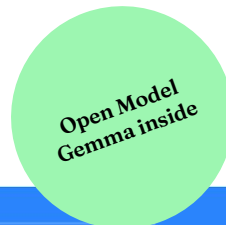




Serverless GPU ou comment déployer facilement son LLM



Julien Landuré | jlandure.dev
Founder & CTO @TechTown
Google Developer Expert Cloud



Introduction



What is an LLM, anyway?

A **Large Language Model** (or **LLMs**) is a type of artificial intelligence model that has been trained on a large dataset, often but not always comprised of human language.

Smaller LLMs are sometimes referred to as **Small Language Models** (or **SLMs**) for short.

Gemini: At The Frontier

Google's most advanced family of AI models, and one of its flagship products.

Runs on Google's servers - no need to worry about infrastructure yourself.

Accessible through products and platforms like **AI Studio**, the **Gemini** website, or simply as an API.



I want to have control over my LLM

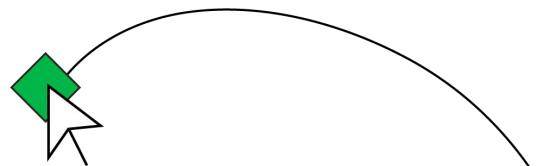
For privacy

To control where I can deploy it

To avoid sending data to an external server

To train on my own data

...Ultimately to build trust with my customers



Gemma: Google's Open Models

Gemma is based on the same research as Gemini, and because it's an open model, you can download it yourself and then deploy it wherever you want.

Gemma is Google's answer for **AI practitioners who want more control.**



Gemma Today



2024

Feb 21, 2024
Gemma
2B, 7B

Apr 5, 2024
Gemma 1.1
2B, 7B

Jun 27, 2024
Gemma 2
9B, 27B

Jul 31, 2024
Gemma 2
2B

Oct 3, 2024
Gemma 2 for Japan
2B

2025

Mar 12, 2025
Gemma 3
1B, 4B, 12B, 27B

core models

Gemma 3



Mar 12, 2025 • 1B, 4B, 12B, 27B

Gemma 3 is the most capable and advanced model of the Gemma family, adding multimodality (text and images), multilinguality (140 languages), a longer context window and more.

Gemma 3 will be available in four sizes (1B, 4B, 12B and 27B) in both pre-trained and instruction-tuned variants.



New in Gemma 3

Multimodality



Features an integrated vision decoder based on SigLIP, allowing Gemma 3 to take images and videos as inputs.

Input	Output
I need to get warm. What button turns up the heat? 	Based on the image, the button that likely turns up the heat is 暖房 (Danbou). "暖房" means "heating" in Japanese. It's the button you'd press to activate the heating function on the air conditioner/climate control system. The button with the plus sign (+) might adjust the temperature after you've selected the heating mode.

New in Gemma 3

Function Calling



Gemma 3's performance around structured inputs and outputs has also been a focus, allowing users to specify function signatures and have Gemma generate function calls.

This has been done in a way that maximizes flexibility - rather than require that you prompt it in a particular way, with specific tokens, it should “just work”.

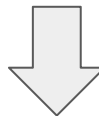
New in Gemma 3

Longer Context Window

Finally, the most asked-for Gemma improvement is here - an expanded context window.

The smallest Gemma 3 1B model has a 32k context window, and its larger siblings have a 128k token context window.

8k



32k

1B

128k

4B, 14B, 27B

New in Gemma 3



Gemma 3n

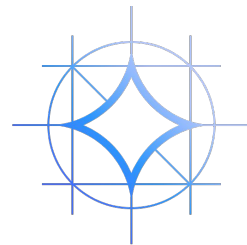
A powerful and efficient open model designed to run locally on phones, tablets, and laptops

- Optimized on-device performance
- Privacy-first, offline-ready
- Multimodal understanding (140 lang, audio, image)
- Dynamic resource usage

The Gemma3n has a 32k context window.

