Fine-Tuning Language Model on Singapore's Cybersecurity Code of Practice

Automated Cybersecurity & Risk Compliance Assessment for Critical Information Infrastructure Owners

Sagar Pratap Singh
MDAI-E Studio Term 1 Report
Industry Partner AETHER by RAiD
19 December 2025

Agenda

1. **Problem Statement**
   Singapore's CCoP 2.0 standards and mandates for CIIOs
2. **Project Objectives**
   Goals and target outcomes
3. **Benefits for CIIOs**
   Value proposition for critical infrastructure owners
4. **Related Works**
   Comparative Research and Learnings
5. **Fine Tuning Model on CCoP 2.0 Standards**
   The case for fine-tuning and methodology
6. **Benchmarks & Evaluation**
   21 benchmarks across 3 tiers
7. **Project Plan**
   Targets and timeline
8. **Updates from Phase 1**
   Establishing ground-truth and local infrastructure
9. **Phase 2 Developments**
   Evaluating baseline model performance
10. **Next Steps**
    Immediate action items

Current Focus: Phase 2 Baseline Evaluation - Ground truth validation and model benchmarking in progress

Presentation Duration: 20 minutes + Q&A

Problem Statement

WHAT IS CYBERSECURITY CODE OF PRACTICE?

CCoP 2.0 is a mandatory regulatory framework issued by the Cyber Security Agency of Singapore (CSA) that prescribes cybersecurity measures for Critical Information Infrastructure Owners (CIIOs) to safeguard systems essential to national functions.

220	11	2022
Security Requirements	Control Domains	Effective Date

CRITICAL INFORMATION INFRASTRUCTURE OWNERS (CIIOs)

Organizations designated by the Commissioner of Cybersecurity as owners of Critical Information Infrastructure (CII) necessary for continuous delivery of essential services for Singapore.

- Energy
- Water
- Healthcare
- Banking & Finance
- Transport
- Government
- Infocomm
- Media
- Security

MANDATE FOR CIIOs

- Conduct cybersecurity audits every 2 years
- Conduct annual cybersecurity risk assessments
- Comply with CCoP 2.0 - approximately 220 security requirements
- Report cybersecurity incidents within prescribed timeframes
- Notify CSA of material changes to CII systems
- Maintain detection mechanisms for cybersecurity threats

Overview of CCoP 2.0 Standards

APPLICABILITY TO DIFFERENT TECHNOLOGIES DEPLOYED IN CII ORGANIZATIONS

IT — Information Technology

Computer networks, servers, cloud platforms, databases, enterprise applications that process and store data

OT — Operational Technology

Industrial control systems, SCADA, PLCs, sensors that monitor and control physical processes in critical infrastructure

220 CLAUSES ACROSS 11 CONTROL SECTIONS

- **60%** of clauses apply to both IT and OT
- **Protect** section has the most clauses (80-90)
- **OT/ICS** is the only exclusively OT section

1. Audit Audit trails, logging, monitoring — 4	**2. Governance** Policies, roles, management oversight — 15-20	**3. Risk Management** Risk assessments, BCP, DR, cloud — 25-30
4. Asset Management Inventory, classification, lifecycle — 8-10	**5. Protect** Network, access, encryption, patching — 80-90	**6. Detect, Respond & Recover** Incident detection, forensics, recovery — 25-30
7. Cybersecurity Awareness Training, awareness, phishing prevention — 8-10	**8. Supply Chain** Vendor security, procurement — 10-12	**9. Third Party** Access controls, contractor security — 12-15
10. OT/ICS Security SCADA, Purdue Model, PLC protection — 35-40	**11. Assurance** Compliance verification, pen testing — 8-10	

Both IT & OT · IT Heavy (60%+) · OT Only

Project Objectives

Fine-tuning LLM on Singapore's CCoP 2.0 for Critical Information Infrastructure compliance

Establish Ground Truth for CCoP 2.0

Create comprehensive evaluation dataset with 118 test cases across 21 benchmarks, validated by domain experts covering all 11sections of CCoP 2.0 (220 clauses)

