UMUTeam at BioASQ 2026: Evaluating Medical-Domain Continual Pretraining for Spanish Clinical Case Report Summarization

BioASQ at CLEF 2026

Tomás Bernal-Beltrán[1], Jorge Gómez-Navalón[1], Ronghao Pan[1], Pedro José Vivancos-Vicente[2] and José Antonio García-Díaz[1,*]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*
[2]*VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia, Carretera de Madrid km 388. Complejo de Espinardo, Murcia, Spain*

Abstract

This paper describes the UMUTeam system submitted to the Spanish subtrack of the MultiClinSum-2 shared task, which is part of the BioASQ workshop organized in the context of CLEF 2026. The task focuses on the automatic summarization of long clinical case reports, requiring systems to generate concise summaries that preserve relevant diagnostic information, clinical findings, treatments, patient evolution and outcomes. Our approach investigates whether medical-domain continual pretraining improves Spanish clinical summarization after supervised instruction tuning. To this end, we compare several pairs of Large Language Models, where each pair consists of a base model and a medically adapted counterpart developed by our research group through continual pretraining on Spanish medical-domain corpora. All models are instruction-tuned using the same Spanish training data, prompt template and experimental configuration, enabling a controlled comparison between base and medical variants. Development results show that the impact of medical-domain continual pretraining is not uniform across model families and sizes. Strong multilingual base models already achieve competitive performance after instruction tuning, but medical variants improve over their base counterparts in specific cases, particularly for Gemma 3 1B and Salamandra 7B. In the official evaluation, Salamandra 7B Medical obtains the best ROUGE-based scores among our submitted runs, supporting the idea that Spanish medical-domain pretraining can be beneficial when combined with a model family well aligned with the target language.

Keywords

Clinical Summarization, Clinical NLP, Instruction Tuning, Continual Pretraining, Large Language Models, Natural Language Processing

1. Introduction

Clinical documents contain essential information for diagnosis, treatment, follow-up and communication between healthcare professionals. However, they are often long, heterogeneous and difficult to review efficiently, since relevant information may be distributed across different sections, time points or narrative fragments. Automatic clinical text summarization aims to reduce this information burden by generating concise summaries that preserve the most relevant medical content [1]. This is particularly important for Electronic Health Records (EHRs), discharge documentation and clinical case reports, where clinicians and researchers need to quickly access the key facts of complex medical histories [2].

Despite its practical relevance, clinical summarization remains a challenging task. Clinical texts contain specialized terminology, abbreviations, implicit temporal relations and domain-specific writing conventions, which make them difficult to process with general-domain summarization systems [2]. In

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
tomas.bernalb@um.es (T. Bernal-Beltrán); jorge.gomezn@um.es (J. Gómez-Navalón); ronghao.pan@um.es (R. Pan); pedro.vivancos@vocali.net (P. J. Vivancos-Vicente); joseantonio.garcia8@um.es (J. A. García-Díaz)
https://www.vocali.net/ (P. J. Vivancos-Vicente)
0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0005-7185-6054 (J. Gómez-Navalón); 0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz)

addition, summarization systems must not only produce fluent text, but also preserve factual correctness, avoid unsupported statements and retain clinically relevant details. These requirements make clinical summarization especially demanding, since omissions, contradictions or hallucinated information may reduce the usefulness and reliability of the generated summaries [3]. The challenge is even greater in languages other than English, where fewer annotated datasets, benchmarks and domain-adapted models are usually available [4].

The MultiClinSum-2 shared task [5], which is part of the BioASQ workshop [6] organized in the context of CLEF 2026, addresses this problem by focusing on the automatic summarization of lengthy clinical case reports across 15 languages. Given a full clinical case report, participating systems must generate a concise summary that preserves the most relevant diagnostic information, medical concepts and clinical details. Each language is evaluated as an independent subtrack, allowing participants to develop either language-specific or multilingual approaches. In this work, we focus on the Spanish subtrack, which provides a suitable setting to study clinical summarization in a non-English language where both medical-domain understanding and language-specific generation quality are required. The task is evaluated using automatic metrics such as ROUGE-2 and BERTScore, complemented by LLM-as-a-judge assessment for qualitative evaluation. This setting is particularly appropriate for analyzing whether Spanish medical-domain pretraining provides an advantage over the corresponding base models after instruction tuning.

Recent work has shown that Transformer-based and Large Language Model (LLM) based approaches can achieve strong results in clinical summarization, especially when models are adapted to the biomedical or clinical domain [7]. More broadly, domain-adaptive pretraining has been shown to improve downstream performance in specialized domains [8]. However, recent evaluations also suggest that biomedical adaptation does not always lead to consistent gains over strong general-purpose models across unseen clinical tasks [9]. This makes controlled comparisons between base and medically adapted models particularly relevant, especially in non-English clinical summarization scenarios where the impact of medical-domain pretraining is still less explored.

