# Is Biomedical Specialization Still Worth It? Insights from Domain-Adaptive Language Modelling with a New French Health Corpus

**Aidan Mannion**[1], **Cécile Macaire**[1], **Armand Violle**[2], **Stéphane Ohayon**[2], **Xavier Tannier**[2], **Didier Schwab**[1], **Lorraine Goeuriot**[1], **François Portet**[1]

[1] Université Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

[2] Sorbonne Université, LIMICS, 15 rue de l'École de Médecine, 75006 Paris, France

**Abstract**

Large language models (LLMs) have demonstrated remarkable capabilities across diverse domains, yet their adaptation to specialized fields remains challenging, particularly for non-English languages. This study investigates domain-adaptive pre-training (DAPT) as a strategy for specializing small to mid-sized LLMs in the French biomedical domain through continued pre-training. We address two key research questions: the viability of specialized continued pre-training for domain adaptation and the relationship between domain-specific performance gains and general capability degradation. Our contributions include the release of a fully open-licensed French biomedical corpus suitable for commercial and open-source applications, the training and release of specialized French biomedical LLMs, and novel insights for DAPT implementation. Our methodology encompasses the collection and refinement of high-quality French biomedical texts, the exploration of causal language modeling approaches using DAPT, and conducting extensive comparative evaluations. Our results cast doubt on the efficacy of DAPT, in contrast to previous works, but we highlight its viability in smaller-scale, resource-constrained scenarios under the right conditions. Our findings further suggest that model merging post-DAPT is essential to mitigate generalization trade-offs, and in some cases even improves performance on specialized tasks at which the DAPT was directed.

**Keywords:** Domain-adaptive pre-training, Biomedical NLP

## 1. Introduction

LLMs are widely recognized as *foundation models* that demonstrate promising general capabilities, often exhibiting emergent reasoning abilities with appropriate prompting (Bommasani et al., 2021). However, achieving high performance and clinical reliability in specialized areas requires thoughtful adaptation. Domain-Adaptive Pre-training (DAPT, Gururangan et al., 2020), also referred to as Continual Pre-Training (CPT, Chen et al., 2025), addresses this by conducting a second phase of pre-training on large, unlabeled, domain-specific text to align the model with the distributional characteristics of text in the target field. This approach aims to capture useful patterns, such as complex medical terminology, that may be inadequately represented in the initial, broad general-purpose training corpus.

The Domain-Adaptive Pre-Training presented in this work is carried out as part of the of the R&D phase of the PARTAGES project, which aims to develop specialized language models for use in the automation of document-processing tasks in the French healthcare system, while releasing the associated resources (models, code, datasets) as freely-available open-source tools.

In this context, we present a new collection of French biomedical corpora that is guaranteed to be fully compatible with all downstream applications from a licensing standpoint, called `PARCOMED` (**PAR**TAGES **C**orpus of **O**pen **ME**dical **D**ocuments). Alongside the corpus, we release a collection of domain-specialized models trained thereon, using Qwen3 (Yang et al., 2025) as a foundation, and reflect on the utility of this kind of continual pre-training as an efficacious strategy going forward.

## 2. Related Work

The application of LLMs to medicine has resulted in several high-profile models, predominantly in English, trained via proprietary or open-source DAPT methodologies, often relying on massive datasets of biomedical literature. Google's Med-PaLM (Singhal et al., 2023), for example, built upon a 540-billion parameter foundation model, achieved state-of-the-art results on medical question-answering benchmarks by combining scaling with prompt tuning strategies. Open-source alternatives have also emerged, focusing on scalability and accessibility, such as BioMedLM (Bolton et al., 2024; 2.7B parameters), BioGPT (Luo et al., 2022; 355M), and MedAlpaca (Han et al., 2023; 7B & 13B). Another significant open-source contribution is MEDITRON (Chen et al., 2023), which scaled medical CPT to 70B parameters using Llama-2 as a backbone, training on a corpus that included PubMed abstracts, full-text papers, and high-quality clinical guidelines. Similarly, BioMistral-7B (Labrak et al., 2024) leveraged the Mistral-7B-Instruct model, supplementing its training with the PubMed Central Open Access



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10617–10633
11–16 May 2026. 

Subset to specialize it for the biomedical domain. These foundational English models, along with related encoder-only models like BioBERT (Lee et al., 2020), established that CPT has the potential to enhance medical-specific language modelling capabilities in certain scenarios.

Despite the reported gains, the necessity of DAPT for highly capable, general-purpose LLMs has been challenged. Recent head-to-head comparisons, using rigorous evaluation protocols that involve optimizing prompts for each model independently and measuring statistical significance, found that most biomedical LLMs failed to consistently improve over their general-domain base models in zero- or few-shot QA tasks (Jeong et al., 2024).

For domains outside of English, such as the French biomedical context, the challenges are magnified by the scarcity of specialized resources. Multilingual generalization remains limited, as performance typically degrades when models are tested on automatically translated benchmarks, as shown by Labrak et al. (2024), who also highlighted that additional pre-training on English medical data has limited benefits for non-English contexts. Addressing the French medical domain specifically, researchers have introduced specialized resources for CPT like the NACHOS corpus (Labrak et al., 2023) and the automatically-translated TransCorpus-bio-fr (Knafou et al., 2025), recognizing that data scarcity is a major hurdle in releasing open-source specialized LLMs in French. A promising direction for more comprehensive evaluations of these strategies is the systematic testing of CPT, SFT (Supervised Fine-Tuning), and combined CPT+SFT approaches, such as the work by Belmadani et al. (2025) on the Mistral-7B architecture.

## 3. The PARCOMED Corpus

### 3.1. Context

The availability of French biomedical data remains a major challenge for improving the multilingual capabilities of large language models (LLMs) in the medical domain.

We introduce and release the PARCOMED corpus, a comprehensive collection of French biomedical texts compiled from a wide range of sources. Although collections of French medical documents, such as NACHOS (Labrak et al., 2023) or Jargon (Segonne et al., 2024) have already been distributed to the community recently, our corpus collection is the result of a greater scrutiny of the licensing term of each source. Thus, in contrast to the collections mentioned above, the PARCOMED corpus is fully compatible with research usage and is also distributed with a version compatible with commercial usage.

The selected datasets for our corpus come from a variety of sources which can be categorized as follows (for readability, citations are provided in Table 1):

- Open-access archives (HAL, HAS, ISTEX, ANSES, QUALISCOPE, CERIMES, CNEDIMTS, ECDC TM).
- Healthcare data such as clinical cases from the E3C, CAS (real, anonymized cases), and FRASIMED (synthetic) corpora, as well as clinical trial protocols (ESSAI).
- Information leaflets for medications (BDPM, EMEA V3).
- Datasets available in literature designed for specific NLP tasks such as machine translation (WMT16, WMT18 Medline), named-entity recognition (QUAERO, DEFT2021, CLEAR, MANTRA GSC), multiple-choice QA (FrenchMedMCQA) and doctor-patient dialogues (MQC, PXCORPUS).
- General knowledge on health and medicine extracted via API requests[1] to French Wikipedia for medecine, pharmacy and biology categories.

### 3.2. Data collection

As mentioned previously, our sources cover diverse biomedical content, including scientific articles, drug leaflets, medical device evaluations, regulatory documents, clinical case reports, and institutional recommendations. In each case, all partitions (train/dev/test) of the datasets were included. We provide two distinct versions of the aggregated dataset, summarized in Table 1: a commercial-use corpus, containing only sources whose licenses permit commercial use, and a research-only corpus, allowing only non-commercial applications. As it can be seen, the corpus is dominated by scientific documents (around 94% of words).

