# Overview of MultiClinSum Task at BioASQ 2025: Evaluation of Clinical Case Summarization Strategies for Multiple Languages: data, evaluation, resources and results.

Miguel Rodríguez-Ortega[1,*], Eduard Rodríguez-López[1], Salvador Lima-López[1], Carlos Escolano[1], Audrey Mash[1], Maite Melero[1], Lorenzo Pratesi[2], Laura Vigil-Gimenez[1,3], Leticia Fernandez[1,4], Eulàlia Farré-Maduell[1] and Martin Krallinger[1]

[1]*Barcelona Supercomputing Center, Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain*
[2]*Translated, Via Indonesia, 23, Rome, Latium 00144, Italy*
[3]*Parc Tauli Hospital Universitari, Parc Taulí, 1, 08208 Sabadell, Barcelona, Spain*
[4]*Hospital Universitari Mútua Terrassa, Plaça del Doctor Robert, 5, 08221 Terrassa, Barcelona, Spain*

**Abstract**

Recent developments in generative AI-based solutions, particularly large language models (LLMs), are enabling high-impact use cases not only in general-domain applications but also in biomedical and clinical domains. Healthcare professionals and researchers face significant challenges related to efficiency, patient safety, and the delivery of high-quality care, largely due to the ever-growing volume of lengthy clinical documents, including electronic health records and clinical case reports. Automatic clinical document summarization, especially across multiple languages beyond English, has the potential to enhance healthcare system efficiency, improve care management, and support clinical research involving unstructured health data. A key requirement for such systems is the ability to generate high-quality summaries that preserve essential clinical insights, such as patient characteristics, diagnoses, therapeutic indications, and outcomes, while minimizing biases. Achieving this requires a robust evaluation framework to benchmark the quality and accuracy of the generated summaries from unstructured health data. To address this need, we organized the MultiClinSum shared task as part of BioASQ/CLEF 2025. This task focused on evaluating automatic summarization of clinical case reports in four languages: English, Spanish, French, and Portuguese. A total of 10 teams submitted 50 runs for the MultiClinSum task, with the majority of top-performing systems adopting abstractive approaches built on decoder-only architectures such as Qwen and MedGemma. These models were often fine-tuned on biomedical corpora using parameter-efficient strategies like LoRA or quantization. A few extractive methods were also explored, they generally achieved lower performance, highlighting the advantage of abstractive techniques for capturing nuanced clinical information. Overall, the participating systems showed promising performance, achieving results of 0.870 BERTScore F1 for English, 0.758 for Spanish, 0.752 for French, and 0.747 for Portuguese. We have publicly released the MultiClinSum corpora to support the development and evaluation of new clinical summarization systems. The dataset includes both a multilingual gold standard of human-written summaries and a large-scale collection derived from the PMC-Patients collection. This resource also includes additional datasets (Romanian, Italian, Dutch, Swedish, Czech, Catalan, Norwegian, Danish, German, Russian and Greek), extending beyond the four languages covered in the shared task. The MultiClinSum resources ae available at: https://zenodo.org/records/15546018

**Keywords**

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

## 1. Introduction

The volume of information available to clinicians and biomedical researchers is growing rapidly, both in terms of biomedical literature and in the form of clinical records [1]. Accurate patient care requires that clinicians are able to efficiently retrieve, interpret, and integrate relevant information

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*
*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)

from multiple data sources. Biomedical researchers need to navigate humongous amounts of literature to generate new hypotheses and stay current with research advancements in their fields. Electronic resources—such as online literature databases and electronic health record (EHR) systems—have been developed to support clinicians and researchers in managing this expanding body of information. For doctors, the widespread adoption of EHRs has significantly increased the clinical documentation workload, contributing directly to rising stress levels and clinician burnout.

Clinicians currently devote substantial time to summarizing large volumes of textual information—whether compiling diagnostic reports, writing progress notes, or synthesizing a patient's treatment history across multiple specialists [2].
Recent studies indicate that physicians may spend up to two hours on documentation for every hour of direct patient care [3]. Even for experienced physicians with extensive expertise, this complex task inherently carries the risk of errors, which can be particularly detrimental in a field where precision is critical [4]. Failing to thoroughly review a patient's detailed clinical history can lead to serious medical errors [5]. These may include misdiagnosis, inappropriate treatment, and life-threatening complications, often resulting from communication failures and a lack of coordination or sharing of relevant information among care team members. Additionally, with the rapidly growing body of clinical literature—particularly case reports—it is becoming increasingly challenging to identify and interpret key medical insights that are relevant for evidence-based medicine and clinical decision support.

The application of automatic text summarization methodologies offers a robust solution to these challenges, through the extraction of concise and coherent representations of extensive clinical texts [6, 7, 1, 8, 9, 10, 11]. Automatic summarization aims to reduce the length of a text while preserving its essential information content. Summarization can be extractive, selecting key sentences from the original text, or abstractive, generating novel sentences that convey the main ideas more concisely. With the advent of Large Language Models (LLMs), especially those fine-tuned for biomedical or clinical language (e.g., BioBERT[12], ClinicalBERT[13], BioGPT[14]), abstractive summarization has become increasingly feasible and effective. The generation of such summaries has been demonstrated to facilitate healthcare professionals in the rapid and effective comprehension and prioritization of patients' clinical histories. Beyond the clinical environment, text summarization has applications in medical education, evidence synthesis, and biomedical research, where fast access to essential information from case reports, EHRs and biomedical literature can accelerate learning and discovery.

Current research have shown that LLMs can produce clinically coherent and semantically accurate summaries of complex medical documents, often approaching or exceeding the quality of human-written summaries. These models are capable of handling nuances in medical language and capturing key clinical insights, such as diagnoses, symptoms, interventions, and outcomes. However, most existing work has focused on English-language texts, and there is a growing recognition of the need to support multilingual clinical summarization. This is particularly important for enabling equitable access to clinical knowledge, supporting international research collaboration, and ensuring that non-English clinical data can be efficiently used in both local and global contexts. In a study by Dave Van Veen et al. [6], the authors demonstrated that domain-adapted LLMs can achieve strong performance across various clinical summarization tasks—including radiology reports, patient-provider dialogues, and progress notes—sometimes even outperforming human-written summaries in terms of completeness and correctness, as assessed in a physician-led reader study. The study highlights both the potential and limitations of LLMs: while these models can reduce documentation burden and enhance clinical efficiency, their performance varies considerably depending on the task and adaptation strategy. The authors also emphasized the critical need to align automatic evaluation metrics with human judgment, a key consideration in clinical contexts where summary quality directly affects patient care.

Building on this line of research, the study presented in [15] demonstrated the effectiveness of

hybrid methods that combine extractive and abstractive techniques for summarizing electronic health records (EHRs), particularly in high-stakes settings such as intensive care units (ICUs). The authors integrated concept-based extraction with a T5-based transformer model to generate daily progress note summaries, which were subsequently used to predict ICU patient length of stay (LOS). By combining these summaries with structured data, their support vector machine (SVM)-based approach achieved a prediction accuracy of 77.5%, outperforming existing systems. This use case—clinical summarization for both human interpretation and machine learning input—underscores the practical value of high-quality summaries in supporting clinical decision-making.

Specific clinical domains, such as radiology, have also seen growing interest in summarization tasks. A recent study addressed key limitations in the field, including the reliance on private datasets and a narrow focus on chest X-ray data [16]. By introducing the RadSum23 shared task at BioNLP 2023—which utilized the MIMIC-III and MIMIC-CXR datasets along with a newly released Stanford test set—researchers expanded the scope of radiology summarization to include multiple imaging modalities and anatomical regions. The task attracted significant participation, with 112 submissions from 11 teams, underscoring the community's commitment to developing robust and generalizable summarization systems. These efforts are essential for advancing clinical NLP research, promoting transparency, and enabling reproducibility in model evaluation.

Beyond clinical applications, summarization research has also progressed through shared evaluation campaigns such as the MSLR2022 shared task, which focused on multi-document summarization for literature reviews [17, 18]. Held as part of the Scholarly Document Processing Workshop at COLING 2022, the task challenged participants to generate coherent summaries from biomedical abstracts, simulating the synthesis typically performed in systematic reviews. By providing datasets from Cochrane Reviews and the $MS^2$ corpus, the shared task showcased both the potential and current limitations of existing summarization methods. The ProbSum 2023 shared task focused on summarizing patients' active diagnoses and problems from electronic health record (EHR) progress notes [19]. The TAC BiomedSumm track asked teams to utilize citation sentences—known as citances—that reference a specific paper for summarization. The task involved three main components: identifying the corresponding text spans in the referenced papers that reflect the citances, classifying those spans into predefined paper facets, and ultimately generating a summary of the referenced papers based on the collective discussion provided by the citances.

