MedGemma-Sum-Pt: A Lightweight Model for Portuguese Clinical Summarization*

Notebook for the BioASQ Lab at CLEF 2025

Elisa Terumi Rubel Schneider[1,*], Fernando Henrique Schneider[2], Emerson Cabrera Paraiso[2,*], Alceu Souza Britto Jr[2] and Rafael Menelau Oliveira Cruz[1,*]

[1]*École de Technologie Supérieure (ÉTS), University of Quebec, 1100 Notre-Dame Street West, Montreal, Canada*
[2]*Pontifícia Universidade Católica do Paraná (PUCPR), Rua Imaculada Conceição, 1155, Curitiba, Brazil*

Abstract

Automatic summarization of clinical case reports is a challenging yet crucial task to support healthcare professionals in rapidly extracting relevant patient information. The scarcity of large-scale domain-specific datasets and pretrained models, combined with the linguistic complexity of medical texts, makes clinical summarization particularly challenging for low-resource languages like Portuguese and motivates the need for efficient adaptation strategies. This paper presents the approach developed by the ÉTS-PUCPR team for the Portuguese subtask of MultiClinSum, a multilingual clinical summarization shared-task. Our methodology explores (i) zero-shot prompting with general-purpose instruction-following models, and (ii) supervised fine-tuning using LoRA, a parameter-efficient fine-tuning technique, on the biomedical language model MedGemma. We compare these strategies to assess their effectiveness for clinical summarization in Portuguese. Our results demonstrated competitive performance in internal evaluations, particularly when compared to zero-shot baseline performances, showing strong semantic similarity with expert summaries as measured by BERTScore. Despite limitations such as a relatively small training dataset, our findings highlight the potential of fine-tuning domain-specific models under resource constraints for low-resource clinical summarization tasks.

Keywords

Large language model, NLP, summarization, clinical cases, fine-tuning

1. Introduction

The increasing availability of multilingual clinical case reports opens new opportunities for developing automatic summarization systems that can assist healthcare professionals in efficiently accessing critical patient information. However, clinical summarization remains a challenging task. Medical texts are complex, sensitive, and often written in highly variable formats. These challenges are amplified in languages with fewer resources, such as Portuguese, where annotated datasets and domain-adapted models are still scarce [1] [2] .

The MultiClinSum task [3], part of the BioASQ Lab at CLEF 2025 [4], addresses these challenges by evaluating systems for automatic summarization of clinical case reports in multiple languages. The task encourages the development of models capable of generating concise, coherent, and medically meaningful summaries across linguistic contexts.

In this paper, we present the approach developed by the ÉTS-PUCPR team for the Portuguese subtask of the MultiClinSum challenge. We explored two strategies: (i) zero-shot prompting using general-purpose instruction-following models, and (ii) supervised fine-tuning on the clinical summarization data

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
elisa.rubel@pucpr.edu.br (E. T. R. Schneider); paraiso@ppgia.pucpr.br (E. C. Paraiso); alceu@ppgia.pucpr.br (A. S. B. Jr); Rafael.Menelau-Cruz@etsmtl.ca (R. M. O. Cruz)
https://www.linkedin.com/in/elisa-terumi-rubel-schneider/ (E. T. R. Schneider)
https://orcid.org/0000-0002-8921-5598 (E. T. R. Schneider); https://orcid.org/0009-0009-4408-4588 (F. H. Schneider); https://orcid.org/0000-0002-6740-7855 (E. C. Paraiso); https://orcid.org/0000-0002-3064-3563 (A. S. B. Jr); https://orcid.org/0000-0001-9446-1040 (R. M. O. Cruz)

released by the challenge organizers. For the latter, we adopt a parameter-efficient fine-tuning approach using LoRA on top of MedGemma [5], a multilingual biomedical language model. All experiments are conducted in a resource-constrained environment to reflect realistic deployment conditions in clinical or academic settings.

This work introduces `MedGemma-Sum-Pt`, a domain-adapted model for summarization in a low-resource language. Our results align with previous studies showing that fine-tuning on in-domain data, such as clinical narratives in Portuguese, tend to outperform general-purpose models due to linguistic complexity and domain-specific terminology [2] [6] [7].

To foster further research, we release our fine-tuned model, addressing the current gap of summarization models for clinical Portuguese: https://huggingface.co/pucpr-br/medgemma-pt-finetuned-multiclinsum.

2. Task Description

The MultiClinSum task, introduced as part of the BioASQ Lab at CLEF 2025, aims to evaluate automatic summarization systems applied to multilingual clinical case reports. Participants are required to generate concise summaries from real-world clinical narratives written in different languages, including English, Spanish, French, and Portuguese [3].

The task is framed as a summarization problem, where systems must produce short, fluent, and medically accurate summaries that capture the essential information from the original clinical text. Each instance consists of a clinical case report as input and a reference summary manually written by domain experts.

The dataset released for the MultiClinSum task, covering all languages, is divided into three subsets:

- A gold-standard set, comprising 592 clinical case reports with summaries manually written and reviewed by medical experts, which serve as high-quality ground truth for training and evaluation;
- A large-scale set, containing 25,902 clinical cases also accompanied by summaries. However, unlike the gold set, these summaries are not expert-verified and may vary in quality;
- A test set with 3,442 clinical cases, used for the official evaluation of system performance. The summaries in this set are withheld from participants to ensure unbiased assessment.