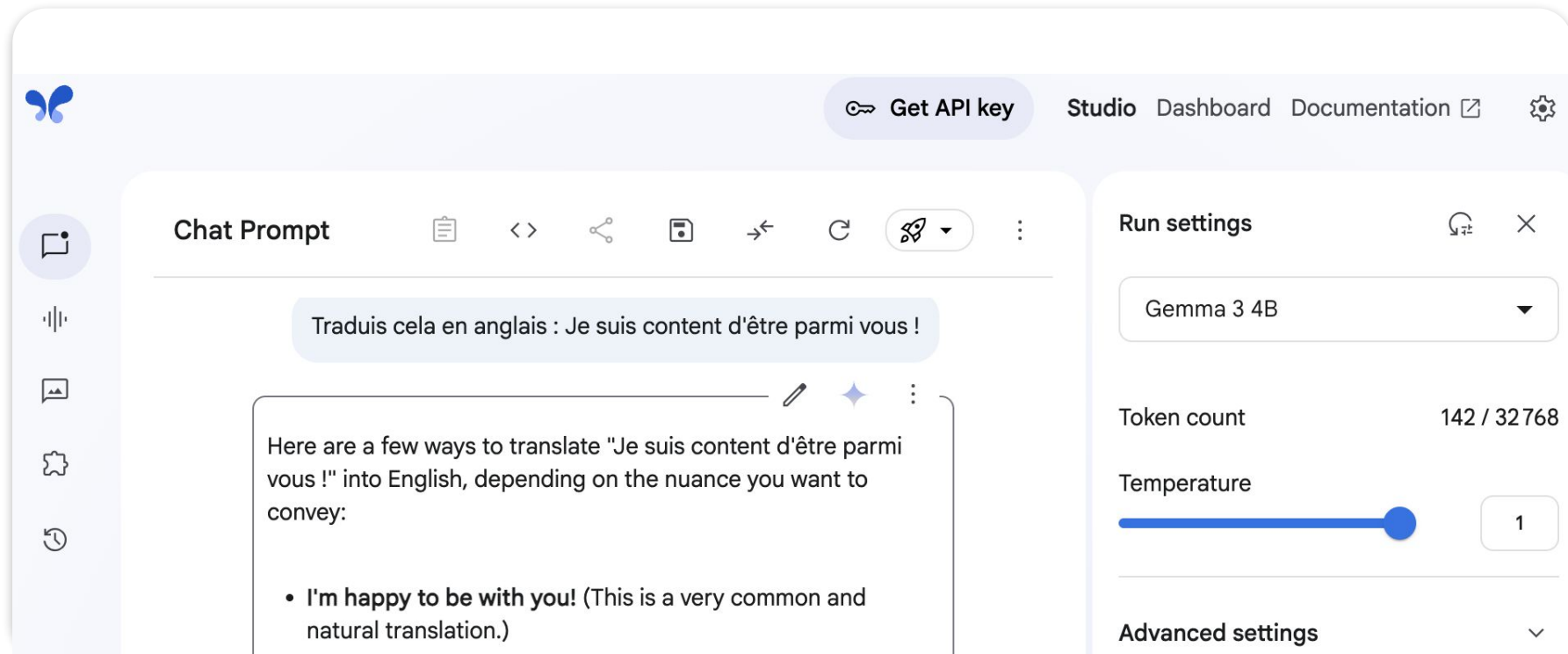
Run on a 2GB Ram system.

Source : <https://apxml.com/posts/gpu-system-requirements-guide-for-gemma-3n>



Gemma 3n

Quick start - Trying out Gemma



The screenshot displays the Google AI Studio interface. At the top, there's a navigation bar with the Google AI logo, a 'Get API key' button, and links to 'Studio', 'Dashboard', and 'Documentation'. Below this, the main area is divided into three sections: a 'Chat Prompt' area on the left, a 'Run settings' panel on the right, and a central output area.

Chat Prompt: The prompt is 'Traduis cela en anglais : Je suis content d'être parmi vous !'. The output shows a translation: 'Here are a few ways to translate "Je suis content d'être parmi vous !" into English, depending on the nuance you want to convey:'. A list of translations is provided, including 'I'm happy to be with you! (This is a very common and natural translation.)'.

Run settings: The model selected is 'Gemma 3 4B'. The token count is '142 / 32768'. The temperature is set to '1'.

