From generic to genius: Fine-tuning LLMs to enhance AI performance and reliability

Jasper Kyle Catapang, MA

Zoom - May 30, 2024

Jasper Kyle Catapang

- NLP Lead at Maya Philippines
- MA in Applied Linguistics, University of Birmingham, UK
- BSc in Computer Science, University of the Philippines
- Developer of GodziLLa-2, Maya's top-ranking open-source LLM
- Publications on code-switching NLP, NLP ethics, social computing
- Industry experience in NLG, conversational agents, Q&A

Outline

- **What are foundation models (FMs)?**
  An overview discussing what foundation models are and how they differ from large language models (LLMs)
- **What are the different LLM-based training techniques?**
  A discussion of various approaches to training LLM models
- **What is GodziLLa-2?**
  A brief description of GodziLLa-2, its training process, capabilities, and accomplishments..

What are foundational models?

What are some features of foundation models?

Foundation models

Foundation models

L _ _ G _ - S _ _ _ E

Foundation models

LARGE-SCALE

Foundation models

_ E R _ _ T _ L E

Foundation models

VERSATILE

Foundation models

G _ _ ER _ _ - P _ _ PO _ _

Foundation models

GENERAL-PURPOSE

Foundation models

Large-scale

Versatile

General-purpose

Foundation models are large, pre-trained AI models designed to be adapted for a wide range of downstream tasks. They serve as the base upon which specialized models can be built.

Why are foundation models important?

Importance of FMs

- **Efficiency**: Reduce the need to train models from scratch.
- **Transfer Learning**: Leverage knowledge from one domain to another.
- **Innovation**: Facilitate the development of new applications and technologies.

What are some examples of foundation models?

Examples of FMs

1. **GPT-4 (OpenAI)**
    a. ~1T - 1.7T parameters (rumored)
    b. Language
    c. Text generation, question answering, translation, etc.
2. **ViT-Large (Google)**
    a. ~300M parameters
    b. Vision
    c. Image classification, object detection, etc.
3. **wav2vec 2.0 (Meta)**
    a. ~300M parameters
    b. Speech
    c. automatic speech recognition (ASR), speaker identification, language identification, etc.

What are large language models then?

