KRYNEX LABS · WHITEPAPER · SEPTEMBER 2026

Lokal oder Cloud? Was ein 15-Euro-Server wirklich kann, gemessen

Online mit allen Links: www.krynexlabs.de/whitepaper/lokal-oder-cloud

1. Kurzfassung

Viele kleine Betriebe stellen sich dieselbe Frage: Lassen wir KI auf einem eigenen Server laufen, damit die Daten im Haus bleiben, oder nutzen wir die großen Cloud-Modelle? Wir haben das nicht abgeschätzt, sondern gemessen. Alle Zahlen in diesem Papier stammen von einem gewöhnlichen Mietserver ohne Grafikkarte, der uns 15 Euro im Monat kostet, brutto und zu einem Sonderangebot.

Das Ergebnis fällt anders aus, als wir erwartet hatten.

- **Spracherkennung läuft lokal sehr gut.** Whisper schrieb 20 Minuten deutsche Sprache mit 2,0 Prozent Wortfehlern mit. Eine Stunde Audio dauerte rund 15 Minuten. Die Aufnahmen verlassen dabei das Haus nicht.
- **Suche im eigenen Wissen läuft lokal sehr gut.** Eine Suche über 1.522 Notizen dauert 14 Millisekunden, mit Vektoren und ohne Grafikkarte.
- **Texte mit Regeln und Zahlen gehören auf diesem Server nicht lokal.** Bei einer Rechnungsprüfung hat kein einziger lokaler Lauf den eingebauten Rechenfehler erkannt, jeder Cloud-Lauf schon. Bei einer Kundenmail hielt das beste lokale Modell 5 von 9 Regeln ein und versprach einen Gutschein, den es nicht geben durfte. Die Cloud-Modelle hielten 9 von 9 ein.
- **Die Cloud ist dabei billiger, als man denkt.** 1.000 solcher Kundenmails kosten mit dem günstigsten getesteten Cloud-Modell rund 26 Cent, mit dem teuersten rund 16 Euro.
- **Eine Kontrollschicht vor jedem KI-Werkzeug kostet fast keine Zeit.** Von Ende zu Ende wartet ein Werkzeugaufruf bei uns rund 10 Millisekunden, davon 6 für die Prüfung samt signierter Protokollzeile. Eine Modellantwort dauert rund 3 bis 14 Sekunden.

Für einen Kleinbetrieb heißt das: Die Frage lautet nicht „lokal oder Cloud“, sondern welche Aufgabe wohin gehört. Kapitel 8 enthält dafür eine Tabelle.

2. Messaufbau

Der Server

Alle Messungen liefen auf demselben Mietserver, auf dem auch unsere Website-Dienste, unser Gedächtnis-Werkzeug und weitere Anwendungen laufen:

Merkmal	Wert
Prozessor	AMD EPYC Milan, 12 virtuelle Kerne, 2,0 GHz
Arbeitsspeicher	23 GB
Grafikkarte	keine
Betriebssystem	Linux, Kernel 6.8, Docker
Miete	15 Euro im Monat brutto (Sonderangebot)

Wir haben bewusst keinen Rechner mit Grafikkarte genommen. Ein Server wie dieser ist das, was ein kleiner Betrieb für wenig Geld bekommt. Wer eine Grafikkarte einplant, bekommt andere Zahlen, vor allem bei den Textmodellen.

Wie wir gemessen haben

Für jede Messung gelten dieselben Regeln.

- **Die Bewertung steht vor dem ersten Lauf fest.** Beim Vergleich lokal gegen Cloud haben wir die Prüfpunkte aufgeschrieben, bevor irgendeine Antwort erzeugt war. Was erst beim Bewerten unklar wurde, steht getrennt im Anhang.
- **Verdeckt bewertet.** Die Antworten aller Modelle wurden gemischt und mit Nummern versehen. Die Zuordnung zu den Modellen haben wir erst nach der Bewertung geöffnet.
- **Wiederholt.** Die Kernmessungen liefen je fünfmal. Wir nennen den Median und die Spannweite vom kleinsten bis zum größten Wert.
- **Unter echten Bedingungen.** Der Server lief im normalen Betrieb weiter. Die Last haben wir vor jedem Lauf mitgeschrieben.

