FULLY OPEN MEDITRON: AUDITABLE AND REPRODUCIBLE DEVELOPMENT OF CLINICAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical decision-support LLMs are either proprietary or open-weight, without the training data and development pipelines required for end-to-end reproducibility and auditing. We investigate how much of the gap between fully-open and open-weight clinical specialists is closeable under full-openness constraints. We introduce MeditronFO (Meditron Fully Open), to our knowledge, the first fully-open pipeline for clinical LLM finetuning. The pipeline creates three new synthetic datasets (340k QA pairs) from clinician-authored adversarial vignettes, clinical practice guidelines and public medical QA, using open-weight gpt-oss-120b for rejection-sampling distillation with pass@8, with generation prompts co-authored by a panel of four clinicians. We apply the pipeline to five fully-open base models from three families and one open-weight control (Gemma-3-27B). On the average of five medical benchmarks, Apertus-70B-MeditronFO reaches 54.56%, against 48.35% for Apertus-70B-Instruct and 63.75% for MedGemma-27B, closing 40% of the gap; the Gemma-3-27B-MeditronFO control reaches 58.65%. We also introduce AutoMOOVE, an automated open-ended clinician preference benchmark, whose judgments agree with blinded clinicians' preferences ($\kappa = 0.564$). The Gemma-3-27B-MeditronFO control is level with or ahead of MedGemma-27B on this benchmark (57.1% win rate under a three-judge majority vote, above parity under two of three judges). We release the data, code and model weights as public artifacts.

1 INTRODUCTION

Medical LLMs are becoming common in high-stakes clinical settings, from clinical decision support to patient-facing applications to clinical documentation (Woo et al., 2025), with little transparency in their creation process (Gilbert et al., 2023; Meskó & Topol, 2023). As LLMs start to impact real-world clinical decisions, recent work calls for explainability and for guidelines for the development of medical LLMs (Gallifant et al., 2025; Mesinovic et al., 2025; Li et al., 2025). Government bodies, such as the FDA, discuss the need for transparency of AI in health care (Warraich et al., 2025). Medical LLMs can be compromised through poisoned training data (Alber et al., 2025), where even small amounts of incorrect information can cause disproportionate misalignment (Betley et al., 2026); without access to both the training data and training pipeline, a comprehensive audit is impossible. But the majority of developed medical LLMs are either open-weight, like MedGemma (Sellergren et al., 2025), Aloe Beta (Garcia-Gasulla et al., 2026), Meditron (Chen et al., 2023b), and BioMistral (Labrak et al., 2024), or proprietary, like Med-PaLM 1 and 2 (Singhal et al., 2023; 2025) and AMIE (Saab et al., 2026). Fully-open LLMs, which release not just their weights but also their training code and data, have emerged, such as Apertus (Hernández-Cano et al., 2025), OLMo (OLMo et al., 2024) and EuroLLM (Ramos et al., 2026), but they have not yet been adapted to medicine.

The evaluation procedures currently used to validate medical LLMs have also raised concerns: benchmarks quickly become saturated (Shao et al., 2025; Bedi et al., 2025), and recent work calls for rethinking medical benchmarks as incomplete (Moreno & Bitterman, 2024; Wang et al., 2026a; Alaa et al., 2025). This issue is further exacerbated by the over-reliance on multiple-choice evaluation. MCQs fail to capture clinically important dimensions, such as contextual awareness, communication, harmlessness, and alignment with guidelines. CRAFT-MD shows that models that ace exam-style QA perform significantly worse on the same cases in open-ended patient conversations (Johri et al., 2025), and MedHELM (Bedi et al., 2026) shows that near-saturated licensing-exam scores coexist

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

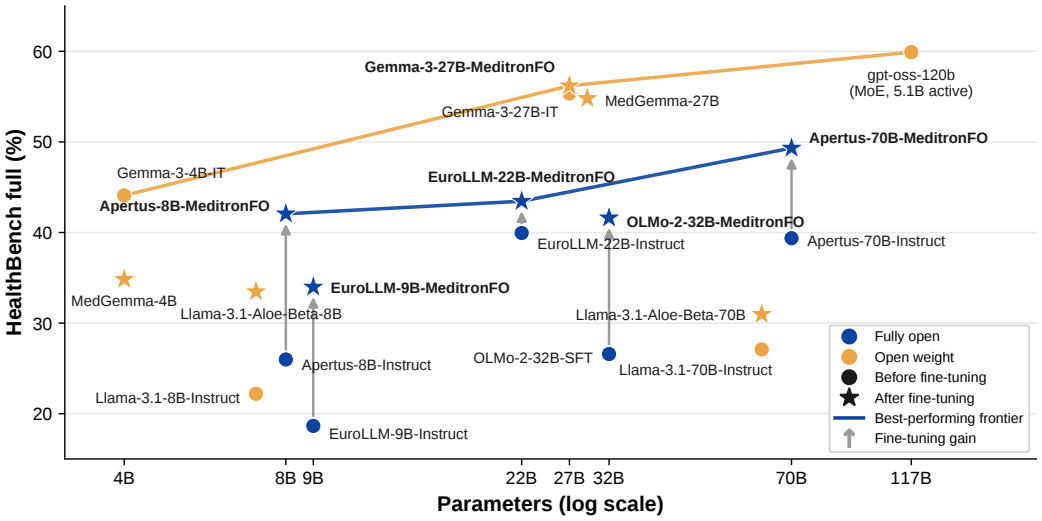

Figure 1: Open models on HealthBench against model size. (Gemma-4-31B grader, HealthBench full)

with markedly weaker performance on clinical decision support and administrative tasks drawn from real practice. Clinical deployment therefore requires evaluation grounded in clinical workflows rather than exam-style recall (Thirunavukarasu et al., 2023). Existing benchmarks also underrepresent low-resource settings (Restrepo et al., 2025), vulnerable populations, and diagnostic reasoning under uncertainty (Kesmen et al., 2026). HealthBench grades open-ended answers against physician-written rubrics (Arora et al., 2025). Clinician platforms such as MOOVE (Sallinen et al., 2025) and MedArena (Wu et al., 2026) instead have clinicians compare the answers of two models and rank the models from these preferences; MedArena's rankings remain stable after controlling for response length and formatting (Wu et al., 2026), but clinician time limits the scale of such evaluations. LLM judges agree with human pairwise preferences about as often as humans agree with each other (Zheng et al., 2023), and we therefore construct an evaluation that automates these pairwise comparisons on open-ended clinical vignettes (Section 3.3).

We introduce **MeditronFO**, the first fully-open pipeline for creating medical LLMs. Our main contributions are: **1) A reproducible and auditable medical corpus (Section 2).** We release a 534k-example training corpus generated from public QA, clinical practice guidelines and clinician-written adversarial vignettes, using prompts co-authored by four clinicians. We release every step with its code and outputs: generation, rejection-sampling distillation, decontamination and audit. **2) The first fully-open medical LLMs (Section 4).** We finetune five fully-open base models from three families into the first fully-open medical LLMs. Apertus-70B-MeditronFO improves over its base on the five-benchmark medical average (48.35% $\rightarrow$ 54.56%) and on HealthBench (39.4% $\rightarrow$ 49.3%), closing 40% of the average gap to the open-weight MedGemma-27B. **3) AutoMOOVE, a clinician-validated open-ended evaluation (Section 3.3).** An LLM judge compares two answers to clinician-written adversarial vignettes and agrees with unanimous blinded clinician preferences ($\kappa = 0.583$; 0.715 without ties; 0.564 and 0.718 for the three-judge majority). Under a majority vote of three LLM judges, Apertus-70B-MeditronFO is preferred over its base (89.7% win rate), and the Gemma-3-27B-MeditronFO control is level with or ahead of MedGemma-27B (57.1%).

2 MeditronFO Development Pipeline

Figure 2 gives an overview of the MeditronFO pipeline. The training corpus combines four datasets. **Curated QA** aggregates seven commonly used public medical QA and MCQA datasets, covering both patient-facing and medical-exam questions: MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), HealthSearchQA (Singhal et al., 2023), LiveQA (Abacha et al., 2017) and AfriMed-QA v1/v2 (Olatunji et al., 2024) for African context inclusion. The other three datasets are generated by an open-weight LLM, gpt-oss-120b (Appendix A). **Synthetic Curated QA** (184.5k examples) expands Curated QA with new exam-style and open-ended questions. **Synthetic**

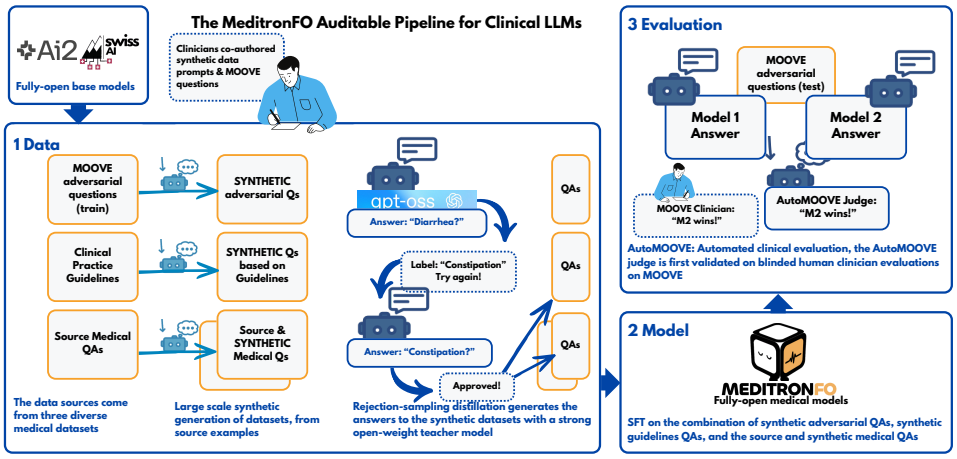

Figure 2: The MeditronFO Auditable and Reproducible Development Pipeline for Clinical LLMs.

vignettes (24.0k) are new adversarial vignettes modelled on those that clinicians wrote for the MOOVE (Massive Open Online Validation and Evaluation) platform (Sallinen et al., 2025): short hypothetical patient cases designed to be adversarial, which may contain a trap or be non-trivial, making them a good basis to push the limits of medical LLMs. We keep the MOOVE vignettes themselves as evaluation material. **Guidelines QA** (129.8k) are questions generated from 14,192 full-text clinical practice guidelines from 16 institutions scraped by Meditron (Chen et al., 2023b), the highest-quality source of medical knowledge. A teacher model writes the answers of all four datasets (Section 2.2), and the corpus is then decontaminated and audited (Section 2.3) before finetuning (Section 2.4).

2.1 Synthetic data generation

We use synthetic data generation for two purposes: the first is to expand Curated QA, and the second is to create two new training datasets based on the MOOVE vignettes and guideline sources. This also serves the purpose of scaling the coverage of the dataset and filling distributional gaps from Curated QA. To create the new questions, we use gpt-oss-120b, a strong and fast open-weight LLM (Agarwal et al., 2025). We follow the method of Sellergren et al. (2025): we sample five questions randomly from the source dataset and use them as examples in the prompt to the model to generate one new example. For Synthetic Curated QA, the five examples come from either the multiple-choice or the open-ended part of Curated QA, and the new question is of the same type. We make sure that the vignettes are seeded from a train split of the MOOVE vignette pool, distinct from the test split and decontaminated against it. To create Guidelines QA, we pass one full guideline and prompt the model to generate ten multiple-choice clinical vignettes answerable from this guideline alone.

Clinician Review. Before generating the final synthetic dataset, we convened a panel discussion of four clinicians to co-author the final prompts used for generation. We showed the clinicians the preliminary prompts and preliminary examples of synthetic questions generated by the pipeline and asked them to rewrite the prompts and surface potential issues. The panel comprised clinicians with expertise across global health, humanitarian response, infectious disease, primary care, pediatrics and surgery. The clinician review was essential and surfaced issues in the prompts: (1) The constraint "avoid controversial, outdated, or unproven treatments" was narrowed, as "outdated" might backfire for low-resource settings. (2) Prompts were modified to ask explicitly for details on context and disease progression. (3) Prompts no longer require questions to be evidence-based; this should only apply to answers. (4) Low-quality sources were excluded from the guidelines, and US-centric phrasing was removed from the prompts. We provide the final prompts in Appendix A.

2.2 Rejection-sampling distillation

Rejection-sampling distillation has shown good results in enhancing open-weight LLMs' capabilities (Yuan et al., 2023; Grattafiori et al., 2024). We use this method to mitigate hallucinations: every item with a gold label (the human label in Curated QA, the generator's label in Synthetic Curated

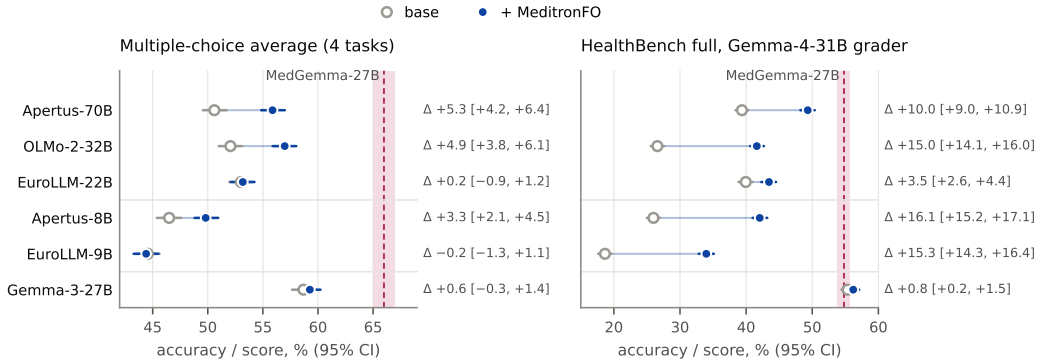

Figure 3: Overall medical benchmark results for MeditronFO models against their base. *Left:* Apertus-70B, Apertus-8B and OLMo-2-32B benefit strongly from finetuning on the multiple-choice average, while EuroLLM and Gemma do not show improvement. *Right:* On the open-ended HealthBench evaluation, all MeditronFO models improve their score.

QA and Guidelines QA) is rejection-sampled with pass@8 at temperature 0.7, and examples never reaching the gold label are dropped (7.7–8.4% of the multiple-choice rows, Appendix Table 3). For the open-ended questions and the Synthetic vignettes we distill without rejection sampling, since there is no gold label. As the teacher model for distillation, we use gpt-oss-120b (Agarwal et al., 2025), selected as a strong open-weight model on medical knowledge. We additionally include a second-teacher ablation in Section 5.2.

2.3 Corpus validation

Decontamination. Medical benchmarks are widely contaminated with training data (Deng et al., 2024), so we apply a two-stage n-gram decontamination to ensure no samples from the training data overlap with the evaluation data. We follow the decontamination methodology of the Apertus tech report (Hernández-Cano et al., 2025); the Synthetic vignettes are also checked against the AutoMOOVE test set (Appendix C). Applying the pipeline to the full 600,361-example corpus removed 4,197 samples (0.70%), almost all from Curated QA (4,113), mostly MedMCQA (2,856) and PubMedQA (1,000) test questions; the three synthetic datasets lose 84 samples in total. Per-benchmark removals, and the rows dropped at each filtering step down to the final 533,888, are in Appendix Tables 4 and 5.

Item-level audit. One clinician audited 60 items of an earlier version of the corpus: 16.7% of the items were rejected. We modified the pipeline to prevent the errors that were spotted: truncated answers were regenerated, and items whose answers never match the gold label are dropped (Section 2.2). Additionally, we use two LLM judges, Gemini 3.6 Flash and Gemma-4-31B, to check a sample of 200 items per dataset; they flag 9.7–12.2% of the sampled items (Appendix Tables 1 and 2).

2.4 Training

We finetune five fully-open base models from three families: Apertus-70B/8B-Instruct (Hernández-Cano et al., 2025), OLMo-2-32B-SFT (OLMo et al., 2024), EuroLLM-22B/9B-Instruct (Ramos et al., 2026), and one open-weight control, Gemma-3-27B-IT (Team et al., 2025), to enable a fairer comparison with open-weight MedGemma (Sellergren et al., 2025). We adapt these base models via supervised finetuning on the MeditronFO corpus while preserving each model's native instruction-tuning format. All finetunes use the same MeditronFO corpus, with answers distilled from gpt-oss-120b (Section 2). Full training infrastructure, optimizer configurations, and per-model hyperparameters are detailed in Appendix D.

