Multilingual Clinical Summarization with LLMs: A Comparative Study of GRPO and Dynamic One-Shot Approaches

BioASQ at CLEF 2026

Ane Varela[1,3,*,†], Arantza Casillas[1,3], Julene Viña[3], Nuria Lebeña[1,3] and Maite Oronoz[1,2]

[1]HiTZ Basque Center for Language Technology (hitz.eus) - Ixa, University of the Basque Country (EHU), Spain
[2]Department of Computer Languages and Systems - EHU
[3]Department of Electricity and Electronics - EHU

Abstract

This paper details the methodologies developed by our group, `ixa-sum`, for the MultiClinSum 2026 initiative, part of the BioASQ shared task at CLEF. This subtask focuses on the generation of multilingual automatic summarization systems for clinical documents, so that extensive clinical narratives can be condensed into concise summaries that preserve critical information. We explore two distinct implementation strategies: a dynamic one-shot system that incorporates contextually similar examples into the prompt and a Group Relative Policy Optimization approach, which optimizes output by integrating multiple performance signals such as ROUGE and entailment-based fact-checking metrics. Using these approaches, we generate summaries for nine languages across two language families, Romance and West Germanic, comparing LLM `Qwen2.5-7B-Instruct` with the new `Gemma-4-E4B-it` model. Our results indicate that both models perform robustly, particularly in English. While traditional and lexical similarity scores are higher for `Qwen`, which obtains $\text{ROUGE}_{\text{Lsum}}$ scores of up to 0.322 and BERTScore metrics of up to 0.882, metrics calculated by LLM-as-a-judge favor the `Gemma` model, which achieves up to 0.962 in Coverage and 0.769 in Faithfulness. This discrepancy highlights the critical importance of multi-faceted evaluation in clinical NLP.

Keywords

summarization, GRPO, dynamic one-shot, entailment

1. Introduction

In critical fields for human development like medicine, Natural Language Processing (NLP) and Large Language Models (LLMs) have demonstrated significant potential in automating and providing assistance in various tasks. One of such tasks is clinical text summarization, the task of synthesizing complex medical documents into more concise but equally relevant information.

The motivation behind the use of LLMs for clinical text summarization is the increasing worry for physician time allocation, with studies showing that, for every hour of direct patient care, physicians may spend up to two hours on clinical documentation [1]. This not only contributes to rising levels of clinician burnout but also makes it easier for said clinicians to fail to thoroughly review a patient's detailed history, which can ultimately lead to an inappropriate diagnosis or treatment. Particularly, efforts have been made on reducing the quantity of clinical text the healthcare workers are subjected to. Research indicates that LLM-generated summaries can match human-produced summaries [2] and, in some cases, even surpass their quality in terms of coherence and structure [3]. This points to a potential application of LLMs to generate clinical summaries.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
ane.varela@ehu.eus (A. Varela)
0009-0000-9426-9681 (A. Varela); 0000-0003-4248-8182 (A. Casillas); 0000-0001-9093-4680 (N. Lebeña); 0000-0001-9097-6047 (M. Oronoz)

On the other hand, the inherent globality of medical knowledge has to be taken into account. Clinical records are produced in the language of each country or region, and public health crises are not bound by linguistic or political boundaries. However, NLP research has historically been focused on English rather than any other language, also in the clinical field [4]. Research shows that, while the availability of non-English resources is slowly growing, mostly for Spanish, German and French, clinical datasets and tools in non-English languages remain scarce. This imbalance is of importance specially for clinical summarization, as a system that performs well on English may fail silently when applied to other languages, although the underlying medical knowledge does not differ.

Addressing this need, the *MultiClinSum-2* shared task [5], organized within the BioASQ workshop in CLEF (Conference and Labs of the Evaluation Forum) [6], provides a standardized benchmark for the evaluation of multilingual clinical summaries. Our work within the *MultiClinSum-2* shared task directly contributes to it with a focus on comparing the efficiency and deployability of two different models and frameworks in two language families, West Germanic and Romance. The main challenges of the *MultiClinSum-2* shared task include:

1. Factuality. LLMs often create artifacts that are not aligned with the facts, which is a critical error in the clinical domain [7].
2. Domain. Clinical language is particularly difficult for NLP, as the clinical vocabulary often includes dense terminology, abbreviations, context-dependent semantics, and other domain-specific complexities [8].
3. Evaluation. The assessment of summary quality is often limited by the use of automatic metrics, which may not fully capture semantic coherence or clinical relevance [9].

Our group, `ixa-sum`, makes a contribution to tackle these challenges by proposing two frameworks for the generation of summaries: dynamic context selection for one-shot prompting of LLMs, and policy alignment based on Group Relative Policy Optimization (GRPO) to bridge the gap between general language understanding and specialized clinical reasoning. We also combine the two frameworks to see whether this would improve results, and see that, although GRPO training of LLMs may improve traditional metrics, a simpler inference based on dynamic one-shot may be optimal for more state-of-the-art evaluation systems. We also highlight differences between languages of the same families and compare two open-source LLM models, obtaining competitive results.

2. Related work

Research regarding medical NLP has greatly increased its prevalence in the last years [10], and clinical document summarization has emerged as one of its most practically significant applications since the rise of LLMs [11]. Early approaches to this task relied on extractive methods, selecting key sentences directly from the source document. While straightforward, these methods are inherently limited, as they cannot condense content beyond what is explicitly stated. The MultiClinSum shared task has highlighted a clear shift toward abstractive techniques, where LLMs generate new text that captures the semantic essence of the clinical narrative [12].

The participating teams in previous editions focused on two main strategies: prompt engineering and model fine-tuning. Given the scale of modern LLMs, adapting them to the clinical domain typically relies on parameter-efficient methods such as Low-Rank Adaptation (LoRA) [13] or quantization [14], that perform fine-tuning in a computationally efficient way. Yet, domain adaptation does not necessarily improve the performance of the model in summarization, as it focuses on training the model to understand clinical text, but not on retaining the most important information or keeping a similar vocabulary in the summary.