Advanced settings: This section is currently collapsed.

Demo



Running AI Models



Running an AI model

This is generally referred to as **inference** - the process of **using a trained model to generate output**.

In response to input, the model will infer a response based on its learned knowledge and patterns.

Hardware

CPUs

General purpose processors that can be used for a wide range of tasks.

Not as efficient as GPUs or TPUs for machine learning tasks due to their architecture.

GPUs

Specialized processors that are designed for parallel processing, making them well-suited for the type of calculations needed for machine learning.

The most common type of accelerator used in machine learning.

Most ML practitioners are already familiar with GPUs.

TPUs

Specialized processors designed specifically for machine learning tasks, invented by Google. Even more efficient than GPUs for certain types of calculations.

Not as widely used as GPUs and may require more specialized knowledge to use effectively.

TPUs are a key differentiating factor for Google Cloud.

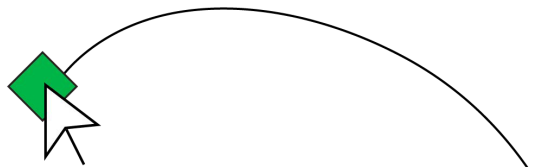
I don't want to manage infra

A lack of skills

A lack of time

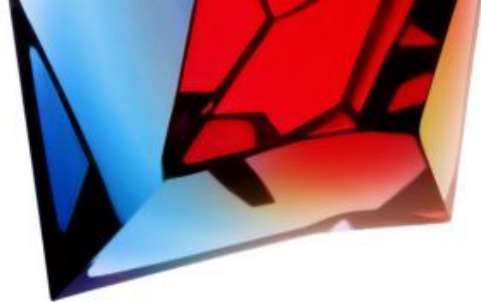
My desire to focus on developing my AI product

...Ultimately I want to build and run quickly for my customers

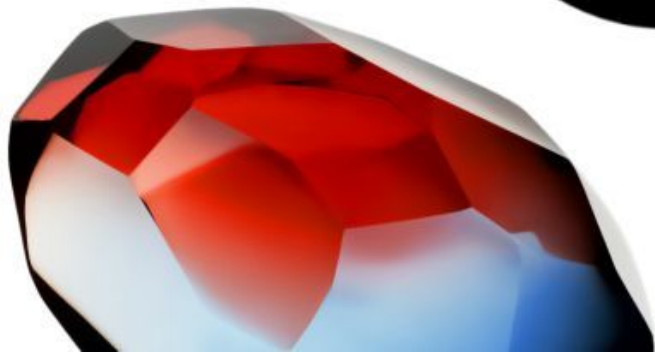




Cloud Run
GPUs



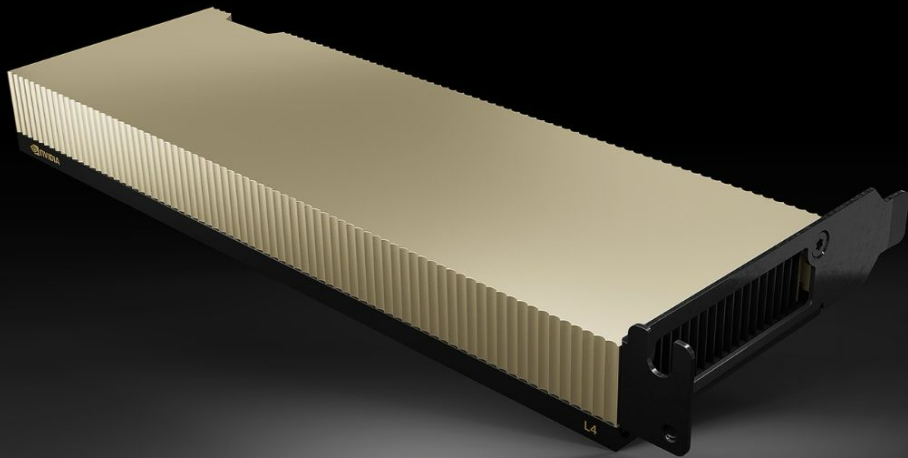
G A



NVIDIA L4 Tensor Core GPU

The breakthrough universal accelerator for
efficient video, AI, and graphics.

