



SUSE



IA Generativa Open Source

Privacidade, segurança,
liberdade e comunidade

Sobre o palestrante

Meu LinkedIn



- Técnico em Informática (IFC)
- Engenheiro de Controle e Automação (UFSC)
- Especialista em Computação Científica para a Indústria (IFSC)
- Experiência como desenvolvedor e líder técnico; produtos de dados e IA
- Engenheiro de Software sênior na SUSE, trabalhando no time SUSE AI.

Sobre a palestra

Linguagem e objetivos

Assumiremos o papel de uma equipe de engenheiros de software. Recebemos o seguinte direcionamento da gestão:

“ Nós ouvimos que a IA Generativa pode trazer valor para nossa organização. Queremos que o time faça algo com IA Generativa para nossa empresa”

Faremos uma análise de como podemos atender ao pedido, criando algo útil e confiável.



IA, IA generativa e os jargões da indústria

Compreendendo o que está sendo pedido

“ IA Generativa é um conjunto de técnicas dentro do campo da IA, usadas para gerar algo novo: imagens, texto e até mesmo vídeo ”

- adaptado do livro “O que é IA Generativa” de Kyle Stratis

Ferramentas como ChatGPT, DeepSeek, Midjourney e similares. Dentro dessas aplicações, temos LLM's e variantes.

Modelos, LLM's e seu uso

Terminologia básica

Representações matemáticas de sistemas, resultados da soma de dados + métodos de treinamento + propósito.

LLM's são modelos de IA Generativa. Organizações podem:

- Criar LLMs (complexidade e custo elevados)
- Usar LLMs em aplicações (onde a maioria compete)

Gerando valor com LLM's

Onde competir, como organização

Como organização, nossa maior chance de competição é nas aplicações.

“Commoditização” de modelos:

- Construir um modelo do zero costuma ser inviável para a maioria
- Ao competir na linha de “aplicações” construímos valor “ao redor” do modelo

Revisando

Seguindo a solicitação

“ Nós ouvimos que a IA Generativa pode trazer valor para nossa organização. Queremos que o time faça algo com IA Generativa para nossa empresa”

- Sabemos o que é IA Generativa e o que são LLMs.
- Sabemos que é mais fácil competir no ramo das aplicações.

Que aplicação devemos criar? Ou ainda: como decidir qual aplicação criar?

Vamos aos dados

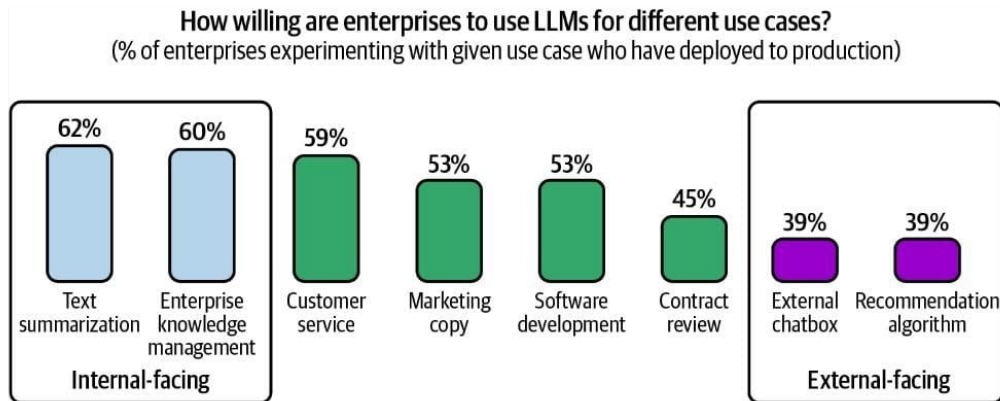
Onde usar LLMs

No livro “AI Engineering” (2025), a autora Chip Huyen mostra que empresas estão mais dispostas a usar IA generativa em aplicações internas.

62% em resumo de textos;

60% para gestão do conhecimento;

Vamos focar nesses casos de uso, então.



