Explore e implante Modelos de Catalogo do Azure AI Foundry

Selecione um modelo usando uma abordagem estruturada.

1. Como a IA pode resolver o meu caso?
   * Defina casos de uso
   * Selecione um modelo
   * Estabeleça a viabilidade
   * Construa um protótipo
2. Como seleciono o modelo “certo” para meu caso de uso?
   * Selecione um modelo
   * Engenharia de prompt
   * Retrieval Augmented Generation
   * Afinação
3. Posso escalar para cargas de trabalho no mundo real?
   * Filtragem de conteúdo
   * Segurança e privacidade de dados
   * Compensações de capacidade e custos
   * Monitoramento

Crie um protótipo para resolver o caso de uso

1. LLM’s e SLM’s
   1. Large Language Models:
      1. GPT-4
      2. Mistral Large
      3. Llama3 70b
      4. Llama 405B
      5. Command R,R+
   2. Small Language Models:
      1. Phi3
      2. Mistral OSS models
      3. Llama3 8b
2. Modalities, Tasks and Tools
   1. Multi-Model:
      1. GPT-4o, Phi3-vision
   2. Image Generation:
      1. Dalle3, Stability AI
   3. Embedding Models:
      1. Ada, Cohere
   4. Function Calling & JSON support
3. Regional and domain specific
   1. Core45 JAIS Arabic languagem LLM
   2. Mistral Large focused on Europena Languages
   3. Nixtla TimeGEN1 – Timeseries forecasting
4. Open and propriety
   1. Premium models frist on Azure:
      1. OpenAi, Mistral Large, Cohere Command R+
   2. 100s Open models from HuggingFace
   3. Open models from Meta, Datbricks, Nvidia

Avalie o desempenho do modelo

Avalie o desempenho do seu modelo em diferentes fases, usando diversas abordagens:

* Referencias de modelo – Compare métricas disponíveis publicamente entre modelos e conjunto s de dados.
* Avaliações manuais – Avalie as respostas do seu modelo
* Métricas tradicionais de aprendizado de máquina – Meça a proporção do número de palavras compartilhadas entre respostas geradas e as respostas verdadeiras.
* Métricas assistidas por IA – Métricas de risco e segurança;   
  Métricas de qualidade de geração

Entenda os benchmarks do modelo

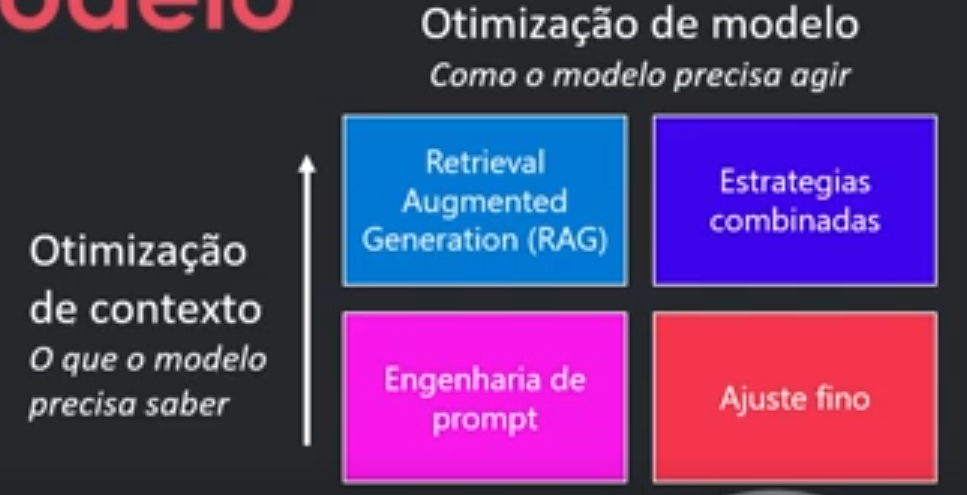
Os conjuntos de dados estão disponíveis publicamente para calcular benchmarks individuais e comparar modelos.

Alguns benchmarks comumente usado são:

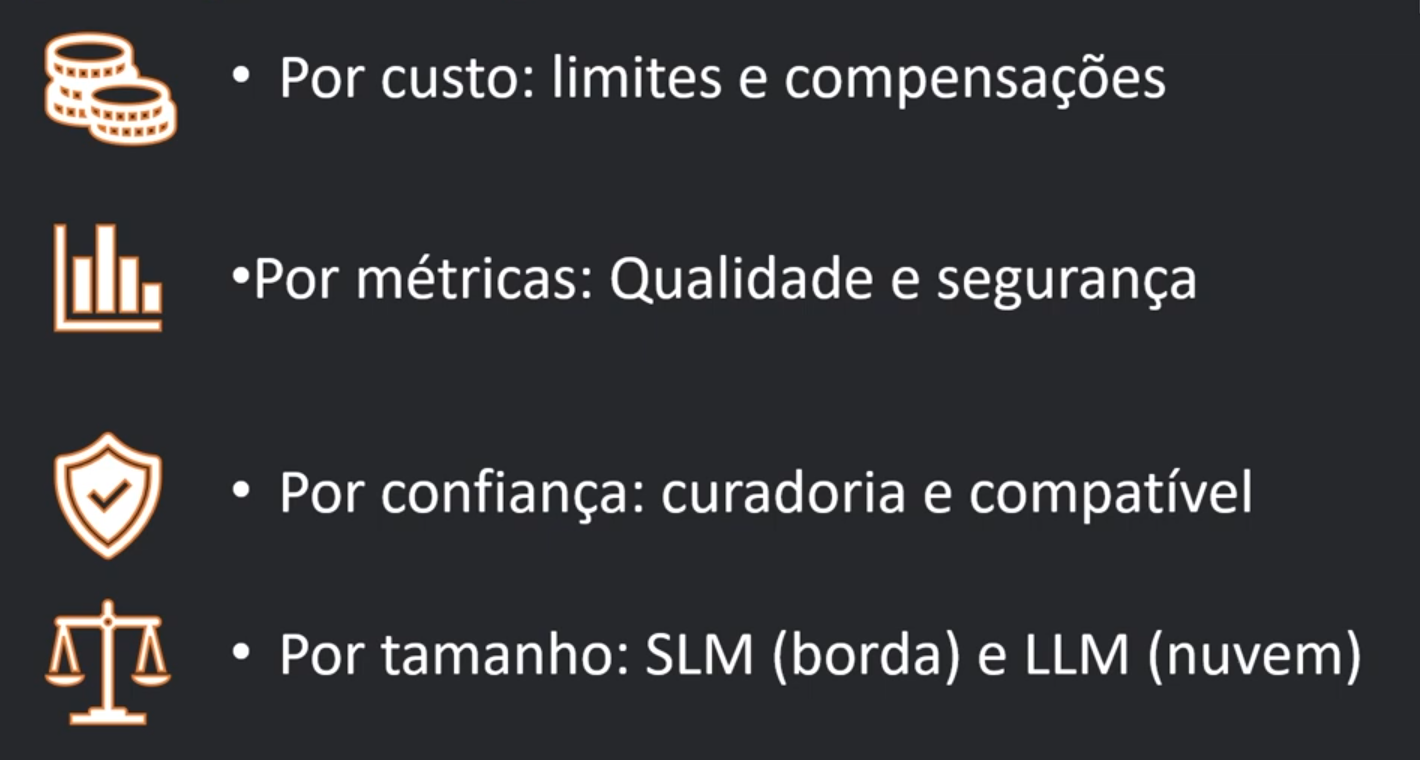
* **Precisão:** Compara o texto gerado pelo modelo com a resposta correta de acordo com o conjunto de dados. O resultado será um se o texto gerado corresponder exatamente à resposta e zero caso contrário.
* **Fluência:** Avalia até que ponto o texto gerado adere às regras gramaticais, estruturas sintáticas e uso apropriado do vocabulário, resultado em respostas linguisticamente corretas e com som natural.
* **Coerência :** Mede se a saída do modelo flui suavemente, se é lida naturalmente e se assemelha à linguagem humana.
* **Similaridade GPT:** Quantifica a semelhança semântica entre uma frase (ou documento) verdadeira a frase de previsão gerada por um modelo de IA

Otimize a precisão após selecionar um modelo

* Otimize o contexto quando o modelo não tiver conhecimento contextual e você quiser maximizar a precisão da resposta.
* Otimize o modelo quando quiser melhorar o formato, o estilho ou a fala da resposta, maximizando a consistência do comportamento.



Otimize o desempenho para se preparar para escala

Materiais de Apoio

Os materiais complementares e de apoio que oferecemos têm como objetivo fornecer informações para facilitar e enriquecer a sua jornada de aprendizado no curso "Explore e implante modelos do catálogo de modelos". Aqui você encontrará links úteis, como slides, repositórios e páginas oficiais, além de dicas sobre como se destacar na DIO e no mercado de trabalho 😉

Recursos Adicionais

Durante este conteúdo, compreendemos os fundamentos da engenharia de prompts. Para ajudá-lo a aprofundar o conhecimento, disponibilizamos a seguir o material complementar contendo os conteúdos e links apresentados no curso:

Slide: Módulo 7 - Otimizando Modelos de Linguagem para Aplicações Generativas de IA.pptx

Dicas e Links Úteis

Para se desenvolver ainda mais e se destacar na DIO e no mercado de trabalho, sugerimos os seguintes recursos:

Artigos e Fórum da DIO: Compartilhe seus conhecimentos e dúvidas através dos artigos (visíveis globalmente na plataforma da DIO) e nos fóruns específicos para cada experiência educacional, como nossos Bootcamps.

Rooms: Participe do Rooms, uma ferramenta de bate-papo em tempo real onde você pode interagir com outros participantes dos nossos Bootcamps, compartilhando dúvidas, dicas e snippets de código.

Exploração na Web: Utilize motores de busca para aprofundar seu conhecimento sobre temas específicos. Páginas como o StackOverflow são recursos valiosos para encontrar soluções e expandir seu entendimento.

Com esses materiais complementares, você estará bem equipado para explorar todo o potencial e se destacar em suas iniciativas. Continue aproveitando as oportunidades de aprendizado, e não hesite em buscar mais conhecimento e compartilhar suas descobertas com a comunidade!