### 3.3. Text cleaning and volume

All documents were preprocessed using a the pipeline inspired by Le et al. (2020), including Unicode conversion and normalization, removal of characters outside standard French encoding, suppression of multiple spaces, and deletion of URLs. The dataset is organized at the document level, where each entry corresponds to a single document (e.g., a Wikipedia page). In total, 906,489 documents were collected from various sources

[1] https://wikipedia-api.readthedocs.io/en/latest/

| Source name | Document type | Commercial | # docs | # words | Reference |
|---|---|---|---|---|---|
| HAL | Scientific | Yes | 26,987 | 703,473,770 | CNRS (2001) |
| HAS | Scientific | Yes | 11,334 | 96,173,390 | Haute Autorité de Santé (HAS) (2021c) |
| ISTEX | Scientific | Yes | 12,179 | 43,138,368 | CNRS (2012) |
| BDPM | Medication | Yes | 11,026 | 20,035,903 | BDPM (2013) |
| WIKIPEDIA | Encyclopedic | Yes | 9,957 | 6,531,021 | Wikipedia Contributors (2025) |
| WMT16 | Scientific | Yes | 587,075 | 6,490,287 | Bojar et al. (2016) |
| EMEA V3 | Medication | Yes | 222,971 | 4,449,136 | Tiedemann (2012) |
| CERIMES | Education | Yes | 22 | 1,715,189 | CERIMES (2003) |
| FRASIMED | Clinical | Yes | 2,048 | 1,322,895 | Zaghir et al. (2024) |
| DEFT2021 | Question Answering | Yes | 271 | 110,641 | Grouin et al. (2021) |
| QUAERO | Scientific | Yes | 2,490 | 71,812 | Névéol et al. (2014) |
| FrenchMedMCQA | Question Answering | Yes | 1,144 | 58,872 | Labrak et al. (2022) |
| CNEDIMTS | Regulation | Yes | 813 | 58,345 | Haute Autorité de Santé (HAS) (2021b) |
| ECDC TM | Other medical | Yes | 2,160 | 42,491 | Steinberger et al. (2014) |
| PXCORPUS | Medication | Yes | 1,414 | 18,372 | Kocabiyikoglu et al. (2022) |
| QUALISCOPE | Regulation | Yes | 298 | 11,736 | Haute Autorité de Santé (HAS) (2021a) |
| MANTRA GSC | Scientific | Yes | 150 | 3,596 | Kors et al. (2015) |
| **Total commercial** | | | **892,343** | **883,706,984** | |
| E3C | Clinical | No | 7,499 | 15,864,637 | Minard et al. (2021) |
| CAS | Clinical | No | 716 | 233,371 | Grabar et al. (2018) |
| CLEAR | Scientific | No | 6 | 226,123 | Grabar and Cardon (2018) |
| ESSAI | Clinical | No | 5,842 | 146,537 | Dalloux et al. (2021) |
| MQC | Dialogue | No | 38 | 15,672 | Laleye et al. (2020) |
| WMT18 Medline | Scientific | No | 49 | 7,719 | Neves et al. (2018) |
| **Total research** | | | **906,489** | **900,199,883** | |

Table 1: Data sources for the PARCOMED corpus.

(see Table 1); the corpus used to train the models was the 892K-document version allowing commercial use.

## 4. Domain-Adaptive Continual Pre-Training for Medical Applications in French

The experimental methodology discussed in this paper proceeds in three main steps: model selection, DAPT, and merging. Firstly, we run a range of baseline evaluations and selected the best-performing generalist foundation models for DAPT (the evaluation protocol is presented in Section 5). We then run Causal Language Modelling on these models, executing the evaluation benchmark at regular intervals. Based on the progression of the averaged evaluation metrics, we select a checkpoint to focus on in the final results. Finally, using Spherical Linear Interpolation (SLERP, Goddard et al. (2024)), we combine the weights of this checkpoint with the base model, in order to investigate the resulting trade-offs in evaluation results. Evaluation results for the selected checkpoint and its corresponding SLERP merge are presented in Section 6.

**Model Selection** The generalist foundation models used in these experiments are from the Qwen3 family.

Having implemented the evaluation of a range of decoder language models on a broad bilingual multi-domain question-answering benchmark, of which a subset is presented in Section 6, the 8B model stood out as the best-performing base LLM, surpassing not only direct competitors from the Llama and Mistral families, but also domain-specific models such as Apollo-7B (Zheng et al., 2024), and BioMistral (Labrak et al., 2024). “Best-performing” in this context refers to the model ranking on a

| Task Group | Topic | Abbreviation | Source | Eval. Metric | # Questions |
|---|---|---|---|---|---|
| (EN/FR)-MEDICAL | Anatomy | Anat. | MMLU | Accuracy | 135 |
| | Clinical Knowledge | C.K. | MMLU | Accuracy | 265 |
| | College Biology | CBio. | MMLU | Accuracy | 144 |
| | College Medicine | CMed. | MMLU | Accuracy | 173 |
| | Health | n/a | MMLU-Pro-X | Exact match | 687 |
| | Medical Genetics | MGen. | MMLU | Accuracy | 100 |
| | Professional Medicine | ProMed. | MMLU | Accuracy | 272 |
| (EN/FR)-OTHER | Business | Bus. | MMLU-Pro-X | Exact match | 789 |
| | Computer Science | CS | MMLU-Pro-X | Exact match | 410 |
| | Economics | Econ. | MMLU-Pro-X | Exact match | 844 |
| | History | Hist. | MMLU-Pro-X | Exact match | 381 |
| | Law | n/a | MMLU-Pro-X | Exact match | 959 |
| | Philosophy | Phil. | MMLU-Pro-X | Exact match | 499 |
| | Psychology | Psych. | MMLU-Pro-X | Exact match | 798 |

Table 2: Groupings, abbreviations, metrics and number of questions for each the QA datasets used for evaluation.

selection of biomedical tasks in French (our target domain), for which more complete results can be found in Appendix A. To investigate the effect of model size on DAPT in this context, we also carry out all of our experiments using three other Qwen3 models, the 0.6B, 1.7B, and 4B variants.

We restrict our attention in this work to the "-Base" variants, which have not undergone instruction tuning. This choice was made in order to more reliably isolate the effects of unsupervised training. In addition, we aim to further fine-tune our domain-specialized models on medical document-processing use cases for which the conversational "chatbot-like" behaviour inculcated by instruction tuning is not necessarily desirable.

### 4.1. Continual Pretraining Setup: PDAPT

After tokenizing the PARCOMED commercial corpus and chunking it into sequences of 2,048 tokens (longer documents were split with an overlap stride of 4 tokens), we continue the pre-training of the four Qwen3 base models for a total of 4,320 update steps. The tokenized corpus contains over 1.95B tokens[2] from a word count of uner 1B, pointing to the large amount of specialized domain-specific terminology contained therein.

This training was carried out with a constant learning rate of $2 \times 10^{-5}$ with no warmup, and an effective batch size (taking into account gradient accumulation and data parallelism) of 1,152 sequences. The full training run thus corresponds to 2.53 epochs over the corpus. The progressive effect of DAPT on downstream task performance is investigated by checkpointing the training state every 720 steps (see Figure 1). Training runs were executed on 48 NVIDIA H100 GPUs on the Jean Zay computing cluster, using BF16 precision and the Fully Sharded Data Parallel framework from PyTorch.

We abbreviate this continual pretraining process as PDAPT (PARCOMED DAPT).

## 5. Evaluation Protocol

The evaluation methodology presented in this paper relies on a set of standardized LLM evaluation benchmarks in both English and French. The specific aims of this evaluation framework are firstly to evaluate whether or not specializing LLMs from the general domain improves their performance on biomedical tasks, and secondly to compare PDAPT model performance on general-purpose benchmarks with their corresponding base models to identify potential degradation due to over-specialization.

