Overview of MultiClinSum-2 at CLEF 2026: Data, Evaluation, and Results for Multilingual Clinical Case Report Summarization Across 15 Languages

Miguel Rodríguez-Ortega[1,*], Eduard Rodríguez-López[1], Salvador Lima-López[1], Audrey Mash[1] and Martin Krallinger[1]

[1]Barcelona Supercomputing Center, Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain

Abstract

Rapid digitalization of healthcare systems has led to an unprecedented accumulation of clinical documentation across diverse sources, including clinical case reports, discharge summaries, laboratory results, and diagnostic documents, generating large volumes of both structured and unstructured data that present significant challenges for efficient processing. Among these, clinical case reports are of particular interest, as they document individual patient histories in rich detail, describing patient demographics, symptoms, clinical findings, procedures, and follow-up information. Despite their clinical value, the dense and lengthy nature of these documents makes efficient information extraction and interpretation a demanding task for healthcare professionals. Recent advances in large language models (LLMs) and generative AI have demonstrated considerable potential for addressing these challenges, enabling automated summarization of clinical documents across multiple languages and supporting clinical research involving unstructured health data. However, the lack of robust multilingual benchmarking frameworks limits the systematic evaluation and comparison of such approaches. To address this gap, we present MultiClinSum-2, the second edition of our shared task on automatic summarization of clinical case reports, expanded from four to 15 languages: English, French, Spanish, Portuguese, Italian, Russian, Catalan, Norwegian, Danish, Romanian, German, Greek, Dutch, Czech, and Swedish. Built on a comprehensive multilingual corpus of full clinical cases paired with reference summaries derived from biomedical literature, the task introduces an enhanced evaluation framework combining classic automatic summarization metrics with LLM-as-judge strategies, enabling a more comprehensive and clinically-oriented assessment of system outputs. In addition, two complementary baseline systems were developed to provide reference points for participating teams. The MultiClinSum-2 resources are available at: *https://zenodo.org/records/19824154*

Keywords

clinical NLP, text summarization, clinical case reports, biomedical adaptation, multilingual, Gen AI, generative pre-trained transformers, LLM-as-a-Judge

1. Introduction

Biomedical and healthcare information constitutes one of the richest and most complex sources of content in scientific research, characterized by highly specialized semantics, domain-specific terminologies, and an unprecedented diversity of document types and formats. This clinical documentation is spread across a wide variety of sources, including scientific literature, discharge summaries, radiology reports, and progress notes, and its volume has grown exponentially over the past decades. This growth has been associated with increased burnout, diminished productivity, and reduced ability to access relevant knowledge in a timely manner [1, 2]. Among these sources, clinical case reports play a particularly central role, containing detailed information about patients, including demographic data, clinical findings, diagnoses, follow-ups, and procedures. The processing of such content is often extensive and complex, contributing to significant information overload among healthcare professionals [2] and presenting considerable time constraints for biomedical researchers who may rely on rule-based approaches to extract and analyze relevant information from these sources.

CLEF 2026 Working Notes, 21 – 24 September 2025, Jena, Germany

*Corresponding author.

miguel.rod.bsc@gmail.com (M. Rodríguez-Ortega); mkrallin@bsc.es (M. Krallinger)

0009-0000-0188-079X (M. Rodríguez-Ortega); 0000-0002-7384-1877 (S. Lima-López); 0000-0002-2646-8782 (M. Krallinger)

In recent years, applications based on large language models (LLMs), generative artificial intelligence (GenAI), and natural language processing (NLP) have emerged as powerful tools for automating clinical data processing. These systems are capable of understanding clinical terminology, vocabulary, and domain-specific semantics, while producing coherent and contextually relevant outputs. As a result, they can considerably reduce healthcare workload and assist professionals in daily tasks such as patient diagnosis support, clinical decision-making, and the conversion of free text into structured representations through the extraction of relevant medical concepts (e.g., diseases, symptoms, findings). Among these NLP-based applications, text summarization has gained particular relevance, driven by the growing volume of clinical information accumulating in EHRs [3, 4]. Clinical text summarization can be described as the process of collecting and synthesizing patient information [5], while preserving key clinical elements such as patient demographics, clinical findings, outcomes, and procedures. Automated LLM-based tools have been shown to outperform clinical professionals in this task, with automatically generated summaries reported to be more accurate and effective than human-written ones [6]. However, robust and faithful evaluation frameworks remain essential, given the intricate and complex nature of medical vocabularies, as well as persistent limitations of LLM-based generation such as hallucinations, omissions, and factual inconsistency [7]. Traditional automatic summarization metrics, while widely used, are often insufficient to capture the full nuance and clinical context required for reliable summary assessment. LLM-as-judge approaches can help fill this gap, offering a complementary evaluation layer to existing automatic metrics.

In the NLP community, several shared tasks have advanced biomedical and clinical text summarization in recent years. The BioLaySumm shared task, held at the BioNLP Workshop in 2023 and 2024, focused on the lay summarization of biomedical research articles, attracting growing participation and a marked shift towards LLM-based approaches in its second edition [8, 9]. The ProbSum 2023 shared task addressed the summarization of patients active diagnoses and problems from EHR progress notes, targeting a clinically critical and underexplored summarization scenario [10]. Other initiatives have extended summarization research to specific clinical domains and document types, such as RadSum23, which focused on multi-modal radiology report summarization [11], and MSLR2022, which addressed multi-document summarization of medical literature reviews [12]. Despite this growing body of work, multilingual clinical case report summarization remains comparatively underexplored, with most existing efforts focused primarily on English-language resources.

To address this gap, we present MultiClinSum-2, the second edition of our shared task on multilingual clinical text summarization, organized as part of the BioASQ Lab at CLEF 2026 [13]. Building on the first edition, MultiClinSum [14], this second edition continues to advance automatic text summarization of clinical case reports, expanding language coverage from four to 15 languages. In addition, we introduce a more refined evaluation methodology in which automatic metrics are complemented by LLM-as-judge strategies, with the aim of capturing deeper clinical dimensions of the summaries generated by participating systems.

2. Task Overview

MultiClinSum-2 is the second edition of the shared task on multilingual text summarization of clinical case reports, organized by the NLP for Biomedical Information Analysis group (NLP4BIA) at the Barcelona Supercomputing Center, and promoted by the European projects DataTools4Heart and AI4HF, in the context of the BioASQ workshop at CLEF 2026. The task challenged participants to develop systems capable of automatically generating concise and clinically faithful summaries of clinical case reports. Given the critical role of case reports in disseminating medical knowledge, such as describing patient presentations, diagnostic processes, and clinical findings, the ability to summarize them accurately and efficiently has direct implications for clinical research and healthcare practice.

Building on the first edition, MultiClinSum [14], MultiClinSum-2 substantially expands the language coverage to 15 languages: English, Spanish, French, Portuguese, Italian, Russian, Catalan, Norwegian, Danish, Romanian, German, Greek, Dutch, Czech, and Swedish. This expansion reflects the need to develop and evaluate summarization approaches that are robust across diverse linguistic contexts, including both high-resource and lower-resource languages in the biomedical domain. Participants were provided with two dataset splits, a train set to develop their systems containing around 26,000 full-text clinical case reports paired with reference summaries, and a test set of 1,000 full-cases to evaluate their systems by generating summaries from them.

The evaluation framework presents another notable novelty for this second edition. The framework combines two complementary assessment strategies. Alongside established automatic metrics also implemented in the first edition (ROUGE-based for lexical overlap and BERTScore for semantic similarity), MultiClinSum-2 introduces a novel LLM-as-judge system, specifically designed to evaluate information retention and loss in generated summaries. This additional evaluation framework assesses dimensions such as clinical coherence, hallucination, completeness, and preservation of key diagnostic and therapeutic information, going beyond surface-level string matching to capture the clinical fidelity of generated summaries.

Task sub-tracks

To promote the development of reliable clinical summarization models across diverse linguistic contexts, MultiClinSum-2 is organized into 15 independent language sub-tracks, one for each of the supported languages: *MultiClinSum2-en* (English), *MultiClinSum2-es* (Spanish), *MultiClinSum2-fr* (French), *MultiClinSum2-pt* (Portuguese), *MultiClinSum2-it* (Italian), *MultiClinSum2-ru* (Russian), *MultiClinSum2-ca* (Catalan), *MultiClinSum2-nb* (Norwegian), *MultiClinSum2-da* (Danish), *MultiClinSum2-ro* (Romanian), *MultiClinSum2-de* (German), *MultiClinSum2-el* (Greek), *MultiClinSum2-nl* (Dutch), *MultiClinSum2-cs* (Czech), and *MultiClinSum2-sv* (Swedish). Each sub-track focused on the automatic summarization of clinical case reports in the given language and is evaluated independently, allowing for a precise assessment of system performance within each linguistic context.

The task adopted a flexible participation mode, allowing teams to choose how many sub-tracks they want to participate in, with no requirement to submit results for all 15 languages. Both monolingual and multilingual modeling strategies are explicitly encouraged, and participants were free to leverage any available external resources, including domain-adapted or general pre-trained language models, biomedical terminologies, or supplementary clinical corpora, being as requirement that all utilized resources were transparently reported. Each team was allowed to submit up to five runs per language sub-track, enabling participants to explore and compare different configurations of their systems within the same evaluation framework.

This sub-track structure reflects two distinct data configurations that participants should be aware of. Four languages—English, Spanish, French, and Portuguese—include a native gold standard component, consisting of clinical case reports originally authored and summarized in that language. The remaining eleven languages rely on high-quality machine-translated versions of the English PMC-Patients subset [15], produced using the SalamandraTA-7B-instruct model; further details are provided in Section 3. Although both configurations were evaluated using the same metrics and no differentiation between sources was provided in the released datasets, this distinction is relevant for participants when designing language-specific or cross-lingual approaches, since translated data may exhibit different linguistic characteristics compared to natively authored clinical texts. Further details about the shared task, including submission guidelines and dataset access, are available on the task website: *https://temu.bsc.es/multiclinsum2/*.

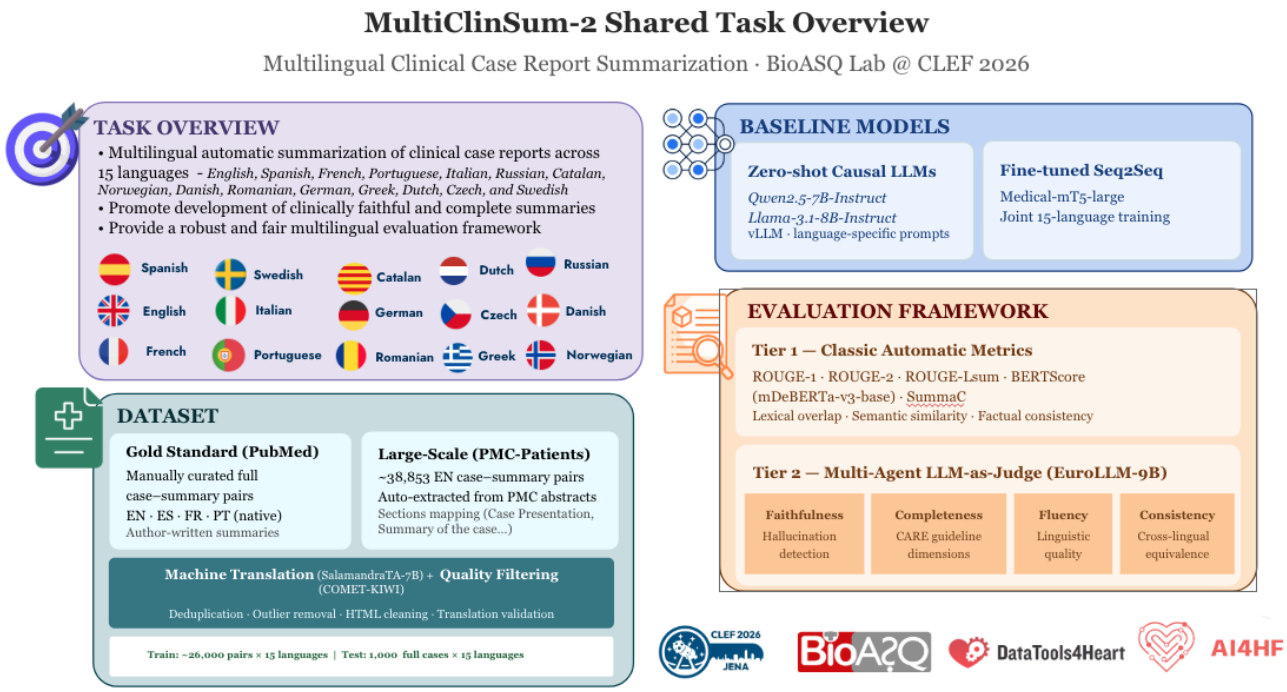

Figure 1: Overview of the MultiClinSum-2 shared task, covering the task objective, dataset construction, baseline models, and evaluation framework.

3. Resources and Datasets of MultiClinSum-2

The MultiClinSum-2 dataset was constructed from two complementary data sources. The first consisted of a curated collection of manually selected clinical case report and author-written summary pairs in English, Spanish, French, and Portuguese, derived from PubMed by applying the publication type filter *"Case Report"*. The second is a large-scale collection of automatically extracted full case-summary pairs in English, constructed from the PMC-Patients subset [15], a dataset of patient summaries and relations extracted from PubMed Central full-text articles, and used for benchmarking retrieval-based clinical decision support systems [16]. For this collection, PMC-Patients patient summaries were used as full cases, while the paired summaries were obtained by identifying and extracting specific sections of the corresponding PubMed abstracts that briefly describe the patient case, such as *"Case Presentation"*, *"Summary of the Case"*, or *"Case Report"* sections.

In terms of data processing and filtering, both collections followed the same procedure established in the first edition of the task [14], where each full case-summary pair underwent a careful set of pre-processing and quality control steps: removal of HTML artifacts, deduplication based on PMIDs and exact string matching, filtering of veterinary cases and case series, and outlier removal using a 3×IQR threshold applied to summary length distributions, compression ratio, and pairs where the summary or full case was excessively long or short. As in the first edition, full case-summary pairs for the remaining task languages were generated through machine translation from their native languages—English, Spanish, French, and Portuguese—using the `SalamandraTA-7B-instruct` model [17], a fine-tuned variant of the Salamandra-7B model, to produce high-quality translations across all 15 supported languages. With respect to translation quality, additional filtering was applied using automatic language identification via the *langdetect* Python package[1], followed by the removal of pairs containing translation errors such as truncated texts, repeated sentences or words, and other translation artifacts that could affect dataset quality. Further details on translation quality as assessed

[1] https://pypi.org/project/langdetect/

via COMET-KIWI scores are provided in Section 3.

The combination of both filtered collections resulted in a training set ranging from 25,749 pairs (Greek) to 26,943 pairs (English) per language across all 15 sub-tracks, reflecting minor variation due to the filtering and validation steps described above. The test set comprises 1,000 full case-summary pairs per language sub-track, derived exclusively from data pairs not released in the first edition, and provided without reference summaries for participant evaluation. Detailed corpus statistics per language sub-track are presented in Table 1. Additionally, Figure 2 shows the word count distribution for both full cases and summaries per language. It can be observed that Romance languages such as Catalan, French, Italian, and Spanish tend to have higher word counts in both full cases and summaries compared to other language groups. This pattern may reflect the naturally more verbose nature of Romance languages, or alternatively, the fact that the machine translation process performed from English into these languages tends to produce longer target texts.

Table 1
Descriptive statistics per language of MultiClinsum-2 dataset.

		Mean Word Count		Mean Sentence Count		
Language	**Nr. Pairs**	**Full Case**	**Summary**	**Full Case**	**Summary**	**Comp. Ratio**
Catalan (ca)	27.699	604.80	118.56	27.49	5.59	0.23 ± 0.13
Czech (cs)	27.488	469.89	86.37	30.78	5.73	0.21 ± 0.13
Danish (da)	27.652	485.14	93.76	28.75	5.70	0.22 ± 0.17
German (de)	27.424	510.20	96.89	30.47	5.82	0.22 ± 0.17
Greek (el)	26.749	547.05	100.84	29.53	5.62	0.22 ± 0.22
English (en)	27.943	528.72	98.75	28.41	5.59	0.22 ± 0.13
Spanish (es)	27.743	591.95	116.16	27.57	5.59	0.23 ± 0.13
French (fr)	27.455	615.43	119.87	27.69	5.58	0.23 ± 0.14
Italian (it)	27.553	572.88	113.19	27.46	5.58	0.23 ± 0.13
Norwegian (nb)	27.609	485.48	95.27	28.28	5.65	0.23 ± 0.61
Dutch (nl)	27.587	511.66	100.19	27.11	5.58	0.23 ± 0.13
Portuguese (pt)	27.593	551.87	108.86	27.22	5.58	0.23 ± 0.13
Romanian (ro)	27.579	551.83	110.71	26.94	5.58	0.23 ± 0.14
Russian (ru)	27.086	481.95	88.55	29.66	5.58	0.21 ± 0.16
Swedish (sv)	27.594	459.82	88.72	27.53	5.59	0.22 ± 0.13

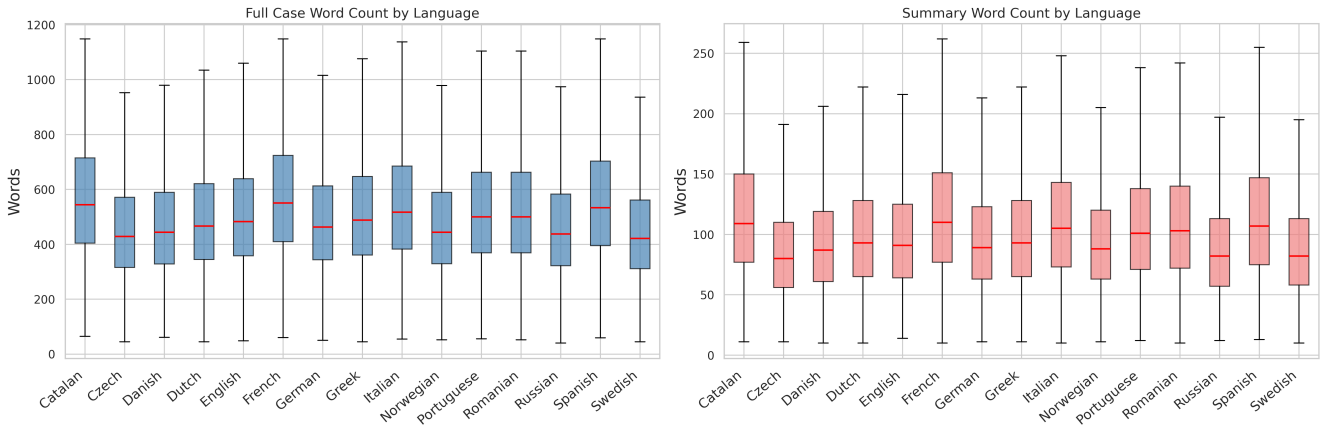

Figure 2: Word count distribution per language for full clinical cases (left) and their corresponding reference summaries (right). The red line indicates the median value. French shows the highest median word count across both full cases and summaries, while Czech and Swedish tend to indicate the lowest.

