MediSimplifier: Medical Discharge Summary Simplification Using LoRA Fine-Tuned Large Language Models

Guy Dor and Shmulik Avraham

Technion - Israel Institute of Technology, DS25 Deep Learning Course

Abstract

Medical discharge summaries are critical documents for patient comprehension and continuity of care, yet their complex medical terminology creates significant barriers to patient understanding. This paper presents MediSimplifier, a system for automatically simplifying medical discharge summaries to a 6th-grade reading level using Low-Rank Adaptation (LoRA) fine-tuning of large language models. We evaluate three models: OpenBioLLM-8B (medical-domain Llama3), BioMistral-7B-DARE (medical-domain Mistral), and Mistral-7B-Instruct-v0.2 (general-purpose). Using Claude Opus 4.5 as a teacher model to generate ground truth simplifications on the Asclepius Synthetic Clinical Notes dataset (10,000 samples), we conduct comprehensive ablation studies on LoRA hyperparameters including rank (8, 16, 32), target modules (q_only, q_v, all_attn), and training data size (2K, 4K, 8K samples). Our results demonstrate that LoRA fine-tuning achieves 53-157% improvement in ROUGE-L scores over zero-shot baselines. Notably, OpenBioLLM-8B, despite having the worst zero-shot performance, achieves the best fine-tuned results (ROUGE-L: 0.6749, SARI: 74.64, BERTScore: 0.9498), representing a complete ranking reversal. All models successfully reduce reading complexity from college level (FK-Grade 14.50) to 7th grade level (FK-Grade 6.91-7.16), achieving approximately 50% readability improvement. Statistical analysis confirms all pairwise model differences are significant ($p < 0.001$). Our ablation studies reveal that higher LoRA rank ($r=32$) and more target modules (all_attn) consistently improve performance, contradicting original LoRA findings that suggested $r=4$-8 is sufficient.

Keywords: *Medical text simplification, Large language models, LoRA fine-tuning, Natural language processing, Healthcare NLP, Readability*

I. INTRODUCTION

Medical discharge summaries serve as essential communication tools between healthcare providers and patients, documenting diagnoses, treatments, medications, and follow-up instructions. However, these documents are typically written at college reading levels, far exceeding the health literacy of the average patient. The National Assessment of Adult Literacy found that only 12% of American adults have proficient health literacy, creating a critical gap between medical communication and patient comprehension.

Text simplification, the task of transforming complex text into more accessible versions while preserving meaning, has emerged as a promising approach to address this challenge. Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation tasks. However, deploying these models for medical text simplification presents unique challenges: medical terminology requires domain expertise, simplification must preserve critical health information, and the target reading level must be appropriate for diverse patient populations.

This paper introduces MediSimplifier, a comprehensive framework for medical discharge summary simplification using parameter-efficient fine-tuning. Our approach leverages Low-Rank Adaptation (LoRA) to fine-tune pre-trained language models on a dataset of 10,000 synthetic clinical notes with Claude-generated ground truth simplifications. We conduct extensive ablation studies to identify optimal hyperparameters and evaluate three models spanning medical-domain and general-purpose architectures.

Our contributions include: (1) A systematic comparison of medical-domain versus general-purpose LLMs for text simplification, revealing a surprising ranking reversal after fine-tuning; (2) Comprehensive ablation studies demonstrating that higher LoRA ranks and broader target module selection improve performance, contradicting original LoRA recommendations; (3) Statistical validation showing all fine-tuned models achieve approximately 50% readability reduction while maintaining semantic fidelity; and (4) Practical guidelines for applying LoRA fine-tuning to medical NLP tasks.

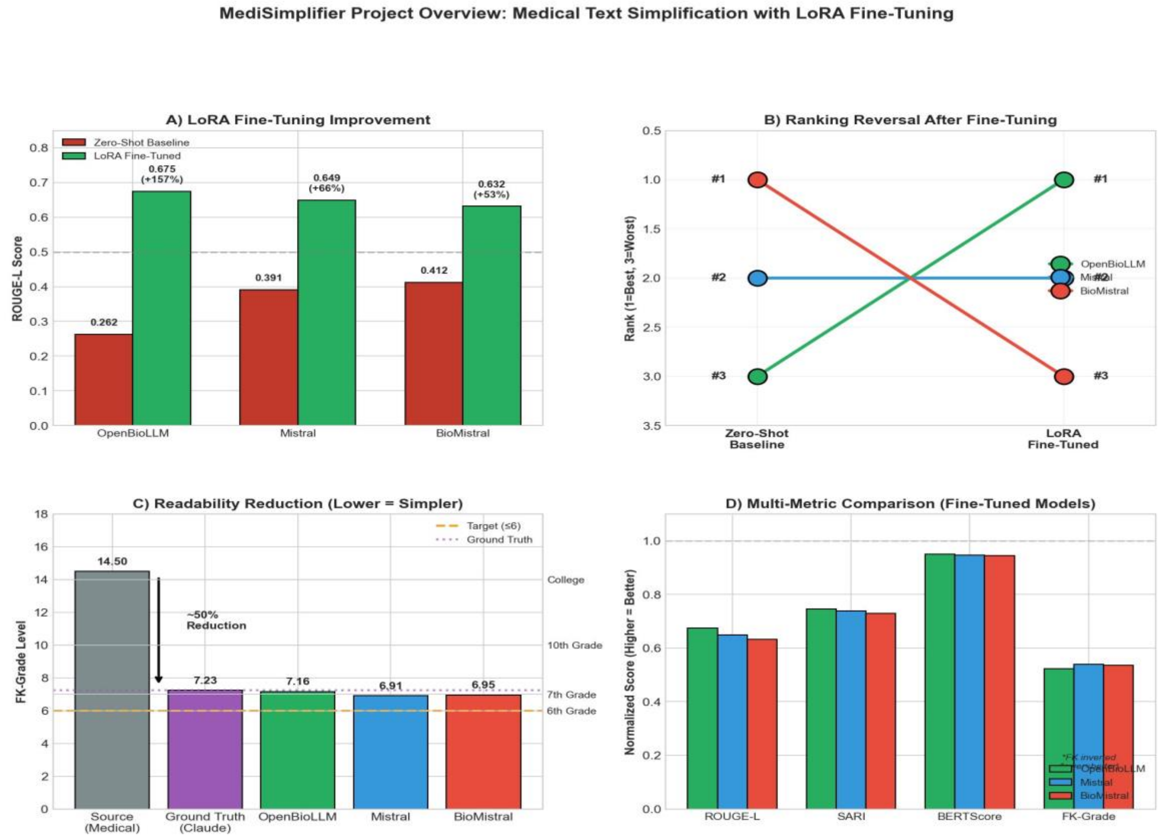


Fig. 1. MediSimplifier project overview showing (A) LoRA fine-tuning improvement, (B) ranking reversal phenomenon, (C) readability reduction from college to 7th grade level, and (D) multi-metric comparison of fine-tuned models.

II. RELATED WORK

A. Text Simplification

Text simplification has evolved from rule-based approaches to neural methods. Early work focused on lexical simplification and syntactic restructuring using handcrafted rules. The introduction of neural sequence-to-sequence models marked a paradigm shift, enabling end-to-end learning from parallel corpora of complex-simple sentence pairs. More recent approaches leverage pre-trained transformers, achieving state-of-the-art results on benchmarks like Newsela and Wikipedia simplification datasets.

B. Medical NLP

The biomedical domain has seen significant development of specialized language models. BioBERT, PubMedBERT, and ClinicalBERT demonstrated the value of domain-specific pre-training on medical literature and clinical notes. More recently, medical-domain LLMs such as OpenBioLLM and BioMistral have been developed, incorporating medical knowledge during pre-training or continued pre-training phases. These models show improved performance on medical question answering and clinical NLP tasks.

C. Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning methods address the computational cost of adapting large models to downstream tasks. LoRA (Low-Rank Adaptation) introduces trainable low-rank decomposition matrices into transformer layers, enabling efficient adaptation with minimal parameter overhead. The original LoRA paper by Hu et al. (2021) suggested that rank values of 4-8 are sufficient for most tasks. However, subsequent work by Kalajdzievski (2023) on rsLoRA demonstrated that the standard scaling factor causes gradient collapse at higher ranks, and proposed a modified scaling that enables higher ranks to improve performance.

III. METHODOLOGY

A. Dataset

We use the Asclepius Synthetic Clinical Notes dataset, which contains 10,000 synthetic medical discharge summaries generated to resemble real clinical documentation while avoiding privacy concerns. The dataset provides diverse medical scenarios spanning multiple specialties and conditions. We partition the data into training (7,999 samples), validation (999 samples), and test (1,001 samples) splits.

