Overview

The **Vector Inference Platform** is a service provided by the AI Engineering team at Vector Institute. It hosts large, state-of-the-art open-source language models that anyone in the Vector community can use freely and easily.

Unlike previous efforts to provide inference services on Vector's compute environment, this platform is a **production-grade, always-available service**. Users do not need to spin up their own models via Slurm jobs or worry about time limits — models remain persistently online.

The source code and technical documentation for this project are available on the GitHub repository.

For the current list of available models and their specifications, visit inference.vectorinstitute.ai.

The AI Engineering team will continue to add new models as the service matures. Feedback and model requests are welcome — contact the AI Engineering team.

Revision #11
Created 2026-01-06 17:01:44 UTC
Updated 2026-05-26 22:09:18 UTC by Amrit Krishnan