Gemma 3 stands out for its performance. Google claims it surpasses well-known models like DeepSeek V3 and OpenAI o3-mini, particularly in tasks requiring text, image, and short video analysis. Its “context window”—the amount of data the model can process at once—reaches 128,000 tokens, equivalent to a 200-page book.

The model is so efficient that its 27-billion-parameter version runs on a single GPU, such as the NVIDIA H100, rather than an entire server cluster.

One standout feature of Gemma 3 is its openness. Unlike closed models such as ChatGPT, Gemma’s code is accessible to everyone via platforms like Google AI Studio, Hugging Face, and Kaggle. The model also supports over 140 languages.