The evaluation is based around the open-source framework "lm-evaluation-harness"[3] (Gao et al., 2024) for few-shot language model assessment, which ensures full reproducibility through open and publicly available datasets. In order to measure the trade-off between specialization and generalization brought about by the DAPT strategy outlined in Section 4, we define four task groups: one in the target domain (medicine) and the target language (French), one in the target domain in a different language (English) and two more that constitute a collection of other specialized domains outside of medicine, in both languages. Each group contains seven tasks, laid out in Table 2. The evaluation datasets themselves are drawn from two sources:

- The MMLU multiple-choice question-answering dataset (Hendrycks et al., 2021),

[2] 1,955,165,272

[3] https://github.com/EleutherAI/lm-evaluation-harness

| Subject → | Anat. | C.K. | CBio. | CMed. | Health | MGen. | ProMed. |
|---|---|---|---|---|---|---|---|
| Qwen3-0.6B-Base | 30.4±4.0 | 49.8±3.1 | 39.6±4.1 | 42.8±3.8 | 15.9±1.4 | 48.0±5.0 | 37.1±2.9 |
| +PDAPT | **39.3±4.2** | **50.2±3.1** | **41.0±4.1** | **49.1±3.8** | **16.9±1.4** | **48.0±5.0** | **44.5±3.0** |
| +SLERP | **43.7±4.3** | 49.8±3.1 | **43.8±4.1** | 42.8±3.8 | **18.0±1.5** | 44.0±5.0 | **37.9±2.9** |
| Qwen3-1.7B-Base | 48.1±4.3 | 59.2±3.0 | 54.2±4.2 | 59.5±3.7 | 25.9±1.7 | 66.0±4.8 | 56.2±3.0 |
| +PDAPT | **58.5±4.3** | **59.2±3.0** | **60.4±4.1** | 57.8±3.8 | 24.3±1.6 | 62.0±4.9 | **57.7±3.0** |
| +SLERP | **51.9±4.3** | **61.1±3.0** | **60.4±4.1** | **61.8±3.7** | **26.2±1.7** | **67.0±4.7** | **59.6±3.0** |
| Qwen3-4B-Base | 54.8±4.3 | 70.9±2.8 | 75.0±3.6 | 68.2±3.6 | 39.3±1.9 | 74.0±4.4 | 69.5±2.8 |
| +PDAPT | **58.5±4.3** | 70.6±2.8 | **77.8±3.5** | **71.1±3.5** | 37.0±1.8 | **81.0±3.9** | **71.7±2.7** |
| +SLERP | **57.0±4.3** | **74.0±2.7** | **78.5±3.4** | **71.1±3.5** | **40.6±1.9** | **79.0±4.1** | **72.4±2.7** |
| Qwen3-8B-Base | 62.2±4.2 | 74.7±2.7 | 87.5±2.8 | 75.7±3.3 | 50.2±1.9 | 80.0±4.0 | 76.5±2.6 |
| +PDAPT | 61.5±4.2 | **76.2±2.6** | 86.8±2.8 | **76.9±3.2** | 45.9±1.9 | 80.0±4.0 | 76.1±2.6 |
| +SLERP | 60.0±4.2 | **77.4±2.6** | 86.8±2.8 | **76.3±3.2** | 49.8±1.9 | 79.0±4.1 | 75.7±2.6 |

Table 3: Comparative accuracy scores for the task group FR-MEDICAL.

| Subject → | Anat. | C.K. | CBio. | CMed. | Health | MGen. | ProMed. |
|---|---|---|---|---|---|---|---|
| Qwen3-0.6B-Base | 47.4±4.3 | 57.0±3.0 | 59.7±4.1 | 52.6±3.8 | 22.4±1.6 | 62.0±4.9 | 55.5±3.0 |
| +PDAPT | 40.7±4.2 | 51.3±3.1 | 59.0±4.1 | 51.4±3.8 | 18.0±1.5 | 52.0±5.0 | 50.7±3.0 |
| +SLERP | 41.5±4.3 | 54.0±3.1 | 59.7±4.1 | **53.2±3.8** | 22.0±1.6 | 59.0±4.9 | 53.7±3.0 |
| Qwen3-1.7B-Base | 59.3±4.2 | 67.9±2.9 | 72.9±3.7 | 68.2±3.6 | 34.6±1.8 | 73.0±4.5 | 64.7±2.9 |
| +PDAPT | 57.8±4.3 | **68.7±2.9** | **74.3±3.7** | 65.9±3.6 | 30.3±1.8 | 69.0±4.6 | 59.9±3.0 |
| +SLERP | 59.3±4.2 | 67.9±2.9 | **73.6±3.7** | 67.1±3.6 | **35.7±1.8** | 71.0±4.6 | 63.6±2.9 |
| Qwen3-4B-Base | 68.1±4.0 | 80.4±2.4 | 84.7±3.0 | 74.0±3.3 | 49.9±1.9 | 81.0±3.9 | 78.3±2.5 |
| +PDAPT | 64.4±4.1 | **80.8±2.4** | **86.1±2.9** | **74.6±3.3** | 46.3±1.9 | **83.0±3.8** | 75.7±2.6 |
| +SLERP | **69.6±4.0** | 80.4±2.4 | **86.1±2.9** | **75.7±3.3** | 48.9±1.9 | **83.0±3.8** | **79.4±2.5** |
| Qwen3-8B-Base | 74.1±3.8 | 80.0±2.5 | 88.9±2.6 | 78.0±3.2 | 55.8±1.9 | 86.0±3.5 | 83.5±2.3 |
| +PDAPT | 70.4±3.9 | 78.9±2.5 | 84.7±3.0 | 76.9±3.2 | 55.0±1.9 | 85.0±3.6 | 80.9±2.4 |
| +SLERP | 72.6±3.9 | **81.5±2.4** | 88.9±2.6 | 76.9±3.2 | **57.1±1.9** | 86.0±3.5 | 81.6±2.4 |

Table 4: Comparative results for the task group EN-MEDICAL.

from which we draw a selection of medical-domain tasks; for French-language evaluation, we reuse the translated versions from Labrak et al. (2024).

- The MMLU-Pro-X dataset (Xuan et al., 2025), a diverse multilingual benchmark built to evaluate the reasoning capacities of LLMs.

We reuse the standard task configuration and metrics for these tasks, as integrated in lm-evaluation-harness: for the medical MMLU tasks, we use few-shot prompting with $n = 3$ and use accuracy as the evaluation metric, while for MMLU-Pro-X, we use $n = 5$ and the *exact-match* metric. The first of these metrics, referred to as "Accuracy" in Table 2, considers a model's answer to be the the string with the highest conditional log probability from a fixed set of possible answer strings. The exact-match metric, on the other hand, only considers the overall highest-probability string to be the answer. In both cases, the aggregate metric corresponds to the percentage of model answers that match the ground-truth label. Each of these metrics is accompanied by a confidence interval based on a bootstrapped standard error measurement implemented via the evaluation harness; as can be seen in Section 6's tables, the smaller dataset sizes for the medical-specific tasks result in wider intervals in general.

As a summary statistic for the general performance tendencies at the level of our four task groups, we calculate an average of these metrics weighted by the number of documents in each dataset. This metric is referred to simply as the "weighted average score" in Section 6.

## 6. Results and Analysis

Figure 1 displays the progression of the weighted average score over the PDAPT training process for each of the four members of the Qwen3 family considered. As the MMLU-Pro-X datasets that make up the "OTHER" task groupings have more difficult questions in larger quantities (they were specifically designed to be more challenging than MMLU), and employ a more demanding evaluation metric (exact-match accuracy), the averages

| Subject → | Bus. | CS | Econ. | Hist. | Law | Phil. | Psych. |
|---|---|---|---|---|---|---|---|
| Qwen3-0.6B-Base | 19.1±1.4 | 19.5±2.0 | 23.5±1.5 | 13.4±1.7 | 7.3±0.8 | 17.0±1.7 | 26.2±1.6 |
| +PDAPT | 15.8±1.3 | 8.8±1.4 | 16.1±1.3 | **14.2±1.8** | **8.4±0.9** | 16.2±1.7 | 18.4±1.4 |
| +SLERP | **19.4±1.4** | 12.7±1.6 | 21.4±1.4 | **15.0±1.8** | 7.2±0.8 | 15.8±1.6 | **26.7±1.6** |
| Qwen3-1.7B-Base | 35.6±1.7 | 27.1±2.2 | 37.2±1.7 | 18.9±2.0 | 8.2±0.9 | 23.0±1.9 | 38.8±1.7 |
| +PDAPT | 26.2±1.6 | 18.8±1.9 | 33.3±1.6 | 17.1±1.9 | 7.5±0.9 | 20.0±1.8 | 32.5±1.7 |
| +SLERP | 33.0±1.7 | **27.6±2.2** | 35.5±1.6 | 17.6±2.0 | **9.1±0.9** | 22.2±1.9 | 37.1±1.7 |
| Qwen3-4B-Base | 50.8±1.8 | 46.1±2.5 | 56.8±1.7 | 34.1±2.4 | 18.6±1.3 | 32.1±2.1 | 54.9±1.8 |
| +PDAPT | 44.4±1.8 | 39.8±2.4 | 53.1±1.7 | 30.4±2.4 | 15.6±1.2 | **35.1±2.1** | 51.6±1.8 |
| +SLERP | **52.7±1.8** | 44.9±2.5 | 56.0±1.7 | **35.2±2.4** | 18.1±1.2 | **33.5±2.1** | 54.5±1.8 |
| Qwen3-8B-Base | 61.5±1.7 | 50.5±2.5 | 62.7±1.7 | 40.2±2.5 | 23.5±1.4 | 4.9±2.2 | 60.5±1.7 |
| +PDAPT | 55.4±1.8 | 45.6±2.5 | 61.1±1.7 | 39.4±2.5 | 21.1±1.3 | 41.7±2.2 | 59.3±1.7 |
| +SLERP | 58.2±1.8 | **53.7±2.5** | **63.6±1.7** | **41.5±2.5** | **23.8±1.4** | **45.5±2.2** | **61.2±1.7** |

