



Relatório de pesquisa 2025

Estado da infraestrutura de IA



Resumo executivo

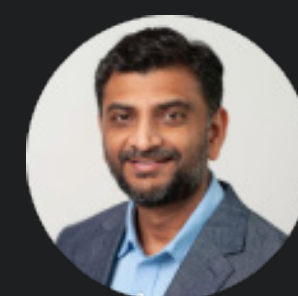
Embora a explosão de interesse na IA tenha gerado bastante experimentação, o foco agora mudou para a aplicação prática. Os líderes de tecnologia estão dando mais importância ao "como" do que ao "se": como integrar, proteger e dimensionar a IA generativa para gerar valor sustentável. Esta fase transformadora exige não só uma abordagem estratégica para o gerenciamento das plataformas, mas também uma implantação econômica.

No entanto, a transformação é complexa. As organizações enfrentam desafios consideráveis, como integração de sistemas legados, migrações complexas e a necessidade de adaptar soluções de IA aos seus próprios contextos corporativos.

Tendo em vista a necessidade crítica de insights baseados em dados, entrevistamos mais de 500 líderes corporativos de todo o mundo. Nossas descobertas revelam

um mercado de nuvem que passa por uma mudança drástica, onde as empresas que investem ativamente em infraestrutura de nuvem com tecnologia de IA estão liderando esse movimento.

Este relatório vai além do hype. Ele dá uma visão clara dos obstáculos práticos e das oportunidades estratégicas que estão surgindo. Oferecemos a pesquisa e os números que os líderes precisam para criar um consenso sobre investimentos essenciais em infraestrutura, capacitando-os a não só criar organizações que de fato sejam impulsionadas pela IA e pela nuvem, mas também a atravessar com confiança este período de transformação.



Nirav Mehta

Google Cloud Sr. Director, Product Management

Principais descobertas

Paisagem

01 A adoção da IA generativa é praticamente universal

98% das organizações estão testando, usando ou desenvolvendo a IA generativa de forma ativa no ambiente de produção

02 O ROI da IA generativa depende de casos de uso internos e externos

64% se concentram na melhora da experiência do cliente e no avanço das operações internas

Desafios

03 Segurança e dados são as maiores preocupações

70% tiveram dificuldades com a governança de dados, a integração de dados aos modelos de IA e a insuficiência de dados de treinamento

04 O custo é uma consideração crítica

83% priorizam a economia ao adotar soluções de infraestrutura de IA generativa

Infraestrutura

05 Uma plataforma de IA robusta é um critério importante na avaliação da infraestrutura

06 A computação de borda está ampliando o alcance da IA

73% consideram importante a implantação de modelos de IA generativa em dispositivos ou sistemas distribuídos

07 A nuvem híbrida oferece flexibilidade e controle

74% preferem uma abordagem de nuvem híbrida para implantações de IA generativa

08 A maioria das organizações depende de soluções de IA generativa dos provedores de nuvem

Uma infraestrutura otimizada é crucial para o sucesso da IA generativa



O sucesso de qualquer iniciativa de IA generativa depende da infraestrutura. Não se trata de apenas adicionar outra carga de trabalho, mas de construir uma base que possa respaldar as demandas exclusivas e em rápida evolução da IA, não importando se o objetivo é aumentar a satisfação do cliente ou reduzir os custos operacionais.

Essa base precisa aguentar várias coisas. Ela tem que ser segura para proteger a propriedade intelectual e os dados do cliente. Ela também precisa ser distribuída, oferecendo serviços gerenciados na nuvem pública e estendendo-se até o data center ou a borda da rede para atender aos requisitos de latência e soberania de dados.

Ela também precisa ser escalonável, eficiente e confiável para lidar com o processamento paralelo de grandes conjuntos de dados de treinamento, bem como com as janelas de contexto estendidas e o processamento em várias etapas de inferência e raciocínio. E, claro, tem que ser econômica.

Parece uma tarefa impossível? Não se você tiver um plano. A IA generativa está redefinindo a estratégia de TI das organizações, e ao que tudo indica não há sinais de que o ritmo vai desacelerar. Esta pesquisa revela em que ponto se encontram em suas respectivas jornadas as empresas líderes de todos os setores. Ao reexaminar a estratégia de infraestrutura em nuvem da organização, pense em quais provedores têm capacidade de atender não apenas a especialistas em IA, mas também aos seus desenvolvedores e suas linhas de negócios.

Descoberta 01

A adoção da IA generativa é quase universal em todos os setores

01

02

03

04

05

06

07

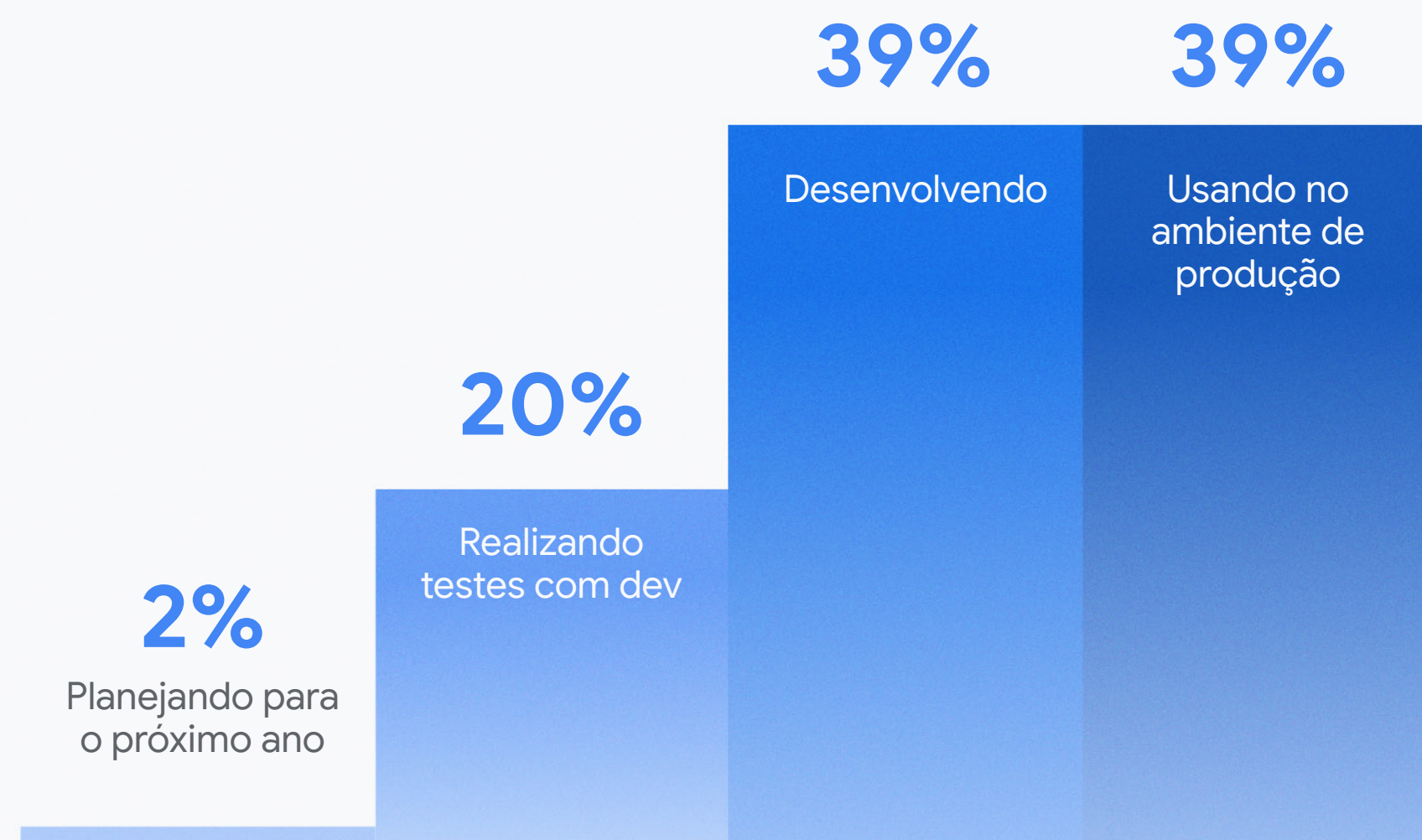
08

No ambiente de trabalho atual, a IA generativa é de extrema importância.

98%

das organizações estão testando, usando ou desenvolvendo a IA generativa de forma ativa no ambiente de produção

Em que estágio está sua organização, em termos de adoção da IA generativa?



79%

dos líderes de tecnologia consideram a IA generativa como sendo pelo menos muito importante para as operações comerciais atuais e futuras de suas organizações

Qual a importância da IA generativa para as operações comerciais atuais e futuras da sua organização?