Nossa proposta

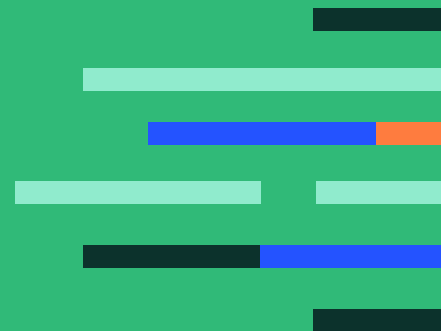
Com base no que aprendemos até então...

Para a demanda dos gestores, faremos a seguinte proposta:

“Vamos construir uma plataforma onde usuários corporativos possam tirar dúvidas sobre processos internos. A plataforma vai usar documentos existentes e IA generativa para responder aos questionamentos dos usuários.”



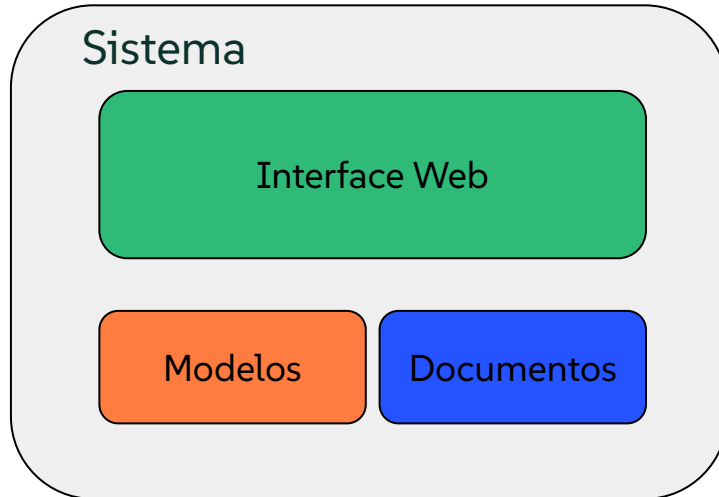
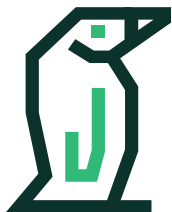
Componentes para nossa solução



Componentes para nossa solução

Nossos gestores aprovaram a ideia de nos deram liberdade para seguirmos em frente

Precisaremos de uma interface para os usuários, um modelo de IA generativa e uma forma de gerir nossos documentos.



Soluções comerciais

E preocupações

Existem várias plataformas de IA no mercado, mas podemos confiar integralmente em serviços de outras organizações?

→ Falta de transparência.

- Qual modelo está sendo executado? Quais políticas estão sendo adotadas nas confecções das respostas? Quais vieses estamos aceitando?

→ Privacidade e segurança da informação.

- Como é feita a gestão de acesso aos meus dados? Quais informações sobre minha organização são coletadas?

Soluções internas

Evitando a confiança cega

Como construiremos nossa solução?

- Desenvolver tudo do zero envolve um esforço enorme
- Numa perspectiva prática, não faz sentido reinventar a roda
- A comunidade de software livre já resolveu esses problemas



Open Source Software

Segurança, privacidade, liberdade e comunidade

- Problema: contratar plataformas que não explicam como operam (software proprietário).
 - Solução: software livre. Você tem o direito de saber como os programas que você usa funcionam.
- Problema: Esforço enorme para construir suas soluções do zero.
 - Solução: software livre. Você usa soluções que uma comunidade de desenvolvedores criou e disponibilizou.

Isso tudo, aliado a um senso de colaboração, possibilita que toda a indústria de software inove de forma acelerada.



Adesão ao Open Source



Our approach ▾

Research ▾

Product experiences ▾

Llama

Blog

Open source AI provides tools for more people to build awesome things and expand our horizons.

By giving more people access to immense opportunities and benefits, open source AI has the power to drive discoveries that enhance our lives and spur economic growth.

