Bridging the Medical Knowledge Gap: 4-bit QLoRA Fine-Tuning of Llama-3-8B for Biomedical Reasoning

Aditya Verma

Abstract

Large language models (LLMs) show promise for biomedical reasoning, but the compute needed to adapt them keeps this work out of reach for many researchers. This study examines whether Meta Llama-3-8B can be adapted for biomedical reasoning on a single NVIDIA T4 GPU (16 GB VRAM) using 4-bit NormalFloat (NF4) quantization and quantized low-rank adaptation (QLoRA). The adapted model, named Bio-Llama-3-Expert, was fine-tuned on the Medical-O1-Reasoning-SFT dataset with approximately 42 million trainable parameters (0.52% of the model) and evaluated on two small, study-level sets. On a 50-question PubMedQA subset (YES/NO/MAYBE), it reached 72.0% accuracy (36/50; 95% CI, 58.3–82.5%), compared with a 42.0% majority-class rate in that subset and a reported 45% for untuned Llama-3-8B with naive prompting. On a separate 69-question reasoning-complexity set, overall accuracy was 79.7% (55/69) and declined from 92.0% on direct-recall questions to 65.0% on multi-hop questions. The MAYBE class was the weakest (3 of 9 correct). The study is limited by small evaluation sets, short training runs, baseline comparisons that are not paired, and the absence of clinical validation. The results are therefore preliminary evidence of computational feasibility, not of clinical readiness.

DOI: 10.5195/pur.2026.166

Introduction

Large language models (LLMs) have advanced natural language processing across many domains, including health care and biomedicine. Systems such as Med-PaLM and Med-PaLM 2 have achieved strong results on medical licensing-style questions (Singhal et al. 2023a, 2023b). Adapting and deploying models of this kind for specialized biomedical use, however, raises three practical problems:

- Computation: State-of-the-art systems are typically trained and served on multi-GPU clusters (for example, A100 or H100 accelerators). The energy and monetary cost of training large language models is substantial (Strubell, Ganesh, and McCallum 2019), which places such systems out of reach for many researchers with limited funding.
- Domain adaptation: General-purpose LLMs draw on a broad range of knowledge, so they may answer complex biomedical questions with generic explanations rather than the precise, evidence-based reasoning that specialized research use requires.
- Memory efficiency: Fine-tuning all parameters of an 8-billion-parameter model demands far more GPU memory and compute than a single low-cost or free-tier GPU can provide.

This paper examines whether domain-specific biomedical reasoning can be improved through 4-bit quantization and parameter-efficient adaptation on resource-constrained hardware. We frame the work as a computational feasibility study and address three research questions:

1. Can Llama-3-8B be adapted for biomedical reasoning using NF4 quantization and QLoRA on a single NVIDIA T4 GPU (16 GB VRAM)?
2. How does the adapted model perform on a PubMedQA subset and on questions grouped by reasoning complexity? Where does it fail?
3. What memory and runtime footprint does this adaptation require?

The study contributes (i) a documented single-GPU QLoRA configuration for biomedical reasoning; (ii) two study-level evaluations reported with exact counts, confidence intervals, and error analysis; and (iii) a discussion of the accessibility, ethical, and governance implications of low-resource medical AI. The name Bio-Llama-3-Expert is a project label. The word “Expert” does not imply expert-level or specialist clinical competence, and this paper makes no such claim.

Related Work

Language Models in Biomedicine

Natural language processing in medicine has progressed from rule-based systems with limited vocabularies (Friedman et al. 1994) to transformer-based language models. Early systems struggled to represent the nuance and complexity of clinical documentation. BERT (Devlin et al. 2019) marked a turning point by showing that a pre-trained model can be fine-tuned to perform well on many downstream tasks. Domain-specific pre-training followed: BioBERT (Lee et al. 2020) was pre-trained on PubMed abstracts and PMC full-text articles, and ClinicalBERT (Alsentzer et al. 2019) on clinical notes from the MIMIC-III database. Both improved performance on biomedical tasks such as named entity recognition, relation extraction, and question answering.

Decoder-only generative models, notably the GPT series (Brown et al. 2020; OpenAI 2023), extended these capabilities to text generation and reasoning. BioGPT (Luo et al. 2022) trained a generative transformer on biomedical literature, and ChatDoctor (Li et al. 2023) adapted a LLaMA model using medical dialogues. At the larger end, Med-PaLM (Singhal et al. 2023a), built on the 540-billion-parameter Flan-PaLM model, and Med-PaLM 2 (Singhal et al. 2023b) achieved high performance on medical licensing-style questions. Both are large proprietary systems that are not openly available for local adaptation.

Parameter-Efficient Fine-Tuning

Hu et al. (2021) proposed Low-Rank Adaptation (LoRA), which freezes the pre-trained weights and injects trainable rank-decomposition matrices into the transformer layers. LoRA matched the quality of full fine-tuning on the tasks the authors studied while training far fewer parameters, which is especially valuable for researchers with limited computational resources.

Dettmers et al. (2023) extended LoRA to QLoRA, which backpropagates through a frozen 4-bit quantized model into LoRA adapters. QLoRA made it possible to fine-tune a 65-billion-parameter model on a single 48 GB GPU and was reported to preserve the performance of 16-bit fine-tuning. It combines three techniques: 4-bit NormalFloat (NF4), a data type designed for normally distributed weights; double quantization, which quantizes the quantization constants to save additional memory; and paged optimizers, which manage memory spikes during training.

The Accessibility Challenge in Medical AI

Parameter-efficient methods allow more groups to adapt large models, but accessibility remains a challenge in medical AI. The strongest medical LLMs are proprietary or restricted-access systems built on large compute clusters (Singhal et al. 2023a; OpenAI 2023), and the cost of training large NLP models is considerable (Strubell, Ganesh, and McCallum 2019). Researchers in low and middle-income countries (LMICs), which account for a large share of the world's population, may therefore have limited access to the resources needed to build and evaluate sophisticated medical AI. Such constraints can hinder exploratory research and local model development, with consequences for global health equity.

One common response is model distillation (Sanh et al. 2019), in which a smaller "student" model is trained to mimic a larger "teacher." Distillation yields more manageable models, but distilled models can underperform their teachers, particularly on complex reasoning tasks. Their training still depends on access to the large teacher model.

