RedHat AI

Red Hat® Enterprise Linux® AI is a foundation model platform to seamlessly develop, test, and run Granite family large language models (LLMs) for enterprise applications.

Red Hat Enterprise Linux AI brings together:

- The Granite family of open source-licensed LLMs, distributed under the Apache-2.0 license with complete transparency on training datasets.
- InstructLab model alignment tools, which open the world of community-developed LLMs to a wide range of users.
- A bootable image of Red Hat Enterprise Linux, including popular AI libraries such as PyTorch and hardware optimized inference for NVIDIA, Intel, and AMD.
- Enterprise-grade technical support and model intellectual property indemnification provided by Red Hat.

URLs:

- Red Hat Enterprise Linux AI
- Red Hat Delivers Accessible, Open Source Generative AI Innovation with Red Hat Enterprise Linux AI

InstructLab

Command-line interface. Use this to chat with the model or train the model (training consumes the taxonomy data)

What are the components of the InstructLab project?

- **Taxonomy**
  InstructLab is driven by taxonomies, which are largely created manually and with care. InstructLab contains a taxonomy tree that lets users create models tuned with human-provided data, which is then enhanced with synthetic data generation.
- **Command-line interface (CLI)**
  The InstructLab CLI lets contributors test their contributions using their laptop or workstation. Community members can use the InstructLab technique to generate a low-fidelity approximation of synthetic data generation and model-instruction tuning without access to specialized hardware.
- **Model training infrastructure**
  Finally, there's the process of creating the enhanced LLMs. It takes GPU-intensive infrastructure to regularly retrain models based on new contributions from the community. IBM donates and maintains the infrastructure necessary to frequently

retrain the InstructLab project’s enhanced models.

How is InstructLab different from retrieval-augmented generation (RAG)?

RAG is a cost-efficient method for supplementing an LLM with domain-specific knowledge that wasn’t part of its pretraining. RAG makes it possible for a chatbot to accurately answer questions related to a specific field or business without retraining the model. Knowledge documents are stored in a vector database, then retrieved in chunks and sent to the model as part of user queries. This is helpful for anyone who wants to add proprietary data to an LLM without giving up control of their information, or who needs an LLM to access timely information.

This is in contrast to the InstructLab method, which sources end-user contributions to support regular builds of an enhanced version of an LLM. InstructLab helps add knowledge and unlock new skills of an LLM.

It’s possible to "supercharge" a RAG process by using the RAG technique on an InstructLab-tuned model.

URLs:

- What is InstructLab? (redhat.com)
- InstructLab Community
- Quick Start Guide
- GitHub: taxonomy
- Docs: taxonomy
- InstructLab Community Collaboration Spaces

Revision #10
Created 10 May 2024 11:35:36 by Admin
Updated 10 May 2024 16:05:42 by Admin