TOTVS

GUIA DE IA

IN DATA WE TRUST

julho/24

GUIA DE IA

Bem-vindo ao Guia de Inteligência Artificial da TOTVS, sua fonte definitiva para entender e implementar tecnologias de IA na sua empresa.

Este guia foi criado com um objetivo: capacitar você e sua empresa a explorar, adotar e maximizar o potencial da IA, transformando desafios em oportunidades de inovação.

Na TOTVS, acreditamos que a IA pode ser um diferencial competitivo crucial, capaz de revolucionar a maneira como os negócios são conduzidos.

POR QUE ESTE GUIA É PARA VOCÊ?

- **Conhecimento prático:** entenda os fundamentos da IA, desde os conceitos básicos até as aplicações mais avançadas.
- **Casos de uso reais:** inspire-se com exemplos concretos de como a IA está sendo utilizada em diversas indústrias para resolver problemas complexos e criar valor.
- **Ferramentas e técnicas:** aprenda sobre as principais ferramentas e técnicas de IA que podem ser aplicadas diretamente no seu ambiente de trabalho.
- **Estratégia e implementação:** descubra os passos práticos para iniciar projetos de IA, desde a identificação de oportunidades até a implementação e avaliação de resultados.

O QUE VOCÊ VAI ENCONTRAR NESTE GUIA:

Tudo sobre IA

1. **O que é Inteligência Artificial e os principais tipos de IA:** uma introdução objetiva sobre o que é IA e as diferentes categorias que a compõem.
2. **O papel dos dados para IA:** entenda a importância dos dados na construção e no treinamento de modelos de IA.
3. **IA Generativa, o "Hype" do momento?** Explore o mundo da IA Generativa, seus principais componentes e suas aplicações mais empolgantes.
4. **Principais casos de uso para IA Generativa:** descubra como diferentes indústrias estão utilizando IA Generativa para transformar seus negócios.
5. **Como tratamos/priorizamos o tema de Dados & IA na TOTVS?** Conheça a abordagem da TOTVS para integrar IA e dados em sua estratégia de negócios.
6. **5 passos para começar um projeto de IA na sua empresa:** um guia passo a passo para você iniciar sua jornada com IA, desde a concepção até a escala.
7. **Glossário:** familiarize-se com os termos e conceitos mais importantes do mundo da IA.

Seja bem-vindo a uma nova era de possibilidades com a Inteligência Artificial.
Vamos começar!

SUMÁRIO

1 O QUE É INTELIGÊNCIA ARTIFICIAL
e os principais tipos de IA

O QUE VOCÊ VAI ENCONTRAR NESTE GUIA

A **Inteligência Artificial (IA)** é um campo da ciência da computação dedicado a criar sistemas que podem executar tarefas que, normalmente, requerem esforço humano. Isso inclui algumas capacidades como o raciocínio, o aprendizado, a percepção de ambiente e até a manipulação de objetos.

Os avanços na IA permitem que máquinas aprendam a partir de experiências, se ajustem a novas entradas de dados e executem tarefas humanas com eficiência. Inclusive, existem formas diferentes aceitas para agrupar e classificar tipos de Inteligência Artificial conhecidos mas, resumidamente, **podemos elencar os seguintes tipos de IA existentes:**

Aprendizado de Máquina (Machine Learning ou Deep Learning): é um campo da IA que foca em interpretar e fazer previsões ou classificações com base nos dados por meio de algoritmos capazes de aprender a partir de dados e fazer previsões ou tomar decisões com base nesses dados.

O Deep Learning pode ser entendido como uma subcategoria do Aprendizado de Máquina, utilizando redes neurais profundas (com muitas camadas de processamento) para modelar e resolver problemas complexos, como o reconhecimento de voz ou na análise de imagens.

A ideia central é desenvolver modelos que possam melhorar seu desempenho ao longo do tempo sem intervenção humana direta na programação das tarefas específicas.

Visão Computacional: é um campo da Inteligência Artificial que envolve o desenvolvimento de técnicas e algoritmos que permitem que computadores interpretem e entendam o conteúdo visual do mundo ao redor. Este campo é fundamental para muitas aplicações práticas onde a percepção visual é

necessária, como em sistemas de segurança, controle de qualidade na manufatura, diagnósticos médicos por imagem, reconhecimento de objetos, e muito mais.

Processamento de Linguagem Natural (PLN): foca na interação entre computadores e humanos usando a linguagem natural. As principais tarefas do PLN incluem tradução automática, reconhecimento de fala, análise de sentimentos, extração de informação, e geração de texto, entre outras. O objetivo do PLN é compreender, interpretar e manipular a linguagem humana de maneira que seja útil para aplicações práticas. A geração de texto é apenas uma das muitas aplicações dentro do PLN.

IA Generativa: é um campo amplo e inclui diferentes técnicas que são usadas para gerar novos dados que podem ser imagens, música, vídeo, e texto, entre outros.

O foco aqui é na criação de conteúdo que é indistinguível do real ou na geração de novos exemplos dentro de um certo domínio de dados.

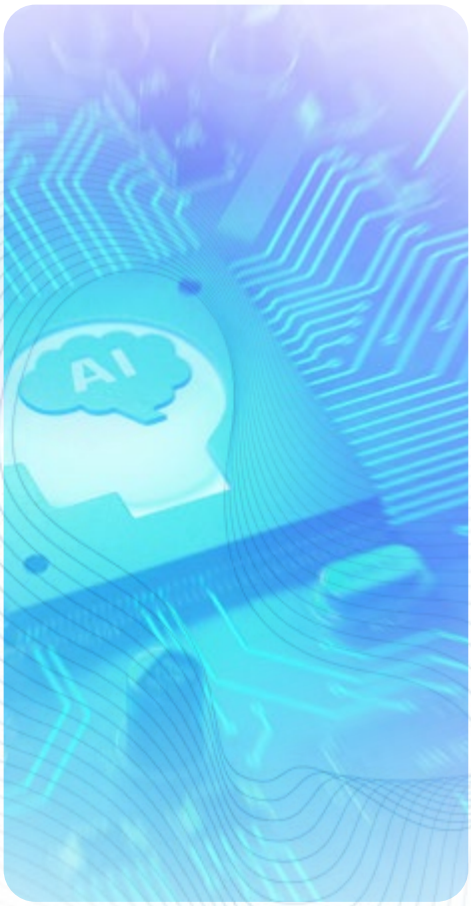
Mas quando nos referimos especificamente a geração de texto, então os modelos utilizados são os chamados LLMs (Large Language Models), que se valem de várias técnicas tanto para compreensão da linguagem humana, quanto para a produção de texto que faça sentido para a compreensão humana.

Vale observar que os **4 tipos acima mencionados podem se interrelacionar** para resolver determinados casos de uso onde se exige diferentes técnicas ou modelos aplicados em conjunto para se chegar ao resultado esperado. Por exemplo, a IA Generativa para geração de texto e interações com seres humanos necessita que previamente os modelos de PLN entrem em campo para determinar o entendimento do tema da interação em andamento.