Research Gaps and Study Goals

Three gaps motivate the present study:

- Clinical biomedical reasoning under tight compute budgets: QLoRA has been demonstrated on general-purpose language tasks (Dettmers et al. 2023), but little published work examines how 4-bit precision and low-rank adaptation trade off against performance for biomedical reasoning on a single low-cost GPU.
- Prompt engineering under resource constraints: Many medical question-answering approaches rely on long context windows or multiple inference passes, which can be impractical on resource-constrained devices. How far domain-specific prompt design can go under such limits is less well characterized.

- Minimum compute for useful biomedical models: Evidence on the minimum computational resources needed to build biomedically useful models is limited, which restricts who can take part in developing medical AI.

This study addresses these gaps by evaluating whether 4-bit QLoRA can adapt a mid-sized model (Llama-3-8B) for biomedical reasoning on one resource-constrained GPU, using domain-specific prompt design in the evaluation, and by reporting the resulting performance, error patterns, and computational footprint.

Technical Methodology

This section describes the model, quantization, adaptation method, and prompt design (Figure 1).

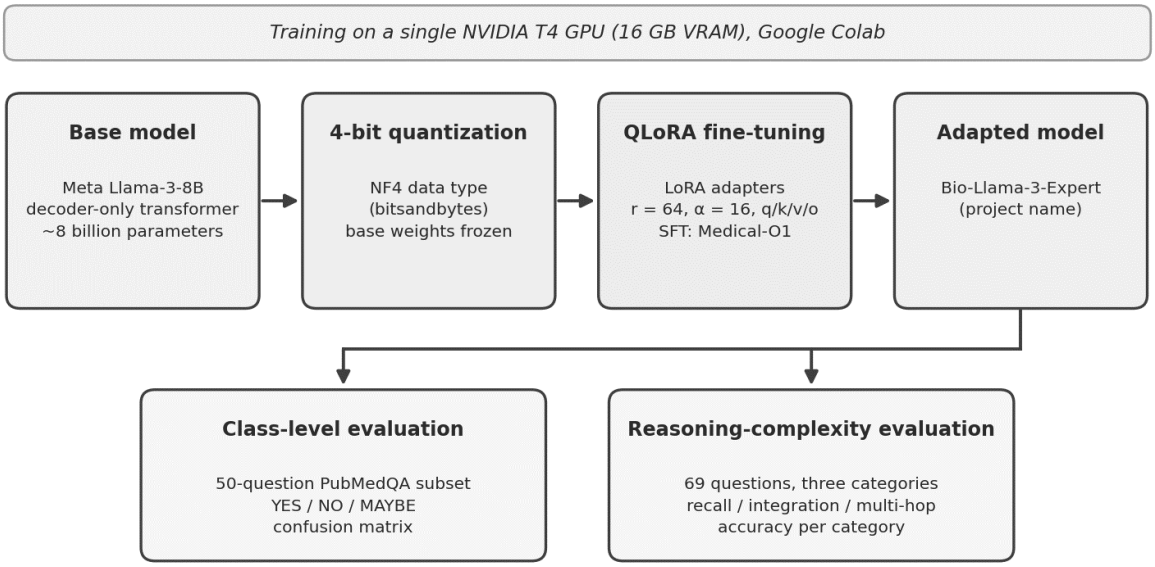


Figure 1. Overview of the study pipeline: 4-bit NF4 quantization of the frozen Llama-3-8B base weights, QLoRA fine-tuning on Medical-O1-Reasoning-SFT, and two study-level evaluations, all on a single NVIDIA T4 GPU.

Technical Stack and Environment
Table 1. Technical stack used in this study.

Component	Role
Base model	Meta Llama-3-8B (Meta AI 2024), a decoder-only transformer
Infrastructure	Google Colab with one NVIDIA T4 GPU (16 GB VRAM)
Framework	PyTorch with Hugging Face Transformers
Unsloth	Memory-efficient fine-tuning library
bitsandbytes	4-bit NormalFloat (NF4) quantization
PEFT	Parameter-efficient fine-tuning (LoRA/QLoRA adapters)
TRL	SFTTrainer for supervised fine-tuning

Quantization Strategy

An 8-billion-parameter model stored in 16-bit precision needs approximately 15 GiB for its weights alone (about 8.03 billion parameters × 2 bytes), which leaves essentially no headroom on a 16 GB T4 once training overhead is added. We therefore quantized the frozen base weights to 4 bits with the NF4 data type, which requires approximately 3.5–4 GiB (parameters × 0.5 bytes), an approximately four-fold reduction in weight storage. This leaves roughly 10 GiB of GPU memory for activations, adapter gradients, and optimizer states. These figures are weight-storage estimates; measured memory use is higher because of quantization constants, embedding layers kept at higher precision, and runtime overhead.

QLoRA Fine-Tuning Architecture

Rather than updating all 8 billion parameters, we trained small low-rank "adapter" matrices while keeping the quantized base weights frozen. For a frozen weight matrix W_0 with d rows and k columns, LoRA learns an update $\Delta W = BA$, where B is a $d \times r$ matrix, A is an $r \times k$ matrix, and the rank r is much smaller than both d and k. The adapted weight is

$$W_{\text{updated}} = W_{\text{frozen}} + (\alpha / r) \cdot BA$$

where W_{frozen} is the 4-bit quantized base weight (dequantized on the fly for computation), A and B are the trainable low-rank matrices, r is the rank, and α is a scaling hyperparameter. We used $r = 64$ and $\alpha = 16$, giving a scaling factor $\alpha/r = 0.25$, applied to the query, key, value, and output projections (q_proj, k_proj, v_proj, o_proj). The number of trainable parameters was 41,943,040 (0.52% of the model). A comparatively high rank increases adapter capacity; the effect of rank was not ablated in this study. This design preserves the pre-trained knowledge encoded in the frozen weights while enabling efficient domain adaptation.

