Konzeptpapier

DSGVO-konforme AI-Coding-Umgebung

Claude Code via AWS Bedrock oder Google Vertex AI

Bewertung, Architektur und Kostenschätzung

Zielsetzung: 2 Entwickler-Teams · 4 Projekte

Stand: Juni 2026

1. Executive Summary

Dieses Papier beschreibt eine DSGVO-konforme Architektur für den Einsatz von Claude als AI-Coding-Assistent in einem Unternehmen mit zwei Entwickler-Teams (insgesamt 5–15 Entwickler) und vier parallel laufenden Projekten. Ziel ist eine Lösung, die Frontier-Qualität (vergleichbar mit Claude Opus 4.7 / GPT-5.5) liefert und gleichzeitig die Anforderungen an Datenresidenz innerhalb der EU sowie Auftragsverarbeitung nach Art. 28 DSGVO erfüllt.

Empfohlen wird der Betrieb von Claude Code über AWS Bedrock in der Region eu-central-1 (Frankfurt) als primäre Option, mit Google Vertex AI in europe-west1 (Belgien) als gleichwertige Alternative. Beide Wege bieten Frontier-Modelle bei vollständiger EU-Datenresidenz, etablierten DPAs und ohne Aufbau eigener GPU-Infrastruktur.

Geschätzte monatliche Kosten: ca. 2.500–4.500 € bei realistischer Nutzung von 10 aktiven Entwicklern, je nach Modell-Mix und Caching-Strategie.

Seit Juni 2026 ist mit Claude Fable 5 erstmals ein Modell der neuen Mythos-Klasse oberhalb von Opus verfügbar. Es bietet nochmals deutlich gesteigerte Fähigkeiten für langlaufende agentische Aufgaben, unterliegt jedoch einer verpflichtenden Datenaufbewahrung von 30 Tagen und ist nicht mit Zero Data Retention kombinierbar. Dieses Papier empfiehlt Fable 5 daher nur als optionale dritte Stufe des Modell-Mixes für zuvor als unkritisch klassifizierte Workloads (siehe Abschnitt 4.4).

2. Ausgangslage und Anforderungen

2.1 Organisatorischer Kontext

- Zwei Entwickler-Teams mit insgesamt 5–15 Personen
- Vier parallele Projekte mit unterschiedlichen Codebasen und Tech-Stacks
- Regelmäßige Nutzung im Daily-Coding-Workflow (kein gelegentlicher Einsatz)
- Standort und Rechtsraum: Deutschland / EU

2.2 Funktionale Anforderungen

- Frontier-Level Coding-Qualität, vergleichbar mit Claude Opus 4.7 oder GPT-5.5
- Integration in den bestehenden Entwickler-Workflow (Terminal, IDE)
- Agentische Fähigkeiten: mehrstufige Tasks, Tool-Use, Refactorings
- Parallele Nutzung durch mehrere Entwickler ohne signifikante Wartezeiten

2.3 Compliance-Anforderungen

- Datenverarbeitung innerhalb der EU (Art. 44–46 DSGVO)
- Auftragsverarbeitungsvertrag (AVV/DPA) gemäß Art. 28 DSGVO
- Keine Nutzung von Kundendaten für Modelltraining
- Nachvollziehbare Audit-Trails und Zugriffskontrollen
- Konfigurierbare Datenretention, idealerweise Zero Data Retention (ZDR) für sensitive Workloads

Hinweis: Zero Data Retention ist nicht für alle Claude-Modelle verfügbar. Modelle der Mythos-Klasse (Claude Fable 5) sind als Covered Models eingestuft und unterliegen einer nicht abwählbaren

Aufbewahrung von 30 Tagen. Die Anforderung „idealerweise ZDR“ ist daher modellspezifisch zu bewerten (siehe Abschnitt 4.4).

3. Bewertete Alternativen

Vor der finalen Empfehlung wurden drei grundsätzliche Architekturoptionen geprüft.

3.1 Self-Hosted Open-Source-Modell

Beispielsweise DeepSeek V3.2/V4, Mistral Large oder Qwen 3 Coder auf gemietetem GPU-Server bei einem deutschen Anbieter.

Vorteile: Vollständige Kontrolle, keine Token-Kosten, Flatrate-Modell, echtes Air-Gap möglich.