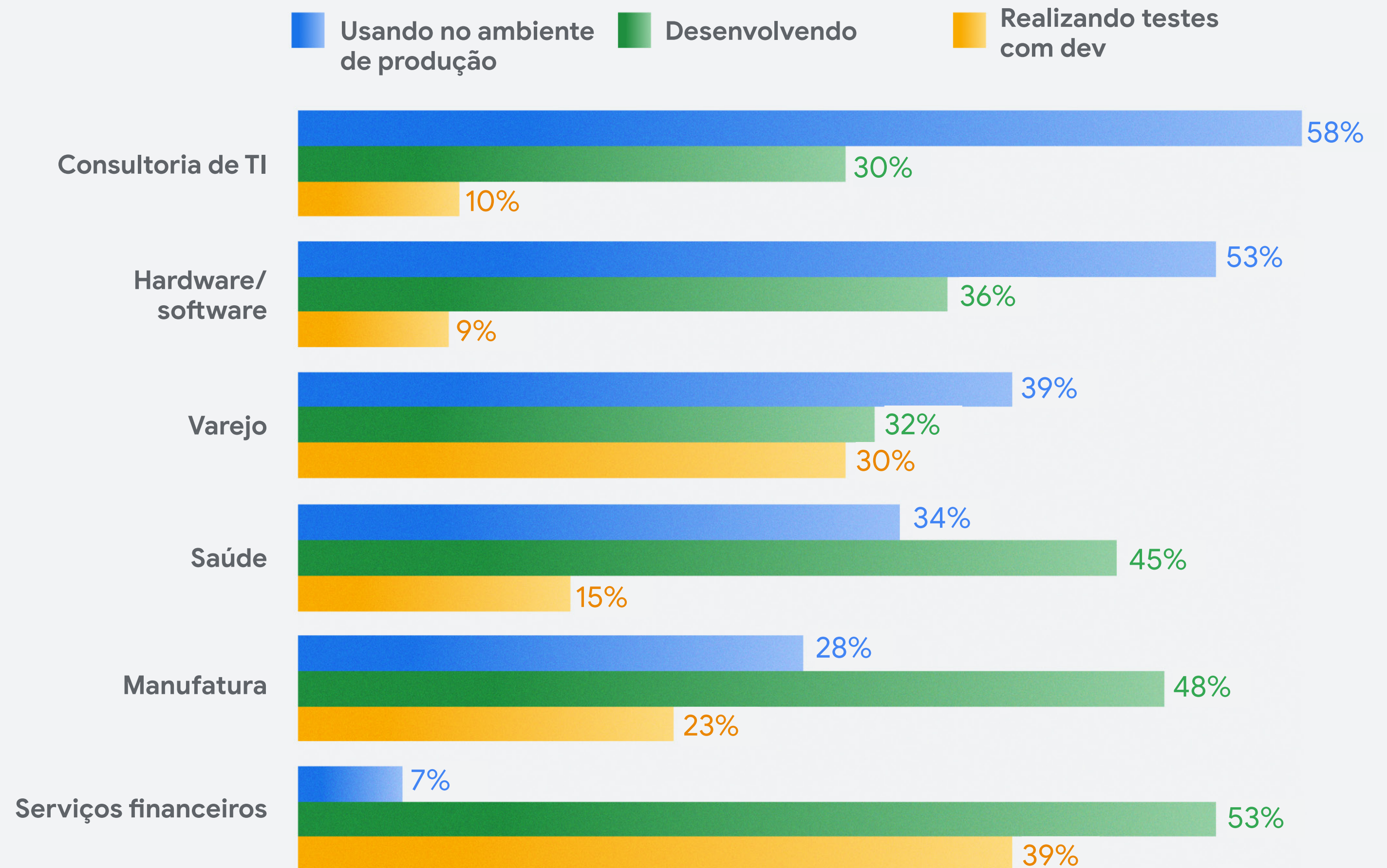
Uma visão sobre os setores

Essas descobertas são consistentes em todos os setores: líderes de tecnologia em geral adotaram a IA generativa em suas organizações e concordam que ela é essencial para os negócios. Vamos analisar os dados em detalhes.

Em todos os setores, a maioria das organizações está desenvolvendo ou implantando a IA generativa na produção, o que mostra que as empresas já estão investindo pesado nessa tecnologia e planejam continuar. Como era de se esperar, as organizações de setores relacionados à tecnologia foram mais rápidas na hora de migrar pilotos de IA generativa e implementá-los na produção e, de modo geral, tiveram taxas de adoção maiores.

No entanto, o imperativo da IA generativa levou quase todas as empresas, independentemente do setor, a investir nessa inovação. Mesmo setores que apresentam menores índices de migração para produção, como serviços financeiros e manufatura, registraram taxas elevadas de testes e desenvolvimento.

Em que estágio está sua organização, em termos de adoção da IA generativa? (por setor)

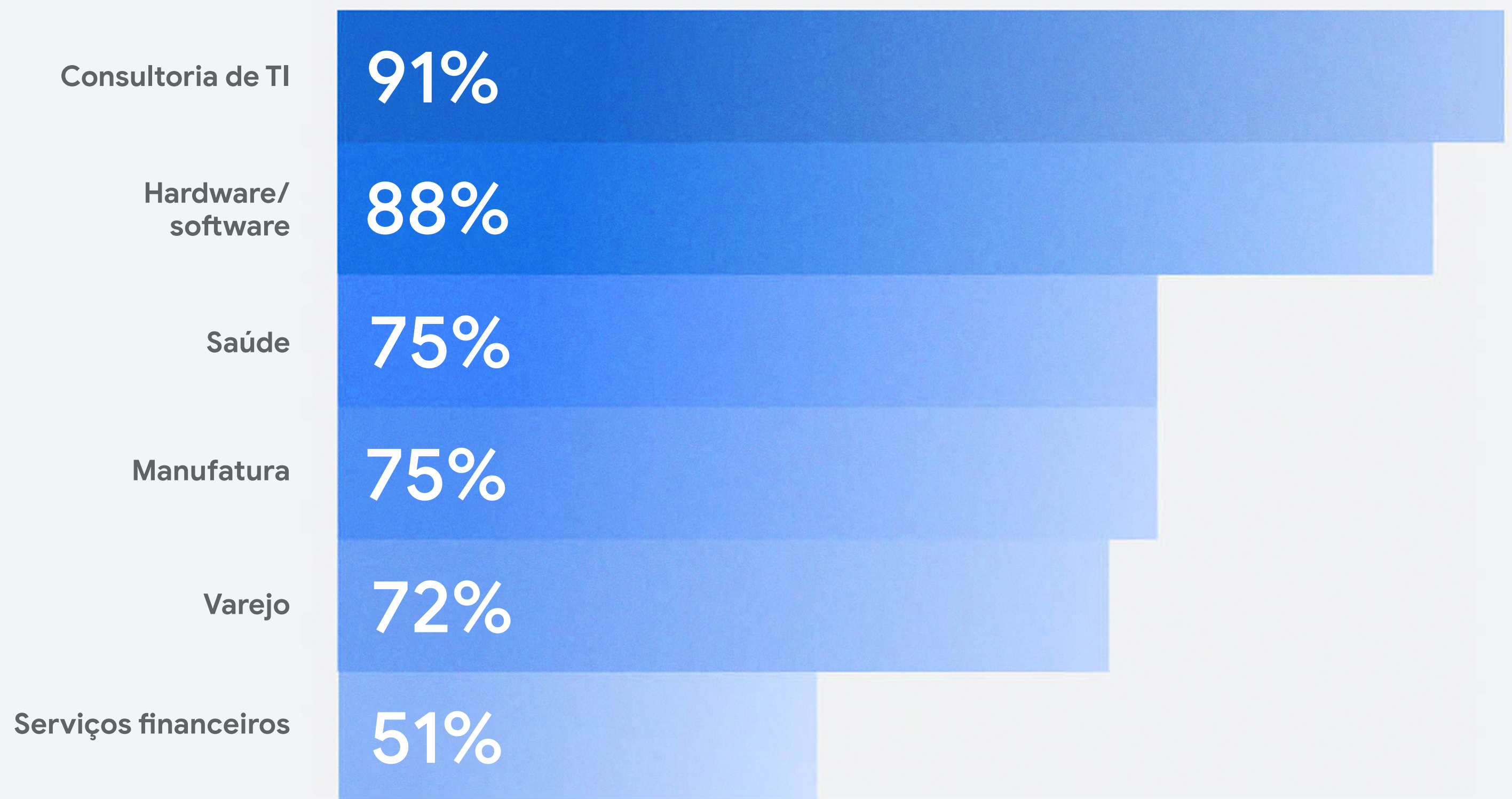


Líderes de tecnologia em todos os setores concordam que a IA generativa é essencial para o sucesso de suas organizações.



Embora impressionantes 79% de todos os líderes acreditem que a IA generativa é muito ou extremamente importante para seus negócios, essa porcentagem é ainda maior para setores relacionados à tecnologia, como consultoria de TI (91%) e hardware/software de computador (88%).

Qual a importância da IA generativa para as operações comerciais atuais e futuras da sua organização? (Extremamente/muito importante por setor)



Por Isso é importante para você?

A IA não é mais um conceito futurista: é um impulsionador de negócios essencial e uma mudança fundamental na forma como as organizações operam. Os líderes de TI deixaram de falar sobre o potencial reconhecido e passaram a desenvolver uma infraestrutura capaz de dar suporte às crescentes demandas das cargas de trabalho de IA.

Com a maioria das organizações desenvolvendo ou usando a IA generativa no ambiente de produção, ela está permitindo que as organizações encontrem novas formas de aumentar a eficiência e desbravem novos caminhos para o sucesso. As decisões de infraestrutura que você tomar hoje vão determinar a capacidade da sua organização de competir em um futuro impulsionado pela IA, independentemente de setor ou região.





