268.

[9] S. Volkova, P. Bautista, A. Hiriyanna, G. Ganberg, I. Erickson, Z. Klinefelter, N. Abele, H.-T. Kao, G. Engberson, Cross-Disciplinary Knowledge Retrieval and Synthesis: A Compound AI Architecture for Scientific Discovery, 2025. doi:10.48550/arXiv.2511.18298. arXiv:2511.18298.

[10] J. Choi, N. Palumbo, P. Chalasani, M. M. Engelhard, S. Jha, A. Kumar, D. Page, MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance, 2024. doi:10.48550/arXiv.2408.01869. arXiv:2408.01869.

[11] G. Carlini, F. Durazzi, D. Remondini, R. Lodi, Designing LLM-powered agents to manage clinical text reports in medical institutions, in: 2025 International Joint Conference on Neural Networks (IJCNN), 2025, pp. 1–8. doi:10.1109/IJCNN64981.2025.11227666.

[12] Z. Chen, J. Chen, Y. G. Wang, Y. Shen, A Human-Centered AI Agent Framework with Large Language Models for Academic Research Tasks, in: X.-L. Mao, Z. Ren, M. Yang (Eds.), Natural Language Processing and Chinese Computing, Springer Nature, Singapore, 2026, pp. 363–374. doi:10.1007/978-981-95-3343-5_28.

[13] D. Yang, J. Zhu, H. Wu, M. Tan, C. Li, M. Yang, CascadeRCG: Retrieval-Augmented Generation for Enhancing Professionalism and Knowledgeability in Online Mental Health Support, in: Companion Proceedings of the ACM on Web Conference 2025, WWW '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1465–1469. doi:10.1145/3701716.3715466.

[14] Ollama, https://ollama.com, 2023.

[15] Llama3:8b, https://ollama.com/library/llama3:8b, 2024.

[16] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, 2023. doi:10.48550/arXiv.2310.06825. arXiv:2310.06825.

[17] J. J. Gagnier, G. Kienle, D. G. Altman, D. Moher, H. Sox, D. Riley, The CARE Guidelines: Consensus-based Clinical Case Reporting Guideline Development, Global Advances in Health and Medicine 2 (2013) 38–43. doi:10.7453/gahmj.2013.008.

[18] spaCy · Industrial-strength Natural Language Processing in Python, https://spacy.io/, 2015.

[19] Medicalai/ClinicalBERT · Hugging Face, https://huggingface.co/medicalai/ClinicalBERT, 2023.

[20] L. Pamio, G. M. Di Nunzio, Comparing CRF vs BERT Models for Named Entity Recognition and Relation Extraction, in: CLEF 2025 Working Notes, CEUR-WS, Madrid, Spain, 2025. Presented at CLEF 2025, 9–12 September 2025.

[21] S. Rhazzafe, S. Colreavy-Donnelly, N. S. Nikolov, Extrasum @ multiclinsum: Extractive summarization of english, spanish, french and portuguese clinical case reports, in: CLEF 2025 Working Notes, CEUR-WS, Madrid, Spain, 2025. Presented at CLEF 2025, 9–12 September 2025.

[22] M. Rodríguez Ortega, E. Rodríguez López, S. Lima López, M. Krallinger, MultiClinSum Dataset: Summarization of Clinical Case Reports in English, Spanish, French and Portuguese, 2025. URL: https://doi.org/10.5281/zenodo.15517617. doi:10.5281/zenodo.15517617.

A. Critic Prompt Templates

A.1. Section Critic Prompt

```
You are a clinical extraction reviewer. Your job is to catch meaningful errors,
not to rewrite or improve extractions.

ONLY flag an issue if ALL three conditions are met:
1. The problem is directly verifiable against the SOURCE CLINICAL CASE REPORT
below.
2. The problem meaningfully changes clinical interpretation (not just style or
phrasing).
3. A clinician would consider it a significant error.

Do NOT flag:
- Paraphrasing or synonyms that preserve clinical meaning
- Omission of details not clinically significant to "{section_name}"
- Formatting or ordering differences
- Information not present in the source text

SECTION BEING EVALUATED: {section_name}

EXTRACTION:
{output}

SOURCE CLINICAL CASE REPORT:
{full_text}

If the extraction is clinically accurate and complete enough for its purpose,
return:
{{"has_issues": false, "issues": []}}

Only if you find a clear, source-verifiable, clinically significant problem,
return:
{{
  "has_issues": true,
  "issues": [
    "<specific issue with direct reference to what the source text says vs.
    what was extracted>"
  ]
}}

Do NOT rewrite the extraction.
Do NOT summarize the case.
Only report issues.
Your response must begin with {{ and end with }}. Do not include any text,
explanation, or formatting outside the JSON object.
```

A.2. Summary Critic Prompt

```
You are a clinical summarization reviewer. Your job is to catch meaningful
errors, not to rewrite or improve summaries.

ONLY flag an issue if ALL three conditions are met:
1. The problem is directly verifiable against the REVIEWED STRUCTURED CONTEXT
below.
2. The problem meaningfully changes clinical interpretation (not just style or
phrasing).
3. A reasonable clinician would consider it a significant error.
```

```
Do NOT flag:
- Paraphrasing or synonyms that preserve clinical meaning
- Omission of minor details not critical to understanding the case
- Stylistic or fluency differences that don't affect clinical accuracy
- Missing information not present in the structured context

FINAL SUMMARY:
{output}

REVIEWED STRUCTURED CONTEXT:
{structured}

If the summary is clinically accurate and sufficiently complete, return:
{{"has_issues": false, "issues": []}}

Only if you find a clear, context-verifiable, clinically significant problem,
return:
{{
  "has_issues": true,
  "issues": [
    "<specific issue with direct reference to what the structured context says
    vs. what the summary states>"
  ]
}}

Do NOT rewrite the summary.
Do NOT summarize the case.
Only report issues.
Your response must begin with {{ and end with }}. Do not include any text,
explanation, or formatting outside the JSON object.
```