Choosing How to Adapt: An Empirical Study on Cross-Lingual Medical Question-Answering Adaptation

Anonymous ACL submission

Abstract

The development of large language models (LLMs) has led to increased focus on their adaptation to specialized domains and languages, yet the effectiveness of domain adaptation strategies remains unclear. We present a study of medical domain adaptation using French medical question answering (QA) as a case study. We compare continual pretraining (CPT), supervised fine-tuning (SFT), and their combination across three model families, multiple sizes, and three initialization types, explicitly disentangling adaptation effects from base model choice. We evaluate both multiple-choice (MCQA) and open-ended QA (OEQA) under greedy and constrained decoding using automatic metrics and LLM-as-a-Judge evaluation. For MCQA, CPT+SFT most often achieves the best scores, but gains over SFT are small and frequently not statistically significant, making SFT a strong and cost-effective default. For OEQA, CPT consistently improves overlap-based metrics, while SFT often degrades generation quality; instruction tuning and CPT+SFT are preferred by LLM-based evaluation. Cross-lingual experiments further show effective transfer from French adaptation to English benchmarks. Overall, we provide practical guidelines for selecting adaptation strategies under computational constraints.

1 Introduction

LLMs are increasingly applied to medical QA and clinical reasoning, where accuracy, robustness, and domain-specific knowledge are critical (Huang et al., 2024). However, most high-performing LLMs are trained primarily on general-domain data, making domain adaptation a necessary step for safe and effective medical deployment. Despite its importance, the effectiveness of medical domain adaptation remains debated. Recent work comparing continual pretraining (CPT) and supervised fine-tuning (SFT) shows that their impact depends strongly on factors such as training scale, data composition, and optimization choices (Christophe et al., 2024; Lu et al., 2025). Although combining CPT and SFT can yield additional gains, prior studies largely report aggregate results, leaving open questions about why certain adaptation strategies succeed or fail. In parallel, Jeong et al. (2024a) challenge common assumptions about medical domain adaptation, showing that biomedical CPT often yields inconsistent or statistically insignificant improvements over base models. Taken together, these findings suggest that adaptation effectiveness is highly context-dependent and sensitive to model initialization and evaluation setup.

Most studies fix the base model initialization, making it difficult to isolate adaptation effects from those of the starting point (Lu et al., 2025; Christophe et al., 2024). In addition, evaluations are predominantly conducted in English, limiting conclusions about generalization to other languages. Finally, prior work focuses mainly on MCQA, where models select one or more answers from a predefined set of options, with limited qualitative analysis, complicating the interpretation of adaptation gains, particularly in light of recent evidence of non-trivial memorization in medical LLMs (Li et al., 2025).

To address these limitations, we conduct a statistically grounded study of medical domain adaptation using French medical QA as a case study. We compare CPT, SFT, and their combination across multiple model families and sizes while explicitly varying base model initialization. Models are evaluated before and after adaptation in both French and English to isolate domain learning from language transfer. Beyond MCQA (predefined answer selection), we evaluate OEQA to assess generative medical reasoning, a setting largely underexplored in prior studies. Our goal is to clarify when and why CPT and SFT are effective for multilingual medical LLM adaptation. Our study is guided by

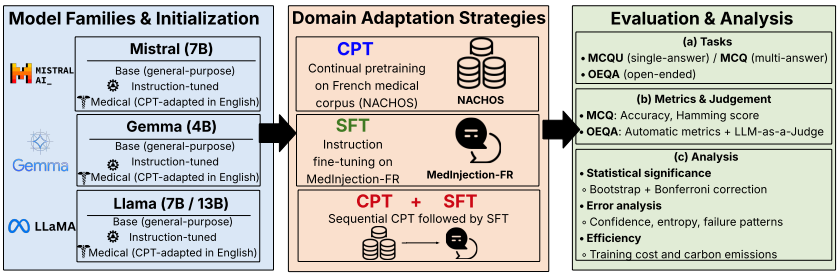

Figure 1: Overview of the experimental pipeline for evaluating medical domain adaptation strategies.

the following research questions:

- **RQ1:** *What are the performance and efficiency trade-offs between CPT and SFT across model families and sizes?*
- **RQ2:** *How does base model initialization influence the effectiveness of CPT and SFT for medical domain adaptation?*
- **RQ3:** *How does French medical adaptation affect cross-lingual transfer to English?*

Our contributions are: (i) we introduce a controlled and reproducible framework to compare medical domain adaptation strategies across model families, sizes, initialization types, and decoding settings; (ii) we provide a statistically grounded analysis of CPT and SFT for medical QA, covering performance trade-offs, error patterns, and cross-lingual transfer to English benchmarks, and publicly release all resources.[1]

2 Related Work

Medical LLM adaptation primarily relies on CPT on domain-specific corpora and SFT on instruction–response data, both shown to support domain transfer (Gururangan et al., 2020; Gema et al., 2024). CPT has been adopted in models such as MediTron (Chen et al., 2023b), BioMistral (Labrak et al., 2024a), PMC-Llama (Wu et al., 2023), and MedGemma (Sellergren et al., 2025). However, recent analyses question the robustness and consistency of CPT gains under stricter evaluation protocols (Jeong et al., 2024a). In parallel, SFT-based models such as ChatDoctor (Li et al., 2023) and MedAlpaca (Han et al., 2023) report strong task-level improvements, though evaluations remain largely in English.

Medical domain adaptation is further challenged in non-English settings due to limited domain-specific resources. Several multilingual medical LLMs have been proposed, including Medical mT5 (García-Ferrero et al., 2024), BiMediX (Pieri et al., 2024), Apollo (Wang et al., 2024a), and MMedLM (Qiu et al., 2024). However, these models are mostly evaluated on translated benchmarks, with limited validation on native-language medical tasks, leaving their language specificities underexplored. Evaluation practices also pose challenges. Widely used benchmarks such as PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022) primarily target English and emphasize multiple-choice formats, which may not fully capture generative medical reasoning, especially in multilingual contexts.

Beyond proposing individual models, recent work has compared adaptation strategies in controlled settings. Christophe et al. (2024) analyze CPT, SFT, and related techniques for clinical LLMs, finding that CPT alone yields limited gains but can amplify performance when combined with instruction tuning. Similarly, Lu et al. (2025) study CPT, SFT, and preference-based optimization across domains, highlighting complex interactions between adaptation methods. However, these studies focus on English and fix the base model initialization. In contrast, in this work we systematically compare CPT, SFT and their combination across multiple model families and initialization points for French medical QA, while also evaluating cross-lingual performance and analyzing adaptation behavior.

3 Experimental Framework

We propose a controlled experimental framework to evaluate medical domain adaptation strategies across architectures, initialization points, and task formats, as illustrated in Figure 1. Our setup explicitly varies (i) the base model and its prior training,

[1] https://anonymous.4open.science/r/MedAdapt-2F23/README.md

(ii) the adaptation strategy, and (iii) the evaluation task and language, in order to isolate the factors that drive adaptation effectiveness.

3.1 Base Models and Adaptation Approaches

Our study focuses on model families for which three complementary initialization states are available: a general-purpose base model, an instruction-tuned variant, and a medically adapted version obtained via CPT. This design allows us to separate the impact of adaptation strategy from that of the starting point.

We consider three model families spanning different sizes, pretraining regimes, and linguistic exposure. Specifically, we include Mistral-7B, Gemma-4B, and Llama models at the 7B and 13B scales. For Mistral-7B, we use Mistral-7B-v0.1 and its instruction-tuned version, and BioMistral-7B, a model adapted to the biomedical domain via CPT (Jiang et al., 2023; Labrak et al., 2024b). For Gemma, we rely on the Gemma-3-4B pretrained and instruction-tuned models, together with MedGemma-3-4B, which incorporates medical pretraining (Team et al., 2025; Sellergren et al., 2025). Finally, for Llama, we include both 7B and 13B variants, using the base and chat versions of Llama-2, as well as their medically adapted counterparts, MediTron-7B and MedLlama-13B (Touvron et al., 2023; Chen et al., 2023a; Wu et al., 2024).

These families differ not only in scale but also in pretraining data and exposure to French. Mistral and Gemma are explicitly multilingual, whereas Llama models are primarily English-centric, although exact language proportions are not disclosed. With the exception of MedGemma, whose medical pretraining corpus is not fully documented, all medical variants rely on PubMed Central as their primary biomedical source[2].

Across all model families and initialization points, we investigate three adaptation strategies: (i) CPT on domain-specific corpora, (ii) SFT on instruction-response pairs, and (iii) a sequential CPT+SFT pipeline.

3.2 Training Data

CPT. We use NACHOS corpus (Labrak et al., 2023), an open-source French medical dataset comprising 4 GB of text collected from French medical websites; full details are provided in Appendix A.

[2] https://pmc.ncbi.nlm.nih.gov/

SFT. We use the train and validation sets of the MedInjection-FR corpus,[3] which contains 543 505 instruction-response pairs. The dataset includes multiple-choice questions with a single **unique** correct answer (MCQU, ∼83%), **multiple** correct answers (MCQ, ∼6%), and OEQAs (∼11%). This mixture allows us to evaluate adaptation effects across both discriminative and generative medical reasoning tasks. Additional dataset details are provided in Appendix B.

3.3 Training Process

To explore the trade-off between computational cost and model plasticity, we adopt contrasting fine-tuning regimes for CPT and SFT. CPT is performed using full-parameter fine-tuning, while SFT relies on parameter-efficient adaptation.

CPT. CPT is performed for three epochs following the setup of Labrak et al. (2024b). Full hyperparameter details are provided in Appendix C.

SFT. We employ DoRA (Weight-Decomposed Low-Rank Adaptation) (Mao et al., 2024), an extension of LoRA (Hu et al., 2022) that decouples magnitude and directional updates. We select DoRA after preliminary experiments showing improved adaptation efficiency and task performance compared to standard LoRA. SFT is run for ten epochs, with hyperparameters reported in Appendix D.

3.4 Evaluation Protocol

Benchmarks. We evaluate all models on MedInjection-FR test set, which consists of 14 533 native French medical examples and 13 293 translated examples derived from established English benchmarks. The test set covers MCQU, MCQ, and OEQA tasks, enabling evaluation of both answer selection and free-form answers. Benchmark sources and translation procedure are detailed in Appendix F.

Prompting Strategy. All evaluations are conducted in a zero-shot setting using greedy, deterministic decoding to ensure reproducibility. For MCQU and MCQ tasks, following Liang et al. (2022); Beeching et al. (2023); Chen et al. (2023a), we restrict the output vocabulary to valid answer options to prevent hallucinated responses. To mitigate position bias, we randomly shuffle answer choices three times and report aggregated results, following

[3] https://huggingface.co/spaces/MedInjection-FR/README

best practices for MCQ evaluation (Pezeshkpour and Hruschka, 2024). Prompt templates are provided in Appendix G.

Evaluation Metrics. For MCQU, we report *Exact Match (EM)*, which measures the proportion of questions for which the predicted answer exactly matches the gold answer. For MCQ, we additionally report the *Hamming score*, which accounts for partial overlap between predicted and reference answer sets and is therefore more informative for multi-answer questions. Formal definitions of both metrics are provided in Appendix E. For OEQA, we rely on both automatic text-based metrics and model-based judgments. We report BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2019) as automatic baselines. Their reliability is assessed by measuring agreement with senior physician annotations on a held-out subset of 500 OEQA instances. Details of the annotation and agreement analysis are provided in Appendix H.

To better align evaluation with human judgment, we complement automatic metrics with an LLM-as-a-Judge approach. On the same annotated subset, LLM-based judges correlate substantially better with human ratings than overlap metrics (Table 6). Qwen-80B yields the highest Pearson correlation, while MedGemma-27B achieves the best F1 with lower self-preference bias and is therefore used as the primary OEQA evaluator (Appendix H).

Statistical Significance and Error Analysis. We assess statistical significance using a percentile bootstrap procedure with 10 000 resamples, following Jeong et al. (2024b). Differences between paired model configurations are considered significant when the associated two-sided p-value is below a predefined threshold α. To control for multiple comparisons, we apply Bonferroni correction, yielding corrected α as specified in Appendix J. In addition, we conduct an error analysis by examining output probabilities, confidence scores, and entropy, enabling us to characterize how CPT and SFT affect uncertainty and error patterns across different base model initializations.

4 Results and Discussion

4.1 MCQA Evaluation

Table 1 reports performance on MCQA across three model families (Gemma-4B, Mistral-7B, Llama-7B-13B), three initialization types (General, Instruct, Medical), and three adaptation strategies (CPT, SFT, CPT+SFT). Results are shown for both MCQs and MCQUs, using EM and Hamming scores for MCQs. All results reported in this table are obtained using constrained decoding. The corresponding results under greedy decoding are provided in Appendix I.

Effectiveness of Adaptation Strategy: A recurring pattern observed throughout the results is:

$$\text{BASE} \ll \text{CPT} \ < \text{SFT} \lesssim \text{CPT+SFT}$$

The strongest performance is most frequently achieved by the adaptation CPT+SFT. Across model families and initialization types, CPT+SFT yields the highest scores in aggregated EM as well as in MCQ and MCQU EM more often than any other strategy.

However, a closer inspection of the results indicates that the gains brought by CPT+SFT over SFT alone are generally limited. When CPT+SFT attains the highest score, the margin over SFT rarely exceeds 1.3 points. In contrast, in configurations where SFT outperforms CPT+SFT, the performance gap is larger. For example, on Llama-7B Instruct, SFT exceeds CPT+SFT by 3.12 points, and a similar pattern is observed for Mistral-7B Instruct, with a gap of 1.44 points in favor of SFT.

Furthermore, the statistical analysis reported in Appendix J shows that, when comparing each adapted model to its corresponding base model, the observed improvements are not always statistically significant. In particular, for Gemma Instruct, neither SFT nor CPT+SFT yields statistically significant gains over the base model. Likewise, for Llama-7B Instruct, the improvement brought by CPT+SFT is not statistically significant. These constitute the only cases in which the improvements of SFT or CPT+SFT over the base model fail to reach statistical significance. Consequently, although CPT+SFT most frequently ranks first, its advantage over SFT is not consistently substantial.

