Enhancing Multilingual Medical Summarization via Contextual Keyword Augmentation[⋆]

Notebook for the UMass BioNLP Lab at CLEF 2025

Md Shahidul Salim[1,2], Lianne Fu[3], Arav Adikesh Ramakrishnan[3], Sunjae Kwon[1,3], Zonghai Yao*[1,3] and Hong Yu*[1,2,3,4]

[1]Center for Healthcare Organization and Implementation Research, VA Bedford Health Care, MA, USA
[2]Miner School of Computer and Information Sciences, University of Massachusetts Lowell, MA, USA
[3]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, MA, USA
[4]Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA

Abstract

This paper presents the work of the UMass BioNLP team for the MultiClinSUM multilingual medical text summarization task. We introduce **MedCOD**, a novel framework that improves multilingual summarization through keyword-based contextual augmentation. MedCOD begins by extracting medical keywords from full clinical texts using the Qwen2.5-14B model. These keywords are translated into five languages—English, Spanish, French, German, and Portuguese—using the NLLB 3.3B model, and validated through back-translation and semantic equivalence checking with Qwen2.5-14B. The resulting multilingual keyword chains are incorporated into prompts as a structured context. We evaluate MedCOD using two open-source large language models, Qwen2.5-14B and Phi-4B, in both zero-shot and fine-tuned settings. We fine-tune the model using parameter-efficient LoRA on the MultiClinSUM training set. Experimental results demonstrate that MedCOD significantly improves summarization quality, especially in non-English languages. Ablation studies show that both prompt-level augmentation and fine-tuning contribute to the observed performance gains.

Keywords

Medical Text Summarization, Multilingual NLP, Contextual Augmentation, Chain of Dictionary

1. Introduction

Electronic Health Records (EHRs) have become an integral part of modern healthcare, serving as a critical medium for enhancing patient engagement and facilitating better communication between healthcare providers and patients. Recognizing the value of EHRs, the Centers for Medicare & Medicaid Services (CMS) Incentive Programs have promoted the meaningful use of EHRs, empowering patients to access and manage their health information electronically. However, the full benefits of EHR accessibility are not uniformly realized, particularly among patients with limited English proficiency [1]. In the United States, approximately 17% of the population identifies as Hispanic, with nearly half possessing limited English skills. This language barrier significantly impedes their ability to comprehend health records, potentially leading to misunderstandings, reduced medication adherence, and poorer health outcomes.

In the domain of medical information processing, a major challenge lies in developing robust text summarization systems capable of effectively condensing complex medical texts [2]. The MultiClinSUM Shared Task provides a valuable platform for addressing this issue by encouraging the development of intelligent multilingual summarization systems. This paper focuses on the MultiClinSUM text

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

[⋆]You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

MdShahidul_Salim@uml.edu (M. S. Salim); lianfu@umass.edu (L. Fu); aravadikeshr@umass.edu (A. A. Ramakrishnan); sunjaekwon@umass.edu (S. Kwon); zonghaiyao@umass.edu (Z. Yao*); Hong_Yu@uml.edu (H. Yu*)
0009-0001-7094-757X (M. S. Salim); 0009-0001-3889-9993 (L. Fu); 0009-0008-7105-3405 (A. A. Ramakrishnan); 0000-0002-5425-6779 (S. Kwon); 0000-0002-5707-8410 (Z. Yao*); 0000-0001-9263-5035 (H. Yu*)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

summarization task [3], where we introduce a novel approach leveraging a medical chain of dictionary to enrich the input with augmented contextual information. This dictionary is constructed by extracting medical keywords from full texts and translating them into multiple languages. The enriched context is then used to enhance the performance of open-source language models in multilingual summarization. While medical text summarization in English is already well-established, our work emphasizes improving summarization quality in other languages, where current models often underperform.

In this study, we have applied our framework named MedCOD, which builds upon Chain-of-Dictionary Prompting (COD) [4] and MedCOD [5]—recent approaches that enhance machine translation by incorporating multilingual dictionaries into large language model (LLM) prompts. COD has demonstrated strong performance on general-domain translation tasks (e.g., FLORES-200), improving translation quality for many language pairs and even outperforming strong baselines like NLLB 3.3B in certain settings. We extend this approach by integrating various prompting methods, using the large language model as a knowledge base (LLM-KB), to enhance English-to-Spanish biomedical translation.

In our framework, we first employed the Qwen2.5-14B [6] model to extract medical keywords from the full text. These extracted keywords were then translated into multiple target languages using the NLLB 3.3B [7] model. To ensure semantic consistency and translation quality, we used the Qwen2.5-14B model to perform equivalence checks between the translated and back-translated keywords. Finally, we evaluated the effectiveness of the overall framework using two open-source large language models (LLMs): Qwen2.5-14B and Phi-4 14B.

2. Related Work

Given that this study encompasses both pipeline development and prompt engineering, we organize the related work accordingly into four subsections: Section 2.1 *Medical summarization*, Section 2.2 *Quantization and Parameter-Efficient Fine-Tuning (PEFT)*, Section 2.3 *Prompt Engineering*, and Section 2.4 *COD and MedCOD in Machine Translation*. We begin by reviewing prior work relevant to the core pipeline, including medical text summarization with LLMs, model efficiency techniques, and advances in prompt design strategies.

2.1. LLMs in Medical Text Summarization

The widespread implementation of EHRs has significantly increased the clinical documentation burden, contributing to rising stress and clinician burnout [2, 8, 9, 10]. Analyzing vast textual data and summarizing key information from EHRs imposes a substantial strain on clinician time. With the advent of large language models (LLMs) like ChatGPT [11], the ability to generate coherent and clinically relevant summaries has improved considerably [12, 13]. While various benchmarks exist for evaluating LLMs on general NLP tasks [14, 15], many fail to capture the nuances of clinical reasoning, terminology, and context. This motivates the development of medically grounded summarization pipelines tailored to real-world use.

2.2. Quantization and Parameter-Efficient Fine-Tuning (PEFT)

Quantization [16] is a model compression technique that reduces inference cost by replacing high-precision floating-point weights (e.g., float32) with lower-precision formats such as int8 or bfloat16. This significantly lowers memory usage and computation time, making LLMs deployable on resource-constrained devices. However, quantization may lead to a drop in accuracy, especially for tasks requiring domain-specific reasoning or subtle semantic understanding—common in medical applications.

