

Inteligência artificial

Cálculo e servidores

Automação de TI

O futuro é programável: como a computação generativa pode reinventar o software



A ideia não surgiu de repente. Surgiu aos poucos, em conversas noturnas no Slack e em bate-papos de corredor, uma reflexão silenciosa sobre como as máquinas raciocinam. Em algum ponto entre o caos dos comandos e as aspirações da automação, um novo conceito começou a tomar forma. Isso poderá redefinir não apenas a inteligência artificial, mas o próprio software.

A premissa é ousada: e se parássemos de tratar grandes modelos de linguagem como

chatbots misteriosos e começássemos a tratá-los como infraestrutura programável? A IBM se refere a essa disciplina emergente como [computação generativa](#), um termo e estrutura desenvolvidos por seus pesquisadores para definir uma nova abordagem para trabalhar com modelos de IA. Trata-se de reestruturar a forma como os modelos de IA são integrados aos sistemas, não como oráculos imprevisíveis, mas como um componente de software controlado e modular. Se tiver sucesso, poderá marcar um ponto de virada para o desenvolvimento da IA, o design de software e a tecnologia empresarial.

[David Cox, diretor da IBM Research](#), disse em entrevista na *IBM Think* que cunhou o termo computação generativa para descrever a mudança que observa no desenvolvimento da inteligência artificial. Não se trata de uma marca nem de um produto. É uma mudança, um movimento para tratar grandes modelos de linguagem não como parceiros de bate-papo inteligentes, mas como elementos programáveis. Esqueça os truques de mágica. Isso é engenharia de software.

"Não é que os LLMs estejam substituindo a programação", disse ele. "É que eles estão se tornando um novo tipo de programação primitiva."

Hoje em dia, interagir com um modelo de linguagem complexo muitas vezes parece invocar um oráculo caprichoso. Altere ligeiramente uma frase em um prompt e a produção sairá do curso. Escreva um prompt do tamanho de um texto e espere, reze, convença. É uma arte, assim como a astrologia é: enigmática, interpretativa e, por vezes, profunda. Mas para bancos, hospitais e governo, o misticismo não é escalável.

"Você digita algo e obtém uma resposta diferente dependendo de como você formulou a pergunta", disse [Ruchir Puri](#), cientista-chefe da IBM Research, em entrevista no *IBM Think*. "É como os primórdios da pesquisa. Ainda estamos na era em que uma vírgula pode alterar a produção. Não se pode gerir uma empresa dessa forma."

Puri descreve um mundo em que as empresas lutam não apenas com a [alucinação](#), mas também com a falta de confiabilidade na forma como os modelos lidam com casos extremos. "Falamos muito sobre alucinação", disse ele, "mas a questão mais profunda é que não há garantia de que os modelos sigam as instruções. Você muda uma palavra em um prompt e não sabe o que vai obter." Isso, argumentou ele, é a antítese da engenharia.

Boletim informativo do setor

As mais recentes tendências em IA, trazidas a você por especialistas

Receba insights selecionados sobre as notícias mais importantes (e intrigantes) sobre IA. Inscreva-se no nosso boletim informativo semanal Think. Consulte a [Declaração de privacidade da IBM](#).

johndoe@yourdomain.com

Subscribe →

Do prompt à programação

Para que fique claro, ninguém está desconsiderando o poder dos modelos modernos. O problema, disse Cox, é a forma como os utilizamos. "Engenharia de prompts não é engenharia. É brincadeira. Precisamos de um sistema em que não tenhamos que torcer para que o modelo faça o que pretendemos e que possamos programá-lo para fazer o que queremos."

A premissa por trás da computação gerativa é simples: tratar o modelo como uma função. Em vez de enterrar instruções em textos prolixos, os desenvolvedores usam um tempo de execução, uma camada de orquestração que divide o prompt em partes atômicas, as encaminha, verifica as condições e reescreve as falhas. A lógica não é apenas implícita; ela é imposta. O controle se torna explícito. A estrutura retorna.

| Do prompt à programação Um blueprint comportamental para máquinas

adicionando à IA." Na prática, isso significa criar sistemas em camadas que dividem tarefas complexas em instruções menores e gerenciáveis, cada uma das quais deve ser verificada antes de prosseguir. "Você pode ter vinte pequenos prompts focados para o modelo, em vez de um único prompt longo e complexo", disse Puri. "Mas agora você pode registrar cada um deles. Você pode tentar novamente. Você pode criar planos de contingência. É disso que as empresas precisam."

