

Um guia executivo para operacionalizar a IA generativa

Da experimentação à implementação:
como usar a IA generativa em cenários
do mundo real



elastic

The Search
AI Company



Índice

Comece sua jornada na IA generativa 3

Parte 1

Compreendendo o panorama da IA generativa 5

O que é a IA generativa? 6

O que é machine learning? 6

O que é um grande modelo de linguagem (LLM)? 6

O que é geração aumentada de recuperação (RAG)? 8

O que é um banco de dados vetorial? 11

O que a IA generativa pode fazer 12

Etapa 0. Descubra o porquê e determine o que é possível 13

Adaptação ao seu setor 16

Telecomunicações 19

Serviços financeiros 20

Varejo 21

Setor automotivo e de manufatura 23

Setor público 24

Comece aqui se você sabe o que
deseja alcançar com a IA generativa



Parte 2

Operacionalizando a IA generativa: 27

um pequeno passo para a máquina, um salto gigantesco para sua organização

Etapa 1. Identifique seu resultado ideal 28

Etapa 2. Saiba qual será o impacto. Meça o sucesso. 29

Etapa 3. Escolha um modelo (caminho a seguir) 30

Etapa 4. Tente rápido, falhe rápido 34

Etapa 5. Governança e operações 36

A questão da segurança dos dados 37

Etapa 6. Defina um cronograma. Coloque padrões de referência. 38

O início de uma nova era 40

Comece sua jornada na IA generativa

A IA generativa foi a tecnologia mais disruptiva que surgiu em 2023. Em todos os setores, prevê-se que a IA generativa moldará praticamente tudo nos próximos anos — mas quantos podem dizer que decifraram o código para colocá-la em funcionamento agora?

Enquanto as empresas estão começando a surfar a onda da IA generativa, algumas já começaram a ver resultados. Por exemplo, na Cisco, os engenheiros de suporte podem encontrar instantaneamente respostas relevantes e resumidas de casos de suporte semelhantes, fóruns de discussão internos e artigos de conhecimento relacionados a problemas de clientes. A Cisco já colheu os benefícios da IA generativa, com 90% das solicitações de suporte resolvidas com sua solução de busca reimaginada e 5.000 horas de engenharia de suporte economizadas por mês.¹

No campo do e-commerce, você provavelmente já viu a IA generativa em ação. A IA generativa pode analisar as compras anteriores, o histórico de navegação e as preferências de um cliente para gerar recomendações de produtos personalizadas com um chatbot. No backend, o uso de IA generativa promete aumentar o engajamento e a retenção do cliente, melhorar a detecção de fraude e muito mais.

Para desmistificar os recursos da IA generativa e decidir como fazê-la trabalhar para você, você precisa de um guia passo a passo sobre como ativar seus dados. Neste e-book, orientaremos você em sua jornada, **partindo da posição de quem pensa na IA e quer adotá-la e chegando a especialista em IA**. Pense nisso como um roadmap para ajudar a revolucionar os resultados dos seus negócios usando a IA generativa.

99%

das organizações acreditam que a IA generativa tem potencial para impulsionar mudanças dentro da organização, sejam elas internas ou externas²

No entanto, apenas

32%

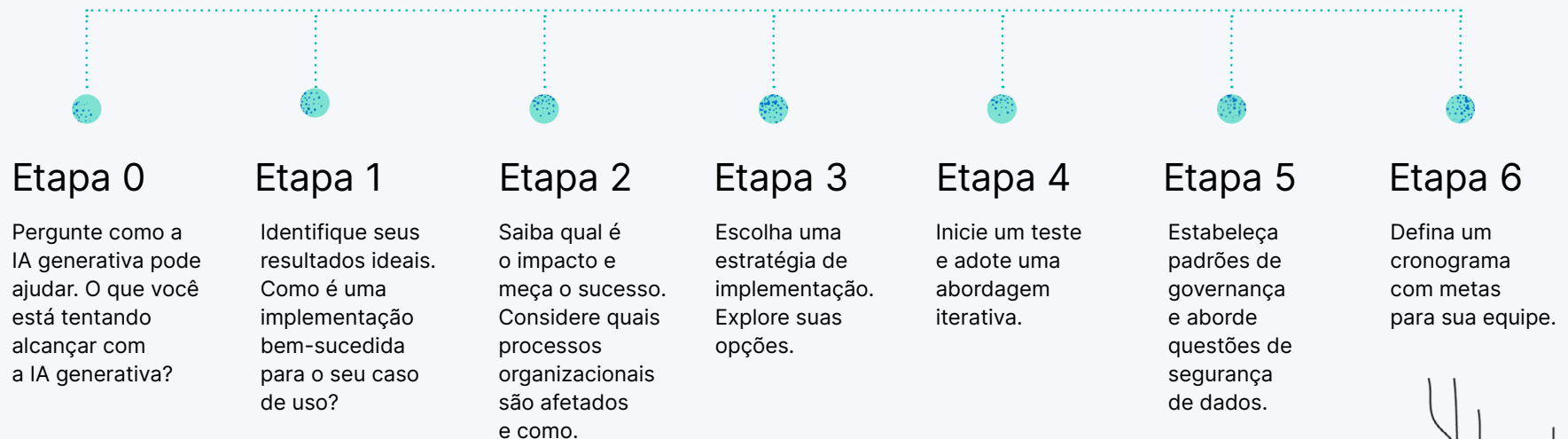
dos líderes estão confiantes em sua capacidade de implementar a IA em suas organizações³

¹ Elastic, A Cisco cria experiências de busca baseadas em IA com a Elastic no Google Cloud.

² Elastic, The Elastic Generative AI Report, (2024)

³ Russell Reynolds, Embracing the Unknown: How Leaders Engage with Generative AI in the Face of Uncertainty, (2024).

Veja o que esperar deste guia:



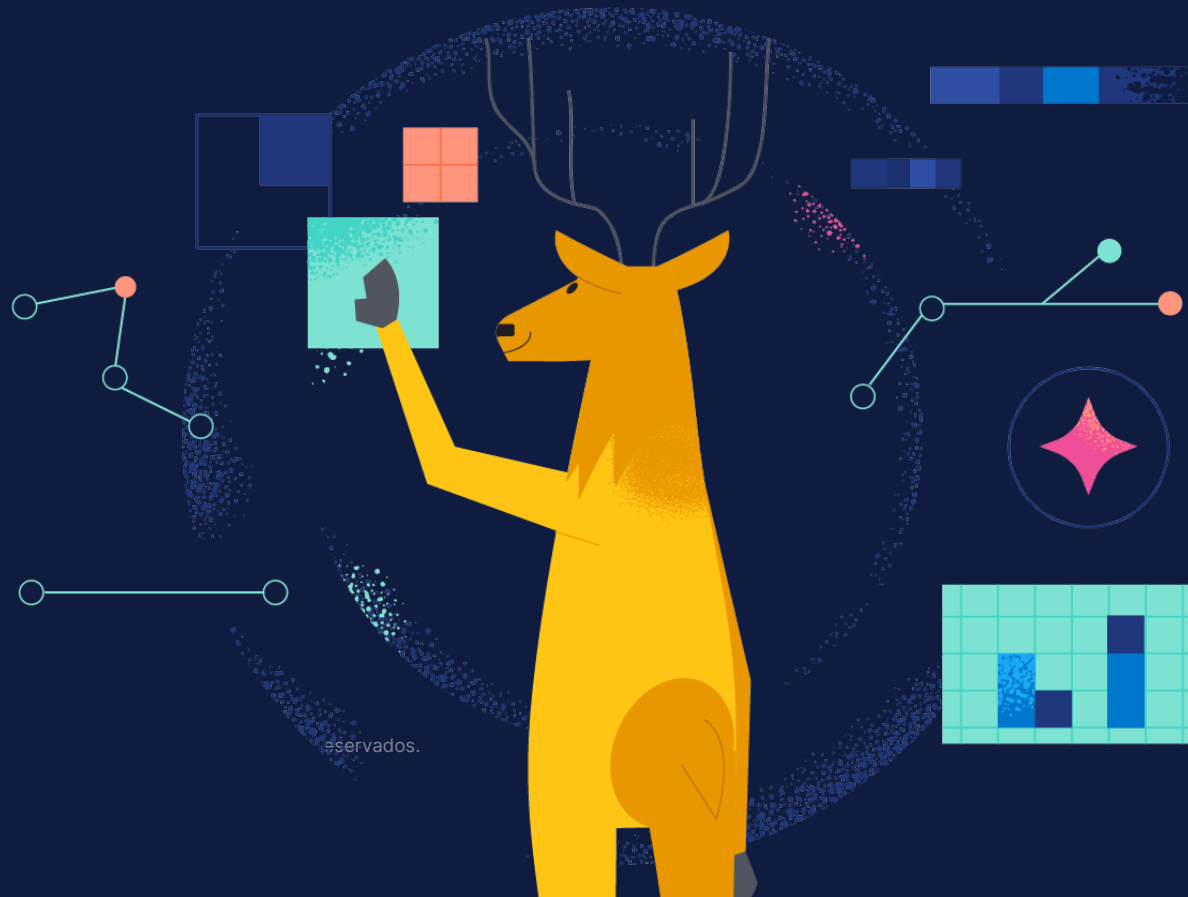
Mas primeiro, vamos relembrar o básico.

[Avance se você sabe o que deseja alcançar com a IA generativa](#)



Parte 1. Compreendendo o panorama da IA generativa

Você não precisa ser um especialista em IA generativa para elaborar um plano para operacionalizá-la. No entanto, ao compreender quais componentes estão em jogo, você pode tomar decisões informadas e estratégicas ao longo do processo. Vamos definir os componentes.



O que é a IA generativa?

