Krynex Labs

WHITEPAPER FÜR KANZLEIEN · NOVEMBER 2026

KI in der Kanzlei, ohne dass Akten das Haus verlassen

Lokale Suche in Akten und Entscheidungen mit Fundstelle, gemessen an 1.440 Bundesgerichtsentscheidungen. Was das Berufsrecht verlangt, welche drei Wege es gibt und was eine KI-Box kostet.

78 %
der Fragen: richtige Entscheidung auf Platz 1, gesucht unter 1.440

92 %
unter den ersten fünf Treffern, jeder mit Seitenangabe

0,11 s
je Suche auf einem Server ohne Grafikkarte, nichts verlässt das Gerät

INHALT

1. Kurzfassung
2. Wo in der Kanzlei die Zeit hängen bleibt
3. Was das Berufsrecht verlangt
4. Drei Wege zu KI in der Kanzlei
5. Was wir gemessen haben
6. Wie das Kanzlei-Modul arbeitet
7. Die KI-Box
8. Acht Fragen an jeden Anbieter lokaler KI
9. So fangen Kanzleien bei uns an
10. Grenzen dieser Messung

Anhang: Methode

1. Kurzfassung

Fast jede Kanzlei hat dasselbe Problem mit KI. Die Werkzeuge wären nützlich, aber Mandatsakten unterliegen dem Berufsgeheimnis, und jeder Cloud-Anbieter wird damit zu einem Dienstleister, den die Kanzlei sorgfältig auswählen, vertraglich verpflichten und im Zweifel um Einwilligung der Mandanten bitten muss. Viele lassen es deshalb ganz. Andere kopieren aus Zeitnot doch etwas in ein Chatfenster.

Es gibt einen dritten Weg: Die KI läuft im eigenen Haus, auf einem Rechner der Kanzlei oder auf einem eigenen Gerät im Serverschrank, der KI-Box. Wir haben gemessen, wie gut das heute für die wichtigste Aufgabe funktioniert, nämlich die richtige Stelle in einem großen Aktenbestand zu finden.

- **Die Suche ist gut genug für den Alltag.** In 1.440 Entscheidungen der Bundesgerichte fand die lokale Suche bei 78 Prozent von 50 Fragen die richtige Entscheidung auf Platz 1 und bei 92 Prozent unter den ersten fünf. Jeder Treffer nennt Dokument und Seite.
- **Dafür braucht es keine Grafikkarte.** Eine Suche dauerte 0,11 Sekunden auf einem gewöhnlichen Server. Nur das erste Einlesen eines großen Bestands braucht Rechenzeit, auf einer KI-Box deutlich weniger als auf einem Bürorechner.
- **Das Modell macht den Unterschied.** Mit einem kleinen, schnellen Suchmodell stand die richtige Entscheidung nur bei 46 Prozent der Fragen oben, mit zwei größeren Modellen bei 76 und 78 Prozent. Welches Modell auf welchem Gerät läuft, entscheidet also über den Nutzen.
- **Antworten bleiben prüfbar.** Lokale Sprachmodelle machen Fehler. Deshalb trägt bei uns jede Antwort eine Fundstelle in den eigenen Akten, und wo die Akten nichts hergeben, soll die Antwort „nicht gefunden“ lauten statt einer Vermutung. In 60 Testfragen kam kein einziges Zitat aus einer Akte, für die die fragende Person nicht freigeschaltet war.
- **Für Antworten zählt die Hardware.** Die Suche läuft auf jedem Bürorechner. Ein Sprachmodell, das daraus Antworten formuliert, braucht auf einem Laptop ohne Grafikkarte rund zwei Minuten je Antwort, auf einer KI-Box rund vier Sekunden (Abschnitt 7).

Dieses Papier beschreibt den Stand im November 2026 ehrlich: Was gemessen ist, steht mit Zahlen da. Was noch entsteht, steht als „in Arbeit“ da. Am 8. November haben wir es um die Messungen zu Rechten, Antworten und Hardware ergänzt. Am 11. November haben wir zwei Angaben korrigiert, die zu günstig waren: die Antwortzeit, wenn drei Personen gleichzeitig fragen, und das Ergebnis der Handprobe (Abschnitte 6 und 7).