Summarization systems remain challenging to develop, as summaries must be not only linguistically fluent and coherent but also medically accurate and relevant. Moreover, most evaluation and benchmarking efforts have focused on English content, leaving the quality and challenges of clinical text summarization in other languages—and comparative performance across multiple languages—largely understudied. This gap motivated the launch of the MultiClinSum shared task on clinical case report summarization, as part of the BioASQ/CLEF 2025 initiative. The inclusion of multilingual clinical case reports (in English, Spanish, French, and Portuguese) for the MultiClinSum task—using both native and automatically translated clinical texts—represents an important step forward. It encourages the development of summarization systems that are robust across languages and sensitive to the unique requirements of the clinical domain.

This paper presents the MultiClinSum shared task, a multilingual biomedical summarization challenge focused on clinical case reports in English, Spanish, Portuguese, and French. The goal is to advance automatic summarization methods capable of addressing the diversity and complexity of clinical narratives across languages. The work is organized as follows: Section 2 outlines the task description and evaluation system; Section 3 details the construction of the corpus and the resources provided to participants; Section 4 describes the systems submitted, their underlying methodologies, and the obtained results; Section 5 provides a comprehensive discussion of the outcomes, including language-specific observations and clinical implications; and Section 6 concludes the paper with final remarks

and potential directions for future work.

## 2. Task description

The MultiClinSum shared task was organized as part of the BioASQ 2025 workshop, held under the CLEF conference—an initiative that supports the systematic evaluation of information access systems, primarily through experimentation with shared tasks. MultiClinSum specifically addressed the automatic generation of clinical case summaries in four languages: English, Spanish, French, and Portuguese. The provided gold standard corpus was constructed from a curated selection of full-text clinical case reports and their corresponding summaries, derived from open-access clinical case report publications. Additionally, a large-scale dataset was release built on the PMC-Patients collection [20], where patients summaries were used as full case while the summaries were extracted from specific sections of the corresponding PubMed abstract. In both datasets, each available full-text and summary pair was translated into the target languages to support multilingual experimentation. For evaluation purposes, the automatically generated summaries submitted by participating teams were compared against human-written summaries provided by the original article authors (or their corresponding translations), using ROUGE-2 scores and BERTScore as evaluation metrics.

The MultiClinSum shared task was organized into three main phases: training, testing, and evaluation/submission. During the training phase, participants were provided with multilingual datasets consisting of clinical case report texts and their corresponding summaries in English, Spanish, Portuguese, and French. These training sets were obtained from both the curated clinical case report collection (gold standard) and the PMC-Patients dataset (large-scale dataset), following a fixed training split comprising 60% of the data. In the test phase, participants received full text clinical case report texts without accompanying summaries, and their systems were expected to generate automatically coherent clinical summaries. Each sub-track focused on one of the four target languages, allowing participants to develop either monolingual systems or multilingual systems capable of processing several languages. The task did not require cross-lingual summarization but emphasized language-specific summary generation.

In the evaluation phase, participating teams submitted their system outputs through the BioASQ online platform, with submissions provided in ZIP format. Each team was allowed to submit up to five runs per language. The evaluation of system-generated summaries was carried out using two automatic metrics: ROUGE-L, which measures lexical overlap, and BERTScore, which assesses semantic similarity A fair and consistent benchmarking process across all participating languages was ensured throughout the evaluation phase.

### 2.1. Subtracks

To promote the development of clinically relevant summarization models across diverse linguistic contexts, the task was organized into separate sub-tracks by language rather than using a single multilingual corpus. This approach allows a more equitable evaluation and enables targeted adaptation to language-specific and data resources.

MultiClinSum was structured into four individual sub-tracks. Each track focuses in the adaptation of automated text summarization systems using clinical case report texts in a specific language: MutiClinSum-en for English, MutiClinSum-es for Spanish, MutiClinSum-fr for French and MutiClinSum-pt for Portuguese.

Participants were allowed to participate in any of the subtracks as they wish, using either monolingual or multilingual models. Notably, each sub-track had a differently-sized dataset.

## 2.2. Evaluation

The use of appropriate performance metrics to evaluate the scientific and technical robustness and relevance of automatic clinical text summarization and generative AI models is essential for an objective and transparent comparative assessment of the generated results by participating teams. Evaluating automatically generated summaries requires careful consideration of the diverse factors of summary quality, including fluency or factual consistency. Numerous metrics have been proposed for this purpose, each with distinct strengths and limitations. Lexical overlap metrics like ROUGE [21] and METEOR [22] measure surface-level n-gram matches. Semantic-oriented metrics such as BERTScore [23] leverage contextual embeddings to assess deeper textual correspondence, while model-based metrics like BARTScore [24] and BLEURT [25] focus on fluency and coherence.

To assess the quality of automatically generated clinical case summaries across all sub-tracks of the MultiClinSum shared task, we employed two complementary evaluation metrics: ROUGE-L-sum [21] and BERTScore [23]. These metrics were selected to capture both surface-level lexical overlap and deeper semantic correspondence between system-generated outputs and reference summaries.
By integrating both evaluation paradigms, we ensure that the assessment protocol reflects both established standards in summarization and emerging best practices for clinical NLP. It also allows systems to be evaluated from more than one single perspective.

### 2.2.1. ROUGE-Lsum

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a standard suite of metrics widely used in summarization research. ROUGE-Lsum is a sentence-level variant of the ROUGE-L metric, adapted for multi-sentence summarization tasks. Rather than computing the longest common subsequence (LCS) over entire documents, ROUGE-Lsum segments both the candidate and reference summaries by sentence boundaries (typically newlines) and computes LCS scores independently for each sentence pair. The final metric aggregates these scores to yield recall, precision, and F1 values, offering a more granular and structurally sensitive assessment than traditional ROUGE-L.

Given it is an exact match measure, its relevance in an abstractive summarization task is challenged. However, clinical summarization brings unique needs, where factual consistency and coverage of key medical entities (e.g., diagnoses, treatments) are critical. Past evaluations on clinical dialogue summarization found that while ROUGE metrics (including ROUGE-Lsum) correlate poorly with human judgments of coherence and clinical utility, they remain useful for benchmarking surface-level content overlap, particularly in extractive settings [26]. Meta-evaluations of faithfulness metrics in hospital-course summarization further highlight that ROUGE-Lsum's sentence-level granularity can help identify localized factual matches, though it fails to capture semantic fidelity or logical flow—gaps often addressed by combining its results with semantic metrics [27].

### 2.2.2. BERTScore

BERTScore offers a semantic evaluation framework based on contextualized token embeddings derived from pre-trained transformer models, such as BERT. Unlike ROUGE, which relies on n-gram matching, BERTScore computes pairwise cosine similarity between all tokens in the candidate and reference texts. Greedy matching is then performed to align semantically similar tokens, from which precision, recall, and F1 scores are computed.

BERTScore has demonstrated superior alignment with human judgment in multiple summarization benchmarks, particularly when reference and candidate summaries use divergent lexical forms to express equivalent content. In the clinical domain, this is especially valuable, as synonymous terminology and paraphrasing are common (e.g., "myocardial infarction" vs. "heart attack").

To evaluate participant entries in MultiClinSum, Facebook's RoBERTa-Large [28][1] was used for English and a Google multilingual BERT model [29] [2] was used for the rest of the languages (Spanish, French and Portuguese).

As part of the task, an official MultiClinSum evaluation script is available on GitHub[3]. After the task results were released, the test set Gold Standard annotations were shared with participating teams to enable them to perform extra experiments and facilitate error analysis of their systems.

