

Science Teaching Assistant: Fine-Tuning Nanochat for Middle/High School Science Education

Team Members: [Name 1], [Name 2], [Name 3], [Name 4]

Course: [Course Code and Name]

Date: December 4, 2025

Abstract

<Team Member 1>

Content Hint: Write a concise 150-200 word summary covering:

- Project objective: Creating a science teaching assistant by fine-tuning nanochat on ScienceQA dataset
 - Key technical approach: Transfer learning using pre-trained 561M parameter model
 - Main methodology: Supervised fine-tuning on 3k-5k science Q&A pairs from elementary to high school level
 - Dataset used: ScienceQA subset with explanations and lectures
 - Training platform: Kaggle free tier GPU
 - Key results: Quantitative improvements (X% accuracy gain over base model) and qualitative improvements (detailed explanations, subject-appropriate responses)
 - Significance: Demonstrates practical application of domain adaptation for educational NLP
-

1. Introduction

<Team Member 1>

1.1 Background and Motivation

<Team Member 1>

Content Hint: Explain the need for AI-powered educational assistants in science education.

Cover:

- Growing demand for personalized learning tools in STEM education
- Challenges students face in accessing quality science tutoring (especially at middle/high school level)
- Recent advances in large language models (LLMs) that make domain-specific assistants feasible
- Limitations of general-purpose LLMs (like ChatGPT) when applied to specific educational contexts without fine-tuning
- The opportunity: nanochat as a minimal, hackable LLM implementation that enables hands-on learning about the full LLM pipeline
- Educational value of building end-to-end NLP systems vs. using black-box APIs

1.2 Importance and Relevance

<Team Member 1>

Content Hint: Justify why this project matters from both educational and technical perspectives:

- **Educational Impact:** Science teaching assistants can provide instant, patient explanations with examples
- **Accessibility:** Free, always-available tutoring for students who lack access to human tutors
- **Technical Relevance:** Demonstrates practical transfer learning and domain adaptation techniques critical in modern NLP
- **Cost-Effectiveness:** Shows how educational institutions with limited budgets can deploy custom AI solutions using free computing resources (Kaggle)
- **Alignment with Course Objectives:** Covers end-to-end NLP pipeline from pre-training concepts through fine-tuning and evaluation

1.3 Objectives and Research Questions

<Team Member 1>

Content Hint: Clearly state the project goals and questions you aim to answer:

Primary Objective: Fine-tune a pre-trained language model (nanochat) to create a science teaching assistant capable of answering elementary through high school science questions with detailed explanations.

Secondary Objectives:

- Understand and document the architecture of modern transformer-based LLMs
- Demonstrate effective dataset preparation and conversion for conversational AI
- Evaluate fine-tuning impact using both quantitative metrics and qualitative analysis
- Deploy an interactive demo showcasing the teaching assistant

Research Questions:

1. How effectively can a general-purpose pre-trained LLM be adapted to a specific educational domain with limited training data (3k-5k samples)?
 2. What performance improvements (accuracy, explanation quality) can be achieved through supervised fine-tuning on domain-specific science Q&A data?
 3. How does the fine-tuned model compare to the base model in providing pedagogically appropriate responses for different grade levels?
 4. What are the practical constraints (compute, cost, time) for implementing such a system using free/low-cost resources?
-

2. Problem Statement

<Team Member 1>

2.1 Problem Definition

<Team Member 1>

Content Hint: Clearly articulate the specific problem this project addresses:

- **Core Problem:** Students need immediate, clear explanations for science concepts across multiple topics (Biology, Chemistry, Physics, Earth Science), but human tutors are not always available
- **Technical Challenge:** General-purpose LLMs lack the focused, pedagogically appropriate responses needed for science education (may provide overly complex answers, lack structured explanations, or fail to ground responses in grade-appropriate language)
- **Gap in Existing Solutions:** Off-the-shelf chatbots aren't optimized for the specific format and style of science education (multiple-choice reasoning, concept explanations with examples, lecture-style content)
- **Target Users:** Middle school (grades 6-8) and high school (grades 9-12) students seeking homework help and concept clarification

2.2 Challenges in Existing Solutions

<Team Member 1>

Content Hint: Discuss limitations of current approaches and why fine-tuning is necessary:

1. General-Purpose LLM Limitations:

- Lack domain-specific knowledge organization
- May provide inconsistent pedagogical quality
- Often generate verbose responses inappropriate for student learning level
- No guarantee of following educational best practices (e.g., providing explanations with examples)

2. Commercial Educational AI Limitations:

- Expensive subscription models limiting accessibility
- Closed-source systems that can't be customized for specific curricula
- Limited transparency in how answers are generated

3. Pre-built API Limitations:

- Using APIs without fine-tuning violates course requirements for implementing custom NLP models
- No learning opportunity about the full NLP pipeline
- Lack of control over model behavior and response style

