parashift

Sovereign European Document AI

WHITEPAPER

Dokumentli® – The Specialized Vision Language Model for Document Intelligence

Dokumentli® achieves higher accuracy and speed than general-purpose frontier models at document processing, at predictable cost, and with the option to run fully air-gapped on a single consumer-grade GPU.

88%

Avg. accuracy (cANLS@0.8), highest value among 7 models tested

2.0×

faster on average than the most accurate reference model

4

Deployment options, from on-prem to API

Air-gapped capable: fully on-prem operation on a single consumer GPU, no data center, no cloud connection required.

EXECUTIVE SUMMARY

A model built to beat general-purpose LLMs at one job

Dokumentli® is Parashift's specialized vision-language flagship AI model for the automated processing of business documents. Unlike general-purpose frontier models, Dokumentli® is trained for one purpose: turning the business documents of European companies and regulated industries into structured data.

In tests across 5 benchmark datasets, Dokumentli® achieves the highest average extraction accuracy of seven models tested and is, at the same time, the fastest model on average in the field. This whitepaper puts these results in context, explains how the model works, honestly scopes its intended use, and describes the available deployment options.

The full benchmark figures, including detailed results by industry and model, are documented in the separate **Parashift Dokumentli® Benchmark Report**; this whitepaper summarizes the headline numbers and points to that report for the detail.

88%
Avg. accuracy (cANLS@0.8)
Highest value among 7 models tested

2.0×
faster on average
than the most accurate of the six reference models

4
Deployment options
On-prem, private cloud, Dokumentli® Cloud API, or Parashift Platform

Predictable
Cost model
Subscription instead of unpredictable pay-per-call billing

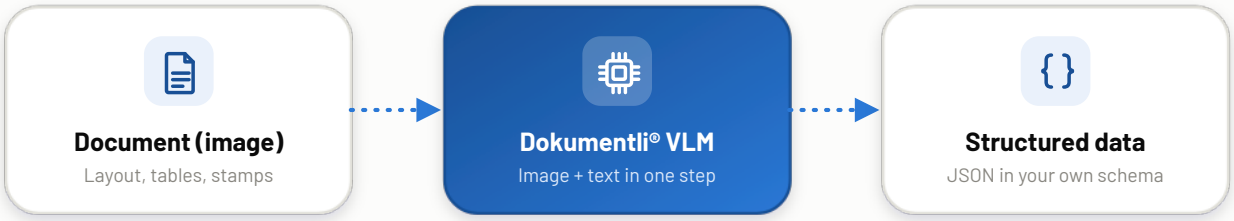
Who this model is built for: European companies and regulated industries with high, recurring document volumes, from financial services and insurance to construction, trade, and logistics.

What this whitepaper covers: what a small special-purpose vision language model is · benchmark headline numbers (accuracy & speed) · why specialization wins on document intelligence tasks · an honest scoping of what the model is not built for · the four deployment options · typical integration patterns · cost logic compared to general-purpose LLM APIs.

FUNDAMENTALS

What is a small special-purpose vision language model?

How Dokumentli® processes a document



A **vision language model (VLM)** processes image and text together. It reads a document page directly as an image, including layout, tables, stamps, and handwriting, and answers a text query about it. A separate OCR pre-processing step is not required for this.

"Special-purpose" means: Dokumentli® is trained specifically for document understanding of European business documents, not for the broadest possible range of tasks.

"Small" means: Dokumentli® is a compact model with significantly fewer parameters than general-purpose frontier models. That reduces the compute cost per document and allows it to run on off-the-shelf PC hardware with a single consumer GPU, all the way to fully air-gapped operation (see page 7).

Image + text in one model: layout, tables, and text content are interpreted together, without a separate OCR pipeline.

Specialized on one domain: trained on real business documents rather than general web text.

Compact: fewer parameters and lower compute and infrastructure needs than general-purpose frontier models.

6×
smaller parameter count for Dokumentli® vs. Gemma 4 26B-A4B, one of the six reference models (see page 5).