Prompt Design

We designed a domain-specific prompt template to elicit structured, evidence-based reasoning followed by an explicit decision. The template consists of a role instruction and two worked exemplars that show the required format (Context, Question, Scientific Reasoning and Answer). The bracketed fields below are placeholders that illustrate the structure; they are not benchmark items.

```
You are a senior Biomedical Researcher. Analyze the following
study abstracts.

Example 1:
Context: [biomedical study context and reported evidence]
Question: [research question]
Scientific Reasoning and Answer: [evidence-based rationale and
decision]

Example 2:
Context: [biomedical intervention, reported outcome, and
relevant safety evidence]
Question: [safety or applicability question]
Scientific Reasoning and Answer: [evidence-based rationale and
decision]
```

Experimental Setup

Each element of the design maps to a research question. Feasibility is addressed by the single-T4 setting, the NF4 plus QLoRA method, and the biomedical training data described below. Performance and failure models are addressed by the two evaluation sets and the error analysis in Results. Footprint is addressed by the recorded runtimes, parameter counts, and memory estimates. Training diagnostics describe optimization behaviour; the two evaluations provide the direct performance measures.

Training Configuration

Supervised fine-tuning used the configuration in Table 2. Two training runs of 150 and 100 steps are summarized in Table 10, and the logged trace in Figures 3–7 spans 300 steps.

Table 2. QLoRA and training configuration.

Parameter	Value	Parameter	Value
Quantization	4-bit NF4	Gradient accumulation steps	8
LoRA rank (r)	64	Effective batch size	16
LoRA alpha (α)	16 (α/r = 0.25)	Learning rate	3e-5
Target modules	q_proj, k_proj, v_proj, o_proj	Scheduler	Cosine, warmup ratio 0.1
Trainable parameters	41,943,040 (0.52%)	Maximum sequence length	2048 tokens
Per-device batch size	2	Optimizer	AdamW 8-bit
Training data	Medical-O1-Reasoning-SFT (English)	Training precision	BF16

Datasets

Table 3. Datasets used in this study.

Dataset	Role	Size	Format
PubMedQA (pqa_labeled)	Evaluation	1,000 expert-annotated instances (official 500/500 split); 50 used here	Question, abstract context, YES/NO/MAYBE label
Medical-O1-Reasoning-SFT	Fine-tuning	About 19.7k rows in the English medical configuration (source-dataset size)	Question, Complex_CoT, Response

PubMedQA

The official PubMedQA release (Jin et al. 2019) contains 1,000 expert-annotated (PQA-L), 61.2k unlabelled, and 211.3k artificially generated question–answer instances. PQA-L is a three-class task with YES, NO, and MAYBE decisions, and the official repository splits it into 500 test and 500 cross-validation examples. This study used 50 questions for class-level evaluation, so the analysis is a study subset, not the full 500-question test split. The reference-label distribution of the subset, derived from the confusion matrix, is 20 YES, 21 NO, and 9 MAYBE.

Medical-O1-Reasoning-SFT

This dataset (FreedomIntelligence 2024; Chen et al. 2024) is a supervised fine-tuning resource for medical reasoning. According to its dataset card, examples were built from medically verifiable problems, with solutions generated and validated using GPT-4o, and the English medical configuration lists about 19.7k rows. The training set used here was drawn from this resource; the number of examples selected is not recorded (see Reproducibility and Reporting Scope).

Preprocessing

Supervised fine-tuning used a maximum sequence length of 2048 tokens and the domain-specific prompt formatting described in prompt design.

Research Ethics

Only publicly available datasets were used. The study involved no human participants and no patient-identifiable data.

Evaluation Design

Two study-level evaluations were run after training (Table 4). Accuracy is the proportion of questions for which the model's answer matched the reference. Because the sets are small, we report exact counts alongside 95% Wilson confidence intervals.

Table 4. Evaluation sets.

Evaluation	Questions	Content	Reported metrics
Class-level	50	PubMedQA (pqa_labeled) subset; YES/NO/MAYBE labels (20/21/9)	Accuracy, confusion matrix, per-class precision/recall/F1
Reasoning complexity	69 (25 / 24 / 20)	Questions grouped into three categories by the number of inferential steps required	Accuracy per category and overall

Why the Evaluation Sets Are Small

The study was scoped as a feasibility investigation on a single, resource-constrained GPU. Its aim was to test whether the quantization and adaptation pipeline produces measurable, interpretable behaviour, not to estimate population-level accuracy. Compact evaluation sets also allowed every error in the class-level set to be reviewed individually (see Error Analysis). The consequence is statistical imprecision: with 50 questions, the 95% interval around an observed accuracy spans roughly ±12 percentage points. Results are therefore treated as preliminary, and differences of a few percentage points are not interpreted.

Selection and Grading

The procedure used to select the 50 PubMedQA questions was not recorded, so no claim of random or stratified sampling is made. Reasoning-complexity categories were defined by the study team according to the number of inferential steps required (see Reasoning-Complexity Analysis); no inter-rater agreement was measured.

Inference Settings

Table 5. Inference configuration recorded in the accompanying inference notebook.

Setting	Value
Model loading	4-bit loading (load_in_4bit=True); bfloat16 dtype; automatic device mapping
Prompt format	Llama-3 chat template applied through the tokenizer; system and user messages
Decoding	Sampling (do_sample=True), temperature 0.7, top-p 0.9
Maximum new tokens	512
Stop token	Tokenizer end-of-sequence token
Output handling	Input prompt removed from the generated text before display
Inference runtime	Kaggle Python 3 environment (Python 3.12)

Because decoding uses sampling, repeated runs can produce different outputs for the same question. The notebook demonstrates inference; it does not contain the training procedure or a routine that maps free-text outputs to YES/NO/MAYBE.

Reproducibility and Reporting Scope

The study documents the hardware, software stack, quantization scheme, LoRA configuration, optimizer and schedule (Table 2), data sources, evaluation structure, and inference settings (Table 5). The following were not archived: exact library and CUDA versions; the training random seed; the dataset revision and the number of Medical-O1 examples used; the tokenizer version and the truncation and deduplication procedures; the item-level lists for the two evaluation sets; the rule for mapping free-text outputs to class labels; the evaluation protocols of the comparison configurations; and measured peak GPU memory. Because the original training script was not preserved, exact reproduction of the reported numbers cannot be guaranteed. The configuration above is intended to support a close re-implementation, and a future public release should pin software versions, seeds, the dataset revision, and item identifiers (see Future Work).

Iterative Development History

Development proceeded through four versions (Table 6). The outcomes listed are qualitative observations made during development, not benchmarked results.