Evaluation is conducted using a combination of automatic metrics, such as ROUGE and BERTScore.

ROUGE [8] measures the lexical overlap between the generated and reference texts, providing an estimate of informativeness based on shared n-grams. It is defined as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}_{\text{match}}(gram_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}(gram_n)} \tag{1}$$

where n is the length of the n-gram, $\text{Count}_{\text{match}}(gram_n)$ is the maximum number of n-grams co-occurring in both the candidate summary and the reference summaries, and $\text{Count}(gram_n)$ is the total number of n-grams in the reference summaries.

BERTScore [9] evaluates semantic similarity by comparing contextualized token embeddings from a pre-trained transformer model. Given a candidate summary $\hat{x} = \{\hat{x}_1, \ldots, \hat{x}_m\}$ and a reference summary $x = \{x_1, \ldots, x_k\}$, we obtain corresponding normalized embeddings $\{\hat{\mathbf{e}}_j\}$ and $\{\mathbf{e}_i\}$, respectively. BERTScore computes token-level cosine similarity using the inner product of embeddings, and derives precision, recall, and F1 as follows:

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{e}_i^\top \hat{\mathbf{e}}_j \tag{2}$$

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{e}_i^\top \hat{\mathbf{e}}_j \tag{3}$$

$$\text{BERTScore}_{\text{F1}} = \frac{2 \times P_{\text{BERT}} \times R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{4}$$

The multilingual nature of the task also encourages research on cross-lingual and language-specific modeling approaches in the clinical domain.

3. Related Work

Few studies have explored LLM-based summarization of clinical texts in Portuguese, a low-resource language that presents several unique challenges. The scarcity of publicly available clinical corpora, complex morphology, frequent use of jargon, acronyms, and institution-specific abbreviations, as well as dialectal variation, all hinder the development of robust NLP tools for tasks such as summarization [1] [2] [10]. Despite the strong performance of large language models in condensing lengthy medical texts while preserving clinically relevant information, their effectiveness in Portuguese clinical texts remains largely underexplored.

Moreover, Portuguese exhibits rich morphology (e.g., gender and number inflections, dropped pronouns), requiring careful natural language generation. Regional variations between European and Brazilian Portuguese further affect vocabulary and syntax. Despite these challenges, initial efforts have shown that specialized models for clinical Portuguese achieve competitive results, indicating the potential of AI solutions for clinical NLP in Portuguese.

A study [6] investigated abstractive summarization for Brazilian Portuguese using deep learning-based approaches. While their results were promising, the study highlighted ongoing challenges related to coherence, grammar, and fluency. Although their work constitutes a preliminary exploration of abstractive summarization in Brazilian Portuguese using neural models, it did not address domain-specific applications such as clinical narratives. Other studies have also highlighted summarization research for Portuguese texts, such as [11] and [12], though these works focus on general-domain texts rather than clinical narratives.

On the clinical domain, a study on summarization for Brazilian Portuguese [1] explored different approaches applied to electronic health records of chronic disease patients, including an unsupervised neural model based on fine-tuned BERT, as well as supervised methods using sequence labeling and dictionary-based techniques. To reduce redundancy in the generated summaries, a semantic similarity method based on Siamese Neural Networks was employed. Results showed that supervised methods achieved better performance, particularly in preserving clinically relevant information, highlighting the importance of domain-specific resources for effective summarization in Portuguese.

Another study on automatic text summarization in Portuguese [13] compared six algorithms, including classical methods (e.g., Luhn), modern neural models (ChatGPT), and a custom-designed approach called the Marques algorithm. The evaluation, conducted on a COVID-19-related document, revealed that the Marques algorithm outperformed others in precision, coherence, cohesion, and processing time. Although not focused on clinical data, this study highlights important considerations for summarization in Portuguese, such as the benefits of domain-adapted models, the importance of comprehensive evaluation (including human assessment), and the challenges of adapting summarization strategies to different text types.

Although several clinical domain language models have been trained in Portuguese, such as MED-LLM-BR [14], BioBERTpt [2], CardioBERTpt [10], gpt2-bio-pt [7], and DepreBERTBR [15], our investigation did not identify any large language model specifically designed or fine-tuned for the task of clinical summarization in Portuguese.

In this work, we address this gap by fine-tuning MedGemma with LoRA on Portuguese clinical summarization data, demonstrating improved performance in generating semantically aligned summaries under realistic resource constraints. Our approach enables efficient domain adaptation and contributes to the community with a publicly available fine-tuned model for Portuguese clinical summarization.