Descoberta 02

O ROI da IA generativa depende de casos de uso internos e externos

A IA generativa proporciona bons resultados em diversos casos. Embora seja natural pensar que "quanto mais, melhor", nossa pesquisa revelou resultados interessantes sobre os pontos mais indicados para direcionar suas iniciativas com a IA generativa.

01 02 03 04 05 06 07 08

As organizações estão usando a IA generativa para melhorar as interações com os clientes, personalizar experiências e criar novas propostas de valor.

Nossa pesquisa mostra que muitos líderes de tecnologia (64%) estão encontrando um equilíbrio entre melhorar a experiência do cliente e otimizar as operações internas. Adotar a IA generativa em casos externos pode ser a diferença entre oferecer uma resposta personalizada e precisa no lugar de uma resposta genérica para um cliente.

Por exemplo, a integração da IA generativa aos dados operacionais da sua organização permite que um representante de atendimento ao cliente ou um chatbot forneçam respostas relevantes e em tempo real, algo muito valorizado pelos clientes. Ao adotar a IA generativa em etapas internas, as organizações podem otimizar processos, impulsionar a produtividade e reduzir os custos operacionais.

Como você classificaria, prioritariamente, o uso atual ou planejado da IA generativa na sua organização?

64%

Voltado tanto
ao ambiente
interno quanto
externo

25%

Voltado
principalmente
ao ambiente
interno

11%

Voltado
principalmente ao
ambiente externo

Três casos de uso principais estão impulsionando a adoção da IA generativa:

01

Análise de dados

É mais do que a geração de relatórios. Trata-se de extrair insights úteis a partir de amplos conjuntos de dados, identificar padrões e permitir a tomada de decisões baseadas em dados, em larga escala e com rapidez.

02

Automação do atendimento ao cliente

Automatizar tarefas repetitivas e otimizar fluxos de trabalho libera capital humano para atividades de maior valor.

03

Automação de processos internos

Vá além dos chatbots básicos. A automação atual envolve a criação de sistemas inteligentes que podem entender consultas detalhadas, fornecer respostas precisas e personalizadas e resolver problemas de modo eficiente.

A análise de dados é a prioridade máxima na implementação da IA generativa, seguida de longe pelo atendimento ao cliente e a automação de processos

Análise de dados

50%

Automação do
atendimento ao
cliente

40%

Automação de
processos internos

40%

Geração de
códigos

25%

Criação de
conteúdo

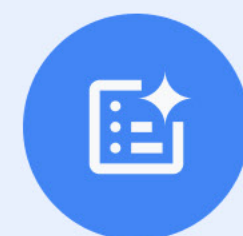
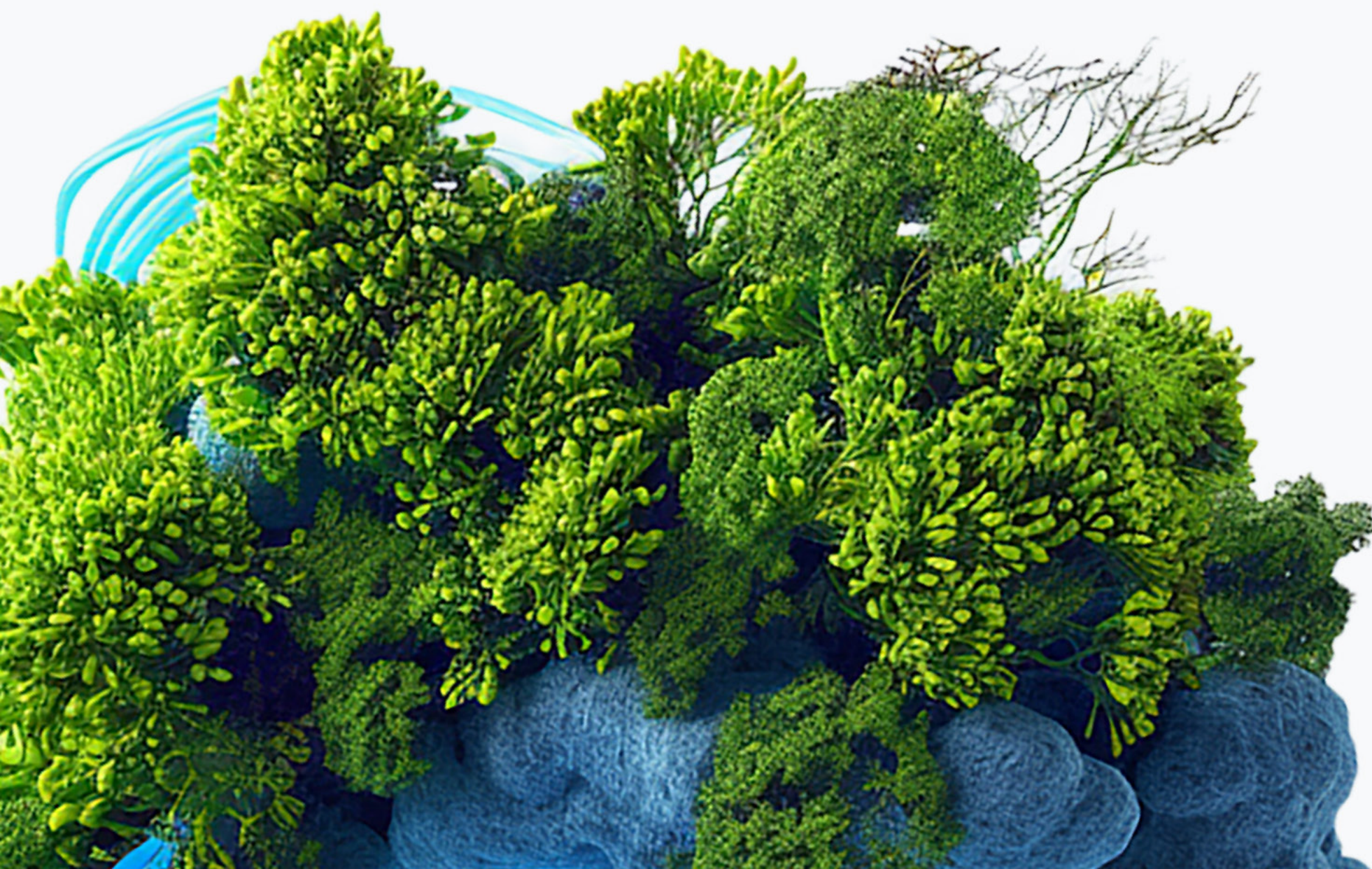
20%

Design e
criatividade

19%

Uma visão sobre os setores

As prioridades para implementação da IA generativa variam bastante entre setores.



Análise de dados e geração de código

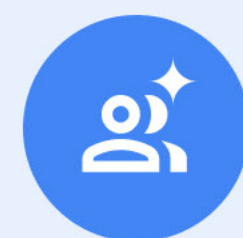
60%

em consultoria de TI

69%

em hardware/software

Essas indústrias também dão maior prioridade à **geração de código** do que outras (47% e 51%)



Automação do atendimento ao cliente

68%

em serviços financeiros

67%

no varejo



Automação de processos internos

55%

em manufatura

54%

em serviços financeiros

Qual dos casos de uso de IA generativa a seguir sua organização considera a maior prioridade para implementação? (por setor)

