LLMs for Cyberattack Detection on UNSW-NB15

University of Tehran — Large Language Models (Spring 2025)

Instructors: Dr. Mohammad-Javad Doosti, Dr. Yadollah Yaghoubzadeh

Mobin Tirafkan, Kiyan Khezri, Mohammad Mehrabian

Sep. 2025

Introduction: Modern Network Security

The Importance

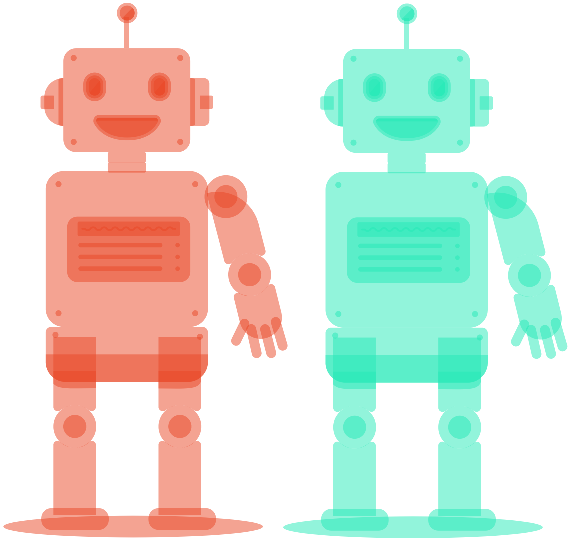
- **The Threat:** We operate in an era of escalating and sophisticated cyber threats
- **The Capability:** The ability to **accurately detect** is fundamental
- **The Goal:** To protect an organization's critical digital assets and infrastructure.

The Challenge

- **Log Volume:** Network logs are increasingly complex and voluminous.
- **Analyst Overload:** Human analysts are overwhelmed by raw data, leading to alert fatigue.
- **Traditional ML Limits:** Often "black-box" decisions without explanation.
- **LLM Direct Use:** LLMs struggle with direct tabular data classification and raise privacy concerns

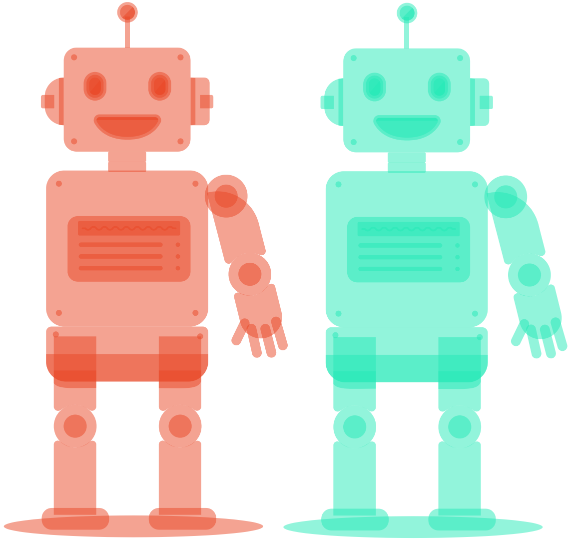
Introduction: From Logs To Explainable Verdict

- Our Proposed Solution: **A multi-agent Framework**
- We introduce a novel framework where **two** specialist agents sequentially transform **raw data** into a final, **explainable verdict**.
- This verdict consists of two key parts:
  - **A classification Label (Attack / Normal)**
  - **An evidence-based Analysis (The "Reason")**
- **The Agents:**
  1. **The Storyteller**
  2. **The Reasoner (a fine-tuned gemma-3-1B Model)**



Introduction: From Logs To Explainable Verdict

- **Key Accomplishments**
  - Our Multi-agent framework achieved a remarkable **93% accuracy on attack predication.**
  - This significantly outperforms larger models:
    - GPT OSS (120B): 59.7% accuracy
    - Qwen 3 (32B)(Reasoning): 58.9% accuracy



Related Works

- Traditional Machine Learning Approaches
    - **Focus:** Primarily on feature engineering and maximizing classification accuracy using numerical data.
    - **Techniques:** Employ models like CNNs, LSTMs, Autoencoders, and XGBoost for feature selection on datasets like UNSW-NB15.
    - **Key Limitation:** While achieving high accuracy, these models often function as "**black boxes,**" lacking the crucial **explainability** needed for an effective security response.

