

Die wahren Kosten von KI

Tokens, Consumption-based Pricing und was Kanzleien jetzt bedenken müssen

von **Thorsten Walter und Dominic Piernot**

24. September 2026, 11:00 Uhr | Lesezeit: 9 Min.

CyberSecurityQuarterly
aus Ausgabe: 3|2026

Kanzleien, die spezialisierte Systeme für künstliche Intelligenz (KI) in der Mandatsarbeit einsetzen, arbeiten häufig noch mit Verträgen auf Fixpreisbasis: monatliche Lizenz pro Nutzer, planbarer Betrag, überschaubares Budget. Dieses Modell ist vertraut. Es entspricht dem, was Sozietäten von klassischer Software aus dem Kanzleialltag kennen. Es spiegelt jedoch nicht die Logik eines stark voranschreitenden Markttrends wider: die der tokenbasierten KI-Systeme. Sogenannte Consumption-Based- oder Token-/Credit-Abrechnungsmodelle sind spätestens seit Claude „Fable“ in aller Munde.

Das ist kein Ausnahmefall – um nur eine Auswahl namhafter Anbieter zu nennen: Harvey, Legora, Libra. Führende Legal-AI-Anbieter sind mit Fixpreisen beziehungsweise Abomodellen gestartet. Jetzt stellen sie um, weil die Abrechnungslogik der zugrundeliegenden Modelle sie dazu zwingt. Wer KI wie klassische Software eingekauft hat, steht plötzlich vor einer Kalkulation, auf die er nicht vorbereitet ist. Die Überraschung kommt nicht aus der Vergangenheit. Sie kommt in dem Moment, in dem der Anbieter das Modell wechselt. Der fortschreitende Drang der Frontier-Modellhersteller wie OpenAI, Anthropic und anderen wirkt sich auf alle Systeme aus.

Vom Fixpreis zur verbrauchsbasierten Abrechnung

Die Kosten klassischer Software errechnen sich nach einem einfachen Prinzip: Anzahl der Nutzer mal Lizenzgebühr ergibt den Monatspreis. Ob jemand die Software täglich oder gar nicht benutzt, spielt für die Rechnung keine Rolle. Dieser Ansatz hat Vorteile. Die Kosten sind transparent und vor allem planbar. Ein Budget lässt sich im Vorhinein kalkulieren und festlegen.

Viele Legal-AI-Anbieter haben dieses Modell übernommen. Als Dauerlösung ist es jedoch nicht angelegt, denn es entspricht nicht der Logik der zugrundeliegenden Technologie. Die Kosten entstehen nicht durch den Zugang zum System, sondern durch jede einzelne Anfrage, jedes analysierte Dokument, jede generierte Antwort. Der Anbieter rechnet nach Verbrauch ab – in den meisten Fällen nach der Menge der verarbeiteten Daten, gemessen in sogenannten Tokens. Eine Kanzlei, die ihre KI kaum nutzt, zahlt wenig. Eine Kanzlei, die sie intensiv nutzt, zahlt entsprechend mehr.

Der Markt bewegt sich deshalb gerade in eine andere Richtung: weg vom Fixpreis, hin zur

verbrauchsbasierten Abrechnung.

Das Problem für Anwaltskanzleien liegt nicht im Modell selbst, sondern in der fehlenden Transparenz. Viele Nutzer wissen nicht, welche Aktionen welche Kosten auslösen. In der Regel gibt es keine transparente Kostenübersicht in Echtzeit. Aber die Rechnung am Monatsende kommt – und führt nicht selten zu unerfreulichen Überraschungen. Ein einfaches Beispiel: Es macht einen erheblichen Unterschied, ob man ein PDF oder ein Bild oder nur einen extrahierten Text in ein Sprachmodell (Large Language Model, LLM) eingibt. Auch die Sprache spielt eine Rolle: Deutschsprachige Texte benötigen für denselben Inhalt in der Regel deutlich mehr Tokens als englische (siehe [hier](#) .

Tokens – die neue Währung

Ein Token ist keine Zeile und kein Wort. Er ist eine Einheit, in die Sprachmodelle Text aufteilen, um ihn verarbeiten zu können. Grob vereinfacht entspricht ein Token etwa vier Zeichen oder drei Viertel eines englischen Worts. Wer eine Frage stellt, sendet Input-Tokens. Wer eine Antwort bekommt, empfängt Output-Tokens. Beide Richtungen kosten, Output in der Regel deutlich mehr als Input.