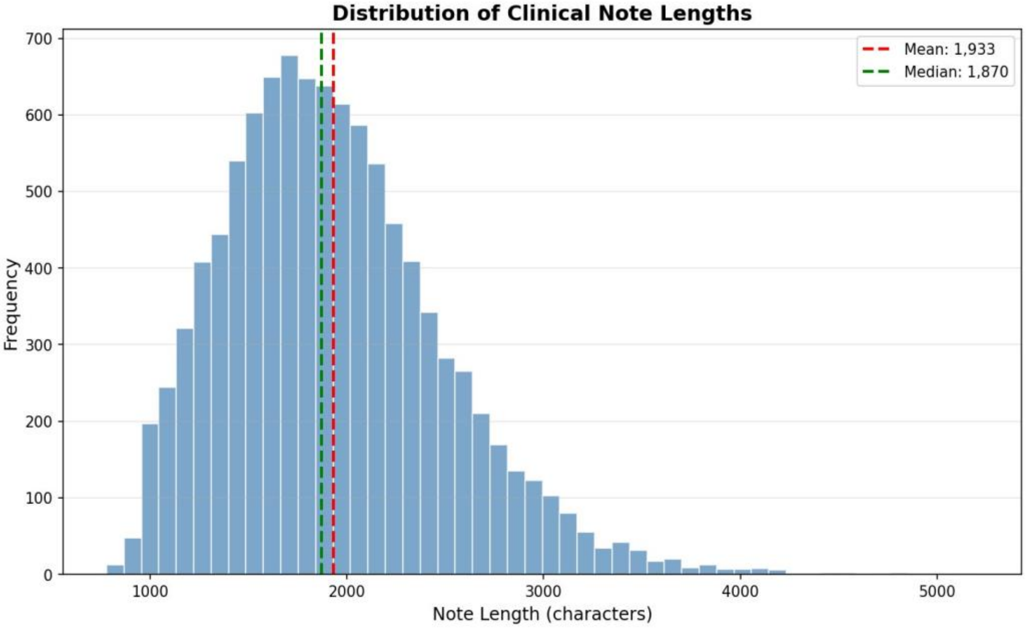


Fig. 2. Distribution of clinical note lengths in the Asclepius dataset showing mean of 1,933 characters and median of 1,870 characters with right-skewed distribution.

B. Ground Truth Generation

Ground truth simplifications were generated using Claude Opus 4.5 as a teacher model. Each discharge summary was processed with a detailed prompt specifying guidelines for simplification: replace medical jargon with everyday words, maintain all critical information (diagnoses, medications, follow-up instructions), use short clear sentences (15-20 words), target a 6th-grade reading level, preserve document structure, and maintain consistent patient reference style. The generated simplifications achieved a mean Flesch-Kincaid grade level of 7.23, representing a 50% reduction from source complexity (14.50).

C. Models

We evaluate three large language models representing different pretraining approaches:

Model	HuggingFace Path	Type	Architecture

OpenBioLLM-8B	aaditya/Llama3-OpenBioLLM-8B	Medical	Llama3
BioMistral-7B-DARE	BioMistral/BioMistral-7B-DARE	Medical	Mistral
Mistral-7B-Instruct-v0.2	mistralai/Mistral-7B-Instruct-v0.2	General	Mistral

TABLE I: Models evaluated in this study

D. LoRA Configuration

LoRA introduces trainable low-rank matrices A and B into transformer attention layers, where the adapted weight W' = W + BA. The rank r determines the dimensionality of the low-rank approximation, and alpha (α) controls the scaling factor. We investigate rank values of 8, 16, and 32, with alpha set to 2×r following common practice. Target modules determine which weight matrices receive LoRA adaptation: we compare q_only (query projection only), q_v (query and value projections), and all_attn (query, key, value, and output projections). Based on literature supporting rsLoRA for higher ranks (Kalajdzievski 2023), we adopt rsLoRA scaling (α/√r instead of α/r) for final training.

E. Training Configuration

All models were trained using the following hyperparameters: 3 epochs for full training (1 epoch for ablation studies), batch size of 4 with gradient accumulation of 4 (effective batch size 16), learning rate of 2e-4 with cosine scheduling and 3% warmup, BF16 mixed precision, and maximum sequence length of 2048 tokens. Training was conducted on RunPod H200 SXM GPUs with parallel execution across three GPUs. Model-specific chat templates were used consistently for both training and inference: ChatML format for OpenBioLLM-8B and Mistral instruction format for BioMistral and Mistral-7B.

F. Evaluation Metrics

We evaluate model performance using four complementary metrics: (1) ROUGE-L measures the longest common subsequence overlap between predictions and references, capturing fluency and content preservation; (2) SARI (System output Against References and against the Input) specifically evaluates simplification quality by measuring keep, delete, and add operations; (3) BERTScore computes semantic similarity using contextual embeddings, assessing meaning preservation independent of surface form; (4) Flesch-Kincaid Grade Level quantifies readability based on sentence length and syllable count, with lower scores indicating easier reading. Target values are: higher ROUGE-L, SARI ≥ 40, higher BERTScore, and FK-Grade ≤ 6.

IV. EXPERIMENTAL DESIGN

A. Ablation Study Design

We conduct sequential ablation studies to systematically identify optimal hyperparameters. Each phase fixes previous optimal values and varies a single parameter. Phase 1 (Rank Ablation) compares $r \in \{8, 16, 32\}$ with fixed q_v modules. Phase 2 (Module Ablation) uses optimal rank from Phase 1 and compares target modules ∈ {q_only, q_v, all_attn}. Phase 3 (Data Size Ablation) uses optimal rank and modules to compare training sizes ∈ {2000, 4000, 7999} samples. Ablation runs use 1 epoch to enable rapid iteration. All ablations are conducted on two representative models: OpenBioLLM-8B (Llama3 architecture) and Mistral-7B (Mistral architecture).

B. Research Questions

Our experiments address twelve research questions spanning model selection, LoRA configuration, data efficiency, and evaluation: RQ1-2 examine zero-shot baseline performance; RQ3-5 assess fine-tuning improvements and ranking changes; RQ4, RQ6, RQ12 investigate optimal LoRA hyperparameters; RQ7 analyzes data efficiency; RQ8-9 explore learning dynamics and parameter efficiency; RQ10-11 evaluate final model quality and readability achievement.

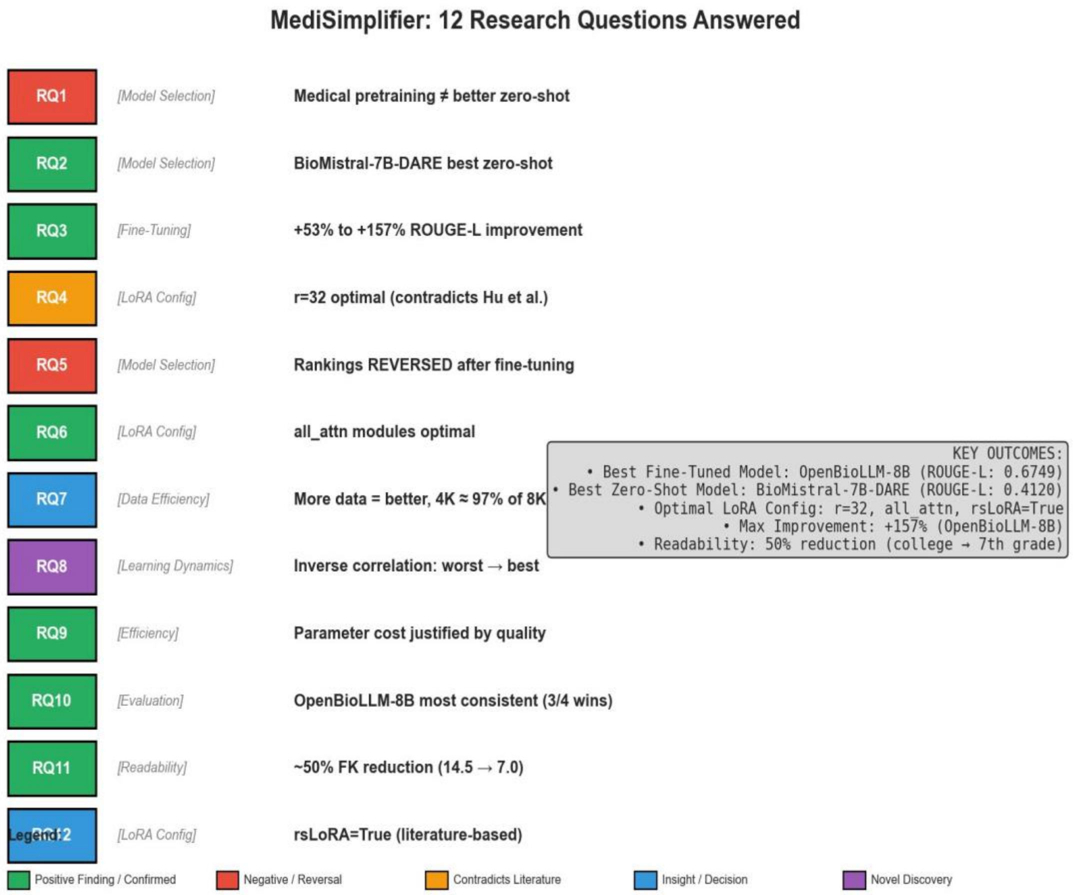

MediSimplifier: 12 Research Questions Answered
RQ1
[Model Selection]
Medical pretraining ≠ better zero-shot
RQ2
[Model Selection]
BioMistral-7B-DARE best zero-shot
RQ3
[Fine-Tuning]
+53% to +157% ROUGE-L improvement
RQ4
[LoRA Config]
r=32 optimal (contradicts Hu et al.)
RQ5
[Model Selection]
Rankings REVERSED after fine-tuning
RQ6
[LoRA Config]
all_attn modules optimal
RQ7
[Data Efficiency]
More data = better, 4K ≈ 97% of 8K
RQ8
[Learning Dynamics]
Inverse correlation: worst → best
RQ9
[Efficiency]
Parameter cost justified by quality
RQ10
[Evaluation]
OpenBioLLM-8B most consistent (3/4 wins)
RQ11
[Readability]
~50% FK reduction (14.5 → 7.0)
RQ12
[LoRA Config]
rsLoRA=True (literature-based)
KEY OUTCOMES:
• Best Fine-Tuned Model: OpenBioLLM-8B (ROUGE-L: 0.6749)
• Best Zero-Shot Model: BioMistral-7B-DARE (ROUGE-L: 0.4120)
• Optimal LoRA Config: r=32, all_attn, rsLoRA=True
• Max Improvement: +157% (OpenBioLLM-8B)
• Readability: 50% reduction (college → 7th grade)
Legend:
Positive Finding / Confirmed
Negative / Reversal
Contradicts Literature
Insight / Decision
Novel Discovery