By contrast, CPT alone exhibits less stable behavior. Although it can improve performance in some rare cases, it can also occasionally degrade performance compared to the base model. Moreover, it is the strategy that most often fails to produce statistically significant improvements over the base. This is the case for 8 models: MedGemma; all Llama-13B variants; Llama-7B GENERAL and INSTRUCT; and Mistral-7B GENERAL and MED-

Model Type	Strategy	MCQ		MCQU	Aggregation	OEQA		
		EM	Hamming	EM	EM	Rouge-L	BERT-F1	Judge
Gemma-4B								
GENERAL	Base	2.24	30.17	27.11	14.68	7.11	46.62	25.09
	CPT	0.73	11.71	25.83	13.28	**10.18**	**49.07**	**25.71**
	SFT	3.73	**43.57**	32.36	18.05	8.23	47.76	21.60
	CPT+SFT	**3.90**	42.81	**32.59**	**18.25**	6.01	45.78	24.80
INSTRUCT	Base	**4.83**	44.01	29.30	17.06	**7.38**	**48.94**	**47.71**
	CPT	3.47	46.00	25.05	14.26	4.57	42.77	13.35
	SFT	3.68	**48.22**	**31.95**	**17.81**	2.20	38.26	20.21
	CPT+SFT	3.42	48.14	30.73	17.07	3.12	40.90	20.76
MEDICAL	Base	1.98	31.46	26.41	14.19	**7.38**	**48.94**	**22.01**
	CPT	1.68	24.04	25.13	13.41	6.77	45.18	1.95
	SFT	**3.54**	**46.22**	30.66	17.10	5.70	45.66	14.23
	CPT+SFT	3.38	43.28	**30.86**	**17.12**	5.04	43.05	11.41
Mistral-7B								
GENERAL	Base	0.37	5.40	28.52	14.44	5.82	44.21	**27.23**
	CPT	3.54	30.50	27.21	15.37	7.22	46.32	24.57
	SFT	5.24	21.62	**32.88**	19.06	**8.83**	**48.02**	22.93
	CPT+SFT	**6.13**	**30.86**	32.29	**19.21**	6.62	46.35	24.89
INSTRUCT	Base	4.86	23.53	24.92	14.89	7.34	49.66	30.14
	CPT	**7.32**	**36.18**	28.79	18.06	**13.51**	**53.87**	**37.59**
	SFT	6.80	23.42	**31.61**	**19.21**	12.41	52.72	17.63
	CPT+SFT	5.45	32.47	30.09	17.77	9.02	48.99	32.13
MEDICAL	Base	2.80	17.47	26.69	14.74	11.34	**51.58**	20.76
	CPT	3.57	24.43	25.73	14.65	**12.41**	51.45	17.89
	SFT	3.36	26.37	31.62	17.49	8.75	47.86	16.72
	CPT+SFT	**4.94**	**27.27**	**32.58**	**18.76**	9.15	48.63	**24.96**
Llama-7B								
GENERAL	Base	1.33	12.01	25.72	13.53	5.05	41.27	9.39
	CPT	1.12	12.86	25.59	13.36	**10.58**	**47.85**	3.78
	SFT	2.66	28.41	28.93	15.80	6.02	44.49	7.67
	CPT+SFT	**3.17**	**46.00**	**29.89**	**16.53**	5.85	44.67	**12.26**
INSTRUCT	Base	3.95	34.43	25.08	14.51	2.57	43.85	25.35
	CPT	3.93	**42.67**	25.07	14.50	11.16	**51.37**	26.06
	SFT	**5.12**	21.06	**29.32**	**17.22**	**11.44**	51.28	12.92
	CPT+SFT	3.13	25.98	25.07	14.10	9.92	50.99	**27.84**
MEDICAL	Base	0.23	2.90	24.43	12.33	5.61	43.25	12.50
	CPT	2.37	28.06	25.60	13.99	**8.00**	**45.79**	13.14
	SFT	3.24	29.10	30.44	16.84	5.40	42.25	9.50
	CPT+SFT	**3.80**	**44.95**	**31.52**	**17.66**	5.87	44.34	**17.39**
Llama-13B								
GENERAL	Base	2.14	21.20	26.11	14.13	2.10	33.25	11.79
	CPT	2.53	19.17	26.99	14.76	**14.12**	**50.36**	5.66
	SFT	**3.54**	**40.49**	30.95	17.24	5.60	43.19	14.88
	CPT+SFT	3.34	29.59	**32.36**	**17.85**	6.30	45.45	**20.38**
INSTRUCT	Base	0.09	29.74	21.52	10.81	3.40	45.31	30.02
	CPT	5.63	**37.40**	25.01	15.32	12.34	**53.07**	**36.19**
	SFT	6.58	23.96	30.20	18.39	11.54	50.94	11.81
	CPT+SFT	**7.77**	25.26	**31.58**	**19.68**	**12.86**	52.46	20.22
MEDICAL	Base	1.77	11.82	24.62	13.19	5.00	42.11	10.86
	CPT	2.26	30.87	24.10	13.18	8.00	45.79	13.39
	SFT	3.12	41.24	30.62	16.87	6.85	45.22	13.77
	CPT+SFT	**3.24**	**45.59**	**32.25**	**17.74**	**8.38**	**46.29**	**19.55**

Table 1: Constrained decoding results (%) for MCQ/MCQU and OEQA. Aggregation corresponds to average EM over MCQ and MCQU. **Bold** denotes the best strategy, and underlining the best initialization.

ICAL. This suggests that representation-level domain adaptation is most effective when paired with task-specific supervision.

Overall, while CPT+SFT ranks first most often, its limited and inconsistent gains over SFT, together with a substantially higher computational cost (Appendix O), make SFT a strong default for medical MCQA, for example, on 7B models, CPT+SFT costs over $1 500 versus $360 for SFT, with a fourfold increase in carbon emissions.

Impact of Model Initialization: The impact of model initialization (General, Instruct, Medical), varies across MCQA metrics and question formats. Considering the overall best scores across all model families, instruction-tuned models dominate the most demanding EM settings: the highest MCQ EM and aggregated MCQA EM scores are both achieved by Llama-13B Instruct, while the best MCQ Hamming score is obtained by Gemma-4B Instruct. In contrast, the best MCQU EM score is achieved by a general Mistral model.

At the family level, the patterns differ. For MCQ EM, the best score within each model family is always obtained by an instruction-tuned variant, confirming that instruction alignment is particularly beneficial for exact multi-label prediction; this advantage is further supported by statistically significant gains when compared to general or medical initializations (Appendix J). For MCQ Hamming, results are more balanced, with the best scores split across initialization types (two instruction-tuned, one general, and one medical).

For MCQU EM, general models most frequently achieve the best performance (three cases), followed by medical models, while instruction-tuned models do not dominate. This indicates that when only a single answer must be selected, performance is driven primarily by answer plausibility ranking, favoring strong language modeling and domain knowledge, while explicit instruction alignment, which mainly benefits structured or multi-label outputs, provides less advantage. Finally, for the aggregated MCQA score, no single initialization consistently dominates: instruction-tuned and general models each obtain the best result in two config-

urations (with ties between them), while medical models lead in one case, and differences across initializations are often not statistically significant.

4.2 OEQA Evaluation

The right side of Table 1 reports OEQA across model families using ROUGE-L, BERTScore and LLM-as-a-Judge . We additionally report BLEU and METEOR in Appendix I, as they reflect similar information as ROUGE-L. Overall, absolute scores remain moderate, reflecting the difficulty of evaluating free-form answer generation. ROUGE-L scores should be interpreted with caution, as they measure surface-level lexical overlap and penalize semantically correct answers that differ in formulation (Yim et al., 2025; Zhu et al., 2025).

Moreover, OEQA represents only 11% of the training data, resulting in a strongly imbalanced supervision signal. Models are therefore adapted to generate short, structured outputs (answer letters in MCQA), which limits OEQA performance.

Effect of Adaptation Strategy: Across model families, SFT often degrades ROUGE-L and BERTScore-F1 compared to base or CPT adapted models, particularly for instruction-tuned and medical variants. This suggests that SFT can overly constrain generation, reducing lexical diversity and semantic overlap in an open-ended setting.

By contrast, CPT is the most consistently beneficial strategy for OEQA. CPT improves ROUGE-L and BERTScore-F1 across most general, instruct, and medical models, with especially strong gains for Mistral and Llama families. These results suggest that domain-adaptive language modeling supports better medical generation than instruction-level supervision alone. Combining CPT with SFT rarely outperforms CPT alone and often leads to intermediate or degraded performance, reflecting the same instability observed in MCQA, but with more pronounced negative effects in OEQA.

In contrast to overlap-based metrics, LLM-as-a-Judge favors CPT+SFT in half of the configurations (6/12), compared to three cases each for the base and CPT models. Gains are most pronounced for Llama-7B, where CPT+SFT consistently outperforms SFT across initializations, and for medical models, where it yields the best or near-best qualitative scores. However, despite these trends, statistically significant improvements over the base model are rare: CPT is significant in only three cases, SFT in two (all involving smaller 4B models), and CPT+SFT never yields statistically significant gains over the base model in OEQA.

Effect of Model Initialization. Initialization effects on OEQA depend strongly on the evaluation metric. For overlap-based metrics, no initialization consistently dominates: ROUGE-L is split between general and instruction-tuned models, while medical models never achieve the top score; BERTScore-F1 is mostly dominated by instruction-tuned models, with a single exception (Gemma).

LLM-as-a-Judge reveals clearer and statistically grounded patterns. When differences are significant, medical models are consistently outperformed across families, particularly under SFT and CPT (Table 9). Comparisons between instruction-tuned and general models are mixed and direction-dependent, with some significant gains under SFT and CPT (notably for Mistral and Llama), but these effects largely disappear under CPT+SFT.

Overall, medical initialization alone does not improve OEQA, while instruction-tuned initialization yields more reliable, yet limited, gains when significant.

5 Cross-Lingual Transfer After French Medical Adaptation

To analyze whether models perform better in English prior to adaptation and how cross-lingual adaptation affects performance, we compute the EM accuracy difference for MCQU benchmarks as the score on French translations minus the score on the corresponding native English datasets. Figure 2 reports averaged results across datasets using constrained decoding; full results for both greedy and constrained decoding are provided in Appendix M.

For the Mistral family, base models consistently perform better on the translated French benchmarks than on the original English data. This trend persists after adaptation with CPT, SFT, and CPT+SFT, with French performance systematically exceeding English, the differences being statistically significant (Table 13).

In contrast, Gemma and Llama families show higher performance on native English benchmarks at the base level, and this advantage remains after adaptation on French data. Moreover, adaptation gains are often larger in English than in French (Table 12), despite all adaptation data being in French.

These results suggest that Mistral models encode French more effectively, whereas Gemma and Llama have stronger English representations.

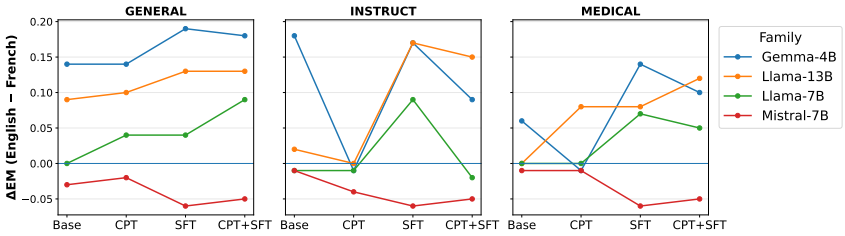

Figure 2: Difference in EM accuracy (ΔEM) between native English MCQU test benchmarks and their French translations across model families and adaptation strategies (constrained decoding).

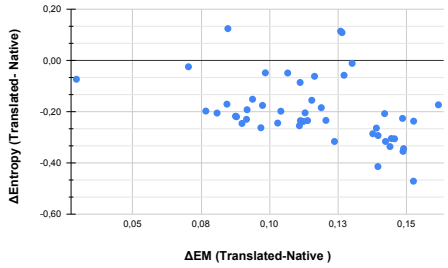

Figure 3: Relationship between accuracy gain (ΔEM) and change in predictive entropy ($\Delta Entropy$) when moving from the translated to native benchmarks. Each point corresponds to a model configuration.

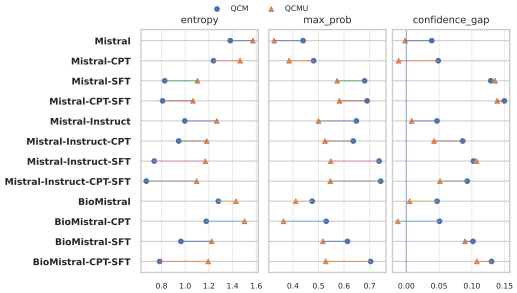

Figure 4: Probability-level metrics for MCQ and MCQU across Mistral variants.

Notably, the improvements observed in both languages indicate effective cross-lingual transfer of medical knowledge: adapting with French medical data improves performance on the original English benchmarks, sometimes more than on their French translations. This supports the complementarity of multilingual medical data, in line with Wang et al. (2024a).

A salient exception is Llama-7B: before adaptation, the base model shows slightly higher performance on French translations than on English, but this difference is not statistically significant (Table 13). After adaptation, English performance surpasses French, suggesting that adaptation amplifies the model's dominant English representations.

6 Effect of Translated Benchmarks on Performance and Confidence

We compare model behavior on a native benchmark, MediQAl (Bazoge, 2025), and a translated benchmark, MedMCQA (Pal et al., 2022), using accuracy and confidence-based metrics. Both benchmarks consist of MCQUs of comparable size, for fair comparison. Although instances are not shared, consistent differences across models are observed.

As shown in Figure 3, all models achieve higher EM accuracy on the translated benchmark. This gain is systematically accompanied by a reduction in predictive entropy, indicating that translated benchmarks induce more confident and less uncertain predictions. The concentration of models in the bottom-right quadrant suggests that translated benchmarks operate in a different evaluation regime, characterized by both higher performance and reduced uncertainty.

Figure 5 further reveals that accuracy gains are often associated with increased confidence on incorrect predictions. Most models exhibit positive shifts in confidence even when wrong, indicating a systematic overconfidence effect induced by the translated benchmark.

Overall, these results show that translated benchmarks are not neutral substitutes for native ones: they tend to inflate performance while also altering model confidence calibration, potentially leading to over-optimistic evaluations.

7 Error Analysis

7.1 Probability-Level Analysis of MCQA

To explain why MCQ is harder than MCQU, we analyze class probability distributions from the Mistral family, selected for its high variance across models and adaptation settings. For each item, we compute entropy, maximum probability, and a confidence gap measuring gold/non-gold separation

(mean gold vs. non-gold probability in MCQ; margin to the second-best option in MCQU). We also report a near-miss rate, defined as cases where all gold answers are ranked in the top-k but the predicted set is incorrect (Figure 4, Appendix K).

Across all variants, MCQ predictions are not more uncertain than MCQU: MCQ exhibits lower entropy and higher maximum probability, indicating confident local rankings. The confidence gap is consistently positive and increases with adaptation, but remains insufficient for exact multi-label generation under greedy decoding, leading to omissions or over-generation.

Adaptation clarifies this effect: SFT strongly improves MCQU, while gains on MCQ remain limited. CPT+SFT primarily increases ranking confidence rather than exact set match, yielding larger confidence gaps without reducing near-miss rates.

7.2 Verbosity Bias in OEQA

To better understand the differences observed between overlap-based metrics and LLM-as-a-Judge evaluations in OEQA, we analyze the length of generated answers across models. The results are reported in Appendix L. We find that CPT-adapted models systematically produce longer responses, with higher mean and median word counts across all model families. This increased verbosity provides a plausible explanation for their strong performance on ROUGE-L and BERTScore-F1, which reward lexical recall and content coverage.

In contrast, instruction-tuned models generate substantially shorter and more controlled answers, particularly under SFT, often producing concise responses with low variance. While this behavior negatively impacts overlap-based metrics, it aligns with higher LLM-as-a-Judge scores, suggesting that concise answers are preferred under LLM evaluation. Finally, SFT exhibits unstable behavior in OEQA, leading either to excessively short outputs or overly long responses depending on model initialization. Overall, these results indicate that OEQA performance is strongly influenced by length biases, and that improvements in automatic metrics may partially reflect increased verbosity rather than improved answer quality.

8 Conclusion

We presented a controlled and statistically grounded study of medical domain adaptation for LLMs using French medical QA, isolating the effects of model initialization, adaptation strategy, decoding, and evaluation. Our results show that adaptation effectiveness is task-dependent and that stronger strategies are not always more cost-effective. We therefore distill practical guidelines for selecting adaptation strategies based on data availability and computational constraints. Given the limited reliability of current OEQA metrics and the small proportion of OEQA supervision, our recommendations primarily emphasize MCQA, with OEQA trends interpreted cautiously.