Diese Unterscheidung ist gerade für Kanzleien besonders relevant. Ein 300-seitiger Datenraum, der vor einer Transaktion analysiert werden soll, enthält Hunderttausende von Input-Tokens. Die Kosten dafür sind berechenbar, wenn man weiß, wie viele Seiten verarbeitet werden.

Schwieriger wird es bei sogenannten Reasoning-Modellen. Reasoning-Modelle denken vor der Antwort nach. Sie durchlaufen interne Rechenschritte – sogenannte Chain-of-Thought-Prozesse –, bevor sie eine Ausgabe produzieren. Diese Denkschritte sind für den Nutzer unsichtbar, generieren aber Tokens. Wer ein Reasoning-Modell einsetzt, um eine einfache Frage zu beantworten, zahlt für Denkleistung, die er für die konkrete Frage unter Umständen gar nicht braucht. Die Kosten sind höher als erwartet, der Mehrwert überschaubar.

Die Kosten für einen Reasoning-Durchlauf können ein Vielfaches einer einfachen Abfrage betragen, auch wenn die Ausgabe für den Nutzer identisch aussieht. Welches Modell für welche Aufgabe eingesetzt wird, ist damit keine technische Entscheidung. Es ist eine Kostenentscheidung.

Versteckte Kostentreiber

Die Tokenkosten für Eingabe und Ausgabe sind nur ein Teil der Rechnung. Daneben gibt es Kostentreiber, die in der initialen Kalkulation kaum auftauchen – und in der Praxis trotzdem wirksam werden.

Falsche Modellwahl ist der naheliegendste davon. Nicht jede Aufgabe braucht das

leistungsstärkste Modell. Ein kurzes Anschreiben benötigt kein Reasoning-Modell. Eine Due-Diligence-Analyse über Hunderte Dokumente hingegen schon. Der Preisunterschied zwischen einem einfachen und einem Reasoning-Modell kann erheblich sein. Wer ohne Differenzierung arbeitet, zahlt permanent zu viel – ohne Qualitätsvorteil.

Agentic AI ist ein weiteres Beispiel. Wer KI nicht als einzelnes Werkzeug nutzt, sondern als verkettete Abfolge von Schritten (Dokument lesen, Informationen extrahieren, Zusammenfassung erstellen, Ergebnis prüfen), multipliziert die Tokenkosten. Jeder Schritt ist eine eigene Anfrage. Jede Anfrage kostet. Ein mehrstufiger Prozess kann das Zehnfache einer einfachen Abfrage kosten.

Wiederholungsschleifen verstärken das Problem. Wenn ein Modell eine Aufgabe nicht korrekt löst und der Prozess automatisch neu startet, läuft die Uhr erneut. Jeder Fehlversuch kostet genauso wie ein erfolgreicher Durchlauf. Systeme ohne Fehlertoleranz und Abbruchbedingungen verursachen unter Umständen erhebliche Mehrkosten.

Schlechtes Prompting ist ein weiterer Faktor. Ungenaue Anweisungen führen zu ungenauen Antworten. Die Nachbearbeitung kostet zusätzliche Tokens – und oft auch Arbeitszeit. Eine Kanzlei, die ihre Prompts nicht strukturiert und testet, zahlt implizit für Ineffizienz.

Hinzu kommt die menschliche Qualitätssicherung. KI-Ausgaben müssen geprüft werden. Für Kanzleien ist das keine Option, sondern Pflicht. Wer Stunden für die Prüfung einer KI-Analyse einrechnet, muss diese Kosten in das Gesamtbild aufnehmen. Die günstige KI-Rechnung kann durch den Prüfaufwand schnell unwirtschaftlich werden.

Vertrag und Mandat

Variable KI-Kosten berühren zwei Vertragsbeziehungen gleichzeitig: die mit dem Anbieter, der die Leistung erbringt – und die mit dem Mandanten, der am Ende für das Ergebnis zahlt. Beide verlangen unterschiedliche Antworten.