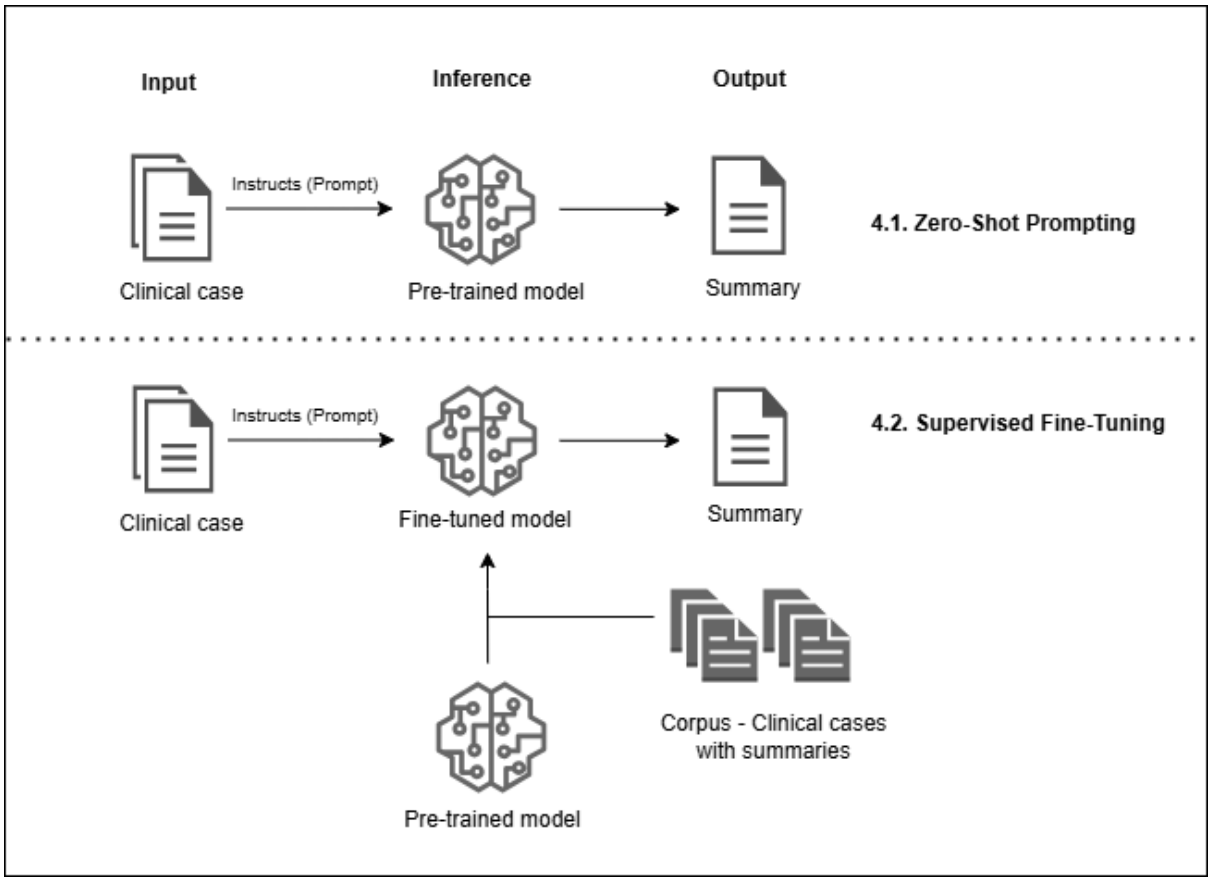

Figure 1: Overview of the two summarization strategies explored in this work: zero-shot prompting and supervised fine-tuning. These correspond to Sections 4.1 and 4.2, respectively.

4. Methodology

To address the Portuguese summarization subtask in MultiClinSum, we experimented with two approaches:

(1) Zero-shot prompting, where the model receives only the input text and an instruction, without seeing any examples; and

(2) Supervised fine-tuning, where the model is explicitly trained on a labeled dataset to optimize its summarization performance.

These strategies are illustrated in Figure 1 and form the basis of the experiments presented in this work[1].

4.1. Zero-Shot Prompting

Our initial experiments focused on evaluating the performance of various open-source language models in a zero-shot setting, to assess their ability to summarize clinical case reports in Portuguese without additional training.

We selected five instruction-following models for this stage, all of which are accessed via the Unsloth repository [16] with 4-bit quantization to enable efficient inference on limited hardware. All selected models are multilingual and capable of understanding and generating text in Portuguese.

- **Gemma-7B-Instruct** [17]: A general-purpose instruction-tuned version of the Gemma 7B model, optimized for dialogue and reasoning tasks.

[1]Few-shot prompting was initially considered but was discarded due to input truncation caused by the limited context window of the models.

- **Llama-3.1-8B-Instruct** [18]: An instruction-tuned variant of LLaMA 3.1 with 8 billion parameters, offering high performance on a wide range of reasoning tasks.
- **Llama-3.2-3B-Instruct** [18]: A lightweight version of LLaMA 3.2 (3B), chosen for its small footprint and fast inference capabilities.
- **Medgemma-4B-Instruct** [5]: A smaller, biomedical-focused variant of Gemma, tuned on clinical and health-related instructions.
- **Qwen2.5-7B-Instruct** [19]: A 7-billion parameter instruction-tuned model based on the Qwen 2.5 architecture, designed for general-purpose reasoning and generation tasks with efficient performance.

These models were selected based on a balance between size, instruction-following capabilities, and accessibility. Our goal was to prioritize lightweight models (3B to 8B parameters) that could be easily deployed on modest hardware, such as edge devices or memory-constrained servers.

Small and mid-sized models typically demonstrate lower performance compared to large-scale models across a wide range of evaluation metrics and benchmarks. In a recent comparison involving 30 benchmark tasks, GPT-4 outperformed lightweight models (with up to 8B parameters) in 26 cases, based on metrics such as ROUGE, accuracy, and task-specific scores [20].