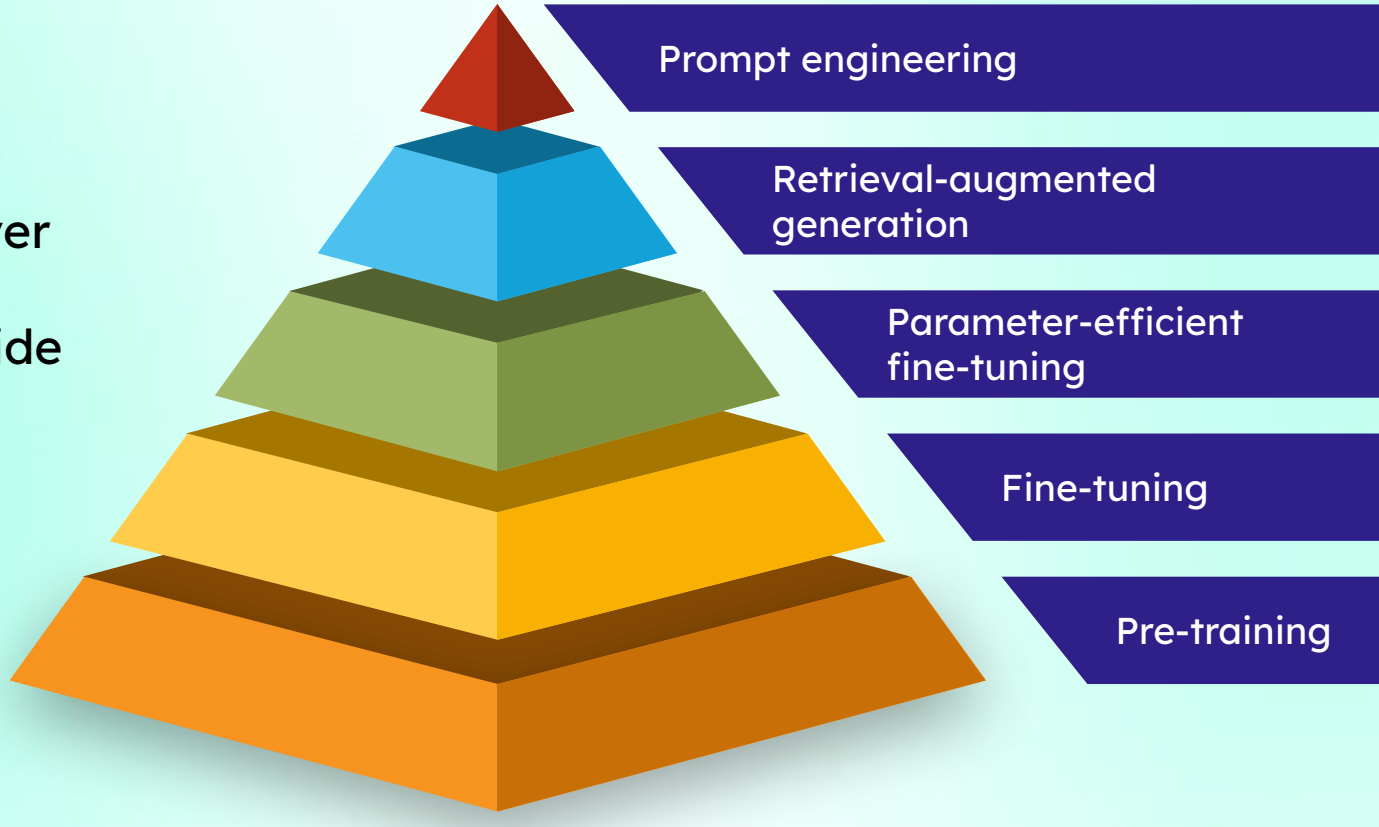
Large language models

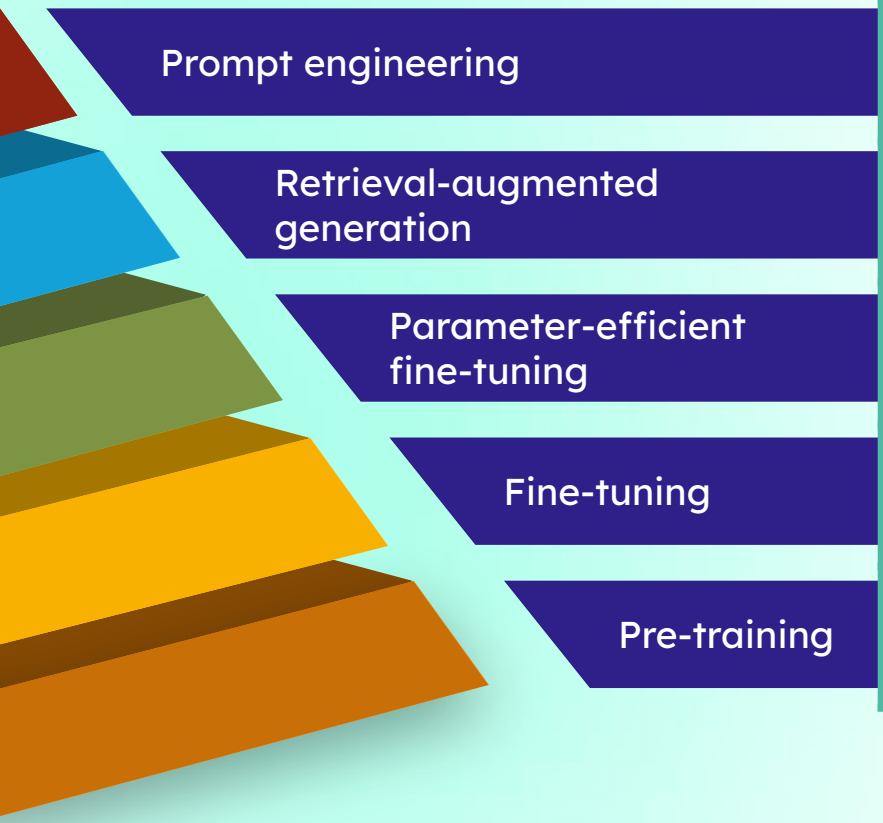
Large language models or LLMs are a specific type of foundation models. LLMs are language-based FMs.

