Retrieve, Generate, Evaluate: A Case Study for Medical Paraphrases Generation with Small Language Models

Ioana Buhnila*♠, **Aman Sinha***♠,♣ and **Mathieu Constant**♠
♠ATILF, CNRS, Université de Lorraine, Nancy, France
♣Institut de Cancérologie, Strasbourg, France
`firstname.lastname@univ-lorraine.fr`

Abstract

Recent surge in the accessibility of large language models (LLMs) to the general population can lead to untrackable use of such models for medical-related recommendations. Language generation via LLMs models has two key problems: firstly, they are prone to hallucination and therefore, for any medical purpose they require scientific and factual grounding; secondly, LLMs pose tremendous challenge to computational resources due to their gigantic model size. In this work, we introduce **pRAGe**, a **p**ipeline for **R**etrieval **A**ugmented **G**eneration and **e**valuation of medical paraphrases generation using Small Language Models (SLM). We study the effectiveness of SLMs and the impact of external knowledge base for medical paraphrase generation in French.

1 Introduction

Large Language Models (LLMs) are used for a big variety of NLP tasks and are found to be very effective, but they exhibit a specific trait that can make them unreliable: *hallucinations* (Zhang et al. 2023; Huang et al. 2023). LLM hallucinations are incorrect output generations that do not correspond to the input prompt, or are not factual information reflecting world knowledge. LLMs such as Chat-GPT (OpenAI, 2022) became widely used by lay people and the risk of incorrect information spreading within medical text generation persists. In very specialized and sensitive fields such as medicine, hallucinations can have dangerous consequences for the patient via wrong prognosis or treatment recommended by LLMs (Umapathi et al., 2023). This leads to the necessity that the information patients receive to be scientifically and factually grounded, either by human expert or external knowledge base. Additionally, recent state-of-the-art (SOTA) results

	Term	Paraphase
fr	hypopnée	respiration partiellement bloquée
en	hypopnea	partially blocked breathing
fr	myasthénie grave	est un trouble qui entraîne une faiblesse musculaire et une fatigue musculaire excessive
en	myasthenia gravis	is a condition that leads to muscle weakness and excessive muscle fatigue
fr	akathisie	agitation intérieure et incapacité à rester assis
en	akathisia	inner restlessness and inability to sit still

Table 1: Examples of medical term paraphrase in French (fr) and its translation in English (en) from RefoMed dataset. Each term is a medical term automatically identified with SNOMED-3.5VF, and corresponding paraphrase represents a sub-sentential paraphrase.

are using LLMs with dozens or hundreds of billion parameters. Their results cannot be easily reproduced due to high costs of GPUs for finetuning and inference. GPT-4 (Achiam et al., 2023) is at the top for many NLG tasks, but the model is not open-source and API usage can become costly during experiments.

Several works such as (LeBlanc et al., 2014; Tavakoly Sany et al., 2020) report frequent misunderstanding which caused due to medical jargon (i.e. medical terms) usage which are highly present in doctor-patient interactions because patients' different health-literacy levels. Such terms are lexical units that designate a concept from a specialised domain (Condamines, 1997). These complex terms have to be adapted or simplified for lay people through paraphrases, short definitions or explanations. We consider that short sequences of words (shorter than a sentence) can provide fast explanations adapted to patients needs of understanding as shown in Table 1.

Therefore, in this work we aim to develop a

---

method of medical term paraphrase generation that can help patients and their families better understand and follow the treatment of their illness. In particular, we focus on the generation of **subsentential paraphrases**, which are defined as simple words or sequences of words (Bouamor et al., 2013; Max et al., 2012) for better understanding of technical terms. We introduce **pRAGe**, an augmented SLM (Small Language Model) pipeline that generates paraphrases and short definitions for medical terms based on patient query. We make use of RALMs (Retrieval Augmented Language Models), which combine a RAG architecture and a language model (LM). RAG (Retrieval Augmented Generation) (Lewis et al., 2020) models help reduce the level of hallucinations generated by LMs by accessing an external knowledge base (KB) to retrieve the answer for an input prompt.

We conduct our RAG experiments on a downstream natural language generation (NLG) task: medical paraphrase generation (Gupta et al., 2018), paired with a question-answering (Q&A) task (Singhal et al., 2023). Our paper focuses on open-source Small Language Models (Schick and Schütze, 2021) and cost efficient quantization methods that can allow our experiments to be easily reproduced by the community. Hence, we aim to answer the following research questions:

- **RQ1**: *How good are open-source small language models (1 to 7B parameters) quantized VS fine-tuned models at medical Q&A task in a RAG system?*

- **RQ2**: *What is the impact of finetuning VS prompting in the medical paraphrase generation task?*

- **RQ3**: *How can we evaluate the quality of RAG systems for medical paraphrase generation?*

The main contributions of our paper are: **(1) pRAGe** (*pipeline for Retrieval Augmented Generation and evaluation*), an open source RAG pipeline for medical paraphrase and explanation generation in a Q&A downstream task making use of only non proprietary LLMs; **(2) RefoMed-KB** (*Medical Paraphrases Knowledge Base*), an French medical knowledge base about scientific medical terms extracted from the RefoMed dataset[1]; **(3)**

pRAGe-FT, a fine-tuned RAG model[2] for the paraphrase generation task in French and an evaluation of its performance on zero-shot inference Q&A task. We share our code, datasets and evaluation metrics with the NLP community to support open-access and reproducible research[3].

2 Related Work

SOTA Large Language Models such as GPT-3.5 (OpenAI, 2022), GPT-4 (Achiam et al., 2023), Mistral and Mixtral (Jiang et al., 2023) and Llama-2 (Touvron et al., 2023) can give impressive results on question-answering tasks (Tan et al., 2023), or radiology reports simplification (Jeblick et al., 2023). However, these models are either not open-source (GPT) or may require access to powerful servers. Smaller models (<7B) are easier and feasible to implement in downstream tasks, but they tend to hallucinate more compared to LLMs. In health related applications, we need to retrieve correct information, thus RAG systems are essential.

Different types of RAG systems have been developed (Gao et al., 2023), going from the original naive RAG (simple structure of a retriever and a generator) (Lewis et al., 2020) to more advanced or modular RAG such as RA-DIT (Lin et al., 2023). There are several types of advanced RAG architectures such as the self-reflective Self-RAG (Asai et al., 2023), Self-BioRAG (Jeong et al., 2024), black box RAG like RePlug (Shi et al., 2023), pretrained model RAG such as REALM (Guu et al., 2020), RETRO (Borgeaud et al., 2022) or ATLAS (Izacard et al., 2022). However, these advanced RAG systems have very complex implementation architectures and require servers with many GPUs. Almanac (Zakka et al., 2024) is a RAG system developed for the clinical domain , but it is not open source and it is available only for English. We are interested in developing easy to implement RAG systems for languages other than English, in our case for French. There are some French language models, such as CamemBERT (Martin et al., 2019) for the general domain and DrBERT (Labrak et al., 2023) for the medical domain, but they are not adapted for text-to-text generation tasks.

