Llama Solution

Resources

Inference Solutions

On Prem:

- TensorRT-LLM
- vLLM
- TGI
- Ollama
- Llama.cpp

Llama Cloud Offerings

- Bedrock
- Azure

Model API Providers

- Together.ai
- Replicate
- Octo.Ai
- AnyScale and others..

Additional Resources

Getting to know Llama 2

Leading LLM App Development Frameworks:

- LangChain
- LlamaIndex

Prompt Engineering:

- Prompt engineering with Llama 2 example notebook
- Prompt engineering with Llama 2 on Deeplearning.AI (DLAI)
- A guide to prompting Llama 2 on Replicate
- 5 Steps to Getting Started with Llama 2
- How to Prompt Llama 2 on Hugging Face

Additional Resources

RAG

- Llama Recipes: Building a Llama 2 chatbot with RAG notebook
- LangChain Chat with Your Data on DLAI
- Hugging Face Massive Text Embedding Benchmark Leaderboard
- Over 60 Vector Stores Supported by LangChain
- Over 130 document loaders supported by LangChain

Additional Resources

Fine-Tuning

- Llama-Recipes
- TorchTune
- HuggingFace SFT and TRL
- Bedrock Fine-Tuning
- Azure Fine-Tuning
- Finetuning Large Language Models on DLAI
- LlamaIndex Fine-tuning

Meta

Additional Resources

Agents

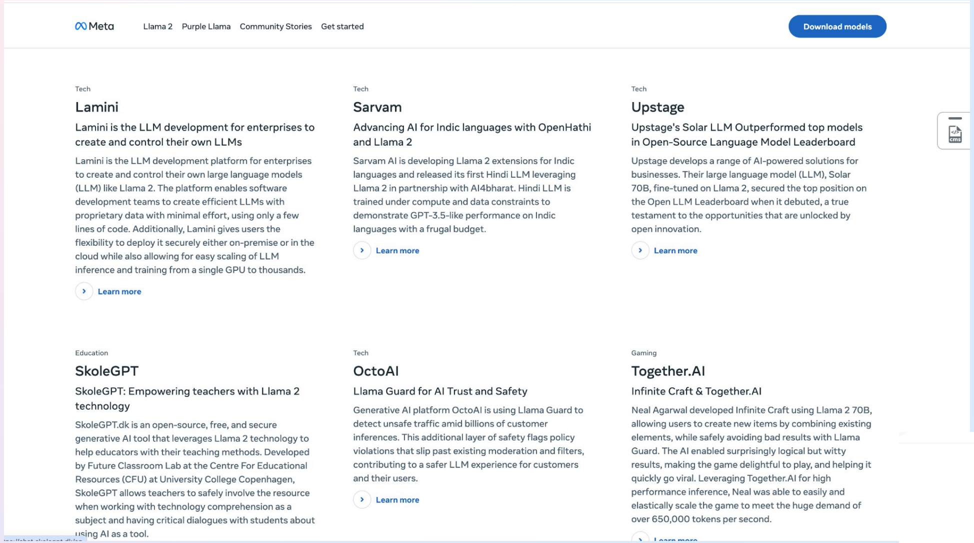
- LLM Powered Autonomous Agents
- LangChain Agent Documentation
- Functions, Tools and Agents with LangChain on DLAI
- What’s next for AI agentic workflows by Andrew Ng
- Introduction to LLM Agents

Additional Resources

LLM App Evaluation

- DeepEval
- UpTrain
- TruLens
- Language Model Evaluation Harness
- HELM (Holistic Evaluation of Language Models)
- Building and Evaluating Advanced RAG on DLAI
- Best Practices for LLM Evaluation of RAG Applications
- Building RAG-based LLM Applications for Production

Community Stories



Tech

Lamini

Lamini is the LLM development for enterprises to create and control their own LLMs

Lamini is the LLM development platform for enterprises to create and control their own large language models (LLM) like Llama 2. The platform enables software development teams to create efficient LLMs with proprietary data with minimal effort, using only a few lines of code. Additionally, Lamini gives users the flexibility to deploy it securely either on-premise or in the cloud while also allowing for easy scaling of LLM inference and training from a single GPU to thousands.

Learn more

Tech

Sarvam

Advancing AI for Indic languages with OpenHathi and Llama 2

Sarvam AI is developing Llama 2 extensions for Indic languages and released its first Hindi LLM leveraging Llama 2 in partnership with AI4bharat. Hindi LLM is trained under compute and data constraints to demonstrate GPT-3.5-like performance on Indic languages with a frugal budget.

