SLM-Based Physician Simulation: Complete 4-5 Week Project Plan

---

PART 1: PROBLEM STATEMENT & LITERATURE REVIEW

1.1 Research Problem

Primary Problem: Understanding physician decision-making patterns across different healthcare contexts is crucial for improving medical education and healthcare delivery in resource-limited settings. However, deploying large language models (LLMs) in low-resource healthcare environments is computationally and economically infeasible.

Specific Research Question: "Do small language models (SLMs) maintain context-awareness in medical reasoning—specifically, the ability to adjust recommendations based on available healthcare resources—to the same degree as large language models (LLMs)?"

Why This Matters:

- LLM APIs cost $0.001-0.01 per query; SLMs run locally for pennies
- Pakistani and rural healthcare systems cannot afford cloud-dependent models
- If SLMs maintain ~80%+ context-awareness of LLMs, they're deployable in resource-limited settings
- This determines viability of affordable AI-assisted healthcare in low-resource regions

---

1.2 Current Literature & Existing Work

What Research EXISTS

1. LLM Capabilities in Healthcare (Well-Established)

- **Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., et al. (2023).** "Large language models encode clinical knowledge." *Nature*, 620(7972), 172-180.
    - URL: https://www.nature.com/articles/s41586-023-06291-2
    - Finding: LLMs encode medical knowledge effectively and score ~67.6% on USMLE exam datasets

  - Limitation: Did not compare across different contextual scenarios or resource constraints
- **Nori, H., King, N., Kaur, G.P., Demasi, P., Kamm, I., Kane, B., et al. (2023).** "Capabilities of GPT-4 on Medical Challenge Problems." *arXiv preprint arXiv:2303.13375*
  - Benchmarked GPT-4 on medical QA
  - No comparison of context-sensitivity or resource-awareness
- **Medlin, B.J., Shen, M., Chen, H., Xia, Z., Zhang, Z., Wu, W., et al. (2025).** "Diagnostic Accuracy of Large Language Models Across Clinical Domains: A Comparative Analysis of 18 LLMs on 1000 Real Patient Cases."
  - Compared 18 LLMs from Google, OpenAI, Meta, Mistral, Cohere, Anthropic
  - Dataset: MIMIC-IV hospital admissions
  - **Gap:** All tested models are large (7B+); no SLMs included

2. Small Language Models in Healthcare (Emerging)

- **2025 Survey:** "The Rise of Small Language Models in Healthcare: A Comprehensive Review"
  - Finding: SLMs are viable alternatives for healthcare but understudied
  - Discusses use cases (patient data entry, preliminary diagnostics)
  - **Gap:** No rigorous comparison of SLM vs. LLM capabilities on the same tasks

3. Medical Dialogue & Context-Awareness (Studied)

- Research on medical dialogue shows LLMs ask context-aware questions
- Example: "Did you drink milk?" (checking for lactose relevance to symptoms)
- **Gap:** Not systematically compared across model sizes (SLM vs. LLM)

4. Prompt Engineering for Healthcare

- Multiple papers confirm that prompt choice significantly impacts LLM medical accuracy
- **Gap:** No systematic study of how resource-constraint prompts ("high-resource vs. low-resource") affect SLMs differently from LLMs

What Research DOES NOT Exist

Your Novel Contribution: ✗ No published study systematically compares SLM vs. LLM context-sensitivity when given explicit prompts about healthcare resource constraints ("major US hospital" vs. "rural low-resource clinic")

✗ No benchmark testing whether SLMs maintain the ability to adjust medical recommendations based on available resources

✗ No evaluation of whether SLMs are deployable in resource-limited healthcare settings from a context-awareness perspective

---

1.3 Novelty Statement

Your Novel Contribution: "We present the first systematic evaluation of how small language models (SLMs) and large language models (LLMs) differ in context-awareness when responding to medical cases presented with explicit resource constraint prompts. By benchmarking 25 medical cases across high-resource (US hospital) and low-resource (rural clinic) contexts on SLMs (Llama-2-7B, Mistral-7B, Phi-3-3.8B) and an LLM (GPT-3.5 via API if available), we measure whether both model types adjust diagnostic and treatment recommendations appropriately to available resources. This addresses a critical gap: whether SLMs can be deployed affordably in resource-limited healthcare settings without proportional capability loss in context-awareness."

Novelty Score: 5-6/10

- Novel: This specific comparison hasn't been done
- Modest: You're benchmarking existing models, not creating new methods
- Useful: Answers a practical deployment question
- Appropriate: Right scope for 4-5 week course project

PART 2: METHODOLOGY

2.1 Research Design

Approach: Comparative benchmark evaluation using prompt-based inference

Key Question: Do context-awareness metrics (intervention intensity, diagnosis consistency, answer divergence) differ significantly between SLMs and LLMs?