Nachteile: Frontier-Qualität erfordert das volle DeepSeek-Modell (685B Parameter), das praxistauglich 8–16 GPUs der 80-GB-Klasse benötigt. Realistische Hardware-Kosten: 5.000–15.000 €/Monat. Zusätzlich erheblicher DevOps-Aufwand für Setup, Updates, Monitoring und Tooling-Eigenentwicklung, da das Ökosystem (Agent-Frameworks, MCP, Subagents) hinter Claude Code und Codex deutlich zurückliegt.

Bewertung: *Nur sinnvoll bei echter Air-Gap-Anforderung oder extremem Token-Volumen. Für den vorliegenden Use Case wirtschaftlich nicht konkurrenzfähig.*

3.2 Claude Pro/Max direkt (ohne Hyperscaler)

Direkte Nutzung der Claude-API oder von Claude Pro/Max ohne EU-Routing.

Vorteile: Einfachste Einrichtung, niedrigste Latenz, voller Funktionsumfang.

Nachteile: Standardmäßig US-basierte Inferenz. Free- und Pro-Tarife enthalten keinen DPA und sind für die Verarbeitung personenbezogener Daten nach Art. 28 DSGVO nicht zulässig.

Bewertung: *Scheidet aufgrund der Compliance-Anforderungen aus.*

3.3 Claude via Hyperscaler mit EU-Region (empfohlen)

AWS Bedrock (Frankfurt) oder Google Vertex AI (Belgien) als verwaltete Plattform für Claude-Modelle. Diese Variante wird im Folgenden detailliert.

4. Empfohlene Architektur

4.1 Komponenten-Überblick

- Claude Code CLI lokal auf den Entwickler-Maschinen
- Cloud-Backend: AWS Bedrock (eu-central-1) oder Google Vertex AI (europe-west1)
- Authentifizierung über IAM-Rollen (AWS) bzw. Service Accounts (GCP)
- Projektspezifische CLAUDE.md-Dateien für Konventionen und Constraints
- Zentrales Logging und Monitoring über CloudWatch bzw. Cloud Logging

4.2 Modell-Mix

Die Empfehlung folgt einem zweistufigen Modell-Mix, der Qualität und Kosten balanciert:

Claude Sonnet 4.6 als Daily Driver für ~80 % aller Coding-Tasks: Autocompletion, Bug-Fixes, Tests, kleinere Refactorings. Preis-Leistungs-Sweet-Spot bei 3 USD Input / 15 USD Output pro Million Tokens.

Claude Opus 4.7 für komplexe Aufgaben: Architektur-Entscheidungen, größere Refactorings, mehrstufige Agent-Workflows, Code-Reviews. Preis 5 USD Input / 25 USD Output pro Million Tokens.

Claude Haiku 4.5 optional für sehr schnelle, einfache Tasks (Inline-Completion, Boilerplate) bei 1 USD Input / 5 USD Output pro Million Tokens.

Claude Fable 5 optional als dritte Stufe für ausgewiesene Langläufer-Aufgaben: codebase-weite Migrationen, mehrstufige autonome Agent-Workflows und komplexe Modernisierungsvorhaben. Preis 10 USD Input / 50 USD Output pro Million Tokens, Kontextfenster bis 1 Million Tokens. Aufgrund der Retention-Eigenschaften (siehe Abschnitt 4.4) ist der Einsatz auf zuvor klassifizierte, unkritische Workloads zu beschränken. Am empfohlenen Grundprinzip ändert sich nichts: Sonnet bleibt Daily Driver, Opus das Werkzeug für komplexe Aufgaben.

4.3 Vergleich der beiden empfohlenen Backends

Kriterium	AWS Bedrock (Frankfurt)	Vertex AI (Belgien)
Region	eu-central-1	europe-west1
Verfügbare Modelle	Opus 4.7, Sonnet 4.6, Haiku 4.5	Opus 4.7, Opus 4.6, Sonnet 4.6, Haiku 4.5
Kontextfenster	Bis 1 Mio. Tokens (Opus 4.6/4.7)	Bis 1 Mio. Tokens (Opus 4.6/4.7)
Preisstruktur	Identisch zu Anthropic-Direkt-API; regionale Endpoints +10 % Aufschlag	Identisch zu Anthropic-Direkt-API; regionale Endpoints +10 % Aufschlag
DPA / Auftragsverarbeitung	AWS DPA + Anthropic Commercial Terms	Google Cloud DPA + Anthropic Commercial Terms
Training auf Kundendaten	Ausgeschlossen	Ausgeschlossen
Latenz aus Deutschland	Sehr niedrig (~15–30 ms)	Niedrig (~25–45 ms)
Integration Claude Code	CLAUDE_CODE_USE_BEDROCK=1	CLAUDE_CODE_USE_VERTEX=1
Default-Quoten Opus	25 RPM (auf 500 RPM erhöhbar)	200 QPM (europe-west1)
Ökosystem-Bonus	Wenn AWS bereits genutzt wird: IAM, VPC, CloudWatch nativ	Wenn GCP bereits genutzt wird: IAM, VPC-SC, Cloud Logging nativ