Table 6. Development history of the project.

Version	Focus	Method	Observed outcome (qualitative)
V1	Phi-3 Mini baseline; initial testing	Base fine-tuning	Limited in complex medical reasoning
V2	Llama-3-8B base; architecture shift	4-bit quantization	Improved logic, but answers remained at a lay level
V3	Bio-instruction tuning	LoRA with OASST (OpenAssistant) instruction data and a biomedical dataset	Markedly more medical terminology
V4	Bio-Llama-3-Expert (final project configuration)	QLoRA with hyperparameter optimization	Improved biomedical reasoning; evaluated in Results

Results

Class-Level Evaluation (50 PubMedQA Questions)

Figure 2 shows the confusion matrix for the 50-question class-level evaluation. The model classified 36 of 50 questions correctly, an accuracy of 72.0% (95% Wilson CI, 58.3–82.5%). Per-class metrics are given in Table 7.

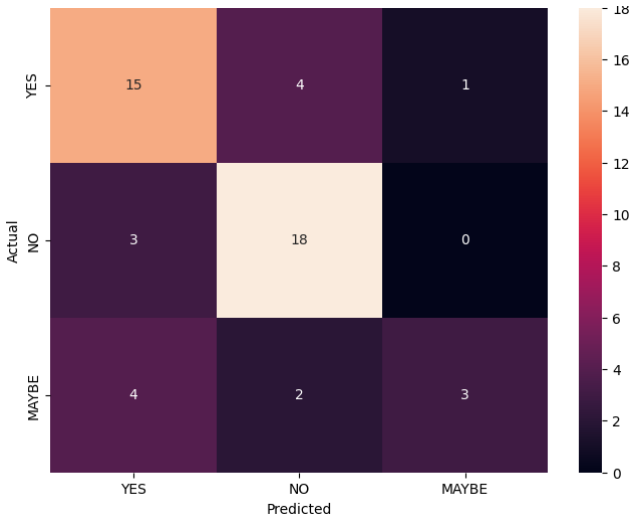


Figure 2. Confusion matrix for the 50-question class-level evaluation (rows: actual label; columns: predicted label). Diagonal cells (15, 18, 3) are correct predictions.

Table 7. Per-class performance derived from the confusion matrix in Figure 2.

Class	Support	Predicted	Precision (%)	Recall (%)	F1 (%)
YES	20	22	68.2	75.0	71.4
NO	21	24	75.0	85.7	80.0
MAYBE	9	4	75.0	33.3	46.2
Macro average	**50**	**50**	**72.7**	**64.7**	**65.9**
Accuracy	**50**	**50**	—	**72.0 (36/50)**	—

Performance was strongest for NO (recall 85.7%, F1 80.0%) and weakest for MAYBE (recall 33.3%, F1 46.2%). Six of the nine MAYBE questions were misclassified, four as YES and two as NO, which accounts for six of the 14 errors (43%). MAYBE precision (75.0%) rests on only four MAYBE predictions and is therefore unstable. The macro-averaged F1 score was 65.9%.

Benchmark Context

The label distribution of this subset (20 YES, 21 NO, 9 MAYBE) differs from that of the full PQA-L set, in which YES is the majority class at roughly 55% (Jin et al. 2019). The majority-class rate in the subset is 42.0% (21/50), and the lower bound of the confidence interval for the model's accuracy (58.3%) lies above it. For orientation only, Jin et al. (2019) reported 68.1% accuracy for their best BioBERT system, 78.0% for single-human performance, and 55.2% for the majority baseline on PQA-L. Because the present subset is small, was not drawn to match the full label distribution, and was scored under a different protocol, those figures are context and not a head-to-head comparison.

Reasoning-Complexity Analysis (69 Questions)

To examine what kinds of tasks the model can handle, we grouped 69 test questions into three categories by the number of inferential steps needed to answer them (Table 8). This set is separate from the PubMedQA subset, so its accuracy is not directly comparable with the 72.0% class-level figure: the two evaluations differ in questions, answer formats, and difficulty.

Table 8. Accuracy by reasoning-complexity category (95% Wilson confidence intervals).

Category	n	Correct	Accuracy (%)	95% CI (%)	Reported baseline
1. Direct recall	25	23	92.0	75.0–97.8	—
2. Integration of related concepts	24	19	79.2	59.5–90.8	—
3. Multi-hop reasoning	20	13	65.0	43.3–81.9	30%*
Overall	**69**	**55**	**79.7**	**68.8–87.5**	—

*Reported baseline for Category 3 from an earlier project evaluation. Its denominator and protocol were not archived, so no paired improvement is calculated.

Category 1 (Direct Recall)

These questions required only direct recall of biomedical information (for example, "Name the mechanism by which metformin works"). The model answered 23 of 25 correctly (92.0%), indicating strong performance on fundamental biomedical knowledge.

Category 2 (Integration)

These questions required integrating two or three related concepts (for example, "In a patient with chronic kidney disease, do you predict higher or lower levels of vitamin D being utilized?"). The model answered 19 of 24 correctly (79.2%), indicating moderate ability on basic clinical reasoning.

Category 3 (Multi-hop Reasoning)

These questions required a multi-hop reasoning chain across several distinct concepts. The model answered 13 of 20 correctly (65.0%). Across the three categories the model answered 55 of 69 questions correctly (79.7%). Accuracy fell as the number of inferential steps rose, from 92.0% to 79.2% to 65.0%, which indicates that multi-step reasoning remained the most challenging. The confidence intervals overlap, in particular between adjacent categories, so the trend is suggestive and not conclusive at this sample size.

Error Analysis

Fourteen of the 50 class-level questions (28.0%) were answered incorrectly. By label, the errors were YES predicted as NO (4), YES as MAYBE (1), NO as YES (3), MAYBE as YES (4), and MAYBE as NO (2) (Figure 2). Review of the incorrect responses identified four recurring patterns. Per-pattern counts were tabulated only for temporal reasoning; the remaining patterns are described qualitatively.

Temporal Reasoning

Four of the 50 questions (8%) were answered incorrectly because the model mishandled temporal relationships, such as how a disease or event evolves and how it relates in time to treatment. For example, when asked about the evolution of the immune response to vaccination, the model confused the acute and memory responses.

