PROJECT PROPOSAL

Ephemeral Polymorphic Defense (EPD)

1. Introduction

The **Ephemeral Polymorphic Defense (EPD)** system represents a paradigm shift in cloud security. Unlike traditional static defense mechanisms that rely on fixed rules and manual intervention, EPD acts as a living, breathing immune system for cloud infrastructure.

By leveraging a triumvirate of specialized AI squads, **Watchers** (Real-time Detection), **Brain** (Cognitive Reasoning), and **Ghost Agents** (Autonomous Interventions). EPD achieves a level of agility and responsiveness that matches the speed and sophistication of modern cyber threats.

This proposal outlines the current capabilities of the EPD system following its successful first verification and details the requirements for the next phase of development: scaling from a successful prototype to an enterprise-grade autonomous defense platform.

2. Executive Summary

The Ephemeral Polymorphic Defense (EPD) system has successfully concluded its first verification phase, marking a critical milestone in autonomous cyber defense. During the "First Test", the system demonstrated a complete end-to-end security loop that operated with 100% autonomy, requiring zero human intervention to detect, analyze, and neutralize simulated zero-day threats.

The operational results validated the core architectural thesis. The system successfully identified high-velocity traffic anomalies, processed them through a cognitive reasoning engine, and deployed targeted remediation solutions in under 2 seconds. This performance was achieved while maintaining a system latency of approximately 12ms per packet analysis, proving that complex AI reasoning can be integrated into real-time security flows without compromising network performance.

Crucially, the test confirmed the harmonious interoperability of the three AI squads. Squad A provided high-recall anomaly detection, Squad B successfully translated technical alerts into strategic action plans, and Squad C executed these plans through ephemeral agents that left no residual footprint. This success provides a solid foundation for the next phase of research, which aims to expand the system's cognitive breadth and deploy it against more diverse and adversarial attack scenarios.

3. Project Scope and Assumptions

To maintain focus on novel contributions to the field of Agentic AI security, this project makes the following key assumption regarding the detection layer:

- **Exclusion of Squad A (The Watchers)**: We assume that the detection of network anomalies (IDS/IPS) is a solved problem within the scope of this research. Existing commercial solutions (e.g., standard ML-based IDS) are sufficient to provide high-fidelity alerts.
- **Focus Area**: This project exclusively targets **Squad B (The Brain)** and **Squad C (The Hands)**. The core research value lies in the *cognitive reasoning* over alerts and the *safe execution* of remediation, rather than the initial detection of the anomaly itself. Future reports and metrics will focus solely on the “Reasoning -> Action” loop.

4. Background and Justification

4.1. The Traditional Layered Approach

Historically, Cloud Native Application Protection Platforms (CNAPP) have relied on a “defense-in-depth” model composed of static, overlapping layers. These include perimeter firewalls, signature-based Intrusion Detection Systems (IDS), and rigid policy enforcement points. While effective against known threats, this architecture suffers from a fundamental rigidity: it requires manual rule updates to address new attack vectors. As noted in recent industry analysis, these centralized tools often succumb to “Context Blindness” and “False Positives,” failing to adapt when attackers shift tactics mid-campaign.

4.2. Agentic Security

The modern security landscape is shifting towards **Agentic AI**—systems where autonomous agents not only detect threats but actively “try to do things” to mitigate them. Unlike passive monitoring tools, these agents possess the agency to execute commands, modify configurations, and interact with the infrastructure directly. Recent frameworks like **MALCDF** (Multi-Agent LLM Cyber Defense Framework) demonstrate the power of this approach, utilizing decentralized squads of agents to detect and analyze threats with greater accuracy (90%) and lower false positives than traditional baselines.

4.3. Capability vs. Control

With the introduction of autonomous agents, the core challenge shifts from *capability* (Can the AI stop the attack?) to *safety* (Can we trust the AI not to break the system?). Unchecked agents introduce new risks, including:

- **Malicious Collaboration**: As explored in **SentinelNet**, compromised agents within a collaborative system can spread misinformation or launch “persuasive attacks” to mislead their peers.

- **Unpredictable Execution**: Developing powerful agents without formal guarantees can lead to disastrous side effects. Principles from **VeriGuard** highlight the need for “Correct-by-Construction” policies, where agent actions are mathematically verified against safety constraints before execution.

EPD directly addresses these concerns. It is not just an agentic system; it is a *safety-first* architecture. By employing **Ephemeral Agents** (Squad C) that self-destruct after a single task and a **Cognitive Brain** (Squad B) trained on secure Q&A datasets, EPD ensures that autonomy never compromises system integrity. The focus of this proposal is to deepen these safety mechanisms, verifying that our "Hands" remain under strict cognitive control.

