Амирхон Абдунабиев

Студент в
The Knowledge Society

ИИ Энтузиаст

Google Developer Group
Editable Location

Эра открытого ИИ:

Развёртываем Gemma локально и строим AI-решения

Build with AI

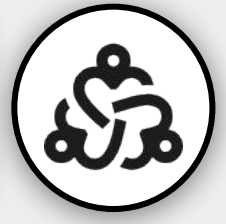
Сегодня поговорим:

1. Эра открытого ИИ
2. Знакомство с Gemma
3. Локальное развертывание Gemma
4. Вызовы и потенциальные возможности
5. Q&A

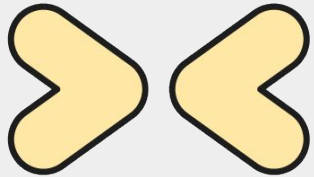
Что такое открытый ИИ?

Прозрачность и конфиденциальность

Доступность и Демократизация



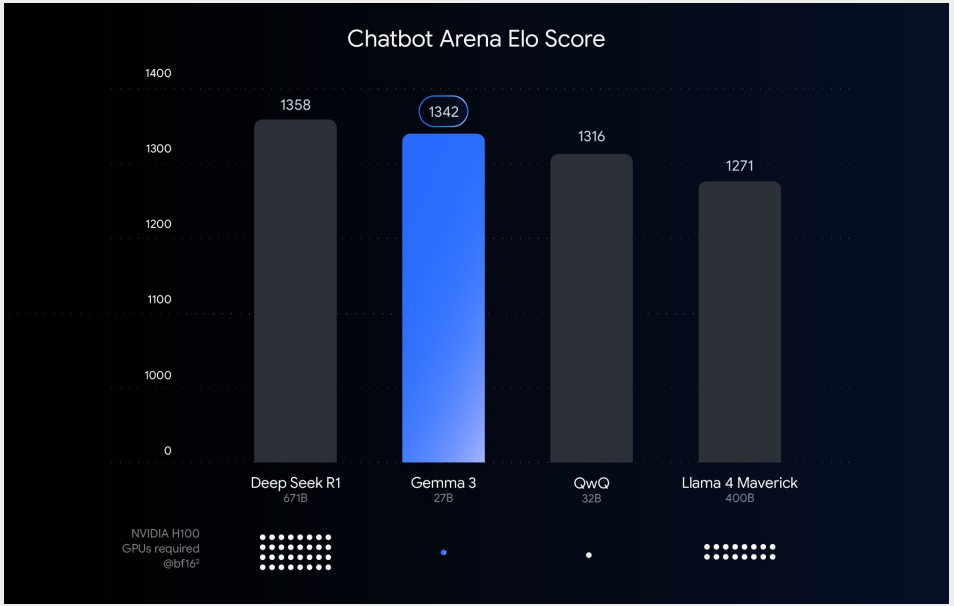
Стимулирование инноваций и сообщества

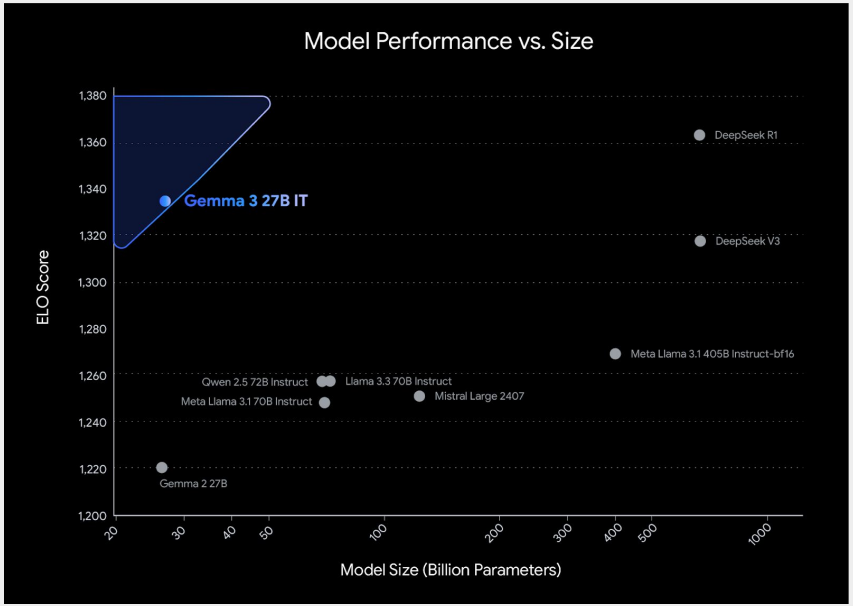


Возможность кастомизации и доработки

Gemma

Google Gemma





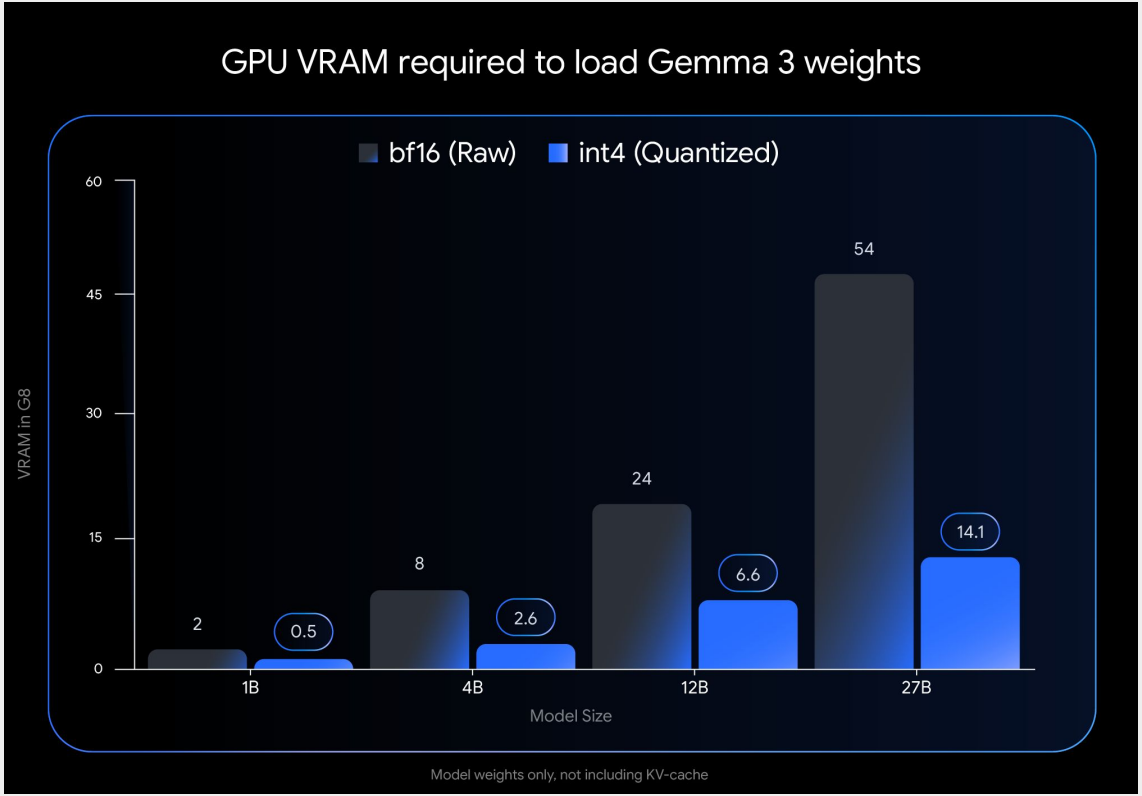
Gemma Family

Локальное развёртывание Google Gemma

Квантизация LLM

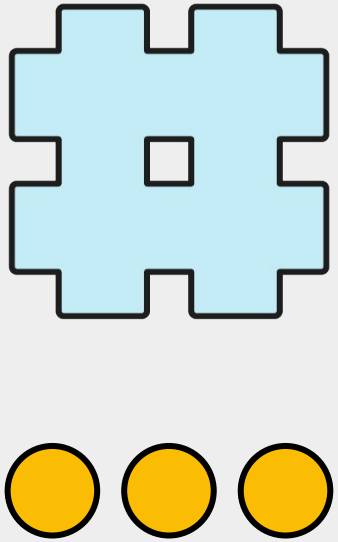
Квантизация — это процесс преобразования этих чисел к более низкому формату

