WHYLABS

# In-Line Security & Safety for Large Language Models

**WhyLabs provides in-line LLM protection against OWASP's Top 10 LLM threats both inside and outside the enterprise.**

Detect and stop unwanted traffic, such as jailbreaks and injections, from reaching your LLMs. Prevent LLM data leakage or hallucinations from degrading your customer's experience.

## Identify & Prevent:

- Prompt Injection
- Insecure Output Handling
- Data Leakage
- Overreliance
- Insecure Plugin Usage
- Toxic Interactions
- Hallucination-like Scenarios


## Guardrails

- Control which prompts and responses are appropriate for your LLM application in real time. Define a set of boundaries that you expect your LLM to stay within, detect problematic prompts and responses based on a range of metrics and block or take other action as needed.

## Tuning

- Tune your protections by logging responses and seeing what actions the reverse proxy would take, once enabled. Look for false positives and tweak the actions taken to protect your LLM model against each type of threat. WhyLabs provides key telemetry data that enables you to build and then measure against baselines over time.

## Highly Scalable & Privacy First

- Because WhyLabs only compares statistical profiles of model behavior over time, it's highly scalable and privacy first. No sampling required!

*"As our team helps enterprises put this powerful technology into practice, safety remains one of the main blocks for widespread adoption. WhyLabs' LangKit is a leap forward for LLM Ops, providing out-of-the-box tools for measuring the quality of LLM outputs and catching issues before they affect tasks downstream."* - Alan Descoins, CTO at Tryolabs

## Internal Use Cases

- Ensure LLM models are providing accurate information to customer support agents.
- Guard against off-topic inquiries to LLM models.
- Protect against rogue employees poisoning your LLM models and causing lost productivity.

## Public Facing Use Cases

- Protect against denial of service attacks.
- Protect your brand by preventing toxic or hallucinatory responses.

## Security Dashboard

- Gain insight on security threats.
- Decide which requests and responses to block, log, or alert on.
- Understand why actions are taken.

## See WhyLabs LLM Security in Action

Contact info@whylabs.ai for a demo.

CBINSIGHTS
Generative AI
50
2023