5. Project Goals and Objectives (SMART)

Building on the successful first validation, the next phase of EPD focuses on transforming the prototype into a hardened, enterprise-ready platform. We have consolidated our focus into a single, high-impact objective that balances advanced capability with rigorous safety:

- **Specific**: Create a unified "Safe-Agent" architecture that simultaneously expands Squad B's reasoning capabilities (to detect malicious collaboration) while enforcing "Correct-by-Construction" constraints on Squad C. This merges the need for *smarter* agents (via larger, multi-turn debate datasets) with the need for *safer* agents (via formal logic verification).
- **Measurable**:
    1. **Safety**: Achieve a **100% Safe-Action Rate** across 1,000 diverse adversarial scenarios, including "hypnotic" jailbreaks.
    2. **Accuracy**: Improve reasoning accuracy from 71.3% (Initial 10%) / 73.0% (Full Dataset) to **>90%** on a dataset of **15,000+ complex security scenarios**.
- **Achievable**: By combining the generation of synthetic "Chain-of-Thought" training data with a runtime logic solver (e.g., Z3) that pre-validates every agent command.
- **Relevant**: Directly addresses the "Capability vs. Control" dilemma, ensuring EPD can handle APTs without becoming a liability itself.

6. Methodology and Approach

This research employs a multi-tiered agentic architecture, treating security operations as a collaborative task between a cognitive reasoner and ephemeral execution agents.

6.1. Cognitive Training (Squad B)

The central intelligence unit ("The Brain") is built upon a Small Language Model (SLM) to balance reasoning capability with low-latency inference.

- **Base Model**: Microsoft Phi-2 (2.7B parameters).

- **Optimization Technique**: Quantized Low-Rank Adaptation (QLoRA). This allows for efficient fine-tuning of the attention layers while freezing the base model weights, preserving general reasoning abilities while injecting domain-specific security knowledge.

- **Training Dataset**: The model was supervised-fine-tuned (SFT) on the **SecQA** dataset, a curated corpus of 242 high-fidelity security decision pairs aggregated from open benchmarks (e.g., CyberMetric, SecEval). The dataset focuses on distinguishing between benign anomalies and malicious indicators (e.g., identifying SQL injection patterns vs. normal traffic spikes).

- **Validation**: The model achieved a reasoning accuracy of **73.0%** on the hold-out test set, demonstrating robust generalization to unseen threat descriptions.

6.2. Polymorphic Execution (Squad C)

Unlike traditional static scripts, the execution layer ("The Hands") utilizes a **Polymorphic Agent Factory** approach to evade detection by adversarial monitoring tools.

- **Agent Foundation**: Llama 3.2 (3B) serving as the execution kernel.
- **Mechanism**:
  - **Ephemerality**: Agents are instantiated for a single atomic task and self-destruct immediately upon completion (`Time-to-Live < 2s`), leaving no memory footprint or traversable graph for attackers to exploit.
  - **Polymorphism**: Prompt templates are dynamically mutated using semantic variation (e.g., changing syntax while preserving intent) to prevent pattern-matching defenses from fingerprinting the defensive actions.
- **Testing Protocol**: The agents were evaluated against a benchmark of **873 unique scenarios**, achieving a **100% Defense Success Rate** in neutralising targets across varying infrastructure contexts (AWS, Linux CLI).

7. Risk Management

Deploying autonomous AI in a security-critical loop carries inherent risks. We have identified key failure modes and established realistic mitigation strategies to ensure operational stability without over-promising perfection.

7.1. Cognitive Hallucination (The "False Positive" Risk)

- **Risk**: No AI model provides 100% accuracy. There is a non-zero probability that Squad B may misinterpret a legitimate administrative action (e.g., a sudden database backup) as a data exfiltration attempts, leading to unnecessary service interruption.
- **Mitigation**: We implement a "Confidence Threshold" mechanism. Remediation plans with an AI confidence score below 95% will default to a "Notify-Only" state, requiring human approval. Furthermore, the proposed dataset expansion (Objective 1) includes negative samples (benign activities that *look* malicious) to specifically improve decision boundary resolution.

7.2. “Friendly Fire” from Polymorphism

- **Risk**: The polymorphic nature of Squad C agents is designed to evade external attackers, but it may effectively look like malware to our *own* ancillary security tools (e.g., legacy EDRs installed on endpoints), causing internal conflicts or “fratricide.”
- **Mitigation**: We will implement cryptographically signed “Agent Passports” While the agent’s behavioral code changes shape, it will carry a verifiable digital signature that authorized internal EDRs can recognize and allow-list.