Beim Anbietervertrag kommt es auf drei Punkte an:

- **Transparenz:** Der Anbieter muss belastbar ausweisen, wie Kosten berechnet werden. Reine Gesamtbeträge auf der Monatsrechnung reichen nicht. Kanzleien brauchen eine aufgeschlüsselte Nutzungshistorie: welche Anfragen, welches Modell, wie viele Tokens, welcher Preis. Nur so lässt sich nachvollziehen, wo die Kosten entstehen – und ob eine Rechnung korrekt ist.
- **Kostendeckel:** Ein Vertrag ohne Ausgabenobergrenze ist für eine Kanzlei mit variablem Verbrauch ein offenes Risiko. Verhandlungspunkt ist ein monatliches oder jährliches Maximum, ab dem der Anbieter auf Mehrkosten verzichtet oder den Dienst vorübergehend drosselt. Viele Anbieter bieten Budget-Alerts an. Diese sollten vertraglich gesichert und nicht nur als technisches Feature behandelt werden.
- **Abrechnungsprüfung:** Kanzleien sollten das Recht haben, Abrechnungen auf Plausibilität zu prüfen – idealerweise mit eigenem Zugang zu den Nutzungsdaten in Echtzeit. Was für klassische IT-Verträge selbstverständlich ist, fehlt bei KI-Anbietern oft noch.

Wie variable KI-Kosten in die Abrechnung gegenüber dem Mandanten einfließen, ist eine eigene und komplexere Frage. Sie betrifft nicht nur einzelne Kostenpositionen, sondern die Grundlogik der Vergütungsmodelle, mit denen Kanzleien arbeiten.

Vergütungsmodelle im Wandel

KI verändert nicht nur die Kosten der Kanzlei. Sie verändert die Logik, nach der Kanzleien abrechnen.

Im klassischen Stundensatzmodell entsteht ein Doppelrisiko. KI reduziert die abrechenbaren Stunden. Gleichzeitig entstehen neue Kosten durch die Nutzung der KI. Wer diese Kosten trägt, ist die entscheidende Frage. Die naheliegende Lösung – Tokenkosten als Auslage an den Mandanten weitergeben – ist rechtlich nicht so einfach, wie sie klingt. Eine Abrechnung als Auslage nach dem Rechtsanwaltsvergütungsgesetz (RVG) scheidet aus. Der Katalog in Teil 7 des Vergütungsverzeichnisses (VV) zum RVG ist abschließend. KI-Nutzungskosten lassen sich in keine der dort genannten Positionen einordnen. Für einen Aufwendungsersatz nach § 670 Bürgerliches Gesetzbuch (BGB) fehlt es an anerkannter Praxis und Rechtsprechung. Bleibt nur eine transparente und – soweit standardmäßig verwendet – den Anforderungen des AGB-Rechts (Allgemeine Geschäftsbedingungen, AGB) standhaltende Regelung in der Mandatsvereinbarung. Wer KI-Kosten ohne belastbare vertragliche Grundlage in Rechnung stellt, riskiert die Uneinbringlichkeit der Kosten. Wer sie nicht weitergibt, trägt sie selbst und verschlechtert seine Marge.

Während das Stundensatzmodell also von zwei Seiten unter Druck steht, liegt beim Pauschalhonorar und bei der RVG-Abrechnung ein Einfrontenproblem vor. Die

Einnahmen sind fix – vertraglich beim Pauschalhonorar, gesetzlich beim RVG. Die Tokenkosten sind es nicht. Ein Mandat mit einfachem Sachverhalt verursacht wenige Tokens. Eines mit Hunderten Dokumenten und mehrstufigen KI-Workflows kann die Marge allein durch Tokenkosten aufzehren. Die Kanzlei trägt das volle Verwendungsrisiko.

Anders als beim Stundensatzmodell wirkt effizienter KI-Einsatz auch in die andere Richtung: Wer die KI gezielt nutzt, verbessert seine Marge – in der Pauschalabrechnung ebenso wie im RVG. Die Einnahmen bleiben fix, der Aufwand sinkt. Dieser Effizienzgewinn ist die größte Renditechance, die KI für Kanzleien bereithält, solange die Tokenkosten unter Kontrolle bleiben.

