Model Card — classifier-ensemble-v3
==========

Training cutoff: 2026-03-01
Adversarial eval: passed (n = 3,420)

Ensemble of 3 classifiers voting on hook / MCP / extension risk classification. FP rates: hook ~8%, mcp ~4%, other ~2%. On-device inference. No prompt or evidence egress.

DRAFT ARTIFACT
This is a preview-build placeholder.
The signed, auditor-attested artifact is delivered from the
Unbound admin Trust Center upon PO signing.

Preview-build hash: (stubbed for prototype)
Generated: 2026-04-17T21:00:01Z

Unbound Security · Trust Center
https://unboundsecurity.ai/trust