MediSimplifier: Medical Text Simplification Using LoRA Fine-Tuning

Technion DS25 Deep Learning Final Project

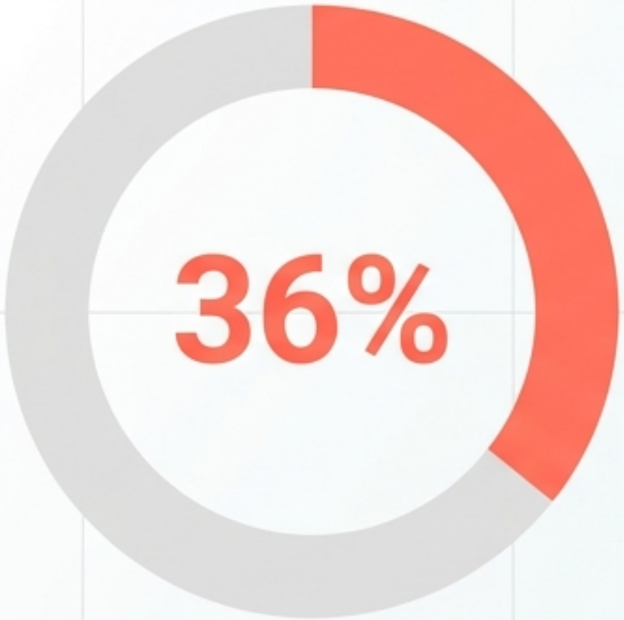
Authors: Guy Dor & Shmulik Avraham

January 2026 | IEEE Report & Project Synthesis

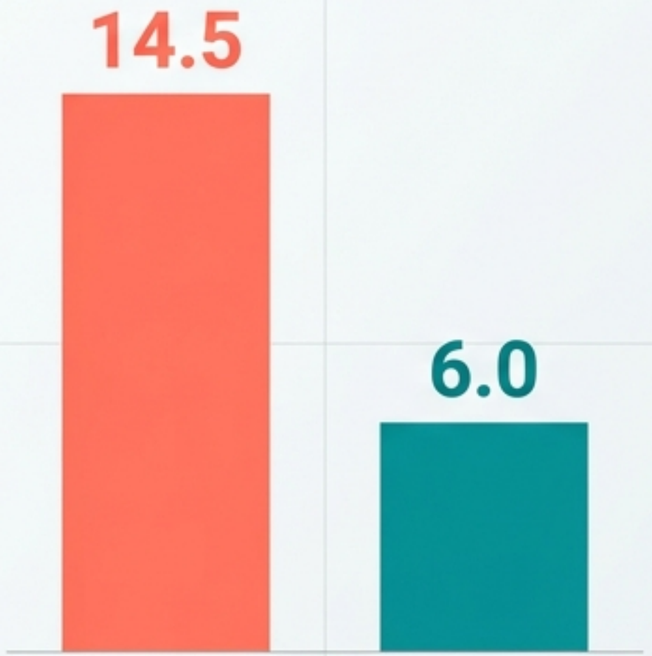
Agenda

1. Problem & Motivation: The Health Literacy Crisis
2. Project Objectives & Research Questions
3. Methodology: Data & Models
4. Baseline Performance (Zero-Shot)
5. The Solution: 4-Phase Pipeline & Ablation
6. Key Results: The Ranking Reversal
7. Analysis: Statistics & Errors
8. Conclusions & Future Work

The Health Literacy Crisis



of US adults have basic or below-basic health literacy.

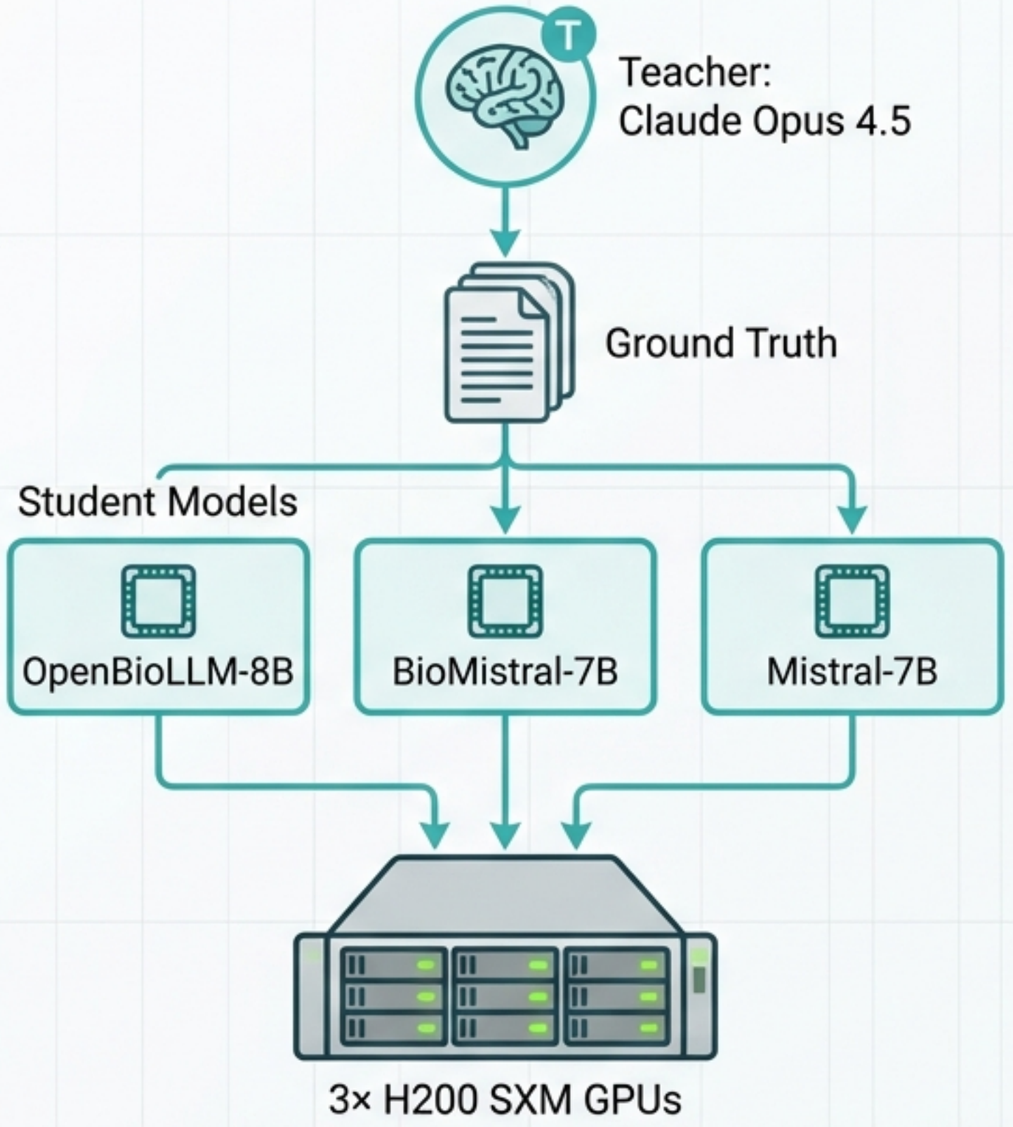


The Gap: Low comprehension → Poor adherence → Higher readmission rates.