Multilingual decoder models, such as Mixtral, Falcon, Bloom, and French models such as Vigogna (based on Vicuna) or Claire are available for
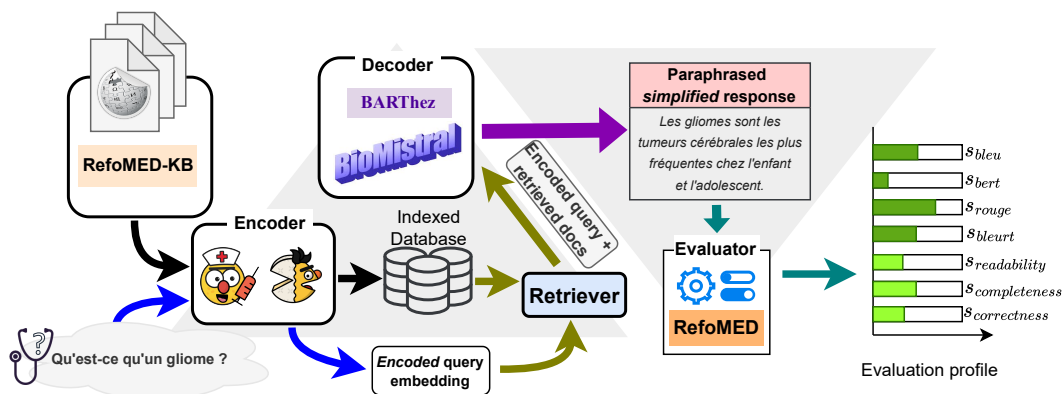

Figure 1: Illustration of pRAGe experimental pipeline. The illustration is intended to read from left to right. Each colored arrow represent a process. The ■ arrow indicates the **creation of indexed database**; the ■ arrow indicates the **encoding of the query**; the ■ arrow represents **retrieval of relevant documents**; the ■ arrow denotes the **generation of simplified paraphrase output** and the ■ arrow indicates the **evaluation step for the generated output** to obtain the evaluation profile of the paraphrase.

general language. As for encoder-decoder French models trained for NLG task (more precisely for summarization tasks), we cite BARThez (Eddine et al., 2020), fine-tuned on BART. Many open-source medical LLMs for NLG tasks are available for English, such as PMC-Llama (Wu et al., 2024), MediTron (Chen et al., 2023) (both based on LLama-2), ClinicalGPT (Wang et al., 2023), but there is no NLG medical model trained exclusively on the French language. Multilingual medical models such as Medical-mT5 (García-Ferrero et al., 2024) and BioMistral (Labrak et al., 2024) are of interest for our research. One important aspect when working with LLMs on medical text is the quality of the generation in order to minimize the hallucination risk. LLMs can generate content according to a specific prompt.

However, we want to improve the quality of the output and obtain a short paraphrase of the term and not a complete description of the term (as shown in Table 1). Thus, we use prompt tuning techniques. However, we need curated sub-sentential paraphrase datasets in medical French for this task. Most of the paraphrase datasets contain only sentential paraphrases from general language in English: MSRP (Dolan et al., 2004), PPDB (Ganitkevitch and Callison-Burch, 2014), PAWS (Zhang et al., 2019b) or multilingual: TaPaCo (Scherrer, 2020) or ParaCotta (Aji et al., 2022).

French sentential paraphrase databases are scarce. One such resource for the medical domain is the WikiLarge FR and CLEAR, a parallel corpus for text simplification built with translation

SLM	FR Encoder-Decoder	
BaseSLM	BARThez-orangesum-abstract	
	BioMistral-7B-SLERP-GPTQ	
pRAGe	**FR Encoder**	**FR Decoder**
BioMistral	DrBERT	BioMistral-7B-SLERP-GPTQ
pRAGe	sent-CamemBERT	BioMistral-7B-SLERP-GPTQ
BARThez	DrBERT	BARThez-orangesum-abstract
pRAGe	sent-CamemBERT	BARThez-orangesum-abstract

Table 2: Configurations of non proprietary French encoders and decoders tested in our experiments.

techniques (Cardon and Grabar, 2020). However, there is little work on sub-sentential paraphrases, as they are difficult to identify. Previous work explored crowdsourcing methods (Tschirsich and Hintz, 2013) or translation techniques (Bouamor et al., 2013) (Zhai et al., 2020). A dataset similar to our work is PARADE (He et al., 2020), containing computer science definition-style paraphrases for English technical concepts extracted from online user-generated flashcards.

3 Methodology

We illustrate our method in Figure 1. **pRAGe** is built on a encoder-retriever-decoder framework. We designed **pRAGe** to embed medical query and to generate an output in a style that *translates* medical knowledge for patients in a simpler language, e.g. *rhizarthrose → arthrose du pouce* (rhizarthrosis → arthrosis of the thumb). Therefore, we pair models for the general language with medical models. We tested different configurations of non proprietary encoders and decoders, as shown in Table 2. We used the gen-

eral French encoder model **sent-CamemBERT** ([Reimers and Gurevych 2019](); [Martin et al. 2020]()) and the domain specific model **DrBERT** ([Labrak et al., 2023]()), a French BERT type model for the medical field. The **pRAGe** pipeline encodes in embeddings the input query and the Wikipedia knowledge base, RefoMed-KB (to be presented in section 3.1). We did prompt engineering to guide the decoder towards the expected output in both experimental settings, base SLM inference and pRAGe pipeline. The task attributed to the SLM is "to answer the user's question with a paraphrase, explanation or short definition" (full prompts in Figure 2 and in Appendix A.1). We used **BARThez-OrangeSum-abstract** (BARThez[4]) ([Eddine et al., 2020]()), a French seq2seq SLM, and **BioMistral-7B-SLERP-GPTQ**[5] ([Labrak et al., 2024]()), a 4-bit precision GPTQ quantized ([Frantar et al., 2022]()) multilingual medical model for training and inference efficiency[6]. We chose the BioMistral-7B-SLERP model ([Shoemake, 1985]()) as it gave the best benchmark results on French datasets, according to the authors of the model ([Labrak et al., 2024]()). We also tested the impact of **finetuning the SLMs** in our **pRAGe** system on an existing sub-sentential paraphrase dataset in medical French, **RefoMed** ([Buhnila, 2023]()). For finetuning we used the **Q-LoRA** method ([Dettmers et al., 2024]()), a computational efficient finetuning method that reduces the number of parameters for BioMistral from 7B to 1,38B parameters. We present the RefoMed dataset in the next section and the evaluation metrics for **pRAGe** generated paraphrases in section 3.2.

3.1 Datasets

We split the RefoMed dataset for finetuning in training, validation and test sets. The validation and test sets were used to built the knowledge base for the RAG system, RefoMed-KB. We present both datasets below.

RefoMED We used an unique open-source dataset of medical sub-sentential paraphrases in French, **RefoMed**[7] ([Buhnila, 2023]()) as input queries and to finetune the SLMs in **pRAGe**. The RefoMed

corpus is made of 6,297 pairs of unique medical terms and their corresponding sub-sentential paraphrases. The source corpora are **ClassYN** ([Todirascu et al., 2012]()) and **CLEAR Cochrane** ([Grabar and Cardon, 2018]()), both comparable corpora of scientific and simplified medical short texts in French. The RefoMed dataset was built by automatically extracting sentences that contain medical terms and paraphrases from the source corpora. The author ([Buhnila, 2023]()) identified medical terms automatically by using a rule-based method and the **SNOMED-3.5VF** French medical terminology ([Cote, 1998]()).The paraphrases were identified with the help of linguistic paraphrase markers such as *c'est-à-dire* ("so called"), *également appelé* ("also called"), *est une maladie* ("is a disease"), and punctuation signs, such as colons and brackets ([Grabar and Hamon 2015](); [Antoine and Grabar 2016](); [Buhnila 2022]()).