2 O PAPEL DOS DADOS PARA IA

O QUE VOCÊ VAI ENCONTRAR NESTE GUIA

Os dados desempenham um papel fundamental na **Inteligência Artificial (IA),** sendo essenciais para treinar, testar e aprimorar modelos de IA. Eles são a base sobre a qual os algoritmos de aprendizado de máquina (machine learning) e aprendizado profundo (deep learning) operam, e a qualidade, quantidade e relevância desses dados têm um impacto direto na eficácia e eficiência dos sistemas de IA. **Nesse sentido, podemos listar alguns papéis essenciais que os dados desempenham em qualquer projeto envolvendo IA:**



Treinamento de Modelos:

- Aprendizado supervisionado: os modelos de IA são treinados usando grandes conjuntos de dados que incluem entradas e saídas esperadas. Esses dados permitem que o modelo aprenda a fazer previsões ou tomar decisões baseadas em exemplos anteriores.
- Aprendizado não supervisionado: em tarefas de aprendizado não supervisionado, os modelos de IA são treinados para identificar padrões e relações em conjuntos de dados sem etiquetas predefinidas.
- Aprendizado por reforço: dados sobre interações com o ambiente e feedback sobre o desempenho do modelo são usados para aprender estratégias ou comportamentos.

Avaliação e validação: os dados são usados para validar e testar a precisão de modelos de IA após o treinamento. Isso geralmente é feito usando um conjunto de dados de teste separado, que não foi visto pelo modelo durante o treinamento, para garantir que o modelo possa generalizar bem para novos dados.

Aprimoramento contínuo: os dados coletados a partir das interações reais dos usuários com sistemas de IA podem ser usados para aprimorar continuamente os modelos. Por exemplo,

erros identificados ou novos exemplos que não foram previstos corretamente podem ser adicionados ao conjunto de treinamento para futuras iterações de aprendizado.

Personalização: dados específicos do usuário permitem que sistemas de IA personalizem suas respostas e recomendações.

Quanto mais dados um sistema tem sobre as preferências, comportamentos e histórico de um usuário, mais precisas e personalizadas podem ser as interações.

Detecção de anomalias: em muitas indústrias, os dados são monitorados por sistemas de IA para detectar padrões anormais ou atividades suspeitas. Isso é fundamental para prevenção de fraudes, manutenção preditiva e monitoramento de segurança.

Simulações e modelagem: dados históricos e simulados são usados para modelar cenários complexos e fazer simulações em ambientes controlados. Isso é amplamente utilizado em campos como meteorologia, economia e planejamento urbano.

A qualidade dos dados é crucial. Dados ruins podem levar ao fenômeno de "garbage in, garbage out", onde modelos de IA treinados com dados de má qualidade produzem resultados inúteis ou incorretos. **Dados precisos, diversificados e representativos são essenciais para o sucesso dos sistemas de IA.**

Os dados não são apenas o combustível para a IA; eles são a espinha dorsal que determina o quão bem um sistema de IA pode operar, adaptar-se e evoluir ao longo do tempo. E justamente para sublinhar a importância desta etapa em qualquer projeto de IA, **aqui na TOTVS costumamos sempre repetir o mantra:**

> **"Sem dados íntegros, acessíveis e em prontidão, não tem IA"**

3 IA GENERATIVA, o “HYPE” do momento?

O QUE VOCÊ VAI ENCONTRAR NESTE GUIA

3.1 O que é?

A essência dos modelos de **IA Generativa** (tanto LLMs para geração de textos, quanto modelos para geração de imagem, por exemplo) reside em sua capacidade de avaliar e manipular probabilidades. Esses modelos utilizam redes neurais avançadas (frequentemente chamadas de redes neurais profundas ou deep learning) para aprender padrões em grandes conjuntos de dados. Eles são treinados para gerar novos dados que se assemelham aos dados de treinamento. **Implicações práticas dessa capacidade:**

Na geração de texto: o modelo usa essas probabilidades para criar sequências de texto coerentes e contextualmente apropriadas. Por exemplo, em uma frase onde o contexto sugere uma conversa sobre o clima, o modelo calculará que palavras relacionadas ao clima são mais prováveis de seguir do que palavras relacionadas a tópicos irrelevantes.

Na diversidade e variedade: a manipulação dessas probabilidades também permite aos modelos de LLM variar o estilo, o tom ou a especificidade do texto gerado. Por exemplo, ajustando a temperatura (um parâmetro que afeta a distribuição de probabilidade durante a geração de texto), pode-se fazer com que o modelo gere respostas mais conservadoras (mais prováveis) ou mais criativas (menos prováveis).

No entendimento e correção: a capacidade de avaliar probabilidades também permite que os LLMs sejam usados para tarefas como completar frases, corrigir gramática, e até mesmo traduzir textos, sempre baseando-se em qual sequência de palavras é estatisticamente mais provável dada uma sequência inicial ou um contexto.

Os modelos de IA Generativa para geração de texto (LLMs) mais avançados utilizam uma arquitetura de rede neural específica conhecida como **“Transformer”** (falaremos mais sobre ela adiante).

Estes modelos são inicialmente pré-treinados em grandes quantidades de texto de forma não supervisionada (aprendendo a prever a próxima palavra em sentenças, por exemplo) e, em seguida, podem ser afinados (fine-tuned) para tarefas específicas.

A **Inteligência Artificial Generativa** ganhou destaque e se tornou o "hit do momento" por várias razões, que refletem tanto os avanços tecnológicos quanto as aplicações práticas e inovadoras. **Aqui estão alguns dos principais fatores que contribuem para sua popularidade:**

Capacidades inovadoras de criação: a IA Generativa é capaz de criar conteúdo novo e único, como textos, imagens, músicas e vídeos, que são frequentemente indistinguíveis dos criados por humanos.

Essa capacidade de "criar" em vez de simplesmente "analisar" ou "prever" captura a imaginação do público e das empresas, sugerindo um futuro onde a IA não é apenas uma ferramenta analítica, mas também um parceiro criativo.

Desenvolvimentos significativos em modelos específicos: alguns Modelos têm demonstrado habilidades extraordinárias em gerar texto coerente e imagens criativas a partir de descrições textuais simples. A habilidade desses modelos de gerar conteúdo detalhado e altamente específico atraiu atenção significativa da mídia e interesse comercial.

Aplicações comerciais atrativas: empresas estão encontrando usos práticos para a IA Generativa que podem transformar indústrias. Isso inclui desde a automação de design gráfico e produção de conteúdo até aplicações personalizadas em moda, publicidade e entretenimento. A capacidade de personalizar produtos e serviços em larga escala, sem os custos associados à criação humana tradicional, é particularmente atraente.

Melhorias na acessibilidade e na usabilidade: as plataformas com interfaces cada vez mais amigáveis para os usuários comuns têm tornado a IA Generativa mais acessível a desenvolvedores e criativos sem formação especializada em IA ou programação.

Isso democratiza o poder da IA Generativa, permitindo a uma gama mais ampla de usuários experimentar e implementar suas próprias ideias criativas.

Impacto cultural e social: a IA Generativa também provocou discussões importantes sobre ética, autoria e criatividade. Questões sobre direitos autorais, autenticidade e o papel da máquina na arte e na criação de conteúdo estimulam debates públicos que aumentam a visibilidade e o fascínio por essas tecnologias.

Evolução da tecnologia e investimento: o avanço contínuo em capacidade computacional e algoritmos de aprendizado profundo tem permitido melhorias constantes nas técnicas de IA Generativa. Além disso, o investimento substancial em pesquisa e desenvolvimento por parte de grandes players tecnológicos, como Google, OpenAI e outros, acelera o desenvolvimento e a aplicação dessas tecnologias.

3.2 Modelos de IA Generativa

3.2.1 LLMs, PLN e IA Generativa

Os modelos de LLM (Linguagem de Grande Escala ou Large Language Models) são na verdade a intersecção entre o Processamento de Linguagem Natural (PLN) e a IA Generativa (IAGen).

No PLN, a ênfase é no "Entendimento": compreender, interpretar e analisar a linguagem humana.

Isso inclui tarefas como:

Reconhecimento de fala, onde o objetivo é converter fala em texto compreensível por máquina.