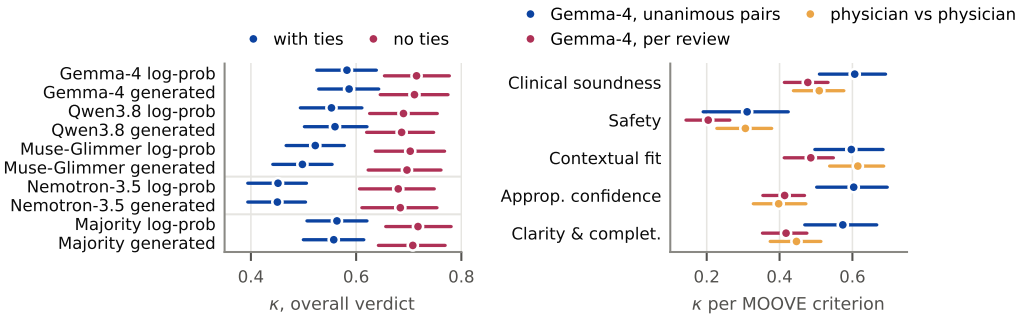

Figure 4: Judge–clinician agreement on the unanimous AutoMOOVE test set ($n = 576$ MOOVE pairs on which all physicians agreed). *Left:* Gemma-4 agrees most with the clinicians and Nemotron-3.5 least; *Majority* is the three-judge vote of Figure 6. *Right:* safety has low agreement both between clinicians and between judge and clinicians.

3 Experimental Setup

For each base, we report the base model and its MeditronFO finetune under the same evaluation settings. We also compare Gemma-3-27B-MeditronFO and Apertus-70B-MeditronFO with MedGemma-27B, the open-weight medical specialist built on Gemma-3-27B. For an overview of the MeditronFO evaluation datasets and evaluation details, including decoding settings, refer to Appendix E.

3.1 Static benchmarks

We evaluate medical knowledge on the test splits of MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019) and the validation split of MedMCQA (Pal et al., 2022) (its test labels are not public), using the held-out MedXpertQA (Zuo et al., 2025) as an out-of-distribution check. To check for catastrophic forgetting, we evaluate on MMLU-Pro (Wang et al., 2024), IFEval (Zhou et al., 2023), and ARC-Challenge (Clark et al., 2018) (Appendix G).

3.2 HealthBench

Medical MCQAs are poorly suited to evaluating models in real-world use, which is why we also evaluate on HealthBench (Arora et al., 2025). HealthBench is an open-ended clinical reasoning evaluation against structured, physician-authored rubrics. We use Gemma-4-31B-it (Team et al., 2026) as an LLM judge to evaluate the models against the rubrics; Appendix H.1 compares this grader with OpenAI's reported scores for gpt-oss-120b.

3.3 AutoMOOVE

MOOVE is a clinician-only platform where clinicians write adversarial vignettes and compare model responses and rate them on a set of established criteria, such as safety, contextual fit, clarity and completeness. This benchmark aims to automate this process with an LLM-as-a-judge. The automated evaluation provides pairwise preference battles between models to compare them on open-ended clinical vignettes.

Method. For each MOOVE adversarial vignette, we generate two model answers, and then pass this triplet in a prompt to our AutoMOOVE judge. The judge has two setups: one where it is prompted to give its overall clinical preference, and the other where it is prompted to give its preference for five clinical criteria: appropriate confidence, clinical soundness, safety, clarity and completeness, and contextual fit. For each preference, the judge is asked to answer either model 1, model 2 or equal. We evaluate two ways to extract the preference: generation and log-prob. Under log-prob, we compare the mean token log-probabilities of the three choices and take the highest as the preference; under generation, we extract it with a regex. We additionally show the judge both answer orders to reduce position bias; in both setups, the preference is Equal when the judge's choice flips with the

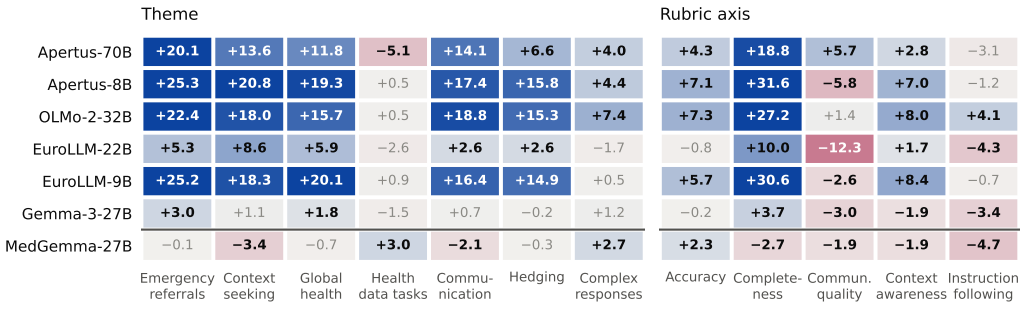

Figure 5: HealthBench Full gain of each MeditronFO finetune over its base, per theme and rubric axis (Gemma-4-31B grader). The last row compares MedGemma-27B with Gemma-3-27B.

order. We report the win rate of the first model: the share of questions it wins, with ties counting half. Main results use the majority vote of three judges, Gemma-4-31B, Qwen3.8-27B (Yang et al., 2025; Qiu et al., 2026) and Muse-Glimmer-30B (Equal when all three differ); the per-criterion and ablation results use Gemma-4-31B, the judge that agrees most with clinicians. Gemma-4-31B and Qwen3.8-27B come from families other than the teacher's, to reduce judge family bias. Gemma-4-31B however shares a family with Gemma-3-27B-MeditronFO and MedGemma-27B, and our third judge, Muse-Glimmer-30B, uses the chat format of the gpt-oss teacher, so we also report every comparison under each judge separately (Table 15). We present judge robustness evaluation in Appendix Figure 9. We use a test set filtered for unanimous clinician preference of 576 pairs, which cover 516 distinct vignettes: the 576 pairs validate the judge, and the 516 vignettes are the questions on which we compare models.

Clinician agreement. We validate the judge against existing blinded clinician comparisons from MOOVE (Figure 4) before using it to evaluate models. To measure alignment with clinicians we restrict the test set to pairs with full agreement between the two or more clinicians who reviewed it. On these 576 pairs, the Gemma-4-31B judge's agreement with the clinicians is moderate with ties (κ 0.583) and substantial without (κ 0.715, 495 pairs), on the scale of Landis & Koch (1977); the three-judge majority used for the main results is close (κ 0.564 and 0.718). Additionally, we measure the agreement of the judges with the clinicians on the clinical criteria preference, and find that except for safety, which sits in fair agreement with the clinicians (κ 0.311, close to the 0.306 agreement between clinicians), the criteria preferences of the judge have a moderate agreement with the clinicians (κ 0.574–0.606). Full validation methodology and per-criterion analysis are in Appendix I.2.

4 Results

4.1 Static benchmarks

We report benchmarking results on medical MCQA tasks in Figure 3, with the full table in Appendix F (Table 9). Finetuning on the MeditronFO corpus significantly raises the average of the four MCQA benchmarks for Apertus-70B ($+5.28$), OLMo-2-32B ($+4.92$) and Apertus-8B ($+3.32$); for EuroLLM-22B, EuroLLM-9B and Gemma-3-27B the change lies within the 95% interval ($+0.17$, -0.15, $+0.58$). Apertus-70B-MeditronFO reaches the highest scores among fully-open medical LLMs. We suspect our hyperparameters are poorly suited to the EuroLLM bases and leave this to future work. General-purpose capability decreases (Appendix G): the general-purpose average falls for four of the six finetunes, mostly through IFEval (-17.6 to -31.4 on Apertus-70B/8B, EuroLLM-9B and Gemma-3-27B), in line with other domain-specific finetuning recipes (Xie et al., 2025).

4.2 HealthBench

We report MeditronFO results on HealthBench-full in Figure 3 (right) and Figure 1, with the full table in Appendix F, and HealthBench-hard in Appendix H. All MeditronFO finetunes improve their score on HealthBench Full with Gemma-4-31B as judge. Apertus-70B-MeditronFO reaches 49.3%, up from 39.4% for Apertus-70B-Instruct; MedGemma-27B reaches 54.8%. Part of this gain comes

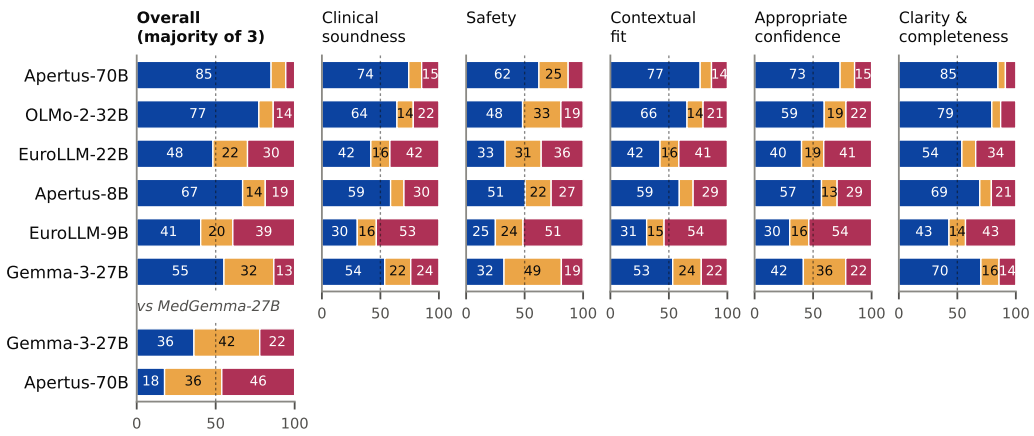

Figure 6: AutoMOOVE win/tie/loss shares of each MeditronFO model on the unanimous test set ($n = 516$): MeditronFO wins (blue), tie (orange), other model wins (red). *Overall*: majority vote of Gemma-4-31B, Qwen3.8-27B and Muse-Glimmer-30B (Equal when all three differ; log-prob, both answer orders) against its base (top) and MedGemma-27B (bottom). *Other panels*: against its base per MOOVE criterion (Gemma-4-31B). Dashed line: 50%.

from longer answers: after the length adjustment of Section 5.4, Apertus-70B-MeditronFO scores 40.3%, against 35.7% for its base and 48.1% for MedGemma-27B, and the gain remains significant for four of the six finetunes.

HealthBench grades every conversation with a theme and every rubric criterion with an axis. Figure 5 breaks the full-set gain down by theme and axis. The main gain is in completeness ($+18.8$ to $+31.6$ on Apertus-70B/8B, OLMo-2 and EuroLLM-9B; $+10.0$ on EuroLLM-22B, $+3.7$ on Gemma-3), followed by accuracy ($+4$ to $+7$ on the same four); communication quality falls on four finetunes and instruction following on two (in line with the IFEval drop). The largest rubric gains are emergency referrals ($+20$ to $+25$), context seeking and global health.

4.3 AUTOMOOVE

Figure 6 reports AutoMOOVE pairwise win rates of each MeditronFO finetune against its base on all 516 unanimous-set questions, under a majority vote of three judges (Gemma-4-31B, Qwen3.8-27B, Muse-Glimmer-30B); per-judge rates are in Table 15. Under the majority vote, five of the six finetunes are preferred over their base, from 59.2% (EuroLLM-22B) to 89.7% (Apertus-70B), and EuroLLM-9B is level with its base. The largest gain is on the largest fully-open base.

The criterion panels of Figure 6 break the Gemma-4-31B preference down by the five MOOVE criteria, using the per-criterion preference rubric validated against clinicians in Figure 4. Averaged over criteria, four finetunes are above parity: Apertus-70B (80.6%), OLMo-2-32B (72.0%), Apertus-8B (65.8%) and Gemma-3-27B (65.0%). EuroLLM-22B (51.7%) is level with its base and EuroLLM-9B (40.3%) is below it. The largest gains are on clarity and completeness. Safety has the most ties: the judge most often rates the two answers as equivalent on safety. Detailed AutoMOOVE pairwise results are in Appendix I.3.

We also compare against MedGemma-27B. Under the majority vote, Gemma-3-27B-MeditronFO reaches a 57.1% win rate and Apertus-70B-MeditronFO 35.8% (Figure 6, bottom). Per judge, Gemma-3-27B-MeditronFO reaches 58.8% under Gemma-4-31B and 52.8% [49.0–56.6] under Qwen3.8-27B; the second interval contains parity. Apertus-70B-MeditronFO reaches 41.4% and 28.9% respectively (its base, Apertus-70B-Instruct, reaches 3.4% and 3.0%; Appendix Table 16). Gemma-3-27B-MeditronFO is therefore level with or ahead of MedGemma-27B, depending on the judge, while Apertus-70B-MeditronFO loses to it under all three judges.

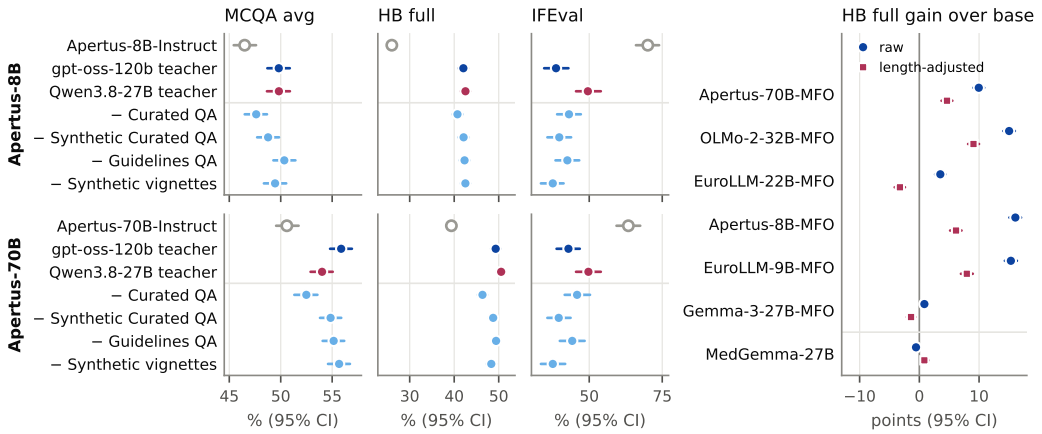

Figure 7: Ablations and the length-adjusted HealthBench gain. *Left:* Apertus-8B (top) and Apertus-70B (bottom) trained on the full corpus answered by gpt-oss-120b or by Qwen3.8-27B, and on the gpt-oss corpus with one component removed (light blue), all other training settings identical. *Right:* HealthBench full gain of each MeditronFO model over its base (MedGemma-27B: over Gemma-3-27B-IT), raw and after the length adjustment (Section 5.4). 95% bootstrap intervals over the benchmark items; they cover evaluation noise only. AutoMOOVE win rates, HB hard, and full tables in Appendix Tables 21 and 22 and Figures 16 and 8.

5 Analysis

5.1 Corpus-component ablations

The corpus-component ablations in Figure 7 (Appendix Table 21) identify which parts of the training mixture drive gains on structured medical QA versus open-ended clinical evaluation. At both model sizes, removing Curated QA has the largest negative impact. Only the Curated QA ablation is clearly significant; the other three stay within or at the edge of the confidence intervals. On AutoMOOVE, only removing the synthetic vignettes at 8B lowers the preference (Pref. -7.2 $[-11.3, -3.1]$); this does not replicate at 70B.

5.2 Second-Teacher Ablation

Figure 7 (teacher rows) compares the two teachers, gpt-oss-120b (reasoning low) and Qwen3.8-27B (thinking off). For Apertus-70B, gpt-oss is better on MCQA (-1.86 for Qwen3.8), and Qwen3.8 is better on HealthBench full ($+1.2$) and IFEval ($+6.8$); for Apertus-8B, both teachers give similar results except on MedMCQA and IFEval. Qwen3.8 answers are also longer (68% more training steps), which is one of the reasons we chose gpt-oss.

5.3 AutoMOOVE judge robustness

Additionally, we analyze each AutoMOOVE judge separately (Appendix I.3, Table 15). All three judges prefer the MeditronFO finetune over its base for Apertus-70B, Apertus-8B, OLMo-2-32B and Gemma-3-27B. For the other two bases the preference depends on the judge: EuroLLM-9B loses to its base under Qwen3.8-27B (34.2%), is level with it under Gemma-4-31B (50.5%), and only Muse-Glimmer-30B prefers it (70.7%). EuroLLM-22B is preferred over its base under Gemma-4-31B (59.0% [55.1–62.8]) and Muse-Glimmer-30B (66.7%), and level with it under Qwen3.8-27B (49.8% [45.9–53.6]). Robustness to tie rules and answer order is in Appendix I.5.

5.4 HealthBench & AutoMOOVE and Answer Length

MeditronFO finetunes answer at greater length than their bases (Appendix Figure 8a and Table 12). As a sensitivity check we apply the length adjustment implemented in OpenAI's simple-evals: 1.47 points are subtracted per 500 characters above 2,000 characters, and added below. About half of the full-set gain survives it (Figure 7, right): Apertus-70B, Apertus-8B, OLMo-2-32B and EuroLLM-9B,

all still significant. EuroLLM-22B and Gemma-3-27B become significant losses, and Gemma-3-27B-MeditronFO's lead over MedGemma-27B ($+1.4$) becomes -2.2.