4. Technical Challenges:

- **Data Formatting:** Converting standard Q&A datasets (like ScienceQA) into conversational formats suitable for LLM fine-tuning
- **Compute Constraints:** Training large models (561M parameters) with limited free-tier GPU resources
- **Evaluation Difficulty:** Assessing both factual correctness and pedagogical appropriateness of generated explanations

3. Project Architecture

<Team Member 2>

3.1 System Overview

<Team Member 2>

Content Hint: Provide a high-level description of the complete system architecture:

- **Input:** User submits a science question (text format, optionally with multiple-choice options)
- **Processing Pipeline:** Question tokenization → Model inference → Response generation → Post-processing
- **Output:** Structured answer with explanation, optionally including background lecture material
- **Components:** (1) Pre-trained base model (nanochat), (2) Fine-tuned adapter/weights, (3) Tokenizer, (4) Inference engine, (5) Web interface (Gradio)

[Insert: System Architecture Diagram] *Figure 1: End-to-end system architecture showing data flow from user input through fine-tuned model to generated response*

3.2 Base Model Architecture: Nanochat

<Team Member 2>

3.2.1 Transformer Foundation

<Team Member 2>

Content Hint: Explain the transformer architecture that underpins nanochat:

- **Model Type:** Decoder-only transformer architecture (similar to GPT family)
- **Parameters:** 561 million trainable parameters (comparable to GPT-2 Medium/Large)
- **Design Philosophy:** Nanochat is built as a minimal, educational implementation focusing on clarity and hackability over maximum performance
- **Key Characteristics:**
 - ~8,000 lines of code (primarily PyTorch with some Rust for tokenization)
 - Full pipeline implementation: tokenization, pre-training, fine-tuning, inference, serving
 - Designed as capstone project for LLM101n course by Eureka Labs
 - Single-file architecture for easier understanding and modification

3.2.2 Model Depth and Layer Configuration

<Team Member 2>

Content Hint: Describe the specific model configuration used:

- **Configuration:** Depth-20 (d20) or Depth-16 (d16) variants available
 - We used: [specify which depth model you downloaded]
- **Layer Structure:** Each layer contains:
 - Multi-head self-attention mechanism (attention heads: [specify])
 - Feed-forward neural network (hidden dimension: [specify])
 - Layer normalization
 - Residual connections
- **Context Window:** [Specify maximum sequence length, typically 2048 or 4096 tokens]
- **Vocabulary Size:** 65,536 tokens (custom BPE tokenizer trained using Rust implementation)

3.2.3 Self-Attention Mechanism

<Team Member 2>

Content Hint: Explain how self-attention works in the model (key technical detail for graduate NLP):

- **Purpose:** Enables the model to weigh the importance of different tokens in the sequence when processing each token
- **Multi-Head Attention:** Uses multiple parallel attention heads to capture different types of relationships:
 - Some heads may focus on syntactic relationships
 - Others may capture semantic similarities
 - Allows model to attend to information from different representation subspaces
- **Mathematical Formulation:** Briefly describe the Query-Key-Value mechanism:
 - Queries (Q), Keys (K), and Values (V) are linear projections of input embeddings
 - Attention scores = $\text{softmax}(QK^T / \sqrt{d_k})$
 - Output = Attention scores \times V
- **Causal Masking:** Decoder-only architecture uses masked attention to prevent attending to future tokens (ensuring autoregressive generation)

3.2.4 Pre-training Details

<Team Member 2>

Content Hint: Describe how the base nanochat model was pre-trained (even though you're downloading weights):

- **Pre-training Corpus:** Trained on 38 billion tokens from diverse web text
- **Objective:** Next-token prediction (language modeling) - learning to predict the next token given previous context
- **Training Infrastructure:** Originally trained on 8×H100 GPUs for ~~33 hours~~ (\$800 cost)
- **Why We Use Pre-trained Weights:** Due to budget constraints, we leverage the ~~(sdobson/nanochat)~~ checkpoint from HuggingFace rather than training from scratch
- **Transfer Learning Rationale:** Pre-trained model has learned general language understanding, grammar, reasoning, and factual knowledge that can be adapted to our science education domain

3.3 Fine-Tuning Architecture Modifications

<Team Member 2>

Content Hint: Explain what changes during fine-tuning vs. pre-training:

- **Approach:** Supervised Fine-Tuning (SFT) - full model fine-tuning (not parameter-efficient methods like LoRA)
- **What Changes:** All model parameters are updated during fine-tuning, but initialized from pre-trained weights rather than random initialization
- **Training Objective:**
 - Shift from general next-token prediction to domain-specific conversational response generation
 - Model learns to map science questions → detailed, pedagogical answers
- **Data Format Adaptation:**
 - Pre-training: Raw text sequences
 - Fine-tuning: Structured conversation format with system prompts, user questions, and assistant responses
- **Key Hyperparameters Modified:**
 - Lower learning rate (5e-5) compared to pre-training to preserve learned features
 - Smaller batch sizes due to GPU memory constraints (batch size: 2, gradient accumulation: 8)
 - Shorter training duration (2-3 epochs vs. many epochs in pre-training)