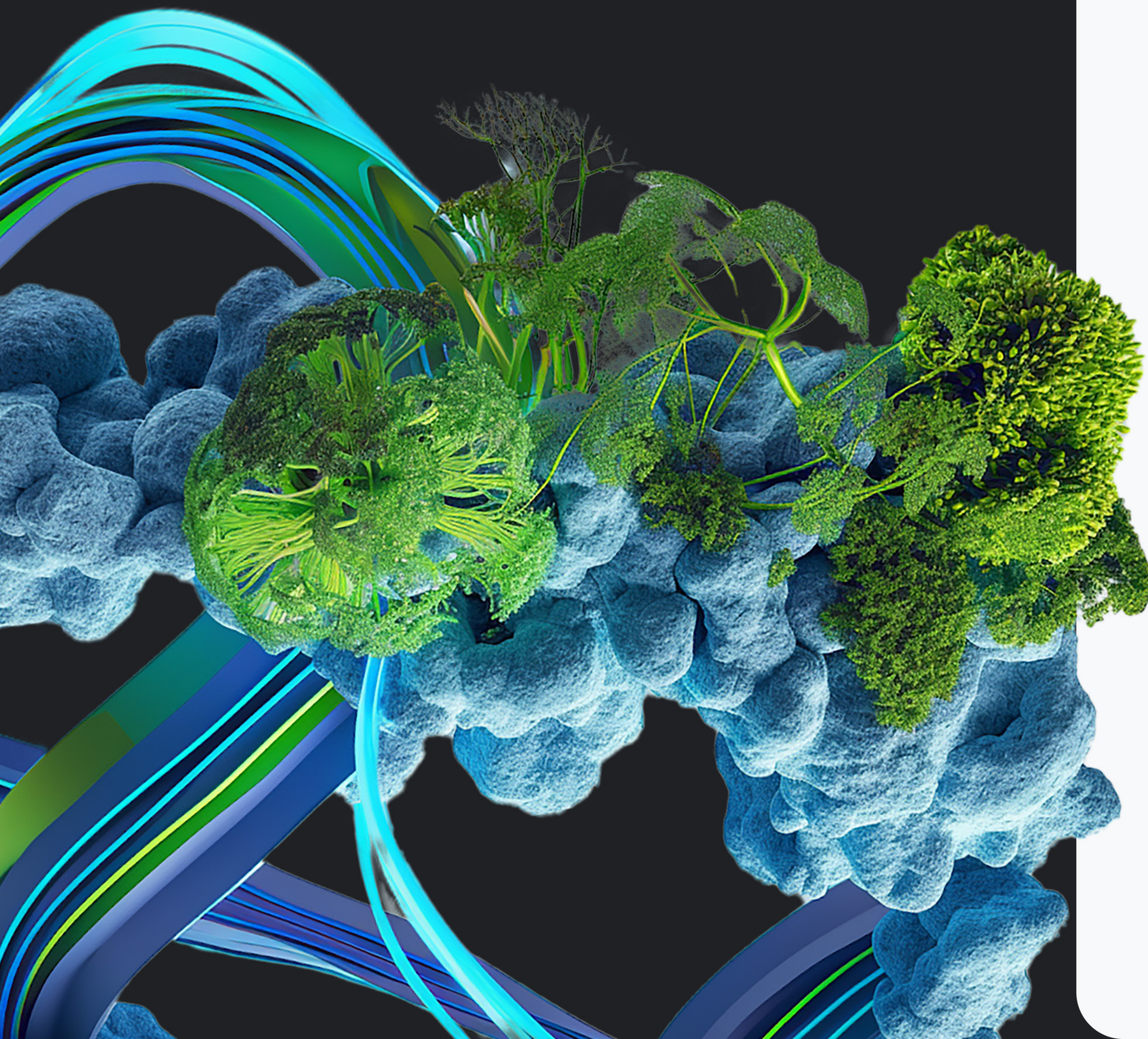
Prioridade máxima de implementação	Consultoria de TI	Hardware/software	Serviços financeiros	Varejo	Saúde	Manufatura	Publicidade, RP
Análise de dados	60%	69%, ↑	47%	49%,	55%,	45%,	4%, ↓
Automação do atendimento ao cliente	22%	25%	68%, ↑	67%, ↑	47%, ↑	30%	8%, ↓
Automação de processos internos	34%	31%	54%, ↑	47%	43%	55%, ↑	0%, ↓
Geração de códigos	47%, ↑	51%, ↑	7%	9%	19%	18%	0%, ↓
Criação de conteúdo	16%	9%	11%	12%	11%	23%	96%, ↑
Design e criatividade	18%	8%	9%	11%	17%	25%	92%, ↑

Por Isso é importante para você?

Para aproveitar ao máximo a IA generativa, as organizações precisam adotar uma abordagem estratégica, integrando a tecnologia aos fluxos de trabalho internos e aos aplicativos voltados para o cliente. Para isso, uma infraestrutura flexível e escalonável que permita diversas cargas de trabalho de IA se faz necessária.

Além disso, embora entender os principais casos de uso no próprio ramo seja importante, saber o que acontece em outros setores pode estimular a inovação e gerar mais valor. Ao descobrir como outros setores estão aproveitando a IA generativa, as organizações podem identificar novas aplicações, expandir suas soluções e conseguir vantagens sobre a concorrência.





REGNOLOGY

No setor de serviços financeiros, a Regnology está aproveitando os serviços do Google Cloud para desenvolver um chatbot de relatórios regulamentares. Esse chatbot foi criado para fornecer respostas precisas a consultas regulamentares com mais rapidez, tanto para usuários internos quanto externos.

Diante das crescentes demandas regulatórias e dos volumes de dados cada vez maiores, a Regnology está ajudando os clientes a otimizar suas operações e permanecer em conformidade com as regulamentações em evolução. Como líder global em relatórios regulamentares, a empresa enfrentou dificuldades com a falta de flexibilidade da sua infraestrutura interna. Um novo chatbot, criado com base nos serviços do Google Cloud, ajuda a gerar rapidamente respostas precisas a consultas regulatórias, de usuários internos e externos.



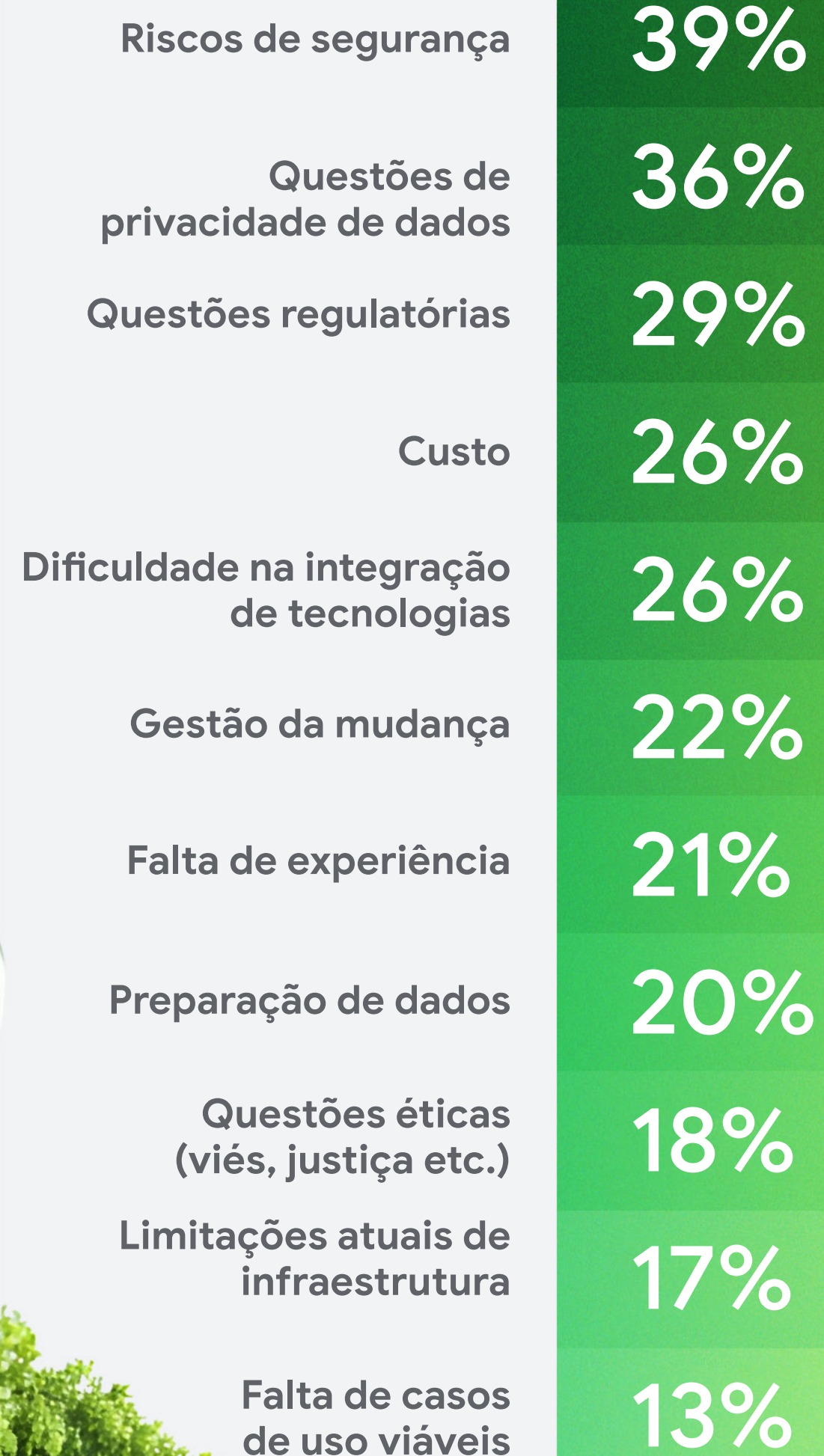
Descoberta 03

Segurança e dados são os principais desafios

A proteção de informações sensíveis continua sendo prioridade máxima. As organizações querem saber como proteger dados reservados mesmo usando modelos de IA externos. Além disso, também querem garantir a conformidade com as regulamentações de proteção de dados, que estão em constante evolução.

01 02 03 04 05 06 07 08

Quais das opções a seguir correspondem aos maiores desafios para a adoção da IA generativa na sua organização?



Nossa pesquisa ressalta as questões críticas de segurança e privacidade que os líderes técnicos enfrentam.

62%

apontam os riscos de segurança e/ou privacidade de dados como o maior desafio

As maiores questões com a adoção da IA generativa são:

01 Riscos de segurança

A proteção de dados privados e da propriedade intelectual é inegociável. A natureza da IA generativa, envolvendo LLMs e amplos conjuntos de dados, apresenta novos desafios de segurança que precisam ser abordados de forma proativa.

39%

02 Questões de privacidade de dados

As organizações precisam garantir que suas implementações de IA generativa protejam dados sensíveis e estejam em conformidade com as regulamentações vigentes, como o Regulamento Geral de Proteção de Dados (GDPR), a Lei de Privacidade do Consumidor da Califórnia (CCPA) e muitas outras.

36%

03 Questões regulatórias

As regulamentações em torno da IA generativa estão aumentando — a [Lei de IA da UE](#) é apenas uma delas — e é importante que as empresas se mantenham em conformidade. Alguns temas sob maior escrutínio regulatório são a capacidade de explicação da IA, a proteção da propriedade intelectual e a defesa contra a desinformação.