Parameter-Efficient Fine-Tuning (PEFT) [?] complements quantization by updating only a small subset of newly added parameters (e.g., LoRA adapters or prefix vectors) while keeping the core model frozen. This approach improves fine-tuning efficiency in both computation and storage, especially for domain adaptation, as it avoids full backpropagation through billions of parameters. Unlike traditional full-model fine-tuning, PEFT enables rapid iteration across downstream tasks and user-specific domains while maintaining generalization.

Although typically studied in isolation, quantization and PEFT are increasingly being used together to build low-resource yet performant LLM-based applications. Recent work demonstrates that when paired carefully, these methods can preserve task-relevant accuracy while reducing both training and inference costs—an important consideration for medical NLP systems deployed in production environments.

2.3. Prompt Engineering

Prompt engineering is the practice of designing input prompts to guide LLM behavior more effectively, and has become essential in aligning model outputs with task requirements—particularly in zero- and few-shot settings. In clinical NLP, carefully crafted prompts can greatly improve performance on tasks like summarization, question answering, and reasoning.

Among prompt engineering methods, Chain-of-Thought (CoT) prompting [11] stands out for its ability to elicit step-by-step reasoning from LLMs. CoT is particularly useful in medical MCQs, where intermediate reasoning steps reflect critical thinking paths. However, studies have shown that different CoT prompts may lead to divergent answers due to variability in reasoning trajectories. To mitigate this, self-consistency [17, 18] aggregates multiple CoT responses and selects the most frequent answer, improving robustness and accuracy.

Recent work also explores prompt personalization (e.g., persona-based prompting), grounding (e.g., incorporating external knowledge), and prompt augmentation (e.g., using keywords or dictionary entries) as techniques to improve LLM alignment with task-specific needs. However, prompt engineering still lacks systematic understanding, especially in specialized domains like healthcare, where task ambiguity and terminology complexity pose unique challenges.

2.4. COD and MedCOD in Machine Translation

Recent advances in multilingual machine translation have significantly enhanced LLM capabilities in low-resource settings. The *No Language Left Behind* project [7] introduced a highly scalable model covering over 200 languages, achieving state-of-the-art results on FLORES-200. Similarly, mBART [19] employed denoising autoencoding for multilingual sequence-to-sequence tasks.

Prompt-based approaches have proven particularly effective for multilingual translation. *Chain-of-Dictionary Prompting* (COD) [20] augments prompts with multilingual dictionaries to guide translation generation. MedCOD [5] extends this to the medical domain by incorporating domain-specific dictionaries and keywords into the prompt, thereby improving terminology accuracy and factual alignment.

Medical machine translation (Med-MT) is uniquely challenging due to the need for precise terminology, context preservation, and limited training data. Early work by Liu and Cai [21] identified common semantic drifts and domain-specific errors when translating EHRs. More recent efforts such as BiomedBench [22] and MeLoT [23] address these limitations by providing multilingual biomedical corpora for systematic evaluation.

Our work extends MedCOD beyond sentence-level translation by integrating it into summarization prompts—effectively unifying translation and summarization via contextual augmentation. This design enables LLMs to generate multilingual medical summaries with improved fluency, fidelity, and task alignment, particularly when paired with keyword filtering and open-source PEFT-tuned models.

3. Methodology

We evaluated open-source LLMs for their effectiveness in medical text summarization across four languages: English, Spanish, French, and Portuguese. Figure 1 provides an overview of our framework, which incorporates the MultiClinSUM dataset, various LLMs, and a specific prompting methodology. Our analysis focuses on the impact of MedCOD, our proposed framework designed to identify the most suitable augmented knowledge (referred to as contextual information throughout this paper) to support