3.4 Data Flow and Processing Pipeline

<Team Member 2>

Content Hint: Trace how data moves through the system:

Training Phase:

- 1. Data Collection:** Download ScienceQA dataset from HuggingFace
- 2. Subset Creation:** Select 3,000-5,000 balanced samples across topics
- 3. Format Conversion:** Transform Q&A pairs into conversational structure:
 - System prompt: "You are a helpful science tutor..."
 - User message: Science question with multiple-choice options
 - Assistant message: Correct answer + explanation + optional lecture content
- 4. Tokenization:** Convert text to token IDs using nanochat's BPE tokenizer
- 5. Batching:** Group samples into batches with padding/truncation (max length: 512 tokens)
- 6. Training Loop:** Forward pass → Loss computation → Backpropagation → Parameter update
- 7. Checkpoint Saving:** Save model weights at regular intervals

Inference Phase:

- 1. User Input:** Science question submitted via web interface
- 2. Tokenization:** Convert question to token IDs
- 3. Model Inference:** Autoregressive generation (token-by-token prediction)
- 4. Decoding:** Convert generated token IDs back to text
- 5. Post-processing:** Format response for display
- 6. UI Rendering:** Show response in Gradio chat interface

[Insert: Data Flow Diagram] *Figure 2: Data processing pipeline from raw ScienceQA data through fine-tuning to inference*

4. Tools and Technologies

<Team Member 2>

4.1 Software Tools

<Team Member 2>

Content Hint: List and justify all software tools used:

Programming Language:

- **Python 3.10+:** Primary language for all implementation
 - Justification: De facto standard for NLP/ML, extensive library support

Core Libraries and Frameworks:

- **PyTorch 2.0+:** Deep learning framework
 - Used for: Model definition, training loops, gradient computation
 - Justification: Nanochat is built on PyTorch; provides flexibility and debugging ease
- **Transformers (HuggingFace):** Model hub and utilities
 - Used for: Loading pre-trained models, tokenizers, training utilities
 - Justification: Industry standard for working with transformer models
- **Datasets (HuggingFace):** Dataset management
 - Used for: Loading ScienceQA, data splitting, preprocessing
 - Justification: Efficient handling of large datasets with built-in caching
- **Accelerate (HuggingFace):** Training optimization
 - Used for: Automatic mixed precision (FP16), gradient accumulation
 - Justification: Enables training larger models on limited GPU memory

Supporting Libraries:

- **NumPy:** Numerical computations
- **Pandas:** Data manipulation and analysis during preprocessing
- **Gradio:** Web interface for demo
 - Justification: Quick deployment of interactive ML demos with minimal code
- **Matplotlib/Seaborn:** Visualization of training metrics

Development Environment:

- **Jupyter Notebook:** Interactive development and experimentation
 - Platform: Kaggle Notebooks / Google Colab
 - Justification: Built-in GPU access, easy sharing, reproducible environment

Version Control:

- **Git + GitHub:** Code versioning and collaboration
 - Repository includes: Training scripts, preprocessing code, evaluation notebooks, README

4.2 Hardware Resources

<Team Member 2>

Content Hint: Detail the compute resources used:

Training Hardware:

- **Platform:** Kaggle Notebooks (Free Tier)
 - **GPU:** NVIDIA Tesla P100 or T4 (16GB VRAM)
 - **GPU Hours Used:** ~3-4 hours for fine-tuning
 - **Memory:** 16GB RAM + 16GB GPU memory
 - **Storage:** 20GB temporary disk space

Alternative Platforms Considered:

- Google Colab Free/Pro: Similar GPU access, used for overflow capacity
- Paperspace Gradient: Evaluated but Kaggle proved sufficient

Compute Optimization Strategies:

- **Mixed Precision Training (FP16):** Reduces memory usage by ~50%
- **Gradient Accumulation:** Simulates larger batch sizes (effective batch: 16) without exceeding memory
- **Gradient Checkpointing:** Trades computation for memory by recomputing activations during backprop
- **Model Precision:** Using float16 instead of float32 during training

Cost Analysis:

- Total compute cost: \$0 (entirely on free tier)
- Time investment: ~4-6 hours total GPU time across all experiments
- Demonstrates feasibility for resource-constrained educational settings

Inference Hardware:

- **Deployment:** Gradio app on Kaggle or Colab
- **Requirements:** CPU sufficient for demo (inference much lighter than training)
- **Response Time:** ~2-3 seconds per query on CPU, <1 second on GPU

5. Methodology

<Team Member 3>

5.1 Dataset Selection and Characteristics

<Team Member 3>

Content Hint: Thoroughly describe the ScienceQA dataset:

Dataset Overview:

- **Name:** ScienceQA
- **Source:** HuggingFace (derek-thomas/ScienceQA)
- **Size:** ~21,000 multimodal multiple-choice questions
- **Coverage:**
 - Subjects: Natural science, language science, social science
 - Grade levels: Elementary through high school
 - Topics: 26 distinct topics
 - Categories: 127 categories
 - Skills: 379 specific skills tested
- **Format:** Multiple-choice questions (typically 4 options) with:
 - Question text
 - Answer choices
 - Correct answer index
 - Detailed explanation (90.5% of questions)
 - Grounded lecture content (83.9% of questions)

Why ScienceQA:

1. **Domain Alignment:** Specifically designed for science education assessment
2. **Pedagogical Quality:** Includes explanations and lectures ideal for teaching assistant persona
3. **Grade Appropriateness:** Covers target audience (middle/high school)
4. **Manageable Size:** Large enough for effective fine-tuning, small enough for free-tier GPUs
5. **Rich Annotations:** Detailed explanations enable learning to generate helpful responses, not just correct answers

Dataset Splits:

- Original: Train (16,000), Validation (2,500), Test (2,500)
- Our Subset: Train (3,000), Validation (500), Test (500)
 - Justification: Subset size balances compute constraints with training effectiveness

5.2 Data Preprocessing

<Team Member 3>

5.2.1 Subset Selection Strategy

<Team Member 3>

Content Hint: Explain how you created the subset:

- **Sampling Method:** Random sampling with fixed seed (42) for reproducibility
- **Balanced Selection:** Ensured representation across:
 - Different subjects (Biology, Chemistry, Physics)
 - Various difficulty levels (elementary → high school)
 - Topics with and without lecture content
- **Size Justification:** 3k-5k samples proven effective for domain adaptation in prior work; fits within free-tier compute budget

5.2.2 Data Cleaning and Quality Checks

<Team Member 3>

Content Hint: Describe preprocessing steps:

- **Text Normalization:**
 - UTF-8 encoding verification
 - Whitespace standardization
 - Removal of malformed entries (if any)
- **Quality Filters:**
 - Removed questions with missing answer keys
 - Verified all multiple-choice options are present
 - Checked for duplicate questions
- **Length Constraints:**
 - Truncated extremely long explanations (>512 tokens)
 - Ensured all samples fit within model's context window

5.2.3 Conversational Format Conversion

<Team Member 3>

Content Hint: Detail the critical transformation from Q&A to conversational format:

Conversion Process:

1. System Prompt Addition:

"You are a helpful science tutor for elementary through high school students.
Explain concepts clearly with examples."

2. Question Formatting:

- Original: Separate question text and choices array
- Converted: Unified format with labeled choices (A, B, C, D)
- Example:

What is photosynthesis?

A. The process plants use to make food
B. The process of cell division
C. The process of breathing
D. The process of digestion

3. Answer Construction:

- Structure: Correct answer + Explanation + Lecture (if available)
- Template:

The correct answer is [Letter]. [Answer text].

Explanation: [Detailed explanation from dataset]

Background: [Lecture content if available]

4. Message Structure:

```
python
{
  "messages": [
    {"role": "system", "content": "..."},  

    {"role": "user", "content": "[formatted question]"},  

    {"role": "assistant", "content": "[formatted answer]"}
  ]
}
```

Rationale: This format teaches the model to respond in a structured, educational manner rather than simply predicting next tokens in isolation.

5.3 Tokenization Process

<Team Member 3>

Content Hint: Explain how text is converted to model inputs:

- **Tokenizer:** Nanochat's custom BPE (Byte-Pair Encoding) tokenizer
 - Vocabulary size: 65,536 tokens
 - Implementation: Rust-based for efficiency
- **Tokenization Steps:**
 1. Concatenate all messages (system + user + assistant) with special tokens
 2. Convert text to token IDs using BPE algorithm
 3. Add special tokens: `<start>`, `<end>`, `<pad>`
 4. Truncate sequences exceeding max length (512 tokens)
 5. Pad shorter sequences to uniform length within batch
- **Attention Masks:** Create binary masks to indicate real vs. padded tokens

5.4 Model Fine-Tuning Configuration

<Team Member 3>

Content Hint: Provide detailed hyperparameters and training setup:

Training Hyperparameters:

- **Learning Rate:** 5e-5
 - Justification: Lower than pre-training (typical: 1e-4) to avoid catastrophic forgetting
- **Batch Size:**
 - Per-device: 2 (limited by 16GB GPU memory)
 - Gradient accumulation steps: 8
 - Effective batch size: 16
- **Epochs:** 2-3
 - Justification: More epochs risk overfitting on small dataset
- **Optimizer:** AdamW
 - Weight decay: 0.01
 - Betas: (0.9, 0.999)
- **Learning Rate Schedule:** Linear warmup (100 steps) + cosine decay
- **Max Sequence Length:** 512 tokens

Training Optimizations:

- **Mixed Precision (FP16):** Enabled
- **Gradient Checkpointing:** Enabled (saves ~30% memory)
- **Gradient Clipping:** Max norm = 1.0 (prevents exploding gradients)

Training Monitoring:

- **Logging Frequency:** Every 50 steps
- **Evaluation Frequency:** Every 200 steps
- **Checkpoint Saving:** Best model (lowest validation loss) + every 200 steps
- **Early Stopping:** Monitor validation loss, stop if no improvement for 3 evaluations

5.5 Training Procedure

<Team Member 3>

Content Hint: Describe the actual training process:

Training Loop:

1. **Initialization:** Load pre-trained weights from [sdobson/nanochat](#)
2. **Data Loading:** Create DataLoader with batching and shuffling
3. **Forward Pass:**
 - Input: Tokenized conversation sequences
 - Output: Next-token predictions for each position
4. **Loss Computation:**
 - Causal language modeling loss (cross-entropy)
 - Only compute loss on assistant response tokens (ignore system/user tokens)
5. **Backward Pass:** Compute gradients via automatic differentiation
6. **Optimization Step:** Update parameters every 8 accumulation steps
7. **Validation:** Evaluate on held-out validation set every 200 steps
8. **Logging:** Record training loss, validation loss, learning rate

Training Duration:

- **Total Time:** ~3-4 hours on Kaggle P100/T4 GPU
- **Steps per Epoch:** ~[calculate: 3000 samples / effective batch 16] = ~188 steps
- **Total Steps:** ~376-564 steps (2-3 epochs)

Convergence Monitoring:

- Track training vs. validation loss to detect overfitting
- Expected pattern: Both losses decrease initially, training continues decreasing while validation plateaus (indicating model is learning)

6. Results

<Team Member 4>

6.1 Quantitative Evaluation

<Team Member 4>

6.1.1 Evaluation Metrics

<Team Member 4>

Content Hint: Define all metrics used:

Primary Metrics:

1. Answer Accuracy:

- Definition: Percentage of test questions where model's response contains the correct answer choice (A, B, C, or D)
- Measurement: Extract answer letter from first 100 characters of generated response, compare to ground truth
- Baseline (Pre-trained): [X%]
- Fine-tuned: [Y%]
- **Improvement:** [Y-X]%

2. Explanation Quality Score (Manual):

- Rubric (0-3 scale):
 - 0: No explanation or incorrect
 - 1: Minimal explanation without examples
 - 2: Clear explanation with some detail
 - 3: Comprehensive explanation with examples and background
- Sample size: 50 randomly selected test questions
- Average score increase: [pre-trained] → [fine-tuned]

3. Response Length:

- Metric: Average number of tokens in generated responses
- Purpose: Ensure model generates detailed explanations, not just short answers
- Target: 100-200 tokens (detailed but not excessively verbose)
- Results: [pre-trained avg] → [fine-tuned avg] tokens

Secondary Metrics:

4. Perplexity:

- Definition: Measure of how "surprised" the model is by the test data (lower = better)
- Formula: $\exp(\text{average cross-entropy loss})$
- Results: [pre-trained] → [fine-tuned]

5. Training Loss Convergence:

- Final training loss: [value]
- Final validation loss: [value]
- Demonstrates: Effective learning without overfitting

6.1.2 Performance Results

<Team Member 4>

Content Hint: Present concrete numbers and comparisons:

[Insert: Results Table]

Metric	Base Model	Fine-tuned Model	Improvement
Answer Accuracy	X%	Y%	+Z%
Avg Explanation Score	A	B	+C
Avg Response Length	M tokens	N tokens	+(N-M)
Perplexity	P	Q	-(P-Q)

Key Findings:

- Fine-tuned model shows **[X%]** improvement in answer accuracy, demonstrating successful domain adaptation
- Explanation quality improved by **[Y points]**, indicating model learned pedagogical response style from ScienceQA's detailed explanations
- Response length increased appropriately, showing model generates more comprehensive answers rather than terse replies
- Perplexity reduction of **[Z]** indicates better fit to science education domain

Performance by Subject Area: *[If you have time to compute this]*

- Biology: **[X%]** accuracy
- Chemistry: **[Y%]** accuracy
- Physics: **[Z%]** accuracy
- Earth Science: **[W%]** accuracy
- Analysis: **[Note any subject-specific strengths/weaknesses]**

6.2 Qualitative Evaluation

<Team Member 4>

6.2.1 Side-by-Side Response Comparison

<Team Member 4>

Content Hint: Provide 2-3 detailed examples showing base vs. fine-tuned responses:

Example 1: Elementary-Level Question

Question:

What is the main source of energy for plants?