In this paper, we present the system developed by UMUTeam for the Spanish subtrack of the MultiClinSum-2 shared task. Our participation is motivated by a broader objective: evaluating a set of medical LLMs developed by our research group against their corresponding base versions in a controlled clinical summarization setting. These medical models were obtained by applying continual pretraining to the original base models using Spanish medical-domain corpora, with the aim of incorporating biomedical knowledge and improving their familiarity with Spanish clinical and medical discourse, including clinical practice guidelines, the Corpus Web Salud Español (CoWeSe) [10] and the Spanish Medical Corpus (SMC) [11] datasets. This paper focuses on their downstream evaluation in the shared task, while Section 3 and Section 4 summarize the main characteristics of the medical adaptation stage needed to interpret the comparison.

To this end, we compare each base model with its medically adapted counterpart under the same experimental conditions. Both variants are instruction-tuned on the Spanish training data provided by the task, using the same training configuration within each model pair. This design controls for model architecture, model size and task-specific training data, allowing us to analyze whether medical-domain continual pretraining provides a better initialization for clinical case report summarization after supervised instruction tuning.

Overall, our experiments provide a controlled comparison across several model families and sizes, including Qwen 3, Gemma 3 and Salamandra models. By evaluating each base model alongside its medically adapted counterpart under the same instruction-tuning conditions, this work aims to assess the practical impact of Spanish medical-domain continual pretraining on clinical case report summarization. This analysis contributes to understanding when domain-adapted LLMs may offer advantages over strong general-purpose models in non-English clinical NLP scenarios.

The remainder of this paper is organized as follows. Section 2 reviews related work on clinical text summarization, biomedical language models, instruction tuning and domain-adaptive pretraining. Section 3 describes the proposed methodology, including the evaluated model pairs and the instruction-tuning strategy. Section 4 outlines the dataset, training configuration and evaluation procedure. Section

5 presents and discusses the results obtained. Finally, Section 6 concludes the paper and outlines potential directions for future work.

2. Background Information

Automatic clinical text summarization has been studied for more than a decade as a way to reduce the information burden associated with EHRs, discharge documentation, clinical case reports and other forms of longitudinal medical text. Clinical documents are often long, heterogeneous and redundant, and relevant information may be distributed across different sections, time points or narrative fragments. Early reviews already identified patient-level summarization as a central challenge for clinical Natural Language Processing (NLP), since clinicians often need concise and reliable overviews of complex medical histories rather than isolated document-level predictions [1]. More recent systematic reviews have confirmed that biomedical and clinical summarization has become an increasingly active research area, while highlighting persistent challenges related to data availability, factual consistency, domain-specific terminology and evaluation [2].

Initial approaches to clinical summarization were mainly extractive, selecting relevant sentences, concepts or sections from the source document through rule-based, statistical or supervised methods. This was useful in the clinical domain because reusing source text helped preserve medical wording and reduced the risk of unsupported generation. However, extractive systems were limited in their ability to synthesize dispersed information, remove redundancy and produce coherent summaries adapted to a specific purpose. Early supervised systems already showed that clinical summarization could benefit from learning-based methods [12], but their outputs remained constrained by the extractive nature of the approach and by the need for annotated data.

Neural sequence-to-sequence models shifted summarization toward abstractive generation, where systems synthesize the input instead of selecting source spans from it. This transition was accelerated by the introduction of the Transformer architecture [13], which became the basis for modern NLP. Encoder-only models such as BERT improved contextual representation learning [14], while encoder-decoder models such as BART [15], T5 [16] and PEGASUS [17] established pre-trained generation models as strong baselines for abstractive summarization. However, applying general-domain summarization models to clinical text remains challenging due to specialized terminology, abbreviations, temporal information and strict factuality requirements.

Transformer-based generative models have also been applied to clinical summarization, especially in the context of discharge summaries. In this setting, one of the main challenges is that the relevant information is often distributed across long clinical narratives, making it difficult for models to process the full input directly. For this reason, previous work has explored different ways of organizing the source documents before generation. For example, chronological segmentation of the hospital course has been shown to be more effective than simply truncating the input to fit the model context window [18]. Subsequent work showed that encoder-decoder Transformer models can generate hospital course summaries of promising quality, although factuality and clinical reliability remain important limitations [19]. Overall, these studies suggest that clinical summarization depends not only on the chosen model, but also on how the clinical text is preprocessed and presented to the model.

At the same time, the specificity of biomedical and clinical language motivated the adaptation of pre-trained language models to the medical domain. In the encoder-based setting, BioBERT [20] demonstrated that further pretraining BERT on biomedical corpora improves performance on biomedical NLP tasks, while ClinicalBERT [21] showed the benefits of pretraining contextual representations on clinical notes. SciBERT [22] extended this idea to scientific text more broadly, showing that in-domain pretraining can improve performance across scientific NLP tasks. Although these models were not primarily designed for summarization, they helped consolidate the idea that domain-specific pretraining can provide a stronger starting point for downstream biomedical and clinical NLP.

The emergence of LLMs has further changed clinical summarization, since these models combine strong generation capabilities, broad linguistic coverage and the ability to follow natural-language

instructions. Adapted LLMs have been evaluated in tasks such as radiology report, patient question, progress note and clinical dialogue summarization. Previous work showed that adapted LLMs can outperform medical experts on several clinical summarization tasks [7]. However, fluent summaries are not necessarily clinically reliable, as they may omit relevant information, introduce unsupported statements or fail to preserve clinically important details [3, 23].

