Efficient Clinical Summarization: LoRA Adaptation of Fine-Tuned Models Matches Full Fine-Tuning Performance

Author: Vivek Tiwari

Email: vivekt@stanford.edu

Abstract

Clinical documentation in Electronic Health Records (EHRs) demands significant human time, and while automated summarization is promising, deployment requires quantifying the performance gap between conversation-based and note-based summarization. Using the AGBonnet dataset (30,000 cases), we systematically evaluate extractive baselines, zero-shot models, and fine-tuning strategies across both modalities. Our key finding shows that Low-Rank Adaptation (LoRA) applied to a fully fine-tuned note-summarization model achieves 98% of full fine-tuning performance on conversations (ROUGE-L: 0.549 vs 0.561) while training only 0.44% of parameters (1.77M vs 406M). This hybrid approach—full fine-tuning on structured notes followed by parameter-efficient adaptation to conversations—offers practical guidance for deploying clinical summarization in resource-constrained environments without sacrificing quality.

1 Introduction

Electronic Health Records (EHRs) capture patient encounters in unstructured text, including medical history, symptoms, diagnoses, treatments, and follow-up plans. While essential for care continuity, documentation places a heavy cognitive burden on physicians, contributing to burnout and reducing time for direct patient care [1]. Automation offers a promising solution.

1.1 Motivation and Problem Statement

Most clinical summarization research focuses on written clinical notes [2], yet real-world documentation begins with doctor–patient conversations captured by transcription or speech-to-text systems. Although large note datasets are available for training, cross-modality transfer from notes to conversations remains underexplored. We study how full fine-tuning and parameter-efficient methods such as Low-Rank Adaptation (LoRA) [3] enable conversational summarization. This leads to the question: How do models perform on conversational versus note input, and can they be efficiently adapted between these modalities?

1.2 Related Work

Prior work addresses summarization of clinical notes [2] and doctor-patient conversations [4, 5]. Krishna et al. fine-tuned BART on conversation transcripts, while Michalopoulos et al. enhanced performance using medical knowledge. These approaches emphasize architectural improvements and assume access to large conversation datasets for full fine-tuning. LoRA [3] enables parameter-efficient adaptation via low-rank updates to frozen models; however, its application to cross-modal clinical transfer remains unexplored.

1.3 Contributions

1. **Cross-Modal Evaluation:** We benchmark seven baselines on conversation- and note-based clinical summarization, revealing a consistent 42–160% ROUGE-L gap across baselines. The best zero-shot model (BART-large) still shows a 62% gap, motivating specialized methods.
2. **Transfer Learning Analysis:** Fully fine-tuned note models provide slightly better (0.7%) result for conversation summarization to direct training, with NOTE→CONV transfer achieving ROUGE-L 0.561 vs. 0.557 for direct conversation training.
3. **Parameter-Efficient Adaptation:** LoRA applied to a note-fine-tuned model reaches 98% of full fine-tuning performance (ROUGE-L 0.549 vs. 0.561) while training only 1.77M parameters (0.44%) compared to 406M (100%), enabling resource-efficient deployment.

2 Methods

2.1 Dataset

We use the AGBonnet [6] dataset derived from PMC-Patients [7], which truncates clinical case reports to a fixed length and generates synthetic doctor–patient conversations with structured JSON summaries. The dataset contains 30,000 examples. We convert JSON summaries to natural language references and split the data 80:10:10 into training (23,977), development (2,997), and test (2,998) sets. We construct two parallel datasets: truncated clinical notes $\rightarrow$ summary and synthetic conversations $\rightarrow$ summary.

Rather than generating new synthetic data, we use AGBonnet to systematically compare conversation- and note-based summarization and evaluate improvements across both modalities.

2.2 Baseline Methods

Extractive Baselines: (1) *Lead-3* selects the first three sentences to measure position bias. (2) *TextRank* [8] applies PageRank with word-overlap similarity for sentence ranking. (3) *LSA* applies TF-IDF with SVD to capture semantic importance.

Zero-Shot Neural Models: We evaluate four pre-trained seq2seq models without fine-tuning: BART-base, BART-large [9], T5-base, and T5-large [10], leveraging summarization pre-training without domain adaptation.

