parashift

Sovereign European Document AI

WHITEPAPER

Dokumentli® – Das spezialisierte Vision-Language-Model für Document Intelligence

Dokumentli® erreicht bei der Dokumentenverarbeitung eine höhere Genauigkeit und Geschwindigkeit als general-purpose Frontier-Modelle, bei planbaren Kosten und der Möglichkeit, vollständig air-gapped auf einer einzelnen Consumer-GPU zu laufen.

88%

Ø Genauigkeit (cANLS@0,8), höchster Wert unter 7 getesteten Modellen

2.0×

schneller im Schnitt als das genaueste Referenzmodell

4

Einsatzoptionen, von On-Prem bis zur API

Air-gapped-fähig: vollständiger On-Prem-Betrieb auf einer einzelnen Consumer-GPU, ohne Rechenzentrum, ohne Cloud-Anbindung.

EXECUTIVE SUMMARY

Ein Modell, das general-purpose LLMs bei einer Aufgabe schlägt

Dokumentli® ist Parashifts spezialisiertes Vision-Language-Flaggschiff-KI-Modell für die automatisierte Verarbeitung von Geschäftsdokumenten. Anders als general-purpose Frontier-Modelle ist Dokumentli® auf einen einzigen Zweck trainiert: Geschäftsdokumente europäischer Unternehmen und regulierter Branchen in strukturierte Daten zu verwandeln.

In Tests auf 5 Benchmark-Datensätzen erreicht Dokumentli® die höchste durchschnittliche Extraktionsgenauigkeit unter sieben getesteten Modellen und ist zugleich im Durchschnitt das schnellste Modell im Feld. Dieses Whitepaper ordnet diese Ergebnisse ein, erklärt die Funktionsweise des Modells, grenzt seinen Einsatzbereich ehrlich ab und beschreibt die verfügbaren Einsatzoptionen.

Die vollständigen Benchmark-Zahlen inklusive Detailergebnissen je Branche und Modell sind im separaten **Parashift Dokumentli® Benchmark Report** dokumentiert; dieses Whitepaper fasst die Kernzahlen zusammen und verweist für die Tiefe auf jenen Report.

88%
Ø Genauigkeit (cANLS@0,8)
Höchster Wert unter 7 getesteten Modellen

2.0×
schneller im Schnitt
als das genaueste der sechs Referenzmodelle

4
Einsatzoptionen
On-Prem, Private Cloud, Dokumentli® Cloud API oder Parashift Plattform

Planbar
Kostenmodell
Abonnement statt unvorhersehbarer Pay-per-Call-Abrechnung

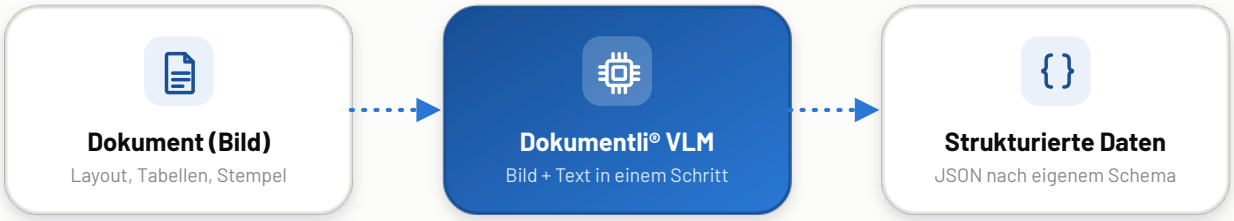
Für wen dieses Modell gebaut ist: europäische Unternehmen und regulierte Branchen mit hohem, wiederkehrendem Dokumentenvolumen, von Finanzdienstleistungen und Versicherungen bis Bau, Handel und Logistik.

Inhalt dieses Whitepapers: Was ein Small Special-Purpose Vision Language Model ist · Benchmark-Kernzahlen (Genauigkeit & Geschwindigkeit) · warum Spezialisierung bei Document-Intelligence-Aufgaben gewinnt · ehrliche Abgrenzung, wofür das Modell nicht gedacht ist · die vier Einsatzoptionen · typische Integrationsmuster · Kostenlogik im Vergleich zu General-Purpose-LLM-APIs.

GRUNDLAGEN

Was ist ein Small Special-Purpose Vision Language Model?

