**IA generativa**

**IA generativa**

Introdução

Imagine a agitação em torno de um recente leilão de arte onde uma pintura, de autoria desconhecida, foi vendida por uma quantia exorbitante. Essa obra-prima foi gerada por inteligência artificial, desafiando nossas percepções de criatividade e o papel da tecnologia nas artes. Eventos como esse destacam o poder transformador da IA ​​Generativa, que está remodelando setores da arte ao desenvolvimento de software.

No cenário tecnológico atual em rápida evolução, entender a IA Generativa não é apenas um trunfo; é uma necessidade, e a capacidade de aproveitar a IA Generativa de forma eficaz abrirá um novo mundo de aplicações e soluções inovadoras.

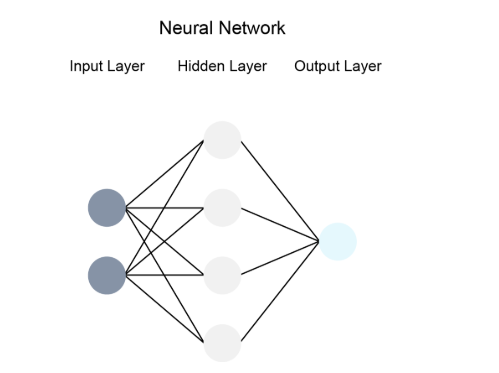
Ao final deste treinamento, você será capaz de:

* Descreva a capacidade da IA ​​generativa de criar conteúdo usando aprendizado de máquina.
* Identifique dois casos de uso do GenAI em desenvolvimento de software e automação de design.
* Liste os desafios na inferência de IA generativa relacionados a dados e diversidade.
* Liste os desafios e limitações da IA ​​generativa, incluindo dados e ética.
* Explique como o preconceito e a imparcialidade no GenAI podem ser abordados para garantir o desempenho ético da IA.
* Liste três técnicas para otimizar recursos de hardware e energia no GenAI.
* Compare a IA generativa e a IA tradicional em termos de abordagens e desafios.
* Compare os requisitos de dados para sistemas de IA tradicional e IA generativa.
* Liste cinco tendências emergentes em Inteligência Artificial (IA).
* Descreva os principais conceitos e tipos de modelos de linguagem de IA.
* Diferencie entre modelos de linguagem grandes e pequenos em aplicações de IA.

**IA generativa**

A IA Generativa, ou GenAI, é capaz de criar novos conteúdos, como texto, imagens, música ou código, aprendendo padrões a partir de dados existentes. Ao contrário da IA ​​tradicional, que se baseia em regras explicitamente programadas, a GenAI utiliza modelos de aprendizado de máquina para compreender e reproduzir os padrões nos dados. Por meio desse processo, a GenAI pode produzir resultados inovadores que não são explicitamente programados, tornando-os valiosos para aplicações que exigem criatividade e variabilidade.

Os sistemas de IA generativa são baseados em redes neurais artificiais (RNAs), que são sistemas computacionais inspirados no cérebro humano. Uma RNA consiste em nós interconectados (neurônios) organizados em camadas: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Pense em um neurônio como um objeto com um valor entre zero e um. Esse valor representa a ativação do neurônio e, quanto maior o número, mais fortemente ele influencia a saída. Para muitos modelos de IA Genética, o valor de entrada é um prompt do usuário. Esses dados são convertidos em valores numéricos que serão processados ​​pelas camadas de neurônios.



A camada de entrada de uma rede neural feedforward é composta por neurônios representados por x1, x2, …xn. Cada um recebe uma característica ou atributo dos dados de entrada. Em uma tarefa como a geração de texto, cada um desses neurônios pode corresponder a uma palavra ou token específico de um prompt.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

As camadas ocultas nessa rede realizam os cálculos principais. Cada neurônio em uma camada oculta recebe entradas dos neurônios da camada anterior e as processa usando uma soma ponderada. Uma soma ponderada é a soma de todas as entradas de um neurônio. Cada entrada é multiplicada por seu peso correspondente, mais um viés (um valor calculado durante o treinamento que ajusta a saída de um neurônio independentemente dos valores de entrada). O próximo passo é uma função de ativação, uma função que determina a saída de um neurônio com base em sua entrada e, em seguida, passa a saída para a próxima camada. Essas camadas ocultas são onde a rede aprende a detectar e abstrair padrões dos dados. Múltiplas camadas ocultas permitem que a rede construa representações cada vez mais complexas dos dados de entrada. Por exemplo, na criação de imagens, as camadas iniciais podem capturar formas ou texturas básicas, enquanto camadas mais profundas capturam recursos mais abstratos, como estilo ou composição.

A camada de saída em uma rede neural feedforward é responsável por gerar as saídas finais, denotadas por y1, y2, …ym. Em modelos GenAI, essa camada gera o conteúdo solicitado pelo usuário. Considerando novamente o exemplo de geração de texto, os neurônios da camada de saída conteriam palavras ou tokens produzidos pela alimentação do prompt de entrada pela rede.

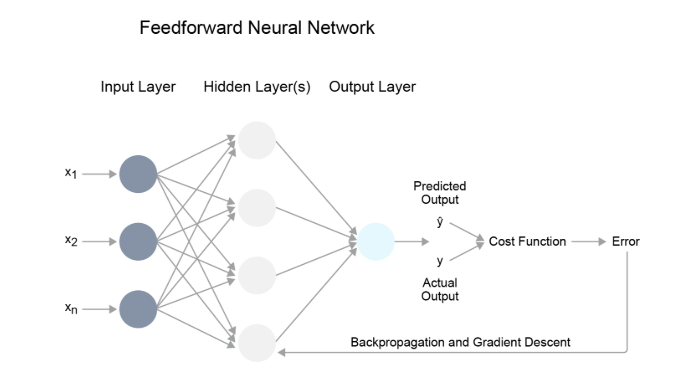
As conexões entre os neurônios têm pesos que são ajustados durante o processo de treinamento. O treinamento de uma rede neural feedforward envolve várias etapas importantes para minimizar erros nas previsões da rede e melhorar seu desempenho.

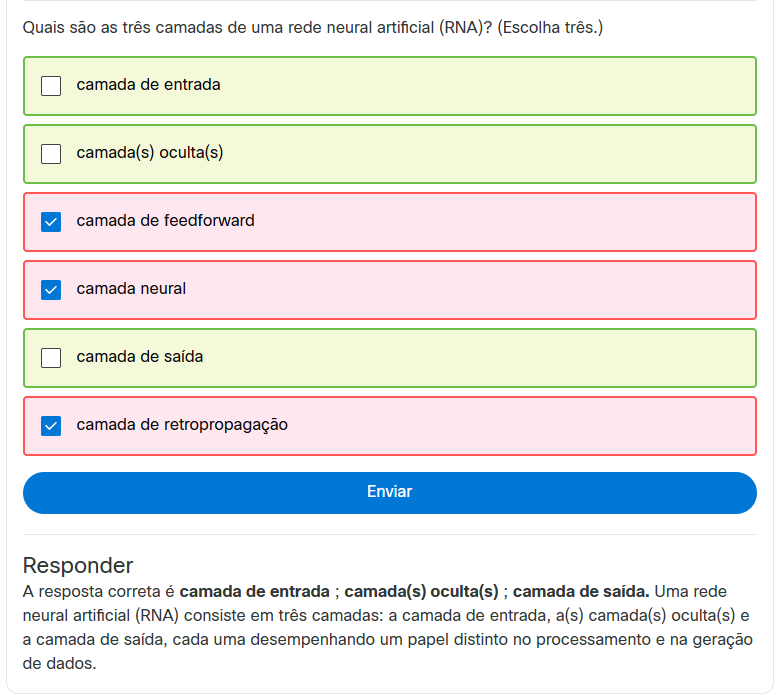
Diagrama, Gráfico de radar

O conteúdo gerado por IA pode estar incorreto.