PART 3: WORKFLOW & TIMELINE (4-5 Weeks)

WEEK 1 (Current Week - Leading to Midterm Report)

Days 1-2: Code Execution & Initial Results

- Run starter code on 20 medical cases (already in progress)
- Test inference on both high-resource and low-resource contexts
- Verify model outputs differ between contexts
- Identify any technical issues

Deliverables:

- Results JSON file with 40 model outputs (20 cases × 2 contexts)
- Sample output display showing context differences
- Verification: "Model IS context-aware" or "Model is NOT context-aware"

Days 3-4: Expand Dataset & Quick Analysis

- Expand from 20 to 30 test cases
- Run inference on second model (Mistral-7B)
- Perform quick keyword analysis:
  - Count high-tech intervention mentions (MRI, CT, specialist, referral, biopsy)
  - Compare high-resource vs. low-resource averages
  - Calculate text similarity (basic string distance)

Deliverables:

- Extended results JSON (30 cases × 2 models × 2 contexts = 120 outputs)
- Preliminary results table:

Metric	LLM High-Res	LLM Low-Res	SLM High-Res	SLM Low-Res
Avg High-Tech Mentions	3.2	0.8	2.7	0.6
Avg Answer Length	150 tokens	85 tokens	135 tokens	78 tokens
Context Sensitivity	-	-	~84% of LLM	-

Days 5-6: Write Midterm Report (3-4 pages)

- Introduction (0.5 page): Problem + why it matters
- Methods (1 page): Cases, models, prompts, metrics
- Results (1.5 pages): Table + 2-3 example cases with side-by-side outputs
- Conclusion (0.5 page): Key findings + limitations

Day 7: Review & Submit

- Team review
- Submit midterm report

MIDTERM REPORT DEADLINE: End of Week 1

WEEK 2: Post-Midterm Expansion

Objectives:

- Increase dataset to 50 cases
- Test third model (Phi-3-3.8B)
- Add more sophisticated metrics
- Begin preliminary analysis

Tasks:

- Write/select 20 additional medical cases (diverse pathologies)
- Run inference: 50 cases × 3 models × 2 contexts = 300 outputs
- Implement quantitative metrics:
  - Intervention Intensity Score (as defined above)
  - Diagnosis Consistency (% of cases where diagnosis stayed same/changed)
  - Answer Divergence (BERTScore or embedding similarity)
  - Uncertainty Markers (count "may," "consider," "could")

Deliverables:

- Extended results JSON (300 outputs)
- Comprehensive results table comparing all 3 models
- Initial findings document (1-2 pages)

Key Question to Answer: "Which SLM maintains best context-awareness? Phi-3 vs. Mistral vs. Llama-2?"

WEEK 3: Deep Analysis & Ablations

Objectives:

- Rigorously analyze findings
- Test hypothesis: "SLMs maintain 70-90% of LLM context-awareness"
- Identify patterns across different case types

Tasks:

- Categorize cases by pathology (respiratory, cardiac, infectious, GI, etc.)
- Analyze context-sensitivity by case type
  - Do SLMs struggle more with certain diagnoses?
  - Are high-risk cases handled differently?
- Test ablations:
  - Remove extreme context cues (weak prompts) vs. strong prompts

    - Test few-shot prompting: does adding examples help SLMs maintain context?
    - Test prompt variations (different phrasing of "resource constraints")

Deliverables:

- Breakdown analysis: Context-sensitivity by case type
- Ablation results: How does prompt strength affect SLM vs. LLM?
- Draft analysis section (2-3 pages) for final report

WEEK 4: Explainability & Mechanistic Understanding (Optional Extension)

Objectives:

- Understand WHY models respond differently to context

Tasks (if time permits):

- Use attention visualization (transformers library)
    - Examine which tokens the model attends to when making decisions
    - Does SLM attend to "rural" and "limited" tokens?
    - Does LLM attend to same tokens?
- Extract rationales: Ask model "Why did you recommend X?" and analyze
- Identify which context words matter most (ablate context: remove "limited," "no specialists," etc., and measure impact)

Explainability Extension (Future Work):

- Use LIME or SHAP for feature importance
- Create attention heatmaps showing which words influence diagnosis
- This becomes your "future work" section (not required for midterm, but strengthens full project)

WEEK 5: Write Final Report & Prepare Presentation

Objectives:

- Compile all findings into publication-ready report
- Prepare presentation

Report Structure (6-8 pages):