8. Implementation Plan and Research Roadmap

To achieve the “Trustworthy Cognitive Autonomy” objective, we propose a systematic study comparing varied model architectures against our current baseline.

8.1. Proposed Model Architecture Search

We will benchmark **7 distinct models** following a graduated parameter progression (from 0.5B to 14B). This “Ladder of Cognition” study aims to identify the precise threshold where security reasoning capabilities emerge and plateau.

Model	Params	Rationale
Qwen2.5-0.5B	0.5B	**Nano Baseline**: Testing if strict logic rules can emerge in <1B parameters.
Llama 3.2-1B	1B	**Mobile Class**: State-of-the-art for extreme edge deployment efficiency.
Gemma 2-2B	2B	**On-Device Standard**: Google’s high-efficiency architecture for local inference.
Phi-3.5-mini	3.8B	**Dense Reasoner**: The successor to our current Phi-2 baseline, known for high data density.
Llama 3.1-8B	8B	**The New Standard**: The defining general-purpose model for the <10B category.
Mistral NeMo	12B	**Mid-Range Bridge**: A collaboration between NVIDIA/Mistral optimizing the 12B sweet spot.
Qwen2.5-14B	14B	**Upper Limit**: The most powerful open-weights model fitting within single-GPU constraints.

8.2. Baselines and Benchmarks

To satisfy Q1 academic reporting standards, we will compare our “Smart & Safe” approach against the current industry standards for reasoning and execution (explicitly excluding detection/IDS baselines):

- **Squad B Baselines (Reasoning & Decision)**:
    1. **Zero-Shot Commercial SOTA (GPT-4o)**: represents the “Upper Bound” of reasoning capability. We aim to demonstrate that our fine-tuned 14B model can match the domain-specific reasoning accuracy of GPT-4o while running at a fraction of the cost and latency.
    2. **Rule-Based SOAR (Static Playbooks)**: The current industry standard. We will compare against rigid “If-Then” logic to quantify the improvement in handling novel/polymorphic threats where static rules fail (Recall vs. Novelty).
- **Squad C Baselines (Execution & Safety)**:
    3. **Static Automation (Ansible/Terraform)**: The standard for automated changes. We will compare our *Polymorphic* agents against these *Static* scripts to demonstrate superior “Evasion” (lower detectability by attacker counter-measures).
    4. **Human Analyst (The “Safety” Gold Standard)**: A human operator is slow (high MTTR) but highly safe. We use this to benchmark our **Safe-Action Rate**, aiming to match human safety levels while exceeding human speed by orders of magnitude.
- **Academic Evaluation Criteria**:
    1. **Reasoning Accuracy**: % of correctly identified threat intents vs. ground truth (Target: >90% on **CyberMetric** & **SecQA**).
    2. **Hallucination Rate**: % of benign anomalies misclassified as threats (Target: <5%).
    3. **Safe-Action Rate**: % of interventions that adhere to verified safety constraints (Target: 100%).
    4. **Operational Latency (MTTR)**: End-to-end time from log ingestion to remediation execution (Target: <1s).

9. Resource Requirements and Budget

9.1. Current Bottleneck

The current research environment (MacBook Pro M2 Max, 32GB RAM) has reached its physical limits for model training.

- **Training Latency**: Fine-tuning Squad B (Phi-2) currently takes **3-4 hours per epoch**. * **Performance Ceiling**: We are capped at **73% accuracy** due to the inability to utilize larger batch sizes or train deeper LoRA ranks effectively.

- **Scalability**: We cannot load models >7B parameters for local training without severe quantization degradation, making the proposed 14B/16B experiments impossible.

9.2. Resource Request: High-Performance Compute

To execute this roadmap and produce a Q1-quality publication, we formally request access to enhanced cloud-based computing resources capable of handling larger scale model training.

- **Requirement**: **High-Memory Accelerated Compute Resources**.
- **Justification**:
    1. **Speed**: Significantly reduce training time (currently 4 hours/epoch), enabling rapid experimentation and hyperparameter tuning.
    2. **Capacity**: Enable the training of larger models (up to 14B parameters) which currently cannot be loaded into local memory without severe performance degradation.
    3. **Scale**: Support the generation and processing of the **15,000+ scenario dataset** required for the “Debate” objective.

10. Conclusion

The EPD first test successfully validated the concept of autonomous defense, but “concept” is not “production.” By upgrading our cognitive engine (Squad B) through the proposed 7-model search and enforcing rigorous safety guarantees (Objective 1), we aim to elevate EPD from a promising prototype to a definitive academic reference architecture for Safe Agentic Security. The requested resources are the only bridge between our current 73% prototype and that >90% industry-ready solution.