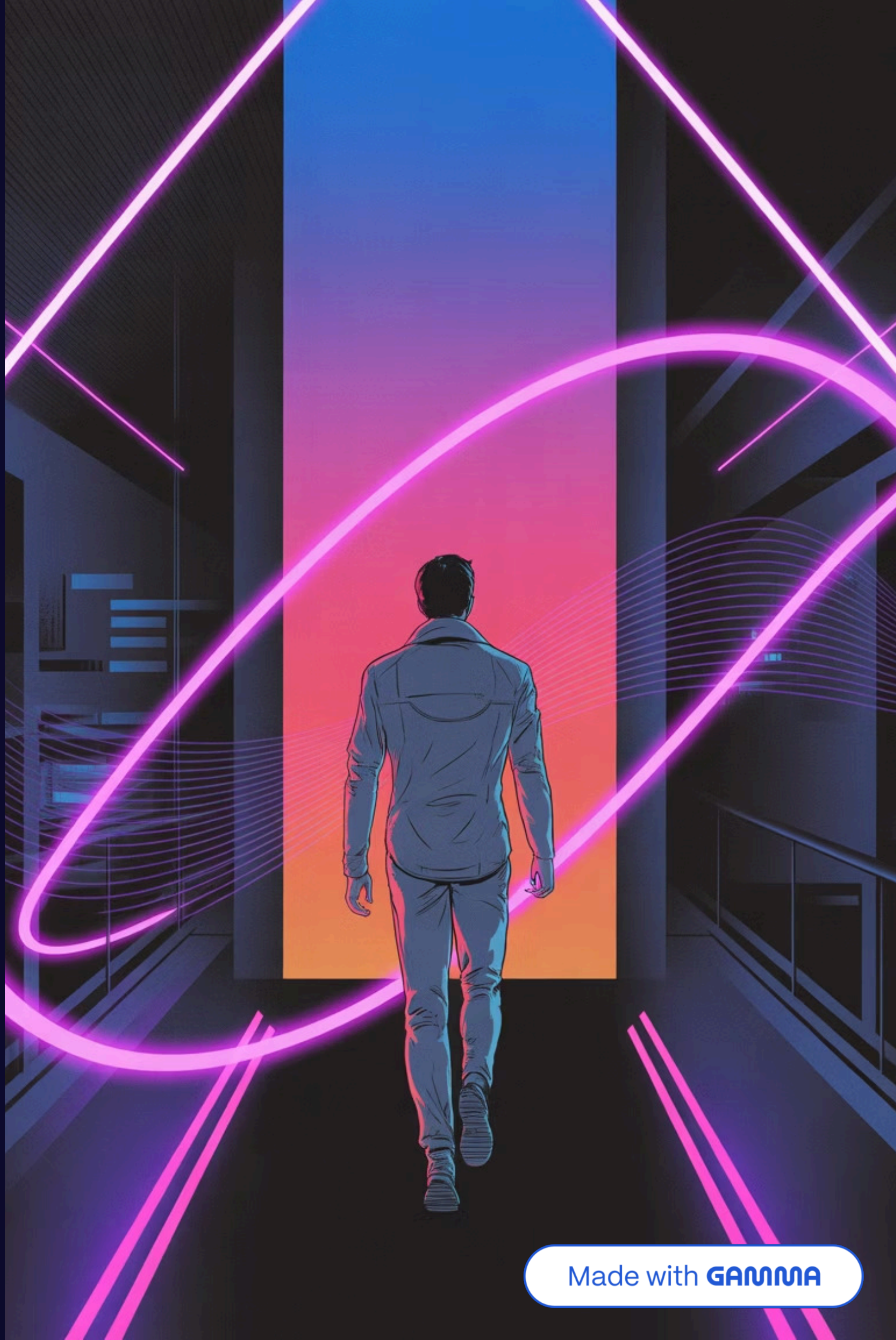
BITTENSOR SUBNET

Gauntlet

A Bittensor Subnet for Classifier Adversarial Robustness

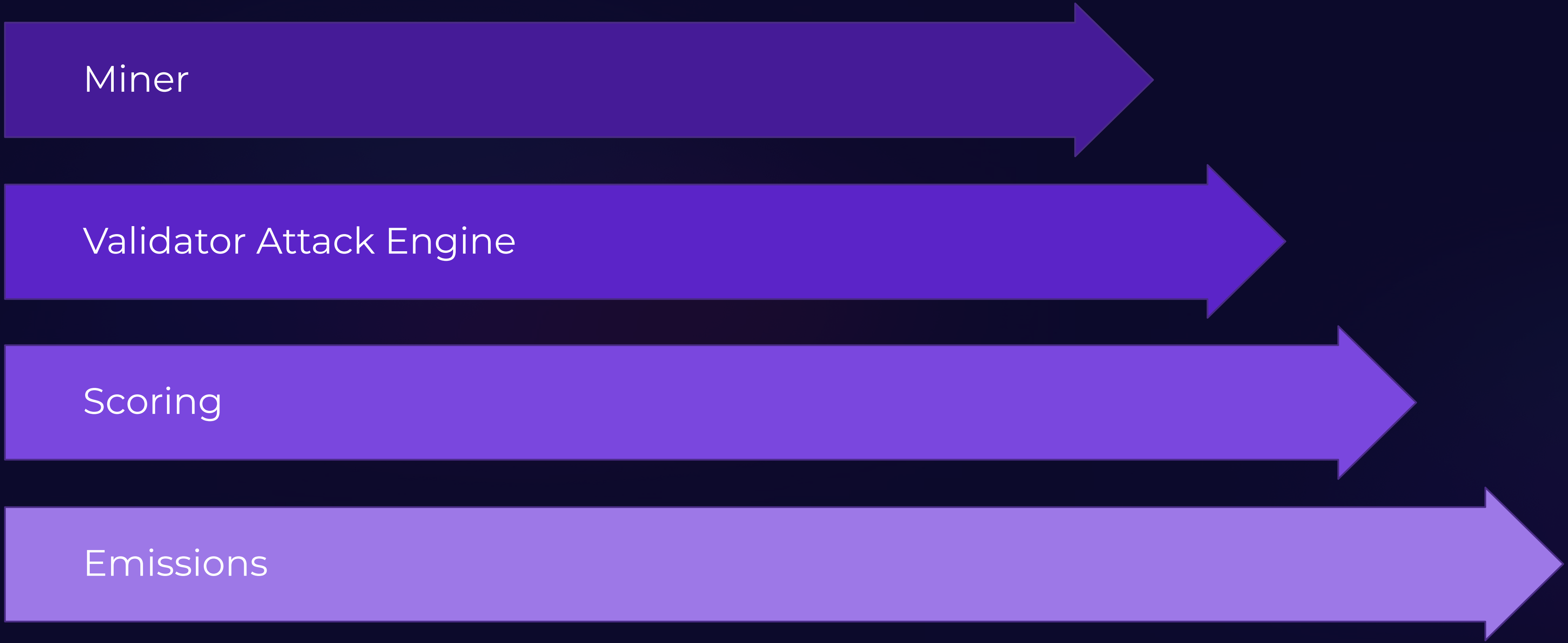
"Proof of Intelligence Through Pressure."



The Problem

- Most AI models are brittle under adversarial attack
- Security testing is centralized and static
- No continuous robustness benchmark
- AI systems are increasingly deployed in high-stakes environments

The Gauntlet Concept



Miner

Validator Attack Engine

Scoring

Emissions

The Gauntlet creates a continuous adversarial loop where **Miners** host robust classifiers, **Validators** generate adversarial attacks to test them, and **Emissions** reward robustness — driving an ever-escalating arms race of intelligence.

Architecture

Hidden Dataset

Hidden evaluation sets ensure miners cannot overfit to known data

Validators

Adaptive attack engine probes for weaknesses

Miners

Host classifiers and respond to queries

Scoring Engine

Epoch-based scoring evaluates performance

Emission Allocation

Rewards distributed based on robustness

Hidden evaluation sets

Adaptive attack engine

Epoch-based scoring

Miner Design

Core Responsibilities

→ Host classifier API

→ Accept batch inputs

→ Return predictions + latency

→ Optimize for robust accuracy

Performance Dimensions

Dimension	Weight
Robust Accuracy	High Weight
Clean Accuracy	Medium
Latency	Medium
Consistency	Medium

Validator Design

1 Generate adversarial attacks

FGSM, PGD, AutoAttack

2 Evaluate accuracy drop

Measure how much model performance degrades

3 Submit perturbations + logs

Full transparency of attack methodology

4 Compete to discover weaknesses

Validators are incentivized to find vulnerabilities

Emission Mechanism

> The scoring formula that drives the Gauntlet economy:

$$Score = \alpha \cdot A_{adv} + \beta \cdot A_{clean} - \gamma \cdot LatencyPenalty$$

$$Emission = \frac{Score^{\tau}}{\sum Score^{\tau}}$$

Robust accuracy weighted highest