We measure AutoMOOVE's length sensitivity as the correlation r between the judge's preference margin and the log word ratio of the two answers, over the 516 unanimous-set questions (Appendix Table 19). Between Gemma-3-27B-MeditronFO and MedGemma-27B both judges show a small length preference ($r = 0.13$ under Gemma-4-31B, 0.14 under Qwen3.8-27B). Against their bases, every finetune writes longer answers. The preference correlates positively with length for OLMo-2-32B and EuroLLM-22B under both judges and for Apertus-8B under Gemma-4-31B, whose preference over their bases is therefore partly a length effect. It is near zero for Gemma-3-27B and Apertus-70B (slightly negative under Qwen3.8-27B), and negative for EuroLLM-9B ($r = -0.14$).

6 Related Work

Medical LLMs. Publicly available medical LLMs are open-weight but not fully-open, including the Llama-based (Grattafiori et al., 2024) Meditron (Chen et al., 2023b; Sallinen et al., 2025) and Med42 (Christophe et al., 2024). HuatuoGPT-II (Chen et al., 2023a) unifies pretraining and finetuning into a single stage, while PMC-LLaMA (Wu et al., 2024) and BioMistral (Labrak et al., 2024) continue pretraining on biomedical corpora before instruction-tuning on aggregated QA benchmarks. Open-weight medical LLMs are not all equal in terms of openness, as the Aloe family goes further in terms of openness than MedGemma and releases training data alongside its alignment and red-teaming efforts, establishing a precedent towards more transparent medical LLMs (Garcia-Gasulla et al., 2026). Closed-data specialists, including the Med-PaLM family (Singhal et al., 2023; 2025), Med-Gemini (Saab et al., 2024) and MedGemma (Sellergren et al., 2025), report strong medical benchmark performance but disclose neither training corpora nor adaptation pipelines.

Open-ended and clinician-grounded evaluation. Recent work addresses the limitations of multiple-choice evaluation through rubric-based protocols: HealthBench (Arora et al., 2025) scores model responses against physician-authored rubrics across thousands of conversational scenarios, and LiveClin (Wang et al., 2026b) introduces a regularly updated benchmark to mitigate contamination. Pairwise preference evaluation has emerged as a complementary paradigm, in platforms such as Chatbot Arena (Zheng et al., 2023), which aggregates large-scale human pairwise judgments into model rankings. LLM judges offer a path to scalable pairwise evaluation: Cohen's κ is a more reliable validation metric than raw percent agreement (Thakur et al., 2025), and a human-likeness test situates a judge's κ within the distribution of per-rater κ values from a human panel (Han et al., 2025). The MOOVE platform (Sallinen et al., 2025) collects pairwise expert preferences over clinical vignettes; AutoMOOVE builds on this by automating the comparison protocol with an LLM-as-a-judge validated against human inter-rater agreement.

7 Discussion

Conclusion. We release the first fully-open development pipeline for clinical LLMs: all of the training data, the decontamination method and results, the label check methods and audit results, and a clinician-validated method for open-ended clinical evaluation. The resulting corpus provides a reusable foundation for training and evaluating future fully-open medical models. Among existing medical LLMs, the MeditronFO models are the only ones built on a fully-open base that release their medical data, the data pipeline and the training code (Appendix Table 23).

Limitations and future directions. The AutoMOOVE agreement with the clinicians' evaluations could be improved by training a judge on the clinicians' evaluations (Yu et al., 2025) or testing a multi-judge setup (Bedi et al., 2026). MeditronFO models' answers are longer; future work should focus on length-controlled teacher answers. We observe general capability degradation; our second teacher, Qwen3.8, loses less, which future work could explore; another direction is multi-teacher distillation, as described in Mai-thinking-1 (The Microsoft AI Team, 2026). Future work should also focus on fully-open self-distillation. Two LLM judges still flag 9.7 to 12.2% of the sampled rows of the final corpus; future work should explore filtering them. RL has been shown to improve the safety and helpfulness of medical LLMs; future work should explore fully-open RL pipelines for medical LLMs (Jagadeesh et al., 2026; Liévin et al., 2026).

AI USE STATEMENT

LLMs in the method. LLMs are part of the pipeline and of the evaluation, as described in the paper: gpt-oss-120b generates the synthetic questions and distills all answers (Section 2); Qwen3.8-27B is the second teacher (Section 5.2); Gemma-4-31B, Qwen3.8-27B and Muse-Glimmer-30B are the AutoMOOVE judges, with further judges compared in Appendix I.2; Gemma-4-31B grades HealthBench; and Gemini 3.6 Flash and Gemma-4-31B screen the corpus (Appendix B). **AI assistance to the authors.** All authors used Claude as a coding assistant for the data, training, evaluation and plotting scripts, and for editing the text of the paper. The research ideas, the clinician audit, the prompt co-authoring by the clinician panel, and the MOOVE vignettes and preferences were done without AI. Generated code was reviewed and run by the authors, every number in the paper was traced to a results file, and citations were checked by hand. The authors take full responsibility for the content of this paper, including any part produced with AI assistance.

ETHICS STATEMENT

Human subjects. Clinicians wrote the MOOVE vignettes and gave the blinded pairwise preferences used to validate AutoMOOVE; four clinicians co-authored the generation prompts, and one clinician audited 60 corpus items. The vignettes are hypothetical cases, and no patient data is involved. The study received exemption from the [omitted for anonymization] and institutional IRB approval from [omitted for anonymization]. Clinicians provided informed consent and participated voluntarily without remuneration. Participation formed part of an academic consortium that co-owns study results. **Data release.** The licenses of the existing assets we use are listed in Appendix M. We release the 516 AutoMOOVE test vignettes. **Intended use and risks.** The models are not intended for clinical use; we provide no evidence for their real-world clinical workflow performance, patient outcomes, calibration, or deployment safety. Their risks are those of medical LLMs in general: confidently incorrect answers, biases inherited from the training data, and substitution for clinical judgment; MeditronFO models also write longer answers than their bases (Section 5.4). **Bias and fairness.** AfriMed-QA and the MOOVE vignettes, written by clinicians across diverse international settings, are our attempt to cover clinical contexts beyond high-income countries, but the evaluation benchmarks remain mostly in English and set in high-income countries. **Conflicts of interest and funding.** Conflicts of interest: none. Funding: withheld for blinded review.

REPRODUCIBILITY STATEMENT

Data: sources and licenses (Appendix M), the generation prompts verbatim (Appendix A), rejection-sampling distillation (Section 2) and decontamination (Appendix C). **Training:** infrastructure, per-model hyperparameters and seeds (Appendix D). **Evaluation:** benchmarks, filters and decoding settings (Appendix E), judge prompts and scoring (Appendix I.1) and judge validation (Appendix I.2). Our intervals are 95% bootstrap intervals over benchmark items and cover evaluation noise only; Appendix Table 22 gives two training runs of the same recipe to show the size of the training noise. **Artifacts:** the supplementary material contains all code and the outputs behind every figure and table; the corpus and model weights will be released with the camera-ready version. Small run-to-run differences come from continuous batching in the vLLM inference engine.

REFERENCES

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pp. 1–12. NIST, 2017.

Emre Can Acikgoz, Osman Batur İnce, Rayene Bench, Arda Anıl Boz, İlker Kesen, Aykut Erdem, and Erkut Erdem. Hippocrates: An open-source framework for advancing large language models in healthcare, 2024. URL https://arxiv.org/abs/2404.16621.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*, 2025.

Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2):618–626, 2025.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 333(4):319–328, 2025.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Holistic evaluation of large language models for medical tasks with medhelm. *Nature medicine*, 32(3):943–951, 2026.

Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589, 2026.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, et al. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*, 2023a.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

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

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, François Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19352–19374. Association for Computational Linguistics, 2025.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719. Association for Computational Linguistics, 2024.

Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, et al. The tripod-llm reporting guideline for studies using large language models. *Nature medicine*, 31(1):60–69, 2025.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, et al. The aloe family recipe for open and specialized healthcare llms. *npj Digital Medicine*, 2026.

Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Steve Han, Gilberto Titericz Junior, Tom Balough, and Wenfei Zhou. Judge's verdict: A comprehensive analysis of llm judge capability through human agreement. *arXiv preprint arXiv:2510.09738*, 2025.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, et al. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*, 2025.

Akshay V Jagadeesh, Rahul K Arora, Khaled Saab, Ali Malik, Mikhail Trofimov, Foivos Tsimpourlas, Johannes Heidecke, and Karan Singhal. Reinforcement learning towards broadly and persistently beneficial models. *arXiv preprint arXiv:2606.24014*, 2026.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577. Association for Computational Linguistics, 2019.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86, 2025.

Yusuf Kesmen, Fay Elhassan, Jiayi Ma, Julien Stalhandske, Yena Chang, David Sasu, Alexandra Kulinkina, Akhil Arora, Lars Klein, and Mary-Anne Hartley. Mobayes: A modular bayesian framework for separating reasoning from language in conversational clinical decision support. *arXiv preprint arXiv:2604.20022*, 2026.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pp. 5848–5864. Association for Computational Linguistics, 2024.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

HongYi Li, Jun-Fen Fu, and Andre Python. Implementing large language models in health care: clinician-focused review with interactive guideline. *Journal of medical Internet research*, 27: e71916, 2025.

Valentin Liévin, Samuel Schmidgall, Tim Strother, Alex Bijamov, Akshay Goel, Anil Palepu, Chunjong Park, Vahid Balazadeh, Min Woo Sun, Marius Guerard, et al. Residencyrl: Reinforcement learning in simulated clinical environments. *arXiv preprint arXiv:2608.07418*, 2026.

Munib Mesinovic, Peter Watkinson, and Tingting Zhu. Explainability in the age of large language models for healthcare. *Communications Engineering*, 4(1):128, 2025.

Bertalan Meskó and Eric J Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

Amy C Moreno and Danielle S Bitterman. Toward clinical-grade evaluation of large language models. *International journal of radiation oncology, biology, physics*, 118(4):916–920, 2024.

Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Chinemelu Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, et al. Afrimed-qa: a pan-african, multi-specialty, medical question-answering benchmark dataset. *arXiv preprint arXiv:2411.15640*, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of qwen3. 8-next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

Miguel Moura Ramos, Duarte M Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M Guerreiro, Ricardo Rei, Nicolas Boizard, et al. Eurollm-22b: Technical report. *arXiv preprint arXiv:2602.05879*, 2026.

David Restrepo, Chenwei Wu, Zhengxu Tang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Cong-Tinh Dao, Jack Gallifant, Robyn Gayle Dychiao, Jose Carlo Artiaga, et al. Multi-ophthalingua: A multilingual benchmark for assessing and debiasing llm ophthalmological qa in lmics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28321–28330, 2025.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David GT Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, et al. Advancing conversational diagnostic ai with multimodal reasoning. *Nature Medicine*, pp. 1–11, 2026.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, et al. Llama-3-meditron: An open-weight suite of medical llms based on llama-3.1. *AAAI*, 2025.

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

Yanjun Shao, Xiangru Tang, Jiwoong Sohn, Jiapeng Chen, Yuxuan Liao, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, et al. Medicalagentsbench for complex medical reasoning: Comparing internalized reasoning models versus externalized agent-based frameworks. *Patterns*, 2025.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 404–430, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9.

The Microsoft AI Team. MAI-Thinking-1: Building a hill-climbing machine. Technical report, Microsoft AI, 2026. URL https://microsoft.ai/wp-content/uploads/2026/06/main_20260602_2.pdf.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Wenxuan Wang, Zizhan Ma, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Wenting Chen, and Linlin Shen. Beyond the leaderboard: Rethinking medical benchmarks for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43078–43123, 2026a.

Xidong Wang, Shuqi Guo, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, and Benyou Wang. Liveclin: A live clinical benchmark without leakage. *arXiv preprint arXiv:2602.16747*, 2026b.

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

Haider J Warraich, Troy Tazbaz, and Robert M Califf. Fda perspective on the regulation of artificial intelligence in health care and biomedicine. *Jama*, 333(3):241–247, 2025.

Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Yanglet Liu, Ahmed Abdelmonsef, Sachin Varghese, and Arnaud Le Hors. The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence. *arXiv preprint arXiv:2403.13784*, 2024.

Brigitte Fong Yeong Woo, Kenrick Cato, Hannah Cho, Sang Bin You, and Jiyoun Song. The use of large language models in clinical documentation: A scoping review. *International Journal of Nursing Studies*, pp. 105322, 2025.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

Eric Wu, Kevin Wu, Jason Hom, Paul H Yi, Angela Zhang, Alejandro Lozano, Jeff Nirschl, Jeff Tangney, Kevin Byram, Braydon Dymm, et al. Medarena: Comparing llms for medicine-in-the-wild clinician preferences. *arXiv preprint arXiv:2603.15677*, 2026.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine*, 8:141, 2025. doi: 10.1038/s41746-025-01533-1.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiaxu Yan, Kaidong Yu, and Xuelong Li. Improve llm-as-a-judge ability as a general ability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14110–14126, 2025.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36. Neural Information Processing Systems Foundation, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

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

APPENDIX CONTENTS

A Synthetic data generation prompts **17**
A.1 Guidelines QA prompt . . . 17
A.2 Synthetic Curated QA prompt . . . 17
A.3 Synthetic vignettes prompt . . . 18

B Item-level audit of the synthetic data **19**

C Decontamination details **19**

D Training details **20**
D.1 Infrastructure and framework . . . 20
D.2 Common training settings . . . 21
D.3 Per-model hyperparameters . . . 21
D.4 Reproducibility artifacts . . . 21
D.5 Compute resources . . . 21

E Evaluation datasets **22**

F Medical benchmark results **22**

G General-purpose benchmarks **22**

H HealthBench **23**
H.1 Grader against OpenAI's reported scores . . . 23
H.2 Answer length . . . 24

I AutoMOOVE **25**
I.1 Judge prompts . . . 25
I.2 Validation against clinicians . . . 27
I.3 Pairwise results . . . 29
I.4 Response length . . . 32
I.5 Judge robustness . . . 33

J Corpus-component ablations **34**

K Second-teacher ablation **34**

L Openness comparison across medical LLMs **35**

M Licenses of existing assets **35**

A SYNTHETIC DATA GENERATION PROMPTS

This appendix documents the exact prompts used by the three synthetic data generation pipelines (Synthetic Curated QA, Guidelines QA, and Synthetic vignettes). Each pipeline uses a component-specific developer message, followed by a user message that injects either few-shot exemplars or a source guideline. Prompts are reproduced verbatim from the generation scripts; placeholders such as `{guideline_text}` and the example slots are filled at runtime.

A.1 GUIDELINES QA PROMPT

The Guidelines QA pipeline seeds generation with one full clinical practice guideline per call and elicits ten multiple-choice vignettes grounded strictly in that document.

Developer Message — Guidelines QA

```
You are an expert medical educator and physician tasked with creating
high-quality, clinically accurate content for a medical exam. Your task
is to generate clinical vignette-style questions along with its correct
answer, based STRICTLY on the provided medical guideline. Focus on
realistic patient presentations (age, symptoms, physical exam findings),
identifying 'red flags', and diagnostic reasoning highlighted in the
text. The timeline and objective progress should always be clear and
detailed in the vignettes. Include clear context about site and where
people travelled etc. Do not include outside information or unproven
treatments.
```

User Message — Guidelines QA

```
Here is the medical guideline:

=== GUIDELINE START ===
{guideline_text}
=== GUIDELINE END ===

Based ONLY on the guideline above, generate exactly 10 unique
MULTIPLE-CHOICE clinical vignette questions and their answers. Each
question should present a realistic patient scenario that tests the
diagnostic or management principles in the text. For each vignette,
provide 4-5 plausible multiple-choice options (A-E). Ensure distractors
represent common diagnostic pitfalls or 'next best steps' that are
incorrect based strictly on the provided guideline.

You MUST format EACH of the 10 items exactly as follows, using these
specific XML tags:

<qa>
<question>
Patient scenario and the specific question here.
A) [Option 1]
B) [Option 2]
C) [Option 3]
D) [Option 4]
</question>
<answer>The rationale explaining your chain of thought without
mentioning the guideline and then Answer: correct answer</answer>
</qa>
```

A.2 SYNTHETIC CURATED QA PROMPT

The Synthetic Curated QA pipeline samples five exemplars without replacement from the curated benchmark pool and produces a single new QA pair per call. The pool is partitioned into labeled (multiple-choice, carrying a `label_letter`) and unlabeled (open-ended) buckets, and the user message is specialized accordingly.