The growth of LLM-based clinical summarization has also been accompanied by the development of more standardized benchmarks. Clinical dialogue summarization, for instance, has received increasing attention through ACI-Bench [24], which focuses on automatic visit note generation from clinical conversations, and the MEDIQA-Chat shared task [25], which addressed the generation of structured clinical notes from doctor-patient conversations. Although these benchmarks differ from clinical case report summarization, they share the same goal of compressing complex clinical information into faithful, useful and well-structured summaries, while also showing the importance of shared tasks and standardized evaluation settings.

A related development is the use of instruction tuning to adapt LLMs to biomedical and clinical tasks. General instruction-tuned models are trained to follow natural-language task descriptions, improving their ability to generalize across tasks and output formats [26, 27]. In the biomedical domain, BioInstruct [28] showed that instruction tuning with biomedical instructions can improve performance in question answering, information extraction and generation tasks. Previous work has also shown that task formulation, instruction diversity and domain relevance can strongly influence downstream performance [29]. For clinical summarization, this paradigm is particularly suitable because the task can be naturally expressed as an instruction: given a clinical report, generate a concise summary that preserves the relevant medical information. Instruction tuning therefore provides a direct mechanism for teaching LLMs the expected summarization behavior, output style and domain-specific constraints.

Domain-adaptive continual pretraining is another common strategy for adapting language models to specialized domains. Gururangan et al. showed that continuing pretraining on domain-specific data, and even on task-specific unlabeled data, can improve downstream performance across different domains [8]. In biomedical and clinical NLP, this has motivated the development of models trained or further pretrained on medical corpora before task-specific fine-tuning. For generative LLMs, this adaptation is often combined with instruction tuning: medical-domain pretraining increases the model's familiarity with biomedical terminology and discourse, while instruction tuning aligns the model with specific tasks and output formats. PMC-LLaMA [30], for example, follows this direction by combining medical-domain pretraining with instruction tuning to build open-source medical LLMs.

However, the benefit of biomedical or clinical adaptation is not guaranteed. Recent evaluations suggest that domain-specific fine-tuning can improve performance on some medical tasks, but may not always generalize better than strong general-purpose models across unseen clinical scenarios [9]. This makes controlled comparisons between base models and medically adapted counterparts especially important. When both models share the same architecture and are instruction-tuned under the same conditions, performance differences can be more directly attributed to the effect of medical continual pretraining. This is particularly relevant for clinical summarization, where improvements may depend not only on recognizing medical terminology, but also on selecting, compressing and verbalizing clinically relevant information.

Most clinical summarization research has focused on English, largely because English has more publicly available clinical datasets and biomedical resources. In contrast, Spanish clinical NLP has historically faced stronger limitations in terms of annotated corpora, domain-specific resources and public benchmarks. Nevertheless, recent work has improved the Spanish biomedical and clinical NLP ecosystem. Carrino et al. introduced large-scale biomedical language models for Spanish and showed that domain-specific pretraining improves performance on clinical NLP tasks [4]. More recent comparative studies of Spanish clinical encoders have further confirmed the importance of evaluating domain-adapted models under controlled conditions [31]. In the generative setting, Medical mT5 extended domain adaptation to a multilingual text-to-text architecture trained on medical corpora in English, Spanish, French and Italian, showing that medical-domain pretraining can also benefit non-English generation tasks [32].

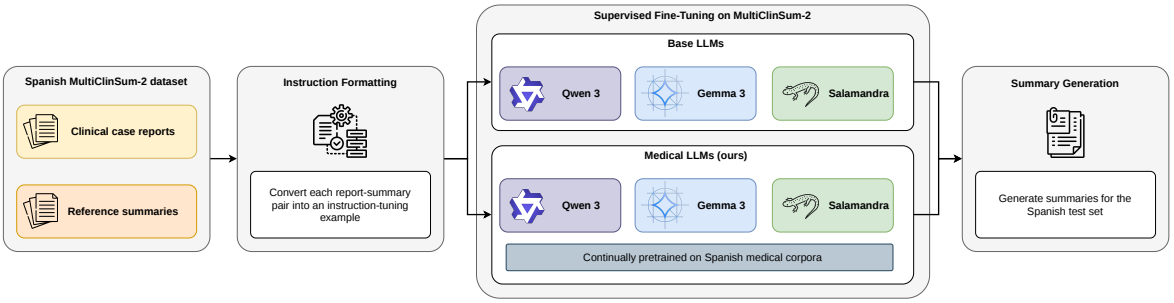

Figure 1: Overview of the proposed system. Spanish clinical case reports and their reference summaries are converted into instruction-tuning examples and used to fine-tune both base LLMs and their medically adapted counterparts. The resulting models generate summaries for the test set, which are then evaluated and compared to analyze the impact of medical-domain continual pretraining.

