Build with AI

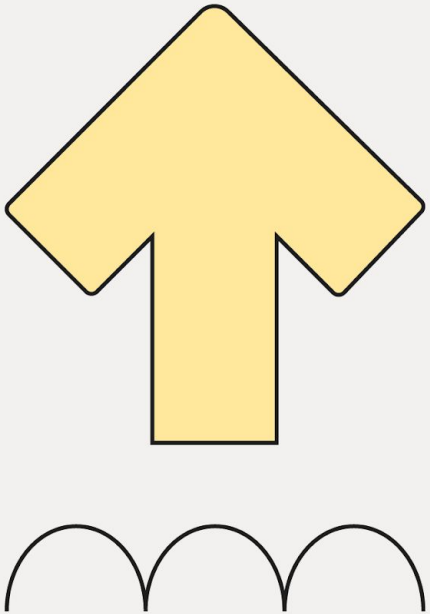
Beja

Massive OCR at the Edge Privacy-First Document Pipelines with Gemma 4 & ADK

Nuno João Andrade
Google Developer Expert

Google Developer Groups

01



Privacy First

The Problem: Sharing the documents out of the privacy of your application and context, may not be allowed.

Google Developer Groups

A new approach

The Legacy Approach

Scaling document processing via cloud-based APIs introduces severe bottlenecks:

- **Latency:** Continuous round-trip network delays.
- **Cost:** High per-API invocation pricing at scale.
- **Privacy:** Sending sensitive corporate records to external servers creates compliance risks.

The Edge Solution

By leveraging Gemma 4's highly optimized, natively multimodal architecture alongside the **Agent Development Kit (ADK)**:

- Deploy sub-1GB memory footprint models (E2B).
- Extract text & structure data entirely offline.
- Ensure zero-latency, privacy-first processing on local hardware.

Google Developer Groups

What is Gemma 4 ?

The **Gemma** model family is a collection of lightweight, open-weight AI models developed by Google DeepMind. Built using the same research and technology as Google's flagship Gemini models, they are designed to bring high-level AI capabilities to local devices, personal computers, and edge environments.

Why local model ?

- **Absolute Data Privacy**
- **Predictable Costs & No API Fees**
- **Full Control & Customization**
- **Zero Latency & Offline Capabilities**
- **Independence & Consistency**

Google Developer Groups

The Evolution to Gemma 4

From text-only generation to unified native multimodality

Gemma 1

FEB 21, 2024

Initial release (2B & 7B). Open weights text models built from Gemini technology.

Gemma 2

JUN 27, 2024

Major architectural update featuring improved efficiency, sliding window attention, and logit soft-capping.

Gemma 3

MAR 10, 2025

Multimodal breakthrough. Models handle text, images, and tools natively.

Gemma 4

MAR 31, 2026

Unified native multimodality (E2B - 31B). Supports text, audio, image inputs with up to a 256K context window.

Google Developer Groups

The Gemma 4 family

Property	E2B	E4B	12B Unified	31B Dense
Total Parameters	2.3B effective *(5.1B with embeddings)*	4.5B effective *(8B with embeddings)*	11.95B	30.7B
Layers	35	42	48	60
Sliding Window	512 tokens	512 tokens	1024 tokens	1024 tokens
Context Length	128K tokens	128K tokens	256K tokens	256K tokens
Vocabulary Size	262K	262K	262K	262K
Supported Modalities	Text, Image, Audio	Text, Image, Audio	Text, Image, Audio	Text, Image
Vision Encoder Parameters	*~150M*	*~150M*	-	*~550M*
Audio Encoder Parameters	*~300M*	*~300M*	-	No Audio

Google Developer Groups

The Local Execution Stack

Powering massive OCR with robust inference and multi-agent orchestrat

Ollama

The standard for developer-friendly local inference. Runs Gemma 4 with a seamless, Docker-like CLI experience, perfect for rapid prototyping.