Unlabeled data only. When only unlabeled medical text is available, CPT yields modest and unstable gains for MCQA and should not be used in isolation. Its benefits mainly appear on OEQA overlap-based metrics, which are sensitive to verbosity and should be interpreted with caution.

Labeled data only. With labeled QA data, SFT provides the best performance–efficiency trade-off for MCQA across all model families. It frequently matches or exceeds CPT+SFT while requiring substantially fewer computational resources, making it the most practical default in this setting.

Labeled and unlabeled data. When both data types are available, CPT+SFT most often achieves the highest MCQA scores, but improvements over SFT are typically small and not consistently statistically significant. Consequently, CPT+SFT is justified only when maximal performance outweighs computational cost.

Initialization and compute considerations. Instruction-tuned models constitute the strongest baseline for French medical MCQA. Medical initialization alone does not reliably improve downstream performance. From a resource perspective, parameter-efficient SFT is by far the most cost-effective strategy, whereas CPT incurs high computational and environmental costs for limited MCQA gains, and CPT+SFT compounds these costs for marginal improvements.

Evaluation and transfer considerations. Finally, we observe strong evaluation effects: adaptation on French medical data transfers to English benchmarks, translated datasets inflate both accuracy and confidence, and OEQA metrics are sensitive to verbosity. These findings highlight the need for task-aware adaptation choices and cautious metric interpretation in multilingual medical LLM evaluation.

9 Limitations

Our evaluation of adaptation strategies faces several limitations. First, we perform an exploratory contamination study to assess possible exposure to NACHOS during pretraining (Appendix P). Although no direct evidence of memorization is observed, likelihood-based tests remain inconclusive due to the lack of a reliable non-member biomedical control corpus, requiring the use of synthetic controls. We therefore treat these results as indicative only and avoid causal conclusions about pretraining inclusion.

Second, our evaluation of OEQA relies on overlap-based metrics, BERTScore, and LLM-as-a-Judge. While these measures capture complementary aspects of answer quality, they do not fully characterize semantic equivalence, clinical correctness, or reasoning validity, and may therefore overlook qualitative differences between correct answers (Yim et al., 2025; Zhu et al., 2025).

Third, while we demonstrate the efficiency of SFT compared to CPT in terms of computational resources, our analysis does not account for the human effort required to create high-quality instruction-tuning datasets. This consideration is particularly relevant for low-resource settings where creating domain-specific instruction data may be costly.

Fourth, we do not include few-shot prompting as an evaluation setting. Our objective is to isolate the effects of parameter-level adaptation strategies under controlled and reproducible conditions. Few-shot prompting introduces additional sources of variance related to example selection, ordering, and prompt design, which would complicate statistical comparison and obscure the interpretation of adaptation gains. Moreover, few-shot prompting assumes access to curated task-specific examples at inference time, which may be unrealistic in medical deployment scenarios. For these reasons, we focus on zero-shot evaluation to ensure fair and stable comparisons across adaptation strategies.

Finally, our study focuses exclusively on CPT and SFT. We do not explore reinforcement learning–based adaptation strategies, such as preference optimization or reward-driven fine-tuning, which may better align models with clinical judgment or evaluation criteria. Investigating how such methods interact with CPT and SFT, particularly under multilingual and domain-specific constraints, constitutes an important direction for future work.

In addition, our findings about the effectiveness of adaptation strategies are specific to the medical domain and French language. The generalizability of these results to other domains or languages, particularly those with different resource constraints or linguistic characteristics, requires further investigation.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Adrien Bazoge. 2025. Mediqal: A french medical question answering dataset for knowledge and reasoning evaluation. *arXiv preprint arXiv:2507.20917*.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard hugging face. *Récupérée mai*, 24:2024.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023a. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

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. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Clement Christophe, Tathagata Raha, Svetlana Maslenkova, Muhammad Umar Salman, Praveenkumar Kanithi, Marco AF Pimentel, and Shadab Khan. 2024. Beyond fine-tuning: Unleashing the potential of continuous pretraining for clinical LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10549–10561, Miami, Florida, USA. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. Medical mt5: An open-source multilingual text-to-text llm for the medical domain. *Preprint*, arXiv:2404.07613.

Aryo Gema, Pasquale Minervini, Luke Daines, Tom Hope, and Beatrice Alex. 2024. Parameter-efficient fine-tuning of LLaMA for the clinical domain. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 91–104, Mexico City, Mexico. Association for Computational Linguistics.

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. *Preprint*, arXiv:2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yining Huang, Keke Tang, Meilian Chen, and Boyuan Wang. 2024. A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv preprint arXiv:2404.15777*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024a. 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.

Daniel P. Jeong, Pranav Mani, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. 2024b. The limited impact of medical adaptation of large language and vision-language models. *Preprint*, arXiv:2411.08870.

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. *Preprint*, arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Zakaria Kaddari and Toumi Bouchentouf. 2022. Frbmedqa: the first french biomedical question answering dataset. *IAES International Journal of Artificial Intelligence*, 11(4):1588.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. In *LOUHI 2022*, Abou Dhabi, United Arab Emirates.

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. *Preprint*, arXiv:2304.00958.

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

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024b. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Anran Li, Lingfei Qian, Mengmeng Du, Yu Yin, Yan Hu, Zihao Sun, Yihang Fu, Erica Stutz, Xuguang Ai, Qianqian Xie, and 1 others. 2025. Memorization in large language models in medicine: Prevalence, characteristics, and implications. *arXiv preprint arXiv:2509.08604*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wei Lu, Rachel K Luu, and Markus J Buehler. 2025. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(1):84.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-QA: A real-world medical Q&A benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 277–294, Bangkok, Thailand. Association for Computational Linguistics.

Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. 2024. DoRA: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11662–11675, Bangkok, Thailand. Association for Computational Linguistics.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix: Bilingual medical mixture of experts llm. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 16984–17002. Association for Computational Linguistics.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. A comprehensive survey of contamination detection methods in large language models. *arXiv preprint arXiv:2404.00699*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Gemma Team, 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 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024a. Apollo: An lightweight multilingual medical llm towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

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*, 31(9):1833–1843.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. *arXiv preprint arXiv:2509.12405*.

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

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274, Miami, Florida, USA. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A CPT Training Corpus : NACHOS Description

The NACHOS corpus is a French medical open-source dataset compiled through extensive web crawling and text collection. While the full corpus spans 7.4 GB of data and contains over one billion words sourced from 24 French-speaking high-quality websites (Labrak et al., 2023), we use in this work its *small* variant, NACHOS$_{small}$. This version consists of approximately 4 GB of data and was obtained by shuffling the full corpus and randomly selecting 25.3 million sentences to ensure homogeneous coverage of data sources.

Note: Full details of the corpus compilation and processing are available in the original paper (Labrak et al., 2023).

A.1 Corpus Composition

The NACHOS corpus encompasses a diverse range of medical textual sources, including:

- Descriptions of diseases and conditions
- Treatment and medication information
- General health-related advice
- Official scientific meeting reports
- Anonymized clinical cases
- Scientific literature
- Theses
- French translation pairs
- University health courses

A.2 Data Sources

The corpus integrates data from multiple sources, with the most significant contributions coming from:

- HAL (638,508,261 words)
- Haute Autorité de Santé (HAS) (113,394,539 words)
- Drug leaflets (74,770,229 words)
- Medical Websites Scraping (60,561,495 words)
- ANSES SAISINE (51,372,932 words)
- Public Drug Database (BDPM) (48,302,695 words)

A.3 Corpus Preparation

The researchers employed several preprocessing steps:

1. Text collection through web scraping, raw textual sources, and optical character recognition (OCR)
2. Sentence splitting using heuristic methods
3. Aggressive filtering to remove short or low-quality sentences
4. Language classification using a custom classifier trained on multilingual corpora

B SFT Training Corpus : MedInjection-Fr Description

B.1 Overview

MedInjection-FR[4] is a large-scale French biomedical instruction dataset composed of native, translated, and synthetic instruction–response pairs. The dataset comprises 571 436 examples spanning MCQUs, MCQs, and OEQAs. The dataset is publicly available; ***however, its primary scholarly reference is not cited in this paper in order to preserve author anonymity during peer review***.

B.2 Data Composition

The dataset consists of 77 247 native examples, 417 674 translated examples, and 76 506 synthetic examples. All data are formatted as instruction–response pairs and normalized to a unified schema, ensuring consistency across heterogeneous sources and supervision types.

B.3 Quality Control for Translated Data

The translated subset was obtained by translating English biomedical instruction datasets into French using two LLMs: `GPT-4o-mini` (Hurst et al., 2024) and `Gemini 2.0 Flash` [5]. Translation quality was evaluated on the WMT 2024 Biomedical Translation Task benchmark (Neves et al., 2024) using BLEU and COMET metrics. `GPT-4o-mini` achieved a BLEU score of 51.01 and a COMET score of 0.8751, while `Gemini 2.0 Flash` achieved a BLEU score of 53.72 and a COMET score of 0.8783. These results are comparable to the best-performing system reported in the shared

[4] https://huggingface.co/spaces/MedInjection-FR/README

[5] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

task (BLEU 53.54, COMET 0.8760), suggesting high semantic fidelity and robust preservation of biomedical terminology in the translated subset.

B.4 Quality Control for Synthetic Data

The synthetic subset was generated using `GPT-4o` from source documents including clinical cases and biomedical abstracts. Each source document was used to generate multiple instructional tasks covering a broad range of biomedical reasoning, such as clinical summarization, factual QA, diagnostic reasoning, treatment suggestion, and classification.

To control generation quality, each synthetic instruction–response pair was evaluated using four independent large language models acting as automatic judges: `GPT-4.1-mini` [6], `Gemini 2.0 Flash`, `MedGemma-27B` (Sellergren et al., 2025), and `Qwen3-Next-80B-A3B-Instruct` (Team, 2025). For MCQAs, evaluators assigned scores on a three-point scale reflecting answer correctness and contextual coherence. For OEQAs, a five-point scale was used to capture varying degrees of factual accuracy and completeness. Only examples meeting predefined minimum quality thresholds across evaluators were retained in the final dataset.

C CPT hyperparameters

Parameter	Value
Learning rate	2e-05 (1e-04 for gemma family)
Train batch size	2 (4 for gema family)
Seed	42
Gradient accumulation steps	2 (16 for gemma family)
Optimizer	AdamW
Weight Decay	0.01
Scheduler	Cosine
Number of epochs	3

Table 2: Hyperparameters used in CPT training

[6] https://openai.com/index/gpt-4-1/

D SFT hyperparameters

Parameter	Value
Rank	16
LoRA Aplha	16
LoRA Dropout	0.05
use_dora	True
Learning rate	2e-05 (1e-04 for gemma family)
Train batch size	4
Evaluation batch size	train_batch_size * 2
Seed	42
WarmUp_ratio	0.05
Gradient accumulation steps	8
Optimizer	AdamW
Scheduler	Cosine
Number of epochs	10
Target Modules	QKVOGUD

Table 3: Hyperparameters used in SFT training

E Evaluation Metrics

We provide here the formal definitions of the evaluation metrics used for MCQU and MCQ evaluation.

Exact Match (EM). Exact Match measures the proportion of predictions that exactly match the gold answer:

$$\text{EM} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i],$$

where N denotes the number of questions, y_i the gold answer, $\hat{y}_i$ the model prediction, and $[\cdot]$ is the indicator function.

Hamming Score. For multi-answer MCQ, we additionally report the Hamming score, which captures partial agreement between predicted and reference label sets:

$$\text{Hamming Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}.$$

This metric rewards partial correctness and is therefore better suited for evaluating multi-label predictions than Exact Match alone.

F Evaluation Benchmarks

The adapted models are evaluated against their corresponding base models using benchmark datasets drawn from the *test split* of MedInjection-FR. The evaluation suite includes both *native French* benchmarks and *translated English* benchmarks. For the translated benchmarks, English test sets were translated into French following the procedure described in section B.3.

The benchmarks cover multiple task formats, including MCQU, MCQ and OEQA. This setup enables a controlled comparison of adaptation effects across both discriminative and generative biomedical reasoning tasks. Table 4 summarizes the datasets used for evaluation and their respective sizes.

Dataset	# Items	Task
NATIVE		
	3 384	MCQ
MediQAl (Bazoge, 2025)	4 343	MCQU
	4 969	OEQA
FrenchMedMCQA (Labrak et al., 2022)	622	MCQ
mlabonne/medical-mcqa-fr[7]	150	MCQ
mlabonne/medical-cases-fr[8]	352	MCQ
FrBMedQA (Kaddari and Bouchentouf, 2022)	187	MCQU
S-Editions[9]	343	OEQA
	183	MCQ
TRANSLATED		
MedQA_4options (Jin et al., 2021)	1 273	MCQU
MedQA_5options (Jin et al., 2021)	1 273	MCQU
PubMedQA (Jin et al., 2019)	500	MCQU
MedMCQA (Pal et al., 2022)	4 183	MCQU
MMLU (Hendrycks et al., 2021)	1 080	MCQU
K-QA (Manes et al., 2024)	201	OEQA
MMLU-PRO (Wang et al., 2024b)	2 333	MCQU
MedXpertQA (Zuo et al., 2025)	2 450	MCQU

Table 4: Evaluation benchmarks used to compare adapted models with their base counterparts. All datasets correspond to the test split of MedInjection-FR.

G Prompt Templates

Overview. We use a unified instruction format across all task types, both for supervised fine-tuning and for zero-shot evaluation. When available, we rely on the native chat templates provided by instruction-tuned models; otherwise, prompts are formatted as plain-text instruction–response pairs.

Shared Structure. All prompts begin with a high-level medical instruction, optionally followed by a contextual passage. The core components are:

1. an instruction describing the task,
2. the question (and answer options when applicable),
3. an optional context section, and
4. a response header indicating where the model output should begin.

Task-Specific Constraints. The only variation across task types lies in the expected response format, which is explicitly stated in the instruction. Table 5 summarizes the templates used for each task.

Canonical Prompt Format. The following abstract template illustrates the prompt structure shared across all tasks:

System prompt (training and evaluation) for MCQ

```
Lire l'instruction médicale suivante et
    fournir une réponse adaptée à la
    situation décrite.
Répondre uniquement avec la lettre
    correspondant à la ou les bonnes
    réponses séparées par des virgules.
    Exemple : A, C, D.
```

System prompt (training and evaluation) for MCQU

```
Lire l'instruction médicale suivante et
    fournir une réponse adaptée à la
    situation décrite.
Répondre uniquement avec la lettre
    correspondant à la bonne réponse.
    Exemple : A.
```

System prompt (training and evaluation) for OEQA

```
Lire l'instruction médicale suivante et
    fournir une réponse adaptée à la
    situation décrite.
```

User prompt (task-dependent)

```
### Instruction:
[Question (+ options for MCQ tasks)]

### Contexte:
[Context, if available]

### Réponse:
```

Chat-Based Formatting. For instruction-tuned models providing an explicit chat interface, the same content is mapped to role-based messages as follows:

- **System**: high-level medical instruction (shared across tasks),
- **User**: task instruction, question, and optional context,
- **Assistant**: model-generated answer.

This formulation ensures consistent supervision and evaluation across models with different input formatting requirements.

Task	Instruction Constraint	Expected Output
MCQU	Respond only with the letter corresponding to the single correct answer.	Single letter (e.g., `A`)
MCQ	Respond only with the letters corresponding to all correct answers, separated by commas.	Comma-separated letters (e.g., `A, C, D`)
OEQA	Provide a free-form medical answer based on the instruction and context.	Unconstrained text

Table 5: Summary of task-specific prompt templates and output constraints.