Wettbewerbsdruck und Preislogik

Das strukturelle Problem betrifft zunächst vor allem Kanzleien mit gewerblichen Mandanten, die Preise und Leistungen aktiv vergleichen. Wenn immer mehr KI-gestützte Kanzleien schneller und günstiger arbeiten, werden diese Mandanten fragen, warum sie für Routinearbeit denselben Stundensatz zahlen sollen wie zuvor. Im klassischen Privatmandatsverhältnis mit seiner hohen Informationsasymmetrie mag dieser Druck langsamer spürbar werden – strukturell ist er angelegt. Effizienzgewinne, die heute in der Marge landen, werden durch Wettbewerb sukzessive auf der Preisseite weitergegeben. Dieser Druck ist bereits spürbar – und er wird zunehmen.

Die Antwort liegt im wertbasierten Vergütungsmodell, also der Abrechnung nach dem wirtschaftlichen Wert des Ergebnisses für den Mandanten, nicht nach dem Zeitaufwand des Anwalts. Abrechnung nach Ergebnis und Expertise – nicht nach Zeit. Das ist das Modell, das zu KI passt. Wer eine Due Diligence in der halben Zeit mit besserer Qualität liefert, sollte dafür nicht weniger verdienen, sondern anders abrechnen. Der Weg dorthin erfordert einen Markt, der dieses Gespräch führen will – und Kanzleien, die es initiieren.

Im Rahmen des Stundensatzmodells bleibt nur die Weitergabe der KI-Kosten an den Mandanten auf Grundlage einer gesondert und transparent ausgewiesenen Position im Mandatsvertrag. Das setzt eine belastbare Kalkulation voraus und erfordert das Gespräch mit dem Mandanten.

Praxisempfehlungen und Ausblick

Der wichtigste Schritt ist Kostenmonitoring von Beginn an. Wer KI einführt, sollte gleichzeitig ein Trackingsystem einrichten. Welches Modell wird wie oft für welche Aufgabe genutzt? Was kostet eine Due-Diligence-Analyse, was eine Vertragsprüfung, was eine Rechtsrecherche? Diese Daten sind die Grundlage für jede weitere Entscheidung.

Pilotphasen mit transparenter Messung der Kosten sind sinnvoll. Eine dreimonatige

Testphase ohne systematische Erfassung der Nutzungsdaten liefert keine belastbaren Erkenntnisse. Die maßgebliche Frage ist nicht, ob die KI gut funktioniert, sondern ob sie wirtschaftlich arbeitet.

Eine interne Leitlinie zur Modellwahl gehört zu den wirksamsten Maßnahmen der Kostenkontrolle (siehe hierzu Abschnitt „Versteckte Kostentreiber“). Wer einmal festlegt, welches Modell für welchen Aufgabentyp gilt, senkt die Kosten dauerhaft – ohne Qualitätsverlust.

KI-Kompetenz ist dabei mehr als ein operatives Hilfsmittel. Wer weiß, welches Modell er für welchen Zweck einsetzt und wie er einen Prompt präzise formuliert, produziert bessere Ergebnisse mit weniger Tokenkosten. Das ist eine anwaltliche Fachkompetenz neuer Art – und sie ist in keinem der klassischen Vergütungsmodelle bisher abgebildet. Genau dort liegt das Argument für wertbasierte Abrechnung: Überlegene Ergebnisse durch besseren KI-Einsatz rechtfertigen eine andere Preislogik – nicht eine niedrigere.

Kanzleien, die in Richtung wertbasierter Vergütung denken, sollten dieses Gespräch mit Mandanten früh und von sich aus führen – nicht als Preisverhandlung, sondern als Transparenzangebot. Grundlage ist nicht der Aufwand, sondern der Nutzen für den Mandanten: der vermiedene Schaden, das gesicherte Ergebnis, die beschleunigte Transaktion. Wer diesen Wert benennen kann, hat das entscheidende Argument für eine andere Preislogik.

Wer dieses Gespräch nicht führt, wird es irgendwann geführt bekommen – auf Initiative des Mandanten.

Vieles spricht dafür, dass die Tokenpreise langfristig sinken – der Wettbewerb zwischen den Anbietern ist intensiv, die Technologie wird effizienter. Gleichzeitig steigt der Verbrauch: mehr Anwendungsfälle, komplexere Workflows, mehr Automatisierung. Die zentrale Frage wird nicht sein, ob KI teuer ist. Sie wird lauten: Was ist uns dieser konkrete Anwendungsfall wert? Wer das heute für seinen Betrieb beantwortet, hat morgen einen Vorteil.