Group Relative Policy Optimization (GRPO), originally introduced by the team that developed DeepSeek [15], is designed to address this gap. Rather than training the model to understand a specific domain, GRPO trains the model by optimizing a reward function that can improve multiple objectives simultaneously [16]. Applied to clinical summarization, GRPO allows a model to be directed toward

outputs that are both fluent and clinically appropriate, with reward functions that can optimize factual consistency, formatting constraints, or lexical similarity.

On the other hand, while policy optimization improves a model's reasoning, the quality of a clinical summary is equally dependent on the context or prompt provided for the summary generation. Zero-shot prompting relies entirely on the model's previous knowledge, which is often insufficient for specialized domains. Few-shot prompting addresses this by providing the model with input-output examples before producing a summary. Nevertheless, a static set of examples cannot cover the full diversity of clinical cases. Dynamic few-shot mitigates this by retrieving the most relevant examples for each specific case directly into the LLM's generation process [17].

In our work, we compare `Qwen2.5-7B-Instruct` and `Gemma-4-E4B-it` across Romance and West Germanic languages, examining summarization quality in a multilingual framework, alongside analysing GRPO training and dynamic prompting as complementary optimization strategies.

3. Materials and methods

3.1. Materials

In this edition of MultiClinSum-2, reference-summary pairs for 15 languages were included. From them, the original native case reports were in English, Spanish, French and Portuguese, and the rest were translated. We focused on two language families: Romance and West Germanic. From them, we particularly explored English, Spanish, and Catalan, due to our interest in Iberian languages.

These are the languages we included in our work:

- Romance languages: Spanish (extra runs), Catalan (extra runs), French, Portuguese, Italian, Romanian.
- West Germanic languages: English (extra runs), Dutch, German.

The characteristics of the training pairs for the selected languages are show in Table 1.

Table 1
Characteristics of the full text-summary pairs for the nine languages (train set).

		N° of instances	N° of characters per instance	N° of words per instance
English	Full text	26943	3514 ± 1539	527 ± 231
	Summary		686 ± 317	99 ± 46
Spanish	Full text	26743	3767 ± 1670	589 ± 259
	Summary		757 ± 358	116 ± 54
Catalan	Full text	26699	3736 ± 1657	602 ± 264
	Summary		749 ± 348	118 ± 55
German	Full text	26424	3928 ± 1848	505 ± 220
	Summary		776 ± 384	96 ± 44
Dutch	Full text	26587	3595 ± 1624	509 ± 225
	Summary		733 ± 336	100 ± 46
French	Full text	26455	3964 ± 1790	612 ± 270
	Summary		788 ± 414	120 ± 55
Italian	Full text	26553	3829 ± 1706	570 ± 251
	Summary		769 ± 356	113 ± 52
Portuguese	Full text	26593	3577 ± 1592	548 ± 243
	Summary		726 ± 337	109 ± 50
Romanian	Full text	26579	3580 ± 1603	549 ± 242
	Summary		730 ± 338	111 ± 51

Upon manual inspection of the provided dataset, we noted a few minor inconsistencies that may impact model performance. Specifically, some instances appear incomplete or contain repetitions of

special characters. There is also inconsistency in the theme of the notes, as, even if most notes concern human medical notes, some of the notes contain veterinary instances, where the subjects are animals. This adds an interesting layer of complexity to the task.

The characteristics of the test set are slightly different, with the full text being slightly longer, particularly in the non-English languages, as it can be seen in Table 2.

Table 2
Characteristics of the full text for the nine languages (test set). All languages have a test set of 1000 instances.

	Nº of characters per instance	Nº of words per instance
English	3799 ± 2309	570 ± 346
Spanish	4321 ± 3498	675 ± 543
Catalan	4279 ± 3412	690 ± 551
German	5110 ± 5844	654 ± 768
Dutch	4033 ± 2915	570 ± 403
French	4654 ± 4031	711 ± 601
Italian	4469 ± 4051	660 ± 572
Portuguese	4204 ± 3891	643 ± 579
Romanian	4202 ± 3728	640 ± 540

3.2. Methods

We use two approaches: (i) a dynamic one-shot framework that augments prompts with contextually relevant examples, and (ii) a Group Relative Policy Optimization (GRPO) method that enhances generation quality through the integration of multiple reward signals. Both strategies are based on `Qwen2.5-7B-Instruct` LLM [18] (which we will call `Qwen`) and additionally, we evaluate its performance against the recently introduced `Gemma-4-E4B-it` model[1] (from now on, `Gemma`).

3.2.1. Dynamic one-shot

Knowing the difference prompting can make when conducting inference of a LLM, we compared several prompting architectures, including *Chain-of-Thought (CoT)* [19] and *Reasoning + Acting (ReAct)* [20].

Multi-step reasoning prompting techniques, such as *CoT* and *ReAct*, require the LLM to generate intermediate reasoning steps prior to producing the final answer, and are shown to improve results in mostly mathematical, logic and decision-making related problems by providing the step-by-step reasoning. This serves the purpose of both making the model slow down to avoid jumping to the wrong conclusion, effectively reducing the problem into more manageable chunks, and making the "thought process" visible so that the user can catch potential logic errors.

However, in our preliminary analyses, empirical testing showed that a simple few-shot or context learning approach yielded the best results, better than those obtained with *CoT* and *ReAct*, even after postprocessing. To refine the prompting method further, we additionally compared two more prompting methods:

- Static Few-Shot: A fixed set of examples is hardcoded into the prompt. While providing a stable baseline, it lacks the flexibility to address diverse input types, which in the clinical case may be relevant to include different medical subfields.
- Dynamic Few-Shot: A specific set of examples is retrieved in real-time based on its semantic similarity to the user's query. Thus, a medical case similar to the one that the LLM needs to create is utilized as an example.