H LLM-as-a-Judge for OEQA

H.1 Overview

This appendix summarizes an evaluation study originally conducted and reported in a separate paper, which we reuse in the present work to motivate the choice of an LLM-based evaluator for open-ended QA (OEQA). ***The citation to the original study is intentionally omitted in this submission in order to preserve author anonymity during peer review***.

H.2 Expert-Annotated Evaluation Set

The evaluation set consists of 500 OEQA instances. We first selected 100 medical questions together with their reference answers from a native French biomedical QA source. For each question, five candidate answers were generated using five different answer-generating LLMs, resulting in a total of 500 model-generated responses. These answers cover a wide range of styles, levels of verbosity, and degrees of medical detail.

All generated answers were annotated by a board-certified physician specialized in neurovascular medicine, with five years of senior clinical experience. Each answer was labeled according to semantic equivalence with the reference answer using a binary decision criterion. To assess annotation consistency, a second clinician independently annotated a random subset of 10 instances. The two annotators agreed on 9 out of 10 cases (90% agreement).

H.3 Candidate LLM Judges

A diverse set of LLMs was evaluated as automatic judges in the original study, including both proprietary and open-source models, as well as general-purpose and medically adapted architectures. The evaluated judges include GPT-5.1[10], Gemini-2.5-Pro (Comanici et al., 2025), Qwen3-Next-80B-A3B-Instruct (Team, 2025), MedGemma-27B (Sellergren et al., 2025), and M-Prometheus-14B (Kim et al., 2024). This selection enables analysis of the impact of model scale, domain specialization, and training objectives on alignment with expert medical judgment.

[10] https://openai.com/index/gpt-5-1/

H.4 Judging Protocol

All candidate judges were evaluated using a strict equivalence-based prompting strategy aligned with the human annotation protocol. Each judge was provided with the medical question, the reference answer, and a model-generated answer, and was required to output a binary decision indicating whether the generated answer was semantically equivalent to the reference. Using an identical formulation for human and automated evaluation ensures conceptual consistency and allows agreement metrics to be interpreted directly.

H.5 Agreement Metrics and Bias Analysis

Judge performance was measured using Pearson correlation, precision, recall, and F1 score with respect to expert annotations. In addition to overall agreement, performance was analyzed separately for each answer-generating model, enabling identification of generator-dependent biases. The analysis shows that none of the evaluated judges is fully invariant to the source of answer generation, with noticeable variations in precision–recall trade-offs across generators.

H.6 Selection of the Primary Evaluator

		Pearson_r	F1
Judges	**GPT5.1**	38,27	45,07
	GEMINI-2.5-PRO	38,81	52,59
	QWEN-80b	**44,77**	60,00
	MEDGEMMA-27B	40,67	**60,50**
	M-PROMETHEUS	34,88	48,56
Metrics	**ROUGE-L**	25,40	-
	METEOR	25.04	-
	BLEU	17,16	-
	BERTScore_F1	14,88	-

Table 6: Agreement between human judgments and automatic evaluation methods for OEQA.

Although Qwen-80B achieved the highest Pearson correlation with expert judgments in the original study, MedGemma-27B consistently obtained the strongest F1 scores (Table 6) and exhibited more stable performance across answer generators. Importantly, MedGemma-27B did not show systematic self-preference toward answers generated by models from the same family and demonstrated reduced sensitivity to stylistic factors such as verbosity. Based on these findings, MedGemma-27B was selected in the original study as the primary LLM-based evaluator for OEQA, and we adopt the same evaluator in the present work.

H.7 Summary

This appendix documents the rationale for selecting MedGemma-27B as an LLM-based evaluator for OEQA, based on an independent evaluation study conducted in prior work.

I MCQA and OEQA Results

I.1 MCQA Greedy Decoding

Table 7 reports performance on MCQA across the studied model families (Gemma-4B, Mistral-7B, Llama-7B, and Llama-13B), three initialization types (General, Instruct, Medical), and four adaptation strategies (Base, CPT, SFT, CPT+SFT). Results are shown for both standard multiple-answer MCQs (MCQ) and single-answer MCQs (MCQU), using Exact Match (EM), Hamming score for MCQ, and aggregated EM. The reported results here are obtained using greedy decoding.

Effectiveness of Adaptation Strategy: Under greedy decoding, SFT clearly dominates all other adaptation strategies across MCQ, MCQU, and aggregated metrics. Unlike constrained decoding, where CPT+SFT often ranks first, greedy decoding exposes a much sharper separation between strategies:

$$\text{BASE} \ll \text{CPT} \ll \text{CPT+SFT} < \text{SFT}$$

Across nearly all model families and initializations, SFT yields the highest MCQ EM, MCQ Hamming, MCQU EM, and aggregated EM. This trend is particularly strong for instruction-tuned models (Mistral, Llama-7B, Llama-13B), where SFT consistently delivers large absolute gains, often by wide margins. As in the constrained decoding setting, when CPT+SFT outperforms SFT, the performance gap is generally smaller than in configurations where SFT outperforms CPT+SFT.

CPT alone remains unstable under greedy decoding. While it sometimes improves over the base model, it frequently underperforms SFT and can even degrade MCQU and aggregated scores. Importantly, CPT+SFT does not systematically improve over SFT in greedy decoding and often performs worse, indicating that the benefits of CPT are largely redundant once task supervision is introduced and decoding constraints are removed.

Overall, greedy decoding amplifies the advantages of task-aligned supervision, making SFT the best adaptation strategy when decoding is unconstrained.

Impact of Model Initialization: Model initialization plays a stronger and more consistent role under greedy decoding than under constrained decoding. Across all families and metrics, instruction-tuned models dominate. General models benefit from SFT but remain consistently below instruction-tuned counterparts. Medical models, while improving with SFT, never achieve the best greedy decoding performance, confirming that domain pretraining alone is insufficient without strong instruction alignment.

This contrasts with constrained decoding, where general and medical models occasionally remain competitive. Under greedy decoding, instruction tuning becomes a necessary condition for strong performance.

MCQA Greedy Decoding Guidelines: For greedy decoding in medical MCQA, the optimal configuration is to start from an instruction-tuned

Model Type	Strategy	MCQ		MCQU	Aggregation	OEQA	
		EM	Hamming	EM	EM	*BLEU*	*METEOR*
Gemma-4B							
GENERAL	Base	0.56	5.15	5.88	3.22	0.92	8.09
	CPT	0.03	0.68	8.53	4.28	**1.68**	**8.35**
	SFT	**1.73**	**19.48**	**19.81**	**10.77**	1.03	7.42
	CPT+SFT	1.67	15.34	19.52	10.60	0.56	6.99
INSTRUCT	Base	**6.81**	**40.75**	28.88	**17.85**	**0.76**	**10.41**
	CPT	0.07	1.72	1.37	0.72	0.46	5.89
	SFT	1.38	10.17	**31.87**	16.63	0.09	5.32
	CPT+SFT	1.22	8.61	30.66	15.94	0.21	6.93
MEDICAL	Base	1.22	11.28	11.47	6.34	0.76	**10.41**
	CPT	1.62	**18.11**	11.18	6.40	**1.04**	5.26
	SFT	1.15	11.51	**17.91**	9.53	0.56	6.98
	CPT+SFT	**1.67**	16.54	17.63	**9.65**	0.34	5.60
Mistral-7B							
GENERAL	Base	0.15	2.90	4.04	2.09	0.66	7.32
	CPT	0.44	3.75	13.71	7.07	0.99	7.69
	SFT	**1.79**	**20.62**	**19.88**	**10.84**	**1.04**	7.59
	CPT+SFT	1.42	17.95	19.57	10.49	0.63	**8.66**
INSTRUCT	Base	3.42	26.94	21.46	12.44	1.12	7.70
	CPT	4.10	29.16	27.63	15.87	**2.34**	**10.90**
	SFT	**11.94**	**47.39**	**31.52**	**21.73**	1.65	7.08
	CPT+SFT	1.85	17.56	29.84	15.85	1.09	9.46
MEDICAL	Base	2.24	19.09	13.39	7.81	1.73	8.39
	CPT	**2.25**	18.39	12.19	7.22	**1.93**	**9.54**
	SFT	1.95	**20.68**	18.47	10.21	1.06	6.75
	CPT+SFT	1.44	16.43	**19.33**	**10.38**	1.05	8.02
Llama-7B							
GENERAL	Base	0.17	2.69	9.46	4.82	0.49	5.91
	CPT	1.13	10.16	5.83	3.48	**1.39**	5.90
	SFT	**1.82**	**24.64**	16.08	8.95	0.51	**7.26**
	CPT+SFT	1.61	17.50	**17.11**	**9.36**	0.52	6.75
INSTRUCT	Base	0.03	0.64	0.04	0.03	0.40	2.51
	CPT	4.92	40.87	0.04	2.48	**1.72**	8.64
	SFT	**7.72**	**42.70**	**29.26**	**18.49**	1.42	6.22
	CPT+SFT	1.03	6.36	0.07	0.55	1.41	**10.07**
MEDICAL	Base	0.14	1.98	0.57	0.35	0.51	**7.06**
	CPT	**2.11**	**24.01**	12.20	7.16	**1.12**	5.69
	SFT	1.12	20.40	17.49	9.30	0.40	5.50
	CPT+SFT	1.67	15.87	**18.29**	**9.98**	0.49	6.70
Llama-13B							
GENERAL	Base	0.29	0.84	11.77	6.03	0.19	1.98
	CPT	**2.81**	21.03	10.49	6.65	**1.85**	7.85
	SFT	2.18	17.08	18.67	10.42	0.42	5.85
	CPT+SFT	1.65	**21.56**	**19.84**	**10.74**	0.59	**8.07**
INSTRUCT	Base	0.00	4.87	0.04	0.02	0.50	3.85
	CPT	4.82	34.09	0.04	2.43	**2.09**	**10.78**
	SFT	10.92	42.98	30.13	20.53	1.52	6.52
	CPT+SFT	**12.57**	**44.02**	**31.48**	**22.02**	1.74	7.32
MEDICAL	Base	0.56	9.41	11.01	5.78	0.48	5.60
	CPT	2.02	**22.25**	11.28	6.65	**1.12**	5.69
	SFT	**2.07**	21.50	18.11	10.09	0.59	**7.37**
	CPT+SFT	1.61	17.64	**19.75**	**10.68**	0.67	6.53

Table 7: Greedy decoding results for MCQ and MCQU and BLEU/METEOR scores for OEQA across model families and adaptation strategies. **Bold** denotes the best strategy and underlining the best initialization.

model and apply SFT only. CPT and CPT+SFT offer no consistent benefit in this setting and can be safely avoided unless constrained decoding is explicitly required.

I.2 OEQA Overlap-based Evaluation

The right part of Table 7 reports the overlap-based metrics BLEU and METEOR. Both metrics exhibit trends consistent with ROUGE-L, with improvements primarily driven by CPT. In a few isolated cases, CPT+SFT yields additional gains on METEOR, but with small differences with when compared with CPT. Regarding model initialization, BLEU and METEOR consistently favor instruction-tuned models as the strongest starting point.

J Statistical Significance

We assess whether observed differences between adaptation strategies and initialization choices are statistically significant using paired bootstrap significance testing. For each comparison, we compute the per-instance score difference (EM for MCQ/MCQU; judge-based correctness for OEQA) and report a two-sided p-value (`p_two_sided`). Statistical significance is determined by comparing this p-value against a predefined threshold α. We report results using a Bonferroni-corrected threshold to control for multiple comparisons.

For comparisons between adaptation strategies within a model family, we perform 12 pairwise tests per family, yielding a corrected threshold of $\alpha_{\text{Bonferroni}} = 0.05/12$. For comparisons between model initialization types under a fixed adaptation strategy, we perform 9 pairwise tests per family, yielding $\alpha_{\text{Bonferroni}} = 0.05/9$. The applied threshold (`alpha_Bonferoni`) and the resulting significance decision (`significant_Bonferroni`) are reported explicitly in Tables 8 and 9.

We define the mean difference as $\Delta = \text{score}(\texttt{model\_a}) - \text{score}(\texttt{model\_b})$ (not shown in the tables). Therefore, if the confidence interval is entirely above zero, `model_a` performs better; if it is entirely below zero, `model_b` performs better. A comparison is considered statistically significant if the corrected decision is TRUE.

J.1 Interpretation of comparison IDs

Each row in Tables 8 and 9 corresponds to a specific pairwise comparison between two models (`model_a` vs. `model_b`). The `id` field encodes the purpose of the comparison: (i) **IDs A–C** compare models *within the same model type* (GENERAL, INSTRUCT, or MEDICAL) in order to quantify the effect of adaptation strategies (CPT, SFT, CPT+SFT) relative to a fixed initialization. (ii) **IDs D–F** compare models *across model types* under a fixed adaptation strategy, in order to identify the most effective initialization point (GENERAL vs. INSTRUCT vs. MEDICAL) for downstream adaptation.

J.2 Decoding conditions

For MCQ/MCQU, Table 8 reports significance results separately for greedy and constrained decoding. For OEQA, Table 9 reports strategy-level comparisons under the evaluation setting used for the main experiments.

K Near-Miss Rates in MCQA

Model	MCQ	MCQU
Mistral	0.203	0.202
Mistral-CPT	0.212	0.203
Mistral-SFT	0.234	0.204
Mistral-CPT-SFT	0.256	0.203
Mistral-Instruct	0.173	0.202
Mistral-Instruct-CPT	0.205	0.202
Mistral-Instruct-SFT	0.174	0.206
Mistral-Instruct-CPT-SFT	0.221	0.192
BioMistral	0.181	0.205
BioMistral-CPT	0.195	0.201
BioMistral-SFT	0.220	0.211
BioMistral-CPT-SFT	0.234	0.204

Table 10: Near-miss rates for MCQ and MCQU across Mistral variants. A near-miss corresponds to cases where all gold answers are ranked within the top-k options but the generated answer does not match the gold label(s). Near-miss rates remain stable across model families and adaptation strategies, indicating that improvements in confidence and ranking do not directly translate into exact prediction.

To better characterize model behavior beyond EM accuracy in MCQA, we analyze *near-miss* predictions. Table 10 shows the near-miss rates obtained for MCQ and MCQU across different Mistral-based model variants and adaptation strategies.