While larger models might offer superior summarization quality, our focus was on finding efficient and practical solutions suitable for researchers and users with limited computational resources, particularly in clinical or academic settings where access to large-scale infrastructure may be constrained.

Smaller models enable real-time summarization directly within healthcare facilities or on portable devices, facilitating faster decision-making and improving patient care without requiring expensive or complex infrastructure.

We used the `FastLanguageModel.from_pretrained` interface from Unsloth to load each model with 4-bit quantization. The models were configured with a context window of 8192 tokens and a generation limit of 512 new tokens. This setup allowed efficient testing without requiring access to high-end GPUs.

4.1.1. Prompt Design

All models were tested using the same prompt, designed to reflect the structure and content expected in a professional clinical summary.

Given the central role of instruction prompts, especially in zero-shot scenarios with general-purpose models, we paid particular attention to the design and refinement of our prompt for this task. In the absence of in-context examples, the prompt alone defines the structure, tone, and granularity of the expected output.

During early experiments, we tested multiple prompt formulations to evaluate their effect on model behavior. Minimal or short prompts (e.g., "Summarize the following clinical case") often produced generic outputs lacking key clinical details. Moreover, without an explicit instruction to reduce the input length, the models tended to generate overly long summaries. The requirement of a 75% reduction in text length helped constrain verbosity and improved focus on relevant findings.

The selected prompt struck a balance between guidance and conciseness. It proved robust across different model architectures and sizes (e.g., 3B to 8B parameters), and consistently improved informativeness and coherence compared to shorter or less explicit alternatives. To emulate expert-authored case summaries, we also encouraged a chronologically coherent narrative that preserves the logical flow of clinical reasoning. We observed that models were particularly sensitive to the presence (or absence) of directives regarding summary length and content types.

The prompt was written in Portuguese to ensure alignment with the language of the input cases:

> Você é um especialista em medicina clínica. Leia o caso clínico e produza um resumo clínico técnico, objetivo, conciso, claro e profissional com as seguintes características:
>
> - Faça uma redução de pelo menos 75% do conteúdo original, mantendo a essência do caso. O texto gerado deve ter entre 5 a 10 frases e entre 50 a 150 palavras.

- Identifique dados principais do paciente (idade, sexo, antecedentes importantes).

- Descreva os sintomas e sinais apresentados, com tempo de evolução.

- Resuma os exames relevantes e os achados principais.

- Aponte as hipóteses diagnósticas e diagnóstico final, se houver.

- Descreva a conduta realizada e a evolução do paciente.

- Escreva um parágrafo coeso e conciso contendo essas informações principais. Não escreva em tópicos, mas em texto fluido.

Escreva com clareza, precisão e fluidez técnica, como um profissional da área da saúde humana, em um texto descritivo com narrativa cronológica simples.

Siga as instruções e forneça o resumo no final.

The English translation of the prompt is provided below for clarity and accessibility.

You are a specialist in clinical medicine. Read the clinical case and produce a technical, objective, concise, clear, and professional clinical summary with the following characteristics:

- Reduce the original content by at least 75%, keeping the essence of the case. The generated text should contain between 5 to 10 sentences and between 50 to 150 words.

- Identify key patient data (age, sex, relevant medical history).

- Describe symptoms and signs presented, including time of onset/evolution.

- Summarize relevant exams and main findings.

- Indicate diagnostic hypotheses and final diagnosis, if available.

- Describe the treatment and the patient's clinical outcome.

- Write a cohesive and concise paragraph containing this core information. Do not use bullet points, but rather a fluid narrative.

Write clearly, precisely, and with technical fluency, as a healthcare professional would, using a simple chronological structure.

Follow the instructions and provide the summary at the end.

The final prompt explicitly instructs the model to produce concise, fluent, and structured summaries emulating clinical documentation style. It specifies essential elements to include, such as patient demographics, symptom evolution, diagnostic reasoning, relevant exams, and treatment outcomes. It also imposes constraints on summary length and discourages bulleted or disjointed text. This design was guided by two main goals: (i) ensuring the inclusion of clinically relevant information; and (ii) enforcing a professional tone and format aligned with medical communication standards.

4.2. Supervised Fine-Tuning

Building on insights from zero-shot evaluation, we fine-tuned MedGemma [5], a domain-adapted language model specialized in clinical and biomedical text generation, to develop a new model named `MedGemma-Sum-Pt`.

The model was fine-tuned to better adapt to the task based on the provided training data and a fixed prompt format. We used the same instruction prompt as in the zero-shot experiment, to ensure that the model internalize the expected output structure, reducing the risk of format mismatch or unpredictable behavior during summarization. The fine-tuning was performed exclusively on the gold-standard dataset provided by the MultiClinSum organizers, containing high-quality expert summaries essential for training accurate clinical summarization models.

