Red Hat
Enterprise Linux AI

Plataforma de modelo de base para IA generativa

Principais benefícios

- Permita que usuários atualizem e aprimorem os Large Language Models (LLM) com o InstructLab
- Alinhe os LLMs com dados proprietários de maneira segura para personalizar os modelos conforme as necessidades dos seus negócios
- Inicie rapidamente com a inteligência artificial generativa (gen AI) e entregue resultados em uma plataforma do Red Hat® Enterprise Linux® confiável e focada em segurança
- Empacotada como imagem de container do Red Hat Enterprise Linux inicializável para instalação e atualizações

Visão geral da solução

O Red Hat Enterprise Linux AI é uma plataforma de modelo de base para gen AI de nível empresarial usada para desenvolver, testar e implantar LLMs para casos de uso empresariais de gen AI.

O Red Hat Enterprise Linux AI reúne:

- A família Granite de LLMs open source.
- O InstructLab, uma ferramenta de alinhamento de modelos que oferece uma abordagem desenvolvida pela comunidade para ajuste fino do LLM.
- Uma imagem inicializável do Red Hat Enterprise Linux, além de bibliotecas e dependências de gen AI, como PyTorch e softwares de drivers aceleradores de IA para NVIDIA, Intel, e AMD
- Suporte técnico de nível empresarial e indenização pela propriedade intelectual do modelo fornecida pela Red Hat.
- O Red Hat Enterprise Linux AI disponibiliza a plataforma confiável do Red Hat Enterprise Linux com os componentes necessários para você começar sua jornada na gen AI e ver resultados.

O futuro da IA é open source e transparente

O Red Hat Enterprise Linux AI inclui um subconjunto dos modelos de código e linguagem Granite open source com indenização total da Red Hat. As soluções Granite open source oferecem às organizações modelos de custo e desempenho otimizados e alinhados com uma ampla variedade de casos de uso de gen AI. Os modelos Granite foram liberados sob a licença do Apache 2.0. Assim como os modelos, os subconjuntos usados para treinamento dos modelos também são transparentes e open source.

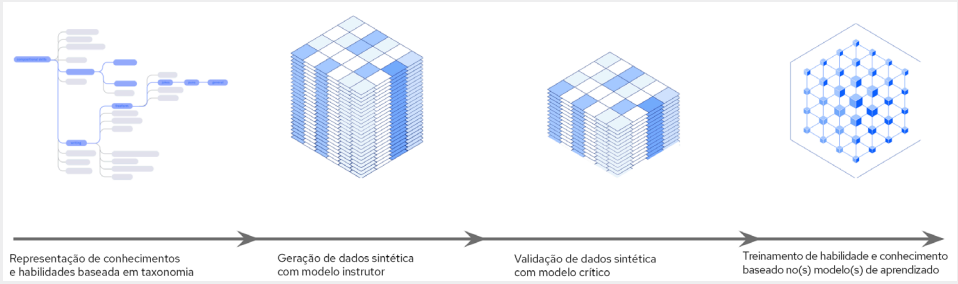
Tabela 1. Modelos Granite no Red Hat Enterprise Linux AI

Modelos de linguagem IBM Granite	Granite-7B-Starter Granite-7B-RedHat-Lab
Modelos de código IBM	Granite-8B-Code-Instruct Granite-8B-Code-Base

Treinamento de modelo de gen AI acessível para um time to value mais rápido

Além dos modelos Granite open source, o Red Hat Enterprise Linux AI também inclui o InstructLab, uma ferramenta de alinhamento de modelo baseada na técnica de alinhamento de larga escala para chatBots (LAB). Com o InstructLab, as equipes das organizações fazem contribuições eficientes para os LLMs com habilidades e conhecimento, personalizando esses modelos para necessidades empresariais específicas.

- Habilidade: domínio de uma capacidade com o objetivo de ensinar um modelo a fazer algo. As habilidades são classificadas em duas categorias.
  - Habilidades composicionais
    - Permitem que os modelos de IA executem tarefas ou funções específicas.
    - São fundamentadas (inclui contexto) ou não (não inclui contexto).
      - Um exemplo fundamentado é a adição de uma habilidade para que o modelo consiga ler uma tabela com formatação em markdown.
      - Um exemplo não fundamentado é a adição de uma habilidade para ensinar o modelo a rimar.
  - Habilidades de base
    - Habilidades como matemática, lógica e codificação.
- Conhecimento: dados e fatos que oferecem um modelo com informações adicionais para responder perguntas com maior precisão.