Quantitative Thresholds

The model sometimes failed to judge whether a numerical value was significant in context. Asked, for example, "Is a serum creatinine of 1.4 mg/dL concerning in a 75-year-old patient?", it applied the same standard threshold across all three patient ages tested without adjusting for age.

Novel Therapy Combinations

The model produced incorrect reasoning on some drug–drug interaction questions and novel therapy combinations that were sparsely represented in the training data, which points to limited transfer of learned patterns to unfamiliar combinations.

Ambiguous Questions

On a high-level medical question for which two major organizations issue differing guidelines, the model selected one answer although the other was equally valid under the other organization's

guideline. This illustrates the difficulty of questions whose correct answer is legitimately uncertain, and it is consistent with the weak MAYBE-class performance in Table 7.

Comparison with Alternative Configurations

We also recorded results for alternative configurations that are feasible on similar hardware (Table 9). These figures come from earlier project evaluations whose denominators and protocols were not archived, and the 45% naive-prompting figure serves as the reference baseline for the class-level evaluation. Differences from the QLoRA result should therefore be read as directional and not as paired percentage-point improvements.

Table 9. Reported accuracy of alternative configurations and of the QLoRA-adapted model.

ID	Configuration	Reported accuracy	Evidence
A	Llama-3-8B, no fine-tuning, naive prompting	45%	Reported (earlier evaluation)
B	Llama-3-8B, no fine-tuning, extended prompt engineering	62%	Reported (earlier evaluation)
C	Smaller 3B-parameter Llama-family model, fine-tuned in full precision	67%	Reported (earlier evaluation)
D	**Bio-Llama-3-Expert: Llama-3-8B, 4-bit NF4 + QLoRA**	**72.0% (36/50)**	**This study, Class-Level Evaluation**

The ordering is consistent with a gradual improvement: extended prompting alone (B) exceeded naive prompting (A), fine-tuning a smaller model (C) exceeded both, and the QLoRA-adapted 8B model (D) reached the highest reported value. The study was not designed to isolate the separate contributions of model scale, prompt design, and fine-tuning, and configurations A–C were not evaluated under a documented protocol shared with D. The comparison supports the feasibility claim but does not establish superiority.

Training Diagnostics

Table 10 summarizes two recorded training runs on the T4, and Figures 3–7 show the logged trace over 300 steps. The runs were labelled “gold-standard” (150 steps) and “sharpening” (100 steps) in the project record.

Table 10. Summary statistics of the two recorded training runs.

Metric	Run 1 (150 steps)	Run 2 (100 steps)
Training loss	1.5178	1.5184
Runtime (s)	1768.87 (about 29.5 min)	1136.53 (about 18.9 min)
Samples per second	0.678	0.704
Steps per second	0.085	0.088
Total FLOPs	2.63×10^{16}	1.70×10^{16}

Loss

Training loss falls from about 1.6–1.7 in the first steps to a band of roughly 1.4–1.55, with large step-to-step fluctuation and no clear sustained decline after about step 50. The run-level averages in Table 10 (1.5178 and 1.5184) sit inside this band (Figure 3).

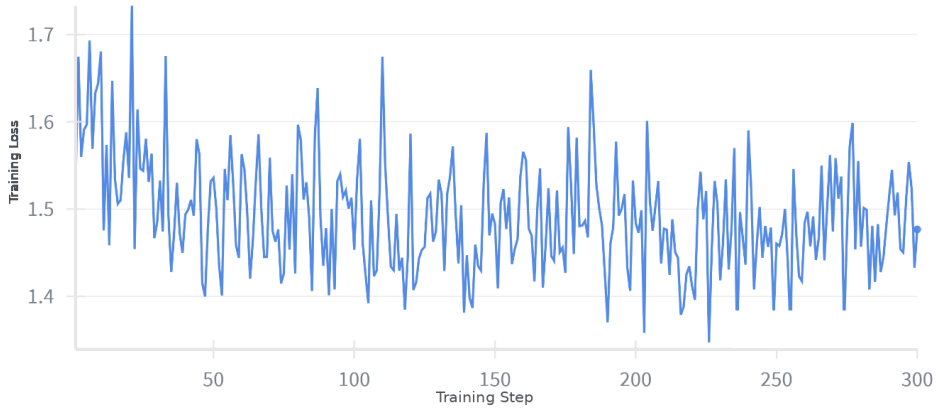


Figure 3. Training loss over the 300-step logged trace.

Learning-rate Schedule

The learning rate rises linearly to its peak of 3×10^{-5} at about step 30 (10% of 300 steps, matching the warmup ratio in Table 2) and then decays along a cosine curve (Figure 4).

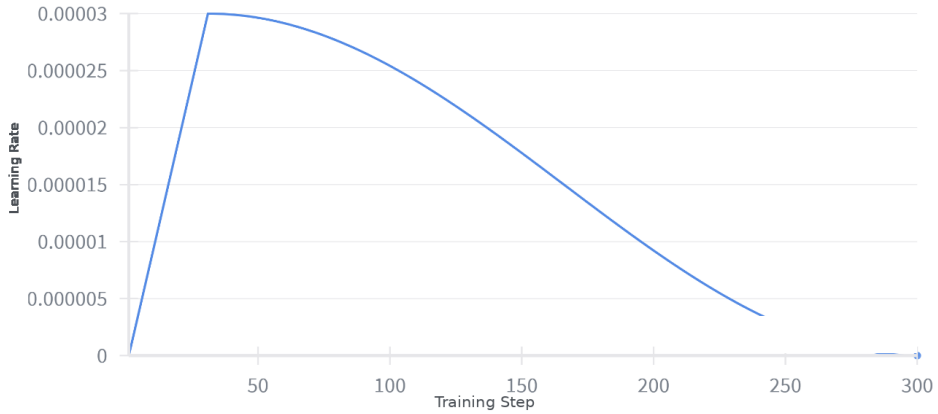


Figure 4. Learning-rate schedule: linear warmup followed by cosine decay.

Gradient Norm

The gradient norm starts high (peaks near 0.6) and settles to a stable range of about 0.25–0.30 within roughly 40 steps, with no divergence (Figure 5).