Análise de sentimentos, que envolve determinar a atitude ou emoção expressa em um texto.

Extração de informações, como identificar pessoas, lugares, datas e outros dados específicos em um texto.

Compreensão de linguagem, onde o modelo avalia e interpreta o significado do texto para responder perguntas ou fornecer insights.

Na IAGen, a ênfase é na "Criação": geração de novo conteúdo textual, de tal forma que não seja apenas uma repetição de informações memorizadas, mas uma recombinação criativa de conhecimentos adquiridos que pode resultar em algo novo e original.

Isso inclui tarefas como:

Geração de texto, onde o modelo produz conteúdo completamente novo, como artigos, histórias ou conversas.

Tradução automática, que embora seja um processo de transformação, envolve a criação de texto em um novo idioma que reflete o significado do original.

Sumarização, que cria uma versão mais curta e concisa de um texto existente, destacando seus pontos principais.

Repare, portanto, que a aplicação de um modelo tipo LLM para interações com seres humanos de forma fluida pressupõe simultaneamente 2 abordagens complementares entre si: o uso de Processamento de Linguagem Natural **(PLN)** para assegurar o entendimento da linguagem humana na interação, seguido da IA Generativa **(IAGen)**, capaz de criar conteúdos ou respostas que sejam criativas, contextualizadas e pertinentes para a compreensão humana.

3.2.2 GANs

GANs são um tipo específico de modelo de IA Generativa. Elas são compostas por duas redes neurais que competem uma com a outra em um jogo teórico:

Gerador: esta rede é responsável por gerar novos dados. O objetivo do gerador é criar dados falsos que sejam indistinguíveis dos reais.

Discriminador: esta rede atua como um crítico ou um juiz. Sua função é distinguir entre os dados reais (dados verdadeiros do conjunto de treinamento) e os dados falsos produzidos pelo gerador.

O processo de treinamento de uma GAN é uma espécie de jogo entre o gerador e o discriminador. O gerador tenta "enganar" o discriminador produzindo dados cada vez mais plausíveis, enquanto o discriminador aprende a ficar melhor em detectar falsificações. Esse processo continua até que o gerador se torne tão bom em simular os dados reais que o discriminador não consiga mais distinguir eficazmente entre o real e o falso.

As GANs têm sido utilizadas em uma variedade de aplicações, tais como:

Geração de imagens artísticas: criar novas imagens que imitam estilos de pinturas famosas ou gerar rostos humanos que não existem.

Melhoria de imagem: aumentar a resolução de imagens de baixa qualidade.

Modelagem de moda: criar novos designs de roupas ou experimentar automaticamente roupas em corpos virtuais.

Simulações: gerar dados de treinamento para outras redes neurais, especialmente em cenários onde os dados reais são escassos ou difíceis de coletar.

3.2.3 Outros modelos

Além das próprias GANs, abaixo alguns outros modelos de IA Generativa também muito utilizados em diferentes aplicações:

VAEs (Modelos Autoencoders Variacionais): para tarefas onde é necessária a geração de novas amostras contínuas dos dados (aplicações de imagens e modelagem de músicas);

MGANs (Redes Adversárias Generativas com Memória): permite armazenar exemplos de dados de treinamento e usar esse conhecimento para gerar novos dados mais consistentes e de alta qualidade (aplicações médicas e científicas);

DBNs (Redes Neurais de Crença Profunda): múltiplas camadas treinadas em um processo não supervisionado, seguido de um ajuste fino supervisionado (aplicações para reconhecimento de padrões e classificação, para gerar novos exemplos);

Modelo Transformer Generativo: baseados na arquitetura "Transformer", são aptos a prever um próximo "token" em uma sequência, como uma próxima palavra em uma frase (aplicações para completar prompt de busca em search engines, criação de artigos, chatbots e código de programação).

3.3 Arquitetura Transformer – e por que ela é tão importante

O termo "Transformer" refere-se a uma arquitetura de modelo de aprendizado de máquina (Machine Learning) que foi introduzida em 2017 por Vaswani et al. no artigo intitulado "Attention is All You Need".

Desde então, essa arquitetura revolucionou o campo do processamento de linguagem natural (PLN) e se tornou a base para a maioria dos LLMs, como o GPT (Generative Pre-trained Transformer) da OpenAI, BERT (Bidirectional Encoder Representations from Transformers) do Google, e muitos outros.

A arquitetura Transformer possibilitou avanços significativos na qualidade e eficiência dos modelos de PLN, tornando possível o treinamento de modelos muito grandes que podem generalizar bem de um conjunto de tarefas de NLP para outro. Essa capacidade generalista é parte do que torna os LLMs baseados em Transformer tão poderosos e versáteis em uma variedade de aplicações de PLN.

3.4 IA Generativa Multimodal

IA Generativa multimodal é um conceito avançado na área de Inteligência Artificial que se refere à capacidade de um sistema de IA de entender e gerar conteúdo em mais de um tipo de modalidade de mídia. Estas modalidades podem incluir texto, imagem, som, vídeo, e outros tipos de dados sensoriais.

Os desafios técnicos são grandes. Desenvolver métodos robustos para analisar e sintetizar informações de diferentes modalidades de dados é desafiador devido à complexidade e à variabilidade das informações que cada modalidade pode conter.

O treinamento de modelos requer grandes conjuntos de dados anotados multimodalmente, o que pode ser difícil e caro de se obter. Além disso, sistemas multimodais podem exigir muito mais recursos de processamento e armazenamento.

Atualmente, todas as soluções de IA Generativa convergem para uma abordagem Multimodal, virando praticamente premissa de qualquer provedor.
O objetivo da IA Generativa multimodal é criar sistemas que não apenas compreendam cada tipo de mídia em isolamento, mas também sejam capazes de integrar e correlacionar informações entre diferentes modalidades de maneira coesa e útil. **Suas características principais são:**

Integração de modalidades: sistemas multimodais são capazes de processar e relacionar informações de várias fontes sensoriais ou tipos de dados ao mesmo tempo. Por exemplo, um sistema pode analisar tanto o texto quanto as imagens em um documento para entender melhor o conteúdo e o contexto.

Geração cruzada de conteúdo: um sistema multimodal pode gerar uma modalidade de conteúdo com base em outra. Por exemplo, pode criar uma descrição textual de uma imagem (uma tarefa conhecida como geração de legenda) ou gerar uma imagem a partir de uma descrição textual.

Enriquecimento de dados: esses sistemas podem enriquecer uma modalidade de dados com informações derivadas de outra. Por exemplo, melhorar a compreensão de um vídeo adicionando metadados textuais extraídos do áudio e da imagem.

Aqui, algumas das principais aplicações da IA Generativa Multimodal:

Geração de conteúdo multimídia: como criar apresentações ou conteúdos para redes sociais que combinam texto, imagem e vídeo de maneira harmoniosa e contextualmente relevante.

Assistentes virtuais avançados: desenvolver assistentes que não só entendem comandos de voz, mas também podem interpretar expressões faciais e linguagem corporal para oferecer respostas mais contextualizadas e humanizadas.

Educação e treinamento: criar materiais de aprendizado que combinam texto, diagramas, e simulações interativas, adaptando-se às respostas e ao engajamento do usuário.

Análise de mídia social: analisar posts que contêm texto, imagens e vídeo para entender melhor as tendências, sentimentos e contextos subjacentes.

Robótica e interfaces homem-máquina: melhorar a interação entre humanos e máquinas permitindo que sistemas robóticos interpretem uma gama mais ampla de comandos sensoriais e respondam de maneira mais natural.