2. Wo in der Kanzlei die Zeit hängen bleibt

Die Kanzleisoftware führt Akten, Fristen und das beA. Was sie selten gut kann, ist **finden**:

- Die Stelle im gegnerischen Schriftsatz von vor zwei Jahren, in der genau dieses Argument schon einmal stand.
- Das eigene Muster für einen Fall, den eine Kollegin vor drei Jahren gewonnen hat.
- Die Entscheidung, die man im Kopf hat, deren Aktenzeichen aber nicht.
- Den Scan vom Posteingang, der in einer anderen Akte abgelegt wurde.

Eine Volltextsuche findet nur, wer das richtige Wort kennt. Eine KI-Suche findet auch, was mit anderen Worten gemeint ist. Damit sie in einer Kanzlei taugt, muss sie drei Dinge zusätzlich können: jeden Treffer belegen, nur zeigen, wofür jemand freigeschaltet ist, und keine Daten nach außen geben.

3. Was das Berufsrecht verlangt

Wir geben keine Rechtsberatung. Die folgenden Punkte sind Orientierung mit Quelle, geprüft am Wortlaut der Gesetze im November 2026.

Verschwiegenheit. Anwälte sind zur Verschwiegenheit über alles verpflichtet, was ihnen in Ausübung des Berufs bekannt geworden ist (§ 43a Abs. 2 BRAO). Wer ein fremdes Geheimnis unbefugt offenbart, macht sich strafbar (§ 203 Abs. 1 Nr. 3 StGB).

Dienstleister sind erlaubt, aber mit Pflichten. Geheimnisse dürfen gegenüber Personen offenbart werden, die an der beruflichen Tätigkeit mitwirken, soweit das für deren Tätigkeit erforderlich ist (§ 203 Abs. 3 StGB). Für Dienstleister regelt § 43e BRAO die Einzelheiten:

Vorgabe	Fundstelle
Zugang nur, soweit für die Dienstleistung erforderlich	§ 43e Abs. 1 BRAO
Dienstleister sorgfältig auswählen, Zusammenarbeit beenden, wenn Vorgaben nicht eingehalten werden	§ 43e Abs. 2 BRAO
Vertrag in Textform, Verpflichtung zur Verschwiegenheit mit Belehrung über die strafrechtlichen Folgen, Regel zu Unterauftragnehmern	§ 43e Abs. 3 BRAO
Dienstleistung im Ausland nur bei vergleichbarem Geheimnisschutz	§ 43e Abs. 4 BRAO
Dient die Dienstleistung unmittelbar einem einzelnen Mandat: Einwilligung des Mandanten	§ 43e Abs. 5 BRAO
Datenschutzrecht gilt daneben, also auch ein Vertrag zur Auftragsverarbeitung	§ 43e Abs. 8 BRAO, Art. 28 DSGVO

Was daraus folgt. Das Gesetz verbietet Cloud-KI nicht. Es macht sie aufwendig: Anbieter prüfen, Verträge schließen, Serverstandorte und Unterauftragnehmer kennen, bei mandatsbezogener Nutzung Einwilligungen einholen. Läuft die KI auf Geräten der Kanzlei, entfällt der Anbieter für die KI selbst. Bleibt ein Dienstleister für Einrichtung und Wartung, und auch der muss nach § 43e BRAO verpflichtet werden. Ob das im Einzelfall genügt, klärt die Kanzlei mit ihrer Kammer oder ihrem Datenschutzbeauftragten.

4. Drei Wege zu KI in der Kanzlei

	Cloud-KI mit Vertrag	Lokal auf einem Rechner	KI-Box im Haus
Für wen	Kanzleien, die Verträge und Einwilligungen ohnehin organisieren	Einzelanwalt, kleines Büro	Kanzlei mit mehreren Personen
Wo die Akten liegen	beim Anbieter, zumindest während der Verarbeitung	auf dem Rechner	zentral auf der Box
Dienstleister nach § 43e BRAO	KI-Anbieter und seine Unterauftragnehmer	nur für Einrichtung und Pflege	nur für Einrichtung und Wartung
Suche mit Fundstelle	je nach Produkt	ja	ja
Antworten formulieren	stärkste Modelle	kleines Modell, kurz und langsam	große Modelle, mehrere Nutzer gleichzeitig
Rechte je Mitarbeiter	je nach Produkt	entfällt, eine Person	ja (in Arbeit, siehe Abschnitt 6)
Kosten	laufend je Nutzer oder Anfrage	vorhandene Hardware	Gerät einmalig, Wartung monatlich