1 GPU
(from 24 GB VRAM) is enough for fully air-gapped operation without a data center or cloud connection.

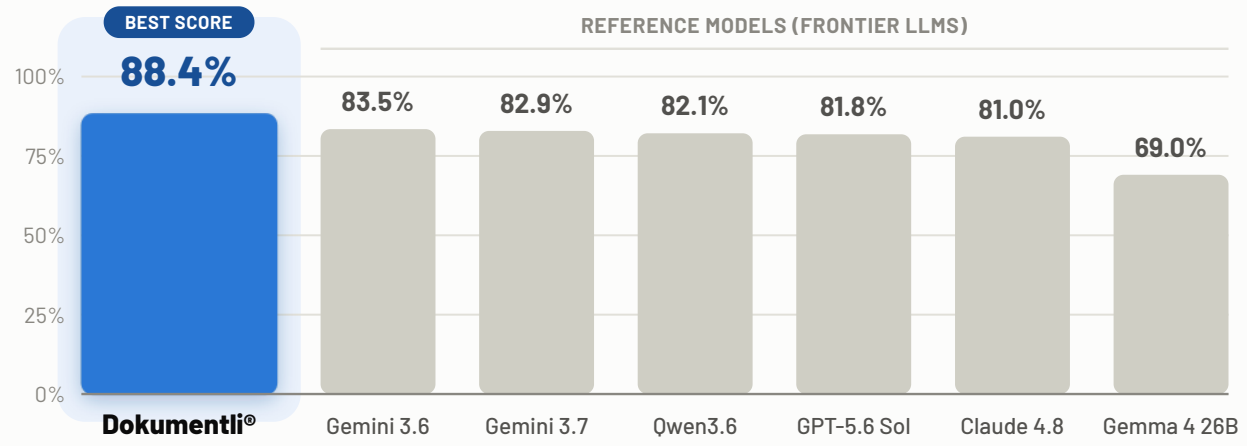
BENCHMARK RESULTS AT A GLANCE

Highest accuracy at the highest speed

In tests across 5 benchmark datasets totaling 3368 real business documents, Dokumentli® was tested against six current reference models: Claude Opus 4.8, GPT-5.6 Sol, Gemini 3.6 Flash, Gemini 3.7 Flash, Qwen3.6-Plus, and the open model Gemma 4 26B-A4B-IT. The quality metric is cANLS (case-insensitive Average Normalized Levenshtein Similarity, threshold τ = 0.8), a common variant of the ANLS industry standard for evaluating document extraction.

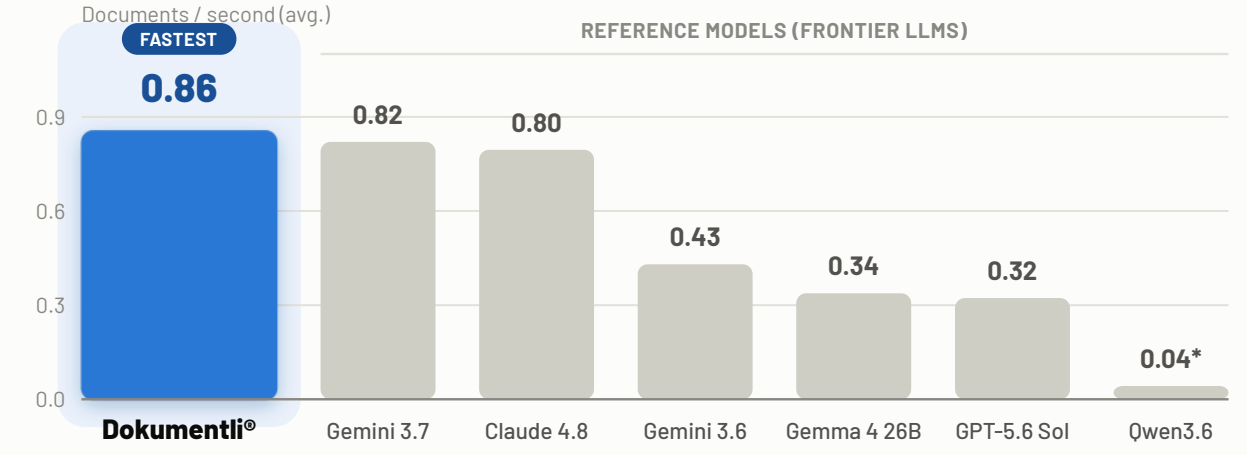
Avg. accuracy (cANLS@0.8) by model, averaged across 5 benchmarks

MORE ACCURACY



Avg. processing speed by model, documents per second

MORE SPEED



Methodology: Dokumentli® measured on an RTX PRO 6000 (Cloud-Run); reference models measured via OpenRouter. *The Qwen3.6 figure is notably low, possibly due to increased platform load during measurement; see the Benchmark Report.