Goal: Automatically simplify medical text to a 6th-grade reading level.

Project Overview

- **Objective:** Automate simplification of clinical notes to 6th-grade reading level.
- **Method:** Low-Rank Adaptation (LoRA) Fine-Tuning.
- **Dataset:** Asclepius-Synthetic-Clinical-Notes (10,000 samples).



12 Research Questions

Zero-Shot & Baselines

RQ1: Medical pretraining advantage?

RQ2: Best zero-shot model?

Fine-Tuning Dynamics

RQ3: Improvement over zero-shot?

RQ5: Ranking changes?

RQ8: Correlation: Baseline vs. Improvement?

Optimization (Ablation)

RQ4: Optimal Rank (r)?

RQ6: Best target modules?

RQ7: Data size impact?

RQ12: rsLoRA necessity?

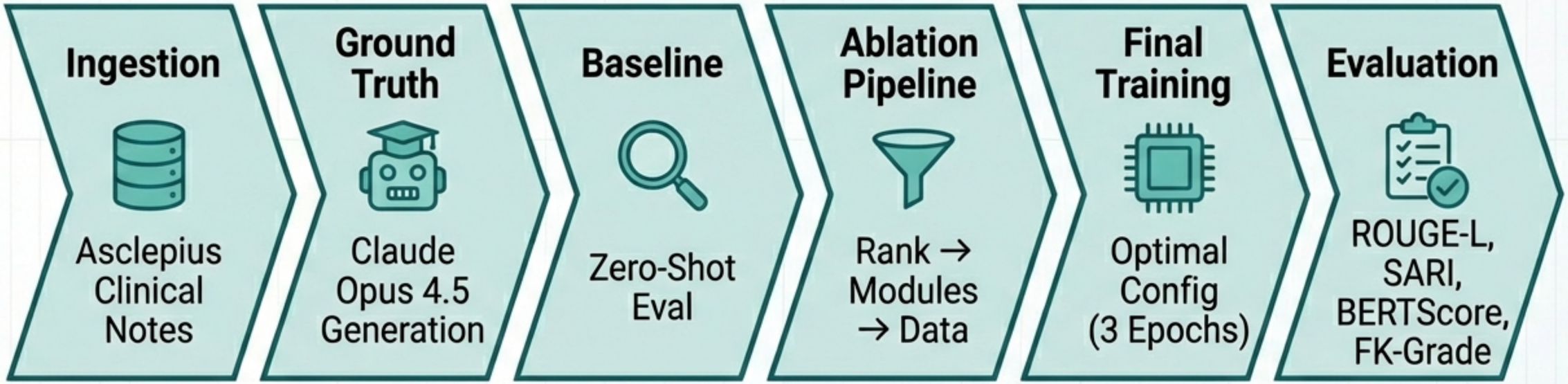
Evaluation

RQ9: Parameter efficiency?

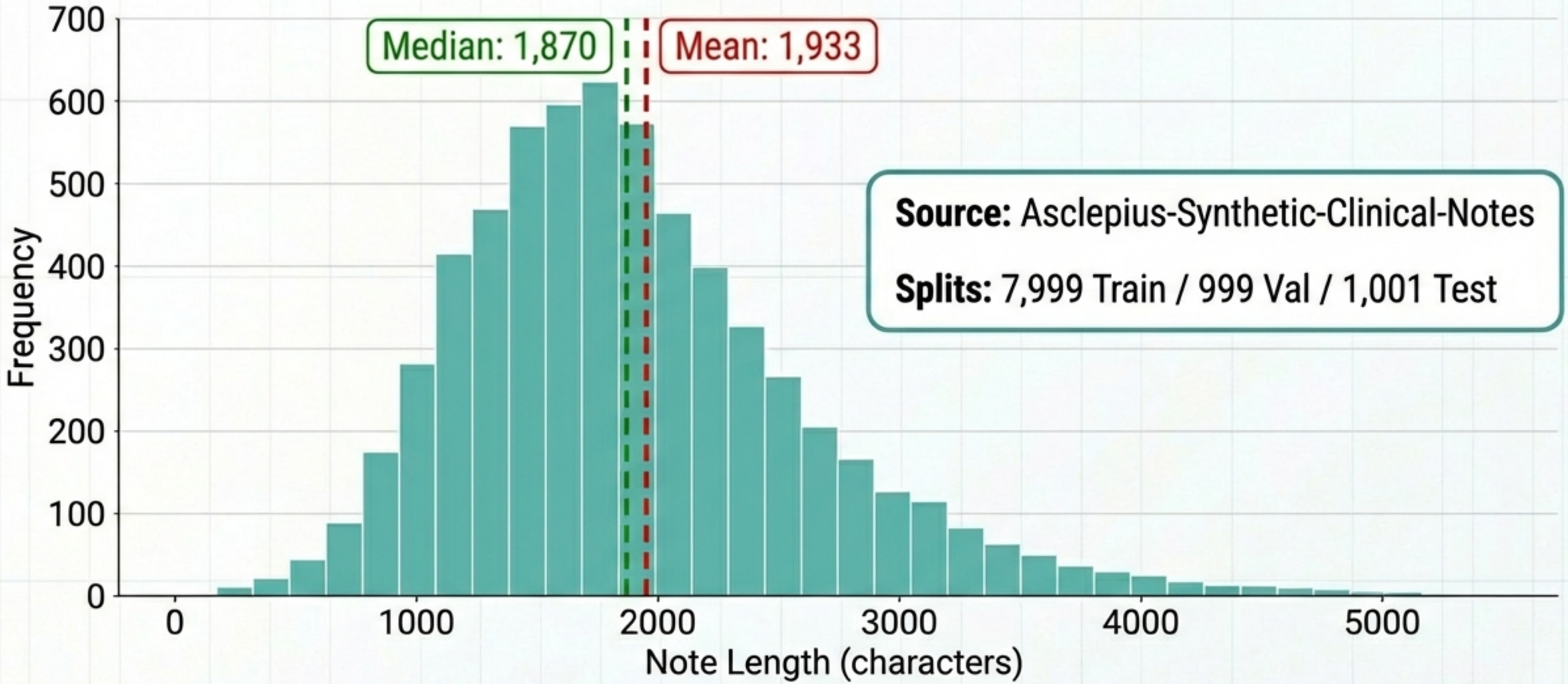
RQ10: Consistency?