29%

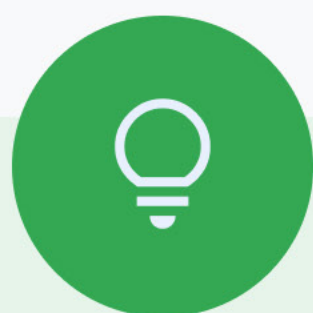
70% das organizações que adotaram a IA generativa tiveram dificuldades com dados.

Isso inclui governança de dados, integração de dados em modelos de IA e dados de treinamento insuficientes que se alinham com as principais descobertas do [Um guia executivo para ajudar você a agregar valor com dados e IA](#).

Em geral, as organizações querem adotar a IA generativa. No entanto, elas enfrentam desafios na hora de passar dos projetos piloto para as implantações de produção. E sabe qual é o principal gargalo? Dados. Nossa pesquisa descobriu que a disponibilidade e a qualidade dos dados, e além disso a plataforma que os armazena, são os fatores mais importantes na implementação da IA generativa.

Para agravar ainda mais esse desafio, há uma necessidade crucial de recursos de linhagem de dados. A linhagem de dados oferece um registro abrangente e auditável da origem, das transformações e dos direitos de uso de um conjunto de dados. Isso permite que as organizações afirmem definitivamente, com justificativa legal, ética e baseada em políticas, seu direito de utilizar dados específicos para uma aplicação de IA específica. Sem uma linhagem de dados verificável, as organizações enfrentam riscos elevados, incluindo possíveis responsabilizações jurídicas, não conformidade regulatória e danos à reputação, o que dificulta a implantação de modelos de IA treinados com dados de procedência incerta.

Disponibilidade e qualidade de dados, uma plataforma de IA robusta, escalabilidade eficiente e viabilidade técnica lideram em importância geral para a nova tecnologia de IA generativa.



As organizações estão preocupadas em manter os dados dos clientes privados e em segurança e conformidade. Já sabemos que, para garantir uma governança de dados de qualidade, as organizações precisam unificar seus dados; no entanto, isso pode exigir uma grande infraestrutura. E talvez não seja surpresa alguma que a qualidade dos dados, uma plataforma robusta e a escalabilidade eficiente sejam os fatores mais importantes na mente dos líderes de tecnologia quando se trata de implementar a IA generativa. Os dados e a infraestrutura, afinal, ainda reinam supremos.

Quais são os 5 fatores mais importantes ao escolher uma tecnologia de IA generativa?

Plataforma de IA robusta

45%

Disponibilidade e qualidade dos dados

45%

Escalonabilidade eficiente

45%

Viabilidade técnica

44%

Por Isso é importante para você?

A criação de uma infraestrutura de IA segura e compatível é algo inegociável. Afinal, os líderes de TI têm que priorizar a segurança e a privacidade dos dados em sua estratégia de IA para reduzir riscos e manter a confiança.

Seja qual for o setor, os dados costumam estar fragmentados em diferentes plataformas e formatos, e, com isso, a implementação de políticas unificadas de segurança e governança se torna um desafio. Para aproveitar ao máximo a IA, é fundamental criar uma só fonte de verdade com uma plataforma unificada em todas as formas de dados, incluindo formatos abertos.

Além disso, uma arquitetura altamente escalonável é crucial, permitindo a unificação de sistemas transacionais e analíticos sem afetar o desempenho.

Com isso, é possível ter acesso aos dados entre esses sistemas, mesmo com cargas de trabalho transacionais exigentes.

À medida que mais dados e aplicativos saem dos data centers locais e se afastam da infraestrutura e dos mecanismos de segurança tradicionais, as abordagens de segurança de dados precisam ser adaptadas à nuvem.

Quer saber mais?

- [Como acelerar a transformação de IA com segurança, privacidade e conformidade](#)
- [Como aproveitar o potencial completo da IA generativa com os bancos de dados operacionais](#)
- [Construindo uma plataforma de dados segura com o Google Cloud](#)



moii.ai

A Moii.AI implementa agentes de IA de visão em CCTV para melhorar a segurança

Com mais de um bilhão de câmeras de CCTV no mundo, a Moii.AI automatiza o processo de revisão e análise com IA, transformando sistemas de câmeras tradicionais em agentes autônomos para criar locais de trabalho mais seguros e produtivos.

“Graças à interface intuitiva do BigQuery, até mesmo engenheiros sem muita experiência em nuvem podem usar a ferramenta já na primeira semana e começar a trabalhar em projetos de dados. Com outra solução, precisaríamos de uma equipe duas vezes maior para manter a infraestrutura que construímos.”

Lakshman Balasubramanian
Head of AI e Co-founder, moii.AI



Descoberta 04

A eficiência de custos é uma consideração importante

A IA exige recursos computacionais significativos, e os custos gerais para implementação e manutenção da IA podem variar amplamente dependendo de fatores como escala, complexidade do modelo, requisitos de dados, manutenção de infraestrutura e talentos atuais.

01 02 03 04 05 06 07 08



Embora o potencial da IA generativa possa ser ilimitado, os orçamentos não são.

83%

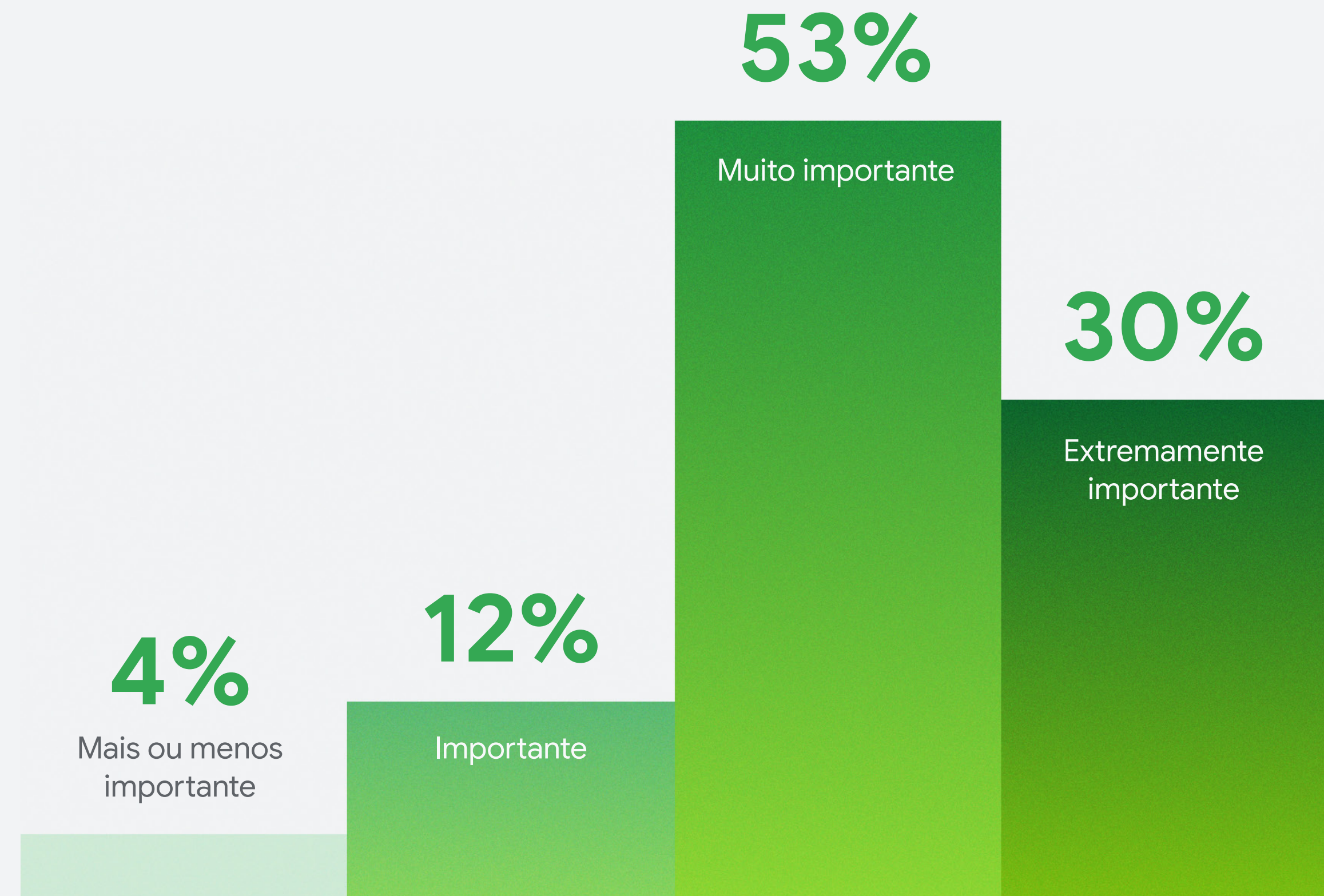
dos líderes de tecnologia acreditam que o custo é um fator essencial na avaliação de soluções

Aproveitar essa tecnologia nova e avançada pode trazer gastos inesperados.

O provedor de nuvem certo pode oferecer soluções econômicas de treinamento, ajuste fino e inferência de cargas de trabalho de IA, incluindo opções para instâncias pontuais, instâncias reservadas e pagamento por uso.

Como vimos na [Descoberta 03](#), o custo das soluções de IA generativa também é um desafio significativo na adoção dessa tecnologia, ocupando o 4º lugar geral.