Related Works

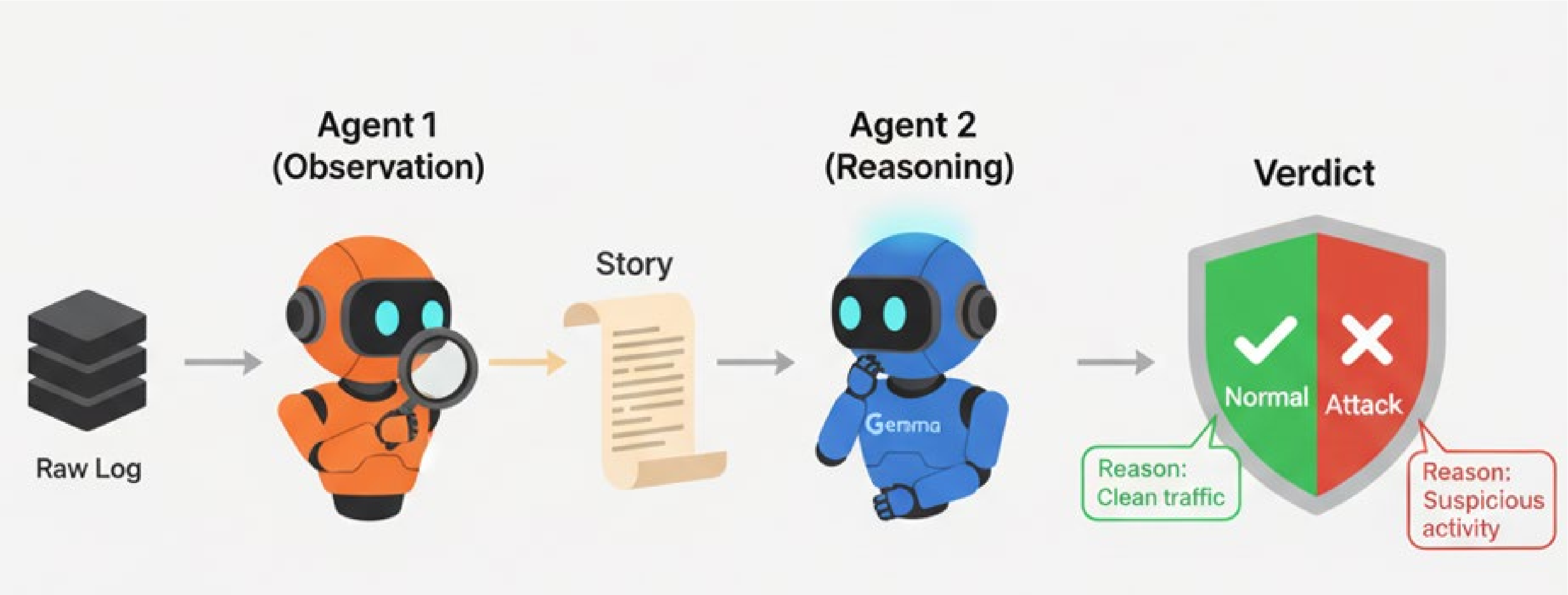
- Recent LLM-based Approaches
  - **Focus:** Exploring the direct use of LLMs for tabular data classification and anomaly detection.
  - **Techniques:**
    - **Prompt Engineering:** Guiding pre-trained models with advanced prompts (e.g., Chain-of-Thought) to analyze tabular data.
    - **Fine-Tuning Strategies:** Optimizing models by enhancing data representation, such as using **decimal truncation** and **randomizing feature order** to improve robustness and generalization.
  - **Key Limitation:** These methods still primarily target **classification accuracy**. They do not emphasize generating human-readable justifications or proposing concrete, actionable steps.

Background: UNSW-NB15 Dataset

- A Benchmark for Network Intrusion Detection
- Dataset Composition and Features
  - real-world normal traffic and nine families of attacks.
  - Each record is explicitly labeled as **Normal (0)** or **Attack (1)**.
  - Comprises **49 features**
  - **Training Set (175,341 records)** and a **Testing Set (82,332 records)**.