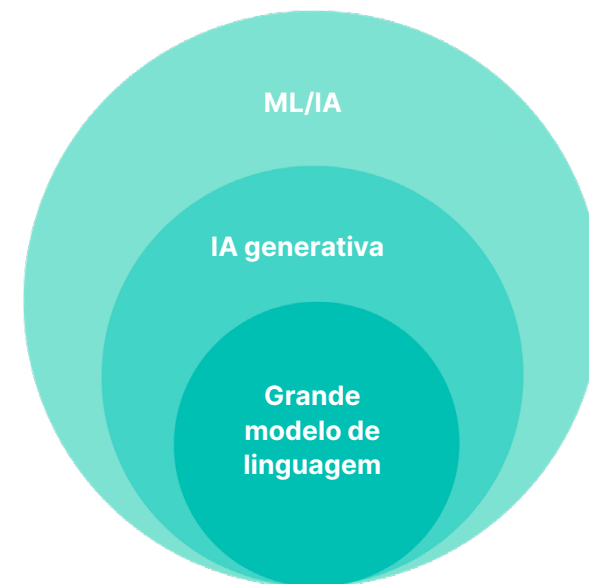
A inteligência artificial generativa (ou IA generativa) refere-se a modelos de aprendizagem profunda que podem gerar resultados quando solicitados. É importante compreender que a funcionalidade generativa dessa tecnologia depende de sua capacidade de prever resultados estatisticamente prováveis, o que é conseguido com a ajuda de machine learning. **Os dados estão no centro da operacionalização da IA generativa e são a chave para o sucesso na implementação e nos resultados.** Veremos mais sobre isso em breve.

O que é machine learning?

O machine learning (ou ML) é um subconjunto da inteligência artificial que usa algoritmos para gerar conhecimento de dados. Esses algoritmos analisam dados e “aprendem” — em um contexto supervisionado, semissupervisionado ou não supervisionado — os padrões e semelhanças que lhes permitem tomar decisões. O machine learning é a tecnologia subjacente que dá à IA generativa, como os grandes modelos de linguagem, a capacidade de “aprender” continuamente.

O que é um grande modelo de linguagem (LLM)?

Um grande modelo de linguagem (ou LLM) é um modelo computacional que se baseia no machine learning. É um tipo de IA generativa que lida especificamente com a linguagem humana. Tendo sido treinado principalmente em vastos conjuntos de dados de linguagem públicos, um LLM pode executar uma variedade de tarefas de processamento de linguagem natural (PLN), incluindo reconhecer, analisar, resumir, prever, traduzir ou gerar texto. No contexto da operacionalização da IA generativa, os LLMs são o que permite que a IA generativa se comunique em linguagem natural (ou humana).





Vamos falar sobre alucinações

Uma alucinação é um resultado incorreto ou enganoso gerado por um LLM. Você provavelmente conhece as respostas às vezes questionáveis do ChatGPT. O resultado parece legítimo... mas será que é mesmo? Se o LLM (o ChatGPT é alimentado por um LLM) não consegue encontrar a resposta, ele tende a inventar uma. É importante considerar esse ponto cego ao discutir o uso de LLMs em aplicações empresariais. Como você garante que os resultados gerados sejam relevantes e precisos? É aqui que entra a geração aumentada de recuperação (RAG).

Você: Quantos dias de folga me restam?



IA: Há 200 dias restantes no ano.

Você: Como faço para consertar meu porteiro eletrônico com câmera que não está se conectando à rede Wi-Fi?

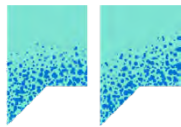


IA: Os melhores porteiros eletrônicos com câmera oferecem gravação 4K e...

O que é geração aumentada de recuperação (RAG)?

Pense na geração aumentada de recuperação (ou RAG) como sua linha de defesa contra alucinações. O resultado gerado por um LLM é aumentado, ou “verificado”, recuperando informações de um conjunto de dados (ou contexto de dados) específico, que você fornece usando busca altamente relevante alimentada por um banco de dados vetorial. Por exemplo, por meio do RAG e em resposta a uma consulta de um usuário, uma organização busca a documentação de suas políticas e fornece respostas relevantes a um LLM para que possa responder a perguntas usando as políticas da organização. Além de ser uma defesa contra alucinações, o **RAG permite que você use a IA generativa com seu conjunto de dados proprietário**, sendo esse seu maior benefício.

No contexto da operacionalização da IA generativa para aplicações de negócios, o RAG é importante por vários motivos: pode fornecer resultados melhores e mais relevantes e oferece uma maneira rápida de inicializar ou utilizar seus próprios dados proprietários. Também é mais econômico do que treinar ou construir seu próprio LLM. Assim, o RAG é a chave para uma integração bem-sucedida da IA generativa. O RAG vai além dos limites dos LLMs gerais para criar “o mecanismo de busca da próxima geração”.

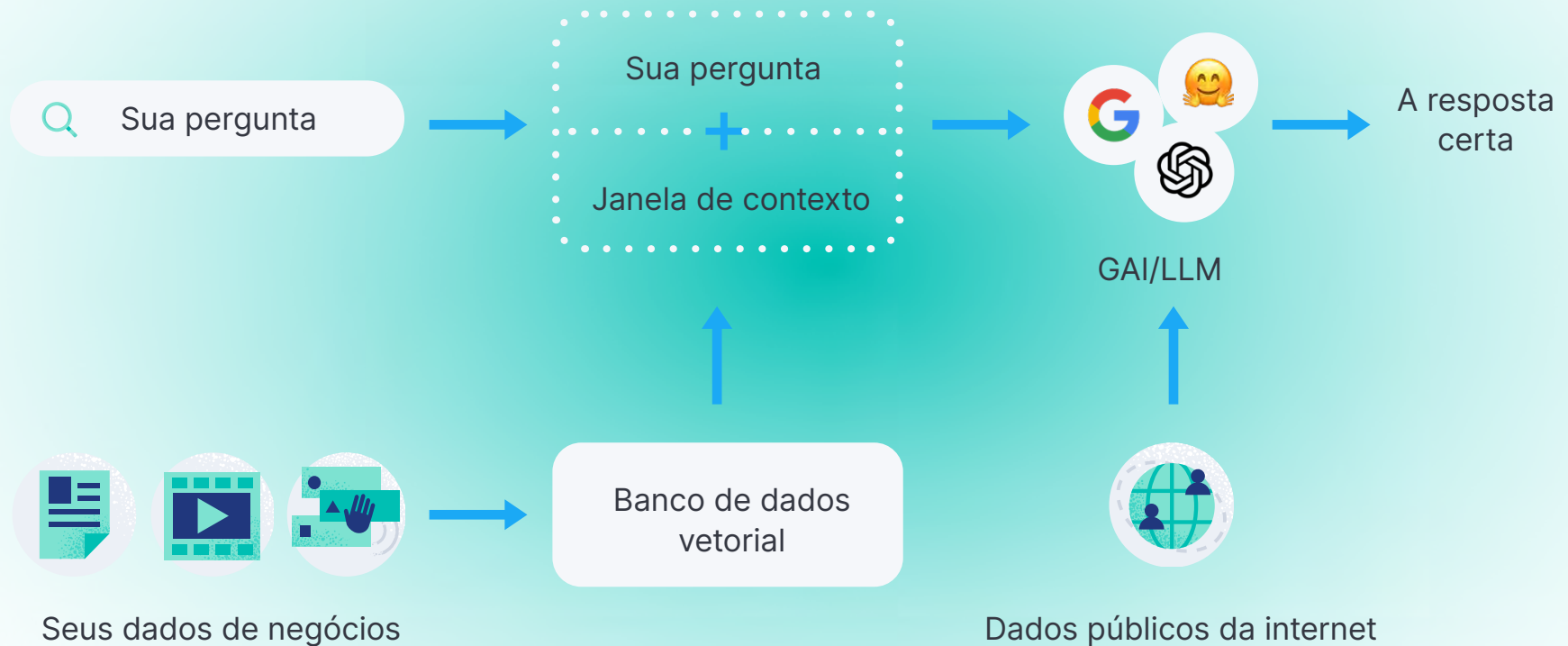


“O RAG ajuda a criar o mecanismo de busca da próxima geração”