Multilingual clinical summarization benchmarks have recently begun to address this gap. MultiClinSum [33] introduced a shared task for summarizing full clinical case reports in multiple languages, including English, Spanish, French and Portuguese. This task is especially relevant because it moves beyond English discharge summaries and clinical dialogues, focusing instead on multilingual clinical case reports. For Spanish, this benchmark provides a valuable opportunity to evaluate whether medical-domain adaptation improves summarization in a setting where both clinical terminology and language-specific generation quality matter.

3. System Overview

The present work is positioned at the intersection of clinical summarization, biomedical instruction tuning and domain-adaptive continual pretraining. Instead of proposing a new summarization architecture, we focus on a controlled comparison between base LLMs and medically adapted versions of the same models. All models are instruction-tuned on the Spanish subset of the MultiClinSum-2 shared task, allowing us to study whether continual pretraining on Spanish medical-domain data provides a better initialization for clinical case report summarization. This design follows the broader literature on domain-adaptive pretraining and biomedical instruction tuning, while addressing a less explored setting: Spanish clinical case report summarization with paired base and medical LLMs.

Figure 1 provides an overview of the proposed system. The pipeline is designed to evaluate whether medical-domain continual pretraining improves the performance of LLMs on Spanish clinical case report summarization after supervised instruction tuning. To this end, each base model is paired with a medically adapted version developed by our research group, and both variants are fine-tuned and evaluated under the same conditions using the Spanish data provided by the shared task. The system follows a controlled pipeline: data preparation, instruction formatting, supervised fine-tuning, summary generation and base-versus-medical comparison.

The first stage of the pipeline consists of preparing the Spanish subset of the task dataset. Each instance consists of a full clinical case report paired with its corresponding reference summary. The source report and its summary are converted into an instruction-following example, where the model is asked to generate a concise and faithful summary of the clinical case while preserving the relevant medical information. The training data are split into training and development partitions, which are used for instruction tuning and model selection, respectively.

The second stage consists of fine-tuning the selected model pairs. For each architecture, we compare the original base model with its medically adapted counterpart. The medical variants were obtained through causal language-model continual pretraining on Spanish medical-domain corpora, including clinical practice guidelines, CoWeSe and SMC datasets, whereas the base versions correspond to the original general-purpose checkpoints. Both variants in each pair are then trained using the same Spanish MultiClinSum-2 data, prompt template and training configuration. This controlled setup is

Table 1
Statistics of the Spanish MultiClinSum-2 dataset used in our experiments. Docs. denotes the number of documents, Rep. W. and Rep. Ch. denote the average number of words and characters in the source reports, respectively, and Sum. W. and Sum. Ch. denote the average number of words and characters in the reference summaries. The official evaluation set does not include reference summaries.

Split	Docs.	Rep. W.	Rep. Ch.	Sum. W.	Sum. Ch.
Official partitions					
Official Training Set	25902	627.74	3996.7	114.97	752.07
Official Test Set	3406	623.19	3967.39	115.23	752.69
Official Evaluation Set	1000	675.41	4321.22	–	–
Internal split of the official training set					
Internal Training Set	20721	628.44	4000.84	115.05	752.94
Internal Development Set	5181	624.96	3980.16	114.64	748.58

intended to isolate, as much as possible, the effect of medical-domain continual pretraining.

After training, the resulting models are used to generate summaries for the Spanish test documents. At inference time, each clinical case report is inserted into the same instruction template used during training, and the model generates the corresponding summary. No additional retrieval component, external knowledge source or post-generation rewriting module is used. Therefore, the generated summaries depend only on the information contained in the input report and on the knowledge encoded in each fine-tuned model.

The final stage consists of evaluating and comparing the generated summaries. Beyond the absolute performance of each model, our main analysis focuses on the comparison between each base model and its medically adapted counterpart. This allows us to study whether Spanish medical-domain continual pretraining provides a consistent advantage for clinical case report summarization after both models have been instruction-tuned on the same task data.

The system was implemented using the HuggingFace Transformers ecosystem. Supervised fine-tuning was carried out with TRL, while PEFT was used to train the models with LoRA adapters in a parameter-efficient setting. The clinical reports were minimally preprocessed to preserve the original medical content while ensuring a consistent input format. The same instruction template was applied during training and inference, so that test reports were presented to the models in the same format used for supervised fine-tuning. Further details about the training configuration, quantization setup, LoRA parameters and decoding strategy are provided in Section 4.

4. Experimental Setup

Our experiments were conducted on the Spanish subset of the MultiClinSum-2 shared task dataset. The dataset includes three official partitions: a training set, a test set with reference summaries and a blind evaluation set for the final submission. Each example in the training and test sets consists of a full clinical case report paired with a reference summary, whereas the evaluation set only contains the source reports. The official training set was further divided into internal training and development partitions, which were used for instruction tuning and model selection, respectively. Since the task focuses on summarizing long clinical narratives, Table 1 reports the number of documents and the average length of the source reports and reference summaries for both the official partitions and the internal training/development split.