Benchmark Baseline Performance

Compare language models on CCoP 2.0 compliance tasks to establish baseline capabilities:

Llama-Primus-Reasoning	GPT-5	DeepSeek-V3
8B cybersecurity specialist	With web-search tool	With web-search tool

Fine-tune Model on CCoP 2.0 Standards

Train Llama-Primus-Reasoning using QLoRA (4-bit quantization) to achieve **at least 50% accuracy** in compliance violation detection across both IT and OT infrastructure

Deploy to Isolated CII Environment

Integrate fine-tuned model with CI/CD pipelines for automated detection of non-compliant source code and infrastructure configurations in air-gapped environments

Project Period: September 2025 – August 2026 | Target: Reduce CIIO compliance timeline from months to hours

Benefits of a CCoP 2.0 Trained Model for CIIOs

Transforming compliance from a months-long burden to continuous automated assurance

Time & Cost Efficiency

Reduces compliance timeline from **months to hours**. Frees security teams for strategic work instead of manual audits.

Automated Compliance Analysis

Gap analysis, preemptive code scanning, and infrastructure review with **at least 50% target accuracy** across 220 clauses.

CI/CD Integration

Catches violations **early in development**. Enables continuous compliance monitoring in deployment pipelines.

Air-Gapped Deployment

Lightweight 8B model runs **locally on-premise**. No sensitive data leaves the secure CII perimeter.

Consistency & Availability

Consistent interpretation of CCoP requirements. Available **24/7 at scale** without human bottlenecks.

Audit Readiness

Prepares evidence for **CSA audits**. Supports both IT and OT infrastructure assessments.

220	IT + OT	>50%	8B
CCoP Clauses Covered	Infrastructure Support	Target Accuracy	Lightweight Model

Related Works: Two Adaptation Strategies

FINE-TUNING — CYBERSECURITY REASONING

Domain-adapted training on curated cybersecurity corpora to improve classification and reasoning on stable technical taxonomies.

WHY FINE-TUNING?

Output space is finite and well-defined
Ground truth labels are stable over time
Evaluation uses objective metrics (accuracy, F1)

EXAMPLE: PRIMUS / Llama-Primus-Reasoning

Maps vulnerability descriptions to structured outputs:

CVE-2021-44228 (Log4Shell)
- CWE-502 (Deserialization of Untrusted Data)
- MITRE ATT&CK T1190 (Exploit Public-Facing App)

RAG — COMPLIANCE & REGULATORY DOMAINS

Retrieval-Augmented Generation grounds outputs in authoritative source documents, addressing hallucination and outdated knowledge.

WHY RETRIEVAL-AUGMENTED GENERATION (RAG) ?

Responses must be traceable to statutes
Citation accuracy is critical for audits
Regulations update; no retraining needed

EXAMPLE: Legal Question-Answering Systems

Retrieves relevant statutory provisions and generates citable responses:

Query: "Is this action permissible?"
- Retrieves relevant statute provisions
- Generates response with explicit citations

Key Insight: Fine-tuning excels at **descriptive classification** (what is this?), while RAG excels at **prescriptive compliance** (what does the regulation say?)
CCoP 2.0 benefits from **both** approaches.

Case for CCoP 2.0: Fine-Tuning + RAG

> **Fine-tuning** improves compliance reasoning, gap identification, and risk justification.
>
> **RAG** provides authoritative clause traceability and audit defensibility. **Both are needed for CCoP 2.0.**

ASPECT	FINE-TUNING STRENGTHS	FINE-TUNING LIMITATIONS	Addressed by RAG?
Compliance Reasoning	Improves scenario interpretation and audit-style reasoning	Hard to justify with exact regulatory text	Yes
Gap Identification	Learns patterns of common control weaknesses	May miss edge cases tied to precise wording	Yes
Remediation Guidance	Generates practical, proportionate recommendations	May lack explicit linkage to mandatory controls	Yes
Regulatory Updates	No dependency on external retrieval systems	Requires retraining when clauses change	Yes
Audit Defensibility	Useful for preliminary analysis and reasoning	Insufficient for evidence-based audits alone	Yes

This Project: Focus on **fine-tuning for compliance reasoning** in Phase 2.