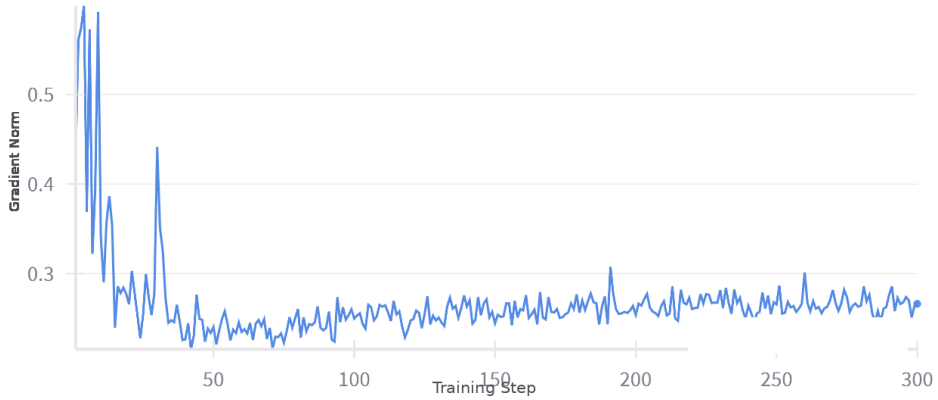


Figure 5. Gradient norm over the logged trace.

Progress Counters

The global step advances linearly to 300, and the epoch counter reaches about 0.87 at step 300, so the plotted run completed less than one pass over its training data (Figures 6 and 7).

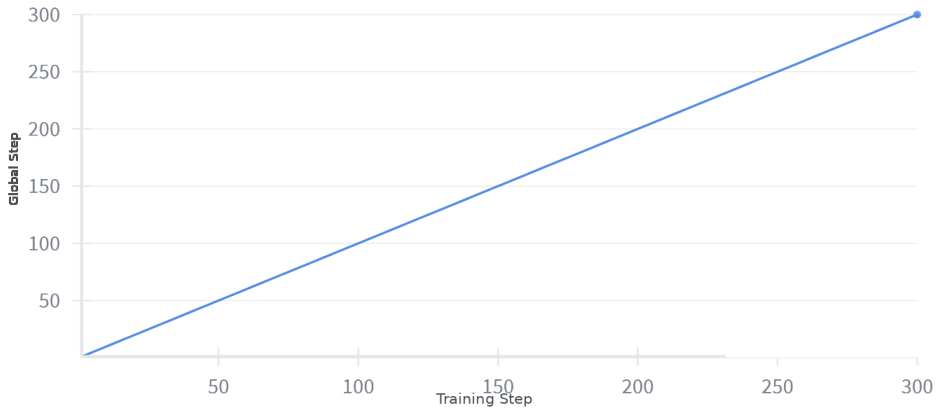


Figure 6. Global step against training step.

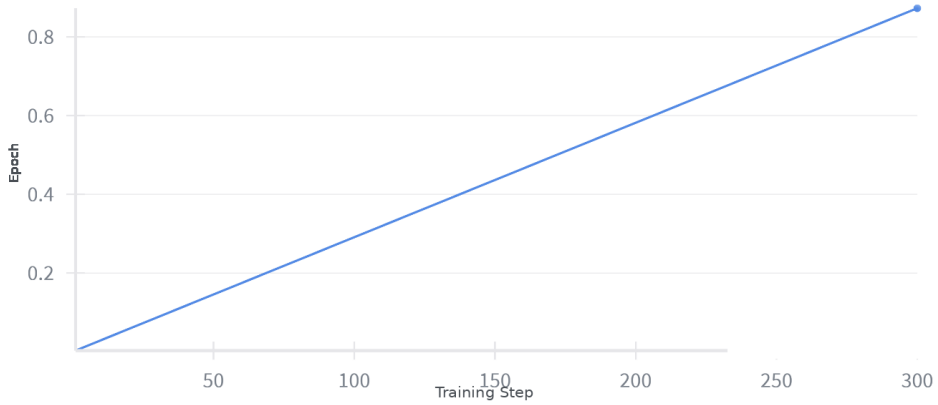


Figure 7. Epoch progress against training step.

Together, these diagnostics indicate stable optimization without divergence. They are not measures of task performance: the modest loss reduction and sub-epoch training suggest a light-touch adaptation, and the class-level and reasoning-complexity evaluations remain the direct performance measures.

Hardware Efficiency

Table 11. Hardware and efficiency summary.

Metric	Value
Hardware	Single NVIDIA T4 GPU, 16 GB VRAM (Google Colab)
Weight storage (theoretical)	About 15 GiB (16-bit) to about 3.5–4 GiB (NF4): roughly a four-fold reduction
Trainable parameters	41,943,040 (0.52% of the model)
Recorded training runtime	1768.87 s (150 steps); 1136.53 s (100 steps)

Qualitative Analysis: Case Study

To illustrate the qualitative change in output, we compared an early configuration (V2) with the final configuration (V4) on one topic: the mechanism of mRNA vaccines. The inference notebook uses the prompt "Explain the mechanism of action of mRNA vaccines and their specific impact on the spike protein of SARS-CoV-2." Table 12 shows the two outputs.

Table 12. Case study: mRNA vaccine mechanism, before and after fine-tuning.

Before fine-tuning (V2)	After fine-tuning (V4)
"mRNA vaccines work by teaching cells to make a protein that triggers an immune response."	"The mRNA construct incorporates N1-methylpseudouridine modifications to enhance stability and reduce innate immune activation. The encoded spike protein contains the 2P mutation (K986P, V987P) to stabilize the prefusion conformation, optimizing epitope presentation for neutralizing antibody generation."
General-audience, high-level explanation	*Technical, mechanism-level explanation*

The V4 output includes four specific points: (1) N1-methylpseudouridine modifications; (2) the 2P mutation nomenclature (K986P, V987P); (3) stabilization of the prefusion conformation; and (4) the implication for neutralizing antibody generation. These statements are consistent with the published literature on mRNA vaccine design. A single example, however, cannot establish general reliability. Generated text can contain errors and should be verified by a qualified expert before any use.