- A. Water
- B. Sunlight
- C. Soil
- D. Air

Base Model Response: [Include actual generated text - typically short, may lack explanation or be off-topic]

Fine-Tuned Model Response: [Include actual generated text - should show clear answer + explanation + example]

Analysis:

- Fine-tuned model provides [specific improvements: e.g., "identifies correct answer explicitly", "includes explanation of photosynthesis process", "uses grade-appropriate language"]
- Demonstrates pedagogical structure learned from training data

Example 2: Advanced High School Question

Question:

During cellular respiration, what is the primary role of the electron transport chain?

- A. Break down glucose into pyruvate
- B. Generate ATP through oxidative phosphorylation
- C. Produce carbon dioxide
- D. Convert NADH to NAD⁺

Base Model Response: [Include actual text]

Fine-Tuned Model Response: [Include actual text]

Analysis:

- [Discuss how fine-tuned model handles more complex scientific concepts]
- [Note if it provides additional context about cellular respiration process]

Example 3: Question Requiring Reasoning

Question:

A ball is thrown upward. What forces act on it at the highest point of its trajectory?

- A. Only gravity
- B. Only air resistance
- C. Gravity and air resistance
- D. No forces act on it

Comparison Analysis: [Discuss how fine-tuned model's reasoning differs from base model]

6.2.2 Error Analysis

<Team Member 4>

Content Hint: Identify and categorize failure modes:

Common Errors Observed:

1. Incorrect Answer Selection:

- Frequency: [X out of 500 test samples]
- Pattern: [e.g., "Confusion on advanced physics concepts", "Misunderstanding of question phrasing"]
- Example case: [Describe specific question where model failed]

2. Incomplete Explanations:

- Frequency: [Y samples]
- Pattern: [e.g., "Provides correct answer but lacks detailed reasoning"]
- Note: Less frequent than base model

3. Off-Topic Responses:

- Frequency: [Z samples]
- Pattern: [e.g., "Occasionally begins answering a different question"]
- Likely cause: Model uncertainty or ambiguous question phrasing

Comparison to Base Model:

- Base model error rate: [%]
- Fine-tuned model error rate: [%]
- Most significant improvement: [Category of errors that decreased most]

6.3 Computational Performance

<Team Member 4>

Content Hint: Report practical performance metrics:

Training Metrics:

- Total training time: [X] hours
- GPU utilization: [average %]
- Memory usage: [peak GB]
- Cost: \$0 (free tier)
- Samples per second: [training throughput]

Inference Metrics:

- Average response time: [X] seconds per query (CPU)
- GPU response time: [Y] seconds per query
- Throughput: [Z] queries per minute
- Practical usability: Suitable for interactive chatbot use

Comparison to Baseline:

- Fine-tuned model has [negligible/minimal] inference overhead compared to base model
 - Demonstrates fine-tuning doesn't significantly impact deployment efficiency
-

7. Discussion

<Team Member 4>

7.1 Interpretation of Results

<Team Member 4>

Content Hint: Provide thoughtful analysis of what the results mean:

Success Indicators:

- **Significant Accuracy Improvement:** The [X%] gain in answer accuracy demonstrates that supervised fine-tuning on domain-specific data successfully adapted the general-purpose language model to science education
- **Enhanced Pedagogical Quality:** Qualitative analysis shows fine-tuned model learned to structure responses educationally (answer → explanation → example), mimicking the teaching style present in ScienceQA dataset
- **Efficient Transfer Learning:** Achieved strong results with only 3k training samples, validating the power of pre-trained language understanding for domain adaptation
- **Cost-Effective Implementation:** Entire project completed on free computing resources proves feasibility for resource-constrained educational settings

Limitations and Boundaries:

- **Domain Scope:** Model specifically tuned for multiple-choice science questions; may not generalize well to open-ended science discussions or other subjects
- **Grade Level Performance:** [If applicable] Model shows better performance on [elementary/middle/high school] questions, possibly due to dataset distribution
- **Explanation Depth:** While improved, some explanations still lack the depth and intuitive examples that expert human tutors provide
- **Factual Accuracy:** Occasional hallucinations or incorrect reasoning persist (though reduced from base model)

7.2 Challenges Encountered

<Team Member 4>

Content Hint: Honestly discuss obstacles and how you overcame them:

Technical Challenges:

1. Data Format Conversion:

- Challenge: ScienceQA's Q&A structure needed transformation into conversational format
- Solution: Developed custom preprocessing pipeline to combine questions, answers, explanations, and lectures into coherent conversational turns
- Learning: Importance of data format alignment between pre-training and fine-tuning objectives

2. GPU Memory Constraints:

- Challenge: 561M parameter model + batch size limitations on 16GB GPU
- Solution: Implemented gradient accumulation (8 steps) and mixed precision training (FP16)
- Trade-off: Longer training time but successfully fit within free-tier resources

3. Evaluation Methodology:

- Challenge: Difficult to automatically assess explanation quality (unlike simple classification accuracy)
- Solution: Combined quantitative metrics (accuracy, perplexity) with manual qualitative evaluation of sample responses
- Learning: Multi-faceted evaluation essential for generative tasks