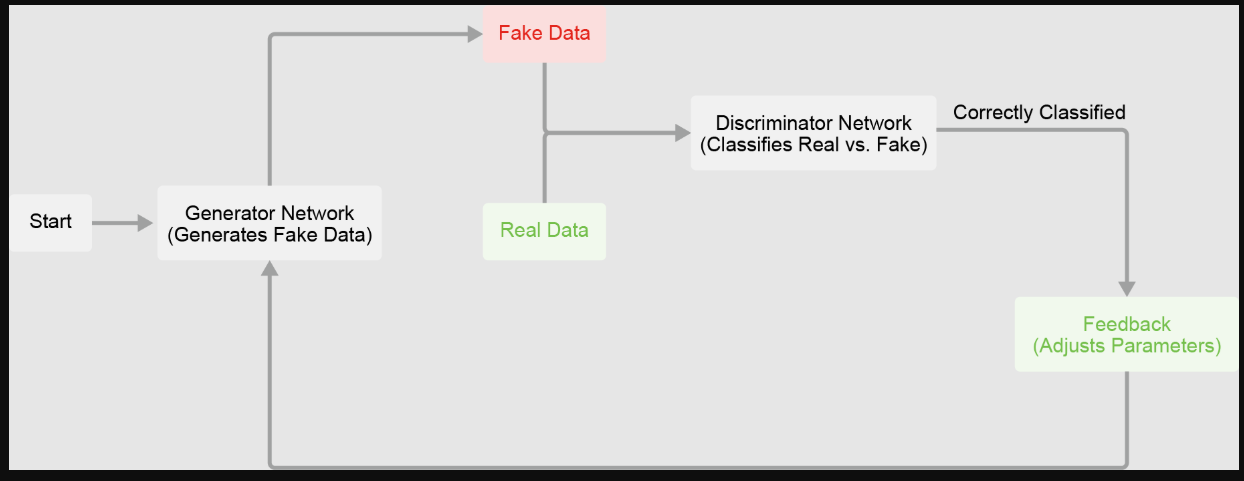
Inicialmente, os dados de entrada são passados ​​pelas camadas da rede em uma passagem direta, onde cada camada processa a entrada e a passa para a próxima camada. Assim que a rede produz uma saída, a diferença entre essa saída prevista e a saída alvo real é calculada, resultando em um erro. Esse erro é medido usando uma função de custo, que quantifica a discrepância entre as saídas previstas e reais.

Os pesos da rede são ajustados finamente usando retropropagação. Esse processo envolve a propagação do erro, calculado como a diferença entre a saída prevista e a saída real, para trás, através da rede. O fluxo de informações em reverso permite que o algoritmo de treinamento determine quanta mudança é necessária para cada peso, de modo que possa reduzir esse erro ou gradiente. A descida do gradiente é uma técnica de otimização que permite que cada peso seja ajustado calculando seu grau de mudança. A magnitude de cada ajuste é determinada pela taxa de aprendizado, um parâmetro que controla o tamanho dos passos dados em direção ao erro mínimo. Após o cálculo desses gradientes, os pesos são reajustados.





Estruturas Adversariais Generativas



Uma Rede Adversarial Generativa (GAN) é uma estrutura de ML para treinar modelos GenAI. GANs consistem em duas redes neurais: um gerador e um discriminador. O gerador cria dados falsos que visam se assemelhar a dados reais, enquanto o discriminador avalia a autenticidade dos dados, distinguindo entre amostras reais e falsas.

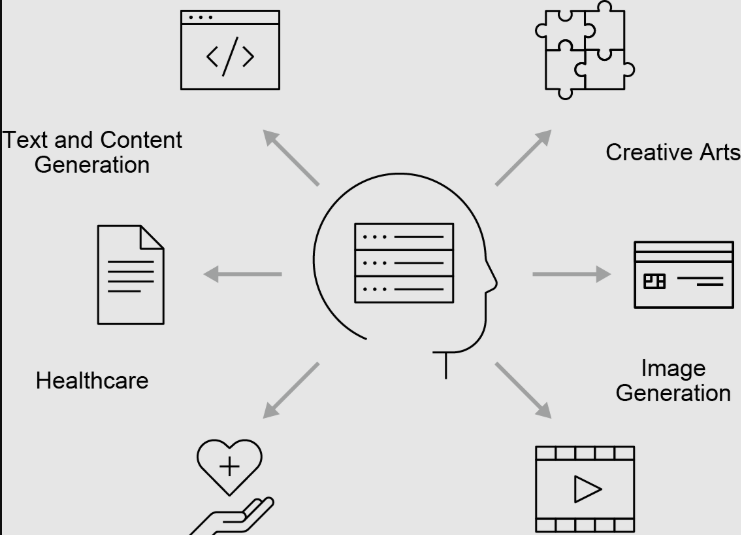
O processo de treinamento de GANs é adversarial, o que significa que as duas redes competem entre si. O objetivo do gerador é produzir dados indistinguíveis dos dados reais. O objetivo do discriminador é identificar com precisão se os dados de entrada são reais ou gerados. Durante o treinamento, o gerador recebe feedback do discriminador, o que o ajuda a melhorar a qualidade dos dados gerados. Esse processo iterativo continua até que o gerador produza dados altamente realistas que possam enganar o discriminador.

Interface gráfica do usuário, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Casos de uso do GenAI**

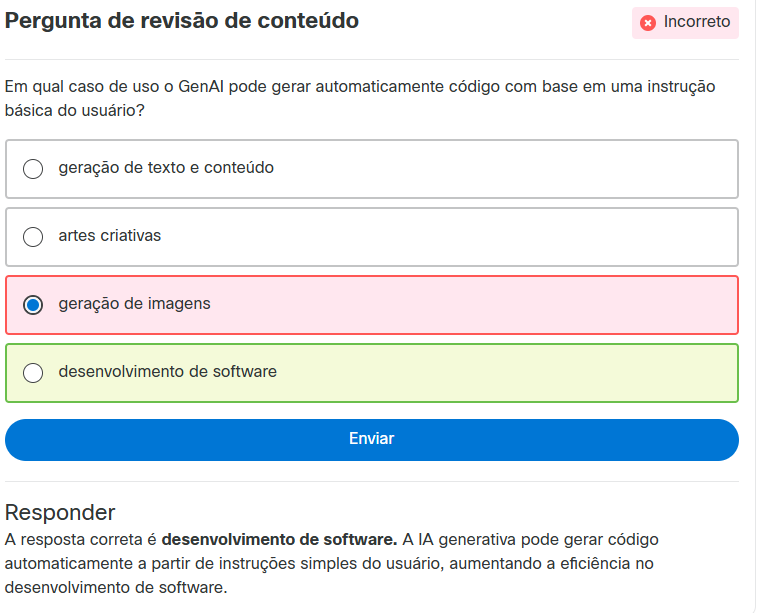
A IA generativa está se tornando cada vez mais popular devido à sua capacidade de auxiliar humanos na realização de diversas tarefas. Mesmo em seus estágios iniciais, a IA generativa auxilia desenvolvedores e engenheiros com design, escrita de código, depuração e outras tarefas nas quais se espera algum tipo de resultado.



Desenvolvedores de software podem interagir com produtos GenAI, como o ChatGPT da OpenAI e o GitHub Copilot, usando linguagem natural em um navegador da web ou em uma extensão para um Ambiente de Desenvolvimento Integrado (IDE) como o Microsoft VS Code. Dentro de um IDE, o GitHub Copilot examina o código e o texto em um arquivo e sugere o preenchimento automático do código que se encaixa logicamente como a próxima linha ou linhas. Um desenvolvedor pode gerar funções inteiras automaticamente simplesmente digitando uma documentação em linha bem elaborada e pressionando uma tecla de atalho que mapeia para aceitar o código sugerido. Além disso, o Copilot possui um chat de texto no qual você pode discutir conceitos e abordagens com o modelo de linguagem. Esses modelos de desenvolvimento também podem gerar documentação com base no código existente, melhorando a legibilidade.

Devido à complexidade de um ambiente, no desenvolvimento de software moderno, os logs de erros podem conter mensagens que não parecem diretamente relacionadas ao que você está tentando capturar. Usando o GenAI, os desenvolvedores podem copiar uma saída de erro ou fazer login nesses modelos para receber a solução para um problema ou uma lista de etapas que podem seguir para uma depuração mais aprofundada.