L OEQA Evaluation: Verbosity Bias

To investigate verbosity bias in OEQA, we compute descriptive statistics of generated answer lengths across all models. Table 11 reports mean, median, and standard deviation of word and character counts over greedy OEQA outputs.

	id	strategy	model_a	model_b	Decoding type	ci95_low	ci95_high	p_two_sided	alpha_Bonferroni	significant_Bonferroni
Gemma 4B										
GENERAL	A1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-pt	constrained	-2.60E-02	-3.48E-03	2.00E-04	4.17E-03	**TRUE**
	A1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-pt	greedy	-6.03E-03	2.71E-02	1.75E-01	4.17E-03	FALSE
	A2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt	constrained	1.57E-02	5.29E-02	4.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt	greedy	6.03E-02	9.37E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt-CPT	constrained	2.34E-02	7.52E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt-CPT	greedy	4.06E-02	9.63E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-pt	constrained	1.32E-02	4.90E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-pt	greedy	6.17E-02	9.54E-02	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	gemma-3-4b-it-CPT	gemma-3-4b-it	constrained	-4.69E-02	-8.61E-03	4.00E-03	4.17E-03	**TRUE**
	B1	CPT	gemma-3-4b-it-CPT	gemma-3-4b-it	greedy	-1.94E-01	-1.47E-01	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it	constrained	-1.25E-02	1.41E-02	9.27E-01	4.17E-03	FALSE
	B2	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it	greedy	-3.54E-02	-2.16E-03	2.66E-02	4.17E-03	FALSE
	B3	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it-CPT	constrained	1.49E-02	4.29E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it-CPT	greedy	1.26E-01	1.78E-01	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	gemma-3-4b-it-SFT	gemma-3-4b-it	constrained	-5.48E-03	2.13E-02	2.77E-01	4.17E-03	FALSE
	B4	SFT	gemma-3-4b-it-SFT	gemma-3-4b-it	greedy	-2.62E-02	3.58E-03	1.23E-01	4.17E-03	FALSE
MEDICAL	C1	CPT	medgemma-4b-pt-CPT	medgemma-4b-pt	constrained	-1.68E-02	1.67E-04	5.60E-02	4.17E-03	FALSE
	C1	CPT	medgemma-4b-pt-CPT	medgemma-4b-pt	greedy	-4.32E-03	9.75E-03	4.60E-01	4.17E-03	FALSE
	C2	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt-CPT	constrained	2.64E-02	4.73E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt-CPT	greedy	2.44E-02	3.98E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt	constrained	1.72E-02	4.13E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt	greedy	2.80E-02	4.25E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	medgemma-4b-pt-SFT	medgemma-4b-pt	constrained	1.74E-02	4.04E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	medgemma-4b-pt-SFT	medgemma-4b-pt	greedy	2.58E-02	4.14E-02	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-it-SFT	constrained	-7.13E-03	1.15E-02	6.83E-01	5.56E-03	FALSE
	D1	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-it-SFT	greedy	-6.92E-02	-4.32E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	gemma-3-4b-pt-SFT	medgemma-4b-pt-SFT	constrained	2.60E-03	1.27E-02	4.80E-03	5.56E-03	**TRUE**
	D2	SFT	gemma-3-4b-pt-SFT	medgemma-4b-pt-SFT	greedy	8.83E-03	1.61E-02	2.00E-04	5.56E-03	**TRUE**
	D3	SFT	gemma-3-4b-it-SFT	medgemma-4b-pt-SFT	constrained	-5.57E-03	1.57E-02	3.13E-01	5.56E-03	FALSE
	D3	SFT	gemma-3-4b-it-SFT	medgemma-4b-pt-SFT	greedy	5.63E-02	8.12E-02	2.00E-04	5.56E-03	**TRUE**
CPT	E1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-it-CPT	constrained	-3.76E-02	2.01E-02	7.90E-01	5.56E-03	FALSE
	E1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-it-CPT	greedy	2.58E-02	4.95E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	gemma-3-4b-pt-CPT	medgemma-4b-pt-CPT	constrained	-2.87E-02	2.19E-02	8.35E-01	5.56E-03	FALSE
	E2	CPT	gemma-3-4b-pt-CPT	medgemma-4b-pt-CPT	greedy	-5.66E-02	1.81E-03	8.56E-02	5.56E-03	FALSE
	E3	CPT	gemma-3-4b-it-CPT	medgemma-4b-pt-CPT	constrained	-2.21E-03	1.59E-02	1.83E-01	5.56E-03	FALSE
	E3	CPT	gemma-3-4b-it-CPT	medgemma-4b-pt-CPT	greedy	-9.23E-02	-3.78E-02	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-it-CPT-SFT	constrained	2.15E-03	2.27E-02	1.92E-02	5.56E-03	FALSE
	F1	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-it-CPT-SFT	greedy	-6.81E-02	-3.28E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	medgemma-4b-pt-CPT-SFT	constrained	2.64E-03	1.80E-02	1.10E-02	5.56E-03	FALSE
	F2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	medgemma-4b-pt-CPT-SFT	greedy	3.47E-03	1.52E-02	1.60E-03	5.56E-03	**TRUE**
	F3	CPT+SFT	gemma-3-4b-it-CPT-SFT	medgemma-4b-pt-CPT-SFT	constrained	-1.27E-02	9.11E-03	7.38E-01	5.56E-03	FALSE
	F3	CPT+SFT	gemma-3-4b-it-CPT-SFT	medgemma-4b-pt-CPT-SFT	greedy	3.84E-02	7.95E-02	2.00E-04	5.56E-03	**TRUE**
Mistral 7B										
GENERAL	A1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-v0.1	constrained	1.63E-03	2.19E-02	2.26E-02	4.17E-03	FALSE
	A1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-v0.1	greedy	3.64E-02	6.00E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1	constrained	3.71E-02	6.85E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1	greedy	6.72E-02	1.01E-01	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1-CPT	constrained	2.80E-02	5.41E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1-CPT	greedy	2.82E-02	4.38E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-v0.1	constrained	3.25E-02	6.88E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-v0.1	greedy	7.41E-02	1.06E-01	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	Mistral-7B-Instruct-v0.1-CPT	Mistral-7B-Instruct-v0.1	constrained	1.76E-02	4.88E-02	2.00E-04	4.17E-03	**TRUE**
	B1	CPT	Mistral-7B-Instruct-v0.1-CPT	Mistral-7B-Instruct-v0.1	greedy	1.67E-02	5.01E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT	constrained	-2.10E-02	1.47E-02	7.64E-01	4.17E-03	FALSE
	B2	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT	greedy	-1.59E-02	1.76E-02	9.83E-01	4.17E-03	FALSE
	B3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1	constrained	1.29E-02	5.00E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1	greedy	1.68E-02	5.21E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Mistral-7B-Instruct-v0.1-SFT	Mistral-7B-Instruct-v0.1	constrained	2.41E-02	6.41E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Mistral-7B-Instruct-v0.1-SFT	Mistral-7B-Instruct-v0.1	greedy	6.71E-02	1.18E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	C1	CPT	BioMistral-7B-CPT	BioMistral-7B	constrained	-1.12E-02	1.00E-02	8.13E-01	4.17E-03	FALSE
	C1	CPT	BioMistral-7B-CPT	BioMistral-7B	greedy	-1.61E-02	2.51E-03	2.85E-01	4.17E-03	FALSE
	C2	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B	constrained	2.63E-02	6.50E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B	greedy	1.33E-02	4.54E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B-CPT	constrained	2.23E-02	7.24E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B-CPT	greedy	1.27E-02	6.02E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	BioMistral-7B-SFT	BioMistral-7B	constrained	1.33E-02	4.67E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	BioMistral-7B-SFT	BioMistral-7B	greedy	1.34E-02	4.18E-02	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-Instruct-v0.1-SFT	constrained	-1.89E-02	2.68E-02	7.28E-01	5.56E-03	FALSE
	D1	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-Instruct-v0.1-SFT	greedy	-1.32E-01	-8.16E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Mistral-7B-Instruct-v0.1-SFT	BioMistral-7B-SFT	constrained	-1.41E-02	4.27E-02	4.35E-01	5.56E-03	FALSE
	D2	SFT	Mistral-7B-Instruct-v0.1-SFT	BioMistral-7B-SFT	greedy	8.72E-02	1.40E-01	2.00E-04	5.56E-03	**TRUE**
	D3	SFT	Mistral-7B-v0.1-SFT	BioMistral-7B-SFT	constrained	5.20E-03	2.93E-02	2.00E-03	5.56E-03	**TRUE**
	D3	SFT	Mistral-7B-v0.1-SFT	BioMistral-7B-SFT	greedy	1.57E-03	1.27E-02	1.66E-02	5.56E-03	FALSE
CPT	E1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-Instruct-v0.1-CPT	constrained	-5.11E-02	-3.56E-04	4.62E-02	5.56E-03	FALSE
	E1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-Instruct-v0.1-CPT	greedy	-1.07E-01	-6.53E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Mistral-7B-v0.1-CPT	BioMistral-7B-CPT	constrained	-9.26E-03	2.57E-02	6.18E-01	5.56E-03	FALSE
	E2	CPT	Mistral-7B-v0.1-CPT	BioMistral-7B-CPT	greedy	-1.44E-02	1.70E-02	6.45E-01	5.56E-03	FALSE
	E3	CPT	Mistral-7B-Instruct-v0.1-CPT	BioMistral-7B-CPT	constrained	4.02E-03	5.94E-02	1.80E-02	5.56E-03	FALSE
	E3	CPT	Mistral-7B-Instruct-v0.1-CPT	BioMistral-7B-CPT	greedy	6.80E-02	1.01E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	constrained	7.69E-03	3.25E-02	1.20E-03	5.56E-03	**TRUE**
	F1	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	greedy	-6.20E-02	-4.21E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	constrained	-9.97E-03	1.59E-02	5.02E-01	5.56E-03	FALSE
	F2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	greedy	-1.08E-02	9.42E-03	6.69E-01	5.56E-03	FALSE
	F3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	constrained	-3.06E-02	-1.28E-03	3.42E-02	5.56E-03	FALSE
	F3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	greedy	3.56E-02	6.81E-02	2.00E-04	5.56E-03	**TRUE**
Llama 7B										
GENERAL	A1	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-hf	constrained	-8.99E-03	4.70E-03	6.17E-01	4.17E-03	FALSE
	A1	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-hf	greedy	-2.29E-02	-4.41E-03	3.00E-03	4.17E-03	**TRUE**
	A2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf-CPT	constrained	2.06E-02	4.76E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf-CPT	greedy	4.28E-02	8.52E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf	constrained	1.81E-02	4.46E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf	greedy	3.37E-02	6.48E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-hf	constrained	1.58E-02	3.86E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-hf	greedy	3.08E-02	6.12E-02	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	Llama-2-7b-chat-hf-CPT	Llama-2-7b-chat-hf	constrained	-7.23E-03	4.04E-03	8.20E-01	4.17E-03	FALSE
	B1	CPT	Llama-2-7b-chat-hf-CPT	Llama-2-7b-chat-hf	greedy	1.86E-02	3.22E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf	constrained	-1.32E-02	2.43E-03	2.78E-01	4.17E-03	FALSE
	B2	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf	greedy	2.64E-03	7.96E-03	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT	constrained	-7.62E-03	2.54E-04	7.10E-02	4.17E-03	FALSE
	B3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT	greedy	-2.69E-02	-1.17E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-7b-chat-hf-SFT	Llama-2-7b-chat-hf	constrained	9.28E-03	4.49E-02	2.00E-03	4.17E-03	**TRUE**
	B4	SFT	Llama-2-7b-chat-hf-SFT	Llama-2-7b-chat-hf	greedy	1.54E-01	2.19E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	C1	CPT	meditron-7b-CPT	meditron-7b	constrained	1.19E-02	2.70E-02	2.00E-04	4.17E-03	**TRUE**
	C1	CPT	meditron-7b-CPT	meditron-7b	greedy	4.73E-02	1.05E-01	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b	constrained	3.78E-02	7.33E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b	greedy	6.80E-02	1.47E-01	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b-CPT	constrained	2.20E-02	5.08E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b-CPT	greedy	1.87E-02	4.21E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	meditron-7b-SFT	meditron-7b	constrained	3.31E-02	6.48E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	meditron-7b-SFT	meditron-7b	greedy	6.23E-02	1.36E-01	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-chat-hf-SFT	constrained	-3.69E-02	1.03E-02	4.19E-01	5.56E-03	FALSE
	D1	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-chat-hf-SFT	greedy	-1.22E-01	-6.88E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	LLama-2-7b-hf-SFT	meditron-7b-SFT	constrained	-1.39E-02	-3.59E-03	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	LLama-2-7b-hf-SFT	meditron-7b-SFT	greedy	-8.30E-03	1.95E-03	2.48E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-7b-chat-hf-SFT	meditron-7b-SFT	constrained	-1.92E-02	2.74E-02	9.14E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-7b-chat-hf-SFT	meditron-7b-SFT	greedy	6.34E-02	1.18E-01	2.00E-04	5.56E-03	**TRUE**
CPT	E1	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-chat-hf-CPT	constrained	-2.48E-02	1.31E-03	8.28E-02	5.56E-03	FALSE
	E1	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-chat-hf-CPT	greedy	-3.69E-03	3.09E-02	1.93E-01	5.56E-03	FALSE
	E2	CPT	Llama-2-7b-hf-CPT	meditron-7b-CPT	constrained	-1.69E-02	-1.09E-03	2.40E-02	5.56E-03	FALSE
	E2	CPT	Llama-2-7b-hf-CPT	meditron-7b-CPT	greedy	-5.63E-02	-2.58E-02	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	Llama-2-7b-chat-hf-CPT	meditron-7b-CPT	constrained	-6.40E-03	1.11E-02	6.02E-01	5.56E-03	FALSE
	E3	CPT	Llama-2-7b-chat-hf-CPT	meditron-7b-CPT	greedy	-8.45E-02	-2.63E-02	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	Llama-2-7b-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT-SFT	constrained	1.27E-02	3.61E-02	4.00E-04	5.56E-03	**TRUE**
	F1	CPT+SFT	Llama-2-7b-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT-SFT	greedy	6.09E-02	1.32E-01	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	meditron-7b-CPT-SFT	constrained	-1.83E-02	-4.73E-03	1.00E-03	5.56E-03	**TRUE**
	F2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	meditron-7b-CPT-SFT	greedy	-1.38E-02	-1.53E-03	8.80E-03	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	meditron-7b-CPT-SFT	constrained	-5.07E-02	-2.24E-02	2.00E-04	5.56E-03	**TRUE**
	F3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	meditron-7b-CPT-SFT	greedy	-1.44E-01	-6.56E-02	2.00E-04	5.56E-03	**TRUE**
Llama 13B										
GENERAL	A1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-hf	constrained	2.54E-03	2.02E-02	1.06E-02	4.17E-03	FALSE
	A1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-hf	greedy	3.67E-03	1.98E-02	4.40E-03	4.17E-03	FALSE
	A2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf	constrained	2.37E-02	5.45E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf	greedy	3.27E-02	5.90E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf-CPT	constrained	1.49E-02	3.87E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf-CPT	greedy	2.07E-02	4.80E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-hf	constrained	1.81E-02	4.28E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-hf	greedy	3.19E-02	5.31E-02	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	Llama-2-13b-chat-hf-CPT	Llama-2-13b-chat-hf	constrained	-1.18E-03	4.97E-02	6.72E-02	4.17E-03	FALSE
	B1	CPT	Llama-2-13b-chat-hf-CPT	Llama-2-13b-chat-hf	greedy	1.92E-02	3.14E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf	constrained	3.87E-02	9.30E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf	greedy	1.82E-01	2.56E-01	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT	constrained	2.54E-02	6.15E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT	greedy	1.58E-01	2.31E-01	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-13b-chat-hf-SFT	Llama-2-13b-chat-hf	constrained	3.11E-02	8.21E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-13b-chat-hf-SFT	Llama-2-13b-chat-hf	greedy	1.72E-01	2.41E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	C1	CPT	MedLLaMA-13B-CPT	MedLLaMA_13B	constrained	-2.15E-02	1.77E-02	7.73E-01	4.17E-03	FALSE
	C1	CPT	MedLLaMA-13B-CPT	MedLLaMA_13B	greedy	-9.55E-03	2.34E-02	2.68E-01	4.17E-03	FALSE
	C2	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA_13B	constrained	3.11E-02	5.74E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA_13B	greedy	3.56E-02	5.96E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA-13B-CPT	constrained	2.34E-02	7.11E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA-13B-CPT	greedy	1.90E-02	6.57E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	MedLLaMA-13B-SFT	MedLLaMA_13B	constrained	2.16E-02	4.90E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	MedLLaMA-13B-SFT	MedLLaMA_13B	greedy	2.84E-02	5.80E-02	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-chat-hf-SFT	constrained	-4.30E-02	7.32E-03	2.15E-01	5.56E-03	FALSE
	D1	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-chat-hf-SFT	greedy	-1.36E-01	-7.93E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Llama-2-13b-hf-SFT	MedLLaMA-13B-SFT	constrained	-3.65E-03	8.03E-03	4.70E-01	5.56E-03	FALSE
	D2	SFT	Llama-2-13b-hf-SFT	MedLLaMA-13B-SFT	greedy	-4.03E-03	7.99E-03	5.32E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-13b-chat-hf-SFT	MedLLaMA-13B-SFT	constrained	-5.69E-03	4.60E-02	1.47E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-13b-chat-hf-SFT	MedLLaMA-13B-SFT	greedy	7.93E-02	1.41E-01	2.00E-04	5.56E-03	**TRUE**
CPT	E1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-chat-hf-CPT	constrained	-1.87E-02	1.02E-02	8.78E-01	5.56E-03	FALSE
	E1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-chat-hf-CPT	greedy	2.31E-02	7.61E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Llama-2-13b-hf-CPT	MedLLaMA-13B-CPT	constrained	2.37E-03	3.99E-02	1.16E-02	5.56E-03	FALSE
	E2	CPT	Llama-2-13b-hf-CPT	MedLLaMA-13B-CPT	greedy	-7.36E-03	2.25E-02	7.03E-01	5.56E-03	FALSE
	E3	CPT	Llama-2-13b-chat-hf-CPT	MedLLaMA-13B-CPT	constrained	-5.92E-03	5.27E-02	2.03E-01	5.56E-03	FALSE
	E3	CPT	Llama-2-13b-chat-hf-CPT	MedLLaMA-13B-CPT	greedy	-5.61E-02	-2.58E-02	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT-SFT	constrained	-4.46E-02	4.48E-03	1.31E-01	5.56E-03	FALSE
	F1	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT-SFT	greedy	-1.48E-01	-8.37E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	constrained	-6.90E-03	8.26E-03	9.14E-01	5.56E-03	FALSE
	F2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	greedy	-7.39E-03	6.54E-03	7.66E-01	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	constrained	-6.57E-03	4.71E-02	1.56E-01	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	greedy	8.45E-02	1.47E-01	2.00E-04	5.56E-03	**TRUE**

