// KI-Hub VON STRUKTURAFLOW

KI PRAKTISCH

Lokale KI für Steuerberater: Mandantendaten bleiben im Haus

Lokale KI läuft auf eigener Hardware — die einzige Architektur, bei der Mandantendaten ins System dürfen, weil sie das Haus nie verlassen. Mit Hardware-Rechnung.

AUTOR
Natascha Reiner

VERÖFFENTLICHT
3. Juli 2026

ONLINE LESEN
https://wissen.strukturaflow.it.com/lokale-ki-fuer-steuerberater-mandantendaten-bleiben-im-haus/

Alle gängigen KI-Tools — ChatGPT, Perplexity, NotebookLM, Copilot — teilen dieselbe Grenze: Mandantendaten dürfen nicht hinein, weil die Daten das Haus verlassen. Lokale KI dreht genau das um.

Bei lokaler KI läuft das Sprachmodell auf Ihrer eigenen Hardware. Das Modell kommt zu den Daten — nicht die Daten zum Modell.

Damit ist es die einzige Architektur, bei der Mandantendaten überhaupt in ein KI-System dürfen: Es gibt keinen US-Anbieter, keinen Drittlandtransfer, kein AVV-Roulette.

Für die Brot-und-Butter-Arbeit — zusammenfassen, extrahieren, klassifizieren, umschreiben, interne Recherche per RAG — sind offene Modelle 2026 gut genug. An die Spitzenmodelle für komplexes Denken reichen sie nicht heran.

Der ehrliche Preis: einmalige Hardware statt laufender Cloud-Gebühren — aber Sie tragen Wartung, Updates und Sicherheit selbst.

Lokale KI ist kein Nischenthema für Technik-Enthusiasten mehr — sie ist Teil der digitalen Souveränität österreichischer Unternehmen. 2026 läuft ein brauchbares Sprachmodell auf einer ordentlichen Workstation — mit voller Datenhoheit, ohne Token-Rechnung, ohne Anbieter im Ausland. Dieser Guide erklärt, was das in der Kanzlei wirklich bedeutet, was es leistet und was nicht, und was es ehrlich kostet.

Was „lokale KI„ bedeutet – und warum das die Spielregeln ändert

Die großen Cloud-Dienste funktionieren nach einem einfachen Prinzip: Sie tippen etwas ein, Ihre Eingabe reist zu einem Rechenzentrum (meist in den USA), wird dort verarbeitet, die Antwort kommt zurück. Das Modell ist riesig, leistungsfähig — und gehört jemand anderem.

Lokale KI kehrt das um. Sie nutzen ein offenes Modell (frei verfügbare Modelle wie Qwen, Gemma, Mistral oder gpt-oss), das Sie auf Ihre eigene Hardware herunterladen und dort betreiben. Ein Werkzeug wie Ollama macht das fast so einfach wie das Installieren eines Programms; eine Oberfläche wie Open WebUI gibt Ihrem Team ein vertrautes Chatfenster, ganz ohne Kommandozeile.

Der entscheidende Satz lautet: Das Modell kommt zu den Daten, nicht die Daten zum Modell. Ihre Anfrage, Ihre Dokumente, Ihre Mandantenakten verlassen die Kanzlei nicht. Es gibt keine Leitung nach Virginia. Was im Haus passiert, bleibt im Haus.

Der eigentliche Grund: § 80 WTBG ohne Kompromiss

Hier liegt der Kern. Die Verschwiegenheitspflicht der Wirtschaftstreuhänder nach § 80 Wirtschaftstreuhandberufsgesetz 2017 erfasst alle anvertrauten Mandantenangelegenheiten samt Betriebs- und Geschäftsgeheimnissen. Genau deshalb gilt für Cloud-KI die eiserne Regel: keine personenbezogenen Mandantendaten ohne geprüften Auftragsverarbeitungsvertrag und belastbare Rechtsgrundlage.

Bei lokaler KI fällt dieser ganze Problemkomplex weg — nicht, weil man die Regel umgeht, sondern weil ihr Auslöser entfällt:

Kein Drittlandtransfer. Es werden keine Daten an einen Anbieter übermittelt, also stellt sich die Frage nach US-Servern und Schutzgarantien gar nicht.

Kein AVV nötig für das Modell selbst. Sie verarbeiten in eigener Verantwortung auf eigener Infrastruktur. Es gibt keinen Auftragsverarbeiter, dessen Vertragsbedingungen Sie prüfen müssten.