O GenAI também está sendo usado para automatizar e aprimorar o processo de design. Modelos de IA podem gerar layouts de design, gráficos e interfaces de usuário com base em requisitos específicos ou preferências estéticas. Por exemplo, as ferramentas de IA da Adobe podem auxiliar designers gerando paletas de cores automaticamente, sugerindo elementos de design e criando variações de um conceito de design. Por exemplo, o recurso de preenchimento generativo do kit de ferramentas da Adobe permite selecionar uma região em uma imagem e instruir a IA a fazer uma adição como "coloque um pinguim aqui" ou algo parecido. Os resultados incríveis são difíceis de serem alcançados por um humano sem treinamento e prática extensivos. Recursos como esse aceleram o processo de design e incentivam a criatividade, fornecendo aos designers novas ideias e inspiração.



**Desafios de inferência de IA generativa**

Modelos de aprendizado de máquina treinados extraem conclusões de novos dados em um processo chamado inferência. O treinamento e o ajuste fino realizados em um modelo se correlacionarão com a qualidade de suas inferências, levando a resultados que variam em alinhamento com as expectativas do usuário. Por exemplo, um modelo treinado com milhões de imagens de vacas pode gerar uma imagem de uma vaca extremamente bem, mas nunca será capaz de renderizar uma imagem de uma girafa.

Para ilustrar o processo de inferência, consulte a figura apresentada. Quando um usuário fornece uma entrada, como "Crie uma obra de arte para mim", o modelo GenAI processa essa entrada por meio de suas camadas, utilizando os padrões aprendidos durante o treinamento. A saída resultante é uma nova obra de arte que combina elementos dos dados de treinamento com um certo grau de variabilidade.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Como foi utilizado um pequeno conjunto de dados de treinamento, a capacidade do modelo de gerar resultados diversos e inovadores é limitada. Essa restrição significa que, ao solicitar ao modelo uma entrada vaga como "Criar arte", você invariavelmente verá elementos de cada dado do conjunto de treinamento refletidos no resultado. A arte resultante combina características das imagens de treinamento, imitando a imaginação com aleatoriedade, mas o resultado não é verdadeiramente imaginativo.

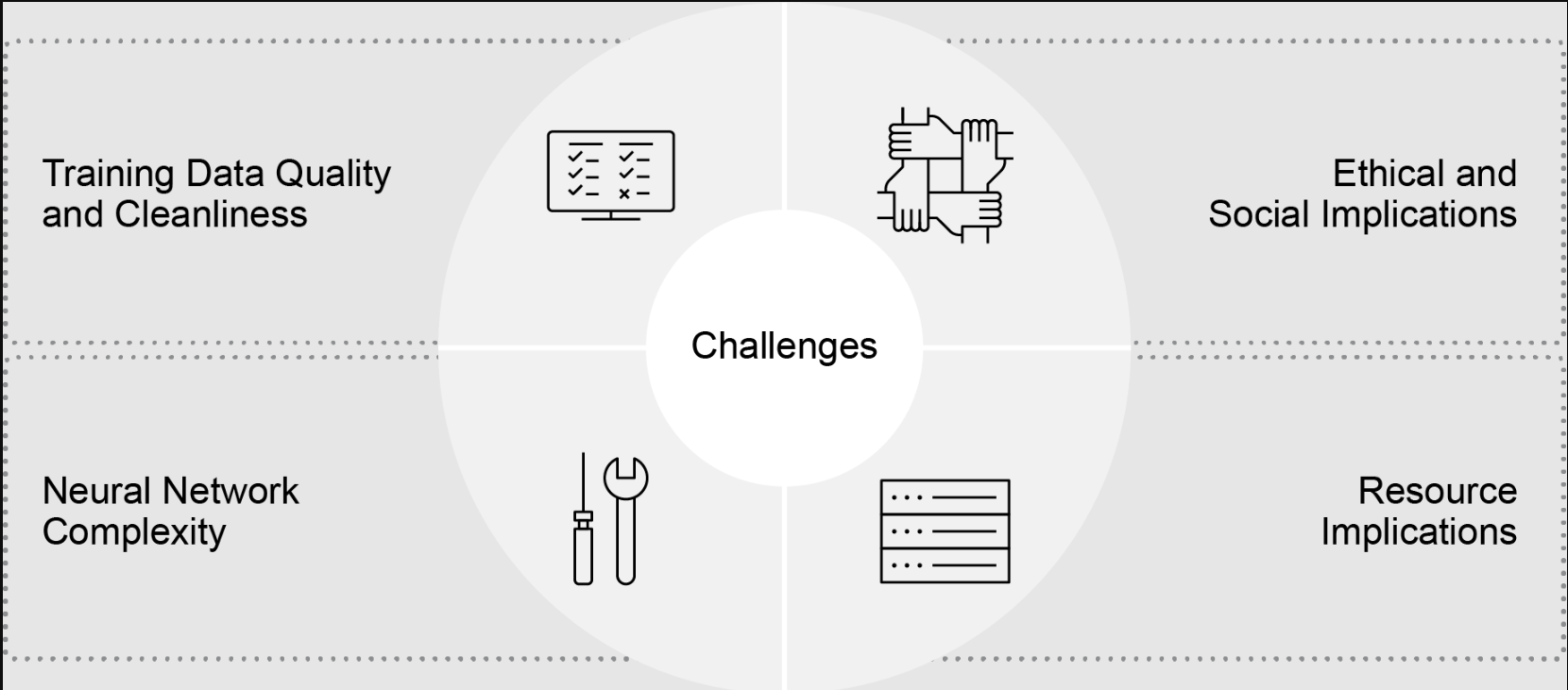
A qualidade das inferências de um modelo generativo determina seu desempenho geral em diferentes sistemas e aplicações. Por exemplo, a geração de imagens em tempo real exige que o modelo crie imagens de alta qualidade rapidamente para uso em realidade virtual ou aplicativos de jogos. O mesmo se aplica à complementação de texto em aplicativos de usuário, como chatbots ou assistentes de escrita; inferências rápidas e precisas são necessárias para fornecer respostas relevantes de forma coerente e rápida.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Desafios e limitações do GenAI

Embora a IA generativa tenha feito progressos significativos na criação de conteúdo crível e humano, ela enfrenta alguns desafios para seu crescimento contínuo.



Texto

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.

