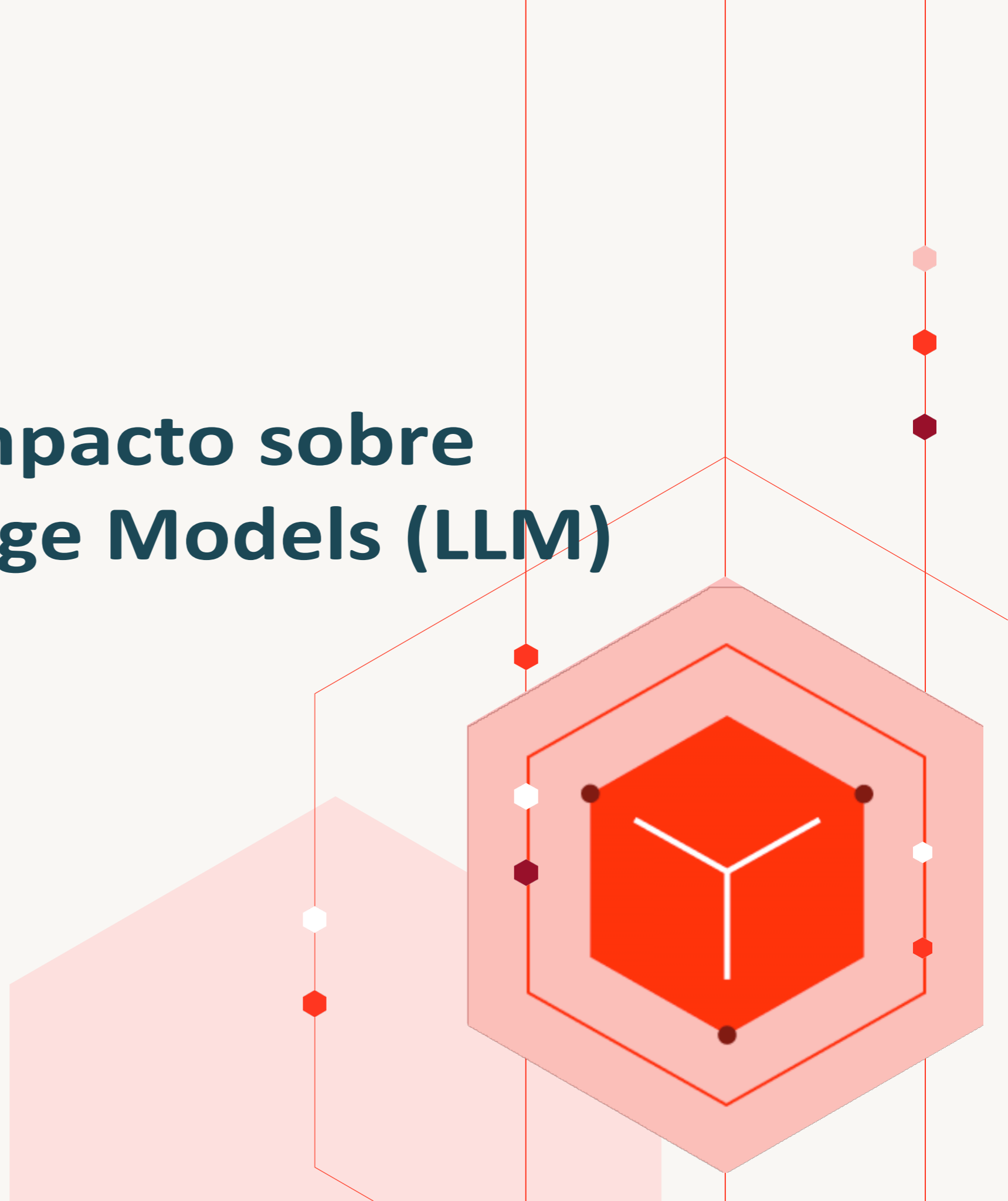


E-BOOK

Um guia compacto sobre Large Language Models (LLM)