3.5 Agentes de IA Generativa

O termo **“Agentes de IA Generativa”** (ou “Generative AI Agents”) refere-se a sistemas de inteligência artificial que utilizam técnicas de IA Generativa para responder ou realizar uma ação a pedido do usuário.

Esses agentes combinam o conhecimento e as habilidades adquiridas por meio do treinamento em vastos conjuntos de dados para gerar respostas, soluções ou produtos que não são simplesmente repetições de exemplos passados, mas sim novas criações baseadas em padrões e informações aprendidas.

Criação de conteúdo: os agentes são capazes de gerar novos conteúdos que não são meras cópias dos dados de treinamento, mas sim recombinações criativas de informações aprendidas.

Autonomia: eles operam com um certo grau de independência, tomando decisões com base nas informações processadas sem intervenção humana constante.

Adaptabilidade: podem adaptar suas criações a diferentes contextos e necessidades, ajustando estilo, tom e complexidade do conteúdo gerado.

A escolha entre chamar esses sistemas de “Agentes de IA Generativa” ou “Agentes de LLM” depende do contexto e do escopo de suas funcionalidades. Ambos os termos têm validade, mas destacam diferentes aspectos dos sistemas:

“Agentes de IA Generativa” ou “Agentes de LLM”?

Agentes de IA Generativa: enfoque mais amplo. Este termo é mais abrangente e inclui qualquer sistema de IA que gera novos conteúdos ou dados que não são simplesmente cópias dos dados de treinamento. Isso inclui, mas não se limita a, texto, imagens, música, e até mesmo comportamentos em ambientes simulados. Diversidade de técnicas: Abrange uma variedade de métodos, incluindo, mas não limitado a, LLMs, podendo portanto incluir modelos para imagens, modelos de música, entre outros.

Agentes de LLM: enfoque específico em linguagem. Este termo se concentra em agentes que utilizam modelos de linguagem de grande escala para processar e gerar linguagem natural. Isso ressalta a capacidade do agente de entender e produzir texto humano de maneira coerente e contextualmente relevante. Limitado a modelos de texto e linguagem.

3.6 Tokens, Embedding e GPUs

Nos casos de uso de IA Generativa, os textos a serem processados pelos modelos de Processamento de Linguagem Natural (PLN) geralmente passam antes por um processo chamado tokenização, onde o texto é dividido em pedaços menores chamados tokens.

Esses tokens podem ser palavras inteiras, partes de palavras (como em tokenização baseada em subpalavras), ou até mesmo símbolos de pontuação.

Funcionam como unidades de processamento.

Na maioria dos modelos de processamento de linguagem natural (PLN), um token é em seguida convertido em um vetor. Esse processo é conhecido como “Embedding”. Esses vetores resultantes, são armazenados em um “banco vetorial”, onde ficam, portanto, acessíveis para que os modelos de PLN possam processá-los. **Veja o exemplo de vetorização da palavra “TOTVS”:**

Vetor “Totvs”=[0.85,–0.22,0.15,0.05,–0.07,0.33,0.18,–0.55]

Neste vetor:

Cada número representa uma característica latente da palavra “Totvs”, aprendida a partir de um grande conjunto de dados de texto que inclui discussões sobre tecnologia, software empresarial, e negócios.

A primeira dimensão (0.85) pode indicar uma forte associação com tecnologia ou inovação.

A última dimensão (-0.55) poderia representar uma distinção de contextos não relacionados, como alimentação ou entretenimento.

Os embeddings são uma parte fundamental dos modelos de linguagem de grande escala (LLMs), pois em qualquer modelo de aprendizado de máquina, incluindo os próprios LLMs, a entrada precisa ser numérica para que possa ser processada por redes neurais.

Sem essa transformação, seria impossível aplicar algoritmos de aprendizado de máquina ao texto.

Os embeddings não apenas convertem texto em números, mas também incorporam informações sobre o significado das palavras e suas relações com outras palavras. Isso é crucial para que os LLMs possam entender e gerar linguagem de maneira eficaz. Por isso, os vetores criados não são aleatórios: eles são projetados de modo que distâncias e direções no espaço vetorial representem relações semânticas e sintáticas para que o modelo de PLN possa interpretá-las de forma lógica.

GPU, que significa Unidade de Processamento Gráfico (do inglês, Graphics Processing Unit), é um componente de hardware especializado em processar operações gráficas e de imagem de maneira rápida e eficiente. GPUs são amplamente conhecidas por sua habilidade em lidar com cálculos complexos e paralelos, o que as torna ideais para jogos, renderização gráfica, e, mais recentemente, aplicações em aprendizado de máquina e processamento de dados.

Elas são especialmente valiosas no treinamento e execução de modelos de IA Generativa devido à sua capacidade de realizar muitas operações em paralelo. Os modelos de IA Generativa modernos, como os baseados em arquiteturas “Transformer”, envolvem muitos cálculos matemáticos simultâneos — especialmente operações de multiplicação de matrizes — que são necessários para processar os tokens. GPUs podem acelerar significativamente esses cálculos devido à sua arquitetura paralela.

Lembrando que, uma vez que cada token (originário de um fragmento do texto a ser processado) precisa ser transformado em um vetor numérico (técnica de embedding), seu processamento é feito por meio de várias camadas do modelo para entender o contexto, gerar previsões ou otimizar a função de perda do modelo.

Nesse contexto, GPUs facilitam esse processo ao lidarem com grandes volumes de tokens (e consequentemente vetores) de forma eficiente, reduzindo o tempo necessário para o treinamento e a inferência.

3.7 Principais técnicas para aumentar a acurácia

Para melhorar a precisão dos modelos de linguagem de grande escala (LLMs), várias técnicas são empregadas. Cada uma delas aborda aspectos diferentes do desempenho e da funcionalidade dos modelos.

3.7.1 RAG (Retrieval Augmented Generation)

É a técnica que combina a geração de texto de um LLM com um sistema de recuperação de informações. O modelo primeiro consulta uma base de dados ou um conjunto de documentos relevantes, em geral vetorizados e armazenados num banco vetorial, para encontrar informações que são usadas como contexto mais preciso na geração de respostas.

No instante da geração de respostas, o modelo consulta documentos relacionados à consulta do usuário e utiliza essas informações para informar e guiar a resposta do modelo, ajudando a aumentar a precisão e relevância factual das saídas.

3.7.2 Injeção de conhecimento

Inserir conhecimento específico no modelo durante o treinamento para ajudá-lo a melhorar em áreas onde pode faltar dados detalhados.

Por exemplo, modelos podem ser pré-treinados com manuais técnicos, livros didáticos ou outros materiais ricos em informações para aumentar seu conhecimento em áreas específicas.

3.7.3 Grounding (ancoragem)

Grounding, ou ancoragem, é o processo de vincular a geração de um LLM a fontes de dados confiáveis e específicas para assegurar que a informação gerada seja precisa e fundamentada na realidade.

Isso pode envolver a integração de bancos de dados, APIs externas ou outros recursos confiáveis diretamente no pipeline de processamento do modelo, para que as respostas geradas reflitam informações verídicas e atualizadas.

3.7.4 Fine Tuning

Fine-tuning (ajuste fino) é o processo de treinar um modelo pré-treinado em um conjunto de dados específico ou tarefas específicas para adaptar suas respostas às necessidades particulares de uma aplicação.

Isso é frequentemente feito treinando o modelo em novos dados que representam o tipo de perguntas ou conteúdo que ele encontrará em uso prático, o que pode melhorar significativamente sua precisão e desempenho em cenários específicos.