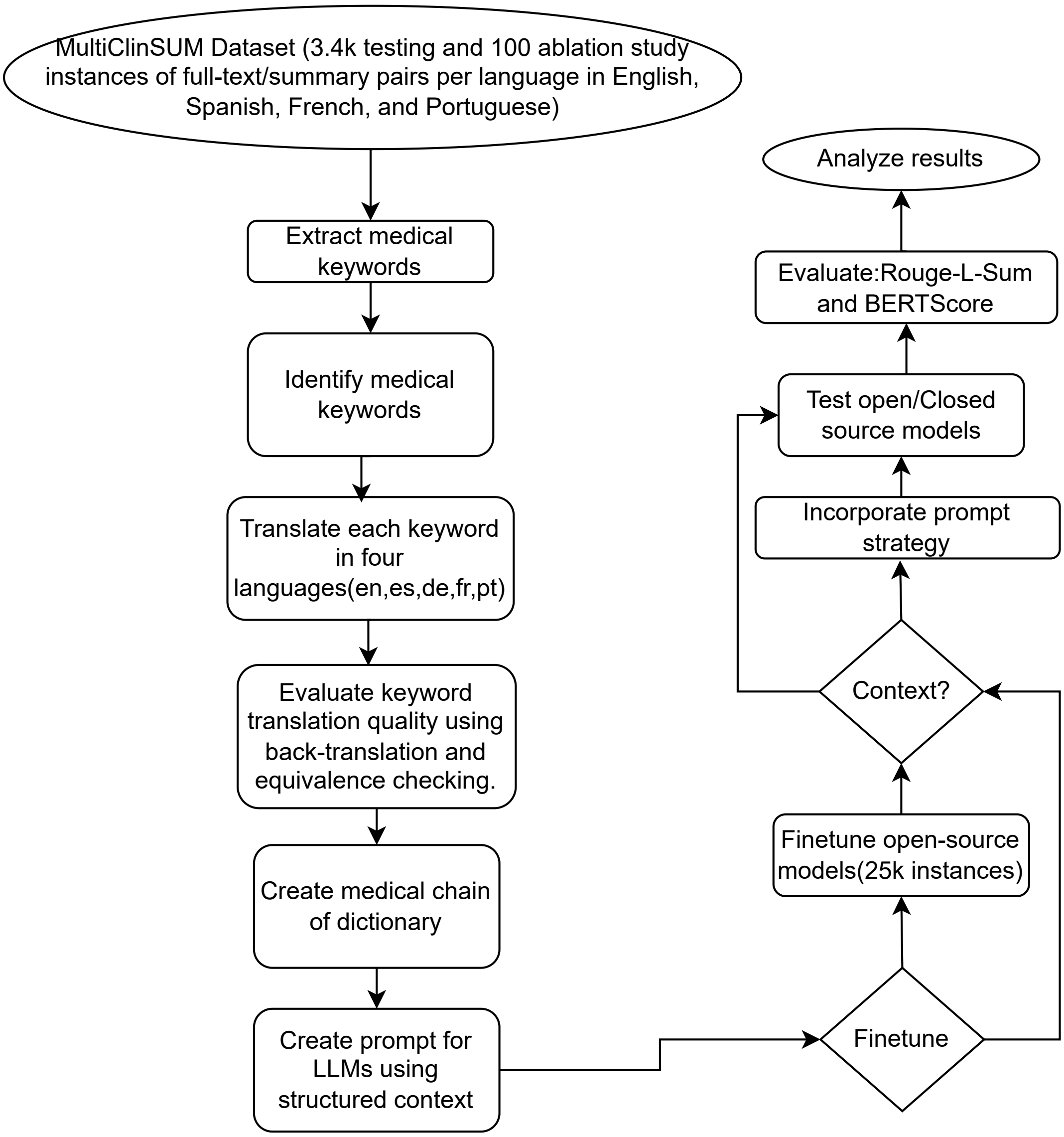
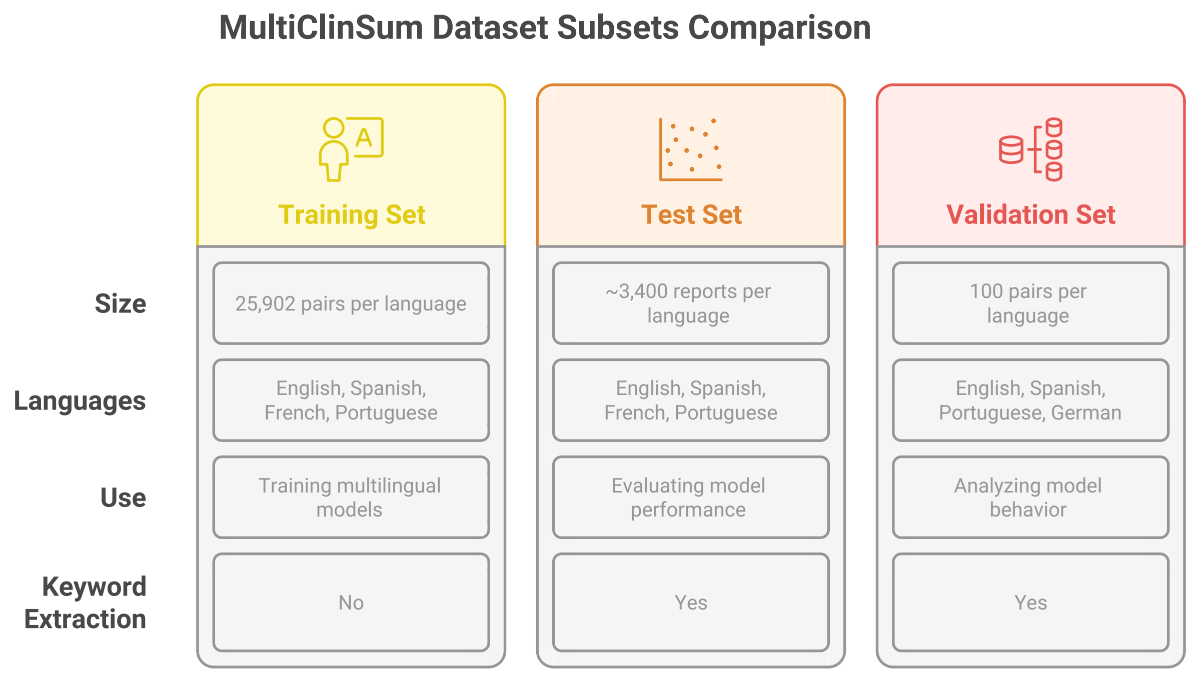

MultiClinSUM Dataset (3.4k testing and 100 ablation study instances of full-text/summary pairs per language in English, Spanish, French, and Portuguese)
Extract medical keywords
Identify medical keywords
Translate each keyword in four languages(en,es,de,fr,pt)
Evaluate keyword translation quality using back-translation and equivalence checking.
Create medical chain of dictionary
Create prompt for LLMs using structured context
Finetune
Finetune open-source models(25k instances)
Context?
Incorporate prompt strategy
Test open/Closed source models
Evaluate:Rouge-L-Sum and BERTScore
Analyze results

Figure 1: Flowchart for illustrating the process of medical dataset preprocessing, structured prompt creation, and model evaluation for fine-tuning.

medical summarization. Furthermore, we explore the role of fine-tuning techniques in enhancing the performance of these models.

3.1. Dataset Preparation and Keyword Extraction

For dataset preparation, we utilized three subsets of the MultiClinSum dataset, which is designed to support multilingual clinical summarization across four primary languages: English, Spanish, French, and Portuguese. The first subset is the large-scale training set, comprising 25,902 full-text and summary pairs distributed across the four languages. This dataset served as the foundation for training our multilingual models. The second subset is the official test set used in the MultiClinSum Shared Task, which includes 3,396 clinical case reports in English, 3,406 in Spanish, 3,469 in French, and 3,442 in Portuguese. This standardized test set enabled consistent evaluation of model performance across languages. Additionally, for our ablation study, we constructed a smaller subset drawn from the gold-



MultiClinSum Dataset Subsets Comparison

	Training Set	Test Set	Validation Set
Size	25,902 pairs per language	~3,400 reports per language	100 pairs per language
Languages	English, Spanish, French, Portuguese	English, Spanish, French, Portuguese	English, Spanish, Portuguese, German
Use	Training multilingual models	Evaluating model performance	Analyzing model behavior
Keyword Extraction	No	Yes	Yes

Figure 2: Dataset preparation and keyword extraction

standard training data—which serves as our validation dataset—consisting of 100 full-text and summary pairs for each language. This subset includes data in English, Spanish, Portuguese, and German, and was used to analyze the effect of language-specific context on model behavior. Following this, we aim to build a medical keyword dictionary; to facilitate this, we extract keywords from each full-text document in both the test and validation subsets using LLM-KB. These extracted keywords were later used to test and validate the models. Figure 2 shows the dataset preparation and keyword extraction.

3.2. Multilingual Keyword Translation and Quality Evaluation

For each medical concept in our test and validation sets, we translated the keywords into five languages—English, German, Spanish, French, and Portuguese. To ensure translation quality, we applied a back-translation technique: each translated keyword was back-translated, and LLM-KB was used to verify semantic equivalence between the original and the back-translated version. This filtering step ensured that only high-quality translations were retained, reducing potential confusion during summarization. The validated translations were then compiled into a multilingual medical keyword dictionary, which served as the foundation for generating structured prompts.