Developer Message — Synthetic Curated QA

```
You are an expert medical educator and physician tasked with creating
high-quality, clinically accurate content for a medical exam. Your task
is to generate a new, unique, clinical vignette-style question along
with its evidence-based correct answer. The timeline and progress
should always be clear and detailed in the vignettes. Include clear
context about site and where people travelled etc. The content must
reflect realistic clinical scenarios, standard-of-care protocols, and
well established medical consensus. Avoid scientifically controversial
treatments. You will be provided with 5 examples. Use them strictly to
understand the desired format, diagnostic difficulty, and clinical
depth. DO NOT copy them. Generate a completely new, scientifically
rigorous question that would be unconditionally approved by a medical
review board.
```

User Message — Synthetic Curated QA

```
Here are example questions and answers to model your format on:

--- Example 1 ---
<question>{question_1}
<answer>{answer_1}

--- Example 2 ---
<question>{question_2}
<answer>{answer_2}

... (5 examples total) ...

# If sampled from the labeled bucket:
Now generate a brand new, unique, and clinically accurate
MULTIPLE-CHOICE medical question and its detailed answer.
Ensure the answer matches the formatting tags above.

# If sampled from the unlabeled bucket:
Now generate a brand new, unique, and clinically accurate
OPEN-ENDED medical question and its detailed answer.
Ensure the answer matches the formatting tags above.
```

A.3 SYNTHETIC VIGNETTES PROMPT

The Synthetic vignettes pipeline samples five exemplar prompts without replacement from the MOOVE training split and generates a single new open-ended clinical scenario per call. Only the question stem is generated; assistant responses are produced downstream.

Developer Message — Synthetic vignettes

```
You are an expert medical educator and physician tasked with creating
high-quality, clinically accurate content. Your task is to generate a
new, unique, and realistic medical scenario or question prompt. The
content must reflect realistic clinical presentations, inquiries from
colleagues, or patient encounters. The timeline and objective progress
should always be clear and detailed. Include clear context about site
and where people travelled etc. You will be provided with 5 examples.
Use them strictly to understand the desired format, diagnostic
difficulty, and clinical depth. DO NOT copy them. Generate a completely
new question that would be unconditionally approved by a medical review
board.
```

User Message — Synthetic vignettes

```
Here are example prompts to model your format and clinical depth on:

--- Example 1 ---
<question>
{prompt_1}
</question>

--- Example 2 ---
<question>
```

```
{prompt_2}
</question>

... (5 examples total) ...

Now, acting as an expert medical educator, generate a brand new, unique,
and clinically accurate medical scenario or question. Wrap your
generated scenario strictly within <question> and </question> tags.
```

B ITEM-LEVEL AUDIT OF THE SYNTHETIC DATA

Table 1: Clinician audit, 60 items of the earlier corpus: items flagged per category. The last columns compare two LLM judges' flags with the clinician's on the same items: flags in common out of the clinician's, and Cohen κ.

	Clinician				Gemini 3.6 Flash	Gemma-4-31B
Category	**Synth. QA**	**Guidel. QA**	**MOOVE**	**All (95% CI)**	**found, κ**	**found, κ**
Factual error	2	2	2	10.0% [5–20]	3/6, 0.23	3/6, 0.29
Unsafe recommendation	2	1	1	6.7% [3–16]	2/4, 0.46	2/4, 0.35
Ambiguous question	7	2	1	16.7% [9–28]	5/10, 0.52	3/10, 0.42
Hallucination	2	1	1	6.7% [3–16]	1/4, 0.20	1/4, 0.09
Label mismatch	not asked				–	–
Rejected	6	2	2	16.7% [9–28]	9/10*	8/10*

* Rejects caught by the screen "any criterion ≤ -1", which passes 33/50 (Gemini) and 39/50 (Gemma) of the accepted items. On the five MOOVE criteria the judges agree with the clinician at weighted κ 0.61 (Gemini) and 0.47 (Gemma).

Table 2: LLM screen of the MeditronFO corpus. % of a random 200 rows per training set flagged in each category, Gemini 3.6 Flash / Gemma-4-31B as judges (Gemma-4-31B graded 199 of the 200 rows for Curated QA and Guidelines QA). *Flagged*: any MOOVE criterion ≤ -1. The category columns are separate flags set by the judge, not a split of *Flagged*, so a category can exceed it. *Corpus*: sets weighted by rows, 95% CI. Intervals on one set are about ± 4 points at 10%.

Set	**Flagged**	**Factual**	**Unsafe**	**Ambiguous**	**Halluc.**	**Label**
Curated QA	7.5 / 9.0	5.0 / 6.5	1.0 / 0.5	3.0 / 4.0	1.0 / 3.5	1.0 / 4.0
Synthetic Curated QA	12.0 / 7.5	9.5 / 6.5	1.5 / 4.5	2.0 / 0.5	4.5 / 5.0	4.0 / 3.5
Guidelines QA	17.5 / 14.1	14.0 / 13.1	3.0 / 9.5	6.0 / 3.5	8.0 / 14.1	5.0 / 4.0
Synthetic vignettes	24.0 / 9.0	26.0 / 8.0	14.0 / 6.5	0.0 / 0.5	9.0 / 4.5	–
Corpus, Gemini	12.2 [9.8–14.7]	9.7 [7.5–11.8]	2.2 [1.3–3.2]	3.2 [1.9–4.6]	4.3 [2.8–5.7]	3.0 [1.7–4.3]
Corpus, Gemma	9.7 [7.5–12.0]	8.2 [6.1–10.2]	4.4 [2.9–5.8]	2.5 [1.3–3.7]	6.6 [4.8–8.5]	3.7 [2.2–5.2]

Table 3: Multiple-choice rows whose label the answering model never reached in 8 samples (pass@8), all rows, % with 95% Wilson CI. *Confident*: all recorded samples chose the same other letter.

		gpt-oss-120b		Qwen3.8-27B	
Set (labels by)	**MCQ rows**	**Missed**	**Confident**	**Missed**	**Confident**
Curated QA (humans)	194 496	8.4 [8.3–8.6]	6.6	5.6 [5.5–5.7]	4.3
Synthetic QA (generator)	180 725	7.7 [7.6–7.8]	6.3	7.2 [7.1–7.3]	6.0
Guidelines QA (generator)	141 413	8.2 [8.1–8.4]	6.2	6.2 [6.1–6.3]	4.3

C DECONTAMINATION DETAILS

We apply a two-stage n-gram and token-alignment decontamination pipeline adapted from Apertus (Hernández-Cano et al., 2025) to the full curated corpus. The reference set aggregates the prompts of all evaluation benchmarks used in this work: MedQA, MedMCQA, PubMedQA, MedXpertQA,

MMLU-Pro, IFEval, ARC-Challenge and HealthBench. The synthetic vignettes are also checked against the MOOVE unanimous set used by AutoMOOVE. MedExpQA (434 rows) is removed from Curated QA before decontamination.

Training samples are tokenized with the Apertus tokenizer. In the first stage, samples sharing any 8-gram with a reference prompt are flagged as candidates. In the second stage, each candidate is token-aligned against the matched reference and removed if the normalized alignment difference is at most $\tau = 0.5$. This filters incidental n-gram overlaps while still catching lightly paraphrased contaminations. For each dataset, the pipeline outputs a decontaminated corpus and a report logging removed samples and their matched references.

Table 4 gives the removals per training set and benchmark.

Table 4: Rows removed by decontamination, per training set and benchmark. A row can match several benchmarks, so the benchmark columns can add up to more than the rows removed. *Rows before*: after dropping MedExpQA's 434 rows from Curated QA. *Other*: MMLU-Pro, ARC, IFEval, HealthBench and the unanimous set).

Training set	Rows before	MedQA	MedMCQA	PubMedQA	MedXpertQA	Other	Removed
Curated QA	216 285	126	2856	1000	58	148	4113
Synthetic Curated QA	214 648	0	36	0	0	23	59
Guidelines QA	145 367	0	0	0	0	7	7
Synthetic vignettes	24 061	0	0	0	0	18	18
Total	600 361	126	2892	1000	58	196	4197

Table 5 gives the rows left after each filtering step, from the corpus before decontamination to the final training set.

Table 5: Rows at each filtering step of the MeditronFO corpus, from the rows before decontamination (Table 4) to the training rows. *Malformed*: rows dropped before answering: multiple-choice questions without options (11 Curated QA, 16,131 Synthetic Curated QA, 6 Guidelines QA), questions missing from the row (202 Curated QA) and unusable questions (1 Curated QA, 3,941 Guidelines QA). *pass@8 miss*: multiple-choice rows whose gold label gpt-oss-120b never reached in 8 samples (Table 3); a truncated sample counts as a miss. Synthetic vignettes have no gold label and are not filtered. Guidelines QA has more rows than ten per guideline because the model does not always return exactly ten items and we keep every item it returns: 30% of the 14,192 guidelines yield 11–13 items, 2% fewer than ten and 71 none.

Training set	Rows before	Decontam.	Malformed	pass@8 miss	Final
Curated QA	216 285	−4 113	−214	−16 420	195 538
Synthetic Curated QA	214 648	−59	−16 131	−13 918	184 540
Guidelines QA	145 367	−7	−3 947	−11 646	129 767
Synthetic vignettes	24 061	−18	–	–	24 043
Total	600 361	−4 197	−20 292	−41 984	533 888

D TRAINING DETAILS

D.1 INFRASTRUCTURE AND FRAMEWORK

All MeditronFO models are trained on a high-performance computing cluster using NVIDIA GH200 Grace Hopper Superchip nodes with 4 GPUs per node. Apertus-70B, OLMo-2-32B and Gemma-3-27B are trained on 16 nodes (64 GH200 GPUs), EuroLLM-22B on 8 nodes (32 GPUs), and the small bases (Apertus-8B, EuroLLM-9B) on 2 nodes (8 GPUs) (Table 6).

Training used the Axolotl framework with PyTorch's `torchrun` launcher and `c10d` rendezvous. The Apertus runs (70B and 8B) used DeepSpeed ZeRO Stage 3 for memory partitioning; all other runs used PyTorch FSDP v2 with transformer-block auto-wrap, sharded state-dict checkpointing, reshard-after-forward, and activation checkpointing. Apertus-70B, OLMo-2-32B, and EuroLLM-22B/9B additionally used the cut-cross-entropy plugin to reduce activation memory at the loss-computation step. All runs used Flash Attention 2 and bfloat16 mixed-precision training.

D.2 COMMON TRAINING SETTINGS

To preserve the alignment work invested in each base, we maintain the instruction-tuning chat template native to each model (ChatML for EuroLLM via explicit override; native templates for all others). All runs share the following settings unless noted otherwise:

- **Sequence length:** 8192 tokens with sample packing. OLMo-2-32B and EuroLLM-9B have a 4,096-token context; packing resets positions per example, so only examples longer than 4,096 tokens (about 2% of rows, mostly synthetic vignettes) go beyond it.
- **Optimizer:** AdamW (fused implementation), $\beta_1 = 0.9$, $\beta_2 = 0.999$ (default) unless otherwise stated.
- **LR scheduler:** cosine decay with warmup.
- **Gradient clipping:** max gradient norm 1.0.
- **Random seeds:** 42 for both model initialization and data shuffling.

D.3 PER-MODEL HYPERPARAMETERS

Per-model settings are summarized in Table 6. Hyperparameters are selected based on each base model's published instruction-tuning recipe where available. Gemma-3-27B is trained with a lower learning rate (1e-6), $\beta_2 = 0.95$ and weight decay 0.1, because Gemma-3 is more sensitive to hyperparameters than the other bases.

Table 6: Per-model training hyperparameters. "Eff. batch" is the effective batch size in sequences (micro-batch $\times$ gradient accumulation $\times$ GPUs). All runs use fused AdamW, cosine LR schedule, sequence length 8192 with sample packing, and seed 42; Gemma-3-27B also offloads activations.

Model	Nodes (GPUs)	Parallelism	Epochs	μBS	Accum.	Eff. batch	LR	Warmup	WD	β_2
Apertus-70B-Instruct	16 (64)	ZeRO-3	1	1	1	64	1e-5	0.10	0.0	0.999
OLMo-2-32B-SFT	16 (64)	FSDP v2	2	1	2	128	5e-6	0.03	0.0	0.999
EuroLLM-22B-Instruct	8 (32)	FSDP v2	2	1	4	128	1e-5	0.10	0.01	0.999
Gemma-3-27B-IT	16 (64)	FSDP v2	2	1	2	128	1e-6	0.10	0.1	0.95
Apertus-8B-Instruct	2 (8)	ZeRO-3	1	2	4	64	1e-5	0.10	0.0	0.999
EuroLLM-9B-Instruct	2 (8)	FSDP v2	1	2	8	128	1e-5	0.10	0.01	0.999

D.4 REPRODUCIBILITY ARTIFACTS

We release the full Axolotl YAML configuration files for each model, the SLURM submission script, and the data preparation pipeline alongside the corpus.

D.5 COMPUTE RESOURCES

Table 7: Compute for the main MeditronFO training runs. Wall-clock time of the completed attempt of each training job, from the cluster's accounting; preempted attempts before it (at most 40 minutes per run) are not counted. GPU-hours = GPUs $\times$ wall-clock hours.

Model	Nodes (GPUs)	Wall-clock (h:mm)	GPU-hours
Apertus-70B-MeditronFO	16 (64)	3:06	199
OLMo-2-32B-MeditronFO	16 (64)	3:31	225
EuroLLM-22B-MeditronFO	8 (32)	4:01	129
Gemma-3-27B-MeditronFO	16 (64)	3:25	219
Apertus-8B-MeditronFO	2 (8)	0:51	7
EuroLLM-9B-MeditronFO	2 (8)	3:07	25
Total (main runs)			804

Table 9: Overall medical benchmarks results for MeditronFO models against their base (best results bolded)

Model	MedMCQA	MedQA	PubMedQA	MedXpertQA	HealthBench	MCQA avg	Avg (5)	MCQA gain
Fully-open base								
Apertus-70B-Instruct	52.8	61.9	74.7	13.0	39.4	50.60	48.35	–
+ MeditronFO	60.2	71.1	74.9	17.3	49.3	55.88	54.56	+5.28 [+4.2, +6.4]
OLMo-2-32B-SFT	59.0	65.8	71.0	12.4	26.6	52.06	46.96	–
+ MeditronFO	60.8	72.3	76.9	17.9	41.6	56.98	53.91	+4.92 [+3.8, +6.1]
EuroLLM-22B-Instruct	55.8	65.3	76.4	14.6	39.9	53.01	50.40	–
+ MeditronFO	55.2	64.7	**77.9**	14.9	43.5	53.18	51.23	+0.17 [−0.9, +1.2]
Small fully-open base								
Apertus-8B-Instruct	47.8	52.1	74.3	11.8	26.0	46.49	42.39	–
+ MeditronFO	52.0	60.7	71.4	15.1	42.1	49.81	48.26	+3.32 [+2.1, +4.5]
EuroLLM-9B-Instruct	45.9	48.0	73.9	10.5	18.7	44.56	39.38	–
+ MeditronFO	47.6	50.4	66.0	13.6	34.0	44.41	42.32	−0.15 [−1.3, +1.1]
Open-weight base								
Gemma-3-27B-IT	64.5	76.9	76.5	16.9	55.4	58.68	58.02	–
+ MeditronFO	63.7	78.2	76.7	18.4	**56.2**	59.26	58.65	+0.58 [−0.3, +1.4]
MedGemma-27B	**73.7**	**87.4**	77.6	**25.2**	54.8	**65.99**	**63.75**	+7.30 [+6.4, +8.2]

E EVALUATION DATASETS

Table 8: Overview of MeditronFO evaluation datasets.

Medical evaluation			General purpose	
Task	**Dataset**	**No. examples**	**Dataset**	**No. examples**
Medical QA (MCQ)	MedQA	1,273	MMLU-Pro	12,032
	MedMCQA	4,183	IFEval	541
	PubMedQA	500	ARC-Challenge	1,172
	MedXpertQA	2,450		
Open-ended clinical	AutoMOOVE (unanimous set)	516		
Open-ended rubric	HealthBench (full)	5,000		

Decoding. Multiple-choice and general-purpose benchmarks are decoded greedily (temperature 0, seed 42). Open-ended answers are sampled at temperature 0.7, the setting of HealthBench's reference evaluation, with seed 0 for HealthBench and seed 42 for AutoMOOVE. The HealthBench grader runs at temperature 0.1. AutoMOOVE judges decode at temperature 0.7 with seed 42; this only affects generated verdicts, since log-prob verdicts read the scores of the three fixed completions.

F MEDICAL BENCHMARK RESULTS

G GENERAL-PURPOSE BENCHMARKS

We treat general-purpose benchmarks as a smoke test for catastrophic forgetting rather than as a primary optimization target. Domain adaptation is expected to trade some broad instruction-following capability for medical specialization (Xie et al., 2025). Table 10 reports MMLU-Pro (5-shot chain of thought), IFEval (prompt-level strict) and ARC-Challenge (`acc_norm`) for every base, its MeditronFO finetune and MedGemma-27B.

