Fine-Tuning LLM on CSA's Cybersecurity Code of Practice (CCoP 2.0) standards for Critical Information Infrastructure (CII)

Problem Statement

- CII organizations and developers struggle to interpret and implement CCoP 2.0's 220 complex, Singapore-specific cybersecurity requirements across both IT and OT/ICS infrastructure—spending months on manual compliance analysis, gap assessments, and policy generation.
- A CCoP 2.0 fine-tuned LLM can automate compliance validation through code scanning, conduct gap analysis against organization's policies and standards, provide real-time regulatory guidance, and reduce compliance achievement time from 12+ months to x months with xx% cost savings (TBC)

Project Objectives

1. Benchmark baseline performance of Llama-Primus-Reasoning model (8B parameters) on CCoP standards to establish current capabilities and identify knowledge gaps
2. Fine-tune Llama-Primus on CCoP standards by creating a comprehensive training dataset and training the model to achieve > 85% accuracy in detecting compliance violations with respect to CCoP (Cybersecurity Code of Practice) standards.
3. Deploy model to isolated environment (mimic CII) and integrate with CI/CD pipelines to detect non-compliant source codes and configurations across application and infrastructure with respect to CCoP standards.

Real-World Application

1. Relevant for CII organizations to conduct gap analysis of their enterprise standards vs all 220 clauses in CCoP.
2. Fine-tuned model suitable for airgapped deployment within CII organization's isolated infrastructure.
3. Automated security scanning of application and infrastructure code and configurations by integrating with their automated CI/CD pipelines.
4. Evaluation of third-party vendor security documentation, contracts, SoC 2 reports and security questionnaires against CCoP requirements.

CCoP Framework Overview

CCoP 2.0 is Singapore's mandatory cybersecurity framework comprising 220 controls across 11 sections (governance, risk management, technical protections, incident response, and OT/ICS security) that Critical Information Infrastructure organizations must implement to protect essential national services from cyber threats.Retry

How CCoP is Organized:

Section	Section Details	Infrastructure Type	Clauses	Training Implication

Section	Section Details	Infrastructure Type	Clauses	Training Implication
1. Audit	Cross-cutting	BOTH IT and OT contexts needed	~4	BOTH IT and OT contexts needed
2. Governance	Cross-cutting	BOTH IT and OT contexts needed	~15-20	BOTH IT and OT contexts needed
3. Risk Management & Resilience	Mostly Cross-cutting (~80%), Some IT-Cloud (~20%)	BOTH IT and OT contexts, plus cloud-specific examples	~25-30	BOTH IT and OT contexts, plus cloud-specific examples
4. Asset Management	Cross-cutting	BOTH IT and OT contexts needed	~8-10	BOTH IT and OT contexts needed
5. Protect	MIXED: ~50-60% IT-specific, ~40-50% Cross-cutting	IT examples primary, but substantial cross-cutting controls (network segmentation, cryptography, logging) apply to both	~80-90	IT examples primary, but substantial cross-cutting controls apply to both
6. Detect, Respond & Recover	Mostly Cross-cutting (~90%), Some IT (~10%)	BOTH IT and OT contexts needed	~25-30	BOTH IT and OT contexts needed
7. Cybersecurity Awareness	Mostly Cross-cutting (~90%), Some IT (~10%)	BOTH IT and OT contexts needed	~8-10	BOTH IT and OT contexts needed
8. Supply Chain Cybersecurity	Cross-cutting	BOTH IT and OT contexts needed	~10-12	BOTH IT and OT contexts needed
9. Third Party Cybersecurity	Cross-cutting	BOTH IT and OT contexts needed	~12-15	BOTH IT and OT contexts needed
10. OT/ICS Security	OT-only	OT examples exclusively (SCADA, PLCs, Purdue Model)	~35-40	OT examples exclusively (SCADA, PLCs, Purdue Model)

Section	Section Details	Infrastructure Type	Clauses	Training Implication
11. Assurance	Mostly Cross-cutting (~90%), Some IT (~10%)	BOTH IT and OT contexts needed	~8-10	BOTH IT and OT contexts needed
TOTAL	All Sections	~60% Cross-cutting, ~25% IT-specific, ~18% OT-specific	~220	Unified training across all infrastructure types

[1]: IT (Information Technology): Traditional enterprise computing systems (servers, databases, cloud, business applications) that process and store data.

[2]: OT (Operational Technology): Industrial control systems (SCADA, PLCs, sensors) that monitor and control physical processes in critical infrastructure like power plants and water facilities.

Training Strategy

Since 60% of CCoP clauses are cross-cutting (apply to both IT and OT), unified training of all 11 sections enables the model to learn relationships between infrastructure types, correctly distinguish when controls apply to IT-only vs OT-only vs both, and deploy as a single production model rather than maintaining separate IT/OT variants. The alternative strategy to train the model sequentially based on IT-only and subsequently OT controls could lead to catastrophic forgetting—if we train IT sections first then fine-tune on OT, the model loses IT knowledge (safety can drop).

Evaluation Benchmarks

The following benchmarks are proposed to conduct evaluation of the baseline Primus model's understanding of CCoP clauses, as well as incremental assessments of the LLM after fine-tuning with CCoP training datasets.