Wie Dokumentli® ein Dokument verarbeitet



Ein **Vision Language Model (VLM)** verarbeitet Bild und Text gemeinsam. Es liest eine Dokumentenseite direkt als Bild, inklusive Layout, Tabellen, Stempel und Handschrift, und beantwortet dazu eine textuelle Anfrage. Ein separater OCR-Vorverarbeitungsschritt ist dafür nicht erforderlich.

„Special-purpose" bedeutet: Dokumentli® ist gezielt auf Dokumentenverständnis für europäische Geschäftsdokumente trainiert, nicht auf ein möglichst breites Aufgabenspektrum.

„Small" bedeutet: Dokumentli® ist ein kompaktes Modell mit deutlich weniger Parametern als general-purpose Frontier-Modelle. Das senkt den Rechenaufwand pro Dokument und erlaubt den Betrieb auf handelsüblicher PC-Hardware mit einer einzelnen Consumer-GPU, bis hin zum Air-Gapped-Betrieb (siehe Seite 7).

Bild + Text in einem Modell: Layout, Tabellen und Textinhalt werden gemeinsam interpretiert, ohne separate OCR-Pipeline.

Auf eine Domäne spezialisiert: Trainingsdaten aus echten Geschäftsdokumenten statt allgemeinem Web-Text.

Kompakt: Weniger Parameter und geringerer Rechen- und Infrastrukturbedarf als general-purpose Frontier-Modelle.

6×
kleinere Parameterzahl von Dokumentli® gegenüber Gemma 4 26B-A4B, einem der sechs Referenzmodelle (siehe Seite 5).

1 GPU
(ab 24 GB VRAM) genügt für vollständigen, air-gapped Betrieb ohne Rechenzentrum oder Cloud-Anbindung.

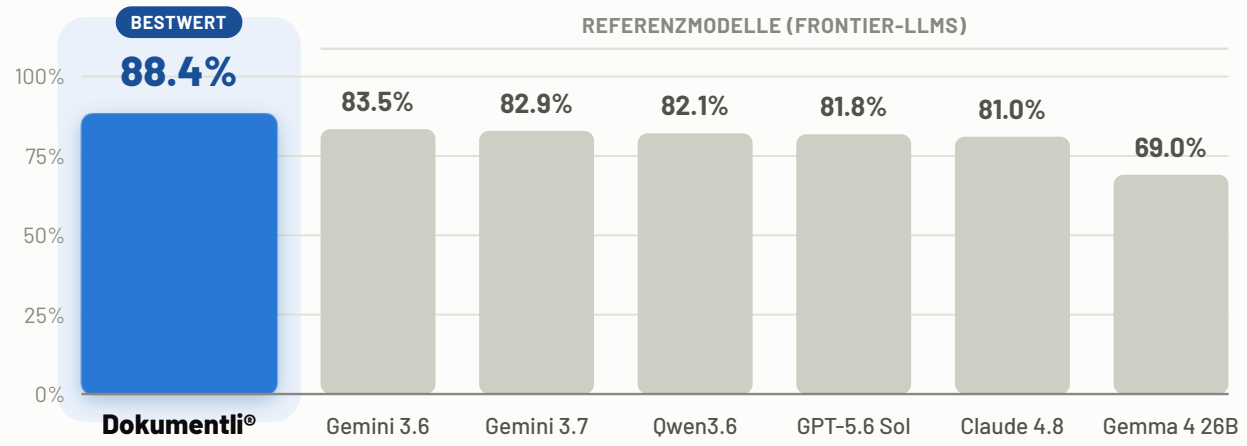
BENCHMARK-ERGEBNISSE IM ÜBERBLICK

Höchste Genauigkeit bei höchster Geschwindigkeit

In Tests auf 5 Benchmark-Datensätzen mit insgesamt 3368 realen Geschäftsdokumenten wurde Dokumentli® gegen sechs aktuelle Referenzmodelle getestet: Claude Opus 4.8, GPT-5.6 Sol, Gemini 3.6 Flash, Gemini 3.7 Flash, Qwen3.6-Plus und das offene Modell Gemma 4 26B-A4B-IT. Gütemaß ist cANLS (case-insensitive Average Normalized Levenshtein Similarity, Schwellenwert τ = 0,8), eine gängige Variante des ANLS-Industriestandards für die Bewertung von Dokumenten-Extraktion.