Table 10: General-purpose benchmark accuracy (%). *Avg*: mean of the three benchmarks. *Gain*: Avg over the base, with a paired 95% bootstrap interval (each benchmark resampled once for both models); the MedGemma gain is measured against Gemma-3-27B-IT. *Unparsed*: MMLU-Pro answers the harness's filter could not read (it takes only "the answer is (X)"), each scored wrong. Best value per column in bold. A single score's 95% interval is about ± 0.9 on MMLU-Pro, ± 4 on IFEval, ± 2.8 on ARC-Challenge and ± 1.7 on the average.

Model	MMLU-Pro	IFEval	ARC-C	Avg	Gain	Unparsed
Fully-open base						
Apertus-70B-Instruct	43.2	63.4	54.3	53.63	–	13.8
+ MeditronFO	41.1	42.9	44.0	42.68	-10.94 [-12.7, -9.3]	5.3
OLMo-2-32B-SFT	41.8	69.3	56.7	55.93	–	5.8
+ MeditronFO	45.1	71.9	55.0	57.34	$+1.41$ [-0.0, $+2.9$]	3.9
EuroLLM-22B-Instruct	31.8	71.2	52.7	51.88	–	40.7
+ MeditronFO	48.0	65.6	51.8	55.12	$+3.24$ [$+1.8$, $+4.7$]	2.0
Small fully-open base						
Apertus-8B-Instruct	34.2	70.1	53.7	52.66	–	11.9
+ MeditronFO	34.5	38.6	44.5	39.22	-13.43 [-15.2, -11.7]	10.1
EuroLLM-9B-Instruct	27.6	59.2	49.9	45.57	–	23.3
+ MeditronFO	31.1	41.6	46.3	39.68	-5.89 [-7.7, -4.1]	11.8
Open-weight base						
Gemma-3-27B-IT	**65.3**	**74.7**	60.4	**66.79**	–	1.5
+ MeditronFO	62.9	54.7	**63.2**	60.29	-6.50 [-7.9, -5.1]	3.8
MedGemma-27B	62.6	73.4	58.8	64.91	-1.88 [-3.2, -0.6]	5.5

IFEval carries the loss. The general-purpose average falls for four of six finetunes, by 10.9 points for Apertus-70B (53.63 $\rightarrow$ 42.68) and 13.4 for Apertus-8B, and most of the fall is IFEval: -20.5 at Apertus-70B, -31.4 at Apertus-8B, -20.0 at Gemma-3-27B, -17.6 at EuroLLM-9B and -5.5 at EuroLLM-22B. Only OLMo-2-32B holds ($+2.6$). A Qwen3.8-27B teacher recovers about a third of the IFEval loss (Table 22: 49.7 and 49.5 against 42.9 and 38.6), so part of the loss comes from the teacher's answer style. ARC-Challenge falls by about 10 points for both Apertus finetunes and by 3.6 for EuroLLM-9B; the other three are within noise or above their base.

The MMLU-Pro gains are mostly answer format. The harness reads only "the answer is (X)", and the bases that miss this format most gain most: EuroLLM-22B-Instruct leaves 40.7% of its answers unparsed and its finetune 2.0%, for a $+16.2$ gain. Where the unparsed share moves by 2 points or less, the gain is small. MMLU-Pro is therefore not evidence that MeditronFO improves general knowledge, and the averages of OLMo-2-32B ($+1.4$, interval touching zero) and EuroLLM-22B ($+3.2$) rise mainly through it.

Medical specialization costs general capability, and our recipe pays more on the same base. MedGemma-27B also scores below its base Gemma-3-27B-IT (-1.9, -3.2 to -0.6), but Gemma-3-27B-MeditronFO loses more (-6.5).

H HealthBench

H.1 Grader against OpenAI's reported scores

The original HealthBench grader is GPT-4.1 (Arora et al., 2025). We grade with Gemma-4-31B, so our scores are not directly comparable to published ones. gpt-oss-120b, our teacher, is the one model with both our grade and OpenAI's: its model card (Agarwal et al., 2025) reports HealthBench full, Hard and Consensus at three reasoning efforts. We run it at low effort, with 4096 tokens for reasoning and answer, at temperature 0.7 and without a system prompt. Hard (1000 conversations) and Consensus (3671 conversations, cluster-level criteria only) are subsets of the full set, so we rescore them from the per-conversation grades of the full run (Table 11).

Against OpenAI's low-effort scores, Gemma-4-31B grades gpt-oss-120b 6.9 points higher on HealthBench full and 11.6 points higher on Hard. On Consensus it grades 6.4 points lower. Our low-effort run scores above OpenAI's high-effort one on both full and Hard, so the difference comes from the grader, not the generation. Gemma-4-31B is more lenient than GPT-4.1 on the specific example-level rubric items. It is stricter on the 34 holistic consensus criteria and marks 83.9% of the consensus-

criterion judgements met (a Consensus score of 84.2). Qwen3-30B-A3B is lenient on both: 16.3 points above OpenAI on full and 27.4 on Hard. So the grader offset differs by criterion type and cannot be removed with one constant. We therefore compare models only under one grader and do not set our scores against published HealthBench numbers. Comparisons within the paper all use Gemma-4-31B and are not affected.

Table 11: gpt-oss-120b on HealthBench (%). OpenAI: Table 3 of the gpt-oss model card, GPT-4.1 grader. Ours: the same model at low reasoning effort (4096 tokens), with the answers graded by Gemma-4-31B and regraded by Qwen3-30B-A3B; Hard and Consensus rescored from the full run. 95% bootstrap intervals over conversations. Δ: Gemma-4-31B minus OpenAI at low effort.

Grader	Reasoning effort	Full	Hard	Consensus
GPT-4.1 (OpenAI)	low	53.0	22.8	90.6
GPT-4.1 (OpenAI)	medium	55.9	26.9	90.8
GPT-4.1 (OpenAI)	high	57.6	30.0	89.9
Gemma-4-31B (ours)	low	59.9 [58.9, 60.9]	34.4 [32.1, 36.6]	84.2 [83.3, 85.1]
Qwen3-30B-A3B (ours)	low	69.3 [68.5, 70.1]	50.2 [48.1, 52.1]	90.1 [89.4, 90.8]
Δ Gemma-4-31B $-$ OpenAI low		+6.9	+11.6	−6.4

H.2 Answer length

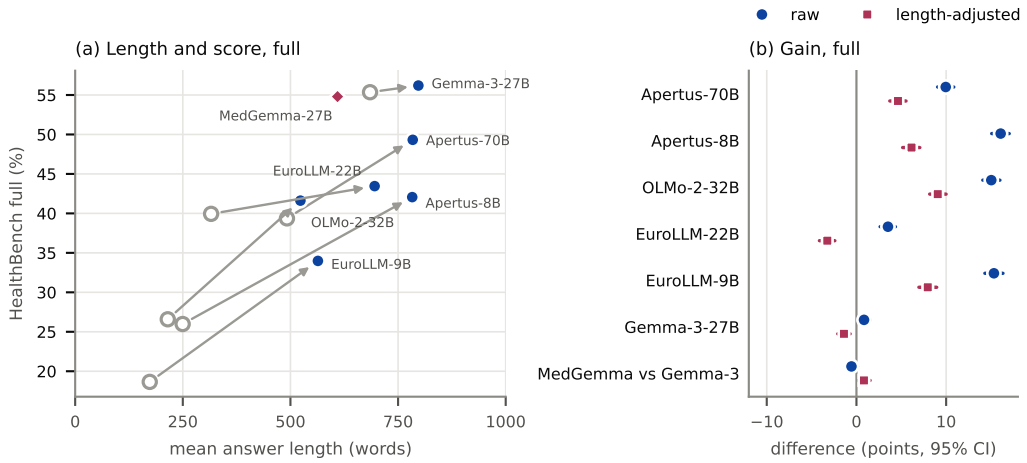

Figure 8: HealthBench gain and answer length (Gemma-4-31B grader). (a) Mean answer length and HealthBench full score, base (hollow) to MeditronFO finetune. (b) Gain of each finetune over its base on HealthBench full, raw and after the length adjustment (Section 5.4); panel (b) is the right panel of Figure 7.

MeditronFO finetunes also write longer HealthBench answers than their bases. Table 12 gives the mean answer length of every base and finetune on HealthBench full and hard, with the raw score and the score after the length adjustment (Section 5.4).

Table 12: HealthBench mean answer length (characters, the unit of the length adjustment), score and length-adjusted score (%) per subset, Gemma-4-31B grader. *Adjusted*: the length adjustment of Section 5.4 (1.47 points per 500 characters from 2,000), applied per answer before the mean. Finetune rows: difference to the base in small type (bold: paired 95% bootstrap interval over conversations excludes 0). A 0.0 on hard is a mean below 0, clipped as in simple-evals.

	Full			**Hard**		
Model	Chars	Score	Adjusted	Chars	Score	Adjusted
Apertus-70B-Instruct	3,250	39.4	35.7	3,371	12.1	8.1
+ MeditronFO	5,061 $+\mathbf{1{,}810}$	49.3 $+\mathbf{10.0}$	40.3 $+\mathbf{4.6}$	4,725 $+\mathbf{1{,}355}$	22.1 $+\mathbf{10.0}$	14.1 $+\mathbf{6.0}$
Apertus-8B-Instruct	1,644	26.0	27.0	1,630	0.0	0.0
+ MeditronFO	5,023 $+\mathbf{3{,}380}$	42.1 $+\mathbf{16.1}$	33.2 $+\mathbf{6.1}$	4,764 $+\mathbf{3{,}134}$	15.7 $+\mathbf{15.7}$	7.6 $+\mathbf{7.6}$
OLMo-2-32B-SFT	1,410	26.6	28.3	1,357	0.0	0.6
+ MeditronFO	3,440 $+\mathbf{2{,}030}$	41.6 $+\mathbf{15.0}$	37.4 $+\mathbf{9.1}$	3,190 $+\mathbf{1{,}832}$	11.8 $+\mathbf{11.8}$	8.3 $+\mathbf{7.7}$
EuroLLM-22B-Instruct	2,233	39.9	39.3	2,165	9.1	8.6
+ MeditronFO	4,534 $+\mathbf{2{,}301}$	43.5 $+\mathbf{3.5}$	36.0 $-\mathbf{3.3}$	4,300 $+\mathbf{2{,}135}$	15.1 $+\mathbf{6.0}$	8.3 -0.3
EuroLLM-9B-Instruct	1,142	18.7	21.2	1,047	0.0	0.0
+ MeditronFO	3,651 $+\mathbf{2{,}509}$	34.0 $+\mathbf{15.3}$	29.1 $+\mathbf{8.0}$	3,409 $+\mathbf{2{,}363}$	6.9 $+\mathbf{6.9}$	2.8 $+\mathbf{2.8}$
Gemma-3-27B-IT	4,749	55.4	47.3	4,531	29.1	21.7
+ MeditronFO	5,509 $+\mathbf{760}$	56.2 $+\mathbf{0.8}$	45.9 $-\mathbf{1.4}$	5,240 $+\mathbf{708}$	29.5 $+0.4$	20.0 $-\mathbf{1.7}$
MedGemma-27B (vs Gemma-3)	4,274 $-\mathbf{475}$	54.8 -0.6	48.1 $+\mathbf{0.8}$	4,021 $-\mathbf{511}$	25.4 $-\mathbf{3.7}$	19.4 $-\mathbf{2.2}$

I AUTOMOOVE

I.1 JUDGE PROMPTS

The two answers are generated without a system prompt of our own (each model's chat template is applied as is; Apertus's template adds its default identity message), at temperature 0.7 with seed 42, and with at most 8,192 new tokens (3,072 for the OLMo-2-32B and EuroLLM-9B models, whose context is 4,096 tokens). The judge receives a single user message: the rubric instructions, then the MOOVE question and the two answers (template below). Every pair is judged in both answer orders. The overall verdict uses the winner rubric. Under log-prob scoring, the judge is not sampled: it reads the message followed by each of the three candidate completions, `{"Winner":"Model 1"}`, `{"Winner":"Model 2"}` and `{"Winner":"Equal"}`, and the completion with the highest mean token log-probability is the verdict. The mean is not a sequence probability, as it includes the shared JSON prefix and the choices differ in token count; generated verdicts give the same agreement with clinicians (κ 0.587 against 0.583). Under generated verdicts, and for the per-criterion rubric, the judge generates its JSON answer at temperature 0.7 with seed 42. Thinking is off for every judge. Prompts are reproduced verbatim from `auto_moove/auto_moove_prompt.py`.

Judge message template

```
{rubric instructions}

--- MEDICAL PROMPT ---
{question}

--- MODEL 1 ANSWER ---
{answer_1}

--- MODEL 2 ANSWER ---
{answer_2}
```

Rubric instructions — winner (overall verdict)

```
You are an expert medical clinician and an objective judge on the MOOVE (Massive Open
↪  Online Validation and Evaluation) platform.
Your task is to evaluate two alternative answers (Model 1 and Model 2) provided for a
↪  given medical prompt.

Declare a winner or tie (Model 1, Model 2, or Equal).
```

```
Output STRICTLY with a JSON in the following format:
{"Winner":"<Model 1 | Model 2 | Equal>"}
```

Rubric instructions — per-criterion preference

```
You are an expert medical clinician and an objective judge on the MOOVE (Massive Open
↪  Online Validation and Evaluation) platform. You are given a medical prompt and two
↪  alternative answers (Model 1 and Model 2).
The reader is a clinician, not a patient, unless the prompt says otherwise.

For each criterion below, compare the two answers directly. Judge every criterion on
↪  its own; do not declare an overall winner.

MOOVE CRITERIA -- for each, say which answer does better on that criterion alone:
"Model 1", "Model 2", or "Equal" if you cannot separate them.
- clinical_soundness: correct, evidence-based, consistent with current guidelines.
- safety: no harmful, dangerous or negligent recommendation.
- contextual_fit: fits the clinical setting, care level and the question actually
↪  asked.
- appropriate_confidence: certainty is calibrated; hedges where evidence is weak;
  asks for missing information instead of guessing.
- clarity_completeness: clear, complete, well organised, no material omission.

Output STRICTLY one JSON object, no prose, no markdown fence:
{"pref": {"clinical_soundness": "Model 1"|"Model 2"|"Equal", "safety": "Model 1"|"Model
↪  2"|"Equal", "contextual_fit": "Model 1"|"Model 2"|"Equal",
↪  "appropriate_confidence": "Model 1"|"Model 2"|"Equal", "clarity_completeness":
↪  "Model 1"|"Model 2"|"Equal"}}
```

I.2 VALIDATION AGAINST CLINICIANS

Table 13: AutoMOOVE judges against clinicians, unanimous set ($n = 576$), the numbers behind Figure 4. *Left:* κ and agreement on the overall verdict, a verdict that changes with the answer order counting as Equal (log-prob and generated); below the first rule, two further judges (thinking off); below the second, the majority of Gemma-4-31B, Qwen3.8-27B and Muse-Glimmer-30B (the verdict at least two share, Equal when all three differ), the vote of Figure 6. Figure 4 leaves out Qwen3-30B-A3B. *Right:* Gemma-4-31B κ per MOOVE criterion (n: pairs, reviews or rater pairs). All columns use the 576 unanimous pairs; *Per review*: each clinician review of those pairs.

	With ties	**No ties**
Human raters (n)	2 per pair: 402; 3+: 174	
Gemma-4-31B, log-prob scoring		
Judge κ	0.583 [0.525–0.638]	0.715 [0.654–0.777]
Agreement	76.4% (576)	86.1% (495)
Gemma-4-31B, generated verdict		
Judge κ	0.587 [0.529–0.643]	0.711 [0.646–0.775]
Agreement	76.6% (576)	85.9% (495)
Qwen3.8-27B, log-prob scoring		
Judge κ	0.553 [0.494–0.611]	0.690 [0.626–0.754]
Agreement	74.5% (576)	84.8% (488)
Qwen3.8-27B, generated verdict		
Judge κ	0.560 [0.502–0.621]	0.686 [0.620–0.747]
Agreement	74.8% (576)	84.7% (489)
Qwen3-30B-A3B, log-prob scoring		
Judge κ	0.507 [0.451–0.563]	0.659 [0.589–0.726]
Agreement	71.2% (576)	83.4% (470)
Qwen3-30B-A3B, generated verdict		
Judge κ	0.515 [0.457–0.568]	0.667 [0.594–0.734]
Agreement	71.5% (576)	83.8% (469)
Muse-Glimmer-30B, log-prob scoring		
Judge κ	0.522 [0.467–0.578]	0.703 [0.636–0.768]
Agreement	71.4% (576)	85.7% (453)
Muse-Glimmer-30B, generated verdict		
Judge κ	0.498 [0.442–0.554]	0.697 [0.623–0.761]
Agreement	69.3% (576)	85.4% (438)
Nemotron-3.5-Lightning-30B-A3B, log-prob		
Judge κ	0.451 [0.394–0.506]	0.680 [0.606–0.749]
Agreement	66.1% (576)	84.3% (428)
Nemotron-3.5-Lightning-30B-A3B, generated		
Judge κ	0.450 [0.394–0.504]	0.684 [0.611–0.754]
Agreement	66.0% (576)	84.5% (426)
Majority of three judges, log-prob scoring		
Judge κ	0.564 [0.507–0.621]	0.718 [0.656–0.781]
Agreement	74.8% (576)	86.2% (480)
Majority of three judges, generated verdict		
Judge κ	0.557 [0.500–0.615]	0.708 [0.642–0.769]
Agreement	74.3% (576)	85.7% (477)

Criterion	**Unanimous pairs**	**Per review**	**Rater–rater**
Clinical soundness	0.606 (188)	0.478 (785)	0.509 (655)
Safety	0.311 (148)	0.204 (785)	0.306 (655)
Contextual fit	0.597 (211)	0.486 (785)	0.615 (655)
Approp. confidence	0.604 (163)	0.414 (782)	0.398 (653)
Clarity & complet.	0.574 (186)	0.418 (784)	0.446 (654)
Overall verdict*	0.587 (576)	0.596 (1363)	–†

* Winner rubric, generated verdict. † 1 by construction: the set keeps only pairs on which all clinicians gave the same overall verdict.