RQ11: Readability target met?

Process Design: From Ingestion to Evaluation



Dataset: Clinical Note Length Distribution



Model Selection

OpenBioLLM-8B

Medical Domain

Llama 3
Architecture

BioMistral-7B-DARE

Medical Domain

Mistral
Architecture

Mistral-7B-Instruct-v0.2

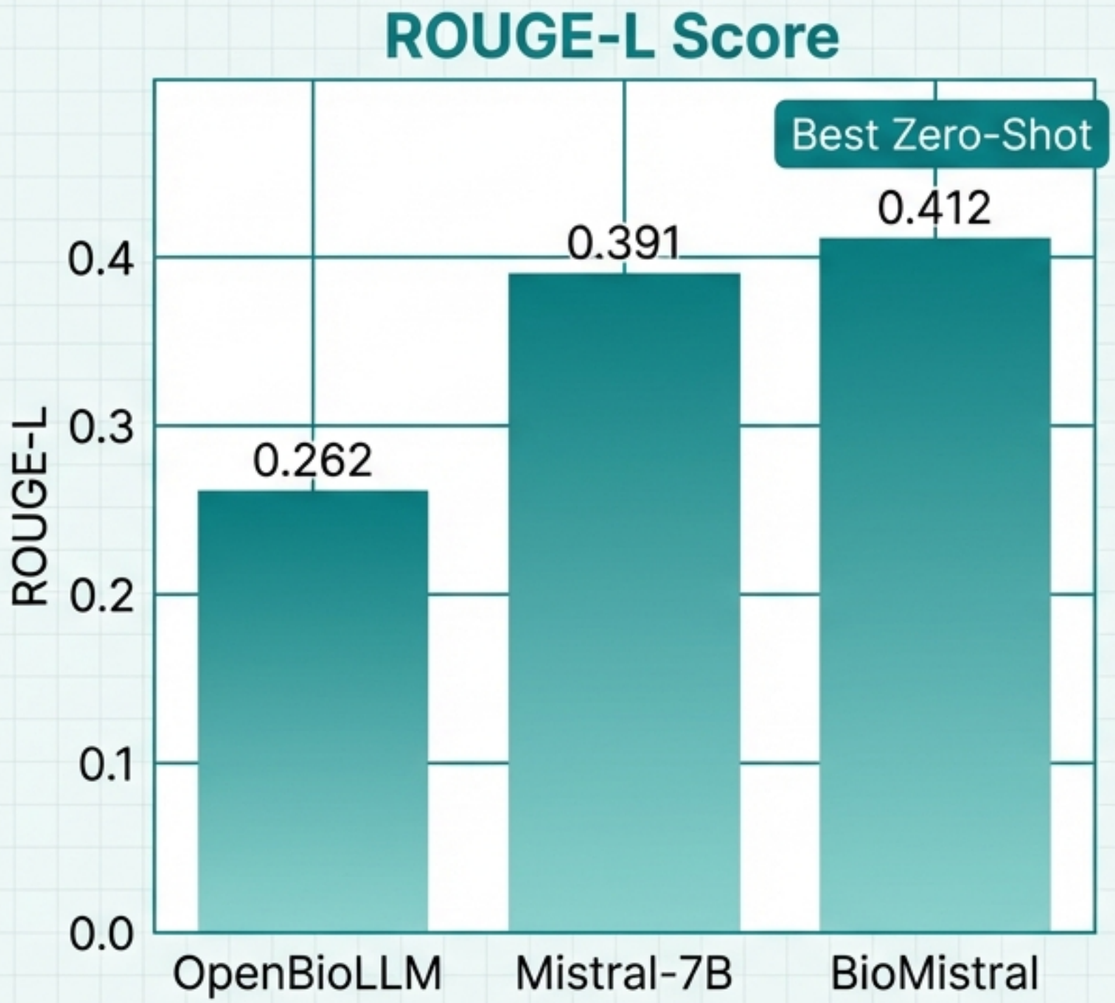
General Purpose

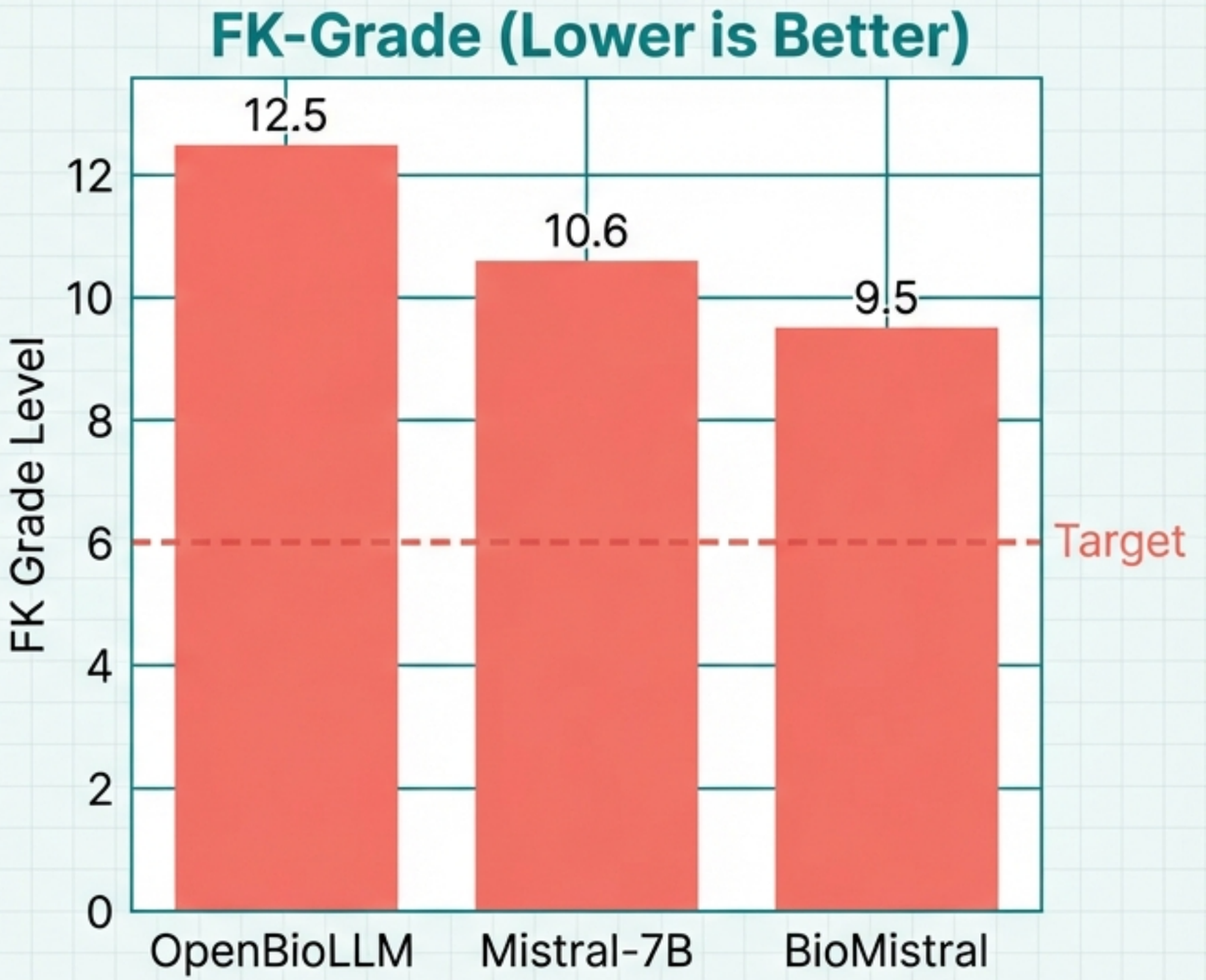
Mistral
Architecture

All models fine-tuned using the same LoRA pipeline for fair comparison.