Category	Benchmark ID	Benchmark Name	Description/Note
1. Compliance Benchmarks	B1	CCoP Interpretation Accuracy	CCoP understanding and regulatory accuracy
	B2	Clause Citation Accuracy	
	B3	Hallucination Rate	0% required
	B4	Singapore Terminology	100% required
	B5	IT vs OT Classification	

Category	Benchmark ID	Benchmark Name	Description/Note
2. Code & Infrastructure Benchmarks	B6	Code Violation Detection	Technical vulnerability detection (SAST, SCA, IaC)
	B7	False Positive Rate	
	B8	IaC Misconfiguration Detection	
3. Advanced Capability Benchmarks	B9	Incident Classification	Broader use cases beyond code scanning
	B10	Gap Analysis Quality	
	B11	Policy Generation Quality	
	B12	Cross-Standard Mapping	
4. Safety & Security Benchmarks	B13	Prompt Injection Resistance	Adversarial robustness
	B14	Jailbreak Resistance	
5. Training Quality Benchmarks	B15	Training Loss	Model learning effectiveness monitoring
	B16	Validation Loss	
	B17	Perplexity Score	
6. Performance Benchmarks	B18	Inference Speed	Operational efficiency
	B19	Memory Usage	

Dataset Requirements

Type	Example Topics / Formats	Total Examples
1. CCoP Compliance Examples	- Q&A covering all 11 sections - Clause citations - Singapore-specific terminology - Hallucination prevention tests - IT vs OT classification scenarios	700
2. Vulnerable & Clean Code	- Code with security vulnerabilities: Python, Java, JavaScript, Go, C++ - Patterns: OWASP Top 10, CWE - Clean code samples for false positive rate testing	1,560

Type	Example Topics / Formats	Total Examples
3. Infrastructure as Code	- Terraform, Kubernetes, CloudFormation - AWS, Azure, GCP - Security misconfigurations and correct baselines	800
4. OT / ICS Specific	- SCADA system cases - PLC code - Industrial protocols - Purdue Model architectures - Secure coding (Section 10)	850
5. Advanced Capabilities	- Incident response scenarios - Gap analysis - Policy generation - Cross-standard mappings (ISO 27001, NIST 800-53, IEC 62443) - Adversarial safety tests	1,360

Total:

- **5,270 examples**
  - *4,850 training + 420 test*

Project Phases Overview

Objective: Incrementally assess Llama Primus Reasoning model's performance in understanding CCoP clauses and detecting non-compliance.

Phase 1: Foundation & Setup

- Set up GPU infrastructure.
- Deploy Llama-Primus-Reasoning.
- Install LoRA fine-tuning framework and safety testing tools.
- Prepare evaluation pipeline and structure CCoP 2.0 documentation for baseline testing.

Phase 2: Quick Baseline Screening

- Test unmodified Primus using 40 screening cases across 6 benchmarks (B1–B6).
- Goal: Determine if the model has a 15–20% baseline understanding.
- **Critical checkpoint:** Proceed only if score > 15% **AND** zero hallucinations are detected.

Phase 3: Comprehensive Baseline

- Conduct detailed evaluation with 170 test cases across 12 benchmarks (B1–B12).
- Identify specific strengths and weaknesses.
- Use results to map out training data requirements based on identified gaps in CCoP understanding.

Phase 4: Small Fine-Tune Test

- Fine-tune using a small dataset (148 examples).
- Validate approach by aiming for >35% improvement with 190 tests (B1–B17).
- **Critical checkpoint:** Confirm fine-tuning effectiveness before building the full dataset.

Phase 5: Full Dataset Creation

- Create a production-ready dataset: 4,850 training examples (all 11 CCoP sections) + 420 comprehensive test cases.
- Validation: Cross-verify all examples for accuracy and completeness vs all 11 CCoP sections.

Phase 6: Comprehensive Fine-Tuning

- Train production model v1.0 using the complete dataset.
- Optimize hyperparameters and monitor training metrics (e.g., loss, perplexity).
- Ensure safety by continuous monitoring and checkpoint management.

Phase 7: Production Validation

- Conduct comprehensive testing across all 19 benchmarks.
- Perform expert review, red team security assessment, and performance profiling.
- **Deployment decision:** Model is production-ready if all must-pass criteria are met (>85% overall score + expert approval).

Key Result

1. Achieve a target weighted average score of >85% across all 19 benchmarks

Learning Objectives

- **LLM Fine-Tuning Pipeline:**
    - Learn LoRA/PEFT techniques
    - Hyperparameter tuning
    - Preventing catastrophic forgetting at scale
- **ML Evaluation Framework Design:**
    - Build a 19-benchmark system
    - Automated testing
    - Expert validation
    - Adversarial robustness
- **Large-Scale Dataset Engineering:**
    - Curate 5,270 examples
    - Quality assurance

  - Zero data leakage

- **Production ML Optimization:**

  - Achieve <5s inference
  - <16GB memory usage
  - Suitability for edge/air-gapped environments

- **Domain-Specific AI Applications:**

  - Apply LLMs to regulated industries
  - Support audit trails, compliance documentation, and explainability

References

1. Singapore Cybersecurity Code of Practice (CCoP) 2.0
2. CSA Singapore: Critical Information Infrastructure (CII) FAQ
3. Llama-Primus-Reasoning on HuggingFace