In order to avoid bias in the LM's finetuning, we split the dataset by unique term entry while staying in the range of the classic 60-20-20 train-validation-test split proportion. This split was important because in the RefoMed dataset we can find multiple paraphrases for one particular term. For instance, the term "placebo" has various paraphrases: (1) "absence d'intervention" / no intervention; (2) "médicament inactif" / inactive drug; (3) "traitement factice" / fake treatment; (4) "par exemple une pilule de sucre" / for example a sugar pill; (5) "aucun traitement" / no treatment. Thus, the resulting split is as follows: 3,981 term-paraphrase pairs for training, 1,063 for validation and 1,253 pairs for testing.

Descriptive statistics for the paraphrases in the RefoMed dataset. We counted the length of the paraphrases in RefoMed. The shortest paraphrase is of 1 word length whereas the longest is 83 words length. The mean and standard deviation is 10.34 and 8.15 respectively. Accounting the above, we consider 10, 25 and 50 word count as the limit of paraphrase generation for our various RAG systems. For our final analysis, we considered token limit of 25 and 50.

RefoMED-KB The next step was to build the knowledge base (KB) for the medical terms from the validation and testing sets. We automatically extracted top-3 Wikipedia articles where the terms appear in the title of the article using the Python *wikipedia* library. We extracted the first 20 lines of

---

[4]For readability reasons, we will hence refer to BARThez-OrangeSum-abstract as BARTHEZ.

[5]BioMistral was pre-trained on 3 bilion tokens data from PubMed Central from Mistral. Less than 1,25% of the data is a GPT-3.5 Turbo automatic translation in French and other 8 languages.

[6]https://huggingface.co/LoneStriker/BioMistral-7B-SLERP-GPTQ

[7]https://github.com/ibuhnila/refomed

each relevant wikipedia page and we obtain a medical knowledge base in French of 20,402 sentences (1,708,034 tokens) about the 1,253 medical terms from the test list.

3.2 Automatic Evaluation

We develop an evaluation method for your system to tackle **(RQ3)**: *How can we evaluate the quality of RAG systems medical paraphrase generations?* Evaluation for the complete RAG framework can be divided into two categories: intrinsic and extrinsic. For extrinsic evaluation, we check for hallucination by evaluating the rate of medical correct answers (Huang et al., 2023). For intrinsic evaluation, we perform manual evaluation by checking the quality of responses generated.

Evaluation metrics. Several metrics are used in SOTA research on text generation evaluation: **ROUGE** (Lin, 2004), calculates the n-grams overlap (recall), **BLEU** (Papineni et al., 2002), computes the number of similar n-grams between the output and the reference (precision), **BERTscore** (Zhang et al., 2019a), compares the embeddings of tokens that match in the output and reference text, while **BLEURT** (Sellam et al., 2020), computes the semantic similarity and lexical difference between them. **MEDCON** (Yim et al., 2023) is a metric that computes the F1-score of the UMLS concepts found both in the output and the reference text (however, available only for English). In our work, we want to evaluate the similarity between the generated output text and the reference text in French. We use the following evaluation metrics: bleu, rouge, bleurt, and bertscore.

RAGref**S** (S) We define a score metric for evaluating the generated response set from pRAGe. For any i$_{th}$ query, if the p_i is the generated response from the RAG pipeline and $\mathcal{R}_i$ is the list of reference paraphrases.

$$ S_\Omega = \frac{\sum_{i=1}^N \mathsf{max}(\{\Omega(p_i, r_{ij}) \forall r_{ij} \in \mathcal{R}_i\})}{N} \quad (1) $$

where N is number of queries and Ω is a lexical or semantic similarity comparison metric such as bleu, rouge, bleurt, bertscore, etc.

3.3 Fine-grained Human Evaluation

We conduct a fine-grained evaluation of the generated paraphrases. Firstly, we automatically evaluate the generation quality with SOTA metrics

(bleu, rouge, bleurt, bertscore) with our metric, RAGrefS (S) as introduced in section 3.2. In addition to the overall generation quality, we study the individual generation quality to obtain more insights. A set of 1200 examples[8] were manually analyzed by 3 French proficient linguist annotators following different criteria:

- readability: scored from *1* to *3*, where *1* means that the generated text is fluent, grammatically correct and easy to understand for laypeople; *2* means the generated text includes invented words, English words or grammatical mistakes, or scientific terms in a grammatically correct context; *3* represents generated text that has the incorrect traits of score *2*, plus scientific terms, rendering the text difficult to understand for laypeople;

- completeness: the generated text represents a full answer, meaning the language model generated a concise answer (score *1* if the text respects this condition, *0* if not). We annotated two types of completeness: **relaxed** - the generated text contains one incomplete sentence or it contains a second incomplete sentence, and **strict** - the generated text contains one syntactically independent sentence;

- correctness: the generated text encompasses the correct medical knowledge and it is in French[9] (score *1* if the two conditions are fulfilled, *0* if not). We considered two types of correctness: **relaxed** - the general meaning of the medical term is comprehensible from the generated text, and **strict** - the exact meaning is both comprehensible and complete.

4 Experiments

In this section, we describe the experiments we conducted to study the comparison between SLMs and pRAGe models for generating paraphrases for medical terms. We also describe the fine-grained evaluation process in the subsections below.

4.1 SLMs Zero-Shot Inference *VS* pRAGe

Firstly, we test SLMs ability to generate medical paraphrases in an zero-shot setting. We chose

---

[8]50 examples from 24 different configurations, as presented in Table 7.

[9]During our validation experiments we noticed the presence of English answers in the generated answers, especially for BioMistral (the model has less then 1% of its training data in French.)