We hypothetise that the dynamic approach is structurally superior to the static one. By fetching examples that are mathematically or linguistically similar to the input, we increase the semantic

[1] https://huggingface.co/google/gemma-4-E4B-it

proximity with the clinical note to be summarized and minimize the reasoning process the model must take. Specifically, text embeddings are extracted using a multilingual sentence transformers model[2] [21], which are then indexed into a vector space to dynamically retrieve the most contextually relevant historical instance based on Euclidean semantic distance. The inference script with the dynamic approach can be found in our GitHub repository[3].

Our preliminary results indicate that dynamic prompting with just one example (dynamic one-shot) yields the best performance. A single, high-quality example provided the template necessary for the model to align its latent representations without over-complicating the context window. For example, test instance 561 in Portuguese selected training instance 24887 as the most similar case, both of them reflecting a case of a female cat.

3.2.2. Group Relative Policy Optimization

To facilitate efficient domain adaptation for clinical summarization, we employ Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) [13]. This was done to avoid traditional fine-tuning, as it is incredibly time- and memory-costly as it involves updating every single one of its billions of parameters. PEFT is the umbrella term for techniques that allow to achieve similar performance by only updating a tiny fraction of the model's weights. LoRA, in turn, is a particular type of PEFT that consists of freezing the original weight matrices of the model; instead of changing those, it aggregates two much smaller matrices that multiply to create a matrix of the same size as the weight matrix, dropping the memory requirements significantly. These fine-tuned weights are merged into the original weights for the final inference.

The LLM was trained using Group Relative Policy Optimization (GRPO). During training, eight candidate responses were generated for each prompt, evaluated against reference outputs, and assigned relative rewards that guided policy optimization. Then, the model is trained based on a reward function that defines the objective of the training. This reward function provides a numerical sign that guides the model toward generating outputs with the format or characteristics that the reward function incentivices. In our case, we compared two reward functions; a simpler one solely based on ROUGE to ensure lexical similarity, and a combination of ROUGE scores and entailment or Natural Language Inference (NLI) scores to prioritize factual consistency were combined and incorporated into the reward function used by GRPO. This is supported by previous work, in which entailment scores are closely aligned with human preference [22].

Taking the full text and generated summary pairs, the NLI model predicts probabilities for three states: entailment, neutrality, and contradiction. To handle text length, we apply a sliding window approach and focus exclusively on the entailment and contradiction scores.

A simple average entailment score would reward summaries that are partially supported but fail to penalize those that introduce contradictions, which would be a critical flaw in the clinical domain. To address this, we design a composite score that rewards broad entailment while explicitly penalizing any contradiction found across the sliding window. Thus, the final score is calculated using the following formula:

$$S_{NLI} = (\text{best\_entailment} - \text{worst\_contradiction}) * \text{mean\_entailment} \quad (1)$$

Where best_entailment is the maximum support found in any window of the note, worst_contradiction penalizes the summary if any contradiction is found, and mean_entailment serves as a weighting factor to reward summaries broadly supported across the context.

The GRPO models were trained using 8 summary generations, per reference, with learning rate 1×10^{-5}, with fp16 bit precision, and batch size 16, saving every 100 steps on an NVIDIA A100. The training script can be found in our GitHub repository[3]. It is important to note that for every GRPO model that is trained, the checkpoint obtaining the best reward is used to create the final summaries. In

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
[3] https://github.com/anevarela/LLM_multi_sum-GRPO

the case of Spanish, this is always the last checkpoint; English tends to have the best performance a bit earlier, and Catalan is more variable.

3.2.3. Runs

Combining the approaches of GRPO and dynamic one-shot prompting, we submitted five runs for English, Spanish and Catalan, summarized in the upper part of Table 3. Rather than treating the runs as independent submissions, they are designed to be compared with one another. Specifically, we want to compare runs 1, 2 and 3 to isolate the contribution of dynamic one-shot prompting and GRPO independently, allowing us to quantify how much each of them contributes to the performance of the different runs. Then, we want to compare runs 3 and 4 to see if two open-source models, similar in size, and identical in prompting technique, perform similarly. Finally, we want to compare runs 1 and 5, in which different GRPO training strategies are used; particularly, run 5 has contradiction penalization, which should, in theory, increase factuality.

Table 3
Summary of the different submitted runs

English (EN), Spanish (ES), Catalan (CA)			
	GRPO	Dynamic one-shot	Model
Run 1	Yes ($ROUGE_2$)	Yes	`Qwen`
Run 2	Yes ($ROUGE_2$)	No	`Qwen`
Run 3	No	Yes	`Qwen`
Run 4	No	Yes	`Gemma`
Run 5	Yes ($ROUGE_2$ + S_{NLI})	Yes	`Qwen`

Dutch (NL), German (DE), French (FR), Portuguese (PT), Italian (IT), Romanian (RO)			
	GRPO	Dynamic one-shot	Model
Run 1	No	Yes	`Qwen`
Run 2	No	Yes	`Gemma`

For the rest of the languages (Dutch and German for the West Germanic languages, and French, Portuguese, Italian and Romanian for the Romance languages), only the untrained runs were applied. Run 1, corresponding to run 3 of English, Spanish, and Catalan, applies dynamic one-shot to the untrained `Qwen` model, while run 2, previously run 4, applies dynamic one-shot to the untrained `Gemma`. Both of these runs are summarized in the lower part of Table 3. The motivation is to compare the quality of the summaries and the characteristics of the tokenization for two models, `Qwen` and `Gemma`, in two different language families, to identify the most prominent and problematic languages between them.

All submitted runs shared some common characteristics. The length of the full texts, as shown in Section 3.1, motivated us to choose the maximum length of the tokenizer at 8,192.

This accounts for higher token-per-word ratios in languages other than English [23], alongside subword fragmentation driven by complex medical terminology [24]. The prompt for the LLMs also specified an approximate length of 100 words for the output summary (see Appendix A).