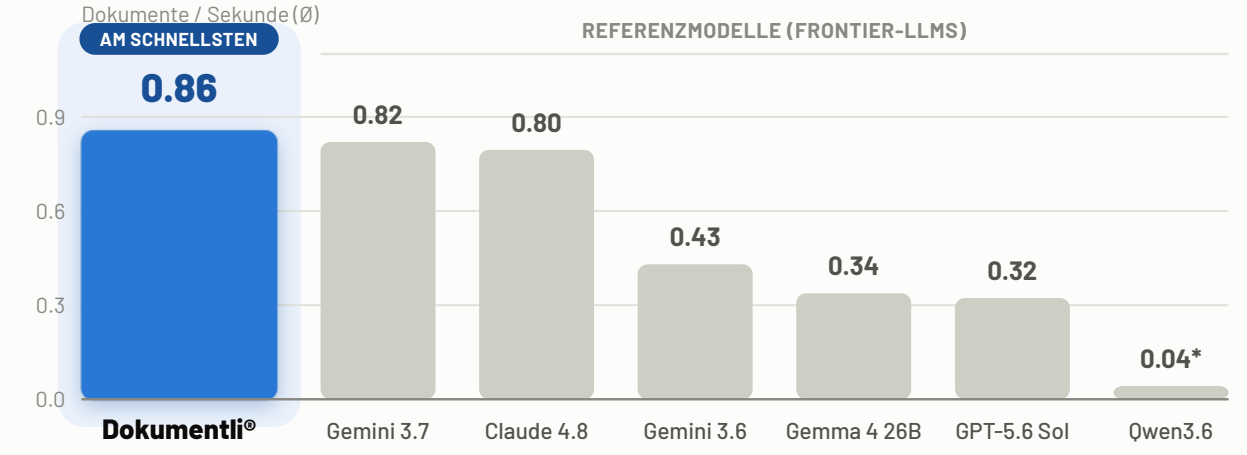
Ø Genauigkeit (cANLS@0,8) je Modell, gemittelt über 5 Benchmarks

MEHR GENAUIGKEIT



Ø Verarbeitungsgeschwindigkeit je Modell, Dokumente pro Sekunde

MEHR TEMPO



Methodik: Dokumentli® gemessen auf einer RTX PRO 6000 (Cloud-Run); Referenzmodelle über OpenRouter. *Der Qwen3.6-Wert ist auffällig niedrig, möglicherweise durch erhöhte Plattformlast während der Messung; Details siehe Benchmark Report.

5/5
Benchmarks, in denen Dokumentli® vor Gemini 3.6 Flash liegt, dem stärksten der sechs Referenzmodelle.

7
getestete Modelle insgesamt, Dokumentli® plus sechs general-purpose Frontier-Modelle als Referenz.

Für Details siehe den Benchmark Report: Ergebnisse je Branche und Modell, Konfidenzintervalle, Positionierungs-Matrizen sowie die vollständige Methodik sind im separaten **Parashift Dokumentli® Benchmark Report** dokumentiert.

EINORDNUNG

Warum ein spezialisiertes Modell general-purpose LLMs schlägt

General-purpose Frontier-Modelle sind darauf trainiert, ein sehr breites Aufgabenspektrum abzudecken. Für die eng definierte Aufgabe „lies dieses Geschäftsdokument und extrahiere diese Felder" ist das ein Trainingsziel, das an der eigentlichen Aufgabe vorbeizielt. Vier Faktoren erklären, warum ein kleineres, spezialisiertes Modell hier in den Benchmarks vorne liegt:

Trainingsdaten-Fokus: Reale, teils unstrukturierte europäische Geschäftsdokumente statt allgemeinem Web-Text, inklusive der Eigenheiten von Rechnungen, Frachtpapieren und Verträgen aus regulierten Branchen.

Architektur für Layout-Verständnis: Bild und Text werden in einem Schritt verarbeitet, ohne verlustbehaftete OCR-Vorstufe, relevant bei Tabellen, mehrspaltigem Layout, Stempeln und handschriftlichen Vermerken.

Geringerer Rechenbedarf: Weniger aktive Parameter pro Anfrage senken Latenz und Infrastrukturkosten pro Dokument.

Nachtrainierbar auf eigene Dokumenttypen DEMNÄCHST VERFÜGBAR, kontinuierliches Feintuning auf unternehmensspezifische Vorlagen und Sonderfälle statt nur generischer Prompt-Anpassung.