Unser Rat ist nicht pauschal „lokal“. Für manche Aufgaben, etwa lange Schriftsätze aus dem Nichts zu entwerfen, sind die großen Cloud-Modelle besser, und das sagen wir dann auch. Für das Suchen und Belegen in den eigenen Akten zeigt unsere Messung, dass lokal reicht.

5. Was wir gemessen haben

Der Bestand

1.400 Entscheidungen der Bundesgerichte aus dem amtlichen Angebot „Rechtsprechung im Internet“, zusammen 10.171 Seiten als PDF, dazu 40 Dokumente mit 200 Seiten als nachgestellte Scans. Gerichtsentscheidungen sind amtliche Werke und damit frei nutzbar (§ 5 UrhG). Echte Akten haben wir nicht verwendet, und das werden wir auch nicht ohne Verpflichtung nach § 203 StGB und einen Vertrag zur Auftragsverarbeitung tun.

Das Einlesen der 10.171 Seiten mit Textebene dauerte 33 Sekunden, der Suchindex mit dem schnellen Suchmodell weitere 15 Sekunden. Jeder Block im Index trägt Dokument und Seite.

Die Fragen

Ein Zufallsgenerator mit festem Startwert hat 50 Stellen aus dem Bestand gezogen: 25 Leitsätze und 25 Absätze aus der Mitte der Entscheidungsgründe. Zu jeder Stelle haben wir eine Frage so formuliert, wie sie in einer Kanzlei gestellt würde, ohne die Wörter der Stelle abzuschreiben. Ein Beispiel:

> *Was muss eine Kanzlei darlegen, damit ihr kein Organisationsverschulden bei der Fristenkontrolle vorgeworfen wird?*

Gezählt wurde, ob die richtige Entscheidung unter den ersten 1, 3, 5 oder 10 Treffern steht. Als Gegenprobe haben wir die Antworten zufällig vertauscht. Dann stand die „richtige“ Entscheidung nur noch bei 4 Prozent der Fragen unter den ersten fünf. Die Messung unterscheidet also echte Treffer von Zufall.

Das Ergebnis

	model2vec	embeddinggemma	bge-m3
richtige Entscheidung auf Platz 1	46 %	**78 %**	76 %
unter den ersten 3	66 %	92 %	**94 %**
unter den ersten 5	78 %	92 %	**94 %**
unter den ersten 10	86 %	94 %	94 %
Leitsätze unter den ersten 5	80 %	96 %	96 %
Stellen aus den Gründen unter den ersten 5	76 %	88 %	**92 %**
mittlere Position (MRR, 1,0 = immer Platz 1)	0,59	**0,85**	0,84
Dauer je Suche (Median)	0,06 s	0,11 s	0,11 s
Ersteinlesen der Vektoren, Server ohne Grafikkarte	15 s	66 min	124 min

Drei Dinge fallen auf.

Das Suchmodell ist der größte Hebel. Zwischen dem kleinen und den beiden größeren Modellen liegen rund 30 Prozentpunkte auf Platz 1. Alle drei laufen lokal, keines kostet Lizenzgebühren.

Die beiden größeren Modelle liegen gleichauf. embeddinggemma setzt eine Entscheidung mehr auf Platz 1, bge-m3 findet eine Stelle mehr unter den ersten fünf. Bei 50 Fragen ist das ein Unterschied von einer Frage und damit Zufall. Entscheiden werden Einlesedauer und Lizenzbedingungen, nicht die Trefferquote.

Stellen tief in den Gründen sind schwerer. Leitsätze fasst ein Gericht selbst zusammen, dort fanden beide größeren Modelle 96 Prozent unter den ersten fünf. Einen Absatz aus der Mitte der Begründung fanden sie zu 88 und 92 Prozent. Das ist der Fall, der in der Praxis zählt, und der Grund, warum jeder Treffer die Seite nennt.