3.7.5 Avaliação e Feedback contínuos

Implementar sistemas de feedback onde os usuários ou avaliadores podem classificar a qualidade e a precisão das respostas do modelo. Esses dados de feedback podem então ser usados para ajustar e melhorar continuamente o modelo, seja por meio de retreinamento ou ajustes finos.

3.8 Alucinação

O termo “alucinação” em Inteligência Artificial Generativa refere-se a quando um modelo de IA gera informações que não são verdadeiras ou que não têm base nos dados de entrada fornecidos. Esse fenômeno é particularmente comum em modelos de linguagem e de imagem, que são projetados para produzir saídas coerentes e convincentes com base no aprendizado de padrões em grandes conjuntos de dados. **Aqui estão alguns pontos importantes sobre o porquê de se falar em alucinação nesse contexto:**

Interpolação vs. Invenção: enquanto os modelos de IA são excelentes para identificar padrões e interpolá-los para criar respostas plausíveis, eles também podem gerar conteúdo completamente novo ou inesperado, que não reflete a realidade ou os dados de entrada. Essa geração pode ser vista como uma “alucinação”, pois o modelo “imagina” detalhes ou informações que não foram especificamente indicados.

Limitações do modelo: os modelos de IA Generativa podem alucinar porque eles não “entendem” o mundo real da mesma maneira que os humanos. Eles operam com base em correlações estatísticas em seus dados de treinamento. Portanto, se algo é comum nos dados de treinamento, pode ser gerado frequentemente, mesmo quando não é relevante ou verdadeiro para uma nova entrada específica.

Qualidade e viés dos dados de treinamento: se os dados de treinamento contiverem erros, imprecisões ou viés, o modelo pode aprender essas características e refleti-las em suas saídas. Isso pode levar a alucinações que são reflexos distorcidos ou exagerados de padrões menos comuns ou atípicos nos dados de treinamento.

Complexidade da geração de conteúdo: em tarefas complexas, como a geração de um artigo ou uma imagem detalhada, os modelos podem precisar “preencher lacunas” onde a informação específica não está disponível ou é ambígua. Isso pode resultar em criações que são logicamente consistentes com os dados de treinamento, mas factualmente incorretas ou enganosas.

Desafios de controle e previsibilidade: gerenciar e prever quando e como as alucinações ocorrem é um desafio técnico significativo. Desenvolvedores e pesquisadores continuam a buscar maneiras de reduzir a frequência e a severidade das alucinações em modelos de IA, por meio de melhorias na arquitetura dos modelos, técnicas de treinamento, e curadoria de dados mais rigorosa.

Portanto, falar em alucinações em IA Generativa é uma maneira de reconhecer e lidar com as limitações e desafios que esses sistemas enfrentam ao tentar criar conteúdo novo e significativo de maneira autônoma.

3.9 Dilemas morais e éticos

A Inteligência Artificial (IA) apresenta uma série de dilemas éticos e morais que são essenciais para serem considerados à medida que a tecnologia continua a evoluir e se integrar mais profundamente na sociedade. **Aqui estão alguns dos principais dilemas éticos e morais enfrentados pela IA atualmente:**

Viés e discriminação: modelos de IA podem perpetuar ou até mesmo amplificar vieses existentes nos dados de treinamento, resultando em preconceitos contra certos grupos. Isso pode levar a decisões discriminatórias em áreas críticas como contratação, empréstimos, justiça e saúde.

Privacidade: sistemas de IA dependem de grandes quantidades de dados, que podem incluir informações pessoais sensíveis. O uso indevido ou a exposição desses dados podem violar a privacidade dos indivíduos.

Transparência e explicabilidade: alguns modelos de IA, especialmente aqueles baseados em aprendizado profundo, são

frequentemente considerados “caixas pretas”, pois suas decisões não são facilmente compreensíveis pelos humanos.

Autonomia e emprego: a automação alimentada por IA tem o potencial de substituir empregos e criar grande impacto na eficiência da cadeia produtiva, levantando preocupações sociais sobre desigualdade e perda de autonomia no trabalho.

Segurança: sistemas de IA podem ser usados para fins maliciosos, ou podem ter falhas de segurança que os tornam vulneráveis a ataques. Isso pode ter graves consequências, especialmente em sistemas críticos como transportes e infraestrutura crítica.

Responsabilidade: determinar a responsabilidade por ações ou decisões tomadas por sistemas de IA é complexo. Veículos autônomos, por exemplo. Isso pode incluir erros que resultam em danos materiais ou mesmo perda de vidas.

Desenvolvimento e uso militar: à medida que modelos de IA estão sendo incorporados em sistemas de armas e estratégias militares, levantam-se questões éticas sobre o papel da IA em conflitos armados em detrimento a princípios elementares e inegociáveis de proteção e preservação da vida humana, por exemplo.

Consentimento e manipulação: sistemas de IA, como algoritmos de recomendação e personalização, podem ser usados para manipular comportamentos e opiniões, levantando preocupações sobre a autonomia e o consentimento dos indivíduos.

Propriedade Intelectual – este problema merece uma abordagem ainda mais detalhada:

- Autoria e criação: quando um sistema de IA gera arte, música, texto ou qualquer outra forma de expressão criativa, surgem questões sobre quem detém os direitos autorais da obra. A legislação atual em muitos países geralmente não reconhece obras criadas por máquinas como passíveis de direitos autorais, uma vez que não têm um “autor” humano.

- Inovações e patentes: sistemas de IA que desenvolvem novas invenções ou descobrem novos processos podem também levantar questões sobre a patenteabilidade dessas inovações. Tradicionalmente, patentes requerem um inventor humano, e muitos escritórios de patentes estão debatendo se invenções geradas por IA podem ser patenteadas.

- Uso de dados para treino: a IA muitas vezes é treinada com vastas quantidades de dados, incluindo dados protegidos por direitos autorais. O uso desses dados pode levantar questões legais sobre violações de direitos autorais, especialmente quando os dados não são usados apenas para treinamento, mas também estão sendo replicados ou imitados pela IA.

- Transparência e rastreabilidade: determinar a origem de uma obra ou invenção criada por IA pode ser desafiador, especialmente quando esses sistemas são caixas-pretas que não revelam facilmente como as decisões são feitas ou de onde os elementos individuais de uma criação vieram.

Todos esses dilemas apontados acima não pretendem constituir uma lista exaustiva, diante de um cenário dinâmico em que nos deparamos com novidades todos os dias. Eles não apenas desafiam os desenvolvedores e usuários de tecnologia de IA, mas também legisladores, reguladores e a sociedade como um todo.

Abordar esses problemas requer um esforço colaborativo para desenvolver diretrizes éticas, regulamentações e políticas que definam o uso responsável da IA.

Aqui na TOTVS, por exemplo, já adotamos um manual de boas práticas para uso responsável de IA e estamos elaborando uma Política definitiva de IA que poderá integrar o Código de Conduta da companhia.

4 CASOS DE USO PARA IA GENERATIVA

O QUE VOCÊ VAI ENCONTRAR NESTE GUIA

Avaliar as aplicações de **Inteligência Artificial Generativa** requer a definição de Key Performance Indicators (KPIs) específicos para cada área de aplicação. A seguir alguns exemplos de casos de uso e sugestões de três KPIs relevantes para cada uma das aplicações listadas, acompanhados de benchmarks*.

4.1 Geração de conteúdo de texto

Exemplo:
modelos como GPT-3 e GPT-4 da OpenAI são usados para gerar conteúdo textual criativo e informativo, incluindo artigos, blogs, scripts, e até livros.

Indústrias aplicáveis:
mídia, publicidade, educação e entretenimento.