Das lässt sich mit harten Zahlen belegen: Dokumentli® ist deutlich kompakter aufgebaut. Gemma 4 26B-A4B, eines der sechs Referenzmodelle, nutzt bei nahezu identischem aktivem Rechenaufwand (rund 4 Mrd. aktive Parameter je Anfrage) ein deutlich grösseres Gesamtmodell, schneidet im Benchmark aber klar schwächer ab, ein direkter Beleg dafür, dass Spezialisierung hier mehr zur Genauigkeit beiträgt als reine Modellgrösse.

+19,4 Pt.
höhere Ø Genauigkeit (cANLS@0,8) von Dokumentli® gegenüber Gemma 4 26B-A4B, bei nahezu identischem aktivem Rechenaufwand je Anfrage.

6×
kleinere Gesamt-Parameterzahl: Dokumentli® statt eines 26-Milliarden-Parameter-Modells wie Gemma 4 26B-A4B.

Generalist (general-purpose LLM)

Viele mögliche Aufgaben: Konversation, Code, Weltwissen, Dokumente. Dokumentenverständnis ist eine von vielen Fähigkeiten, nicht der Trainingsschwerpunkt.

Spezialist (Dokumentli®)

Eine Aufgabe: Geschäftsdokumente lesen, klassifizieren, strukturieren. Trainingsdaten und Architektur sind vollständig darauf ausgerichtet.

ABGRENZUNG

Was Dokumentli® nicht kann

Die Spezialisierung, die Dokumentli® bei Document-Intelligence-Aufgaben stark macht, ist zugleich eine bewusste Grenze: Dokumentli® ist kein general-purpose Assistent. Ausserhalb der Dokumentenverarbeitung ist ein general-purpose Modell die richtige Wahl. Wo die Aufgabe eng umrissen und hochvolumig ist, gilt dieselbe Logik: Ein spezialisiertes Modell schlägt in der Regel den Generalisten.

AUFGABE	DOKUMENTLI®	GENERAL-PURPOSE FRONTIER-MODELL
Dokumente klassifizieren & Felder extrahieren	✓ Kernaufgabe, dafür trainiert	✗ Fähig, aber nicht spezialisiert (siehe Benchmarks)
Freies Reasoning & offene Konversation	✗ Nicht vorgesehen	✓ Kernaufgabe
Programmierung / Code-Generierung	✗ Nicht vorgesehen	✓ Kernaufgabe
Allgemeine Wissensfragen ausserhalb des Dokumentinhalts	✗ Nicht vorgesehen	✓ Kernaufgabe
Freies kreatives Schreiben	✗ Nicht vorgesehen	✓ Kernaufgabe
Mehrstufige, dokumentfremde Agenten-Aufgaben	✗ Nicht vorgesehen	✓ Kernaufgabe

Dokumentli® einsetzen, wenn…

- hohe Volumen an Geschäftsdokumenten klassifiziert oder Felder daraus extrahiert werden müssen
- Datenhoheit, Air-Gapped-Betrieb oder planbare Kosten wichtig sind
- strukturierte, schemakonforme Ausgaben statt freier Text benötigt werden

General-purpose LLM einsetzen, wenn…

- die Aufgabe offenes Reasoning, Konversation oder Code betrifft
- allgemeines Weltwissen ausserhalb des Dokumentinhalts gefragt ist
- es um freies kreatives Schreiben oder dokumentfremde Agenten-Aufgaben geht

Komplementär, nicht konkurrierend: Unternehmen setzen Dokumentli® gezielt für die Dokumenteneingangsverarbeitung ein und kombinieren es bei Bedarf mit general-purpose Modellen für nachgelagerte, dokumentfremde Schritte. Beides lässt sich über die eigene Workflow-Lösung orchestrieren (siehe Seite 7 und 9).

EINSATZOPTIONEN

Drei Wege, dieselbe Dokumentli® API zu betreiben

On-Prem, Private Cloud und die Dokumentli® Cloud API sind drei Varianten **derselben API**, sie unterscheiden sich nur darin, wer den Dokumentli® Node betreibt und wo. Wer heute gegen die Dokumentli® Cloud API integriert, kann später verlustfrei auf On-Prem oder Private Cloud wechseln, ohne die eigene Integration anzupassen, kein Lock-in. Die vierte Einsatzoption, die Parashift Plattform Integration mit einer eigenen, umfassenderen API, ist auf der nächsten Seite beschrieben.