Training techniques for LLM-based architectures

LLM training hierarchy

⬆️ easier to do
⬆️ more people implement it
⬇️ more control over data in model
⬇️ harder to override inserted data





	parameters modified	purpose	implementation difficulty
Prompt engineering	none	better responses	🌕🌑🌑🌑🌑
Retrieval-augmented generation	none	incorporate external data	🌕🌕🌓🌑🌑
Parameter-efficient fine-tuning	a small subset	efficient task adaptation	🌕🌕🌕🌕🌑
Fine-tuning	all	task-specific adaptation	🌕🌕🌕🌕🌓
Pre-training	all	train from scratch	🌕🌕🌕🌕🌕

What is GodziLLa-2? 

GodziLLa-2

GodziLLa-2 is an experimental large language model developed by Maya Philippines. It is based on Meta AI's Llama 2 and aims to push the boundaries of instruction-tuned LLMs. It is a composite model consisting of various Maya-curated text datasets and public text datasets made possible through low-rank adapters.

GodziLLa-2 is an experimental large language model developed by Maya Philippines. It is based on Meta AI's Llama 2 and aims to push the boundaries of instruction-tuned LLMs. It is a composite model consisting of various Maya-curated text datasets and public text datasets made possible through low-rank adapters.

-GodziLLa-2, May 22, 2024

What's on your mind?

GodziLLa-2 is an **experimental** large language model developed by Maya Philippines. It is based on Meta AI's **Llama 2** and aims to push the boundaries of **instruction-tuned** LLMs. It is a **composite** model consisting of various **Maya-curated text datasets** and public text datasets made possible through **low-rank adapters**.

-GodziLLa-2, May 22, 2024

What's on your mind?

Llama 2

Experimental

Instruction-tuned

Composite

Maya-curated text datasets

Low-rank adapters

Llama 2

Experimental

Instruction-tuned

Composite

Maya-curated text datasets

Low-rank adapters

- Meta released Llama 2 in July 2023
- Three LLM sizes: 7B, 13B, 70B
- Caused a surge in top-performing open-source LLMs