Table 14: Three-way agreement (%) of each judge with the physicians' unanimous verdict under other tie rules and answer orders ($n = 576$; Equal is a class). First row: the default setting of Figure 4, where a verdict that changes with the answer order counts as Equal. *Both orders averaged*: the judge's probabilities averaged over both orders (log-prob only). Equal thresholds apply to log-prob scoring only. *Ties dropped*: pairs where physicians and judge both picked a winner (n below). *Random*: one order per pair, drawn at random (seed 42). Judge Equal calls in brackets where the rule changes them by more than 10. *Majority*: the verdict shared by at least two of Gemma-4-31B, Qwen3.8-27B and Muse-Glimmer-30B under the same rule (Equal when all three differ). κ with 95% intervals in Figure 9.

	Gemma-4-31B		Qwen3.8-27B		Qwen3-30B-A3B		GPT-5.6 Sol	Muse-Glimmer-30B		Nemotron-3.5-L.		Majority	
Setting	**log-prob**	**generate**	**log-prob**	**generate**	**log-prob**	**generate**	**generate**	**log-prob**	**generate**	**log-prob**	**generate**	**log-prob**	**generate**
Order flip counts as Equal (default)	76.4	76.6	74.5	74.8	71.2	71.5	65.5	71.4	69.3	66.1	66.0	74.8	74.3
Judge Equal calls	61	63	69	70	90	92	68	112	129	134	136	79	84
Tie handling													
Both orders averaged	79.0 (14)	–	76.6 (7)	–	75.0 (29)	–	–	75.5 (57)	–	75.9 (8)	–	77.8 (24)	–
Equal threshold 0.1	79.2 (15)	–	77.1 (33)	–	74.7 (31)	–	–	70.5	–	67.5 (114)	–	77.6 (51)	–
Equal threshold 0.3	79.0 (29)	–	53.1 (280)	–	72.6 (63)	–	–	43.8 (339)	–	21.0 (481)	–	56.6 (254)	–
Ties dropped	86.1 (495)	85.9 (495)	84.8 (488)	84.7 (489)	83.4 (470)	83.8 (469)	74.4 (488)	85.7 (453)	85.4 (438)	84.3 (428)	84.5 (426)	86.2 (480)	85.7 (477)
Answer order													
AB only	79.2 (16)	79.5 (21)	77.8 (10)	77.4 (10)	75.9 (28)	76.4 (27)	69.1 (12)	75.5 (56)	73.4 (67)	75.0 (12)	75.2 (10)	77.6 (28)	77.3 (30)
BA only	78.0 (12)	78.0 (21)	76.6 (4)	76.7 (4)	74.5 (27)	74.5 (29)	68.1 (13)	75.2 (63)	74.8 (69)	71.2 (7)	71.4 (8)	78.0 (23)	77.6 (29)
Random order per pair	78.8 (12)	78.8 (16)	76.6 (4)	76.2 (4)	75.9 (26)	76.4 (29)	68.6 (11)	76.4 (59)	75.3 (69)	71.5 (7)	71.7 (6)	77.4 (25)	76.9 (29)
κ, default	0.583	0.587	0.553	0.560	0.507	0.515	0.395	0.522	0.498	0.451	0.450	0.564	0.557
κ, random order	0.601	0.603	0.554	0.547	0.553	0.564	0.408	0.580	0.567	0.460	0.463	0.581	0.574
Macro-F1, default	0.633	0.644	0.621	0.633	0.601	0.608	0.548	0.615	0.601	0.567	0.566	0.629	0.632
Macro-F1, random order	0.611	0.632	0.591	0.589	0.640	0.648	0.568	0.674	0.667	0.507	0.524	0.632	0.643

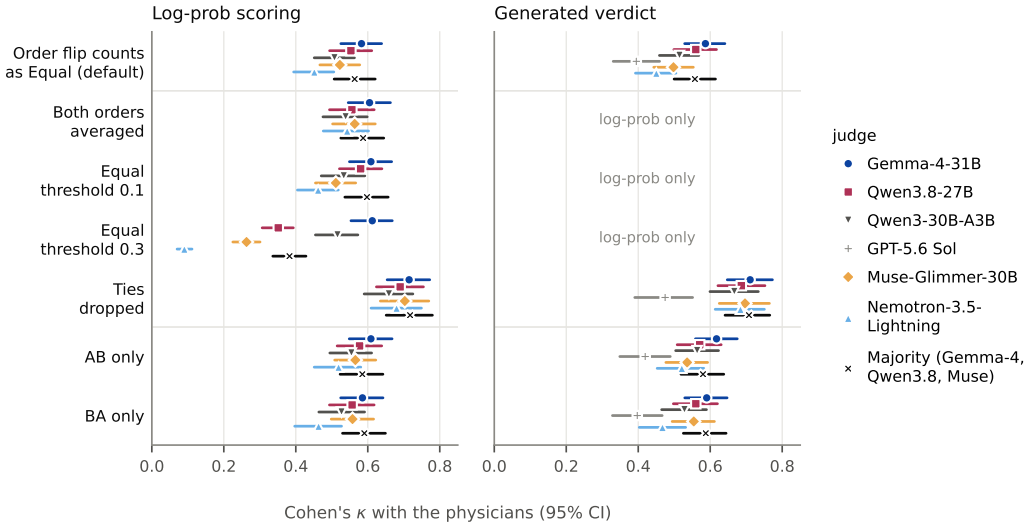

Figure 9: AutoMOOVE judge agreement with clinicians robustness under the tie rules and answer orders. ($n = 576$ pairs) *Order flip counts as Equal (default)*: each pair is judged with the answers in both orders, and a verdict that changes with the order counts as Equal. *Both orders averaged*: the judge's probabilities are averaged over both orders (log-prob only). *Equal threshold t*: Equal is chosen when its probability is within t of the more likely answer (log-prob only). *AB/BA only*: a single order. *Majority*: as in Figure 4, taken after each rule is applied to Gemma-4, Qwen3.8 and Muse-Glimmer. Macro-F1 in Figure 10, values in Table 14. *Left:* Most of the judges show high robustness under the different settings, with Nemotron-3.5 consistently showing the least robustness, and Gemma-4 consistently showing the highest. *Right:* Frontier LLM GPT-5.6 Sol performs worse on the agreement task than all open-source models tested.

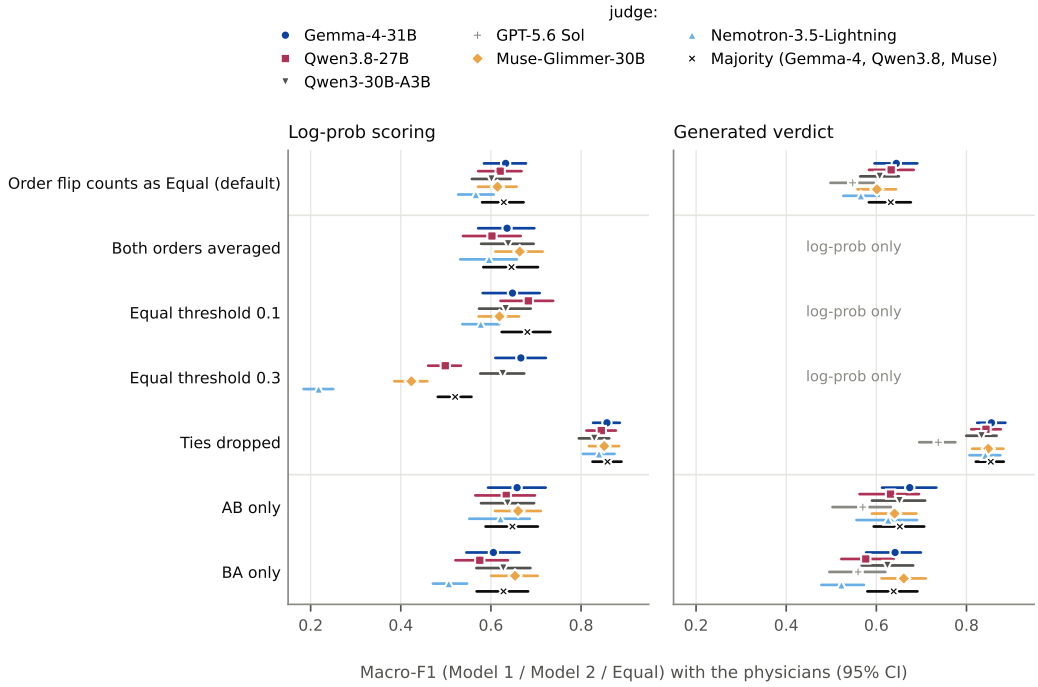

Figure 10: Figure 9 with macro-F1 in place of Cohen's κ: the mean of the per-class F1 over Model 1, Model 2 and Equal ($n = 576$ pairs, 95% bootstrap intervals). *Ties dropped* averages the two winner classes only, as no Equal verdicts remain.

I.3 Pairwise results

Table 15: AutoMOOVE results per judge, unanimous set, winner rubric, log-prob scoring. *Win*, *Tie*, *Loss* and *Win rate*: Gemma-4-31B judge. *Qwen3.8*, *Muse*: win rate with the same answers judged by Qwen3.8-27B and by Muse-Glimmer-30B (thinking off). *Majority*: the three-judge majority vote of Figure 6 (Equal when all three differ). *Win rate*: wins plus half the ties, 95% bootstrap interval over questions. Ties when the judge changes its mind when the order changes.

MeditronFO	Against	Win	Tie	Loss	Win rate	Qwen3.8	Muse	Majority	N
Apertus-70B	Apertus-70B-Instruct	442	35	39	89.1 [86.5–91.4]	84.5 [81.5–87.2]	90.1 [88.0–92.2]	89.7 [87.4–91.9]	516
Apertus-8B	Apertus-8B-Instruct	357	63	96	75.3 [71.9–78.7]	61.7 [57.7–65.6]	81.2 [78.3–84.0]	74.2 [70.8–77.6]	516
OLMo-2-32B	OLMo-2-32B-SFT	406	37	73	82.3 [79.0–85.3]	72.1 [68.5–75.6]	85.5 [82.8–88.0]	81.9 [78.7–84.8]	516
EuroLLM-22B	EuroLLM-22B-Instruct	258	93	165	59.0 [55.1–62.8]	49.8 [45.9–53.6]	66.7 [63.5–69.8]	59.2 [55.4–62.9]	516
EuroLLM-9B	EuroLLM-9B-Instruct	216	89	211	50.5 [46.6–54.5]	34.2 [30.6–38.1]	70.7 [67.3–73.9]	50.7 [46.7–54.7]	516
Gemma-3-27B	Gemma-3-27B-IT	324	106	86	73.1 [69.8–76.3]	70.3 [67.1–73.6]	58.3 [55.8–60.9]	71.0 [67.9–74.0]	516
Gemma-3-27B	MedGemma-27B	228	151	137	58.8 [55.2–62.4]	52.8 [49.0–56.6]	54.2 [51.5–56.7]	57.1 [53.8–60.3]	516
Apertus-70B	MedGemma-27B	161	105	250	41.4 [37.5–45.1]	28.9 [25.5–32.3]	41.6 [39.1–44.2]	35.8 [32.5–39.1]	516

Table 16: AutoMOOVE against MedGemma-27B: each MeditronFO model next to its base, in the setting of Table 15 (unanimous set, winner rubric, log-prob scoring, an order flip counts as a tie). *Win*, *Tie*, *Loss* and *Win rate*: Gemma-4-31B judge; *Qwen3.8*, *Muse*: win rate under Qwen3.8-27B and Muse-Glimmer-30B; *Majority*: three-judge majority vote. Win rate: wins plus half the ties, 95% bootstrap interval over questions. The same MedGemma-27B answers are used in every row.

Model vs MedGemma-27B	Win	Tie	Loss	Win rate	Qwen3.8	Muse	Majority	N
Gemma-3-27B-MeditronFO	228	151	137	58.8 [55.2–62.4]	52.8 [49.0–56.6]	54.2 [51.5–56.7]	57.1 [53.8–60.3]	516
Gemma-3-27B-IT	103	173	240	36.7 [33.5–40.1]	29.3 [26.1–32.5]	44.9 [42.7–47.2]	35.1 [32.3–38.0]	516
Apertus-70B-MeditronFO	161	105	250	41.4 [37.5–45.1]	28.9 [25.5–32.3]	41.6 [39.1–44.2]	35.8 [32.5–39.1]	516
Apertus-70B-Instruct	11	13	492	3.4 [2.1–4.8]	3.0 [1.7–4.6]	3.7 [2.6–4.9]	2.6 [1.6–3.9]	516

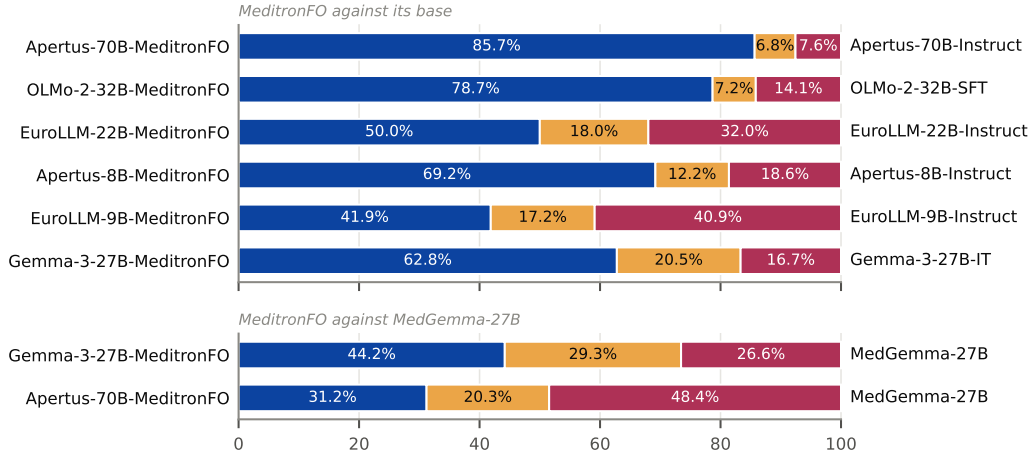

Figure 11: Figure 6 for Gemma-4-31B alone (winner rubric, log-prob scoring, both answer orders; a tie when its verdict changes with the answer order). Win rates, ties counting half: 89.1, 82.3, 59.0, 75.3, 50.5 and 73.1% against the bases (top to bottom), 58.8 and 41.4% against MedGemma-27B.

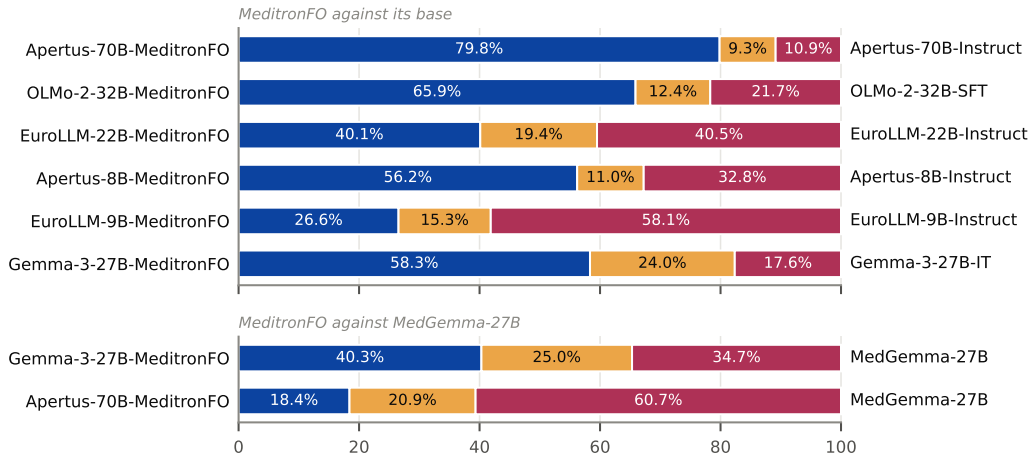

Figure 12: Figure 6 for Qwen3.8-27B alone (winner rubric, log-prob scoring, both answer orders; a tie when its verdict changes with the answer order). Win rates, ties counting half: 84.5, 72.1, 49.8, 61.7, 34.2 and 70.3% against the bases (top to bottom), 52.8 and 28.9% against MedGemma-27B.