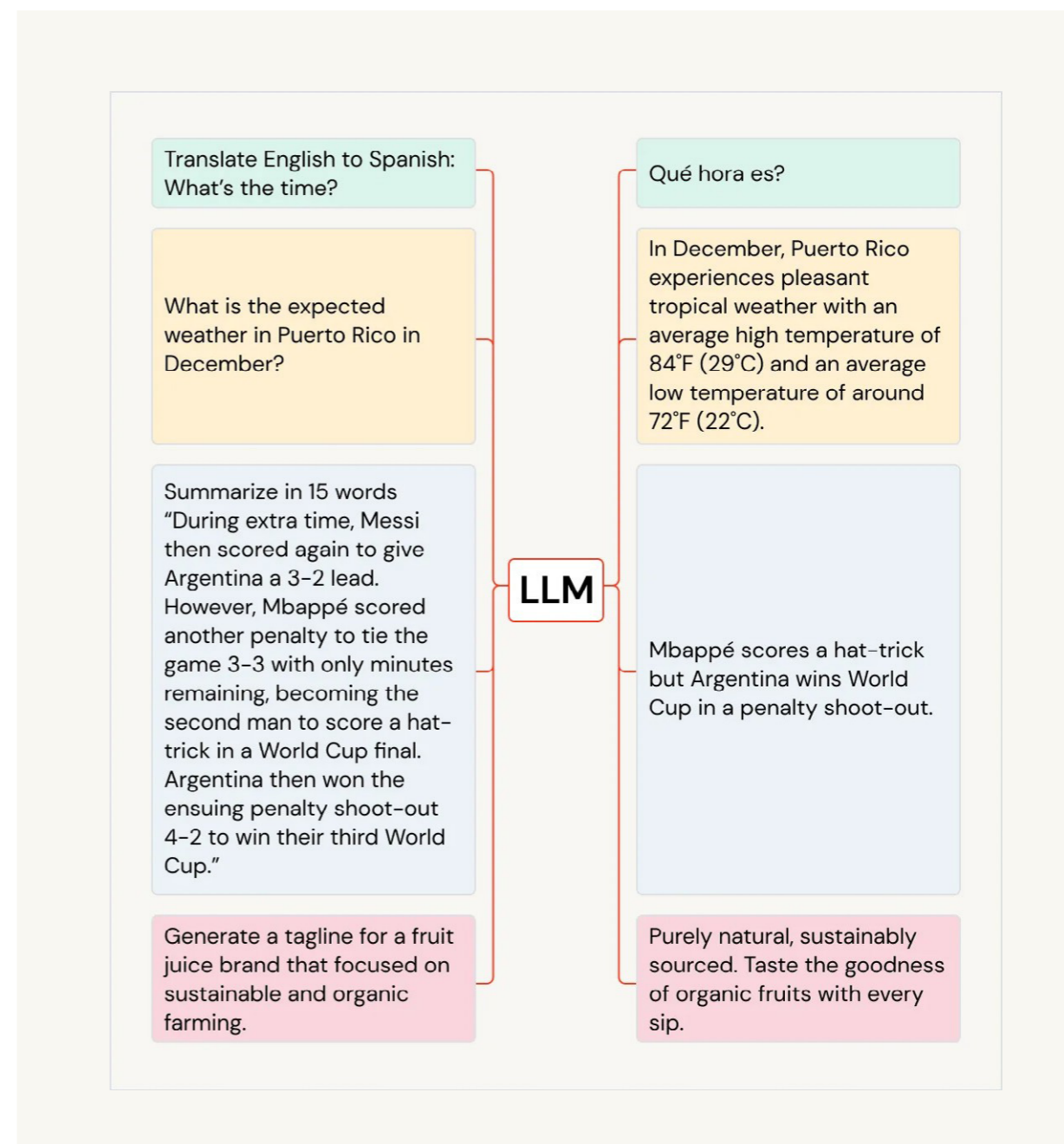
PARTE 1

Introdução

Definição de LLM (tradução livre: grandes modelos de linguagem)

LLMs são sistemas de IA desenvolvidos para processar e analisar enormes quantidades de dados de linguagem natural e, em seguida, usar essas informações para gerar respostas às solicitações dos usuários. Esses sistemas são treinados em grandes conjuntos de dados usando algoritmos avançados de machine learning para aprender os padrões e as estruturas da linguagem humana e são capazes de gerar respostas em linguagem natural a uma ampla variedade de contribuições escritas. Os grandes modelos de linguagem estão se tornando cada vez mais importantes em uma variedade de aplicativos, como processamento de linguagem natural, tradução automática, geração de código e texto e muito mais.

Embora este guia tenha como foco os modelos de linguagem, é fundamental compreender que eles representam apenas um elemento dentro do vasto espectro da IA generativa. Outras implementações notáveis de IA generativa incluem projetos como a geração de arte a partir de texto, áudio e vídeo, e certamente muitas outras novidades surgirão em breve.