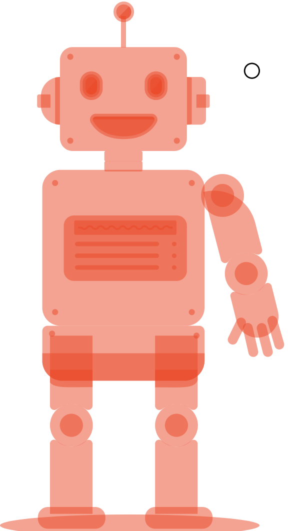
dur	proto	service	sbytes	dbytes	sttl	sload	attack_cat	label
0.012947	tcp	-	2766	24004	31	1670811.875	**Normal**	**0**
0.031951	tcp	-	1540	1644	31	361553.625	**Normal**	**0**
0.005483	tcp	http	1040	824	31	1327740.25	**Normal**	**0**
0.004066	tcp	http	1040	824	31	1790457.375	**Normal**	**0**
0.132404	tcp	-	4862	77276	31	290323.5625	**Normal**	**0**
4.413976	tcp	ftp	1284	1638	62	2231.094971	**Exploits**	**1**
0.623141	tcp	smtp	28292	1936	62	353666.3438	**Exploits**	**1**
1.360272	tcp	ftp	1210	1662	62	6792.758789	**Exploits**	**1**
1.252438	tcp	ftp-data	364	740	62	2037.625854	**Exploits**	**1**
7.717545	tcp	smtp	3235	2048	254	3242.481934	**DoS**	**1**

Methodology: A Multi-Agent Framework Overview



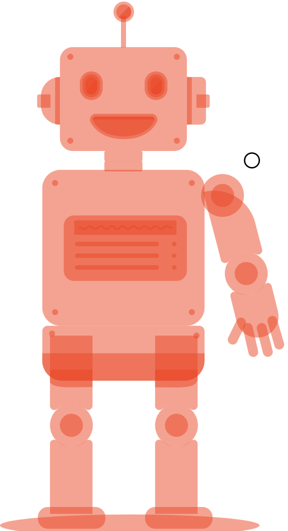
Agent 1: The Log Storyteller

- Objective: From Raw Numbers to a Rich Narrative
    - To transform a single, cryptic log entry (a row of numbers and codes) into an analytical, human-readable "story."
    - The goal is to provide context, not just data.
- Core Mechanism: Contextualization via Baseline Comparison
    - Agent 1 compares each feature from an incoming log against a pre-computed statistical baseline derived from **"Normal" traffic**.
    - This process turns raw data into meaningful insights by quantifying the **deviation from the norm.**



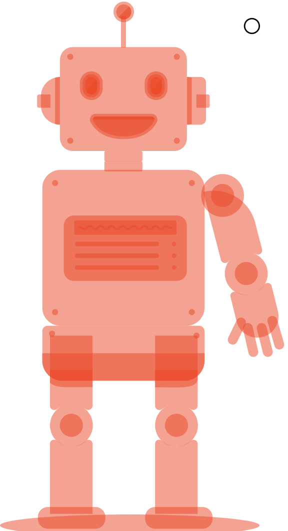
Agent 1: The Log Storyteller

- **Implementation:** A Two-Step Process
- **Step 1 (Offline):** Building the "Normal" Baseline
    - First we selected 21 important features from all dataset features.
    - Then, we performed a statistical analysis exclusively on the **"Normal"** records from the training set.
    - This process involved computing and storing key statistics for our 21 selected features:
        - **For Numerical Features:** We calculated the **Mean**, Min, and Max.
        - **For Categorical Features:** We calculated the **frequency distribution** of each value (e.g., tcp appears in 75% of normal traffic).
    - This baseline serves as the algorithm's "memory" of what constitutes usual network behavior.