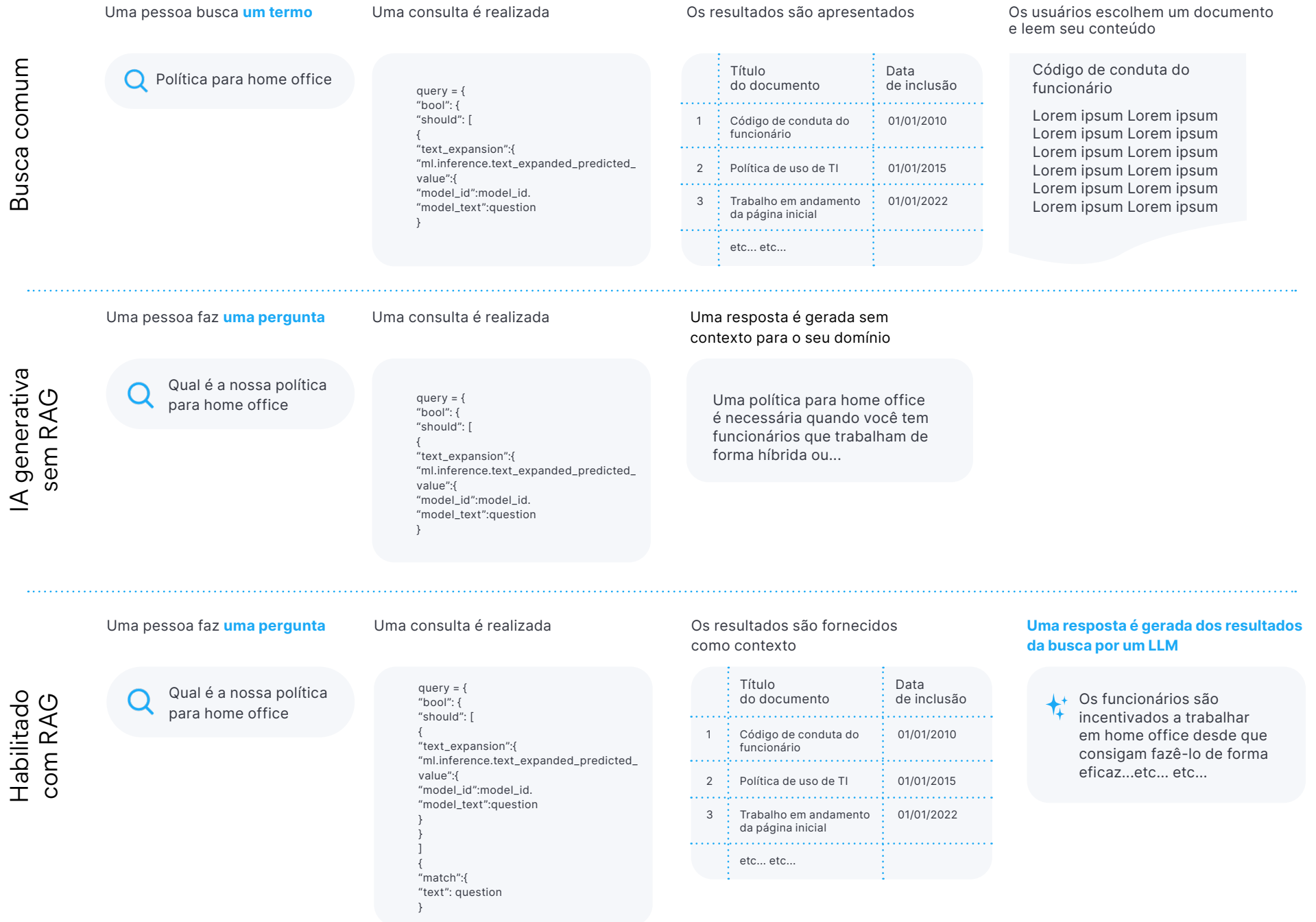
— Baha Azarmi
Vice-presidente de engenharia global de clientes da Elastic

Geração aumentada de recuperação (RAG)

O RAG permite que você use a IA generativa com seu conjunto de dados proprietário.



O RAG é uma nova maneira de responder às perguntas dos usuários:



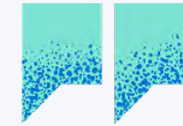
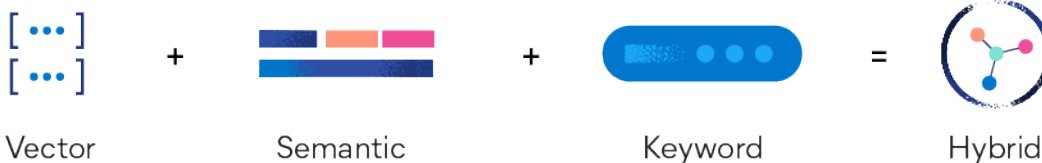
O que é um banco de dados vetorial?

Um banco de dados vetorial armazena embeddings vetoriais, ou seja, representações numéricas de palavras, imagens ou vídeo. Esses embeddings são multidimensionais e permitem a busca semântica, um tipo de busca que procura a intenção e o significado contextual de uma consulta. Por outro lado, uma busca textual procura apenas resultados que correspondam às palavras-chave na consulta de busca.

No contexto do RAG, um banco de dados vetorial permite uma busca semântica rápida com base no prompt fornecido à IA generativa. Isso é o que torna o RAG possível.

A IA generativa é boa em PLN, mas a busca tradicional por palavras-chave não é capaz de usar linguagem natural e fornecer os melhores resultados para alimentar a IA generativa. Portanto, com um banco de dados vetorial alimentando a IA generativa com resultados de busca semanticamente semelhantes ao prompt original, a IA generativa pode gerar respostas mais relevantes. Pense nos bancos de dados vetoriais como o banco de conhecimento que permite que a IA generativa responda a perguntas com informações precisas.

No entanto, a IA generativa não se limita a bancos de dados vetoriais. Usando o RAG, a IA generativa pode acessar bancos de dados relacionais, bancos de dados gráficos, bancos de dados baseados em documentos ou até mesmo mecanismos de busca por palavra-chave. O melhor banco de dados para você geralmente depende da natureza dos dados, dos algoritmos específicos usados e dos requisitos de desempenho do sistema. Por exemplo, bancos de dados relacionais podem ser usados para dados estruturados, enquanto bancos de dados gráficos são adequados para dados com relações complexas e mecanismos de busca tradicionais para busca de texto completo.



“Todos os caminhos levam à busca híbrida.”

— Serena Chou, diretora de gerenciamento de produtos da Elastic

Embora a **busca semântica** forneça resultados que correspondem ao significado de uma consulta, a **busca por palavra-chave** ainda desempenha um papel importante na correspondência de resultados a palavras-chave exatas das consultas. A busca híbrida é a prática de usar tecnologias de busca vetorial (frequentemente usada para busca semântica) e por palavra-chave para fornecer o que geralmente são os resultados mais relevantes para uma IA generativa.

Resumindo: uma solução de busca híbrida provavelmente fornecerá os resultados mais relevantes para experiências de IA generativa na sua organização.

O que a IA generativa pode fazer

Já falamos muito sobre as tecnologias subjacentes e os conceitos fundamentais, mas o que exatamente a IA generativa pode fazer?



Criar

Ao aprender padrões em seus dados de treinamento, a IA generativa pode “criar” ou *gerar* resultados. Ela itera nos dados existentes para produzir novas ideias, imagens, insights e muito mais.



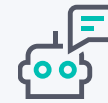
Resumir

Graças a seus recursos de processamento de linguagem natural, a IA generativa pode analisar um texto e resumir-lo. Precisa ler documentos extensos em um curto espaço de tempo? A IA generativa ajuda.



Descobrir

A chave para a IA generativa é a tecnologia de busca subjacente. Ela permite que a ferramenta de IA generativa receba uma consulta, busque em um vasto conjunto de dados privados ou públicos e produza uma resposta.



Automatizar

Digamos que sua organização use duas plataformas de nuvem diferentes para serviços diferentes. Cada plataforma de nuvem gerará logs em diferentes formatos. Ao transformar automaticamente esses dados no mesmo formato e mapeá-los com IA, sua equipe pode resumir e fazer perguntas sobre os dados com a IA generativa. Sua equipe de TI pode se concentrar no monitoramento e no gerenciamento dos seus sistemas em vez de realizar tarefas trabalhosas.

Etapa 0. Descubra o porquê e determine o que é possível

Com tantas maneiras possíveis de agregar valor, é crucial pensar em um ponto de partida. Como você aproveita o poder da IA generativa para capacitar sua equipe, atender à rápida evolução das expectativas dos clientes e impulsionar sua empresa a novos patamares? Concentrando-se na área em que a IA generativa pode agregar mais valor à sua organização.

Considere estas questões:

1

Que problema estou tentando resolver?

Existe alguma área na sua empresa que seja particularmente ineficiente? Onde seus funcionários estão gastando um tempo significativo em tarefas repetitivas? Eles estão constantemente buscando informações existentes em bancos de dados internos ou em mecanismos externos?



Por exemplo:

Seus funcionários têm dificuldade para encontrar informações, sejam atualizações de projetos ou informações relacionadas a RH? Sua equipe de segurança tem tarefas que poderia automatizar para ter mais tempo para assumir uma postura mais proativa? Há entropia nos processos de atendimento ao cliente porque os engenheiros nem sempre se comunicam em tempo real com os agentes de atendimento ao cliente sobre problemas conhecidos e correções recentes?

Lembre-se:

Você precisará trabalhar com a equipe cujo fluxo de trabalho impactará. Por exemplo, se decidir atualizar um processo de RH ineficiente, ter a adesão do RH desde o início será importante para obter a aprovação das partes interessadas e reforçar o apoio.

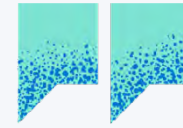
2 Esse é um problema que posso resolver com uma base de conhecimento?

Uma base de conhecimento é uma coleção de conteúdos informativos, sejam artigos de suporte ou processos internos.



Considere de qual conteúdo você precisa extrair as respostas e se esse conteúdo pode ser explorado para resolver seu problema com mais eficiência. É suficiente automatizar e personalizar as respostas?



Por exemplo: Você identificou que os funcionários passam uma quantidade considerável de horas de trabalho procurando informações relacionadas ao RH. Sua equipe não tem um especialista em RH, então os funcionários estão sendo redirecionados para a intranet da empresa, onde acabam tendo de ler diversos documentos de políticas para, digamos, descobrir quantos dias de férias eles ainda têm no ano. Para resolver esse problema, você precisará de uma base de conhecimento, como seus documentos de políticas de RH, bem como acesso aos dados pessoais dos funcionários para personalizar a resposta.



“Quando você puxa um fio, milhares [de fios] continuam aparecendo.”

Baha Azarmi,
vice-presidente
de engenharia global
de clientes da Elastic

O passo 0 é entender quais processos estão limitando a produtividade. Ao eliminar as tarefas rotineiras, você libera espaço para seus funcionários serem criativos. Uma implementação bem executada é responsável. **Investir nos seus funcionários, qualificá-los adequadamente e planejar fluxos de trabalho e processos ajustados são partes cruciais de um casamento bem-sucedido com a IA generativa.**



Ao procurar o porquê, **seja simples e específico**. Identificar qual caso de uso de IA generativa você deseja abordar primeiro é um excelente passo inicial na operacionalização da IA generativa. Então, projetos menores vão preparar você para uma implementação eficaz.

Por exemplo, no cenário anterior de RH, a IA generativa poderia atender a vários casos de uso:

1

Descoberta

Um funcionário consulta a interface: quantos dias de férias me restam em um ano? Para responder, a IA precisa realizar uma busca e apresentar os documentos relevantes para a consulta, puxando documentos de políticas de RH e o cadastro do funcionário.