Llama 2

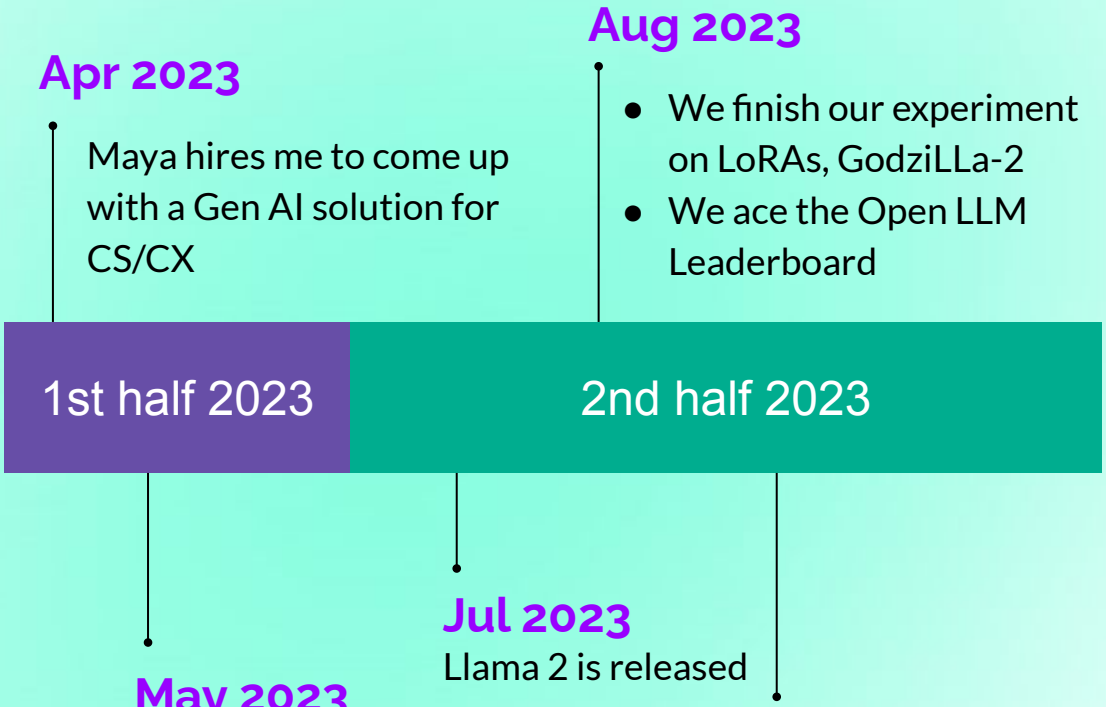
Experimental

Instruction-tuned

Composite

Maya-curated text datasets

Low-rank adapters



Llama 2

Experimental

Instruction-tuned

Composite

Maya-curated text datasets

Low-rank adapters

Instruction-tuned LLMs are a *subset* of fine-tuned LLMs. As the name suggests, it's fine-tuned on instruction data.

Example:
"Write a summary of X, make it concise and simple. Don't use jargon or any complicated words.

Llama 2

Experimental

Instruction-tuned

Composite

Maya-curated text datasets

Low-rank adapters

A composite is any model made by merging or combining smaller modules.

An example of a composite architecture is a mixture-of-experts (MoE) architecture. MoEs are similar to ensemble models in classical machine learning.

Examples of MoE architectures include Mixtral and GPT-4 (rumored)

