BIZCOMMUNITY

Meta enters the advanced cloud LLM fight with Llama 3

By Lindsey Schutters

18 Apr 2024

Meta has released its latest large language model (LLM), Meta Llama 3, aiming to enhance its AI offerings and shore up its position in the cloud computing market. The company emphasises an open-source approach to foster a collaborative AI ecosystem for developers. Meta Llama 3 will be available on popular cloud platforms including Amazon's AWS, Google Cloud, and Microsoft Azure.

Meta's latest LLM Llama 3 is taking aim at Gemini Advanced and OpenAI's GPT-4

The new LLM features two models – one with 8 billion parameters and another with 70 billion. These massive parameter counts enable Meta's AI to excel in reasoning, coding, and following instructions with what the company is calling "unprecedented capabilities".

The development team used a standard decoder-only transformer architecture for Llama 3 but made key improvements over Llama 2. A tokenizer with a 128,000 token vocabulary improves the way the model encodes language, boosting performance.

Google renames Bard, goes all in on Gemini with new subscription plan
Lindsey Schutters 8 Feb 2024

Llama 3 also includes grouped query attention (GQA) across both the 8bn and 70bn sizes to streamline calculations. Models are trained on sequences of 8,192 tokens, with masking ensuring focus remains within defined boundaries.

Meta collected a large amount of high-quality training data for Llama 3, with over 15trn tokens from public sources – seven times bigger than Llama 2's dataset and four times more focused on code.

More than 5% of this dataset is in languages other than English (over 30 in total), making a solid foundation for strong multilingual capabilities (though non-English performance is not yet as good as English standards).

Data filtering

To ensure quality, Meta created complex data-filtering pipelines. Heuristic filters, NSFW filters, deduplication methods, and text classifiers that estimate data quality all contribute.

Interestingly, even previous versions of Llama were already surprisingly good at finding high-quality data, with Llama 2 used to help create classifiers for Llama 3.

Optimised pretraining makes full use of Llama 3's data abundance through precise scaling laws for downstream tasks. This guided data blending for the best performance across trivia, STEM, coding, history, and more.

Crucially, scaling laws also help estimate the performance of the largest models before they're fully trained, ensuring high results across various applications.

Observations indicate that while the 'optimal' training compute for an 8bn parameter model was initially believed to be around 200bn tokens, the model showed improvement even with much more data.

Both models continued to show log-linear improvement up to 15trn tokens.

While larger models can achieve the same performance as smaller ones with less training, smaller models are often favoured for their efficiency during inference.

Responsibility at scale

Meta used a combination of data, model, and pipeline parallelisation to train the biggest Llama 3 models.

The team achieved over 400 TFLOPS per GPU on 16,000 GPUs at the same time, using two custom-made clusters of 24,000 GPUs each. A new training stack was essential for keeping the GPU usage high, and for automatically detecting, handling, and fixing errors.

Meta sees Llama 3 as the basis of wider systems that developers can customise for their specific requirements. Instruction fine-tuning is still essential for safety, with in-house and external red-teaming (thorough adversarial testing).

Llama Guard models offer prompt/response safety and can be readily fine-tuned for application-specific needs.

It also uses the new MLCommons taxonomy for industry standardisation efforts. Moreover, CyberSecEval 2 and Code Shield reduce problems with unsafe code generation and abuse risks.

ABOUT LINDSEY SCHUTTERS

Lindsey is the editor for ICT, Construction&Engineering and Energy&Mining at Bizcommunity

- #Computex2024: AMD and Qualcomm ignite the AI revolution at Intel's expense - 3 Jun 2024
- DCDT overhauls radio frequency spectrum policy - 31 May 2024
- Vodacom goes to war against spectrum pooling - 30 May 2024
- Icasa extends deadline for digital migration regulations review - 27 May 2024
- HPE takes aim at Cisco, emphasises partner ecosystem and AI focus - 24 May 2024

View my profile and articles...