5/5
Benchmarks where Dokumentli® leads Gemini 3.6 Flash, the strongest of the six reference models.

7
models tested in total, Dokumentli® plus six general-purpose frontier models as reference.

For details, see the Benchmark Report: results by industry and model, confidence intervals, positioning matrices, and the full methodology are documented in the separate **Parashift Dokumentli® Benchmark Report**.

CONTEXT

Why a specialized model beats general-purpose LLMs

General-purpose frontier models are trained to cover a very broad range of tasks. For the narrowly defined task of "read this business document and extract these fields", that is a training objective that misses the actual task. Four factors explain why a smaller, specialized model comes out ahead in the benchmarks:

Training-data focus: real, partly unstructured European business documents rather than general web text, including the specifics of invoices, freight papers, and contracts from regulated industries.

Architecture built for layout understanding: image and text are processed in a single pass, without a lossy OCR pre-stage, relevant for tables, multi-column layouts, stamps, and handwritten notes.

Lower compute requirement: fewer active parameters per request reduce latency and infrastructure cost per document.

Fine-tunable on your own document types COMING SOON, continuous fine-tuning on company-specific templates and edge cases, rather than relying only on generic prompt adjustments.

This can be shown with hard numbers: Dokumentli® is built to be significantly more compact. Gemma 4 26B-A4B, one of the six reference models, uses a much larger overall model at nearly identical active compute (roughly 4B active parameters per request), yet scores clearly lower in the benchmark, direct evidence that specialization contributes more to accuracy here than raw model size.

+19.4 pts
higher avg. accuracy (cANLS@0.8) of Dokumentli® over Gemma 4 26B-A4B, at nearly identical active compute per request.

6×
smaller total parameter count: Dokumentli® vs. a 26-billion-parameter model like Gemma 4 26B-A4B.

Generalist (general-purpose LLM)

Many possible tasks: conversation, code, world knowledge, documents. Document understanding is one of many capabilities, not the training focus.

Specialist (Dokumentli®)

One task: read, classify, and structure business documents. Training data and architecture are entirely built around it.

BOUNDARIES

What Dokumentli® is not built for

The specialization that makes Dokumentli® strong at document intelligence tasks is, at the same time, a deliberate boundary: Dokumentli® is not a general-purpose assistant. Outside document processing a general-purpose model is the right choice. Where the task is narrow and high-volume, the same logic applies: a specialized model will usually beat the generalist.

TASK	DOKUMENTLI®	GENERAL-PURPOSE FRONTIER MODEL
Classifying documents & extracting fields	✓ Core task, trained for this	⊗ Capable, but not specialized (see benchmarks)
Open-ended reasoning & conversation	⊗ Not intended	✓ Core task
Coding / code generation	⊗ Not intended	✓ Core task
General knowledge questions beyond the document content	⊗ Not intended	✓ Core task
Open-ended creative writing	⊗ Not intended	✓ Core task
Multi-step, non-document agentic tasks	⊗ Not intended	✓ Core task

Use Dokumentli® when...

- high volumes of business documents need to be classified or have fields extracted from them
- data sovereignty, air-gapped operation, or predictable cost matter
- structured, schema-conformant output is needed rather than free text

Use a general-purpose LLM when...

- the task involves open-ended reasoning, conversation, or code
- general world knowledge beyond the document content is required
- it's about open-ended creative writing or non-document agentic tasks

Complementary, not competing: companies deploy Dokumentli® specifically for inbound document processing and combine it, where needed, with general-purpose models for downstream, non-document steps. Both patterns can be orchestrated via your own workflow solution (see pages 7 and 9).

DEPLOYMENT OPTIONS

Three Ways to Run the Same Dokumentli® API

On-prem, private cloud, and the Dokumentli® Cloud API are three variants of **the same API**, they differ only in who runs the Dokumentli® Node and where. Teams integrating against the Dokumentli® Cloud API today can later move to on-prem or private cloud without changing their integration, no lock-in. The fourth deployment option, Parashift Platform Integration with its own, more comprehensive API, is described on the next page.

The model and the Node component: Dokumentli® is the AI model itself. For easy setup on your own or rented hardware, Parashift additionally ships the **Dokumentli® Node**, a lightweight component that packages the model and lets you get it running quickly, without building your own deployment setup. In all three options below, the same Node runs behind the same API, only who operates it and where changes.