As shown in Table 1, the source reports are considerably longer than the reference summaries. On average, the official training and test reports contain 627.74 and 623.19 words, respectively, whereas their summaries contain 114.97 and 115.23 words. This reflects the compression-oriented nature of the task, where models must identify the most relevant clinical information from lengthy case reports and generate concise summaries without losing essential medical content. The official evaluation set is

slightly longer on average, with 675.41 words per report, but remains within a comparable length range. The internal training and development partitions preserve similar length distributions to the official training set, making the development split suitable for monitoring model behavior before generating summaries for the official test and evaluation sets.

The experimental comparison includes six pairs of LLMs. Each pair consists of a base model and a medically adapted version of the same model developed by our research group (UMUTeam) through continual pretraining on Spanish medical-domain corpora. We evaluated two models from the Qwen 3 family [34], namely the 4B and 8B variants, which belong to a multilingual family of dense and mixture-of-experts LLMs. We also included two Gemma 3 models [35], the 1B and 4B variants, as representatives of Google's family of lightweight open models. Finally, we evaluated Salamandra 2B and Salamandra 7B, two models from a multilingual family specifically designed with strong support for European languages [36]. This paired design allows us to compare models with the same architecture and size while changing only whether they were previously exposed to Spanish medical-domain data before task-specific instruction tuning.

All models were adapted to the task using supervised instruction tuning. Each training instance was formatted as a Spanish instruction-following example, where the model was asked to generate a concise and accurate summary of the clinical case while preserving the essential medical information. A single prompt template was used for all models in order to keep the comparison between base and medically adapted variants controlled. Listing 1 shows the original Spanish prompt template used in our experiments.

Listing 1: Original Spanish prompt template used for instruction tuning and inference.

```
Eres un asistente especializado en resumen clínico.

Resume el siguiente caso clínico en español. El resumen debe ser conciso,
clínicamente preciso y conservar la información médica esencial: diagnóstico,
hallazgos clínicos relevantes, pruebas, tratamientos, evolución y desenlace.

Caso clínico:
{fulltext}

Resumen:
```

For readability, Listing 2 shows the English translation of the prompt. This translation is provided only to help international readers understand the instruction; all experiments were conducted using the original Spanish prompt shown in Listing 1.

Listing 2: English translation of the original Spanish prompt template.

```
You are an assistant specialized in clinical summarization.

Summarize the following clinical case in Spanish. The summary must be concise,
clinically accurate, and preserve the essential medical information: diagnosis,
relevant clinical findings, tests, treatments, patient evolution, and outcome.

Clinical case:
{fulltext}

Summary:
```

As shown in Listing 1, the prompt explicitly encourages the model to retain the diagnosis, relevant clinical findings, diagnostic tests, treatments, patient evolution and outcome. During training, the reference summary was appended after the prompt and used as the target response. During inference, only the prompt and the source clinical case were provided to the model, which then generated the corresponding summary. The same internal training partition and prompt template were used for all models, ensuring a controlled comparison between base and medically adapted variants.

The system was implemented using the HuggingFace Transformers ecosystem, together with TRL for supervised fine-tuning and PEFT for parameter-efficient adaptation. Models were loaded with 4-bit quantization and fine-tuned using LoRA adapters. The LoRA configuration was applied to the attention projection layers, specifically `q_proj`, `k_proj`, `v_proj` and `o_proj`, with `lora_r=16`, `lora_alpha=32` and `lora_dropout=0.05`. Training was performed for three epochs with a maximum sequence length of 2048 tokens and without sequence packing. We used a per-device batch size of 4 and gradient accumulation over 2 steps, resulting in an effective batch size of 8. The learning rate was set to `2e-5`, with a weight decay of 0.05 and a warmup ratio of 0.03. This configuration was kept fixed across all model pairs to make the comparison between base and medical variants as consistent as possible.

After instruction tuning, the models were used to generate summaries for both the Spanish test documents and the official evaluation documents. Inference was performed using the same instruction template shown in Listing 1, so each clinical case report was presented to the model in the same format used during supervised fine-tuning. Text generation was carried out deterministically using greedy decoding, with `do_sample=False` and `num_beams=1`. This strategy was adopted to reduce variability across generations and encourage stable and reproducible outputs. The maximum generation length was limited to 256 new tokens, which was intended to allow complete summaries while preventing unnecessarily long outputs.

No retrieval component, external knowledge source or post-generation rewriting module was used during inference. Therefore, the generated summaries depend only on the input clinical case report and on the information encoded in each fine-tuned model. The summaries generated for the official evaluation set were saved following the submission format required by the shared task organizers.

5. Results

This section reports the results obtained by the evaluated systems from two complementary perspectives. First, we present the development evaluation carried out during system construction, using the internal development split and the reference-based test set released by the organizers. This analysis is intended to study the behavior of the different model pairs after instruction tuning and to compare each base model with its medically adapted counterpart under controlled conditions. Second, we report the official leaderboard results obtained on the blind evaluation set, which reflect the final performance of our submitted systems in the MultiClinSum-2 Spanish subtrack.