Implementation: Extractive methods use `sumy` and `nltk`; neural models use HuggingFace Transformers with default inference (beam width = 4). All code is implemented in Python 3.10.

2.3 Learning-Based Approaches

Based on baseline results showing BART-large's superior zero-shot performance, we focus subsequent experiments on this architecture (406M parameters).

Full Fine-Tuning: We fine-tune all model parameters using AdamW [11] optimizer (learning rate: 3e-5, batch size: 16, gradient accumulation: 2) B. We train for 5 epochs measuring summarization quality at each epoch.

LoRA Adaptation: We apply Low-Rank Adaptation to attention weights (query, key, value projections) with rank $r = 8$ and $\alpha = 16$. This introduces approximately 1.77M trainable parameters (0.44% of total), keeping the base model frozen.

Transfer Learning Configurations: We explore multiple training schedules:

- *Direct training (Lora or Full Fine Tune):* Train only on the target task (notes or conversations)
- *Sequential transfer(Lora or Full Fine Tune):* Train on notes, then continue on conversations
- *Hybrid approach:* Full fine-tuning on notes, then LoRA adapt to conversations

2.4 Evaluation Metrics

We report ROUGE-1, ROUGE-2, and ROUGE-L F1 scores [12]. ROUGE-L serves as our primary metric as it captures overall summary quality better than unigram/bigram scores alone. All metrics are computed using the `rouge-score` library with default parameters.

3 Experiments and Results

3.1 Baseline Performance Analysis

Table 1 presents zero-shot and unsupervised method performance across both modalities. We calculate the performance gap as $\frac{\text{Notes}-\text{Conversations}}{\text{Conversations}} \times 100\%$.

Key Observations:

Table 1: Baseline Performance: Zero-shot and unsupervised methods.

Category	Method	Conversations			Notes			Gap
		R-1	R-2	R-L	R-1	R-2	R-L	
Extractive	Lead-3	0.084	0.022	0.063	0.236	0.063	0.164	160%
	TextRank	0.130	0.035	0.095	0.333	0.152	0.225	137%
	LSA	0.267	0.100	0.178	0.360	0.178	0.253	42%
Zero-Shot	BART-base	0.238	0.092	0.156	0.446	0.251	0.315	102%
	T5-base	0.262	0.110	0.188	0.414	0.237	0.316	68%
	BART-large	**0.282**	**0.120**	**0.199**	**0.426**	**0.244**	**0.322**	**62%**
	T5-large	0.279	0.118	0.200	0.400	0.228	0.313	57%

Table 2: Full Fine-Tuning: Direct training versus transfer learning using BART-large.

Task	Configuration	Note Ep.	Conv Ep.	R-1	R-2	R-L
A. Clinical Notes Summarization (evaluated on note test set)						
Zero-shot	BART-large (pre-trained)	—	—	0.426	0.244	0.322
Direct	NOTE (1 epoch)	1	—	0.699	0.573	0.633
Direct	NOTE (2 epochs)	2	—	0.702	0.577	0.637
Direct	NOTE (3 epochs)	3	—	0.706	0.584	0.642
Direct	**NOTE (4-5 epochs)**	4-5	—	**0.709**	**0.587**	**0.645**
B. Conversation Summarization (evaluated on conversation test set)						
Zero-shot	BART-large (pre-trained)	—	—	0.282	0.120	0.199
Direct	CONV (1 epoch)	—	1	0.632	0.462	0.550
Direct	CONV (2 epochs)	—	2	0.632	0.463	0.549
Direct	CONV (3 epochs)	—	3	0.637	0.469	0.555
Direct	CONV (4 epochs)	—	4	0.640	0.470	0.558
Direct	CONV (5 epochs)	—	5	0.640	0.470	0.557
Transfer	NOTE(5)→CONV(1)	5	1	0.633	0.464	0.552
Transfer	NOTE(5)→CONV(2)	5	2	0.640	0.470	0.558
Transfer	NOTE(5)→CONV(3)	5	3	0.640	0.472	0.558
Transfer	**NOTE(5)→CONV(4-5)**	5	4-5	**0.643**	**0.474**	**0.561**

1. Cross-Modal Gap: All methods degrade on conversations vs. notes (42% LSA–160% Lead-3). Even BART-large (zero-shot) shows a 62% gap, reflecting fundamental modality differences beyond pre-training. LSA minimizes the gap (42%) by modeling semantics, but its ROUGE-L (0.178) is far below neural models.