Essa estrutura também abre caminho para testes e validação, dois princípios que estiveram ausentes por muito tempo da IA generativa. "Você pode escrever asserções sobre o comportamento do LLM da mesma forma que faz com o código", disse Cox. "E se você não obtiver o comportamento desejado, pode pedir ao modelo para tentar novamente ou direcionar para uma sub-rotina diferente."

Essa ideia se torna particularmente poderosa quando aplicada à segurança. Puri afirma que frequentemente ouve de CTOs que gostam do potencial do agente de IA, mas hesitam devido à sua imprevisibilidade. "Eles têm medo de deixá-los fazer qualquer coisa por conta própria." E se eles alucinarem? E se eles enviarem a mensagem errada ou aprovarem a transação errada?"

Para responder a isso, a computação generativa introduz ferramentas como detecção de alucinação, validação de contexto e processamento com reconhecimento de conformidade. "Com nosso tempo de execução", disse Cox, "você pode interpor um modelo guardião, que verifica a produção do modelo principal. Se algo parecer suspeito, ele pode sinalizar ou pedir para tentar novamente."

Esse tipo de sobreposição de camadas permite um nível de reprodutibilidade e confiabilidade que a engenharia de prompts atual não consegue proporcionar. Os desenvolvedores podem combinar código tradicional com respostas dos LLM, incorporando a produção em sistemas maiores sem perder o controle.

"Não é um chatbot", disse Cox. "Faz parte do seu stack de softwares. Você testa da mesma forma que testa qualquer outro módulo."



Um blueprint comportamental para máquinas

Segundo Cox, este momento se compara a épocas anteriores na computação. Na década de 1980, a introdução de padrões de projeto de software, como o [Model-View-Controller \(MVC\)](#), permitiu que os desenvolvedores separassem a lógica da interface, criando uma base modular e reutilizável para o desenvolvimento de aplicações. Ele acredita que a computação generativa representa um ponto de inflexão semelhante.

"Vamos encontrar padrões", disse ele. "Assim como o MVC se tornou omnipresente no desenvolvimento de IU, veremos uma estrutura para a orquestração de LLMs. Este é o início de uma nova camada na stack de software."

Essa visão de estrutura está na base de grande parte do movimento da computação generativa. Em vez de tentar entender cada neurônio em um modelo de linguagem complexo, os desenvolvedores criam mecanismos de proteção que se alinham às restrições da empresa. "Criamos responsabilidade", disse Puri.

Transparência, disse Cox, não precisa significar simplicidade. "O motor do seu carro é complicado", disse ele. "Mas está construído dentro de uma estrutura de segurança. Quando algo quebra, você tem procedimentos a seguir. É isso que queremos para a IA. Não mistério. Engenharia."

Em termos técnicos, isso significa expor as etapas intermediárias da tomada de decisão de um modelo. O tempo de execução usado na computação generativa pode gerar registros, anexar metadados e realizar validação em cada etapa.

"É a explicação como funcionalidade", disse Cox. "Não como uma reflexão tardia."

[Os modelos Granite](#) da IBM já foram ajustados para suportar esse tipo de orquestração modular. Eles são otimizados para inferência rápida e com uso eficiente de memória, permitindo muitas consultas pequenas em vez de um único prompt massivo. Isso os torna muito adequados para uma abordagem orientada por tempo de execução.

"Você pode pensar neles como blocos de construção", disse Puri. "Em vez de tentar fazer tudo de uma vez, nós os chamamos várias vezes para subtarefas específicas." É mais rápido, mais barato e mais confiável."

Os benefícios não são apenas técnicos, mas também organizacionais. Em um projeto piloto, um cliente corporativo usou a computação generativa para criar um pipeline de classificação de documentos. Em vez de confiarem em um único prompt para resumir um documento jurídico, eles dividiram a tarefa em nove etapas: classificação, segmentação, extração, validação, avaliação de risco, resumo, formatação, análise e aprovação.

"Cada etapa foi isolada e monitorada", disse Cox. "Se algo falhar, poderá ser retentado ou corrigido. Não seria possível fazer isso com um único prompt."

Puri acredita que esse tipo de estrutura se tornará a norma. "Vamos parar de pensar nos LLMs como algo mágico que resolve tudo de ponta a ponta e começar a tratá-los como infraestrutura", disse ele. "Não se trata de substituir os desenvolvedores. Se trata de dar a eles novas ferramentas."

Uma dessas ferramentas, disse Cox, é o [LLM intrínseco](#), um novo conceito onde funções especiais do modelo são expostas diretamente ao tempo de execução, permitindo uma integração mais profunda e uma adaptação em tempo real. "Você pode conectar um

adaptador que altera o comportamento do modelo", disse ele. "Isso permite mudar o tom, reduzir o risco e até mesmo detectar a alucinação na hora."