5.1. Development Results

The development evaluation was carried out to analyze the effect of instruction tuning and medical-domain continual pretraining before considering the official leaderboard results. These scores correspond to two reference-based evaluation sets available during system development: the internal development split created from the official training data and the official test set, which was released with reference summaries. The main goal of this analysis is to compare each base model with its medically adapted counterpart under the same training and inference configuration.

Table 2 reports the ROUGE scores obtained by all evaluated models on these two development evaluation sets. We include ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-Lsum. For each model pair and evaluation split, the best value for each metric is highlighted in bold.

The best overall results in this development evaluation are obtained by the base Qwen 3 8B model, which achieves the highest scores on both the internal development split and the official test set across all metrics. On the test set, this model obtains 0.4149 ROUGE-1, 0.1897 ROUGE-2, 0.2721 ROUGE-L and 0.2733 ROUGE-Lsum. The base Gemma 3 4B and base Qwen 3 4B models also obtain competitive results, which indicates that strong multilingual base models can already provide robust performance for Spanish clinical case report summarization after instruction tuning.

The pairwise comparison between base and medically adapted models shows a nuanced effect of medical-domain continual pretraining. The medical variant improves over the base model for Gemma

Table 2
Development evaluation results obtained by the evaluated systems on the internal development split and the reference-based test set released by the organizers. R-1, R-2, R-L and R-Ls denote ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-Lsum, respectively. For each model pair and evaluation split, the best value for each metric is highlighted in bold.

Model	Variant	Split	R-1	R-2	R-L	R-Ls
Qwen 3 4B	Base	Dev	0.4081	**0.1839**	**0.2664**	**0.2675**
	Medical		**0.4082**	0.1837	0.2660	0.2669
	Base	Test	**0.4077**	**0.1831**	**0.2654**	**0.2667**
	Medical		0.4069	0.1828	0.2644	0.2654
Qwen 3 8B	Base	Dev	**0.4158**	**0.1914**	**0.2736**	**0.2745**
	Medical		0.2357	0.0935	0.1492	0.1505
	Base	Test	**0.4149**	**0.1897**	**0.2721**	**0.2733**
	Medical		0.2528	0.1011	0.1599	0.1616
Gemma 3 1B	Base	Dev	0.3859	0.1698	0.2510	0.2514
	Medical		**0.3923**	**0.1729**	**0.2546**	**0.2553**
	Base	Test	0.3874	0.1697	0.2518	0.2528
	Medical		**0.3896**	**0.1708**	**0.2532**	**0.2541**
Gemma 3 4B	Base	Dev	**0.4067**	**0.1867**	0.2648	0.2652
	Medical		0.4062	0.1859	**0.2652**	**0.2656**
	Base	Test	**0.4087**	**0.1883**	**0.2668**	**0.2678**
	Medical		0.4070	0.1856	0.2645	0.2656
Salamandra 2B	Base	Dev	**0.3122**	**0.0948**	**0.1856**	**0.1869**
	Medical		0.2644	0.0624	0.1560	0.1569
	Base	Test	**0.3136**	**0.0952**	**0.1868**	**0.1886**
	Medical		0.2639	0.0615	0.1553	0.1565
Salamandra 7B	Base	Dev	0.2941	0.0810	0.1716	0.1730
	Medical		**0.3226**	**0.1024**	**0.1896**	**0.1936**
	Base	Test	0.2934	0.0801	0.1707	0.1728
	Medical		**0.3244**	**0.1022**	**0.1902**	**0.1947**

3 1B on both evaluation sets, although the gains are modest. A clearer improvement is observed for Salamandra 7B, where the medical model consistently outperforms its base counterpart across all metrics and both splits. This result is particularly relevant because Salamandra is a multilingual model family with strong support for Spanish and other European languages, suggesting that medical-domain continual pretraining may be especially beneficial when applied to models whose linguistic pretraining is already well aligned with the target language.

For the remaining model pairs, the gains from medical-domain adaptation are less evident. In the Qwen family, the base models obtain the best results on the reference-based test set. However, the difference is very small for Qwen 3 4B, where both variants behave almost identically, suggesting that instruction tuning on the task data may already be sufficient to adapt this model to the summarization setting. The larger gap observed for Qwen 3 8B indicates that the interaction between continual pretraining and subsequent instruction tuning can vary substantially across models, even within the same family.

A similar interpretation applies to the Gemma and Salamandra families. In Gemma, the medical 1B model improves over the base version, while the 4B base model remains stronger on the test set. In Salamandra, the 2B medical variant underperforms the base model, whereas the 7B medical variant obtains clear and consistent improvements. These differences suggest that the effect of medical-domain continual pretraining is not only domain-dependent, but also architecture- and scale-dependent. In particular, larger or more linguistically aligned models may be better able to exploit the additional Spanish medical knowledge introduced during continual pretraining.

Table 3
Official results obtained by our submitted runs in the Spanish subtrack of MultiClinSum-2. R-1, R-2 and R-Ls denote ROUGE-1, ROUGE-2 and ROUGE-Lsum, respectively. BS denotes BERTScore. Faith., Comp., Flu. and Cons. denote faithfulness, completeness, fluency and consistency, respectively. The best value among our submitted runs for each metric is highlighted in bold.