Machine Translation Process

The non-native language versions of the corpus were produced from the native English, Spanish, French, and Portuguese cases through machine translation. Translation was carried out with SalamandraTA-7B-instruct [17], an instruction-tuned multilingual translation model derived from the Salamandra LLM family. Each full case-summary pair was translated line by line: every non-empty line was wrapped in a short instruction prompt of the form *"Translate the following text from <source> into <target>"* and decoded independently, preserving the original line structure so that source and translation remain aligned at the segment level. Decoding used beam search with five beams and greedy selection (no sampling). To suppress the degenerate repetition loops that long clinical narratives can trigger, we applied a repetition penalty of 1.15 and a no-repeat n-gram constraint of order 4, and capped the number of generated tokens at twice the source length (and at the model's context limit). Outputs that did not terminate with an end-of-sequence token, that showed internal repetition, or whose length deviated strongly from the source were logged for inspection.

Because no human reference translations exist for these language pairs, we assessed translation quality with a two-stage automatic protocol. First, a *length-ratio screening* compared the character length of each translated segment against its source, using language-pair-specific expansion/compression bounds (e.g. a wider tolerance for German, which tends to expand, than for closely related Romance pairs). Segments outside these bounds, empty lines, and source-target line-count mismatches were flagged as candidate translation errors. Second, we estimated semantic adequacy with COMET-KIWI [18], a reference-free quality-estimation metric that scores each source-translation pair directly, which is well suited to our setting where gold translations are unavailable. Given the size of the large-scale PMC-Patients collection, COMET-KIWI was computed on a fixed random sample (500 files per language pair; seed 42) rather than exhaustively, whereas the smaller native gold-standard set was scored in full; the PMC-Patients figures are therefore sample-based quality estimates.

Across the machine-translated corpus, per-language-pair mean COMET-KIWI scores fell in the range of approximately 0.68 to 0.81 (overall mean ≈ 0.76), with the majority of segments scoring above the 0.70 adequacy threshold. Quarlity was consistent across the corpus components: the gold-standard native cases (translated out of English, Spanish, French, and Portuguese) scored between 0.68 and 0.81 and the English PMC-Patients pairs between 0.73 and 0.78. Translation quality was highest for typologically close pairs (e.g. Romance source languages into Italian) and lowest for the most distant target languages, with Greek and Russian consistently receiving the lowest scores. Files containing length-ratio anomalies or low COMET-KIWI scores were re-translated and re-scored; the majority of flagged cases were resolved or improved by retranslation, and the remainder were retained only where retranslation did not degrade quality. The resulting filtered set constitutes the machine-translated portion of the MultiClinSum-2 corpus.

4. Participation overview

Overall, this second edition has received a satisfactory level of participation, with a broader engagement reflecting the expanded multilingual coverage of this new edition. A total of 9 teams submitted at least one run of their predictions, resulting in 75 runs across all sub-tracks, with each team allowed to submit up to 5 runs per sub-track. As expected, the highest participation was observed in the English language, with 7 teams submitting 20 runs, followed by Spanish (5 teams, 14 runs) and Portuguese (5 teams, 9 runs). Catalan and Dutch also attracted considerable attention, with 4 and 3 teams respectively. Norwegian was the only language that received no submissions, although full coverage was ensured through the baseline systems. Table 2 presents the participation received per language sub-track.

Table 2
Participation overview of MultiClinSum-2 task per team and language subtrack. Numbers indicate runs submitted; — indicates no participation. Language codes correspond to: EN = English, ES = Spanish, FR = French, PT = Portuguese, IT = Italian, RU = Russian, CA = Catalan, NO = Norwegian, DA = Danish, RO = Romanian, DE = German, EL = Greek, NL = Dutch, CS = Czech, SV = Swedish.

Team	EN	ES	FR	PT	IT	RU	CA	NO	DA	RO	DE	EL	NL	CS	SV	Total
ixa-sum	5	5	2	2	2	—	5	—	—	2	2	—	2	—	—	**27**
MediScribes	5	—	—	—	—	—	—	—	—	—	—	—	5	—	—	**10**
NLP4Health	3	3	3	2	—	—	—	—	—	—	—	—	—	—	—	**11**
DACHausa	1	1	1	1	1	1	1	—	1	1	1	1	1	1	1	**14**
InfoLab-FEUP	3	—	—	3	—	—	—	—	—	—	—	—	—	—	—	**6**
PMH-Aqeel	1	1	—	1	1	1	1	—	—	—	—	—	—	—	1	**7**
UMUTeam	—	4	—	—	—	—	—	—	—	—	—	—	—	—	—	**4**
Aleyna	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	**2**
RaRaMi	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	**1**
Total runs	**20**	**14**	**6**	**9**	**4**	**2**	**7**	**0**	**1**	**3**	**3**	**1**	**7**	**1**	**2**	**75**

5. Baseline Systems

5.1. Overview and Motivation

With the aim of providing participants with performance benchmarks for evaluating submitted systems, two complementary baseline systems were developed for MultiClinSum-2. These models were designed to explore different strategies in clinical text summarization, enabling meaningful comparisons across methodological approaches while analyzing potential trade-offs between model complexity, domain adaptation, and computational requirements. They serve multiple purposes within the shared task framework: first, they establish performance thresholds that participating teams should aim to exceed, helping to measure the level of the task difficulty across the different language sub-tracks; and second, they offer a controlled comparison between generalist models and domain-adapted alternatives, highlighting the value of medical pre-training for summarization tasks.

The two systems consisted of, first a zero-shot inference with causal LLMs, where two open-source LLMs below the 8 billion parameters, were evaluated in a zero-shot summarization scenario using language-specific prompts. The study behind this approach was to measure the capabilities of modern GenAI LLMs when applied to clinical summarization without any specification about the training task, only relying on the model's general knowledge and instruction-following capabilities acquired during pre-training. While, on the other hand, a seq2seq encoder-decoder transformer model was fine-tuned to summarization using the training data. The selected model was a Medical-mT5-large [19] of 738M parameters pre-trained on biomedical and clinical corpora in English, Spanish, French, and Italian. This system aligns more to a domain-adapted approach that combines the benefits of medical pre-training with task-specific fine-tuning, offering a much smaller size model yet potentially more clinically-aware alternative to the general domain models. More details about each baselines are detailed in the sections below.

The two systems consisted of, first, a zero-shot inference approach with causal LLMs, where two open-source instruction-tuned models with fewer than 8 billion parameters were evaluated in a zero-shot summarization scenario. The motivation behind this approach was to measure the capabilities of modern generative LLMs when applied to clinical summarization without any task-specific training, relying exclusively on the models' general knowledge and instruction-following capabilities. The second system consisted of a seq2seq encoder-decoder transformer model fine-tuned for summarization using the MultiClinSum-2 training data. For this case, the selected model was Medical-mT5-large [20], a model of 738M parameters pre-trained on biomedical and clinical corpora in English, Spanish,

French, and Italian. This system was more aligned with a domain-adapted approach that combines the benefits of medical pre-training with task-specific fine-tuning, offering a considerably more compact yet potentially more clinically-aware alternative to the general-domain causal LLMs. Further details about each baseline system are provided in the sections below.

5.2. Zero-shot Inference with Causal LLMs

Two state-of-the-art open-source instruction-tuned causal language models were evaluated: Qwen2.5-7B-Instruct [21] (7 billion parameters) and Llama-3.1-8B-Instruct [22] (8 billion parameters). Both models are decoder-only transformers pre-trained on large multilingual corpora and subsequently trained for instruction-following through supervised fine-tuning and reinforcement learning from human feedback (RLHF). Neither model was fine-tuned on the clinical full case-summary pairs, only a purely zero-shot setting was applied, relying entirely on their general instruction-following capabilities. Inference was performed using vLLM [23], an optimized engine that uses *PagedAttention* mechanism to maximize GPU memory efficiency and throughput. Generation parameters were kept identical for both models, a maximum of 1,024 new tokens and temperature $T = 0.3$.

The prompting strategy consisted of a system turn and a user turn, applied via the `apply_chat_template` method of the HuggingFace tokenizer using each model native chat template. The system prompt instructed the model, in English, to act as a clinical summarization assistant, generating a concise and faithful summary, write exclusively in the same language as the input, and returning only the generated summary without any additional explanation. The user turn contains the instruction and the full clinical case text, both written in the target language of each sample, for example, the Catalan prompt reads *"Si us plau, resumiu el següent cas clínic:"* followed by the full case text, while the Russian prompt uses the Cyrillic equivalent. Native-language instructions were defined for all 15 task languages, without few-shot examples, chain-of-thought instructions, or task-specific demonstrations provided.

5.3. Fine-tuning of Medical-mT5-large: The MCS-mT5 Baseline

For this second baseline, Medical-mT5-large [20], a multilingual encoder-decoder transformer of 738M parameters built on top of mT5-large [24] and further pre-trained on biomedical and clinical corpora in English, Spanish, French, and Italian. The model was fine-tuned on the MultiClinSum-2 training data across all 15 languages, producing the *MCS-mT5* baseline. from which two checkpoints were evaluated: *MCS-mT5-1epoch* and *MCS-mT5-2epoch*, corresponding to one and two training epochs respectively. The decision of choosing an encoder-decoder architecture was motivated for abstractive summarization aspect: the encoder compresses the full clinical case into a contextual representation, while the decoder generates the summary auto-regressively, a more suitable approach than causal decoding for length-controlled, faithful text generation. The medical pre-training provides the model with domain-specific vocabulary and clinical reasoning patterns that general models might show limited capabilities, while keeping it substantially more compact than the 7 to 8Bparameter causal baselines. Despite the medical pre-training covered only four languages, the hypothesis supporting the fine-tuning strategy on the full 15-language MultiClinSum-2 pairs would enable the model to transfer cross-lingual summarization capabilities to other languages, leveraging the shared multilingual representations inherited from mT5. It should be noted that source sequences were truncated to 512 tokens, which represents a known limitation given the average length of clinical case reports in the dataset; this constraint was imposed by GPU memory requirements during fine-tuning. More details about the fine-tuning procedure are listed below.

- *Multilingual joint training*: Rather than training a unique model per language, we fine-tuned a single model joining all full case-summary pairs of all the 15 languages. Training and validation

samples from every language were pooled into a single dataset and fed to the model without any explicit language identifier in the input. Each full case source text was initiated with the T5-style task prefix *"summarize: "*, following the convention established during the original T5 pre-training, after which the full clinical case text follows. This approach encourages the model to learn a language-agnostic summarization strategy while relying on its multilingual representations to produce output in the correct target language.

- *Hyperparameters*: Model training was set for 2 epochs with a learning rate of 5×10^{-5} and a linear warmup over the first 500 steps, followed by linear decay. We used a per-device batch size of 4 with gradient accumulation over 2 steps, yielding an effective batch size of 8 per GPU, with a weight decay of 0.01. The best checkpoint was selected based on ROUGE-2 on the validation set, evaluated at the end of each epoch. For summary generation, both during validation and at inference, we used a maximum output length of 256 tokens, a length penalty of 2.0, and a no-repeat n-gram constraint of order 3 to reduce repetition. Source sequences were truncated to 512 tokens and target sequences to 256 tokens.
- *Training data and splits*: The complete MultiClinSum-2 training set, covering all 15 languages, was combined and randomly split into a 90 per cent training split and a 10 per cent validation subsets using a fixed random seed (seed=42) to ensure reproducibility. The held-out test set was used exclusively for final evaluation and was never used during training.

6. Evaluation Framework

The evaluation of automatically-generated summaries, obtained by both participants and baselines systems, requires a comprehensive, fair, and clinically meaningful assessment framework. For that purpose, a two-tier evaluation approach was designed, combining established automatic metrics with a novel LLM-as-judge strategy. The two evaluation components were deliberately kept independent, with each metric reported as a separate score. This design allows the participants to inspect the trade-offs between lexical overlap, semantic fidelity, and higher-order clinical quality dimensions. Figure 3 illustrates the different evaluation levels with their corresponding metrics.

6.1. Classic Automatic Metrics

The first evaluation level relies on two well-established reference-based metrics that operate by comparing the generated summary against a gold standard reference.

ROUGE-N. ROUGE-1, ROUGE-2 [25] were employed to measure the overlap of unigrams and bigrams between the generated and reference summaries, providing insight into how well the generated output captures key phrases and clinical terminology present in the reference. The `rouge_score` library was used for computation, with stemming disabled to preserve full compatibility across all 15 task languages.

ROUGE-Lsum. ROUGE-Lsum was computed in place of the standard sentence-level ROUGE-L, as the former splits both the generated and reference summaries on sentence boundaries before computing the longest common subsequence, making it more appropriate for multi-sentence clinical summaries where sentence structure carries meaningful information. Tokenisation follows the library's default whitespace-and-punctuation scheme, which provides a consistent and language-agnostic baseline across all 15 supported scripts, including non-Latin writing systems such as Cyrillic and Greek.

BERTScore. To capture semantic similarity beyond surface n-gram overlap, BERTScore [26] F1 was computed using `mdeberta-v3-base` [27] as the backbone model. This model was chosen for its strong multilingual coverage, which encompasses all 15 task languages within a shared token embedding space. Input sequences are truncated to 512 tokens before embedding. Baseline rescaling is disabled because it requires pre-computed language-specific baselines that are not available for all task languages.

6.2. Multi-Agent LLM-as-Judge Evaluation

Reference-based metrics are known to correlate imperfectly with human judgements of clinical summary quality, particularly for faithfulness and completeness, where a fluent hallucinated summary can achieve high lexical overlap scores. To address this limitation, an LLM-as-judge evaluation system based on EuroLLM-9B [28] model was developed, in which a judge LLM independently evaluates each generated summary across four clinically relevant dimensions.

The system consisted of four different agents, sharing the same underlying judge model: *faithfulness* agent, *completeness* agent, *fluency* agent, and *consistency* agent. The model was loaded once at the start of evaluation via vLLM and reused across all samples and languages, so that only a single set of weights resides in GPU memory throughout the run. Model inference was set as fully deterministic ($T = 0$, greedy decoding) to ensure that evaluation scores are reproducible across independent runs. Each agent could generate up to 1,024 new tokens per response, a brief reasoning paragraph, and a list of quality flags. If the model output cannot be parsed as valid JSON format, the system retries up to two additional iterations before falling back to a score of zero and a `llm_output_malformed` flag, so that a siensuring that malformed responses are never silently propagated without annotation. Also, prompts exceeding the model's context length are truncated from the left, preserving the JSON instruction at the end of the prompt that directs the output format.

Faithfulness Agent The faithfulness judge assesses whether every claim in the generated summary can be traced back to the original clinical case report, with no hallucinated clinical details. Only the original full case and the generated summary were provided to this agent, the reference summary was omitted to prevent the agent from rewarding reference similarity rather than factual accuracy. Faithfulness was rated on an integer scale from 0 (completely hallucinated) to 10 (every statement directly supported by the source), and the agent is required to list any specific hallucinated claims as structured flags.

Completeness Agent The completeness agent evaluates how thoroughly the generated summary covers the key clinical information present in the original full case, using the CARE (CAse REport) guidelines detailed by [29] as the evaluation criteria. Six clinically detailed dimensions were scored independently, each on a 0–10 scale:

1. *Patient demographics* (age, gender, ethnicity, occupation)
2. *Clinical presentation* (main symptoms, medical history, comorbidities)
3. *Diagnosis* (diagnostic methods, reasoning, differential diagnosis)
4. *Intervention* (treatment type, dosage, duration, and rationale for changes)
5. *Outcome* (clinical course, follow-up results, adverse events)
6. *Follow-up* (adherence, tolerability, and clinician- and patient-assessed outcomes).

The six dimension scores were averaged to produce a single completeness score. This fine-grained decomposition allows further detailed analysis to identify which clinical aspects were systematically under-represented in generated summaries.

Fluency Agent The fluency agent evaluates the linguistic quality of the generated summary, independently of its factual content. Only the generated summary was provided, neither the source full case nor the reference summary is shown, so the agent cannot penalize for missing clinical content. Grammatical correctness, readability, and natural language flow was assessed on a 0–10 scale. The agent is explicitly instructed to evaluate in the language of the text being assessed, as identified in the prompt header.

Consistency Agent The consistency agent evaluates factual equivalence across the different language versions of the summary generated for the same clinical case. Rather than assessing translation exactness, the agent evaluates whether all language versions contain the same diagnosis, interventions, outcomes, and patient details, allowing for natural variation in phrasing and level of detail. All available language versions of a given case summary were presented to the agent in a single prompt, and a single cross-lingual consistency score was assigned on a 0–10 scale. This dimension is unique to the proposed framework and is motivated by the multilingual nature of the shared task, so that, a system that produces high-quality summaries in high-resource languages but systematically omits clinical details in lower-resource ones would be penalised by this metric, even if per-language reference-based scores remain acceptable.

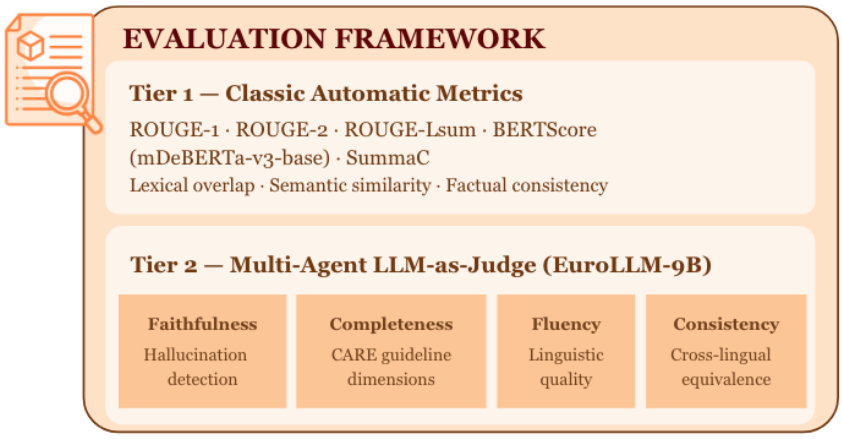

Figure 3: Overview of the two-tier evaluation framework proposed for MultiClinSum-2.

6.3. Score Normalisation and Aggregation

All LLM judge scores are normalized to $[0, 1]$ before reporting the results table by dividing the raw integer score (0–10) by 10. BERTScore F1 is already in $[0, 1]$ by definition; ROUGE scores are also F1 values in $[0, 1]$. For each team-language combination, per-sample scores are aggregated into a system-level summary report containing the mean, standard deviation, minimum, and maximum for each of the eight metrics independently. No composite ranking score is defined, and the leaderboard displays all eight metrics as separate columns, giving participants and readers full visibility into the performance profile of each system.

7. MultiClinSum-2 Participant Systems

This section presents a systematic description of the methodological approaches implemented by participant teams during the shared task, examining the different strategies applied across the 9 systems. For each team, a dedicated subsection details the summarization strategy, model architecture, additional clinical knowledge sources incorporated, and prompt engineering or optimisation mechanisms employed. Table 21 in Appendix C provides a summary table of the participant systems based on these characteristics.