Finally, to ensure maximum coherence, some general processing steps were applied in all languages and runs. In the input, deduplication was used to remove artifacts likely caused by automatic translation, and we adjusted input lengths to capture full prompts while balancing memory constraints. In the output, conditional regeneration was applied if any tokenization error, alphabetical issue, or repetition was detected. In cases of persistent errors, such as empty outputs, the system defaulted to returning the original clinical note.

4. Results

System-generated summaries of the 1000 instances of the test set are evaluated by the shared task organizers against expert-curated gold standard references using a multi-dimensional assessment with

metrics calculated between 0 and 1. Three evaluation frameworks are included: a traditional metric based on ROUGE [25], a semantic similarity-based framework relying on transformers with BERTScore [26], and a LLM-as-a-judge paradigm using EuroLLM[4] [27] that focuses in four metrics normally evaluated by humans: Faithfulness, Completeness, Fluency and Consistency. The definition for each metric can be found in the MultiClinSum-2 overview [5].

4.1. English, Spanish and Catalan

On the three languages where the 5 runs were submitted, comparing the different evaluation frameworks yields different conclusions. For instance, we can see in Table 4 that evaluating the traditional metrics reliant on ROUGE shows that run 2 is superior in English and in Spanish. This run corresponds to the Qwen model being trained on GRPO in ROUGE metrics and prompted in a zero-shot setting. Thus, it would make sense that this run, and run 1, which corresponds to the same model being prompted with a dynamic example, yield the best results in this particular metrics. Run 5, also trained on GRPO (both with ROUGE and NLI entailment) and with dynamic prompting, obtains good results as well. In Catalan, run 1 is superior to the others, generally speaking, which suggests that, for lower-resource languages, extra information or examples are beneficial. The similarity-based metric sheds a similar light, with run 3, an untrained Qwen model being dynamically prompted on one-shot, performing better than in the other metrics, in all three languages.

Table 4
Traditional ($ROUGE_1$, $ROUGE_2$ and $ROUGE_{Lsum}$) and similarity-based (BERTScore) evaluation metrics for English, Spanish and Catalan on the submitted runs. The boldface underlined number corresponds to the first best metric, the boldface corresponds to the second best metric, and the underlined corresponds to the third best metric.

Language	Run	$ROUGE_1$	$ROUGE_2$	$ROUGE_{Lsum}$	BERTScore
English	Run 1	**0.433**	**0.216**	**0.318**	**0.875**
	Run 2	**0.440**	**0.220**	**0.322**	**0.876**
	Run 3	0.381	0.155	0.267	0.870
	Run 4	0.372	0.138	0.255	0.868
	Run 5	0.412	0.184	0.283	0.869
Spanish	Run 1	**0.463**	**0.232**	**0.314**	**0.881**
	Run 2	**0.467**	**0.235**	**0.314**	**0.882**
	Run 3	0.400	0.171	0.271	0.871
	Run 4	0.398	0.155	0.257	0.870
	Run 5	0.436	0.206	0.293	0.868
Catalan	Run 1	**0.458**	**0.233**	**0.306**	**0.881**
	Run 2	**0.456**	**0.231**	**0.301**	**0.882**
	Run 3	0.402	0.172	0.264	0.870
	Run 4	0.375	0.150	0.239	0.848
	Run 5	0.407	0.186	0.263	0.866

Until now, run 3 and 4, relying on untrained models with dynamic one-shot prompting, have not obtained particularly good results. However, when looking at the results in Table 5, which show the results for LLM-as-a-judge calculated metrics, a different reality is shown. In particular, run 4, which corresponds to the newly released Gemma, obtains good results in most of these metrics. In general, the Completeness is where our runs excel at; Faithfulness is not a bad metric either. These two metrics are, arguably, the most important for medicine-based summarization, where clinician-to-clinician summaries do not have to be as simple or fluent as patient-oriented summaries. On the other hand, the Consistency metric, which obtains lower scores for our runs, highlights the high variability in summary quality between different languages.

Before submission, English produced the cleanest outputs, Spanish required moderate postprocessing, and Catalan showed the most severe issues, even with Qwen, suggesting Catalan is particularly challeng-

[4] https://huggingface.co/blog/eurollm-team/eurollm-9b

Table 5
LLM-as-a-judge evaluation metrics for English, Spanish and Catalan on the submitted runs. The boldface underlined number corresponds to the first best metric, the boldface corresponds to the second best metric, and the underlined corresponds to the third best metric.

Language	Run	Faithfulness	Completeness	Fluency	Consistency
English	Run 1	0.726	0.813	0.700	0.698
	Run 2	0.732	0.832	0.700	0.698
	Run 3	**0.761**	**0.945**	**0.701**	**0.699**
	Run 4	**0.769**	**0.962**	**0.704**	**0.699**
	Run 5	0.729	0.862	0.700	**0.699**
Spanish	Run 1	0.709	0.786	0.700	0.647
	Run 2	0.713	0.796	0.700	0.649
	Run 3	**0.733**	**0.894**	**0.701**	**0.652**
	Run 4	**0.750**	**0.920**	**0.701**	**0.655**
	Run 5	0.704	0.602	0.700	0.609
Catalan	Run 1	0.702	0.833	**0.700**	**0.640**
	Run 2	**0.705**	0.867	**0.700**	0.627
	Run 3	**0.705**	**0.872**	**0.700**	**0.640**
	Run 4	**0.714**	**0.873**	0.693	0.613
	Run 5	0.697	0.726	0.699	0.553

ing, likely due to its lower representation in pretraining data. The most problematic setup was dynamic inference with Gemma, which consistently required fallback strategies and heavy postprocessing.

Our best-performing model in preliminary analysis was combining $ROUGE_2$ GRPO with dynamic one-shot. In contrast, on the competition, the GRPO model works best with zero-shot inference, with good balanced performance across all metrics.