RAG-based grounding identified as future work for audit defensibility.

Fine-Tuning Methodology

WHAT IS **QLoRA ?**

Quantized Low-Rank Adaptation - A Parameter-Efficient Fine-Tuning (PEFT) technique that combines 4-bit quantization with Low-Rank Adapters (LoRA) to drastically cut resource usage while preserving model performance.

WHY QLoRA FOR CCoP 2.0?

Enables **cost-efficient, lightweight offline deployment** for air-gapped and on-premise infrastructure - exactly what CIIOs need for their isolated CII environments.

Fine-Tuning Pipeline

1. DATA
Curate CCoP 2.0 training dataset

2. QUANTIZE
4-bit NF4 quantization

3. TRAIN
LoRA adapters on frozen weights

4. EVALUATE
21 custom benchmarks

5. DEPLOY
Air-gapped CII environment

MEMORY EFFICIENT
Fine-tune 65B models on single 48GB GPU

COST EFFECTIVE
3-4x memory reduction vs full precision

HIGH ACCURACY
97-99% performance retention

AIR-GAP READY
Offline deployment for CII infrastructure

Benchmarks & Evaluation Framework

21 BENCHMARKS BY TYPE

CLASSIFICATION (4 benchmarks)

B1 CCoP Applicability Determination
B2 Compliance Status Classification
B4 IT/OT Infrastructure Classification
B5 Control Requirement Comprehension

REASONING (15 benchmarks)

B3 Conditional Compliance
B6 Control Intent Recognition
B7 Gap Identification Quality
B8 Gap Prioritization
B9 Risk Identification
B10 Risk Justification Coherence
B11 Risk Severity Assessment
B12 Audit Awareness
B13 Audit Planning
B14 Remediation Quality
B15 Remediation Prioritization
B16 Residual Risk Awareness
B17 Governance Understanding
B18 Cross-Scenario Consistency
B19 Cross-Domain Consistency

SAFETY (2 benchmarks)

B20 Over-Specification Avoidance
B21 Hallucination Detection

TIER 1: DETERMINISTIC

Rule-based, binary outcomes

Label matching + automated key fact extraction. No human review required.

TIER 2: SEMANTIC REASONING

Similarity + fact recall + rubric

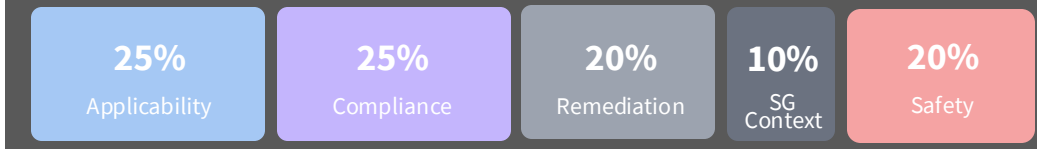
Embedding-based semantic matching, reasoning dimension scores. 20% expert validation.

TIER 3: LLM-AS-JUDGE

Rubric-guided LLM evaluation

Hallucination detection, forbidden claim checks. Single fabrication = failure.

EVALUATION CATEGORY WEIGHTS



Hybrid approach: automated metrics + LLM-as-Judge + human-in-the-loop validation by domain experts

Establishing the Ground Truth

PHASE 2

THE CHALLENGE

No publicly available benchmark datasets exist for Singapore's CCoP 2.0. A custom ground-truth dataset must be constructed through multi-stage validation.

Ground Truth Creation Process

1. AI GENERATION
Claude AI generates 118 test cases across 21 benchmarks

2. HUMAN REVIEW
Verify clauses against official CCoP 2.0 PDF

3. VALIDATION
Schema alignment, automated checks, JSONL format

4. EXPERT REVIEW
CCOP Practitioner Validation (PENDING)

118 Test Cases

21 Benchmarks

100% CCoP Coverage

612 Key Facts

TIER 1: CLASSIFICATION

29 Cases (25%)

- **B1** Applicability (8)
- **B2** Compliance (7)
- **B4** IT/OT Systems (7)
- **B5** Control Requirements (7)