Team ixa-sum [30] This team explored two main strategies applied in two language sub-tracks subsets, romance languages (Spanish, Catalan, French, Portuguese, Italian, Romanian) and west germanic (English, Dutch, German). One of them consisted on a dynamic one-shot prompting approach,

which automatically identifies and retrieves contextually similar clinical examples from a reference pool and injects them into the LLM's prompt to guide the style and structure of the generated text. The second strategy utilized Group Relative Policy Optimization (GRPO), a reinforcement learning alignment technique that optimizes a reward function composed of multiple critical performance signals, including lexical similarity metrics like ROUGE and semantic alignment via entailment-based fact-checking models. Both strategies were evaluated using two open-source models, Qwen2.5-7B-Instruct and Gemma-4-E4B-it, with an ablation study performed for English, Spanish and Catalan. The findings highlight a performance trade-off: while the Qwen-driven models score higher on traditional automated metrics (such as ROUGE-Lsum and BERTScore), the Gemma models optimized via alignment score significantly better under LLM-as-a-judge evaluations focusing on coverage and faithfulness.

Team NLP4Health [31] This team designed approaches based on three different strategies: a prompt-only Qwen2.5-7B system, a Qwen2.5-1.5B model adapted with QLoRA, and a Qwen2.5-7B model adapted with QLoRA, all evaluated under a resource-constrained setting using a single NVIDIA RTX A4000 GPU with 16GB of memory. The prompt-only configuration relies on a carefully engineered instruction template that constrains the output to a single paragraph of 3 to 5 sentences, explicitly instructs the model not to invent details, and requires the preservation of key clinical elements such as patient age, sex, medication names and outcome. For long clinical reports, a balanced excerpt strategy is applied, combining the beginning, middle and end of the source document to ensure that patient presentation as well as treatment and outcome information are both captured within the model context window. The QLoRA-adapted runs fine-tune low-rank adapters on top of a quantized base model using language-specific training examples for each of the four covered languages.

Team DACHausa [32] This team approach consisted of fully fine-tuning an mT5-small model of 300M parameters on 334,024 full case-summary pairs across all 15 languages, using language identifier prefixes appended to each source full case for explicit cross-lingual adaptation. The language identifiers, such as [en], [es] or [fr], were treated as prefixes tokenized by the existing mT5 SentencePiece vocabulary. To handle the large dataset efficiently within memory constraints, a file-indexed streaming dataset was implemented, loading text on-the-fly during training rather than loading the full dataset into memory. The model was trained using the *AdamW* optimizer with a linear learning rate decay, distributing training across two NVIDIA RTX A6000 GPUs in full 32-bit floating-point precision. A per-language stratified random split was applied to the full training dataset, which resulted in 334,024 training pairs and 37,121 validation pairs, with the best checkpoint selected by validation BERTScore. At inference time, beam search decoding with four beams was used to generate summaries of up to 256 tokens.

Team InfoLab-FEUP [33] The InfoLab-FEUP system proposed an agentic summarization pipeline that processed each clinical case report through a series of structured steps. First, four specialized extraction agents decompose the case report into clinically relevant sections (patient characteristics, diagnostics, interventions and outcomes), followed by an iterative agent-critic loop that validates the extracted content against the source full case and triggers revisions when clinically significant errors are detected, up to a maximum of three iterations. The extracted sections are then enriched with biomedical concepts identified by QuickUMLS, which maps text spans to standardized UMLS terms, augmented in some cases with their corresponding definitions. Finally, a summary generation agent produces the clinical summary from the enriched structured context. The entire pipeline was implemented locally powered by Llama3-8B-Instruct via Ollama, without fine-tuning. For the particular Portuguese language sub-track, two strategies were compared: a direct pipeline using a Portuguese QuickUMLS index, and a translation-assisted pipeline that translates the target full case into English, applies the described English pipeline, and translated the output summary into Portuguese.

Team RaRaMi [34] The RaRaMi system proposes a QLoRA-based parameter-efficient fine-tuning pipeline for the Romanian language sub-track, developed entirely on a single GPU under constrained computational resources. Three instruction-tuned multilingual backbones of parameter size range of 1B–8B were used, including Llama-3.2-1B-Instruct, Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct. These models were fine-tuned under identical conditions, varying only backbone identity and training subset size. To ensure fair cross-backbone comparability, three strictly nested and hash-verified training subsets of 1,000, 6,000 and full records were employed as the Romanian training corpus. All models were trained using the same Romanian prompt template, QLoRA hyperparameters and greedy decoding strategy, with a completion-only loss mask applied so that the language modeling loss is computed exclusively over the generated summary tokens. Based on the performance results, the team submitted the system corresponding to Llama-3.1-8B-Instruct fine-tuned on 6,000 samples, which metrics were computed on a fixed development subset of 100 samples.

Team UMUTeam [35] This team system investigated whether medical-domain continual pre-training models improves Spanish clinical case report summarization after supervised instruction tuning. To this end, six pairs of models are evaluated, where each pair consisted of a general-domain base model and a medically adapted version developed by the authors through continual pretraining on Spanish medical-domain corpora. The evaluated model families are Qwen 3 (4B and 8B), Gemma 3 (1B and 4B) and Salamandra (2B and 7B). Both variants in each pair are fine-tuned under identical conditions using LoRA adapters on the Spanish language sub-track training data, with the same prompt template, hyperparameters and greedy decoding strategy, so that performance differences can be attributed specifically to the effect of medical-domain continual pretraining. The submitted runs correspond to Qwen 3 8B and Salamandra 7B, in both their base and medically adapted versions, selected based on development set performance.

Team MediScribes [36] The MediScribes system focused on studying the incorporation of knowledge graphs as a factuality improvement of automatically generated summaries from LLMs. They submitted results for English and Dutch language sub-tracks comparing five methods, all using GPT-5.1 as the underlying model: a direct text-to-summary baseline with two-shot prompting, two knowledge graph approaches based on a custom FHIR-light ontology (Text-KG and Sentence-KG) that convert the clinical report into RDF triples before generating the summary, and two GraphRAG-based approaches (GraphRAG-KG and FGraphRAG-KG) that automatically construct knowledge graphs through entity and relation extraction. The FHIR-light ontology is a reduced subset of the FHIR standard designed for clinical summarization, containing ten core classes covering patient demographics, conditions, observations, medications, procedures and care plans. For Dutch, language-specific spaCy models and Dutch prompt templates are used. No fine-tuning or training data is used in any of the submitted runs, all methods operate entirely at inference time.

8. Results

System label notation. Throughout this section, the four baseline systems are referred to using the following notation: *Qwen2.5-7B-Ins (ZS)* and *Llama3.1-8B-Ins (ZS)* denote the zero-shot Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct baselines respectively, where *ZS* indicates that no task-specific fine-tuning was performed; *MCS-mT5-1epoch* and *MCS-mT5-2epoch* refer to the Medical-mT5-large model fine-tuned on the MultiClinSum-2 training data for one and two epochs respectively. Participant systems are identified by their submission team name and run number as provided at registration time. This section reports an overview of system performance, displaying only the top-3 best performing systems per language sub-track as shown in Table 3. Complete per-language leaderboards are provided in Appendix A. Additionally, language-specific pairs of full case and generated summaries that obtained notable results are shown in Appendix B.

Across the sub-tracks with the highest participation, team ixa-sum [30] achieved robust performance across high-resource languages, leading in BERTScore across English (0.876), Spanish (0.882), Catalan (0.882), and French (0.873). Their GRPO-based system combined with a dynamic one-shot prompting strategy, outperformed both zero-shot and fine-tuned MCS-mT5-large baseline systems and the rest of the teams on ROUGE-Lsum and BERTScore metrics. Portuguese results showed an exception, where team NLP4Health [31] run1 obtained highest BERTScore (0.870), barely outperforming team ixa-sum, with their carefully engineered prompt-only Qwen2.5-7B configuration. In English, where 24 runs were submitted across 8 teams, ixa-sum run2 obtained the highest ROUGE-Lsum (0.322) and BERTScore (0.876). However, the KG-based summarization system employed by team MediScribes [36], achieved the best performance in *Completeness* and *Faithfulness* LLM-as-judge dimensions, demonstrating greater results compared to all other systems on these clinical quality dimensions despite lower lexical overlap scores. This observed pattern suggests that different methodological choices benefit different aspects of summary quality.

For mid-resource languages such as Italian, Dutch, and Romanian, both *Qwen2.5-7B-Ins (ZS)* and *Llama3.1-8B-Ins (ZS)* baseline models performed notably across these sub-tracks. In Dutch, MediScribes run3, using a FHIR-light knowledge graph approach with GPT-5.1, achieved the highest *Completeness* score (0.983) despite a lower BERTScore (0.862). In Romanian, only ixa-sum and RaRaMi [34] teams submitted predictions, from which ixa-sum lead on BERTScore (0.873) and most LLM-as-judge metrics, while RaRaMi system designed for Romanian with QLoRA fine-tuning, obtained slightly lower results on the automated metrics, even though its *Faithfulness* (0.700) remains competitive.

For the German sub-track, the fine-tuned baseline model *MCS-mT5-2epoch* achieved notable lexical results, obtaining the highest ROUGE-1 (0.307), only surpassed by team ixa-sum in BERTScore (0.856) and *Consistency* (0.652) metrics. Team DACHausa [32] was the only participant to submit results in all 15 languages, with the exception of Norwegian. Their fully fine-tuned mT5-small system, achieved competitive results, however was consistently outperformed by zero-shot causal LLMs baselines, indicating a potential trade-off between broad multilingual coverage and per-language performance when using a lighter model. For lower-resource sub-tracks such Russian, Greek, Czech, Danish, Norwegian, and Swedish, *Llama3.1-8B-Ins (ZS)* achieved the highest ROUGE-Lsum and BERTScore in most cases, outperforming the single submitted system of DACHausa across all metrics in languages such as Czech (ROUGE-Lsum: 0.220 vs. 0.192) and Danish (0.238 vs. 0.208). This suggests that zero-shot instruction-tuned model capabilities provide a strong baseline that fine-tuned smaller models struggle to overcome.

Regarding the LLM judge dimensions (*Faithfulness*, *Completeness*, *Fluency*, and *Consistency*), several patterns were observed across languages that in some cases contradict or support the results obtained in automatic metrics. The first observation is that *MCS-mT5* baseline systems consistently achieved substantially lower scores across all four judge dimensions compared to their BERTScore rankings. This particular evident in Italian, where *MCS-mT5-1epoch* obtains very low values in *Faithfulness* (0.030), *Fluency* (0.048), and *Consistency* (0.000), despite a competitive BERTScore of 0.866. This finding may be associated with the known limitation of encoder-decoder models family to produce degenerate outputs in languages not seen during pre-training, an effect amplified by the 512-token full case truncation applied during fine-tuning and inference. Second, *Llama3.1-8B-Ins (ZS)* consistently showed low *Consistency* scores across all languages, more notably in Norwegian (0.107), Czech (0.279), and Spanish (0.345), suggesting a potential weakness of this zero-shot prompting strategies, where the model can produce semantically accurate summaries but lack factual equivalence across language versions of the same clinical case. Finally, it was observed that participant systems that incorporate additional clinical information to their models, such as MediScribes or NLP4Health teams, consistently achieved higher *Completeness* scores relative to their BERTScore rankings, suggesting that key clinical dimensions defined in CARE guideline, might be fulfilled when integrating this kind of knowledge.

Table 3
Top-3 performing teams per language, ranked by BERTScore (LLM-as-judge metrics: EuroLLM-9B). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Cons. = Consistency. Bold indicates best value per language and metric. Language codes correspond to: EN = English, ES = Spanish, FR = French, PT = Portuguese, IT = Italian, RU = Russian, CA = Catalan, NO = Norwegian, DA = Danish, RO = Romanian, DE = German, EL = Greek, NL = Dutch, CS = Czech, SV = Swedish.

Lang	Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Cons.
	ixa-sum	2	**0.440 (0.220)**	**0.322**	**0.876**	0.732	0.832	0.700	0.698
EN	ixa-sum	1	0.433 (0.216)	0.319	0.875	0.726	0.813	0.700	0.698
	Llama3.1-8B-Ins (ZS)	1	0.398 (0.168)	0.272	0.873	**0.763**	**0.921**	**0.700**	**0.698**
	ixa-sum	2	**0.467 (0.235)**	**0.314**	**0.882**	0.713	0.796	0.700	**0.649**
ES	ixa-sum	1	0.463 (0.232)	**0.314**	0.881	0.709	0.786	0.700	0.647
	Llama3.1-8B-Ins (ZS)	1	0.439 (0.205)	0.289	0.876	**0.728**	**0.858**	**0.700**	0.345
	ixa-sum	2	0.456 (0.232)	0.301	**0.882**	0.705	0.867	**0.700**	**0.627**
CA	ixa-sum	1	**0.458 (0.233)**	**0.306**	0.881	0.702	0.833	**0.700**	0.640
	Llama3.1-8B-Ins (ZS)	1	0.430 (0.202)	0.278	0.873	**0.711**	**0.891**	0.699	0.447
	ixa-sum	2	0.415 (0.172)	0.256	**0.873**	**0.742**	**0.904**	**0.702**	**0.667**
FR	ixa-sum	1	0.410 (0.180)	0.263	0.872	0.727	0.866	0.701	0.654
	Llama3.1-8B-Ins (ZS)	1	**0.432 (0.201)**	**0.274**	0.871	0.728	0.809	0.697	0.477
	NLP4Health	1	0.384 (0.156)	0.257	**0.870**	0.738	0.891	**0.701**	0.636
PT	ixa-sum	1	0.381 (0.152)	0.256	0.870	0.741	**0.915**	0.700	**0.652**
	Llama3.1-8B-Ins (ZS)	1	**0.409 (0.178)**	**0.270**	0.869	**0.746**	0.884	0.700	0.418
	Llama3.1-8B-Ins (ZS)	1	**0.374 (0.145)**	**0.255**	**0.865**	0.717	0.881	0.699	0.430
NL	ixa-sum	2	0.350 (0.115)	0.236	0.862	0.732	0.949	**0.712**	0.665
	MediScribes	3	0.336 (0.109)	0.223	0.862	**0.747**	**0.983**	0.708	**0.669**
	Llama3.1-8B-Ins (ZS)	1	**0.379 (0.155)**	**0.251**	**0.871**	0.720	0.865	**0.701**	0.550
IT	ixa-sum	1	0.358 (0.136)	0.242	0.870	0.718	0.890	0.700	**0.647**
	MCS-mT5-1epoch	1	0.346 (0.142)	0.225	0.866	0.030	0.513	0.048	0.000
	Llama3.1-8B-Ins (ZS)	1	**0.383 (0.166)**	**0.262**	**0.874**	0.738	**0.924**	0.699	0.431
RO	ixa-sum	1	0.358 (0.146)	0.251	0.873	0.737	0.911	**0.700**	0.569
	MCS-mT5-2epoch	1	0.355 (0.152)	0.238	0.869	**0.565**	0.569	0.529	**0.431**
	ixa-sum	1	0.297 (0.094)	**0.205**	**0.856**	0.721	0.827	**0.699**	**0.652**
DE	MCS-mT5-2epoch	1	**0.307 (0.111)**	0.199	0.854	0.544	0.481	0.469	0.388
	MCS-mT5-1epoch	1	0.306 (0.110)	0.198	0.854	**0.566**	**0.481**	0.506	0.401
	MCS-mT5-2epoch	1	0.259 (0.069)	0.251	**0.860**	0.560	0.442	0.202	0.293
RU	MCS-mT5-1epoch	1	0.248 (0.063)	0.239	0.859	0.550	0.428	0.343	0.296
	PsychAI-Discovery-Lab	1	**0.291 (0.066)**	**0.285**	0.858	**0.705**	**0.575**	**0.697**	**0.387**
	Llama3.1-8B-Ins (ZS)	1	**0.355 (0.129)**	**0.220**	**0.866**	0.699	0.865	0.681	0.279
CS	MCS-mT5-2epoch	1	0.333 (0.118)	0.200	0.861	0.543	0.476	0.166	0.306
	MCS-mT5-1epoch	1	0.333 (0.117)	0.199	0.861	**0.567**	**0.483**	**0.255**	**0.340**
	Llama3.1-8B-Ins (ZS)	1	**0.347 (0.132)**	**0.238**	**0.859**	0.722	0.860	0.651	0.306
DA	MCS-mT5-1epoch	1	0.326 (0.132)	0.214	0.857	0.544	0.507	0.524	**0.387**
	MCS-mT5-2epoch	1	0.325 (0.131)	0.213	0.857	**0.537**	**0.532**	**0.518**	0.373
	MCS-mT5-2epoch	1	0.262 (0.069)	0.254	**0.868**	0.611	0.562	0.592	**0.328**
EL	MCS-mT5-1epoch	1	**0.264 (0.070)**	**0.256**	0.868	0.104	0.560	0.046	0.000
	DACHausa	1	0.309 (0.070)	0.304	0.847	**0.654**	**0.578**	**0.695**	0.282
	Llama3.1-8B-Ins (ZS)	1	**0.351 (0.134)**	**0.244**	**0.862**	0.709	0.834	0.591	0.107
NO	MCS-mT5-2epoch	1	0.329 (0.134)	0.221	0.857	0.546	0.523	0.521	0.382
	MCS-mT5-1epoch	1	0.328 (0.133)	0.219	0.857	**0.553**	**0.505**	**0.526**	**0.385**
	Llama3.1-8B-Ins (ZS)	1	**0.345 (0.133)**	**0.233**	**0.861**	0.715	0.891	0.699	0.354
SV	PsychAI-Discovery-Lab	1	0.327 (0.115)	0.227	0.858	0.710	0.779	0.697	0.609
	Qwen2.5-7B-Ins (ZS)	1	0.333 (0.125)	0.226	0.856	**0.720**	**0.898**	**0.700**	**0.665**

9. Additional Analysis

9.1. Analysis of Machine Translation Systems on Summarization Metrics

A large proportion of the full case-summary pairs released in MultiClinSum-2 were obtained through automatic machine translation, where English full case-summary pairs were translated into the remaining task languages. While quality filtering steps were applied to minimize translation artifacts, including language identification and removal of truncated or repetitive outputs, the selection of translation system may still influence the linguistic characteristics of the translated clinical cases and, consequently, the performance of summarization systems evaluated on them. To study this effect, an alternative set of translations was obtained for a subset of eight languages (Spanish, French, Portuguese, Italian, Dutch, Swedish, Czech, and Romanian), using a second machine translation system (*Translated*). Inference was computed using three baseline models: *Llama3.1-8B-Ins (ZS)*, *Qwen2.5-7B-Ins (ZS)*, and *MCS-mT5-2epoch*. Evaluation scores were then averaged across all eight languages and compared against the results obtained with the primary translation system (SalamandraTA-7B-instruct) used in the task release.

Figure 4 shows the comparison of the results between the two translation systems—System 1 (SalamandraTA-7B-instruct) and System 2 (alternative translation)—across the entire set of evaluation metrics for each baseline model. The preliminary results obtained revealed a high level of agreement between the two systems across all metrics and all three baselines, with small differences in both automatic metrics (ROUGE-1, ROUGE-2, ROUGE-Lsum, BERTScore) and LLM-as-judge dimensions (*Faithfulness*, *Completeness*, *Fluency*, *Consistency*). The most notable difference was observed in the Consistency dimension for *LLaMA-3.1-8B (ZS)*, where System 2 scores slightly higher, and in the *MCS-mT5-2epoch*, where a more significant separation in *Faithfulness* was observed, suggesting that the fine-tuned encoder-decoder model may be more sensitive to translation style differences than the zero-shot causal LLMs. Overall, the close alignment between the two systems across baselines suggests that the summarization performance observed in MultiClinSum-2 was not affected by the choice of machine translation system and that the employed translation pipeline does not introduce biases that would affect the results of the evaluation.