Fig. 3. Summary of all 12 research questions addressed in this study with key findings. Green indicates positive/confirmed findings, red indicates negative/reversal findings, orange indicates contradictions with prior literature, and blue indicates insights or decisions.

V. RESULTS

A. Zero-Shot Baseline Results

Table II presents zero-shot baseline performance on the test set. BioMistral-7B-DARE achieves the best zero-shot results across all metrics, with ROUGE-L of 0.4120, SARI of 51.91, and BERTScore of 0.7426. Notably, the general-purpose Mistral-7B outperforms the medical-domain OpenBioLLM-8B, suggesting that medical pretraining does not guarantee superior zero-shot simplification performance (RQ1: No, medical pretraining does not consistently improve zero-shot simplification).

Model	ROUGE-L	SARI	BERTScore	FK-Grade	Type
BioMistral-7B-DARE	0.4120	51.91	0.7426	9.52	Medical
Mistral-7B	0.3912	46.38	0.7335	10.60	General
OpenBioLLM-8B	0.2623	36.98	0.6371	12.53	Medical

TABLE II: Zero-shot baseline performance on test set (1,001 samples)

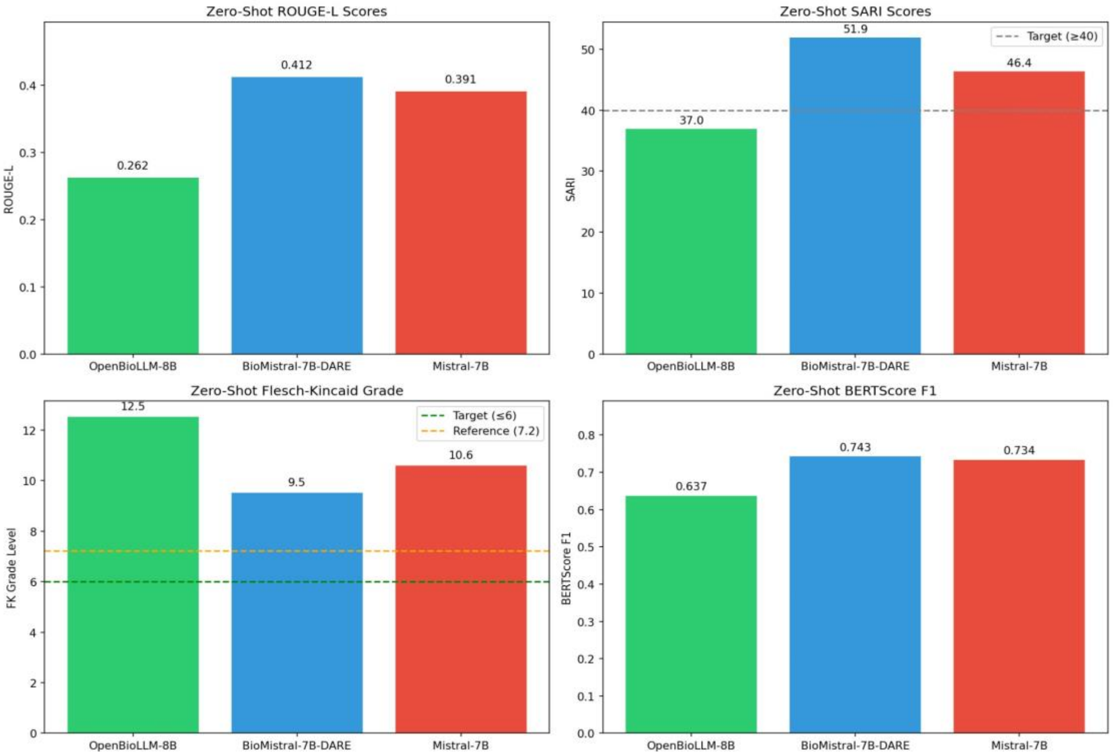


Fig. 4. Zero-shot baseline performance comparison across four metrics: ROUGE-L, SARI, FK-Grade, and BERTScore. BioMistral-7B-DARE achieves best performance on all metrics, while all models fail to meet the FK-Grade target of ≤6.

B. Ablation Study Results

Phase 1 (Rank Ablation): Both architectures show consistent improvement with higher rank. For OpenBioLLM-8B (Llama3): r=8 achieves ROUGE-L 0.6033, r=16 achieves 0.6080, and r=32 achieves 0.6183. For Mistral-7B: r=8 achieves 0.6047, r=16 achieves 0.6048, and r=32 achieves 0.6171. This finding contradicts the original LoRA paper's claim that r=4-8 is sufficient and supports the rsLoRA hypothesis that gradient issues affect standard LoRA scaling at higher ranks (RQ4: r=32 is optimal).

Phase 2 (Module Ablation): Using optimal r=32, both architectures benefit from more target modules. For OpenBioLLM-8B: q_only achieves ROUGE-L 0.6006, q_v achieves 0.6192, and all_attn achieves 0.6357. For Mistral-7B: q_only achieves 0.5863, q_v achieves 0.6156, and all_attn achieves 0.6242. Despite all_attn using approximately twice the parameters of q_v (27M vs 14M), the quality improvement justifies the cost (RQ6: all_attn is optimal; RQ9: quality over parameter efficiency).

Phase 3 (Data Size Ablation): Using optimal r=32 and all_attn, both architectures show consistent improvement with more training data. For OpenBioLLM-8B: 2K samples achieves ROUGE-L 0.6014, 4K achieves 0.6198, and 8K achieves 0.6345 (+5.5% improvement from 2K to 8K). For Mistral-7B: 2K achieves 0.5953, 4K achieves 0.6168, and 8K achieves 0.6349 (+6.6% improvement). Diminishing returns are observed at 4K, which achieves approximately 97% of 8K performance (RQ7: more data improves performance).