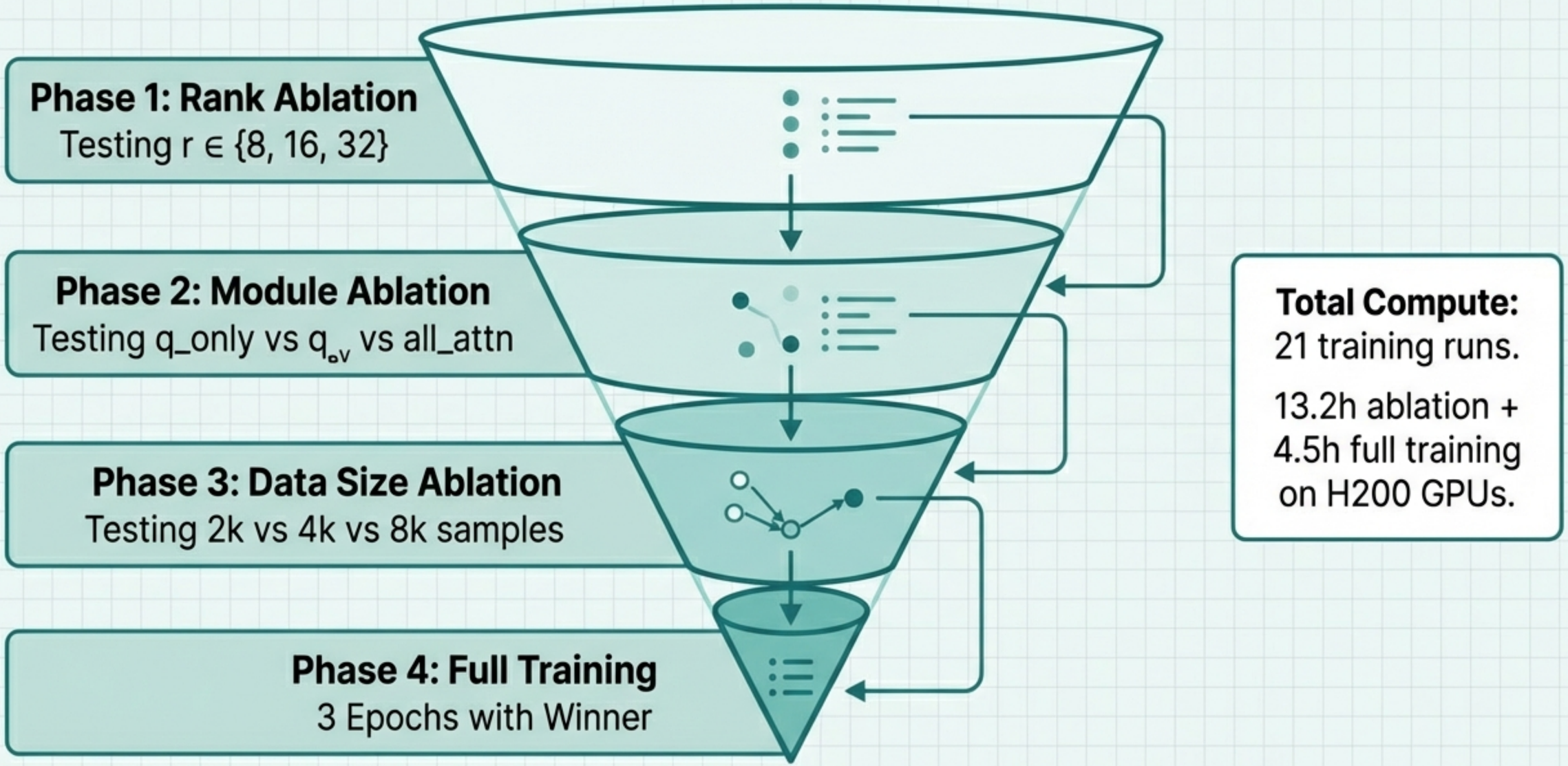
Baseline Results: Zero-Shot Performance

Medical Pretraining ≠ Better Zero-Shot Simplification





4-Phase Pipeline Architecture

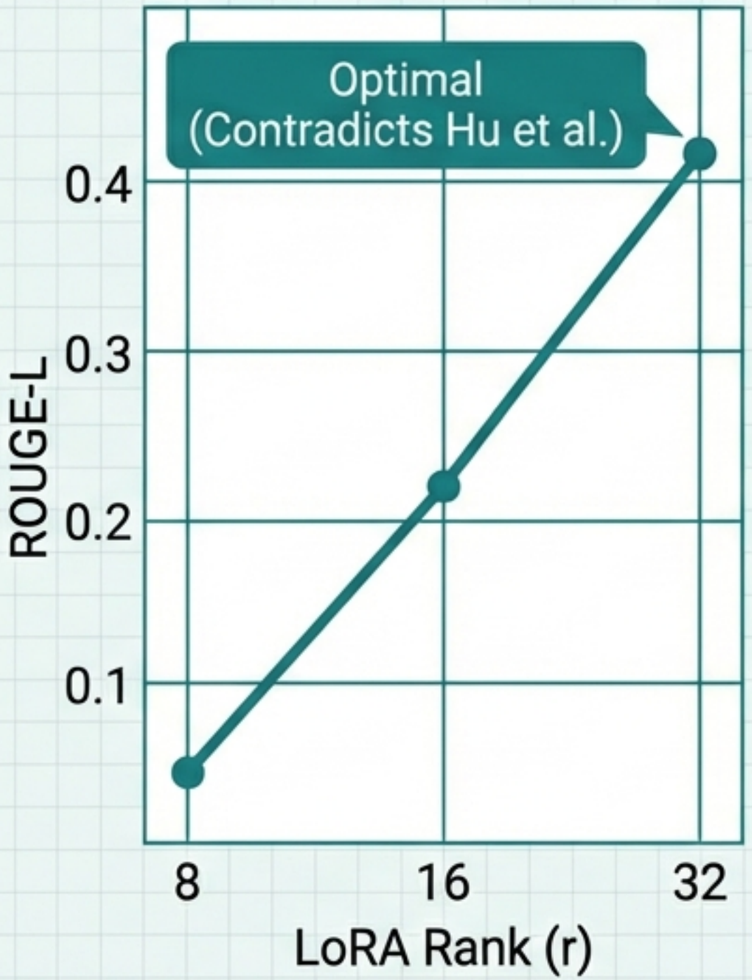


Total Compute:
21 training runs.

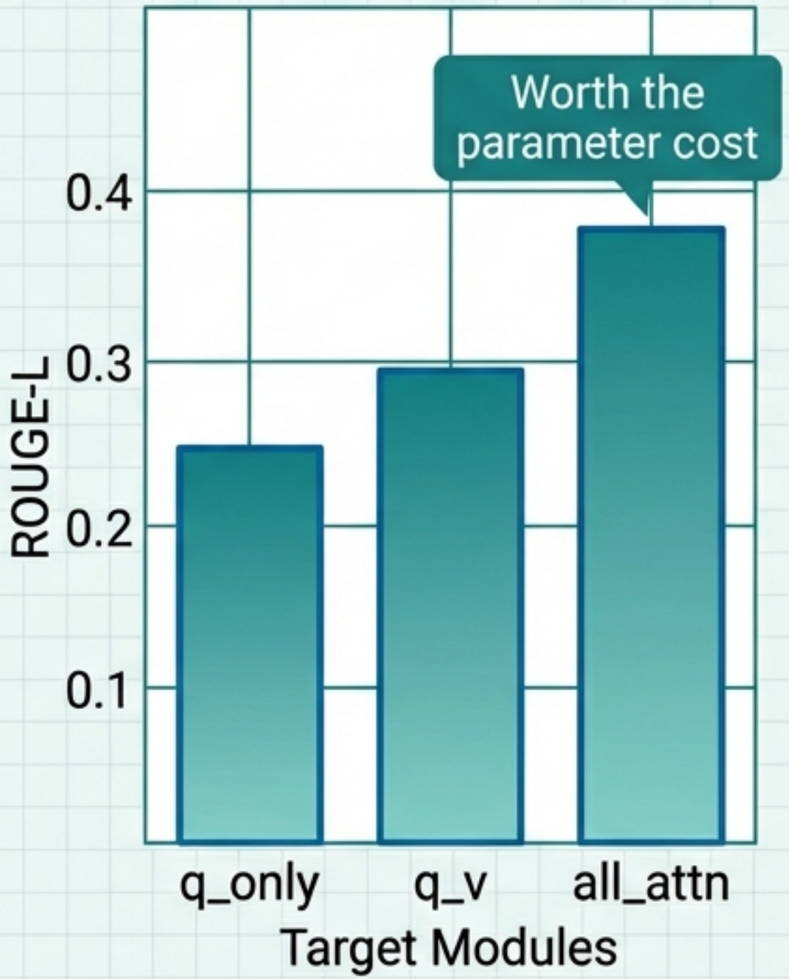
13.2h ablation + 4.5h full training on H200 GPUs.