On-Prem

MAXIMUM DATA SOVEREIGNTY

- **Node run by:** you, in your own data center
- **Hardware:** 1 consumer GPU (from 24 GB VRAM)
- **Typically suited for:** highest compliance requirements, air-gapped

Private Cloud

DEDICATED TENANT

- **Node run by:** an isolated cloud instance, region/zone of your choice
- **Hardware:** provided by the cloud provider, not by Parashift
- **Typically suited for:** regulated companies with their own cloud strategy

Dokumentli® Cloud API

MANAGED ACCESS COMING SOON

DIRECT MODEL CLOUD API ACCESS

- **Node run by:** fully managed by Parashift
- **Hardware:** none
- **Typically suited for:** teams embedding Dokumentli® directly into their own application

Identical API across all three options: integrate once, switch operating mode anytime, from the Cloud API to on-prem or back, without changing your own integration.

VERY IMPORTANT

Open for any Agentic Automation and Workflow Solution

On-prem, private cloud, and the Dokumentli® Cloud API all connect to virtually any document or workflow automation product on the market, whether RPA, iPaaS, ERP, DMS, or a self-built application. The only requirement is that the solution can make an API call; no lock-in to a specific platform is required. This is exactly where these three options differ from the lock-in of a closed platform.

DEPLOYMENT OPTIONS

The Fourth Option: Parashift Platform Integration

The Parashift Platform Integration is a distinct, fourth way to use Dokumentli®, with an **own, more comprehensive API** than the unified Dokumentli® Cloud API of the three options on the previous page. Rather than only classifying documents and extracting fields, it covers the full document pipeline, from ingestion to automation.

Parashift Platform Integration

COMING SOON

DOCUMENT INTELLIGENCE AI GUARDRAIL PLATFORM

- **Operated:** Parashift's European sovereign cloud platform
- **Data sovereignty:** high, compliance zones EU, Switzerland, Germany
- **Additional features:** ingestion, separation, validation, automation, 30+ connectors
- **Typically suited for:** teams who want to use the full document pipeline, not just the extraction step

Cost control

Rather than running on shared Parashift compute capacity, the platform can also be operated against your own **Dokumentli® Node**, giving you full control over infrastructure cost.

Performance control

Dedicated compute instead of shared infrastructure: no "noisy neighbors" sharing compute with other tenants.

Different from the Cloud API: the Platform Integration uses its own, more comprehensive API with pipeline features (validation UI, routing, audit trail), not the unified Dokumentli® Cloud API of the three options on page 7.

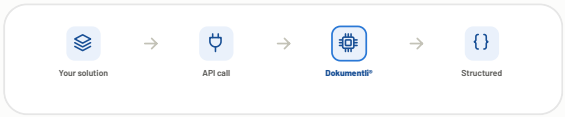
USAGE & INTEGRATION

Fits into your existing stack and into the Parashift Platform

Dokumentli® is deliberately built so that, regardless of the chosen deployment option, it fits into virtually any existing document or workflow automation landscape. Any solution that can make an API call can use Dokumentli® (see below). On top of that, Dokumentli® can be integrated in two fundamental ways.

DIRECT INTEGRATION

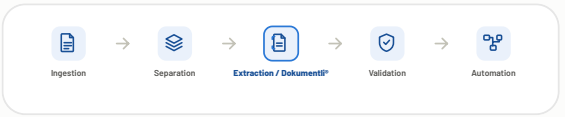
Programmatic, into any existing solution



- **Full control:** Dokumentli® as a building block in your existing stack or your own application.
- **Your own schema:** the request specifies the desired data format.
- **Vendor-neutral:** no lock-in to a specific platform, just an API call away.

OPTIONAL: WORKFLOW SOLUTION

Additionally orchestrated via the Parashift Platform



- **For anyone without their own orchestration:** embedded validation UI, human-in-the-loop review without building your own front end.
- **30+ connectors** to common business systems, without building your own middleware.
- **Governance included:** confidence scores, routing thresholds, audit trail, 2-/3-way matching.

In practice, this is not an either-or choice: existing document and workflow automation solutions integrate Dokumentli® directly via the API (left diagram), while teams without their own orchestration can additionally use the pre-built Parashift Platform, whose extraction step internally runs on the same Dokumentli® (right diagram). Both patterns can be combined on the same Dokumentli® instance.