Keine versteckten Einschränkungen. Keine „Datenregion greift hier leider nicht"-Fußnote, kein Trainingsvorbehalt, kein stilles Mitlesen.

Das heißt nicht, dass lokale KI automatisch „DSGVO-fertig„ ist. Sie verarbeiten weiterhin personenbezogene Daten und brauchen die normalen technischen und organisatorischen Maßnahmen: Zugriffsschutz, Verschlüsselung, Backups, ein sauberes Berechtigungskonzept. Aber der schwierigste Teil — die rechtssichere Datenweitergabe an einen externen KI-Dienst — entfällt vollständig.

Aus „Das dürfen wir nicht eingeben" wird „Das läuft bei uns im Haus.„

Damit ist lokale KI die einzige Architektur, bei der ein Steuerberater echte Mandantendaten durch ein KI-System laufen lassen kann, ohne mit der Verschwiegenheitspflicht in Konflikt zu geraten. Mehr Hintergrund dazu in unserem Leitfaden KI DSGVO-konform einsetzen.

Was eine lokale KI in der Kanzlei wirklich leistet

Die ehrliche Einordnung vorweg: Ein lokal betriebenes Modell ist nicht so brillant wie das jeweils neueste Spitzenmodell aus der Cloud. Für tiefes, mehrstufiges Schlussfolgern oder kreative Höchstleistung merkt man den Unterschied.

Für die wiederkehrende Kanzleiarbeit dagegen kaum.

Diese Aufgaben erledigt ein gutes lokales Modell zuverlässig:

Zusammenfassen: lange Schreiben, Protokolle, Aktenkonvolute auf das Wesentliche bringen — auch mit echtem Mandantenbezug, weil nichts das Haus verlässt.

Extrahieren: strukturierte Daten aus Belegen, Verträgen oder Schreiben herausziehen.

Klassifizieren: Eingangspost dem richtigen Mandat oder der richtigen Kategorie zuordnen.

Umformulieren: Aktenvermerke, Mandantenbriefe und Erklärtexte aus Stichworten entwerfen — diesmal direkt mit den realen Fallangaben.

Interne Recherche: ein durchsuchbares Kanzlei-Wissen aufbauen (das sogenannte RAG-Prinzip), das Fragen aus Ihren eigenen Unterlagen mit Quellenverweis beantwortet — inklusive interner, vertraulicher Dokumente.

Der qualitative Sprung gegenüber den Cloud-Tools liegt im letzten Punkt: Während Perplexity oder NotebookLM nur mit anonymisiertem oder öffentlichem Material gefüttert werden dürfen, darf in die lokale KI auch das hinein, was sonst nie ein KI-System sehen dürfte — weil es Ihr Server ist.

Was lokale KI nicht leistet: Sie kennt das offene Web nicht (eine tagesaktuelle Gesetzesänderung müssen Sie selbst einspeisen), sie erreicht nicht das Spitzenniveau der größten Cloud-Modelle, und sie erzeugt — wie jedes Sprachmodell — gelegentlich überzeugend klingende, aber falsche Aussagen (Halluzinationen). Die fachliche Prüfung bleibt Ihre Aufgabe.

Wie groß ist der Rückstand wirklich?

Und dieser Rückstand ist kleiner, als die Schlagzeilen vermuten lassen. Das Forschungsinstitut Epoch AI misst laufend, wie weit die besten frei verfügbaren Modelle hinter den teuersten Spitzenmodellen liegen — nicht in Prozentpunkten auf irgendeinem Benchmark, sondern in Zeit: Wie viele Monate dauert es, bis ein offenes Modell das Niveau erreicht, das ein geschlossenes Spitzenmodell heute schon hat?

Ende 2024 lag dieser Abstand noch bei rund einem Jahr. 2026 sind es im Schnitt nur noch wenige Monate — je nach Aufgabe grob vier bis sieben.

Für die Kanzlei ändert das die Perspektive auf den Preis: Bei der Cloud zahlen Sie dauerhaft einen Aufpreis für einen Vorsprung von wenigen Monaten — einen Vorsprung, der für das Zusammenfassen, Extrahieren und Umformulieren ohnehin keine Rolle spielt. Wo es tatsächlich auf das absolute Spitzenniveau ankommt, greifen Sie gezielt zum Cloud-Modell.