4.2. Romance languages

For these languages, the main focus was on the comparison of metrics in the different languages. We only compare the dynamic runs corresponding to no pre-training of the LLM model, comparing Qwen and Gemma, as shown previously in Table 3.

The traditional and similarity-based metrics in Figure 1 show that there is no clear difference in performance between the different languages in the case of BERTScore, while in ROUGE, the superiority of Spanish, French and Catalan is evident.

We also see a clear trend of Qwen performing slightly better than Gemma, with some exceptions in French. These exceptions that make French perform better than expected may be due to the fact that it was one of the native languages of the full text-summary pairs, which suggests that translation may also degrade performance.

Things change for the evaluation based on judge LLM in Figure 2, where Gemma clearly performs better than Qwen in Faithfulness and Completeness, with Fluency and Consistency being more balanced to Qwen. In general, the number of tokens per word seems not to have a great influence in the performance, although Italian, Catalan and Romanian do have a worse performance, particularly in Consistency.

Qwen generally produced summaries with no fallbacks or postprocessing needed in French, Italian, Portuguese, and Romanian, while Gemma degraded substantially, especially in Portuguese and Romanian. French and Italian showed the same trend but less severely. This indicates that Gemma is less robust than Qwen in lower-resource Romance languages.

We can see that, in general, the Qwen run is superior to the Gemma one, although this trend is not as strong for non-traditional metrics and sometimes flips in, particularly, LLM-as-a-judge systems.

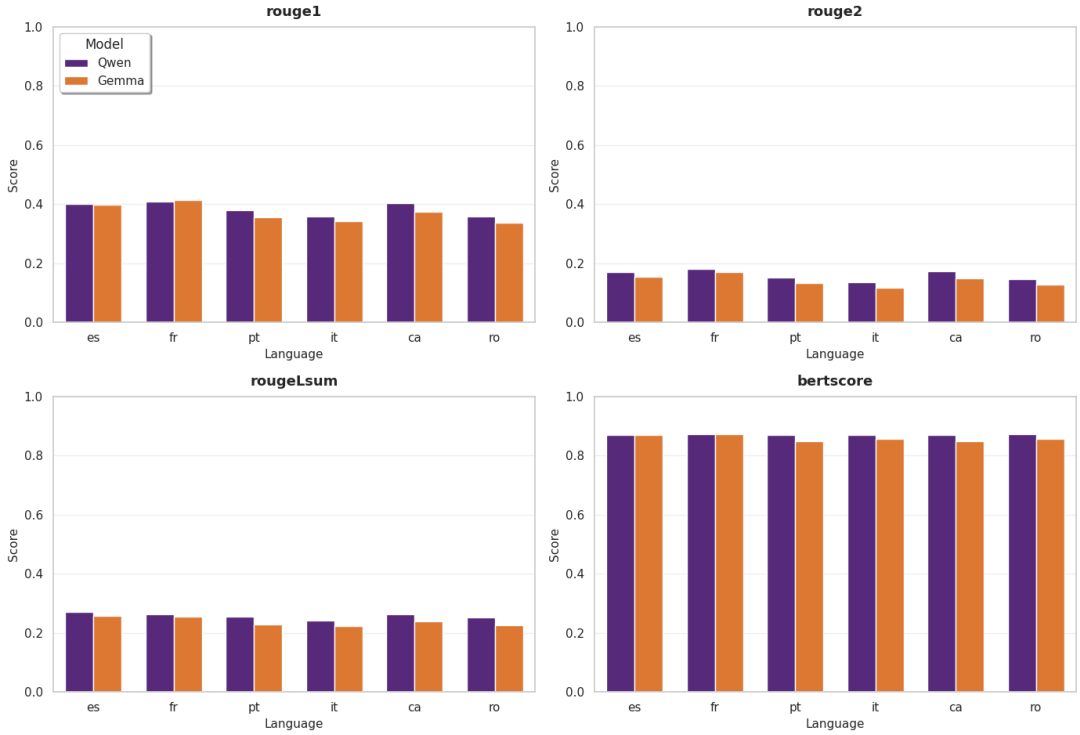

Figure 1: Traditional and similarity-based evaluation metrics comparison for Romance languages on the submitted runs.