Logistical Challenges:

4. Platform Session Timeouts:

- Challenge: Kaggle notebooks disconnect after inactivity; long training runs risk interruption
- Solution: Implemented frequent checkpointing; trained during periods of active monitoring
- Prevention: Saved intermediate checkpoints to persistent storage

5. Team Coordination:

- Challenge: Dividing work on interconnected NLP pipeline
- Solution: Clear module boundaries (data prep | training | evaluation | demo), regular sync meetings
- Tool: Git for version control and collaboration

7.3 Risks and Mitigation Strategies

<Team Member 4>

Content Hint: Discuss potential risks that could have derailed the project:

Risk 1: Insufficient Training Data

- Concern: 3k samples might be too small for effective fine-tuning
- Mitigation: Pre-trained model's strong foundation made small-sample fine-tuning viable; validated with baseline comparisons
- Result: Risk did not materialize; 3k samples proved sufficient

Risk 2: Overfitting

- Concern: Small dataset could lead to memorization rather than learning
- Mitigation: Used only 2-3 epochs, monitored validation loss, early stopping if divergence detected
- Result: [Training/validation loss curves showed appropriate generalization]

Risk 3: Compute Resource Exhaustion

- Concern: Exceeding Kaggle's 30 hrs/week GPU quota
- Mitigation: Optimized training config for efficiency, planned experiments carefully, had backup Colab Pro option
- Result: Completed within free-tier limits

Risk 4: Poor Model Performance

- Concern: Fine-tuning might not significantly improve over base model
- Mitigation: Chose high-quality dataset (ScienceQA), validated approach through literature review
- Result: Achieved measurable improvements, validating approach

7.4 Lessons Learned

<Team Member 4>

Content Hint: Reflect on key takeaways from the project:

Technical Insights:

1. **Pre-training is Powerful:** Even a 561M parameter model pre-trained on general text provides strong baseline that can be effectively specialized
2. **Data Quality > Quantity:** 3k high-quality Q&A pairs with explanations proved more valuable than potentially larger but lower-quality datasets
3. **Conversational Format Matters:** Structuring training data as conversations (system/user/assistant) effectively teaches response style and format
4. **Evaluation Complexity:** Assessing generative models requires combining automated metrics with human evaluation

Practical Insights:

1. **Free Resources are Viable:** Kaggle free tier + HuggingFace hub enabled complete project at zero cost
2. **Time Management:** Training takes hours; plan for experimentation and iteration in project timeline
3. **Documentation is Critical:** Jupyter notebooks with markdown explanations make reproduction and collaboration much easier
4. **Incremental Development:** Start with small subset (e.g., 500 samples) to validate pipeline before full training

7.5 Future Work and Improvements

<Team Member 4>

Content Hint: Propose concrete next steps:

Short-Term Improvements:

- 1. Expand Training Data:** Increase to 10k-15k samples for potentially better performance
- 2. Multi-Turn Conversations:** Extend to handle follow-up questions and clarifications
- 3. Subject-Specific Models:** Train separate models for Biology, Chemistry, Physics for deeper specialization
- 4. Enhanced Evaluation:** Develop automated explanation quality metrics using reference-based scoring

Medium-Term Extensions:

- 1. Adaptive Difficulty:** Implement mechanisms to adjust explanation complexity based on student grade level or prior questions
- 2. Quiz Generation:** Extend model to generate practice questions, not just answer them
- 3. Multimodal Integration:** Incorporate image understanding for diagrams, graphs, experimental setups (using vision encoders)
- 4. Error Correction Learning:** Fine-tune further on students' common misconceptions

Long-Term Vision:

- 1. Deployment:** Host permanently on HuggingFace Spaces or similar platform for actual student use
 - 2. User Feedback Loop:** Collect student ratings on explanation quality, retrain periodically
 - 3. Curriculum Alignment:** Tailor model to specific curriculum standards (e.g., NGSS)
 - 4. Personalization:** Adapt to individual student learning styles and knowledge gaps
-

8. Conclusion

<Team Member 1>

8.1 Project Summary

<Team Member 1>

Content Hint: Synthesize the entire project in 1-2 paragraphs:

This project successfully demonstrated the adaptation of a general-purpose pre-trained language model (nanochat, 561M parameters) into a specialized science teaching assistant through supervised fine-tuning on the ScienceQA dataset. By fine-tuning on approximately 3,000 carefully formatted question-answer pairs with detailed explanations, we achieved [X%] improvement in answer accuracy and significant enhancement in explanation quality compared to the base model. The complete pipeline—from dataset preprocessing through model fine-tuning to interactive demo deployment—was implemented using only free-tier computing resources (Kaggle notebooks), demonstrating the accessibility of modern NLP techniques for educational applications.