Learn More





From 0 to
an instance
with a GPU
and drivers
installed in...

5s

Google Cloud

Google Cloud

Serverless GPU



Pay-per-second billing

Scale to zero

Rapid startup and scaling (less than 5s)

Full streaming support

Support Gemma

Support any model



Limited to NVIDIA L4 GPUs

Up to 4 GPUs per instance

Min of 4 CPUs and 16Gib RAM 

Max instances depending of your GPU quotas

Resources

Memory

16 GiB

Memory to allocate to each instance of this container.

CPU

4

Number of vCPUs allocated to each instance of this container.



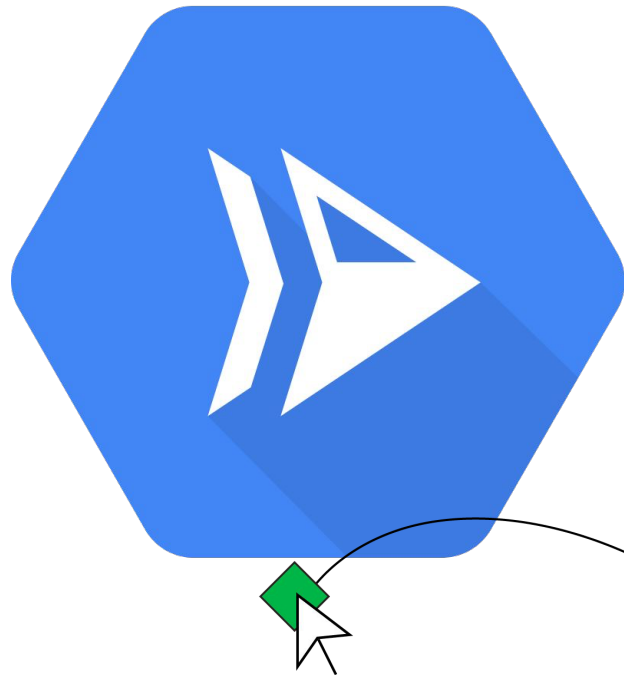
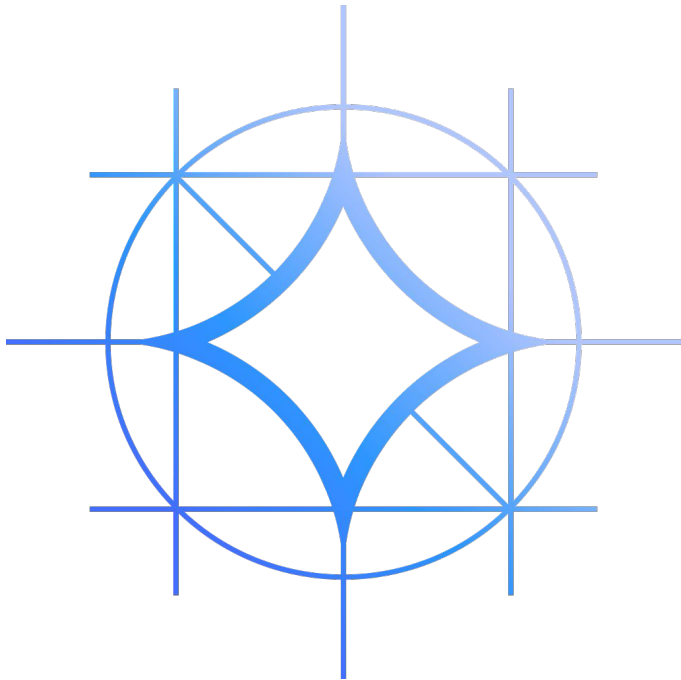
GPU

Attach GPUs to this container

Demo



I love it



Responsible AI



ShieldGemma

Supported Safety Checks

Harm Type	Text Content
Harassment	✓
Hate Speech	✓
Dangerous Content	✓
Sexually Explicit Content	✓





goo.gle/gemma

goo.gle/gemma-discord



Serverless GPU ou comment déployer facilement son LLM



Julien Landuré
Founder & CTO @TechTown
Google Developer Expert Cloud