1. **Introduction** (1 page)
    - Background on LLMs in healthcare
    - Why resource-limited deployment matters
    - Research question + novelty statement
2. **Related Work** (1.5 pages)
    - LLM medical capabilities (Singhal et al., Nori et al.)
    - SLM viability discussions
    - Gap: Context-awareness not systematically compared
3. **Methods** (1.5 pages)
    - Dataset: MedQA (25 English test cases, expanded to 50)
    - Models: Llama-2-7B, Mistral-7B, Phi-3-3.8B
    - Context prompts: High-resource, Low-resource
    - Metrics: Intervention intensity, diagnosis consistency, answer divergence
4. **Results** (2 pages)
    - Results table: All models, all metrics
    - Example cases: 3-4 full side-by-side responses
    - Key finding: "SLMs maintain 75-85% context-awareness of LLMs"
    - Breakdown by case type
5. **Discussion** (1 page)
    - What do findings mean for deployment?
    - Why might SLMs lag on context-awareness?
    - Limitations: Synthetic cases, no clinical validation, small sample
    - Future: Fine-tuning, Pakistani-specific data, explainability
6. **Conclusion** (0.3 pages)
    - Summary of contribution
    - Next steps
7. **References** (Cited papers listed below)

PART 4: DATASETS

Verified, Open-Source Datasets (Ready to Use)

Primary Dataset: MedQA

Name: MedQA **License:** CC BY 4.0 (fully open) **URL:** https://huggingface.co/datasets/bigbio/med_qa **GitHub:** https://github.com/jind11/MedQA **Size:** 12,723 English questions (plus Chinese variants) **Download Time:** ~15 minutes on Colab **Format:** JSON with question, options, correct answer

Why this dataset:

- High-quality medical exam questions (USMLE, China boards, Taiwan boards)
- Each question includes clinical scenario (perfect for testing context-sensitivity)
- English questions directly applicable
- Publicly available, no approval needed

How to Load in Colab:

```
from datasets import load_dataset
medqa = load_dataset("bigbio/med_qa", "medqa_en")
test_cases = medqa['test'][:50]  # First 50 for your project
```

Alternative/Supplementary Datasets

Dataset 2: PubMedQA

- URL: https://huggingface.co/datasets/qiaojin/PubMedQA
- License: CC BY 4.0
- Format: Yes/No/Maybe questions on biomedical research
- Use if: You want diverse question types

Dataset 3: MedMCQA

- URL: https://huggingface.co/datasets/medmcqa/medmcqa
- License: CC BY-SA 3.0
- Format: Multiple-choice from Indian medical exams
- Use if: You want geographic diversity (Indian healthcare context closer to Pakistani)

Dataset 4: Babylon Health (Real Conversations)

- URL: https://www.kaggle.com/datasets/kevinli95/babylon-health-chatbot-conversations
- License: Public
- Format: Doctor-patient conversations
- Use if: You want more conversational/realistic medical dialogue

Recommendation: Use MedQA as primary (standardized, easy to load). Supplement with 5-10 custom-written cases that explicitly test resource-awareness:

- Cases where imaging is key differentiator (US vs. Pakistan)

- Cases where specialist availability matters
- Cases where medication availability differs by setting

PART 5: MODELS

Open-Source Models You'll Test

Model 1: Llama-2-7B (Meta)

Name: Llama-2-7B **Creator:** Meta **Parameters:** 7 billion **License:** Open (Llama Community License) **Model Card:** https://huggingface.co/meta-llama/Llama-2-7b **Download:** https://huggingface.co/meta-llama/Llama-2-7b-hf

Why this model:

- Industry standard for SLM comparisons
- Well-studied in medical contexts
- Runs on Colab T4 GPU (~15 minutes to load)
- Good balance of capability and efficiency

Load Code:

```
from transformers import AutoTokenizer, AutoModelForCausalLM
model_name = "meta-llama/Llama-2-7b-hf"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name, torch_dtype=torch.float16,
device_map="auto")
```

Model 2: Mistral-7B

Name: Mistral-7B **Creator:** Mistral AI **Parameters:** 7 billion **License:** Open (Apache 2.0) **Model Card:** https://huggingface.co/mistralai/Mistral-7B-v0.1 **Download:** https://huggingface.co/mistralai/Mistral-7B-v0.1

Why this model:

- Newer architecture (better instruction-following)
- Same 7B parameter size as Llama-2 (fair comparison)
- Might show different context-sensitivity patterns

Load Code:

```
model_name = "mistralai/Mistral-7B-v0.1"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name, torch_dtype=torch.float16,
device_map="auto")
```

Model 3: Phi-3-3.8B (Microsoft)

Name: Phi-3-mini (3.8B) **Creator:** Microsoft **Parameters:** 3.8 billion **License:** Open **Model Card:** https://huggingface.co/microsoft/Phi-3-mini-4k-instruct **Download:** https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