Das Einlesen kostet einmal Zeit. Die Suche selbst ist schnell. Die Vektoren für einen großen Bestand zu berechnen, dauert auf einem Server ohne Grafikkarte mit den größeren Modellen ein bis zwei Stunden. Danach kommen nur neue Dokumente dazu. Auf einer KI-Box mit Grafikkarte geht das deutlich schneller, die genaue Zahl messen wir als Nächstes.

6. Wie das Kanzlei-Modul arbeitet

Die Suche oben ist Teil von Cyberbrain, unserer Software für lokale Suche und Gedächtnis. Den Kern gibt es öffentlich, der Quelltext liegt auf GitHub. Für Kanzleien bauen wir darauf das Kanzlei-Modul. Der Stand im November 2026:

Baustein	Was er tut	Stand
Einlesen	PDF je Seite mit Fundstelle, auch Scans mit Texterkennung	**fertig**, gemessen
Word-Dokumente	DOCX mit Absatz einlesen	in Arbeit
Suche mit Fundstelle	Treffer mit Dokument und Seite	**fertig**, gemessen (Abschnitt 5)
Signiertes Protokoll	jede Änderung am Bestand mit Zeitpunkt, signiert und verkettet, sodass Lücken auffallen	**fertig**
Nachweisbar löschen	Dokument mit allen Ableitungen löschen, dazu ein signierter Löschbericht	**fertig**
Suchprotokoll	wer hat wann in welcher Akte gesucht, signiert, mit Löschfrist	**gebaut**, erscheint mit der nächsten Version
Akten und Rechte	jede Akte ein eigener Bereich, Zugriff je Person oder Gruppe, Sperrvermerk bei Interessenkonflikt	**fertig**, geprüft
Antworten nur aus den Akten	jeder Satz mit Zitat, Antworten mit erfundenen Zitaten werden verworfen, sonst „In den freigegebenen Akten nicht gefunden“	**gemessen**, maschinell vorgeprüft, Prüfung mit einem Anwalt bei jeder Einrichtung
Hardware	Antwortzeit auf Laptop, PC mit Grafikkarte und KI-Box, mit einem und mit drei Nutzern	**gemessen** (Abschnitt 7)

Rechte gelten überall. Wer für eine Akte nicht freigeschaltet ist, soll dort nichts sehen: keinen Treffer, keine Trefferzahl, keinen Ausschnitt, keine Antwort. Das haben wir geprüft, auch über die Schnittstelle, über die Mitarbeiter auf die KI-Box zugreifen. Ein Prüfprogramm vergleicht die eingerichteten Rechte mit der Rechteliste der Kanzlei und meldet Abweichungen.

Antworten, gemessen. Wir haben 60 Fragen gestellt, als Mitarbeiterin, die nur eine von zwei Akten sehen darf: 25 Fragen zu ihrer Akte, 25 zu der gesperrten und 10 zu Mandatsdetails, die in keinem Dokument stehen. Kein Zitat und kein Ausschnitt kam aus der gesperrten Akte. Bei ihrer eigenen Akte belegte die Antwort in 20 von 25 Fällen die gesuchte Entscheidung (Modell Gemma 4 mit 12 Milliarden Parametern). Bei den erfundenen Mandatsdetails kam neunmal „nicht gefunden“. Einmal antwortete das Modell mit einer echten, wörtlich zitierten Stelle aus einer fremden Entscheidung, als gehöre sie zum gefragten Mandat. Solche Fälle fängt die Zitatprüfung allein nicht ab, weil das Zitat stimmt. Bei einer Handprobe an 20 weiteren Antworten haben wir dasselbe Muster dreimal gefunden: eine wörtlich richtige, freigegebene Stelle, die zur gefragten Sache nicht passt. Deshalb haben wir eine zusätzliche Prüfung gebaut, ob die Quelle zum gefragten Mandat gehört; in der Nachmessung verwirft sie zwei der drei Fälle, der dritte bleibt eine Aufgabe. Bei jeder Einrichtung prüfen wir außerdem 20 Antworten gemeinsam mit einem Anwalt der Kanzlei.