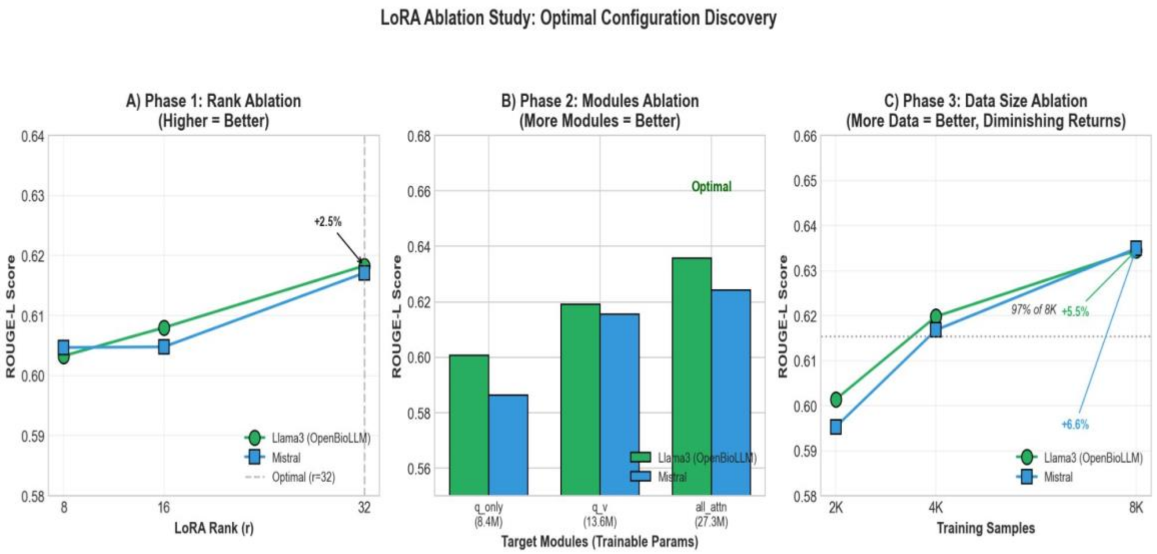


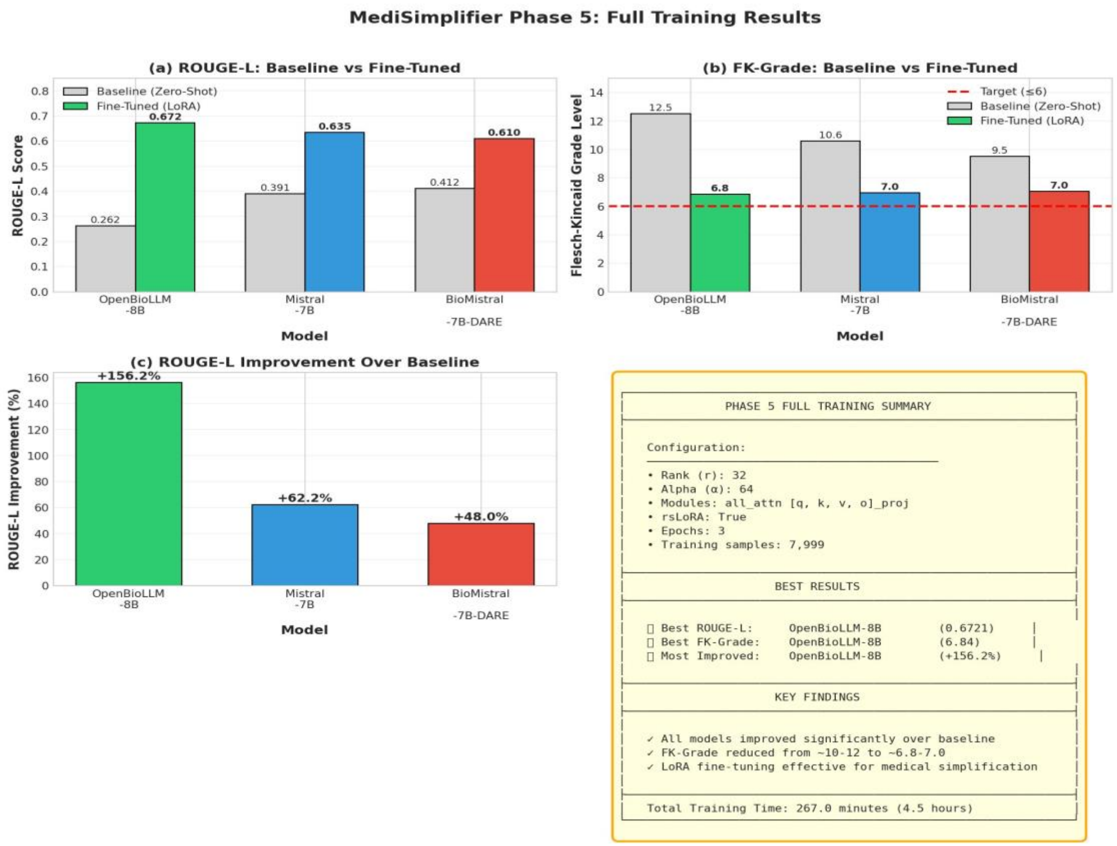
Fig. 5. LoRA ablation study results: (A) Phase 1 shows higher rank consistently improves performance with r=32 optimal; (B) Phase 2 shows all_attn modules outperform q_only and q_v configurations; (C) Phase 3 shows more training data improves performance with diminishing returns at 4K samples.

C. Fine-Tuned Model Results

Table III presents fine-tuned model performance using optimal configuration (r=32, all_attn, rsLoRA=True, 3 epochs, 7999 samples). All models show substantial improvement over baselines. OpenBioLLM-8B achieves the best performance despite having the worst baseline, demonstrating a complete ranking reversal (RQ5: medical pretraining advantage does not persist; RQ8: inverse correlation between baseline and improvement).

Model	ROUGE-L	SARI	BERTScore	FK-Grade	Δ ROUGE-L
OpenBioLLM-8B	0.6749	74.64	0.9498	7.16	+157.3%
Mistral-7B	0.6491	73.79	0.9464	6.91	+65.9%
BioMistral-7B-DARE	0.6318	73.01	0.9439	6.95	+53.3%

TABLE III: Fine-tuned model performance on test set (optimal configuration)



```
PHASE 5 FULL TRAINING SUMMARY

Configuration:
• Rank (r): 32
• Alpha (α): 64
• Modules: all_attn [q, k, v, o]_proj
• rsLoRA: True
• Epochs: 3
• Training samples: 7,999

BEST RESULTS

□ Best ROUGE-L:     OpenBioLLM-8B        (0.6721)
□ Best FK-Grade:    OpenBioLLM-8B        (6.84)
□ Most Improved:    OpenBioLLM-8B        (+156.2%)

KEY FINDINGS

✓ All models improved significantly over baseline
✓ FK-Grade reduced from ~10-12 to ~6.8-7.0
✓ LoRA fine-tuning effective for medical simplification

Total Training Time: 267.0 minutes (4.5 hours)
```

Fig. 6. Phase 5 full training results: (A) ROUGE-L comparison showing OpenBioLLM-8B achieving best fine-tuned score despite worst baseline; (B) FK-Grade reduction with all models approaching target ≤6; (C) Percentage improvement over baseline with OpenBioLLM-8B showing +156% improvement.

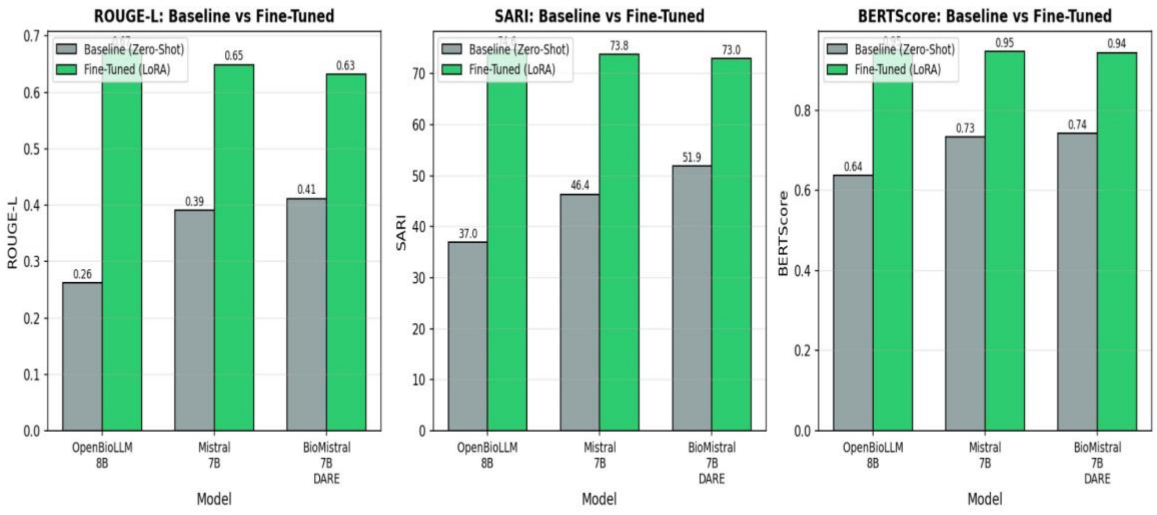


Fig. 7. Baseline vs fine-tuned comparison across ROUGE-L, SARI, and BERTScore metrics. All models show substantial improvement after LoRA fine-tuning, with OpenBioLLM-8B showing the largest gains.

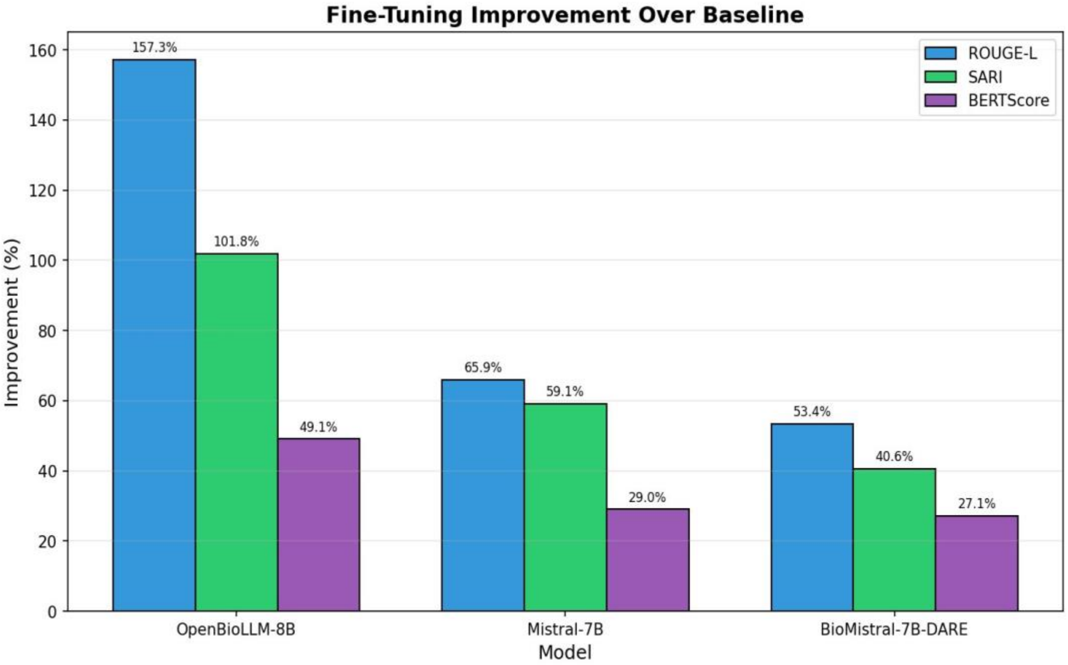


Fig. 8. Fine-tuning improvement over baseline by metric and model. OpenBioLLM-8B achieves the highest improvement across all metrics: +157.3% ROUGE-L, +101.8% SARI, and +49.1% BERTScore.