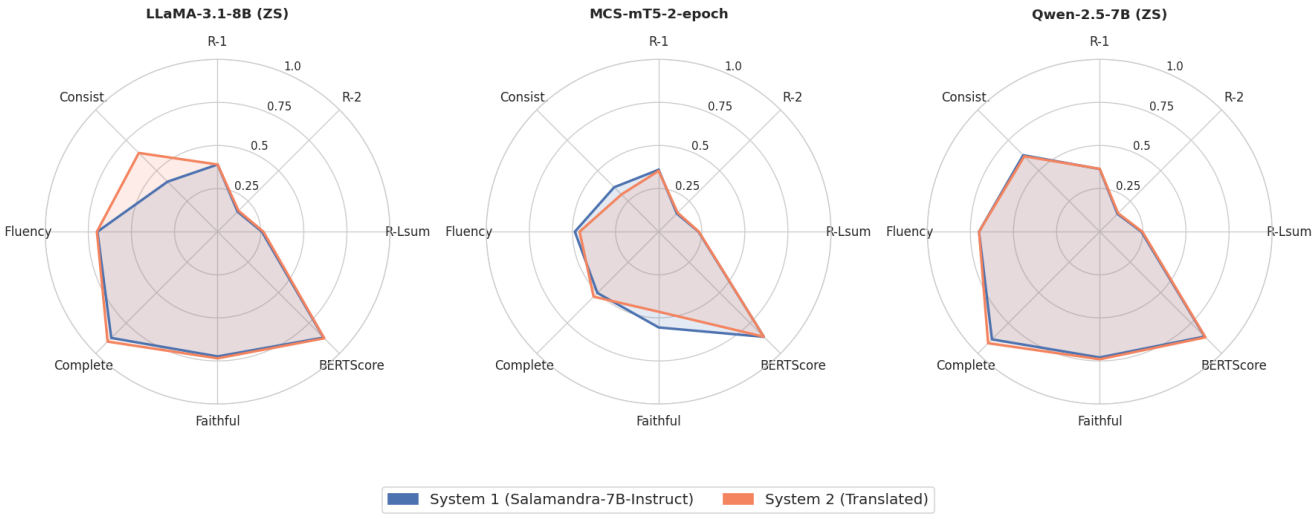

Figure 4: Chart comparison of evaluation metrics for three baseline systems (LLaMA-3.1-8B (ZS), MCS-mT5-2epoch, and Qwen2.5-7B (ZS)) when applied to clinical case reports translated by two different machine translation systems: System 1 (SalamandraTA-7B-Instruct, primary translation used in the official task release) and System 2 (alternative translation system).

9.2. Impact of Full Case Length on Summarization Performance

The length of a clinical case report may have an influence on the quality of the generated summary and, consequently, the evaluation results obtained. To investigate this relationship, an analysis was conducted by grouping test set full cases into five length intervals based on word count quintiles. The resulting ranges and their corresponding word count limits are as follows: *Very Short* ($\leq$300 words, 3,006 cases), *Short* (301–432 words, 2,992 cases), *Medium* (433–584 words, 3,006 cases), *Long* (585–803 words, 2,991 cases), and *Very Long* (>803 words, 2,998 cases). This interval strategy ensures an approximately balanced number of cases per category, enabling fair comparisons across length groups. Figure 5 shows the percentage distribution of the clinical cases per length interval for each language sub-track, where romance languages such as French, Catalan, and Spanish tend to have a higher proportion of shorter cases than Germanic and Nordic languages.

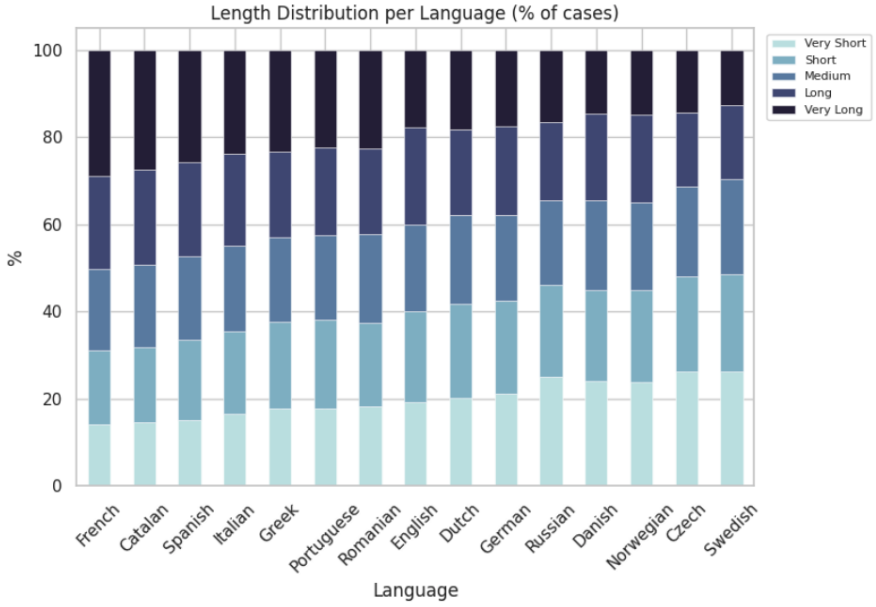

Figure 5: Percentage distribution of full case length intervals (Very Short, Short, Medium, Long, Very Long) per language sub-track in the MultiClinSum-2 test set. Each interval was defined by word count quintiles computed over the full test set distribution.

Evaluation metrics were then averaged and compared across the five length intervals, using only the baseline system outputs. The results, shown in Figure 6, reveal several interesting and complementary patterns across the two evaluation approaches. Regarding automatic metrics, models performance followed the expected trend, ROUGE-1, ROUGE-2, and ROUGE-Lsum had the better results at 'Medium' length cases and drop progressively for 'Long' and 'Very Long' cases (ROUGE-Lsum: 0.254 for 'Very Short', 0.249 for 'Medium', falling to 0.229 for 'Very Long'). This is consistent with the known behaviour of lexical overlap metrics, which are sensitive to the increasing difficulty of capturing all relevant content as the source document grows longer, as well as to the greater diversity of phrasing in longer clinical narratives. BERTScore follows a similar pattern, with a notable drop for 'Very Long' cases (0.867 for 'Medium' and Long vs. 0.833 for 'Very Long'), suggesting that semantic alignment also decreases for longest full cases. For the fine-tuned MCS-mT5 baseline, this effect may be due to the 512-token source truncation applied during both training and inference, which particularly affects longer clinical cases by cutting off potentially critical clinical information.

The LLM-as-judge dimensions however, reveal a more detailed and, in some cases, unexpected outcome. *Faithfulness* decreases with case length (0.648 for 'Very Short' to 0.608 for 'Very Long'), suggesting that longer clinical cases present higher risks of hallucination or unsupported claims in

the generated summary, likely because the model must compress a larger and more complex clinical narrative. *Completeness*, in contrast, increases with length (0.667 for 'Very Short', rising to 0.704 for 'Long') before dropping slightly for 'Very Long' cases (0.681). This suggests that medium/long clinical cases, which tend to follow more structured reporting conventions and contain richer clinical detail across all CARE dimensions, are actually summarized more completely than very short ones, a finding that would harder to detect from ROUGE or BERTScore alone. *Fluency* and *Consistency* remain relatively stable across all length intervals (*Fluency*: 0.562–0.579; *Consistency*: 0.404–0.434), indicating that document length primarily affects the clinical content quality of generated summaries rather than their linguistic coherence or cross-lingual factual equivalence.

Overall, these findings highlight an important difference in how length affects different aspects of summary quality, while longer cases are harder to summarize faithfully and obtain lower scores on lexical and semantic metrics, they may offer more complete clinical coverage up to a certain length threshold. This underscores the importance of the multi-dimensional evaluation framework introduced in this edition, as a single metric would provide an incomplete picture of summarization quality across the full distribution of clinical case lengths.

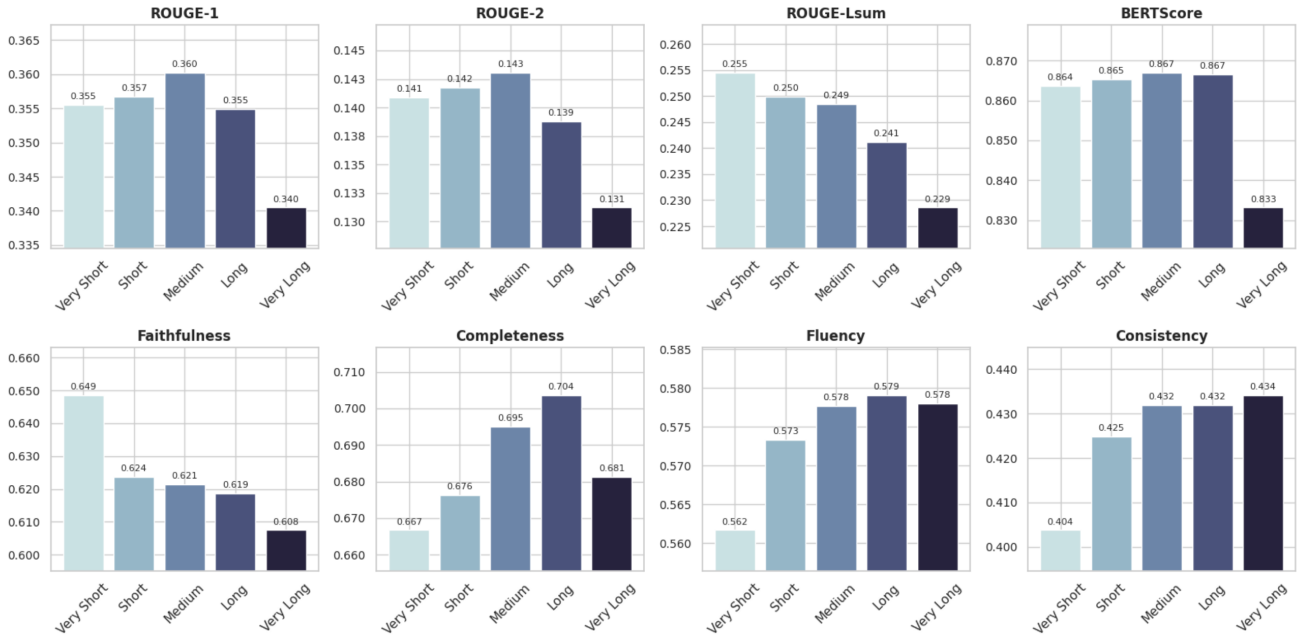

Figure 6: Average evaluation metrics per full case length intervals across all baseline system outputs. Scores are averaged across all 15 language sub-tracks.

9.3. Factuality Evaluation with SummaC

As a complementary factuality evaluation independent of the LLM-as-judge pipeline, *SummaC* [37] metric was computed across all participant and baseline runs. Specifically, a *SummaC* zero-shot (ZS) variant was applied using a multilingual DeBERTa-v3-large model [38] as the underlying NLI backbone, measuring the degree to which each generated summary is entailed by its source clinical full case. Positive scores indicate strong factual consistency between the summary and the source; scores near zero reflect a neutral alignment; and negative scores suggest the presence of claims that contradict or are unsupported by the source, which might indicate potential hallucinations detected.

Table 4 reports the overall statistics grouped by baseline models or participant system. Both systems groups showed negative mean and median SummaC scores on average, potentially indicating that generated summaries are not fully consistent with their source clinical cases as assessed by NLI-based

entailment. Participant systems showed higher mean (-0.108 vs. -0.152) and median (-0.138 vs. -0.208) scores than baseline systems, suggesting that fine-tuned and task-adapted approaches tend to produce more factually supported summaries than zero-shot causal LLMs and fine-tuned medical domain models on average. This is consistent with the idea that instruction-following models generating abstractive summaries may introduce paraphrasing and reformulation that NLI models interpret as factual inconsistency, particularly given the complexity of clinical terminology.

Table 5 shows the *SummaC* ranking results for each team revealing a clear difference compared to the BERTScore leaderboard. Teams UMUTeam and DACHausa achieved the highest *SummaC* scores (0.128 and 0.119 respectively), while their systems ranked lower on automatic and LLM-as-judge metrics. This suggests that their outputs are more closely entailed by the source case based on the model of choice to compute the results. In contrast, teams such as ixa-sum or MediScribes, which were top-performing teams across most sub-tracks on BERTScore and *Faithfulness* scores, ranked 8th (mean -0.229) and 11th (mean -0.242) respectively, indicating that their abstractive outputs, while semantically close to reference summaries, may contain rephrasings that NLI models identified as inconsistent.

It should be noted that *SummaC* scores in the clinical domain are expected to be lower than in general-domain summarization, as clinical summaries tend to paraphrase and restructure information rather than reproducing it literally, a property that NLI models may incorrectly penalize as factual inconsistency. Additionally, the results may be particularly sensitive to the choice of NLI language model employed. These results should therefore not be interpreted as an absolute measure of factual correctness, but rather as a relative evaluation for comparing the level of faithfulness across different methodologies.

Table 4
Descriptive statistics of *SummaC* scores by system type.

System type	Nr. of runs	Mean	Median	std	min	max
Participants	48	-0.108	-0.137	0.184	-0.444	0.199
Baselines	46	-0.152	-0.207	0.132	-0.347	0.139

Table 5
SummaC scores by team, ranked by mean score.

Rank	Team	Langs.	Mean	Median	std	min	max
1	UMUTeam	1	**0.1280**	0.1280	—	0.1280	0.1280
2	DACHausa	14	0.1185	0.1196	0.0417	0.0370	0.1997
3	MCS-mT5-2epoch	15	0.0230	0.0159	0.0472	-0.0562	0.1389
4	RaRaMi	1	0.0019	0.0019	—	0.0019	0.0019
5	NLP4Health	4	-0.1390	-0.1218	0.0554	-0.2148	-0.0975
6	PsychAI-Discovery-Lab	7	-0.1858	-0.2290	0.0828	-0.2829	-0.0430
7	ixa-sum	9	-0.2288	-0.2822	0.1238	-0.3347	0.0012
8	Llama3.1-8B-Ins (ZS)	15	-0.2326	-0.2516	0.0594	-0.3469	-0.1073
9	Qwen2.5-7B-Ins (ZS)	15	-0.2407	-0.2518	0.0337	-0.2809	-0.1682
10	MediScribes	2	-0.2416	-0.2416	0.0717	-0.2923	-0.1909
11	Aleyna	1	-0.4207	-0.4207	—	-0.4207	-0.4207
12	InfoLab-FEUP	2	-0.4327	-0.4327	0.0161	-0.4442	-0.4213

9.4. Multilingual Embedding Space Analysis

With the objective of characterizing the released task dataset, an embedding study of the full cases derived from the PMC-Patients was conducted. The same model used to compute BERTScore,

`mDeBERTa-v3-base`, was selected for embedding extraction, producing a 768-dimensional representation for each full case. Due to computational and visualization constraints, only the 1,000 full case test set was used for this analysis. Embeddings were subsequently reduced to two dimensions via UMAP [39] for visualization purposes. Figure 7 shows the resulting 2-dimensional embedding space, with each point coloured by language.

The visualization shows clear language family clustering patterns. Romance languages such as Spanish, Portuguese, Catalan, Romanian, French, and Italian, formed a dense, well-defined cluster in the lower-right region of the space, reflecting their shared lexical and semantic properties. Slavic languages like Czech and Russian, appear in close proximity in the lower-left region. The Germanic and Nordic languages—Dutch, Norwegian, Swedish, and Danish—form a broader overlapping cluster in the central region, with a mixed structure. In contrast, several languages occupy clearly isolated locations: English appears as a distinct cluster in the top of the image, Greek forms a well-separated cluster in the upper-left, and German is positioned distinctly in the right of the space.

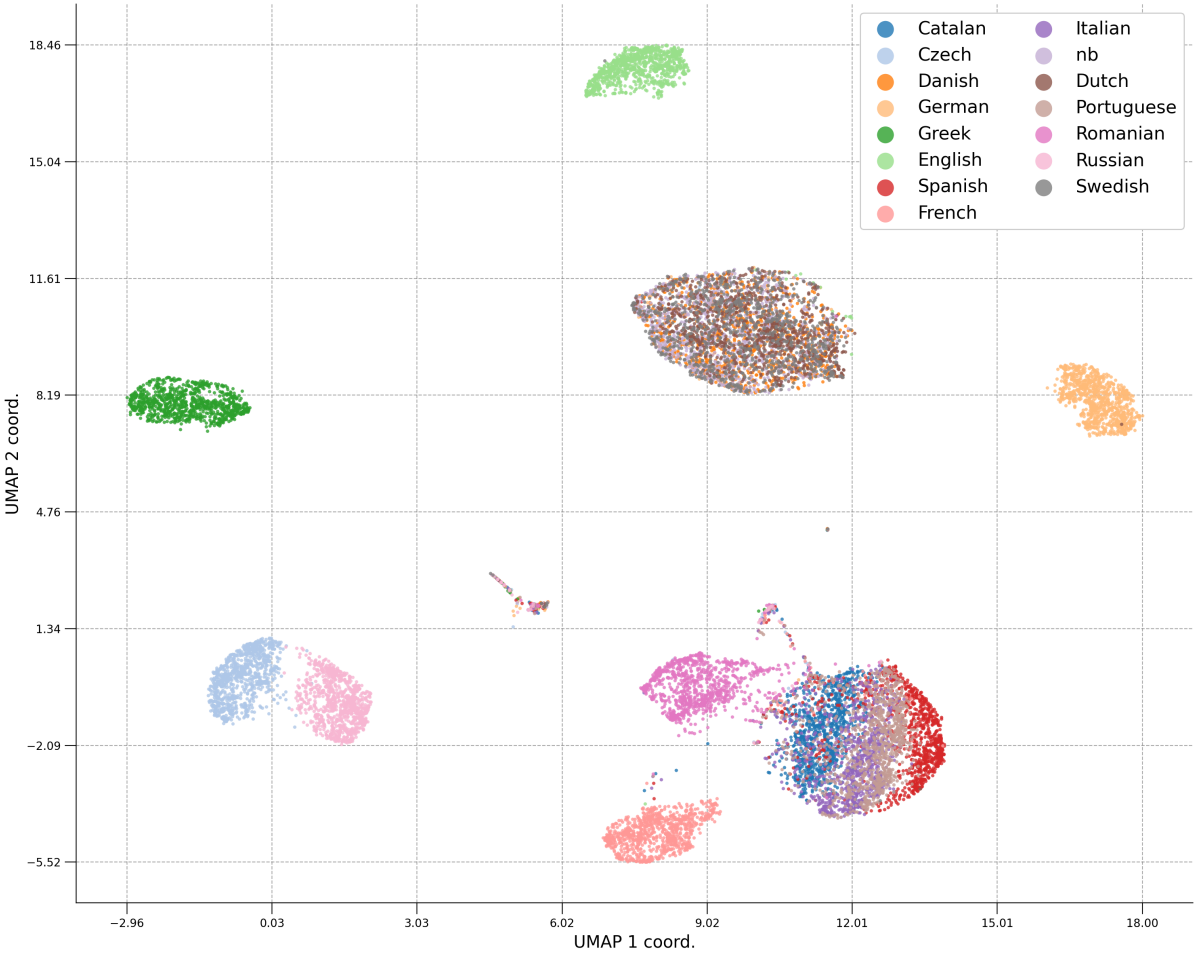

Figure 7: UMAP coordinates projection of mDeBERTa-v3-base language model embeddings of test set full clinical cases, coloured by language. Each point represents one clinical case reduced to 2 dimensions.