2. Pre-trained Strength: BART-large reaches ROUGE-L 0.322 on notes without fine-tuning, confirming large neural models excel at language tasks.

3.2 Full Fine-Tuning Results

Table 2 presents systematic fine-tuning experiments across both modalities and transfer learning configurations.

Key Observations:

1. Finetuning Improves both Results and Gap: Notes reach ROUGE-L 0.645 after 4–5 epochs, a 100% gain over zero-shot (0.322). Conversations peak at 0.557–0.558 in 4–5 epochs, cutting the gap to 15.8%.

2. Transfer Equivalent to Direct: NOTE→CONV models score 0.561 vs. 0.557 for direct training. Transfer further shrinks the note–conversation gap from 15.8% to 15.0%, showing structured note knowledge transfers to conversational contexts.

3. Early Convergence: Both direct and transfer plateau after epoch 3–4, indicating efficient convergence for resource-limited training.

Table 3: Comparison of training strategies for conversation summarization. The hybrid approach (full FT on notes + LoRA on conversations) achieves 98% of full fine-tuning performance with only 0.44% trainable parameters. See appendix for Lora Direct C

Category	Configuration	Note Train.	Conv Train.	Params	R-L
Baseline	Zero-shot BART-large	—	—	0	0.199
Full Fine-Tuning (100% parameters, 406M)					
Direct FT	CONV only (5 epochs)	None	Full FT	406M	0.557
Transfer FT	NOTE(5)→CONV(5)	Full FT	Full FT	406M	**0.561**
Parameter-Efficient Fine-Tuning (0.44% parameters, 1.77M)					
LoRA Direct	NOTE+CONV (LoRA)	LoRA (4ep)	LoRA (1ep)	1.77M	0.495
Hybrid	**FT-NOTE→LoRA-CONV**	**Full FT (5ep)**	**LoRA (2ep)**	**1.77M**	**0.549**
Performance: Hybrid achieves 98% of Transfer FT (0.549 vs 0.561) with 99.6% fewer trainable parameters					

3.3 LoRA Adaptation: Matching Full Fine-Tuning with 0.44% Parameters

Table 3 presents our central finding: LoRA adaptation applied to a fully fine-tuned note model achieves performance comparable to full fine-tuning while training only 0.44% of parameters.

Key Observations:

1. Hybrid Efficiency: Full fine-tuning on notes + LoRA adaptation to conversations yields ROUGE-L 0.549—just 2% below full fine-tuning (0.561) while training 99.6% fewer parameters (1.77M vs. 406M). A strong note model thus enables parameter-efficient conversation adaptation.

2. LoRA-Only Limits: LoRA from scratch across modalities scores ROUGE-L 0.495—10% below hybrid. Rank-8 adapters (1.77M params) lack capacity to learn clinical summarization fully but excel at task-specific adaptation when built on a strong base.

3. Fast Convergence: Table 4 shows LoRA converges within 2 epochs at ROUGE-L 0.549, with minimal gains thereafter.

Table 4: LoRA on fully fine-tuned note model: Per epoch performance on conversations

Epoch	R-1	R-2	R-L	vs. Full FT
1	0.630	0.457	0.545	-2.9%
2	**0.633**	**0.460**	**0.549**	-2.1%
3	0.632	0.460	0.549	-2.1%
4	0.633	0.460	0.549	-2.1%
5	0.628	0.455	0.542	-3.4%
Best LoRA (epoch 2): 0.549 vs Full FT: 0.561 = 98% performance				

3.4 Qualitative Analysis

We manually examined outputs from both Transfer FT and Hybrid LoRA models across test cases with varying ROUGE scores. For high-scoring examples (ROUGE-L > 0.9), both approaches yield nearly identical summaries. In moderate cases (ROUGE-L 0.5–0.7), the Hybrid LoRA model sometimes includes slightly more or fewer details but consistently preserves clinical accuracy. For low-scoring cases (ROUGE-L < 0.2), both models struggle, typically when reference summaries are extremely brief or conversations lack clear clinical structure. Overall, the 2% ROUGE-L difference corresponds to minimal qualitative variation in practice. Detailed examples are provided in Appendix A. **We also observe that truncation during conversation generation can omit information present in the reference summary, leading the model to hallucinate patient age or other details when such details are missing.**

4 Discussion

4.1 Why Does Transfer Learning Work?

Clinical notes provide superior initialization for conversation summarization by concentrating medical terminology, diagnoses, and treatments (entity recognition), encoding standardized symptom→diagnosis→treatment reasoning (causal structure learning), and condensing encounters to highlight essential details (information compression). These mechanisms transfer directly to conversations, enabling models to filter scattered medical facts from dialogue noise more effectively than direct training.

