Who Is Really Answering? Auditing LLMs from the Outside

One-page report | CRANE North East Regional Cyber Security Seminar

Speaker: Dr Eaby Babu **Chair:** Dr Biju Issac **Date and format:** 2 November 2026 | Northumbria University and online

15	144	21/21
experiments	open-weight models	model-swap tests

Overview

Dr Eaby Babu presented a year-long investigation of how to audit large language models (LLMs) using only API-visible behaviour, without access to proprietary weights, logs or vendor cooperation. Introduced by Dr Biju Issac, the talk addressed three assurance questions: which model is responding, what can be inferred about its architecture, and whether it follows trusted instructions when faced with attacks. Experiments covered 144 open-weight and eight closed-box models, alongside commercial API evaluations.

Principal research findings

Model identification and substitution. Analysis of streaming responses combined 30 timing and 15 response features, yielding around 72% model-identification success in the reported setting. All 21 model-swap tests detected an endpoint change. However, network latency and deployment conditions significantly distorted timing patterns, limiting reliable transfer from controlled environments to production APIs.

Architectural inference. Structured string and capacity probes offered clues to model family and construction. The speaker reported about 99% identification of positional-encoding class and 70% model-family inference in specific controlled tests. Internal properties, including layers, width and attention heads, were harder to estimate accurately; the results represent statistical predictions, not verified architecture disclosures.

Prompt-injection resilience. The Prompt Injection Resistance and Recovery (PIRR) benchmark examined malicious instructions hidden in untrusted documents, persistence across turns, attempted secret leakage and recovery. Tests spanned 20 attack families and three mitigation settings. No canary-secret leakage was observed in the reported trials; mitigation generally improved outcomes, while tool-calling actions remained a notable risk. Results were specific to the tested models and configurations.

Live demonstration and discussion

The locally hosted Inference Hub integrated model comparisons, architectural probes, swap monitoring and PIRR evaluations. During an LLM Hunt demonstration, Claude Haiku 4.5 was correctly ranked as the most likely model from a live trace, but at only about 35% confidence, underscoring uncertainty. The dashboard displayed token timing, latency and connection measurements. Audience questions, including those from Professor Thomas Gross, explored rapid model evolution, common assurance standards, interception and data-exfiltration risks, evaluation design and compute demands. The team described local GPUs, cloud models and access to the Dawn supercomputer.

Implications and conclusion

Organisations integrating third-party LLM APIs should independently check model provenance, monitor endpoint changes, and test prompt-injection defences across the full tool-enabled workflow. The research demonstrated the value of external auditing while making its limits clear: network effects complicate fingerprinting, and observable behaviour cannot conclusively reveal internal model design. The session supported CRANE's aim of making advanced cybersecurity research accessible and fostering regional collaboration.