Breve histórico e resumo do desenvolvimento dos LLMs

Décadas de 1950 a 1990

Foram feitas tentativas iniciais para criar regras rígidas para as linguagens e seguir passos lógicos para realizar tarefas como traduzir frases de um idioma para outro.

Embora esse método funcionasse em alguns casos, estava limitado a tarefas bem definidas das quais o sistema tinha conhecimento.

Década de 1990

Os modelos de linguagem começaram a evoluir para modelos estatísticos, e os padrões linguísticos começaram a ser analisados, mas projetos em larga escala eram limitados pela capacidade de processamento de dados.

Anos 2000

Os avanços em machine learning aumentaram a complexidade dos modelos de linguagem, e a ampla adoção da internet forneceu uma grande quantidade de dados de treinamento.

2012

Os avanços em arquiteturas de deep learning e conjuntos de dados maiores levaram ao desenvolvimento do GPT (Transformadores Pré-treinados Generativos).

2018

O Google apresentou o BERT (Bidirectional Encoder Representations from Transformers), que foi um grande salto na arquitetura e abriu caminho para futuros grandes modelos de linguagem.

2020

A OpenAI lançou o GPT-3, que se tornou o maior modelo com 175 bilhões de parâmetros e estabeleceu um novo referencial de desempenho para tarefas relacionadas à linguagem.

2022

O ChatGPT foi lançado, transformando o GPT-3 e modelos semelhantes em um serviço amplamente acessível aos usuários por meio de uma interface web, o que iniciou um aumento significativo na conscientização pública sobre LLMs e IA generativa.

2023

Os LLMs de código aberto começam a apresentar resultados cada vez mais impressionantes, com lançamentos como Dolly 2.0, LLaMA, Alpaca e Vicuna.

O GPT-4 também é lançado, estabelecendo um novo referencial tanto em tamanho de parâmetros quanto em desempenho.

PARTE 2

Compreendendo os grandes modelos de linguagem (LLMs)

O que são modelos de linguagem e como eles funcionam?

Os grandes modelos de linguagem são sistemas avançados de inteligência artificial que recebem entradas e geram respostas semelhantes às de seres humanos em forma de texto. Eles funcionam primeiro analisando enormes quantidades de dados e criando uma estrutura interna que modela os conjuntos de dados de linguagem natural nos quais foram treinados. Uma vez que essa estrutura interna tenha sido desenvolvida, os modelos podem receber entradas na forma de linguagem natural e produzir uma resposta adequada.

Se eles existem há tantos anos, por que só agora estão ganhando destaque?

Alguns avanços recentes trouxeram grande destaque à IA generativa e aos grandes modelos de linguagem:

▶ AVANÇOS EM TÉCNICAS

Nos últimos anos, houve avanços significativos nas técnicas usadas para treinar esses modelos, resultando em grandes melhorias de desempenho.

Notavelmente, um dos maiores saltos de desempenho veio da integração do feedback humano diretamente no processo de treinamento.

▶ MAIOR ACESSIBILIDADE

O lançamento do ChatGPT abriu as portas para qualquer pessoa com acesso à internet interagir com um dos LLMs mais avançados por meio de uma interface web simples. Isso trouxe os impressionantes avanços dos LLMs para o centro das atenções, uma vez que anteriormente esses modelos mais poderosos estavam disponíveis apenas para pesquisadores com recursos significativos e conhecimento técnico profundo.

▶ AUMENTO DA POTÊNCIA COMPUTACIONAL

A disponibilidade de recursos de computação mais poderosos, como unidades de processamento gráfico (GPUs), e melhores técnicas de processamento de dados permitiu que os pesquisadores treinassem modelos muito maiores, melhorando o desempenho desses modelos de linguagem.

▶ MELHORIA DOS DADOS DE TREINAMENTO

À medida que progredimos na coleta e análise de grandes volumes de dados, o desempenho dos modelos melhorou drasticamente. Na verdade, a Databricks demonstrou que é possível obter resultados incríveis treinando um modelo relativamente pequeno com um conjunto de dados de alta qualidade com o **Dolly 2.0** (e também lançamos o conjunto de dados com o conjunto de dados databricks-dolly-15k <http://databricks/databricks-dolly-15k>).

Então, para que as organizações estão usando grandes modelos de linguagem?

