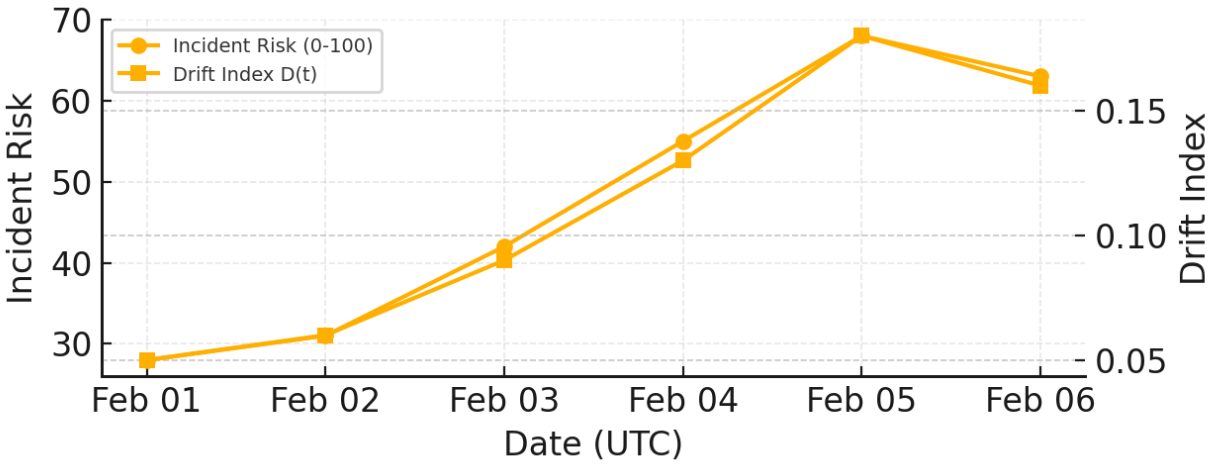
Agent Stability Monitor — Evidenzpaket

3-seitiger Evidenz-Export für output-only Stabilitätsmonitoring eines LLM-Agenten. Dieses Dokument enthält anonymisierte, synthetische Beispieldaten, um Format und Inhalt eines kundenfähigen Exports zu demonstrieren.

Stabilitätsstatus	AMBER (DriftWatch)
Audit-Fenster (UTC)	2026-02-01 → 2026-02-06
Scope	Ein Agent, Black-Box API, output-only Diagnostik; 6 Tage
Top-KPIs	Incident Risk / Drift Index / Evidence Completeness
Payload SHA-256	e04e201e12395ba5…

KPI	Aktuell	Trend (ggü. Vortag)	Lesart
Incident Risk (0–100)	63	↓	Komposit-Risiko für Instabilität/Incident innerhalb des Horizonts
Drift Index D(t)	0.16	↓	Trajektorien-Drift aus output-only Observables
Evidence Completeness	94%	↑	Abdeckung der erforderlichen Evidenz-Felder in diesem Export



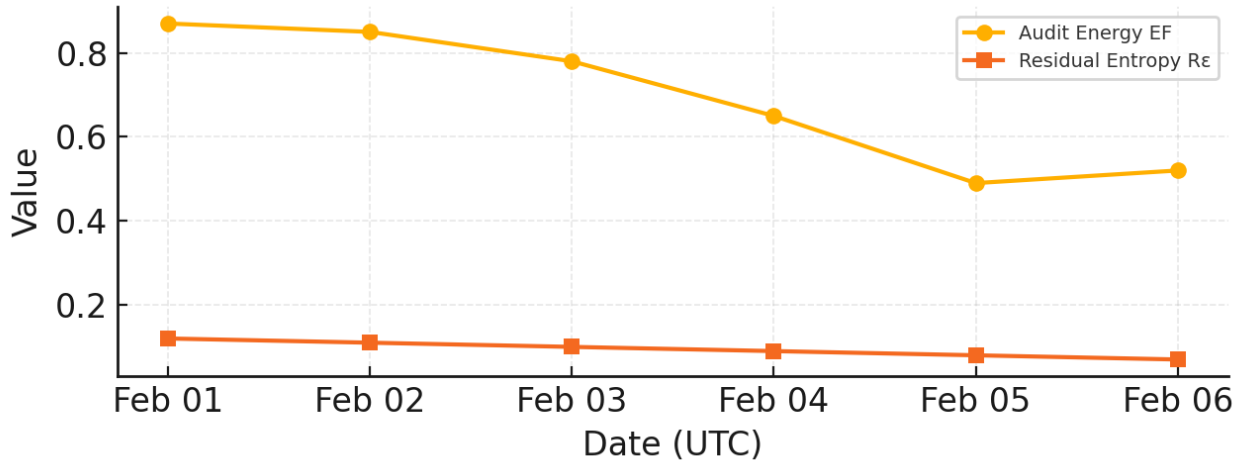
Kernbefunde (nicht-normativ)

• Drift steigt über das Fenster an; Peak am 2026-02-05 (Change-Point bestätigt).
• Incident Risk steigt parallel und erholt sich nach Rollback/Fallback teilweise.
• Evidence Completeness bleibt hoch; einige optionale Telemetrie-Felder sind bewusst nicht enthalten.