3.3. Medical Dictionary Construction and Prompt Engineering

We experimented with various prompting strategies, incorporating information from LLM-KB to assess which types of structured input contributed most to summarization performance. Ultimately, we selected MedCOD as our final prompting method, as it consistently delivered the most accurate results. These structured prompts enriched the input with contextual information, enabling the model to better understand sentence meaning and structure, thereby improving the quality of multilingual summarization. The prompts are presented in Table 1.

3.4. Fine-tuning with Low-Rank Adaptation (LoRA) and hyperparameters tuning

As illustrated in Figure 1, the MedCOD framework primarily functions as a prompting strategy that supplies external knowledge to LLMs. Beyond prompting, we further enhance the performance of

Table 1
Prompt Structures for Contextual Information

Prompt Structure
- <Concept X>: <Auxiliary language 1>: <Word X in auxiliary-language 1>, <Auxiliary language 2>: <Word X in auxiliary-language 2>.
- <Concept Y>: <Auxiliary language 1>: <Word Y in auxiliary-language 1>, <Auxiliary language 2>: <Word Y in auxiliary-language 2>.

open-source LLMs by fine-tuning them to better leverage the contextual information provided. To achieve this, we use LoRA [24], a lightweight and efficient fine-tuning technique that significantly reduces the number of trainable parameters. LoRA works by introducing a small set of trainable weights into the model while keeping the original parameters frozen. This method not only accelerates training and reduces memory usage but also produces compact model weight files—typically only a few hundred megabytes—making them easier to store, distribute, and deploy.

Instruction-tuned format (SFT mode), where each training sample consists of structured dialogue with distinct roles (e.g., system, user, assistant). The model is trained to predict only the assistant response, with loss masking applied to exclude input tokens from gradient updates. This is suitable for instruction-following tasks.

4. Results and Analyses

4.1. Experiment Settings

4.1.1. Datasets

In this study, we utilized the MultiClinSum dataset [3], which contains clinical case reports designed for multilingual summarization tasks across four languages: English, Spanish, French, and Portuguese. The dataset comprises 3,396 full-text clinical cases in English, 3,406 in Spanish, 3,469 in French, and 3,442 in Portuguese. We also used the MultiClinSum large-scale training datasets for model training across four languages. Each dataset contains 25,902 full-text and summary pairs in English , Spanish, French, and Portuguese. This rich multilingual dataset enables effective training, fine-tuning, and comprehensive evaluation of summarization models in the medical domain, supporting cross-lingual benchmarking and demonstrating the generalizability of our proposed framework. For the ablation study, we selected a small subset from the gold-standard training dataset in English, Spanish, German, and Portuguese, with each language containing 592 full-text and summary pairs. The test set size was intentionally kept small due to the high computational cost of applying our MedCOD framework to each instance using LLM-KB—a process that is both resource-intensive and time-consuming.

4.1.2. Models

We evaluated a range of open-source LLMs for multilingual medical summarization. Our experiments included Phi-4 (14B) [25], Qwen2.5-14B [6], and GPT-4o Mini [26] as baseline models. Additionally, we employed NLLB-200 3.3B [7] as a translation model to support and enhance the prompting methods used in our MedCOD framework.

Phi-4 (14B) Developed by Microsoft, Phi-4 is a 14-billion-parameter LLM that emphasizes high data quality through extensive use of synthetic data during training. Unlike earlier versions that relied heavily on distillation from teacher models like GPT-4, Phi-4 surpasses its predecessor in STEM-related question-answering tasks. This performance gain is primarily due to improved data generation and post-training techniques, while its architecture remains largely consistent with Phi-3.

Qwen2.5-14B Qwen2.5 is an advanced LLM series released by Alibaba Cloud, significantly enhanced through both pre-training and post-training phases. The pre-training data was expanded from 7 trillion to 18 trillion tokens, resulting in notable improvements in common sense reasoning, domain-specific knowledge, and general performance. Post-training involved over 1 million supervised samples and multi-stage reinforcement learning. Qwen2.5 includes multiple model sizes, with the 14B and 7B versions publicly available as open-weight models.

GPT-4o Mini GPT-4o Mini is a compact version of OpenAI's GPT-4o, designed for efficient deployment with minimal performance compromise. Released in July 2024, it supports both text and image inputs and maintains strong performance across various benchmarks, including MMLU (82%), MGSM (87.0%), and HumanEval (87.2%). With a 128K token context window and robust multilingual capabilities, GPT-4o Mini is well-suited for lightweight, multimodal reasoning tasks.

NLLB-200 3.3B Developed by Meta AI, NLLB-200 3.3B is a multilingual translation model capable of translating across 200 languages, including many low-resource languages. Leveraging a Sparsely Gated Mixture of Experts architecture, it achieves a 44% improvement in BLEU score over previous models, according to the FLORES-200 benchmark. NLLB-200 emphasizes both translation quality and safety, playing a key role in enhancing cross-lingual understanding within our framework.

4.1.3. Evaluation

We evaluated the summarization performance using two main metrics: ROUGE-L-Sum [27] and BERTScore [28].

ROUGE-L-Sum measures the overlap between generated summaries and reference summaries, focusing on the longest common subsequence to capture sentence-level similarity. It is particularly suitable for extractive summarization tasks, where key sentences from the original text are selected and combined. The calculation for ROUGE-L-Sum is built upon the standard ROUGE-L formulas for recall, precision, and F1-score, which are applied to each pair of sentences. For a given reference sentence (r) and a candidate sentence (c), the formulas are:

Recall (R_{lcs}) Measures what fraction of the reference sentence is captured in the candidate sentence.

$$R_{\text{lcs}} = \frac{\text{LCS}(r, c)}{\text{length}(r)}$$

where $\text{LCS}(r, c)$ is the length of the longest common subsequence of words, and $\text{length}(r)$ is the number of words in the reference sentence.

Precision (P_{lcs}) Measures what fraction of the candidate sentence is relevant compared to the reference sentence.

$$P_{\text{lcs}} = \frac{\text{LCS}(r, c)}{\text{length}(c)}$$

where $\text{length}(c)$ is the number of words in the candidate sentence.

F1-Score (F_{lcs}) The harmonic mean of recall and precision, providing a single, balanced score.

$$F_{\text{lcs}} = \frac{2 \cdot R_{\text{lcs}} \cdot P_{\text{lcs}}}{R_{\text{lcs}} + P_{\text{lcs}}}$$

BERTScore evaluates semantic similarity between the generated and reference summaries using contextual embeddings from pre-trained BERT models. Unlike ROUGE, it captures meaning beyond exact word matches. It reports three components: Precision (relevance of generated content), Recall (coverage of reference content), and F1-Score (harmonic mean of precision and recall), providing a balanced assessment of summary quality.

