

14 of 14 controlled failures diagnosed

A dated production-path benchmark achieved 100.0% strict and semantic accuracy on the covered Kubernetes failure corpus. This is bounded, reproducible evidence - not a claim that every Kubernetes incident will be diagnosed correctly.

Published: 23 August 2026

Strict accuracy	Semantic accuracy	Transport reliability	Publication gate
14/14 (100.0%)	14/14 (100.0%)	14/14 (100.0%)	PASS (threshold 95%)

Evidence validity

Item	Recorded value
Run time (UTC)	2026-08-23T05:33:10.380289+00:00
Snapshot age	24.7 seconds (maximum accepted: 180 seconds)
Application commit	7144d8a881936185e427aa15da4a3096e9618088
Agent / chart	2.0.32 / 0.1.32
Kubernetes server	v1.36.1

Covered scenarios

Scenario	Expected diagnosis	Result
Nonexistent registry/image	Image pull failure	PASS
Readiness wrong port	Probe failure	PASS
Liveness kill loop	Crash loop + probe failure	PASS
Node selector mismatch	Node selector failure	PASS
Insufficient resources	Scheduling/capacity failure	PASS
Missing ConfigMap	Missing configuration	PASS
Missing Secret	Missing secret	PASS
Unbound PVC	Storage/PVC binding failure	PASS
Init container crash	Init container failure	PASS

Scenario	Expected diagnosis	Result
DNS failure	DNS/dependency failure	PASS
Wrong command	Invalid runtime configuration	PASS
CPU throttling	CPU throttling/resource pressure	PASS
Missing ServiceAccount	Missing identity dependency	PASS
NetworkPolicy egress block	Network policy/dependency failure	PASS

Scoring and limits

- Strict accuracy requires the expected structured diagnosis identifier.
- Semantic accuracy accepts the documented equivalent when multiple related causes are returned.
- The gate also requires fresh evidence, successful API transport and remediation evidence quality.
- The corpus is controlled and intentionally bounded. It does not measure every production topology, intermittent dependency or multi-cause incident.

REPRODUCE

Review the full methodology and per-scenario evidence at krevopilot.com/docs-benchmark.html. The public page records the run, versions, scoring method and explicit limitations.