Qual a importância da relação custo-benefício no seu processo de tomada de decisão para soluções de infraestrutura de IA generativa?



Embora o custo associado às soluções de IA generativa seja uma consideração importante, os benefícios dessa tecnologia também impactam positivamente os resultados financeiros da organização.



Desde melhorar a produtividade, aumentar as vendas e reduzir os custos operacionais, vários casos de uso da IA generativa proporcionam uma economia de custos.

Onde você espera o maior ROI da IA generativa?

Aumento da produtividade dos funcionários

Melhora da satisfação e do engajamento do cliente

Simplificação de processo e fluxos de trabalho

Melhora da competitividade e ganhos de participação de mercado

Aceleração do crescimento de receitas

Aumento das vendas e receitas

Redução dos custos operacionais

22%

21%

20%

18%

14%

13%

13%

Por Isso é importante para você?

A IA generativa simplifica processos e automatiza tarefas, criando eficiências que ajudam as organizações a reduzir custos. No entanto, apesar das eficiências disponíveis, quantificar o valor comercial da IA generativa pode ser desafiador. E isso ocorre porque, além da economia de custos em toda a organização, a IA generativa pode gerar gastos cada vez maiores com tecnologia, incluindo custos computacionais, desenvolvimento e gerenciamento de modelos e gerenciamento de dados. É por isso que a otimização de custos é uma consideração crucial.



Melhore a eficiência de custos da sua IA

Para gerenciar efetivamente os custos da IA generativa, é necessário uma abordagem estratégica que vai além de simplesmente escolher os serviços de menor preço. Aqui estão três áreas importantes em que você deve se concentrar para alcançar a melhor relação custo-benefício em sua infraestrutura de IA.

Ao considerar cuidadosamente esses fatores e selecionar um provedor de nuvem comprometido com a otimização de custos, as organizações podem liberar todo o potencial da IA generativa sem incorrer em despesas insustentáveis.



Dimensione os recursos conforme necessário, com controle granular

Aproveite plataformas de nuvem que oferecem controle detalhado sobre recursos computacionais para evitar o excesso de provisionamento. Procure recursos como dimensionamento automático, que ajusta dinamicamente os recursos com base nas demandas de carga de trabalho em tempo real e permite instâncias de computação heterogêneas (por exemplo, CPUs, GPUs, TPUs) para atender aos requisitos específicos do modelo. Isso minimiza o desperdício de recursos e garante que você pague apenas pelo que realmente usa.



Aproveite as pilhas de hardware e software otimizadas para IA

Escolha provedores de infraestrutura com histórico comprovado de otimização do desempenho de IA. Isso inclui o uso de aceleradores de hardware especializados (como TPUs ou GPUs projetados para IA) e o aproveitamento de estruturas de software altamente otimizadas (por exemplo, JAX, TensorFlow, PyTorch) que maximizam a utilização do hardware. Essas otimizações podem reduzir significativamente os tempos de treinamento e inferência, levando a economias substanciais de custos.



Implemente o gerenciamento e o agendamento inteligentes de recursos

Conheça ferramentas avançadas de gerenciamento de recursos que automatizam a alocação e o agendamento das cargas de trabalho de IA. Recursos como agendamento dinâmico de carga de trabalho, instâncias preemptivas (quando apropriado para cargas de trabalho tolerantes a falhas) e mecanismos inteligentes de armazenamento em cache podem reduzir significativamente o tempo de inatividade de recursos e otimizar a utilização geral da infraestrutura, resultando em custos operacionais mais baixos.

Como as FinOps geradas com IA generativa podem ajudar você a otimizar os retornos sobre os investimentos em tecnologia

As FinOps trazem disciplina para a nuvem ao alinhar equipes de tecnologia, finanças e negócios com uma estrutura operacional abrangente focada em gerar retornos econômicos sobre investimentos em tecnologia, com organizações reduzindo seus gastos com nuvem em até 30%.

Avalie o nível de preparação da empresa para adotar a IA generativa com a estrutura de Cloud FinOps pensada para IA generativa.



nuro

Nuro, uma empresa de tecnologia de direção autônoma, está usando o Google Cloud para desenvolver o Nuro Driver, seu sistema de direção autônoma de ponta, ao mesmo tempo em que melhora consideravelmente a relação custo-benefício dos treinamentos. Usando as TPUs do Google Cloud, a Nuro agora pode processar petabytes de dados de direção do mundo real e treinar modelos de IA para produzir mapas em escala urbana e detectar obstáculos na estrada duas vezes mais rápido, sem aumentar seus custos.



Descoberta 05

Uma plataforma de IA robusta é um critério importante na avaliação da infraestrutura

Os entrevistados querem tudo - desempenho, escala, eficiência - em uma só plataforma que possa treinar, implantar e gerenciar a infraestrutura de IA.

01 02 03 04 05 06 07 08

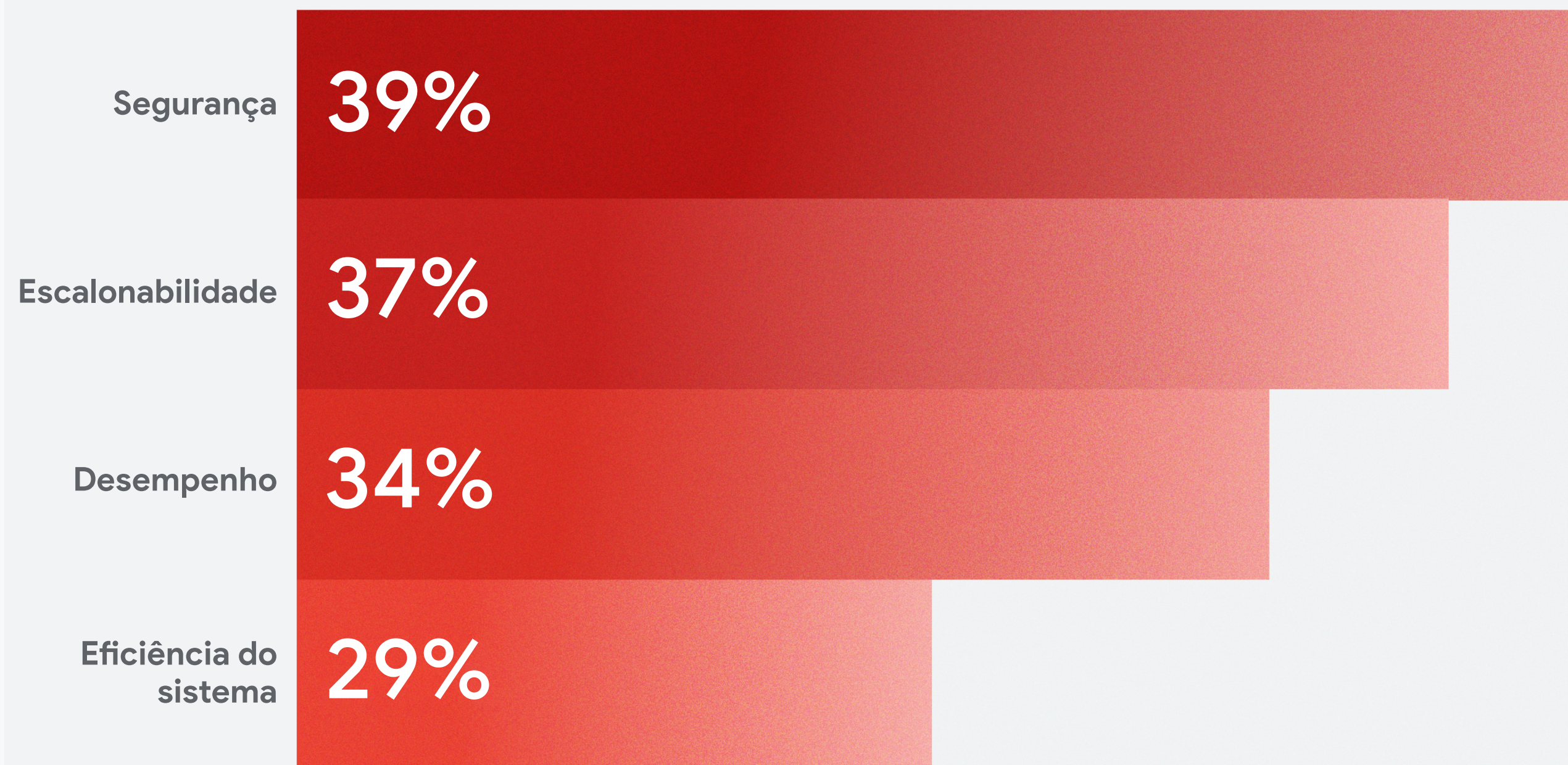
45% dos entrevistados
querem uma plataforma
de IA robusta

Não é de se admirar que, com a velocidade impressionante de mudanças nos modelos, aplicativos e arquitetura da IA generativa, as organizações estejam priorizando plataformas capazes de acompanhar o ritmo (confira na [Descoberta 03](#) os resultados da pesquisa).

Fatores importantes de infraestrutura

Poucas organizações estão treinando modelos de fundação em larga escala. Só que a escala e o desempenho são fatores importantes em todas as geografias, tanto para quem está disponibilizando modelos rapidamente para milhares de usuários finais internos como para quem está lidando com milhões de clientes. Como as plataformas de IA precisam lidar com comunicação rápida entre agentes e um volume ainda maior de solicitações de inferência, as demandas de desempenho e escalabilidade se intensificam. Uma plataforma robusta precisa permitir esses tipos de cargas de trabalho, independentemente de a arquitetura ser containerizada ou não, com inovação constante em computação, armazenamento e dimensionamento de rede, compartilhamento de recursos, acesso e cálculo de rotas.

