

Llama-3.2-1B-Instruct

Introduction

The 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. They outperform many of the available open source and closed chat models on common industry benchmarks.

Available NPU Models

Base Model

llama3.2-1B-prefill-ax630c

The **Base Model** providing a 128 context window and a maximum output of 1,024 tokens.

Support Platforms: LLM630 Compute Kit, Module LLM, and Module LLM Kit

- 128 context window
- 1,024 max output tokens
- ttft 891ms
- avg-token/s 4.48

Install

```
apt install llm-model-llama3.2-1b-prefill-ax630c
```

- [Download llm-model-llama3.2-1b-prefill-ax630c](#)

Long-Context Model

llama3.2-1B-p256-ax630c

The **Long-Context Model** Compared to the **Base Model**, it provides extended context capabilities, offering a 256 context window and a maximum of 1,024 output tokens.

Support Platforms: LLM630 Compute Kit, Module LLM, and Module LLM Kit

- 256 context window
- 1,024 max output tokens
- ttft 2601.11ms
- avg-token/s 4.49

Install

```
apt install llm-model-llama3.2-1b-p256-ax630c
```

- [Download llm-model-llama3.2-1b-p256-ax630c](#)