Model	R-1	R-2	R-Ls	BS	Faith.	Comp.	Flu.	Cons.
Qwen 3 8B Base (run1)	0.397	0.177	0.261	**0.862**	0.703	0.602	0.685	**0.450**
Qwen 3 8B Medical (run2)	0.395	0.175	0.259	0.861	**0.704**	0.602	**0.688**	0.444
Salamandra 7B Medical (run3)	**0.401**	**0.182**	**0.266**	0.862	0.695	0.591	0.684	0.448
Salamandra 7B Base (run4)	0.252	0.099	0.161	0.592	0.692	**0.619**	0.685	0.394

Overall, these development results suggest that continual pretraining on Spanish medical-domain resources can be useful for Spanish clinical summarization, but its benefits are not uniform across all model families and sizes. This result should be interpreted in light of the experimental setting: the evaluated models are already strong multilingual LLMs and all of them were further instruction-tuned on Spanish clinical summarization data. Under this setup, part of the domain adaptation may be learned during supervised instruction tuning itself, reducing the observable gap between base and medically adapted variants. Even so, the improvements observed for Gemma 3 1B and especially Salamandra 7B indicate that Spanish medical-domain pretraining can provide advantages in specific configurations. These findings support the need for controlled pairwise evaluations, rather than assuming either that medical adaptation will always improve performance or that strong base models are always sufficient.

5.2. Official Evaluation Results

The official evaluation of the submitted systems was carried out by the MultiClinSum-2 organizers using the shared task evaluation protocol. Unlike the development results reported above, which were computed locally on reference-based evaluation sets available during system development, the official results correspond to the evaluation of the submitted runs on the blind evaluation set. This evaluation provides the reference point for assessing the final behavior of our systems in the Spanish subtrack.

Table 3 reports the official scores obtained by our four submitted runs in the Spanish subtrack. The submitted systems correspond to the two model pairs that were selected according to the development evaluation: Qwen 3 8B, which achieved the strongest overall development results, and Salamandra 7B, which showed the clearest improvement from medical-domain adaptation.

As shown in Table 3, the official results are broadly consistent with the trends observed during development. Among our submitted systems, Salamandra 7B Medical obtains the best lexical-overlap scores, reaching the highest ROUGE-1, ROUGE-2 and ROUGE-Lsum values. Qwen 3 8B Base obtains the best BERTScore and consistency score among our runs, while Qwen 3 8B Medical achieves the highest faithfulness and fluency values. Salamandra 7B Base obtains the highest completeness score, but its ROUGE and BERTScore values are substantially lower than those of its medically adapted counterpart.

These results reinforce the main observation from the development evaluation: the effect of medical-domain continual pretraining depends strongly on the underlying model. For Qwen 3 8B, the base model remains slightly stronger than the medical variant in ROUGE and BERTScore, although the differences are small. This suggests that, for this large multilingual model, instruction tuning on the task data may already provide enough adaptation to the Spanish clinical summarization setting. In contrast, the Salamandra 7B pair shows a clear benefit from medical-domain continual pretraining. The medical variant substantially improves over the base model in ROUGE-1, ROUGE-2, ROUGE-Lsum, BERTScore and consistency. This supports the idea that Spanish medical-domain pretraining can be particularly useful when applied to a model family whose multilingual pretraining is already well aligned with Spanish and related European languages.

6. Conclusion and Future Work

This paper presents the UMUTeam system for the Spanish subtrack of the MultiClinSum-2 shared task. Our approach focuses on evaluating whether medical-domain continual pretraining provides an advantage for Spanish clinical case report summarization after supervised instruction tuning. To this end, we compared several pairs of LLMs, where each pair consisted of a base model and a medically adapted counterpart developed by our research group. All models were instruction-tuned using the same Spanish training data, prompt template and experimental configuration, enabling a controlled comparison between base and medically adapted variants.

The development results showed that the effect of medical-domain continual pretraining is not uniform across model families and sizes. Strong multilingual base models, particularly Qwen 3 8B, already achieved robust results after instruction tuning, suggesting that task-specific supervised adaptation can reduce the gap between base and medically adapted models. Nevertheless, the medical variants improved over their base counterparts in specific cases, especially for Gemma 3 1B and Salamandra 7B. The official evaluation further supported this observation: Salamandra 7B Medical obtained the best ROUGE-based scores among our submitted runs and clearly outperformed the corresponding Salamandra 7B Base model. This suggests that Spanish medical-domain pretraining can provide useful gains when the underlying model is well aligned with the target language and is able to exploit the additional domain knowledge.

Overall, our findings indicate that medical-domain continual pretraining can be beneficial for Spanish clinical summarization, but its impact depends on the interaction between the base model, the Spanish medical resources used for adaptation, and the subsequent instruction-tuning stage. In our setting, the amount and composition of the continual-pretraining data, based on clinical practice guidelines, CoWeSe and SMC datasets, may have limited the observable differences between base and adapted models, especially for already strong multilingual LLMs. Even so, the improvements observed for some configurations suggest that the adaptation is having an effect and that further scaling the medical-domain pretraining process could lead to stronger and more consistent gains.