Uma imagem contendo Texto

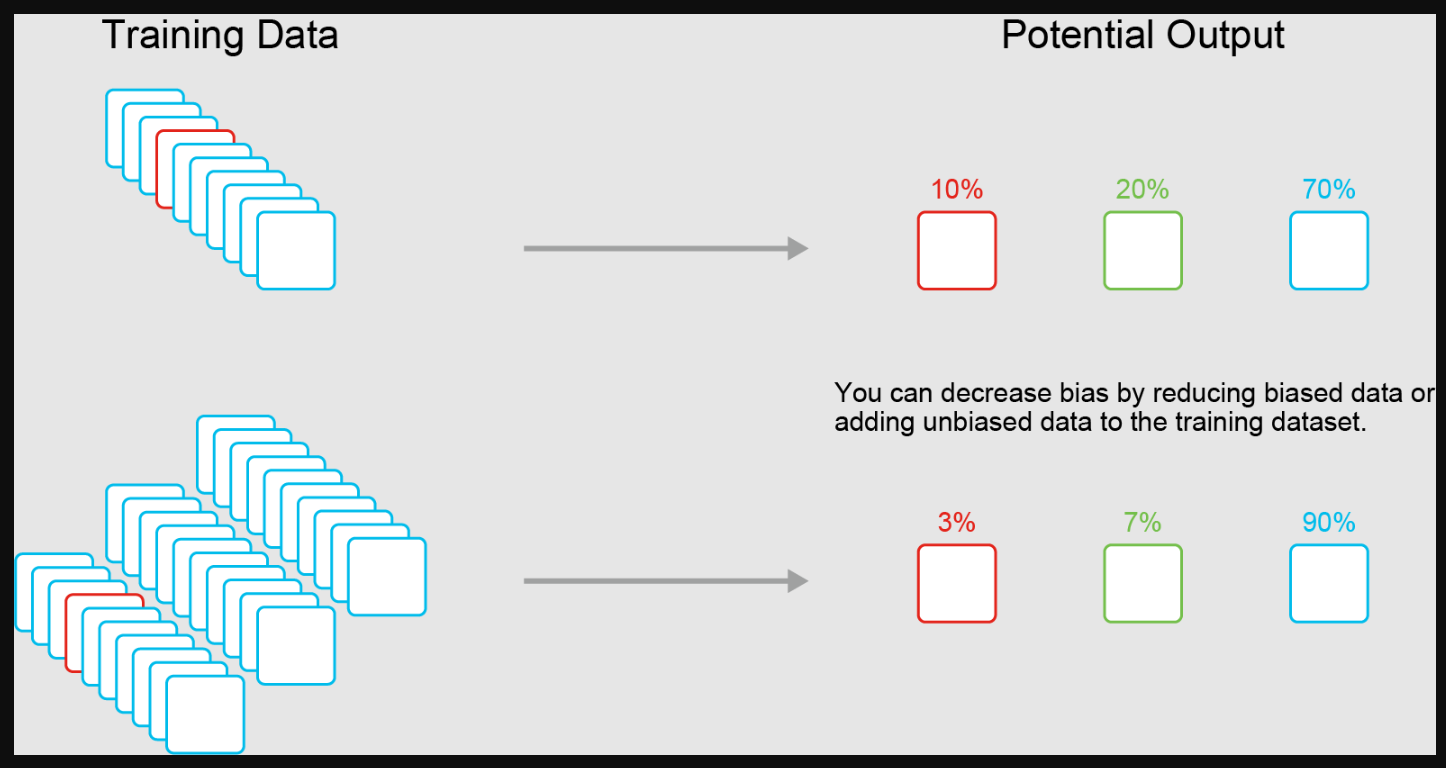
O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Preconceito e imparcialidade da GenAI**

O viés em sistemas de IA pode assumir formas como discriminação racial ou de gênero, que pode reforçar estereótipos culturais e levar a um tratamento desigual. Modelos generativos podem ser tendenciosos e injustos porque herdam os vieses presentes em seus dados de treinamento. Um modelo de linguagem treinado em dados tendenciosos pode produzir conteúdo que reflete esses mesmos vieses. Compreender e identificar vieses em modelos generativos é essencial para garantir que os modelos de IA funcionem de acordo com padrões éticos. Na figura a seguir, os quadrados vermelhos representam dados tendenciosos em um conjunto de dados de treinamento, enquanto os quadrados azuis representam dados limpos. A saída de um modelo treinado em cada representação de um conjunto de dados de treinamento produzirá saídas que podem ser tendenciosas, um pouco tendenciosas ou nada tendenciosas. Você pode diminuir o viés reduzindo os dados tendenciosos ou adicionando dados não tendenciosos ao conjunto de dados de treinamento.



Garantir justiça e equidade em modelos generativos requer diversas técnicas para reduzir vieses. Uma delas é o aumento de dados, que envolve a adição de amostras diferentes e representativas ao conjunto de dados de treinamento. Além disso, restrições de justiça e algoritmos de detecção de vieses podem ser incorporados durante o processo de treinamento do modelo para identificar e corrigir quaisquer padrões de vieses que possam existir nele. O monitoramento e a avaliação contínuos dos resultados do modelo devem ser realizados para abordar quaisquer novos vieses que possam surgir antes de sua implantação.

As implicações éticas de resultados tendenciosos de IA generativa também devem ser consideradas na implantação. Considerações éticas incluem o potencial impacto social da IA ​​generativa, como a disseminação de desinformação por meio de deepfakes ou a criação de conteúdo prejudicial. Desenvolvedores e partes interessadas devem abordar a justiça e a responsabilização em sistemas de IA para construir confiança e aceitação entre os usuários. A adesão às diretrizes e padrões éticos envolve o estabelecimento de estruturas éticas claras para o desenvolvimento e a implantação de IA, a promoção da transparência no design e no treinamento de modelos e o fomento da colaboração entre pesquisadores, especialistas em ética e formuladores de políticas de IA.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Otimização de Recursos GenAI

Otimizar o uso de recursos de hardware, como GPUs e TPUs, é essencial para inferência eficiente em modelos GenAI. Técnicas como poda de modelo, quantização e processamento paralelo podem melhorar o desempenho reduzindo a carga computacional. A poda de modelo envolve a remoção de parâmetros desnecessários da rede neural, tornando o modelo menor e mais eficiente sem comprometer significativamente a precisão. A quantização reduz a precisão dos pesos do modelo, produzindo cálculos de menor precisão que requerem menos energia e memória. Por fim, o processamento paralelo é o processamento simultâneo de múltiplas tarefas, ou a divisão de uma tarefa em várias threads. Com o processamento paralelo, cada recurso computacional trabalha em sua parcela da tarefa geral, acelerando drasticamente o tempo de computação.

Tabela

O conteúdo gerado por IA pode estar incorreto.

Gerenciar o consumo de energia de modelos generativos durante a inferência é crucial para minimizar o impacto ambiental e garantir a implantação sustentável da IA. Modelos generativos podem consumir quantidades substanciais de energia, contribuindo para a pegada de carbono das aplicações de IA. Equilibrar desempenho e eficiência energética envolve otimizar algoritmos e hardware para reduzir o consumo de energia. Data centers que executam cargas de trabalho de IA generativa podem se beneficiar de práticas de eficiência energética, como o uso de fontes de energia renováveis ​​e a implementação de soluções de resfriamento que reduzem o consumo geral de energia.

Otimizar a infraestrutura em nuvem ou local pode reduzir os custos de implantação e manutenção de modelos GenAI. Técnicas como computação sem servidor e inferência distribuída podem ajudar a alcançar a relação custo-benefício, dimensionando recursos dinamicamente com base na demanda. A computação sem servidor permite que os desenvolvedores executem código sem provisionar ou gerenciar servidores, enquanto a inferência distribuída utiliza múltiplas máquinas para lidar com tarefas de inferência em larga escala, o que melhora a eficiência e reduz custos.

Interface gráfica do usuário, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**IA generativa vs. IA tradicional**

A IA tradicional e a GenAI têm abordagens técnicas muito diferentes. Cada uma oferece vantagens únicas, mas também enfrenta desafios e limitações distintos.

**IA tradicional**

A IA tradicional baseia-se principalmente em sistemas baseados em regras. O modelo tradicional de IA utiliza símbolos, também chamados de tokens, para representar algum elemento de conhecimento, além de regras para analisar esses símbolos e chegar a conclusões. Essa abordagem, com regras bem definidas, explícitas e transparentes para a tomada de decisões, é relativamente mais fácil de entender e depurar.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Um exemplo de aplicação tradicional de IA é um sistema financeiro baseado em regras, projetado para detectar transações fraudulentas. O sistema se baseia em um conjunto de regras que representam padrões conhecidos que indicam alguma probabilidade de atividade fraudulenta. Uma transação, como uma compra de US$ 2.500 em uma loja específica na cidade de Nova York em 5 de janeiro de 2024, chega em formato JSON e é enviada ao mecanismo de inferência. O mecanismo de inferência usa suas diversas regras de lógica de negócios para analisar os detalhes da transação e produzir um veredito.

O mecanismo de inferência deve utilizar uma base de conhecimento para gerar sua saída. A base de conhecimento contém um amplo conjunto de registros ou transações históricas, padrões de fraude conhecidos, regras internas específicas e predefinidas para identificar atividades estranhas ou suspeitas e até mesmo regras relacionadas a requisitos regulatórios.

Alguns exemplos dessas regras podem incluir:

* SE o valor for > $ 2.000, ENTÃO marque como potencialmente fraudulento
* SE a localização não estiver a 80 km da localização típica do usuário, ENTÃO marque como potencialmente fraudulento
* SE o comerciante não estiver na lista de comerciantes frequentes, ENTÃO marque como potencialmente fraudulento

Cada regra na base de conhecimento é avaliada em relação à compra recente. Em seguida, os resultados da análise, incluindo uma enumeração detalhada de cada regra e a descrição exata de quais partes da regra falharam, são enviados de volta ao mecanismo de inferência. A saída gerada pelo mecanismo de inferência inclui uma decisão final sobre a probabilidade de a transação ser fraudulenta, bem como um resumo do motivo pelo qual a transação foi sinalizada e uma pontuação de confiança.