The project addressed real challenges in science education (availability of patient, knowledgeable tutoring) while providing hands-on experience with the complete machine learning workflow: data preparation, model architecture understanding, transfer learning, evaluation methodology, and deployment. Key technical contributions include effective data format conversion from Q&A to conversational structure, optimization strategies for resource-constrained training, and multi-faceted evaluation combining quantitative metrics with qualitative assessment. The resulting model serves as both a functional prototype for educational AI and a learning artifact demonstrating core NLP concepts taught in graduate coursework.

8.2 Impact and Significance

<Team Member 1>

Content Hint: Emphasize broader implications:

Educational Impact:

- Demonstrates feasibility of custom AI tutors for specific subjects and grade levels
- Provides template for educators and institutions to build similar tools without expensive infrastructure
- Shows that effective domain adaptation requires relatively small amounts of high-quality training data

Technical Contributions:

- Validates transfer learning approach for educational NLP applications
- Documents practical implementation details (hyperparameters, compute requirements, troubleshooting) valuable for future projects
- Provides open-source example of end-to-end LLM fine-tuning pipeline

Course Learning Objectives Met:

- ✓ Implemented advanced NLP model (transformer-based LLM)
- ✓ Integrated multiple NLP concepts (tokenization, attention mechanisms, language modeling, fine-tuning)
- ✓ Used real-world dataset (50k+ sentences equivalent in ScienceQA)
- ✓ Evaluated using multiple metrics (accuracy, perplexity, qualitative analysis)
- ✓ Demonstrated model architecture and pipeline design
- ✓ Compared base vs. fine-tuned model performance

8.3 Final Reflections

<Team Member 1>

Content Hint: Personal/team reflection on the project experience:

This project reinforced the power and accessibility of modern NLP tools while highlighting the importance of careful data preparation and thoughtful evaluation. The hands-on experience of adapting a sophisticated language model to a specific domain provided deeper understanding of concepts like transfer learning, fine-tuning dynamics, and the interplay between model capacity and training data quality. The constraints we faced—limited compute resources, small training set, tight timeline—ultimately strengthened our problem-solving skills and forced creative solutions (gradient accumulation, efficient data sampling, multi-stage evaluation).

Working with nanochat's minimalist, educational-focused implementation offered invaluable insight into LLM internals that would be obscured in production frameworks. The project succeeded in its dual goals: creating a functional science teaching assistant prototype and deepening our understanding of the full NLP pipeline from pre-training concepts through deployment. The zero-cost implementation proves that cutting-edge NLP capabilities are increasingly accessible, democratizing AI development for educational and research purposes.

9. References

<All Team Members>

Content Hint: Cite all sources in IEEE format. Include at minimum:

Primary Sources:

1. Andrej Karpathy, "Nanochat: A Minimal LLM Implementation," GitHub repository, 2025. [Online]. Available: <https://github.com/karpathy/nanochat>
2. P. Lu et al., "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.
3. ScienceQA Dataset, HuggingFace Datasets, 2022. [Online]. Available: <https://huggingface.co/datasets/derek-thomas/ScienceQA>

Technical Background: 4. A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. 5. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019. 6. T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

Tools and Libraries: 7. A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019. 8. T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," *Proceedings of EMNLP: System Demonstrations*, 2020. 9. Gradio: Open-Source Python Library for Machine Learning Demos, 2021. [Online]. Available: <https://gradio.app>

Fine-Tuning and Transfer Learning: 10. J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *Proceedings of ACL*, 2018. 11. N. Houlsby et al., "Parameter-Efficient Transfer Learning for NLP," *Proceedings of ICML*, 2019.

Educational AI: 12. [Add any papers on educational AI/tutoring systems you reference] 13. [Add domain-specific references relevant to science education]

Appendix A: Code Repository

GitHub Repository: [Insert link to your project repository]

Repository Contents:

- `/notebooks`: Jupyter notebooks for data preprocessing, training, evaluation, demo
 - `/data`: Dataset preparation scripts and sample data
 - `/models`: Model configuration files and saved checkpoints
 - `/evaluation`: Evaluation scripts and results
 - `README.md`: Setup instructions and project overview
 - `requirements.txt`: Python dependencies
-

Appendix B: Demo Access

Interactive Demo: [Insert link to Gradio demo if deployed]

Demo Instructions:

1. Navigate to [demo URL]
2. Enter a science question (optionally in multiple-choice format)
3. Click "Submit" to receive AI-generated answer and explanation
4. Explore example questions provided in the interface

Demo Video: [Optional: Link to screen recording demonstrating the teaching assistant]

Team Member Assignments Summary:

- **Team Member 1:** Abstract, Introduction (all subsections), Problem Statement (all subsections), Conclusion (all subsections)
- **Team Member 2:** Project Architecture (all subsections), Tools and Technologies (all subsections)
- **Team Member 3:** Methodology (all subsections)
- **Team Member 4:** Results (all subsections), Discussion (all subsections)

Total Sections per Member: Approximately 8-10 subsections each, balancing technical depth and writing load.

End of Report Skeleton