Primärempfehlung: AWS Bedrock in Frankfurt, sofern keine starke Bindung an GCP besteht. Vorteile: niedrigste Latenz aus Deutschland, breite Modellverfügbarkeit, etabliertes Ökosystem in deutschen Unternehmen. Vertex AI ist eine vollwertige Alternative mit nahezu identischen Konditionen, insbesondere wenn das Unternehmen bereits auf GCP setzt.

4.4 Sonderfall: Claude Fable 5 und die Mythos-Klasse

Mit Claude Fable 5 hat Anthropic im Juni 2026 die erste allgemein verfügbare Modellklasse oberhalb von Opus eingeführt. Das Modell ist auf langlaufende, autonome Aufgaben ausgelegt und erreicht

insbesondere bei codebase-weiten Migrationen und mehrstufigen Agent-Workflows Ergebnisse, die mit Opus 4.8 erheblich mehr Steuerungsaufwand erfordern würden.

Compliance-seitig unterscheidet sich Fable 5 grundlegend von den übrigen Claude-Modellen. Es ist als Covered Model eingestuft: Prompts und Outputs werden zum Betrieb von Safety-Classifiern bis zu 30 Tage aufbewahrt und anschließend gelöscht. Eine Nutzung der Daten für das Modelltraining findet nicht statt. Zero Data Retention ist für dieses Modell nicht verfügbar, auch nicht über AWS Bedrock oder Google Vertex AI. Die Verfügbarkeit in den EU-Regionen der Hyperscaler ist zum Zeitpunkt der Projektumsetzung zu prüfen.

Für die Bewertung bedeutet das eine neue Abwägung, die es in dieser Form bisher nicht gab: maximale Modellfähigkeit und minimale Datenhaltung schließen sich erstmals gegenseitig aus. Daraus leiten sich zwei Konsequenzen ab. Erstens ist vor dem Einsatz von Fable 5 eine Workload-Klassifizierung durchzuführen: Projekte mit personenbezogenen Daten, Mandantendaten oder vertraglichen ZDR-Zusagen sind vom Einsatz auszuschließen und verbleiben auf Opus 4.8 und Sonnet 4.6. Zweitens sind die Retention-Eigenschaften eines Modells künftig als reguläres Auswahlkriterium neben Qualität und Preis zu dokumentieren.

Für das in diesem Papier betrachtete Anforderungsprofil bleibt die Empfehlung aus Abschnitt 4.2 unverändert gültig. Fable 5 ist eine optionale Erweiterung für unkritische Langläufer-Aufgaben, kein Ersatz für den zweistufigen Modell-Mix. Bei konsequentem Einsatz nur für geeignete Aufgaben bleibt auch der Kostenrahmen aus Kapitel 6 weitgehend stabil, da der höhere Tokenpreis (Faktor 2 gegenüber Opus) durch den geringen Anteil solcher Aufgaben am Gesamtvolumen kompensiert wird.

5. Begründung der Empfehlung

5.1 Compliance vollständig abgedeckt

EU-Datenresidenz: Die Inferenz erfolgt physisch in der EU (Frankfurt bzw. Belgien). Keine Datenübermittlung an US-Server.

Auftragsverarbeitung: AWS und Google Cloud bieten standardmäßig DPAs nach Art. 28 DSGVO. Anthropics Commercial Terms (Stand 1.1.2026) sind automatisch in Enterprise-Verträge eingebettet und decken Subprozessor-Kontrollen, Löschverpflichtungen und Sicherheitsmaßnahmen nach Art. 32 DSGVO ab.

Kein Training auf Kundendaten: Vertraglich ausgeschlossen sowohl bei AWS/GCP als auch bei Anthropic.