## 3. Corpus and resources

For the MultiClinSum task, publicly available clinical case reports were used [30]—a specialized medical publication type of key clinical relevance, which often resembles the kind of medical information found in discharge summaries. Specific guidelines and reporting recommendations for clinical case reports have been established, stating that a case report should follow a format that includes demographic information, medical history, presenting concerns, clinical findings, diagnoses, interventions, outcomes (including adverse events), and follow-up for a given patient [31]. Clinical case reports are a type of medical scientific publication that describe an individual patient's medical history, detailing symptoms, clinical findings, treatments, diagnostic reasoning, and other relevant medical information [32]. In some cases, these reports focus on rare diseases, unexpected treatment responses, or emerging health threats, playing a crucial role in the dissemination of medical knowledge. Such texts became particularly important during the COVID-19 pandemic, serving as a valuable source for studying symptoms and disease progression. Public literature repositories such as PubMed Central (PMC) are key resources for clinical case report publications, providing clinicians and researchers with access to a large number of case studies. The widely used PubMed database, for example, contains nearly 2.5 million citations corresponding to clinical case reports dating from 1.846 to the present.

An illustrative example of a clinical case report, along with its corresponding summary for all available languages, is presented in figures 1, 2, 3, and 4.

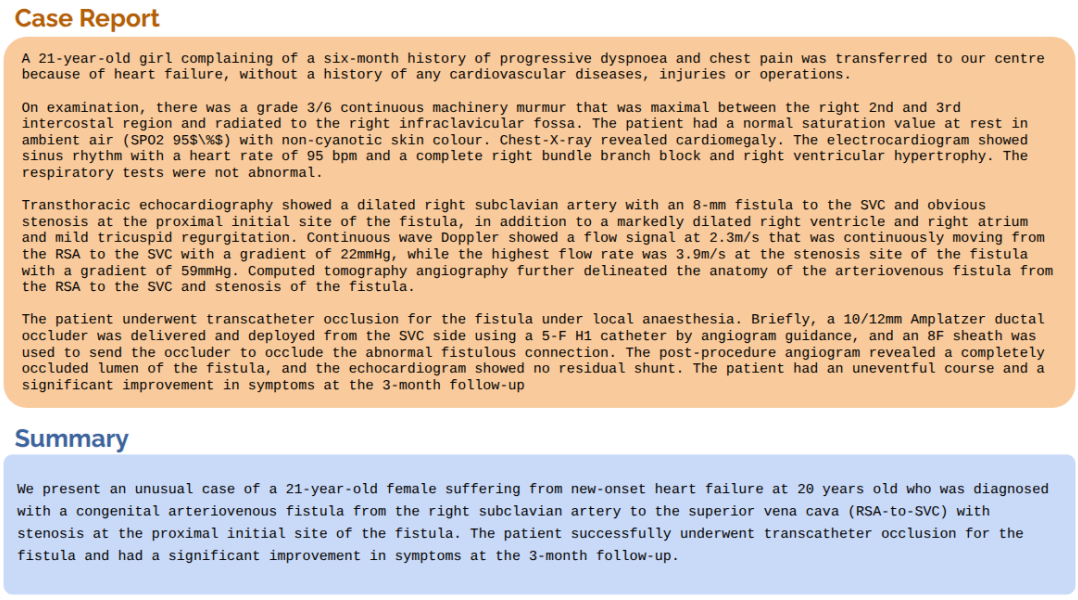


**Figure 1:** Example of a clinical case report and its corresponding summary in English.

[1]https://huggingface.co/FacebookAI/roberta-large
[2]https://huggingface.co/google-bert/bert-base-multilingual-cased
[3]https://github.com/nlp4bia-bsc/MultiClinSumEval.git

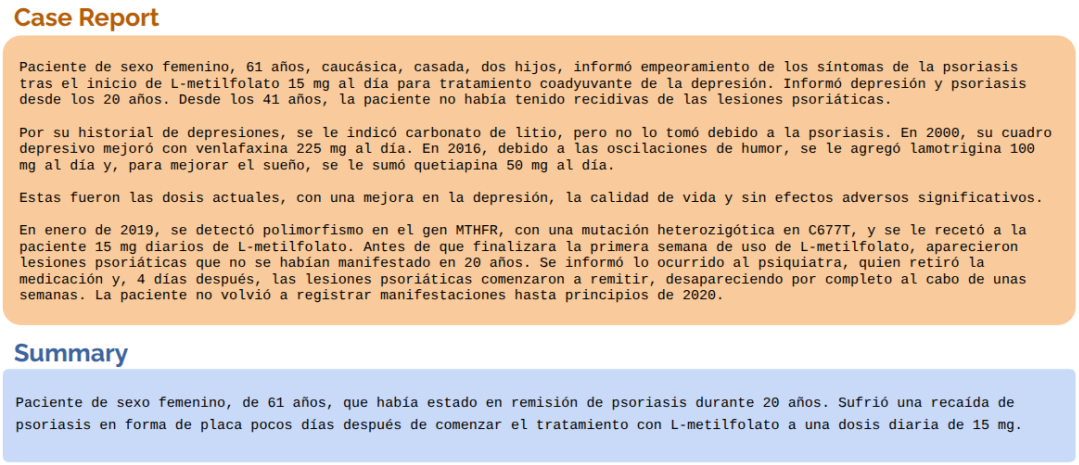


**Figure 2:** Example of a clinical case report and its corresponding summary in Spanish.

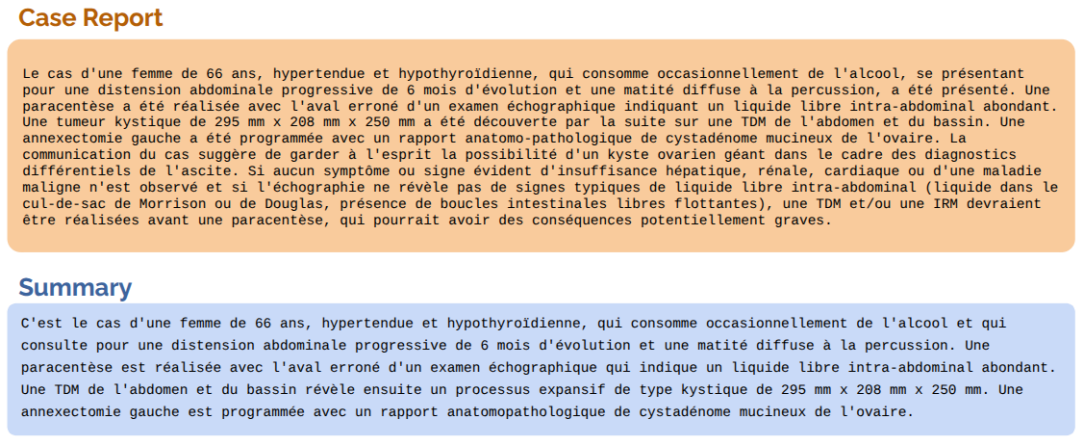


**Figure 3:** Example of a clinical case report and its corresponding summary in French.

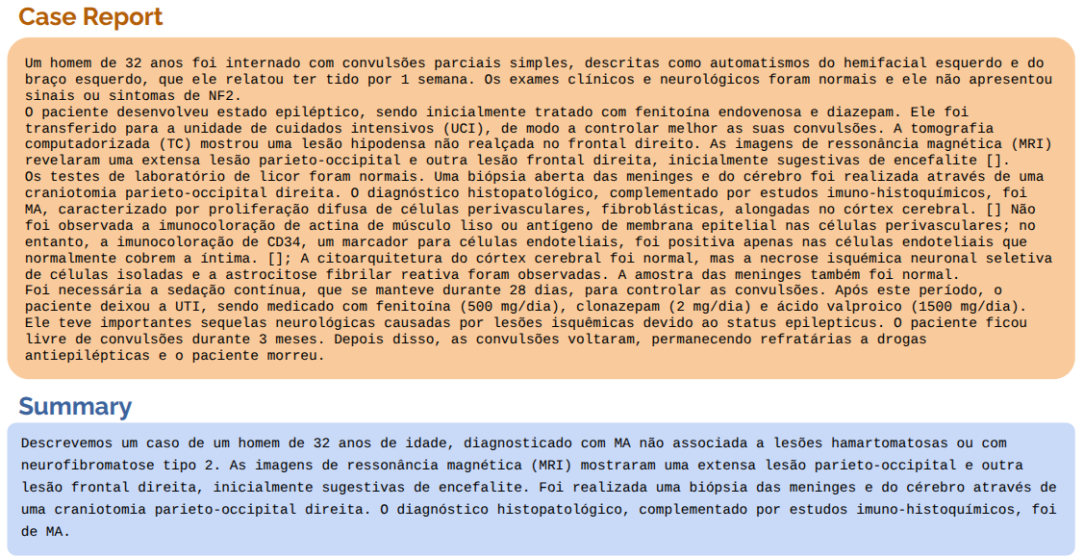


**Figure 4:** Example of a clinical case report and its corresponding summary in Portuguese.

The usual length of the clinical case reports can vary depending on the complexity, structure of the

case, or journal publication instructions, but they typically range from 1,000 to 3,000 words. Based on an internal analysis of the datasets provided in the task, we found a clinical case report mean length of 2,480 words.

