



OPINIÃO

Latência é o novo gargalo da IA em produção

Everton Fernandes (*)

Durante os últimos anos, a corrida da inteligência artificial foi impulsionada pela evolução dos modelos.

A discussão se concentrava em parâmetros, capacidade de processamento e precisão das respostas. À medida que a IA entra definitivamente na rotina das empresas, porém, uma nova prioridade começa a surgir no sentido de garantir respostas rápidas, previsíveis e economicamente viáveis.

Essa mudança acompanha a evolução do próprio mercado. Segundo o *Gartner*, os investimentos globais em inteligência artificial devem atingir **US\$ 2,52 trilhões em 2026**, impulsionados principalmente pela expansão da infraestrutura necessária para colocar aplicações de IA em produção.

Esse movimento também altera a forma como as empresas investem. O *Gartner* projeta que **55% dos investimentos em infraestrutura otimizada para IA em 2026 serão destinados à inferência**, superando pela primeira vez os gastos com treinamento de modelos. Em outras palavras, o maior desafio deixa de ser construir modelos mais inteligentes para executá-los de forma eficiente.

É nesse contexto que a latência assume um papel estratégico. Aplicações como copilots internos, assistentes corporativos e agentes operacionais precisam responder em tempo real para gerar valor ao usuário. Uma resposta precisa, mas entregue com atraso, compromete a experiência, reduz a produtividade e dificulta a adoção da solução.

Por isso, decisões de arquitetura ganham um peso cada vez maior. Inferência distribuída, modelos menores combinados com mecanismos inteligentes de roteamento, cache de contexto e otimização dos pipelines de dados deixam de ser boas práticas e passam a atuar como componentes essenciais para equilibrar desempenho, escalabilidade e custo.

O cache de contexto ilustra bem essa transformação. Em vez de reprocessar continuamente informações já conhecidas, a aplicação reutiliza estados computacionais previamente gerados, reduzindo a latência e o consumo de recursos. Essa abordagem vem ganhando

importância justamente porque melhora simultaneamente a experiência do usuário e a eficiência operacional.

Outro aspecto frequentemente negligenciado está na camada de dados. Em muitos projetos, o maior tempo de espera não está na geração da resposta pelo modelo, mas na recuperação, movimentação e preparação das informações necessárias para produzir um resultado relevante. Isso faz com que a arquitetura de dados tenha impacto direto na latência percebida pelo usuário.

Essa evolução também muda a forma de medir o desempenho das aplicações. Além da qualidade das respostas, indicadores como Time to First Token (TTFT) e End-to-End Latency passam a ser métricas essenciais para avaliar a experiência do usuário. Em aplicações corporativas, reduzir algumas centenas de milissegundos pode representar ganhos relevantes de produtividade quando milhares de solicitações são processadas diariamente.

O foco crescente na inferência também explica a expansão da infraestrutura global para IA. Segundo a *McKinsey & Company*, as cargas de trabalho de inferência deverão representar mais de 40% da demanda dos data centers até 2030, crescendo em ritmo acelerado à medida que a IA passa a fazer parte das operações das empresas.

Esse conjunto de fatores reforça o quanto é necessário aplicar uma mudança primordial: a inteligência do modelo continua sendo fundamental, mas não é apenas isso que a torna suficiente. O sucesso das aplicações corporativas dependerá cada vez mais da integração entre infraestrutura, arquitetura de dados, otimização da inferência e eficiência operacional.

A pergunta mais relevante para líderes de tecnologia e arquitetos de IA deve ser outra: quantos projetos deixam de gerar valor não pela qualidade do modelo escolhido, mas porque a latência e o custo por requisição tornam sua operação inviável em escala?

Amadurecemos em IA, então a vantagem competitiva tende a migrar dos modelos para a arquitetura que os sustenta. E é justamente nessa transição que a latência se transforma de um detalhe técnico para fator estratégico de negócio.

(*) Sr. Manager de Engenharia de Soluções na Cloudera.