Zertifizierungen: Anthropic verfügt über ISO 27001:2022, ISO 42001 und SOC 2 Type I & II. AWS und GCP entsprechend ISO 27001, ISO 27017, ISO 27018, C5 (BSI), TISAX (relevant für Automotive).

Modellspezifische Datenretention: Die Aussagen dieses Abschnitts gelten uneingeschränkt für Claude Opus, Sonnet und Haiku. Für Claude Fable 5 gilt abweichend eine verpflichtende Aufbewahrung von 30 Tagen ohne ZDR-Option (siehe Abschnitt 4.4). Die Datenschutz-Folgenabschätzung und das Verarbeitungsverzeichnis sind bei Einsatz von Fable 5 entsprechend zu ergänzen.

5.2 Frontier-Qualität ohne eigenes Infrastruktur-Investment

Claude Opus 4.7 erreicht 87,6 % auf SWE-bench Verified und 64,3 % auf SWE-bench Pro – das ist State of the Art im Coding-Bereich. Diese Qualität ist auf eigener Hardware nur mit dem vollen DeepSeek V4-Modell und Multi-GPU-Setups erreichbar, deren Aufbau und Betrieb erheblichen Aufwand bedeuten.

5.3 Skalierbarkeit

Pay-as-you-go ohne Mindestabnahme. Bei wechselnder Last (Urlaubszeiten, Projekt-Sprints) skalieren die Kosten automatisch mit. Es entstehen keine Leerkosten für ungenutzte GPU-Kapazität.

5.4 Tool-Ökosystem

Claude Code ist als CLI-Agent mit nativer Bedrock- und Vertex-Integration verfügbar. MCP-Server, Subagents, Hooks und das gesamte Anthropic-Ökosystem stehen ohne Eigenentwicklung zur Verfügung. Bei Self-Hosted-Lösungen müssten viele dieser Komponenten selbst implementiert werden.

5.5 Reversibilität

Sollte sich die Anforderungslage ändern (z. B. Wechsel zu einem anderen Modell-Anbieter oder Migration zu eigener Infrastruktur), ist der Wechsel mit überschaubarem Aufwand möglich. Die CLAUDE.md-Konventionen, Prompts und Workflows bleiben portabel.

6. Kostenschätzung

6.1 Annahmen

- Team-Größe: 10 aktiv nutzende Entwickler (Mittelwert aus 5–15)
- Nutzungstage: 20 Arbeitstage pro Monat
- Aktive Coding-Zeit pro Entwickler: 4–6 Stunden pro Tag
- Modell-Mix: 80 % Sonnet 4.6, 20 % Opus 4.7
- Caching-Effektivität: ca. 40 % der Input-Tokens werden über Prompt-Caching wiederverwendet (Rabatt bis 90 %)

6.2 Token-Schätzung pro Entwickler

Erfahrungswerte aus dem täglichen Claude-Code-Einsatz: ein aktiver Entwickler verbraucht typischerweise 8–15 Millionen Tokens pro Arbeitstag, aufgeteilt etwa 70 % Input (Code-Kontext, Dateien, Tool-Outputs) und 30 % Output (generierter Code, Erklärungen).

Verbrauchsstufe	Tokens / Tag	Tokens / Monat	Pro Entw. / Mo.
Leichte Nutzung	8 Mio.	160 Mio.	~180 €
Mittlere Nutzung	12 Mio.	240 Mio.	~270 €
Intensive Nutzung	15 Mio.	300 Mio.	~340 €

Werte verstehen sich nach Caching-Rabatten, Modell-Mix (80/20 Sonnet/Opus) und 10 % Regional-Aufschlag, gerundet.

6.3 Gesamtkosten für 10 Entwickler

Szenario	Pro Monat	Pro Jahr
Minimum (5 Entw., leicht)	~900 €	~11.000 €
Realistisch (10 Entw., mittel)	**~2.700 €**	**~32.000 €**
Maximum (15 Entw., intensiv)	~5.100 €	~61.000 €

Hinzu kommen die ohnehin meist vorhandenen AWS- bzw. GCP-Grundgebühren (CloudWatch/Logging, Datenübertragung), die im niedrigen zweistelligen Bereich pro Monat liegen.