### 3.1. Dataset creation process

Two distinct methodologies were employed in the data selection and construction of the task dataset: one focused on creating a gold standard dataset of native clinical case report and summary pairs, and another based on the PMC-Patients subset.

For the gold standard dataset, full case-summary pairs were manually selected independently for each language, resulting in 640 pairs in English, 277 in Spanish, 100 in Portuguese and 100 in French. These clinical cases were primarily sourced from the PubMed database, using filters based on publication type *"case reports"* and publication languages.

The construction of the large-scale dataset started with the retrieval of the publicly available PMC-Patients subset [20], consisting on clinical summaries of patients in English that have been extracted from PubMed Central (PMC) full-text articles repository. To obtain the corresponding author-generated summaries, the abstracts of these PMC articles were processed to isolate specific text fragments which briefly describe the patient clinical case. This extraction was done by manually filtering abstract sections of interest (e.g., clinical case, presentation of the case, case presentation). Following this process, the filtered abstract sections were mapped to their corresponding PMC-Patient record, thereby generating a total of 43,750 full case-summary pairs. As last step, the pairs were pre-processed involving three major steps: text pre-processing, de-duplicating records, and removing noisy data. Text pre-processing was limited to deleting HTML artifacts from the documents, as well as figure and table references mentioned within the case summaries. Duplicate entries—both within the large-scale dataset and across the native gold standard—were removed based on PMC IDs, as well as exact string matching of either cases or summaries, reducing the dataset to 40,650 samples. Additional filters eliminated veterinary cases and case series using substring heuristics, further narrowing the dataset to 40,063 samples.

While the large-scale dataset created from PMC abstracts and case descriptions provided substantial coverage, its construction introduced the challenge of variability in the quality and structure of author-generated summaries. In clinical writing, summarization practices can vary significantly from one case to another—some authors include extensive details to be noteworthy, while others could be more selective and concise. This inconsistency can introduce noise in both model training and evaluation phases. To prevent this variability, an outlier detection strategy based on content length distribution was performed. Specifically, we identified and remove cases considered statistical outliers using a 3×IQR (interquartile range) threshold applied to three features: the relative length of the summary compared to the full case text, the absolute word count of the summary, and the number of sentences. All three of these distributions were long-tailed, so one of their boundaries was set qualitatively. Once these data filtering phase was completed under the aforementioned conditions, the resulting large-scale dataset comprised a total of 38,853 full case-summary pairs in English. Figure 5 shows the whole process of the large-scale dataset creation.

### 3.2. Machine translation process and quality control of translations

To generate a larger multilingual MultiClinSum dataset, machine translation techniques were used, employing the SalamandraTA-7B-instruct architecture[4], a fine-tuned version of the Salamandra large language model[33] that was optimized for translation across 35 official European languages. All

[4]https://huggingface.co/BSC-LT/salamandraTA-7b-instruct

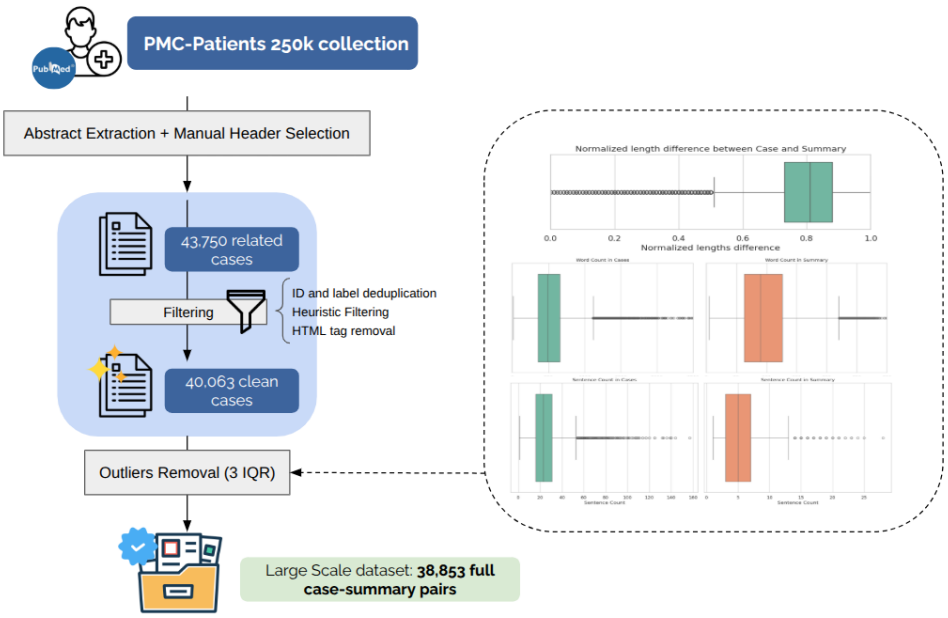


**Figure 5:** Workflow process for the large-scale dataset creation.

examples were produced using the Transformers library[5] with the beam search decoding algorithm with 5 beams.

Since the original dataset contains only monolingual examples, the quality of the generated translations was assessed through quality estimation using COMET-KIWI[34]. This metric, ranging from 0 to 1, computes sentence embeddings for both the input sentence and its translated counterpart, followed by calculating the similarity between the resulting vectors. Higher similarity scores indicate greater semantic alignment between the sentences, suggesting that the meaning of the original input has been effectively preserved. The results show an average score of 0.81 across the four languages tested, ranging from 0.75 for Portuguese–English translations to 0.82 for English–Spanish. Across all language pairs, 91% of the documents achieved a score of 0.7 or higher, indicating robust performance across the entire dataset.

### 3.3. Corpus statistics

MultiClinSum corpus text statistics were calculated for each language for both gold standard and large-scale created datasets. These statistics include the total number of sentences and total tokens, as well as their average value across the corresponding dataset. Tables 1 and 2 illustrate the computed statistics of full cases and summaries repectively. The English sub-track reveals the lowest number of sentences per document in comparison to the other languages, both in relation to the gold standard and the large-scale dataset.

### 3.4. MultiClinSum additional resources

In order to carry out additional analyses beyond the MultiClinSum shared task settings and the four main target languages used, we also generated additional alternative translations. These serve to further characterize the impact of different machine translation systems, as well as to provide a scenario for data augmentation. To ensure high-quality medical text translations for the MultiClinSum task additional dataset, the shared task organizers selected Translated's Lara system, in addition to leveraging the

[5]https://github.com/huggingface/transformers

previously described in-house translation tools developed at BSC. Translated generously provided these translations free of charge in support of the task. Lara had previously been employed in clinical NLP applications within the DataTools4Heart project[6], where it demonstrated strong performance in complex clinical translation scenarios and was validated by professional medical translators. Thus additional alternative translations were provided for clinical cases in English, Spanish, French and Portuguese (MultiClinSum data augmentation collection). Moreover, extended MultiClinSum corpora were generated for the following languages: Romanian, Italian, Dutch, Swedish, Czech, Catalan, Norwegian, Danish, German, Russian and Greek. Some of the author-provided clinical case summaries used for the MultiClinSum task are, from a clinical perspective, more complete, comprehensive, and focused specifically on patient information, rather than on general motivation, relevance, or background of the discussed diseases. To assess the clinical quality and completeness of these case report summaries, clinical experts were asked to classify them using specific quality labels (very complete, complete, and partially complete), as well as labels reflecting the particular clinical information covered (e.g., patient demographics, clinical presentation, diagnosis, intervention and treatment, outcome, and follow-up). This additional set of clinical label classifications will be released alongside the test set clinical case reports to support further research and analysis of clinical summarization approaches.