4.2 Why Does LoRA on Fine-Tuned Notes Succeed?

Our hybrid approach (full fine-tuning on notes followed by LoRA on conversations) succeeds because clinical summarization decomposes into two capabilities: domain knowledge, requiring substantial capacity to learn medical vocabulary, reasoning patterns, and diagnostic relationships (captured through full fine-tuning), and modality adaptation, requiring lightweight adjustment to handle dialogue turns, speaker attribution, and temporal progression (captured through LoRA). LoRA-only training underperforms because limited capacity cannot simultaneously acquire both domain knowledge and modality adaptation.

4.3 Limitations and Future Work

Our study is limited by reliance on synthetic conversations, which may not capture the complexity of real clinical dialogue, highlighting the need for validation on real-world data. **We also observe hallucinations linked to data quality, underscoring the importance of evaluating longer, more accurate conversations**. Results are based on single-seed experiments without statistical testing, and our focus on BART-large reflects resource constraints, meaning larger models may behave differently. Moreover, ROUGE scores may not fully capture clinical correctness, necessitating expert human evaluation. Finally, practical deployment challenges—including ASR errors, privacy concerns, and EHR integration—remain unaddressed.

4.4 Clinical Implications

Our results support a practical deployment strategy for clinical summarization. Models can be trained on abundant clinical note datasets and later adapted to scarce conversation data using lightweight LoRA updates. This two-stage approach reduces data dependence and computational cost, enabling deployment in resource-constrained settings. LoRA also allows rapid updates as clinical practices or terminology evolve.

5 Conclusion

We conducted a systematic study of clinical summarization across conversation and note modalities, showing that a hybrid strategy—full fine-tuning on notes followed by LoRA adaptation to conversations—achieves 98% of full fine-tuning performance (ROUGE-L: 0.549 vs 0.561) while training only 0.44% of parameters (1.77M vs 406M). This approach addresses the large modality performance gap (42–160% across baselines, 62% for the best zero-shot model) and computational constraints in healthcare. Our experiments demonstrate that knowledge learned from structured notes transfers to conversations, reducing the gap from 15.8% to 15.0%, and that summarization decomposes into domain knowledge (requiring full capacity) and modality adaptation (requiring lightweight adjustments), enabling efficient fine-tuning. Future work should validate these results on real doctor-patient dialogues, include human evaluations, and test whether similar transfer patterns extend to larger models and other tasks such as diagnosis prediction or treatment recommendation.

Contributions and AI Tool Usage

Vivek Tiwari was solely responsible for all aspects of this project, including: dataset selection and preprocessing, experimental design and implementation, baseline and neural model development, training and evaluation of all models, results analysis, and report preparation.

AI Tool Usage: GitHub Copilot and Claude (v4.5) were used to assist with code generation, debugging, documentation, and report editing. All generated code and text were reviewed, modified as necessary, and validated by the author to ensure correctness and alignment with project objectives.

References

[1] Ming Tai-Seale, Elizabeth C Dillon, Yuting Yang, Robyn Nordgren, Renae L Steinberg, Tracey Nauenberg, Dominick L Frosch, Trisha Lee, Rachel Gilmore, Anna Meehan, et al. Measuring documentation burden in healthcare. *Journal of General Internal Medicine*, 32, 2017.

[2] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[4] Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. Leveraging pre-trained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*, 2021.

[5] Gagandeep Singh Thomas Lin George Michalopoulos, Kyle Williams. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. 2022.

[6] Antoine Bonnet. AGBonnet: Augmented Clinical Notes Dataset, 2024.

[7] Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. PMC-Patients: A Large-scale Dataset of Patient Notes and Relations Extracted from Case Reports in PubMed Central, 2022.

[8] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text, 2004.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[10] Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

[12] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries, 2004.