Die Messungen im Überblick

Kapitel	Messung	Datum	Wiederholungen
3	Whisper, fünf Modellgrößen	23.09. und 28.09.2026	Genauigkeit 5 Läufe, Tempo 3 Läufe
4	Ollama, drei Qwen-Modelle	23.09. und 27.09.2026	5 Läufe
5	Lokal gegen Cloud, zwei Aufgaben	27.09.2026	5 Läufe je Modell und Aufgabe
6	Gedächtnis, LongMemEval-S	26.09.2026	470 Fragen, Tempo 20 Anfragen
7	AgentGuard, Kontrollschicht	27.09.2026	5 Läufe

3. Sprache zu Text: Whisper

Whisper ist ein Spracherkennungsmodell von OpenAI. Code und Gewichte stehen unter MIT-Lizenz und laufen auf dem eigenen Server. Wir haben sechs Modellgrößen mit faster-whisper gemessen, einer schnelleren Neuimplementierung.

Testmaterial: 20,4 Minuten deutsche Sprache aus unseren eigenen Erklärfilmen, rund 2.300 Wörter, mit viel Zahlen, Firmennamen und Anglizismen. Den gesprochenen Text kennen wir Wort für Wort, weil die Filme aus einem festen Text vertont wurden.

Modell	Wortfehler, Standard	Wortfehler, mit Stillefilter und ohne Vortext	Eine Stunde Audio dauert	Arbeitsspeicher
tiny	12,1 %	13,6 %	4,3 min	0,3 GB
base	6,8 %	7,5 %	6,2 min	0,4 GB
small	3,9 %	3,8 %	5,9 min (4 Threads)	0,7 GB
medium	2,7 %	3,2 %	29 min	1,6 GB
large-v3-turbo	6,5 % (4,2–9,7 %)	**2,0 %**	**15 min**	1,7 GB

Drei Befunde sind für den Betrieb wichtig.

Die Einstellungen zählen mehr als die Modellgröße. Mit den Standardeinstellungen schnitten die großen Modelle schlechter ab als small und medium. Der Grund waren erfundene Wörter in Pausen und fehlende Sätze, keine Verhörer. Zwei Einstellungen behoben das: ein Stillefilter vor der Erkennung und das Abschalten des Vortexts aus dem vorigen Abschnitt (`condition_on_previous_text=False`).

In Stille erfindet Whisper Text. Ohne Stillefilter schrieben die großen Modelle auf 30 Sekunden reine Stille zum Beispiel „Untertitelung des ZDF, 2020“. Mit Stillefilter erfand keines der Modelle etwas. Ein Transkript bleibt trotzdem ein Entwurf. Wo der genaue Wortlaut zählt, prüft ein Mensch gegen die Aufnahme.

Mehr Kerne sind nicht schneller. small brauchte mit 4 Threads weniger als halb so lange wie mit 12. Das sollte man an der eigenen Hardware ausprobieren.

Zum Vergleich die Cloud: OpenAI nannte am 23.09.2026 für gpt-4o-mini-transcribe 0,003 und für gpt-transcribe 0,0045 US-Dollar je Minute Audio. Eine Stunde kostet dort also 18 bis 27 US-Cent, und die Aufnahme geht an OpenAI. Lokal kostet eine Stunde Audio mit large-v3-turbo 15 Minuten Rechenzeit auf dem ganzen Server, umgerechnet höchstens 0,5 Cent.