Das Modell und die Node-Komponente: Dokumentli® ist das KI-Modell selbst. Für den einfachen Betrieb auf eigener oder gemieteter Hardware liefert Parashift zusätzlich den **Dokumentli® Node**, eine leichtgewichtige Komponente, die das Modell verpackt und die Einrichtung in kurzer Zeit ermöglicht, ohne dass ein eigenes Deployment-Setup gebaut werden muss. In allen drei Optionen unten läuft derselbe Node hinter derselben API, nur der Betreiber und Standort ändern sich.

On-Prem

MAXIMALE DATENHOHEIT

- **Node betrieben von:** Ihnen, im eigenen Rechenzentrum
- **Hardware:** 1 Consumer-GPU (ab 24 GB VRAM)
- **Typisch für:** höchste Compliance-Anforderungen, air-gapped

Private Cloud

DEDIZIERTER TENANT

- **Node betrieben von:** isolierte Cloud-Instanz, Region/Zone Ihrer Wahl
- **Hardware:** vom Cloud-Anbieter bereitgestellt, nicht von Parashift
- **Typisch für:** regulierte Unternehmen mit eigener Cloud-Strategie

Dokumentli® Cloud API

VERWALTETER ZUGRIFF DEMNÄCHST

DIREKTER MODELL-CLOUD-API-ZUGRIFF

- **Node betrieben von:** vollständig durch Parashift verwaltet
- **Hardware:** keine
- **Typisch für:** Teams, die Dokumentli® direkt in eine eigene Applikation einbetten

Identische API in allen drei Optionen: einmal integrieren, jederzeit die Betriebsart wechseln, von der Cloud API zu On-Prem oder zurück, ohne die eigene Integration anzupassen.

SEHR WICHTIG

Offen für jede Agentic-Automatisierungs- und Workflow-Lösung

On-Prem, Private Cloud und die Dokumentli® Cloud API lassen sich mit praktisch jeder Dokumenten- oder Workflow-Automatisierungslösung am Markt verbinden, ob RPA, iPaaS, ERP-, DMS- oder eine selbst entwickelte Applikation. Voraussetzung ist lediglich, dass die jeweilige Lösung einen API-Aufruf ausführen kann; eine Bindung an eine bestimmte Plattform ist nicht erforderlich. Das ist zugleich genau der Punkt, an dem sich diese drei Optionen vom Lock-in einer geschlossenen Plattform unterscheiden.

EINSATZOPTIONEN

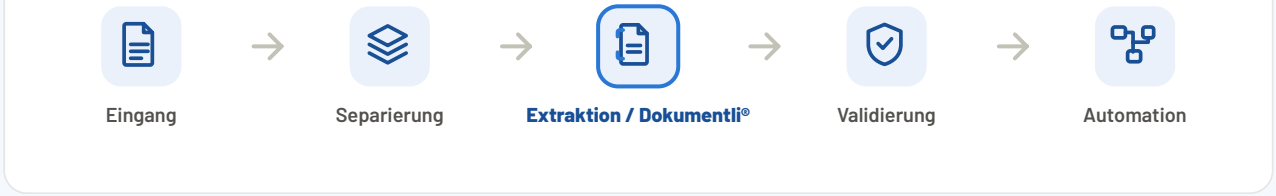
Die vierte Option: Parashift Plattform Integration

Die Parashift Plattform Integration ist ein eigenständiger, vierter Weg, Dokumentli® zu nutzen, mit einer **eigenen, umfassenderen API** als die einheitliche Dokumentli® Cloud API der drei Optionen auf der vorherigen Seite. Statt nur Dokumente zu klassifizieren und Felder zu extrahieren, deckt sie die gesamte Dokumenten-Pipeline ab, von der Ingestion bis zur Automatisierung.

Parashift Plattform Integration

DEMNÄCHST VERFÜGBAR

DOCUMENT INTELLIGENCE AI GUARDRAIL PLATFORM



- **Betriebsort:** Parashifts europäische Sovereign-Cloud-Plattform
- **Datenhoheit:** hoch, Compliance-Zonen EU, Schweiz, Deutschland
- **Zusatzfunktionen:** Ingestion, Separation, Validierung, Automatisierung, 30+ Konnektoren
- **Typisch für:** Teams, die die gesamte Dokumenten-Pipeline nutzen möchten, statt nur den Extraktionsschritt

Kostenkontrolle

Statt auf geteilter Parashift-Rechenkapazität zu laufen, lässt sich die Plattform auch gegen den eigenen **Dokumentli® Node** betreiben, volle Kontrolle über die Infrastrukturkosten.