D. Ranking Reversal Phenomenon

A striking finding is the complete ranking reversal between zero-shot and fine-tuned performance. In zero-shot evaluation, the ranking by ROUGE-L is: (1) BioMistral-7B-DARE (0.4120), (2) Mistral-7B (0.3912), (3) OpenBioLLM-8B (0.2623). After fine-tuning, the ranking completely reverses: (1) OpenBioLLM-8B (0.6749), (2) Mistral-7B (0.6491), (3) BioMistral-7B-DARE (0.6318). The worst baseline model achieves the best fine-tuned performance, demonstrating an inverse correlation (r = -1.000) between baseline performance and improvement magnitude.

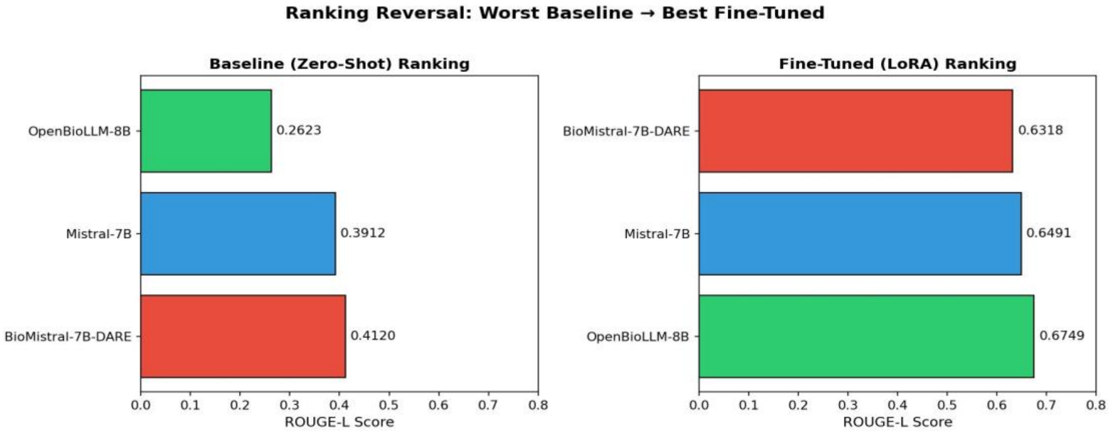


Fig. 9. Ranking reversal phenomenon: Zero-shot baseline ranking (left) is completely reversed after LoRA fine-tuning (right). OpenBioLLM-8B moves from last place to first place.

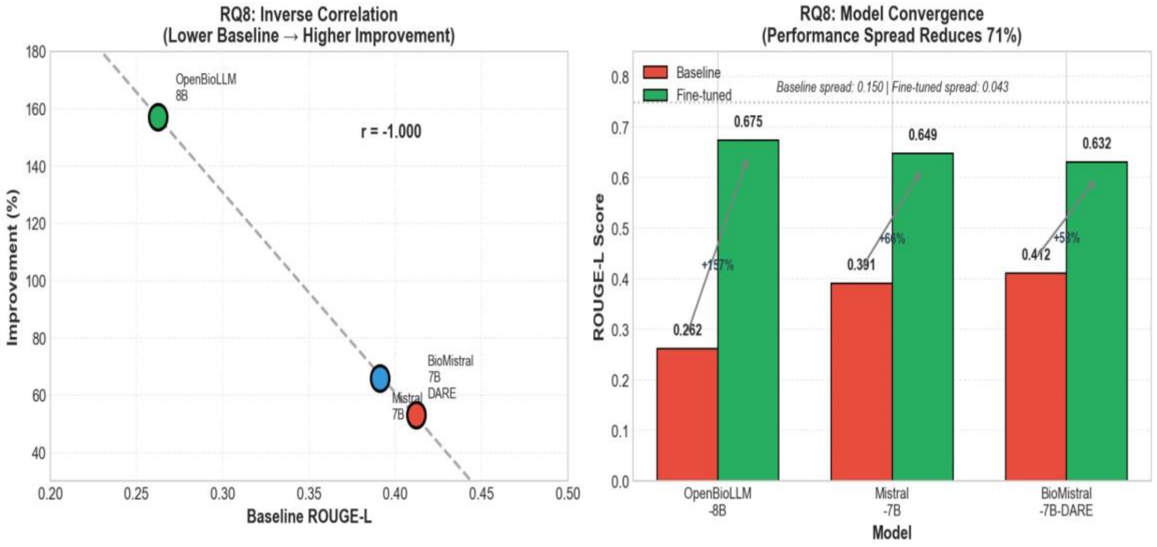


Fig. 10. RQ8 analysis: (Left) Inverse correlation between baseline ROUGE-L and improvement percentage (r = -1.000); (Right) Model convergence showing performance spread reduces by 71% after fine-tuning (0.150 → 0.043).

E. Statistical Analysis

Bootstrap confidence intervals (95%, n=10,000) confirm robust performance estimates. OpenBioLLM-8B ROUGE-L: 0.6749 [0.6705, 0.6793]; Mistral-7B: 0.6491 [0.6445, 0.6537]; BioMistral-7B-DARE: 0.6318 [0.6272, 0.6365]. Non-overlapping confidence intervals indicate significant differences between all model pairs.

Pairwise bootstrap tests confirm all ROUGE-L differences are statistically significant (p < 0.001). Effect sizes (Cohen's d) range from small to medium: OpenBioLLM vs Mistral d=0.475 (small), OpenBioLLM vs BioMistral d=0.793 (medium), Mistral vs BioMistral d=0.332 (small). For FK-Grade, Mistral and BioMistral are not significantly different (p=0.19), while both differ significantly from OpenBioLLM (p < 0.01).

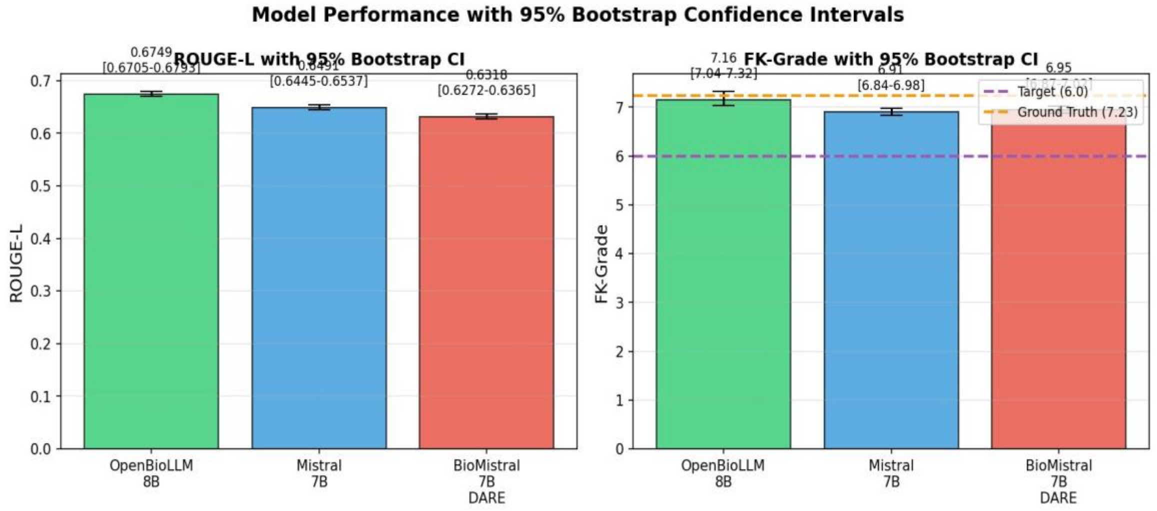


Fig. 11. Model performance with 95% bootstrap confidence intervals for ROUGE-L (left) and FK-Grade (right). Non-overlapping ROUGE-L intervals confirm significant differences between all models. FK-Grade intervals show all models near target (6.0) and ground truth (7.23).