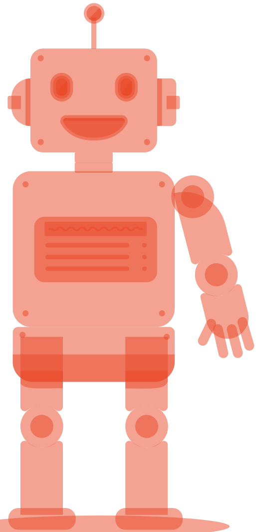
Agent 1: The Log Storyteller

- **Step 2 (Online):** The Story-Generation Algorithm
- For each new log, the Python algorithm compares its feature values against the stored baseline statistics:
    - **Numerical Comparison:** The log's value is compared to the stored Mean. The algorithm calculates the ratio (value / mean) to generate dynamic, qualitative phrases like ~88% below the Usual or several-fold higher.
    - **Categorical Comparison:** The log's value (e.g., udp) is looked up in the stored frequency distribution. The algorithm then uses this percentage to generate phrases like observed in ~22% of the Usual set.



Agent 1: The Log Storyteller

- **An Example:**
  - Input (A Single Row from UNSW-NB15):

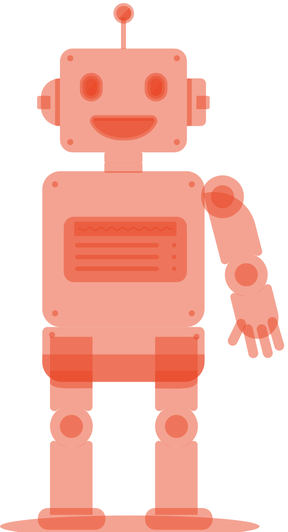


Input (Features)

dur: 0.02595, **proto**: tcp, **service**: -, **state**: FIN, **spkts**: 48, **dpkts**: 50, **sbytes**: 2974, **dbytes**: 30506, **rate**: 3737.957655, **sttl**: 31, **dttl**: 29, **sload**: 898034.6875, **dload**: 9216493, **sloss**: 7, **dloss**: 18, **sinpkt**: 0.556935, **dinpkt**: 0.550848, **sjit**: 37.603086, **djit**: 39.067097, **swin**: 255, **stcpb**: 2509677024, **dtcpb**: 364595143, **dwin**: 255, **tcprtt**: 0.000768, **synack**: 0.000603, **ackdat**: 0.000165, **smean**: 62, **dmean**: 610, **trans_depth**: 0, **response_body_len**: 0, **ct_srv_src**: 11, **ct_state_ttl**: 0, **ct_dst_ltm**: 2, **ct_src_dport_ltm**: 1, **ct_dst_sport_ltm**: 1, **ct_dst_src_ltm**: 2, **is_ftp_login**: 0, **ct_ftp_cmd**: 0, **ct_flw_http_mthd**: 0, **ct_src_ltm**: 6, **ct_srv_dst**: 7, **is_sm_ips_ports**: 0

Agent 1: The Log Storyteller

- **An Example:**
  - Output (The Generated "Story"):



Output (The Generated "Story")

This record describes a single network flow using tcp (Transmission Control Protocol; observed in ~75% of the Usual set); the application service was identified as http (web (unencrypted)) (observed in ~11% of the Usual set), and the connection state recorded as FIN (graceful close observed). As a combination, the proto↔service pairing is pairing frequency not available. The flow duration was ~1% below the Usual. source→destination data volume was ~61% below the Usual and destination→source data volume was ~46% below the Usual. source→destination packet count was ~47% below the Usual and destination→source packet count was ~29% below the Usual. source→destination transfer rate was very lower than the Usual and destination→source transfer rate was ~94% below the Usual. source→destination hop limit (TTL) was ~75% below the Usual and destination→source hop limit (TTL) was ~73% below the Usual. source→destination packet loss was ~30% below the Usual and destination→source packet loss was ~45% below the Usual. source TCP window was ~+33% vs the Usual. destination TCP window was ~+41% vs the Usual. source TCP base sequence was ~77% below the Usual. destination TCP base sequence was ~+66% vs the Usual. SYN-to-SYN-ACK time was ≈ 0.51 ms. SYN-ACK-to-ACK time was ≈ 0.14 ms. TCP handshake round-trip time was ≈ 0.66 ms.