10. Conclusions

This paper presents MultiClinSum-2, the second edition of the shared task on multilingual automatic summarization of clinical case reports, organized as part of the BioASQ Lab at CLEF 2026. Overall, the task received satisfactory participation, with 9 teams submitting a total of 89 runs across 15 language sub-tracks, reflecting a broad and diverse engagement with the multilingual clinical summarization

challenge.

Two key novelties distinguish this second edition from the first edition. The first is a substantial expansion of language coverage, scaling the task from four to 15 languages and encompassing a wide range of linguistic contexts, from high-resource languages such as English, Spanish, French, and Portuguese, to lower-resource ones such as Norwegian, Greek, Czech, and Swedish. To support this expansion, a large and carefully curated multilingual dataset was constructed and publicly released, comprising full clinical case reports paired with corresponding reference summaries in each of the 15 supported languages, derived from both an author-written gold standard collection and a large-scale PMC-Patients subset in English, and translated to the remaining languages. The second key novelty is the introduction of an enhanced evaluation framework that complements established automatic metrics (ROUGE and BERTScore) with a novel multi-agent LLM-as-judge system. This system assesses generated summaries across four clinically relevant dimensions: Faithfulness, Completeness, Fluency, and Consistency, providing a more granular and clinically meaningful analysis of summary quality. Additionally, two complementary baseline systems were developed and released: a zero-shot causal LLM approach and a fine-tuned encoder-decoder medical model, with the aim of serving as reference points for participating teams and future benchmarking efforts.

Participating systems exhibited a rich diversity of methodological approaches, ranging from reinforcement learning strategies via GRPO and carefully engineered prompting strategies, to parameter-efficient fine-tuning with QLoRA and knowledge-augmented pipelines incorporating structured biomedical resources such as UMLS and clinical knowledge graphs. The use of state-of-the-art open-source model families—including Qwen, Gemma, and Llama—reflects the rapid evolution of the LLM landscape and its growing applicability to clinical NLP tasks, highlighting the community interest in advancing in the topic of multilingual clinical summarization.

In terms of future editions and further work, several directions are currently under consideration. A first line of development involves the incorporation of entity-aware evaluation mechanisms, with the objective of assessing the preservation of specific clinical concepts—such as diagnoses, medications, and procedures, previously extracted using NLP techniques such as Named Entity Recognition (NER)—enabling a more fine-grained assessment of clinical content coverage. In addition to this, a more systematic hallucination detection process is planned to be integrated in the evaluation framework, with the aim of categorizing factual errors in generated summaries such as fabricated diagnoses or unsupported clinical claims. A second direction concerns the quality assessment of reference summaries themselves. To this end, reference summaries are currently being manually annotated by clinical experts using the Prodigy annotation tool [40], who assess whether each summary adequately covers key clinical dimensions including patient demographics, clinical presentation, diagnosis and diagnostic process, intervention and treatment, outcome, and follow-up. This annotation effort aims to produce a higher-quality gold standard that can support more reliable evaluation in future editions of the task. An example of the annotation interface is shown in Figure 8 in the Appendix D. Additionally, the analysis of potential demographic and gender biases in the task corpus is planned. From a resource perspective, future editions will consider incorporating new available PMC-Patients data not covered in the present release, while also exploring the incorporation of new languages semantically distinct such as Chinese, which would extend the task linguistic contexts.

Acknowledgments

The MultiClinSum-2 track was funded by Spanish and European projects such as DataTools4Heart (Grant Agreement No. 101057849), AI4HF (Grant Agreement No. 101080430). Salvador Lima-López fellowship within the “Generación D” initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1). Funded by the European Union

NextGenerationEU funds, through PRTR. We would also like to acknowledge the BioASQ organizers, Anastasios Nentidis and Anastasia Krithara for their technical support.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 4.6 for grammar checks and sentence rephrasing. This was followed by a review and edit of the content, with the author(s) taking full responsibility for the publication.

References

[1] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, J. Mostafa, A systematic review of automatic text summarization for biomedical literature and ehrs, Journal of the American Medical Informatics Association 28 (2021) 2287–2297.

[2] M. Murad, B. Vaa Stelling, C. West, et al., Measuring documentation burden in healthcare, J Gen Intern Med. 39 (2024) 2837–2848.

[3] L. Bednarczyk, D. Reichenpfader, C. Gaudet-Blavignac, A. K. Ette, J. Zaghir, Y. Zheng, A. Bensahla, M. Bjelogrlic, C. Lovis, Scientific evidence for clinical text summarization using large language models: Scoping review, J Med Internet Res 27 (2025) e68998. URL: https://www.jmir.org/2025/1/e68998. doi:10.2196/68998.

[4] A. Chaves, C. Kesiku, B. Garcia-Zapirain, Automatic text summarization of biomedical text data: A systematic review, Information 13 (2022). URL: https://www.mdpi.com/2078-2489/13/8/393. doi:10.3390/info13080393.

[5] D. Keszthelyi, C. Gaudet-Blavignac, M. Bjelogrlic, C. Lovis, Patient information summarization in clinical settings: Scoping review, JMIR Med Inform. 28 (2023).

[6] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerova, et al., Clinical text summarization: adapting large language models can outperform human experts, Research square (2023) rs–3.

[7] E. Croxford, Y. Gao, N. Pellegrino, et al., Current and future state of evaluation of large language models for medical summarization tasks., npj Health Syst. 2 (2025).

[8] T. Goldsack, Z. Luo, Q. Xie, C. Scarton, M. Shardlow, S. Ananiadou, C. Lin, Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles, in: D. Demner-fushman, S. Ananiadou, K. Cohen (Eds.), Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 468–477. URL: https://aclanthology.org/2023.bionlp-1.44/. doi:10.18653/v1/2023.bionlp-1.44.

[9] T. Goldsack, C. Scarton, M. Shardlow, C. Lin, Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles, in: D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, J. Tsujii (Eds.), Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 122–131. URL: https://aclanthology.org/2024.bionlp-1.10/. doi:10.18653/v1/2024.bionlp-1.10.

[10] Y. Gao, D. Dligach, T. Miller, M. Afshar, Overview of the problem list summarization (ProbSum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes, in: D. Demner-fushman, S. Ananiadou, K. Cohen (Eds.), Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 461–467. URL: https://aclanthology.org/2023.bionlp-1.43/. doi:10.18653/v1/2023.bionlp-1.43.

[11] J.-B. Delbrouck, M. Varma, P. Chambon, C. Langlotz, Overview of the RadSum23 shared task on multi-modal and multi-anatomical radiology report summarization, in: D. Demner-fushman, S. Ananiadou, K. Cohen (Eds.), Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Lin-

guistics, Toronto, Canada, 2023, pp. 478–482. URL: https://aclanthology.org/2023.bionlp-1.45/. doi:10.18653/v1/2023.bionlp-1.45.

[12] L. L. Wang, J. DeYoung, B. Wallace, Overview of MSLR2022: A shared task on multi-document summarization for literature reviews, in: A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, M. Shmueli-Scheuer, A. de Waard, L. L. Wang (Eds.), Proceedings of the Third Workshop on Scholarly Document Processing, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 175–180. URL: https://aclanthology.org/2022.sdp-1.20/.

[13] 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, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), CLEF 2026 Working Notes, 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.

[14] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Giménez, L. Fernandez, E. Farré-Maduell, M. Krallinger, Overview of multiclinsum task at bioasq 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results, in: CLEF, 2025.

[15] Z. Zhao, Q. Jin, F. Chen, T. Peng, F. Yu, PMC-Patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems, https://huggingface.co/datasets/zhengyun21/PMC-Patients, 2022. HuggingFace dataset.

[16] Z. Zhao, Q. Jin, F. Chen, T. Peng, S. Yu, A large-scale dataset of patient summaries for retrieval-based clinical decision support systems, Scientific Data 10 (2023). URL: http://dx.doi.org/10.1038/s41597-023-02814-8. doi:10.1038/s41597-023-02814-8.

[17] J. G. Gilabert, X. Liao, S. D. Dalt, E. Bohman, A. Mash, F. D. L. Fornaciari, I. Baucells, J. Llop, M. C. Argote, C. Escolano, M. Melero, From salamandra to salamandrata: Bsc submission for wmt25 general machine translation shared task, 2025. URL: https://arxiv.org/abs/2508.12774. arXiv:2508.12774.

[18] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, A. F. T. Martins, CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task, in: P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, M. Zampieri (Eds.), Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 634–645. URL: https://aclanthology.org/2022.wmt-1.60/.

[19] I. García-Ferrero, R. Agerri, A. A. Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, A. Zaninello, Medical mt5: An open-source multilingual text-to-text llm for the medical domain, 2024. URL: https://arxiv.org/abs/2404.07613. arXiv:2404.07613.

[20] I. García-Ferrero, R. Agerri, A. A. Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, A. Zaninello, Medical mt5: An open-source multilingual text-to-text llm for the medical domain, 2024. URL: https://arxiv.org/abs/2404.07613. arXiv:2404.07613.

[21] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[22] A. Dubey, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[23] W. Kwon, Z. Li, S. Zhen, L. Zheng, C. Cen, H. Zhang, J. McCauley, J. E. Anderson, I. Stoica, J. E. Gonzalez, Efficient memory management for large language model serving with pagedattention,

in: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.
[24] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, 2021. URL: https://arxiv.org/abs/2010.11934. arXiv:2010.11934.
[25] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
[26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020.
[27] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.
[28] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. C. de Souza, A. Birch, A. F. T. Martins, Eurollm: Multilingual language models for europe, 2024. URL: https://arxiv.org/abs/2409.16235. arXiv:2409.16235.
[29] 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.
[30] A. Varela, A. Casillas, J. Viña, N. Lebeña, M. Oronoz, Multilingual clinical summarization with LLMs: A comparative study of GRPO and dynamic one-shot approaches, in: CLEF 2026 Working Notes, 2026.
[31] M. Roy, D. Das, NLP4Health at MultiClinSum-2 2026: Multilingual clinical case report summarization with qwen prompting and QLoRA, in: CLEF 2026 Working Notes, 2026.
[32] I. R. Abdullahi, I. Abdulmumin, How far can a lightweight multilingual model go on clinical summarization?, in: CLEF 2026 Working Notes, 2026.
[33] A. Oliveira, T. Rodrigues, C. Teixeira Lopes, InfoLab-FEUP at MultiClinSum-2: Agentic multilingual clinical case summarization with biomedical concept enrichment, in: CLEF 2026 Working Notes, 2026.
[34] R. Marin, M. Lacusteanu, R.-A. Iancu, RaRaMi at MultiClinSum-2: Cross-backbone QLoRA fine-tuning for romanian clinical case-report summarization, in: CLEF 2026 Working Notes, 2026.
[35] T. Bernal-Beltrán, J. Gómez-Navalón, R. Pan, P. J. Vivancos-Vicente, J. A. García-Díaz, UMUTeam at BioASQ 2026: Evaluating medical-domain continual pretraining for spanish clinical case report summarization, in: CLEF 2026 Working Notes, 2026.
[36] M. H. T. de Boer, Q. T. S. Smit, R. M. Bakker, MediScribes at MultiClinSUM2: A dutch-english clinical summarization system, in: CLEF 2026 Working Notes, 2026.
[37] P. Laban, T. Schnabel, P. N. Bennett, M. A. Hearst, Summac: Re-visiting nli-based models for inconsistency detection in summarization, 2021. URL: https://arxiv.org/abs/2111.09525. arXiv:2111.09525.
[38] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. URL: https://arxiv.org/abs/2111.09543. arXiv:2111.09543.
[39] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).
[40] I. Montani, M. Honnibal, Prodigy: A modern and scriptable annotation tool, 2018. URL: https://prodi.gy.

A. Per-language Leaderboard Tables

Table 6
Per-language leaderboard for English (EN), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
ixa-sum	2	**0.440 (0.220)**	**0.322**	**0.876**	0.732	0.832	0.700	0.698
ixa-sum	1	0.433 (0.216)	0.319	0.875	0.726	0.813	0.700	0.698
Llama3.1-8B-Ins (ZS)	1	0.398 (0.168)	0.272	0.873	0.763	0.921	0.700	0.500
Qwen2.5-7B-Ins (ZS)	1	0.387 (0.162)	0.268	0.871	0.766	0.931	0.701	0.699
ixa-sum	3	0.381 (0.155)	0.267	0.870	0.761	0.945	0.701	0.699
ixa-sum	5	0.412 (0.184)	0.284	0.869	0.729	0.862	0.700	**0.699**
NLP4Health	1	0.382 (0.155)	0.265	0.869	0.765	0.939	0.701	0.698
ixa-sum	4	0.372 (0.138)	0.255	0.868	0.769	**0.962**	**0.704**	0.699
MediScribes	3	0.373 (0.142)	0.249	0.866	**0.782**	**0.964**	0.702	0.697
MediScribes	1	0.349 (0.129)	0.230	0.864	0.766	0.919	0.700	0.690
NLP4Health	2	0.355 (0.144)	0.240	0.864	0.701	0.767	0.698	0.664
InfoLab-FEUP	1	0.353 (0.133)	0.227	0.862	0.758	0.885	0.700	0.694
InfoLab-FEUP	2	0.354 (0.133)	0.227	0.862	0.760	0.890	0.700	0.693
MediScribes	4	0.300 (0.107)	0.199	0.862	**0.804**	0.943	0.701	**0.706**
MCS-mT5-1epoch	1	0.356 (0.163)	0.242	0.862	0.526	0.501	0.481	0.469
InfoLab-FEUP	3	0.347 (0.127)	0.222	0.861	0.755	0.860	**0.707**	0.696
MCS-mT5-2epoch	1	0.354 (0.162)	0.241	0.860	0.497	0.480	0.461	0.456
MediScribes	2	0.321 (0.115)	0.217	0.859	0.719	0.776	0.698	0.661
NLP4Health	3	0.343 (0.116)	0.239	0.856	0.699	0.742	0.700	0.696
PsychAI-Discovery-Lab	1	0.332 (0.130)	0.262	0.854	0.771	0.893	0.698	0.662
Aleyna	1	0.328 (0.081)	0.215	0.853	0.669	0.597	0.693	0.608
Aleyna	2	0.325 (0.080)	0.217	0.852	0.669	0.596	0.695	0.631
MediScribes	5	0.221 (0.062)	0.159	0.846	0.733	0.710	0.700	0.702
DACHausa	1	0.338 (0.153)	0.243	0.837	0.638	0.529	0.681	0.483

Table 7
Per-language leaderboard for Spanish (ES), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
ixa-sum	2	**0.467 (0.235)**	0.314	**0.882**	0.713	0.796	0.700	**0.649**
ixa-sum	1	0.463 (0.232)	**0.314**	0.881	0.709	0.786	0.700	0.647
Llama3.1-8B-Ins (ZS)	1	0.439 (0.205)	0.289	0.876	0.728	0.858	0.700	0.345
NLP4Health	1	0.406 (0.176)	0.273	0.873	0.722	0.839	**0.700**	0.630
Qwen2.5-7B-Ins (ZS)	1	0.412 (0.180)	0.274	0.871	0.732	0.873	0.700	0.642
ixa-sum	3	0.401 (0.171)	0.271	0.871	0.733	0.894	**0.701**	0.652
ixa-sum	4	0.398 (0.155)	0.257	0.870	**0.750**	**0.920**	**0.701**	**0.655**
MCS-mT5-1epoch	1	0.403 (0.179)	0.258	0.868	0.571	0.500	0.536	0.404
MCS-mT5-2epoch	1	0.403 (0.181)	0.259	0.868	0.555	0.490	0.508	0.388
ixa-sum	5	0.436 (0.206)	0.293	0.868	0.704	0.602	0.700	0.609
NLP4Health	2	0.383 (0.164)	0.257	0.866	0.678	0.634	0.698	0.370
PsychAI-Discovery-Lab	1	0.392 (0.164)	0.280	0.864	0.712	0.761	0.698	0.554
UMU team	1	0.397 (0.177)	0.261	0.862	0.703	0.602	0.685	0.450
UMU team	3	0.401 (0.182)	0.266	0.862	0.695	0.591	0.684	0.448
UMU team	2	0.395 (0.175)	0.259	0.861	0.704	0.602	0.688	0.444
NLP4Health	3	0.335 (0.098)	0.221	0.845	0.632	0.575	0.700	0.529
DACHausa	1	0.372 (0.159)	0.254	0.837	0.630	0.507	0.690	0.361
UMU team	4	0.252 (0.099)	0.161	0.592	0.692	0.619	0.685	0.394

Table 8
Per-language leaderboard for Catalan (CA), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
ixa-sum	2	**0.456 (0.232)**	0.301	**0.882**	0.705	0.867	**0.700**	**0.627**
ixa-sum	1	**0.458 (0.233)**	**0.306**	0.881	0.702	0.833	**0.700**	0.640
Llama3.1-8B-Ins (ZS)	1	0.430 (0.202)	0.278	0.873	**0.711**	**0.891**	0.699	0.447
ixa-sum	3	0.402 (0.172)	0.264	0.871	0.705	0.872	**0.700**	0.640
Qwen2.5-7B-Ins (ZS)	1	0.405 (0.181)	0.267	0.867	0.710	**0.890**	**0.700**	0.626
MCS-mT5-1epoch	1	0.398 (0.178)	0.249	0.867	0.569	0.584	0.529	0.411
MCS-mT5-2epoch	1	0.395 (0.176)	0.247	0.866	0.539	0.582	0.456	0.366
ixa-sum	5	0.407 (0.186)	0.263	0.866	0.697	0.726	0.699	0.553
ixa-sum	4	0.375 (0.150)	0.239	0.848	**0.714**	**0.873**	0.693	0.613
DACHausa	1	0.367 (0.160)	0.247	0.840	0.658	0.689	0.694	0.338
PsychAI-Discovery-Lab	1	0.262 (0.096)	0.195	0.838	0.706	0.811	0.688	0.449

Table 9
Per-language leaderboard for French (FR), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
ixa-sum	2	0.415 (0.172)	0.256	**0.873**	**0.742**	**0.904**	**0.702**	**0.667**
ixa-sum	1	0.410 (0.180)	0.263	0.872	0.727	0.866	0.701	0.654
Llama3.1-8B-Ins (ZS)	1	**0.432 (0.201)**	**0.274**	0.871	0.728	0.809	0.697	0.477
NLP4Health	1	0.405 (0.178)	0.261	0.871	0.724	0.828	0.700	0.646
Qwen2.5-7B-Ins (ZS)	1	0.422 (0.190)	0.269	0.868	0.731	0.850	0.700	0.649
MCS-mT5-2epoch	1	0.399 (0.175)	0.245	0.866	0.572	0.502	0.540	0.373
MCS-mT5-1epoch	1	0.397 (0.172)	0.243	0.866	0.582	0.491	0.552	0.386
NLP4Health	2	0.356 (0.142)	0.224	0.860	0.675	0.634	0.688	0.339
DACHausa	1	0.370 (0.159)	0.245	0.841	0.649	0.519	0.694	0.312
NLP4Health	3	0.329 (0.101)	0.210	0.839	0.632	0.575	0.700	0.444