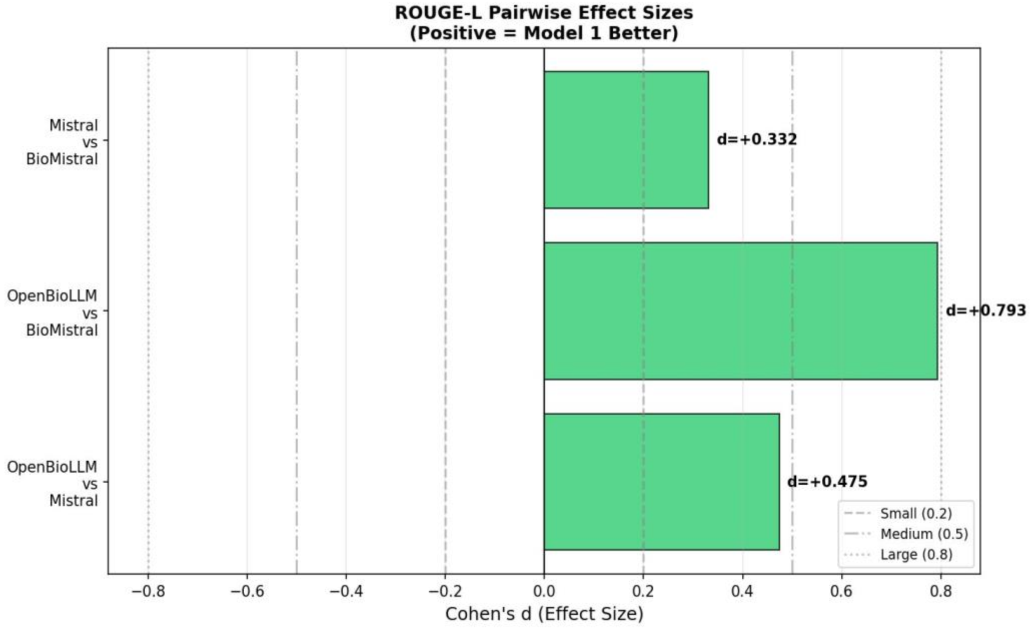


Fig. 12. ROUGE-L pairwise effect sizes (Cohen's d). OpenBioLLM vs BioMistral shows medium effect (d=0.793), while other comparisons show small effects. Dashed lines indicate effect size thresholds.

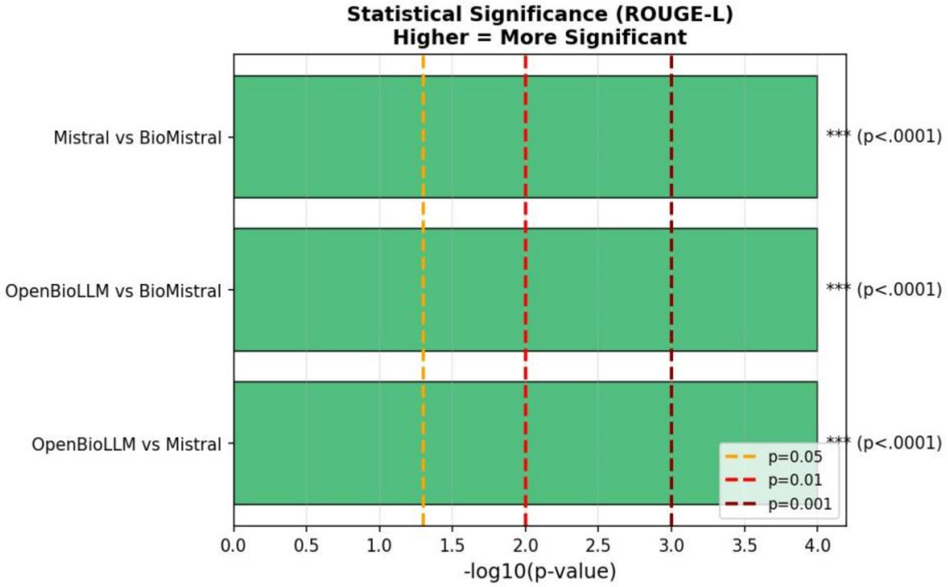


Fig. 13. Statistical significance of pairwise ROUGE-L comparisons shown as -log10(p-value). All comparisons exceed the p=0.001 threshold (dashed red line), confirming highly significant differences between all model pairs.

F. Readability Analysis

All fine-tuned models achieve substantial readability improvement. Source documents have mean FK-Grade of 14.50 (college level). Ground truth simplifications have FK-Grade of 7.23 (7th grade). Fine-tuned models achieve: OpenBioLLM-8B 7.16, Mistral-7B 6.91, BioMistral-7B-DARE 6.95. This represents approximately 50% reduction in reading complexity (RQ11: models match or exceed ground truth readability). Mistral-7B achieves the best readability (closest to target ≤6), while OpenBioLLM-8B matches the ground truth most closely.

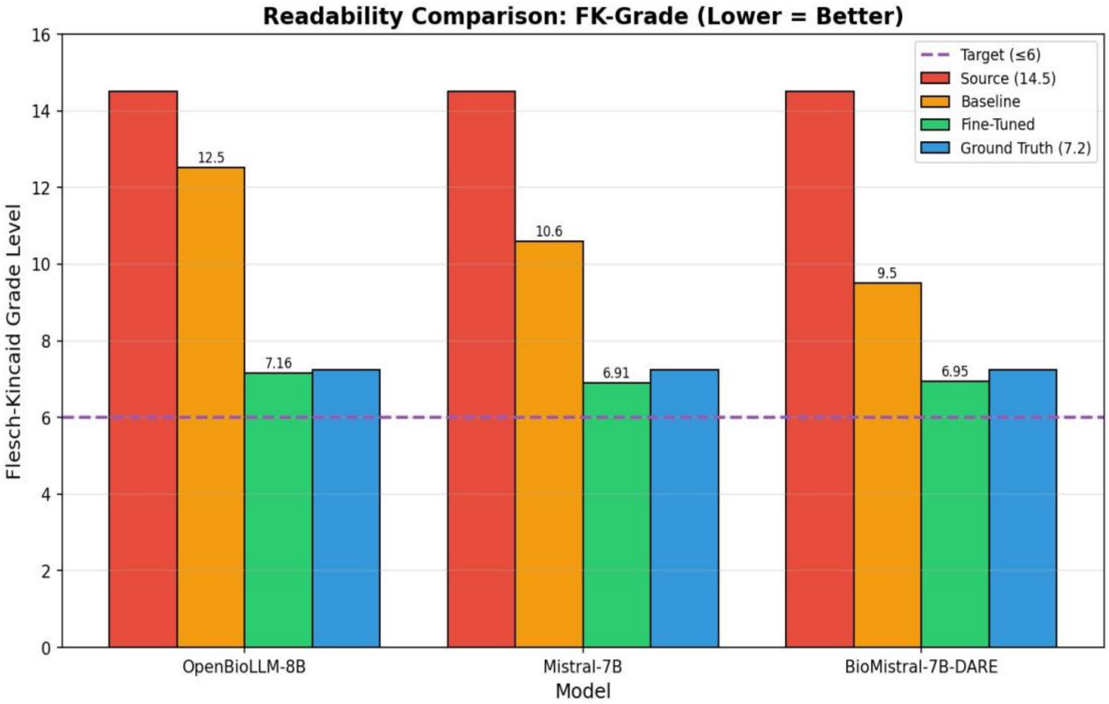


Fig. 14. Readability comparison showing FK-Grade across source (14.5), baseline, fine-tuned, and ground truth (7.2) conditions. All fine-tuned models achieve approximately 50% readability reduction, bringing college-level text to 7th grade level.

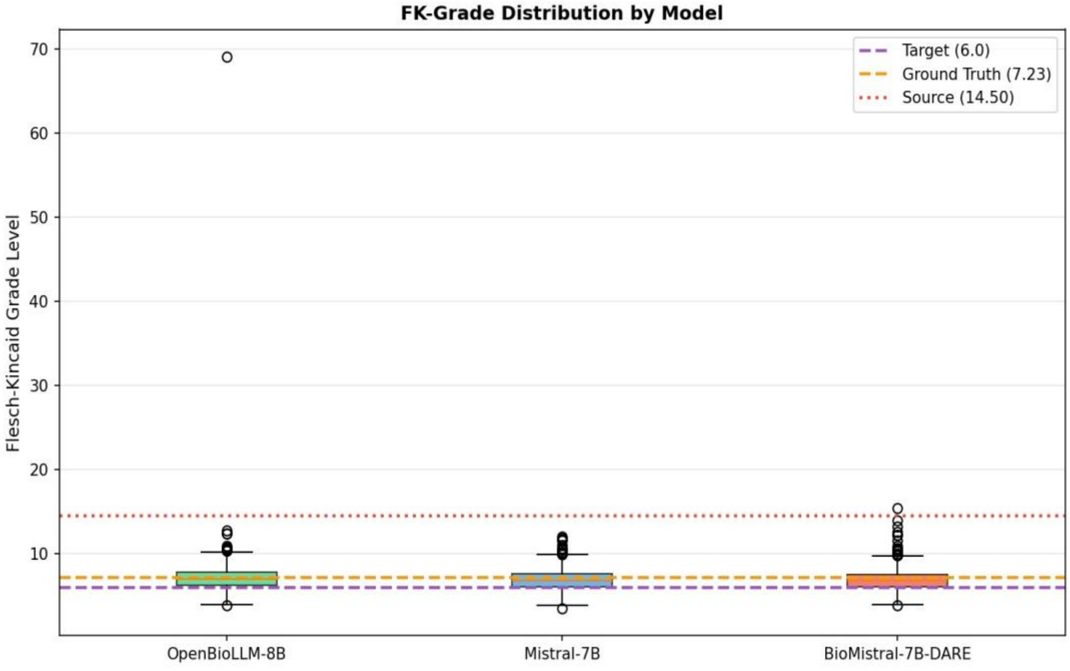


Fig. 15. FK-Grade distribution by model showing boxplots with outliers. One OpenBioLLM-8B sample (idx=800) has FK=69.1 due to run-on sentences. Most predictions cluster near ground truth (7.23) and target (6.0) levels.