Einordnung: Die Genauigkeit haben wir fünfmal gemessen, am 23. und am 28. September 2026, mit Beam-Suche (Breite 5) und fest eingestellter Sprache Deutsch. Mit den empfohlenen Einstellungen waren alle fünf Transkripte von large-v3-turbo Zeichen für Zeichen gleich, jedes Mal 2,0 Prozent. Streuung gab es nur bei large-v3-turbo mit den Standardeinstellungen: 4,2 bis 9,7 Prozent, Median 6,5. Dort versucht Whisper misslungene Abschnitte mit höherer Temperatur neu, und dabei wird gewürfelt. Die erfundenen Schleifen am Ende zweier Nachrichten waren jedes Mal verschieden lang. Die Tabelle zeigt deshalb bei dieser einen Zeile Median und Spannweite, bei allen anderen war die Fehlerquote in jedem Lauf gleich. Das Material ist sauber gesprochen, von einer Stimme, ohne Störgeräusche. Telefonaufnahmen, Besprechungen und Dialekt schneiden schlechter ab. Die Werte sind eine Untergrenze.

4. Textmodelle lokal: Ollama

Ollama lädt offene Sprachmodelle herunter und stellt sie über eine Schnittstelle bereit. Wir haben drei Größen von Qwen2.5 gemessen, jeweils in der verbreiteten 4-Bit-Fassung (Q4_K_M), nur auf dem Prozessor.

Modell	Größe im Speicher	Schreibtempo	Kurze Mail, gesamt	Laden beim ersten Aufruf
Qwen2.5 1,5B	1,0 GB	73,6 Token/s (71,4–78,8)	2,5 s	6,0 s
Qwen2.5 3B	1,9 GB	43,1 Token/s (37,3–44,1)	3,7 s	7,9 s
Qwen2.5 7B	4,7 GB	19,1 Token/s (18,2–20,2)	8,5 s	20,9 s

Median aus 5 Läufen, Spannweite in Klammern. Kurze Mail mit rund 160 bis 180 Token Antwort. Messung vom 27.09.2026.

Ein Token ist grob eine Silbe. 19 Token pro Sekunde beim 7B-Modell sind etwas langsamer, als ein Mensch liest. Für eine kurze Antwort reicht das.

Lange Texte sind das eigentliche Problem. Zum Einlesen eines Blogartikels mit 4.015 Token brauchte das 7B-Modell 86 Sekunden, bis zur fertigen Zusammenfassung rund 100 Sekunden. In dieser Zeit waren rund sechs der zwölf Kerne voll belegt. Wer Verträge oder lange Mails lokal zusammenfassen will, plant mit Minuten je Dokument.

5. Lokal oder Cloud: dieselbe Aufgabe

Tempo allein sagt wenig. Die entscheidende Frage ist, ob die Antwort stimmt. Wir haben deshalb zwei typische Büroaufgaben an sieben Modelle gegeben, je fünfmal, mit wörtlich derselben Anweisung.

Lokal: Qwen2.5 mit 1,5, 3 und 7 Milliarden Parametern über Ollama. **Cloud:** Claude Haiku 4.5 und Claude Opus 5.5 von Anthropic, GPT-6 Luna und GPT-6 Sol von OpenAI, alle über OpenRouter.

Aufgabe 1: Rechnungsprüfung

Ein Elektriker hat ein Angebot über vier Positionen gemacht. Die Rechnung weicht an drei Stellen ab: mehr Stunden zu höherem Satz (plus 166 Euro), eine zweite Anfahrt (plus 35 Euro) und ein Rechenfehler, denn 3 × 42,50 Euro sind nicht 137,50 Euro, sondern 127,50 Euro. Alles andere stimmt. Das Modell soll jede Abweichung mit Betrag nennen.

Modell	Abweichungen gefunden (von 3)	Falsche Behauptungen in 5 Läufen	Zeit je Aufgabe	Kosten je Aufgabe
Qwen2.5 1,5B, lokal	1 (0–2)	19	10,1 s	lokal, siehe unten
Qwen2.5 3B, lokal	1 (1–2)	9	2,1 s	lokal
Qwen2.5 7B, lokal	2 (1–2)	6	12,3 s	lokal
Claude Haiku 4.5	3 (1–3)	0	2,8 s	0,16 Cent
GPT-6 Luna	3 (3–3)	0	5,8 s	0,04 Cent
GPT-6 Sol	3 (3–3)	0	7,4 s	0,54 Cent
Claude Opus 5.5	3 (3–3)	0	14,3 s	2,28 Cent

Median aus 5 Läufen, Spannweite in Klammern. Kosten laut OpenRouter in US-Dollar, umgerechnet mit dem EZB-Referenzkurs vom 25.09.2026 (1 Euro = 1,1403 US-Dollar).