The calculation of BERTScore involves generating contextual embeddings for each token, computing their similarity, and then aggregating these values into precision and recall scores.

Let the reference sentence be a sequence of tokens $x = \langle x_1, x_2, \ldots, x_k \rangle$ and the candidate sentence be $\hat{x} = \langle \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_l \rangle$.

Contextual Embeddings Both sentences are passed through a pre-trained BERT model to obtain a sequence of contextual embedding vectors for each token. Let the embedding for token x_i be denoted as $\mathbf{x}_i$ and for $\hat{x}_j$ as $\hat{\mathbf{x}}_j$.

Similarity Matrix The cosine similarity is calculated for every pair of tokens between the reference and candidate sentences, creating a similarity matrix. The cosine similarity between two embedding vectors $\mathbf{a}$ and $\mathbf{b}$ measures their alignment.

Using the similarity scores, BERTScore computes recall, precision, and an F1-score through a greedy matching process.

Recall (R_{BERT}) Recall measures how well the candidate sentence captures the content of the reference sentence. For each token in the reference sentence, the metric finds the most similar token in the candidate sentence based on their embedding similarity. The recall score is the average of these maximum similarity scores.

The equation for recall is:

$$R_{\text{BERT}} = \frac{1}{k} \sum_{i=1}^{k} \max_{j=1,\ldots,l} (\mathbf{x}_i^T \hat{\mathbf{x}}_j)$$

Assuming the embedding vectors are normalized, their dot product $\mathbf{x}_i^T \hat{\mathbf{x}}_j$ is equivalent to their cosine similarity.

Precision (P_{BERT}) Precision measures how relevant the tokens in the candidate sentence are with respect to the reference sentence. For each token in the candidate sentence, it finds the most similar token in the reference sentence. Precision is the average of these maximum similarity values.

The equation for precision is:

$$P_{\text{BERT}} = \frac{1}{l} \sum_{j=1}^{l} \max_{i=1,\ldots,k} (\mathbf{x}_i^T \hat{\mathbf{x}}_j)$$

F1-Score (F_{BERT}) The F1-score is the harmonic mean of precision and recall, providing a single, balanced metric that reflects both accuracy and completeness.

The equation for the F1-score is:

$$F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

4.2. MultiClinSUM Official Evaluations

In this section, we present the official evaluation results of our system on the MultiClinSum dataset for summarizing clinical case reports in English, Spanish, French, and Portuguese. We submitted only one result after the deadline, which we included in the Table 2.

Table 2
Evaluation results on the MultiClinSum dataset across four languages using BERTScore and ROUGE metrics.

Language	BERTScore P	BERTScore R	BERTScore F1	ROUGE P	ROUGE R	ROUGE F1
English	0.8759	0.8412	0.8577	0.4029	0.2243	0.2574
Spanish	0.7551	0.7046	0.7283	0.4056	0.2067	0.2534
French	0.7269	0.6924	0.7082	0.3504	0.2114	0.2391
Portuguese	0.7493	0.6975	0.7217	0.3913	0.1892	0.2358

4.3. Ablation Study

We conducted an ablation study using 100 gold-standard annotated instances in four languages: English, Spanish, French, and Portuguese. The experiments were run on two different instruction-tuned models—Qwen2.5, Phi-4, and a proprietary GPT-4o-mini model—and evaluated using ROUGE-L and BERTScore metrics (Precision, Recall, F1).

Table 3
Comparison of summarization models across four languages using ROUGE-L (F1), ROUGE Precision, ROUGE Recall, and BERTScore metrics.

Model	Language	Finetuned	Context	ROUGE-P	ROUGE-R	ROUGE-L	BERTScore_P	BERTScore_R	BERTScore_F
GPT-4o-mini	English	No	-	0.265	0.239	0.2514	0.7656	0.8327	0.7974
	Spanish	No	-	0.272	0.249	0.2596	0.7739	0.8267	0.7992
	French	No	-	0.248	0.228	0.2377	0.7813	0.8265	0.8031
	Portuguese	No	-	0.236	0.213	0.2237	0.7649	0.8178	0.7902
Qwen2.5	English	No	No	0.271	0.256	0.2632	0.7863	0.7914	0.7883
		Yes	Yes	0.275	0.264	0.2697	0.7635	0.8136	0.7869
		Yes	No	0.268	0.262	0.2649	0.7849	0.7723	0.7772
		No	Yes	0.237	0.222	0.2288	0.7603	0.7691	0.7637
	Spanish	Yes	Yes	0.266	0.252	0.2590	0.7721	0.8100	0.7903
		Yes	No	0.270	0.259	0.2648	0.7863	0.7803	0.7825
		No	Yes	0.239	0.229	0.2337	0.7712	0.7608	0.7650
		No	No	0.056	0.051	0.0531	0.6273	0.5774	0.6010
	French	Yes	Yes	0.222	0.212	0.2171	0.7644	0.7826	0.7725
		No	Yes	0.179	0.170	0.1747	0.7596	0.7212	0.7383
		Yes	No	0.223	0.215	0.2192	0.7235	0.7102	0.7152
		No	No	0.057	0.051	0.0535	0.6318	0.5873	0.6084
	Portuguese	Yes	Yes	0.223	0.211	0.2169	0.7634	0.7892	0.7755
		Yes	No	0.253	0.241	0.2469	0.7792	0.7702	0.7740
		No	Yes	0.212	0.198	0.2050	0.7635	0.7537	0.7577
		No	No	0.105	0.096	0.1009	0.6668	0.6345	0.6493
Phi-4	English	Yes	No	0.260	0.243	0.2510	0.7610	0.8175	0.7873
		Yes	Yes	0.258	0.242	0.2503	0.7582	0.8150	0.7847
		No	No	0.270	0.252	0.2605	0.7798	0.7846	0.7814
		No	Yes	0.260	0.246	0.2527	0.7655	0.7774	0.7704
	Spanish	Yes	Yes	0.260	0.247	0.2529	0.7669	0.8002	0.7826
		Yes	No	0.241	0.229	0.2344	0.7641	0.7983	0.7800
		No	Yes	0.239	0.226	0.2323	0.7628	0.7654	0.7632
		No	No	0.145	0.134	0.1390	0.6846	0.6615	0.6719
	French	Yes	No	0.255	0.236	0.2450	0.7805	0.7963	0.7877
		Yes	Yes	0.240	0.225	0.2324	0.7736	0.7846	0.7784
		No	Yes	0.202	0.189	0.1954	0.7475	0.7379	0.7418
		No	No	0.139	0.126	0.1326	0.6971	0.6676	0.6808
	Portuguese	Yes	No	0.225	0.209	0.2170	0.7527	0.7896	0.7702
		Yes	Yes	0.220	0.208	0.2140	0.7373	0.7816	0.7584
		No	Yes	0.168	0.153	0.1601	0.7254	0.6950	0.7083
		No	No	0.145	0.134	0.1393	0.6957	0.6723	0.6819