KPIs recomendados:
- Taxa de Engajamento do Usuário - mede a interação dos usuários com o conteúdo gerado.
- Precisão do Conteúdo - avalia a correção factual e relevância do conteúdo.
- Eficiência na Geração - tempo necessário para gerar conteúdo.

Benchmarks:
- Taxa de engajamento - +50% em relação ao conteúdo não gerado por IA.
- Precisão do conteúdo - +90% de precisão factual.
- Eficiência na geração: redução de 50% no tempo de produção comparado ao processo manual.

4.2 Geração de imagens e arte

Exemplo:
modelos como DALL-E e outros baseados em técnicas de Redes Generativas Adversariais (GANs) que criam imagens artísticas ou fotográficas a partir de descrições textuais.

Indústrias aplicáveis:
arte, design de moda, arquitetura e publicidade.

KPIs recomendados:
- Qualidade visual - Avaliada por especialistas ou por meio de feedback do usuário.
- Originalidade - medida de unicidade em relação a um conjunto de dados de referência.
- Taxa de aceitação pelo cliente - percentual de imagens aceitas pelos clientes na primeira submissão.

Benchmarks:
- Qualidade visual - 85% de avaliações positivas.
- Originalidade - índice de novidade acima de 80%.
- Taxa de aceitação - superior a 75% na primeira submissão.

4.3 Assistentes virtuais avançados

Exemplo:
chatbots e assistentes virtuais que usam IA Generativa para fornecer respostas mais naturais e contextualizadas.

Indústrias aplicáveis:
atendimento ao cliente, saúde, finanças, e comércio eletrônico.

KPIs recomendados:
- Taxa de resolução no primeiro contato (FCR) - resoluções sem transferência para suporte humano.
- Satisfação do cliente - medido por meio de pesquisas pós-interação.
- Taxa de contenção - chamadas contidas dentro do sistema de IA.

Benchmarks:
- FCR - superior a 70%.
- Satisfação do cliente - mais de 80% de satisfação.
- Taxa de contenção - mais de 75%.

4.4 Geração de código automatizado

Exemplo:
ferramentas como GitHub Copilot que sugerem ou geram código fonte para auxiliar programadores.

Indústrias aplicáveis:
desenvolvimento de software e educação em programação.

KPIs recomendados:
- Precisão do código - medida de erros ou bugs por linhas de código.
- Eficiência do desenvolvimento - redução no tempo de desenvolvimento.
- Satisfação do desenvolvedor - avaliada por meio de feedback direto dos programadores.

Benchmarks:
- Precisão do código - menos de 0.5% de erros.
- Eficiência do desenvolvimento - redução de 50% no tempo.
- Satisfação do desenvolvedor - mais de 85% de satisfação.

Estas aplicações demonstram a versatilidade e o potencial transformador da IA Generativa, permitindo a criação e a inovação em uma escala que era inimaginável até recentemente.

*Os benchmarks apresentados são referências iniciais e gerais, podendo variar significativamente entre diferentes segmentos. Os dados estão sujeitos a atualizações à medida que novos testes e estudos sobre esses casos de uso se tornam disponíveis no mercado, aprimorando a precisão das estimativas.

5 Como tratamos/ priorizamos o tema de DADOS & IA NA TOTVS?

O QUE VOCÊ VAI ENCONTRAR NESTE GUIA

Na TOTVS somos orientados pela construção de valor sobre dados e chamamos isso de **"Cultura Data-Value-Driven".** Aqui, um datalake para centralizar e armazenar volumes infinitos de dados jamais será o protagonista da estratégia. O protagonista é o caso de uso, que se vale dos dados para destravar valor percebido para nossos clientes.

A IA tem um papel fundamental nessa construção de valor, seja aplicada aos dados para alcançarmos ganhos de eficiência operacional, seja aplicada aos dados para criar valor percebido ao mercado em forma de novas funções em nossos softwares de gestão, de soluções financeiras ou de otimização de marketing e vendas. Portanto, aqui na TOTVS, Dados & IA são o meio e não o fim.

Isso exige uma mentalidade adequada de toda a organização, com comportamentos e atitudes permanentes dos nossos colaboradores nas diferentes unidades de negócio e áreas, em direção a uma cultura orientada a dados. **Em nosso playbook de Dados & IA, estas são as premissas da TOTVS:**

5.1 Caso de uso em primeiro lugar

De nada adianta ter dados ou aplicar IA se não há construção de valor objetiva (o petróleo só tem valor quando vira gasolina disponível na bomba de gasolina). Portanto colocamos o caso de uso e o valor percebido como critério central primário em qualquer projeto.

5.2 Plataforma de dados a serviço de todos

Somos orientados pela construção de valor sobre dados, incorporando uma verdadeira cultura "Data-Value-Driven". A plataforma self-service centraliza os dados, garantindo ao mesmo tempo escalabilidade e segurança. Nosso lema: sem dados íntegros, acessíveis e em prontidão, não tem IA.

5.3 Governança federada

Viabilizamos uma abordagem descentralizada e flexível entre diferentes domínios autônomos (áreas ou unidades da companhia que utilizam dados em larga escala nos seus próprios casos de uso), com diferentes características entre si. Asseguramos os guardrails para garantirmos diligência em nossas iniciativas, sem perdermos a agilidade e velocidade nos negócios.

5.4 IA escalável, segura e com propósito específico

Facilitamos para todos os produtos e serviços TOTVS os melhores recursos de Inteligência Artificial disponíveis, **organizados em 3 grandes pilares:**

Content&Knowledge: casos de uso associados a aplicação de IA que se valem de bancos de conhecimento ou conteúdo;

Build&Code: casos de uso associados a aplicação de IA dedicadas à criação e/ou revisão de Código Fonte em diferentes linguagens;

Live&Action: casos de uso associados a aplicação de IA que exigem consultas ou processamento de dados em tempo real;

5.5 DTA

A orquestração técnica de aplicações de IA nesses 3 pilares materializam a plataforma que batizamos de **DTA (Digital Trusted Advisor)**. Essa plataforma busca essencialmente facilitar, de forma escalável e segura, as **3 principais etapas de viabilização de uma aplicação:**

- Acessar aos principais **modelos de IA disponíveis no mercado;**
- Assegurar **governança no uso desses modelos;**
- Garantir o **controle de consumo orçamentário,** com o custo das requisições de tokens dos diferentes modelos usados.

Com o TOTVS DTA, nós viabilizamos a IA para aprimorarmos a nossa eficiência operacional e a capacidade de tomada de decisão de nossos clientes.

6 Os 5 passos para começar um projeto de IA NA SUA EMPRESA

O QUE VOCÊ VAI ENCONTRAR NESTE GUIA

1º Identificação de necessidades e oportunidades

Avalie onde a IA Generativa pode agregar valor à sua empresa. Isso pode incluir automação de tarefas criativas, personalização de produtos ou serviços, e inovação em processos existentes. **Um bom modelo para servir como guia desta etapa envolve o seguinte:**

1. Defina qual a dor (ou pain points) que se pretende resolver e/ou as respectivas frustrações e/ou sentimentos negativos decorrentes dessa dor;

2. Aponte agora as principais implicações objetivas para o seu negócio (exemplo: ineficiência operacional, tempo excessivo consumido por posições sênior em tarefa operacional, insatisfação dos clientes ou prospects, etc.);

3. Defina a persona para quem se pretenda resolver o problema por meio de uma solução baseada em IA Generativa, assim fica mais fácil projetar uma abordagem suficientemente adequada para o perfil de usuário pretendido (ex: advogados do departamento jurídico).

4. Proponha a hipótese de solução, discriminando detalhadamente o que se espera ser executado pela aplicação de IA Generativa, ou seja, os "jobs to be done". Aqui é hora de especificar cada ação esperada durante toda jornada da solução até seu desfecho.