G. Model Performance Comparison

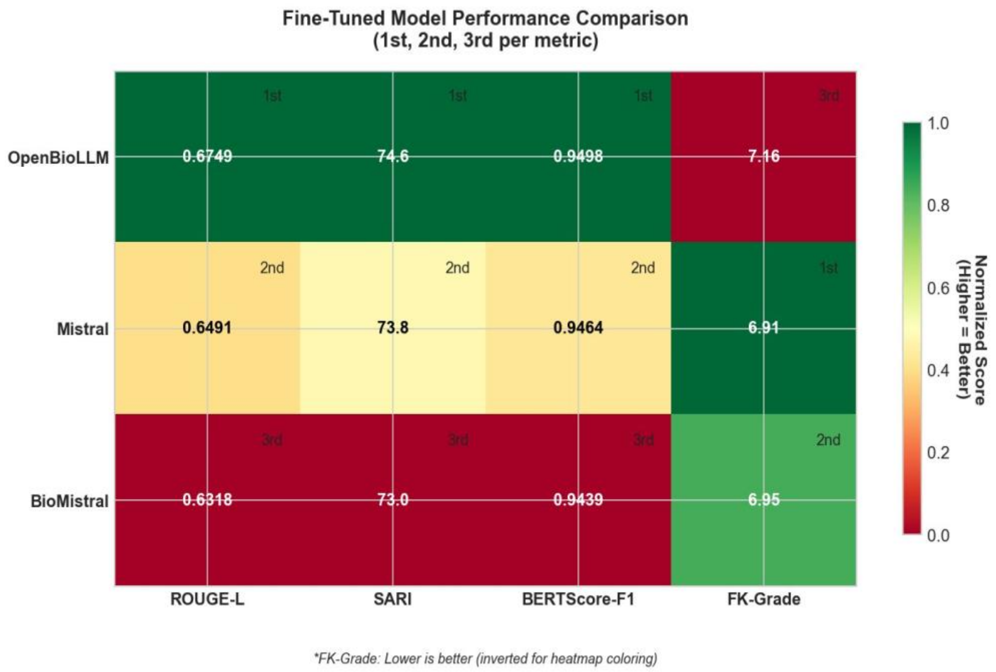


Fig. 16. Fine-tuned model performance comparison heatmap with rankings per metric. OpenBioLLM-8B ranks 1st on ROUGE-L, SARI, and BERTScore but 3rd on FK-Grade. Mistral-7B achieves best readability (1st on FK-Grade).

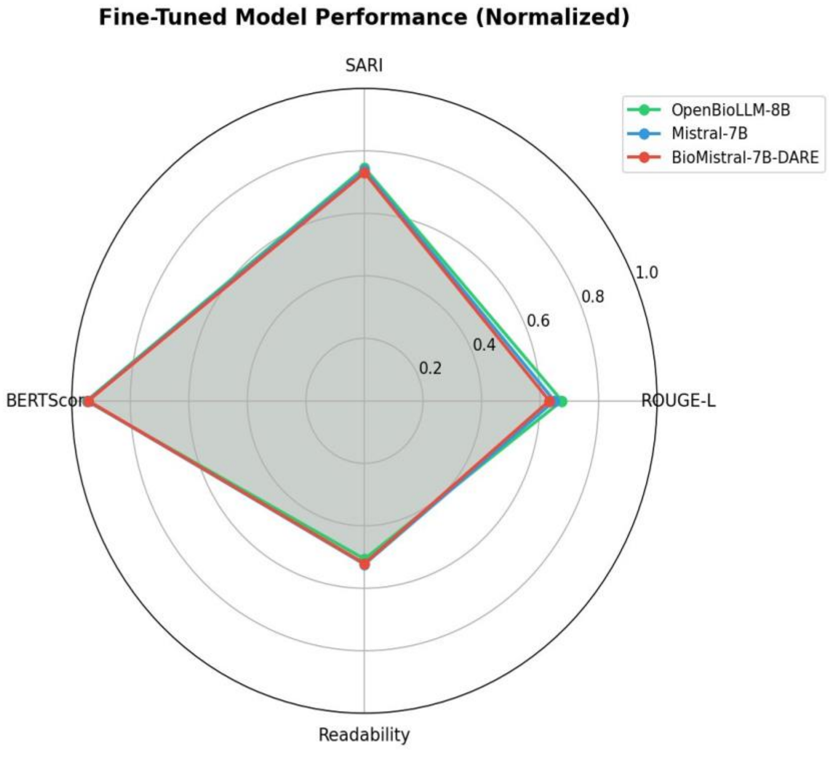


Fig. 17. Normalized performance radar chart comparing fine-tuned models across four metrics. OpenBioLLM-8B (green) shows largest area, indicating best overall performance. All models show similar profiles with slight variations.

H. Error Analysis

Error analysis reveals that approximately 95% of samples across all models produce satisfactory simplifications with no significant errors. The most common error types are: low ROUGE-L scores (2-4% of samples), high FK-Grade (1-2%), and over-compression (1-2%). OpenBioLLM-8B achieves the best ROUGE-L distribution with the highest proportion of samples scoring above 0.5. One notable outlier in OpenBioLLM-8B (sample 800) achieved FK=69.1 due to valid vocabulary simplification but extremely long run-on sentences, representing only 0.1% of samples with minimal impact on aggregate statistics.

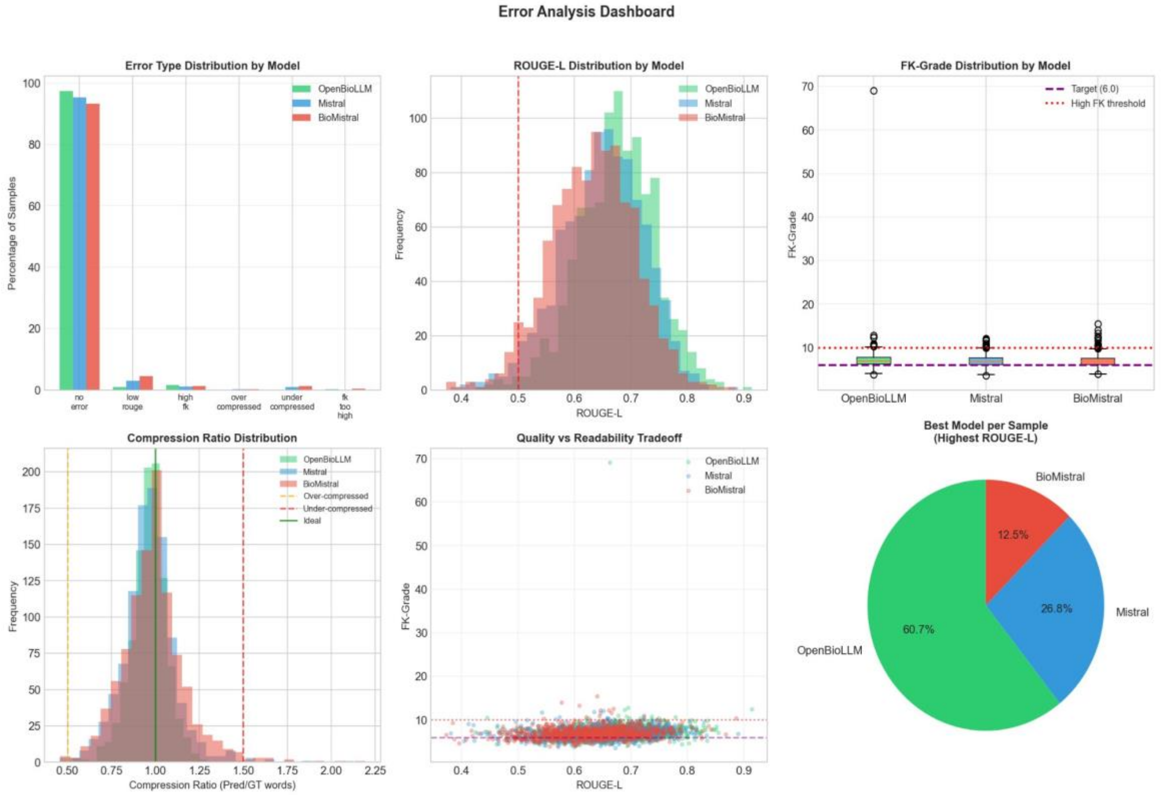


Fig. 18. Error analysis dashboard showing: error type distribution by model, ROUGE-L distribution histograms, FK-Grade boxplots, compression ratio distribution, quality vs readability tradeoff scatter, and pie chart of best model per sample. OpenBioLLM-8B is best for 60.7% of individual samples.