6.4 Kostenoptimierungs-Hebel

Prompt-Caching: Bis zu 90 % Ersparnis auf wiederverwendete Input-Tokens (z. B. Systemprompts, Datei-Kontexte). Bei Claude Code automatisch aktiv.

Modell-Routing: Konsequente Trennung: Sonnet 4.6 als Default, Opus 4.7 nur für ausgewiesen komplexe Tasks. Spart 40–60 % gegenüber Opus-only-Nutzung.

Batch-Inferenz: 50 % Rabatt auf nicht-zeitkritische Workloads (z. B. nächtliche Code-Reviews, Doku-Generierung, Test-Erstellung).

Budget-Alerts: Per Cost Anomaly Detection (AWS) bzw. Budget Alerts (GCP) frühzeitige Warnung bei ungewöhnlichem Verbrauch.

Provisioned Throughput: Bei dauerhaft hohem konstantem Verbrauch (ab ca. 8.000 € / Monat) lohnt der Wechsel zu reservierter Kapazität – nicht relevant für die hier projektierte Größenordnung.

7. Vergleich zu Self-Hosting DeepSeek

Zur Einordnung der Empfehlung wird ein realistisches Self-Hosting-Szenario mit DeepSeek V3.2 oder V4 auf eigener GPU-Infrastruktur gegenübergestellt.

Aspekt	AWS Bedrock EU (empfohlen)	Self-Hosted DeepSeek
Coding-Qualität	Frontier (Opus 4.7)	Volles Modell: nahe Frontier; destillierte Varianten: deutlich schwächer
Hardware-Kosten	Keine	5.000–15.000 €/Monat (volles Modell, 8–16 GPUs)
DevOps-Aufwand	Minimal (Konfiguration)	Hoch (Setup, Tuning, Updates, Monitoring): ca. 0,3–0,5 FTE
Setup-Dauer	1–2 Wochen (inkl. DPA)	4–8 Wochen (Hardware-Beschaffung, Konfiguration, Testing)
Tool-Ökosystem	Claude Code, MCP, Subagents nativ	Begrenzt, viele Eigenentwicklungen nötig
Skalierbarkeit	Elastisch (Pay-as-you-go)	Fix (Hardware-gebunden)

DSGVO-Konformität	Erfüllt (EU-Region + DPA)	Erfüllt (vollständige Eigenkontrolle)
Air-Gap-fähig	Nein	Ja

Im vorliegenden Anforderungsprofil (DSGVO ist Treiber, keine Air-Gap-Pflicht, 5–15 Entwickler, Frontier-Qualität gewünscht) ist die Hyperscaler-Variante in jeder Hinsicht wirtschaftlicher und schneller einsatzbereit. Self-Hosting bleibt eine Option für spezielle Folge-Szenarien (z. B. Mandanten mit Air-Gap-Auflagen).

8. Umsetzungsplan

8.1 Phasen

Phase 1 (Woche 1): Vertragliche Grundlagen

- AWS-Account in eu-central-1 anlegen bzw. existierenden nutzen
- Bedrock-Zugang für Claude-Modelle in der Console aktivieren
- DPA mit Anthropic über das Enterprise-Team abschließen
- ggf. Zero Data Retention vereinbaren

Phase 2 (Woche 2): Technisches Setup

- IAM-Rollen und -Policies für die Entwickler-Teams definieren
- Default-Quoten prüfen und ggf. Erhöhung beantragen (Opus von 25 auf 500 RPM)
- CloudWatch-Logging und Budget-Alerts konfigurieren
- Claude Code auf den Entwickler-Maschinen installieren und auf Bedrock-Backend umstellen

Phase 3 (Woche 3): Pilotbetrieb

- Start mit 2–3 Pilotnutzern pro Team
- Projektspezifische CLAUDE.md-Dateien erstellen und committen
- Erste Kostenmessung und Nutzungsanalyse

Phase 4 (ab Woche 4): Rollout

- Schrittweise Erweiterung auf alle Entwickler beider Teams
- Schulung zu effektivem Prompting und Modell-Auswahl
- Monatliches Reporting zu Nutzung, Kosten und Qualität

8.2 Konfigurationsbeispiel

Aktivierung von Claude Code mit Bedrock-Backend in Frankfurt:

```
export CLAUDE_CODE_USE_BEDROCK=1 export AWS_REGION=eu-central-1 export AWS_PROFILE=dev-team-alpha claude
```