Table 5: Comparative exact-match scores for the task group FR-OTHER.

| Subject → | Bus. | CS | Econ. | Hist. | Law | Phil. | Psych. |
|---|---|---|---|---|---|---|---|
| Qwen3-0.6B-Base | 28.8±1.6 | 26.1±2.2 | 31.5±1.6 | 16.0±1.9 | 11.2±1.0 | 19.4±1.8 | 36.6±1.7 |
| +PDAPT | 18.1±1.4 | 20.5±2.0 | 25.7±1.5 | 15.5±1.9 | 10.3±1.0 | 15.8±1.6 | 27.7±1.6 |
| +SLERP | 27.0±1.6 | 24.9±2.1 | 30.7±1.6 | **17.1±1.9** | 10.4±1.0 | 19.0±1.8 | 34.5±1.7 |
| Qwen3-1.7B-Base | 37.1±1.7 | 39.0±2.4 | 46.0±1.7 | 27.6±2.3 | 14.6±1.1 | 34.1±2.1 | 47.0±1.8 |
| +PDAPT | 33.2±1.7 | 34.4±2.3 | 43.8±1.7 | 25.5±2.2 | **14.8±1.1** | 29.1±2.0 | 42.5±1.8 |
| +SLERP | 42.6±1.8 | 39.0±2.4 | 45.4±1.7 | 25.5±2.2 | 14.2±1.1 | 32.7±2.1 | 45.9±1.8 |
| Qwen3-4B-Base | 57.3±1.8 | 53.2±2.5 | 63.5±1.7 | 38.6±2.5 | 25.2±1.4 | 42.1±2.2 | 61.8±1.7 |
| +PDAPT | 52.1±1.8 | 48.0±2.5 | 62.1±1.7 | **39.6±2.5** | 19.2±1.3 | 40.9±2.2 | 59.1±1.7 |
| +SLERP | 56.3±1.8 | 51.0±2.5 | **65.0±1.6** | **39.4±2.5** | 22.7±1.4 | **43.3±2.2** | 61.3±1.7 |
| Qwen3-8B-Base | 62.5±1.7 | 60.2±2.4 | 68.2±1.6 | 48.3±2.6 | 29.6±1.5 | 49.3±2.2 | 67.2±1.7 |
| +PDAPT | 60.7±1.7 | 59.8±2.4 | 68.2±1.6 | **49.9±2.6** | 27.7±1.4 | 47.1±2.2 | 66.0±1.7 |
| +SLERP | **62.6±1.7** | 58.3±2.4 | 67.4±1.6 | **49.9±2.6** | **29.7±1.5** | **50.7±2.2** | 66.2±1.7 |

Table 6: Comparative results for the task group EN-OTHER.

are significantly lower than for the medical-domain tasks.

We can see from these charts that the overall impact of PDAPT is minimal, with changes in the average becoming less pronounced as model size increases. As would be expected, performance on the non-medical tasks decreases the more the models are exposed to the PARCOMED corpus, although this is not necessarily accompanied by increases in medical-domain performance, and many of the averages trend back downward in the latter part of the training. The only aspect of this that stands out as a potential avenue for improvement is the slight increase in the FR-MEDICAL average early in training for the smallest model, Qwen3-0.6B-Base. Indeed, on further inspection, the 1440-step checkpoint gives us the greatest number of per-task improvements across all models. It is thus these checkpoints for which the SLERP merging was carried out, and for which the task-by-task results are presented.

**Baseline Evaluations** For the task group that represents the target domain for the work in this project, FR-MEDICAL, we present a range of accuracy results for open-source LLMs in Appendix A. These results provide a baseline reference for the performance metrics presented in Table 3, by showing the metrics for both generalist and specialist models, with and without supervised training. As they were beyond the range of parameter counts being considered for continual pre-training in this project, the 14B and 32B Qwen3 variants and GPT-oss-20B models are included for reference only.

**Domain Adaptation Experiments** Side-by-side comparisons of the Qwen3 base models and their domain-adapted counterparts are presented for each task group as follows: Tables 3 and 4 show results for the medical domain and Tables 5 and 6 for the other specializations, for French and English respectively. We highlight in **bold** results where the specialized models improved on the performance of the base model. Green cells denote statistically significant increases (i.e. non-overlapping standard-error confidence intervals) and red cells statistically significant declines.

We thus make 56 comparisons per task domain: for FR-MEDICAL, there were 8 statistically significant changes, of which only 1 was negative. For the

other groups, the picture is somewhat more bleak - there was only one significant positive change (the model Qwen3-8B-Base+PDAPT+SLERP on the MMLU-Pro-X Business dataset in English). However, it is worth noting that these declines apply to the PDAPT models only: once model merging is carried out, there are no longer any statistically significant decreases in performance for any of the base models across any of the task groups, while the performance on our actual target domain (FR-MEDICAL) remains elevated, particularly for the smaller models.

These results are summarized in the chart shown in Figure 2 - as observed in Figure 1, we can see that there is little significant change in performance at the aggregate level - the per-task improvements in the FR-MEDICAL group appear to be cancelled out by concomitant losses in accuracy when averaged. This suggests that DAPT is better approached in an even more specialized manner, at the level of medical subjects, and motivates further exploration into more granular experiments within the medical NLP domain.

## 7. Conclusion

This work introduces the PARCOMED corpus, the first French biomedical corpus collection with full licensing compatibility for all downstream applications, addressing a gap in openly-available domain-specific resources. Accompanying this corpus, we release the Qwen3-PDAPT collection, a series of decoder-only language models based on the Qwen3 pre-trained foundation models, which we hope will aid the research community in systematically assessing the capabilities of smaller language models in French biomedical contexts. Through extensive experimentation and analysis encompassing multiple domains, we offer actionable insights and practical recommendations for researchers and practitioners working to adapt language models for healthcare applications in languages beyond English.

All datasets, models, training code and evaluation framework configurations used in this work, along with more extensive fully reproducible benchmarking experiments, will be made freely available online.

As generalist language models continue to improve in quality and breadth, the marginal benefits of domain-adaptive pre-training are likely to diminish even further. Nevertheless, given the more pronounced improvements we observed with smaller-scale decoder models, we advocate for continued investigation of DAPT in resource-constrained environments, where energy efficiency considerations are paramount; this is a particularly salient concern in healthcare settings that face stringent limitations