Estudantes estão perdendo a capacidade de raciocinar à medida em que usam cada vez mais IA

Estudantes em todo o mundo estão usando modelos de IA para escrever trabalhos e fazer exercícios. Há muitos casos de fraude, mas muitas vezes as escolas permitem que os alunos façam algum uso dessas ferramentas.

Vivaldo José Breternitz (*)

No entanto, especialistas estão alertando que o uso de IA no processo de escrita não ameaça apenas as habilidades dos alunos nesse campo, mas toda a sua capacidade de pensar.

“Escrever é uma tecnologia para pensar”, disse o psicólogo cognitivo Ronald T. Kellogg ao *New York Times* em um artigo escrito por Dana Goldstein sobre as consequências de evitar o trabalho árduo de escrever. A escrita, observa o artigo, ajuda a desenvolver vários tipos de faculdades cognitivas, incluindo a construção da nossa memória de trabalho, habilidades de planejamento executivo e metacognição, a capacidade monitorarmos e autorregularmos nossos processos cognitivos.

Kellogg afirma que redigir um texto pode ser tão mentalmente desgastante quanto cavar uma vala é fisicamente exaustivo. É preciso forçar o cérebro para encontrar as palavras certas e organizar ideias de modo que façam sentido para os que forem ler o texto.

Em outras palavras, para escrever é preciso esforço, e quando se introduz uma ferramenta que suprime o esforço, deixa-se de exercitar o cérebro.

Um estudo do MIT, realizado no início deste ano por uma equipe liderada pela pesquisadora Nataliya Kosmyna, descobriu que pessoas que usaram IA enquanto escreviam um ensaio apresentaram menor atividade cerebral em comparação com aquelas que não usaram. Além disso, os que escreveram com IA tiveram dificuldade em lembrar trechos dos ensaios que acabaram de produzir. Além disso, sua atividade cerebral permaneceu em nível reduzido mesmo quando foram posteriormente convidados a escrever novamente sem ajuda da IA.

Como observa o artigo do *NYT*, poucas pessoas caviariam uma vala de bom grado, e o mesmo vale para escrever. Mas não há parte do processo de escrita que possa ser terceirizada para uma IA sem algum custo, alertou Steve Graham, psicólogo educacional da Universidade Estadual do Arizona.

Muitas escolas, de todos os níveis, estão tentando conter o uso dessa tecnologia pelos alunos, voltando a exigir que trabalhos sejam escritos à mão, frequentemente nas salas de aula, e fazendo avaliações orais.

A Ivy League é um grupo de oito universidades dos Estados Unidos conhecidas por sua excelência acadêmica, prestígio his-



Firm_CANVA

tórico e seletividade. Uma delas, Princeton, aboliu seu Código de Honra que permitia aos alunos fazer provas sem supervisão, por ter detectado casos de uso indevido de IA.

Felizmente, diz Goldstein, alguns estudantes já estão percebendo que usar IA indiscriminadamente prejudica-os, como disse um deles ao *NYT*. IA é uma ferramenta extremamente poderosa, mas o conhecimento acerca de seus perigos deve ser sempre divulgado.

(*) Doutor em Ciências pela Universidade de São Paulo, é professor, consultor e diretor do Fórum Brasileiro de Internet das Coisas – vnjitz@gmail.com.

Inscrições para última turma do ano do curso gratuito de Desenvolvedor de Games vão até 9 de setembro

Trabalhar em um mercado que combina criatividade, tecnologia e inovação ganha um incentivo da Prefeitura de São Paulo, que abriu inscrições para mais 160 vagas gratuitas no curso de Desenvolvedor de Games. Esta é a última turma da formação de 2026 que já abriu 680 vagas ao longo do ano.

Os interessados têm até o dia 9 de setembro para se inscrever pelo site. O curso é voltado a pessoas sem conhecimentos prévios sobre games, mas com interesse em ingressar nesta indústria que, segundo a Associação Brasileira das Desenvolvedoras de Jogos Digitais (Abragames), cresceu 695% em uma década, alcançando 1.042 estúdios de desenvolvedores de jogos em 2024, ante 150 em 2014. O curso tem atraído um público diversificado: entre os alunos matriculados, 19% são mulheres e 44% têm mais de 31 anos e apenas 16% têm entre 16 e 18 anos.