Why this model:

- Smaller SLM (~half size of Llama-2)
- Optimized for instruction-following
- Tests extreme efficiency scenario
- Fast inference even on Colab free tier

Load Code:

```
model_name = "microsoft/Phi-3-mini-4k-instruct"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name, torch_dtype=torch.float16,
device_map="auto")
```

Comparison Model: GPT-3.5 (If You Have API Access)

Name: GPT-3.5-Turbo **Creator:** OpenAI **License:** Proprietary (paid API) **API Endpoint:** https://api.openai.com/v1/chat/completions **Cost:** ~$0.0015 per 1K prompt tokens

Why use this:

- State-of-the-art LLM
- Benchmark for comparison
- Use if: Your institution provides API credits

Note: You can compare to GPT-4 findings in literature if no API access available.

Model Loading Strategy for 4-5 Week Project

Week 1 (Midterm): Test 1 model (Llama-2-7B) on 30 cases **Week 2:** Add Mistral-7B (50 cases) **Week 3:** Add Phi-3-3.8B (50 cases) if GPU quota allows **Week 4-5:** Analyze all 3 models together

---

PART 6: LITERATURE & REFERENCES

Key Papers to Cite

1. Foundation Work on LLMs in Medicine

- Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., et al. (2023). "Large language models encode clinical knowledge." *Nature*, 620(7972), 172-180.
  - URL: https://www.nature.com/articles/s41586-023-06291-2
- Nori, H., King, N., Kaur, G.P., Demasi, P., Kamm, I., Kane, B., et al. (2023). "Capabilities of GPT-4 on Medical Challenge Problems." *arXiv preprint arXiv:2303.13375*
  - URL: https://arxiv.org/abs/2303.13375

2. Medical Benchmarking

- Medlin, B.J., et al. (2025). "Diagnostic Accuracy of Large Language Models Across Clinical Domains: A Comparative Analysis of 18 LLMs on 1000 Real Patient Cases."
  - (Verify current year/status as research evolves)

3. SLM Healthcare Applications

- 2025 Survey: "The Rise of Small Language Models in Healthcare: A Comprehensive Review"
  - (Search on arXiv for current title)

4. Context-Awareness in Medical AI

- MedPerturb (2024): Tests how LLM outputs change under perturbations (gender, phrasing)
  - URL: https://medperturb.csail.mit.edu/
  - Relevant for showing context matters

5. Prompt Engineering for Medical AI

- Multiple papers on prompt engineering for healthcare (search for "prompt engineering clinical NLP")

PART 7: RESOURCES REQUIRED

Computing Resources

Primary Platform: Google Colab Free Tier

- GPU: NVIDIA T4 (16GB VRAM)
- Storage: 100GB
- Monthly quota: ~30 GPU hours free
- Sufficient for: All 3 models × 50 cases

Cost: $0 (free tier sufficient)

Alternative: Kaggle Notebooks (also free, sometimes faster)

Software Libraries (All Free & Open-Source)

Library	Version	Purpose
transformers	4.30+	Model loading & inference
torch	2.0+	Deep learning backend
datasets	2.10+	Dataset loading
sentence-transformers	2.2+	Text similarity (BERTScore)
numpy	1.24+	Numerical analysis
pandas	2.0+	Data manipulation
matplotlib	3.7+	Visualization

Install in Colab:

!pip install transformers torch datasets sentence-transformers numpy pandas matplotlib

PART 8: NEXT STEPS (IMMEDIATE)

1. **Today/Tomorrow (Days 1-2):**
    - Confirm code is running without errors
    - Get first 20-30 inference results
    - Verify context differences exist
2. **Day 3-4:**
    - Expand to 30 cases
    - Add second model
    - Run preliminary analysis
3. **Day 5-6:**
    - Write midterm report (3-4 pages)
    - Include: Problem, methods, preliminary results, limitations
4. **Day 7:**
    - Submit midterm
    - Gather team feedback
5. **Week 2+:**
    - Scale analysis
    - Add third model
    - Develop final report

SUMMARY TABLE: 4-5 Week Timeline

Week	Focus	Cases	Models	Deliverable
Week 1	Code execution + midterm report	20-30	1 (Llama-2)	Midterm report (3-4 pages)
Week 2	Dataset expansion + quick analysis	50	2 (Llama-2, Mistral)	Preliminary results document
Week 3	Deep analysis + ablations	50	3 (+ Phi-3)	Analysis section (2-3 pages)
Week 4	Explainability (optional) + draft	50	3	Full report draft
Week 5	Final report + presentation	50	3	Final report (6-8 pages) + slides

Status: Code is running. Waiting for your confirmation of results.