- Zero-config model pulling
- Built-in model registry
- OpenAI-compatible REST API

Production Engines

When moving from prototype to massive scale, specialized inference engines ensure you maximize your hardware utilization and throughput.

- **vLLM:** High-throughput GPU serving
- **Llama.cpp:** Bare-metal Edge efficiency
- **MLX:** Native Apple Silicon optimization

ADK Orchestration

The Agent Development Kit wraps raw local inference endpoints into complex, scalable, and structured multi-agent OCR workflows.

- Multi-agent routing logic
- Native tool & function calling
- Structured JSON data extraction

Google Developer Groups

Why ADK over Raw SDKs?

Moving from basic inference APIs to robust, tool-using OCR agents

</> Building with Raw APIs

Directly interfacing with model endpoints (like OpenAI-compatible SDKs) leaves the heavy lifting to the developer:

- **Manual Context Limits:** You must manually track token counts and truncate chat history as large OCR documents fill the window.
- **Fragile Tooling:** Relying on the model to output perfect JSON strings, requiring custom parser regex and manual retry loops.
- **Stateless Execution:** Every step requires a freshly assembled request, mixing HTTP infrastructure with business logic.

The ADK Advantage

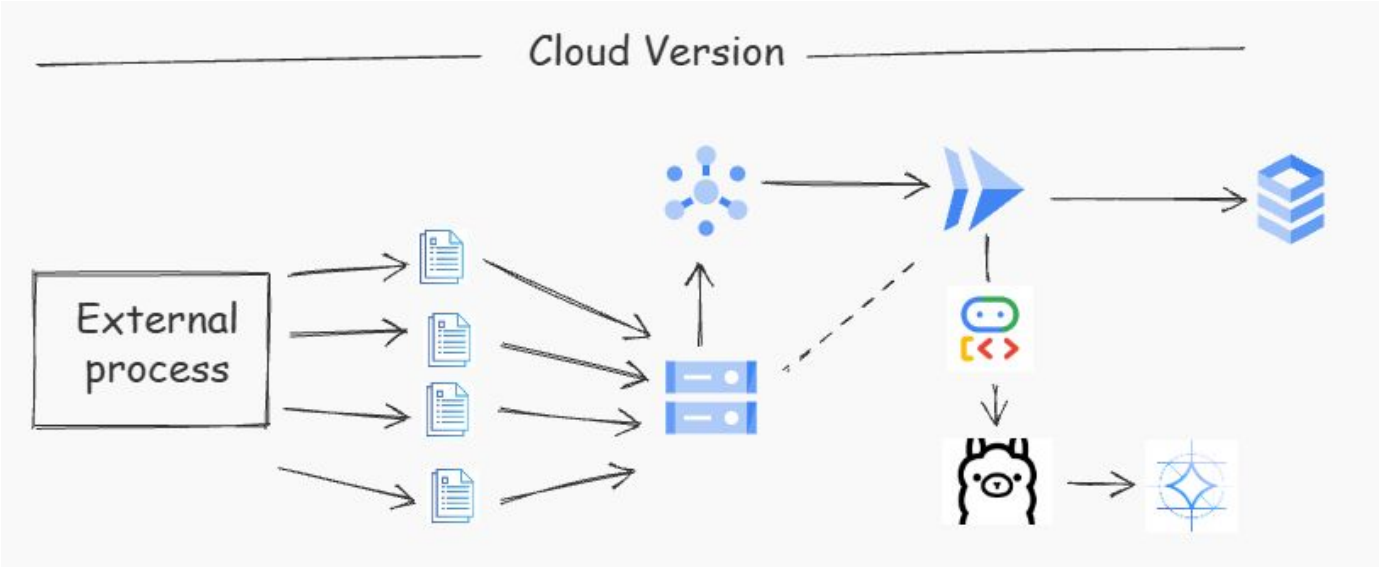
ADK abstracts the infrastructure, providing native primitives for multi-agent workflows out of the box:

- **Managed Memory:** Built-in systems automatically summarize or slide context windows as large document payloads grow.
- **Native Function Calling:** Define tools as standard code functions; ADK handles the schema generation, execution, and error recovery natively.
- **Agentic Routing:** Seamlessly pass state between specialized OCR agents (e.g., Vision Agent to Validation Agent).

