**Important things to remember about Gemma 2**

Gemma 2 is a high-performance, efficient, and open model designed for both developers and researchers, offering significant advancements over its predecessor. Here are the key highlights:

**1. Performance & Efficiency:**

- Available in 9 billion (9B) and 27 billion (27B) parameter sizes.

- The 27B model offers performance comparable to models more than twice its size.

- Designed to be efficient at inference, running on a single NVIDIA H100 GPU or TPU host, which reduces deployment costs.

- Optimized for fast inference across a range of hardware, including high-end desktops, gaming laptops, and cloud setups.

**2. Architecture & Cost Savings:**

- Gemma 2 is built on a redesigned architecture that balances performance with efficiency.

- The 27B model runs efficiently at full precision on powerful GPUs or TPUs, offering cost savings while maintaining high performance.

**3. Developer & Researcher-Friendly:**

- Available under a commercially-friendly license, making it accessible for research and commercialization.

- Broad compatibility with major AI frameworks like Hugging Face Transformers, JAX, PyTorch, and TensorFlow, with integrations into tools like Keras, vLLM, and Gemma.cpp.

- Deployment is easy through platforms like Google Cloud Vertex AI and Hugging Face.

**4. Integration & Deployment:**

- Gemma 2 is optimized for NVIDIA TensorRT-LLM and is being fine-tuned for NVIDIA's NeMo.

- Developers can use the new **Gemma Cookbook** for practical examples and fine-tuning guidance.

- Starting next month, it will be available for easy deployment on Google Cloud's Vertex AI.

**5. Safety & Responsible AI:**

- Extensive safety processes were followed during training, including filtering pre-training data and rigorous testing for biases and risks.

- Developers can access tools like the **Responsible Generative AI Toolkit** and **LLM Comparator** to evaluate and compare models for safety and representational fairness.

- **SynthID**, a text watermarking technology, is being open-sourced to ensure the responsible use of the models.

**6. Project Support:**

- Gemma has already led to innovative projects, such as Navarasa, which used Gemma to create a model focused on India's linguistic diversity.

- Upcoming specialized variants of Gemma 2, like a 2.6B parameter model, will offer further flexibility between performance and accessibility.