> **Por exemplo:**
> em um departamento jurídico, um "Assistente de Revisão de Contratos" baseado em IA Generativa deve: possuir uma interface amigável o upload do arquivo a ser revisado, depois confirme o conteúdo do arquivo recebido que será revisado, em seguida proceda a revisão consultando uma base de conhecimento previamente disponível, para depois apontar os respectivos itens que são sensíveis na negociação ou que estão em desconformidade com a política comercial ou compliance, depois disponibilize o arquivo revisado em um formato pré-estabelecido e, finalmente, solicite um feedback do usuário em relação a qualidade da revisão constatada a fim de retroalimentar o conhecimento)

5. Eleja os KPIs que reflitam adequadamente o impacto de uma solução sobre as implicações identificadas no item anterior;

Repare que nesse primeiro Passo não estamos entrando em temas técnicos da solução. **O objetivo nesta etapa é mapear fielmente a regra de negócio pretendida,** ou seja, as necessidades, a partir das dores identificadas. O conceito "jobs to be done" (desenvolvido em Harvard pelo Prof. Clayton M. Christensen) se propõe a nos ajudar a focar primariamente no mapeamento do problema, ou seja, "o que se pretende exatamente resolver", para só depois prosseguirmos com o "como resolver".

2º

Capacitação e formação de equipes

Invista na capacitação de sua equipe para entender e aplicar tecnologias de IA Generativa. Isso pode envolver treinamentos internos ou parcerias com instituições educacionais.

As alternativas para se encontrar a melhor abordagem são múltiplas. É possível tirar proveito de modelos de IA Generativa totalmente prontos para uso e que se prestam especificamente a um caso de uso (os chamados "Agentes"), ou usar modelos pré-treinados para serem adaptados a um caso de uso específico e receberem um refinamento do treinamento sobre os seus dados corporativos, exclusivamente.

Os modelos podem ser encontrados, por exemplo, nos principais serviços de Cloud do mercado ou em plataformas colaborativas onde modelos open source pré-treinados estão disponíveis.

O melhor caminho a seguir dependerá da sua estratégia de longo-prazo e do nível de especialização do seu time.

Seu time precisa, portanto, estar familiarizado com as alternativas de mercado, ou pelo menos conhecer muito bem aquilo que seu fornecedor de Cloud disponibiliza. Contar com o apoio de uma consultoria especializada homologada pela plataforma utilizada pode ser um caminho alternativo, especialmente para a etapa de prototipagem e setup de infraestrutura.

3º

Seleção de Tecnologia e Ferramentas

Nesta etapa é hora de propor detalhadamente a hipótese de solução, sempre baseada nas definições do Passo 1 acima. Escolha as ferramentas e plataformas de IA Generativa que melhor se adequem às suas necessidades específicas. Isso pode variar desde soluções prontas até desenvolvimento personalizado com a ajuda de especialistas em IA. **Uma boa abordagem para essa etapa deve levar em consideração:**

1. Identifique os principais vetores que impactarão no custo de desenvolvimento e na manutenção da solução: se for possível estimar os custos envolvidos e projetar uma eventual margem de contribuição proporcionada pela solução, tanto melhor. Repare que aqui é fundamental levar em conta a estimativa de requisição de tokens junto ao modelo de IA Generativa, pois cada requisição representará centavos que, na escala, terão impacto determinante no custeio total da aplicação (e encontrar alternativas que aliviam o volume de requisições, bem como ter uma boa negociação com seu vendedor, são premissas para um projeto bem sucedido). Aprofundar a análise financeira, estimando sua viabilidade a partir de VPL, TIR e Payback é de grande valia neste momento para se evitar um gasto de energia importante da sua equipe em um projeto que já se mostra economicamente inviável na largada;

2. Defina a sua estratégia para utilização do modelo de IA Generativa: **de maneira simplificada existem 3 grandes estratégias possíveis.**

Estratégia "Taker"

A solução se vale de um "Agente de IA" totalmente pronto e treinado para um caso de uso específico. É a forma mais rápida e "plug-and-play" de se adotar uma solução de IA Generativa, no entanto ainda existe uma baixa oferta de Agentes totalmente prontos e é preciso confirmar se os Agentes disponíveis a que você tem acesso são perfeitamente adequados ao seu caso de uso;

Estratégia "Maker"

A solução utiliza o framework de algum vendor de mercado (um fornecedor de serviço Cloud, por exemplo), para acessar modelos de IA Generativa pré-treinados (de propriedade do próprio vendor ou de terceiros), para adaptá-los a um treinamento sobre uma base de dados específica da sua empresa. Há que se ter um time capacitado para essa tarefa, naturalmente, mas é possível encontrar apoio consultivo externo de empresas especializadas;

Estratégia "Shaper"

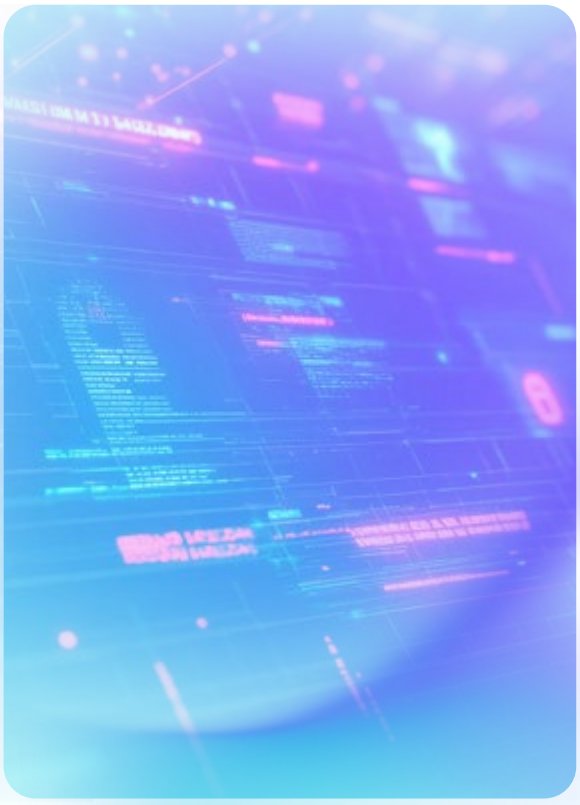
A solução pretendida é tão única e estratégica que você decide criar seu próprio modelo de IA Generativa do zero. Essa abordagem é altamente complexa e só tem pertinência no caso de se visualizar uma oportunidade de mercado única, onde uma solução totalmente proprietária poderia significar uma vantagem competitiva determinante para um movimento de go-to-market, ofertando essa solução em escala a um mercado endereçável relevante.

3. Defina a sua fundação de dados: não existe "almoço grátis" numa jornada com IA, e o ponto de partida é garantir a disponibilidade dos dados proveniente de todas as fontes envolvidas em um repositório pré-definido, que será acessado pelo modelo de maneira permanente.

Os dados precisam estar íntegros, acessíveis, seguros e em prontidão em um data warehouse, datalake, data mart ou arquitetura mais conveniente ao projeto.

Tanto melhor se essa arquitetura estiver pré-definida e já em produção para toda sua empresa, a serviço deste e de qualquer outro futuro caso de uso com IA

4. Levante os requisitos de infraestrutura: para suportar sua arquitetura e garantir a melhor performance para servir sua aplicação, há que se contratar recursos compatíveis que precisam ser previamente detalhados junto ao seu fornecedor de Cloud. Como já apontamos acima, este é um dos principais vetores de custo do projeto, sendo determinante para seu sucesso.



