4 considerations for choosing the right AI model

Enterprise organizations are under increasing pressure to realize value with artificial intelligence (AI) and generative artificial intelligence (gen AI) as the technology continues to evolve. A crucial step in that process is selecting the right gen AI model, which can determine how effectively an organization can use AI to fulfill its strategic goals. This checklist provides 4 key considerations when selecting an AI model, and outlines how Red Hat® Enterprise Linux® AI can provide the models you need to build tailored solutions for your unique use cases.

1 What size model should my organization use?

Choosing the right model size for your organization and its specific use cases is a crucial place to start, as it can directly affect the cost of running that model, the skills required to manage it, and the accuracy of its results.

Using a larger model with more parameters can increase capacity for evaluating complex scenarios. However, larger models (and the data that powers them) require increased computational resources and complex infrastructure, which can significantly increase operating costs and the skills needed to manage it.

Those costs and required skills can become even more pronounced when you need to tailor a larger model with data relevant to your enterprise, use cases, or industry.

Using smaller, more specialized models, such as the Granite family large language models (LLMs) included with Red Hat Enterprise Linux AI, is a more cost-effective approach, as they can be trained on your enterprise data to address your specific needs without the intensive levels of costs, skills, and time.

2 How transparent is my model?

Another key factor that can affect the accuracy and efficiency of your gen AI model is the data used to train it, including the data's origin, ownership, quality, accessibility, reliability, relevance, diversity, and volume.

Using a model that is not transparent about the data that was used to train it opens your organization up to significant risks. This could include inaccurate results due to inaccurate data, and even more significantly, the potential legal consequences of using a model that has been trained on copyrighted material.

Red Hat recommends using models with transparent data sources and sets, such as the Granite family of models included with Red Hat Enterprise Linux AI that provide complete training dataset and model content transparency and are fully indemnified and supported by Red Hat.

This allows your organization to move its gen AI initiatives forward confidently with the assurance that the model is backed by a credible source that guarantees the model was built and trained ethically with all regulations in mind, and is proven to safeguard sensitive data.

3 Should I use open source or proprietary models?

Several types of licensing exist for LLMs, including open source licenses (examples include Apache 2.0, general public licenses, and more) and commercial licenses that are usually proprietary or subscription-based.

Whether the model you choose is open source or proprietary can have a great influence on a number of key outcomes, including:

- Operational efficiency and costs.
- Allowed levels of customization.
- Ownership of applications and services built with that model and any data fed into it.

Proprietary models can be more expensive to run, more difficult to train/tune to your use cases or industry, limited in customizability, and lacking in the efficiency offered by smaller, open source models that benefit from community adoption and contribution. Additionally, using a commercial license may require you to pay the owner of the model if you want to sell a service or application you built using the model, or could give them ownership of any data fed into the model.

The open source licensed Granite LLMs offered through Red Hat Enterprise Linux AI provide operational and financial efficiency, and permit you to use, modify, and distribute models liberally, all while allowing you to retain full ownership of your data.

4 How should I train my model?

Tailoring your model can help you meet the needs of specialized use cases with efficiency, accuracy, and minimized costs; and how you adapt that model to your needs can have a significant effect on the efficiency and effectiveness of that model-tuning.

Red Hat Enterprise Linux AI offers accessible model alignment tools through a supported, lifecycle-based distribution of InstructLab.

The InstructLab training method for LLMs addresses the challenges of developing these models and the scaling challenges seen in traditional LLM training. This is achieved with a taxonomy-guided synthetic data generation process and a multiphase tuning framework to improve model performance, and an open source community that allows developers to collectively contribute new skills and knowledge to any model.

Using the InstructLab model alignment tools:

- Provides you with a cost-effective solution for improving the alignment of LLMs.
- Drives improvements to open source models with a community approach.
- Democratizes the process of customizing models with unique data sets through accessible tools, training, and community knowledge.
- Allows you to retain full ownership of training data sets and model IP.

Learn more

Read this e-book to learn how Red Hat can help you get started with your AI innovation.

Get started

Speak to a Red Hatter to discuss how Red Hat can help your organization get started with AI.