Ablation Study Results

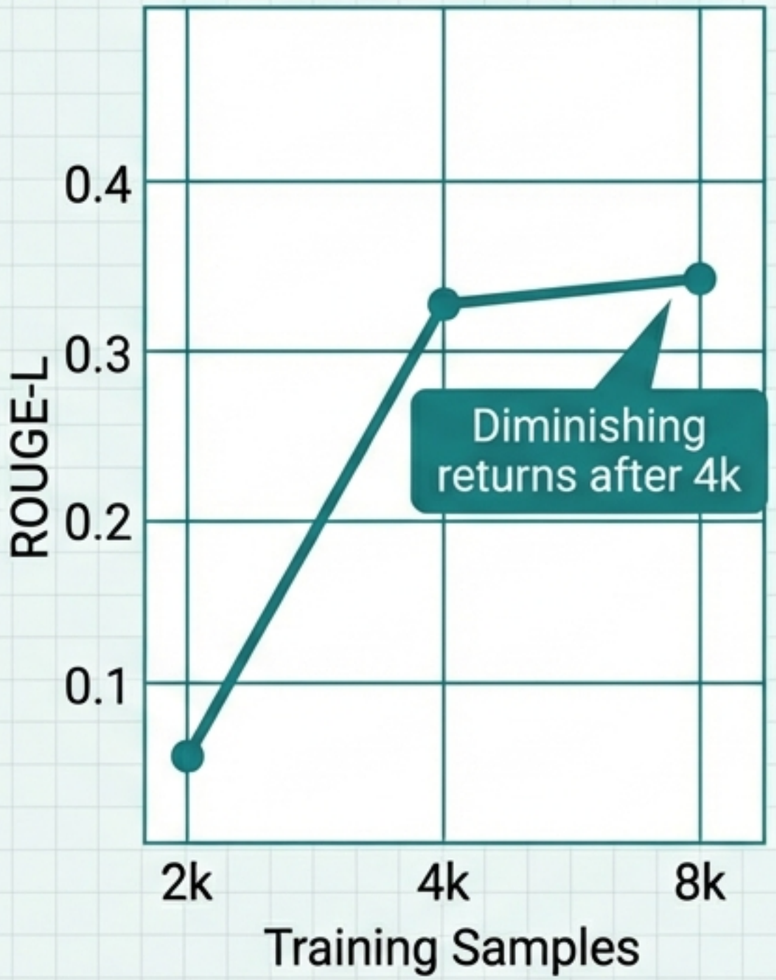
Rank Ablation



Module Ablation



Data Size Ablation

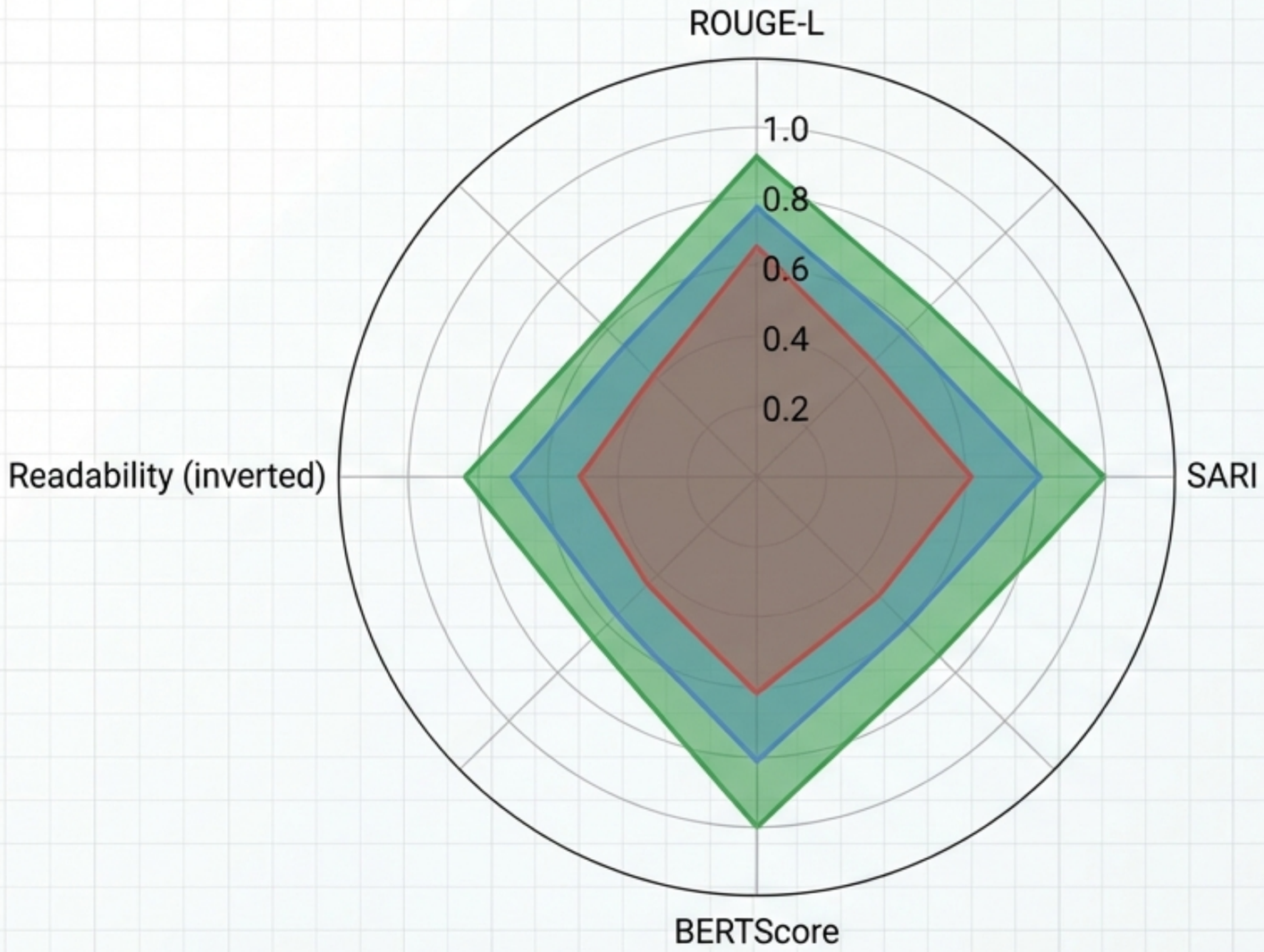


Full Training Configuration

The “Gold” Configuration

- **Rank (r)**: 32
- **Alpha (α)**: 64
- **Modules**: all_attn (q, k, v, o projections)
- **Method**: rsLoRA (Rank Stabilized LoRA - Crucial for stability)
- **Data**: 7,999 samples
- **Epochs**: 3