Die deutlichste Zahl steht nicht in der Tabelle: **Den Rechenfehler hat kein einziger der 15 lokalen Läufe als Rechenfehler erkannt. Alle 20 Cloud-Läufe haben ihn erkannt.** Die lokalen Modelle haben dafür Fehler erfunden. Das 7B-Modell behauptete zum Beispiel, die Umsatzsteuer fehle auf der Rechnung. Sie steht dort. Eine solche Antwort ist schlimmer als keine, weil sie glaubwürdig klingt.

Aufgabe 2: Kundenmail

Eine Kundin beschwert sich: Die Gartenbank kam 12 Werktage zu spät, die Lehne hat einen Riss, sie will die Versandkosten zurück, dazu einen Gutschein, und auf der Rechnung steht ein anderer Preis als im Shop. Das Modell bekommt fünf Regeln des Betriebs, etwa: Versandkosten ab 10 Werktagen Verzug erstatten, Fotos anfordern, keine Gutscheine ohne Freigabe der Teamleitung, Rechnungsfragen an die Buchhaltung.

Wir haben 9 Punkte geprüft, darunter jede der Regeln, die Sie-Form, die Länge und ob das Modell Zusagen erfindet.

Modell	Punkte (von 9)	Zeit je Mail	Kosten je Mail
Qwen2.5 1,5B, lokal	3 (2–3)	3,8 s	lokal
Qwen2.5 3B, lokal	3 (3–4)	5,5 s	lokal
Qwen2.5 7B, lokal	5 (4–7)	8,8 s	lokal
Claude Haiku 4.5	9 (9–9)	3,4 s	0,18 Cent
GPT-6 Luna	9 (8–9)	5,2 s	0,03 Cent
GPT-6 Sol	9 (9–9)	6,7 s	0,38 Cent
Claude Opus 5.5	9 (9–9)	10,7 s	1,56 Cent

Das 7B-Modell hat in vier von fünf Läufen einen Gutschein in Aussicht gestellt oder zugesagt, obwohl die Regel das verbietet. In allen fünf Läufen stand mindestens eine erfundene Zusage oder falsche Tatsache in der Mail. Die Rechnungsfrage hat es in keinem Lauf an die Buchhaltung verwiesen. Kein Cloud-Modell hat eine Zusage erfunden.

Was es kostet

Für die Cloud zahlen Sie je Aufruf. Für den eigenen Server zahlen Sie die Miete, egal wie viel er rechnet. Um beides vergleichbar zu machen, rechnen wir die Miete auf Sekunden um: 15 Euro im Monat sind 0,00058 Cent je Sekunde für den ganzen Server. Das ist eine Obergrenze, denn auf dem Server läuft anderes mit.

1.000 Kundenmails	Kosten	Qualität (Punkte von 9)
Qwen2.5 7B, lokal	rund 5 Cent Serverzeit	5
GPT-6 Luna	rund 26 Cent	9
Claude Haiku 4.5	rund 1,75 Euro	9
GPT-6 Sol	rund 3,77 Euro	9
Claude Opus 5.5	rund 15,60 Euro	9

Lokal ist also billiger, aber der Abstand zur günstigsten Cloud liegt bei rund 20 Cent je 1.000 Mails. Unsere 15 Euro waren ein Sonderangebot. Selbst bei der zehnfachen Miete, also 150 Euro im Monat, kosteten 1.000 Mails lokal rund 50 Cent Serverzeit. Die Rechnung kippt also nicht am Mietpreis, sondern an der Qualität. Dafür bekommt man in der Cloud Antworten, die man verschicken kann. Die lokalen müsste jemand Satz für Satz prüfen, und diese Arbeitszeit kostet mehr als jede Cloud-Rechnung.