Esses avanços podem mudar a forma como o software é escrito. Cox imagina IDEs [IDEs](#) que incluem modelos de orquestração com tempo de execução para LLMs, testes unitários que validam os prompts e sistemas de controle de versão que rastreiam o comportamento do modelo.

"Os engenheiros de software terão que aprender novas habilidades", disse ele. "Mas os fundamentos ainda estão lá: inputs, produção, acerto, observabilidade. Não estamos abandonando a engenharia de software. Estamos atualizando-a."

Os pesquisadores preveem que a computação generativa irá se estender para além de seu atual nicho de casos de uso. À medida que o campo amadurece, novas camadas de abstração, novos padrões e novas funções profissionais surgirão.

Ele faz uma pausa por um instante. "Passamos uma década aprendendo como fazer esses sistemas parecerem inteligentes", disse ele. "Agora temos que ensiná-los a se comportar."

AI Academy



Do piloto à produção: gerando ROI com IA generativa

Saiba como a sua organização pode aproveitar o poder das soluções orientadas por IA em escala para reinventar e transformar seus negócios de maneiras que realmente façam diferença.

[Acessar o episódio →](#)



Relatório

AI in action

Pesquisamos duas mil organizações sobre suas iniciativas de IA para descobrir o que está funcionando, o que não está e como é possível progredir.

[Leia o relatório](#)



Recursos

Insight

Escale os recursos de IA da sua empresa

Libere quatro para escalar a IA com

uma base de dados sólida.

[Leia os insights](#)



Guia

Guia do CEO sobre IA generativa

Aprenda como os CEOs podem equilibrar o valor que a IA generativa pode criar com o investimento que ela exige e os riscos que ela introduz.

[Leia o guia](#)



Relatório

IA em ação 2024

Entrevistamos duas mil organizações a respeito de suas iniciativas de IA para descobrir o que está funcionando, o que não está e como se preparar.

[Leia o relatório](#)



Ebook

Apresentando a IA ágil: um guia prático

Aprenda uma abordagem de IA ágil que permite que as organizações inovem rapidamente e reduzam o risco de falhas.

[Leia o e-book](#)



Ebook

Libere o poder da IA generativa + ML

Saiba como incorporar IA generativa, aprendizado de máquina e modelos de base em suas operações de negócios para melhorar o desempenho.

[Leia o e-book](#)



Guia

Coloque a IA para trabalhar: como gerar ROI com a IA generativa

Quer ter mais retorno sobre seus investimentos em IA? Saiba como o dimensionamento da IA generativa em áreas importantes promove mudanças, ajudando suas melhores mentes a criar e oferecer soluções

novas e inovadoras.

[Leia o guia](#)



Vídeo

A ascensão da IA generativa e o que isso significa para os negócios

Aprenda sobre a história da IA e explore o que o futuro reserva para as empresas que consideram a adoção da IA.

[Assista o vídeo](#)



Guia

Como prosperar nesta nova era da IA com confiança e convicção

Aprofunde-se nos três elementos críticos de uma estratégia de IA forte: gerar vantagem competitiva, escalar a IA em toda a empresa e avançar na IA confiável.

[Leia o guia](#)

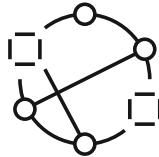
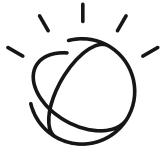


1 / 4



Soluções relacionadas

IBM Watson - Acelerando o seu futuro



IBM watsonx.ai

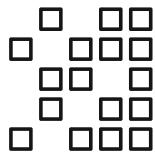
Treine, valide, ajuste e implemente recursos de IA generativa, modelos de base e recursos de aprendizado de máquina com o IBM watsonx.ai, um estúdio empresarial de última geração para construtores de IA. Crie aplicações de IA em uma fração do tempo com uma fração dos dados.

[Conheça o watsonx.ai →](#)

Soluções de inteligência artificial

Use a IA a serviço de sua empresa com a experiência e o portfólio de soluções líder do setor da IBM à sua disposição.

[Explore as soluções de IA →](#)



Consultoria e serviços em IA

Reinvente os fluxos de trabalho e operações críticos adicionando IA para maximizar experiências, tomadas de decisão em tempo real e valor de negócios.

[Explore os serviços de IA →](#)

Dê o próximo passo

Tenha acesso completo aos recursos que abrangem o ciclo de vida de desenvolvimento da IA. Produza soluções avançadas de IA com interfaces fáceis de usar, fluxos de trabalhos e acesso a APIs e SDKs padrão do setor.

[Explore o WatsonX.ai](#)



Agende uma demonstração em tempo real