2

Sumarização

Indo um passo além, a IA generativa pode analisar os documentos e resumi-los para o funcionário em uma resposta de conversação. “Você tem 10 dias de férias restantes para este ano e quatro dias de folga flexíveis que você pode usar. Confira a página da intranet sobre políticas de folga remunerada.”

3

Criação e automação

O chatbot poderia poupar o tempo dos gerentes criando uma resposta para aprovar ou negar uma solicitação de uso de férias e até mesmo explicar o motivo. Também poderia criar convites de calendário e registrar a solicitação de folga no sistema.

Adaptação ao seu setor

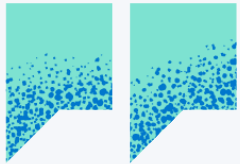
Copiloto, assistente, bot — a IA generativa assume várias “formas” que fornecem serviços valiosos de aumento de produtividade em uma variedade de campos e setores. Como copiloto de segurança e observabilidade ou com apps internos e externos, a IA generativa ajuda as empresas a aumentar a eficiência, atualizar seus esforços de segurança, melhorar a experiência do cliente e acelerar sua diferenciação competitiva.

Ao usar dados para revelar as melhores respostas de IA e aproveitar o poder da tecnologia de busca subjacente, **as empresas podem diminuir o tempo gasto em tarefas repetitivas, reduzir os tempos de resposta e, de modo geral, melhorar a produtividade.** Adicionando o RAG à equação, você pode aproveitar seus dados proprietários para obter respostas de IA generativa que sejam seguras, respeitando as permissões nos níveis de documento e de usuário.

De repente, você está olhando para o próximo nível de velocidade e relevância — a mesma velocidade e relevância que os clientes cada vez mais experientes em tecnologia esperam. Garantir que seus serviços correspondam a essas expectativas é fundamental. O mesmo acontece com o uso de IA generativa das formas mais relevantes para a experiência do usuário. Não há nada pior do que investir em soluções que ninguém usa.

A IA generativa é mais comumente integrada à infraestrutura de TI de uma empresa como um assistente de IA ou um copiloto de segurança e/ou observabilidade.





Todas as empresas têm oportunidades com a IA generativa porque ela é fundamentalmente uma maneira muito mais humana e intuitiva de extrair informações dos sistemas de informação.

 **Ash Kulkarni**

CEO da Elastic





Assistentes de IA

Aproveite ao máximo as habilidades de conversação da IA generativa para funcionários e clientes com apps internos e externos. Os assistentes de IA fornecem a cada usuário ajuda flexível, adaptável e pessoal como um especialista disponível, um personal shopper ou até mesmo um agendador.



Copiloto de segurança e observabilidade

Aumente seus recursos de observabilidade e segurança com copilotos de IA generativa. Projetados para trabalhar em conjunto com equipes de TI, os copilotos de IA generativa funcionam como parceiros especializados na resolução de problemas. Por exemplo, você pode solicitar ao seu copiloto uma descrição detalhada do motivo pelo qual um alerta de segurança foi disparado e obter etapas recomendadas para fazer a triagem e remediar o ataque (com base em ataques semelhantes anteriores que sua organização encontrou). Esse tipo de prompt pode gerar um runbook dinâmico para uma organização.

Essas integrações possibilitam que empresas de todos os setores aumentem seu potencial de personalização, automação e produtividade, levando a **três grandes casos de uso de IA generativa**:

Melhorar a resiliência operacional

A resiliência operacional é vital para manter os sistemas funcionando sem problemas. Com o impulso da IA generativa, as equipes de TI podem acelerar a análise da causa raiz, correlacionar mais dados em todos os ambientes para identificar problemas com mais rapidez e ter à mão uma ferramenta de descoberta dedicada para acelerar suas respostas, tudo em favor da continuidade dos negócios.

Aprimorar a experiência do cliente

A satisfação do cliente é um ponto central de toda empresa. A IA generativa fornece às suas equipes as ferramentas para resolver problemas com mais rapidez e obter as informações necessárias, além de oferecer aos clientes atenção personalizada e acesso rápido a informações relevantes. O resultado? Melhores experiências para o cliente e melhores resultados para os negócios.

Mitigar o risco à segurança

Conforme o mundo digital evolui a um ritmo vertiginoso, surgem novas ameaças sofisticadas à segurança. Lidar com elas requer medidas dinâmicas e proativas, sem falar na expertise para responder às ameaças e gerenciá-las. A IA generativa pode não apenas reforçar sua equipe e operações de segurança, mas também automatizar alertas e manter uma postura proativa.

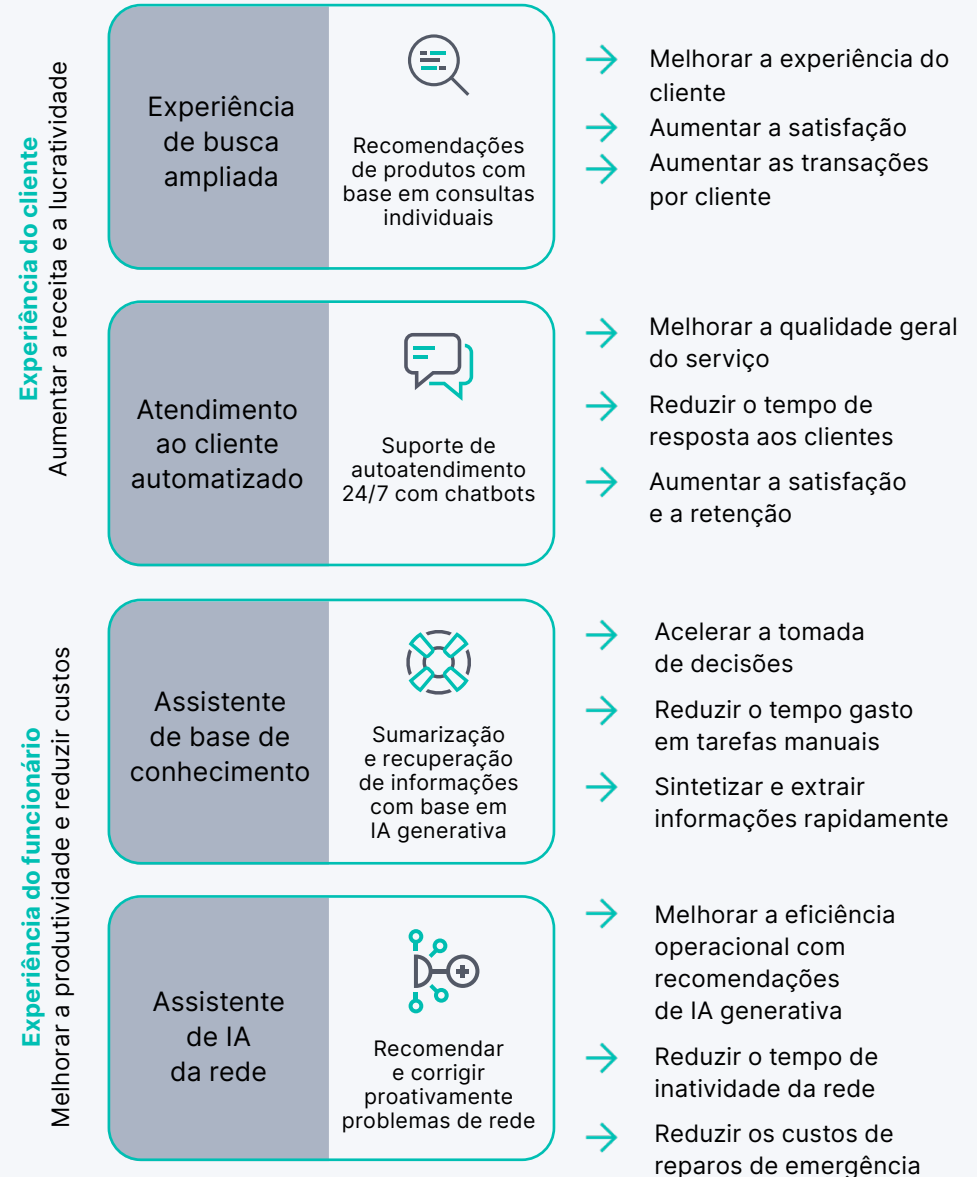
Em todos os setores, a IA generativa pode amplificar as experiências existentes dos funcionários e dos clientes, fornecendo respostas personalizadas, relevantes e prescritivas para as consultas. Independentemente do seu setor de atuação, há uma maneira de operacionalizar a IA generativa para potencializar sua busca e revelar novos recursos nos seus dados.



Telecomunicações

Para as empresas de telecomunicações, projeta-se que a IA generativa crie um valor econômico superior a US\$ 60 bilhões.⁴ Com a IA generativa, as empresas de telecomunicações podem possibilitar que seus funcionários e clientes consultem seu site ou base de conhecimento interna do local de trabalho para obter respostas personalizadas e relevantes rapidamente. O resultado? Melhor atendimento ao cliente e maior produtividade.

⁴ McKinsey, *Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom*, (2024).