Our training strategy utilized a parameter-efficient fine-tuning (PEFT) technique, LoRA (Low-Rank Adaptation) [21], to efficiently update a subset of model parameters. Unlike traditional fine-tuning, which updates all model weights and requires substantial computational resources, LoRA introduces trainable low-rank matrices into specific layers (e.g., attention and language layers) while keeping the original model weights frozen. This significantly reduces the number of trainable parameters and

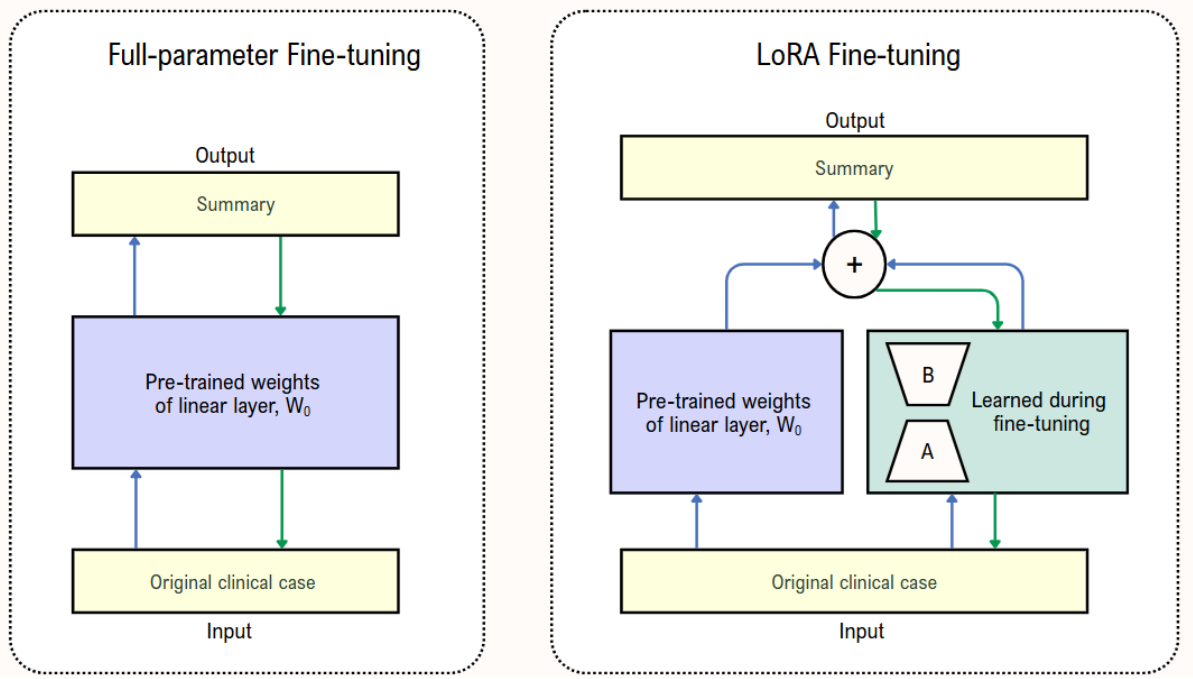

Figure 2: Comparison between full fine-tuning and LoRA adaptation for the summarization task. LoRA achieves comparable performance with fewer trainable parameters and lower memory usage by adapting selected layers through low-rank updates.

accelerates training. As illustrated in Figure 2, this approach enables a compact and efficient adaptation process.

We configured LoRA with a rank of $r = 8$, which balances parameter efficiency and expressiveness, as evidenced by [21]. Adaptation modules were injected into the query and value projection matrices within the attention layers of the transformer, which have been shown to be particularly effective for fine-tuning language models on downstream NLP tasks. Since the task is purely textual, all vision-related components of the architecture were disabled. Additionally, 4-bit quantization was employed to optimize memory usage during training, allowing training on modest hardware without significantly degrading model accuracy.

Training was conducted using the `SFTTrainer` framework [22], with the following hyperparameters: a per-device batch size of 4 with gradient accumulation over 4 steps, a learning rate of `2e-4`, and a total of 3 training epochs (375 optimization steps). To reduce memory consumption, we used the `adamw_8bit` optimizer, combined with a linear learning rate scheduler and a brief warm-up phase of 5 steps. The entire training process took approximately 6.4 hours on a single NVIDIA T4 GPU (16 GB), with a peak memory usage of around 11 GB. This demonstrates the feasibility of fine-tuning using affordable, widely accessible hardware rather than relying on high-end infrastructure.

4.3. Evaluation Strategy

To assess the quality of summaries produced by both zero-shot prompting and fine-tuned generation, we set aside the first 50 examples from our 592-report gold-standard training set. These 50 cases were excluded from fine-tuning and used solely as an internal validation set for comparative evaluation.

We adopted the automatic evaluation metrics proposed by the MultiClinSum task organizers: ROUGE and BERTScore, which provide surface-level and semantic similarity measurements, respectively. For BERTScore, we used the `biobertpt-all` model [2], a Portuguese clinical and biomedical BERT model, ensuring domain and language alignment with the evaluation setting.

In our internal evaluation, we consider the performance of instruction-tuned models in the zero-shot setting as baselines, against which we compare improvements achieved through supervised fine-tuning.

5. Results and Discussion

5.1. Internal Evaluation Results

We evaluated both strategies, zero-shot prompting (our baselines) and supervised fine-tuning (`MedGemma-Sum-Pt`), on our internal validation set of 50 gold-standard examples held out from training. Table 1 summarizes the corresponding ROUGE and BERTScore results.