Automated label-based accuracy scoring

TIER 2: REASONING

79 Cases (67%)

- **B3** Conditional Compliance
- **B6** Control Intent
- **B7-B8** Gap Analysis
- **B9-B11** Risk Assessment
- **B12-B13** Audit Awareness
- **B14-B16** Remediation
- **B17-B18** Governance
- **B19** Consistency

Semantic similarity + key fact recall + expert validation (20%)

TIER 3: SAFETY

10 Cases (8%)

- **B20** Over-Specification (3)
- **B21** Hallucination (7)

Critical: Single fabrication = failure

LLM-as-Judge + human validation

Current Status: Ground truth dataset sent to CCoP 2.0 compliance practitioner for expert validation. Awaiting approval before finalizing baseline evaluation.

Project Plan

September 2025 – August 2026 | 12-month development timeline

1

SEP – OCT 2025

Phase 1: Foundation

Local infrastructure setup
Model configuration (QLoRA)
Evaluation framework design

COMPLETED

2

NOV 2025 – JAN 2026

Phase 2: Baseline

Ground truth establishment
21 benchmark evaluation
Model comparison study

IN PROGRESS

3

FEB – MAY 2026

Phase 3: Fine-tuning

QLoRA training on CCoP 2.0
Iterative improvement cycles
Target: 85% or higher accuracy

PLANNED

4

JUN – AUG 2026

Phase 4: Deployment

Air-gapped environment setup
CI/CD pipeline integration
Production validation

PLANNED

Problem Identification, Research & Plan
Dec 2025

Ground Truth Finalized
Jan 2026

Fine-tuned Model Ready
May 2026

Production Deployment
Aug 2026

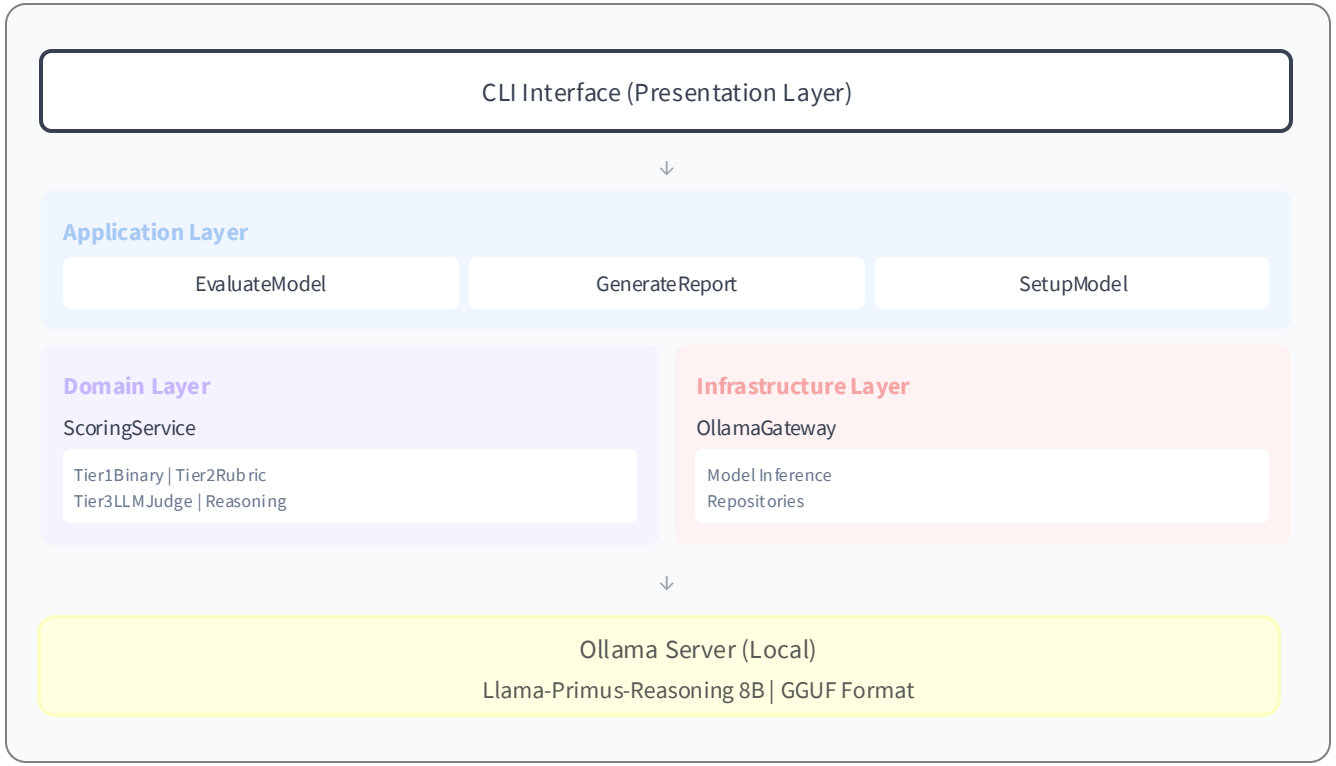
Current Status: Phase 2 in progress | 17/21 benchmarks evaluated | Awaiting domain expert validation