**Table 1**
Full case statistics of gold standard and large-scale datasets.

| Dataset | Language | Nr. full cases | Nr. sentences | Nr. tokens |
|---|---|---|---|---|
| Gold Standard | EN | 988 | 27,315 | 539,571 |
| | ES | 998 | 26,827 | 599,698 |
| | FR | 1061 | 36,880 | 710,594 |
| | PT | 1034 | 31,562 | 607,356 |
| Large Scale | EN | 28,902 | 814,896 | 14,852,763 |
| | ES | 28,902 | 839,035 | 17,194,901 |
| | FR | 28,902 | 1,028,741 | 18,769,036 |
| | PT | 28,902 | 921,674 | 16,511,501 |

**Table 2**
Summary statistics of gold standard and large-scale datasets.

| Dataset | Language | Nr. summaries | Nr. sentences | Nr. tokens |
|---|---|---|---|---|
| Gold Standard | EN | 988 | 5,515 | 103,198 |
| | ES | 998 | 5,478 | 116,613 |
| | FR | 1061 | 6,439 | 127,856 |
| | PT | 1034 | 6,090 | 115,315 |
| Large Scale | EN | 28,902 | 156,519 | 2,809,401 |
| | ES | 28,902 | 155,670 | 3,218,657 |
| | FR | 28,902 | 166,198 | 3,455,175 |
| | PT | 28,902 | 162,017 | 3,010,877 |

# 4. Results

## 4.1. Participation overview

In general, there has been a very satisfactory participation in the task with promising results in each of the sub-tracks. A total of 60 teams registered for the MultiClinSum task, of which 10 teams submitted at least one run for a given sub-track as presented in Table 3. Specifically, 8 teams participated in the

[6] https://www.datatools4heart.eu/

English sub-track, 6 teams in the Spanish, 5 teams in French, and 6 in Portuguese. Each team was allowed to submit up to 5 runs per sub-track. As expected, the best results were obtained in the English sub-track (MultiClinSum-en), which had the highest level of participation. Nevertheless, the others sub-tracks were quite well represented in terms of both participation and novel methodologies applied.

**Table 3**
Overview of the participating teams in MultiClinSum. The Affiliation column, A/I stands for academic or industry institution.

| Team name | Language | Nr. runs | Affiliation | Ref. |
|---|---|---|---|---|
| chatzimina | English | 1 | ICS-FORTH (I) | N/A |
| pjmathematician | English | 5 | Netaji Subhas University of Technology, New Delhi, India (A) | [35] |
| | Spanish | 5 | | |
| | French | 3 | | |
| | Portuguese | 3 | | |
| BU Team | English | 1 | Bournemouth University, UK (A) | [36] |
| | Spanish | 1 | | |
| | French | 1 | | |
| | Portuguese | 1 | | |
| seemdog | English | 1 | AITRICS (I) | N/A |
| MaLei | English | 1 | University of Manchester, UK (A) | [37] |
| ExtraSum | English | 4 | University of Limerick, Ireland (A) | [38] |
| | Spanish | 4 | | |
| | French | 2 | | |
| | Portuguese | 2 | | |
| grazhdanski | Spanish | 1 | N/A | [39] |
| JohannaUE | English | 2 | Universidad Europea de Valencia, Spain (A) | [40] |
| | Spanish | 4 | | |
| | French | 1 | | |
| | Portuguese | 2 | | |
| ETS-PUCPR | Portuguese | 1 | University of Quebec, Montreal, Canada (A) | [41] |
| MedCOD | English | 1 | N/A | [42] |
| | Spanish | 1 | | |
| | French | 1 | | |
| | Portuguese | 1 | | |

## 4.2. System results

The results for the MultiClinSum for each sub-track are shown in tables 4, 5, 6, 7. Across the four languages, a few teams' system consistently demonstrated strong performance in both lexical and semantic evaluation metrics. The most consistent team was *pjmathematician* [35], which achieved top or near-top BERTScore F1 values across all languages and ranked first in ROUGE-F1 in the French and Portuguese sub-tracks. In the English sub-track, team *seemdog* led overall with the highest BERT-F1 (0.870), showcasing particularly effective semantic understanding in this language. Meanwhile, in the Spanish track, team *grazhdanski* [39] outperformed the rest with the best BERTScore and strong ROUGE-F1.

Other teams also showed competitive results in specific areas. For instance, *MaLei* [37] and *MedCOD* [42] ranked highly in ROUGE precision scores, particularly in English, indicating strong surface-level

alignment with the test set summaries. Team *BU Team* [36] performed consistently across all sub tracks, with particularly strong BERT recall in French and Portuguese. Some teams, such as *ExtraSum* [38], achieved high lexical precision (ROUGE-P) but not so well in terms of semantic similarity, highlighting the difficulty of achieving good results in both. Overall, the results show the complexity of clinical summarization across languages and suggest that systems prioritizing semantic representation, such as those performing better in terms of BERTScore, tend to generalize more effectively across linguistic contexts.

**Table 4**
Results of the MultiClinSum-en sub-track in English. The best result for each metric is bolded, and the second-best is underlined.

| Team | BERT-P | BERT-R | BERT-F1 | ROUGE-P | ROUGE-R | ROUGE-F1 |
|---|---|---|---|---|---|---|
| BU team | 0.855 | <u>0.857</u> | 0.856 | 0.275 | **0.275** | 0.259 |
| chatzimina | 0.816 | 0.804 | 0.810 | 0.104 | 0.062 | 0.070 |
| JohannaUE | 0.846 | 0.834 | 0.840 | 0.207 | 0.154 | 0.167 |
| MaLei | <u>0.878</u> | 0.832 | 0.855 | **0.465** | <u>0.247</u> | **0.308** |
| pjmathematician | <u>0.878</u> | 0.845 | <u>0.861</u> | <u>0.405</u> | 0.231 | <u>0.274</u> |
| MedCOD | 0.876 | 0.841 | 0.858 | 0.403 | 0.224 | 0.257 |
| seemdog | **0.880** | **0.861** | **0.870** | 0.340 | 0.240 | 0.267 |
| ExtraSum | 0.867 | 0.824 | 0.845 | 0.386 | 0.177 | 0.231 |

**Table 5**
Results of the MultiClinSum-es sub-track in Spanish. The best result for each metric is bolded, and the second-best is underlined.

| Team | BERT-P | BERT-R | BERT-F1 | ROUGE-P | ROUGE-R | ROUGE-F1 |
|---|---|---|---|---|---|---|
| BU team | 0.724 | 0.735 | 0.729 | 0.261 | **0.290** | 0.259 |
| grazhdanski | **0.770** | **0.747** | **0.758** | 0.364 | <u>0.267</u> | <u>0.290</u> |
| JohannaUE | 0.692 | 0.713 | 0.702 | 0.203 | 0.285 | 0.218 |
| pjmathematician | <u>0.755</u> | <u>0.719</u> | <u>0.735</u> | <u>0.389</u> | 0.243 | **0.260** |
| MedCOD | <u>0.755</u> | 0.705 | 0.728 | **0.406** | 0.207 | 0.253 |
| ExtraSum | 0.741 | 0.688 | 0.713 | 0.388 | 0.187 | 0.240 |

**Table 6**
Results of the MultiClinSum-fr sub-track in French. The best result for each metric is bolded, and the second-best is underlined.

| Team | BERT-P | BERT-R | BERT-F1 | ROUGE-P | ROUGE-R | ROUGE-F1 |
|---|---|---|---|---|---|---|
| BU team | 0.725 | **0.740** | <u>0.731</u> | 0.241 | **0.289** | 0.247 |
| JohannaUE | 0.716 | 0.710 | 0.712 | 0.211 | 0.214 | 0.197 |
| pjmathematician | **0.770** | <u>0.736</u> | **0.752** | <u>0.370</u> | 0.246 | **0.277** |
| MedCOD | 0.727 | 0.692 | 0.708 | 0.350 | 0.211 | 0.239 |
| ExtraSum | <u>0.742</u> | 0.692 | 0.716 | **0.374** | 0.176 | <u>0.227</u> |

## 4.3. Methodologies by team

This section provides an overview of the methodologies employed by participating teams in the MultiClinSum shared task. Despite the diversity of approaches, most systems were built upon large language models, with varying strategies in pre-processing, fine-tuning, and generation. Participants explored both extractive and abstractive paradigms, leveraging encoder-decoder and decoder-only architectures, as well as monolingual and multilingual setups. Below, we briefly summarize the main

**Table 7**
Results of the MultiClinSum-pt sub-track in Portuguese. The best result for each metric is bolded, and the second-best is underlined.

| Team | BERT-P | BERT-R | BERT-F1 | ROUGE-P | ROUGE-R | ROUGE-F1 |
|---|---|---|---|---|---|---|
| BU team | 0.724 | <u>0.732</u> | 0.727 | 0.250 | **0.270** | 0.244 |
| ETS-PUCPR | 0.740 | **0.735** | <u>0.737</u> | 0.280 | <u>0.259</u> | 0.249 |
| JohannaUE | 0.701 | 0.680 | 0.689 | 0.230 | 0.197 | 0.197 |
| pjmathematician | **0.767** | <u>0.729</u> | **0.747** | <u>0.368</u> | 0.239 | **0.273** |
| MedCOD | <u>0.749</u> | 0.698 | 0.722 | **0.391** | 0.189 | 0.236 |
| ExtraSum | 0.739 | 0.685 | 0.711 | 0.375 | 0.176 | <u>0.227</u> |

methodological choices made by each team.