Table 1
Internal evaluation results on 50 held-out clinical case summaries.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore (F1)
`Gemma-7B-Instruct`	0.4220	0.1915	0.2832	0.6877
`Llama-3.1-8B-Instruct`	0.4363	0.1941	0.2808	0.7002
`Llama-3.2-3B-Instruct`	0.4160	0.1685	0.2664	0.6826
`MedGemma-4B-Instruct`	0.4165	0.1658	0.2650	0.6946
`Qwen2.5-7B-Instruct`	0.4175	0.1613	0.2662	0.6879
`MedGemma-Sum-Pt`	**0.5009**	**0.2979**	**0.3768**	**0.7158**

As we can see from the internal evaluation results, the zero-shot models achieved reasonable performance when guided by a well-crafted prompt, especially the `Llama-3.1-8B-Instruct`, which slightly outperformed others in terms of ROUGE and BERTScore. Despite the `Llama-3.1-8B-Instruct` model's slightly better zero-shot performance, we chose to fine-tune `MedGemma` due to its smaller size (4B) and native biomedical specialization. Our goal was to develop a lightweight yet effective model that could be fine-tuned and deployed with limited computational resources. Even without fine-tuning, `MedGemma` showed competitive results, and its domain adaptation made it a natural candidate for clinical tasks.

While the zero-shot models demonstrated competitive results with well-designed prompts, our fine-tuned `MedGemma-Sum-Pt` model consistently outperformed all baselines across all metrics. The improvement in ROUGE-2 and ROUGE-L scores suggests that fine-tuning helped the model better capture both local and structural coherence of the summaries. Likewise, the increase in BERTScore F1 demonstrates an enhanced semantic alignment with the expert references.

To illustrate, we selected an example from our test set in which the fine-tuned model achieved 0.43 in ROUGE-L and 0.776 in BERTScore, while the zero-shot model obtained 0.26 in ROUGE-L and 0.668 in BERTScore (as shown in Figure 3). In this case, the fine-tuned model demonstrated superior lexical and semantic similarity to the reference summary. Qualitatively, the summary generated by the fine-tuned model more accurately preserved essential clinical information, such as the patient's age, renal transplant history, absence of CMV infection prior to the episode, decreased visual acuity in the left eye, and treatment with ganciclovir. In contrast, the zero-shot model included content not present in the reference summary, such as bilateral eye involvement (whereas the summary refers only to the left eye) and herpes simplex infection. In this example, the fine-tuned model produced a more concise and focused summary that preserved the case's key points, which explains its higher metric scores.

Across other examples, common zero-shot errors included additional clinical information not present in the reference summaries, omission of relevant lab findings, and inconsistent clinical timelines. In contrast, the fine-tuned model showed higher fidelity to the reference data.

This stage helped us identify the limitations of zero-shot summarization with compact models, motivating the transition to a fine-tuning approach. Based on these findings, we selected the `MedGemma-Sum-Pt` model, our best-performing approach, for the official submission to the MultiClinSum task.

5.2. Official Results

In this section, we present the official results of our system on the shared task.

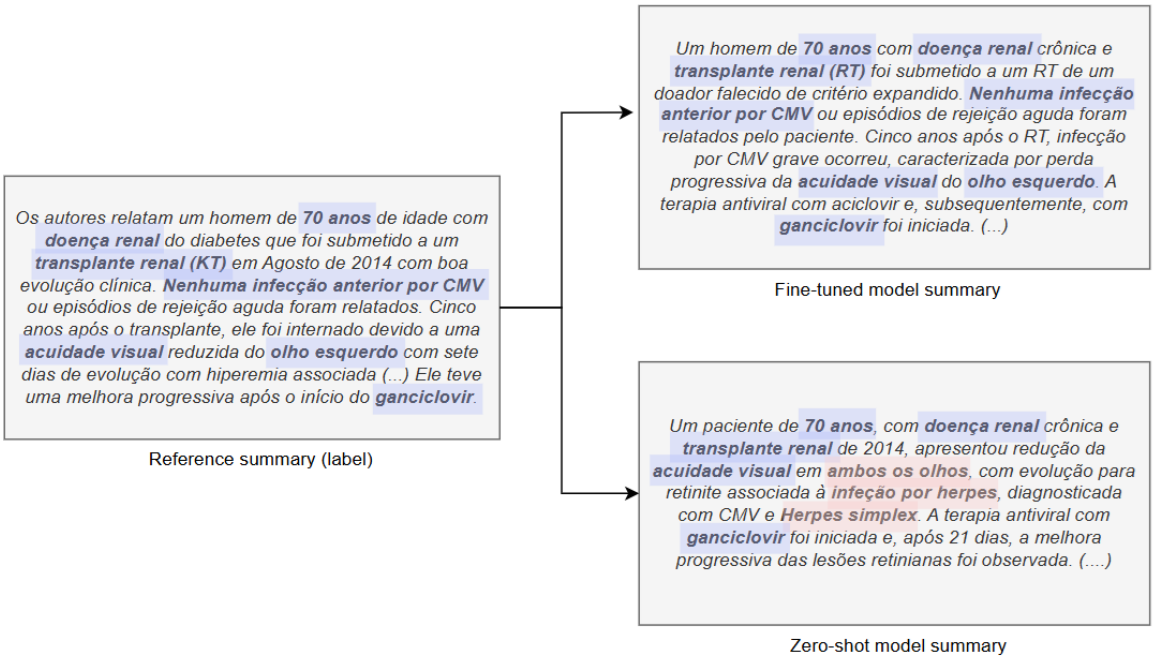

Figure 3: Comparison between summaries generated by the fine-tuned and zero-shot models for a selected clinical case in Portuguese. Clinically relevant information preserved from the reference summary is highlighted in blue. Content generated by the model that does not appear in the reference is highlighted in red.