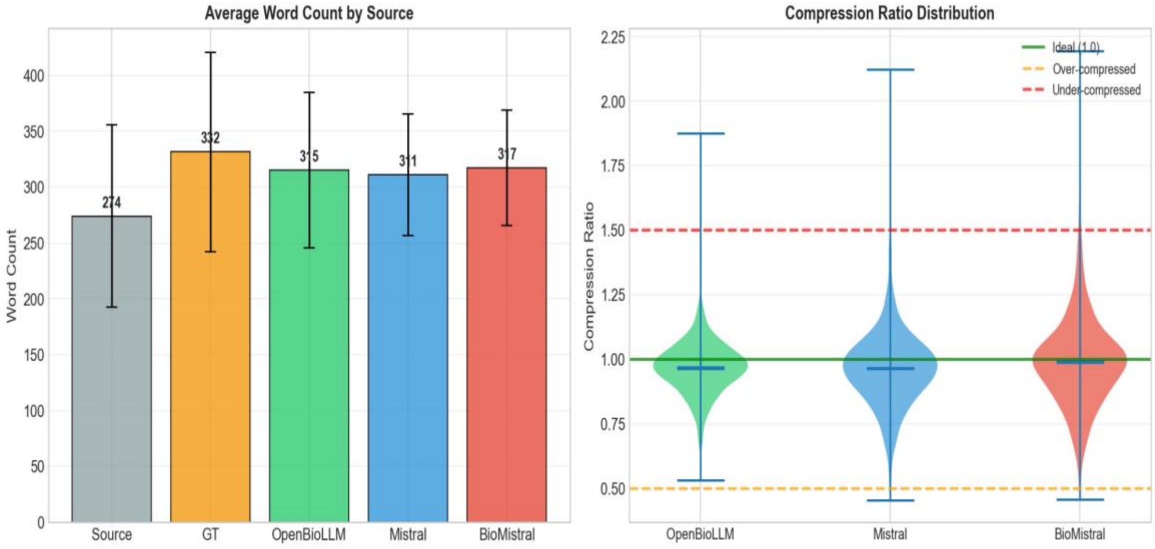


Fig. 19. Word count analysis: (Left) Average word counts showing model predictions (311-317 words) closely match ground truth (332 words) while expanding from source (274 words); (Right) Compression ratio distribution showing most samples near ideal ratio of 1.0.

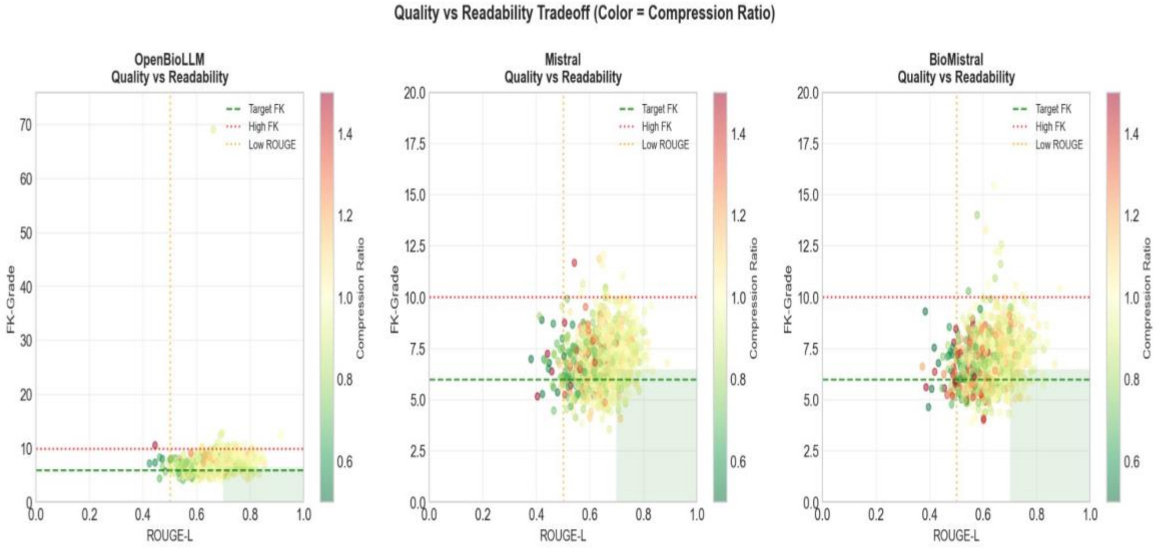


Fig. 20. Quality vs readability tradeoff scatter plots by model showing ROUGE-L vs FK-Grade with compression ratio as color. Green shaded regions indicate target zones (ROUGE-L > 0.5, FK-Grade < 10). OpenBioLLM-8B shows tightest clustering in optimal region.

VI. DISCUSSION

A. Ranking Reversal Phenomenon

The most striking finding is the complete ranking reversal between zero-shot and fine-tuned performance. OpenBioLLM-8B, despite achieving the worst zero-shot results (ROUGE-L 0.2623), achieves the best fine-tuned performance (ROUGE-L 0.6749), representing a 157% improvement. Conversely, BioMistral-7B-DARE achieves the best zero-shot results (0.4120) but the worst fine-tuned results (0.6318), representing only 53% improvement.

This inverse correlation suggests a floor effect: models with strong zero-shot capabilities have less room for improvement through fine-tuning. Alternatively, OpenBioLLM's Llama3 architecture may have higher learning capacity or better alignment with the fine-tuning objective. The practical implication is that zero-shot performance is not a reliable predictor of fine-tuned performance for this task.

B. LoRA Hyperparameter Insights

Our ablation studies reveal that higher LoRA rank consistently improves performance, contradicting the original LoRA paper's finding that r=4-8 is sufficient. This discrepancy may be explained by the rsLoRA hypothesis: standard LoRA scaling (α/r) causes gradient collapse at higher ranks, masking potential benefits. Our use of rsLoRA scaling (α/√r) may enable the observed improvements at r=32.

Similarly, using all attention projection layers (all_attn) outperforms more selective configurations (q_only, q_v) despite the doubled parameter count. This aligns with recent recommendations from Raschka (2023) and Unsloth, suggesting that the original LoRA configuration (Wq + Wv only) may be suboptimal. The practical implication is that practitioners should consider higher ranks and broader target modules when computational resources permit.

C. Medical Domain Considerations

Medical pretraining does not provide consistent advantages for text simplification. In zero-shot evaluation, BioMistral outperforms other models, but the general-purpose Mistral-7B is competitive. After fine-tuning, the medical-domain advantage disappears entirely, with the Llama3-based OpenBioLLM achieving best results. This suggests that the specific task of simplification may not require deep medical knowledge embedded during pretraining, as long as sufficient domain-specific fine-tuning data is available.

D. Limitations

Several limitations should be noted. First, we use synthetic clinical notes rather than real patient data, which may not capture all complexities of actual discharge summaries. Second, ground truth simplifications were generated by a single teacher model (Claude), potentially introducing bias. Third, automatic metrics may not fully capture clinical accuracy or patient comprehension. Fourth, we did not evaluate on external test sets or conduct human evaluation studies. Finally, our findings are specific to the 7-8B parameter scale and may not generalize to larger models.

VII. CONCLUSION

This paper presented MediSimplifier, a comprehensive framework for medical discharge summary simplification using LoRA fine-tuned language models. Through systematic evaluation of three models and extensive ablation studies, we demonstrate that: (1) LoRA fine-tuning achieves 53-157% improvement over zero-shot baselines; (2) Zero-shot performance inversely predicts fine-tuned performance, with the worst baseline model achieving best final results; (3) Higher LoRA ranks (r=32) and broader target modules (all_attn) consistently improve performance; (4) All fine-tuned models achieve approximately 50% readability reduction, from college level to 7th grade level; (5) Medical pretraining advantages disappear after domain-specific fine-tuning.

These findings provide practical guidance for practitioners applying LoRA to medical NLP tasks: consider higher ranks than originally recommended, target all attention layers when resources permit, and do not assume medical pretraining will provide advantages after fine-tuning. Future work should validate these findings on real clinical data, conduct human evaluation of simplification quality, and explore larger model scales.

REFERENCES

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.

[2] D. Kalajdzievski, "Rank Stabilization for Scaling Large Language Models with LoRA," arXiv preprint arXiv:2312.03732, 2023.

[3] S. Raschka, "Finetuning Large Language Models," Lightning AI Blog, 2023.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.

[5] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare, vol. 3, no. 1, pp. 1-23, 2021.

[6] W. Xu, C. Callison-Burch, and C. Napoles, "Problems in Current Text Simplification Research: New Data Can Help," Transactions of the Association for Computational Linguistics, vol. 3, pp. 283-297, 2015.

[7] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," arXiv preprint arXiv:1904.09675, 2019.

[8] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Text Summarization Branches Out, 2004, pp. 74-81.

[9] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing Statistical Machine Translation for Text Simplification," Transactions of the Association for Computational Linguistics, vol. 4, pp. 401-415, 2016.

[10] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas for Navy enlisted personnel," Naval Technical Training Command, Millington TN Research Branch, 1975.