A Qualitative Examples

Table 5, 6 and 7 provides 3 examples comparing outputs from Transfer FT (full fine-tuning on notes then conversations) and Hybrid LoRA (full fine-tuning on notes + LoRA adaptation to conversations) approaches.

Conversation idx: 64230	Doctor: Good morning, can you tell me your name and what brings you here today? Patient: Good morning, my name is [Name]. I was admitted to the emergency room with mental confusion and ataxia. Doctor: Okay, I see. And can you tell me a little bit about your medical history? Patient: Yes, I'm a smoker and I have been diagnosed with chronic obstructive pulmonary disease (COPD) before. Doctor: I see. And what happened before you were admitted to the emergency room? Patient: I just felt confused and had trouble walking. [... conversation continues ...]	
Reference Summary	69-year-old male presented with mental confusion and ataxia. Medical history includes chronic obstructive pulmonary disease (copd). Primary complaint was mental confusion. Diagnosed with small cell lung cancer, small cell lung cancer. Patient was treated with palliative chemotherapy for small cell lung cancer at carboplatin AUC 5 on day 1 and etoposide 100mg/m2 on days 1 to 3.	
	Transfer FT	**Hybrid LoRA**
Generated Summary	72-year-old male presented with mental confusion and ataxia. Medical history includes chronic obstructive pulmonary disease (copd). Primary complaint was mental confusion. Diagnosed with small cell lung cancer. Patient was treated with Palliative chemotherapy for small cell carcinoma at Carboplatin AUC 5 on day 1 and etoposide 100mg/m2 on days 1 to 3.	72-year-old male presented with mental confusion and ataxia. Medical history includes chronic obstructive pulmonary disease (copd). Primary complaint was mental confusion. Diagnosed with small cell lung cancer. Patient was treated with Palliative chemotherapy for small cell carcinoma at Carboplatin AUC 5 on day 1 and etoposide 100mg/m2 on days 1 to 3.
ROUGE-F1	0.943	0.943
Summary Length	52 words	52 words

Table 5: High-scoring example (idx: 64230): Both models produce identical summaries with ROUGE-F1 of 0.943.

Conversation-idx: 130273	Doctor: Hello, how are you feeling today? Patient: I'm feeling okay, doctor. Doctor: Can you tell me why you're here today? Patient: I'm here for a routine check-up. Doctor: Alright, let's start with your medical history. Have you had any significant surgeries or medical procedures recently? Patient: No, I haven't had any surgeries or procedures. [... conversation continues ...]	
Reference Summary	52-year-old male was evaluated.	
	Transfer FT	**Hybrid LoRA**
Generated Summary	42-year-old female was evaluated. Medical history includes no significant medical or surgical history. Diagnosed with celiac artery originated as a ventral branch of the abdominal aorta.	40-year-old male presented with routine check-up. Medical history includes no significant surgeries or medical procedures. Diagnosed with celiac artery, hepatic artery. Patient was discharged patient was discharged from the hospital.
ROUGE-F1	0.138	0.138
Summary Length	26 words	30 words

Table 6: Low-scoring example (idx: 130273): Both models struggle with minimal reference summary (5 words), achieving identical ROUGE-F1 of 0.138.

B Experimental Configuration

All experiments were conducted on NVIDIA A100 GPUs using PyTorch 2.0+ with BFloat16 precision. The base model is `facebook/bart-large-cnn` (406M parameters). Table 8 lists all hyperparameters used in our experiments.

Hardware: All models were trained on NVIDIA A100 (40GB) GPUs. Full fine-tuning used gradient checkpointing to fit the 406M parameter model in memory. LoRA adaptation required significantly less memory due to frozen base model weights.

Training Strategy: For full fine-tuning, we trained separate models on notes and conversations for 5 epochs each. For transfer learning, we first fine-tuned on notes for 5 epochs, then continued training on conversations for 5 additional epochs. For the hybrid LoRA approach, we fully fine-tuned on notes for 5 epochs, then applied LoRA adapters and trained on conversations for 5 epochs.