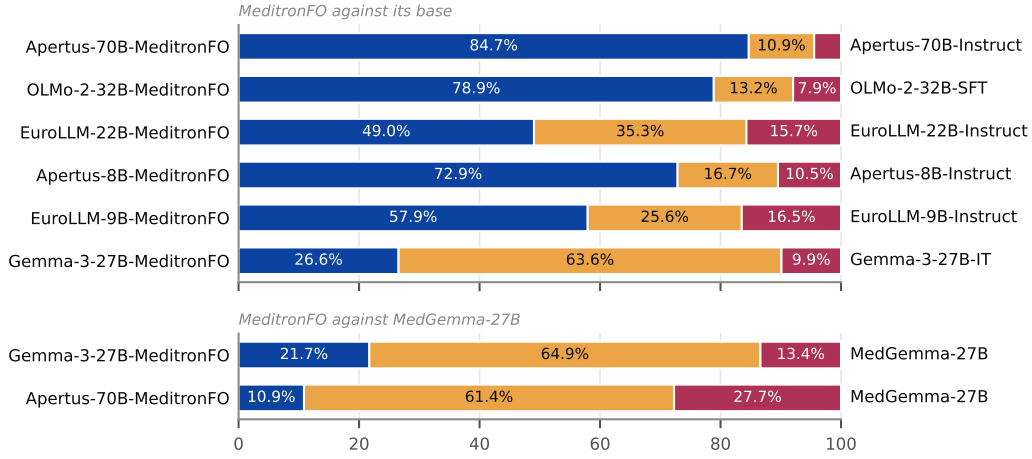

Figure 13: Figure 6 for Muse-Glimmer-30B alone (winner rubric, log-prob scoring, both answer orders; a tie when its verdict changes with the answer order). Win rates, ties counting half: 90.1, 85.5, 66.7, 81.2, 70.7 and 58.3% against the bases (top to bottom), 54.2 and 41.6% against MedGemma-27B. Muse-Glimmer-30B changes its verdict with the answer order on 11–65% of questions, so its bars hold many ties.

Table 17: Per-criterion preference (the rubric validated in Figure 4; numbers behind Figures 6 and 14), unanimous set, Gemma-4-31B judge, generated verdict, both orders. MeditronFO win rate (%) against its base on each MOOVE criterion: wins plus half the ties, an order disagreement counting as a tie; *Mean*: the per-question mean of the five; 95% bootstrap intervals over the questions. N: questions with all five criteria parsed in both orders. Below the rule, the Apertus-8B ablations of Table 21, also against Apertus-8B-Instruct.

MeditronFO	Clinical	Safety	Context	Confidence	Clarity	Mean	N
Apertus-70B	79.7 [76.4–82.7]	74.7 [71.5–77.6]	81.6 [78.4–84.4]	79.3 [76.0–82.2]	87.9 [85.2–90.4]	80.6 [77.6–83.3]	516
OLMo-2-32B	71.1 [67.4–74.5]	64.6 [61.2–67.9]	72.4 [68.8–75.7]	68.8 [65.2–72.3]	83.2 [80.0–86.1]	72.0 [68.8–75.1]	516
EuroLLM-22B	50.0 [45.9–54.0]	48.7 [45.1–52.3]	50.5 [46.3–54.5]	49.6 [45.5–53.6]	59.8 [55.7–63.8]	51.7 [47.9–55.4]	516
Apertus-8B	64.4 [60.4–68.3]	61.6 [57.9–65.3]	64.8 [60.9–68.6]	63.9 [59.8–67.7]	74.1 [70.6–77.7]	65.8 [62.1–69.3]	516
EuroLLM-9B	38.4 [34.5–42.3]	36.8 [33.1–40.5]	38.5 [34.6–42.4]	38.2 [34.3–42.2]	49.9 [45.7–54.0]	40.3 [36.7–44.1]	516
Gemma-3-27B	64.8 [61.3–68.4]	56.9 [54.0–60.0]	65.5 [62.0–69.0]	59.8 [56.5–63.2]	77.9 [75.1–80.9]	65.0 [62.1–67.9]	516
Apertus-8B − Curated QA	61.8 [58.2–65.7]	59.0 [55.5–62.6]	63.0 [59.4–66.8]	61.7 [58.1–65.5]	73.0 [69.4–76.6]	63.7 [60.3–67.3]	516
Apertus-8B − Synthetic Curated QA	62.5 [58.4–66.4]	59.9 [56.1–63.6]	63.3 [59.3–67.2]	61.6 [57.6–65.5]	74.5 [70.8–78.0]	64.4 [60.6–68.0]	516
Apertus-8B − Guidelines QA	60.4 [56.5–64.1]	58.8 [54.9–62.4]	61.2 [57.3–65.1]	60.1 [56.1–64.0]	72.9 [69.2–76.4]	62.7 [58.9–66.3]	516
Apertus-8B − Synthetic vignettes	57.2 [53.4–61.0]	54.3 [50.7–57.9]	57.4 [53.6–61.3]	56.3 [52.4–60.1]	67.9 [64.2–71.7]	58.6 [55.1–62.2]	516

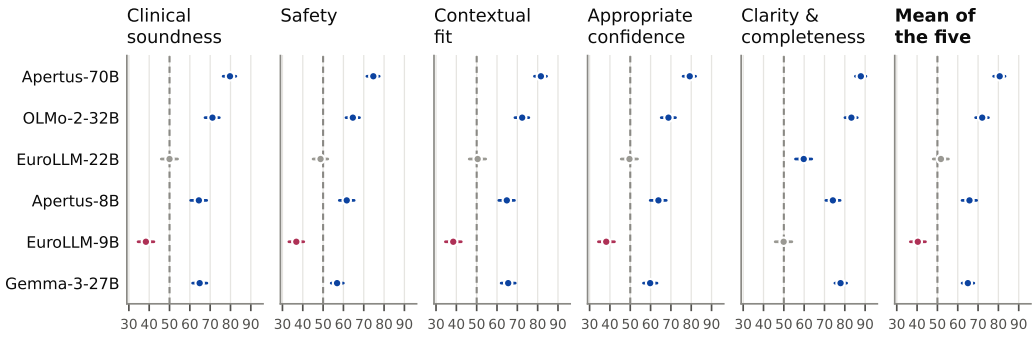

Figure 14: Per-criterion AutoMOOVE win rate for MeditronFO models versus corresponding base, with intervals. Win rate of each MeditronFO model against its base on each MOOVE criterion and on their mean (ties count half; an answer-order disagreement counts as a tie), Gemma-4-31B judge, per-criterion preference rubric (the rubric validated in Figure 4), 95% bootstrap intervals over the 516 questions. The dashed line is parity; blue where the interval is above 50%, grey where it contains 50%. The win, tie and loss shares behind these rates are in Figure 6.

I.4 RESPONSE LENGTH

LLM-as-a-judge evaluations are known to be sensitive to response length. To check that the AutoMOOVE comparison between Gemma-3-27B-MeditronFO and MedGemma-27B is not driven by length, we measure both models' answers on all 516 unanimous-set questions (Table 18). The two models write similar lengths: MeditronFO answers average 5,533 characters against 5,478 for MedGemma-27B, 11% longer in tokens (paired ratio 1.11). Both judges slightly prefer the longer answer: the correlation between the preference margin and the log word ratio is $r = 0.13$ under Gemma-4-31B and $r = 0.14$ under Qwen3.8-27B (Table 19). On length-matched pairs (answers within 20% of each other in length), the MeditronFO win rate is 56.9% under Gemma-4-31B and 48.1% under Qwen3.8-27B, close to the rates on all pairs (58.8% and 52.8%).

Table 18: Response length, Gemma-3-27B-MeditronFO against MedGemma-27B, 516 unanimous-set questions. Tokens: Gemma-3 tokenizer. Paired ratio: geometric mean of the per-question token ratio, bootstrap interval.

	Characters	**Tokens**	**At 8192-token limit**
Gemma-3-27B-MeditronFO, mean	5533	1334	0
MedGemma-27B, mean	5478	1240	1
Paired ratio, MeditronFO / MedGemma (95% CI)		1.11 [1.07–1.16]	

We run the same check for each MeditronFO finetune against its base (Table 19), measuring length in words. Every finetune writes longer answers than its base: the median word ratio is 1.20 for Apertus-70B and 1.27 for Gemma-3-27B, and 2.37 to 3.93 for the other four. For these four, few questions have similar-length answers (6 to 50), so the length-matched check has little power. Apertus-70B and Gemma-3-27B-MeditronFO win about as often on similar-length answers as on all questions (85.8% against 89.1% and 69.8% against 73.1% under Gemma-4-31B). Apertus-8B, OLMo-2-32B and EuroLLM-22B-MeditronFO win less often on similar-length answers (59.5%, 74.0% and 45.5% under Gemma-4-31B, against 75.3%, 82.3% and 59.0% on all questions), and the preference correlates with length ($r = 0.14$ to 0.16), so their results against the base partly reflect length. EuroLLM-9B-MeditronFO writes the longer answer on 99% of questions, and its correlation is negative ($r = -0.14$): 58 of its answers stop at the 3,072-token limit, against 25 for OLMo-2-32B-MeditronFO and none for the bases.

Table 19: AutoMOOVE win rate (%) of each MeditronFO finetune against its base by answer length, 516 unanimous-set questions, ties as in Table 15. *Ratio*: median per-question word ratio, finetune over base. *Longer*: questions where the finetune's answer is longer. *Similar*: word ratio between 0.8 and 1.25, the length-matched pairs above; *>25% longer*: the finetune's answer is more than 25% longer. Questions in brackets; 95% bootstrap interval over questions. r: correlation between the judge's preference margin and the log word ratio, with its 95% bootstrap interval. Last row: Gemma-3-27B-MeditronFO against MedGemma-27B.

			Gemma-4-31B				**Qwen3.8-27B**			
MeditronFO vs base	**Ratio**	**Longer**	**All**	**Similar**	**>25% longer**	r	**All**	**Similar**	**>25% longer**	r
Apertus-70B	1.20	59%	89.1	85.8 [78.4–92.0] (88)	91.7 (248)	−0.01 [−0.10, 0.07]	84.5	82.4	85.1	−0.12 [−0.21, −0.02]
Apertus-8B	2.44	90%	75.3	59.5 [46.4–73.8] (42)	78.6 (441)	0.14 [0.04, 0.25]	61.7	47.6	63.7	0.08 [−0.01, 0.16]
OLMo-2-32B	2.55	94%	82.3	74.0 [62.0–85.0] (50)	83.2 (450)	0.15 [0.04, 0.26]	72.1	63.0	73.3	0.19 [0.07, 0.31]
EuroLLM-22B	2.37	89%	59.0	45.5 [31.8–59.1] (44)	61.8 (431)	0.16 [0.07, 0.24]	49.8	52.3	50.6	0.12 [0.01, 0.22]
EuroLLM-9B	3.93	99%	50.5	58.3 [24.8–91.7] (6)	50.2 (505)	−0.14 [−0.23, −0.06]	34.2	16.7	34.2	−0.11 [−0.21, −0.02]
Gemma-3-27B	1.27	84%	73.1	69.8 [64.7–74.9] (217)	76.6 (274)	0.06 [−0.03, 0.15]	70.3	65.4	74.5	0.07 [−0.01, 0.16]
Gemma-3-27B vs MedGemma-27B	1.12	66%	58.8	56.9 [51.5–62.0] (233)	65.3 (199)	0.13 [0.04, 0.23]	52.8	48.1	62.6	0.14 [0.04, 0.24]

I.5 JUDGE ROBUSTNESS

Table 20: AutoMOOVE win rate (%) of MeditronFO against MedGemma-27B under other judges, tie rules and answer orders, 516 unanimous-set questions. First row: each judge alone in the setting of Table 15; Figure 6 instead shows the three-judge majority. *Both orders averaged*: the judge's probabilities averaged over both orders. *Equal threshold*: Equal wins when its probability is within this margin of the more likely answer. *AB*: MeditronFO shown first; *BA*: MedGemma shown first; *random*: one order per question, drawn at random (seed 42). Ties in brackets after the rate where the rule creates more than 10.

	Apertus-70B-MeditronFO				Gemma-3-27B-MeditronFO			
Setting	**Gemma-4-31B**	**Qwen3.8-27B**	**Muse-Glimmer**	**Nemotron-3.5-L.**	**Gemma-4-31B**	**Qwen3.8-27B**	**Muse-Glimmer**	**Nemotron-3.5-L.**
Order flip counts as Equal (default)	41.4 [37.5–45.1] (105)	28.9 [25.5–32.3] (108)	41.6 [39.1–44.2] (317)	71.7 [69.1–74.5] (204)	58.8 [55.2–62.4] (151)	52.8 [49.0–56.6] (129)	54.2 [51.5–56.7] (335)	75.7 [73.2–78.2] (199)
Tie handling								
Both orders averaged	41.9	29.0	38.3 (73)	79.8	60.0 (21)	53.1	55.9 (155)	83.4 (11)
Equal threshold 0.1	41.8	28.6 (15)	37.4 (222)	70.9 (178)	60.0 (25)	51.9 (34)	51.9 (310)	66.6 (267)
Equal threshold 0.3	41.4 (19)	39.0 (382)	45.3 (450)	50.2 (514)	58.5 (60)	48.2 (453)	49.8 (474)	50.2 (514)
Ties dropped	41.7	28.8	36.3	80.2	60.4	53.1	58.4	84.2
Answer order								
AB only	46.2 [41.7–50.4]	23.3 [19.9–26.8]	64.6 [60.9–68.1]	57.4 [53.4–61.7]	68.7 [64.8–72.5]	48.6 [44.4–53.1]	74.4 [71.2–77.4]	63.0 [58.9–66.9]
BA only	36.8 [32.7–40.8]	34.6 [30.3–38.6]	18.2 [15.3–21.4]	86.6 [83.6–89.4]	50.9 [46.6–55.1]	57.1 [52.8–61.3]	38.5 [35.0–42.0]	89.2 [86.6–91.8]
Random order per question	40.7 [36.3–44.9]	29.9 [26.0–34.0]	43.0 [39.0–46.9]	73.4 [69.7–76.9]	59.6 [55.4–63.8]	52.6 [48.4–56.8]	58.2 [54.7–61.7]	76.1 [72.2–79.5]
Both orders agree (%)	80.8	79.8	42.6	61.2	72.9	76.4	50.8	62.8

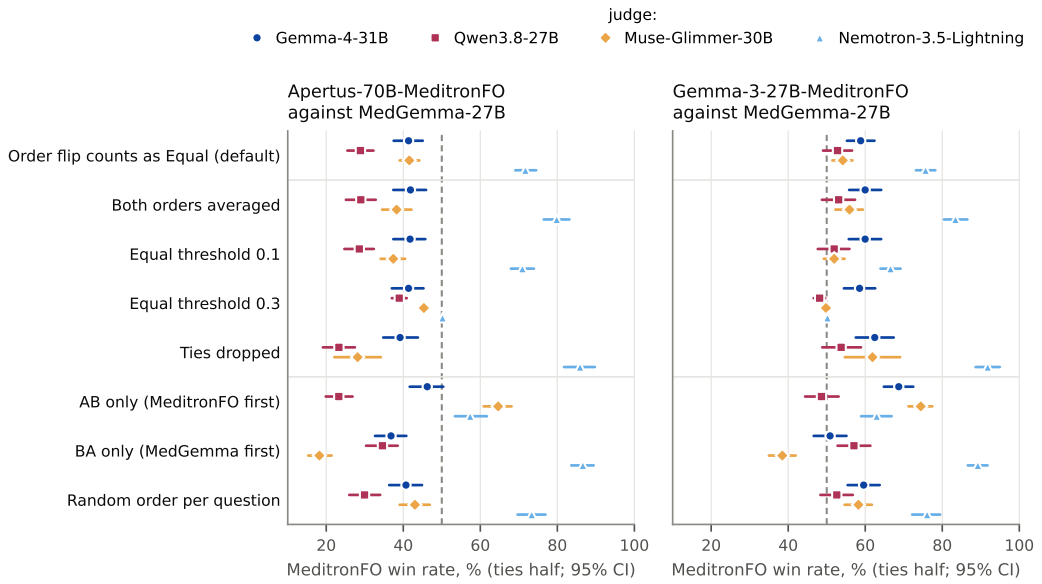

Figure 15: AutoMOOVE judge robustness. Win rate of Apertus-70B-MeditronFO (left) and Gemma-3-27B-MeditronFO (right) against MedGemma-27B under four judges, tie rules and answer orders, 516 unanimous-set questions, 95% bootstrap intervals. First row: each judge alone in the setting of Table 15 (an order flip counts as Equal, log-prob scoring); Figure 6 instead shows the three-judge majority. *Equal threshold*: Equal wins when its probability is within this margin of the more likely answer. *AB*: MeditronFO shown first; *BA*: MedGemma shown first. Dashed line: parity. Full table in Table 20.