A IA tradicional é adequada para ambientes altamente estruturados. No entanto, esses sistemas costumam ser severamente limitados em termos de flexibilidade e escalabilidade. Quase sempre, eles exigirão atualizações manuais na lista de regras e lógicas que regem sua tomada de decisão, caso o sistema precise ser alterado para lidar com cenários novos ou inesperados. À medida que o número de regras e cenários aumenta, o número de interações entre essas regras pode crescer exponencialmente, e o sistema se torna mais desafiador de gerenciar.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

**IA de geração**

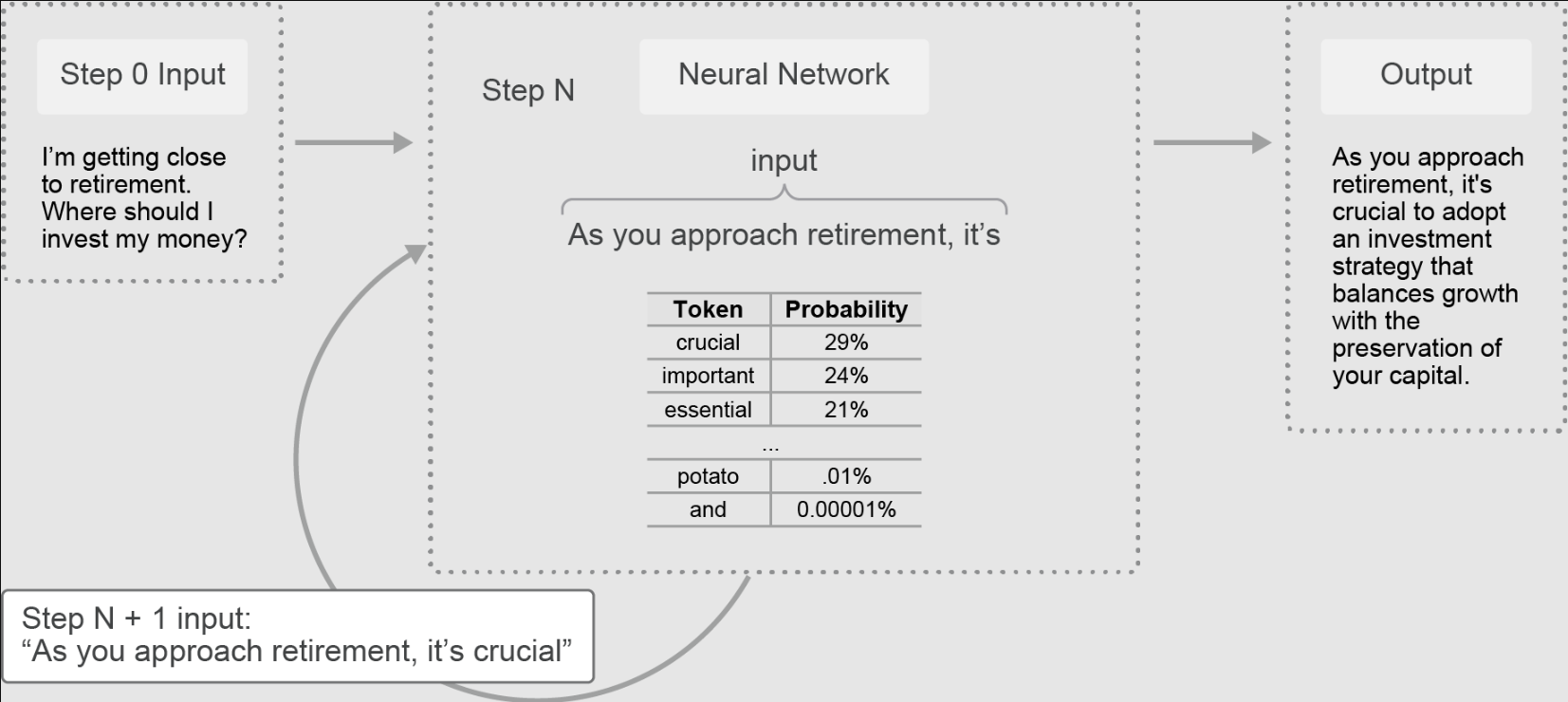
Por outro lado, o objetivo da GenAI é criar algum tipo de conteúdo ou tomada de decisão inovadora utilizando imensos conjuntos de dados. A GenAI utiliza redes neurais e aprendizado de máquina para atingir esse objetivo. Esse tipo de IA supera os humanos em ambientes extremamente complexos, onde o número de variáveis ​​é maior do que o que normalmente seria gerenciado por regras distintas.

Em vez de ser usado em um sistema de detecção de transações fraudulentas, o GenAI teria maior probabilidade de ser usado como um chatbot de consultor financeiro. Se você tentar codificar um aplicativo de IA tradicional para fornecer aconselhamento financeiro, o grande número de variáveis ​​envolvidas tornaria quase impossível abordá-las de forma abrangente. O GenAI e as redes neurais ainda têm o mesmo nível de complexidade subjacente. No entanto, com o GenAI, o foco do desenvolvedor muda de listar regras específicas para ajustar o modelo com base em entradas para atingir uma baixa taxa de erro. No final, o desenvolvedor pode não ser capaz de explicar como o modelo chegou à sua saída, mas o sistema ainda pode fazê-lo descobrindo semelhanças entre dados históricos e atuais.

O processo de utilização de um modelo GenAI começa com uma interação, geralmente por meio de um prompt fornecido pelo usuário. Esse prompt é transformado em tokens por meio de um processo chamado tokenização. Um token pode ser uma palavra, um fragmento de uma palavra ou até mesmo um único caractere, dependendo do esquema de tokenização utilizado pelo modelo.

Os tokens de entrada iniciais são processados ​​pelo modelo para gerar uma distribuição de probabilidade para o próximo token possível. Esse processo de geração é iterativo, onde cada novo token gerado é influenciado tanto pelo prompt original quanto por todos os tokens gerados anteriormente. Após a geração de cada token, o modelo o incorpora à sequência e usa a sequência atualizada como entrada para a próxima iteração. Esse processo iterativo continua até que:

* O modelo gera um token especial (frequentemente chamado de token de "fim de sequência") que sinaliza a conclusão da saída
* Outro critério de parada é atendido, como atingir um limite de token predefinido



A GenAI oferece vantagens significativas, como notável criatividade e inovação, adaptabilidade a novas tarefas com retreinamento mínimo e personalização eficaz. No entanto, a complexidade inerente desses modelos, embora permita resultados sofisticados, também pode levar à imprevisibilidade. Os modelos GenAI podem herdar vieses presentes nos dados de treinamento, exigindo atualização contínua e seleção cuidadosa dos dados para garantir relevância e imparcialidade. Além disso, a natureza intensiva em recursos do treinamento desses modelos apresenta desafios significativos.

Diagrama, Diagrama de Venn

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Requisitos de dados da GenAI vs. IA tradicional**

Devido às suas abordagens técnicas e domínios de aplicação, a IA tradicional e a IA generativa têm requisitos de dados muito diferentes.

**IA tradicional**

Os sistemas tradicionais de IA dependem fortemente de dados explícitos e estruturados, além de regras predefinidas. Eles são projetados para operar dentro de limites bem definidos e, frequentemente, exigem entradas de dados claras e concisas. Quando o sistema de IA precisa tomar uma decisão, ele consulta esses dados estruturados para aplicar as regras relevantes e chegar a conclusões.