4.3.1. Effect of MedCOD Context on Base LLM Performance

To evaluate the utility of MedCOD as an external contextual augmentation method, we compare the performance of base (non-finetuned) LLMs with and without contextual input. Across multiple languages and models, we consistently observe that the inclusion of MedCOD context improves summarization performance in terms of BERTScore and ROUGE-L. For instance, using Qwen2.5 in French, the BERTScore_F increases from 0.6084 (no context) to 0.7383 (with MedCOD context), and in Portuguese, from 0.6493 to 0.7577. Similarly, Phi-4 in Spanish improves from 0.6719 to 0.7632 with context. These results highlight MedCOD's effectiveness in guiding LLMs during summarization, especially for languages where models may lack sufficient domain coverage or training representation.

However, for English—the dominant language in most training corpora—LLMs already demonstrate strong capabilities in medical summarization. This may stem from data contamination (no one knows exactly which data has been used to train these models) or simply reflect the fact that current models are already sufficiently capable for such tasks in English. As a result, additional training or knowledge augmentation (e.g., using MedCOD) offers limited benefit. For example, in Phi-4, the difference in BERTScore_F between using context (0.7704) and no context (0.7814) is marginal, suggesting saturation in model performance.

4.3.2. Effect of MedCOD Context on finetuned LLM Performance

To further understand the impact of MedCOD beyond zero-shot settings, we assess its effect on finetuned LLMs. We finetuned our LLMs on a 25K supervised summarization dataset provided by MultiClinSUM, which covers diverse clinical narratives across multiple languages. Despite being trained on this substantial dataset, the addition of MedCOD context during inference still leads to measurable improvements or maintains strong performance across various languages. For instance, in Qwen2.5 (Spanish), the BERTScore_F improves from 0.7825 (finetuned, no context) to 0.7903 (finetuned, with context). Similarly, Phi-4 (French) shows BERTScore_F of 0.7784 with context and 0.7877 without, indicating competitive performance. In Portuguese, Qwen2.5 improves from 0.7740 to 0.7755 when MedCOD is provided. While these gains are smaller than those seen in base models, the results suggest that MedCOD can act as a valuable auxiliary signal, especially in complex domains like medical summarization where implicit knowledge and subtle cues are important. Even after extensive finetuning, contextual cues from MedCOD help reinforce or clarify information, potentially addressing coverage gaps not captured in the training data.

5. Discussion

In this study, we introduced MedCOD, a multilingual keyword-based contextual augmentation framework designed to enhance medical text summarization in low-resource language settings. Through extensive experiments on the MultiClinSUM dataset, we identified a number of key insights that illuminate both the strengths and limitations of current open-source LLMs for multilingual medical summarization.

Performance Saturation in English. For English—the dominant language in most LLM pretraining corpora—existing models already exhibit strong performance in medical summarization. As shown in Table 3, the base Qwen2.5 model, without context or fine-tuning, achieves a BERTScore$_{\mathrm{F}}$ of 0.7883, and fine-tuning brings only a marginal improvement (0.7869). Similarly, Phi-4 achieves 0.7873 in the fine-tuned setting without MedCOD context. These results suggest that English summarization tasks have reached a saturation point for current LLMs. The effectiveness may stem from data contamination (since the actual pretraining corpus is unknown for these models) or simply from the inherent advantages in English-centric training pipelines. As a result, additional knowledge augmentation—either via MedCOD or parameter-efficient fine-tuning—yields minimal or even negative gains.

Challenges and Error Patterns in Non-English Settings. In contrast, performance in Spanish, French, and Portuguese is consistently lower, especially in zero-shot scenarios. For example, the Portuguese summarization task under the base Qwen2.5 setting (no fine-tuning, no context) yields a BERTScore$_{\mathrm{F}}$ of only 0.6493, while French yields 0.6084 in the same configuration. This aligns with known disparities in language representation across pretraining corpora, where non-English languages—particularly Portuguese and French—are significantly underrepresented [7, 29, 30]. Consequently, models often fail to follow instructions or generate outputs in the target language, instead defaulting to English. We observed several failure cases in which the models produced fluent summaries in English, despite being explicitly prompted in Spanish, French, or Portuguese, as shown in Table 4.

Table 4
Examples of summarization failures in multilingual settings where the model defaulted to English despite being prompted in the target language.

Language	Full Text Snippet	Generated Summary (English)	Failure Type
Spanish	*La paciente era una madre de 54 años... llegó al hospital con una masa vaginal...*	*The patient, a 54-year-old woman from a rural area, presented with a vaginal mass of 3 years' duration...*	Output in wrong language (English)
French	*Une femme de 80 ans... douleurs aiguës dans sa cuisse droite...*	*An 80-year-old woman admitted to the emergency department due to significant weight loss and acute thigh pain...*	Output in wrong language (English)
Portuguese	*Homem de 56 anos com febres intermitentes e história de hipertensão...*	*Here is a concise and clear summary of the medical text: A 56-year-old man with intermittent fevers and prior hypertension...*	Partial translation: starts in PT, switches to EN

These behavior patterns directly affect reference-based metrics like ROUGE and BERTScore, leading to abnormally low scores due to mismatched output languages.

Benefits of MedCOD and Fine-Tuning in Non-English Settings. Both MedCOD and fine-tuning significantly enhance performance in non-English summarization tasks, primarily by improving language adherence and fluency. For instance, Qwen2.5 in Portuguese improves from 0.6493 (no context, no fine-tuning) to 0.7577 when only MedCOD is used. Likewise, Spanish performance with Phi-4 improves from 0.6719 to 0.7632 in the same comparison. These gains are primarily due to better alignment with the target language, as shown in the Table 5.