As future work, we plan to extend the continual pretraining stage with a larger and more diverse collection of Spanish medical and clinical documents. This would allow the adapted models to acquire broader domain coverage and potentially improve their ability to summarize specialized clinical content. We also plan to analyze different adaptation strategies, including alternative instruction-tuning configurations, decoding strategies and model-selection criteria based on task-oriented summarization metrics.

Finally, future evaluations should go beyond lexical-overlap metrics by incorporating clinical factuality, completeness and expert-based assessment, since ROUGE and BERTScore may not fully capture the clinical adequacy and usefulness of generated summaries.

Acknowledgments

This work is part of the research projects Late4PoliticES (PID2022-138099OB-I00) and MedicaLLM (CPP2024-011574) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF EU/FEDER UE)- a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammatical and spelling correction. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. Pivovarov, N. Elhadad, Automated methods for the summarization of electronic health records, Journal of the American Medical Informatics Association 22 (2015) 938–947.

[2] 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.

[3] G. Adams, J. Zuckerg, N. Elhadad, A meta-evaluation of faithfulness metrics for long-form hospital-course summarization, in: Machine Learning for Healthcare Conference, PMLR, 2023, pp. 2–30.

[4] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical nlp in spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pp. 193–199.

[5] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, A. Mash, M. Krallinger, Overview of MultiClinSum-2 at CLEF 2026: Data, Evaluation, and Results for Multilingual Clinical Case Report Summarization Across 15 Languages, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2025 Working Notes, 2026.

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

[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] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8342–8360.

[9] F. J. Dorfner, A. Dada, F. Busch, M. R. Makowski, T. Han, D. Truhn, J. Kleesiek, M. Sushil, L. C. Adams, K. K. Bressem, Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks, Journal of the American Medical Informatics Association 32 (2025) 1015–1024.

[10] C. P. Carrino, J. Armengol-Estapé, O. d. G. Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, M. Villegas, Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models, arXiv preprint arXiv:2109.07765 (2021).

[11] L. Dionis, G. Alvaro, M. Dylan, B. Daniel, SpanishMedicaLLM, https://huggingface.co/datasets/somosnlp/SMC, 2024. Dataset on Hugging Face.

[12] J. Liang, C.-H. Tsou, A. Poddar, A novel system for extractive clinical note summarization using ehr data, in: Proceedings of the 2nd clinical natural language processing workshop, 2019, pp. 46–54.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation,

and comprehension, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 7871–7880.
[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.
[17] J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International conference on machine learning, PMLR, 2020, pp. 11328–11339.
[18] V. Hartman, T. R. Campion, A day-to-day approach for automating the hospital course section of the discharge summary, AMIA Summits on Translational Science Proceedings 2022 (2022) 216.
[19] V. C. Hartman, S. S. Bapat, M. G. Weiner, B. B. Navi, E. T. Sholle, T. R. Campion Jr, A method to automate the discharge summary hospital course for neurology patients, Journal of the American Medical Informatics Association 30 (2023) 1995–2003.
[20] 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 (2020) 1234–1240.
[21] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, in: Proceedings of the 2nd clinical natural language processing workshop, 2019, pp. 72–78.
[22] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 3615–3620.
[23] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, et al., A framework for human evaluation of large language models in healthcare derived from literature review, NPJ digital medicine 7 (2024) 258.
[24] W.-w. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, M. Yetisgen, Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, Scientific data 10 (2023) 586.
[25] A. B. Abacha, W.-w. Yim, G. Adams, N. Snider, M. Yetisgen-Yildiz, Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations, in: Proceedings of the 5th Clinical Natural Language Processing Workshop, 2023, pp. 503–513.
[26] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, arXiv preprint arXiv:2109.01652 (2021).
[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.
[28] H. Tran, Z. Yang, Z. Yao, H. Yu, Bioinstruct: instruction tuning of large language models for biomedical natural language processing, Journal of the American Medical Informatics Association 31 (2024) 1821–1832.
[29] O. Rohanian, M. Nouriborji, S. Kouchaki, F. Nooralahzadeh, L. Clifton, D. A. Clifton, Exploring the effectiveness of instruction tuning in biomedical language processing, Artificial intelligence in medicine 158 (2024) 103007.
[30] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, Y. Wang, Pmc-llama: toward building open-source language models for medicine, Journal of the American Medical Informatics Association 31 (2024) 1833–1843.
[31] G. García Subies, et al., A comparative analysis of spanish clinical encoder-based models on named entity recognition and classification tasks, Journal of the American Medical Informatics Association 31 (2024) 2137–2148. doi:10.1093/jamia/ocae074.
[32] I. García-Ferrero, R. Agerri, A. Atutxa Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, A. Zaninello, Medical mT5: An open-source multilingual text-to-text LLM for the medical domain, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and

Evaluation, ELRA and ICCL, 2024, pp. 11165–11177.

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

[34] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[35] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, 2025. URL: https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[36] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. Da Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, et al., Salamandra technical report, arXiv preprint arXiv:2502.08489 (2025).