Conversation-idx: 174088	Doctor: Hello, how are you feeling today? I see that you presented for evaluation of a new asymptomatic bump on your left arm. Patient: Yes, doctor. It's a new lesion that wasn't present during my skin examination six months ago. Doctor: I see. Just to clarify, you mentioned that you previously had three basal cell carcinomas and one squamous cell carcinoma excised on different parts of your body. Is that correct? Patient: Yes, that's right. The basal cell carcinomas were on my left temple, left side of my upper lip, and left mid back 32 years, four years and two years ago, respectively. The s...	
Reference Summary	68-year-old male presented with evaluation of a new asymptomatic bump on his left arm. Medical history includes severe acne vulgaris as an adolescent, hypercholesterolemia, hypertension, prostate cancer diagnosed one year ago. Primary complaint was asymptomatic bump in the left arm near the elbow for new. Diagnosed with nodular basal cell carcinoma. Patient was treated with Mupirocin two percent ointment for post-biopsy care at Two percent. Patient underwent robotic-assisted laparoscopic prostatectomy, bilateral pelvic lymph node dissection for prostate cancer. Surgical outcome was no evidence...	
	Transfer FT	**Hybrid LoRA**
Generated Summary	68-year-old male presented with evaluation of a new asymptomatic bump on his left arm. Medical history includes severe acne vulgaris as an adolescent, hypercholesterolemia, hypertension, prostate cancer diagnosed one year ago. Primary complaint was new lesion on left arm in the left arm near the elbow for new. Diagnosed with nodular basal cell carcinoma. Patient was treated with Mupirocin two perc...	71-year-old male presented with evaluation of a new asymptomatic bump on the left arm. Medical history includes three basal cell carcinomas and one squamous cell carcinoma excised on different parts of the body, actinic keratoses treated with liquid nitrogen cryotherapy, severe acne vulgaris as an adolescent, hypercholesterolemia, hypertension, prostate cancer diagnosed one year ago. Primary compl...
ROUGE-F1	0.967	0.522
Summary Length	95 words	103 words

Table 7: Moderate-scoring example (idx: 174088): Transfer FT achieves higher ROUGE (0.967) while Hybrid LoRA includes more historical details, resulting in lower overlap (0.522) despite clinical completeness.

C LoRA Transfer Learning Results

Table 9 presents detailed results from our transfer learning experiments using LoRA adaptation. We compare a baseline approach (LoRA only on conversations for 5 epochs) against progressive transfer learning strategies where we first train on notes for varying epochs (1-4) and then continue with LoRA adaptation on conversations for the remaining epochs (4-1), maintaining a total of 5 epochs of conversation training.

Table 8: Hyperparameters for BART-large fine-tuning and LoRA adaptation

Category	Parameter	Full FT	LoRA
LoRA-Specific Configuration			
	Rank (r)	—	8
	Alpha (α)	—	16
	Dropout	—	0.1
	Target modules	—	q_proj, k_proj, v_proj
	Trainable parameters	406M	1.77M
Training Hyperparameters			
	Optimizer	AdamW (fused) [11]	
	Learning rate	3e-5	2e-4
	Batch size (per device)	16	16
	Gradient accumulation	2	2
	Effective batch size	32	32
	Warmup ratio	0.1	0.1
	Weight decay	0.01	0.01
	Max gradient norm	1.0	1.0
	Precision	BFloat16	
	Gradient checkpointing	Yes	No
	Epochs (notes)	5	5
	Epochs (conversations)	5	5
Generation Parameters			
	Beam width	3	
	Length penalty	1.5	
	Max length	150 tokens	
	Min length	40 tokens	
	No-repeat n-gram	3	
	Early stopping	True	
Data Configuration			
	Max input length	1024 tokens	
	Train/Dev/Test split	23,977 / 2,997 / 2,998	
	Evaluation batch size	32	

Table 9: Transfer Learning with LoRA: Conversation Task Performance

Experiment	NOTE ep	CONV ep	Trans	R-1	R-2	R-L
BART-large (zero-shot)	—	—	—	0.282	—	—
Baseline (no transfer)	—	5	No	0.578	0.404	0.479
Transfer (1 ep NOTE)	1	4	Yes	0.574	0.400	0.472
Transfer (2 ep NOTE)	2	3	Yes	0.575	0.401	0.473
Transfer (3 ep NOTE)	3	2	Yes	0.582	0.406	0.477
Transfer (4 ep NOTE)	4	1	Yes	0.593	0.417	0.495