Die ganze Messung mit 40 Cloud-Aufrufen hat 0,29 US-Dollar gekostet.

Die Cloud-Beträge sind die Kosten, die OpenRouter je Aufruf in US-Dollar meldet, und zwar ohne Steuern: Laut OpenRouter kommt Umsatzsteuer, wo das Recht es verlangt, erst beim Kauf des Guthabens dazu. Unsere 15 Euro Miete sind dagegen brutto. Mit 19 Prozent obendrauf

kosten 1.000 Mails mit GPT-6 Luna rund 31 statt 26 Cent. An den Größenordnungen ändert das nichts.

Wichtig zur Einordnung: Die Cloud-Aufrufe liefen über die API, nicht über die Apps. Bei den Cloud-Anbietern gehen die Daten in die USA, sofern Sie keine EU-Verarbeitung vereinbaren. Das ist der eigentliche Preis der Cloud, und er steht in keiner Tabelle. Kapitel 8 geht darauf ein.

6. Gedächtnis für KI-Agenten

Ein KI-Agent vergisst nach jeder Sitzung alles, was nicht aufgeschrieben ist. Ab einer gewissen Menge liegen die Notizen in einem Speicher mit Suche. Die Suche entscheidet dann, was der Agent weiß.

Wir haben auf dem öffentlichen Datensatz LongMemEval-S (ICLR 2025, 470 Fragen) gemessen, ob die richtige Gesprächssitzung gefunden wird. Verglichen haben wir eine reine Wortsuche mit einer hybriden Suche aus Wörtern und Vektoren. Das Vektormodell (potion-multilingual-128M) läuft ohne Grafikkarte.

	Richtige Sitzung auf Platz 1	Alle nötigen Sitzungen unter den ersten 10 (Fragen über mehrere Sitzungen)
Hybrid	84,3 %	89,3 %
Nur Wortsuche	84,3 %	80,2 %

Suche	rund 50 Notizen	1.522 Notizen
Nur Wortsuche	9,3 ms	13,8 ms
Hybrid, Modell bleibt geladen	7,8 ms	13,6 ms
Hybrid, Modell wird je Anfrage geladen	1.500 ms	1.473 ms

Für den Betrieb heißt das: Suche im eigenen Wissen ist eine Aufgabe, die ein kleiner Server mühelos lokal erledigt. Die Wortsuche trägt den Großteil. Vektoren helfen, wenn eine Antwort über mehrere Notizen verteilt ist, und kosten nichts extra, solange das Modell geladen bleibt.

Gemessen ist hier nur, ob die richtige Notiz gefunden wird, nicht ob der Agent daraus richtig antwortet.

7. Kontrolle kostet fast nichts

Sobald ein KI-Agent Werkzeuge bedient, also Dateien schreibt, Befehle ausführt oder Mails verschickt, braucht es eine Stelle, die jeden Schritt prüft und festhält. Bei uns ist das AgentGuard. Jeder Werkzeugaufruf geht vorher dort durch: Er wird gegen Regeln geprüft, auf Schlüssel und versteckte Anweisungen im Inhalt durchsucht und in einem Protokoll festgehalten, dessen Einträge verkettet und signiert sind.

Die naheliegende Sorge ist, dass das den Agenten ausbremst. Wir haben es an einer eigenen Kopie mit denselben Einstellungen wie im Betrieb gemessen, je fünfmal.

Schritt	Dauer (Median)	Langsamste 1 %
Entscheidung im Programm, ohne Inhalt	0,07 ms	0,21 ms
Entscheidung mit Inhaltsprüfung (Mail, 600 Zeichen)	0,19 ms	0,41 ms
dazu dauerhafte Protokollzeile in der Datenbank	1,18 ms	1,77 ms
Prüfung über das Netz, gemessen am Server	2,5 bis 3,0 ms	3,9 bis 4,5 ms
Werkzeugaufruf von Ende zu Ende, wie im Betrieb	**10,3 ms**	13 bis 14 ms
davon durch die Prüfung	6,0 ms	9 bis 10 ms

Median aus 5 Läufen. Die letzten beiden Zeilen messen den echten Weg: Claude Code startet vor jedem Werkzeug unser Gedächtnis-Werkzeug Cyberbrain als eigenen Prozess, und dieses fragt AgentGuard. Verglichen mit demselben Aufruf ohne Prüfung. Gemessen am 28.09.2026 mit Cyberbrain 0.7.2.