Discussion

Principal Findings

- Feasibility: A Llama-3-8B model was adapted with NF4 quantization and QLoRA on a single NVIDIA T4 GPU, training only 0.52% of the parameters. The recorded runs took about 29.5 and 18.9 minutes.
- Performance: The model reached 72.0% accuracy (95% CI, 58.3–82.5%) on the 50-question PubMedQA subset, above the 42.0% majority-class rate of that subset, and 79.7% (95% CI, 68.8–87.5%) on the 69-question reasoning-complexity set, with accuracy decreasing as reasoning depth increased.
- Failure modes: Errors clustered in the MAYBE class, temporal reasoning, age- and context-dependent thresholds, novel drug combinations, and questions with legitimately divergent guidelines.

Interpretation

The results support the claim that a mid-sized open model can be adapted for biomedical question answering on a single low-cost GPU. They do not support claims of superiority over other systems or of clinical readiness. The reported comparison configurations (Table 9) were not evaluated under a shared, documented protocol, and the PubMedQA reference figures from the original benchmark paper cannot be compared directly with a 50-question subset that has a different label distribution. The MAYBE-class result is a reminder that calibrated uncertainty is a distinct capability from picking a definite answer, and it matters in clinical contexts.

Technical Trade-offs and Design Decisions

The 4-bit NF4 configuration reduces weight storage roughly four-fold (from about 15 GiB to about 3.5–4 GiB), which is what allows an 8-billion-parameter model to be adapted within 16 GB. The cost of quantization for this task was not measured, because no 8-bit or 16-bit comparison was run. The LoRA settings ($r = 64$, $\alpha = 16$, scaling factor 0.25) were adopted as the project configuration through iterative development; rank, alpha, and target-module choices were not systematically ablated, so no performance or cost multiplier is claimed for them. Such ablations are a priority for future work.

Computational Efficiency

The recorded training runs were measured directly on the single T4: 1768.87 s for the 150-step run and 1136.53 s for the 100-step run. At these speeds a researcher could complete several short adaptation experiments in a day on the same hardware, although session limits on free-tier platforms may constrain this in practice. The study did not run a full-precision fine-tuning baseline or price a cloud alternative, so no speedup or cost-saving multiplier is claimed.

Reproducibility

The study was designed to support reproducibility by documenting the hardware, training configuration, datasets, evaluation structure, and inference settings. It relies on widely available tools (PyTorch, Hugging Face Transformers, Unsloth, and PEFT) and on public datasets (PubMedQA and Medical-O1-Reasoning-SFT). The items listed in Reproducibility and Reporting Scope were not

archived, so exact reproduction of the reported numbers cannot be guaranteed; a close re-implementation is nonetheless feasible from the information provided.

Limitations

- Small evaluation sets: The evaluations use 50 and 69 questions, with 95% intervals of roughly ±12 and ±9 percentage points. The 50-question subset was not selected by a recorded sampling procedure, and its label distribution differs from that of the full PQA-L set. Larger, independently sampled evaluations would give more reliable estimates.
- Unpaired baselines: The comparison configurations in Table 9 and the Category 3 baseline were not evaluated under a documented shared protocol, so paired improvements cannot be calculated.
- Short training and synthetic training data: Training runs were short (at most 300 logged steps, under one epoch), and the number of training examples used is not recorded. Medical-O1-Reasoning-SFT responses were generated and validated with GPT-4o, so errors and biases of that model may propagate.
- No ablations: The effects of bit width, LoRA rank, and prompt design were not isolated.
- Uncertainty handling: The confusion matrix shows weak performance on the MAYBE class (3 of 9 correct), which limits utility where uncertainty must be expressed.
- Evaluation procedure: Decoding used sampling (temperature 0.7), so outputs can vary between runs, and the answer-mapping rule was not archived. Reasoning-complexity categories were defined by the study team, and no inter-rater agreement was measured.
- Scope of data: All evaluation was text-only and English-only.
- No clinical validation: The model has not been validated in clinical settings and would require careful clinical validation and regulatory approval before any patient-facing use.

Future Work

- Repeat the evaluation on larger, independently sampled sets with paired baselines, and release item identifiers, seeds, the dataset revision, pinned software versions, and answer-parsing code.
- Ablate bit width (4-, 8-, and 16-bit), LoRA rank and alpha, target modules, and prompt design; measure peak memory and energy use.
- Compare against openly available biomedical models such as BioGPT and ChatDoctor.
- Extend to multilingual medical adaptation (for example, Hindi, Swahili, and Arabic) and to multimodal inputs such as images, laboratory values, and clinical notes.
- Integrate retrieval-augmented generation for access to current literature.
- Explore federated learning so that institutions can improve models collaboratively while keeping patient data local, and test whether adapters trained for one medical specialty can be composed with adapters from another.
- Conduct independent clinical evaluation and safety assessment with clinicians and, where appropriate, patient representatives.

Social Impact and Ethical Considerations

Global Health Equity

Bio-Llama-3-Expert is a research prototype, but the question it addresses, who can build and evaluate medical AI, has broader implications. Development of large medical AI systems has been concentrated in well-resourced institutions in high-income countries, while researchers in LMICs are more often users than developers of these technologies. This concentration matters for health equity. Models trained primarily on data from a few populations may generalize poorly to others and may widen existing disparities. Without the resources to develop and validate models on local data, researchers in LMICs may depend on models that do not reflect local epidemiology, genetic diversity, or health-system realities.

This study shows that a single NVIDIA T4 GPU is enough to run a first adaptation experiment on an 8-billion-parameter model. That lowers one hardware barrier to experimentation. It does not, by itself, establish clinical readiness or suitability for patient care.

Potential Impact on Healthcare Delivery

The World Health Organization's 2023 health workforce support and safeguards list identifies 55 countries facing the most pressing health workforce challenges related to universal health coverage, including low densities of doctors, nurses, and midwives (World Health Organization 2023). AI-based decision-support tools could give non-specialist health workers additional informational support in such settings, but only with appropriate validation, clinical governance, and qualified human oversight. As an illustration, a locally hosted biomedical language model could support literature interpretation and educational activities in a rural clinic with limited specialist access. Bio-Llama-3-Expert has not been validated for such uses, does not retrieve current literature, and must not substitute for the judgment of a qualified health professional.

Ethical Considerations and Responsible Deployment

Validation and Labelling

Broader accessibility raises the risk that tools are deployed before they are adequately validated. Accuracy on biomedical question-answering benchmarks does not establish that a model is safe or effective with patients. We suggest a minimum framework for responsible deployment: (1) mandatory clinical validation studies in representative patient populations before any patient-facing use; (2) labelling of outputs as "decision support" and not as a "diagnostic tool"; and (3) integration into existing clinical governance, with accountability and error-reporting mechanisms.