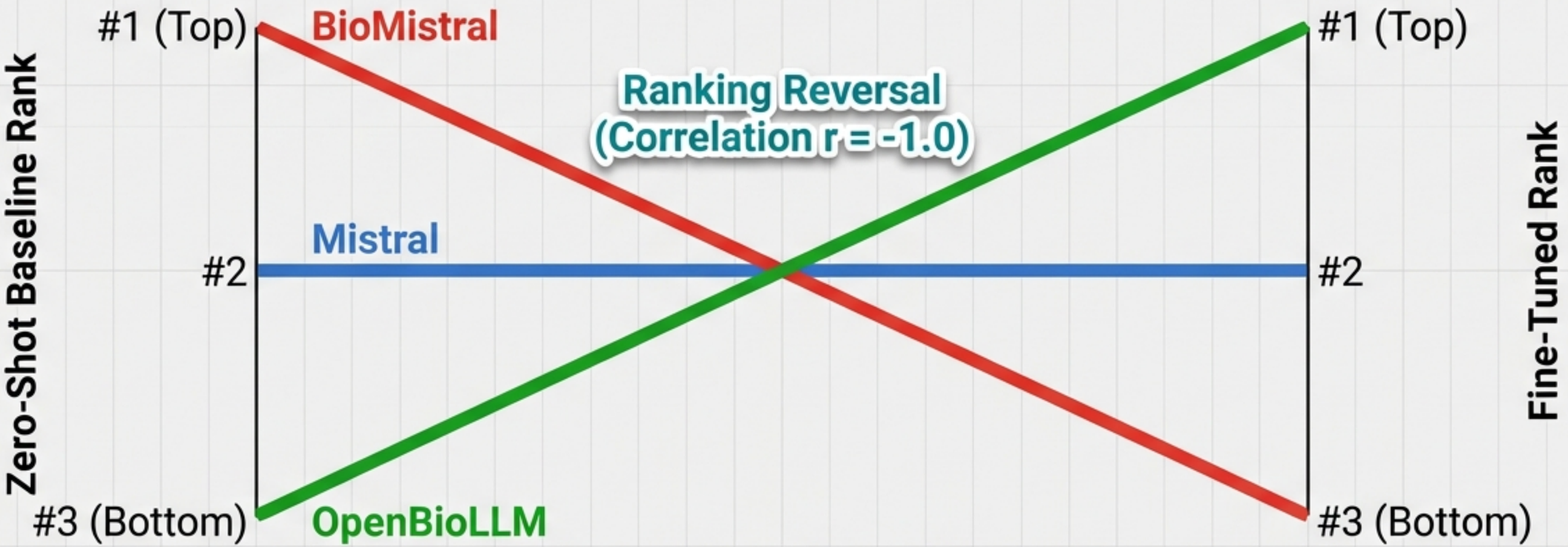
Model Performance Comparison (Post-Training)



Leaderboard (ROUGE-L)

1. **OpenBioLLM-8B**: 0.675 ROUGE-L
2. Mistral-7B: 0.649 ROUGE-L
3. BioMistral-7B: 0.632 ROUGE-L

KEY FINDING: Ranking Reversal



The ‘Underdog’ had the highest capacity to learn.

Project Results Overview

Improvement

+157% ↗