	Model Setup		Tokens=25				Tokens=50			
			bert	bleurt	bleu-1	rouge-1	bert	bleurt	bleu-1	rouge-1
						w/o FINE TUNING				
SLM	BARTHEZ		$0.63_{0.03}$	$0.10_{0.10}$	$0.04_{0.06}$	$0.07_{0.08}$	$0.63_{0.03}$	$0.10_{0.10}$	$0.04_{0.06}$	$0.07_{0.08}$
	BIOMISTRAL		$\mathbf{0.70}_{0.06}$	$\mathbf{0.15}_{0.15}$	$\mathbf{0.11}_{0.12}$	$\mathbf{0.20}_{0.16}$	$\mathbf{0.68}_{0.06}$	$\mathbf{0.16}_{0.15}$	$\mathbf{0.08}_{0.08}$	$\mathbf{0.18}_{0.13}$
pRAGe	camemBERT	BARTHEZ	$0.65_{0.05}$	$0.07_{0.09}$	$0.05_{0.07}$	$0.12_{0.10}$	$0.65_{0.05}$	$0.11_{0.11}$	$0.05_{0.06}$	$0.12_{0.10}$
	DrBERT	BARTHEZ	$0.64_{0.03}$	$0.02_{0.06}$	$0.04_{0.06}$	$0.10_{0.09}$	$0.65_{0.04}$	$0.05_{0.07}$	$0.05_{0.06}$	$0.11_{0.09}$
	camemBERT	BIOMISTRAL	$\mathbf{0.69}_{0.06}$	$\mathbf{0.14}_{0.15}$	$\mathbf{0.12}_{0.14}$	$\mathbf{0.19}_{0.17}$	$\mathbf{0.68}_{0.06}$	$\mathbf{0.17}_{0.15}$	$\mathbf{0.08}_{0.09}$	$\mathbf{0.18}_{0.14}$
	DrBERT	BIOMISTRAL	$\mathbf{0.69}_{0.06}$	$\mathbf{0.14}_{0.15}$	$0.11_{0.12}$	$0.18_{0.17}$	$\mathbf{0.68}_{0.06}$	$\mathbf{0.17}_{0.15}$	$\mathbf{0.08}_{0.08}$	$0.17_{0.13}$
						w/ FINE TUNING				
SLM★	BARTHEZ		$0.62_{0.02}$	$0.05_{0.08}$	$0.06_{0.07}$	$0.11_{0.08}$	$0.63_{0.03}$	$0.09_{0.09}$	$0.07_{0.07}$	$0.12_{0.08}$
	BIOMISTRAL		$\mathbf{0.72}_{0.07}$	$\mathbf{0.15}_{0.17}$	$\mathbf{0.14}_{0.13}$	$\mathbf{0.22}_{0.17}$	$\mathbf{0.69}_{0.07}$	$\mathbf{0.16}_{0.16}$	$\mathbf{0.10}_{0.10}$	$\mathbf{0.18}_{0.13}$
pRAGe★	camemBERT	BARTHEZ	$\mathbf{0.65}_{0.05}$	$0.05_{0.09}$	$\mathbf{0.06}_{0.07}$	$0.12_{0.10}$	$\mathbf{0.64}_{0.05}$	$0.10_{0.10}$	$\mathbf{0.06}_{0.07}$	$\mathbf{0.12}_{0.10}$
	DrBERT	BARTHEZ	$0.64_{0.03}$	$0.01_{0.04}$	$\mathbf{0.06}_{0.07}$	$\mathbf{0.13}_{0.10}$	$\mathbf{0.64}_{0.04}$	$0.05_{0.07}$	$0.05_{0.06}$	$\mathbf{0.12}_{0.09}$
	camemBERT	BIOMISTRAL	$0.60_{0.04}$	$\mathbf{0.13}_{0.11}$	$0.03_{0.03}$	$0.09_{0.06}$	$0.60_{0.05}$	$\mathbf{0.16}_{0.11}$	$0.03_{0.03}$	$0.09_{0.06}$
	DrBERT	BIOMISTRAL	$0.59_{0.04}$	$0.12_{0.15}$	$0.03_{0.03}$	$0.08_{0.06}$	$0.60_{0.04}$	$0.14_{0.15}$	$0.03_{0.02}$	$0.08_{0.06}$

Table 3: Automatic Evaluation Metric Comparison of BaseSLMs with pRAGe setups on test set. Top scores for each model setups are shown in **bold**.

this method as it reflects real-life usage of language models by laypeople or patients. We test this method to answer (**RQ1**): *How good are open-source small language models and quantized models (1B-7B) at medical Q&A alone VS in a RAG system?* We are interested in analyzing if the LM's parametric knowledge (learned during the pre-training phase) is sufficient for generating accurate paraphrases, explanations or short definitions of medical terms. We compare these results with the settings with added non-parametric knowledge, the RefoMed-KB corpus. For this, we test a GPTQ quantized version of BioMistral, BioMistral-7B-SLERP-GPTQ4 (Labrak et al., 2024), and BARThez in an inference alone setting (Eddine et al., 2020). Furthermore, we test these two SLMs integrated in the pRAGe pipeline as decoders.

4.2 Vanilla Inference *VS* Finetuning

RAG systems are useful in the mitigation of hallucinations, as they give extra-knowledge to the LM. However, we want to test to what extent finetuning helps the LM generate a more accurate medical paraphrases in the pRAGe pipeline (added knowledge through RefoMed-KB) and inference alone (parametric knowledge only) (**RQ2**). We therefore test the two SLMs in two settings: non-fine-tuned (**NonFT**) and fine-tuned (**FT**) on the RefoMed paraphrase dataset.

4.3 Implementation Details

We tested two different lengths for the generated text: 25 and 50 tokens. For the inference setting, we used a simple prompt in French (Figure 2) for the Base SLM and a RAG adapted prompt in French for our pRAGe pipeline (Appendix A.1). We decided to use the prompts in French, as initial test experiments with English prompts generated French-English text.

> Expliquez-moi le terme médical
> en mots simples, par une paraphrase
> ou une courte définition :

Figure 2: Our prompt template in French for inference.

We present the results of our experiments and our fine-grained analysis in the results section below.

5 Results and Discussion

We present the summarized automatic evaluation in Table 3 and complete automatic evaluation table in Appendix A.2 (See Table 7). The automatic evaluation shows that BIOMISTRAL SLM reponses are more semantically and lexically related to gold paraphrase compared to BARTHEZ SLM. Further, both SLMs benefit from finetuning. Further, we notice that BIOMISTRAL pRAGe setups are obtained lower scores with finetuning. On the contrary, BARTHEZ pRAGe setups overall stay unaffected by fine tuning. This observation can be attributed to

the fact that in pRAGe setups models are restricted by the external knowledge base whereas in the case of SLM only setup, the models are free to generate anything and therefore, can be prone to hallucinations.

Next, we analyse the results observed from the manual evaluation, presented in Table 4. The human annotation of 50 examples in 24 different configurations (1200 samples)[10] shows that our fine-tuned version of BIOMISTRAL in inference alone and integrated in the pRAGe pipeline is the best model for **short answers** (**90%** strict correctness). Base BIOMISTRAL is the best model for **longer answers** in inference alone setting (**94%** strict correctness). The generated medical paraphrases and explanations should share correct medical knowledge, be informative and concise. This last trait is essential to our study, as we test the generation SLMs for a patient oriented downstream application. In this sense, we further analyzed the generated text according to different criteria:

Correctness of the medical knowledge. The best model, BIOMISTRAL, non fine-tuned, generated an English word in a French sentence, as seen in example [1]. After finetuning, the model generates a correct answer in French [2][11]. In the context of the pRAGe pipeline, BIOMISTRAL gives as well a full French answer [3].

- **non fine-tuned** - *[1] Asthme: maladie où les airways se ferment et se contractent, faisant du bruitlors de l'inspiration et de la respiration* (Asthma: a disease in which the airways close and contract, making noise during inspiration and breathing)

- **fine-tuned** - *[2] maladies respiratoires chroniques et maladies rares respiratoires* (chronic respiratory diseases and rare respiratory diseases)

- **pRAGe (CamemBERT)** - *[3] Asthme: maladie qui fait ressentir des difficultés à respirer, souvent accompagnée de toux et de sifflements.* (Asthma: a disease that makes breathing difficult, often accompanied by coughing and wheezing.)

---

[10]See Table 7 for all 24 configurations analyzed.

[11]The query contained a list of terms (*asthme , mucoviscidose , ventilation mécanique* - asthma , cystic fibrosis , mechanical ventilation), thus explaining the plural form of the generated text.

While the hallucination (Huang et al., 2023) percentage is low for BIOMISTRAL alone in zero-shot setting (6% for fine-tuned and 4% for vanilla, in the best settings[12]), BARThez is not adapted for zero-shot inference, as it is a summarization model. In most analyzed cases, it only summarizes the prompt, as the following example shows: *Expliquez-moi le terme médical en mots simples : phobie* (Explain the medical term in simple words: phobia). However, in pRAGe setting, BARThez is capable of summarizing the retrieved documents in a coherent answer. Nevertheless, its hallucination rate is still higher: 50% for both finetuned and vanilla. This can be explained by the fact that the model is not adapted for Q&A tasks, contrary to BIOMISTRAL. In pRAGe, BIOMISTRAL had a lower hallucination rate (10% in fine-tuned version and 8% in vanilla version).