O MYCIN, um sistema médico especializado, processa dados de pacientes usando regras de diagnóstico predefinidas e recomenda um conjunto de tratamentos aplicáveis. O conjunto de dados históricos inclui muitos detalhes clinicamente relevantes, incluindo informações do paciente como idade, sexo, sintomas, histórico médico, resultados de exames e diagnósticos confirmados. Quando um novo paciente apresenta sintomas problemáticos, o sistema compara os dados do novo paciente com esse conjunto de dados históricos. O sistema de IA usa esses dados estruturados armazenados em sua base de conhecimento para aplicar as regras predefinidas e sugerir um diagnóstico para o novo paciente com base em padrões semelhantes.

Tabela

O conteúdo gerado por IA pode estar incorreto.

Devido a essa dependência de dados bem definidos e estruturados, os sistemas tradicionais de IA exigem menos dados em geral e são mais fáceis de gerenciar. Eles podem apresentar desempenho superior mesmo utilizando conjuntos de dados menores e formatos de dados mais simples. No entanto, esses conjuntos de dados pequenos também limitam a flexibilidade de um modelo tradicional de IA e sua capacidade de expansão fácil. Os sistemas tradicionais de IA têm dificuldade para processar dados ambíguos e não conseguem lidar eficientemente com dados não estruturados, como texto livre ou imagens complexas. Essa limitação impõe um limite rígido ao número de aplicações relevantes de IA tradicional e as limita a domínios bem definidos nos quais os dados são previsíveis e consistentes.

**IA generativa**

Em contraste, o GenAI precisa de grandes quantidades de dados para treinamento, a fim de aprender e gerar novos resultados realistas. Esses enormes conjuntos de dados diversos permitem que os modelos GenAI entendam e imitem padrões e estruturas complexas. Por exemplo, modelos GenAI como Redes Adversariais Generativas (GANs) podem usar conjuntos de dados pictóricos para aprender as características gerais de gatos e seus ambientes. Em seguida, o modelo pode ser usado para gerar imagens que incluem gatos com fundos realistas.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

No entanto, como o conjunto de dados mostrado categoriza essas imagens simplesmente como "Gato Doméstico", quaisquer imagens desses gatos que o modelo gerar provavelmente não apresentarão alguns detalhes. As imagens de gatos podem ter o formato e o tamanho corretos, mas podem não se assemelhar a nenhuma raça específica. Ou podem se parecer com uma raça inexistente devido à mistura de características de diferentes raças. O modelo com o conjunto de dados mostrado não consegue distinguir entre raças nem identificá-las pelo nome. Para treinar um modelo GenAI para gerar imagens de gatos de uma única raça específica ou de várias raças, um desenvolvedor precisa expô-lo a milhares de imagens de várias raças, juntamente com metadados que descrevam de alguma forma os dados que o modelo está visualizando.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Este conjunto de dados, com rótulos e outros metadados incluídos, ajuda o modelo a diferenciar diferentes raças, aprendendo suas características únicas. Ao vincular essas características visuais a diversos metadados sobre as raças, a capacidade do modelo de reconhecê-las ou gerá-las com precisão melhora drasticamente. As imagens geradas podem mostrar com mais clareza a pelagem manchada característica de um gato-de-bengala ou o corpo sem pelos de um gato-esfinge. O modelo agora poderá produzir imagens muito mais detalhadas e realistas de uma raça de gato selecionada.

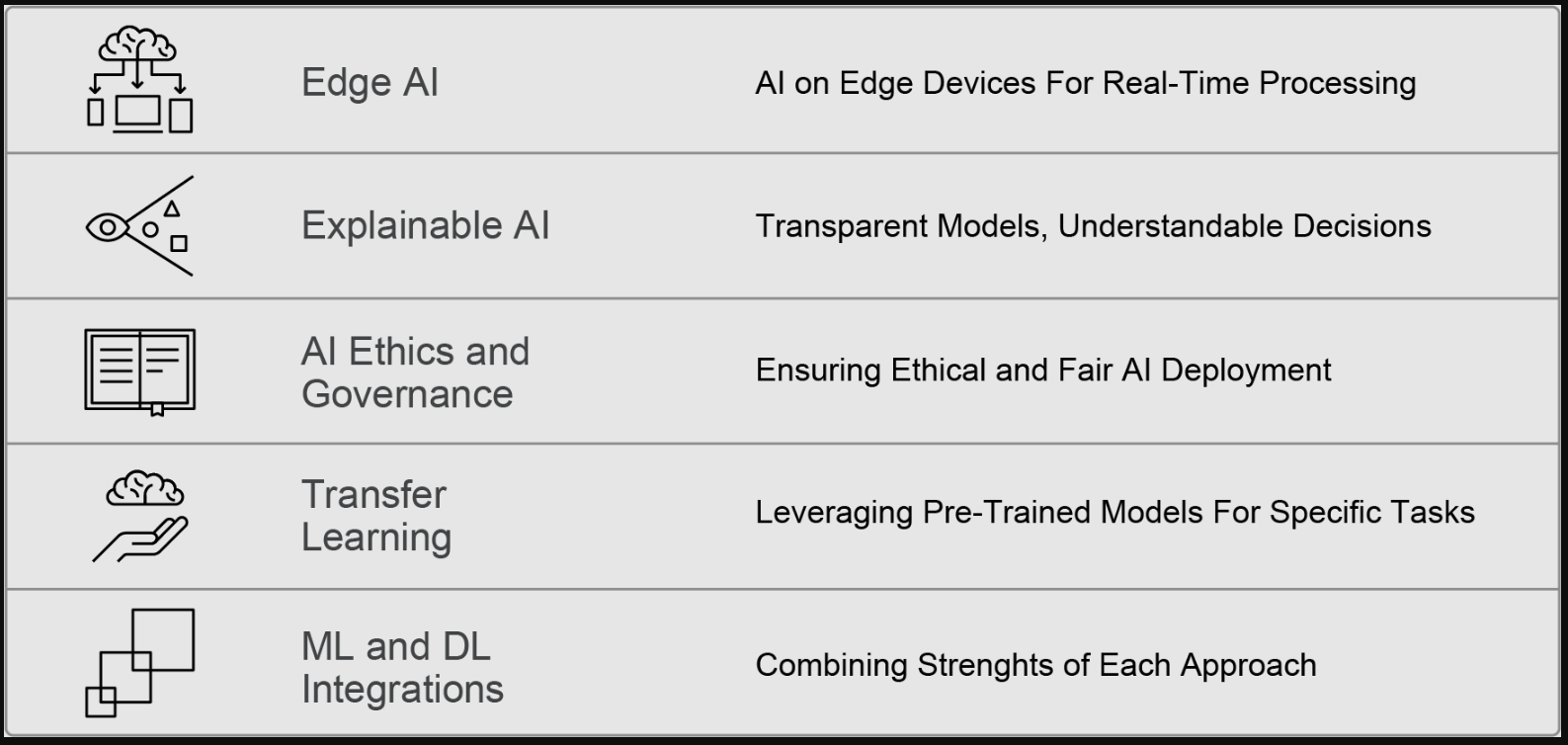
A capacidade de lidar com dados não estruturados é uma vantagem significativa do GenAI. Esses modelos criam uma ampla gama de saídas em formatos tão díspares quanto texto e imagens, até mesmo música ou código. No entanto, essa vantagem acarreta altos requisitos de dados. Os modelos GenAI precisam de conjuntos de dados extensos para funcionar de forma eficaz, e dados de treinamento vagos ou de baixa qualidade levarão a saídas de baixa qualidade. Esse requisito pode ser uma barreira significativa, pois a aquisição e o processamento de conjuntos de dados tão grandes exigem recursos substanciais em termos de tempo, armazenamento e poder computacional.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Tendências futuras em IA

O futuro da IA ​​é marcado por avanços significativos e aplicações em evolução que prometem remodelar vários setores e aspectos da vida diária.