Serviços financeiros

Com a IA generativa, as empresas de serviços financeiros podem personalizar ainda mais as experiências dos seus clientes e funcionários. Projeta-se que as melhorias na experiência do cliente, na prevenção de fraude e na automação criem um valor econômico superior a US\$ 250 bilhões para o setor de serviços financeiros.⁵

⁵McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Experiência do cliente

Aumentar a receita e a lucratividade

Assistente de banco de varejo



Sumarização e recuperação de informações com base em IA generativa

- Expandir a visibilidade das finanças pessoais
- Fornecer ofertas personalizadas para maior conversão
- Aumentar a satisfação e a retenção

Atendimento ao cliente aprimorado



Recomendar e corrigir proativamente problemas de rede

- Melhorar a qualidade geral do serviço
- Reduzir o tempo de resposta aos clientes
- Aumentar a retenção

Experiência do funcionário

Melhorar a produtividade e reduzir custos e riscos

Sumarização da detecção de fraude



Resumo de transações/detecção de anomalia e próxima melhor ação

- Melhorar a precisão e a velocidade da detecção de fraude
- Reduzir custos automatizando tarefas
- Reduzir perdas financeiras

Assistente virtual



Sumarização e recuperação de informações com base em PLN

- Acelerar a tomada de decisões
- Reduzir o tempo gasto em tarefas manuais
- Sintetizar e extrair informações rapidamente



Varejo


Mais atraente para o varejo, a IA generativa promete aumentar a retenção de clientes, melhorando a relevância da busca, recomendando produtos adicionais e enviando acompanhamentos personalizados em todos os canais. Você já recebeu um daqueles emails “Você esqueceu seus produtos no carrinho”? A IA pode automatizá-los e aprimorá-los para fornecer melhores recomendações e mais personalização na descoberta de produtos.

Seja criando experiências de próxima geração para os clientes para impulsionar as vendas no e-commerce ou capacitando os funcionários com a tecnologia mais recente para melhorar a produtividade, projeta-se que a IA generativa crie um valor econômico superior a US\$ 240 bilhões para os varejistas.⁵

⁵McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Experiência do cliente Aumentar a receita e a lucratividade


Busca e descoberta personalizadas de produtos



Resposta a perguntas, experiência de busca personalizada

- Melhorar as taxas de conversão do site
- Aumentar as transações por cliente
- Aumentar a satisfação

Atendimento ao cliente aprimorado




Interações de autoatendimento com chatbots

- Reduzir o tempo de resposta aos clientes
- Melhorar o serviço, reduzir o churn
- Aumentar a retenção

Experiência do funcionário Melhorar a produtividade e reduzir custos


Atendimento ao cliente aprimorado



Experiência e interação aprimoradas do agente

- Resolução no primeiro contato
- Integração mais rápida
- Redução da rotatividade dos agentes

Manutenção preditiva



Avaliar a integridade de sistemas críticos para priorizar tarefas críticas de manutenção

- Reduzir o tempo de inatividade de equipamentos e sistemas
- Reduzir os custos de reparos de emergência
- Melhorar a eficiência operacional

Estudo de caso: HSE

A HSE é uma das marcas líderes no setor de comércio ao vivo na Europa.⁶

“Para a Home Shopping Europe [HSE], o sucesso comercial começa com a personalização e a relevância do site.”

Peter Strasser

Desenvolvedor de software da HSE



A oportunidade

Como qualquer negócio de e-commerce, o recurso de busca é fundamental para a experiência do cliente e as vendas. A HSE precisa atender às jornadas do cliente originadas em vários canais, resultando em diversos termos de busca que refletem onde o cliente foi apresentado ao produto.

A HSE usou IA generativa e LLMs para extrair o significado semântico de uma consulta do cliente e gerar resultados que complementam a correspondência tradicional de palavras-chave.



O resultado

A HSE observou um **aumento de 4% na taxa de cliques** e um **aumento de 8% na satisfação do cliente** graças à maior precisão e relevância nos resultados de busca.



Insight

Concentre-se em uma área que você já quer melhorar, como a experiência de busca do cliente. Veja como você pode integrar a IA generativa para elevar a experiência a um patamar superior com personalização e relevância.

⁶ Elastic, HSE aumenta a satisfação do cliente e reduz o tempo de manutenção em 42%, usando o Elasticsearch na AWS, (2024).



Setor automotivo e de manufatura

Cada etapa do processo do setor automotivo e de manufatura pode ser simplificada com a IA, proporcionando um valor econômico projetado superior a US\$ 170 bilhões.⁵ A IA generativa tem o potencial de transformar o setor, desde a inovação na pesquisa e desenvolvimento de produtos até estratégias personalizadas de retenção de clientes. Carros voadores? Talvez!

⁵McKinsey, The economic potential of generative AI: The next productivity frontier, (2023).

Experiência do cliente

Aumentar a receita e a lucratividade

Manuais digitais interativos



Assistente virtual de produto

- Respostas em tempo real sobre recursos de produtos, manutenção e solução de problemas
- Reduzir as consultas de suporte
- Melhorar a satisfação

Atendimento ao cliente aprimorado



Interações de autoatendimento com chatbots

- Reduzir o tempo de resposta aos clientes
- Melhorar o serviço, reduzir o churn
- Aumentar a retenção

Experiência do funcionário

Melhorar a produtividade e reduzir custos

Otimização da tecnologia operacional



Manutenção preditiva: sumarização de problemas e resoluções

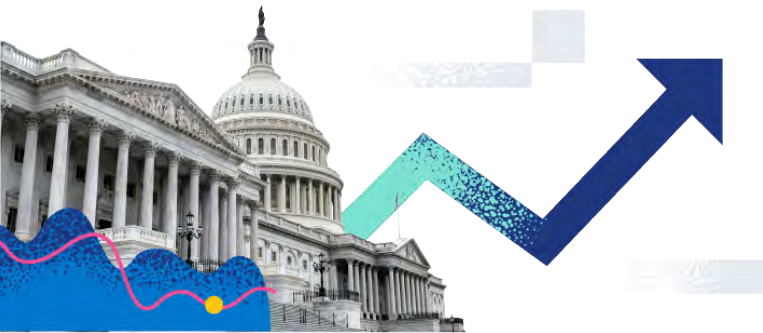
- Identificar e resolver problemas rapidamente
- Melhorar a eficiência operacional e a tomada de decisão
- Reduzir os custos de fabricação

Análise de sentimentos dos produtos



Resumir a porcentagem dos produtos e recomendar melhorias

- Melhorar as ofertas de produtos com o ponto de vista do cliente
- Diminuir o tempo de percepção de benefícios de novas ofertas de produtos



Setor público

A IA generativa pode acelerar significativamente os resultados da missão, melhorar os serviços aos cidadãos e conectar melhor os analistas governamentais e os profissionais de segurança aos dados certos no momento certo, conectando de forma segura a IA generativa aos dados das agências.



Redução da carga de trabalho
Automatizar processos e fluxos de trabalho manuais



Conformidade
Habilitar o acesso aos dados baseado em função



Consciência situacional em tempo real
Tomar decisões mais precisas



Produtividade dos funcionários
Encontrar as informações certas no momento certo



Experiências dos cidadãos
Construir a confiança por meio de interações digitais personalizadas



Serviços públicos
Aumentar a acessibilidade e as opções de autoatendimento



Inteligência dinâmica
Acelerar a busca e os insights sobre missões



Segurança cibernética
Realizar avaliações e análises de risco em tempo real

As aplicações voltadas para os cidadãos incluem:

- Acesso personalizado a serviços públicos
- Experiências online simplificadas para os cidadãos
- Maior acessibilidade e opções de autoatendimento

As aplicações voltadas para os funcionários incluem:

- Investigações e inteligência mais precisas
- Maior produtividade ao automatizar processos e fluxos de trabalho manuais
- Processos de compras mais eficientes

Estudo de caso: Relativity

A Relativity ajuda corporações, escritórios de advocacia e agências a armazenar e utilizar dados para descoberta eletrônica e busca jurídica.⁷



A oportunidade

A Relativity precisava consolidar seus dados e, ao mesmo tempo, manter uma postura de segurança em primeiro lugar. Como resultado dessa explosão de dados, fontes e complexidade, as abordagens tradicionais de busca por palavras-chave eram ineficazes. Aí chegou o RAG.



O resultado

Juntos, com o RAG e um banco de dados vetorial, a Relativity implementou experiências de busca baseadas em dados proprietários e ofereceu aos usuários uma experiência de busca rápida, relevante e precisa. Sua solução de IA generativa atende a padrões de conformidade como PCI, DSS, SOC2 e HIPAA.



Insight

Construa a solução tendo em mente a escalabilidade. Começar aos poucos pode ajudar a identificar os recursos da IA generativa e aprimorar as aplicações mais relevantes. Depois de encontrar o ponto ideal, o céu é o limite.

“O maior desafio que os clientes da Relativity enfrentam atualmente é a explosão de dados provenientes de fontes de dados heterogêneas. O desafio é realmente agravado pelas diferenças nos dados gerados de diferentes modos de comunicação.”

Brittany Roush
Gerente sênior de produto

⁷Elastic, Relativity usa Elasticsearch e OpenAI do Azure para criar experiências de busca futurísticas, hoje (2024).

Agora você entende o imenso potencial econômico da IA generativa em todos os setores. Você pode até ter alguns casos de uso potenciais em mente. Esperamos que você também tenha o seu “porquê”.



No entanto, a implementação da IA generativa pode parecer um processo árduo e disruptivo. Existem preocupações com a privacidade, algum trabalho braçal a ser feito no domínio da conformidade e mudanças na forma como as pessoas realizam seus trabalhos. A operacionalização responsável requer treinamento, melhoria de competências e uma reorganização parcial da força de trabalho.

Apesar desses desafios, o valor que a IA generativa pode proporcionar a uma empresa é inegável. A implementação é inevitável para manter a competitividade. As boas notícias? Você pode alcançar uma rápida percepção dos benefícios com testes que não estão completamente prontos para produção. Em outras palavras, é hora de começar.