ROUGE-L increase for OpenBioLLM.

Readability

14.5 (College)

↓

7.16 (7th Grade)

50% Reduction

Consistency

> 0.94

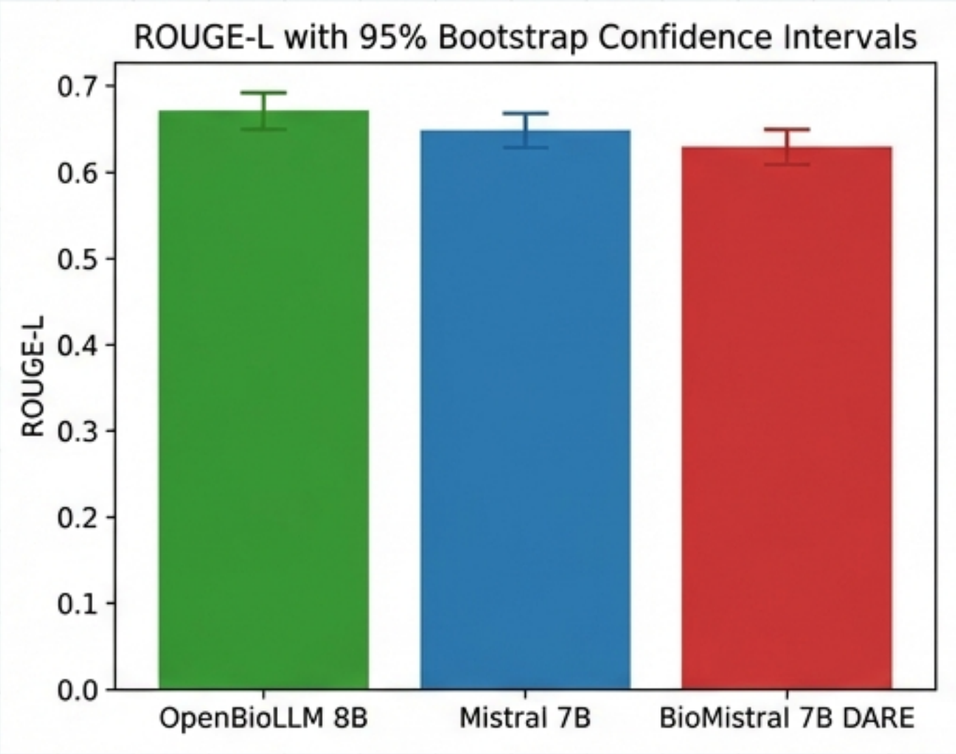
BERTScore across all models.

SARI

74.64

Exceeded target of 40.

Statistical Analysis



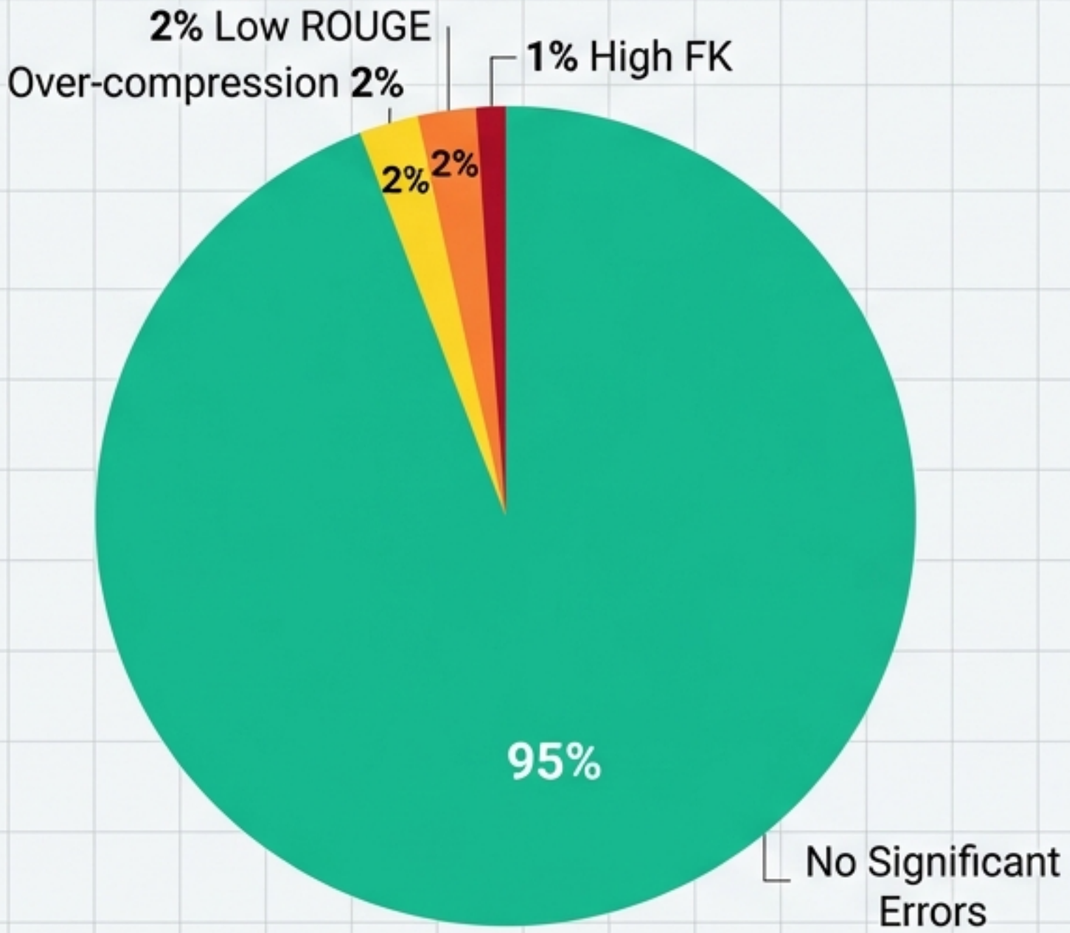
No overlap = Statistically Significant Differences ($p < 0.001$)

