ModelAudit — AI Model Security Scanner

Static security analysis for AI/ML model files. Detects malicious code, backdoors, and supply chain risks before models reach production — without ever loading or executing them.

The Problem

Organizations increasingly rely on pre-trained AI models from external sources — Hugging Face, cloud registries, vendor deliverables, and internal teams. These model files can contain executable code, embedded credentials, and hidden backdoors that traditional security tools are not designed to catch. A single compromised model file can lead to arbitrary code execution, data exfiltration, or persistent access the moment it is deserialized in your environment.

How ModelAudit Works

ModelAudit performs static analysis on model files, inspecting their contents byte-by-byte without executing any code. It identifies threats across the full spectrum of ML model formats, from high-risk serialization formats like Pickle and PyTorch to safer alternatives like SafeTensors. **30 specialized scanners** cover the formats and frameworks used across the ML ecosystem:

Risk Level	Formats
High	Pickle, PyTorch, Joblib, NumPy
Medium	TensorFlow, Keras, ONNX, XGBoost
Low	SafeTensors, GGUF/GGML, JAX/Flax, TFLite, TensorRT, PaddlePaddle, OpenVINO

Additional scanners cover archive formats (ZIP, TAR, 7-Zip, OCI layers), configuration files, and model metadata.

What It Detects

- **Code execution attempts** — dangerous operations embedded in serialized model files
- **Model backdoors** — hidden functionality and anomalous weight distributions
- **Embedded secrets** — API keys, tokens, and credentials in model weights or metadata
- **Network indicators** — URLs, IPs, and communication patterns suggesting data exfiltration
- **Archive exploits** — path traversal and symlink attacks in compressed model packages
- **Unsafe ML operations** — risky framework-specific constructs (Lambda layers, custom ops, JIT code)
- **Supply chain risks** — tampering indicators, suspicious configurations, known CVEs

Findings are classified by severity (Critical, Warning, Info) with context-aware assessments to minimize false positives.

Integration & Capabilities

ModelAudit fits into existing security and MLOps workflows:

Capability	Detail
CLI	Scan local files, directories, or remote sources from the command line
Remote registries	Scan directly from Hugging Face Hub, AWS S3, Google Cloud Storage, MLflow, JFrog Artifactory, and DVC
CI/CD pipelines	Deterministic exit codes (`0` clean, `1` issues, `2` errors) for automated gating
Output formats	Text, JSON, and SARIF (integrates with GitHub Advanced Security, VS Code)
SBOM generation	CycloneDX v1.6 ML Bill of Materials for supply chain compliance
CVE awareness	Flags known vulnerabilities in model serialization libraries
Large model support	Streaming mode for scanning models of any size with minimal disk usage
Docker	Containerized scanning with no local installation required
Platform support	Python 3.10–3.13 on Linux, macOS, and Windows

Getting Started

```
pip install modelaudit[all]

modelaudit ./models/
modelaudit https://huggingface.co/your-org/your-model
modelaudit s3://your-bucket/models/ --format sarif --output results.sarif
```

Website: promptfoo.dev/docs/model-audit | **Contact:** promptfoo.dev