Parte 2:

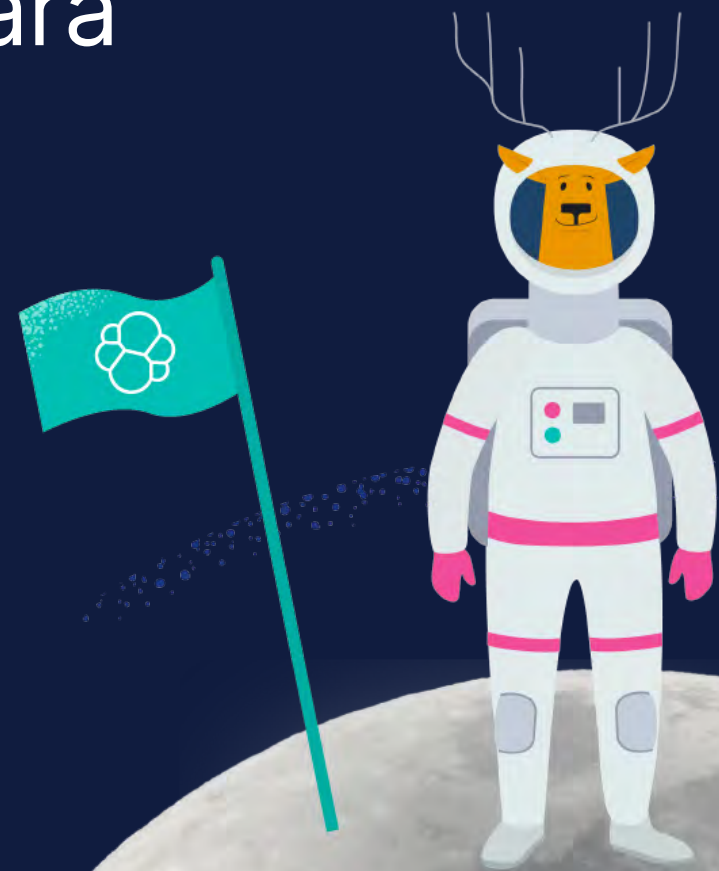
Operacionalizando a IA generativa

Um pequeno passo para a máquina, um salto gigantesco para sua organização

A operacionalização da IA generativa não acontece de uma só vez. É uma abordagem iterativa que requer planejamento e um resultado claro. Ao começar com um único projeto de IA generativa — um pequeno passo —, você capacita sua equipe pela inevitável curva de aprendizado e permite que eles pratiquem os processos, ajustem a tecnologia e resolvam as preocupações, preparando-os e à sua empresa para o sucesso — o cobiçado salto gigantesco.

Agora, veja como fazer isso acontecer.

Um salto à frente
graças ao RAG!



Etapa Identifique seu resultado ideal

1

Você se concentrou em um problema. Você sabe que está procurando otimizar um processo ineficiente. Agora você precisa pensar em como os usuários vão interagir com a sua solução. Você está ampliando uma aplicação de busca ou um chatbot? Está procurando uma nova maneira de interagir com suas equipes ou com clientes?

Seu processo de pensamento pode ser mais ou menos assim:

- ||→ Você está buscando mais retenção de clientes.
- ||→ Você decidiu implementar uma aplicação personalizada de busca e descoberta de produtos.
- ||→ Você cria métricas para o sucesso. Pense nisso como o seu “subporquê”.

Você quer utilizar a IA generativa. Por quê? Para personalizar a busca e a descoberta de produtos. Por quê? **Este é o seu resultado ideal:** ao interagir dessa nova maneira com nossos dados, os clientes encontrarão facilmente os produtos de que precisam e descobrirão produtos que possivelmente lhes interessem, com base em seu histórico de busca e localização. Como resultado, a retenção de clientes melhorará.

- ||→ Agora você inicia a grande tarefa de operacionalizar seu primeiro projeto de IA generativa.



Pergunte-se o seguinte:

Quais ações e resultados podem ser criados a partir dessa nova forma de interagir com seus dados?

A resposta a essa pergunta ajudará a definir suas metas. Identificar o resultado ideal determina o que é “bom” para seu projeto e, em uma escala maior, o que é bom para sua empresa.

Etapas **2** Saiba qual será o impacto. Meça o sucesso.

Para ter sucesso na operacionalização da IA generativa, você precisará estabelecer um conjunto de KPIs que ajudem a medir o que significa ser “bom” para você. Compreender como a IA generativa está mudando a produtividade na sua organização é apenas um dos muitos indicadores de desempenho.

Outros podem incluir um aumento na satisfação do cliente, medido por avaliações em um contexto de suporte ao cliente, uma diminuição nos tíquetes de suporte ou resoluções mais ágeis. Dependendo do caso de uso que você estiver testando, será necessário estabelecer indicadores de desempenho correspondentes. Incorporá-los em cada etapa do processo de teste é vital para compreender o progresso que você e sua equipe estão fazendo.

Indicadores básicos de desempenho

1

Impacto na produtividade

Meça as mudanças na produtividade resultantes do seu caso de uso. Compare o tempo necessário para concluir determinadas tarefas com o tempo necessário sem o uso de IA generativa.

2

Escalabilidade

Avalie se o modelo se adapta bem a um aumento no uso e na demanda. Ele ainda está funcionando de maneira confiável e precisa?

3

Resultados financeiros

Avalie o impacto que a implementação da IA generativa teve em termos de custos empresariais. Você pode querer incluir certas métricas de negócios nesta análise, como o número de reclamações de clientes registradas ou quaisquer alterações em suas vendas.

4

Conformidade

Monitore continuamente a adesão da IA generativa às regulamentações de privacidade de dados.

5

Satisfação do cliente

Analise métricas de negócios como churn, aumento de vendas e manutenção da fidelidade à marca, e examine o feedback dos clientes.

Use esses indicadores para determinar se um projeto é viável, prático, escalável e acessível. Esses indicadores ajudarão você a determinar seu ROI e poderão ser ampliados no futuro à medida que você expandir seus casos de uso.

Etapa Escolha um modelo (caminho a seguir)

3

Como você constrói uma arquitetura de IA generativa que atenda às suas necessidades de negócios? Muitos fatores influenciarão a escolha que você faz: custo, idioma, seu ecossistema de TI, seus recursos e cronograma de implantação, regulamentações de privacidade de dados e governança. Por esse motivo, adotar uma postura restrita — começando com um caso de uso simples e específico — é crucial.

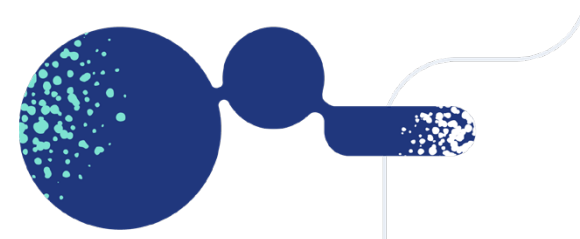
Para operacionalizar a IA generativa, você precisará destes componentes:

- ||→ **Uma infraestrutura de nuvem totalmente gerenciada** aumentará a agilidade, melhorará a eficiência de custos e reduzirá o desperdício de recursos. Chips e hardware estão evoluindo a uma velocidade vertiginosa. Se você investir na construção do seu próprio datacenter de IA, ele poderá se tornar obsoleto em alguns meses.
- ||→ **Um LLM** será a base que permitirá que a IA generativa se comunique em linguagem natural e a compreenda.
- ||→ **Uma plataforma de dados** que inclua busca vetorial, híbrida e tradicional por palavra-chave que possa ser usada para enriquecer o LLM com o contexto certo dos seus dados proprietários.
- ||→ **APIs abrangentes** que permitam enriquecer e transmitir seus dados ao LLM e ao seu mecanismo de busca.

Os ingredientes de que as empresas precisam para busca com IA



A maneira como você combinar esses componentes — se você ajustará seu próprio modelo, trará seu próprio banco de dados vetorial, trará seu próprio modelo ou qualquer combinação disso — determinará seu esquema de implementação, afetando seu cronograma, a complexidade de seu teste e se você pode precisar complementar sua equipe.



Pré-treinar um LLM

Esta abordagem exige muitos recursos e envolve começar do zero, treinando um grande modelo de linguagem em um grande conjunto de dados.

Ajustar um modelo

Esta abordagem usa um LLM existente com seu mecanismo de busca e um banco de dados vetorial para fornecer contexto aos seus dados proprietários.

RAG

Este processo usa um LLM pré-treinado existente e um conjunto de técnicas para ajustar o modelo para atender às suas necessidades.

Custo

\$\$\$\$

\$\$\$

\$\$

Tempo de implantação

Longo, implantado em meses

Moderado, implantado em semanas

Curto, implantado em dias

Privacidade dos dados

Você tem um conjunto de dados grande o suficiente para fornecer material de aprendizagem significativo ao LLM? Caso contrário, você precisará de dados públicos. Você combinará dados públicos e privados?

Você tem um conjunto de dados grande o suficiente para fornecer material de aprendizagem significativo ao LLM? Caso contrário, você precisará de dados públicos. Você combinará dados públicos e privados?

Esta abordagem permite que você mantenha seus dados privados.