Für alles andere gilt: Das offene Modell von heute ist das, was das teure Modell vor einem halben Jahr war — mit dem entscheidenden Unterschied, dass es bei Ihnen im Haus läuft.

Die ehrliche Hardware- und Kostenrechnung

Diesen Teil schreibt fast niemand offen — also tun wir es.

Was Sie brauchen, hängt von der Modellgröße ab (gemessen in „Parametern“, grob: mehr Parameter = mehr Können, aber auch mehr Speicherbedarf):

MODELLKLASSE	REALISTISCHE HARDWARE	EIGNUNG IN DER KANZLEI
7–8 Mrd. Parameter	Ordentliche Workstation, ~8–16 GB RAM, auch ohne Grafikkarte (dann langsamer)	Einstieg: Zusammenfassen, Klassifizieren, einfache Texte
13–14 Mrd. Parameter	16 GB RAM + Grafikkarte mit 8–12 GB Speicher	Der Sweet Spot für die meisten Kanzlei-Aufgaben
30 Mrd. Parameter	Grafikkarte mit ~24 GB Speicher	Höhere Qualität, kleiner Server sinnvoll
70 Mrd. Parameter	Server mit 40 GB+ Grafikspeicher	Für Kanzleien selten nötig

Eine Technik namens Quantisierung halbiert den Speicherbedarf bei nur geringem Qualitätsverlust — die Standardeinstellung der meisten Werkzeuge.

Realistisch für eine kleine bis mittlere Kanzlei: Ein 14-Mrd.-Parameter-Modell auf einer Workstation mit guter Grafikkarte deckt den Großteil der Aufgaben ab und liefert für den Alltag eine Qualität, die sich von den Cloud-Tools kaum unterscheidet.

Wer es zunächst kleiner halten will, kann ein 7-Mrd.-Modell sogar auf einem europäischen Server (etwa in Deutschland) ab rund 50 € im Monat betreiben — langsamer, aber für Hintergrundaufgaben brauchbar.

Die eigentliche Rechnung ist eine Capex-gegen-Opex-Entscheidung: Cloud-KI kostet pro Nutzer und Monat, dauerhaft, und die Daten gehen außer Haus. Lokale KI kostet einmalig Hardware (eine gute Workstation oder ein kleiner Server bewegt sich im niedrigen vierstelligen Bereich) plus Strom — und die Daten bleiben da.

Bei mehreren Nutzern und sensiblem Datenbestand kippt diese Rechnung über die Jahre klar zugunsten der lokalen Variante.

Der Preis dafür ist nicht finanziell: Sie übernehmen Verantwortung für Betrieb, Updates und Sicherheit — es gibt keinen Microsoft-Support, an den Sie sich lehnen.

Welche Modelle 2026 – und warum das die kleinere Frage ist

Die offene Modelllandschaft bewegt sich monatlich. Aktuell relevant für den lokalen Einsatz mit guter Deutsch-Fähigkeit:

Qwen (Alibaba) — der pragmatische Standard 2026: stark mehrsprachig, in vielen Größen verfügbar, freie Lizenz.

Gemma (Google) — sehr breite Sprachunterstützung, läuft schon in kleineren Größen ordentlich.

Mistral (Frankreich/EU) — europäische Modellgewichte, solide im Deutschen — relevant, wenn die Herkunft des Modells für Sie zählt.

gpt-oss (OpenAI) — die offenen Modelle von OpenAI, die kleinere Variante läuft bereits mit überschaubarem Speicher.

Wichtiger als die Tagesform eines einzelnen Modells ist diese Eigenschaft: Das Modell ist austauschbar. Kommt nächsten Monat ein besseres heraus, tauschen Sie es — die Architektur, Ihre Daten und Ihre Abläufe bleiben gleich. Sie binden sich nicht an einen Anbieter.

Genau diese Unabhängigkeit ist der strategische Wert hinter dem ganzen Ansatz.

Wie der Einstieg realistisch aussieht

Lokale KI ist kein Großprojekt, mit dem man das Jahr blockiert. Der vernünftige Weg ist klein und konkret:

Schritt 1 — Ein Anwendungsfall, eine Maschine. Beginnen Sie mit einer Workstation und einer einzigen Aufgabe, die echten Mandantenbezug hat und heute Zeit kostet — etwa das Zusammenfassen umfangreicher Eingangsschreiben.