WORKS WITH

RPA · iPaaS · ERP · DMS · CRM · Email inboxes · Custom applications

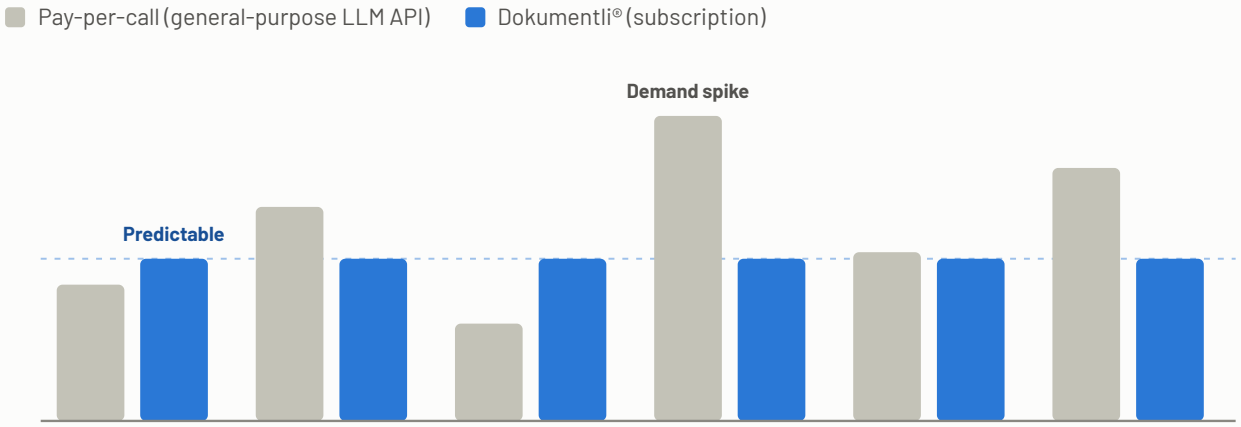
COST LOGIC

Why Dokumentli® is cheaper than general-purpose LLM APIs

Cloud providers generally charge per tokens processed. For companies with high and fluctuating document volumes, that means costs scale directly with volume, with no predictable ceiling, including during demand spikes.

Cost over time, schematic illustration (no specific amounts)

Illustrative: pay-per-call cost swings with volume and demand spikes; a page-allowance subscription stays constant and predictable.



Pay-per-call

General-purpose LLM APIs: costs rise directly and without a ceiling as volume and token consumption grow.

Capacity-based

Dokumentli®: a subscription with a page allowance, predictable in advance, independent of demand spikes.

Subscription tiers: Dokumentli® is available in multiple subscription tiers, sized by monthly document volume, from pilot deployments to high-volume enterprise use. Every tier includes **4 base model updates per year**, so accuracy and capability improvements ship automatically, without a new contract.

parashift

A specialized model for a clearly defined task

CONCLUSION

Dokumentli® achieves the highest average extraction accuracy among seven models tested, and is at the same time the fastest model on average in the field, at lower compute cost than general-purpose frontier models require for comparable tasks.

- **More accurate** than general-purpose LLMs
- **Faster** than general-purpose LLMs
- **Cheaper** to operate
- **Air-gapped** on a single GPU

Outside document processing, a general-purpose frontier model remains the right choice. Dokumentli® is deliberately specialized on document intelligence, and is the model of choice for that task. Anyone who wants to test the model on their own documents can do so at **playground.parashift.io**.

Sources & methodology: The benchmark headline numbers on page 4 are taken from the **Parashift Dokumentli® Benchmark Report**: 5 benchmarks, 3368 documents, seven models tested, quality metric cANLS (case-insensitive Average Normalized Levenshtein Similarity, threshold $\tau = 0.8$), a common variant of the ANLS industry standard, as a macro-average. The full methodology, confidence intervals, and detailed results by industry and model are documented in that report. Statements on architecture, training-data focus, and deployment options are based on Parashift's product documentation (dokumentli-docs.parashift.io) and internal product materials. The cost comparison on page 10 describes the billing logic qualitatively, without stating specific tariff amounts. This whitepaper (v1.7) is a snapshot as of 2026-08-27.

Read the Benchmark Report

Full results by industry and model, confidence intervals, and methodology.

Test the model

Try it on your own documents at **playground.parashift.io**.

Documentation

Technical details at **dokumentli-docs.parashift.io**.