Llama 2

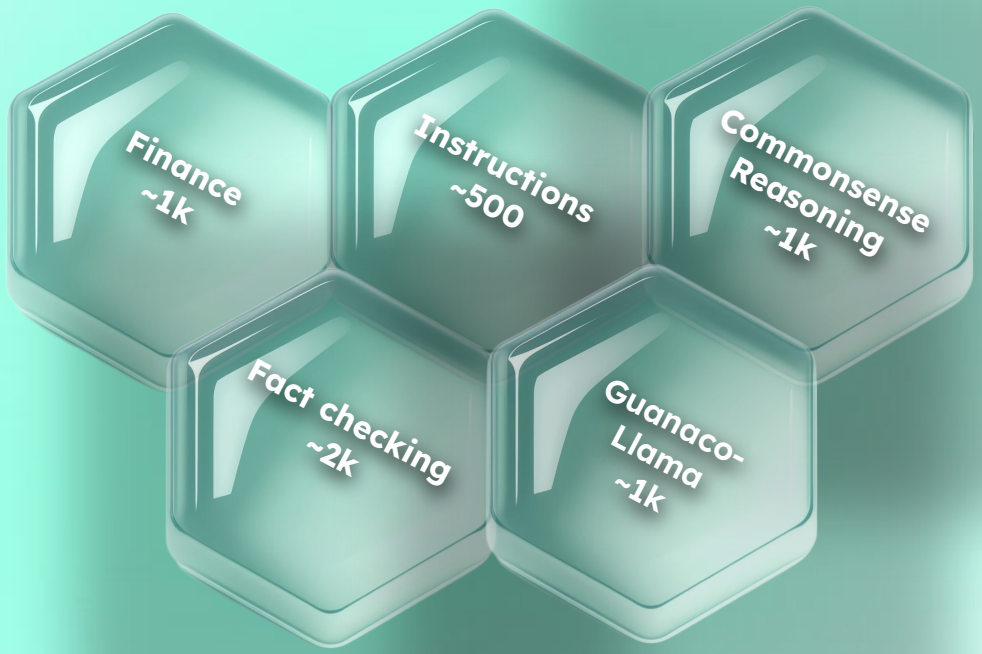
Experimental

Instruction-tuned

Composite

Maya-curated text datasets

Low-rank adapters



* text datasets curated by Maya DS/AI

Llama 2

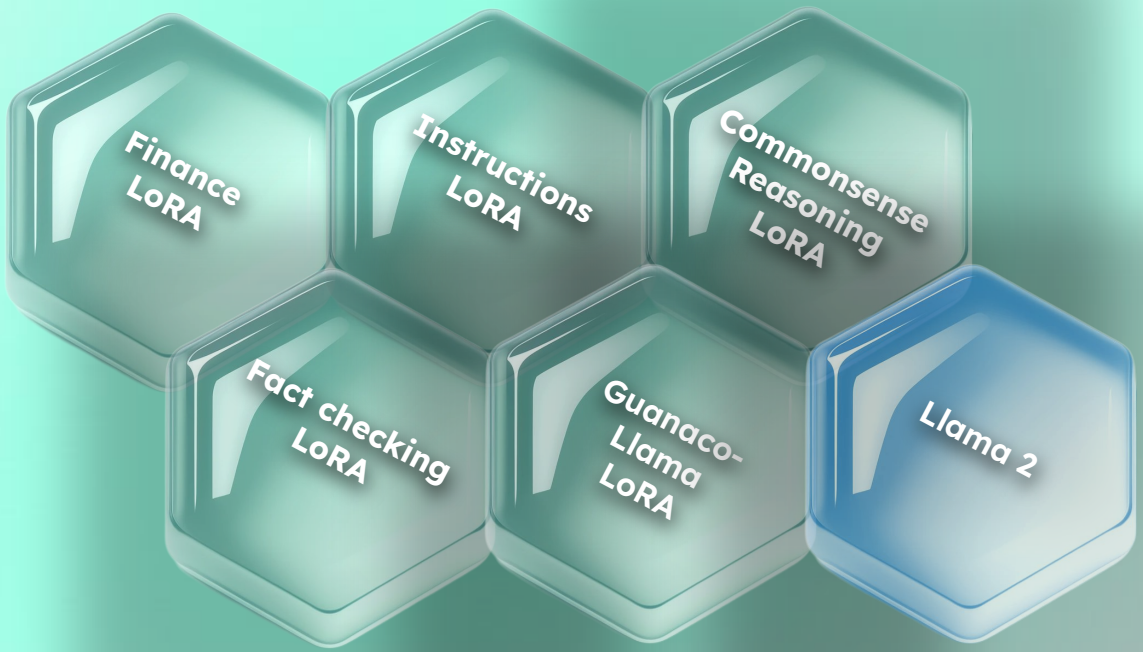
Experimental

Instruction-tuned

Composite

Maya-curated text datasets

Low-rank adapters



* all five (5) LoRA adapters were merged into the base model

What can GodziLLa-2 do?