Table 5
Effect of MedCOD on Spanish Clinical Note Summarization

Aspect	Without MedCOD (Baseline)	With MedCOD (Improved)
Source (excerpt)	*Un hombre de 27 años, fumador sin antecedentes patológicos, sufrió un accidente en motocicleta sin casco. Presentó una lesión craneal penetrante con hemorragia intracraneal. Fue trasladado inconsciente a un hospital mejor equipado y se le administraron anticonvulsivos en el postoperatorio.*	
Generated Summary	*Un hombre de 27 años... hemorragia intracraneal... tratamiento con intubación, fluidos y antibióticos... fue trasladado a un hospital de primera categoría.*	*Un hombre de 27 años... presentó lesión penetrante en la cabeza... fue trasladado inconsciente a un hospital de primera categoría, libre de intubación... el paciente fue puesto en anticonvulsivos con buen resultado.*
Fluency & Coherence	Moderate; partially disjoint and overly technical list format	High; fluent transitions and cohesive structure
Clinical Relevance	Generic; lacks focus on critical intervention outcomes	Specific; includes key events like anticonvulsant treatment and surgical outcome
Language Alignment	Acceptable Spanish but includes translation artifacts	Native-like phrasing with proper medical expressions

Fine-tuning provides stronger improvements by explicitly exposing the model to task-specific and language-specific patterns. However, even without training, MedCOD achieves meaningful zero-shot gains through prompt-level augmentation. This is especially important in settings where computational resources are limited, making full fine-tuning impractical. MedCOD achieves these gains by inserting target-language medical keywords as context, which serves as both a language anchor and a domain signal. This aligns with observations in prompt-based adaptation literature [31, 32, 4].

MedCOD Combined with Fine-Tuning Achieves Best Results. The combination of MedCOD and fine-tuning yields the best performance across most non-English tasks. For example, in the Spanish task, Qwen2.5 with both fine-tuning and MedCOD achieves a BERTScore$_F$ of 0.7903, outperforming all other configurations. This supports our hypothesis that MedCOD provides complementary contextual grounding that reinforces the representations learned during fine-tuning. As described in prior work on prompting [4], providing high-salience, low-ambiguity input tokens can help direct LLM attention and reduce reasoning drift. In our setting, MedCOD's multilingual keyword chains offer just such signal, particularly effective when LLMs face under-represented linguistic domains which show in table 6.

Table 6
Example of Fine-Tuning (FT) versus FT+MedCOD impact on summary quality

Aspect	FT-Only Summary	FT + MedCOD Summary
Language	Portuguese	Portuguese
Structure	Concise but incomplete; lacks detailed clinical timeline and interventions.	Detailed and well-organized; preserves clinical history, echocardiographic data, and treatment chronology.
Factuality	Omits key facts such as hypertension history, valve vegetation, and surgical treatment.	Includes specific medical facts: patient demographics, valve regurgitation, vegetation size, and antibiotic treatment.
Example snippet	*"Paciente com febre e dor, tratado genericamente."*	*"Paciente de 56 anos com hipertensão, vegetação valvular detectada e cirurgia realizada."*

Limitations and Future Work. Despite promising results, several limitations remain:

- **Information Overload in Input.** MedCOD expands the prompt with multiple language tags and keyword chains, which may overwhelm the model. In some cases, we observed long, disorganized outputs—likely caused by the model attending to less relevant context tokens. We refer to this as "COD explosion." Future work can explore input filtering strategies, such as perplexity-guided token selection or keyword salience ranking, as discussed in [33, 34, 35].
- **Minimal Gains in English.** As noted earlier, injecting additional task-specific knowledge (via MedCOD or LoRA) in English settings offers little benefit. Worse, overly long prompts may distract from key content or introduce inconsistencies. This suggests a need for adaptive prompting or context compression strategies. Additionally, test-time adaptation methods [36]—e.g., self-consistency [37], self-refinement [38], test-time training [39]—may yield more value in saturated English domains.
- **Unexplored Factors.** Due to time and resource constraints, several areas remain under-investigated: (1) The discrepancy between ROUGE and BERTScore in some languages (e.g., Spanish vs. French) in shown in table. (2) Comparison between monolingual and multilingual fine-tuning setups; (3) More analysis about whether MedCOD improves not just linguistic consistency but also content structure or factuality.

Looking ahead, we see significant potential for MedCOD to support a range of patient-centered multilingual clinical NLP applications. Specifically, the ability to generate accurate and readable summaries across multiple languages can benefit real-world scenarios such as: (1) Patient-facing summaries, which simplify complex clinical language to improve health literacy; (2) Discharge summaries, which offer clear and concise overviews of hospital visits to support care transitions; (3) Medical literature summarization, which distills key findings and methodologies from multilingual scientific publications; (4) Multilingual clinical communication, where summarization combined with translation facilitates cross-lingual understanding of medical records; and (5) Telemedicine and remote consultations, where concise summaries of patient data support efficient triage and diagnostic workflows. Indeed, our motivation for participating in this shared task was to lay the groundwork for such patient-oriented multilingual applications [1]. While the scope of this work was constrained by the competition's structure, we believe that the dataset provided by the organizers serves as a strong and practical foundation for future research aimed at advancing these patient-centered multilingual application scenarios. Moreover, our findings in this task align with a broader and well-documented challenge: many patient-centered tasks suffer from imbalanced training data across languages, resulting in relatively strong performance in English but substantial degradation in other widely spoken languages, such as Spanish, French, and Portuguese. This disparity is particularly concerning given that patients who are not native English speakers often have the greatest need for accessible, high-quality clinical NLP tools. Our future work will therefore focus on extending our prior patient-centered BioNLP research [40, 41, 42, 43, 44]—largely concentrated in English—to underrepresented languages by leveraging this resource, with the goal of developing equitable, multilingual medical summarization systems that are not only clinically accurate but also understandable and actionable for patients across diverse linguistic backgrounds.

6. Conclusion

In this work, we introduced MedCOD, a multilingual, keyword-based contextual augmentation framework aimed at enhancing medical text summarization in low-resource language settings. Our comprehensive experiments on the MultiClinSUM dataset, covering English, Spanish, French, and Portuguese, demonstrate that MedCOD improves performance, particularly in non-English languages where baseline models often struggle. While existing open-source LLMs, such as Qwen2.5 and Phi-4, already achieve strong results in English—with minimal improvements from fine-tuning or context—our findings reveal substantial gaps in Spanish, French, and Portuguese. These include frequent failures in language adherence and factual completeness, which directly affect summarization quality and evaluation metrics.

