Kempner INSTITUTE

For the Study of Natural & Artificial Intelligence

HARVARD UNIVERSITY

WORKSHOPS @ KEMPNER

LARGE LANGUAGE MODEL DISTRIBUTED INFERENCE VIRTUAL WORKSHOP

March 26 @ 9:30am - 12:30pm

This hands-on, virtual workshop covers high-performance LLM inference using vLLM. You will work with leading open-weight models like Qwen and LLaMA to gain practical, real-world experience hosting and running inference for large language models that exceed the memory capacity of a single GPU.

What will attendees learn from this workshop?

- Basics of distributed computing for large language model inference
- Using vLLM, a popular library for LLM inference, to host a server
- Setting up, prompting, and extracting logits from large language models on an HPC cluster
- Using offline batch inference with large language models

Presenter

- Naeem Khoshnevis, Lead ML Research Engineer and Pytorch Ambassador

Who can attend?

- Any Harvard-affiliates with an FASRC account, with priority given to Kempner community members.

Prerequisites

- FAS Research Computing account
- Familiarity with Python
- Familiarity with large language models
- Familiarity with HPC cluster
- Completion of pre-work

Location

- Virtual via Zoom (link provided after registration is confirmed).

Please register as soon as possible (by 3/2 at latest) as space is limited.

For more information visit

kempnerinstitute.harvard.edu/kempner-community/community-calendar

More workshops coming!

Questions? Contact kempnereducation@harvard.edu