**Team ExtraSum**
This team compared four existing extractive summarization approaches: graph-based, concept-based, topic-based and cluster-based techniques. The approach emphasizes factual consistency by preserving original sentences and highlights language-specific performance trends (e.g., tokenizer artifacts in Spanish). Clustering-based selection outperforms other extractive approaches in both ROUGE and BERTScore, yet still falls short when comparing against abstractive approaches.

**Team ÉTS-PUCPR**
This team present MedGemma-Sum-Pt, a lightweight model for automatic summarization of Portuguese clinical case reports. Addressing the challenges posed by limited annotated resources and the linguistic complexity of medical Portuguese, the authors explore two strategies: (i) zero-shot prompting using instruction-tuned multilingual language models, and (ii) supervised fine-tuning of the domain-specific MedGemma model using LoRA, a parameter-efficient adaptation method. The fine-tuned model demonstrates superior performance in internal and official evaluations, particularly in terms of semantic fidelity as measured by BERTScore. Despite being trained on a relatively small expert-annotated dataset and deployed using modest computational resources, MedGemma-Sum-Pt outperforms all tested zero-shot baselines. The authors release the model publicly to support further research in clinical NLP for low-resource languages, highlighting the viability of domain-adapted, compact models for real-world medical applications.

**Team MedCOD**
The authors of this team proposed MedCOD, a contextual augmentation framework to enhance multilingual medical summarization. The approach begins by extracting medical keywords from full clinical texts using the Qwen2.5-14B model. These keywords are then translated into five target languages (EN, ES, FR, DE, PT) using the NLLB-3.3B model, and validated through back-translation and semantic equivalence checking. The validated multilingual keywords are incorporated into structured prompts to provide contextual input for LLMs. Experiments were conducted using Qwen2.5-14B and Phi-4B in both zero-shot and fine-tuned settings, and fine-tuning was performed using LoRA (Low-Rank Adaptation) dataset. The study shows that combining MedCOD prompting with fine-tuning leads to improved summarization quality, particularly in non-English languages. Notably, even without fine-tuning, MedCOD prompts alone provided substantial gains in languages like Portuguese and French, highlighting its effectiveness as a lightweight adaptation method.

**Team grazhdanski**
In this case, the author explores the use of Group Relative Policy Optimization (GRPO), a reinforcement learning approach, for summarizing Spanish clinical case reports within the MultiClinSum shared task at CLEF 2025. Several adaptation strategies are evaluated using LLaMA 3.1 8B-Instruct, including supervised fine-tuning, domain adaptation, and GRPO-based training. The GRPO-trained

model, optimized with ROUGE-L and BERTScore reward functions, outperforms both fine-tuned and zero-shot baselines, achieving the best performance on the official test set. The study highlights the potential of GRPO for improving clinical summarization in low-resource settings.

**Team JohannaUE**
This team present Agentic MCS, a multilingual clinical summarization framework combining extractive and abstractive techniques using a modular LangGraph-based architecture with components such as NER-guided entity preservation, knowledge graphs, and fine-tuned or prompted language models. Evaluated across four languages, it achieves strong semantic performance (BERTScore) and highlights the trade-off with lexical overlap (ROUGE). The study demonstrates the effectiveness of hybrid, agent-based pipelines for domain-specific biomedical summarization.

**Team BU Team**
The authors of this team propose a two-stage distillation framework for multilingual clinical summarization using Qwen2.5 models family[43]. In the first stage, a teacher model (Qwen-2.5-72B-instruct[7]) extracts key clinical information; in the second stage, the student model (Qwen2.5-0.5B-Instruct[8]) is trained with a dual loss that supervises both summarization and alignment to the extracted information. The purpose of this secondary task is not for inference but rather to force the model to represent the relevant clinical concept and therefore have a greater chance of including them in the summary.

**Team MaLei**
MaLei team implemented a prompt-based framework using Iterative Self-Prompting (ISP) to guide large language models, specifically GPT-4 and GPT-4o. Their approach involved crafting a meta-prompt that combined Chain-of-Thought instructions, clinical perspectives, and evaluation-based feedback, which was refined iteratively using few-shot examples and model-generated reflections. They used BERTScore and ROUGE-L to monitor performance across refinement epochs, optimizing prompt versions until improvements plateaued. The final prompt version was applied to generate summaries on the English test set, with additional regeneration steps to enforce conciseness where needed.

**Team pjmathematician**
This team employed a multilingual summarization approach based on fine-tuned Qwen family language models, enhanced through Low-Rank Adaptation (LoRA). They developed an automated prompt optimization framework in which a "judge" model iteratively refined a system prompt for a "worker" model, aiming to maximize ROUGE scores. The final optimized prompt emphasized extractive fidelity, closely mirroring the terminology and structure of the source clinical reports.

## 4.4. Methodologies by approach

This subsection presents an in-depth analysis of the system performance in the shared task, examining how different methodological choices impacted summarization quality across languages. Given the diversity of participating teams' approaches, we analyze results along several key aspects: the type of summarization strategy, model size and architecture, prompting techniques, and the influence of the automatic translation systems. Table 8 provides a summary of the participant systems based on these characteristics.

**Performance by summarization type**
The shared task attracted a wide range of approaches, including extractive (e.g., *ExtraSum* [38]), abstractive (e.g., *MaLei* [37], *MedCOD* [42]). Team *pjmathematician* [35] used a decoder-based model, which implies abstractive summarization, yet the automatically optimized prompt stated the summaries should be extractive. One of the systems proposed by team *JohannaUE* [40], had a similar idea as the

[7]https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[8]https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct

previous, but did so in a more defined manner (as opposed to stating it in a prompt) where there was a first step to rank the most relevant sentences using BM25 with subsequent LLM processing for quality assurance and abstractive condensations.

Abstractive systems generally outperformed extractive ones in BERT and ROUGE scores across all languages. The extractive system *ExtraSum* was consistently outperformed by LLM-based abstractive systems, despite showing competitive ROUGE precision in some languages (e.g., French, Portuguese).

Extractive methods were also at a disadvantage given that the gold standard summaries had not been formulated with extractive constraints in mind, resulting in a mismatch between the content selected by extractive systems and the more paraphrastic or restructured gold summaries. This inherent limitation reduced their ability to match the semantic richness captured by abstractive models, particularly those leveraging LLMs capable of cross-lingual generalization and deeper contextual understanding.

**Performance by model size**

Larger models (e.g., *pjmathematician*'s 32B Qwen3 [35]) generally achieved the highest ROUGE-F1 and BERT-F1 scores across most tracks. However, smaller models with tailored fine-tuning (e.g., *BU team*'s 0.5B Qwen2.5 with distillation and quantization [36]) performed competitively, indicating the effectiveness of lightweight architectures when combined with targeted training strategies. In particular, the two-stage distillation framework enabled the smaller model to internalize domain-relevant concepts, improving content selection and summary coherence without increasing inference-time complexity.

Although performance did not scale linearly with model size, larger models generally produced better results even with less sophisticated approaches. However, the performance gains were not substantial enough to render smaller models obsolete, highlighting their potential for deployment in low-resource or high-latency environments.

**Performance by model architecture**

A notable trend in recent summarization research is the increasing dominance of decoder-only architectures over the traditional encoder-decoder models like BART and T5. While encoder-decoder frameworks were initially considered the gold standard for sequence-to-sequence tasks due to their explicit separation of input comprehension and output generation, the landscape has shifted as large-scale pretraining and instruction tuning have unlocked the generative power of decoder-only models.