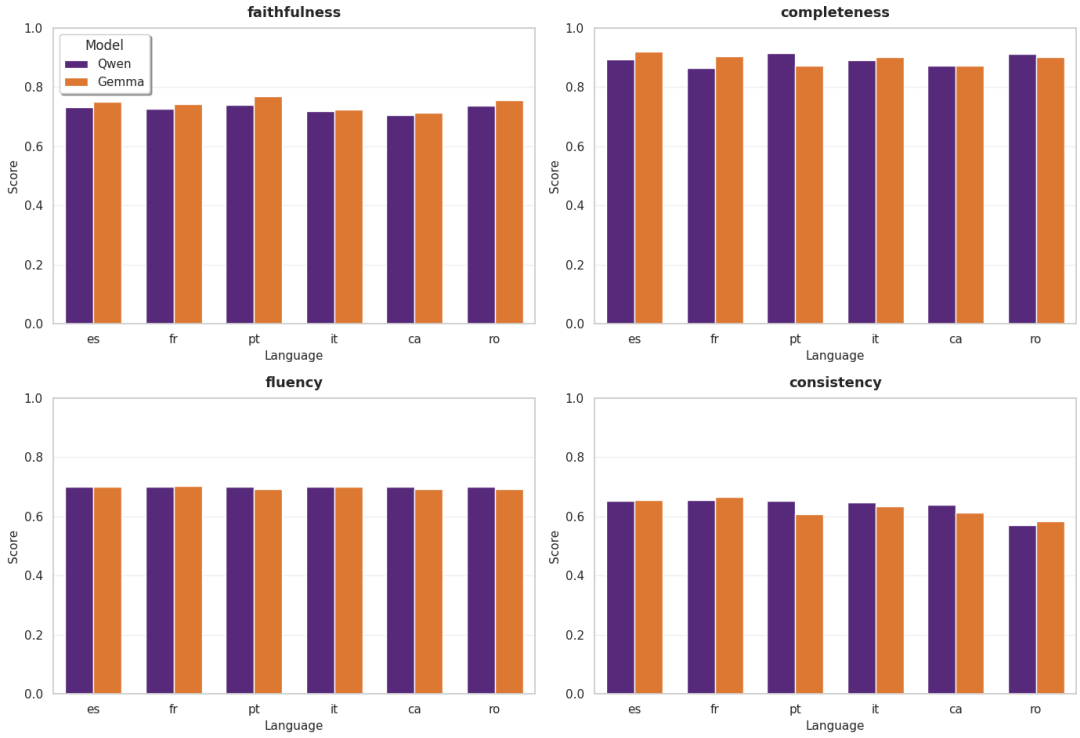

Figure 2: LLM-based evaluation metrics comparison for Romance languages on the submitted runs.

4.3. West Germanic languages

Comparing between the three West Germanic languages in all three evaluation frameworks, the differences can be seen in Figures 3 and 4. In this instance, the differences are clear between languages

and models, with English performing better than the other two, as a general norm. Except in the LLM-based evaluation, Qwen seems to outperform Gemma as well, as it is also seen in the Romance languages. This suggests that, although Gemma may not paraphrase, it retains information in a satisfactory manner. This happens for both Romance and West Germanic languages.

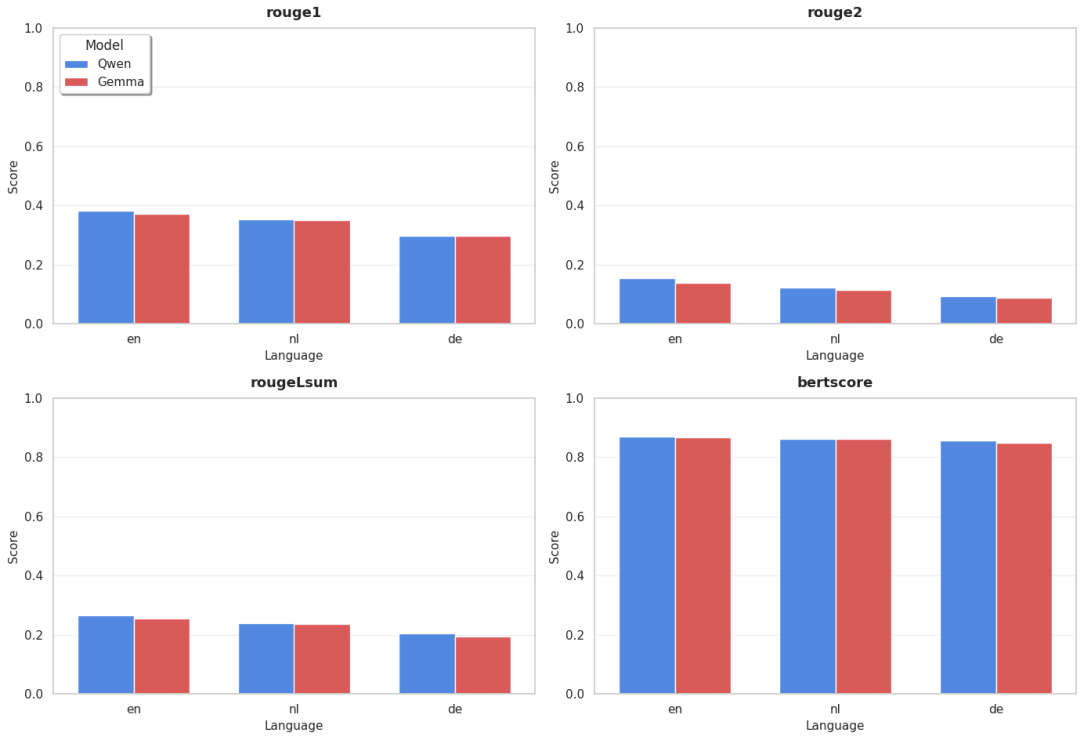

Figure 3: Traditional and similarity-based evaluation metrics comparison for West Germanic languages on the submitted runs.

In the same way, English required few fallback and postprocessing strategies with both Qwen and Gemma, Dutch showed only minor degradation with Gemma, while German was the most problematic language overall, including for Qwen. This suggests that input length constraints contributed to generation failures, maybe also due to the subtokenization problem.

5. Conclusions

This work presents a comprehensive evaluation of optimization strategies for medical text summarization across two language groups: Romance and West Germanic languages. Our findings yield several key insights into the trade-offs between optimization paradigms, resource allocation, and evaluation methodologies.

While GRPO outperforms dynamic one-shot prompting on the traditional metric ROUGE, this trend reverses in LLM-judge evaluations. This divergence highlights the gap between literal text overlap, mostly used in metrics like ROUGE, and semantic accuracy, emphasizing the need for advanced evaluation frameworks. Furthermore, dynamic one-shot relies on simple vector database lookups for example matching, whereas GRPO requires full model fine-tuning. Since our one-shot approach utilizes significantly fewer tokens than traditional 3–5 shot prompting, it remains a highly efficient alternative when operational costs outweigh marginal accuracy gains.