Learn more

Tech

Upstage

Upstage's Solar LLM Outperformed top models in Open-Source Language Model Leaderboard

Upstage develops a range of AI-powered solutions for businesses. Their large language model (LLM), Solar 70B, fine-tuned on Llama 2, secured the top position on the Open LLM Leaderboard when it debuted, a true testament to the opportunities that are unlocked by open innovation.

Learn more

Education

SkoleGPT

SkoleGPT: Empowering teachers with Llama 2 technology

SkoleGPT.dk is an open-source, free, and secure generative AI tool that leverages Llama 2 technology to help educators with their teaching methods. Developed by Future Classroom Lab at the Centre For Educational Resources (CFU) at University College Copenhagen, SkoleGPT allows teachers to safely involve the resource when working with technology comprehension as a subject and having critical dialogues with students about using AI as a tool.

Tech

OctoAI

Llama Guard for AI Trust and Safety

Generative AI platform OctoAI is using Llama Guard to detect unsafe traffic amid billions of customer inferences. This additional layer of safety flags policy violations that slip past existing moderation and filters, contributing to a safer LLM experience for customers and their users.

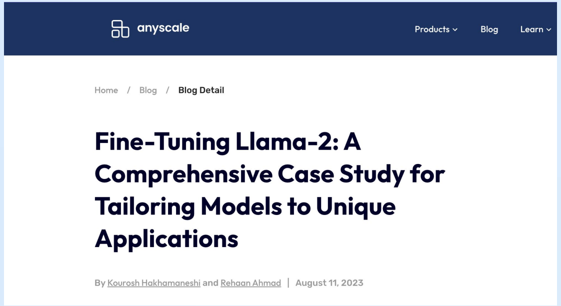
Learn more

Gaming

Together.AI

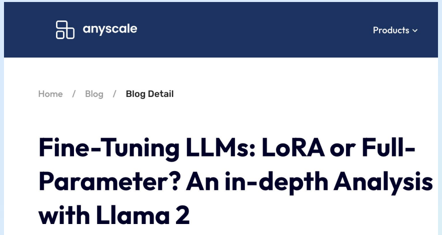
Infinite Craft & Together.AI

Neal Agarwal developed Infinite Craft using Llama 2 70B, allowing users to create new items by combining existing elements, while safely avoiding bad results with Llama Guard. The AI enabled surprisingly logical but witty results, making the game delightful to play, and helping it quickly go viral. Leveraging Together.AI for high performance inference, Neal was able to easily and elastically scale the game to meet the huge demand of over 650,000 tokens per second.

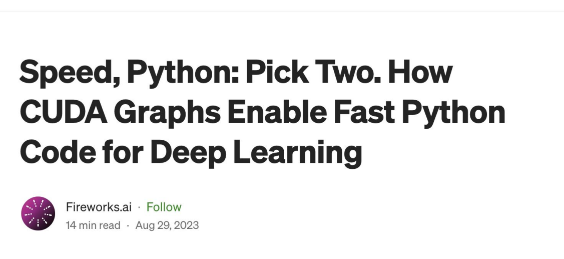


Fine-Tuning Llama-2: A Comprehensive Case Study for Tailoring Models to Unique Applications

By Kourosh Hakhamaneshi and Rehaan Ahmad | August 11, 2023

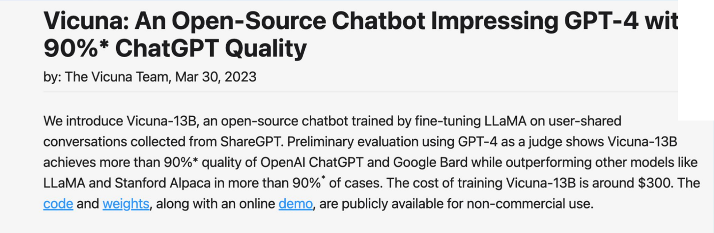


Fine-Tuning LLMs: LoRA or Full-Parameter? An in-depth Analysis with Llama 2



Speed, Python: Pick Two. How CUDA Graphs Enable Fast Python Code for Deep Learning

Fireworks.ai · Follow
14 min read · Aug 29, 2023



Vicuna: An Open-Source Chatbot Impressing GPT-4 wit 90%* ChatGPT Quality

by: The Vicuna Team, Mar 30, 2023

We introduce Vicuna-13B, an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. Preliminary evaluation using GPT-4 as a judge shows Vicuna-13B achieves more than 90%* quality of OpenAI ChatGPT and Google Bard while outperforming other models like LLaMA and Stanford Alpaca in more than 90%* of cases. The cost of training Vicuna-13B is around $300. The code and weights, along with an online demo, are publicly available for non-commercial use.

Meta