Evaluation

- **Baseline Performance with In-Context Learning (ICL)**
  - We first evaluated the Reasoner agent using only ICL to measure the impact of the "Story" format without fine-tuning.
  - **Raw Logs:** Prompting a large LLM with raw log data resulted in random, coin-flip performance (~50% Accuracy).
  - **Story (Zero-shot):** Simply converting the log to a "Story" immediately boosted accuracy significantly.
  - **Story (Few-shot):** Adding a few examples provided the best ICL results, but performance was still not sufficient for a reliable security tool.

Comparison to prior work (UNSW-NB15)

Gemma-4b-it

In-Context Learning (ICL) Results

Gemma-3 4B IT performance with different log formats

Performance Comparison

Gemma-3 4B IT (Raw Logs)

⚠ **Random Performance (50% accuracy)**

Model performs at chance level - equivalent to coin flipping

Precision, Recall, F1-Score are meaningless at this performance level

Gemma-3 4B IT (Story - Zero-shot)

Accuracy	Precision	Recall	F1-Score
64.9%	65.6%	63.1%	64.4%

Gemma-3 4B IT (Story - Few-shot)

Accuracy	Precision	Recall	F1-Score
67.1%	60.3%	100.0%	75.2%

Gemma-7b-it

Dataset	Accuracy	Precision	Recall	F1 Score	Remarks
CICIDS2017	0.6600	**0.7222**	0.5200	0.6047	Experiment 1, 2[a]
CICIDS2017	**0.7600**	0.6970	**0.9200**	**0.7931**	Experiment 2, 3[b]
CICIDS2017	0.6800	0.6216	**0.9200**	0.7419	Experiment 3[c]
CICIDS2017	0.5000	0.000	0.000	0.000	Experiment 1[d]
KDD Cup 1999	0.7600	**1.0000**	0.5200	0.6842	Experiment 1, 2[a]
KDD Cup 1999	**0.9800**	**1.0000**	0.9600	**0.9796**	Experiment 2, 3[b]
KDD Cup 1999	0.7800	0.6944	**1.0000**	0.8197	Experiment 3[c]
KDD Cup 1999	0.5000	0.000	0.000	0.000	Experiment 1[d]
UNSW-NB15	0.4200	0.3750	0.2400	0.2927	Experiment 1, 2[a]
UNSW-NB15	0.6000	0.6471	0.4400	0.5238	Experiment 2, 3[b]
UNSW-NB15	**0.6400**	**0.6842**	**0.5200**	**0.5909**	Experiment 3[c]
UNSW-NB15	0.5000	0.000	0.000	0.000	Experiment 1[d]

Zhao, X., Leng, X., Wang, L., ***et al.*** *"Efficient anomaly detection in tabular cybersecurity data using large language models."* **Scientific Reports** *15, 3344 (2025). https://www.nature.com/articles/s41598-025-88050-z*

ICL Performance Analysis

Key Insights

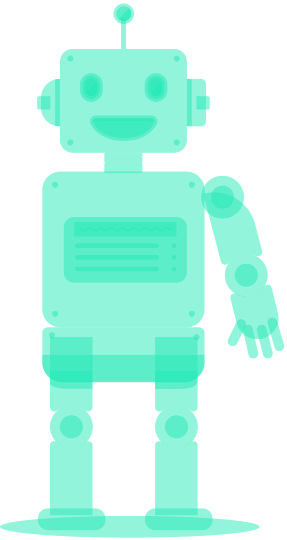
- Raw logs: Random performance (50% - coin flip level)
- Story format (zero-shot): +14.9% accuracy improvement
- Story format (few-shot): +17.1% accuracy improvement
- Best ICL F1-Score: 75.2% with few-shot story format

ICL Detailed Performance Analysis

METHOD	INPUT FORMAT	ACCURACY	PRECISION	RECALL	F1-SCORE	STATUS
Raw Logs	Raw Network Logs	50.0%	Random	Random	Random	Random Performance
Zero-shot	Structured Story	64.9%	65.6%	63.1%	64.4%	Improved
Few-shot	Structured Story	67.1%	60.3%	100.0%	75.2%	Best ICL