Our ablation studies in English, Spanish and Catalan revealed that utilizing standalone GRPO yielded better performance than combining it with dynamic prompting elements. Comparing the GRPO trained with only ROUGE to the one also trained with entailment, the entailment-based run showed a slight LLM-judge metric improvement in English. However, this benefit did not translate to Spanish or Catalan due to the current scarcity of high-quality multilingual NLI resources. Additionally, a limitation

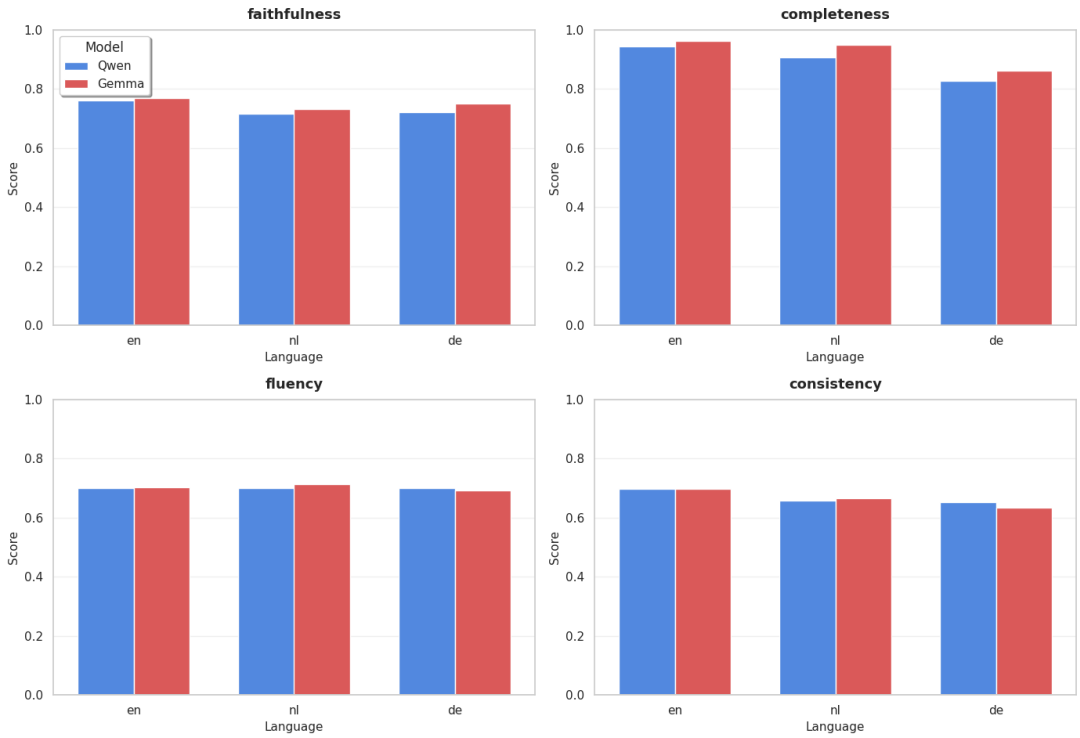

Figure 4: LLM-based evaluation metrics comparison for West Germanic languages on the submitted runs.

of this evaluation is the lack of a baseline comparison against an untrained model using a simpler prompt structure, which prevents us from isolating the exact performance gains attributable to the used approaches.

For the same language, `Qwen` generally outperformed `Gemma` across most tasks. The notable exception occurred during LLM-as-a-judge evaluations, where `Gemma` demonstrated superior performance specifically in the categories of Faithfulness and Completeness. Both models are under 8B parameters, highlighting the viability of open-source, relatively small models in medical settings.

While our pipeline achieved strong results, it highlighted clear avenues for refinement, particularly regarding evaluation and multilingual scaling. Particularly, a primary bottleneck in expanding entailment-guided GRPO to non-English languages was the lack of medical-domain specific NLI models. This lack of support, particularly for lower-resource languages like Catalan, remains a challenge that requires targeted dataset curation. Furthermore, while our post-processing pipeline successfully mitigated severe structural failures, residual formatting and extraction errors persisted. Particularly for the `Gemma` model, which necessitated more postprocessing, more robust, natively multilingual generation strategies are required.

Despite these limitations, our approaches achieved highly competitive results.

Acknowledgments

Ane Varela is funded by an FPU predoctoral fellowship from the MICIU (FPU24/00333). This work has been partially supported by the HiTZ Center and the following projects: (i) HumanAIze (AIA2025-163322-C61) and (ii) EDHIA (PID2022-136522OB-C22) MCIN/AEI/10.13039/501100011033 projects and (iii) the European Research Council under Horizon Europe, grant number 101293013, ELEXAI project. Furthermore, the authors acknowledge the technical and human support provided by the DIPC Supercomputing Center, where computational resources were provided by the Hyperion cluster.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Gemini in order to: Check grammar and spelling, Paraphrase and Reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Sinsky, L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty, G. Blike, Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties, Annals of Internal Medicine 165 (2016) 753–760. doi:10.7326/m16-0961.

[2] J. D. Oliveira, H. D. P. Santos, A. H. D. P. S. Ulbrich, J. C. Couto, M. Arocha, J. Santos, M. M. Costa, D. Faccio, F. O. Tabalipa, R. F. Nogueira, Development and evaluation of a clinical note summarization system using large language models, Communications Medicine 5 (2025). doi:10.1038/s43856-025-01091-3.

[3] B. Janota, K. Janota, Application of artificial intelligence (AI) in the creation of discharge summaries in psychiatric clinics, The International Journal of Psychiatry in Medicine 60 (2024) 330–337. doi:10.1177/00912174241284730.

[4] A. Shaitarova, J. Zaghir, A. Lavelli, M. Krauthammer, F. Rinaldi, Exploring the latest highlights in medical natural language processing across multiple languages: A survey, Yearbook of Medical Informatics 32 (2023) 230–243. doi:10.1055/s-0043-1768726.

[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 2026 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, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), CLEF 2026 Working Notes, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[7] P. Chung, A. Swaminathan, A. Goodell, Y. Kim, M. Reincke, L. Han, B. Deverett, M. A. Sadeghi, A. b. El Ariss, M. Ghanem, D. Seong, A. Lee, C. Coombes, B. Bradshaw, M. Sufian, H. J. Hong, T. Nguyen, M. Rasouli, K. Kamra, M. Burbridge, J. McAvoy, R. Saffary, S. P. Ma, D. Dash, J. Xie, E. Wang, C. Schmiesing, N. Shah, N. Aghaeepour, MIMIC-III-Ext-VeriFact-BHC: Labeled Propositions From Brief Hospital Course Summaries for Long-form Clinical Text Evaluation, PhysioNet (2025). doi:10.13026/abat-g475, version 1.0.0.