Google Developer Groups

Cloud Solution

Example infrastructure on the cloud (basic infra in github)



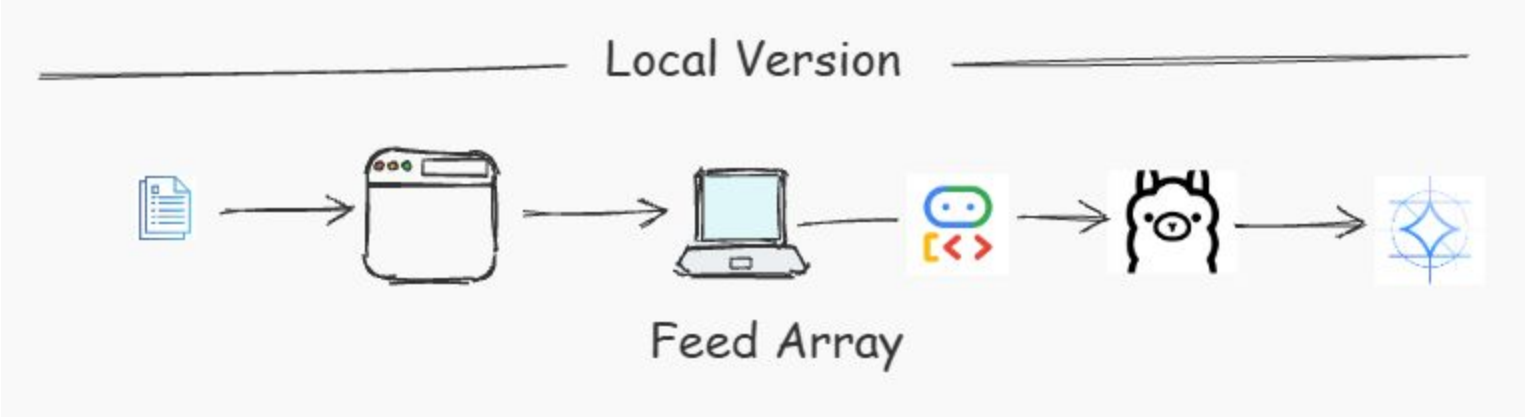
https://drawit.nja.dev/?gallery=ocr.gemma.series

For reference only, there are missing some security elements

Google Developer Groups

Local Solution

Example infrastructure on the local (full code present in github)

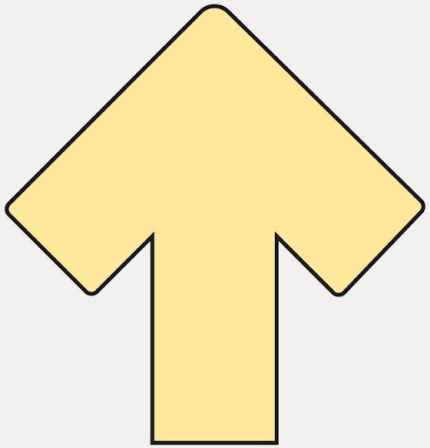


https://drawit.nja.dev/?gallery=ocr.gemma.series

For reference only, there are missing some security elements

Google Developer Groups

02



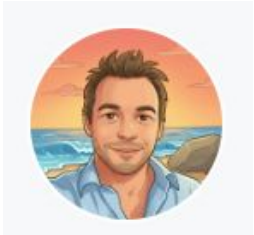
Demo

Google
Developer
Groups

Thank you.

LinkedIn

https://github.com/nuno-joao-andrade-dev



https://nja.dev

https://drawit.nja.dev

Google Developer Groups