Schritt 2 — Vertraute Oberfläche fürs Team. Über eine Chat-Oberfläche wie Open WebUI nutzt das Team das lokale Modell wie einen gewohnten Chatbot — ohne technische Hürde, ohne Kommandozeile.

Schritt 3 — Wissen anbinden. Im nächsten Schritt verbinden Sie das Modell mit Ihren eigenen Unterlagen (RAG), sodass es Fragen aus dem internen Kanzleiwissen beantwortet — mit Verweis auf die Quelle.

Schritt 4 — Bei Bedarf zum kleinen Server wachsen. Bewährt sich der Ansatz, wandert das Modell von der Einzel-Workstation auf einen kleinen Server, auf den mehrere Mitarbeiter zugreifen.

Das ist der Punkt, an dem sich die Frage stellt, ob Sie das intern aufbauen oder mit einem Partner umsetzen, der die Einrichtung, Absicherung und Anbindung sauber macht. Eine lokale KI ist schnell zum Laufen gebracht — sie richtig und betriebssicher aufzusetzen, ist die eigentliche Arbeit.

Wo lokale KI an Grenzen stößt

Damit der Blick ehrlich bleibt:

Kein Spitzenniveau. Für die anspruchsvollsten Denkaufgaben bleibt ein aktuelles Cloud-Spitzenmodell überlegen. Lokale Modelle decken den Alltag ab, nicht die Champions League.

Sie tragen die Verantwortung. Betrieb, Updates, Sicherung, Ausfallsicherheit — das alles liegt bei Ihnen oder Ihrem Dienstleister. Das ist der Preis der Unabhängigkeit.

Einmalige Einstiegshürde. Hardware muss angeschafft und eingerichtet werden. Der Komfort eines „Häkchen setzen, fertig„ der Cloud-Dienste fehlt.

Kein Webwissen out of the box. Tagesaktuelles müssen Sie selbst zuführen — hier ergänzen sich lokale KI und ein Web-Werkzeug wie Perplexity sinnvoll.

Praxis-Einschätzung – für wen lohnt sich lokale KI?

Lohnt sich, wenn:

- Sie regelmäßig mit hochsensiblen Mandantendaten arbeiten und KI dort einsetzen wollen, wo Cloud-Tools an § 80 WTBG scheitern
- Digitale Souveränität und Unabhängigkeit von US-Anbietern für Sie strategisch zählen
- mehrere Mitarbeiter KI nutzen sollen und die laufenden Cloud-Kosten ins Gewicht fallen
- jemand im Haus oder ein Partner die Infrastruktur betreuen kann

Lohnt sich weniger, wenn:

- Sie KI nur gelegentlich für allgemeine, unkritische Aufgaben brauchen — dann ist ein abgesichertes Cloud-Tool einfacher
- niemand für Betrieb und Wartung zur Verfügung steht und kein Dienstleister eingebunden wird
- Sie ausschließlich auf das absolute Leistungsmaximum eines einzelnen Modells angewiesen sind

Für die meisten Kanzleien ist die ehrliche Antwort: ein abgesichertes Cloud-Tool für das Unkritische — und eine lokale KI für genau das, was das Haus nicht verlassen darf. Diese Kombination ist kein Widerspruch, sondern die durchdachte Aufteilung.

Nächste Schritte – der souveräne Weg zur KI in Ihrer Kanzlei

Lokale KI ist kein Verzicht auf Komfort, sondern eine Entscheidung für Datenhoheit. Für eine verschwiegenheitspflichtige Kanzlei ist sie oft nicht die exotische, sondern die naheliegende Variante — gerade dort, wo Mandantendaten im Spiel sind.

Welche Prozesse sich dafür eignen, welche Modellgröße zu Ihrem Bedarf passt und wie der Einstieg betriebssicher gelingt, klären wir im kostenlosen Erstgespräch vor Ort oder online. Für Steuerkanzleien, die den Schritt zur vollständig selbstgehosteten KI gehen wollen, zeigen wir auch den konkreten Aufbau.

NÄCHSTER SCHRITT

Mehr praktische KI-Anleitungen für KMU

Dieser Artikel ist Teil des KI-Hubs von Strukturaflow — einer deutschsprachigen Plattform für den praktischen KI-Einsatz in kleinen und mittleren Unternehmen.

https://wissen.strukturaflow.it.com