This shift is also reflected in the implementations presented. Of all 8, one is an encoder-only, while the rest are all decoder-only.

**Prompt engineering strategies**

The increasing dominance of decoder-only architectures in the task of summarization has placed prompt engineering at the forefront of model control and output quality. Unlike encoder-decoder frameworks, which typically rely on supervised fine-tuning, decoder-only models such as GPT, Qwen, or LLaMA depend heavily on prompt design to steer generation. As a result, the quality of prompts emerged as a key differentiator among submissions, with participants adopting diverse strategies to best guide the model towards the expected results.

One notable approach was automated prompt optimization, exemplified by *pjmathematician* [35], who leveraged a fully automated pipeline based on Qwen3 (32B) to generate prompts without manual intervention. This allowed for scalable multilingual summarization and removed potential biases. *MaLei* [37] applied a related strategy, employing *Iterative Self-Prompting (ISP)* with GPT-4, where prompts were refined across multiple rounds to enhance semantic alignment and factuality, particularly for English clinical summaries. Both papers include the final, optimized system prompts.

Contextual prompt augmentation was another effective method. Team *MedCOD* [42] automatically extracted medical keywords and injected them into the prompts to enrich domain specificity across five languages (they added German). This structured augmentation yielded significant improvements in non-English texts when compared to simpler summarization prompts, increasing F1 scores for both BertScore and Rouge by around 0.15. The authors attribute this increase to both a language anchor and a domain signal.

A more modular and dynamic strategy was adopted by *JohannaUE* [40], who introduced hybrid prompt structures. Their multi-agent pipeline integrated Named Entity Recognition (NER) outputs and knowledge graph summaries into the prompt, aligning extracted factual content with the abstractive generation phase. This design allowed for prompt composition to adapt based on both language and content type.

Both *ÉTS-PUCPR* [41] and *grazhdanski* [39] tried to force their models to focus on relevant clinical information. They both assigned a role (clinical assistant), goal (summarization) and structural constraints (tone and format), with the former also adding some examples, making it few-shot as opposed to zero-shot.

Interestingly, some teams—such as *BU Team* and *ExtraSum*—chose not to use prompt engineering and instead relied on lightweight fine-tuning of models. Their results, while competitive in certain languages, highlighted the trade-offs between static model tuning and dynamic prompt-based control.

Overall, the submissions demonstrated that prompt engineering can not only complement fine-tuning but in some cases substitute it.

**Effect of automatic Machine Translation**

The use of automatic machine translation to generate data in Spanish, French and Portuguese from the English PMC-Patients subset of clinical cases is likely to impact system performance across the translation languages. While translation ensured consistent content, it may have introduced issues such as unnatural phrasing, tokenization inconsistencies, or loss of medical detail, particularly notable in languages with less comprehensive medical translation resources. These difficulties could have impacted both model training and evaluation metrics, especially ROUGE, which is sensitive to lexical overlap. However, the solid performance observed in some teams submission (e.g., *ÉTS-PUCPR*, *grazhdanski*, *MedCOD*) suggests that with appropriate adaptation methods, translated data can still support competitive summarization performance in low-resource languages.

**Table 8**
Summary of participants methodologies.

| Team Name | Sum. type | Model name | Params (B) | Finetuned | Prompt engineered | Use Large-scale | Domain | Lang. |
|---|---|---|---|---|---|---|---|---|
| BU team | Abstractive | Qwen2.5-0.5B | 0.5 | Yes (Distillation, Quantization) | Yes (persona-based) | No | General | Multilingual |
| ExtraSum | Extractive | BERT | 0.1 | No | No | Yes | General | Multilingual |
| ÉTS-PUCPR | Abstractive | MedGemma-pt | 4 | Yes (LoRA, Quantization) | Yes | No | Clinical | Portuguese |
| JohannaUE | Hybrid | Various | - | - | - | - | - | - |
| MaLei | Abstractive | GPT4 | proprietary | No | Yes (Iterative Self-Prompting) | No | General | Multilingual |
| MedCOD | Abstractive | Qwen2.5 / Phi-4 | 14 | Yes (LoRA, Quantization) | Yes (Contextual Information) | Yes | General | Multilingual |
| grazhdanski | Abstractive | LLaMA3.1-Instruct | 8 | Yes (Reinforcement Learning) | Yes (persona-based) | No | General | Multilingual |
| pjmathematician | Extractive | Qwen3 | 32 | Yes (LoRA) | Yes (Iterative Self-Prompting) | No | General | Multilingual |

## 5. Discussion

The MultiClinSum shared task has led to the development and evaluation of novel systems for automatic clinical case summary generation for multiple languages, that is four languages: English, Spanish, French, and Portuguese. Together, these languages represent over 1.198 million native speakers worldwide, complementing most previous medical summarization shared tasks and efforts, which have primarily focused only on English content. Since many medical records and a significant number of clinical case reports are not written in English, it is becoming increasingly important to promote the development of NLP and generative AI systems that can handle non-English data or are inherently capable of multilingual processing.

Overall, the MultiClinSum task received a wide range of solutions and strategies explored by the participating teams with a total of ten team submissions. The evaluation was conducted using both lexical (ROUGE-Lsum) and semantic (BERTScore) metrics, reflecting the complex requirements of clinical summarization. Across all languages, the best-performing systems demonstrated strong semantic alignment with gold standard summaries, as captured by BERTScore-F1, while some systems exhibited a trade-off with surface-level overlap (ROUGE scores).

In the English sub-track, *pjmathematician* [35] achieved the best BERTScore-F1 score, while *seemdog* led in ROUGE metrics. The *BU Team* [36] also performed strongly with a Qwen2.5-based distillation framework, consistently ranking among the top. In Spanish, team *grazhdanski* [39] obtained the highest overall scores using GRPO, a reinforcement learning method optimized with ROUGE-L and BERTScore rewards. For French, the *BU Team* [36] again led, confirming the robustness of their multilingual training strategy. For Portuguese, *Team ÉTS-PUCPR* [41] outperformed others in semantic metrics using a fine-tuned, domain-specific model (MedGemma) adapted with LoRA, while *Salim's* MedCOD [42] framework also showed competitive results through keyword-based prompting.

Top-performing systems combined efficient adaptation techniques (e.g., LoRA, prompting, distillation) with strong domain alignment, illustrating the importance of lightweight, semantically focused models in clinical summarization across the languages. Systems' performance varied noticeably across languages sub-tracks, reflecting both resource availability and linguistic complexity. English, benefiting from abundant training data and model support, showed overall higher scores for both BERTScore and ROUGE metrics. Spanish and French sub-tracks demonstrated decent performance, with certain systems like GRPO (Spanish) and MedCOD (French) achieving strong results despite limited resources. Portuguese showed the largest performance gap, though domain-specific fine-tuning (as in *Team ÉTS-PUCPR's* MedGemma-Sum-Pt [41]) closed the gap significantly in semantic metrics. These differences highlight the importance of language-specific strategies, as well as the advantages of domain adaptation and efficient fine-tuning, particularly in under-resourced clinical settings.

From a clinical perspective, automatically generated summaries must contain key medical information essential for decision-making and case understanding. This includes the patient's primary diagnosis, relevant clinical observations, treatments, outcomes, and follow-up recommendations. In the context of case reports, an accurate summarisation of these elements is essential to ensure that clinicians can rapidly assess the core narrative without missing critical details. So systems that focus on these aspects through approaches detailed in previous sections, demonstrate remarkable potential for real-world integration into clinical documentation and decision-support workflows.

Clinical case reports can vary widely, covering heterogeneous scenarios from rare genetic disorders to complex surgeries and emerging diseases. This particular issue presents a challenge to automatic summarisation systems, requiring them to be capable of handling diverse contents, terminologies, and levels of detail. The capture of key clinical insights without the loss of unusual details requires adaptable models and robust domain understanding, especially in highly specialized or uncommon cases.

With regard to the author provided summaries, while these are valuable for creating gold standard datasets, they often vary in quality and consistency. The presence of a diversity of writing styles, emphasis, and completeness may result in redundancy in the summaries, the introduction of subjective interpretations, or an emphasis on narrative flow rather than structured clinical content, which may have an impact on the training and evaluation processes. These inconsistencies highlight the need for careful curation and potentially supplementary annotation to ensure reliable benchmarks for automatic summarization models.

