# NetworkLLM: Intent-Aware Intrusion Detection Using Large Language Models Grounded in Protocol Semantics

## Abstract

Traditional intrusion detection systems (IDS) are constrained by their reliance on signatures or statistical anomaly detection, which often fail to capture the intent behind network traffic. As cyber threats evolve, adversaries increasingly exploit legitimate protocols in semantically novel ways (e.g., DNS tunnelling, TLS padding exploits), evading byte-level detection.  
  
This project proposes NetworkLLM, a framework that employs Large Language Models (LLMs) to infer the semantic intent of network packets and flows. By combining field-aware tokenisation, RFC-grounded retrieval, and LLM-guided exploration, the system will move from surface anomaly spotting towards intent-aware, explainable detection. The outcome will be an IDS capable of recognising not only that a packet is suspicious, but why, grounded in protocol semantics, and mapping it to adversarial tactics (e.g., MITRE ATT&CK).

## Background and Motivation

Gap: Existing IDS (Snort, Suricata, ML-based detectors) identify anomalies or signatures but lack semantic reasoning. They answer “what is odd?” but not “what was the traffic trying to achieve?”  
  
Opportunity: LLMs trained on structured representations can interpret packets as messages, making it possible to infer intent (e.g., handshake vs exploit).  
  
Research context:  
- LLM-guided fuzzing (ChatAFL, NDSS 2024): LLMs can learn protocol grammars and generate meaningful traffic.  
- PROSPER (HotNets 2023): LLMs can extract protocol state machines from RFCs.  
- TrafficLLM (2025): Domain-specific tokenisation boosts performance over raw bytes.  
- LLMcap (2024): Self-supervised LLMs detect anomalies in unlabeled PCAPs.  
- Packet Field Tree (ACM 2024): Enables field-aware representations.  
  
Together, these point towards a paradigm shift: intent-aware intrusion detection.

## Research Questions

1. How can network traffic be represented as language-like tokens that capture both byte-level robustness and semantic field structure?  
2. Can LLMs reliably infer intent classes (e.g., benign handshake, probe, exploit, exfiltration, C2)?  
3. How does grounding LLMs in RFC/state machine knowledge improve accuracy and reduce hallucinations?  
4. Can LLMs autonomously explore protocol space (via fuzzing) to discover new attack behaviours and generate adversarial training data?  
5. How do intent-aware IDS outputs impact SOC analysts’ efficiency, false positive rates, and trust in automated alerts?

## Objectives

- Develop a field-aware tokenisation framework for network traffic.  
- Train and evaluate intent classifiers (LLM-based, grounded in RFCs).  
- Integrate RFC-RAG retrieval for explainable IDS decisions.  
- Build a self-supervised anomaly detector for zero-day intent discovery.  
- Use LLM-guided fuzzing to stress-test and adversarially train the IDS.  
- Demonstrate operational integration in SOC pipelines (Elastic/Wazuh).

## Methodology

Datasets:  
- Benchmark: CIC-IDS2017, UNSW-NB15.  
- Modern: MAWIFlow 2025 (backbone), Gotham 2025 (IoT), CESNET TimeSeries 2025 (encrypted).  
- Specialised: Darknet/Tor (BCCC-DarkNet-2025), ICS honeypot traces (LLMPot).  
  
Work Packages:  
1. Representation Learning (Months 1–6): Develop field-aware tokeniser; ablation vs raw bytes.  
2. Intent Inference (Months 6–12): Fine-tune LLMs (DistilBERT, domain-adapted) on tokenised traffic.  
3. Grounded Explainability (Months 12–18): Build RFC/spec RAG index; integrate FSM validation.  
4. Zero-Day Detection (Months 18–24): Self-supervised anomaly detection on unlabeled PCAPs.  
5. Exploration & Robustness (Months 24–30): LLM-guided fuzzing harness; generate hard negatives.  
6. Operational Deployment (Months 30–36): Integrate into SOC workflows; human-in-the-loop evaluation.  
  
Evaluation Metrics:  
- Detection: PR-AUC, F1, ROC-AUC.  
- Explainability: % correct RFC citations, hallucination rate.  
- Exploration: State coverage, new protocol paths found.  
- Operational: Analyst false positive reduction, trust ratings, mean time to triage.

## Expected Contributions

- A novel intent-aware IDS framework using LLMs for semantic packet analysis.  
- A field-aware traffic representation optimised for LLM processing.  
- An RFC-grounded explanation system, reducing hallucinations in AI-driven IDS.  
- A methodology for LLM-guided fuzzing and adversarial hard-negative mining.  
- A deployment-ready prototype integrated into SOC pipelines.  
- Theoretical framework for defining and evaluating network intent in cybersecurity.

## Timeline (3 Years)

- Year 1: Baselines, tokenisation framework, initial LLM classifier.  
- Year 2: RFC-RAG grounding, anomaly detection, protocol fuzzing harness.  
- Year 3: SOC integration, evaluation with analysts, thesis writing.

## Impact

Academic: Advances at the intersection of NLP, security, and formal methods.  
Industrial: Direct application to SOC automation, reducing alert fatigue and improving explainability.  
Societal: Improves trust and resilience of cybersecurity infrastructure against zero-day exploits and adversarial tactics.