Phase 1: Infrastructure & Model Configuration

Local inference setup for Phase 2 baseline evaluation using clean architecture pattern

MODEL INFERENCE APPLICATION ARCHITECTURE

Phase 1 local deployment validates methodology before cloud scaling



10.1 HARDWARE (LOCAL)

M3	16GB	Metal
Apple Silicon	Unified Memory	GPU Accel

MODEL SPECIFICATION

Base Model	**Llama-Primus-Reasoning**
Parameters	**8 Billion**
Quantization	**4-bit NF4 (QLoRA)**
Format	**GGUF (Apple Silicon)**
Framework	**llama.cpp + Ollama**

BASELINE EVALUATION PERFORMANCE

118	~83s	~2-2.5 hrs
Test Cases	Per Test Case	Total Runtime

Phase 2: Baseline Eval Preliminary Results

48%
Weighted Score

17/21 benchmarks evaluated*
97 test cases executed

**4 benchmarks pending human expert evaluation as they require interpretive judgement to assess whether responses would pass regulatory scrutiny*

STRENGTHS: Reasoning & Classification

Compliance Classification Accuracy: **69%**
Responsibility Attribution: **69%**
Gap Identification & Prioritization: **60%**

THREE CRITICAL FAILURE MODES IDENTIFIED

A. Hallucination — B21: 22%

Model **fabricates specific technical details** - inventing password lengths, SIEM vendor requirements, downtime limits that do not exist in CCoP 2.0.

IMPACT: Unsafe for production without hallucination mitigation

B. IT/OT Classification — B4: 21%

Cannot distinguish IT vs OT infrastructure. Lacks Singapore's **Critical Infrastructure taxonomy** and OT-specific terminology.

IMPACT: Cannot map systems to correct CCoP control sets

C. Control Application — B6: 21%

Struggles to explain **why controls exist** and what security objectives they serve. Limits ability to provide meaningful guidance

IMPACT: Learned auditor skill gap, not knowledge gap

Key Insight: Model has strong meta-reasoning but weak factual grounding—**a 2.0x capability gap**.
Fine-tuning must inject **CCoP-specific knowledge**: anchor reasoning to actual clauses to stop fabrication,
then teach technical details (control mechanics, IT/OT classification) so the model's reasoning produces accurate compliance advice.

Next Steps

Immediate actions and Phase 3 preparation

IMMEDIATE ACTIONS (Dec 2025 - Jan 2026)

1

Complete Benchmark Evaluation

Finish remaining 4 benchmarks (17/21 completed)

Run full evaluation suite on all 3 models (Llama-Primus, GPT-5, DeepSeek-V3)

2

Domain Expert Validation

Incorporate expert feedback into ground truth dataset

Target: Ground truth finalized by **Jan 2026**

PHASE 3 PREPARATION (Feb 2026)

3

Fine-tuning Data Preparation

Curate training dataset from validated ground truth

Design QLoRA training configuration

Set up training infrastructure

KEY MILESTONES

Ground Truth Finalized
January 2026

→

Fine-tuning Begins
February 2026

→

Model Ready
May 2026

Thank you.