A formação de 180 horas aborda as diferentes etapas da criação de jogos digitais, do desenvolvimento de conceitos, narrativas e mecânicas à produção e apresentação de um jogo como projeto final. Também aborda questões comerciais como distribuição, animação, roteiro, áudio e empreendedorismo. Com duração de 12 meses, o curso é realizado presencialmente, uma vez por semana, na região central da capital paulista, e conta

Sergio Barzaghi/SECOM



também com módulos online, com turmas em todos os turnos, inclusive aos sábados.

Além da grade curricular, o curso promove uma imersão prática e conectada ao mercado. Somente em 2026, foram realizadas 47 atividades complementares, sendo 20 mentorias coletivas exclusivas para alunos matriculados, 12 meetups e 15 workshops sobre temas como dublagem para jogos, criação de músicas para jogos e como se preparar para trabalhar com games.

Como se inscrever

O processo de seleção priorizará moradores da cidade de São Paulo a partir de 16 anos e com Ensino Médio completo ou em andamento. Metade das vagas será reservada a estudantes e egressos de escolas públicas ou bolsistas integrais da rede particular de ensino. O processo seletivo prevê prova teórica online em 12 de setembro, com as aulas iniciando em 20 de outubro de 2026 (adesampa.com.br/hubsampagames).

News@TI

Copiloto de RH que automatiza processos e antecipa riscos para RH estratégico

A HRtech brasileira Joinin, acaba de anunciar o lançamento da Rub.IA, sua inteligência artificial desenvolvida para transformar o grande volume de informações geradas pelo RH de grandes empresas em processos estratégicos. A solução realiza análises preditivas e recomendações que apoiam a tomada de decisão na gestão de pessoas, além de atuar como copiloto do colaborador que deseja tirar dúvidas e consultar informações relacionadas ao RH. A Rub.IA funciona como uma “parceira” inteligente capaz de interpretar dados de diferentes áreas da gestão de pessoas e transformá-los em informações acionáveis. A proposta é atuar nas duas pontas e apoiando 4 stakeholders: alta liderança, gestores, time de RH e também o colaborador que deseja consultar informações ou tirar dúvidas. A Rub.IA torna o acesso às informações mais simples, e por meio de uma interface intuitiva, líderes, profissionais de RH e colaboradores podem realizar perguntas em linguagem natural e obter respostas rápidas, relatórios, análises e recomendações geradas pela inteligência artificial. O colaborador pode, por exemplo, tirar dúvidas e consultar seus direitos, conforme a política da empresa, sem necessidade de ir ao RH (www.joinin.com.br).

Editorias

Economia: Nelson Tucci (nelson.tucci@netjen.com.br)
Mercado/Negócios/Tecnologia/Agronegócios/ Espaço empresarial: Ricardo Souza (ricardosouza@netjen.com.br);
Livros: Ralph Peter (ralphpeter@agenteliterarioralph.com.br)
Comercial: comercial@netjen.com.br
Publicidade Legal: lilian@netjen.com.br

Webmaster/TI: Fabio Nader; *Editoração Eletrônica:* Ricardo Souza.
Revisão: Maria Cecília Camargo; *Serviço informativo:* Agências Brasil, Senado, Câmara, EBC, ANSA.

Artigos e colunas são de inteira responsabilidade de seus autores, que não recebem remuneração direta do jornal.

Jornal Empresas & Negócios Ltda

Administração, Publicidade e Redação: Rua Joel Jorge de Melo, 468, cj. 71 – Vila Mariana – São Paulo – SP – CEP.: 04128-080
Telefone: (11) 3106-4171 – E-mail: netjen@netjen.com.br
Site: (www.netjen.com.br). CNPJ: 05.687.343/0001-90
JUCESP, Nire 35218211731 (6/6/2003)
Matriculado no 3º Registro Civil de Pessoa Jurídica sob nº 103.