Based on our internal evaluation, we selected our fine-tuned `MedGemma-Sum-Pt` model as our official submission to the Portuguese track of the MultiClinSum task. The model demonstrated the most consistent performance across all evaluation metrics when compared to zero-shot baselines.

The submitted system was evaluated on the hidden test set by the task organizers using the official metrics: ROUGE and BERTScore. Our official results are shown in Table 2.

Table 2
Official competition evaluation results for the ÉTS-PUCPR team.

Metric	Value
ROUGE-F1	0.2490
ROUGE Precision	0.2802
ROUGE Recall	0.2590
BERTScore-F1	0.7370
BERTScore Precision	0.7403
BERTScore Recall	0.7351

These results show that the model maintained a strong semantic alignment with the gold summaries (as reflected by the BERTScore), although lexical overlap (measured by ROUGE) was more modest. This suggests frequent paraphrasing or alternative expressions, which is common in clinical language. Since ROUGE relies on n-gram overlap, it may underestimate quality when summaries are semantically correct but lexically diverse.

At the time of writing, comparative results from other participants have not been made publicly available, which limits a more detailed contextualization of our system's performance within the competition.

5.3. Discussion

Our findings suggest that even small instruction-tuned models can perform competitively in clinical summarization tasks when fine-tuned with domain-specific data and guided by carefully crafted

prompts. The MedGemma-Sum-Pt model delivered strong semantic results with minimal infrastructure, highlighting the feasibility of low-cost clinical NLP solutions.

Our model demonstrated close alignment with expert-authored summaries in terms of meaning and clinical relevance, as indicated by the high BERTScore. This metric captures meaning similarity beyond exact word matches, suggesting that MedGemma-Sum-Pt was able to convey clinical information reliably even when using different phrasing. The BERTScore results include precision, recall, and F1 metrics, which together provide a more nuanced understanding of model performance. In our results, the close values of precision and recall suggest that MedGemma-Sum-Pt maintains a good balance between generating accurate and comprehensive clinical summaries.

In contrast, ROUGE scores were more modest, reflecting less lexical overlap. This discrepancy may be explained by the abstractive nature of the model's outputs, which often generate paraphrases or reformulations that preserve meaning but diverge lexically from the reference summaries. Additionally, the model may have learned to prioritize semantic coherence over surface-level n-gram overlap, which aligns with the goals of clinical summarization but can penalize ROUGE scores. This behavior is consistent with findings in the literature, such as [23], where abstractive models evaluated on clinical texts in English showed similarly low ROUGE scores despite strong semantic fidelity as captured by BERTScore. The persistence of this pattern in our Portuguese-language setting suggests that the trade-off between lexical overlap and semantic alignment may be a broader characteristic of abstractive summarization.

Furthermore, these findings highlight the importance of combining lexical and semantic metrics for evaluating clinical summarization, where preserving meaning is paramount.

Our findings also underscore the central role of instruction prompts in clinical summarization with language models. The prompt used in this work was carefully crafted to guide model behavior and ensure consistency, particularly in zero-shot settings. While effective, this manually designed prompt may limit generalization to other formats or clinical domains. Future work should explore systematic prompt optimization or adaptation techniques, including prompt tuning and evaluation of diverse prompt formulations, to reduce reliance on handcrafted instructions and improve robustness across diverse inputs.

Overall, this work presents a practical approach to clinical summarization in Portuguese using parameter-efficient fine-tuning. While our setup prioritizes accessibility and resource efficiency, further studies comparing performance with larger models are needed to fully assess trade-offs in summarization quality. These findings contribute to ongoing efforts toward scalable, language-specific NLP solutions in the clinical domain.

6. Conclusions

This work presented a practical approach for clinical summarization in Portuguese by fine-tuning a compact, biomedical-specialized model, MedGemma. We showed that even with limited computational resources and a relatively small training dataset, it is possible to achieve competitive results, particularly in semantic alignment with expert summaries.

Our internal evaluation indicated the advantage of fine-tuning over zero-shot strategies, highlighting the importance of domain adaptation for clinical NLP tasks. Furthermore, the combined use of BERTScore and ROUGE metrics provided a complementary view of summary quality, capturing both semantic fidelity and lexical overlap.

Our participation in the MultiClinSum challenge underscores the potential of compact, domain-adapted models for multilingual clinical summarization and reinforces their viability for real-world deployment in medical settings.

7. Limitations

We acknowledge limitations related to the size of the dataset used for training and evaluation scope. Future work should explore larger datasets and more comprehensive external evaluations. Specifically, we plan to incorporate the larger Portuguese clinical dataset (25,902 examples) into the training process to assess whether a broader data foundation improves summarization quality and robustness. Moreover, while our results show promise, the current evaluation does not include direct comparisons with other participating systems in the MultiClinSum challenge. As official rankings and results become available, we intend to conduct a more thorough comparative analysis to better contextualize our model's performance.