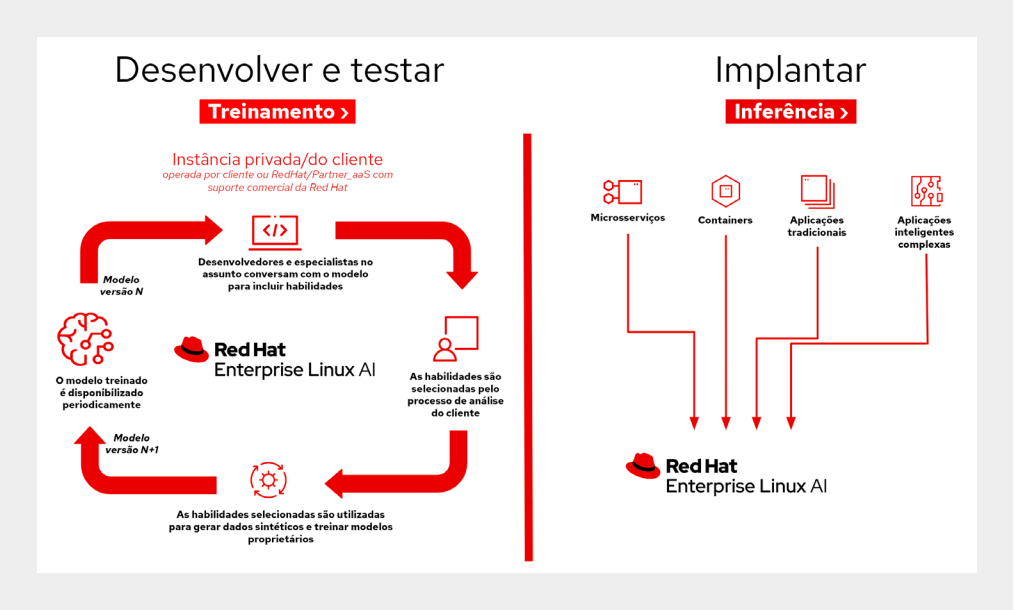


A imagem acima representa o fluxo de trabalho para ajuste fino de modelos do InstructLab.

1. Contribuições de habilidades e conhecimento são alocadas em um repositório de dados baseado em taxonomia.
2. Com esses dados de taxonomia, é gerada uma quantidade significativa de dados sintéticos para produzir um conjunto de dados grande o suficiente para atualizar e alterar um LLM.
3. Os dados sintéticos gerados são analisados, validados e lapidados por um modelo crítico.
4. O modelo é treinado com dados sintéticos baseados em entradas geradas manualmente.

Red Hat
Enterprise Linux AI

O InstructLab é acessível para desenvolvedores e especialistas que às vezes não têm a expertise em ciência de dados necessária para realizar ajustes finos em LLMs. Com a metodologia do InstructLab, as equipes adicionam dados (ou habilidades particularmente adequadas às necessidades do caso de uso empresarial) ao modelo de treinamento escolhido, colaborando entre si e possibilitando um time to value (TTV) mais rápido.



Treine e implante onde estiver

O Red Hat Enterprise Linux AI ajuda organizações a acelerar o processo de migração, da prova de conceitos a implantações baseadas em servidor de produção, oferecendo todas as ferramentas necessárias e as habilidades para treinar, ajustar e implantar esses modelos onde os dados estão, em qualquer lugar da nuvem híbrida. Assim, os modelos implantados podem ser usados por vários serviços e aplicações na empresa.

Com as organizações prontas, o Red Hat Enterprise Linux AI também fornece acesso ao Red Hat OpenShift® AI, com treinamento, ajuste e disponibilização desses modelos em escala em um ambiente de clusters distribuídos adotando os mesmos modelos Granite e abordagem do InstructLab usados na implantação do Red Hat Enterprise Linux AI.

Funcionalidades e benefícios

Funcionalidades	Benefícios
Modelos de código e linguagem Granite com suporte completo e open source sob licença do Apache 2.0.	LLMs open source e transparentes, com dados de treinamento amplamente acessíveis, melhoram a transparência de dados e abordam preocupações éticas sobre conteúdos e fontes de dados, reduzindo, por fim, riscos gerais de negócios.

Red Hat Enterprise Linux AI

Datasheet

Funcionalidades	Benefícios
Indenização de PI dos modelos Granite	A indenização pelos modelos Granite no Red Hat Enterprise Linux AI reflete a forte confiança que a Red Hat e a IBM têm no rigor de desenvolvimento e teste desses modelos. Essa indenização oferece aos clientes mais segurança e capacitação para usar os modelos Granite, aumentando a confiança no compromisso da Red Hat com o sucesso.
Ferramentas de alinhamento do LLM do InstructLab para ajuste fino de modelos acessível e escalável	O InstructLab disponibiliza um método acessível de ajuste fino dos LLMs, reduzindo a necessidade de know-how profundo em ciência de dados e viabilizando a contribuição de várias funções dentro da empresa. Isso possibilita a adoção de gen AI pela sua empresa, acelerando o time to value e maximizando seu retorno sobre o investimento.
Instâncias de runtime de modelo otimizadas e inicializáveis	O Red Hat Enterprise Linux AI é disponibilizado como uma imagem de container inicializável, um método de implantação chamado modo de imagem no Red Hat Enterprise Linux. Essa tecnologia reduz a complexidade de instalação, configuração e atualização, simplificando os processos de gerenciamento de configuração e alteração.
Dependências de pacotes de gen AI e drivers de softwares para hardware de IA	Comece agora mesmo a usar a gen AI com um conjunto de ferramentas abrangente, incluindo pacotes e drivers essenciais, como os drivers PyTorch, vLLM e NVIDIA. Assim, você se prepara para resolver os casos de uso empresariais de gen AI desde o início.

Sobre a Red Hat

A Red Hat é a líder mundial em soluções de software open source empresariais e utiliza uma abordagem impulsionada pela comunidade para oferecer tecnologias confiáveis e de alto desempenho em Linux, nuvem híbrida, containers e Kubernetes. A Red Hat ajuda os clientes a desenvolver aplicações nativas em nuvem, integrar aplicações de TI novas e existentes e automatizar e gerenciar ambientes complexos. Parceira de confiança das empresas da Fortune 500, a Red Hat oferece serviços de consultoria, treinamento e suporte premiados, compartilhando os benefícios da inovação open source com todos os setores. A Red Hat é um hub que conecta uma rede global de empresas, parceiros e comunidades, ajudando organizações a crescer, se transformar e se preparar para o futuro digital.

facebook.com/redhatinc
@redhatbr
linkedin.com/company/red-hat-brasil

AMÉRICA LATINA
+54 11 4329 7300
latammktg@redhat.com

BRASIL
+55 11 3629 6000
marketing-br@redhat.com

br.redhat.com
#1384734_0924