Adversarial performance is the primary driver of rewards

Temperature sharpens competition

The τ parameter concentrates emissions toward top performers

CORE THESIS

Why This Is Proof of Intelligence

"Intelligence that survives attack."

Requires adversarial training

Models must be deliberately hardened against attack vectors to earn emissions

Resists adaptive gradient attacks

Robustness must hold against evolving, sophisticated attack strategies

Penalizes gradient masking

Superficial defenses that hide gradients are detected and punished

Continuous competitive pressure

The adversarial arms race never stops — only the truly robust survive

Epoch Flow

01

Sample hidden batch

Draw evaluation samples from the hidden dataset that miners have never seen

02

Query miner

Send the batch to the miner's classifier API and collect predictions

03

Generate adversarial samples

Validators craft adversarial perturbations targeting the miner's model

04

Measure clean & adversarial accuracy

Compare performance on original vs. perturbed inputs

05

Compute score

Apply the scoring formula to determine the miner's epoch performance

06

Distribute emissions

Allocate rewards proportional to normalized scores across all miners

Market Rationale

The Opportunity

- AI security is underdeveloped
- Enterprises need robustness certification
- No decentralized robustness oracle

Future Potential

AI insurance input

Robustness scores as underwriting data for AI liability coverage

Security scoring API

Enterprise-grade adversarial robustness assessments on demand

On-chain robustness oracle

Decentralized, verifiable AI security benchmarks for the ecosystem

Why This Belongs on Bittensor

→ Incentivized competition

→ Emissions reward measurable performance

→ Adversarial co-evolution

→ Decentralized red-teaming

Gauntlet

Run the Gauntlet.

Continuous adversarial benchmarking

Proof of resilience

The security layer for AI