Table 10
Per-language leaderboard for Portuguese (PT), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
NLP4Health	1	0.384 (0.156)	0.257	**0.870**	0.738	0.891	**0.701**	0.636
ixa-sum	1	0.381 (0.152)	0.256	0.870	0.741	**0.915**	0.700	**0.652**
Llama3.1-8B-Ins (ZS)	1	**0.409 (0.178)**	**0.270**	0.869	0.746	0.884	0.700	0.418
MCS-mT5-1epoch	1	0.377 (0.159)	0.243	0.865	0.574	0.484	0.540	0.376
MCS-mT5-2epoch	1	0.377 (0.160)	0.244	0.865	0.560	0.487	0.524	0.384
Qwen2.5-7B-Ins (ZS)	1	0.388 (0.161)	0.259	0.865	0.754	0.899	0.700	0.649
NLP4Health	2	0.356 (0.141)	0.238	0.864	0.676	0.651	0.697	0.372
InfoLab-FEUP	3	0.361 (0.135)	0.229	0.863	0.744	0.827	0.700	0.575
InfoLab-FEUP	1	0.358 (0.132)	0.227	0.862	0.745	0.842	0.700	0.585
InfoLab-FEUP	2	0.336 (0.104)	0.212	0.858	0.727	0.797	0.699	0.546
ixa-sum	2	0.355 (0.132)	0.229	0.847	**0.770**	0.871	0.691	0.606
PsychAI-Discovery-Lab	1	0.244 (0.080)	0.181	0.835	0.697	0.597	0.677	0.437
DACHausa	1	0.343 (0.141)	0.238	0.835	0.609	0.472	0.685	0.351

Table 11
Per-language leaderboard for Dutch (NL), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
Llama3.1-8B-Ins (ZS)	1	**0.374 (0.145)**	**0.255**	**0.865**	0.717	0.881	0.699	0.430
ixa-sum	2	0.350 (0.115)	0.236	0.862	0.732	0.949	**0.712**	0.665
MediScribes	3	0.336 (0.109)	0.223	0.862	0.747	**0.983**	0.708	0.669
ixa-sum	1	0.353 (0.122)	0.239	0.862	0.715	0.908	0.700	0.659
Qwen2.5-7B-Ins (ZS)	1	0.355 (0.129)	0.243	0.860	0.721	0.887	**0.701**	0.646
MediScribes	4	0.227 (0.073)	0.164	0.858	**0.755**	0.938	0.700	**0.699**
MCS-mT5-2epoch	1	0.342 (0.135)	0.226	0.855	0.551	0.487	0.545	0.391
MCS-mT5-1epoch	1	0.343 (0.134)	0.226	0.855	0.558	0.477	0.554	0.424
MediScribes	1	0.135 (0.037)	0.093	0.845	0.736	0.939	**0.701**	0.666
MediScribes	5	0.233 (0.055)	0.164	0.843	0.700	0.680	0.699	**0.667**
MediScribes	2	0.131 (0.034)	0.093	0.841	0.695	0.740	0.697	0.589
DACHausa	1	0.317 (0.121)	0.223	0.826	0.620	0.444	0.685	0.346

Table 12
Per-language leaderboard for Italian (IT), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
Llama3.1-8B-Ins (ZS)	1	**0.379 (0.155)**	**0.251**	**0.871**	0.720	0.865	**0.701**	0.550
ixa-sum	1	0.358 (0.136)	0.242	0.870	0.718	0.890	0.700	**0.647**
MCS-mT5-1epoch	1	0.346 (0.142)	0.225	0.866	0.030	0.513	0.048	0.000
MCS-mT5-2epoch	1	0.346 (0.142)	0.225	0.866	0.563	0.491	0.552	0.355
Qwen2.5-7B-Ins (ZS)	1	0.357 (0.136)	0.238	0.862	0.722	**0.881**	0.700	0.643
PsychAI-Discovery-Lab	1	0.327 (0.121)	0.240	0.860	0.705	0.786	0.697	0.559
ixa-sum	2	0.343 (0.117)	0.223	0.858	**0.725**	**0.903**	**0.701**	0.635
DACHausa	1	0.316 (0.128)	0.220	0.835	0.592	0.445	0.693	0.327

Table 13
Per-language leaderboard for Romanian (RO), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
Llama3.1-8B-Ins (ZS)	1	**0.383 (0.166)**	**0.262**	**0.874**	0.738	**0.924**	0.699	0.431
ixa-sum	1	0.358 (0.146)	0.251	0.873	0.737	0.911	**0.700**	0.569
MCS-mT5-2epoch	1	0.355 (0.152)	0.238	0.869	0.565	0.569	0.529	0.326
MCS-mT5-1epoch	1	0.353 (0.152)	0.236	0.868	0.565	0.538	0.540	0.330
Qwen2.5-7B-Ins (ZS)	1	0.370 (0.156)	0.253	0.866	0.747	0.915	**0.700**	0.565
RaRaMi	1	0.339 (0.150)	0.241	0.861	0.700	0.677	0.683	0.304
ixa-sum	2	0.337 (0.129)	0.226	0.857	**0.757**	0.902	0.693	**0.583**
DACHausa	1	0.335 (0.145)	0.239	0.845	0.634	0.604	0.691	0.294

Table 14
Per-language leaderboard for German (DE), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
ixa-sum	1	0.297 (0.094)	**0.205**	**0.856**	0.721	0.827	**0.699**	**0.652**
MCS-mT5-2epoch	1	**0.307 (0.111)**	0.199	0.854	0.544	0.481	0.469	0.388
MCS-mT5-1epoch	1	0.306 (0.110)	0.198	0.854	0.566	0.481	0.506	0.401
ixa-sum	2	0.297 (0.088)	0.194	0.848	**0.750**	**0.861**	0.693	0.634
Llama3.1-8B-Ins (ZS)	1	0.326 (0.115)	0.216	0.842	0.722	0.817	**0.700**	0.412
Qwen2.5-7B-Ins (ZS)	1	0.310 (0.102)	0.206	0.835	0.728	0.845	0.700	0.640
DACHausa	1	0.272 (0.097)	0.191	0.822	0.625	0.456	0.691	0.340

Table 15
Per-language leaderboard for Russian (RU), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
MCS-mT5-2epoch	1	0.259 (0.069)	0.251	**0.860**	0.560	0.442	0.202	0.293
MCS-mT5-1epoch	1	0.248 (0.063)	0.239	0.859	0.550	0.428	0.343	0.296
PsychAI-Discovery-Lab	1	0.291 (0.066)	0.285	0.858	**0.705**	**0.575**	**0.697**	0.387
Llama3.1-8B-Ins (ZS)	1	**0.393 (0.096)**	**0.384**	0.845	0.710	0.681	**0.700**	0.386
DACHausa	1	0.319 (0.073)	0.312	0.829	0.628	0.425	0.677	0.268
Qwen2.5-7B-Ins (ZS)	1	0.321 (0.081)	0.309	0.810	**0.722**	**0.778**	**0.700**	**0.483**

Table 16
Per-language leaderboard for Czech (CS), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
Llama3.1-8B-Ins (ZS)	1	**0.355 (0.129)**	**0.220**	**0.866**	0.699	0.865	0.681	0.279
MCS-mT5-2epoch	1	0.333 (0.118)	0.200	0.861	0.543	0.476	0.166	0.306
MCS-mT5-1epoch	1	0.333 (0.117)	0.199	0.861	0.567	**0.483**	0.255	0.340
Qwen2.5-7B-Ins (ZS)	1	0.273 (0.091)	0.171	0.838	**0.713**	**0.867**	**0.700**	**0.562**
DACHausa	1	0.299 (0.102)	0.192	0.835	0.632	0.453	0.678	0.288

Table 17
Per-language leaderboard for Danish (DA), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
Llama3.1-8B-Ins (ZS)	1	**0.347 (0.132)**	**0.238**	**0.859**	0.722	0.860	0.651	0.306
MCS-mT5-1epoch	1	0.326 (0.132)	0.214	0.857	0.544	0.507	0.524	0.387
MCS-mT5-2epoch	1	0.325 (0.131)	0.213	0.857	0.537	0.532	0.518	0.373
Qwen2.5-7B-Ins (ZS)	1	0.335 (0.126)	0.228	0.856	**0.731**	**0.873**	**0.700**	**0.657**
DACHausa	1	0.294 (0.118)	0.208	0.826	0.631	0.582	0.680	0.354

Table 18
Per-language leaderboard for Greek (EL), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
MCS-mT5-2epoch	1	0.262 (0.069)	0.254	**0.868**	0.611	0.562	0.592	0.328
MCS-mT5-1epoch	1	0.264 (0.070)	0.256	0.868	0.104	0.560	0.046	0.000
DACHausa	1	0.309 (0.070)	0.304	0.847	0.654	0.578	**0.695**	0.282
Llama3.1-8B-Ins (ZS)	1	**0.343 (0.092)**	**0.330**	0.829	0.697	**0.844**	**0.700**	0.323
Qwen2.5-7B-Ins (ZS)	1	0.298 (0.080)	0.287	0.774	**0.707**	0.818	**0.700**	**0.460**

Table 19
Per-language leaderboard for Norwegian (NO), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
Llama3.1-8B-Ins (ZS)	1	**0.351 (0.134)**	**0.244**	**0.862**	0.709	0.834	0.591	0.107
MCS-mT5-2epoch	1	0.329 (0.134)	0.221	0.857	0.546	0.523	0.521	0.382
MCS-mT5-1epoch	1	0.328 (0.133)	0.219	0.857	0.553	0.505	0.526	**0.385**
Qwen2.5-7B-Ins (ZS)	1	0.337 (0.121)	0.231	0.856	**0.714**	**0.847**	**0.700**	**0.648**

Table 20
Per-language leaderboard for Swedish (SV), ranked by BERTScore (judge metrics: EuroLLM). R-1 = ROUGE-1, R-2 = ROUGE-2, R-L = ROUGE-Lsum, Faith. = Faithfulness, Compl. = Completeness, Fluen. = Fluency, Consist. = Consistency.

Team Name	Run	R-1 (R-2)	R-L	BERTScore	Faith.	Compl.	Fluen.	Consist.
Llama3.1-8B-Ins (ZS)	1	**0.345 (0.133)**	**0.233**	**0.861**	0.715	0.891	0.699	0.354
PsychAI-Discovery-Lab	1	0.327 (0.115)	0.227	0.858	0.710	0.779	0.697	0.609
Qwen2.5-7B-Ins (ZS)	1	0.333 (0.125)	0.226	0.856	**0.720**	**0.898**	**0.700**	**0.665**
MCS-mT5-1epoch	1	0.324 (0.131)	0.211	0.855	0.558	0.525	0.530	0.407
MCS-mT5-2epoch	1	0.322 (0.131)	0.211	0.855	0.538	0.520	0.524	0.395
DACHausa	1	0.295 (0.118)	0.206	0.828	0.640	0.576	0.692	0.364

B. Generated Summary Examples Across Language Sub-tracks

B.1. Catalan (CA)

Full Case (Catalan)

Una dona de 60 anys es va sotmetre a una resecció total del tumor d'un meningioma del lòbul frontal dret que es trobava a la convexitat del cervell el setembre de 2018, i el diagnòstic patològic va ser un meningioma atípic. L'octubre de 2021, la pacient va ser ingressada a l'hospital amb tos i dificultat per respirar. Una tomografia computada (TC) va revelar la presència de grans masses a la cavitat toràcica i abdominal dreta, i no es va trobar cap recurrència en la ressonància magnètica craniocerebral (MRI). Les característiques histològiques del tumor pulmonar eren similars a les de la massa cerebral, i les cèl·lules tumorals eren positives per vimentina, EMA i Ki-67, i negatives per TTF1, PD-L1, P40 i cromogranina. Basant-se en aquestes troballes de la biòpsia i la immunohistoquímica (IHC), les masses es van identificar com meningiomes metastàtics. La pacient va ser tractada amb l'agent anti-PD-1 camrelizumab (200 mg, Dia 1) combinat amb l'agent anti-VEGF anlotinib (10 mg, Dies 1-14) cada 3 setmanes sense tractament de radiació. Després de dos cicles d'aquest règim, els símptomes de la pacient es van resoldre completament sense altres esdeveniments adversos, i la TC va revelar que el tumor s'havia encongit significativament en > 80%. La biòpsia realitzada després dels dos cicles de tractament i l'anàlisi d'immunohistoquímica van revelar que la infiltració de limfòcits T CD4+ i CD8+ i macròfags CD68+ en el microentorn del tumor havia augmentat significativament en comparació amb abans del tractament. El nombre de limfòcits T CD4+ i CD8+ a la sang perifèrica va continuar augmentant a mesura que el tumor es va encongir. Aquestes troballes indiquen que el tractament combinat anti-PD-1 i anti-VEGF estimula les cèl·lules immunitàries de la sang perifèrica per a matar les cèl·lules de meningioma metastàtic. La pacient ha rebut onze cicles del tractament cada 3 setmanes des del 21 d'octubre de 2021 fins al 21 de juny de 2022 sense progressió del tumor.

Reference Summary

Una dona de 60 anys va acceptar la resecció total del tumor d'un meningioma del lòbul frontal dret el setembre de 2018. L'octubre de 2021, la pacient va ser ingressada a l'hospital a causa de la tos i la manca d'aire i se li va diagnosticar meningiomes metastàtics. La tomografia computada (TC) va revelar la presència de grans masses a la cavitat toràcica i abdominal dreta. Després de dos cicles de tractament anti-PD-1 i anti-VEGF, els símptomes es van alleujar i el tumor es va necrotitzar. Seguiment fins al 21 de juny de 2022, la pacient ha rebut onze cicles del tractament cada 3 setmanes sense progressió del tumor.

Generated Summary – Qwen-2.5-Ins (ZS) | BERTScore 0.910

Una dona de 60 anys va rebre una resecció de meningioma atípic del lòbul frontal dret en setembre de 2018. En octubre de 2021, va ser ingressada amb tos i dificultat respiratriu. Una TC va revelar grans masses a la cavitat toràcica i abdominal dreta, identificades com meningiomes metastàtics basant-se en característiques histològiques i IHC similars als del meningioma cerebral anterior. Va rebre tractament amb camrelizumab i anlotinib sense radiació, i després de dos cicles, els símptoms es van resoldre i el tumor s'havia encongit significativament. L'anàlisi d'immunohistoquímica va mostrar augment de limfòcits T CD4+ i CD8+ i macròfags CD68+ en el microentorn del tumor. La pacient ha rebut onze cicles del tractament sense progressió del tumor fins al 21 de juny de 2022.

B.2. English (EN)

Full Case (English)

A 20-year-old man presented with mild dyspnea and was diagnosed with a common cold at the first visit to a clinic. However, severe dyspnea persisted after 3 days. Enhanced computed tomography (CT) scan revealed a mediastinal mass in the subcarinal space, which compressed the right pulmonary artery and a delayed right pulmonary vein and airway enhancement at the tracheal bifurcation (-A, B). The mass was diagnosed as a mediastinal cyst and bronchoscopy was attempted; however, the patient was unable to be placed in the supine position due to severe respiratory distress. The patient was then transported to our hospital by ambulance due to the need for emergent surgery. Despite the fact that transportation took only two hours, his white blood cell count increased rapidly and symptoms became progressively worse, which suggested that the patient's condition was emergent. Signed consent was obtained for all procedures. To ensure that the mass was a simple cyst without septum and mural nodule, magnetic resonance imaging (MRI) was performed as the patient's condition allowed; the image showed a two-layered simple cyst, which indicated an infection or bleeding inside the cyst (-C, D). Since the symptoms developed and exacerbated rapidly, we performed emergent surgery. After the induction of general anesthesia with "stand-by" extracorporeal membrane oxygenation, video-assisted thoracic surgery (VATS) was performed. First, we punctured the cyst and aspirated white pus; then, we performed cyst wall fenestration on the subcarinal lesion and superior mediastinum. The bottom layer fluid inside the cyst contained both pus and blood, which were compatible with MRI findings, while continuous bleeding from the cyst wall was not observed (). Performing complete cyst resection was difficult owing to severe adhesion of the cyst to surrounding organs such as both main bronchi and pericardium. After surgery, the symptoms resolved immediately and completely. The postoperative course was uneventful and the patient was discharged on the 15th postoperative day. Pathological examination of the cyst wall revealed an inflamed bronchogenic cyst with the findings of bronchial gland, cartilage with infiltration of inflammatory cells, and no malignancy (). Six months after the operation, no sign of recurrence was observed.

Reference Summary

A 20-year-old man presented with severe dyspnea. Imaging revealed a mediastinal cyst in the subcarinal space compressing his right pulmonary artery and airway, which was later diagnosed as a bronchogenic cyst. Due to quick symptom exacerbation, emergent cyst wall fenestration was performed through video-assisted thoracic surgery with "stand-by" extracorporeal membrane oxygenation. Complete cyst resection was difficult owing to adhesion of the cyst to the surrounding organs. The symptoms immediately resolved after surgery and the postoperative course was uneventful.

Generated Summary – InfoLab-FEUP | BERTScore: 0.900

A 20-year-old male was diagnosed with an inflamed bronchogenic cyst in the mediastinum. Treatment consisted of attempted bronchoscopy, emergent surgery, induction of general anesthesia, "stand-by" extracorporeal membrane oxygenation, video-assisted thoracic surgery (VATS), puncture and aspiration of cyst contents, and cyst wall fenestration on the subcarinal lesion and superior mediastinum. The patient underwent complete cyst resection, which was difficult due to severe adhesions. Following surgery, symptoms resolved immediately and completely, with no signs of recurrence observed six months postoperatively.

B.3. Spanish (ES)

Full Case (Spanish)

Un hombre caucásico de 75 años de edad con antecedentes de artritis reumatoide e insuficiencia cardíaca congestiva se presentó con un absceso abdominal subcutáneo recurrente con drenaje espontáneo de pus. Se le realizó una colecistectomía laparoscópica por colecistitis aguda calculosa hace 3 años. El examen físico reveló enrojecimiento y sensibilidad periumbilical con un tracto percutáneo drenante. Los análisis de laboratorio revelaron un recuento de glóbulos blancos de 13 900 por milímetro cúbico. Una tomografía computarizada (TC) mostró una cavidad en la pared abdominal periumbilical con contraste periférico, junto a un cuerpo extraño calcificado entre las vainas del músculo recto. Se realizó una exploración de la herida bajo anestesia general con drenaje de la cavidad, extracción del cuerpo extraño y cierre de la hoja rectal anterior sobre un catéter de drenaje. En el examen patológico, el cuerpo extraño resultó ser un cálculo biliar. Se perdió en el puerto periumbilical durante el procedimiento en el que se produjo la perforación de la vesícula biliar. Se encontraron bacterias de E. coli en la matriz microbiológica. Se continuó el tratamiento antibiótico con coamoxiclav durante 14 días. El paciente fue dado de alta 9 días después de la operación con una herida limpia, y continuó visitando la clínica ambulatoria para el seguimiento de la herida durante 8 semanas. El seguimiento fue sin incidentes.