Table 8: Significance testing for MCQ/MCQU comparisons. reported separately for greedy and constrained decoding. Each row reports a paired bootstrap test between `model_a` and `model_b`. including the 95% confidence interval of the mean EM difference. the two-sided p-value. and the Bonferroni-adjusted threshold with the resulting decision. IDs A–C compare adaptation strategies within the same model type; IDs D–F compare model initializations across types.

	id	strategy	model_a	model_b	ci95_low	ci95_high	p_two_sided	alpha_Bonferoni	significant_Bonferroni
Gemma 4B									
GENERAL	**A1**	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-pt	-5.05E-02	3.69E-02	5.14E-01	4.17E-03	FALSE
	A2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt	-2.25E-01	1.65E-01	8.77E-01	4.17E-03	FALSE
	A3	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt-CPT	-1.74E-01	1.31E-01	8.73E-01	4.17E-03	FALSE
	A4	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-pt	-2.11E-01	1.41E-01	8.06E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	gemma-3-4b-it-CPT	gemma-3-4b-it	-4.58E-01	-2.75E-01	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it	-4.58E-01	-1.41E-01	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it-CPT	-4.94E-03	1.41E-01	6.44E-02	4.17E-03	FALSE
	B4	SFT	gemma-3-4b-it-SFT	gemma-3-4b-it	-3.81E-01	-2.06E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	**C1**	CPT	medgemma-4b-pt-CPT	medgemma-4b-pt	-3.46E-01	-9.28E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt-CPT	3.71E-02	1.63E-01	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt	-2.79E-01	3.51E-02	1.20E-01	4.17E-03	FALSE
	C4	SFT	medgemma-4b-pt-SFT	medgemma-4b-pt	-1.88E-01	-1.94E-04	3.96E-02	4.17E-03	FALSE
SFT	**D1**	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-it-SFT	-3.40E-02	6.17E-02	7.03E-01	5.56E-03	FALSE
	D2	SFT	gemma-3-4b-pt-SFT	medgemma-4b-pt-SFT	4.05E-02	9.53E-02	2.00E-04	5.56E-03	**TRUE**
	D3	SFT	gemma-3-4b-it-SFT	medgemma-4b-pt-SFT	2.34E-02	9.61E-02	2.00E-04	5.56E-03	**TRUE**
CPT	**E1**	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-it-CPT	4.04E-03	2.43E-01	4.24E-02	5.56E-03	FALSE
	E2	CPT	gemma-3-4b-pt-CPT	medgemma-4b-pt-CPT	9.44E-02	4.14E-01	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	gemma-3-4b-it-CPT	medgemma-4b-pt-CPT	6.99E-02	1.63E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-it-CPT-SFT	5.49E-03	7.53E-02	6.80E-03	5.56E-03	FALSE
	F2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	medgemma-4b-pt-CPT-SFT	7.33E-02	1.95E-01	2.00E-04	5.56E-03	**TRUE**
	F3	CPT+SFT	gemma-3-4b-it-CPT-SFT	medgemma-4b-pt-CPT-SFT	3.56E-02	1.55E-01	2.00E-04	5.56E-03	**TRUE**
Mistral 7B									
GENERAL	**A1**	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-v0.1	-1.74E-01	5.41E-02	6.35E-01	4.17E-03	FALSE
	A2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1	-2.09E-01	1.37E-01	8.37E-01	4.17E-03	FALSE
	A3	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1-CPT	-8.42E-02	9.06E-02	9.22E-01	4.17E-03	FALSE
	A4	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-v0.1	-2.32E-01	1.02E-01	6.17E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	Mistral-7B-Instruct-v0.1-CPT	Mistral-7B-Instruct-v0.1	-4.68E-02	1.69E-01	1.48E-01	4.17E-03	FALSE
	B2	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT	-7.19E-02	-3.74E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1	-1.15E-01	1.34E-01	7.76E-01	4.17E-03	FALSE
	B4	SFT	Mistral-7B-Instruct-v0.1-SFT	Mistral-7B-Instruct-v0.1	-2.64E-01	1.30E-02	1.29E-01	4.17E-03	FALSE
MEDICAL	**C1**	CPT	BioMistral-7B-CPT	BioMistral-7B	-1.74E-01	7.18E-02	6.52E-01	4.17E-03	FALSE
	C2	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B	-3.40E-02	1.03E-01	2.42E-01	4.17E-03	FALSE
	C3	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B-CPT	1.04E-02	1.31E-01	9.00E-03	4.17E-03	FALSE
	C4	SFT	BioMistral-7B-SFT	BioMistral-7B	-1.41E-01	4.33E-02	4.24E-01	4.17E-03	FALSE
SFT	**D1**	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-Instruct-v0.1-SFT	2.14E-02	9.83E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Mistral-7B-Instruct-v0.1-SFT	BioMistral-7B-SFT	-3.04E-02	4.71E-02	7.02E-01	5.56E-03	FALSE
	D3	SFT	Mistral-7B-v0.1-SFT	BioMistral-7B-SFT	4.48E-02	7.93E-02	2.00E-04	5.56E-03	**TRUE**
CPT	**E1**	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-Instruct-v0.1-CPT	-1.51E-01	-9.28E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Mistral-7B-v0.1-CPT	BioMistral-7B-CPT	-2.97E-02	1.67E-01	2.44E-01	5.56E-03	FALSE
	E3	CPT	Mistral-7B-Instruct-v0.1-CPT	BioMistral-7B-CPT	1.02E-01	3.12E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	-1.25E-01	-1.04E-02	7.40E-03	5.56E-03	FALSE
	F2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	-2.23E-02	3.51E-02	8.31E-01	5.56E-03	FALSE
	F3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	4.40E-02	1.07E-01	2.00E-04	5.56E-03	**TRUE**
LLAMA-7 FAMILY									
GENERAL	**A1**	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-hf	-1.17E-01	2.53E-03	7.46E-02	4.17E-03	FALSE
	A2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf-CPT	4.46E-02	1.40E-01	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf	-5.92E-02	9.36E-02	5.18E-01	4.17E-03	FALSE
	A4	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-hf	-9.58E-02	5.00E-02	8.01E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	Llama-2-7b-chat-hf-CPT	Llama-2-7b-chat-hf	-1.05E-01	7.92E-02	7.39E-01	4.17E-03	FALSE
	B2	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf	-4.91E-02	7.67E-02	4.88E-01	4.17E-03	FALSE
	B3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT	-3.13E-02	6.42E-02	3.87E-01	4.17E-03	FALSE
	B4	SFT	Llama-2-7b-chat-hf-SFT	Llama-2-7b-chat-hf	-2.67E-01	2.10E-03	8.08E-02	4.17E-03	FALSE
MEDICAL	**C1**	CPT	meditron-7b-CPT	meditron-7b	-8.36E-03	1.95E-02	4.18E-01	4.17E-03	FALSE
	C2	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b	3.69E-03	9.40E-02	4.08E-02	4.17E-03	FALSE
	C3	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b-CPT	-1.49E-02	9.99E-02	1.30E-01	4.17E-03	FALSE
	C4	SFT	meditron-7b-SFT	meditron-7b	-9.43E-02	3.43E-02	4.08E-01	4.17E-03	FALSE
SFT	**D1**	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-chat-hf-SFT	-9.81E-02	-2.37E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	LLama-2-7b-hf-SFT	meditron-7b-SFT	-3.18E-02	-4.98E-03	7.80E-03	5.56E-03	FALSE
	D3	SFT	Llama-2-7b-chat-hf-SFT	meditron-7b-SFT	-4.93E-03	8.30E-02	1.49E-01	5.56E-03	FALSE
CPT	**E1**	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-chat-hf-CPT	-2.96E-01	-1.31E-01	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Llama-2-7b-hf-CPT	meditron-7b-CPT	-1.37E-01	-3.73E-02	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	Llama-2-7b-chat-hf-CPT	meditron-7b-CPT	8.64E-02	1.72E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT-SFT	-2.83E-01	-5.47E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	meditron-7b-CPT-SFT	-7.39E-02	-3.20E-02	2.00E-04	5.56E-03	**TRUE**
	F3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	meditron-7b-CPT-SFT	6.85E-03	2.06E-01	7.00E-03	5.56E-03	FALSE
Llama 13B									
GENERAL	**A1**	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-hf	-9.64E-02	-2.62E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf	-1.25E-02	1.84E-01	1.19E-01	4.17E-03	FALSE
	A3	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf-CPT	6.32E-02	2.31E-01	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-hf	-3.35E-02	9.53E-02	4.87E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	Llama-2-13b-chat-hf-CPT	Llama-2-13b-chat-hf	-1.03E-02	1.32E-01	1.44E-01	4.17E-03	FALSE
	B2	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf	-3.25E-01	6.12E-02	4.22E-01	4.17E-03	FALSE
	B3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT	-3.19E-01	-6.80E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-13b-chat-hf-SFT	Llama-2-13b-chat-hf	-3.78E-01	-4.44E-03	4.28E-02	4.17E-03	FALSE
MEDICAL	**C1**	CPT	MedLLaMA-13B-CPT	MedLLaMA_13B	8.59E-03	3.82E-02	9.60E-03	4.17E-03	FALSE
	C2	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA_13B	-7.65E-03	1.81E-01	1.50E-01	4.17E-03	FALSE
	C3	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA-13B-CPT	-2.00E-02	1.43E-01	1.48E-01	4.17E-03	FALSE
	C4	SFT	MedLLaMA-13B-SFT	MedLLaMA_13B	-4.70E-02	9.93E-02	3.78E-01	4.17E-03	FALSE
SFT	**D1**	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-chat-hf-SFT	5.63E-03	5.57E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Llama-2-13b-hf-SFT	MedLLaMA-13B-SFT	-2.49E-02	4.98E-02	6.34E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-13b-chat-hf-SFT	MedLLaMA-13B-SFT	-7.22E-02	1.13E-02	6.27E-01	5.56E-03	FALSE
CPT	**E1**	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-chat-hf-CPT	-4.04E-01	-2.38E-01	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Llama-2-13b-hf-CPT	MedLLaMA-13B-CPT	-9.08E-02	-6.25E-02	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	Llama-2-13b-chat-hf-CPT	MedLLaMA-13B-CPT	1.63E-01	3.21E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT-SFT	-4.38E-02	4.50E-02	8.76E-01	5.56E-03	FALSE
	F2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	-2.24E-02	3.90E-02	5.73E-01	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	-1.39E-02	2.73E-02	5.62E-01	5.56E-03	FALSE

Table 9: Significance testing for OEQA comparisons. Each row reports a paired bootstrap test between `model_a` and `model_b`, including the 95% confidence interval of the mean difference (`ci95_low`, `ci95_high`), the two-sided p-value, and the Bonferroni-adjusted threshold (`alpha_Bonferoni`) with the resulting decision (`significant_Bonferroni`). IDs A–C compare adaptation strategies within the same model type; IDs D–F compare model initializations across types.