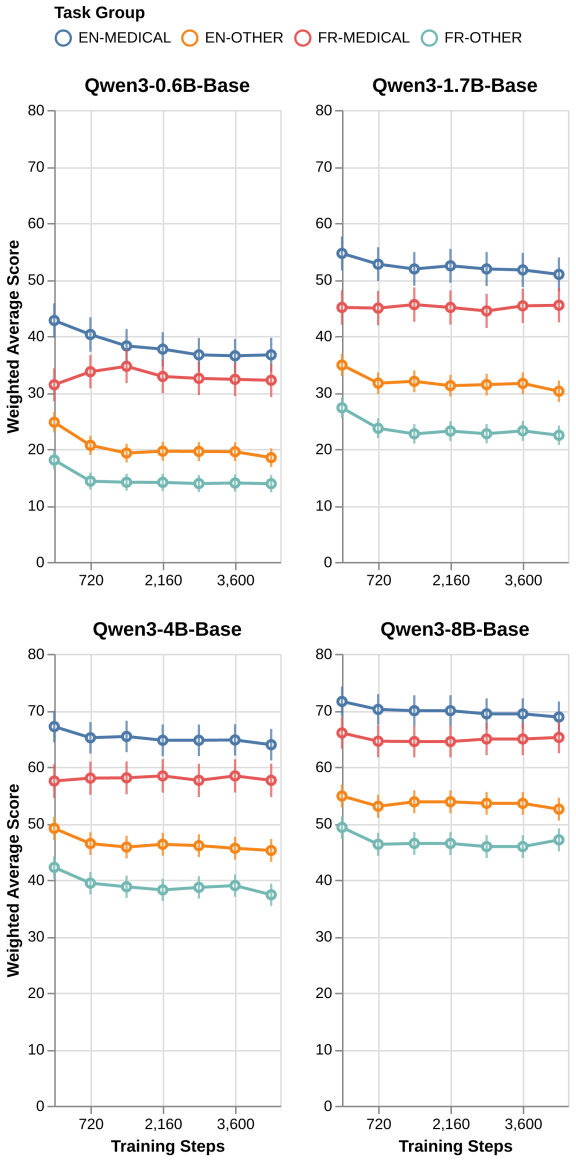


Figure 1: Progression of evaluation scores on the four task groups.

on computational resources and restrictions on utilizing external model providers. Furthermore, highly specialized pre-training targeting narrow biomedical subdomains may yield more substantial performance gains than broad domain adaptation.

Finally, our results demonstrate that merging domain-adapted models with their generalist base counterparts is not merely an optional enhancement but a fundamental requirement for maintaining balanced capabilities across both specialized and general language tasks.

Code for the experiments carried out in this work can be found at https://github.com/PARTAGES-dev/partages-llm; the PARCOMED corpus is available at https://huggingface.co/HealthDataHub.

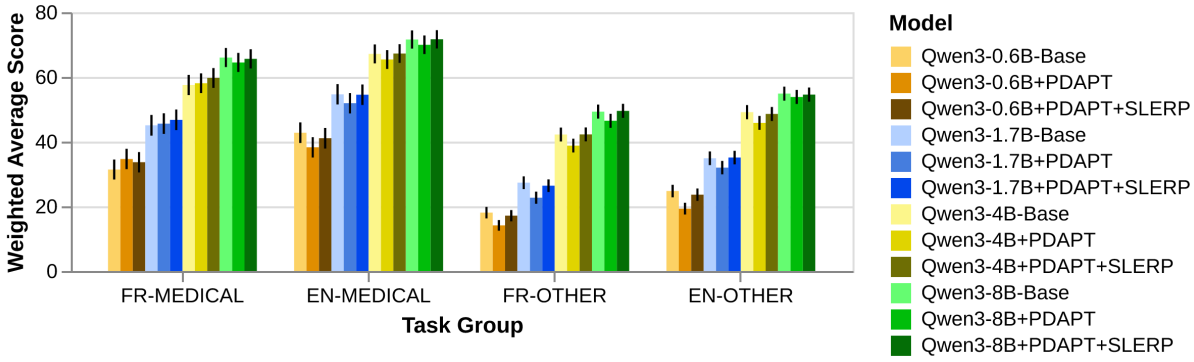


Figure 2: Comparison of group-level averages for base vs. specialized models, along with the SLERP merge between them. Black bars represent combined confidence intervals.

### 7.1. Limitations

Our investigation focuses exclusively on causal language modeling as the pre-training objective and employs only the Qwen3 model family, without exploring supervised fine-tuning approaches that might complement domain adaptation. The absence of publicly available technical specifications for Qwen3 constrained our ability to select optimal hyperparameters and training configurations for continued pre-training. Moreover, our evaluation benchmark, while comprehensive, does not encompass the full breadth of medical subdomains and specialist topics necessary to thoroughly characterize the performance trade-offs inherent in domain-specific adaptation.

Although the SLERP model merging method was chosen for these experiments due to the fact that we are merging models with very similar loss trajectories, other merging techniques, notably DARE and/or TIES (Yadav et al., 2023), represent promising alternatives to this method and have recently showed to leads to improved merged general and specialized models in specialized domains (Ueda et al., 2026).

### 7.2. Future Work

As suggested in Section 6, advancements of this work will involve the investigation of more targeted specialization by partitioning our pre-training corpora according to biomedical subtopics and applying selective continued pre-training to specific thematic subsets. Exploring the effects of instruction fine-tuning on our domain-adapted models represents another crucial direction, as this technique may help reconcile specialized domain knowledge with general conversational capabilities. Finally, we intend to expand our evaluation methodology beyond academic question-answering tasks to include practical document-processing and automation workflows encountered in real-world healthcare operations, thereby better assessing the utility of our models for applied clinical and administrative applications.

## 8. Acknowledgements

This work was carried out as part of the PARTAGES project, winner of the Bpifrance France 2030 call for proposals “Digital Commons for Generative Artificial Intelligence”. It was also partially supported by the French National Research Agency (ANR) through the MIAI “AI & Language” chair (ANR-23-IACL-0006). This work was performed using HPC resources from GENCI at IDRIS under allocation 2025-A0181016171 on the Jean Zay supercomputer.

## 9. References

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, and Vinko Sabolčec. 2025. Apertus: Democratizing open and compliant llms for global language environments.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025.

SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

Ikram Belmadani, Benoit Favre, Richard Dufour, Frédéric Bechet, and Carlos Ramisch. 2025. Adaptation des connaissances médicales pour les grands modèles de langue : Stratégies et analyse comparative. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 50–72, Marseille, France. Association pour le Traitement Automatique des Langues.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, and Li Fei-Fei. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv*.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Xin Zhao, Zhicheng Dou, Jiaxin Mao, Yankai Lin, Ruihua Song, Jun Xu, Xu Chen, Rui Yan, Zhewei Wei, Di Hu, Wenbing Huang, and Ji-Rong Wen. 2025. Towards effective and efficient continual pretraining of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5779–5795, Vienna, Austria. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. _eprint: 2311.16079.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Nathan Godey, Wissam Antoun, Rian Touchent, Rachel Bawden, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2025. Gaperon: A peppered english-french generative language model suite.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, and Arthur Hinsvark. 2024. The llama 3 herd of models.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. _eprint: 2304.08247.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane

Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and Gaël Liu. 2025. Gemma 3 technical report.

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. _eprint: 2402.10373.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbe, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *LREC*.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. Ho So, and J. Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 2020*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, 23(6).

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, and Samuel Schmidgall. 2025. Medgemma technical report.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Kentaro Ueda, François Portet, Hirohiko Suwa, and Keiichi Yasumoto. 2026. Merging continual pretraining models for domain-specialized llms: A case study in finance. In *Proceedings of LREC 2026*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pages 7093–7115. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2024. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts.

## 10. Language Resource References

BDPM. 2013. Base de Données Publique des Médicaments (BDPM). https://base-donnees-publique.medicaments.gouv.fr/.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics.

CERIMES. 2003. Centre de ressources et d'information sur les multimédias pour

l'enseignement supérieur (CERIMES). http://www.cerimes.fr.

CNRS. 2001. HAL (archive ouverte). https://hal.science.

CNRS. 2012. Infrastructure de services pour la fouille de texte (ISTEX). https://www.istex.fr.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in french and brazilian portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Natalia Grabar and Rémi Cardon. 2018. CLEAR-Simple Corpus for Medical French. In *ATA*, Tilburg, Netherlands.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. Cas: French corpus with clinical cases. In *LOUHI 2018-The Ninth International Workshop on Health Text Mining and Information Analysis*, pages 1–7.

Cyril Grouin, Natalia Grabar, and Gabriel Illouz. 2021. Classification de cas cliniques et évaluation automatique de réponses d'étudiants: présentation de la campagne deft 2021 (clinical cases classification and automatic evaluation of student answers: Presentation of the deft 2021 challenge). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 1–13.

Haute Autorité de Santé (HAS). 2021a. Base sur la qualité et la sécurité des soins (QualiScope). https://www.data.gouv.fr/datasets/base-sur-la-qualite-et-la-securite-des-soins-anciennement-scope-sante/informations.

Haute Autorité de Santé (HAS). 2021b. Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé (CNEDiMTS). https://www.data.gouv.fr/datasets/evaluation-des-dispositifs-medicaux.