[8] S. Bandare, Natural language processing with AI for unstructured claims data, American Journal of Bioinformatics 7 (2026) 13–25. doi:10.71465/ajb.3561.

[9] L. Bednarczyk, D. Reichenpfader, C. Gaudet-Blavignac, A. K. Ette, J. Zaghir, Y. Zheng, A. Bensahla, M. Bjelogrlic, C. Lovis, Scientific evidence for clinical text summarization using Large Language Models: Scoping review, J Med Internet Res 27 (2025) e68998. doi:10.2196/68998.

[10] X. Chen, H. Xie, F. L. Wang, Z. Liu, J. Xu, T. Hao, A bibliometric analysis of natural language processing in medical research, BMC Medical Informatics and Decision Making 18 (2018). doi:10.1186/s12911-018-0594-x.

[11] A. N. Khoruzhaya, M. D. Varyukhina, R. A. Erizhokov, I. A. Blokhin, R. V. Reshetnikov, M. R. Kodenko, A. P. Pamova, T. A. Burtsev, K. M. Arzamasov, O. V. Omelyanskaya, A. V. Vladzymyrskyy, Y. A. Vasilev, MEDAI-LLM-SUMM: a reporting checklist for medical text summarization studies

using large language models, Frontiers in Digital Health 8 (2026). doi:10.3389/fdgth.2026.1761601.

[12] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, C. Escolano, A. Mash, M. Melero, L. Pratesi, L. Vigil-Gimenez, L. Fernandez-Lopez, 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: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 19–40.

[13] J. E. 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. arXiv:2106.09685.

[14] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. G. Howard, H. Adam, D. Kalenichenko, Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. arXiv:1712.05877.

[15] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. arXiv:2402.03300.

[16] U. Devulapalli, A. Satsangi, A. Narayan, Fine-tuning large language models for structured clinical report generation using GRPO, Proceedings of the AAAI Symposium Series 7 (2025) 496–500. doi:10.1609/aaaiss.v7i1.36923.

[17] L. He, Y. Li, B. Li, E. H. Cui, D. Wang, DSS-Prompt: Dynamic-static synergistic prompting for few-shot class-incremental learning, in: Proceedings of the 33rd ACM International Conference on Multimedia, MM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2703–2712. doi:10.1145/3746027.3754719.

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

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022, pp. 24824 – 24837.

[20] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: Synergizing reasoning and acting in language models, 2023. arXiv:2210.03629.

[21] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. arXiv:1908.10084.

[22] X. Tang, A. Cohan, M. Gerstein, Aligning factual consistency for clinical studies summarization through reinforcement learning, in: T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, A. Rumshisky (Eds.), Proceedings of the 5th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 48–58. doi:10.18653/v1/2023.clinicalnlp-1.7.

[23] O. Ahia, S. Kumar, H. Gonen, J. Kasai, D. Mortensen, N. Smith, Y. Tsvetkov, Do All Languages Cost the Same? Tokenization in the era of commercial language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, p. 9904–9923. doi:10.18653/v1/2023.emnlp-main.614.

[24] G. Balde, S. Roy, M. Mondal, N. Ganguly, Evaluation of LLMs in medical text summarization: The role of vocabulary adaptation in high OOV settings, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 22989–23004. doi:10.18653/v1/2025.findings-acl.1179.

[25] S. Abdul Samad, R. Sushma, G. Bharathi Mohan, P. Samuji, S. Repakula, S. R. Kothamasu, Advancing abstractive summarization: Evaluating GPT-2, BART, T5-Small, and Pegasus models with baseline in ROUGE and BLEU metrics, in: S. M. Basha, H. Taherdoost, C. Zanchettin (Eds.), Innovations in Cybersecurity and Data Science, Springer Nature Singapore, Singapore, 2024, pp. 119–131.

[26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, CoRR (2019). arXiv:1904.09675.
[27] P. H. Martins, J. Alves, P. Fernandes, N. M. Guerreiro, R. Rei, A. Farajian, M. Klimaszewski, D. M. Alves, J. Pombal, N. Boizard, M. Faysse, P. Colombo, F. Yvon, B. Haddow, J. G. C. de Souza, A. Birch, A. F. T. Martins, EuroLLM-9B: Technical Report, 2025. arXiv:2506.04079.

A. Prompts used for the inference

Note: only the prompts for English, Spanish and Catalan are shown. The rest of the languages are omitted for brevity, but all the prompts were translated to the corresponding languages.

System prompts

```
{
    "es": "Eres un asistente médico experto.",
    "en": "You are an expert medical assistant.",
    "ca": "Ets un assistent mèdic expert."
}
```

User prompts (Note: Ahora/Now/Ara is only used if there is any dynamic context being provided).

```
{
    "es": "Ahora, genera un resumen clínico de unas 100 palabras.
    NOTA:
    [note]
    RESUMEN:
    ",
    "en": "Now, generate a brief clinical summary of around 100 words.
    NOTE:
    [note]
    SUMMARY:",
    "ca": "Ara, genera un resum clínic d'unes 100 paraules.
    NOTA:
    [note]
    RESUM:
    "
}
```

One-shot prompting (before user prompt):

```
if self.lang == "es":
    header = "Aquí tienes un ejemplo previo similar:\n\n"
    label_n, label_s = "Nota", "Resumen"
elif self.lang == "ca":
    header = "Aquí tens un exemple previ similar:\n\n"
    label_n, label_s = "Nota", "Resum"
elif self.lang == "en":
    header = "Here is a similar previous example:\n\n"
    label_n, label_s = "Note", "Summary"
context = header + f"{label_n}: {fulltext_note}\n{label_s}: {summary}\n\n"
```