Project Plan
Chain-of-Thought Reasoning for LLM-Based Radiology Report Generation

Prepared for: Internship Cohort
Date: 29 July 2025

Table of Contents

1 Project Overview

This internship project aims to build a state-of-the-art radiology report generation system that enhances factual accuracy and clinical relevance through Chain-of-Thought (CoT) reasoning. By explicitly generating intermediate reasoning steps, the model can better align with human diagnostic workflows, reduce hallucinations, and provide transparent explanations.

- Key Objectives:
    - • Understand the theoretical underpinnings of CoT reasoning in Large Language Models (LLMs).
    - • Construct a reproducible end-to-end pipeline—from data ingestion to report generation.
    - • Implement multiple CoT techniques (prefix, self-consistency, RLHF) and benchmark against baselines.
    - • Deploy a privacy-compliant prototype with visual explanations for clinical stakeholders.

2 Prerequisites & Onboarding

2.1 Technical Stack

- Python 3.10 + Conda (base environment)
- PyTorch 2.3 with CUDA 12 (AMP enabled)
- Hugging Face Transformers & PEFT libraries
- Docker 24 + docker-compose
- Git / GitHub Enterprise
- Optional: JAX 0.4 for experimental TPU runs

2.2 Compliance & Data Privacy (Not Needed)

You must complete the “Human Subjects Research & HIPAA” CITI module before accessing MIMIC-CXR. Store certificates in the project SharePoint. IU X-Ray is license-free but still requires PHI scrubbing.

2.3 Compute Resources

2.4 Environment Setup Deliverable

Submit a screenshot of `nvidia-smi` and `pytest` pass for the project repo by the end of Week 0.

3 Literature & Conceptual Foundations

3.1 Core LLM Papers

- Vaswani et al., “Attention Is All You Need”, 2017.
- Ouyang et al., “Training language models to follow instructions with human feedback”, 2022.
- Wei et al., “Chain-of-Thought Prompting Elicits Reasoning”, 2022.
- Wang et al., “Self-Consistency Improves Chain-of-Thought Reasoning”, 2023.

3.2 Radiology-Specific Research

- Delbrouck et al., “Physician-Driven Radiology Report Generation with Vision-Language Models”, ICLR 2024.
- Gan et al., “RadGraph: Information Extraction from Radiology Reports”, 2021.
- Xiao et al., “Automatic Radiology Report Generation – A Survey”, MedIA 2025.

3.3 Coding Specific Research

- https://www.datacamp.com/tutorial/chain-of-thought-prompting
- https://huggingface.co/learn/llm-course/en/chapter12/1
-

3.4 Learning Deliverable

Compose a 2-page critical review summarizing strengths, weaknesses, and open questions by end of Week 2.

4 Dataset Pipeline

4.1 Data Acquisition

Request MIMIC-CXR access via PhysioNet, sign DUA, and download via AWS S3 CLI. IU X-Ray can be fetched directly from Indiana University servers.

4.2 Pre-Processing Steps

12. Convert DICOM to PNG (512×512) using dcmtk.
13. Apply CLAHE contrast normalization.
14. Strip patient identifiers from DICOM tags.
15. Link reports, isolate “Findings” + “Impression” sections.
16. Generate RadGraph annotations with official toolkit.

4.3 Data Split Strategy

Use patient-level stratified split (70/10/20). Verify no leakage via hashed MRN IDs.

4.4 Artifact Deliverables

Upload a versioned JSONL manifest to DVC remote containing the tuple: `{"image_path": "...", "report": "...", "radgraph": {...}}`.

5 Baseline Model Development

5.1 Architecture

Visual Encoder: Swin-V2-B, frozen for first 5 epochs then unfrozen with LR 1e-5.

Language Decoder: Llama-3-Instruct-7B with LoRA (rank 8, α = 32, dropout 0.05).

Fusion Mechanism: BLIP-2 Q-Former (12 layers, 64 hidden size).

5.2 Training Details

- Epochs: 10 (early-stop on val CheXbert-F1).
- Batch size: 16 (gradient accumulation ×4).
- Optimizer: AdamW, LR = 3e-5, weight decay 1e-4.
- Scheduler: Cosine with warmup 500 steps.
- Loss: Cross-entropy on report tokens (ignore reasoning for baseline).

5.3 Evaluation Metrics

- Textual overlap: BLEU-4, ROUGE-L, CIDEr.
- Clinical accuracy: CheXbert F1, RadGraph F1.
- Explainability: Average cosine similarity between Grad-CAM heatmap and Lung Mask.

6 Integrating Chain-of-Thought Reasoning

6.1 Reasoning Annotation Methods

- RadGraph Triplet Extraction → textualize as `<lesion> in <location>`.
- GPT-4o Distillation Prompt: "Describe each abnormality and image evidence in one sentence."
- Manual Verification: Radiologist validates 10% sample.

6.2 Training Strategies

- Prefix-CoT: Concatenate reasoning before report; joint supervision.
- Dual-Decoder: Separate heads; KL penalty to keep shared features.
- RLHF: Reward = 0.7 × RadGraph Alignment + 0.3 × CheXbert-Gain.

6.3 Inference Decoding

Use step-wise decoding (generate reasoning → feed as context). For self-consistency, sample K = 8 beams and pick majority vote.

7 Evaluation & Error Analysis

After training, compare the baseline and CoT models using the metrics in Section 5.3. For qualitative analysis, label at least 100 random test cases as Acceptable / Partial / Incorrect.

- Common Failure Categories:

    - • Hallucinated finding (not present in image).
    - • Missed critical finding (present but omitted).
    - • Wrong anatomical location.

8 Deployment Prototype & DevOps

- Containerize inference with NVIDIA Triton to support model parallelism.
- API contract: `POST /generate` with Base64-encoded PNG returns `{ "reasoning": […], "report": "…" }`.
- Frontend: React (Vite) + Cornerstone 3D viewer + heat-map overlay.
- On-premise deployment with VPN access to comply with PHI regulations.
- Prometheus + Grafana dashboards for latency and GPU utilisation.

9 Timeline & Milestones

17. Week 0 Prerequisites & environment setup complete.
18. Weeks 1-2 Literature review; deliver digest.
19. Weeks 3-4 Dataset pipeline ready; JSONL manifest committed.
20. Weeks 5-6 Baseline model trained & evaluated.
21. Weeks 7-9 CoT integration; ablation study finished.
22. Week 10 Error analysis & slide deck.
23. Weeks 11-12 Deployment prototype & final presentation.

10 Resources & Further Reading

- Awesome-Radiology-Report-Generation GitHub list.
- Hugging Face PEFT documentation.
- RSNA 2024 guidelines on generative-AI safety.
- Docker "Dive" tool for image inspection.

Appendices

Appendix A MIMIC-CXR Access Checklist

1. Register PhysioNet account. 2. Complete CITI "Data or Specimens Only Research" course. 3. Electronically sign the DUA. 4. Await approval (~2 business days).

Appendix B Ethical Considerations

Explainability is critical in clinical settings. Always show reasoning and heat-maps to end-users. Avoid deploying unverifiable black-box models.

Appendix C Cost Estimate

Assuming 200 GPU hours on A100 (USD 0.90/hr) and storage of 1 TB (USD 23/mo), total infrastructure cost ≈ USD 203.