Precisão e relevância

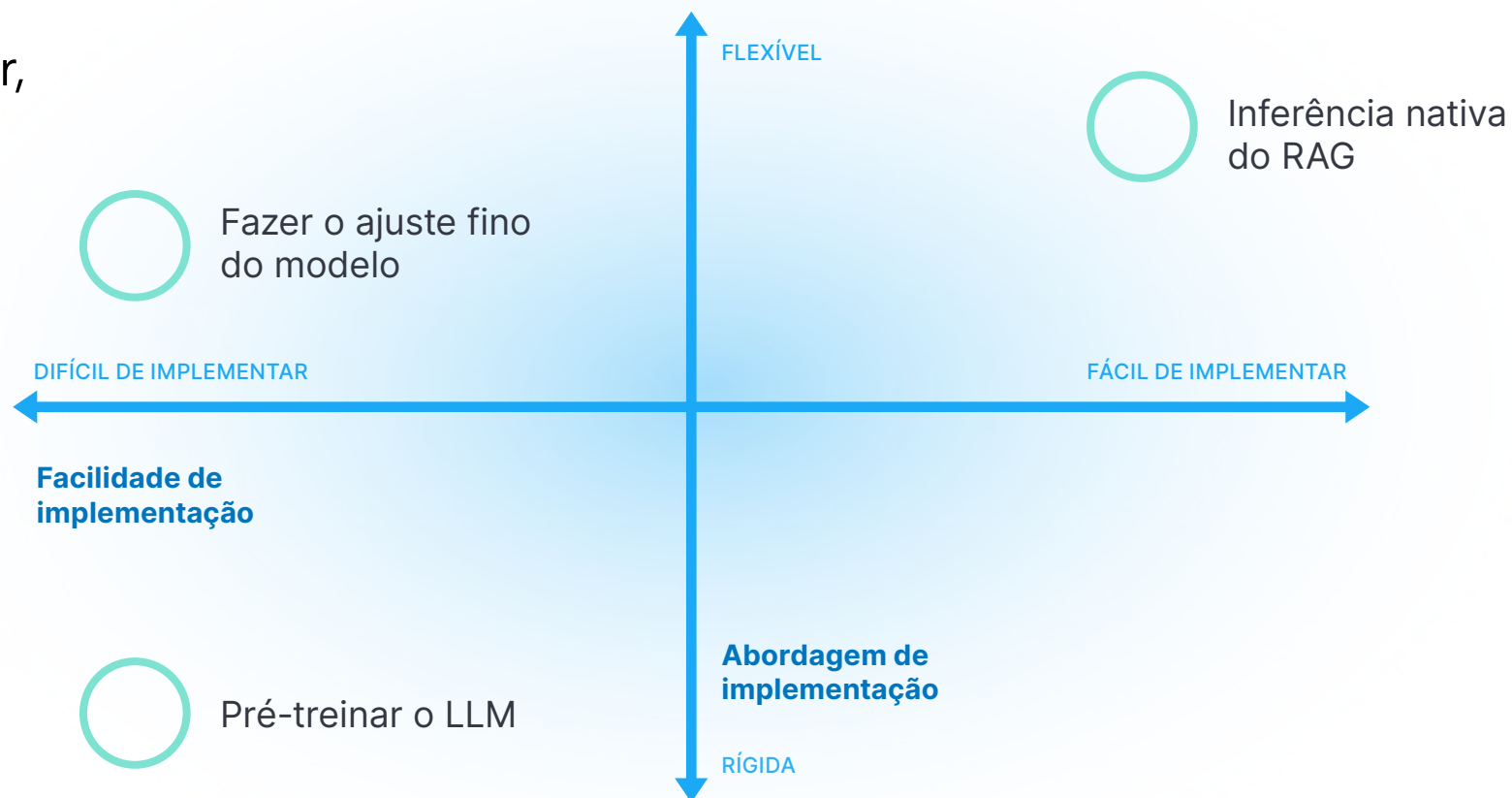
Difícil de garantir consistentemente

Mais fácil de garantir a precisão e a relevância para as tarefas específicas para as quais o modelo foi ajustado.

O benefício do RAG reside especialmente na sua capacidade de “detectar” alucinações citando fontes ou informando o usuário quando ele não tem a resposta.

Considere estas opções:

Pré-treinar,
ajustar,
RAG



Escolhendo o caminho certo a seguir

Não existe um caminho certo para todos. Portanto, verifique suas decisões em relação às metas delineadas nas etapas anteriores e se essas metas estão alinhadas com as partes envolvidas. Em última análise, você precisa de um caminho que possa descrever claramente para as partes interessadas da empresa e sua equipe.

A menos que você esteja buscando uma revisão mais extensa e um projeto de maior escala, construir um LLM do zero e fazer o ajuste fino consumirá recursos demais. Haverá muitas perguntas que precisarão ser respondidas: você precisa de um banco de dados vetorial para complementar seu mecanismo de busca? Você pode atualizar seu mecanismo de busca, criar e armazenar embeddings nele e criar a lógica para continuar a melhorar sua busca? Como você expande isso para um sistema de recomendações?

Para soluções baseadas em nuvem com busca híbrida e busca semântica, pode ser bastante simples. Ao se conectar a um LLM existente, você pode usar o RAG para criar uma experiência de busca mais relevante para seus clientes.

Veja o que considerar:

||→ Faça um balanço do que você já tem no seu ambiente de TI. Muitas vezes, reestruturar a infraestrutura não é necessário.

||→ Considere soluções prontas para uso, como um LLM, um banco de dados vetorial e um pacote completo.

Pró: tecnologias de caixa preta provavelmente farão com que você entre em operação com mais rapidez.

Contra: você não pode redimensionar da mesma maneira porque elas são minimamente customizáveis.

||→ Encontre um produto complementar que ofereça flexibilidade e um mínimo de interrupções. Convém avaliar a relevância e o desempenho da busca e talvez trocar de modelo para ver qual funciona melhor para você.

Etapa Tente rápido, falhe rápido

4

Um ecossistema digital em rápida evolução significa que há muitas partes móveis num projeto de IA generativa. Há um limite para o controle que você pode ter em um LLM pré-treinado, bem como uma quantidade limitada de flexibilidade para manipular sua arquitetura.

Este é o momento de adotar uma abordagem iterativa: você tem seu caso de uso, estabeleceu os resultados desejados, definiu KPIs e considerou como implementar seu projeto de IA generativa.



Lembre-se

Basicamente, operacionalizar a IA generativa significa obter respostas dos *seus* dados. Certifique-se de ter a conformidade sempre em mente. Você está configurando um teste que pode comprometer suas políticas de privacidade? Esse é um teste de baixo risco?

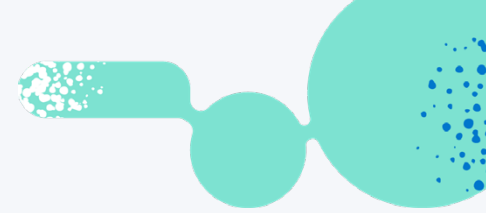


Neste estágio, você quer...

- ||→ **Criar um ciclo de feedback.** Estabeleça quem reporta o que para quem e identifique as principais partes interessadas no projeto.
- ||→ **Enriquecer seu LLM.** Certifique-se de que seu LLM tenha acesso às informações corretas armazenadas em um banco de dados vetorial. Um banco de dados vetorial permitirá que você forneça rapidamente as informações mais relevantes para enriquecer seu LLM.
- ||→ **Ajustar a experiência do usuário.** Trabalhe em uma interface intuitiva e continue testando-a. Em última análise, a IA generativa existe para servir seus funcionários e clientes. A construção de uma interface adequada à aplicação e ao usuário é fundamental para um projeto de IA generativa bem-sucedido e garante sua escalabilidade.
- ||→ **Estabelecer uma arquitetura de referência que possa ser redimensionada.** Enquanto você testa seu projeto de IA generativa, fique de olho no panorama geral. Como ficará sua arquitetura quando você redimensionar o projeto e como ficará quando você expandir para outros casos de uso?

Por exemplo, se construir um banco de dados vetorial do zero parece uma tarefa pesada, você pode procurar um banco de dados para download — sim, eles existem. Com esse banco de dados vetorial, você pode ter acesso ao próximo nível: a busca híbrida. O uso da busca semântica nas suas aplicações de busca permite testar o protótipo do seu projeto de IA de próxima geração. Este é um exemplo de como é importante começar pequeno e iterar.





Etapa Governança e operações

5

Os projetos de IA generativa trazem seu próprio conjunto de desafios: desde a privacidade e a conformidade dos dados até considerações éticas, controle de qualidade e gerenciamento de riscos. Você precisa antecipar possíveis obstáculos e garantir que seu projeto esteja alinhado com seus objetivos de negócios.

Como parte da sua análise de governança e operações, você precisará considerar uma série de elementos:

- ||→ **Gestão de custos.** A cobrança é feita por milhar de tokens; um custo para os prompts e outro para as respostas.
- ||→ **Logging.** Você precisará registrar todas as respostas para ver a comunicação entre seu modelo e seu cliente para fins de controle de qualidade.
- ||→ **Estabelecer o sentimento das respostas.** Determine o sentimento das respostas do LLM para que elas correspondam à marca e ao tom de voz da sua empresa (outra etapa importante do controle de qualidade).
- ||→ **Monitorar alucinações.** As alucinações incluem informações incorretas ou enganosas, mas também podem incluir discurso de ódio e comportamento antissocial de um chatbot.
- ||→ **Sinalizar respostas inconclusivas.** Monitorar a qualidade e a relevância das respostas é vital para o controle de qualidade. Esta é uma oportunidade para entender quais aplicações exigirão mais envolvimento humano do que outras e planejar adequadamente quando chegar a hora de expandir.

Sobre o viés na IA

Os modelos de IA generativa dependem dos dados nos quais são treinados. Se os dados de treinamento contiverem vieses e limitações, eles serão refletidos nos resultados.

As organizações podem mitigar esses riscos considerando e limitando cuidadosamente os dados nos quais seus modelos são treinados ou usando modelos especializados e customizados específicos para suas necessidades. Dito isto, os humanos que programam essa tecnologia ou que fazem a curadoria dos dados nos quais o modelo é treinado também têm vieses.

O viés, como em qualquer contexto, é difícil de erradicar. Isso não significa que as organizações não devam tentar enfrentar esse desafio e educar os usuários para que exerçam um certo pensamento crítico como parte da solução.

Além disso, espere o envolvimento da sua equipe jurídica e leve em consideração o trabalho deles na sua prova de conceito. Embora o envolvimento deles possa parecer retardar uma fase de testes, é crucial estabelecer processos de revisão que sejam completos e eficientes para uma implementação responsável, ética e dentro dos padrões de conformidade.

A questão da segurança dos dados

Com ameaças à segurança afetando as organizações todos os dias, a segurança dos dados é fundamental. Seus clientes confiam seus dados a você — e é por isso que muitas empresas operam a partir de um framework zero trust. Zero trust se baseia no princípio de que usuários e dispositivos nunca devem ser automaticamente ou implicitamente confiáveis, seja dentro ou fora do perímetro da rede de uma organização.