Reference Summary

Un hombre caucásico de 75 años de edad con antecedentes de artritis reumatoide e insuficiencia cardiaca congestiva presentó un absceso recurrente subcutáneo en la pared abdominal con drenaje espontáneo de pus. Se le realizó una colecistectomía laparoscópica por colecistitis calculosa aguda hace 3 años con una recuperación sin incidentes y rápida. Una tomografía computarizada mostró una cavidad en la pared abdominal peri-umbilical con contraste periférico, junto a un cuerpo extraño calcificado entre las láminas del músculo recto. Se realizó una exploración de la herida bajo anestesia general con drenaje de la cavidad, extracción del cuerpo extraño y cierre de la lámina rectal anterior sobre un catéter de drenaje. El cuerpo extraño resultó ser un cálculo biliar que se había perdido en el puerto peri-umbilical durante el procedimiento. Se continuó el tratamiento antibiótico con co-amoxiclav durante 14 días. El paciente fue dado de alta 9 días después de la operación con una herida limpia.

Generated Summary – UMUTeam | BERTScore: 0.953

Un hombre caucásico de 75 años de edad con antecedentes de artritis reumatoide e insuficiencia cardíaca congestiva se presentó con un absceso abdominal subcutáneo recurrente con drenaje espontáneo de pus. Se le realizó una colecistectomía laparoscópica por colecistitis aguda calculosa hace 3 años. El examen físico reveló enrojecimiento y sensibilidad periumbilical con un tracto percutáneo drenante. Los análisis de laboratorio revelaron un recuento de glóbulos blancos de 13 900 por milímetro cúbico. Una tomografía computarizada mostró una cavidad en la pared abdominal periumbilical con contraste periférico, junto a un cuerpo extraño calcificado entre las vainas del músculo recto. Se realizó una exploración de la herida bajo anestesia general con drenaje de la cavidad, extracción del cuerpo extraño y cierre de la hoja rectal anterior sobre un catéter de drenaje. En el examen patológico

B.4. French (FR)

Full Case (French)

Une femme de 87 ans a présenté une paraplégie progressive (c'est-à-dire jambe gauche 3/5 et jambe droite 2/5) de 5 jours. La patiente a commencé un traitement par Rocephin pour une infection des voies urinaires, mais une angiographie par tomodensitométrie de la tête et du cou et une imagerie par résonance magnétique de la colonne thoracique ont montré une maladie métastatique osseuse étendue avec une tumeur épidurale à T4-T7 avec une compression sévère de la moelle épinière []. Elle a subi une laminectomie T3-T8 pour la résection de la tumeur. Préopérativement, elle a reçu une immunoglobuline intraveineuse, une transfusion de plaquettes, du plasma frais congelé et de la vitamine K pour normaliser son rapport international normalisé (RIN). La pathologie a révélé un sarcome myéloïde avec différenciation monocytaire et coloration immunohistochimique CD68-positive [et]. Postopérativement, elle a retrouvé une certaine force dans ses membres inférieurs. Cependant, une fois qu'une biopsie de la moelle osseuse a révélé une atteinte diffuse avec sarcome myéloïde et différenciation monocytaire, elle a décidé de ne pas poursuivre le traitement (c'est-à-dire la chimiothérapie ou la radiothérapie) et est décédée 4 mois plus tard.

Reference Summary

Une femme de 87 ans a présenté une paraplégie progressive de 5 jours. L'imagerie par résonance magnétique (IRM) a révélé une tumeur épidurale de T4 à T7 avec compression de la moelle épinière. Lorsqu'elle a subi une laminectomie pour la résection de la tumeur, la pathologie a révélé un sarcome myéloïde avec différenciation monocytaire. Bien qu'elle se soit améliorée après l'opération, elle a choisi de poursuivre les soins palliatifs et est décédée 4 mois plus tard.

Generated Summary – ixa-sum | BERTScore: 0.910

Résumé : Une femme de 87 ans a été hospitalisée pour une paraplégie progressive due à une tumeur épidurale à T4-T7. Après une laminectomie T3-T8, elle a retrouvé une certaine force dans ses membres inférieurs. Cependant, une biopsie de la moelle osseuse a révélé une atteinte diffuse par sarcome myéloïde différencié monocytaire, conduisant à une décision de ne pas poursuivre le traitement. La patiente est décédée 4 mois plus tard.

B.5. Portuguese (PT)

Full Case (Portuguese)

Um homem de 66 anos de idade, sem histórico médico anterior, foi admitido no ED para um primeiro episódio de dor epigástrica grave de início rápido. Apareceu em repouso sem sintomas associados, tais como dispneia ou vômito. Ele relatou tomar anti-inflamatórios não esteróides durante 1 semana antes da admissão para uma epicondilite. Na admissão, a sua dor tinha diminuído rapidamente e o exame físico foi inconclusivo: pressão arterial, temperatura, frequência cardíaca e enchimento capilar foram normais. A auscultação torácica e a palpação abdominal revelaram apenas uma ligeira dor epigástrica. O ECG e o ensaio de troponina foram normais, assim como a hemoglobina, a bilirrubina, as enzimas hepáticas e os níveis de lipase. Como parte da política do nosso departamento, foi realizado um POCUS essencialmente para excluir a colelitíase. Foi realizado com um Philips CX50 com uma sonda abdominal de 3,5-5 MHz. Não havia cálculos biliares; no entanto, o POCUS revelou uma massa localizada entre o baço e o rim esquerdo com medidas de 80 * 74 mm. A massa era vascular com um sinal de Doppler de cor forte e um trombo intraluminal que explica porque o sinal de Doppler não preenche todo o lúmen. A velocidade máxima no aneurisma foi de apenas 18 cm/s devido ao grande diâmetro. A aorta foi explorada do diafragma à bifurcação. Não foi encontrado outro aneurisma. Pequenas quantidades de fluido peritoneal livre foram detetadas em torno do baço. Tomografia computadorizada urgente diagnosticou um aneurisma gigante fissuroso da artéria esplênica. Uma angiografia seletiva foi então realizada sob anestesia geral; uma exclusão completa do aneurisma foi obtida com embolização de cola/lipiodol. O curso pós-intervenção foi sem intercorrências e o paciente foi dispensado para casa 3 dias depois. O paciente permaneceu livre de quaisquer complicações da embolização 6 meses após o procedimento.

Reference Summary

Um homem de 66 anos de idade, sem histórico médico anterior, foi admitido no Serviço de Urgência por uma dor epigástrica de início rápido. Ele relatou ter tomado anti-inflamatórios não esteróides durante 1 semana antes da admissão. A sua dor tinha diminuído rapidamente e o exame físico foi inconclusivo. O ECG e as amostras de sangue foram normais. O POCUS revelou uma massa vascular localizada entre o baço e o rim esquerdo com 80 * 74 mm associada a pequenas quantidades de fluido peritoneal livre. A tomografia computadorizada diagnosticou um aneurisma gigante fissuroso da artéria esplénica. O aneurisma foi gerido emergentemente por exclusão endovascular por embolização seletiva da artéria esplénica. O curso pós-intervenção foi sem incidentes e o paciente foi dispensado em casa 3 dias depois. O paciente permaneceu livre de quaisquer complicações da embolização 6 meses após o procedimento.

Generated Summary– NLP4Health | BERTScore: 0.910

Um homem de 66 anos apresentou-se com dor epigástrica grave, sem sinais de gravidade clínica. Após exames laboratoriais e ecocardiograma de ponto de cuidado essencial, um POCUS suspeitou de uma massa entre o baço e o rim esquerdo, que foi confirmada por tomografia computadorizada como um aneurisma gigante fissuroso da artéria esplênica. Após angiografia e embolização com cola/lipiodol, o paciente teve um curso pós-intervenção sem complicações e foi dispensado 3 dias após o procedimento, mantendo-se bem 6 meses após a intervenção.

B.6. Italian (IT)

Full Case (Italian)

Una donna indiana di 50 anni affetta da diabete e ipertensione si è presentata con un disagio addominale superiore destro, perdita di appetito e febbre di basso grado. Non aveva precedenti di ittero. L'esame ha rivelato un'epatomegalia e una dolorabilità nell'ipocondrio destro. Un'ecografia addominale ha rivelato una lesione di 11 × 7,5 × 11,2 cm, prevalentemente anecoica, a parete spessa, derivante dalla superficie inferiore del fegato, con pochi echi interni di basso livello e un'area ecogenica omogenea nella porzione superolaterale della lesione. La cistifellea non poteva essere visualizzata separatamente e le altre viscere addominali erano normali. I suoi test di funzionalità epatica erano normali e i test sierologici per amebiasi e idatidosi erano negativi. La sua radiografia del torace era normale. È stata fatta una diagnosi di ascesso epatico e l'aspirazione della lesione ha rivelato un fluido torbido. Tuttavia, la lesione non si è risolta con ripetute aspirazioni e antibiotici. Una tomografia computerizzata a contrasto (CECT) ha rivelato una lesione cistica lobata ben definita con una componente solida nella parte superiore della lesione. La figura e la figura mostrano la vicinanza della lesione al colon e al duodeno. La lesione è stata interpretata come una cisti non parassitaria infetta e sono state effettuate l'evacuazione e la rimozione della cisti. L'esame istopatologico della parete della cisti escissa ha rivelato un cistoadenoma biliare con atipia nucleare. Successivamente, è stata effettuata un'asportazione radicale della lesione con un'asportazione parziale del fegato. La diagnosi istologica finale è stata di cistadenocarcinoma infiltrante del fegato.

Reference Summary

Una donna indiana di 50 anni si è presentata con dolore e gonfiore nell'ipocondrio destro. Un'ecografia ha rivelato una lesione cistica derivante dalla fossa della cistifellea. Questa lesione è stata inizialmente gestita con aspirazione e antibiotici dal medico curante. La paziente è stata sottoposta a trattamento chirurgico perché l'ascesso non si è risolto con un trattamento conservativo. È stata fatta una diagnosi di cisti infetta non parassitaria e la cisti è stata rimossa. Un esame istopatologico della parete della cisti asportata ha mostrato un cistadenocarcinoma. La paziente è stata successivamente sottoposta a un'asportazione chirurgica della lesione, che si è rivelata un successo.

Generated Summary – ixa-sum | BERTScore: 0.911

Una donna indiana di 50 anni con diabete e ipertensione si è presentata con un disagio addominale superiore destro, perdita di appetito e febbre di basso grado. L'esame ha rivelato un'epatomegalia e dolorabilità nell'ipocondrio destro. Un'ecografia addominale ha identificato una lesione cistica lobata con componente solida, interpretata come un ascesso epatico. La lesione non si è risolta con antibiotici e aspirazioni, quindi è stata effettuata una CECT che ha confermato una cisti non parassitaria infetta. La cisti è stata evaccata e rimossa chirurgicamente, rivelando un cistoadenoma biliare con atipia nucleare. La diagnosi finale istopatologica è stata di cistadenocarcinoma infiltrante del fegato, con asportazione radicale della lesione e parziale del fegato.

B.7. Dutch (NL)

Full Case (Dutch)

Een 15-jarige jongen werd met een voorgeschiedenis van acute lage rugpijn en zwakte van de onderste ledematen naar de spoedgevallendienst verwezen na het optillen van een zwaar Indiaas muziekinstrument (Dholak). De patiënt had urine-retentie ontwikkeld sinds het moment van de verwonding. De patiënt had gevoelloosheid, tintelingen in beide onderste ledematen en was niet in staat om te staan of te lopen. Onderzoek onthulde Grade 0 kracht van bilaterale enkel dorsiflexoren, extensor hallucis longus en extensor digitorum. Hij had Grade 3 kracht van bilaterale enkel plantar flexoren. Een enkel schrikreflex was afwezig. Hij had verminderde sensaties in L5 en sacrale dermatomen. Bulbocavernosus reflex was afwezig. Een spoed MRI werd uitgevoerd, die een breuk van de ledematen van de L4-wervel met een verplaatst fragment in het ruggenmerg onthulde, waardoor de cauda equina ernstig werd gecomprimeerd (en). Spoedchirurgie werd uitgevoerd in de vorm van een microscopische decompressie van de middenlijn van de L3-L4. Alle losse en uitgestoten schijffragmenten werden verwijderd. Een bot bult werd teruggeduwd naar de positie met behulp van een hockeystick. De patiënt vertoonde tekenen van neurologisch herstel in de onmiddellijke post-operatieve periode. De patiënt had binnen 1 maand na de operatie een volledig neurologisch herstel inclusief blaascontrole. Hij had geen functionele tekorten tijdens de follow-up na 3-6 maanden. Dynamische röntgenfoto's aan het einde van 6 maanden vertoonden geen enkel teken van instabiliteit ().

Reference Summary

Een 15-jarige jongen werd met een voorgeschiedenis van acute lage rugpijn en bilaterale radiculaire pijn in de onderste ledematen met zwakte, na het optillen van een zwaar Indiaas muziekinstrument (Dholak), voor noodgevallen naar de spoedgevallendienst verwezen. Onderzoek toonde L5- en S1-zwakte met bilaterale afwezigheid van enkelreflexen. De bulbocavernosusreflex was afwezig. Er werd een MRI-scan uitgevoerd, waaruit een fractuur van de cephalad-kant van het L4-wervellichaam met een verplaatst fragment in het ruggenmergkanaal bleek, waardoor de cauda equina werd samengedrukt. Er werd noodchirurgie uitgevoerd in de vorm van een microscopische decompressie van de L3-L4-middenlijn. De patiënt herstelde binnen een maand na de operatie volledig op neurologisch vlak, inclusief de blaascontrole. De patiënt vertoonde geen functionele tekorten tijdens de follow-up na 3 en 6 maanden. Dynamische röntgenfoto's aan het einde van 6 maanden toonden geen enkel teken van instabiliteit.

Generated Summary – MediScribes | BERTScore: 0.891

De patiënt is een 15-jarige jongen. Hij presenteerde zich met acute lage rugpijn, zwakte van de onderste ledematen, gevoelloosheid en tintelingen, urine-retentie en was niet in staat om te staan of te lopen. Klinisch onderzoek toonde graad 0 kracht in bilaterale enkel dorsiflexoren/extensoren, graad 3 kracht in enkel plantairflexoren, afwezige enkelreflex, verminderde sensaties in L5 en sacrale dermatomen en afwezige bulbocavernosusreflex; beeldvorming toonde een breuk van de L4-wervel met verplaatst fragment dat de cauda equina ernstig comprimeerde. De patiënt onderging een uitgevoerde microscopische decompressie van de middenlijn L3-L4 met verwijdering van schijffragmenten en terugduwen van de botbult. Er was volledig neurologisch herstel inclusief blaascontrole binnen 1 maand na de operatie en geen functionele tekorten na 3–6 maanden.

B.8. German (DE)

Full Case (German)

Eine 41-jährige Frau, die sich über Beschwerden im Oberbauchbereich beklagte, kam in die Ambulanz. Es wurde eine Gastroskopie durchgeführt und ein Polypen wurde gefunden. Die histologische Untersuchung ergab, dass es sich nicht um ein bösartiges Geschwür handelte. Wenn ein Polypen im Magen gefunden wurde, wurde eine Koloskopie durchgeführt. Während der Untersuchung wurde ein kleiner fester Gegenstand beobachtet, der durch die Darmwand hervorragte (). Wir versuchten, ihn mit einer endoskopischen Schlinge () herauszuziehen. Er war jedoch fest an der Darmwand befestigt. Um keine Schäden am Darm zu verursachen, beschlossen wir, unsere Bemühungen einzustellen. Der Rest der Koloskopie verlief ohne Zwischenfälle. CT-Scan zeigte einen IUD-ähnlichen Fremdkörper in „T"-Form (). Der längste Teil ragte durch die Wand der Milzfalte des Darms und der Querabschnitt befand sich in der Bauchhöhle. Nach diesen CT-Befunden informierte uns die Patientin darüber, dass sie vor fast zehn Jahren eine IUD eingesetzt bekommen hatte. Zwei Jahre nach dem Eingriff brachte sie auf natürliche Weise ein Kind zur Welt. Die Patientin gab zu, dass sie sich wegen der IUD nicht medizinisch beraten ließ, weil sie dachte, dass sie herausfallen würde. Das Gerät wurde an der Stelle der splenischen Flexur mit einem laparoskopischen Ansatz gefunden (). Der längste Teil des IUD war die penetrierende Wand des Colon. Eine Mini-Laparotomie mit einer Breite von drei Zentimetern wurde im oberen linken Quadranten durchgeführt, das IUD wurde entfernt und der Darmschaden wurde repariert. Der postoperative Verlauf verlief ohne Zwischenfälle. Der Patient wurde nach drei Tagen aus dem Krankenhaus entlassen und die Behandlung wurde ambulant fortgesetzt.

Reference Summary

Eine 41-jährige Patientin kam mit Beschwerden im Oberbauchbereich in die Ambulanz. Ein Arzt führte eine Gastroskopie durch und fand einen Polypen. Bei einer Koloskopie wurde ein kleiner fester Gegenstand entdeckt, der durch die Darmwand ragte. Die CT-Untersuchung ergab ein IUP-ähnliches Fremdkörper mit "T"-Form, dessen längste Seite durch die Darmwand ragte. Nach diesen CT-Befunden informierte uns die Patientin, dass sie vor fast zehn Jahren ein IUP eingesetzt bekommen hatte. Wir fanden das Gerät mit dem laparoskopischen Ansatz. Das IUP wurde entfernt und der Darmschaden wurde während der Mini-Laparotomie repariert.

Generated Summary – Llama-3.1-8B-Ins (ZS) | BERTScore: 0.905

Eine 41-jährige Frau klagte über Beschwerden im Oberbauchbereich und wurde in die Ambulanz gebracht. Bei der Gastroskopie wurde ein Polypen im Magen gefunden, der jedoch nicht bösartig war. Bei der anschließenden Koloskopie wurde ein kleiner fester Gegenstand beobachtet, der durch die Darmwand hervorragte. Ein CT-Scan zeigte einen IUD-ähnlichen Fremdkörper in "T"-Form, der durch die Darmwand ragte. Die Patientin gab an, vor zehn Jahren einen IUD eingesetzt zu haben, den sie nicht entfernen ließ. Mit einem laparoskopischen Ansatz wurde das IUD an der Stelle der splenischen Flexur gefunden und entfernt. Eine Mini-Laparotomie wurde durchgeführt, um den Darmschaden zu reparieren. Der postoperative Verlauf war ohne Zwischenfälle und die Patientin wurde nach drei Tagen entlassen.