## 6. Conclusions and outlook

The MultiClinSum shared task addressed the growing need for effective automatic summarization of clinical case reports across multiple languages. By providing a multilingual dataset covering English, Spanish, French, and Portuguese languages, the task enabled the community to explore and evaluate a wide range of summarization strategies using both gold standard and large-scale datasets. Participating teams employed diverse methodologies, from prompt optimization with powerful LLMs to fine-tuning multilingual architectures. The partipant system results showed strong performance overall, with English models achieving the highest BERTScore F1, and competitive results observed for the other languages. By releasing the MultiClinSum resources to the public, we aim to promote further research in this field, fostering innovation in multilingual clinical summarization NLP techniques and supporting the development of tools that can improve healthcare and medical research on a global scale.

As with discharge summaries, clinical case reports are particularly enriched with specific types of clinical entities or concepts relevant to patient demographics (e.g., age, gender, or occupation), clinical presentation (e.g., findings, signs, and symptoms), diagnosis (e.g., disorders and diseases), intervention (e.g., clinical procedures and medications), as well as outcome and follow-up. Therefore, further examination of the impact and relevance of these clinical entities for automatic summarization approaches should be considered in future research scenarios and summarization system developments. The release of semantically enriched or entity-annotated full clinical case reports and their corresponding summaries might constitute a valuable resource to foster such future developments and is planned for upcoming MultiClinSum data releases. A more granular evaluation setting, or specific subtasks based on the type of case report or clinical specialty, could provide deeper insights into clinical case summarization performance. Other aspects that would require further analysis relate to the robustness and potential biases of automatic clinical text summarization strategies with respect to patient sex or gender.

We foresee that the MultiClinSum task may help promote future initiatives and efforts to implement and evaluate clinical text summarization solutions across multiple languages—not only in English, Spanish, French, and Portuguese. There is a pressing need to systematically collect and release multilingual full-text clinical documents along with their corresponding summaries, in order to enable more effective benchmarking and exploitation of generative AI and large language models (LLMs) for text summarization tasks.

## Acknowledgments

The MultiClinSum track was funded by Spanish and European projects such as DataTools4Heart (Grant Agreement No. 101057849), AI4HF (Grant Agreement No. 101080430). This publication is part of the R &D &I project TED2021-129974B-C22, funded by MICIU/AEI /10.13039 /501100011033 and by the European Union NextGenerationEU/PRTR (BARITONE (Proyectos de Transición Ecológica y Transición Digital 2021). We would also like to acknowledge the scientific committee members, Sophia Ananiadou, Horacio Saggion and Simon Mille for their valuable feedback and suggestions regarding the task settings and evaluation scenarios as well as the BioASQ organizers and specially Anastasios Nentidis for their technical support.

## Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI-GPT-4.1 to: enhance the grammar and paraphrasing. This was followed by a review and edit of the content, with the author(s) taking full responsibility for the publication.

## References


[1] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, G. Del Fiol, Text summarization in the biomedical domain: a systematic review of recent research, Journal of biomedical informatics 52 (2014) 457–467.

[2] J. F. Golob Jr, J. J. Como, J. A. Claridge, The painful truth: The documentation burden of a trauma surgeon, Journal of Trauma and Acute Care Surgery 80 (2016) 742–747.

[3] C. Sinsky, L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty, G. Blike, Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties, Annals of internal medicine 165 (2016) 753–760.

[4] T. R. Yackel, P. J. Embi, Unintended errors with ehr-based result management: a case series, Journal of the American Medical Informatics Association 17 (2010) 104–107.

[5] R. J. FitzGerald, Medication errors: the importance of an accurate drug history, British journal of clinical pharmacology 67 (2009) 671–675.

[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] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, et al., Adapted large language models can outperform medical experts in clinical text summarization, Nature medicine 30 (2024) 1134–1142.

[8] 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, Journal of Medical Internet Research 27 (2025) e68998.

[9] A. Chaves, C. Kesiku, B. Garcia-Zapirain, Automatic text summarization of biomedical text data: a systematic review, Information 13 (2022) 393.

[10] 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.

[11] D. Keszthelyi, C. Gaudet-Blavignac, M. Bjelogrlic, C. Lovis, et al., Patient information summarization in clinical settings: scoping review, JMIR Medical Informatics 11 (2023) e44639.

[12] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. URL: http://dx.doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682.

[13] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2020. URL: https://arxiv.org/abs/1904.05342. arXiv:1904.05342.

[14] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics 23 (2022). URL: http://dx.doi.org/10.1093/bib/bbac409. doi:10.1093/bib/bbac409.

[15] S. Rhazzafe, F. Caraffini, S. Colreavy-Donnelly, Y. Dhassi, S. Kuhn, N. S. Nikolov, Hybrid summarization of medical records for predicting length of stay in the intensive care unit, Applied Sciences (2024). URL: https://www.mdpi.com/2076-3417/14/13/5809. doi:10.3390/app14135809.

[16] 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: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 2023, pp. 478–482.

[17] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, L. L. Wang, Ms^2: Multi-document summarization of medical studies, in: EMNLP, 2021.
[18] B. C. Wallace, S. Saha, F. Soboczenski, I. J. Marshall, Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization, AMIA Annual Symposium abs/2008.11293 (2020).
[19] Y. Gao, D. Dligach, T. Miller, M. M. Churpek, 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: Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2023, 2023, p. 461.
[20] Z. Zhao, Q. Jin, F. Chen, T. Peng, S. Yu, Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems, arXiv preprint arXiv:2202.13876 (2022).
[21] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
[22] A. Lavie, A. Agarwal, Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07, Association for Computational Linguistics, USA, 2007, p. 228–231.
[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020.
[24] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, 2021. URL: https://arxiv.org/abs/2106.11520. arXiv:2106.11520.
[25] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704/. doi:10.18653/v1/2020.acl-main.704.
[26] D. Fraile Navarro, E. Coiera, T. Hambly, Z. Triplett, N. Asif, A. Susanto, A. Chowdhury, A. Lorenzo, M. Dras, S. Berkovsky, Expert evaluation of large language models for clinical dialogue summarization, Scientific Reports 15 (2025). doi:10.1038/s41598-024-84850-x.
[27] G. Adams, J. Zucker, N. Elhadad, A meta-evaluation of faithfulness metrics for long-form hospital-course summarization, 2023. URL: https://arxiv.org/abs/2303.03948. arXiv:2303.03948.
[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[29] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[30] T. Nissen, R. Wynn, The clinical case report: a review of its merits and limitations, BMC research notes 7 (2014) 1–7.
[31] 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.
[32] M. Kidd, C. Hubbard, Introducing journal of medical case reports, 2007.
[33] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, M. Villegas, Salamandra technical report, 2025. URL: https://arxiv.org/abs/2502.08489. arXiv:2502.08489.
[34] 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/.

[35] P. Vachharajani., pjmathematician at multiclinsum 2025: A novel automated prompt optimization framework for multilingual clinical summarization., 2025.

[36] J. C. Nicolay Rusnachenko, Xiaoxiao Liu, J. J. Zhang, Using decoder-based distillation for enhancing multilingual clinical case report summarization, 2025.

[37] Y. M. N. Libo Ren, L. Han., Malei at multiclinsum: Summarisation of clinical documents using perspective-aware iterative self-prompting with llms., 2025.

[38] S. C.-D. Soukaina Rhazzafe, N. S. Nikolov., Multiclinsum: Extractive summarization of english, spanish, french and portuguese clinical case reports, 2025.

[39] G. Grazhdanski., Group relative policy optimization for spanish clinical case report summarization., 2025.

[40] J. Angulo, V. Y. Agentic., Agentic mcs: A multilingual clinical summarization framework., 2025.

[41] E. C. P. A. S. B. J. Elisa Terumi Rubel Schneider, Fernando Henrique Schneider, R. M. O. Cruz., Medgemma-sum-pt: A lightweight model for portuguese clinical summarization, 2025.

[42] A. A. R. S. K. Z. Y. Md Shahidul Salim, Lianne Fu, H. Yu., Enhancing multilingual medical summarization via contextual keyword augmentation., 2025.

[43] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, 2025. URL: https://arxiv.org/abs/2412.15115. arXiv:2412.15115.