Alternativ für Vertex AI in Belgien:

```
export CLAUDE_CODE_USE_VERTEX=1 export ANTHROPIC_VERTEX_PROJECT_ID=mein-gcp-projekt export CLOUD_ML_REGION=europe-west1 claude
```

9. Risiken und Gegenmaßnahmen

Risiko	Bewertung	Gegenmaßnahme
Kosten-Drift bei Vielnutzern	Mittel	Budget-Alerts, monatliches Reporting, Schulung zu Modell-Mix
Quota-Limits bei Spitzenlast	Niedrig–Mittel	Frühzeitige Quota-Erhöhung beantragen; ggf. zweites Account-Profil
Vendor-Lock-in	Mittel	Workflows in portabler Form (CLAUDE.md, MCP) halten; alternativ Vertex AI als Backup
US Cloud Act bei AWS	Niedrig (juristisch umstritten)	Bei höchster Sensibilität: Mandanten-Daten pseudonymisieren oder zu European Sovereign Cloud wechseln, sobald Claude dort verfügbar
Modell-Deprecation	Niedrig	Anthropic kündigt Deprecations mit Vorlauf an; Migration über Model-ID-Wechsel meist trivial
Retention-Pflicht bei Covered Models (Fable 5)	Mittel	Workload-Klassifizierung vor Einsatz; kritische Projekte ausschließlich auf Opus/Sonnet mit ZDR; Retention-Eigenschaften je Modell im Verarbeitungsverzeichnis dokumentieren

10. Zusammenfassende Empfehlung

Für den vorliegenden Use Case (zwei Entwickler-Teams, vier Projekte, 5–15 Entwickler, DSGVO-Anforderung, Frontier-Qualität gewünscht) wird der Einsatz von Claude Code via AWS Bedrock in der Region Frankfurt (eu-central-1) empfohlen. Vertex AI in Belgien (europe-west1) ist eine vollwertige Alternative, sofern die Unternehmens-IT bereits auf Google Cloud setzt.

Die Lösung erfüllt alle Compliance-Anforderungen ohne den erheblichen finanziellen und operativen Aufwand einer Self-Hosting-Architektur. Die geschätzten Kosten von rund 2.700 €/Monat für eine realistische Nutzung durch 10 Entwickler stehen in einem deutlich günstigeren Verhältnis zum Produktivitätsgewinn als der Aufbau und Betrieb einer eigenen GPU-Infrastruktur (ab 5.000 €/Monat reine Hardware plus DevOps-Aufwand).

Der Einsatz von Claude Fable 5 wird als optionale dritte Stufe empfohlen, beschränkt auf zuvor klassifizierte, unkritische Workloads mit hohem Autonomiebedarf. Für alle Workloads mit personenbezogenen Daten oder ZDR-Anforderung bleiben Opus 4.8 und Sonnet 4.6 die einzigen geeigneten Modelle. Die Modellauswahl ist damit ab sofort auch eine dokumentationspflichtige Datenschutzentscheidung.

Der Einstieg ist innerhalb von 2–4 Wochen produktiv möglich. Bei sich ändernden Anforderungen (z. B. echte Air-Gap-Pflicht) bleibt eine spätere Migration zu Self-Hosting offen, da Workflows und Konventionen portabel bleiben.

Anhang: Datenquellen

- Anthropic Pricing & Data Residency: platform.claude.com/docs/en/about-claude/pricing
- AWS Bedrock Pricing & Regionen: aws.amazon.com/bedrock/pricing
- Google Vertex AI – Claude-Modelle: docs.cloud.google.com/vertex-ai/generative-ai/docs/partner-models/claude
- Anthropic DPA und Compliance: claude.com/regional-compliance
- Anthropic: Claude Fable 5 und Claude Mythos 5 (Ankündigung Juni 2026): anthropic.com/news/claude-fable-5-mythos-5
- Anthropic: Modellspezifische Anforderungen zur Datenaufbewahrung (Covered Models): platform.claude.com/docs/de/about-claude/models/introducing-claude-fable-5-and-claude-mythos-5
- Benchmarks: SWE-bench Verified und Pro (Stand Mai 2026), Terminal-Bench 2.0

Alle Preisangaben in USD wurden zum Stand Mai 2026 mit einem indikativen Kurs von 1 USD ≈ 0,92 EUR umgerechnet und gerundet. Verbindliche Preise siehe jeweilige Anbieter-Seiten.