

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 DLA](#)
- [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](#)

Additional Resources

Agents

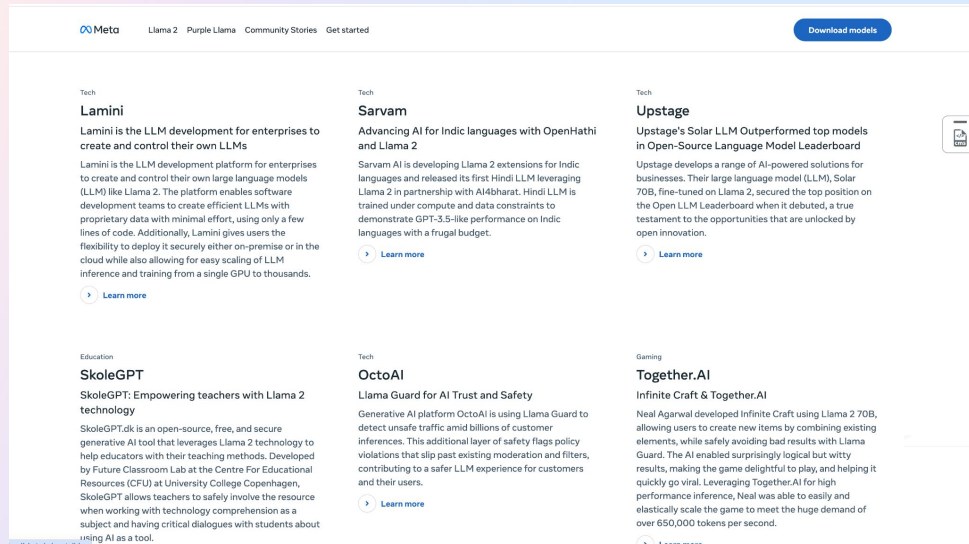
- [LLM Powered Autonomous Agents](#)
- [LangChain Agent Documentation](#)
- [Functions, Tools and Agents with LangChain on DLAI](#)
- [What's next for AI agentic workflows by Andrew Ng](#)
- [Introduction to LLM Agents](#)

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](#)

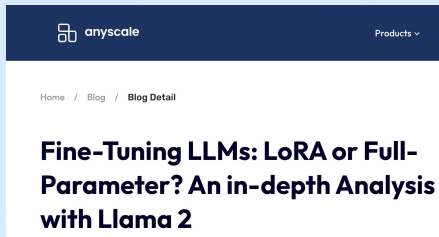
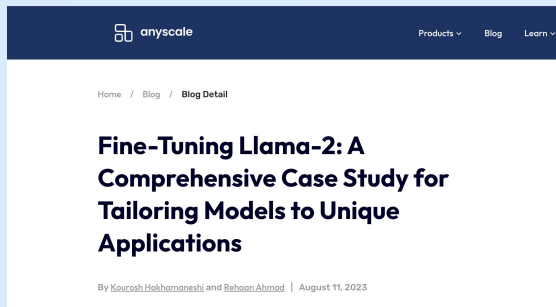
Community Stories



Vicuna: An Open-Source Chatbot Impressing GPT-4 with 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.



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



Fireworks.ai · Follow

14 min read · Aug 29, 2023