We also observed sensitivity to prompt phrasing during generation, highlighting the need for systematic prompt optimization strategies or prompt tuning techniques to reduce variability in output quality. This issue was particularly evident in our zero-shot experiments.

Acknowledgments

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq), under Project 441610/2023-4. And the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant program (RGPIN-2021-04130).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT to assist with grammar and spelling checks. After using this service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. E. S. e. Oliveira, C. M. C. M. Barra, S. S. A. Hasan, Assembling natural language processing resources to perform the summarization of clinical narratives, Tese (doutorado), Pontifícia Universidade Católica do Paraná, Curitiba, 2020. URL: https://archivum.grupomarista.org.br/pergamumweb/vinculos/000094/0000943a.pdf, acesso em: 21 dez. 2020.

[2] E. T. R. Schneider, J. V. A. de Souza, J. Knafou, L. E. S. e. Oliveira, J. Copara, Y. B. Gumiel, L. F. A. d. Oliveira, E. C. Paraiso, D. Teodoro, C. M. C. M. Barra, BioBERTpt - a Portuguese neural language model for clinical named entity recognition, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 65–72. URL: https://www.aclweb.org/anthology/2020.clinicalnlp-1.7.

[3] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Gimenez, 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: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[4] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, F. N. Maria Di Nunzio, Giorgio, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[5] Google, Medgemma hugging face, https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4, 2025. Accessed: 2025-06-16].

[6] P. H. Paiola, Sumarização abstrativa de textos em português utilizando aprendizado de máquina, Dissertação de mestrado, Universidade Estadual Paulista (Unesp), Bauru, SP, 2022. URL: http://hdl.handle.net/11449/236858, orientador: João Paulo Papa.

[7] E. T. R. Schneider, J. V. A. de Souza, Y. B. Gumiel, C. Moro, E. C. Paraiso, A gpt-2 language model for biomedical texts in portuguese, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), 2021, pp. 474–479. doi:10.1109/CBMS52027.2021.00056.

[8] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[9] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[10] E. T. R. Schneider, Y. B. Gumiel, J. V. A. de Souza, L. Mie Mukai, L. Emanuel Silva e Oliveira, M. de Sa Rebelo, M. Antonio Gutierrez, J. Eduardo Krieger, D. Teodoro, C. Moro, E. C. Paraiso, Cardiobertpt: Transformer-based models for cardiology language representation in portuguese, in: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), 2023, pp. 378–381. doi:10.1109/CBMS58004.2023.00247.

[11] T. A. S. Pardo, Sumarização automática: principais conceitos e sistemas para o português brasileiro, 2008.

[12] O. d. Souza, H. R. Tabosa, D. M. d. Oliveira, M. H. d. S. Oliveira, Um método de sumarização automática de textos através de dados estatísticos e processamento de linguagem natural, Informação amp;amp; Sociedade: Estudos 27 (2017). URL: https://periodicos.ufpb.br/index.php/ies/article/view/32571.

[13] M. M. R. Silva, Sumarização automática de textos: Desafios e avanços em técnicas de processamento de linguagem natural, 2023. Monografia de Graduação.

[14] J. G. de Souza Pinto, A. R. de Freitas, A. C. G. Martins, C. M. R. Sawazaki, C. Vidal, L. E. S. e Oliveira, Developing resource-efficient clinical llms for brazilian portuguese, in: Proceedings of the 34th Brazilian Conference on Intelligent Systems (BRACIS), 2024. In press.

[15] A. D. R. Herculano, DepreBERTBR: um modelo de linguagem pré-treinado para o domínio da depressão no idioma português brasileiro, 2024. Accessed: 2025-06-17.

[16] M. H. Daniel Han, U. team, Unsloth, 2023. URL: http://github.com/unslothai/unsloth.

[17] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. C. et al., Gemma: Open models based on gemini research and technology, 2024. URL: https://arxiv.org/abs/2403.08295. arXiv:2403.08295.

[18] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. K. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[19] Qwen, A. Yang, B. Yang, B. Zhang, B. H. et al., Qwen2.5 technical report, 2025. URL: https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

[20] J. Zhao, T. Wang, W. Abid, G. Angus, A. Garg, J. Kinnison, A. Sherstinsky, P. Molino, T. Addair, D. Rishi, Lora land: 310 fine-tuned llms that rival gpt-4, a technical report, 2024. URL: https://arxiv.org/abs/2405.00732. arXiv:2405.00732.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[22] H. Face, Sfttrainer: Supervised fine-tuning for language models, 2023. URL: https://huggingface.co/docs/trl/main/en/sft_trainer, accessed: 16 June 2025.

[23] C. Izuchukwu, H. Wimmer, C. Redman, A comparison of large language models for oncology clinical text summarization, Journal of Information Systems Applied Research and Analytics (JISARA) 18 (2025) 20–29. ISSN: 1946-1836.

A. Online Resources

To support further research in clinical natural language processing, particularly within Portuguese-language healthcare contexts, our `MedGemma-Sum-Pt` model is publicly available on Hugging Face at https://huggingface.co/pucpr-br/medgemma-pt-finetuned-multiclinsum, released under a research-only license.

As it was built upon the original MedGemma model, its use is restricted to academic and non-commercial purposes, in accordance with the original licensing constraints.