Haute Autorité de Santé (HAS). 2021c. Haute Autorité de Santé. https://www.has-sante.fr.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Julien Knafou, Luc Mottin, Anaïs Mottaz, Alexandre Flament, and Patrick Ruch. 2025. TransBERT: A Framework for Synthetic Translation in Domain-Specific Language Modeling. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. In *LREC 2022*.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickaël Rouvier. 2022. Frenchmedmcqa: A french multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46.

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. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper*, Toronto, Canada. Association for Computational Linguistics.

Fréjus AA Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 2020. A french medical conversations corpus annotated for a virtual patient dialogue system. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 574–580.

Anne-Lyse Minard, Roberto Zanoli, Begoña Altuna, Manuela Speranza, Bernardo Magnini, and Alberto Lavelli. 2021. European Clinical Case Corpus (E3C-Corpus). Dataset. Version 2.0.0. Licensed under CC BY-NC 4.0.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus: A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, pages 24–30.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

Vincent Segonne, Aidan Mannion, Laura Cristina Alonzo Canul, Alexandre Audibert, Xingyu Liu, Cécile Macaire, Adrien Pupier, Yongxin Zhou, Mathilde Aguiar, Felix Herron, Magali Norré, Massih-Reza Amini, Pierrette Bouillon, Iris Eshkol-Taravella, Emmanuelle Esperança-Rodier, Thomas François, Lorraine Goeuriot, Jérôme Goulian, Mathieu Lafourcade, Benjamin Lecouteux, François Portet, Fabien Ringeval, Vincent Vandeghinste, Maximin Coavoux, Marco Dinarelli, and Didier Schwab. 2024. Jargon: A Suite of Language Models and Evaluation Tasks for French Specialized Domains. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9463–9476, Turin, Italy.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. *Language resources and evaluation*, 48(4):679–707.

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Wikipedia Contributors. 2025. Wikipédia, l'encyclopédie libre – Médecine, Pharmacie, Biologie. https://fr.wikipedia.org.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. 2025. Mmluprox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

Jamil Zaghir, Mina Bjelogrlic, Jean-Philippe Goldman, Soukaïna Aananou, Christophe Gaudet-Blavignac, and Christian Lovis. 2024. Frasimed: A clinical french annotated resource produced through crosslingual bert-based annotation projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7450–7460.

# A. Appendix: Benchmarking Results

| Model | Average Ranking |
|---|---|
| Qwen3-32B | 1.57 |
| Qwen3-14B | 2.43 |
| MedGemma-27B-it | 2.71 |
| Qwen3-8B-Base | 3.71 |
| Qwen3-8B | 4.86 |
| Qwen3-4B | 6.86 |
| Qwen3-4B-Base | 7.43 |
| Apertus-8B-Instruct | 11.14 |
| Apertus-8B | 11.71 |
| Llama3.1-8B | 12.57 |
| GPT-OSS-20B | 13.00 |
| Apollo-7B | 13.57 |
| Llama3.1-8B-Instruct | 13.71 |
| EuroLLM-9B-Instruct | 13.86 |
| MedGemma-4B-it | 15.14 |
| Qwen3-1.7B-Base | 15.43 |
| EuroLLM-9B | 16.43 |
| Mistral-7B-Instruct-v0.3 | 18.00 |
| SmolLM-3B | 18.29 |
| Mistral-7B-v0.3 | 18.43 |
| Llama3.2-1B | 20.14 |
| Olmo3-7B | 21.43 |
| Qwen3-1.7B | 21.71 |
| Gaperon-8B | 22.57 |
| BioMistral-7B | 24.14 |
| MedGemma-4B-pt | 25.14 |
| Qwen3-0.6B-Base | 26.71 |
| Qwen3-0.6B | 27.00 |
| Llama3.2-1B-Instruct | 27.57 |
| Gemma3-1B-it | 29.00 |
| Gemma3-1B-pt | 32.00 |
| Gaperon-1B | 32.29 |
| Gemma3-270m-it | 32.43 |
| Baguettotron | 33.14 |
| Gemma3-270m-pt | 33.86 |

Table 7: Average ranking of all 35 models across all seven French-language tasks.

This section lays out the full results of the comparative benchmark used for model selection, i.e. the FR-MEDICAL task group. Tables 8-11 detail the evaluation metrics, alongside the corresponding EN-MEDICAL scores for reference. The top FR-MEDICAL result is highlighted in green and B text denotes all scores whose confidence intervals overlap with it.

In addition to the Qwen3 models presented in Section 4, our evaluations include models from Google's Gemma3 (Kamath et al., 2025) and MedGemma (Sellergren et al., 2025) families, Meta's LLama3 family (Grattafiori et al., 2024), and Mistral's 7B models (Jiang et al., 2023). As mentioned in Section 4, we include specialised biomedical models Apollo-7B (Zheng et al., 2024), and BioMistral (Labrak et al., 2024). Alongside these models that are trained at private companies and research labs and are open-source only in the sense that their weights are freely available online, we include fully open models (base and instruction-tuned) EuroLLM-9B (Martins et al., 2025), Apertus-8B (Apertus et al., 2025), Gaperon-8B (Godey et al., 2025), and SmolLM-3B (Bakouch et al., 2025).

In addition, we include results from GPT-OSS-20B (OpenAI, 2025), although on inspection of its outputs it appears that additional safety guardrail training inhibits the propensity of this particular model to give explicit answers to many medical questions, meaning that it does not surpass even 7-9B models despite its size.

Table 7 shows the average rank of each model across the seven tasks.

The benchmark results underline the centrality of model size as a determining factor in performance, unsurprising for knowledge-base tasks such as these, with the top three models being the three largest in terms of parameter count. However, it also underlines the aforementioned ability of the Qwen3 models to punch above their weight, with Qwen3-8B-Base coming within a statistically insignificant margin of these larger models in four out of seven tasks, and the Qwen3-4B models outperforming all of the 7-9B models apart from its own larger version.

Given resource and operational constraints in the PARTAGES project, the largest models considered for continual pre-training on `PARCOMED` were in the 7-9B range.

As we use an evaluation setup that directly accesses the log-likelihood distributions output by the decoders, we can see that instruction tuning is not necessarily helpful in this benchmark, although variations in the amount and type of supervised fine-tuning involved in building the instruction-tuned version from the base version also seem to play a role. As noted previously, the Qwen3 models' "-Base" versions rank higher than their instruction-tuned counterparts (apart from the case of the 4B model) while the instruction-tuned Gemma3 and MedGemma instruction-tuned models (denoted by the "-it" suffix) all significantly improve on the performance of the corresponding base models ("-pt" suffix).