J CORPUS-COMPONENT ABLATIONS

Table 21: Corpus-component ablations on Apertus-8B and Apertus-70B. All runs of one model size use identical training settings; only the indicated corpus component is removed. *Steps*: optimizer steps, which fall with the corpus. *HealthBench*: full set, Gemma-4-31B grader. *AutoMOOVE*: win rate against the Instruct model of the same size (unanimous set, Gemma-4-31B judge); *Pref.*: mean per-criterion preference win rate against the same model (the rubric of Figure 6), same questions and judge. Best MCQA value per column and model size in bold. The 70B full-corpus row is the first training run of Table 22; its second run is shown for the training noise. The last block gives each ablation minus the full-corpus model with a paired 95% bootstrap interval (both models resampled on the same questions or conversations); these cover evaluation noise only.

Configuration	Steps	MedMCQA	MedQA	PubMedQA	MedXpertQA	MCQA avg	HealthBench	AutoMOOVE	Pref.
Apertus-8B-Instruct	–	47.8	52.1	**74.3**	11.8	46.49	26.0	–	–
Corpus component ablations (one component removed at a time)									
Full gpt-oss corpus	723	52.0	60.7	71.4	**15.1**	49.81	42.1	75.3	65.8
− Curated QA	560	50.3	56.1	69.0	**15.1**	47.62	40.7	74.5	63.7
− Synthetic Curated QA	402	52.6	59.2	69.8	13.6	48.77	42.1	75.8	64.4
− Guidelines QA	604	**53.1**	**61.8**	72.8	13.7	**50.35**	42.3	73.8	62.7
− Synthetic vignettes	603	52.5	61.0	69.8	14.4	49.45	42.5	71.3	58.6
Apertus-70B-Instruct	–	52.8	61.9	74.7	13.0	50.60	39.4	–	–
Full gpt-oss corpus	725	**60.2**	**71.1**	74.9	17.3	**55.88**	49.3	89.1	80.6
second training run	723	59.4	69.6	75.5	16.4	55.22	49.5	–	–
− Curated QA	560	56.4	65.4	72.2	16.0	52.49	46.4	88.1	80.9
− Synthetic Curated QA	402	59.0	70.2	73.6	16.5	54.85	48.8	89.5	80.8
− Guidelines QA	604	58.8	70.1	**75.6**	16.0	55.15	49.4	87.7	79.1
− Synthetic vignettes	603	60.1	70.0	74.7	**18.0**	55.68	48.3	88.1	79.5
Change against the full corpus, paired 95% interval									
Apertus-8B									
− Curated QA						-2.19 [-3.26, -1.08]	-1.3 [-2.1, -0.6]	-0.8 [-4.5, $+2.7$]	-2.1 [-5.9, $+1.8$]
− Synthetic Curated QA						-1.04 [-2.09, $+0.02$]	$+0.1$ [-0.7, $+0.8$]	$+0.5$ [-3.1, $+3.9$]	-1.4 [-5.2, $+2.2$]
− Guidelines QA						$+0.54$ [-0.53, $+1.66$]	$+0.3$ [-0.5, $+1.0$]	-1.5 [-4.8, $+1.8$]	-3.1 [-6.8, $+0.6$]
− Synthetic vignettes						-0.36 [-1.34, $+0.76$]	$+0.5$ [-0.3, $+1.2$]	-4.0 [-7.9, -0.1]	-7.2 [-11.3, -3.1]
Apertus-70B									
Second training run						-0.65 [-1.58, $+0.26$]	$+0.2$ [-0.5, $+0.9$]		
− Curated QA						-3.39 [-4.40, -2.41]	-3.0 [-3.7, -2.2]	-1.0 [-3.7, $+1.7$]	$+0.3$ [-2.8, $+3.3$]
− Synthetic Curated QA						-1.03 [-2.03, -0.07]	-0.6 [-1.3, $+0.1$]	$+0.5$ [-2.3, $+3.1$]	$+0.1$ [-2.8, $+3.0$]
− Guidelines QA						-0.73 [-1.75, $+0.26$]	$+0.1$ [-0.7, $+0.8$]	-1.4 [-4.0, $+1.3$]	-1.6 [-4.5, $+1.2$]
− Synthetic vignettes						-0.20 [-1.16, $+0.83$]	-1.0 [-1.7, -0.3]	-1.0 [-4.1, $+2.0$]	-1.1 [-4.2, $+2.0$]

K SECOND-TEACHER ABLATION

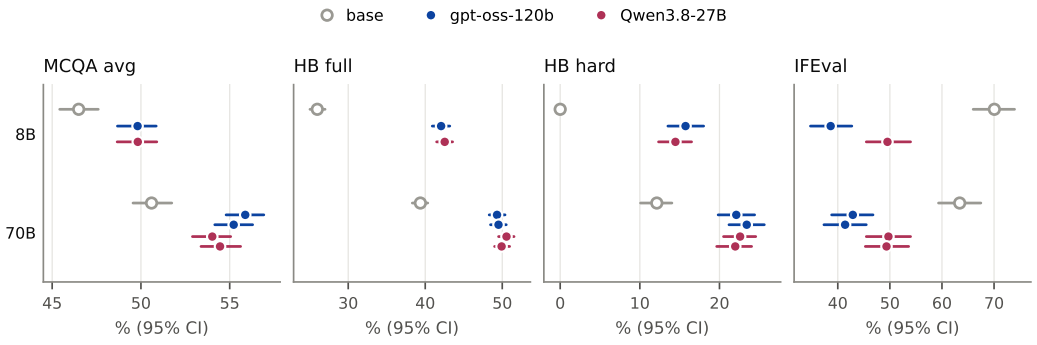

Figure 16: Second-teacher ablation, HealthBench full and hard. Apertus-8B and -70B trained on the same questions answered by gpt-oss-120b or Qwen3.8-27B; at 70B each teacher has two training runs, drawn one above the other. HB: HealthBench full and hard, Gemma-4-31B grader. 95% bootstrap intervals over the benchmark items; they cover evaluation noise only, and the gap between the two runs of one recipe shows the training noise. Full table in Appendix Table 22.

Table 22: Second-teacher ablation. *Rows*: training rows kept after pass@k filtering. *HB*: HealthBench full and hard, Gemma-4-31B grader. Replicate rows are a second training run of the same recipe on 8 nodes. A single score's 95% interval is about ±1 point on HB full and ±2 on hard. The last block gives the teacher gap and the run-to-run difference with a paired 95% bootstrap interval (both models resampled on the same items); a single-task difference is in bold where its interval excludes zero.

Model	Rows	Steps	MedMCQA	MedQA	PubMedQA	MedXpertQA	MCQA avg	HB full	HB hard	IFEval
Apertus-8B-Instruct	–	–	47.8	52.1	74.3	11.8	46.49	26.0	0.0	70.1
8B, gpt-oss	533 888	723	52.0	60.7	71.4	15.1	49.81	42.1	15.7	38.6
8B, Qwen3.8	543 135	1217	54.0	61.0	69.9	14.3	49.82	42.5	14.5	49.5
Apertus-70B-Instruct	–	–	52.8	61.9	74.7	13.0	50.60	39.4	12.1	63.4
70B, gpt-oss	533 888	725	60.2	71.1	74.9	17.3	55.88	49.3	22.1	42.9
70B, Qwen3.8	543 135	1227	58.0	71.2	68.5	18.3	54.02	50.6	22.6	49.7
gpt-oss, replicate	533 888	723	59.4	69.6	75.5	16.4	55.22	49.5	23.4	41.4
Qwen3.8, replicate	543 135	1217	58.2	71.2	70.0	18.4	54.46	49.9	22.0	49.4
Differences, paired 95% interval										
Qwen3.8 − gpt-oss, 8B			+2.0	+0.3	−1.5	−0.8	+0.01 [−1.18, +1.16]	+0.5 [−0.4, +1.4]	−1.3 [−3.3, +1.0]	+10.9 [+6.3, +15.3]
Qwen3.8 − gpt-oss, 70B			−2.2	+0.2	−6.4	+1.1	−1.86 [−3.06, −0.71]	+1.2 [+0.4, +2.1]	+0.5 [−1.6, +2.7]	+6.8 [+2.4, +11.3]
gpt-oss, replicate − first run			−0.8	−1.5	+0.6	−0.9	−0.65 [−1.62, +0.24]	+0.2 [−0.5, +0.9]	+1.3 [−0.5, +3.1]	−1.5 [−4.6, +1.7]
Qwen3.8, replicate − first run			+0.2	+0.0	+1.5	+0.1	+0.44 [−0.59, +1.53]	−0.6 [−1.3, +0.0]	−0.6 [−2.3, +1.0]	−0.4 [−3.7, +3.0]

L OPENNESS COMPARISON ACROSS MEDICAL LLMS

We compare all medical LLMs discussed in this work on the openness dimensions of Table 23, which follow the Model Openness Framework (MOF) (White et al., 2024): whether the base model is fully-open, whether the medical training data, the data pipeline and the training code are released, and the license of the released weights. As argued in Section 1, releasing weights alone does not constitute full openness: the pretraining data and training recipe of the base model determine what clinical knowledge and biases the model has absorbed, yet these are undisclosed for every major open-weight base used in prior medical LLM work (LLaMA, Llama 2, Llama 3.1, Mistral, Qwen2.5, Phi-3.5, Gemma 3). The MeditronFO finetunes of Apertus, OLMo-2 and EuroLLM are the only models that are open on every dimension.

Table 23: Openness across all medical LLMs discussed in this work. **O**=permissive open license, **C**=community or commercially usable license with restrictions, **IC**=inherited C license, and **R**=research-only, or proprietary license. Models released in several sizes or on several bases list each base; a license entry "IC/O" covers variants on different bases.

Model	Base	Fully-open Base	Open Medical Data	Open Data Pipeline	Open Training Code	License
Closed specialists						
MedPaLM (Singhal et al., 2023)	PaLM	✗	✗	✗	✗	R
MedPaLM-2 (Singhal et al., 2025)	PaLM 2	✗	✗	✗	✗	R
Med-Gemini (Saab et al., 2024)	Gemini	✗	✗	✗	✗	R
Open-weight medical specialists						
PMC-LLaMA (Wu et al., 2024)	LLaMA	✗	✗	✗	✗	R
Meditron-70B (Chen et al., 2023b)	Llama 2	✗	✗	✗	✗	IC
Meditron-3 (Sallinen et al., 2025)	Llama 3.1	✗	✗	✗	✗	IC
MedGemma-27B (Sellergren et al., 2025)	Gemma-3-27B	✗	✗	✗	✗	C
BioMistral (Labrak et al., 2024)	Mistral	✗	✓	✗	✗	O
HuatuoGPT-o1 (Chen et al., 2024)	Llama 3.1 / Qwen2.5	✗	✓	✗	✓	IC/O
Aloe Beta (Garcia-Gasulla et al., 2026)	Llama 3.1 / Qwen2.5	✗	✓	✗	✗	R
Hippocrates (Acikgoz et al., 2024)	Llama 2 / Mistral	✗	✓	✗	✓	IC/O
MediPhi (Corbeil et al., 2025)	Phi-3.5	✗	✗	✗	✗	O
MeditronFO (this work)						
Gemma-3-27B-MeditronFO	Gemma-3-27B	✗	✓	✓	✓	IC
Apertus-70B-MeditronFO	Apertus-70B	✓	✓	✓	✓	O
Apertus-8B-MeditronFO	Apertus-8B	✓	✓	✓	✓	O
OLMo-2-32B-MeditronFO	OLMo-2-32B	✓	✓	✓	✓	O
EuroLLM-22B-MeditronFO	EuroLLM-22B	✓	✓	✓	✓	O
EuroLLM-9B-MeditronFO	EuroLLM-9B	✓	✓	✓	✓	O

M LICENSES OF EXISTING ASSETS

Table 24 lists each existing asset used in this work, with its originating reference, version where applicable, public URL, and license. All assets are used in accordance with their respective licenses and restricted to research purposes consistent with the originating works.

Table 24: Licenses of existing assets used in this work. Datasets and models are listed with their originating reference, version, public URL, and license. † License terms verified at time of writing; downstream users should re-verify upstream terms before redistribution.

Asset	Reference	URL	License
Source QA datasets (training)			
MedQA	Jin et al. [2021]	https://github.com/jind11/MedQA	MIT
MedMCQA	Pal et al. [2022]	https://medmcqa.github.io	MIT
PubMedQA	Jin et al. [2019]	https://pubmedqa.github.io	MIT
HealthSearchQA	Singhal et al. [2023]	via Med-PaLM release	CC BY 4.0
LiveQA-Med	Abacha et al. [2017]	https://github.com/abachaa/LiveQA_MedicalTask_TREC2017	CC BY 4.0
AfriMed-QA v1/v2	Olatunji et al. [2024]	https://huggingface.co/datasets/intronhealth/afrimedqa_v2	CC BY 4.0†
Evaluation benchmarks			
MedXpertQA	Zuo et al. [2025]	https://huggingface.co/datasets/TsinghuaC3I/MedXpertQA	MIT†
MMLU-Pro	Wang et al. [2024]	https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro	MIT
IFEval	Zhou et al. [2023]	https://github.com/google-research/google-research/tree/master/instruction_following_eval	Apache 2.0
ARC-Challenge	Clark et al. [2018]	https://allenai.org/data/arc	CC BY-SA 4.0
HealthBench	Arora et al. [2025]	https://github.com/openai/simple-evals	MIT†
Base models (finetuned)			
Apertus-70B-Instruct	Hernández-Cano et al. [2025]	https://huggingface.co/swiss-ai/Apertus-70B-Instruct-2509	Apache 2.0
Apertus-8B-Instruct	Hernández-Cano et al. [2025]	https://huggingface.co/swiss-ai/Apertus-8B-Instruct-2509	Apache 2.0
OLMo-2-32B-SFT	OLMo et al. [2024]	https://huggingface.co/allenai/OLMo-2-0325-32B-SFT	Apache 2.0
EuroLLM-22B-Instruct	Ramos et al. [2026]	https://huggingface.co/utter-project/EuroLLM-22B-Instruct	Apache 2.0†
EuroLLM-9B-Instruct	Ramos et al. [2026]	https://huggingface.co/utter-project/EuroLLM-9B-Instruct	Apache 2.0
Gemma-3-27B-IT	Team et al. [2025]	https://huggingface.co/google/gemma-3-27b-it	Gemma Terms of Use
Reference / baseline models (evaluation only)			
MedGemma-27B	Sellergren et al. [2025]	https://huggingface.co/google/medgemma-27b-text-it	Health AI Developer Foundations TOS
Llama-3.1-70B-Meditron	Sallinen et al. [2025]	https://huggingface.co/OpenMeditron/Meditron3-70B	Llama 3.1 Community License
MediPhi	Corbeil et al. [2025]	https://huggingface.co/microsoft/MediPhi	MIT†
Qwen3-30B-A3B-Instruct-2507	Yang et al. [2025]	https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507	Apache 2.0
gpt-oss-120b	Agarwal et al. [2025]	https://huggingface.co/openai/gpt-oss-120b	Apache 2.0
Judges and graders (evaluation and corpus audit only)			
Gemma-4-31B-it	—	https://huggingface.co/google/gemma-4-31B-it	Apache 2.0
Qwen3.8-27B	—	https://huggingface.co/Qwen/Qwen3.8-27B	Apache 2.0
Muse-Glimmer-30B	—	https://huggingface.co/meta-models/Muse-Glimmer-30B	Apache 2.0
Nemotron-3.5-Lightning-30B-A3B	—	https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16	OpenMDW 1.1
Gemini 3.6 Flash (corpus audit)	—	—	Proprietary API terms
GPT-5.6 Sol (judge comparison)	—	—	Proprietary API terms
Frameworks and infrastructure			
Axolotl	—	https://github.com/axolotl-ai-cloud/axolotl	Apache 2.0
PyTorch (FSDP v2)	—	https://pytorch.org	BSD 3-Clause
DeepSpeed (ZeRO-3)	—	https://github.com/deepspeedai/DeepSpeed	Apache 2.0
FlashAttention 2	—	https://github.com/Dao-AILab/flash-attention	BSD 3-Clause