Zum Vergleich: Die Cloud-Modelle in Kapitel 5 brauchten für eine Antwort rund 3 bis 14 Sekunden. Die Kontrolle macht davon weniger als ein halbes Prozent aus.

Die Zahl von Ende zu Ende ist die ehrliche. Bei der ersten Messung am Vormittag wartete ein Werkzeugaufruf 15 Millisekunden, 11 davon für die Prüfung, obwohl der Server selbst in unter 3 Millisekunden entscheidet. Der Rest entstand beim Aufrufer: Er las vor jeder Anfrage den kompletten Zertifikatsspeicher des Systems ein, obwohl die Verbindung gar nicht verschlüsselt ist. Das kostete 4 bis 5 Millisekunden. Bemerkt haben wir es erst, als wir den ganzen Weg gemessen haben und nicht nur den Server. Behoben haben wir es am selben Tag, das ist Cyberbrain 0.7.1. Die Tabelle zeigt die Nachmessung mit dieser Korrektur: 10,3 statt 15 Millisekunden.

Zwei Dinge haben wir dabei gelernt, und beide sprechen für das Messen.

Eine schöne Zahl kann irreführen. Ältere Messungen nannten 67 Mikrosekunden für die Entscheidung. Die Zahl stimmt, aber sie enthält weder die Datenbank noch das Netz noch den Aufrufer. Wer nur sie nennt, unterschätzt die Wartezeit eines Werkzeugaufrufs um mehr als das Hundertfünfzigfache.

Messen findet Fehler. Bei der ersten Messung am 25.09. wurde die Entscheidung ab der 10.000. Aktion zwanzigmal langsamer. Die Ursache war eine Aufräumroutine, die mit jeder Aktion die ganze Warteliste durchlief. Nach der Korrektur bleibt die Dauer bis 30.000 Aktionen flach bei 0,07 ms. Die Gegenprobe mit dem alten Stand zeigt den Einbruch in allen fünf Läufen, die Messung erkennt ihn also.

Eine zweite Protokollzeile fällt nur an, wenn eine Prüfung im Beobachtungsmodus etwas beanstandet hätte. Im Betrieb war das in den 24 Stunden vor der Messung bei 7 von 995 Werkzeugaufrufen der Fall (0,7 Prozent), jedes Mal wegen eines Zugangsdatums oder einer Bankverbindung im Inhalt. Diese Aufrufe dauern etwa 0,5 bis 1 Millisekunde länger.

8. Was das für die Entscheidung heißt

Aufgabe	Unsere Empfehlung für einen Server wie diesen	Warum
Aufnahmen verschriftlichen (Besprechung, Diktat)	lokal	2,0 % Wortfehler, Aufnahmen bleiben im Haus, Stillefilter einschalten
Suche im eigenen Wissen, Gedächtnis für Agenten	lokal	14 ms je Suche, kein Unterschied zur Cloud nötig
Kundenmails, Angebote, Texte nach Regeln	Cloud	lokal 5 von 9 Regeln, erfundene Zusagen
Rechnungen und Zahlen prüfen	Cloud	lokal kein Rechenfehler erkannt, erfundene Fehler
Lange Dokumente zusammenfassen	Cloud, oder lokal mit Geduld	lokal rund 100 s je Dokument
Kontrolle vor jedem Werkzeugaufruf	lokal	rund 10 ms je Aufruf, das Protokoll gehört ins Haus

Wenn Daten das Haus nicht verlassen dürfen. Dann ist die Cloud-Spalte nicht einfach zu wählen. Drei Wege bleiben. Man kann Daten vor dem Senden pseudonymisieren, also Namen und Kennungen ersetzen. Man kann einen Anbieter mit Verarbeitung in der EU und einem Vertrag zur Auftragsverarbeitung wählen. Oder man plant für Textaufgaben einen Server mit Grafikkarte und größeren Modellen ein und misst ihn genauso nach. Unsere Messung sagt nur, dass es mit einem 7B-Modell auf dem Prozessor nicht reicht.