[1] https://temu.bsc.es/multiclinsum/

MedCOD addresses these challenges by incorporating task-relevant, target-language medical keywords into the input, acting as both a domain signal and a language anchor. This strategy improves fluency, coherence, and clinical relevance of the generated summaries, even in zero-shot scenarios where fine-tuning is not feasible. Furthermore, combining MedCOD with fine-tuning yields the best performance across all non-English tasks. For instance, in Spanish, Qwen2.5's BERTScore$_\text{F}$ improves from 0.7650 (no context) to 0.7903 with the addition of MedCOD context and fine-tuning. These results confirm that MedCOD provides a practical and effective solution for overcoming the multilingual limitations of current LLMs and supports the development of equitable, language-inclusive clinical NLP systems.

Acknowledgments

This research was supported by the U.S. Department of Veterans Affairs (VA) under award VA HSR IIR 24-083 (I01HX003969). The content is solely the responsibility of the authors and does not necessarily represent the views of the VA or the U.S. government.

Declaration on Generative AI

During the preparation of this work, the author(s) used AI in order to perform grammar and spelling checks. No generative AI tools were used to produce images or other creative content. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] W. Liu, S. Cai, Translating electronic health record notes from english to spanish: A preliminary study, in: Proceedings of BioNLP 15, 2015, pp. 134–140.

[2] 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.

[3] 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 Working Notes, CEUR Workshop Proceedings, 2025. To appear.

[4] H. Lu, H. Yang, H. Huang, D. Zhang, W. Lam, F. Wei, Chain-of-dictionary prompting elicits translation in large language models, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 958–976.

[5] M. S. Salim, L. Fu, A. A. Ramakrishnan, Z. Yao, Enhancing english-to-spanish medical translation of large language models using enriched chain-of-dictionary framework, 2025. Under review.

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

[7] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022). URL: https://arxiv.org/abs/2207.04672.

[8] E. Gesner, P. Gazarian, P. Dykes, The burden and burnout in documenting patient care: an integrative literature review, MEDINFO 2019: Health and Wellbeing e-Networks for All (2019) 1194–1198.

[9] R. M. Ratwani, E. Savage, A. Will, R. Arnold, S. Khairat, K. Miller, R. J. Fairbanks, M. Hodgkins, A. Z. Hettinger, A usability and safety analysis of electronic health records: a multi-center study, Journal of the American Medical Informatics Association 25 (2018) 1197–1201.

[10] J. M. Ehrenfeld, J. P. Wanderer, Technology as friend or foe? do electronic health records increase burnout?, Current Opinion in Anesthesiology 31 (2018) 357–360.

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[12] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2025. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

[13] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL: https://arxiv.org/abs/2303.12712. arXiv:2303.12712.

[14] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, 2023. URL: https://arxiv.org/abs/2211.09110. arXiv:2211.09110.

[15] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[16] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko, Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. URL: https://arxiv.org/abs/1712.05877. arXiv:1712.05877.

[17] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, E. Horvitz, Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023. URL: https://arxiv.org/abs/2311.16452. arXiv:2311.16452.

[18] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (2023) 172–180.

[19] Y. Liu, et al., Multilingual denoising pre-training for neural machine translation, arXiv preprint arXiv:2001.08210 (2020). URL: https://arxiv.org/abs/2001.08210.

[20] H. Lu, H. Yang, H. Huang, D. Zhang, W. Lam, F. Wei, Chain-of-dictionary prompting elicits translation in large language models, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 958–976.

[21] W. Liu, S. Cai, Translating electronic health record notes from english to spanish: A preliminary study, in: BioNLP 2015, 2015, pp. 134–140.

[22] H. Chintakunta, M. Zhang, S. Shaar, et al., Multilingual biomedical translation benchmarks, arXiv preprint arXiv:2301.02500 (2023). URL: https://arxiv.org/abs/2301.02500.

[23] S. Khare, B. Gholami, et al., Melot: A medical language translation dataset, arXiv preprint arXiv:2307.07955 (2023). URL: https://arxiv.org/abs/2307.07955.

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

[25] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, Y. Zhang, Phi-4 technical report, 2024. arXiv:2412.08905.

[26] OpenAI, Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

[27] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization

Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675. `arXiv:1904.09675`.

[29] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model (2023).

[30] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., Opt: Open pre-trained transformer language models, arXiv preprint arXiv:2205.01068 (2022).

[31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM computing surveys 55 (2023) 1–35.

[32] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, et al., Language models are multilingual chain-of-thought reasoners, arXiv preprint arXiv:2210.03057 (2022).

[33] W. S. Jang, S. Sultana, Z. Yao, H. Tran, Z. Yang, S. Kwon, H. Yu, Enhancing llms for identifying and prioritizing important medical jargons from electronic health record notes utilizing data augmentation, arXiv preprint arXiv:2502.16022 (2025).

[34] J. Chen, E. Druhl, B. Polepalli Ramesh, T. K. Houston, C. A. Brandt, D. M. Zulman, V. G. Vimalananda, S. Malkani, H. Yu, A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews, Journal of medical Internet research 20 (2018) e26.

[35] J. Chen, J. Zheng, H. Yu, et al., Finding important terms for patients in their electronic health records: a learning-to-rank approach using expert annotations, JMIR medical informatics 4 (2016) e6373.

[36] C. Snell, J. Lee, K. Xu, A. Kumar, Scaling llm test-time compute optimally can be more effective than scaling model parameters, arXiv preprint arXiv:2408.03314 (2024).

[37] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, arXiv preprint arXiv:2203.11171 (2022).

[38] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al., Self-refine: Iterative refinement with self-feedback, Advances in Neural Information Processing Systems 36 (2023) 46534–46594.

[39] J. Hu, Z. Zhang, G. Chen, X. Wen, C. Shuai, W. Luo, B. Xiao, Y. Li, M. Tan, Test-time learning for large language models, arXiv preprint arXiv:2505.20633 (2025).

[40] H. Tran, Z. Yao, L. Li, H. Yu, Readctrl: Personalizing text generation with readability-controlled instruction learning, arXiv preprint arXiv:2406.09205 (2024).

[41] S. Kwon, Z. Yao, H. S. Jordan, D. A. Levy, B. Corner, H. Yu, Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2022, 2022, p. 11733.

[42] Z. Yao, N. S. Kantu, G. Wei, H. Tran, Z. Duan, S. Kwon, Z. Yang, H. Yu, et al., Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp, arXiv preprint arXiv:2312.15561 (2023).

[43] P. Cai, Z. Yao, F. Liu, D. Wang, M. Reilly, H. Zhou, L. Li, Y. Cao, A. Kapoor, A. Bajracharya, et al., Paniniqa: Enhancing patient education through interactive question answering, Transactions of the Association for Computational Linguistics 11 (2023) 1518–1536.

[44] Z. Yao, H. Yu, A survey on llm-based multi-agent ai hospital (2025).