Keine Verbindung nach außen. So bauen wir es: Das Modul braucht im Betrieb kein Internet, auch die Lizenzprüfung läuft offline. Updates spielt die Kanzlei oder wir auf ihre Anforderung ein.

Keine Fachsoftware-Kopie. Aktenführung, Fristenkalender und beA bleiben in der Kanzleisoftware. Das Modul liest Dokumente ein und durchsucht sie. Es entscheidet nichts und versendet nichts.

Mitbestimmung. Ein Suchprotokoll ist eine technische Einrichtung, die das Verhalten von Beschäftigten erfassen kann. Wo es einen Betriebsrat gibt, ist er zu beteiligen (§ 87 Abs. 1 Nr. 6 BetrVG). Wir liefern die Beschreibung dafür mit.

Zu den Scans: Die Texterkennung ist eingebaut und erkennt in unseren nachgestellten Scans auch Paragrafenzeichen und Normzitate. Zahlen dazu veröffentlichen wir erst, wenn wir echte Papierscans mit Stempeln, Faxköpfen und Handschrift gemessen haben.

7. Die KI-Box

Die KI-Box ist ein eingerichtetes Gerät in der Kanzlei, auf dem Sprachmodell, Suche und Spracherkennung laufen. Welche Hardware es wird, entscheidet nicht ein Katalog, sondern ein Probelauf mit den Aufgaben der Kanzlei.

Schritt	Was passiert	Preis
Probelauf	20 echte Aufgaben aus dem Kanzleialltag laufen mehrfach auf lokalen Modellen, die Ergebnisse prüfen wir gegen das, was richtig ist. Am Ende steht in Zahlen, was lokal zuverlässig läuft und welche Hardware dafür nötig ist, auch wenn die Antwort „lohnt sich nicht“ lautet.	1.490 € netto, rund zwei Wochen, beim Kauf voll angerechnet
Einrichtung	Gerät mit Anmeldung, nur im Kanzleinetz erreichbar, Suche über die eigenen Dokumente, Spracherkennung für Diktate, ein halber Tag Schulung, Verpflichtung nach § 43e BRAO und Vertrag zur Auftragsverarbeitung	ab 1.900 € netto, Hardware zum Einkaufspreis
Wartung	Updates, Wechsel auf bessere Modelle, Prüfung der Datensicherung, Fernwartung nur auf Anforderung und protokolliert	ab 290 € netto im Monat

Welche Geräte reichen

Die Suche mit Fundstelle läuft auf jedem Bürorechner, auch ohne Grafikkarte. Für Antworten in ganzen Sätzen muss ein Sprachmodell je Frage rund 2.000 bis 3.000 Wörter aus den Akten lesen, bevor es schreibt. Genau dieses Lesen ist ohne Grafikkarte langsam.

Gerät	Suche	Antwort je Frage	Einschätzung
Laptop oder PC mit Intel- oder AMD-Prozessor ab 2021, 32 GB Arbeitsspeicher, ohne Grafikkarte	unter 1 s	rund 2 Minuten (gemessen, nachgestellt)	für gelegentliche Fragen einer Person, nicht für den Alltag
PC mit eigener Grafikkarte ab 16 GB Grafikspeicher	unter 1 s	wenige Sekunden (geschätzt)	gut für eine Person
Mac mit M3 Pro oder M4 Pro, 24 GB Arbeitsspeicher	unter 1 s	rund 25 bis 40 s, mit Max-Chip rund 15 s (abgeleitet)	gut für eine Person
Mac mit M3 oder M4 im Grundmodell	unter 1 s	rund 1 Minute (abgeleitet)	kaum schneller als ohne Grafikkarte
KI-Box im Büro	unter 1 s	rund 4 s, bei drei gleichzeitigen Nutzern rund 6 s (gemessen)	die Lösung für mehrere Personen

32 GB Arbeitsspeicher, weil das Modell allein rund 8 GB belegt und daneben Kanzleisoftware, Office und Browser laufen. Mit 16 GB wird es eng. Ein Mac teilt seinen Arbeitsspeicher mit dem Grafikchip, deshalb zählen dort 24 GB. Gemessen haben wir auf einem PC mit Grafikkarte (24 GB), den Laptop ohne Grafikkarte haben wir auf demselben PC nachgestellt, mit acht Rechenkernen und ohne Grafikkarte. Ein echter Laptop kann schneller oder langsamer sein. Die Mac-Werte haben wir aus veröffentlichten Messungen des Projekts llama.cpp, auf dem unsere Software für Sprachmodelle aufbaut, auf unser Modell und unsere Fragen umgerechnet.