🦜 DeepSeek-R1 is now live and open source, rivaling OpenAI's Model o1. Available on web,

deepseek

Into the unknown

Google "We Have No Moat, And Neither Does OpenAI" // Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI

← Back to Articles

Open-R1: a fully open reproduction of DeepSeek-R1

Published January 28, 2025

[Update on GitHub](#)



[@eliebak](#)

Elie Bakouch



[@lvwerra](#)

Leandro von Werra



[@lewtun](#)

Lewis Tunstall



SUSE

Open Source

Como inovar em comunidade

O Open Source permite que o código possa ser lido, compreendido, modificado e redistribuído livremente. São várias licenças, com restrições diferentes.

Comunidades crescem ao redor do software livre.

IA Open Source implica em mais ferramentas e aplicações para desenvolvedores e para toda a comunidade.



Usando Open Source

A comunidade de software livre conta com componentes que nos atendem perfeitamente

UI para chatbots → Open WebUI

Gestão de LLMs → Ollama

LLMs → Open Weight, Open Source → Llama, Qwen, R1...

Documentos → RAG → Banco de dados vetorial → Milvus

Todos os softwares listados são livres, e você pode usá-los para o que for.

E ainda atendem a critérios rigorosos de segurança, privacidade e escalabilidade!



Milvus

Banco de Dados Vetorial



- Armazena informações como vetores
- Busca por similaridade excelente, basta observar vetores próximos

Usamos bancos de dados vetoriais para implementarmos uma RAG.

RAG provê mais conhecimento para LLM's, sem retreinamento.

Outra alternativa Open Source: Weaviate



Ollama

Gestão de LLMs

- Interface comum para modelos
- Ciclo de vida dos modelos
- Catálogo de modelos para download
- Sem interface, sem autenticação, “básico”

Outra alternativa Open Source: vLLM



Open WebUI

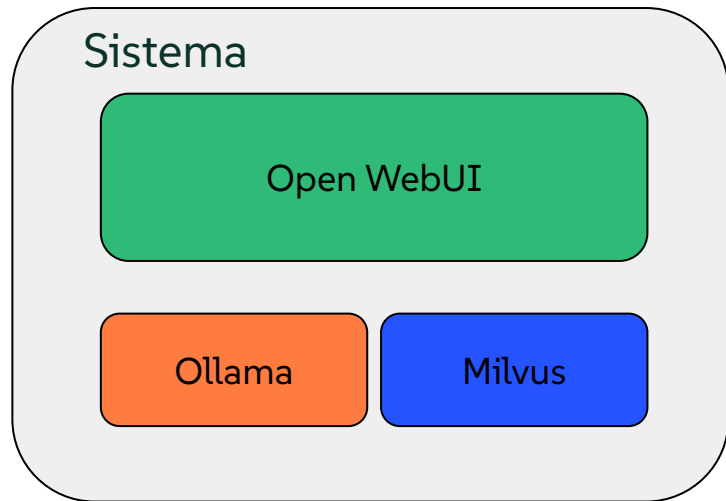
- Interface Web com gestão de chats, workspaces e usuários
- Permite utilizar RAGs
- Camada de controle de acesso
- Permite criar chats personalizados usando modelos e RAGs
- Extensível por meio de “pipelines”