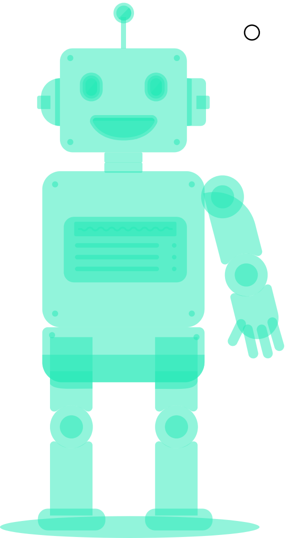
Agent 2: The Attack Reasoner

- **Objective:** Moving from "What" to "Why"
  - To go beyond a simple Attack/Normal label and provide a concise, evidence-based **"Reason"** for each classification.
  - This transforms a black-box detection into an explainable analysis.



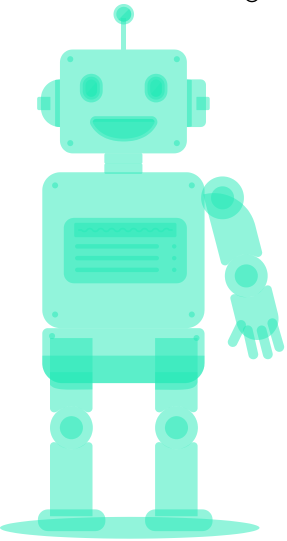
Agent 2: The Attack Reasoner

- **Implementation:**
- **Step1**:Curating a Diverse Set of Stories
  - To ensure quality and prevent redundancy, we first clustered all "Stories" using sentence embeddings.
  - We then sampled from each cluster to create a diverse and representative training set of ~4,000 high-quality examples.
  - To clarify, this curated set is **highly efficient**, constituting just **3%** of the total records in the original training dataset.



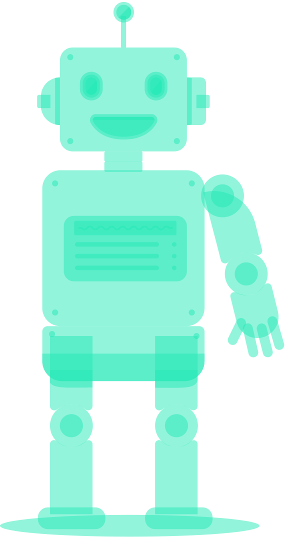
Agent 2: The Attack Reasoner

- **Step 2:** Generating the Reasoning Dataset (The "Teacher")
  - We feed the powerful "Teacher" model (e.g., GPT-5) the **curated stories** along with their correct ground-truth **“attack category“**.
  - The Teacher's task is to act as a security expert and generate a high-quality, professional **"Reason"** for each story.
  - The output is our final training dataset of **Story → {Reason, label}.**



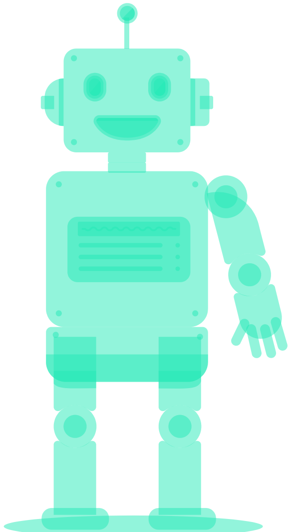
Agent 2: The Attack Reasoner

- **Step 3:** Fine-Tuning the Specialist (The "Student“)
  - The curated **Story → {Reason, label}** dataset is used to fine-tune the smaller, efficient "Student" model (Gemma 3 1B & 4B).
  - The Student model learns to mimic the expert reasoning patterns of the Teacher, becoming a highly specialized agent for this task.



