David Pierce, MPA

BLUF External Observability, Internal Explainability and Automation drive AI Safety

Executive Summary: Affecting Enterprise AI/ML Observability & Safety is simultaneously simple in concept and complex in execution. The high level can be easily summarized by the development and integration of the four product categories listed below, with the Core Services and Capabilities referenced

1) Per-Application Baselines:

- Allow for use of ‘known good’ and ‘known bad’ examples in Real-Time Monitoring
- Allow for use across deployment patterns (e.g. chatbot, AI agent monitoring, etc)
- Extensible for monitoring of other modalities (e.g. Audio, Image, etc) via abstraction layer

2) Real-Time Monitoring:

- Leverages efficient stepwise evaluation against known bad/good examples
- Configurable with both lightweight and heavyweight non-conformity predictions
- Configurable Application-Specific Thresholds, Sanitization, Post-Processing, and Forecasting
- Technology Agnostic Deployment of Functions, Microservices, and Key-Value Pair DB

3) Automated Playbooks:

- Drag’n’Drop (e.g. from HuggingFace) Whitebox Attack Suite w/ Configurable Test Criteria
- Allows for Continual Validation of Baselines & Monitoring by Offensive Cyber Teams
- Pairs with Customized Kill-Chain & Defense Plan informed by MITRE ATLAS Framework
- Informed by SOTA Research Publications from Carnegie Mellon[1], DeepMind[1] & Anthropic[2,3]

4) Customizable Instrumentation:

- Customized Transformer Heads act as Linear Probes to understand per-layer activations
- Configurable Expansion of Model Residuals into High Dimensional Vector Space
- Allows for Automated Alerting and Reporting of activations based on feature space
- Informed by SOTA Research Publications[4] and from ML Alignment & Theory Scholars[5]

Core Competencies:

- External Observability - ‘Managing Alerts at-Scale by Exception’
- Internal Explainability - ‘Detailed Audit of AI Decisions’
- Per-Component Securability - ‘Minimizing Residual Risk’

Paired w/ Core Services:

- AI Software Integration & Strategic Planning
- Data Pipeline & Platform Modernization
- Cross-Domain AI Upskilling for Existing Teams
- Audit & Regulatory Support for Business

Citations
1) Universal and Transferable Vulnerabilities
2) Introducing Hidden & Persistent Backdoors
3) Universal Vulnerability & Toxicity Scaling
4) Mediating LLM Refusal via Orthogonalization
5)ML Alignment & Theory Scholars Program