Weiterführend auf unserer Website:

- Pseudonymisierung vor dem LLM
- AVV mit KI-Anbietern
- On-Premise oder Cloud-LLM
- Claude oder ChatGPT im Betrieb

9. Grenzen der Messung

- **Ein Server, keine Grafikkarte.** Mit Grafikkarte laufen Textmodelle um ein Vielfaches schneller, und größere Modelle werden möglich. Das haben wir nicht gemessen.
- **Eine Modellfamilie lokal.** Wir haben Qwen2.5 gemessen. Andere offene Modelle gleicher Größe können besser oder schlechter abschneiden.
- **Zwei Aufgaben.** Rechnung und Mail stehen für viele Büroaufgaben, aber nicht für alle. Die Ergebnisse waren so deutlich, dass wir kein anderes Bild erwarten. Bewiesen ist das nicht.
- **Keine Temperatur gesetzt.** Alle Modelle liefen mit ihrer Standardeinstellung.
- **Verdeckt, aber nicht unerkennbar.** Stil und Länge verraten manchmal die Modellfamilie. Die maschinell prüfbaren Punkte (Länge, Sie-Form, Mailadresse) lief ein Skript.

- **Nebenlast.** Der Server lief im normalen Betrieb. Die lokalen Läufe hatten eine Last von 1,4 bis 6,7, zum Großteil durch Ollama selbst.
- **Preise sind Stand September 2026.** Cloud-Preise ändern sich oft.

Anhang: Methode

Die Rohdaten und Messskripte veröffentlichen wir nicht. Aufbau, Bewertungsregeln und Versionen stehen unten so vollständig, dass sich jede Messung nachbauen lässt.

Bewertung Rechnung (vorab festgelegt)

- R1: Position 1 als Abweichung genannt und 166,00 Euro genannt
- R2: Position 2 als Abweichung genannt und 35,00 Euro genannt
- R3: Position 4 als Rechenfehler erkannt und 10,00 Euro Differenz genannt
- Falsche Behauptungen getrennt gezählt: erfundene Abweichung, falscher Betrag, falsche Folgezahl oder „alles stimmt“

„Streng“ zählt R3 nur, wenn der Rechenfehler als solcher benannt ist. Die milde Zählung, die auch „plus 10 Euro gegenüber dem Angebot“ gelten lässt, steht in den Rohdaten. Sie ändert am Bild nichts: Mild kamen die lokalen Modelle im Median auf 2, 1 und 3 Punkte.

Bewertung Kundenmail (vorab festgelegt)

Höchstens 150 Wörter · Sie-Form · Rechnungsfrage an die Buchhaltung verwiesen · Versandkosten 19,90 Euro zugesagt · Fotos angefordert · Wahl Ersatzteil oder Gutschrift · kein Gutschein zugesagt · keine erfundenen Zusagen · verständliches, freundliches Deutsch.

Versionen

Ollama 0.33.3, Qwen2.5 Instruct in Q4_K_M. faster-whisper 1.2.1, 8-Bit. Cloud-Modelle über OpenRouter am 27.09.2026: anthropic/claude-opus-5.5, anthropic/claude-haiku-4.5, openai/gpt-6-sol, openai/gpt-6-luna. AgentGuard Stand c5b2c45. Cyberbrain 0.6.1 (Gedächtnis-Messung), 0.7.0 und 0.7.2 (Hook-Messung).

Krynex Labs UG (haftungsbeschränkt) i. G. Alle Messungen September 2026 auf eigener Infrastruktur. Jede Zahl ist an den Rohdaten oder der genannten Quelle geprüft.

Krynex Labs UG (haftungsbeschränkt) i. G. · www.krynexlabs.de