Agent 2: The Attack Reasoner

- An Example:



Input (A "Story" from Agent 1):

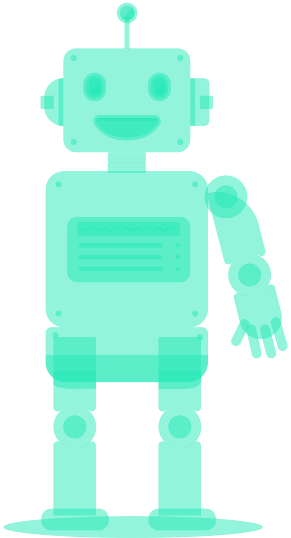
This record describes a single network flow using tcp (Transmission Control Protocol; observed in ~75% of the Usual set); no specific application service was identified (observed in ~74% of the Usual set), and the connection state recorded as FIN (graceful close observed). As a combination, the proto↔service pairing is pairing not applicable. The flow duration was very lower than the Usual. source→destination data volume was ~27% below the Usual and destination→source data volume was ~+63% vs the Usual. source→destination packet count was ~+111% vs the Usual and destination→source packet count was ~+97% vs the Usual. source→destination transfer rate was very lower than the Usual and destination→source transfer rate was ≈ 6.7× the Usual (many times higher). source→destination hop limit (TTL) was ~75% below the Usual and destination→source hop limit (TTL) was ~73% below the Usual. source→destination packet loss was ~+63% vs the Usual and destination→source packet loss was ~+99% vs the Usual. source TCP window was ~+33% vs the Usual. destination TCP window was ~+41% vs the Usual. source TCP base sequence was ~+65% vs the Usual. destination TCP base sequence was ~76% below the Usual. SYN-to-SYN-ACK time was ≈ 0.60 ms. SYN-ACK-to-ACK time was ≈ 0.17 ms. TCP handshake round-trip time was ≈ 0.77 ms.

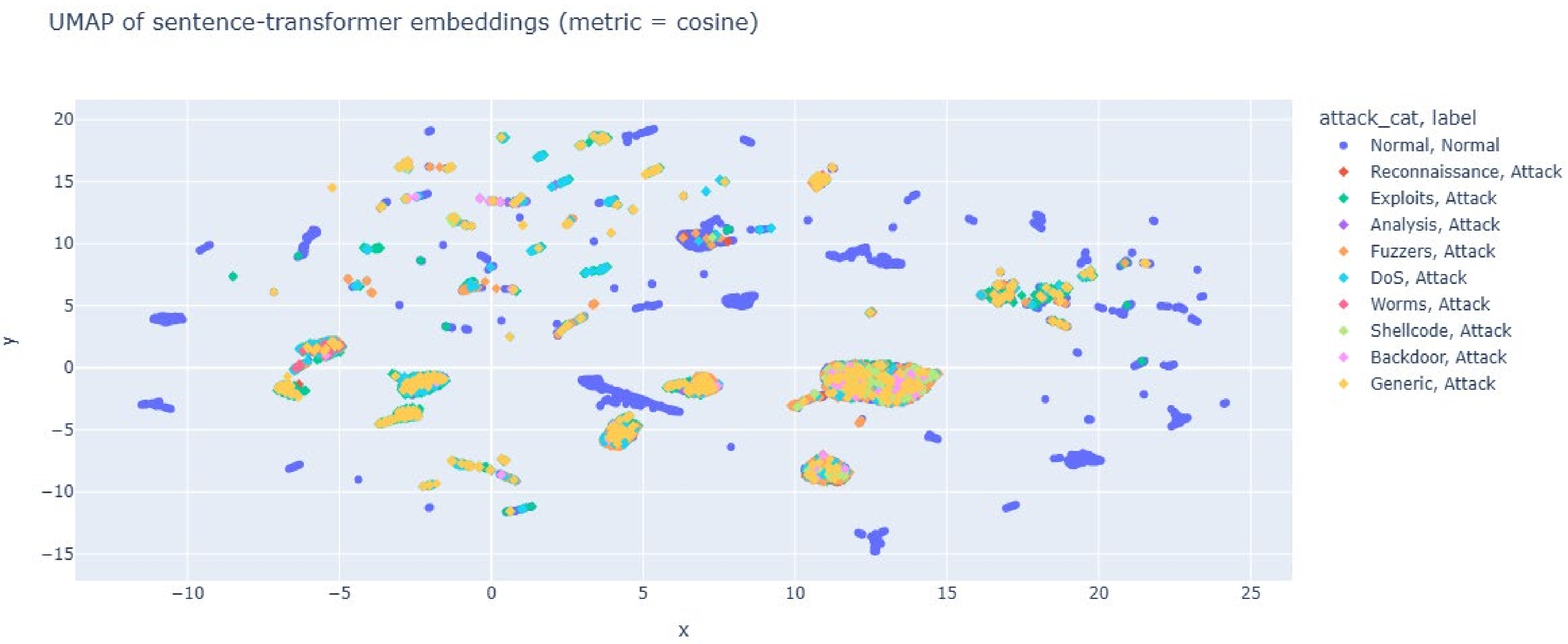
Agent 2: The Attack Reasoner

- An Example:

Output (The final JSON from Agent 2):

"reason": "Despite rate and TTL anomalies, the graceful TCP close and healthy handshake timings suggest a legitimate, non-malicious data exchange.",
"label": "normal"




UMAP of sentence-transformer embeddings (metric = cosine)
attack_cat, label
Normal, Normal
Reconnaissance, Attack
Exploits, Attack
Analysis, Attack
Fuzzers, Attack
DoS, Attack
Worms, Attack
Shellcode, Attack
Backdoor, Attack
Generic, Attack
x
y
−10
−5
0
5
10
15
20
25
20
15
10
5
0
−5
−10
−15

Evaluation: Fine Tuning Results

Accuracy

Best: Gemma 3 1B

Gemma 3 1B	93.0%
GPT OSS 120B	59.7%
Qwen 32B	58.9%

F1 Score

Best: Gemma 3 1B

Gemma 3 1B	93.0%
GPT OSS 120B	58.1%
Qwen 32B	65.4%

Precision

Best: Gemma 3 1B

Gemma 3 1B	93.2%
GPT OSS 120B	60.5%
Qwen 32B	56.5%

Recall

Best: Gemma 3 1B

Gemma 3 1B	92.8%
GPT OSS 120B	55.8%
Qwen 32B	77.8%

Comparison to prior work (UNSW-NB15)

Model Performance Comparison

UNSW-NB15 Binary Classification Results

MODEL	SIZE	TEST	ACCURACY	PRECISION	RECALL	F1-SCORE
Zhao et al. (2024) Gemma-2 2B IT + LoRA	2.0B	300	86.67	82.50	91.67	86.84
Our Approach Gemma-3 1B IT + LoRA	1.0B (-50%)	1,000 (+233%)	93.00 +6.33	93.20 +10.70	92.80 +1.13	93.00 +6.16

Best Performance: Our approach outperforms across all metrics **Efficiency:** 50% smaller model with superior results

Zhao et al., Biomimetics (MDPI), 2024: https://doi.org/10.3390/biomimetics9110708

Intrusion Class Breakdown — Strengths & Gaps

Gemma-3 1B IT (finetuned) — 3% training subset, test n=1,000.

Performance by Attack Type

ATTACK TYPE	TOTAL	TRUE POSITIVE	FALSE NEGATIVE	ACCURACY
Analysis	6	4	2	66.7%
Backdoor	7	7	0	100.0%
Dos	53	52	1	98.1%
Exploits	152	146	6	96.1%
Fuzzers	68	45	23	66.2%
Generic	165	165	0	100.0%
Reconnaissance	42	42	0	100.0%
Shellcode	7	3	4	42.9%
Normal (TN/FP)	500	466	34	93.2%

Conclusion

- Summary of Our Contribution
  - We successfully developed a novel **2-agent framework**, "From Logs to Explainable Verdict," that transforms cryptic network logs into clear, evidence-based security insights.
  - Our hybrid approach effectively combines a **rule-based algorithm** (Agent 1) for data contextualization with a **fine-tuned small LLM** (Agent 2) for expert-level reasoning.
- Key Findings
  - The **"Story" format** is critical: it bridges the semantic gap, making log data intelligible to LLMs and dramatically improving performance.
  - Our fine-tuned **Gemma 3 1B** model achieved **~93% balanced accuracy and F1-score**, vastly outperforming massive, general-purpose LLMs.
  - This high performance was achieved by fine-tuning on just **3%** of the training data, proving the efficiency of our curation method.
  - The final model is highly **reliable**, with low rates of both false positives and false negatives, making it suitable for real-world use.