Para otimizar a segurança, você pode:

1. **Adotar uma abordagem com RAG.** Os modelos com RAG aproveitam mecanismos de recuperação para entender melhor o contexto do prompt de entrada, o que leva a respostas mais contextualmente apropriadas que omitem detalhes sensíveis. Usar o RAG com uma plataforma de dados que tenha segurança baseada em função e em nível de documento garantirá que as permissões sejam respeitadas.
2. **Investir na sua solução de observabilidade ou expandi-la.** Resolva a questão da confiança. Siga a trilha dos dados com recursos de monitoramento e monitore as respostas produzidas pelo seu projeto de IA generativa. Para onde estão indo seus dados e o que a IA generativa está dizendo aos seus clientes?

Em última análise, trazer a IA generativa para o seu ecossistema exigirá que você estabeleça novos protocolos operacionais e, conseqüentemente, novas políticas. Com processos mais eficientes e receitas mais elevadas, o tempo poupado na execução de tarefas rotineiras pode ser redirecionado para esses esforços.

Etapas Defina um cronograma. Coloque padrões de referência.

6

Determine um período — por exemplo, um trimestre. Dentro desse período, estabeleça metas nas marcas de 30 e 90 dias. Use o trimestre para provar o valor do seu caso de uso impulsionado pela IA generativa.



No dia 30, convém que você tenha lançado seu primeiro teste. Como ficou?

Você escolheu um caso de uso

Você designou uma pequena equipe para a tarefa

Você facilitou o acesso a sessões de treinamento conforme necessário

Você estabeleceu os resultados desejados

Você construiu uma interface de protótipo



No dia 90, você estará pronto(a) para lançar seu primeiro caso de uso. Como ficou?

Você abriu o teste para algumas pessoas internas

Você testou, ajustou e mediu os resultados gerados

Você monitorou continuamente a maneira como os usuários interagem com a interface

Você estabeleceu um conjunto de diretrizes para o que constituem resultados de qualidade

Você coletou dados sobre algumas métricas importantes de desempenho

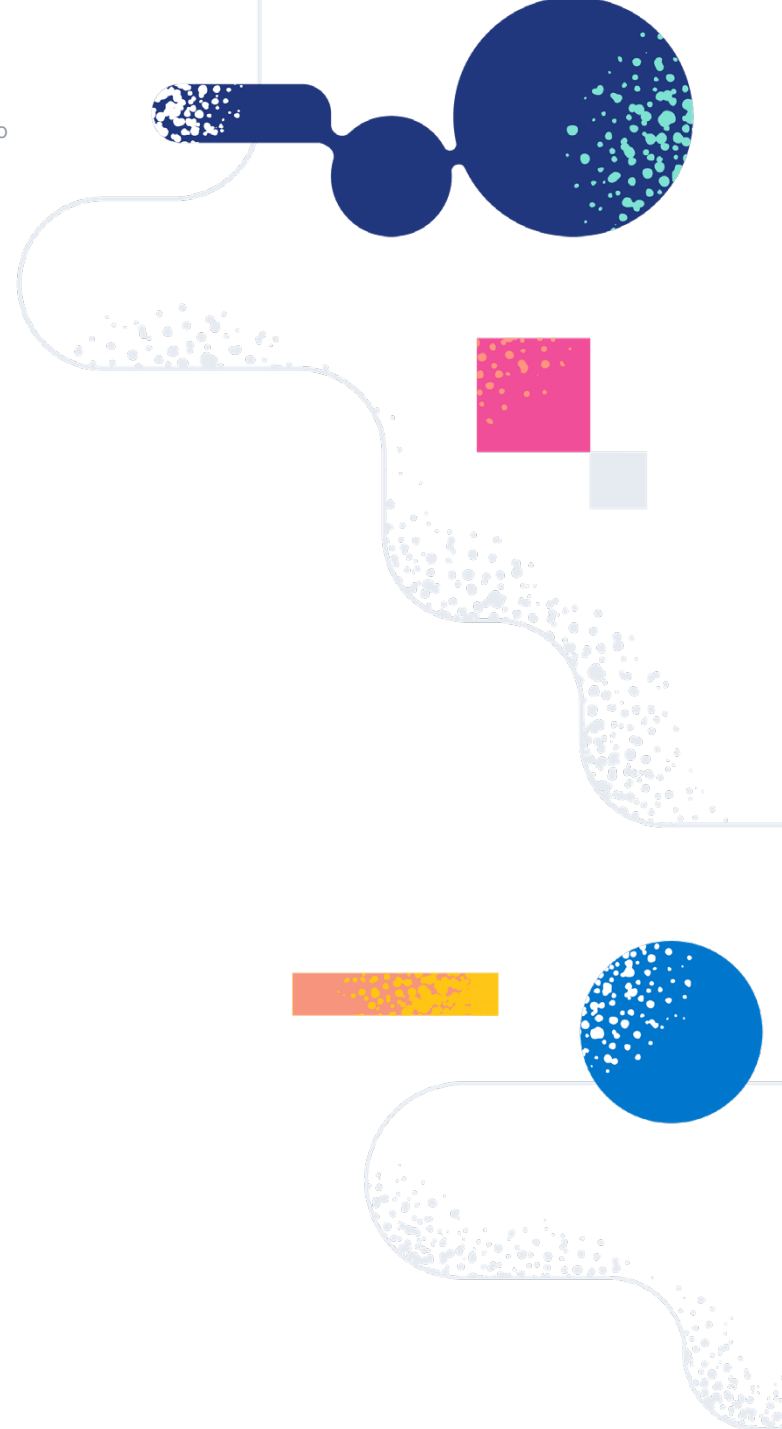
Você mediu o valor da iniciativa

Estas tarefas devem servir como referências aproximadas. As necessidades específicas da sua empresa — a composição da sua equipe e a tecnologia em que estão trabalhando ou adicionando à sua stack — afetarão a velocidade com que você pode implantar seu primeiro caso de uso e coletar insights.

Neste estágio, considere o seguinte:

1. **A taxa de erro.** Meça a taxa de erro. A IA generativa está produzindo resultados corretos e relevantes? Isso é crucial para o ajuste fino da IA generativa.
2. **Tempo e custo do treinamento.** Meça o tempo e os recursos necessários para treinar seu modelo. Isso ajuda a garantir um período de testes eficiente e, portanto, uma operacionalização mais rápida.
3. **Intervenção humana.** A IA generativa funciona apenas com um ser humano no circuito? Quanta supervisão é necessária para manter a confiabilidade e a precisão?
4. **Tempo de resposta e qualidade dos resultados.** Meça a rapidez com que a IA generativa fornece resultados e compare a qualidade dos resultados com um conjunto de regras ou diretrizes estabelecidas.

E assim você estará com tudo pronto para operacionalizar e ampliar seu sucesso.



O início de uma nova era

Muitos líderes do setor já estão percebendo os benefícios da IA generativa — e muitos outros estão tentando replicar esse sucesso e acompanhar as mudanças nas expectativas dos clientes. A inovação na IA generativa está avançando rapidamente. Mas você não pode ir a lugar nenhum sem o básico.

Criar estratégias para as formas mais alinhadas de implementar a IA generativa pode ajudar você a aproveitar o poder dos seus dados sem se distrair com inovações interessantes, mas irrelevantes. Para operacionalizar a IA generativa da forma mais eficaz, dedique tempo e recursos para implementá-la em estágios. Integrar novas tecnologias com um propósito é a receita para você ter o maior retorno do seu investimento. Além do mais, a customização e a adaptação das ferramentas de IA para atender às suas necessidades operacionais garante relevância e eficácia — que é a “chamada” da IA generativa em primeiro lugar.

Tenha em mente a necessidade de implementar de forma responsável a IA generativa, desde a proteção da privacidade dos dados e a segurança até a sensibilidade e a ética. Além de agregar valores que podem chegar à casa dos trilhões de dólares à economia global, a IA generativa apresenta uma oportunidade para democratizar a força de trabalho e melhorar as competências dos trabalhadores. Posicione-se com pioneirismo não apenas em IA generativa, mas também no desenvolvimento de novos processos de negócios para sua empresa.



Agora vamos começar.

Identificar seu primeiro caso de uso de IA generativa é um esforço entre equipes. Você precisa que suas equipes de segurança, TI, desenvolvimento e linha de negócios trabalhem juntas desde o primeiro dia.

Veja como a Elastic pode ajudar:

Com sua equipe de segurança

Você pode aumentar a produtividade dos profissionais e reduzir riscos. A operacionalização da IA generativa para seus casos de uso de segurança começa com [uma abordagem unificada em uma plataforma aberta](#).

Você pode aproveitar o poder da IA generativa com o Elastic Security para criar uma experiência adaptada às necessidades da sua equipe de segurança.

Conheça o Elastic Security

Com sua equipe de SRE e operações de TI

Ofereça a seus SREs e engenheiros uma interface interativa em linguagem natural para eles terem acesso a informações mais relevantes com mais rapidez. Descubra como combinar a IA conversacional com o Elastic Observability e machine learning avançado para ter uma [experiência de chat interativo com reconhecimento de contexto](#) baseada nos seus dados proprietários e runbooks.

Conheça o Elastic Observability

Com sua equipe de desenvolvimento

Você pode aprimorar o kit de ferramentas da sua equipe de desenvolvimento para ajudá-los a fornecer suporte ao cliente por meio de opções de autoatendimento, como chatbots altamente personalizados com o Elastic Search. Disponibilize aos seus agentes de atendimento ao cliente as mesmas excelentes ferramentas de busca para resolver casos rapidamente, incluindo experiências de IA generativa que os ajudam a encontrar respostas em fontes de dados diferentes. Descubra como [implementar uma busca poderosa para sua base de conhecimento](#).

Conheça o Elastic Search

