MODEL ASSURANCE SOLUTION

Determine where a model came from. *Confirm it has not changed.*

The Model Assurance Solution verifies the AI model inside a supplied system. It assesses the model once at source selection, confirms the installed model matches in an installation audit, and then monitors it continuously as an ongoing service while the system is in operation.

INSPECTION POINTS · EXAMPLE: VENDOR-SUPPLIED DRONE

Three inspection points. One baseline.

01 · SOURCE SELECTION

Evaluate the offered model

- Vendor submits the model
- Full assessment: weight analysis and behavioral fingerprint
- Model approved or rejected

METHOD · ● BEHAVIORAL ● WEIGHT
OUTPUT · EVIDENCE REPORT
RUNS · ONCE

02 · INSTALLATION AUDIT

Verify the installed model

- Drone delivered with model on board
- Behavioral fingerprint taken
- Matched to the stage 01 evidence report

METHOD · ● BEHAVIORAL ONLY
OUTPUT · MATCH OR MISMATCH
RUNS · ONCE, AT DELIVERY

03 · CONTINUOUS MONITORING

Monitor in operation

- Service enabled on the drone, offline
- Behavioral fingerprint, ongoing
- Drift or change flagged locally

METHOD · ● BEHAVIORAL ONLY
OUTPUT · STABILITY STATUS
RUNS · ONGOING SERVICE

Weight analysis runs only at stage 01. Stages 02 and 03 match against the behavioral fingerprint in the stage 01 evidence report. Reference catalog updates are small vector files that can be moved across an air gap.

STAGE 01 · SOURCE SELECTION

A detailed provenance inspection.

1. **Behavioral fingerprint** VAIL
   How the model maps fixed inputs to outputs. Catches behavioral divergence weight analysis misses.
2. **Weight-derived signals** CISCO MPK
   Classifies direct descent, fine-tune, quantization, merger, or independent training.
3. **Tokenizer comparison** CISCO MPK
   Tokenizer structure and vocabulary overlap.
4. **Architecture metadata** CISCO MPK
   Layer layout, dimensions, configuration.

OUTPUT

Evidence report

Each measurement compared against a reference catalog of known models.

- ● Behavioral neighbors
- ● Weight-lineage matches
- ● Tokenizer and architecture matches
- ◐ Combined reading

BEHAVIORAL FINGERPRINT RETAINED AS THE BASELINE FOR STAGES 02–03

SPECIFICATION	MODEL ASSURANCE SOLUTION
FUNCTION	Model lineage and behavioral identity verification
APPLIES TO	Open-weight language models, including models embedded in delivered systems
METHODS	Behavioral fingerprinting (VAIL) at all stages; weight analysis (Cisco Model Provenance Kit) at source selection
REFERENCE	Catalog of known models; updates delivered as offline vector files
INSPECTION POINTS	01 Source selection 02 Installation audit 03 Continuous monitoring (ongoing service)
OUTPUT	Evidence report; recorded baseline per model
DEPLOYMENT	On-premises, air-gapped, or on-device
NETWORK	No outbound connections; weights do not leave the environment
RECORDS	Supports provenance records and AIBOM documentation

PUBLISHED RESEARCH

Validated, published methods.

cisco | VAIL

JOINT STUDY · CISCO AND VAIL · AUGUST 2026

Models don't have passports: AI lineage crosses organizational and geographic borders

Amy Chang, Ankit Garg (Cisco) · Manish Shah, Jonah Leshin (VAIL)

Applies both methods to NVIDIA Nemotron models, some documented as post-trained from Alibaba Qwen base weights. Each method independently recovers the documented lineage. Detection evidence, not proof of derivation.

	WEIGHT	BEHAVIORAL
Qwen-neighbor enrichment	1.74×	1.89×
Lineage order, p	0.00001	0.00002
Catalog size	184	1,159

VAIL · ACM CAIS 2026 · SYSTEM DEMONSTRATIONS

Behavioral Fingerprints for LLM Endpoint Stability and Identity

Jonah Leshin, Manish Shah, Ian Timmis, Daniel Kang

Periodically fingerprints a model from a fixed prompt set and compares output distributions over time. The method used for installation audit and continuous monitoring.

arxiv.org/abs/2603.19022

CHANGE TO THE MODEL	DETECTED
Model family or version	Yes
Inference stack	Yes
Quantization or parameters	Yes

JOINT STUDY · LUCID COMPUTING, LUNAL, VAIL · ICDS 2025

Hardware-Rooted Trust Anchors for Sovereign AI Processing

Milos Borenovic, Connor Dunlop, José Dunia (Lucid Computing) · Amean Asad (Lunal) · Manish Shah, Jonah Leshin (VAIL)

Uses VAIL's input–output fingerprint for runtime model integrity, alongside hardware trusted execution and location attestation. Stable across quantization; separates independently trained models.

	RESULT
Base vs. instruction-tuned	0.96–0.98
Qwen3-32B vs. quantized	> 0.91
Qwen3-32B vs. unrelated	0.63–0.84
Derivation threshold	0.85