Aqui estão apenas alguns exemplos de casos de uso comuns para grandes modelos de linguagem:

▶ CHATBOTS E ASSISTENTES VIRTUAIS

Uma das implementações mais comuns, os LLMs podem ser usados por organizações para fornecer ajuda em tarefas como suporte ao cliente, solução de problemas ou até mesmo para ter conversas abertas com prompts fornecidos pelo usuário.

▶ GERAÇÃO DE CÓDIGO E DEPURAÇÃO

Os LLMs podem ser treinados com grandes volumes de exemplos de código e fornecer trechos úteis de código como resposta a solicitações escritas em linguagem natural. Com as técnicas apropriadas, os LLMs também podem ser desenvolvidos de forma a fazer referência a outros dados relevantes que talvez não tenham sido treinados, como a documentação de uma empresa, para fornecer respostas mais precisas.

▶ ANÁLISE DE SENTIMENTO

Frequentemente, uma tarefa difícil de quantificar, os LLMs podem ajudar a analisar emoções e opiniões a partir de um texto. Isso pode ajudar as organizações a coletarem os dados e o feedback necessários para melhorar a satisfação dos clientes.

▶ CLASSIFICAÇÃO E AGRUPAMENTO DE TEXTO

A capacidade de categorizar e classificar grandes volumes de dados permite a identificação de temas e tendências comuns, apoiando a tomada de decisões informadas e estratégias mais direcionadas.

▶ TRADUÇÃO DE IDIOMAS

Globalize todo o seu conteúdo sem horas de trabalho árduo simplesmente alimentando suas páginas da web por meio dos LLMs apropriados e traduzindo-os para diferentes idiomas. À medida que mais LLMs são treinados em outros idiomas, a qualidade e a disponibilidade continuarão melhorando.

▶ RESUMO E PARAFRASEAMENTO

Chamadas ou reuniões de clientes completas podem ser resumidas de forma eficiente para que outras pessoas possam digerir o conteúdo mais facilmente. Os LLMs podem pegar grandes volumes de texto e resumir apenas os bytes mais importantes.

▶ GERAÇÃO DE CONTEÚDO

Comece com um prompt detalhado e deixe um LLM desenvolver um esboço para você. Em seguida, continue com esses prompts e os LLMs podem gerar um primeiro rascunho para você desenvolver. Use-os para criar ideias e faça perguntas ao LLM para ajudar a se inspirar.

Observação: a maioria dos LLMs *não* é treinada para ser uma máquina de fatos. Eles sabem como usar a linguagem, mas podem não saber quem ganhou o grande evento esportivo do ano passado. É sempre importante verificar os fatos e entender as respostas antes de usá-las como referência.

PARTE 3

Aplicação de grandes modelos de linguagem

Existem alguns caminhos que você pode seguir ao procurar aplicar grandes modelos de linguagem para seu caso de uso específico. Em termos gerais, você pode dividi-los em duas categorias, mas há alguma sobreposição entre elas. Vamos abordar brevemente as vantagens e desvantagens de cada uma e em quais cenários cada uma se encaixa melhor.

Serviços proprietários

Como o primeiro serviço amplamente disponível alimentado por LLM, o ChatGPT da OpenAI foi o catalisador explosivo que trouxe os LLMs para o mainstream. O ChatGPT fornece uma interface de usuário (ou API) em que os usuários podem enviar prompts para muitos modelos (GPT-3.5, GPT-4 e outros) e geralmente obter uma resposta rápida. Eles estão entre os modelos de maior desempenho, treinados em conjuntos de dados enormes, e são capazes de realizar tarefas extremamente complexas tanto do ponto de vista técnico, como geração de código, quanto do ponto de vista criativo, como escrever poesia em um estilo específico.

A desvantagem desses serviços é a quantidade absolutamente enorme de recursos computacionais necessários não apenas para treiná-los (a OpenAI afirmou que o GPT-4 custou mais de US\$ 100 milhões para desenvolver), mas também para fornecer as respostas. Por esse motivo, esses modelos extremamente grandes provavelmente sempre estarão sob o controle de organizações

e exigirão que você envie seus dados para seus servidores a fim de interagir com seus modelos de linguagem. Isso levanta preocupações com privacidade e segurança, e também sujeita os usuários a modelos “caixa preta”, sobre cujos treinamentos e limites eles não têm controle. Além disso, devido aos recursos computacionais necessários, esses serviços não são gratuitos além de um uso muito limitado, então o custo se torna um fator ao aplicá-los em grande escala.

Resumindo: serviços proprietários são ótimos para usar se você tiver tarefas muito complexas, tiver disposição para compartilhar seus dados com terceiros e quiser incorrer em custos ao operar em escala significativa.