However, we observed that BIOMISTRAL alone outperforms its pRAGe counterpart in both fine-tuning and zero-shot setting (+4% improvement). This results could be explained by the work of Mallen et al. (2023) on the impact of parametric an non-parametric knowledge of GPT language models (LMs) in vanilla and RAG setting depending on fact popularity. The authors' experiments showed that non-parametric memories are more effective for less popular facts (end-tail knowledge) than base LMs. However, they also proved that non-parametric memories can mislead LMs with less factual information. In our experiments, we hypothesize that the popularity of the medical knowledge to be paraphrased and the relevance of the information retrieved from RefoMed-KB can explain the lower performance of BIOMISTRAL in the pRAGe pipeline. Further qualitative analysis is needed to test this hypothesis for medical knowledge.

Short and concise answers. We analyzed if the generated medical paraphrase or explanation is accurate and concise in two settings: a constraint of 25 and 50 tokens of generated text. In a 50 token setting, the human analysis shows that the best model is BIOMISTRAL accuracy in correct answers in a relaxed and strict setting (96% ; 94%), while the second best is its fine-tuned version (94% ; 90%). However, the model is not as good in strict correctness in a 25 token setting (68%). Our fine-tuned version of BIOMISTRAL is better (**90%**, up to **22%** increase in performance) on strict correct-

---

[12]The best settings are the relaxed correctness values from Token=50, Table 4.

	w/o FINE TUNING					w/ FINE TUNING				
	readability (↓)	completeness%(↑)		correctness%(↑)		readability (↓)	completeness%(↑)		correctness%(↑)	
		STRICT	RELAX	STRICT	RELAX		STRICT	RELAX	STRICT	RELAX
SLMs BARTHEZ	1.22	**100**	**100**	0	0	<u>1.36</u>	0	0	0	0
	1.20	**100**	**100**	0	0	1.42	0	0	0	0
BIOMISTRAL	**1.08**	10	20	<u>68</u>	96	1.34	<u>16</u>	<u>20</u>	90	94
	1.10	18	96	**94**	96	1.5	**24**	**42**	90	94
pRAGe camemBERT BARTHEZ	1.22	56	64	42	46	1.22	14	14	38	42
	1.26	**96**	96	46	50	1.34	<u>70</u>	<u>76</u>	48	50
DrBERT BARTHEZ	**1**	18	68	0	0	**1.04**	60	60	0	0
	1.46	<u>94</u>	94	0	0	<u>1.08</u>	90	92	0	0
camemBERT BIOMISTRAL	1.10	27	33	<u>82</u>	88	1.40	33	48	<u>81</u>	90
	1.06	37	**100**	88	90	1.56	10	33	90	92
DrBERT BIOMISTRAL	<u>1.04</u>	14	24	46	84	1.20	34	38	74	88
	1.08	32	<u>98</u>	88	<u>88</u>	1.50	14	32	72	<u>88</u>

Table 4: Manual evaluation comparison of BaseSLMs with pRAGe models for subset of test set. The gray highlighted rows correspond to token=50 generation and rest of the rows correspond to token=25.Top scores for each model setups are shown in **bold** and second highest score is <u>underlined</u>.

ness and conciseness in a 25 tokens setting. In the **pRAGe** pipeline, CamemBERT `BIOMISTRAL`, both non fine-tuned and finetuned, gave better answers in terms of strict correctness for the 25 token setting (**82%**; **81%**).

Readability for laypeople. We analyzed how the `readability` score was influenced by our different setting experiments. In the short answer setting (25 token), the `readability` is better with Base `BIOMISTRAL` (**1.08**, lower is better). However, even if the readability is good, the answers are incomplete (10% `completeness-strict`). Our fine-tuned version of Base `BIOMISTRAL` improves the completeness of the answer (16% `completeness-strict`), while the best pRAGe model, CamemBERT `BIOMISTRAL` fine-tuned, increases it even further up to **33%**. Thus, we see a **+23%** improvement in performance (in `completeness-strict`) with our fine-tuned model in pRAGe.

One aspect that explains the decrease in `readability` in our fine-tuned models is the higher use of medical terms in the generated answers, as the fine-tuning step with the RefoMed dataset focuses on medical terms. As the goal of this study was to give short and concise paraphrases to a user query, we see that there are advantages of the fine-tuned model: it generates subsentential paraphrases, thus shorter and complete units of meaning (in the 25 tokens setting). Moreover, the fine-tuned model also generates simplified text generations, as observed in the following example where the medical term "osteophyte" is explained by using a subsentential paraphrase in very simple language: "deposits of bone tissue that form on the edges of bones" (original in French *ostéophyte -*

	Krippendorff's alpha(nominal)	% agreement
token=25		
w/o FINE TUNING		
completeness-STRICT	0.879	98%
completeness-RELAX	0.649	90%
readability	1.555	78%
w/ FINE TUNING		
completeness-STRICT	-0.076	84%
completeness-RELAX	-0.1	80%
readability	0.132	68%
token=50		
w/o FINE TUNING		
completeness-STRICT	0.66	98%
completeness-RELAX	0.003	64%
readability	0.105	70%
w/ FINE TUNING		
completeness-STRICT	0.105	70%
completeness-RELAX	1.151	64%
readability	-0.529	2%

Table 5: Inter-annotator agreement analysis

des dépôts de tissu osseux qui se forment sur les bords des os).

Inter-annotator agreement score. We computed a Krippendorff's alpha score (Krippendorff, 2018) for two criteria of the human evaluation: `completeness` and `readability`. The annotation was conducted by 3 French linguists annotators: 1 linguist completed a full annotation and 2 other linguists contributed to the second annotation (one annotated Token=25 and the other Token=50 length paraphrases). We show the Krippendorff's alpha score and the percentage agreement in Table 5. The inter-annotator agreement is highest for `completeness-strict`, for both lengths (98% agreement), showing syntactic analysis is an easy task for the annotators. However, regarding `readability`, it is more difficult for the two annotators to agree (78% to 68% agreement), meaning
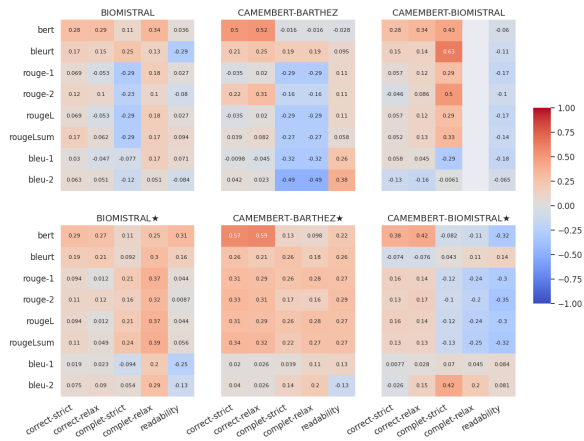
Figure 3: Correlation Heatmap between Automatic evaluation metrics (y-axis) and Manual evaluation metrics (x-axis). The ★ symbol denotes configurations with finetuned SLM.

that the medical knowledge of the annotators can influence the readability level of annotations.