GodziLLa-2 features

Instruction following

Executing complex tasks with precision and reliability

Question answering

Expertly answering questions across a wide range of topics and contexts

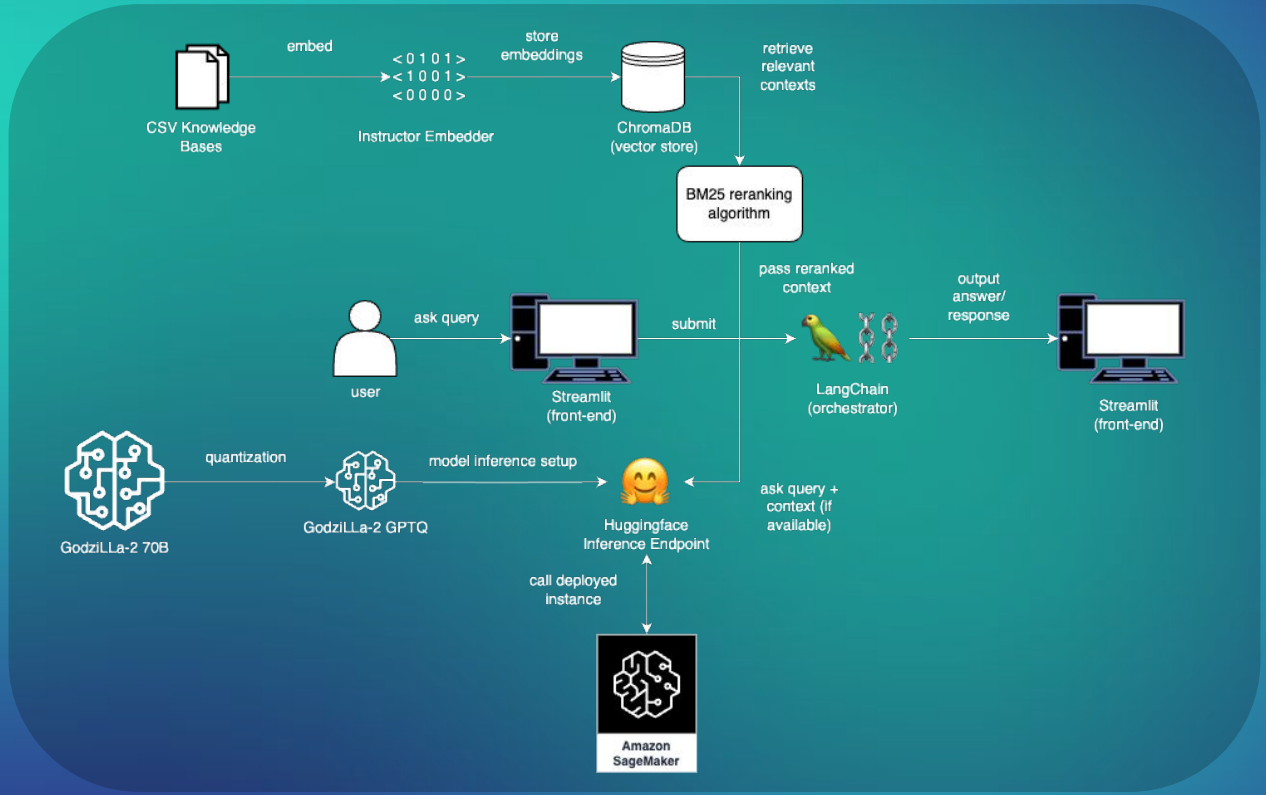
Enhanced truthfulness

Less tendency to generate false information, ensuring accuracy and promotes trustworthiness

Basic reasoning

Performing basic reasoning tasks, drawing logical conclusions effectively

GodziLLa-2 Demo



What did GodziLLa-2 achieve?

GodziLLa-2 feats (Nov 2023)

1. Godzilla 2 70B debuts at 2nd place worldwide in the newly updated Open LLM Leaderboard.
2. Godzilla 2 70B beats GPT-3.5 (ChatGPT) in terms of average performance and the HellaSwag benchmark (87.53 > 85.5).
3. Godzilla 2 70B outperforms GPT-3.5 (ChatGPT) and GPT-4 on the TruthfulQA benchmark (61.54 for G2-70B, 47 for GPT-3.5, 59 for GPT-4).
4. Godzilla 2 70B is on par with GPT-3.5 (ChatGPT) on the MMLU benchmark (<0.12%).

Highlights

Highlights

- Foundation Models: Large, pre-trained AI models adaptable for various tasks; versatile and efficient.
- LLMs: Specialized foundation models optimized for language tasks.
- LLM Training: Techniques include prompt engineering, fine-tuning, and pre-training.
- GodziLLa-2: Experimental LLM by Maya Philippines, based on Llama 2, excelling in instruction-tuned tasks.
- Achievements: Ranked 2nd on Open LLM Leaderboard, outperforming GPT-3.5 in key benchmarks.
- You can leverage different training techniques and curate your own data to create your own LLMs that are sure to top the charts!

From generic to genius:
Fine-tuning LLMs to enhance AI performance and reliability

Jasper Kyle Catapang, MA

jcatapang
jaspercatapang
NLPinas
jasperkylecatapang@gmail.com

PACLIC38

38th Pacific Asia Conference on
Language, Information and Computation

Tokyo, Japan · 2024 December 7-9

東京外国語大学
Tokyo University of Foreign Studies

IMPORTANT DATES

31 JUL 2024	31 AUG 2024	30 SEP 2024	31 OCT 2024	30 NOV 2024	7-9 DEC 2024
Deadline of Submission of Papers	Notification of Acceptance	Deadline for Presenters' Registration	Deadline for Camera-Ready Papers	Deadline for Participants' Registration	The Conference

Questions?

Thank you very much.
See you next week!