Die KI-Box ist der klare Gewinner. Mit einer Grafikkarte dieser Klasse kam eine Antwort in rund vier Sekunden (Median 3,7 s, höchstens 5,6 s). Fragen drei Personen gleichzeitig, dauert es länger: im Median 6,4 s, höchstens 11,4 s. Kein Arbeitsplatzrechner muss aufgerüstet werden. Die Rechner der Mitarbeiter greifen mit ihrer eigenen Anmeldung auf die Box zu, die Rechte je Akte gelten dabei unverändert. Ein eigener Rechner mit Grafikkarte oder ein Mac mit Pro- oder Max-Chip ist eine Lösung für Einzelanwälte, die vor allem selbst arbeiten.

Kleinere Modelle mit 4 Milliarden Parametern antworten auch ohne Grafikkarte in rund 40 Sekunden, belegten in unserer Messung aber nur selten die richtige Entscheidung. Wir empfehlen sie deshalb nicht.

KI-Boxen richten wir bereits bei Kunden ein und betreuen sie im laufenden Betrieb. Für Kanzleien kommen die Rechte je Akte, das signierte Protokoll und die Antworten mit Fundstelle dazu.

Die Box wird gekauft wie eine Maschine. Größere Hardware, etwa ein Server mit mehreren Grafikkarten, kann über unseren Partner PrimeLine Solutions geleast werden. Eine Gebühr je Person oder je Anfrage gibt es nicht. Die Modelle stehen unter Lizenzen, die gewerbliche Nutzung erlauben.

8. Acht Fragen an jeden Anbieter lokaler KI

Ob von uns oder von jemand anderem: Diese Fragen sollte jede Kanzlei vor dem Kauf stellen und sich die Antwort schriftlich geben lassen.

1. **Welche Daten verlassen das Gerät, und wann?** Auch Nutzungsstatistiken, Fehlerberichte und Lizenzprüfungen zählen.
2. **Läuft das System ohne Internet?** Oder muss es sich regelmäßig beim Hersteller melden, um weiterzulaufen?
3. **Wie gut findet die Suche?** Gibt es eine Messung mit offengelegter Methode, oder nur ein Werbeversprechen?
4. **Woher stammt eine Antwort?** Zeigt das System für jede Aussage Dokument und Seite?
5. **Was passiert, wenn die Akten keine Antwort hergeben?** Sagt das System „nicht gefunden“, oder formuliert es etwas Plausibles?
6. **Wer sieht was?** Gelten Rechte je Akte auch für Trefferzahlen und Ausschnitte, nicht nur für das geöffnete Dokument?
7. **Welche Bestandteile stecken drin?** Gibt es eine Stückliste der Software (SBOM) mit Lizenzen?
8. **Was steht im Vertrag nach § 43e BRAO?** Wer wartet, wer hat Zugriff, wer sind die Unterauftragnehmer?

9. So fangen Kanzleien bei uns an

1. **Erstgespräch, 30 Minuten, kostenlos.** Wo bleibt die Zeit hängen, und welche Aufgaben kämen für lokale KI in Frage?
2. **Probelauf der KI-Box.** Zwanzig eigene Aufgaben, Ergebnis in Zahlen.
3. **Gerät und Einrichtung**, wenn der Probelauf es trägt. Suche mit Fundstelle über die eigenen Dokumente ab dem ersten Tag.
4. **Kanzlei-Modul einrichten.** Bevor das Modul im Alltag genutzt wird, prüfen wir mit einem Anwalt der Kanzlei 20 Antworten aus den eigenen Akten. Echte Akten kommen erst mit Verpflichtung nach § 43e BRAO und Vertrag zur Auftragsverarbeitung auf das Gerät.

