Meta unveils Llama 3.2: Smaller AI models for edge and mobile devices

capacitymedia.com/article/meta-llama-3-2

September 26, 2024

Meta has unveiled new versions of its popular open source AI model Llama, with small and medium-sized models capable of powering workloads on edge and mobile devices.

Llama 3.2 models were shown at the company's annual Meta Connect event. They're capable of supporting multilingual text generation and vision applications like image recognition.

"This is our first open source, multimodal model, and it's going to enable a lot of interesting applications that require visual understanding," said Mark Zuckerberg, CEO of Meta.

New Llamas join the flock

Llama 3.2 follows Llama 3.1 model, the massive open source model released in late July.

The prior Llama model is the largest open-source AI model in history, standing at a whopping 405 billion parameters — parameters are the adjustable variables within an AI model that help it learn patterns from data. The size reflects the complexity and capacity of the AI to understand and generate human-like text.

The new Llama models unveiled at Meta Connect 2024 are much smaller in size. Meta stated that they chose to design smaller models because not all researchers have the significant computing resources and expertise needed to run a model as massive as Llama 3.1.

LLama 3.2 comes in a variety of sizes, ranging from one billion parameters to 90 billion.

They're split into two segments - the small size (1B and 3B) which are designed to be lightweight and can handle only text inputs. These can fit onto edge and mobile devices, enabling them to process inputs on-device.

The 1B and 3B Llama 3.2 models can support up to 128K tokens (~ 96,240 words) and achieve state-of-the-art performance for use cases like summarisation, instruction following, and rewriting tasks run at the edge.

Meta suggested that the ability to run the models locally is more secure as data is not sent into the cloud and processing is done fast enabling responses to "feel instantaneous."

The smaller models are enabled to run on hardware from Qualcomm and MediaTek and are specially optimised to run on Arm-based processors.

The medium-sized systems, which are 11 and 90 billion parameters in size, are multimodal meaning they're capable of processing inputs beyond text, such as visual inputs like images.

The larger-sized 3.2 models can take in both image and text prompts while also understanding and reasoning better based on the combination of inputs.

For example, the medium-sized Llama 3.2 models can be used for use cases like understanding charts and graphs, enabling businesses to use them to gain insights about sales figures on financial statements.

Performance: LLama 3.2 beats OpenAI and Anthropic models

In terms of performance, Meta's new Llama 3.2 models are competitive with industry-leading systems from Anthropic and OpenAI.

The 3B model outperforms Google's Gemma 2 2.6B and Microsoft's Phi 3.5-mini on tasks like instruction following and content summarisation.

The largest of the models, the 90B version, outperforms both Claude 3-Haiku and GPT-4o-mini on a variety of benchmarks, including the popular MMLU test, an industry-leading evaluation tool for AI models.

Vision instruction-tuned benchmarks

Modality	Benchmark	Llama 3.2 11B	Llama 3.2 90B	Claude 3 - Haiku	GPT-4o-mini
Image	MMMU	50.7	60.3	50.2	59.4
	MMMU-Pro, Standard	33.0	45.2	27.3	42.3
	MMMU-Pro, Vision	23.7	33.8	20.1	36.5
	MathVista	51.5	57.3	46.4	56.7
	ChartQA	83.4	85.5	81.7	—
	AI2 Diagram	91.1	92.3	86.7	—
	DocVQA	88.4	90.1	88.8	—
	VQAv2	75.2	78.1	—	—
Text	MMLU	73.0	86.0	75.2	82.0
	MATH	51.9	68.0	38.9	70.2
	GPQA	32.8	46.7	33.3	40.2
	MGSM	68.9	86.9	75.1	87.0

Llama 3.2's safety features

Since Meta's LLama models are accessible to anyone, Meta has moved to ensure the models are safe and secure.

Building on previous safeguards, Meta has introduced a new Guard feature to support the image understanding for medium-sized models.

Also introduced were a series of filters, preventing certain text and impact outputs from occurring to specific prompts.

The smaller scale Llama 3.2 models feature an optimised Llama Guard which reduces them further. Llama Guard 3 1B is essentially a "pruned" iteration of the one billion 3.2 version but shrunk down to be more basic in function but also smaller in size — from 2,858 MB down to just 438 MB, enabling it to fit on consumer-grade USB sticks.

How to access Llama 3.2 models

The new Llama 3.2 models are open source, meaning anyone can download them and use them to power AI applications.

The models can be downloaded directly from llama.com and Hugging Face, the popular open source repository platform.

Llama 3.2 models can also be accessed through a variety of cloud partners, including Google Cloud, AWS, Nvidia, Microsoft Azure, and Grow among many others.

Figures published in early September suggested that demand for Meta's Llama models from cloud users grew 10 times from January to July — which will likely increase further in the wake of the new 3.2 line of models.

Meta partner Together AI is offering free access to the vision version of Llama 3.2 11B on its platform through the end of the year

Vipul Ved Prakash, founder and CEO of Together AI said the new multimodal models will further accelerate the growth of open-source AI among developers and enterprises.

"We're thrilled to partner with Meta to offer developers free access to the Llama 3.2 vision model and to be one of the first API providers for Llama Stack," Prakash said.

"With Together AI's support for Llama models and Llama Stack, developers and enterprises can experiment, build, and scale multimodal applications with the best performance, accuracy, and cost."