Performance-Kontrolle

Dedizierte Rechenkapazität statt geteilter Infrastruktur: keine „Noisy Neighbors", die sich Compute mit anderen Mandanten teilen.

Anders als die Cloud API: die Plattform-Integration nutzt eine eigene, umfassendere API mit Pipeline-Funktionen (Validierungs-UI, Routing, Audit-Trail), nicht die einheitliche Dokumentli® Cloud API der drei Optionen auf Seite 7.

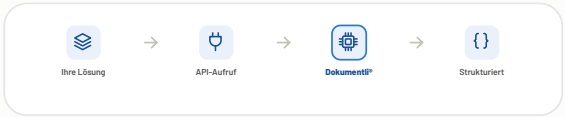
NUTZUNG & INTEGRATION

Passt sich in Ihren bestehenden Stack und in die Parashift Plattform ein

Dokumentli® ist bewusst so gebaut, dass es sich unabhängig von der gewählten Einsatzoption in praktisch jede bestehende Dokumenten- oder Workflow-Automatisierungslandschaft einfügt. Jede Lösung, die einen API-Aufruf ausführen kann, kann Dokumentli® nutzen (siehe unten). Zusätzlich lässt sich Dokumentli® auf zwei grundsätzliche Arten einbinden.

DIREKTE EINBINDUNG

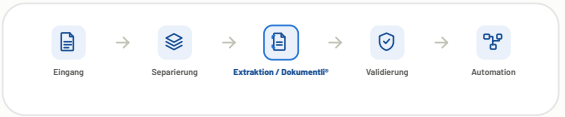
Programmatisch, in jede bestehende Lösung



- **Volle Kontrolle:** Dokumentli® als Baustein in Ihrem bestehenden Stack oder einer eigenen Applikation.
- **Eigenes Schema:** die Anfrage gibt das gewünschte Datenformat vor.
- **Anbieterneutral:** keine Bindung an eine bestimmte Plattform, nur ein API-Aufruf nötig.

OPTIONAL: WORKFLOW-LÖSUNG

Zusätzlich orchestriert über die Parashift Plattform



- **Für alle, die keine eigene Orchestrierung betreiben möchten:** eingebettete Validierungs-Oberfläche, Human-in-the-Loop ohne eigene Frontend-Entwicklung.
- **30+ Konnektoren** zu gängigen Business-Systemen, ohne eigene Middleware.
- **Governance inklusive:** Konfidenzwerte, Routing-Schwellen, Audit-Trail, 2-/3-Way-Match.

In der Praxis ist die Wahl keine Entweder-Oder-Entscheidung: bestehende Dokumenten- und Workflow-Automatisierungslösungen binden Dokumentli® direkt über die API ein (linkes Diagramm), während Teams ohne eigene Orchestrierung zusätzlich die vorgefertigte Parashift Plattform nutzen können, deren Extraktionsschritt intern auf demselben Dokumentli® läuft (rechtes Diagramm). Beide Muster lassen sich auf derselben Dokumentli®-Instanz kombinieren.