Quais são os fatores mais importantes ao avaliar um sistema de infraestrutura de IA generativa?



Por Isso é importante para você?

Uma plataforma de IA robusta pode dar suporte a diversos usuários, de desenvolvedores a cientistas e operadores de TI. Ela pode dimensionar e otimizar arquiteturas e estruturas em evolução e também pode lidar com todo o ciclo de vida da IA, desde o desenvolvimento, passando pela implantação, até o gerenciamento.



Infraestrutura inovadora

A IA popularizou as GPUs. No entanto, sua infraestrutura principal também deve ter a flexibilidade para oferecer o melhor custo-benefício para cada caso, incluindo CPUs e silício personalizado. Os dados são essenciais para a IA; a plataforma precisa oferecer vários mecanismos de armazenamento para ingerir, armazenar e entregar dados. Por exemplo, sistemas de arquivos paralelos são o armazenamento mais recomendado para certos treinamentos de IA, mas o armazenamento de objetos e o armazenamento em bloco podem ser muito mais eficientes quando os dados de treinamento são estáticos. É necessário considerar a rede de acordo com cada host, entre os hosts de um cluster, entre redes de nuvem, entre nuvens públicas e nas nuvens privadas.



Software dinâmico e integrado

À medida que as implementações de IA generativa evoluem, o software da plataforma de IA também precisa evoluir. Por exemplo, janelas de contexto de comprimento variável exigem um balanceamento de carga de rede mais inteligente para rotear corretamente as solicitações. O raciocínio em várias etapas pode exigir vários hosts de computação para manipular o processamento de modo dinâmico. O alto custo de infraestrutura de grandes trabalhos de IA exige que o software minimize as interrupções inevitáveis com observabilidade expansiva, pontos de verificação mais sofisticados e até mesmo que os clusters de IA operem em modo degradado.



Atende a múltiplas funções

Algumas organizações desejam acesso total à infraestrutura de IA porque querem implementar novas técnicas, estruturas e modelos para maximizar os recursos de hardware. Outras preferem pegar os modelos atuais e refiná-los para seus casos de uso específicos — e entendem o controle da infraestrutura como algo secundário. E tem aquela empresa que quer simplesmente acessar modelos de IA pré-configurados usando a API para integrar a um aplicativo. Todas essas organizações se beneficiam quando as diferentes modalidades operam a partir da mesma infraestrutura escalonável, eficiente e confiável.



Em um cenário digital congestionado, a Moloco oferece soluções de publicidade geradas com IA, usando dados primários para ajudar empresas a segmentar e adquirir usuários de alto valor com base no comportamento do consumidor em tempo real. Em última análise, o que a empresa faz é proporcionar maiores taxas de conversão e retorno do investimento.

A Moloco aproveita previsões de uma dúzia de redes neurais profundas com uma plataforma que ingere 10 petabytes de dados por dia a uma taxa máxima de 10,5 milhões de consultas por segundo. Ela conta com o Google Kubernetes Engine para lidar com essa enorme carga de dados e otimizar a eficiência do serviço, mantendo a relação custo-benefício.



Descoberta 06

A computação de borda está ampliando o alcance da IA

01 02 03 04 05 06 07 08



A IA generativa é utilizada em um número cada vez maior de contextos, e muitos ocorrem fora dos tradicionais data centers centralizados.

A computação de borda inclui dispositivos de IoT (como geladeiras inteligentes), veículos autônomos e dispositivos móveis. Ela distribui o processamento, reduz a latência e exige recursos computacionais eficientes. Essa distribuição computacional também pode melhorar o desempenho da rede e aumentar a adesão aos regulamentos de privacidade de dados (como GDPR e CCPA), permitindo o tratamento local de dados e minimizando a transmissão de informações sensíveis.

Quase todos os setores já dependem, em certo grau, de IoT ou dispositivos móveis para dar suporte a aplicativos em tempo real e manter as operações em execução. Na área da saúde, manufatura e varejo, a computação de borda usa e processa informações de sensores para alimentar análises, manter o controle de qualidade e até mesmo operar máquinas de modo autônomo.

Casos específicos:



Processamento local de sinais vitais dos pacientes a partir de dispositivos médicos para detecção imediata de anomalias e notificação à equipe



Verificação em tempo real da conformidade dos equipamentos de proteção individual (EPI) entre os profissionais de saúde por meio da implantação de modelos de visão computacional na borda

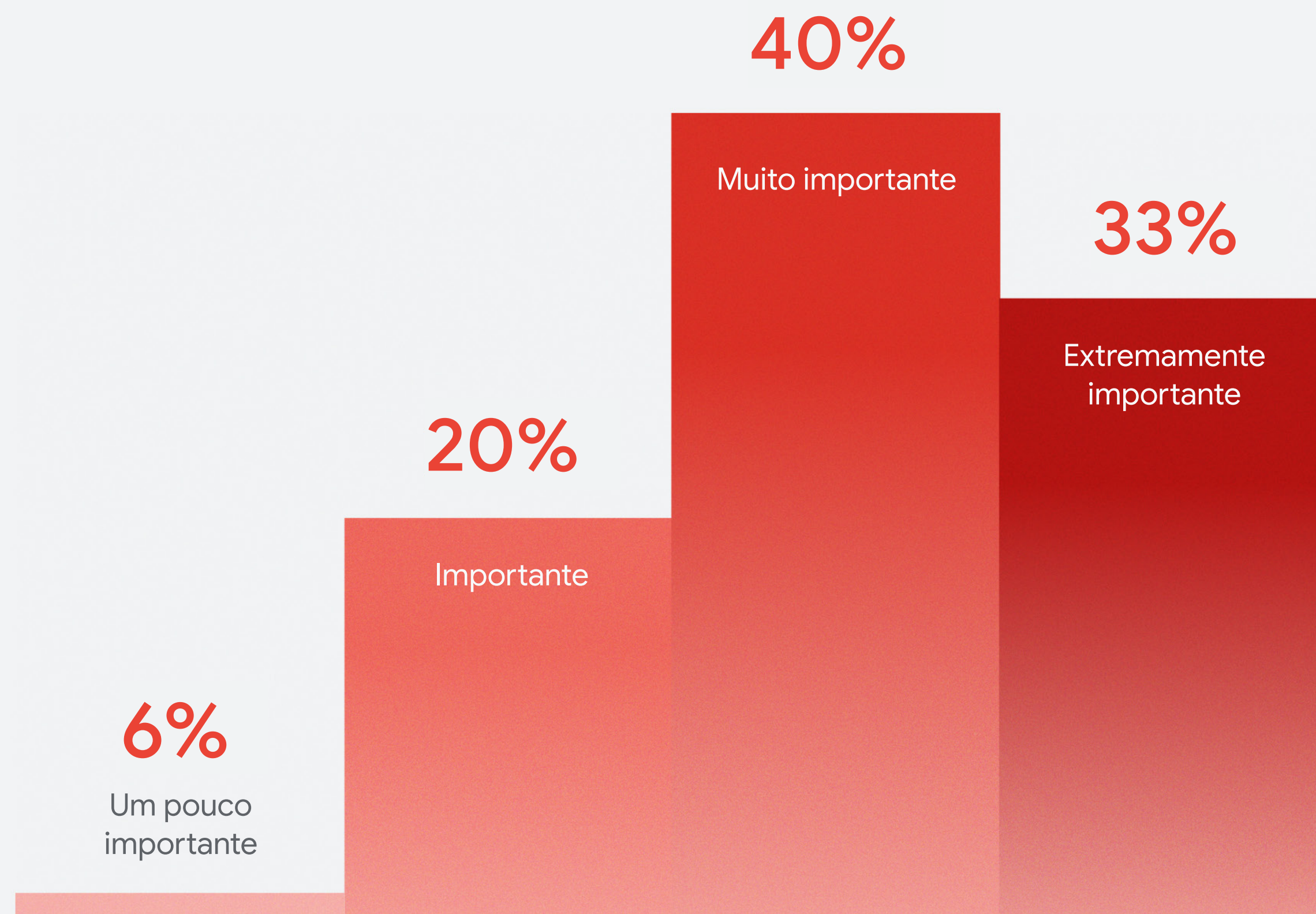


Manutenção preditiva de equipamentos industriais baseada na análise de dados de sensores



Gerenciamento de inventário em tempo real usando análise de vídeo baseada na borda

Qual a importância de implantar modelos de IA generativa na borda (por exemplo, em dispositivos IoT ou dispositivos móveis) para sua organização?