10. Grenzen dieser Messung

- **Entscheidungen sind keine Akten.** Gerichtsentscheidungen sind sauber gesetzt und einheitlich aufgebaut. Schriftsätze, Korrespondenz und Scans sind unordentlicher. Wir erwarten dort eher schlechtere Werte und messen bei jeder Einrichtung an den Akten der Kanzlei nach.
- **50 Fragen.** Das reicht, um große Unterschiede zwischen Modellen zu sehen. Ein Unterschied von wenigen Prozentpunkten wäre bei dieser Zahl Zufall.
- **Die Fragen stammen von uns.** Wir haben sie ohne die Wörter der Stelle formuliert, aber wir kannten die Stelle. Fragen echter Anwälte können anders ausfallen.
- **Seite genau.** Oft stand oben eine andere Stelle derselben Entscheidung, die denselben Punkt behandelt. Wir haben das nicht als Treffer der Seite gezählt, als Treffer der Entscheidung schon.
- **Antworten nur maschinell gezählt.** Gemessen ist, ob eine Antwort die gesuchte Entscheidung wörtlich zitiert und nichts aus gesperrten Akten enthält. Ob jeder Satz inhaltlich stimmt, haben wir an 20 Antworten maschinell vorgeprüft, mit einem anderen Modell als dem antwortenden. Die Prüfung mit einem Anwalt gehört zu jeder Einrichtung.
- **Hardware auf einem Gerät.** Alle Zeiten stammen von einem PC mit Grafikkarte (24 GB). Den Laptop ohne Grafikkarte haben wir darauf nachgestellt. Karten mit 16 GB und Macs messen wir noch nach.

Anhang: Methode

Bestand: 1.400 Entscheidungen, mit festem Startwert zufällig gezogen aus rechtsprechung-im-internet.de, gesetzt in Liberation Serif, 10.171 Seiten. 40 Bild-PDFs mit 200 Seiten, nachgestellt mit 200 dpi, leichter Drehung, Unschärfe, Rauschen und JPEG-Kompression. Prüfsummen aller Dateien liegen vor.

Fragen: 50 Stellen, gezogen mit festem Startwert, je zur Hälfte Leitsätze und Absätze aus den Gründen. Fragen von Hand formuliert. Treffer: richtige Entscheidung unter den ersten k Ergebnissen. MRR: Mittelwert aus 1 geteilt durch die Position des ersten richtigen Treffers.

Suchmodelle: model2vec (256 Dimensionen, in Cyberbrain eingebaut), embeddinggemma 300m in 4-Bit-Quantisierung (768 Dimensionen) und bge-m3 (1.024 Dimensionen), beide über Ollama auf demselben Gerät. Suche kombiniert aus Stichwort- und Vektorsuche, je Anfrage zehn Treffer.

Gerät: Mietserver mit 12 Threads (AMD EPYC, 2,0 GHz) und 23 GB Arbeitsspeicher, ohne Grafikkarte. Cyberbrain 0.10.0.

Antworten und Hardware (8. November 2026): derselbe Bestand, geteilt in zwei Akten, Abruf über die Schnittstelle mit der Anmeldung einer Person, die nur eine Akte sehen darf, je Frage acht Abschnitte. Antwortmodelle über Ollama, Temperatur 0. Verglichen: Gemma 4 mit 4 und 12 Milliarden Parametern, Qwen 3 mit 8, 14 und 27 Milliarden, Phi-4 und Gemma 3. Gewählt wurde Gemma 4 12B in 4-Bit-Quantisierung: 11 von 12 Fragen mit belegter Entscheidung, wie das größte getestete Modell, aber dreimal so schnell. PC mit AMD Ryzen 9 5900X, 64 GB Arbeitsspeicher und NVIDIA RTX 3090 (24 GB). Laptop nachgestellt mit acht Rechenkernen ohne Grafikkarte: Median 116 Sekunden je Antwort.

Krynex Labs UG (haftungsbeschränkt). Alle Messungen im November 2026 auf eigener Infrastruktur. Jede Zahl ist an den Rohdaten oder der genannten Quelle geprüft. Dieses Papier ist keine Rechtsberatung.

Läuft das auch mit euren Akten?

Kostenloses Erstgespräch, 30 Minuten: www.krynexlabs.de/kontakt. Danach zeigt ein Probelauf mit 20 eigenen Aufgaben, was auf einem Gerät im Haus zuverlässig läuft.