Várias tendências emergentes estão moldando o futuro da IA, impulsionando a inovação e criando novas oportunidades:

* **IA de Borda:** A implantação de modelos de IA em dispositivos de borda, como smartphones e dispositivos IoT, permite o processamento de dados e a tomada de decisões em tempo real sem depender da infraestrutura de nuvem. Essa tendência é crucial para aplicativos que exigem baixa latência e alta privacidade.
* **IA explicável:** À medida que os sistemas de IA se tornam mais complexos, há uma necessidade crescente de IA explicável para garantir transparência e confiança. A pesquisa nessa área se concentra no desenvolvimento de modelos que forneçam explicações claras e compreensíveis para suas decisões.
* **Ética e Governança da IA:** Garantir o desenvolvimento e a implantação éticos da IA ​​está se tornando cada vez mais importante. O desenvolvimento responsável envolve lidar com vieses nos dados de treinamento, garantir a imparcialidade e estabelecer diretrizes para o uso responsável da IA.
* **Aprendizado por Transferência:** O aprendizado por transferência permite que modelos utilizem conhecimento pré-treinado em grandes conjuntos de dados e o apliquem a tarefas específicas com menos dados e tempo de treinamento. Essa abordagem torna os modelos de aprendizado profundo mais eficientes e escaláveis.
* **Integração de ML e DL:** A integração de técnicas de aprendizado de máquina e aprendizado profundo pode levar a soluções de IA mais robustas, combinando os pontos fortes de ambas as abordagens. Por exemplo, o uso de aprendizado de máquina para seleção de recursos e aprendizado profundo para extração e previsão de recursos pode aprimorar o desempenho e a versatilidade do modelo. Essa abordagem híbrida permite um processamento de dados mais eficiente e maior precisão em tarefas complexas.

Interface gráfica do usuário, Aplicativo

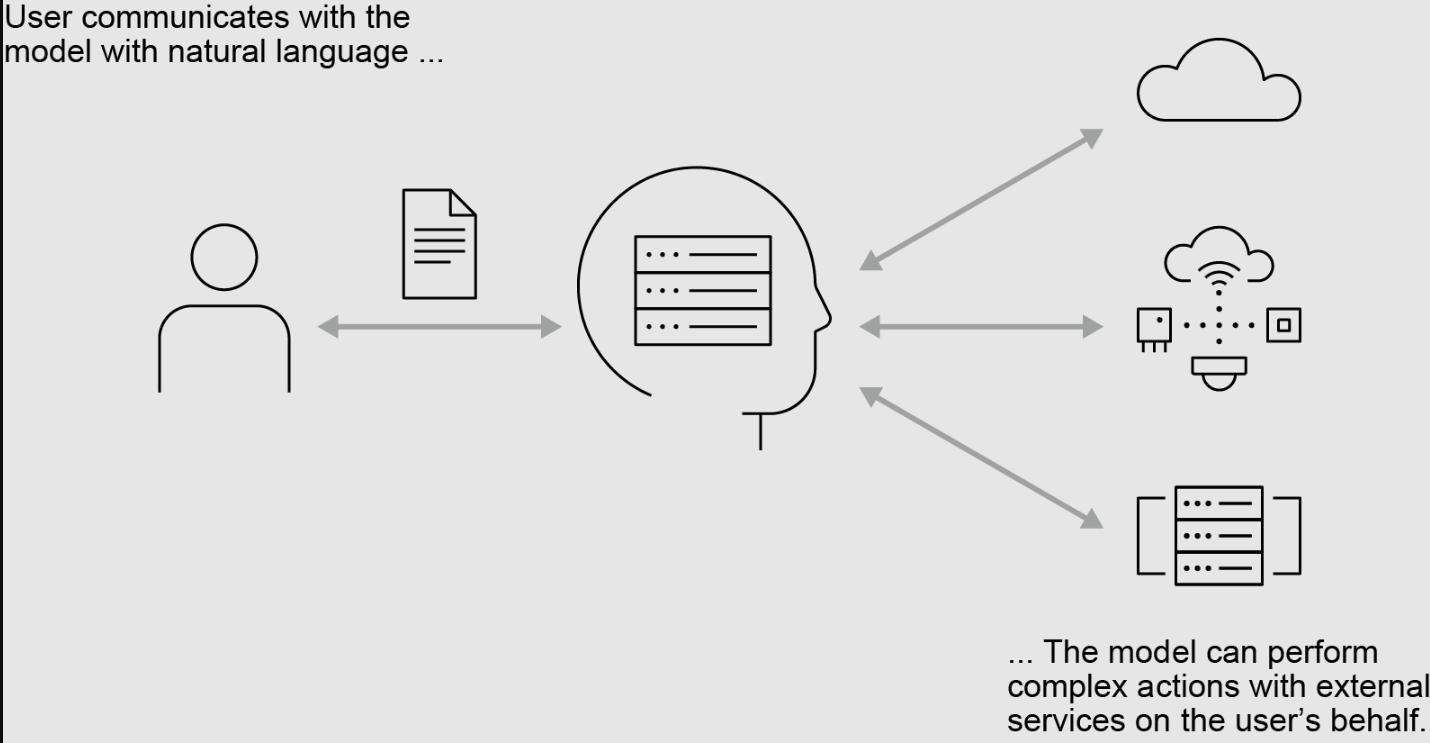
O conteúdo gerado por IA pode estar incorreto.

**Modelos de Linguagem de IA**

Modelos de linguagem são sistemas sofisticados de IA projetados para compreender entradas em linguagem humana, gerar saídas em linguagem humana e manipular a linguagem. Esses modelos utilizam técnicas de aprendizado profundo para processar e produzir texto ou voz, possibilitando uma ampla gama de aplicações, como chatbots, sumarização de texto e serviços de tradução. Ao analisar grandes quantidades de dados de linguagem, os modelos de linguagem podem executar tarefas complexas que imitam a compreensão e a geração da linguagem humana.

**Conceitos Básicos**

A razão pela qual os modelos de linguagem são tão importantes no campo da IA ​​é que os humanos falam línguas diferentes. É importante notar que a única maneira pela qual a maioria das pessoas consegue interagir com IA avançada é comunicando-se com ela usando sua própria linguagem. Um usuário solicita a um modelo que execute alguma ação em seu nome. Portanto, os modelos de linguagem podem aceitar entradas e gerar saídas tanto em formato de texto quanto falado.



Modelos de linguagem podem ser projetados para interagir com diversos sistemas e serviços externos por meio de Interfaces de Programação de Aplicativos (APIs). Um modelo de linguagem devidamente treinado pode obter informações de sua interação com um usuário e usá-las para consultar um banco de dados externo, enviar um comando para um dispositivo doméstico inteligente ou recuperar informações de um serviço online. Por exemplo, após um modelo receber algumas permissões, senhas ou tokens de segurança, ele pode atuar em nome do usuário para:

* Acesse a API de um serviço de reserva
* Pesquisar horários disponíveis
* Conclua o processo de reserva sem muita intervenção do usuário

A combinação do processamento de linguagem natural e a integração com outras tecnologias aprimoram a funcionalidade e a utilidade dos sistemas de IA, aproximando-os da compreensão e da atuação efetivas das intenções humanas em diversos domínios. Os sistemas de IA tornam-se capazes de atuar em objetivos complexos e multietapas, tornando-os ferramentas poderosas para automatizar tarefas e fornecer assistência inteligente e integrada. É provável que, no futuro, as integrações de modelos de linguagem se tornem mais complexas e interajam com outros tipos de IA para realizar feitos incríveis com mais rapidez do que os humanos jamais poderiam esperar.

**Tipos de modelos de linguagem**

Existem duas categorias distintas dentro do domínio da IA ​​generativa: modelos de linguagem grande (LLMs) e modelos de linguagem pequena (SLMs), cada um com características e aplicações únicas.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Os LLMs são os mais avançados dos dois tipos de modelos de linguagem. Eles exigem conjuntos de dados muito maiores para serem treinados adequadamente. Eles podem possuir bilhões de parâmetros dentro das redes neurais das quais dependem para viabilizar sua abordagem de aprendizado profundo. Essa abordagem permite que os modelos gerem textos coerentes e contextualmente relevantes para diversas tarefas. Por exemplo, o ChatGPT da OpenAI, com seus 175 bilhões de parâmetros, é um LLM capaz de realizar diversas tarefas de linguagem, incluindo escrever redações, responder perguntas e gerar respostas conversacionais.

SLMs são um tipo mais simples de modelo de linguagem que requer um número menor de parâmetros e contém menos complexidade computacional. Embora também possam ser treinados usando técnicas de aprendizado profundo, os SLMs geralmente são otimizados para tarefas específicas, com um limite menor quanto ao que está dentro do escopo e o que não está.