| Type | Model | Accuracy (%)↑ | | | |
|---|---|---|---|---|---|
| | | Anatomy | | Clinical Knowledge | |
| | | EN | FR | EN | FR |
| Base | Gemma3-270m-pt | 16.30±3.2 | 16.30±3.2 | 28.68±2.8 | 23.77±2.6 |
| | Baguettotron | 20.00±3.5 | 26.67±3.8 | 24.53±2.6 | 24.53±2.6 |
| | Qwen3-0.6B-Base | 47.41±4.3 | 29.63±3.9 | 56.98±3.0 | 49.81±3.1 |
| | Llama3.2-1B | 40.74±4.2 | 26.67±3.8 | 30.57±2.8 | 27.55±2.8 |
| | Gemma3-1B-pt | 32.59±4.0 | 28.89±3.9 | 21.89±2.5 | 23.77±2.6 |
| | Gaperon-1B | 17.04±3.2 | 19.26±3.4 | 26.79±2.7 | 32.08±2.9 |
| | Qwen3-1.7B-Base | 59.26±4.2 | 48.15±4.3 | 67.92±2.9 | 59.25±3.0 |
| | Qwen3-4B-Base | 68.15±4.0 | 54.81±4.3 | 80.38±2.4 | 70.94±2.8 |
| | Olmo3-7B | 57.78±4.3 | 44.44±4.3 | 69.43±2.8 | 55.85±3.1 |
| | Mistral-7B-v0.3 | 61.48±4.2 | 55.56±4.3 | 65.66±2.9 | 59.25±3.0 |
| | Qwen3-8B-Base | 74.07±3.8 | **62.22±4.2** | 80.00±2.5 | 74.72±2.7 |
| | Llama3.1-8B | 60.74±4.2 | 57.78±4.3 | 72.08±2.8 | 62.26±3.0 |
| | Apertus-8B | 60.00±4.2 | 54.81±4.3 | 73.21±2.7 | 66.42±2.9 |
| | Gaperon-8B | 54.81±4.3 | 40.00±4.2 | 55.47±3.1 | 54.72±3.1 |
| | EuroLLM-9B | 57.04±4.3 | 56.30±4.3 | 61.51±3.0 | 59.25±3.0 |
| Instruct | Gemma3-270m-it | 26.67±3.8 | 22.96±3.6 | 25.28±2.7 | 28.30±2.8 |
| | Qwen3-0.6B | 48.15±4.3 | 45.19±4.3 | 50.19±3.1 | 44.53±3.1 |
| | Llama3.2-1B-Instruct | 51.85±4.3 | 44.44±4.3 | 47.17±3.1 | 38.87±3.0 |
| | Gemma3-1B-it | 41.48±4.3 | 34.07±4.1 | 43.77±3.1 | 43.77±3.1 |
| | Qwen3-1.7B | 57.04±4.3 | 40.74±4.2 | 63.02±3.0 | 55.09±3.1 |
| | SmolLM-3B | 54.81±4.3 | 45.93±4.3 | 67.55±2.9 | 60.75±3.0 |
| | Qwen3-4B | 63.70±4.2 | 57.78±4.3 | 76.23±2.6 | 68.30±2.9 |
| | Mistral-7B-Instruct-v0.3 | 64.44±4.1 | 44.44±4.3 | 68.30±2.9 | 60.38±3.0 |
| | Qwen3-8B | 73.33±3.8 | **62.96±4.2** | 77.36±2.6 | 74.72±2.7 |
| | Llama3.1-8B-Instruct | 51.85±4.3 | 44.44±4.3 | 47.17±3.1 | 38.87±3.0 |
| | Apertus-8B-Instruct | 59.26±4.2 | **60.74±4.2** | 70.94±2.8 | 61.89±3.0 |
| | EuroLLM-9B-Instruct | 59.26±4.2 | 57.04±4.3 | 60.38±3.0 | 59.62±3.0 |
| | Qwen3-14B | 79.26±3.5 | **67.41±4.0** | 81.51±2.4 | **78.49±2.5** |
| | GPT-OSS-20B | 48.89±4.3 | 60.00±4.2 | 66.42±2.9 | 62.26±3.0 |
| | Qwen3-32B | 80.00±3.5 | **67.41±4.0** | 86.42±2.1 | **80.38±2.4** |
| BioMed | MedGemma-4B-pt | 48.89±4.3 | 42.96±4.3 | 52.83±3.1 | 48.68±3.1 |
| | MedGemma-4B-it | 54.81±4.3 | 52.59±4.3 | 61.51±3.0 | 63.02±3.0 |
| | Apollo-7B | 59.26±4.2 | 47.41±4.3 | 69.43±2.8 | 61.51±3.0 |
| | BioMistral-7B | 46.67±4.3 | 42.96±4.3 | 63.77±3.0 | 55.47±3.1 |
| | MedGemma-27B-it | 79.26±3.5 | **68.89±4.0** | 81.89±2.4 | **76.98±2.6** |

Table 8: Accuracy (mean) and standard error on MMLU medical benchmarks.

| Type | Model | Accuracy (%)↑ | | | |
|---|---|---|---|---|---|
| | | College Biology | | College Medecine | |
| | | EN | FR | EN | FR |
| Base | Gemma3-270m-pt | 20.83±3.4 | 21.53±3.4 | 23.70±3.2 | 19.65±3.0 |
| | Baguettotron | 26.39±3.7 | 18.06±3.2 | 27.17±3.4 | 23.12±3.2 |
| | Qwen3-0.6B-Base | 59.72±4.1 | 39.58±4.1 | 52.60±3.8 | 40.46±3.7 |
| | Llama3.2-1B | 77.08±3.5 | 62.50±4.1 | 64.74±3.6 | 59.54±3.7 |
| | Gemma3-1B-pt | 25.69±3.7 | 27.08±3.7 | 21.39±3.1 | 21.97±3.2 |
| | Gaperon-1B | 22.92±3.5 | 20.14±3.4 | 20.81±3.1 | 21.39±3.1 |
| | Qwen3-1.7B-Base | 72.92±3.7 | 54.17±4.2 | 68.21±3.6 | 59.54±3.7 |
| | Qwen3-4B-Base | 84.72±3.0 | 75.00±3.6 | 73.99±3.3 | 68.21±3.6 |
| | Olmo3-7B | 76.39±3.6 | 53.47±4.2 | 70.52±3.5 | 50.87±3.8 |
| | Mistral-7B-v0.3 | 70.14±3.8 | 54.86±4.2 | 63.01±3.7 | 53.18±3.8 |
| | Qwen3-8B-Base | 88.89±2.6 | **87.50±2.8** | 78.03±3.2 | **75.72±3.3** |
| | Llama3.1-8B | 77.08±3.5 | 62.50±4.1 | 64.74±3.6 | 59.54±3.7 |
| | Apertus-8B | 72.92±3.7 | 68.75±3.9 | 61.85±3.7 | 57.80±3.8 |
| | Gaperon-8B | 62.50±4.0 | 52.78±4.2 | 47.40±3.8 | 45.09±3.8 |
| | EuroLLM-9B | 66.67±3.9 | 66.67±3.9 | 54.34±3.8 | 52.02±3.8 |
| Instruct | Gemma3-270m-it | 34.72±4.0 | 22.92±3.5 | 21.39±3.1 | 21.39±3.1 |
| | Qwen3-0.6B | 56.25±4.1 | 34.03±4.0 | 47.40±3.8 | 38.15±3.7 |
| | Llama3.2-1B-Instruct | 46.53±4.2 | 34.03±4.0 | 36.99±3.7 | 28.32±3.4 |
| | Gemma3-1B-it | 34.72±4.0 | 36.11±4.0 | 39.31±3.7 | 38.73±3.7 |
| | Qwen3-1.7B | 67.36±3.9 | 49.31±4.2 | 61.85±3.7 | 56.07±3.8 |
| | SmolLM-3B | 72.92±3.7 | 59.72±4.1 | 64.16±3.7 | 55.49±3.8 |
| | Qwen3-4B | 84.72±3.0 | 73.61±3.7 | 70.52±3.5 | **73.99±3.3** |
| | Mistral-7B-Instruct-v0.3 | 72.22±3.7 | 59.03±4.1 | 60.12±3.7 | 54.34±3.8 |
| | Qwen3-8B | 88.19±2.7 | 85.42±3.0 | 79.19±3.1 | **73.41±3.4** |
| | Llama3.1-8B-Instruct | 81.94±3.2 | 67.36±3.9 | 65.32±3.6 | 63.01±3.7 |
| | Apertus-8B-Instruct | 75.00±3.6 | 67.36±3.9 | 61.27±3.7 | 59.54±3.7 |
| | EuroLLM-9B-Instruct | 71.53±3.8 | 71.53±3.8 | 53.76±3.8 | 54.34±3.8 |
| | Qwen3-14B | 92.36±2.2 | **89.58±2.6** | 80.92±3.0 | **78.03±3.2** |
| | GPT-OSS-20B | 70.83±3.8 | 69.44±3.9 | 54.34±3.8 | 56.07±3.8 |
| | Qwen3-32B | 90.28±2.5 | 90.97±2.4 | 82.08±2.9 | **77.46±3.2** |
| BioMed | MedGemma-4B-pt | 59.03±4.1 | 48.61±4.2 | 49.71±3.8 | 42.77±3.8 |
| | MedGemma-4B-it | 68.75±3.9 | 54.86±4.2 | 55.49±3.8 | 56.65±3.8 |
| | Apollo-7B | 77.78±3.5 | 62.50±4.1 | 59.54±3.7 | 58.38±3.8 |
| | BioMistral-7B | 59.03±4.1 | 47.22±4.2 | 53.76±3.8 | 50.29±3.8 |
| | MedGemma-27B-it | 85.42±3.0 | **86.81±2.8** | 72.83±3.4 | **73.99±3.3** |