4º

Desenvolvimento e implementação de projetos piloto

Implemente projetos piloto para testar a eficácia da IA Generativa em pequena escala antes de uma implementação mais ampla. Isso ajuda a identificar desafios e ajustar estratégias de forma eficiente.

Aqui é o momento de se viabilizar um protótipo funcional, usando os modelos de IA Generativa disponíveis na prateleira dos principais vendors de mercado.

O Protótipo precisa ser útil para validar o conceito numa versão "MVP" (Mínimo Produto Viável), de tal forma que seja possível, mesmo que em pequena escala, confirmar as hipóteses de funcionamento planejadas e avaliar o impacto nos indicadores de sucesso pré-definidos.

5º

Avaliação de resultados e escala

Avalie os resultados dos projetos piloto e refine as abordagens conforme necessário antes de prosseguir. Os indicadores de sucesso pré-definidos precisam refletir se as implicações da "dor" de negócio original estão sendo endereçadas (exemplo: eficiência operacional obtida, custos envolvidos, satisfação do cliente, percentual de encurtamento ou simplificação de um workflow, etc.).

Além da avaliação dos benefícios obtidos, ainda é importante considerar na sua avaliação os seguintes aspectos:

Avaliação de custos envolvidos: licença, infraestrutura, operação, equipe dedicada, etc.;

Retorno do investimento: taxa e tempo de retorno, considerando-se a margem proporcionada ao longo do tempo do projeto em operação;

Impacto dos riscos associados: implicações éticas envolvidas nas respostas da IA Generativa, vazamento de dados corporativos, etc.;

Alinhamento com a estratégia corporativa: considere se o projeto está em linha com a agenda estratégica da Alta Gestão ou Conselho de Administração (exemplo: Centralidade no Cliente, Corte de Custos, etc.);

Flexibilidade de expandir para novas oportunidades: avalie se o seu projeto, se bem-sucedido, pode depois ser estendido para outras oportunidades adjacentes com relativa facilidade;

Feedback de usuários: considere as manifestações de seus clientes ou usuários que ficaram expostos à aplicação durante o período de testes.

Uma vez validada a eficácia, comece a escalar o uso de IA Generativa em toda a organização para maximizar o impacto. Em resumo, estes passos fornecem uma estrutura para incorporar com sucesso a IA Generativa em sua empresa, impulsionando a inovação e a competitividade no mercado.

7 GLOSSÁRIO

A

Aprendizado por Reforço

Modelo de IA em que agentes aprendem a tomar decisões por meio de tentativa e erro para maximizar a soma de recompensas recebidas.

B

Backpropagation

Algoritmo fundamental para treinar redes neurais, ajustando os pesos para minimizar a diferença entre as previsões e os resultados reais.

Bias

Tendência ou erro sistemático nos dados ou no algoritmo que pode resultar em previsões injustas ou incorretas.

Big Data

Conjuntos de dados extremamente grandes que desafiam processos analíticos convencionais devido ao seu volume, velocidade e variedade.

D

Data Augmentation

Técnica de aumentar o volume de dados de treinamento através da introdução de cópias modificadas para aumentar a diversidade e robustez do modelo.

Deep Learning (Aprendizado Profundo)

Subset de aprendizado de máquina que usa redes neurais com várias camadas para aprender representações de dados com vários níveis de abstração.

Deep Learning Frameworks

Plataformas e bibliotecas que facilitam a construção, treinamento e validação de modelos de aprendizado profundo.

DeepFake

Técnica que utiliza aprendizado profundo para criar vídeos ou áudios falsificados altamente realistas, muitas vezes para simular a aparência e a voz de pessoas reais.

Dropout

Método de regularização em redes neurais que envolve a desativação aleatória de neurônios durante o treinamento para prevenir o overfitting.

E

Ética em IA

Campo de estudo dedicado às questões morais que surgem com o desenvolvimento e a utilização de tecnologias de IA, focando em viés, privacidade, responsabilidade e impacto social.

F

F1 Score

Métrica que combina precision e recall em uma única medida harmônica, especialmente útil quando as classes são desbalanceadas.

G

GANs (Redes Generativas Adversárias)

Arquitetura de IA em que duas redes neurais, uma geradora e uma discriminadora, são treinadas simultaneamente para gerar novos dados realistas.

I

IA forte

IA teórica com habilidade de realizar qualquer tarefa cognitiva que um humano possa, com potencial para raciocínio e entendimento abstrato.

IA Generativa

Subcampo da IA que foca na criação de conteúdo novo e original, como texto, imagens e sons, usando modelos que aprendem de dados existentes.

IA Geral (General A.I.)

Tipo de inteligência artificial capaz de entender, aprender e aplicar conhecimento em uma ampla gama de tarefas, simular a capacidade cognitiva humana, podendo teoricamente realizar qualquer tarefa intelectual que um ser humano pode fazer, adaptando-se a novos contextos sem intervenção humana direta.

IA Superinteligente

Forma hipotética de inteligência artificial que supera a capacidade intelectual humana em todas as áreas relevantes, incluindo criatividade, conhecimento geral e raciocínio social.

L

Loss Function

Função usada durante o treinamento de um modelo de IA para quantificar o erro entre as previsões do modelo e os resultados esperados.

M

Machine Learning (Aprendizado de Máquina)

Campo da IA que permite a sistemas aprender e melhorar a partir de experiências sem serem explicitamente programados.

Modelagem de música

Aplicação de IA que utiliza modelos para aprender estilos e padrões musicais e gerar novas composições musicais.

N

NLP (Processamento de Linguagem Natural)

Subcampo da IA que se concentra no desenvolvimento de sistemas capazes de entender e responder à linguagem humana de forma útil.

O

Overfitting

Situação em que um modelo de aprendizado de máquina se ajusta demais aos dados de treinamento, perdendo a capacidade de generalizar para novos dados.

P

Precision

Proporção dos verdadeiros positivos em relação a todas as previsões positivas feitas pelo modelo de IA.

Propriedade intelectual

Direitos associados às invenções e criações no domínio da IA, incluindo questões de autoria e titularidade de obras criadas por máquinas.

R

Recall

Proporção dos verdadeiros positivos identificados pelo modelo em relação ao total de casos que são realmente positivos.

Redes neurais

Estruturas computacionais que simulam a maneira como os neurônios humanos interagem, projetadas para reconhecer padrões e realizar tarefas específicas.

S

Segurança em IA

Área focada em garantir que sistemas de IA sejam seguros e estáveis, protegidos contra ataques maliciosos e falhas técnicas.

Síntese de imagem

Geração de imagens visuais a partir de modelos de IA que aprendem a partir de grandes quantidades de dados visuais.

Síntese de texto

Geração de texto natural e coerente a partir de dados estruturados ou não estruturados usando modelos de IA.

Síntese de voz

Tecnologia que converte texto em fala natural, imitando a voz humana, através de técnicas de IA.

T

Tokenization

Processo de dividir um texto em pedaços menores, chamados tokens, preparando-os para processamento em tarefas de NLP.

Transfer learning

Prática de reutilizar um modelo pré-treinado em uma tarefa para acelerar o treinamento em outra tarefa relacionada.

Transformers

Arquitetura baseada em mecanismos de atenção, super eficaz em tarefas de NLP para entender contextos complexos em textos.

U

Underfitting

Condição onde um modelo de IA é demasiado simples para capturar adequadamente a complexidade dos dados de treinamento, resultando em desempenho inadequado.

V

VAEs (Autoencoders Variacionais)

Tipo de rede neural que aprende a representar dados em um espaço latente menor, permitindo a geração de novos dados que são variações dos dados originais.


TOTVS