O tamanho reduzido e a eficiência de um SLM são alcançados por meio de técnicas como destilação de conhecimento, poda e quantização, que permitem que os SLMs operem eficientemente com recursos limitados. A destilação de conhecimento transfere o conhecimento de um LLM pré-treinado para um modelo menor, capturando suas principais capacidades sem toda a complexidade. A poda remove partes menos úteis do modelo, e a quantização reduz a precisão de seus pesos, reduzindo ainda mais seu tamanho e os requisitos de recursos.

Os SLMs geralmente não exigem os extensos conjuntos de dados e o poder computacional dos LLMs. Portanto, são mais acessíveis para aplicações práticas com recursos limitados. Um modelo de chatbot personalizado, treinado em um conjunto de dados específico de atendimento ao cliente com alguns milhões de parâmetros, exemplifica um SLM, fornecendo respostas direcionadas sem os extensos requisitos de recursos de um LLM.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

LLMs vs. SLMs

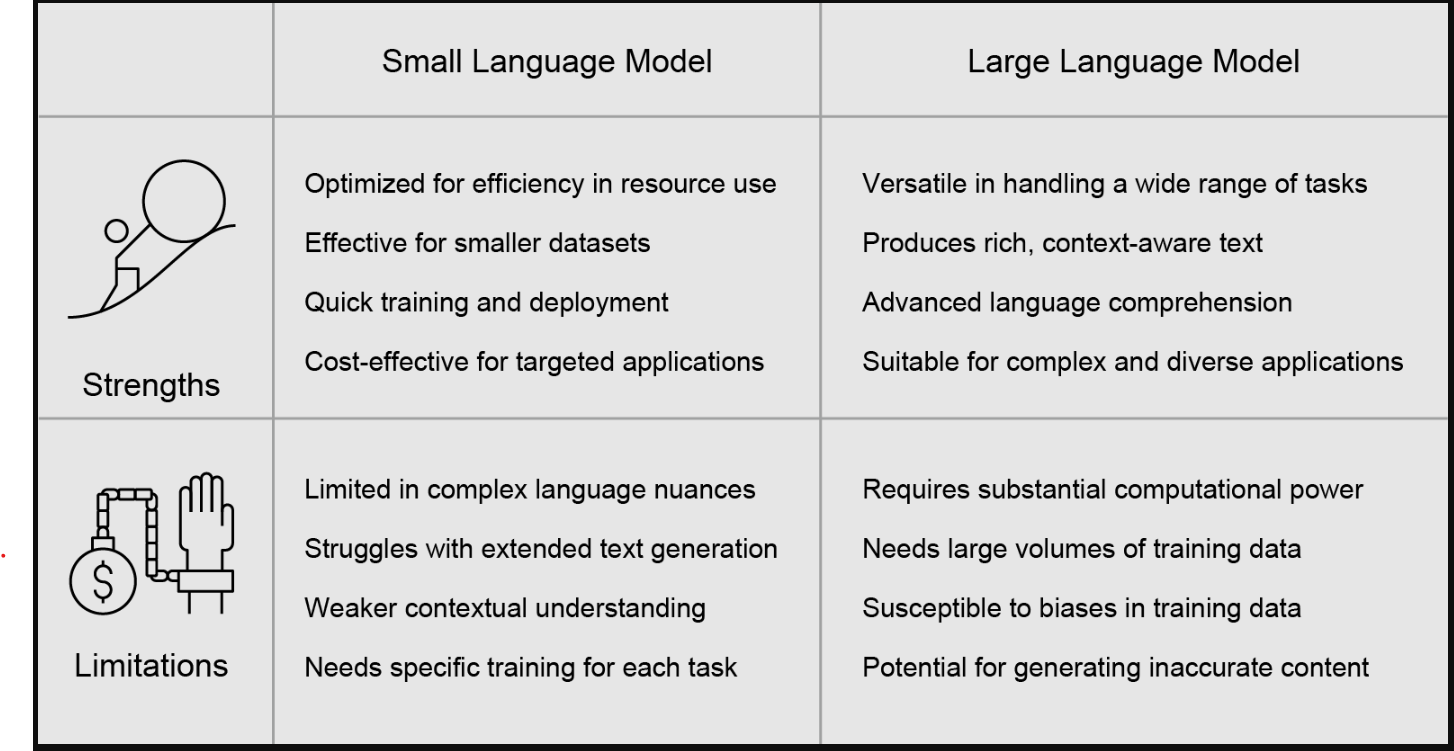
LLMs vs. SLMs

Os tomadores de decisão precisam entender a diferença entre modelos de linguagem grande (LLMs) e modelos de linguagem pequena (SLMs) para que possam estar informados ao decidir que tipo de aplicação de IA desejam desenvolver. Conhecer os pontos fortes e as limitações de cada tipo de modelo ajuda a selecionar a abordagem certa para necessidades específicas. Esse conhecimento fundamental garante que as organizações possam otimizar recursos e tomar decisões informadas que se alinhem aos seus objetivos específicos.

A arquitetura mais simples dos SLMs permite sua implantação em ambientes com recursos limitados, como sistemas embarcados ou aplicativos móveis. Além disso, os SLMs são mais adequados para aplicativos que exigem mais segurança devido às suas bases de código menores e menos superfícies com potencial para violações de segurança. Eles também são melhores para o meio ambiente, consumindo menos energia e memória. Essa eficiência também os torna ideais para computação de ponta e aplicativos em tempo real. Como os SLMs consomem menos recursos computacionais, eles oferecem economia significativa de custos e menor impacto ambiental.

Os SLMs são mais fáceis de treinar, o que os torna adequados para aplicações práticas nas quais eficiência e velocidade são primordiais. São mais aplicáveis ​​a tarefas que não exigem uma compreensão profunda do texto lido e que carecem de geração de texto extensiva. Por exemplo, os SLMs podem classificar rapidamente e-mails como spam ou não, ou extrair palavras-chave de textos curtos. Sua implantação em cenários onde é necessário um processamento rápido e dinâmico reforça ainda mais sua utilidade em aplicações específicas e direcionadas.

Por outro lado, os LLMs são caracterizados por seu treinamento em vastos conjuntos de dados. Essa característica os permite lidar com uma ampla gama de tarefas linguísticas. Os grandes conjuntos de dados permitem que eles generalizem e abstraiam significados, gerem textos complexos e diversos, entendam o contexto em que uma entrada foi fornecida e executem raciocínio em várias etapas. Eles provavelmente terão um bom desempenho em diversas tarefas sem treinamento extensivo específico para cada tarefa. Os LLMs foram criados para gerar respostas semelhantes às humanas em aplicativos de bate-papo, traduzir textos e escrever ensaios. Essa versatilidade os torna adequados para uma ampla gama de aplicações, desde atendimento automatizado ao cliente até geração de conteúdo criativo. A capacidade de gerar respostas diferenciadas e sensíveis ao contexto distingue ainda mais os LLMs, tornando-os indispensáveis ​​para aplicações sofisticadas de IA que exigem níveis mais altos de precisão e complexidade.



Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Resumo

Neste treinamento, você explorou o fascinante mundo da IA ​​Generativa, começando com sua capacidade de criar arte e estendendo-se ao seu potencial de automatizar o desenvolvimento de software. Você viu em primeira mão como a IA está expandindo os limites do que as máquinas podem realizar e adquiriu insights cruciais sobre o funcionamento da IA ​​Generativa, incluindo redes neurais e Redes Adversariais Generativas (GANs).

Agora que você concluiu este treinamento, considere as seguintes questões:

* Como a IA Generativa pode ser aplicada em seu campo ou setor específico para aumentar a criatividade e a eficiência?
* Quais são as considerações éticas que você deve ter em mente ao implantar sistemas de IA no seu trabalho?
* Como você planeja abordar possíveis vieses em modelos de IA com os quais você pode trabalhar no futuro?
* Reflita sobre um problema real que você enfrenta no seu ambiente profissional. Como a IA Generativa pode oferecer uma solução?