Effect Size (Cohen's d):

OpenBioLLM vs BioMistral: d=0.793 (Medium)
OpenBioLLM vs Mistral: d=0.475 (Small)

These effect sizes quantify the magnitude of the differences observed in the confidence interval chart.

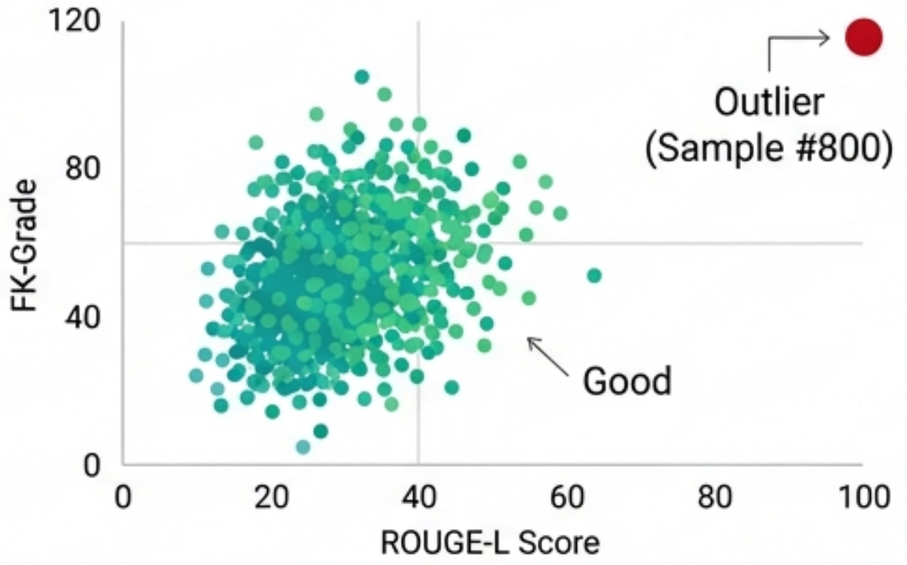
Error Analysis Dashboard



The Outlier

Sample #800: FK Grade = 69.1

Cause: Failed to break up run-on sentences. Occurred in only 0.1% of data.



Research Questions: Answered

RQ1
Medical pretraining?
No benefit for
Fine-tuning.

RQ2
BioMistral-7B-DARE
best zero-shot

RQ3
LoRA Impact?
+157%
Improvement.

RQ4
Optimal Rank?
r=32.
Contradicts Literature

RQ5
Ranking?
Reversed.

RQ6
Modules?
all_attn optimal.

RQ7
More data = better,
4K ≈ 97% of 8K

RQ8
Inverse correlation:
worst → best

RQ9
Parameter cost
justified by quality

RQ10
OpenBioLLM-8B
most consistent
(3/4 wins)

RQ11
~50% FK reduction
(14.5 → 7.0)

RQ12
rsLoRA?
Required for
stability.

Confirmed/Pesitive Finding
Reversed/Negative Finding
Contradicts Literature
Neutral/Other

Key Takeaways

LoRA Works. ~92% average improvement with only 0.38% trainable parameters.

Domain Value Unlocked. Medical pretraining wasn’t visible in Zero-Shot, but dominated after Fine-Tuning.

Readability Bridged. Successfully moved from College Level (14.5) to 7th Grade (7.1).

MediSimplifier: Generalization to Real Clinical Data (MIMIC-IV)

Validation Dataset: Real-World ICU

Source: 95 Real Discharge Summaries

Origin: MIMIC-IV (Beth Israel Deaconess Medical Center)

Status: De-identified Patient Records

10,389 **Avg. Characters per Note**

5x longer than synthetic training data

Target: Reduce reading level from **Grade 11 to Grade 6.**

Cross-Dataset Consistency (OpenBioLLM-8B + LoRA)

50.7

SARI Score (Real Data)

vs 28.4 baseline (+78% improvement maintained)

Scorecard Table

Metric	Training Data	Real MIMIC-IV	Status
FK Grade	7.16	6.52	✓ Better on real data
Base→LoRA Δ	+157% ROUGE-L	**+659%** ROUGE-L	✓ Larger improvement
Semantic	0.95 BERTScore	0.88 BERTScore	✓ High preservation

Sample Transformation

Diff

INPUT (Raw Clinical Note)

Pt w hx Klatskins tumor (T4 N1 M0) s/p L hepatic trisegmentectomy w RNY intrahepatic HJ reconstruction c/b bile leak/abscess

OUTPUT (MediSimplifier)

The patient has bile duct cancer called Klatskin tumor. He had surgery to remove part of his liver and reconnect the bile ducts. After surgery, he had a bile leak that caused an infection.

Reading Level Achieved: **Grade 6.5**

KEY TAKEAWAY: Fine-tuned models successfully generalize from synthetic training data to real-world clinical text, achieving target readability on complex ICU discharge summaries without retraining.

Final Performance Summary

Champion Card

OpenBioLLM-8B

ROUGE-L:
0.6749
(+157%)

SARI:
74.64

BERTScore:
0.95

FK-Grade:
7.16

The clear winner across all quality metrics.