Automatic evaluation VS Manual evaluation. In Figure 3 (which merges values from Table 3 and Table 4), we show the correlation heatmap between automatic metric scores and manual metrics scores for *best* SLM and pRAGe configurations. The top row corresponds to configuration without finetuning and the bottom row corresponds to configurations with finetuned SLM backbone. `rouge` metrics positively corrected with `correctness` for the generation with finetuning which is intuitive as finetuning allows the pRAGe setups to generate in a more similar style as the gold reference (RefoMed). `bert` and `bleurt` scores remain unaffected with finetuning as the semantic similarity of generation can saturate unless lexical similarity increases. Finally, we notice the `readability` aspect of SLMs is differently affected from pRAGe setup as it is not as trivial as lexical closeness.

6 Conclusion and Future Work

We presented pRAGe, a pipeline for retrieval, generation and evaluation of medical paraphrases in a patient/lay person oriented downstream application. We showed that finetuning `BIOMISTRAL` with the RefoMed dataset increases the ability of the model to generate short, concise and correct subsentential paraphrases. The pRAGe pipeline helps increase the scientific grounding of text generation via LLMs for medical domain. Our work intends to help bridge the gap between scientific medical knowledge and lay people.

Further work will include testing another French language models, such as Vigogne or Claire in similar settings, finetuned and in pRAGe. We are also planning to conduct an extensive annotation campaign with specialists of the medical domain to asses a fair inter-agreement score on `correctness` and further explore how the generation is affected by the choice of knowledge base used.

Ethics Statement

Acknowledgments

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alham Fikri Aji, Tirana Noor Fatyanosa, Radityo Eko Prasojo, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2022. Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. *arXiv preprint arXiv:2205.04651*.

Edwide Antoine and Natalia Grabar. 2016. Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert. In *TALN 2016: Traitement Automatique des Langues Naturelles*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Houda Bouamor, Auréelien Max, and Anne Vilnat. 2013. Multitechnique paraphrase alignment: A contribution to pinpointing sub-sentential paraphrases. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):1–27.

Ioana Buhnila. 2022. Le rôle des marqueurs et indicateurs dans l'analyse lexicale et sémantico-pragmatique de reformulations médicales. In *SHS Web of Conferences*, volume 138, page 10005. EDP Sciences.

Ioana Buhnila. 2023. *Une méthode automatique de construction de corpus de reformulation*. Ph.D. thesis, Université de Strasbourg.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Josette Condamines, Anne ; Rebeyrolle. 1997. Point de vue en langue spécialisée. *Meta*, 42(1):174–184.

Roger A Cote. 1998. Systematized nomenclature of human and veterinary medicine: Snomed international. version 3.5. *Northfield, IL: College of American Pathologists*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *LREC*, pages 4276–4283. Citeseer.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, et al. 2024. Medical mt5: An open-source multilingual text-to-text llm for the medical domain. *arXiv preprint arXiv:2404.07613*.

Natalia Grabar and Rémi Cardon. 2018. Clear-simple corpus for medical french. In *ATA*.

Natalia Grabar and Thierry Hamon. 2015. Extraction automatique de paraphrases grand public pour les termes médicaux. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 182–195.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. Parade: A new dataset for paraphrase identification requiring computer science domain knowledge. *arXiv preprint arXiv:2010.03725*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2023. Chatgpt makes

medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, pages 1–9.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *arXiv preprint arXiv:2401.15269*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Thomas W LeBlanc, Ashley Hesson, Andrew Williams, Chris Feudtner, Margaret Holmes-Rovner, Lillie D Williamson, and Peter A Ubel. 2014. Patient understanding of medical jargon: a survey study of us medical students. *Patient education and counseling*, 95(2):238–242.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain. Association for Computational Linguistics, 2004*, page 74–81.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language mode. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Aurélien Max, Houda Bouamor, and Anne Vilnat. 2012. Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 721–731.

OpenAI. 2022. Openai: Introducing chatgpt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *https://arxiv.org/abs/1908.10084*.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Seyedeh Belin Tavakoly Sany, Fatemeh Behzhad, Gordon Ferns, and Nooshin Peyman. 2020. Communication skills training for physicians improves health literacy and medical outcomes among patients with hypertension: a randomized controlled trial. *BMC health services research*, 20:1–10.

Amalia Todirascu, Sebastian Padó, Jennifer Krisch, Max Kisselew, and Ulrich Heid. 2012. French and german corpora for audience-based text type classification. In *LREC*, volume 2012, pages 1591–1597.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Martin Tschirsich and Gerold Hintz. 2013. Leveraging crowdsourcing for paraphrase recognition. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 205–213.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

Yuming Zhai, Lufei Liu, Xinyi Zhong, Gbariel Illouz, and Anne Vilnat. 2020. Building an english-chinese parallel corpus annotated with sub-sentential translation techniques. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4024–4033.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

> Vous êtes un expert en médecine. Utilisez les informations suivantes pour répondre à la question de l'utilisateur par une paraphrase, une explication ou une courte définition.
> Si vous ne connaissez pas la réponse, dites simplement que vous ne savez pas, n'essayez pas d'inventer une réponse.
> **Contexte**: context
> **Question**: question
> Ne renvoyez que la réponse utile. La réponse doit être claire, concise et facile à comprendre pour le grand public.
> **Réponse utile** :

Table 6: Initial prompt for base SLM experiments.