Model Type	Strategy	*mean_words*	*std_words*	*median_words*	*mean_chars*	*std_chars*	*median_chars*
				Gemma-4B			
GENERAL	**Base**	243,30	123,00	288,00	1 616,51	798,74	1,927,00
	CPT	107,81	122,82	37,00	720,79	806,61	245,00
	SFT	176,70	**127,77**	204,00	1 182,16	**877,20**	1 153,00
	CPT+SFT	**279,84**	87,77	**300,00**	**1 884,34**	649,64	**2 076,00**
INSTRUCT	**Base**	261,62	67,98	282,00	1 819,58	473,70	1 985,00
	CPT	**266,87**	98,66	**286,00**	1 763,62	576,08	1 878,00
	SFT	243,03	65,58	261,00	**3 513,80**	827,10	**3 326,00**
	CPT+SFT	183,16	**98,86**	207,00	3 497,45	**1 680,82**	2 632,00
MEDICAL	**Base**	208,62	122,41	205,00	1 371,82	793,35	1 431,00
	CPT	216,10	**142,15**	264,00	1 308,81	**861,75**	1 661,00
	SFT	271,22	97,47	292,00	1 796,17	705,16	**1 982,00**
	CPT+SFT	**282,47**	48,88	**283,00**	**1 825,36**	519,84	1 954,00
				Mistral-7B			
GENERAL	**Base**	212,56	58,84	224,00	1 466,40	329,83	1 502,00
	CPT	173,15	84,30	193,00	1 102,15	536,40	1,321,50
	SFT	130,36	**90,43**	138,00	884,67	**603,37**	906,00
	CPT+SFT	**226,60**	31,85	**229,00**	**1 508,78**	236,78	**1 527,00**
INSTRUCT	**Base**	134,18	74,11	125,00	876,12	476,77	812,00
	CPT	67,79	75,91	37,00	447,32	481,36	244,00
	SFT	19,59	14,19	19,00	138,75	100,30	136,00
	CPT+SFT	**168,09**	**77,91**	**199,00**	**1 112,24**	**499,35**	**1 314,00**
MEDICAL	**Base**	66,26	76,56	37,00	443,89	500,84	250,00
	CPT	99,03	**102,06**	41,00	651,69	650,72	279,00
	SFT	128,83	98,75	**159,00**	855,12	**657,31**	**950,00**
	CPT+SFT	**132,20**	91,72	129,00	**909,46**	619,60	859,00
				Llama-7B			
GENERAL	**Base**	206,13	67,03	222,00	1 358,82	392,59	**1 459,00**
	CPT	41,99	**78,06**	8,00	269,34	**490,77**	56,00
	SFT	**219,26**	35,68	220,00	**1 399,00**	266,52	1 441,00
	CPT+SFT	217,98	50,08	**224,00**	1 376,29	353,15	1 442,00
INSTRUCT	**Base**	**233,79**	66,58	**244,00**	**1 513,45**	433,69	**1 586,00**
	CPT	72,29	**87,21**	26,00	483,67	572,37	180,00
	SFT	18,24	14,33	18,00	126,03	93,60	127,00
	CPT+SFT	123,65	86,95	106,00	859,20	**581,43**	784,00
MEDICAL	**Base**	**227,34**	40,78	**230,00**	**1 498,30**	211,08	**1 522,50**
	CPT	130,92	**104,72**	128,00	847,07	**671,76**	1 008,00
	SFT	204,04	40,98	209,00	1 350,16	345,45	1 431,00
	CPT+SFT	216,96	38,84	220,00	1 416,50	301,18	1 466,00
				Llama-13B			
GENERAL	**Base**	168,15	44,63	146,00	1 144,75	260,69	1 020,00
	CPT	26,19	**53,89**	9,00	173,89	**351,27**	58,00
	SFT	218,99	36,11	219,00	**1 440,23**	300,92	**1 523,00**
	CPT+SFT	**225,12**	36,23	**228,00**	1 429,25	280,81	1 482,00
INSTRUCT	**Base**	**226,32**	60,72	**235,00**	**1 471,46**	396,55	**1 529,00**
	CPT	79,32	**80,61**	45,00	528,55	**526,68**	303,00
	SFT	17,34	10,39	18,00	121,18	74,56	129,00
	CPT+SFT	19,79	17,18	19,00	141,49	132,88	136,00
MEDICAL	**Base**	**217,49**	47,81	**221,00**	**1 429,19**	269,92	1 459,00
	CPT	65,17	**92,92**	13,00	431,03	**597,64**	92,00
	SFT	206,21	53,97	219,00	1 394,11	392,06	**1 505,00**
	CPT+SFT	215,44	37,48	220,00	1 351,46	345,39	1 402,50

Table 11: Output length statistics for OEQA generations across model families, initialization types (GENERAL/INSTRUCT/MEDICAL), and adaptation strategies (Base, CPT, SFT, CPT+SFT). We report the mean, standard deviation, and median number of words and characters per generated answer. **Bold** values highlight, within each block, the maximum value for the corresponding statistic.

M English vs. French Benchmarks: Full Numeric Results

Model Type	Strategy	MCQU-FR		MCQU-EN	
		Greedy	Constrained	Greedy	Constrained
		EM		EM	
Gemma-4B					
GENERAL	**Base**	**5.76**	26.63	1.60	**41.22**
	CPT	8.54	25.60	**12.77**	**40.10**
	SFT	19.92	32.82	**51.60**	**51.60**
	CPT+SFT	19.55	32.73	**51.42**	**51.42**
INSTRUCT	**Base**	29.38	29.76	**47.89**	**47.94**
	CPT	1.36	**24.28**	**1.39**	23.43
	SFT	32.38	32.46	**48.74**	**48.74**
	CPT+SFT	30.31	30.38	**39.17**	**39.17**
MEDICAL	**Base**	**11.50**	26.43	0.04	**32.47**
	CPT	**10.94**	**24.71**	7.64	23.85
	SFT	17.81	30.77	**45.03**	**45.03**
	CPT+SFT	17.41	30.50	**40.14**	**40.14**
Mistral-7B					
GENERAL	**Base**	**4.51**	**28.96**	1.34	26.15
	CPT	**13.84**	**27.15**	6.00	25.20
	SFT	**19.88**	**32.98**	7.77	27.00
	CPT+SFT	**19.47**	**32.22**	9.58	27.39
INSTRUCT	**Base**	**21.58**	**25.51**	5.96	25.10
	CPT	**28.65**	**29.69**	6.90	25.51
	SFT	**31.64**	**31.74**	7.18	26.38
	CPT+SFT	**29.94**	**30.11**	6.75	25.21
MEDICAL	**Base**	**13.52**	**26.88**	5.45	25.69
	CPT	**12.10**	**25.49**	6.05	24.32
	SFT	**18.45**	**31.64**	7.10	26.28
	CPT+SFT	**19.30**	**32.33**	7.43	26.90
Llama-7B					
GENERAL	**Base**	**9.38**	**25.48**	3.46	25.17
	CPT	6.00	25.27	**21.14**	**28.56**
	SFT	15.74	28.61	**32.86**	**32.87**
	CPT+SFT	16.97	29.77	**39.25**	**39.25**
INSTRUCT	**Base**	0.00	**24.34**	0.00	23.44
	CPT	0.00	**24.29**	0.00	23.49
	SFT	29.52	29.58	**38.73**	**38.73**
	CPT+SFT	0.00	**24.54**	0.00	23.45
MEDICAL	**Base**	0.35	**24.19**	**0.98**	23.68
	CPT	**11.97**	**25.14**	0.09	24.96
	SFT	17.38	30.27	**36.80**	**36.80**
	CPT+SFT	18.29	31.60	**36.53**	**36.53**
Llama-13B					
GENERAL	**Base**	11.20	25.51	**16.55**	**34.83**
	CPT	10.68	26.64	**29.79**	**37.21**
	SFT	17.30	30.18	**43.22**	**43.22**
	CPT+SFT	18.27	31.41	**43.62**	**43.62**
INSTRUCT	**Base**	0.00	21.68	0.00	**23.60**
	CPT	0.00	24.57	0.00	**24.85**
	SFT	30.04	30.10	**46.99**	**46.99**
	CPT+SFT	31.42	31.51	**46.57**	**46.57**
MEDICAL	**Base**	10.31	23.97	**12.48**	**24.28**
	CPT	**10.22**	23.37	10.01	**30.87**
	SFT	16.73	29.88	**37.71**	**37.71**
	CPT+SFT	18.24	31.32	**42.73**	**42.73**

Table 12: Cross-lingual comparison between native English MCQU benchmarks (MCQU-EN) and their French translations (MCQU-FR), reported as EM (%). Results are shown for both greedy and constrained decoding. For each row and decoding type, bold values indicate the higher EM between MCQU-FR and MCQU-EN.

The main paper reports averaged results using *constrained decoding* (Figure 2). Here, we provide the complete *numeric* EM results for both *greedy* and *constrained* decoding on the native English MCQU benchmarks (MCQU-EN) and their French translations (MCQU-FR).

M.1 Greedy decoding analysis

The greedy decoding results reported in Table 12 exhibit the same overall tendencies as those observed under constrained decoding in section 5. For the Mistral family, greedy decoding consistently yields higher performance on the French translations than on the original English benchmarks, both before and after adaptation. Conversely, Gemma and Llama models generally perform better on native English benchmarks under greedy decoding, and this advantage is preserved after French medical adaptation.

As with constrained decoding, adaptation on French medical data improves performance in both languages under greedy decoding, indicating effective cross-lingual transfer. While absolute EM scores differ between decoding strategies, greedy decoding generally producing lower scores, the relative ordering between English and French benchmarks and the direction of adaptation effects remain consistent. These results suggest that the cross-lingual patterns reported in the main paper are robust to the choice of decoding strategy.

M.2 Significance testing (English vs. French)

To assess whether the English–French performance gaps are statistically significant, we perform paired significance testing *separately for each model configuration*, i.e., for each combination of (model family/type, adaptation strategy, decoding type). For each configuration, we compute the per-item EM difference between MCQU-EN and MCQU-FR on matched translated instances, and estimate a 95% confidence interval for the mean difference together with a two-sided p-value. Because each test compares a model strictly with itself across languages and each English–French pair is independent of the others, we do not apply a Bonferroni correction. Table 13 reports the resulting confidence intervals and significance decisions.

Model Type	Strategy	Model	Decoding Type	ci95_low	ci95_high	p_two_sided	Significant
			Gemma-4B				
GENERAL	Base	gemma-3-4b-pt	greedy	2,57E-02	6,42E-02	2,00E-04	TRUE
	Base	gemma-3-4b-pt	constrained	-1,97E-01	-9,55E-02	2,00E-04	TRUE
	CPT	gemma-3-4b-pt-CPT	greedy	-6,57E-02	-1,71E-02	2,20E-03	TRUE
	CPT	gemma-3-4b-pt-CPT	constrained	-1,86E-01	-1,01E-01	2,00E-04	TRUE
	CPT+SFT	gemma-3-4b-CPT-SFT	greedy	-4,01E-01	-2,18E-01	2,00E-04	TRUE
	CPT+SFT	gemma-3-4b-CPT-SFT	constrained	-2,48E-01	-1,14E-01	2,00E-04	TRUE
	SFT	gemma-3-4b-pt-SFT	greedy	-3,98E-01	-2,19E-01	2,00E-04	TRUE
	SFT	gemma-3-4b-pt-SFT	constrained	-2,50E-01	-1,17E-01	2,00E-04	TRUE
INSTRUCT	Base	gemma-3-4b-it	greedy	-2,31E-01	-1,36E-01	2,00E-04	TRUE
	Base	gemma-3-4b-it	constrained	-2,27E-01	-1,33E-01	2,00E-04	TRUE
	CPT	gemma-3-4b-it-CPT	greedy	-8,82E-03	7,42E-03	9,70E-01	FALSE
	CPT	gemma-3-4b-it-CPT	constrained	-1,89E-02	4,55E-02	6,71E-01	FALSE
	CPT+SFT	gemma-3-4b-it-CPT-SFT	greedy	-1,14E-01	-6,32E-02	2,00E-04	TRUE
	CPT+SFT	gemma-3-4b-it-CPT-SFT	constrained	-1,14E-01	-6,21E-02	2,00E-04	TRUE
	SFT	gemma-3-4b-it-SFT	greedy	-2,22E-01	-9,37E-02	2,00E-04	TRUE
	SFT	gemma-3-4b-it-SFT	constrained	-2,20E-01	-9,53E-02	2,00E-04	TRUE
MEDICAL	Base	medgemma-4b-pt	greedy	6,05E-02	2,00E-01	2,00E-04	TRUE
	Base	medgemma-4b-pt	constrained	-1,05E-01	-6,63E-03	2,94E-02	TRUE
	CPT	medgemma-4b-pt-CPT	greedy	-8,53E-03	9,14E-02	1,71E-01	FALSE
	CPT	medgemma-4b-pt-CPT	constrained	-1,45E-02	4,11E-02	6,29E-01	FALSE
	CPT+SFT	medgemma-4b-pt-CPT-SFT	greedy	-2,92E-01	-1,50E-01	2,00E-04	TRUE
	CPT+SFT	medgemma-4b-pt-CPT-SFT	constrained	-1,39E-01	-4,81E-02	4,00E-04	TRUE
	SFT	medgemma-4b-pt-SFT	greedy	-3,43E-01	-1,90E-01	2,00E-04	TRUE
	SFT	medgemma-4b-pt-SFT	constrained	-1,93E-01	-8,67E-02	2,00E-04	TRUE
			Mistral-7B				
GENERAL	Base	Mistral-7B-v0.1	greedy	-1,62E-02	1,20E-01	6,98E-01	FALSE
	Base	Mistral-7B-v0.1	constrained	8,39E-02	1,70E-01	2,00E-04	TRUE
	CPT	Mistral-7B-v0.1-CPT	greedy	3,64E-02	1,48E-01	2,00E-04	TRUE
	CPT	Mistral-7B-v0.1-CPT	constrained	7,59E-02	1,43E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	greedy	5,94E-02	1,61E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	constrained	4,90E-02	1,04E-01	2,00E-04	TRUE
	SFT	Mistral-7B-v0.1-SFT	greedy	7,77E-02	1,92E-01	2,00E-04	TRUE
	SFT	Mistral-7B-v0.1-SFT	constrained	5,74E-02	1,29E-01	2,00E-04	TRUE
INSTRUCT	Base	Mistral-7B-Instruct-v0.1	greedy	1,30E-01	1,83E-01	2,00E-04	TRUE
	Base	Mistral-7B-Instruct-v0.1	constrained	1,21E-01	1,60E-01	2,00E-04	TRUE
	CPT	Mistral-7B-Instruct-v0.1-CPT	greedy	1,74E-01	2,62E-01	2,00E-04	TRUE
	CPT	Mistral-7B-Instruct-v0.1-CPT	constrained	9,71E-02	1,73E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	greedy	1,83E-01	2,94E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	constrained	6,43E-02	1,42E-01	2,00E-04	TRUE
	SFT	Mistral-7B-Instruct-v0.1-SFT	greedy	1,92E-01	3,03E-01	2,00E-04	TRUE
	SFT	Mistral-7B-Instruct-v0.1-SFT	constrained	7,81E-02	1,56E-01	2,00E-04	TRUE
MEDICAL	Base	BioMistral-7B	greedy	4,88E-02	1,29E-01	2,00E-04	TRUE
	Base	BioMistral-7B	constrained	1,21E-01	1,63E-01	2,00E-04	TRUE
	CPT	BioMistral-7B-CPT	greedy	3,47E-02	9,30E-02	2,00E-04	TRUE
	CPT	BioMistral-7B-CPT	constrained	9,18E-02	1,20E-01	2,00E-04	TRUE
	CPT+SFT	BioMistral-7B-CPT-SFT	greedy	6,82E-02	2,06E-01	2,00E-04	TRUE
	CPT+SFT	BioMistral-7B-CPT-SFT	constrained	8,89E-02	1,88E-01	2,00E-04	TRUE
	SFT	BioMistral-7B-SFT	greedy	6,71E-02	1,94E-01	2,00E-04	TRUE
	SFT	BioMistral-7B-SFT	constrained	8,33E-02	1,68E-01	2,00E-04	TRUE
			Llama-7B				
GENERAL	Base	Llama-2-7b-hf	greedy	1,22E-02	1,24E-01	6,60E-03	TRUE
	Base	Llama-2-7b-hf	constrained	-2,49E-02	3,53E-02	8,46E-01	FALSE
	CPT	Llama-2-7b-hf-CPT	greedy	-2,23E-01	-6,14E-02	4,00E-03	TRUE
	CPT	Llama-2-7b-hf-CPT	constrained	-2,27E-02	1,19E-01	3,12E-01	FALSE
	CPT+SFT	LLama-2-7b-hf-CPT-SFT	greedy	-2,93E-01	-1,33E-01	2,00E-04	TRUE
	CPT+SFT	LLama-2-7b-hf-CPT-SFT	constrained	-1,49E-01	-3,08E-02	4,20E-03	TRUE
	SFT	LLama-2-7b-hf-SFT	greedy	-2,27E-01	-9,34E-02	4,00E-04	TRUE
	SFT	LLama-2-7b-hf-SFT	constrained	-8,38E-02	7,10E-03	8,72E-02	FALSE
INSTRUCT	Base	Llama-2-7b-chat-hf	greedy	0,00E+00	0,00E+00	1,00E+00	FALSE
	Base	Llama-2-7b-chat-hf	constrained	1,84E-01	3,09E-01	2,00E-04	TRUE
	CPT	Llama-2-7b-chat-hf-CPT	greedy	0,00E+00	1,02E-04	7,23E-01	FALSE
	CPT	Llama-2-7b-chat-hf-CPT	constrained	5,27E-02	2,03E-01	2,00E-04	TRUE
	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	greedy	0,00E+00	6,80E-05	7,01E-01	FALSE
	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	constrained	8,66E-02	1,40E-01	2,00E-04	TRUE
	SFT	Llama-2-7b-chat-hf-SFT	greedy	-1,32E-01	-3,94E-02	1,20E-03	TRUE
	SFT	Llama-2-7b-chat-hf-SFT	constrained	-1,31E-01	-3,80E-02	1,20E-03	TRUE
MEDICAL	Base	meditron-7b	greedy	-1,03E-02	-2,51E-03	1,00E-03	TRUE
	Base	meditron-7b	constrained	-1,42E-02	3,19E-02	5,00E-01	FALSE
	CPT	meditron-7b-CPT	greedy	-6,55E-03	1,08E-01	3,98E-01	FALSE
	CPT	meditron-7b-CPT	constrained	2,15E-03	6,87E-02	3,58E-02	TRUE
	CPT+SFT	meditron-7b-CPT-SFT	greedy	-2,61E-01	-7,38E-02	1,80E-03	TRUE
	CPT+SFT	meditron-7b-CPT-SFT	constrained	-1,05E-01	2,86E-02	1,86E-01	FALSE
	SFT	meditron-7b-SFT	greedy	-2,64E-01	-1,02E-01	2,00E-04	TRUE
	SFT	meditron-7b-SFT	constrained	-1,15E-01	-3,59E-03	4,00E-02	TRUE
			Llama-13B				
GENERAL	Base	Llama-2-13b-hf	greedy	-1,31E-01	5,40E-02	2,74E-01	FALSE
	Base	Llama-2-13b-hf	constrained	-1,35E-01	-4,79E-02	4,00E-04	TRUE
	CPT	Llama-2-13b-hf-CPT	greedy	-2,65E-01	-1,00E-01	2,00E-04	TRUE
	CPT	Llama-2-13b-hf-CPT	constrained	-1,09E-01	2,30E-02	1,58E-01	FALSE
	CPT+SFT	Llama-2-13b-hf-CPT-SFT	greedy	-3,31E-01	-1,52E-01	2,00E-04	TRUE
	CPT+SFT	Llama-2-13b-hf-CPT-SFT	constrained	-1,82E-01	-5,08E-02	1,60E-03	TRUE
	SFT	Llama-2-13b-hf-SFT	greedy	-3,30E-01	-1,68E-01	2,00E-04	TRUE
	SFT	Llama-2-13b-hf-SFT	constrained	-1,87E-01	-6,61E-02	2,00E-04	TRUE
INSTRUCT	Base	Llama-2-13b-chat-hf	greedy	0,00E+00	0,00E+00	1,00E+00	FALSE
	Base	Llama-2-13b-chat-hf	constrained	-4,79E-02	2,61E-03	1,06E-01	FALSE
	CPT	Llama-2-13b-chat-hf-CPT	greedy	0,00E+00	0,00E+00	1,00E+00	FALSE
	CPT	Llama-2-13b-chat-hf-CPT	constrained	-2,29E-02	2,51E-02	7,65E-01	FALSE
	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	greedy	-2,09E-01	-7,94E-02	4,00E-04	TRUE
	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	constrained	-2,06E-01	-7,95E-02	2,00E-04	TRUE
	SFT	Llama-2-13b-chat-hf-SFT	greedy	-2,21E-01	-1,15E-01	2,00E-04	TRUE
	SFT	Llama-2-13b-chat-hf-SFT	constrained	-2,20E-01	-1,16E-01	2,00E-04	TRUE
MEDICAL	Base	MedLLaMA_13B	greedy	-8,31E-02	4,03E-02	4,90E-01	FALSE
	Base	MedLLaMA_13B	constrained	-2,46E-02	1,98E-02	7,65E-01	FALSE
	CPT	MedLLaMA-13B-CPT	greedy	-3,43E-02	3,73E-02	8,95E-01	FALSE
	CPT	MedLLaMA-13B-CPT	constrained	-1,10E-01	-4,51E-02	2,00E-04	TRUE
	CPT+SFT	MedLLaMA-13B-CPT-SFT	greedy	-3,14E-01	-1,60E-01	2,00E-04	TRUE
	CPT+SFT	MedLLaMA-13B-CPT-SFT	constrained	-1,68E-01	-5,56E-02	1,40E-03	TRUE
	SFT	MedLLaMA-13B-SFT	greedy	-2,76E-01	-1,29E-01	2,00E-04	TRUE
	SFT	MedLLaMA-13B-SFT	constrained	-1,27E-01	-2,23E-02	7,20E-03	TRUE

