TextSummarizer.ai Documentation

Model access: [LLama 3.2]

Model Release Date: Sept 25, 2024

Model Information: The Meta Llama 3.2 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction-tuned generative models in 1B and 3B sizes (text in/text out). The Llama 3.2 instruction-tuned text only models are optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks.

Model Architecture: Llama 3.2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

	Training Data	Params	Input modalities	Output modalities	Context Length	GQA	Shared Embeddings	Token count	Knowledge cutoff
Llama 3.2 (text only)	A new mix of publicly available online data.	1B (1.23B)	Multilingual Text	Multilingual Text and code	128k	Yes	Yes	Up to 9T tokens	December 2023
			3B (3.21B)	Multilingual Text	Multilingual Text and code				

Supported Languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai are officially supported. Llama 3.2 has been trained on a broader collection of languages than these 8 supported languages.

Steps to setup application in Local Env:

Step 1: Clone the code from the github repo https://github.com/ashitosh098/textSummarizer

Step 2: In the base directory open terminal and run command **npm i** or **npm install**

Step 3: Run the command **npm run dev** and the application will run on your localhost.

Pricing of the Model:

The pricing for Llama-3.2-3B-Instruct is as follows:

- $0.03 per million input tokens
- $0.05 per million output tokens.

Training Energy Use: Training utilized a cumulative of 916k GPU hours of computation on H100-80GB (TDP of 700W) type hardware, per the table below. Training time is the total GPU time required for training each model and power consumption is the peak power capacity per GPU device used, adjusted for power usage efficiency.