Google Gemma 3 (Quantized)



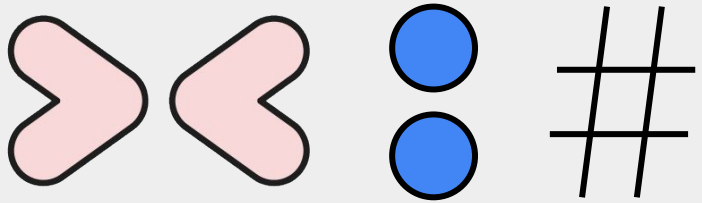
Технические требования Gemma 3

Parameters	Full 32bit	BF16 (16-bit)	SFP8 (8-bit)	Q4_0 (4-bit)	INT4 (4-bit)
Gemma 3 1B (*text only*)	4 GB	1.5 GB	1.1 GB	892 MB	861 MB
Gemma 3 4B	16 GB	6.4 GB	4.4 GB	3.4 GB	3.2 GB
Gemma 3 12B	48 GB	20 GB	12.2 GB	8.7 GB	8.2 GB
Gemma 3 27B	108 GB	46.4 GB	29.1 GB	21 GB	19.9 GB



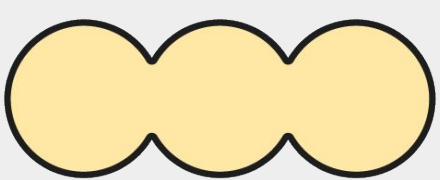
Локальное развёртывание Google Gemma

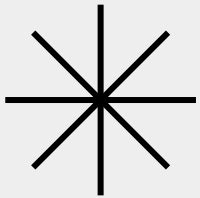
Вызовы и потенциальные возможности



Вызовы:

- Зависимость от мощности оборудования
- Сложность масштабирования
- Потребление ресурсов
- Распространение вредоносного контента





Возможности

- Постоянное повышение доступности
- Инновации в инструментах развёртывания
- Развитие специализированных моделей
- Рост сообщества
- Новые бизнес-модели

Спасибо за ваше внимание!

Q&A