Table 13: Paired significance testing between MCQU-EN and MCQU-FR for each model configuration (model, strategy, and decoding type). Reported values are the 95% confidence interval of the mean EM difference and the corresponding two-sided two-sided p-value; `Significant` indicates whether the difference is statistically significant. We define the difference as $(\mathrm{FR} - \mathrm{EN})$, such that positive values indicate higher performance in French.

N Effect of Translated Benchmarks on Performance and Confidence

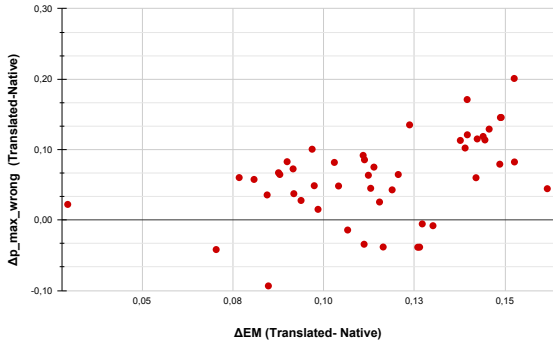

Figure 5: Relationship between accuracy gain (ΔEM) and change in confidence on incorrect predictions ($\Delta p_{\max,\text{wrong}}$) between the translated and native benchmarks. Positive values of $\Delta p_{\max,\text{wrong}}$ indicate increased confidence on incorrect predictions.

O Computational Resources and Environmental Impact

Table 14 summarizes the computational resources and environmental impact associated with each adaptation strategy, aggregated by model size. For clarity and conciseness, we do not report the consumption of each individual training run. Instead, we provide a representative summary per model size and per adaptation strategy. In total, 36 training runs were performed across all experiments. The CPT+SFT strategy is not reported as a separate entry in the table, as its computational cost and environmental impact correspond to the sum of the CPT and SFT phases. Reporting CPT and SFT independently therefore fully characterizes the overall resource usage of the combined strategy.

We report, for each configuration, the dataset size, number of epochs, batch size, GPU type, GPU memory, number of GPUs, total training time, estimated carbon emissions (in gCO_2e), and estimated monetary cost (in USD). Carbon emissions and cost estimates are derived from documented power consumption profiles and usage costs of the underlying high-performance computing infrastructure. ***To preserve author anonymity during peer review, the specific computing facility is not disclosed***.

O.1 Analysis

Overall, the results highlight a clear contrast between CPT and SFT in terms of computational cost and environmental impact. CPT is consistently the most resource-intensive strategy, driven by large-scale datasets, longer effective compute time, and high degrees of GPU parallelism.

In contrast, SFT incurs substantially lower emissions and monetary costs across all model sizes. This difference is not only due to the smaller dataset size, but also to the use of parameter-efficient fine-tuning: SFT is implemented with DoRA adapters rather than full weight updates, significantly reducing both memory usage and energy consumption. Despite longer wall-clock durations in some configurations, the overall compute footprint of SFT remains markedly lower than that of CPT.

As model size increases, CPT costs grow rapidly, particularly for the 13B setting, where energy consumption and carbon emissions increase sharply. SFT, while also scaling with model size, remains comparatively efficient due to its parameter-efficient design. These findings underscore the importance of adaptation strategies that balance performance gains with computational and environmental sustainability.

P Pretraining Data Contamination Study: Was NACHOS Seen During Pretraining?

Because most of the base models we evaluate (Gemma, MedGemma, Mistral, and Llama) do not disclose their full pretraining mixtures, we conducted a small contamination study to probe whether the French biomedical NACHOS corpus may have been included (or partially included) in their pretraining data. This appendix reports two complementary, lightweight detection protocols inspired by the broader literature on memorization and pretraining-data detection in large language models (Ravaut et al., 2024)

P.1 Protocol 1: Prefix–Continuation Reproduction + Likelihood Heuristics

Idea. If a model has memorized (or near-memorized) training documents, conditioning on a prefix may lead it to reproduce the exact continuation, or to assign a noticeably higher likelihood to the true continuation than to a perturbed version. This is conceptually related to training-data extraction / memorization diagnostics used in prior work (Carlini et al., 2021).

Implementation. Using a sample of n=1915 NACHOS documents, we split each document into a *prefix* (first 400 characters) and a *continuation*

Model Size	Strategy	Dataset size (KB)	Epochs	Batch-size	Type of GPU	Memory per GPU (GB)	Number of GPUs	Training time (hours)	Emissions (g CO2e)	Cost (USD)
4B	CPT	4 000 000	3	4	NVIDIA A100	80	24	80	49 344	1 824.62
	SFT	369	10	4	NVIDIA H100	80	3	146	11 256.6	832.48
7B	CPT	4 000 000	3	2	NVIDIA A100	80	32	40	32 896	1 216.42
	SFT	369	10	4	NVIDIA H100	80	1	190	4 883	361.12
13B	CPT	4 000 000	3	2	NVIDIA H100	80	32	100	82 240	6 082.08
	SFT	369	10	4	NVIDIA H100	80	6	122	18 812.4	1 391.27

Table 14: Summary of computational resources and environmental impact for different adaptation strategies, aggregated by model size. Reported values correspond to a representative training configuration per strategy. CPT+SFT costs are obtained by summing CPT and SFT.

(rest). For each sampled document, we: (i) generate up to 200 new tokens from the prefix (greedy decoding), and compute ROUGE-L between the generated continuation and the gold continuation; (ii) compute the length of the longest common prefix (LCP) between generated and gold continuations; (iii) compute the perplexity of the gold continuation conditioned on the prefix, and compare it to the perplexity of a lightly perturbed continuation (character swaps + whitespace noise), reporting the ratio $\mathrm{PPL}(\text{gold})/\mathrm{PPL}(\text{perturbed})$. We flag a case as "suspicious" if any of the following holds: ROUGE-L ≥ 0.7, LCP ≥ 200 characters, or $\mathrm{PPL}(\text{gold})/\mathrm{PPL}(\text{perturbed}) \leq 0.85$.

Results. Across models, ROUGE-L remained very low and we observed no exact continuation matches, which does *not* support verbatim memorization of long continuations under this setup. However, the fraction of items flagged as "suspicious" is extremely high (0.82–0.96), which indicates that our heuristic is likely over-sensitive (in particular, the perturbation and/or the chosen ratio threshold may dominate the flagging decision).

- Llama-2-7B: ROUGE-L $= 0.031$, exact matches $= 0/1915$, suspicious fraction $= 0.959$.
- Llama-2-13B: ROUGE-L $= 0.020$, exact matches $= 0/1915$, suspicious fraction $= 0.964$.
- Mistral-7B: ROUGE-L $= 0.014$, exact matches $= 0/1915$, suspicious fraction $= 0.944$.
- MedGemma-4B: ROUGE-L $= 0.018$, exact matches $= 0/1915$, suspicious fraction $= 0.821$.
- Gemma-3-4B: ROUGE-L $= 0.019$, exact matches $= 0/1915$, suspicious fraction $= 0.835$.

Interpretation. Given the near-zero reproduction scores (ROUGE-L, exact match) but massive "suspicious" rates, this first protocol is inconclusive as a contamination detector in our setting: it does not show direct copying, and the likelihood-based heuristic is too unstable without a careful calibration procedure and stronger perturbations/-controls. This is consistent with known difficulties of turning likelihood signals into reliable membership decisions without explicit calibration (Yeom et al., 2018).

P.2 Protocol 2: DC-PDD (Divergence-based Calibration Pretraining Data Detection)

Idea. We also tested a dedicated pretraining-data detection score, DC-PDD, which estimates a per-text statistic $\beta(x)$ combining (i) the model probability of next tokens and (ii) reference token frequencies estimated from a large background corpus D' (here, French OSCAR[11]). The method is designed to be more robust than raw perplexity by incorporating a calibration term from D' (Zhang et al., 2024).

Implementation. For each model, we first build a tokenizer-specific unigram table $p(v; D')$ from OSCAR-FR (streaming counts, capped number of documents), then compute DC-PDD $\beta(x)$ on: (i) 1,000 NACHOS samples, and (ii) a *synthetic control* set ("non-member") of biomedical texts generated to be *unlikely* to appear in any public pretraining mixture. We report distributional statistics (median, p75/p90/p95, mean, std) for both sets and a separation diagnostic $\Delta_{\text{median}} = \text{median}(\beta_{\text{nachos}}) - \text{median}(\beta_{\text{control}})$.

Why synthetic controls? (Major limitation) Gemma-family models were released recently (June 2025), and we could not reliably curate a sufficiently large set of *web-native biomedical French texts written after the model release date* to serve

[11] https://oscar-project.org/

as a credible "definitely-non-member" control. As a consequence, we used synthetic biomedical controls, which weakens the study: synthetic controls differ from natural corpora in style and token statistics, and thus may artificially inflate separation (or mask it), independently of membership. We therefore treat DC-PDD results as *indicative only*, not as evidence of true pretraining inclusion.

Results. DC-PDD yields consistently *lower* scores on NACHOS than on the synthetic controls (negative Δ), suggesting the models assign *slightly* more "in-distribution" likelihood structure to NACHOS than to the synthetic texts. The separation is small for Mistral/Llama and somewhat larger for Gemma/MedGemma:

- Mistral-7B: $\Delta_{\text{median}} \approx -3.23 \times 10^{-4}$.
- Llama-2-7B: $\Delta_{\text{median}} \approx -3.58 \times 10^{-4}$.
- Llama-2-13B: $\Delta_{\text{median}} \approx -2.94 \times 10^{-4}$.
- Gemma-3-4B: $\Delta_{\text{median}} \approx -8.20 \times 10^{-4}$.
- MedGemma-4B: $\Delta_{\text{median}} \approx -6.77 \times 10^{-4}$.

Interpretation. While DC-PDD produces a consistent ordering (Nachos $<$ Control), this cannot be confidently attributed to pretraining membership because our control set is synthetic and therefore not distribution-matched. In other words, the observed separation may reflect *domain/style differences* rather than exposure during pretraining. As prior work emphasizes, robust pretraining-data detection typically requires carefully constructed controls and/or calibrated baselines (e.g., Min-K% variants, calibrated likelihood tests), which we could not fully satisfy here (Zhang et al., 2024).

P.3 Summary and Takeaways

Overall, these experiments do not provide strong evidence for (or against) NACHOS being included in the undisclosed pretraining mixtures: (i) we do not observe continuation copying under our greedy prefix–continuation setup; (ii) DC-PDD shows a small but consistent separation between NACHOS and synthetic controls, but the lack of a reliable post-release, naturally occurring biomedical control corpus makes the conclusion weak. We therefore report these results for transparency, but we do not use them to support any causal claim about pretraining contamination in the main analysis.