Outra alternativa Open Source: GPT4All

oi



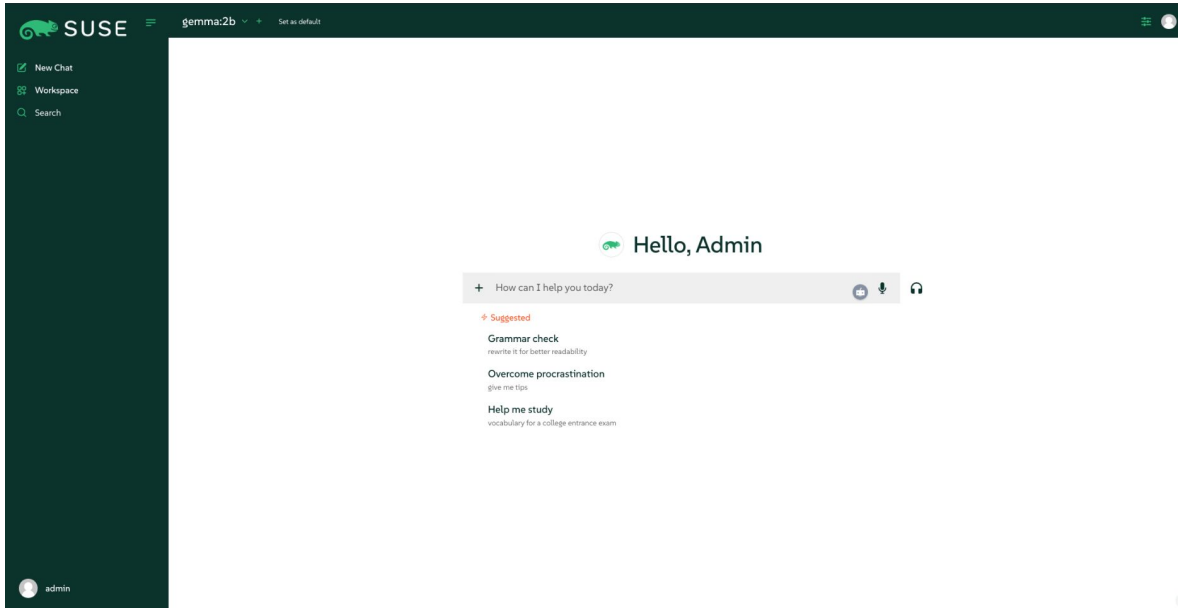
Open Source AI Stack



Você pode, agora mesmo, instalar todos esses componentes e começar a usar uma aplicação de IA generativa com todas as capacidades que discutimos até agora!

A performance é consideravelmente superior com aceleração de vídeo.

Demonstração



Navegando pela interface web do Open WebUI.

É simples criar uma base de conhecimentos para nossos modelos de IA generativa.

Preocupações adicionais

Tudo que vimos até aqui, porém em produção

- Escalabilidade (horizontal, vertical) Kubernetes, Helm Charts, Rancher
- Observabilidade (métricas, logs, custo etc) com OpenTelemetry
- Performance e aceleração com GPUs
- “Guardrails”, restrições para modelos
- Controle de autenticação e autorização integrados
- Técnicas de fine-tuning para LLMs
- ... e mais!



SUSE AI

Escalando nossa solução

Os componentes apresentados fazem parte da SUSE AI.

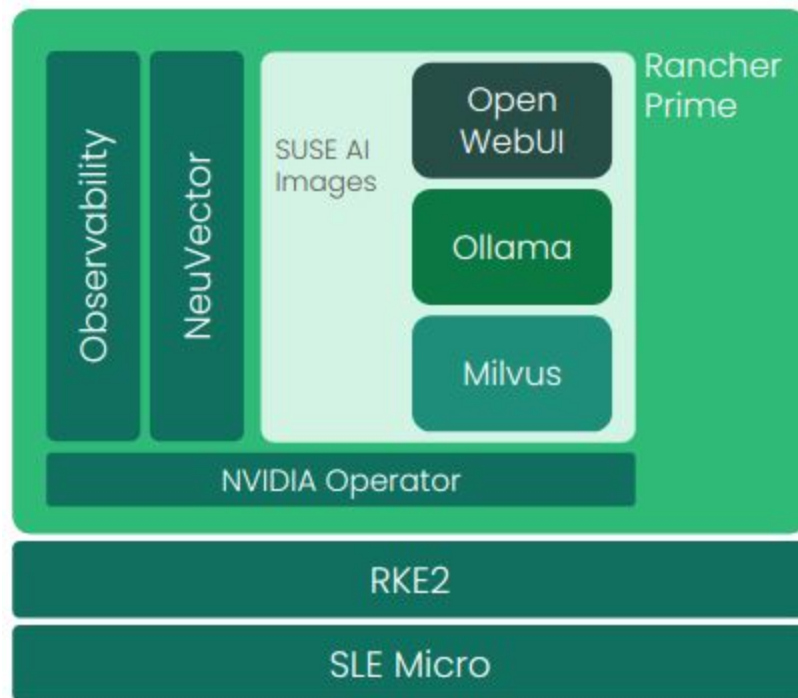
Nós disponibilizamos:

- Uma fundação sólida para você criar e escalar sua solução de IA na nuvem;
- Uma cadeia produtiva segura no software!
- Integrações com ferramentas de segurança e observabilidade;
- Suporte técnico e correções de vulnerabilidades.

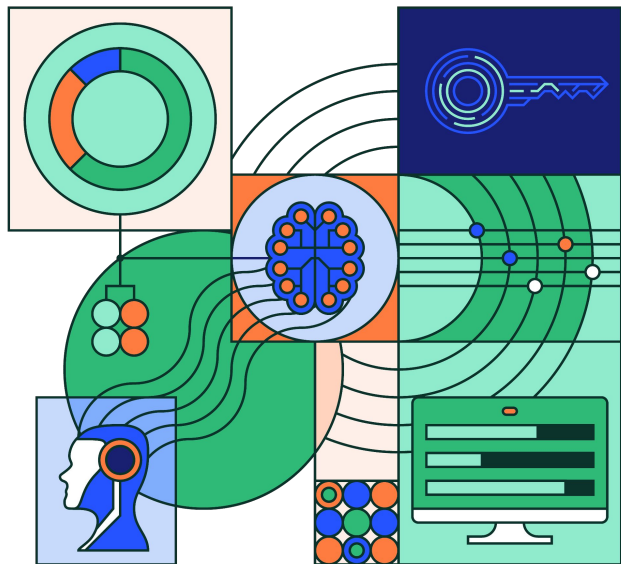
A SUSE AI é uma plataforma que acelera a adoção do código aberto!



SUSE AI



SUSE AI



<https://apps.rancher.io/stacks/suse-ai>



stacks / suse-ai



SUSE AI SUSE AI
Last updated an hour ago

SUSE AI provides an open infrastructure platform to deploy and run GenAI Workloads that provides choice in deployment options, choice in LLMs, and choice in AI tools and components. With SUSE AI, you get seamless deployment of sanitized AI tools and components. And because SUSE AI is secure by design, you get real-time threat tracking and full-stack observability across AI applications, as well as a FIPS validated common criteria certified platform that supports confidential computing.

Applications

Dependencies of this stack



Milvus
Last updated 18 hours ago

[Helm chart](#)

Milvus is a high-performance vector database built for scale. It is used by AI applications to organize and search through large amount of unstructured data, such as text, images, and multi-modal information.



Ollama
Last updated 8 hours ago

[Helm chart](#)

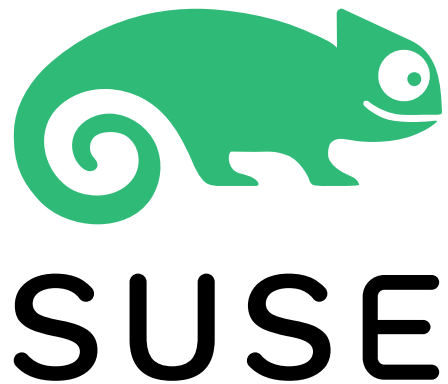
Get up and running with Llama 3.2, Mistral, Gemma 2, and other large language models.



Open WebUI
Last updated an hour ago

[Helm chart](#)

Open WebUI is an extensible, feature-rich, and user-friendly self-hosted WebUI designed to operate entirely offline. It supports various LLM runners, including Ollama and OpenAI-compatible APIs.



Estamos
contratando!





SUSE