FUNKTIONIERT MIT

RPA · iPaaS · ERP · DMS · CRM · E-Mail-Postfächer · Eigene Applikationen

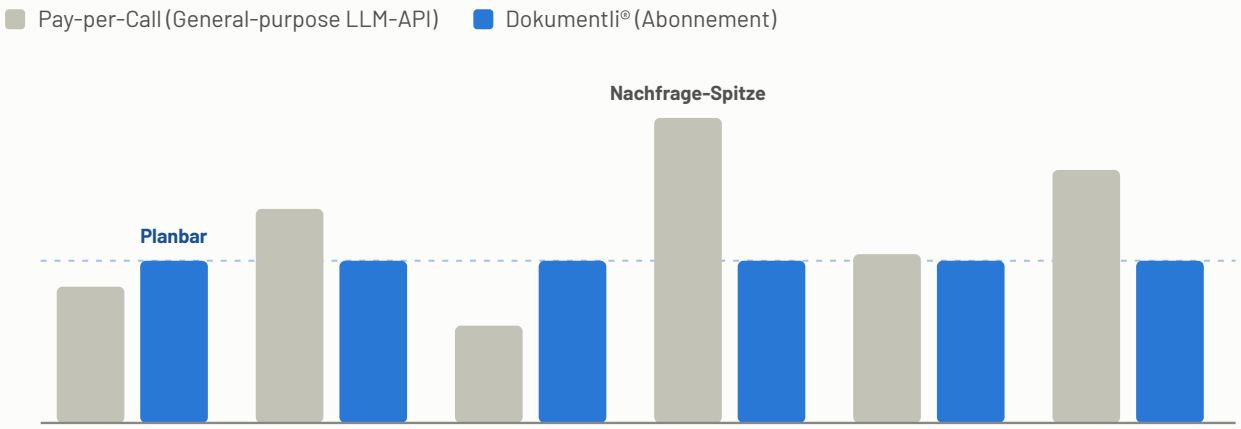
KOSTENLOGIK

Warum Dokumentli® günstiger ist als general-purpose LLM-APIs

Cloud-Anbieter berechnen in der Regel nach verarbeiteten Tokens. Für Unternehmen mit hohem und schwankendem Dokumentenvolumen bedeutet das: Die Kosten skalieren direkt und ohne planbare Obergrenze mit dem Volumen, inklusive Nachfrage-Spitzen.

Kostenverlauf über die Zeit, schematische Darstellung (ohne konkrete Beträge)

Illustrativ: Pay-per-Call-Kosten schwanken mit Volumen und Nachfrage-Spitzen; ein Seitenkontingent-Abonnement bleibt konstant und planbar.



Pay-per-Call
General-purpose LLM-APIs: Kosten steigen direkt und ohne Obergrenze mit Volumen und Tokenverbrauch.

Kapazitätsbasiert
Dokumentli®: Abonnement mit Seitenkontingent, im Voraus planbar, unabhängig von Nachfrage-Spitzen.

Abonnement-Stufen: Dokumentli® wird in mehreren Abonnement-Stufen angeboten, gestaffelt nach monatlichem Dokumentenvolumen, von Pilotprojekten bis zum Hochvolumen-Einsatz im Unternehmen. Jede Stufe umfasst **4 Basismodell-Updates pro Jahr**, sodass Genauigkeits- und Funktionsverbesserungen automatisch einfliessen, ohne neuen Vertrag.

FAZIT & QUELLEN

Ein spezialisiertes Modell für eine klar definierte Aufgabe

FAZIT

Dokumentli® erreicht die höchste durchschnittliche Extraktionsgenauigkeit unter sieben getesteten Modellen und ist zugleich im Durchschnitt das schnellste Modell im Feld, bei geringerem Rechenaufwand als general-purpose Frontier-Modelle für vergleichbare Aufgaben.

- **Genauer** als general-purpose LLMs
- **Schneller** als general-purpose LLMs
- **Günstiger** im Betrieb
- **Air-Gapped** auf einer GPU

Ausserhalb der Dokumentenverarbeitung bleibt ein general-purpose Frontier-Modell die richtige Wahl. Dokumentli® ist bewusst auf Document Intelligence spezialisiert und dafür das Modell der Wahl. Wer das Modell an eigenen Dokumenten testen möchte, kann dies unter **playground.parashift.io** tun.

Quellen & Methodik: Die Benchmark-Kennzahlen auf Seite 4 stammen aus dem **Parashift Dokumentli® Benchmark Report**: 5 Benchmarks, 3368 Dokumente, sieben getestete Modelle, Gütemaß cANLS (case-insensitive Average Normalized Levenshtein Similarity, Schwellenwert $\tau = 0{,}8$) als Makro-Mittelwert. Die vollständige Methodik, Konfidenzintervalle und Detailergebnisse je Branche und Modell sind in jenem Report dokumentiert. Angaben zu Architektur, Trainingsdatenfokus und Einsatzoptionen basieren auf der Produktdokumentation von Parashift (dokumentli-docs.parashift.io) sowie internen Produktunterlagen. Der Kostenvergleich auf Seite 10 beschreibt die Abrechnungslogik qualitativ, ohne konkrete Tarifbeträge zu nennen. Dieses Whitepaper (v1.7) ist eine Momentaufnahme mit Stand 27.08.2026.

Benchmark Report lesen

Vollständige Ergebnisse je Branche und Modell, Konfidenzintervalle und Methodik.

Modell testen

Eigene Dokumente ausprobieren unter **playground.parashift.io**.

Dokumentation

Technische Details unter **dokumentli-docs.parashift.io**.