Invarianten & Drift-Signale

Diese Seite fasst Stabilitätsinvarianten und beobachtete Änderungen zusammen. Invarianten sind fail-closed: fehlende Evidenz führt für den Check zu UNDECIDED.

Invariant / Signal	Schwelle (Policy)	Beobachtet (Fenster)	Status
Audit Energy EF	EF > 0 (fail-closed)	min=0.49, max=0.87	PASS
Residual Entropy Rε	Rε ≤ 0.20 (watch >0.15)	start=0.12 → end=0.07	PASS
Semantic ΔH (embedding shift)	|ΔH| ≤ 0.05	max=0.041	PASS
Probe Robustness η	η ≥ 0.80	min=0.83	PASS
Residual feature rate φ_res	0.03 ≤ φ_res ≤ 0.07	0.031–0.066	PASS
Control latency τc (p95)	τc ≤ 3.0s	2.7s	PASS



Detektierte Änderungen

• 2026-02-05: bestätigter Regime-Shift (RSE) via Change-Point auf Embedding-Trajektorie.
• Probe Robustness sinkt im gleichen Zeitraum (höhere Sensitivität auf Korrekturprobes).
• EF sinkt, bleibt aber positiv; damit bleibt das System unter aktueller Policy auditierbar.

Marker, Logs & Einschränkungen

Marker werden aus output-only Observables abgeleitet und dienen Monitoring und Audit-Evidenz — nicht Claim-Upgrades, Benchmarking oder Personalentscheidungen.

Zeit (UTC)	Run ID	Marker	Detail
2026-02-03 16:22	RUN-26-0203-A17	CCE-precursor	Contradiction-rate ↑ (window=5), probe recovery latency +1.2 turns
2026-02-04 09:10	RUN-26-0204-C04	RSE-weak	Embedding change-point candidate (below confirm threshold)
2026-02-05 11:30	RUN-26-0205-F11	RSE-confirmed	Change-point confirmed (m=2); drift index crossed 0.15
2026-02-05 11:31	RUN-26-0205-F11	Probe sensitivity spike	Output invariance ↓ (similarity 0.42) under standardized correction probe
2026-02-05 11:34	RUN-26-0205-F11	Rollback executed	Fallback policy activated; tool-routing restricted; τc=2.7s (p95)
2026-02-06 15:55	RUN-26-0206-B09	Stabilization observed	EF recovered above 0.5; drift index decreased from 0.18 → 0.16

Evidence-Completeness Aufschlüsselung

Artefakt	Coverage	Status
Run transcripts (redacted)	100%	OK
Timestamps & run IDs	100%	OK
Model & prompt version hashes	100%	OK
Tool-call metadata (name, latency, error)	80%	Partial
Sampling metadata (temperature, top_p)	60%	Partial
Ground-truth labels for business outcomes	0%	Not in scope

Limitations (Output-only)

• Kein Zugriff auf Model-Interna (Weights, Activations, Training Data). Diagnostik basiert ausschließlich auf Transcript + Interface-Metadaten.
• Dieses Evidenzpaket bewertet *nicht* Faktizität, Value-Alignment oder Safety. Niedriger Score ≠ korrekt/sicher.
• Diagnostics können durch adaptive Modelle evadable sein; abrupte Failures können ohne Vorzeichen auftreten.
• Aussagen sind auf das Audit-Fenster und die in diesem Export eingefrorenen Config-Hashes begrenzt.

RIO Sperrvermerk (Validity & Use Restrictions)
Gültigkeitsstufe: G1 (lokal / agent-spezifisch). Default is local validity; upgrades require explicit release.
Nicht zulässig: (1) automatische Hochskalierung zu systemischen Aussagen, (2) politische Legitimation, (3) asymmetrischer Vergleich mit anderen Agents/Teams, (4) impliziter Machttransfer (z.B. HR-/Performance-Maßnahmen).
Eskalation: Wenn dieses Artefakt in Kontexten wie Vendor-Selection, externe Compliance-Reports oder öffentliche Kommunikation genutzt werden soll: erneuter Regime-Check (RRO) + explizite Transfer-Note erforderlich.