Modelos de código aberto

A outra opção para modelos de linguagem é recorrer à comunidade de código aberto, onde houve um crescimento igualmente explosivo nos últimos anos. Comunidades como a [Hugging Face](#) reúnem centenas de milhares de modelos de contribuidores que podem ajudar a resolver muitos casos de uso específicos, como geração de texto, resumo e classificação. A comunidade de código aberto está rapidamente alcançando o desempenho dos modelos proprietários, mas ainda não conseguiu igualar o desempenho de algo como o GPT-4.

Atualmente, requer um pouco mais de esforço para pegar um modelo de código aberto e começar a usá-lo, mas o progresso está ocorrendo muito rapidamente para torná-los mais acessíveis aos usuários. Na Databricks, por exemplo, fizemos **melhorias em frameworks de código aberto** como o MLflow para tornar muito fácil para alguém com um pouco de experiência em Python pegar qualquer modelo transformador da Hugging Face e usá-lo como um objeto Python. Muitas vezes, você pode encontrar um modelo de código aberto que resolve seu problema específico e que é várias **ordens de grandeza** menor que o ChatGPT, permitindo que você traga o modelo para seu ambiente e hospede-o você mesmo. Isso significa que você pode manter os dados sob seu controle para preocupações com privacidade e governança, além de gerenciar seus custos.

Outra grande vantagem de usar modelos de código aberto é a capacidade de ajustá-los aos seus próprios dados. Como você não está lidando com uma caixa preta de um serviço proprietário, existem técnicas que permitem pegar modelos de código aberto e treiná-los com seus dados específicos, melhorando significativamente o desempenho deles em seu domínio específico. Acreditamos que o futuro dos modelos de linguagem seguirá nessa direção, à medida que mais organizações desejem ter controle total e compreensão de seus LLMs.

Conclusão e diretrizes gerais

Em última análise, cada organização terá desafios únicos a superar, e não existe uma abordagem única para os LLMs. À medida que o mundo se torna mais orientado a dados, tudo, incluindo os LLMs, dependerá de uma base sólida de dados. Os LLMs são ferramentas incríveis, mas devem ser usados e implementados sobre essa base sólida de dados. A Databricks oferece tanto essa base sólida de dados quanto as ferramentas integradas para permitir que você use e ajuste os LLMs no seu domínio.

PARTE 4

E agora, o que fazer se eu quiser começar a usar LLMs?

Isso depende de onde você está em sua jornada. Felizmente, temos algumas opções para você.


Se você deseja se aprofundar um pouco mais nos LLMs, mas ainda não quer fazer isso por conta própria, pode assistir a uma das apresentações sob demanda de um dos desenvolvedores e palestrantes mais talentosos da Databricks sobre esses conceitos em mais detalhes, durante a palestra “[Crie seu próprio grande modelo de linguagem como Dolly](#)”.

Se você quiser se aprofundar um pouco mais e expandir seus conhecimentos e compreensão dos fundamentos dos LLMs, recomendamos conferir nosso [curso sobre LLMs](#). Você aprenderá como desenvolver aplicativos prontos para produção com LLMs e se aprofundará na teoria por trás dos modelos de fundação.

Se suas mãos já estão tremendo de emoção e você já tem algum conhecimento prático de Python e Databricks, forneceremos alguns ótimos exemplos com código de exemplo que podem ajudar você a começar a trabalhar com LLMs imediatamente.



Introdução à PNL usando pipelines transformadores do Hugging Face



Ajuste de grandes modelos de linguagem com Hugging Face e DeepSpeed



Apresentando funções de IA: Integração de LLMs com o Databricks SQL



Sobre a Databricks

A Databricks é a empresa de dados e IA. Mais de 9.000 organizações em todo o mundo, incluindo a Comcast, Condé Nast e mais de 50% da Fortune 500, contam com a Plataforma Databricks Lakehouse para unificar seus dados, análises e IA. A Databricks tem sede em São Francisco, com escritórios em todo o mundo.

Fundada pelos criadores originais do Apache Spark™, Delta Lake e MLflow, a Databricks tem como missão ajudar as equipes de dados a resolver os problemas mais difíceis do mundo. Para saber mais, siga a Databricks no [Twitter](#), [LinkedIn](#) e [Facebook](#).

EXPERIMENTE GRÁTIS

Entre em contato conosco para ver uma demonstração:
databricks.com/contact