B.9. Romanian (RO)

Full Case (Romanian)

Un bărbat de 26 de ani, de origine poloneză, anterior sănătos, a prezentat o durere de zece zile la piciorul stâng și la partea inferioară a spatelui. Aceasta a fost urmată de dificultate la urinare și scăderea senzației în ambele picioare. Examinarea a relevat semne piramidale și o slăbiciune ușoară a ambelor membre inferioare. Avea o implicare timpurie a sfincterului, care a necesitat cateterizare. Imagistica prin rezonanță magnetică (MRI) a creierului a fost normală. Cu toate acestea, cea a măduvei spinării a evidențiat multiple leziuni intradurale ale măduvei spinării, atât intra- cât și extra-medulare, care se extind de la măduva cervicală până la rădăcinile cordonului cauda equina. A existat, de asemenea, o syrinx care se extindea de la T1 la T10. Nu au existat leziuni intracraniene. A fost efectuată o laminectomie T12-L1. Au fost observate multiple tumori intradurale, extra- și intramedulare. A fost prelevat un eșantion de biopsie dintr-o tumoare intramedulară exofitică, care a fost confirmată a fi un ependimom. Histologia a arătat o tumoare moderat celulară cu rozete perivasculare tipice. Ki-67 a arătat un nivel scăzut și/sau moderat de reînnoire celulară, sugerând un ependimom tipic primar. Nu a existat nicio dovadă de schimbare anaplazică. După operație, pacientul s-a deteriorat, cu un nivel senzorial la T4. A fost administrată radioterapie cranio-spinală post-operatorie, dar nu a existat nicio îmbunătățire clinică la nivelul membrelor inferioare.

Reference Summary

Un bărbat de 26 de ani, de origine poloneză, a prezentat o durere de zece zile în piciorul stâng și în partea de jos a spatelui. Aceasta a fost urmată de dificultate în urinarea și o scădere a senzației în ambele picioare. Examinarea a relevat semne piramidale și o slăbiciune ușoară în ambele membre inferioare. A avut o implicare timpurie a sfincterului, care a necesitat cateterizare. Imagistica prin rezonanță magnetică a creierului a fost normală. Cu toate acestea, cea a măduvei spinării a relevat multiple leziuni intradurale ale măduvei, atât intra- cât și extra-medulare, care se extindeau de la măduva cervicală până la rădăcinile cauda equina. A fost efectuată o laminectomie T12-L1. Au fost observate multiple tumori intradurale, extra- și intra-medulare. După operație, pacientul s-a deteriorat, cu un nivel senzorial la T4. A fost administrată radioterapie cranio-spinală post-operatorie, dar nu a existat nicio îmbunătățire clinică la nivelul membrelor inferioare.

Generated Summary – RaRaMi | BERTScore: 0.930

Un bărbat de 26 de ani a prezentat o durere la piciorul stâng și la partea inferioară a spatelui, urmată de dificultate la urinare și scăderea senzației în ambele picioare. Examinarea a revelat semne piramidale și o slăbiciune ușoară a ambelor membre inferioare. MRI a evidențiat multiple leziuni intradurale ale măduvei spinării, atât intra- cât și extra-medulare, de la măduva cervicală până la rădăcinile cordonului cauda equina, precum și o syrinx de la T1 la T10. A fost realizată o laminectomie T12-L1 și prelucrat un eșantion de biopsie dintr-o tumoare intramedulară exofitică confirmată a fi un ependimom. Pacientul s-a deteriorat după operație, având un nivel senzorial la T4. A fost administrată radioterapie cranio-spinală post-operatorie, dar nu s-a observat îmbunătățire la nivelul membrelor inferioare.

B.10. Czech (CS)

Full Case (Czech)

Zprávy o případu 64letého pacienta, který podstoupil implantaci TPP před 21 lety (1999) k léčbě primární osteoartrózy (OA) pravého kyčelního kloubu. Po 21 letech sledování se u pacienta nerozvinuly žádné komplikace a uvedl skóre SF12 32 a skóre Oxfordského kyčelního kloubu 47/48. TPP se ukázala jako vynikající terapeutická možnost a alternativa ke standardní totální náhradě kyčelního kloubu (THR) pro mladé pacienty s symptomatickou osteoartrózou kyčelního kloubu, protože obnovuje biomechanické a fyziologické zatížení nativního kyčelního kloubu. Jsou poskytnuty jeho současné RTG () a CT () snímky.

Reference Summary

Zprávu jsme podali o 64letém pacientovi, který před 21 lety (1999) podstoupil implantaci tlačné destičky (TPP) k léčbě primární osteoartritidy (OA) pravého kyčelního kloubu. Při 21letém sledování se u něj nerozvinuly žádné pooperační komplikace a uvedl skóre SF12 32 a skóre Oxfordského kyčelního kloubu 47/48.

Generated Summary – Llama3.1-8B-Ins (ZS) | BERTScore: 0.925

Pacient, který podstoupil implantaci TPP před 21 lety (1999) k léčbě primární osteoartrózy pravého kyčelního kloubu, neměl žádné komplikace a dosáhl skóre SF12 32 a skóre Oxfordského kyčelního kloubu 47/48.

B.11. Danish (DA)

Full Case (Danish)

En 40-årig kvinde blev indlagt på vores hospital på grund af en gradvist forstørret masse over fire år af hendes forreste brystvæg. Siden oktober 2008 havde hun oplevet intermitterende smerter i det forreste bryst, og smerterne intensiveredes i august 2012. Den fysiske undersøgelse afslørede en varm 10 × 8 × 6 cm stor masse fastgjort til det øvre brystben, og den var øm ved palpation. Der blev ikke konstateret pulsation. Computertomografi afslørede en osteolytisk læsion med diskret forkalkning i brystbensmarven. Tumoren strakte sig over den ødelagte cortex til de parietale og viscerale bløde dele og involverede noget af ribbenbrusk og det meste af brystbenet. Ifølge de rekonstruerede billeder af CT-scanningen af brystet blev der fremstillet en individuelt tilpasset rustfri stålplade i samme form som patientens thorakale knoglestruktur ved hjælp af den øvre sternum, ribbenbuen og begge sternoclavicular led til rekonstruktionen. Placeringen og fikseringen af pladen var enkel og uden problemer. Pladen blev fastgjort med en klemme og skruer til de resterende ribben og kraveben. Operationen var vellykket, og rekonstruktionen af brystvæggen var tilfredsstillende både med hensyn til udseende og funktion. Efter operationen var der ingen komplikationer, og med et kropsbælte blev patienten udskrevet på den 14. postoperative dag. Den histologiske undersøgelse af det kirurgiske materiale bekræftede diagnosen chondromyxoid fibroma. Ni måneder efter den første operation, oplevede patienten forværret brystsmerter og lokal udsættelse af pladen over det øvre brystben. Røntgen afslørede forskydning af pladen langs de venstre 1. til 3. ribben og et brud i den højre plade-clavikulær forbindelse, to uger senere, udviklede et lignende brud sig på venstre side. Den komplette kirurgiske fjernelse af pladen måtte udføres gennem det oprindelige snit. Rekonstruktion af brystvæggen blev derefter foretaget med et titaniumnet. Nettet blev fastgjort direkte til manubrium og ribbenbrusk og trukket mod hver ribbenstub. Et blødt væv blev syet direkte. Efter operationen, blev der ikke observeret paradoksale bevægelser af brystkassen under vejrtrækning.

Reference Summary

En 40-årig kvinde præsenterede en gradvist forstørret masse på hendes forreste brystvæg. Computertomografi afslørede en osteolytisk læsion med diskret forkalkning i brystknoglen. Tumoren strakte sig over den ødelagte cortex til det parietale og viscerale bløde aspekt, og involverede noget af ribbenbrusk og det meste af brystbenet. En delvis resektion af brystbenet blev udført med succes og en individuel specifik rustfri stålplade blev brugt til at rekonstruere den forreste brystvæg. Det tidlige resultat var godt, men ni måneder efter den første operation blev der fundet brud på pladen ved den bilaterale plade-clavikulær forbindelse. Pladen måtte fjernes, og et titaniumnet blev brugt til at rekonstruere brystvæggen. Patienten har været sygdomsfri i mere end 18 måneder efter den anden operation.

Generated Summary – DACHausa | BERTScore: 0.912

En 40-årig kvinde blev indlagt på vores hospital på grund af en gradvist forstørret masse over fire år af hendes forreste brystvæg. Computertomografi afslørede en osteolytisk læsion med diskret forkalkning i brystbensmarven. Tumoren strakte sig over den ødelagte cortex til de parietale og viscerale bløde dele og involverede noget af ribbenbrusk og det meste af brystbenet. Røntgen afslørede forskydning af pladen langs de venstre 1. til 3. ribben og et brud i den højre plade-clavikulær forbindelse, to uger senere udviklede et lignende brud sig på venstre side. Den komplette kirurgiske fjernelse af pladen måtte udføres gennem det oprindelige snit.

B.12. Norwegian (NO)

Full Case (Norwegian)

En 6 år gammel gutt av indo-aryansk etnisitet fra den nordlige delen av India ble presentert med episoder av akutt magesmerte og abdominal distensjon med oppkast og manglende passasje av avføring. Han benektet en historie med å ta noen medisiner i det siste og hans familiehistorie var ikke medvirkende. En fysisk undersøkelse var normal unntatt for svak blekhet. Det var ingen hepatosplenomegali. Hans fullstendige blodtellinger, erytrocyttsedimentasjonshastighet (ESR), og serumelektrolytter var innenfor normale grenser. På ultralyd ble innvollsinngrep mistenkt sammen med forstørrede flere mesenteriske lymfeknuter. Ved laparotomi ble en liten tarminngrep funnet som var ureducerbar, så en reseksjonsanastomose ble utført og ilealsegmentet (spissen av innvollsinngrepet) ble sendt for histopatologisk undersøkelse. Vi mottok en del av ilealsegmentet som målte 3 × 1,5 cm; de eksterne og snittede overflatene var umerkbare. På histopatologi viste snitt fra spissen av intussusception omfattende gastrisk metaplasi av slimhinnen med lange, spiralformede forgrenende kjertler som inneholdt både rikelige hoved- og parietalceller. Etter et uproblematisk postoperativt utfall, ble han restituert tilfredsstillende og er for tiden under oppfølging.

Reference Summary

En 6 år gammel gutt av indo-aryansk etnisitet fra India ble presentert med episoder av akutt magesmerte og oppblåsthet med oppkast og manglende passasje av avføring. Ved ultralyd ble innvollsoppbygging mistenkt. En laparotomi ble utført og det ileale segmentet (spissen av innvollsoppbygging) ble sendt for histopatologisk undersøkelse. Ved histopatologi viste snitt fra spissen av innvollsoppbygging omfattende gastrisk metaplasi av slimhinnen.

Generated Summary – Llama3.1-8B-Ins (ZS) | BERTScore: 0.920

En 6 år gammel gutt fra den nordlige delen av India ble presentert med episoder av akutt magesmerte og abdominal distensjon med oppkast og manglende passasje av avføring. Ved laparotomi ble en liten tarminngrep funnet, som ble reseksjonert og ilealsegmentet sendt for histopatologisk undersøkelse. Histopatologien viste gastrisk metaplasi av slimhinnen med lange, spiralformede forgrenende kjertler som inneholdt både hoved- og parietalceller.

B.13. Swedish (SV)

Full Case (Swedish)

En 15-årig pojke kom till akuten med en historia av akut smärta i ländryggen och svaghet i underbenen efter att ha lyft ett tungt indiskt musikinstrument (Dholak). Patienten hade utvecklat urinretention sedan skadan. Patienten hade domningar, stickningar i båda underbenen och kunde inte stå eller gå. Undersökningen visade grad 0 styrka i de bilaterala dorsiflexorerna i fotleden, extensor hallucis longus och extensor digitorum. Han hade grad 3 styrka i de bilaterala plantarflexorerna i fotleden. Fotledsryckning var frånvarande bilateralt. Han hade minskad känsel i L5 och sakrala dermatomer. Bulbocavernosusreflexen var frånvarande. Akut MRT-undersökning gjordes, som visade en fraktur i den främre delen av ländryggen med ett fragment som hade förskjutits in i ryggradskanalen och som kraftigt komprimerade cauda equina (och). Akut operation gjordes i form av en interlaminär mikroskopisk dekompression i mitten av L3-L4. Alla lösa och utskjutande diskfragment avlägsnades. En benig knöl trycktes tillbaka till sin position med hjälp av ett hockeystav. Patienten visade tecken på neurologisk återhämtning under den omedelbara postoperativa perioden. Patienten hade fullständig neurologisk återhämtning, inklusive blåstömningskontroll inom en månad efter operationen. Han hade inga funktionella brister under uppföljningen efter 3-6 månader. Dynamiska röntgenbilder tagna vid slutet av 6 månader visade inga tecken på instabilitet ().

Reference Summary

En 15-årig pojke kom till akuten med en historia av akut smärta i ländryggen och smärta i båda nedre extremiteterna med svaghet efter att ha lyft ett tungt indiskt musikinstrument (Dholak). Detta var associerat med urinretention och domningar i perinealregionen. Undersökningen visade svaghet i L5 och S1 med frånvaro av fotledsrörelser bilateralt. Bulbocavernosusreflexen var frånvarande. En akut MRT-undersökning gjordes, som visade en fraktur i den övre delen av L4-kotan med ett fragment som hade förskjutits in i ryggradskanalen och orsakade kompression av cauda equina. Akut kirurgi gjordes i form av en mikroskopisk dekompression i mitten av L3-L4-nivån. Patienten hade en fullständig neurologisk återhämtning inklusive blåskontroll inom en månad efter operationen. Patienten hade inga funktionella underskott under uppföljningen vid 3 och 6 månader. Dynamiska röntgenbilder tagna i slutet av 6 månader visade inte några tecken på instabilitet.

Generated Summary – PsychAI-Discovery-Lab | BERTScore: 0.902

I en 15-årig pojke kom till akuten med akut smärta i ländryggen och svaghet i underbenen efter att ha lyft ett tungt indiskt musikinstrument (Dholak). Han hade urinretention, domningar, stickningar i båda underbenen och kunde inte stå eller gå.
Undersökningen visade grad 0 styrke i dorsiflexorerna och grad 3 styrke i plantarflexorerna. Fotledsryckning var frånvarande bilateralt. Han hade minskad känsel i L5 och sakrala dermatomer. Bulbocavernosusreflexen var frånvarande.
Akut MRT-undersökning visade en fraktur i den främre delen av ländryggen med ett fragment som komprimerade cauda equina. Akut operation gjordes med interlaminär mikroskopisk dekompression vid L3-L4. Alla lösa diskfragment avlägsnades och benig knöl trycktes tillbaka till sin position.
Patienten visade

C. Overview of Participant Systems and Methodologies

Table 21: Comparison of system configurations and methodological approaches across participating teams in the MultiClinSum-2 shared task.

Team	Methodology	Model	Params	Fine-tuned	Prompt Engineering	Train data	Model Domain	Lang.
InfoLab-FEUP	Agentic pipeline: section extraction, critic loop, UMLS enrichment	Llama-3.1-8B-Instruct	8B	—	Role-based prompting	No training data (zero-shot)	General	en, pt
NLP4Health	Prompt engineering + QLoRA fine-tuning	Qwen2.5 (1.5B / 7B)	1.5B, 7B	QLoRA	Role-based, length constraints	MultiClinSum-2 (en, es, fr, pt)	General	en, es, pt, fr
ixa-sum	Dynamic one-shot prompting + GRPO RL fine-tuning	Qwen2.5-7B-Instruct / Gemma-4-E4B	7B, 4B	GRPO + LoRA	Dynamic one-shot, role-based	MultiClinSum-2 training data (9 languages)	General	9 languages
RaRaMi	QLoRA fine-tuning on 4-bit quantized instruction-tuned backbones	Qwen2.5-7B / Llama-3.1-8B / Llama-3.2-1B	1B, 7B, 8B	QLoRA	Role-based, Romanian prompt	26,579 Romanian pairs	General	ro
UMUTeam	Continual medical pretraining + supervised instruction tuning	Qwen3, Gemma3, Salamandra	1B–8B	Continual pretraining + LoRA	Standardized prompt in Spanish	Spanish medical corpora + MultiClinSum-2 (20,721 pairs)	Medical	es
DACHausa	Full fine-tuning of multilingual encoder-decoder with language tags	mT5-small	300M	Full fine-tuning	—	334,024 multilingual pairs	General	14 languages
MediScribes	KG-grounded summarization (RDF triples, GraphRAG) + LLM prompting	GPT-5.1 (API)	—	—	2-shot, role-based, structured output	Few-shot examples only	General	en, nl

D. Manual Summary Quality Assessment with Prodigy Tool

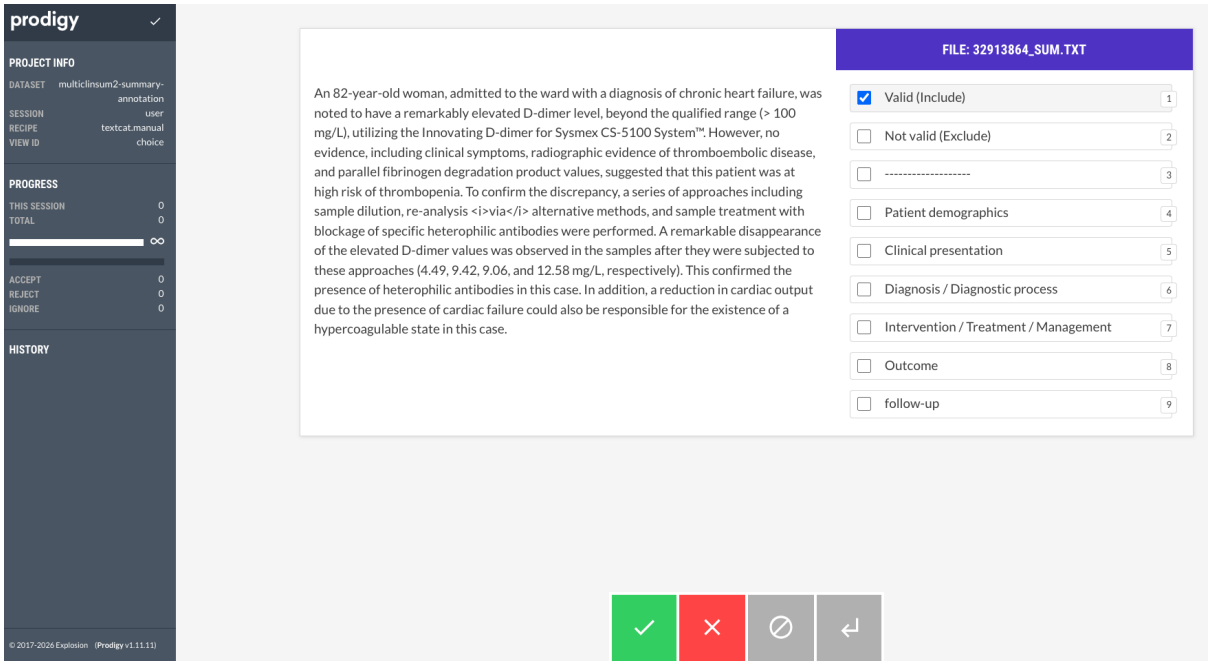

Figure 8: Example of the Prodigy manual annotation interface for clinical summary quality assessment. Clinical annotators evaluate each reference summary using predefined labels covering overall validity and key clinical dimensions such as patient demographics, clinical presentation, diagnosis, intervention, outcome, and follow-up.