Setup	TOKEN	-bleu	-bert	-bleurt	-rouge1	-rouge2	-rougL	-rougeLsum	-bleu-p1	-bleu-p2	-bleu-p3	-bleu-p4
							w/o FINE TUNING					
BARTHEZ	25	$0.00_{0.00}$	$0.63_{0.03}$	$0.10_{0.10}$	$0.07_{0.08}$	$0.01_{0.03}$	$0.07_{0.08}$	$0.06_{0.07}$	$0.04_{0.06}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
	50	$0.00_{0.00}$	$0.63_{0.03}$	$0.10_{0.10}$	$0.07_{0.08}$	$0.01_{0.03}$	$0.07_{0.08}$	$0.06_{0.07}$	$0.04_{0.06}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
BIOMISTRAL	25	$0.00_{0.02}$	$0.70_{0.06}$	$0.15_{0.15}$	$0.20_{0.16}$	$0.07_{0.12}$	$0.20_{0.16}$	$0.17_{0.14}$	$0.11_{0.12}$	$0.03_{0.08}$	$0.01_{0.06}$	$0.01_{0.05}$
	50	$0.00_{0.03}$	$0.68_{0.06}$	$0.16_{0.15}$	$0.18_{0.13}$	$0.06_{0.09}$	$0.18_{0.13}$	$0.14_{0.11}$	$0.08_{0.08}$	$0.02_{0.05}$	$0.01_{0.03}$	$0.00_{0.03}$
CAMEMBERT+BARTHEZ	25	$0.00_{0.00}$	$0.65_{0.05}$	$0.07_{0.09}$	$0.12_{0.10}$	$0.02_{0.05}$	$0.12_{0.10}$	$0.10_{0.08}$	$0.05_{0.07}$	$0.00_{0.02}$	$0.00_{0.01}$	$0.00_{0.00}$
	50	$0.00_{0.00}$	$0.65_{0.05}$	$0.11_{0.11}$	$0.12_{0.10}$	$0.02_{0.05}$	$0.12_{0.10}$	$0.10_{0.08}$	$0.05_{0.06}$	$0.00_{0.02}$	$0.00_{0.01}$	$0.00_{0.00}$
DRBERT+BARTHEZ	25	$0.00_{0.00}$	$0.64_{0.03}$	$0.02_{0.06}$	$0.10_{0.09}$	$0.00_{0.02}$	$0.10_{0.09}$	$0.08_{0.06}$	$0.04_{0.06}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
	50	$0.00_{0.00}$	$0.65_{0.04}$	$0.05_{0.07}$	$0.11_{0.09}$	$0.01_{0.02}$	$0.11_{0.09}$	$0.09_{0.07}$	$0.05_{0.06}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
CAMEMBERT+BIOMISTRAL	25	$0.01_{0.06}$	$0.69_{0.06}$	$0.14_{0.15}$	$0.19_{0.17}$	$0.08_{0.14}$	$0.19_{0.17}$	$0.17_{0.15}$	$0.12_{0.14}$	$0.04_{0.12}$	$0.02_{0.11}$	$0.02_{0.10}$
	50	$0.00_{0.03}$	$0.68_{0.06}$	$0.17_{0.15}$	$0.18_{0.14}$	$0.06_{0.10}$	$0.18_{0.14}$	$0.15_{0.12}$	$0.08_{0.09}$	$0.02_{0.05}$	$0.01_{0.04}$	$0.01_{0.04}$
DRBERT+BIOMISTRAL	25	$0.00_{0.02}$	$0.69_{0.06}$	$0.14_{0.15}$	$0.18_{0.17}$	$0.07_{0.13}$	$0.18_{0.17}$	$0.16_{0.16}$	$0.11_{0.12}$	$0.03_{0.08}$	$0.02_{0.06}$	$0.01_{0.05}$
	50	$0.00_{0.02}$	$0.68_{0.06}$	$0.17_{0.15}$	$0.17_{0.13}$	$0.05_{0.09}$	$0.17_{0.13}$	$0.14_{0.12}$	$0.08_{0.08}$	$0.02_{0.05}$	$0.01_{0.04}$	$0.00_{0.03}$
							w/ FINE TUNING					
BARTHEZ★	25	$0.00_{0.00}$	$0.62_{0.02}$	$0.05_{0.08}$	$0.11_{0.08}$	$0.01_{0.02}$	$0.11_{0.08}$	$0.09_{0.06}$	$0.06_{0.07}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
	50	$0.00_{0.01}$	$0.63_{0.03}$	$0.09_{0.09}$	$0.12_{0.08}$	$0.01_{0.04}$	$0.12_{0.08}$	$0.09_{0.06}$	$0.07_{0.07}$	$0.01_{0.02}$	$0.00_{0.02}$	$0.00_{0.01}$
BIOMISTRAL★	25	$0.00_{0.00}$	$0.72_{0.07}$	$0.15_{0.17}$	$0.22_{0.17}$	$0.09_{0.13}$	$0.22_{0.17}$	$0.20_{0.16}$	$0.14_{0.13}$	$0.04_{0.07}$	$0.01_{0.04}$	$0.00_{0.02}$
	50	$0.00_{0.00}$	$0.69_{0.07}$	$0.16_{0.16}$	$0.18_{0.13}$	$0.07_{0.10}$	$0.18_{0.13}$	$0.16_{0.12}$	$0.10_{0.10}$	$0.02_{0.04}$	$0.01_{0.02}$	$0.00_{0.01}$
CAMEMBERT+BARTHEZ★	25	$0.00_{0.00}$	$0.65_{0.05}$	$0.05_{0.09}$	$0.12_{0.10}$	$0.02_{0.05}$	$0.12_{0.10}$	$0.10_{0.08}$	$0.06_{0.07}$	$0.01_{0.02}$	$0.00_{0.01}$	$0.00_{0.00}$
	50	$0.00_{0.01}$	$0.64_{0.05}$	$0.10_{0.10}$	$0.12_{0.10}$	$0.02_{0.05}$	$0.12_{0.10}$	$0.10_{0.08}$	$0.06_{0.07}$	$0.01_{0.02}$	$0.00_{0.01}$	$0.00_{0.01}$
DRBERT+BARTHEZ★	25	$0.00_{0.00}$	$0.64_{0.03}$	$0.01_{0.04}$	$0.13_{0.10}$	$0.01_{0.03}$	$0.13_{0.10}$	$0.10_{0.07}$	$0.06_{0.07}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
	50	$0.00_{0.00}$	$0.64_{0.04}$	$0.05_{0.07}$	$0.12_{0.09}$	$0.01_{0.03}$	$0.12_{0.09}$	$0.09_{0.06}$	$0.05_{0.06}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
CAMEMBERT+BIOMISTRAL★	25	$0.00_{0.00}$	$0.60_{0.04}$	$0.13_{0.11}$	$0.09_{0.06}$	$0.02_{0.03}$	$0.09_{0.06}$	$0.07_{0.05}$	$0.03_{0.03}$	$0.01_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
	50	$0.00_{0.00}$	$0.60_{0.05}$	$0.16_{0.11}$	$0.09_{0.06}$	$0.02_{0.03}$	$0.09_{0.06}$	$0.07_{0.05}$	$0.03_{0.03}$	$0.01_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
DRBERT+BIOMISTRAL★	25	$0.00_{0.00}$	$0.59_{0.04}$	$0.12_{0.15}$	$0.08_{0.06}$	$0.02_{0.02}$	$0.08_{0.06}$	$0.07_{0.04}$	$0.03_{0.03}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$
	50	$0.00_{0.00}$	$0.60_{0.04}$	$0.14_{0.15}$	$0.08_{0.06}$	$0.02_{0.02}$	$0.08_{0.06}$	$0.07_{0.04}$	$0.03_{0.02}$	$0.00_{0.01}$	$0.00_{0.00}$	$0.00_{0.00}$

Table 7: Detailed Experiment report on <u>test set</u> following equation 1 for each metric. The ★ symbol denotes configurations with finetuned SLM.

A Appendix

A.1 Sample prompt for pRAGe

We use the following prompt (See Table 6) to start our experiment and then tune it.

A.2 Complete Automatic Evaluation

We provide in Table 7 detail results for all the configuration we consider during our experiments.

A.3 Manual Evaluation Examples

We provide in Table 8, 9 samples from BIOMISTRAL and BARTHEZ SLMs and pRAGe configurations and their manual evaluation annotation which are used to create the final table (See Table 4).