Table 9: Accuracy (mean) and standard error on MMLU medical benchmarks.

| Type | Model | Accuracy (%)↑ | | | |
|---|---|---|---|---|---|
| | | Medical Genetics | | Professional Medecine | |
| | | EN | FR | EN | FR |
| Base | Gemma3-270m-pt | 25.00±4.4 | 23.00±4.2 | 43.38±3.0 | 23.16±2.6 |
| | Baguettotron | 30.00±4.6 | 24.00±4.3 | 44.85±3.0 | 20.96±2.5 |
| | Qwen3-0.6B-Base | 62.00±4.9 | 47.00±5.0 | 55.51±3.0 | 37.13±2.9 |
| | Llama3.2-1B | 80.00±4.0 | 67.00±4.7 | 69.85±2.8 | 55.51±3.0 |
| | Gemma3-1B-pt | 27.00±4.5 | 26.00±4.4 | 30.51±2.8 | 24.63±2.6 |
| | Gaperon-1B | 30.00±4.6 | 27.00±4.5 | 44.85±3.0 | 30.88±2.8 |
| | Qwen3-1.7B-Base | 73.00±4.5 | 66.00±4.8 | 64.71±2.9 | 56.25±3.0 |
| | Qwen3-4B-Base | 81.00±3.9 | 74.00±4.4 | 78.31±2.5 | 69.49±2.8 |
| | Olmo3-7B | 76.00±4.3 | 63.00±4.9 | 63.60±2.9 | 48.90±3.0 |
| | Mistral-7B-v0.3 | 73.00±4.5 | 59.00±4.9 | 65.07±2.9 | 54.41±3.0 |
| | Qwen3-8B-Base | 86.00±3.5 | **80.00±4.0** | 83.46±2.3 | 76.47±2.6 |
| | Llama3.1-8B | 80.00±4.0 | 67.00±4.7 | 69.85±2.8 | 55.51±3.0 |
| | Apertus-8B | 67.00±4.7 | 64.00±4.8 | 60.29±3.0 | 59.56±3.0 |
| | Gaperon-8B | 62.50±4.0 | 65.00±4.8 | 48.53±3.0 | 42.28±3.0 |
| | EuroLLM-9B | 69.00±4.7 | 62.00±4.9 | 55.51±3.0 | 55.88±3.0 |
| Instruct | Gemma3-270m-it | 27.00±4.5 | 28.00±4.5 | 36.76±2.9 | 18.75±2.4 |
| | Qwen3-0.6B | 54.00±5.0 | 38.00±4.9 | 41.18±3.0 | 33.46±2.9 |
| | Llama3.2-1B-Instruct | 57.00±5.0 | 46.00±5.0 | 51.84±3.0 | 26.47±2.7 |
| | Gemma3-1B-it | 41.00±4.9 | 44.00±5.0 | 27.21±2.7 | 20.96±2.5 |
| | Qwen3-1.7B | 75.00±4.4 | 62.00±4.9 | 58.09±3.0 | 48.16±3.0 |
| | SmolLM-3B | 67.00±4.7 | 58.00±5.0 | 55.88±3.0 | 49.26±3.0 |
| | Qwen3-4B | 80.00±4.0 | 69.00±4.6 | 76.10±2.6 | 68.38±2.8 |
| | Mistral-7B-Instruct-v0.3 | 72.22±3.7 | 60.00±4.9 | 62.50±2.9 | 52.94±3.0 |
| | Qwen3-8B | 86.00±3.5 | **80.00±4.0** | 82.35±2.3 | 72.79±2.7 |
| | Llama3.1-8B-Instruct | 84.00±3.7 | 68.00±4.7 | 76.47±2.6 | 61.40±3.0 |
| | Apertus-8B-Instruct | 65.00±4.8 | 62.00±4.9 | 62.87±2.9 | 60.66±3.0 |
| | EuroLLM-9B-Instruct | 63.00±4.9 | 64.00±4.8 | 59.56±3.0 | 56.25±3.0 |
| | Qwen3-14B | 89.00±3.1 | **80.00±4.0** | 83.46±2.3 | **78.68±2.5** |
| | GPT-OSS-20B | 64.00±4.8 | 64.00±4.8 | 59.19±3.0 | 52.94±3.0 |
| | Qwen3-32B | 94.00±2.4 | **85.00±3.6** | 85.29±2.2 | **84.19±2.2** |
| BioMed | MedGemma-4B-pt | 55.00±5.0 | 48.00±5.0 | 38.24±3.0 | 35.29±2.9 |
| | MedGemma-4B-it | 67.00±4.7 | 65.00±4.8 | 62.13±2.9 | 51.47±3.0 |
| | Apollo-7B | 77.00±4.2 | 67.00±4.7 | 68.01±2.8 | 59.93±3.0 |
| | BioMistral-7B | 65.00±4.8 | 48.00±5.0 | 55.15±3.0 | 46.32±3.0 |
| | MedGemma-27B-it | 87.00±3.4 | **86.00±3.5** | 83.09±2.3 | **78.68±2.5** |

Table 10: Accuracy (mean) and standard error on MMLU medical benchmarks.

| Type | Model | Accuracy (%)↑ | |
|---|---|---|---|
| | | EN | FR |
| Base | Gemma3-270m-pt | 9.02±1.1 | 9.46±1.1 |
| | Baguettotron | 10.92±1.2 | 6.26±0.9 |
| | Qwen3-0.6B-Base | 22.42±1.6 | 15.87±1.4 |
| | Llama3.2-1B | 14.85±1.4 | 10.77±1.2 |
| | Gemma3-1B-pt | 10.92±1.2 | 9.75±1.1 |
| | Gaperon-1B | 12.81±1.3 | 7.57±1.0 |
| | Qwen3-1.7B-Base | 34.64±1.8 | 25.91±1.7 |
| | Qwen3-4B-Base | 49.93±1.9 | 39.30±1.9 |
| | Olmo3-7B | 34.93±1.8 | 18.20±1.5 |
| | Mistral-7B-v0.3 | 35.52±1.8 | 27.22±1.7 |
| | Qwen3-8B-Base | 55.75±1.9 | 50.22±1.9 |
| | Llama3.1-8B | 43.38±1.9 | 28.53±1.7 |
| | Apertus-8B | 34.64±1.8 | 36.10±1.8 |
| | Gaperon-8B | 26.64±1.7 | 20.52±1.5 |
| | EuroLLM-9B | 33.92±1.8 | 29.11±1.7 |
| Instruct | Gemma3-270m-it | 9.90±1.1 | 10.63±1.2 |
| | Qwen3-0.6B | 22.27±1.6 | 14.41±1.3 |
| | Qwen3-1.7B | 35.52±1.8 | 25.91±1.7 |
| | Llama3.2-1B-Instruct | 22.13±1.6 | 16.45±1.4 |
| | Gemma3-1B-it | 16.16±1.4 | 13.68±1.3 |
| | SmolLM-3B | 36.97±1.8 | 29.40±1.7 |
| | Qwen3-4B | 52.11±1.9 | 41.63±1.9 |
| | Mistral-7B-Instruct-v0.3 | 40.90±1.9 | 34.21±1.8 |
| | Qwen3-8B | 57.50±1.9 | 48.18±1.9 |
| | Llama3.1-8B-Instruct | 49.64±1.9 | 36.39±1.8 |
| | Apertus-8B-Instruct | 41.48±1.9 | 41.92±1.9 |
| | EuroLLM-9B-Instruct | 36.97±1.8 | 30.86±1.8 |
| | Qwen3-14B | 63.03±1.8 | 56.48±1.9 |
| | GPT-OSS-20B | 37.26±1.9 | 35.81±1.8 |
| | Qwen3-32B | 68.85±1.8 | **66.08±1.8** |
| BioMed | MedGemma-4B-pt | 20.67±1.5 | 18.49±1.5 |
| | MedGemma-4B-it | 41.48±1.9 | 31.30±1.8 |
| | Apollo-7B | 43.67±1.9 | 30.71±1.8 |
| | BioMistral-7B | 32.17±1.8 | 20.23±1.5 |
| | MedGemma-27B-it | 65.36±1.8 | 61.86±1.9 |

Table 11: Accuracy (mean) and standard error on the MMLU-Pro-X task *Health*.