

CLAUDE OPUS 5 MAY FALL BACK TO OPUS 4.8 FOR FLAGGED CYBER REQUESTS

OPERATIONAL SIGNAL

DOMAIN: CAP — Capability | CONFIDENCE LEVEL: Medium-High

INSTITUTIONAL DIMENSIONS: Authority | Attribution | Visibility | Escalation | Continuity

Signal ID	Platform	Source	Audience	Status	Date
SIG-CAP-013	Claude Opus	Link	Security IT Risk Ops	Pilot	2026-07-29

CAPABILITY SHIFT

Anthropic released Claude Opus 5 on 24 July 2026. Its cyber classifiers block a narrow range of requests, including binary-based vulnerability scanning, penetration testing and exploit generation. On Claude.ai, Claude Code and Claude Cowork, flagged requests fall back to Opus 4.8 by default. On the Claude API, organisations can choose whether to enable automatic fallback, which is in beta.

ORGANISATIONAL CONSEQUENCE

A request made to Opus 5 may be answered by Opus 4.8. This can affect workflows that depend on a specific model's behaviour, evaluation or approval. API responses identify the model that answered, but Anthropic's announcement does not explain how fallback is presented on its first-party user surfaces. Approved Cyber Verification Program organisations may encounter fewer blocks for legitimate dual-use security work.

WHAT TO WATCH

- Opus 4.8 appearing where Opus 5 was requested
- Legitimate security tasks receiving refusals or fallback responses
- Security teams reporting different results across Claude organisations

WHY THIS MATTERS NOW

Opus 5 became available on 24 July 2026. Fallback is active by default on Claude.ai, Claude Code and Claude Cowork. API customers choose whether to enable it.

WHAT TO REVIEW

- Which workflows require Opus 5
- Whether Opus 4.8 is an approved fallback
- Whether API fallback is enabled
- Whether logs identify the serving model
- Whether security teams need Cyber Verification Program access

GOVERNANCE CONTEXT

The model selected may not be the model that answers. On Anthropic's first-party surfaces, flagged cyber requests can fall back by default. API customers control fallback and can inspect the serving model. Organisations should define where fallback is acceptable.

Scope Boundary: This artefact translates observed platform capability changes into operational considerations. It does not constitute policy, legal advice, or enforcement guidance.