	TOKEN		Base SLM : BIOMISTRAL	
			NON-FINETUNED	FINETUNED
SLM	25		**PROMPT + ostéophyte**	
		fr	Un ostéophyte est une formation osseuse qui pousse à partir d'une articulation ou d	des dépôts de tissu osseux qui se forment sur les bords des os[1
		en	(An osteophyte is a bony formation that grows from a joint or)	(deposits of bone tissue that form on the edges of bones[1])
		[TAG]	Cr-R:1; Cr-S:1; Cm-R:0; Cm-S:0; Rd:2.	Cr-R:1; Cr-S:1; Cm-R:1. Cm-S:0; Rd:1.
			PROMPT + asthme, mucoviscidose, ventilation mécanique	
		fr	Asthme: maladie où les airways se ferment et se contractent, faisant du bruit lors	maladies respiratoires chroniques et maladies rares respiratoires (M)
		en	(Asthma: a disease in which the airways close and contract, making noise as)	(chronic respiratory diseases and rare respiratory diseases (M)
		[TAG]	Cr-R:1; Cr-S:0; Cm-R:0; Cm-S:0; Rd:1.	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1.
	50		**PROMPT + ostéophyte**	
		fr	Un ostéophyte est une formation osseuse qui pousse à partir d'une articulation ou d'une jointure. Il peut être formé en réponse à une blessure ou à une inflammation, mais	des dépôts de tissu osseux qui se forment sur les bords des os [1] [2] [3] [4] [5] [6] [7]
		en	(An osteophyte is a bone formation that grows from a joint or articulation. It may be formed in response to injury or inflammation, but)	(deposits of bone tissue that form on the edges of bones [1] [2] [3] [4] [5] [6] [7])
		[TAG]	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:2.	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1.
			PROMPT + asthme, mucoviscidose, ventilation mécanique	
		fr	Asthme: maladie où les airways se ferment et se contractent, faisant du bruit lors de l'inspiration et de la respiration. Mucoviscidose: maladie où	maladies respiratoires chroniques et maladies rares respiratoires (MRR) (MRR) (MRR) (MRR) (M
		en	(Asthma: a disease in which the airways close and contract, making noise when breathing in and out. Cystic fibrosis: a disease in which)	(chronic respiratory diseases and rare respiratory diseases (MRR) (MRR) (MRR) (MRR) (MRR) (MRR) (M
		[TAG]	Cr-R:1; Cr-S:0; Cm-R:1; Cm-S:0; Rd:1.	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1.
			pRAGe : BIOMISTRAL	
C.BERT	25		**PROMPT + ostéophyte**	
		fr	Un ostéophyte est une formation osseuse qui pousse autour d'une articulation. Il	est une formation osseuse qui se développe sur les os entourant une articulation.
		en	(An osteophyte is a bone formation that grows around a joint. It)	(is a bone formation that develops on the bones surrounding a joint.)
		[TAG]	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:2.	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:1; Rd:1.
			PROMPT + asthme, mucoviscidose, ventilation mécanique	
		fr	Asthme: maladie qui fait ressentir des difficultés à respirer, souvent accompagnée de	est une maladie génétique qui affecte les glandes salivaires et les glandes lacry
		en	(Asthma: a disease that makes it difficult to breathe, often accompanied b)	(is a genetic disease that affects the salivary and lacrimal glan)
		[TAG]	Cr-R:1; Cr-S:1; Cm-R:0; Cm-S:0; Rd:1.	Cr-R:1; Cr-S:0; Cm-R:0; Cm-S:0; Rd:1.
	50		**PROMPT + ostéophyte**	
		fr	Un ostéophyte est une formation osseuse qui pousse autour d'une articulation. Il peut être causé par une maladie appelée arthrose, qui fait que de la matière osseuse	est une formation osseuse qui se développe sur les os entourant une articulation. Les ostéophytes peuvent être symptomatiques ou asymptomatiques. Lors
		en	(An osteophyte is a bone formation that grows around a joint. It can be caused by a disease called arthritis, which causes bone material to)	(is a bone formation that develops on the bones surrounding a joint. Osteophytes may be symptomatic or asymptomatic. At)
		[TAG]	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:2.	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1.
			PROMPT + asthme, mucoviscidose, ventilation mécanique	
		fr	Asthme: maladie qui fait ressentir des difficultés à respirer, souvent accompagnée de toux et de sifflements. Mucoviscidose: maladie génétique qui affect	est une maladie génétique qui affecte les glandes salivaires et les glandes lacrymales, provoquant une production excessive de mucus. Cette maladie peut également affecter les voies
		en	(Asthma: a disease that makes breathing difficult, often accompanied by coughing and wheezing. Cystic fibrosis: genetic disease that affects)	(is a genetic disorder that affects the salivary and lacrimal glands, causing excessive mucus production. This disease can also affect the [tracks])
		[TAG]	Cr-R:1; Cr-S:1; Cm-R:0; Cm-S:0; Rd:1.	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1.

Table 8: Examples of generated answers from the different BioMistral configurations. The human annotation tags stand for our evaluation metrics: correctness-relaxed (Cr-R), correctness-strict (Cr-S), completeness-relaxed (Cm-R), completeness-strict (Cm-S), and readability (Rd), and *C.BERT* for CamemBERT sentence-embedding. PROMPT refers to the prompt syntax in Fig.2

	TOKEN		pRAGe : BARTHEZ	
			NON-FINETUNED	FINETUNED
C.BERT	25		**PROMPT + ostéophyte**	
		fr	L'arthrose est une affection dégénérative des articulations qui fait que de la matière osseuse est anormalement produite	L'ostéophytose désigne le phénomène d'apparition d'un ostéophyte sur une articulation, lors
		en	(Osteoarthritis is a degenerative joint disorder in which bone material is abnormally produced.)	(Osteophytosis refers to the phenomenon of the appearance of an osteophyte on a joint, when)
		[TAG]	Cr-R:0; Cr-S:0; Cm-R:1; Cm-S:0; Rd:1.	Cr-R:1; Cr-S:1; Cm-R:0; Cm-S:0; Rd:2.
			PROMPT + asthme, mucoviscidose, ventilation mécanique	
		fr	Ces descriptions se focalisent le plus souvent sur les problèmes digestifs, stéatorrhée et troubles pancréatiques,	Notes et références Voir aussi Articles connexes BradypnéeDyspnéePolypnéeT
		en	(These descriptions most often focus on digestive problems, steatorrhea and pancreatic disorders,)	(Notes and references See also Related articles BradypneaDyspneaPolypneaT)
		[TAG]	Cr-R:0; Cr-S:0; Cm-R:0; Cm-S:0; Rd:2.	Cr-R:0; Cr-S:0; Cm-R:0; Cm-S:0; Rd:3.
	50		**PROMPT + ostéophyte**	
		fr	L'arthrose est une affection dégénérative des articulations qui fait que de la matière osseuse est anormalement produite à l'extérieur de l'os.	L'ostéophytose est une affection dégénérative des articulations qui fait que de la matière osseuse est anormalement produite à l'extérieur de l'os, au détriment de la membrane blanchâtre fibreuse qui entoure l'os
		en	(Osteoarthritis is a degenerative joint disease in which abnormal bone material is produced outside the bone.)	(Osteophytosis is a degenerative joint disorder in which bone material is abnormally produced on the outside of the bone, to the detriment of the whitish fibrous membrane that surrounds the bone.
		[TAG]	Cr-R:1; Cr-S:0; Cm-R:1; Cm-S:1; Rd:1.	Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:1; Rd:2.
			PROMPT + asthme, mucoviscidose, ventilation mécanique	
		fr	Ces descriptions se focalisent le plus souvent sur les problèmes digestifs, stéatorrhée et troubles pancréatiques.	Notes et références Voir aussi Articles connexes BradypnéeDyspnéePolypnéeTachypnéeSyndrome d'apnées du sommeilApnée (sport)
		en	(These descriptions most often focus on digestive problems, steatorrhea and pancreatic disorders.)	(Notes and references See also Related articles BradypneaDyspneaPolypneaTachypneaSleep apnea syndromeApnea (sport)))
		[TAG]	Cr-R:0; Cr-S:0; Cm-R:1; Cm-S:1; Rd:2.	Cr-R:0; Cr-S:0; Cm-R:0; Cm-S:0; Rd:3.

Table 9: Examples of generated answers from BARTHEZ in the CamemBERT pRAGe setup. The human annotation tags stand for our evaluation metrics: correctness-relaxed (Cr-R), correctness-strict (Cr-S), completeness-relaxed (Cm-R), completeness-strict (Cm-S), and readability (Rd), and *C.BERT* for CamemBERT sentence-embedding.PROMPT refers to the prompt syntax in Fig.2