Bias. Making model development possible for more people is a positive goal, but if base models and training data reflect the dominant patterns of medical research and practice, wider access also spreads those biases. PubMedQA is built from PubMed abstracts, most of which appear in English-language journals, so medical knowledge from other regions, including parts of Asia and Africa, may be under-represented. The Medical-O1 responses were generated and validated with GPT-4o, so that model's errors and biases may also propagate. Multilingual medical datasets and validation across diverse patient populations are needed, and future studies should test whether parameter-efficient fine-tuning can adapt medical LLMs to local languages and contexts.

Governance and Dual Use

Building on an openly available base model (Meta Llama 3) and describing the method openly allows others to reproduce and extend the work, but it also allows uses that the authors cannot anticipate or monitor. We support community-led governance structures for medical AI in which researchers, clinicians, patients, and ethicists develop standards and norms for responsible design and use.

Environmental Considerations

Environmental cost is an important consideration in machine learning (Strubell, Ganesh, and McCallum 2019). This study documents the use of a single T4 GPU and the runtimes of two training runs. Energy consumption was not measured, and no full-fine-tuning comparison was run, so the study does not quantify an energy or emissions advantage. Local deployment may reduce reliance on cloud inference, but whether it lowers net energy use depends on utilization and was not assessed here.

Workforce and Medical Education

Accessible AI has implications for medical education and the health workforce. We view such systems as tools that augment, and do not replace, human expertise. Systems like Bio-Llama-3-Expert could serve as interactive study aids that explain complex biomedical concepts, and the case study in Qualitative Analysis: Case Study hints at that potential. A single example cannot establish educational value, so controlled studies are needed. For practicing physicians in generalist or underserved settings, AI systems could offer quick access to specialty knowledge, which may be most useful for rare diseases and newer treatments.

Growing reliance on AI could erode clinical skills and judgment if it is not managed. Medical education should therefore include AI literacy, an understanding of algorithmic bias, and training in critically evaluating the limitations and possible errors of AI outputs, so that future clinicians can work alongside these systems while retaining independent judgment.

Policy Recommendations

Drawing on this work, we offer the following recommendations for the fair development and use of medical AI:

- Funding: Funding agencies should value research on access and deployability in resource-limited settings. Criteria that reward only peak benchmark performance can favor systems that require hardware unavailable in many settings.
- Benchmarks: Medical AI benchmarks should report computational efficiency and accessibility alongside accuracy, so that predictive performance is considered together with the resources needed to achieve and deploy it.
- Regulation: Regulatory frameworks should be developed with health professionals and patient advocates from LMICs. Existing regulations largely reflect the settings of high-income countries and may create barriers elsewhere.
- Open development: Open-source medical AI should be supported through funding, recognition, and targeted incentives. Concentrating capability in a few proprietary systems creates dependencies that can entrench global inequality.

- Local capacity: Global collaborations should prioritize building local capability (training programs, infrastructure, and support for local datasets and model validation) over simply importing finished technology.

Conclusion

This study demonstrates the feasibility of adapting Llama-3-8B for biomedical reasoning using 4-bit NF4 quantization and QLoRA on a single NVIDIA T4 GPU, with only 0.52% of parameters trained and recorded training runs of under 30 minutes. On a 50-question PubMedQA subset the adapted model reached 72.0% accuracy (95% CI, 58.3–82.5%), and on a separate 69-question reasoning-complexity set it reached 79.7%, with accuracy declining as reasoning depth increased. These findings support further investigation of resource-efficient approaches to biomedical language modelling. However, the small evaluation sets, unpaired baselines, short training, benchmark-specific nature of the experiments, and absence of clinical validation limit how far the findings can be generalized to real-world clinical settings. Future work should evaluate the approach on larger and more diverse datasets with paired baselines, ablate the key design choices, and pursue independent medical evaluation and clinically relevant validation.

References

Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. "Publicly Available Clinical BERT Embeddings." In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-1909.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Chen, Junying, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. "HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs." arXiv:2412.18925. Preprint, arXiv. https://arxiv.org/abs/2412.18925.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. "QLoRA: Efficient Finetuning of Quantized LLMs." *Advances in Neural Information Processing Systems* 36. https://arxiv.org/abs/2305.14314.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 4171–86. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

FreedomIntelligence. 2024. *Medical-O1-Reasoning-SFT*. Dataset. Hugging Face. https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT.

Friedman, Carol, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. "A General Natural-Language Text Processor for Clinical Radiology." *Journal of the American Medical Informatics Association* 1 (2): 161–74. https://doi.org/10.1136/jamia.1994.95236146.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv:2106.09685. Preprint, arXiv. https://arxiv.org/abs/2106.09685.

Jin, Qiao, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. "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)*, 2567–77. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1259.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." *Bioinformatics* 36 (4): 1234–40. https://doi.org/10.1093/bioinformatics/btz682.

Li, Yunxiang, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge." arXiv:2303.14070. Preprint, arXiv. https://arxiv.org/abs/2303.14070.

Luo, Renqian, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. "BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining." *Briefings in Bioinformatics* 23 (6): bbac409. https://doi.org/10.1093/bib/bbac409.

Meta AI. 2024. "Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date." Meta AI Blog. https://ai.meta.com/blog/meta-llama-3/.

OpenAI. 2023. "GPT-4 Technical Report." arXiv:2303.08774. Preprint, arXiv. https://arxiv.org/abs/2303.08774.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." arXiv:1910.01108. Preprint, arXiv. https://arxiv.org/abs/1910.01108.

Singhal, Karan, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. 2023a. "Large Language Models Encode Clinical Knowledge." *Nature* 620 (7972): 172–80. https://doi.org/10.1038/s41586-023-06291-2.

Singhal, Karan, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, et al. 2023b. "Towards Expert-Level Medical Question Answering with Large Language Models." arXiv:2305.09617. Preprint, arXiv. https://arxiv.org/abs/2305.09617.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–50. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355.

World Health Organization. 2023. *WHO Health Workforce Support and Safeguards List 2023*. https://www.who.int/publications/i/item/9789240069787.