73%

das organizações dizem que a computação de borda é muito ou extremamente importante

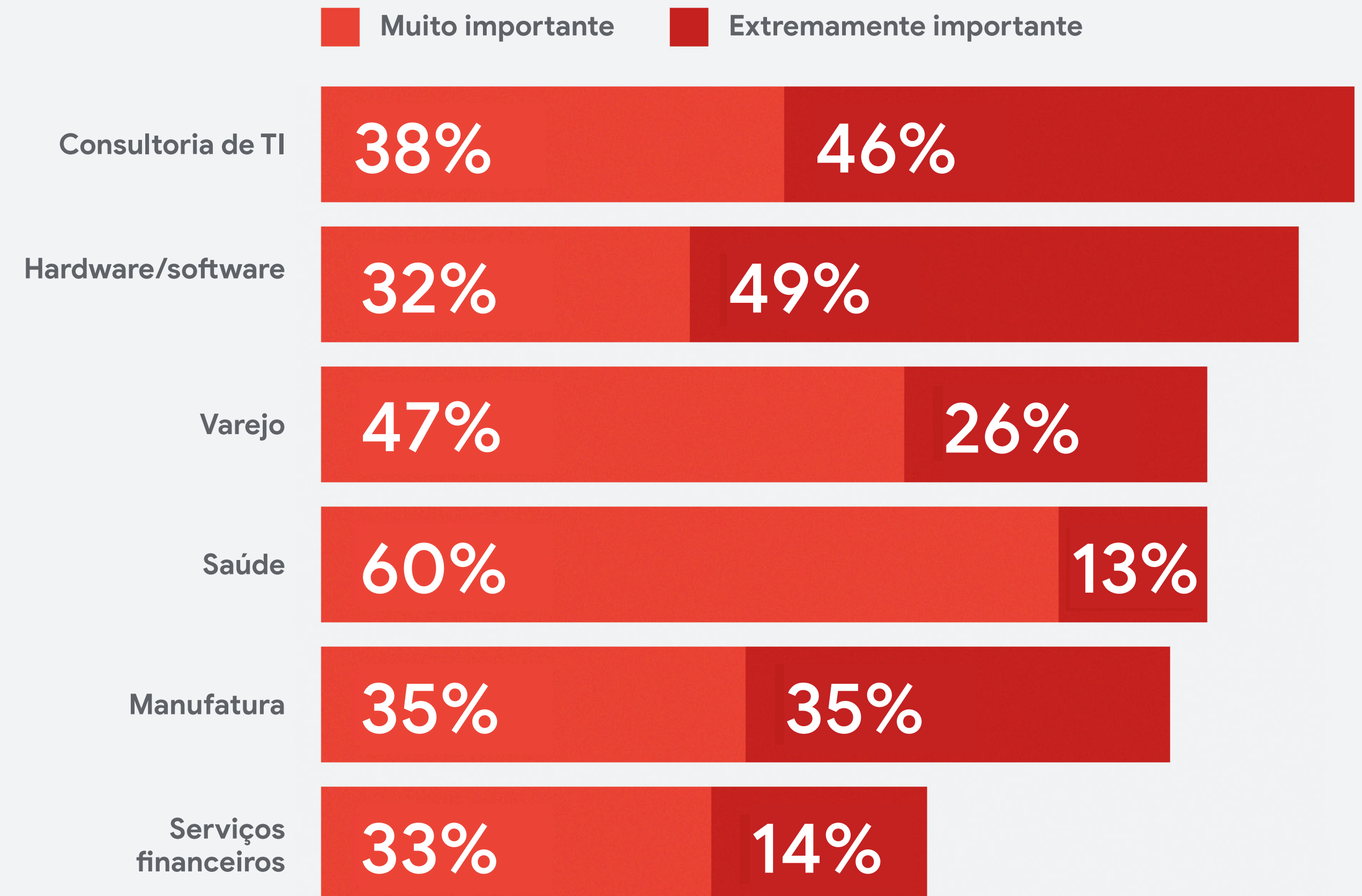


Uma análise de setores e regiões



Como seria de se esperar, a implantação de IA generativa na borda é importante para setores como manufatura (70%) e saúde (74%), mas ainda mais para empresas do setor de hardware ou software computacional (81%) e, especialmente, para empresas de consultoria em informática/TI (85%).

Qual a importância de implantar modelos de IA generativa na borda (por exemplo, em dispositivos IoT ou dispositivos móveis) para sua organização? (por setor)



Por Isso é importante para você?

Apesar da crescente importância, modernizar e aproveitar a IA na borda, manter o controle sobre dados essenciais e gerenciar a complexidade de múltiplas implantações na borda apresentam desafios significativos.

Não é surpresa, então, que haja uma demanda crescente por soluções baseadas na nuvem que permitam, em tempo real, a implantação de modelos de IA generativa na borda para aplicativos.

À medida que as expectativas dos clientes por interações em tempo real continuam a aumentar, também aumentará a necessidade de computação de borda e da infraestrutura que torna o processamento rápido da IA generativa uma realidade confiável.

Desafios com múltiplas implantações de borda:



Restrições de recursos e conectividade

Dispositivos de borda têm poder de computação limitado e frequentemente enfrentam conectividade intermitente, dificultando a implantação e o gerenciamento eficaz de modelos complexos de IA. Isso requer otimização especializada e recursos off-line robustos.



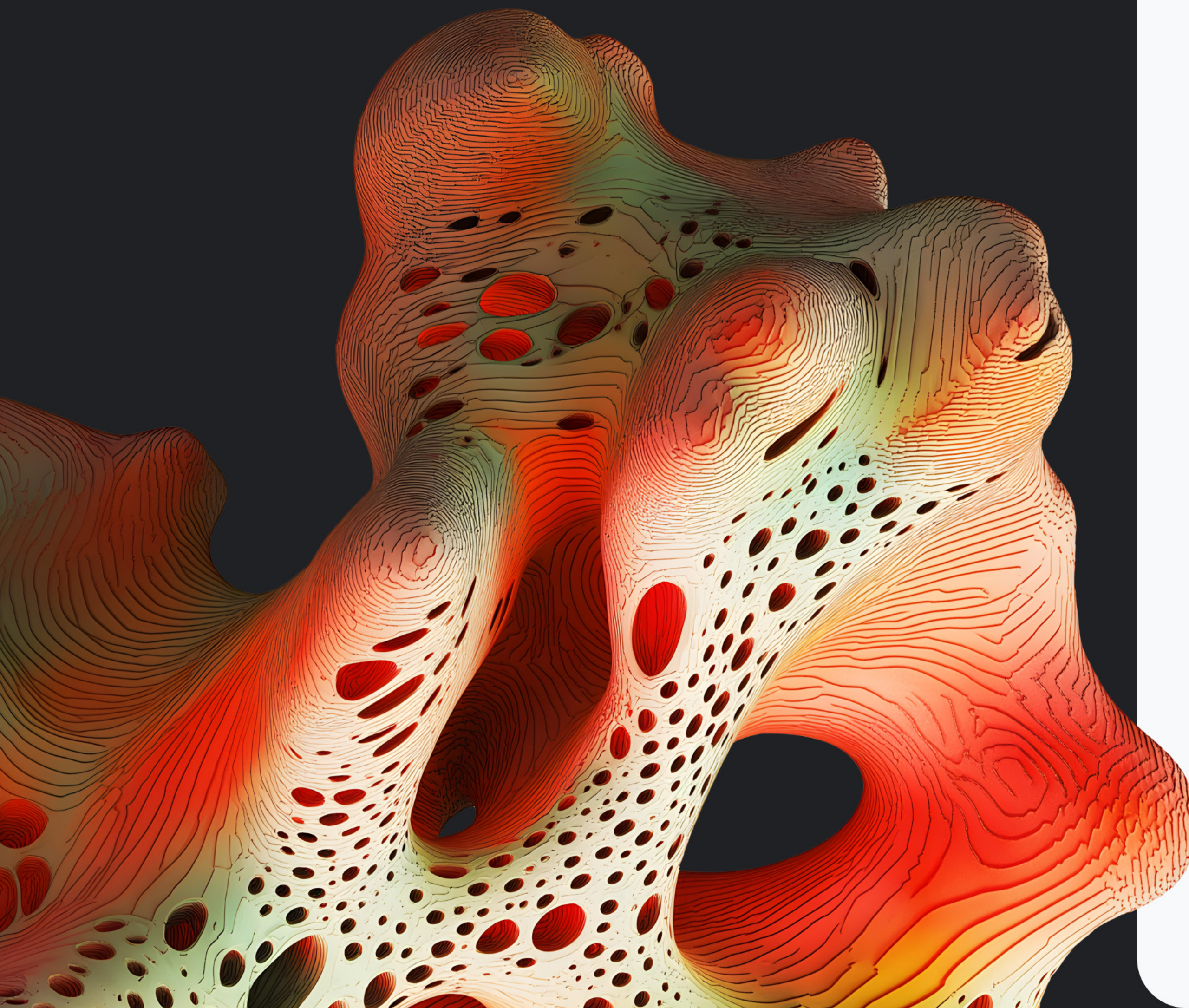
Segurança e governança de dados

Proteger dados sensíveis na borda, garantir a qualidade e a conformidade dos dados e manter o controle sobre a soberania de dados em locais distribuídos apresenta obstáculos significativos de segurança e governança.



Complexidade de implantação e gerenciamento

Escalar implantações de IA para vários dispositivos de borda geograficamente dispersos, com monitoramento, manutenção, controle de versão e orquestração, introduz uma complexidade significativa que exige automação e ferramentas de gerenciamento sofisticadas.



Varejistas globais como o McDonald's planejam implementar o Google Distributed Cloud para fornecer IA a milhares de restaurantes, habilitando novas plataformas para clientes e trazendo armazenamento de informações e computação de alta potência para restaurantes. Com esses recursos de computação de borda e IA distribuída, o McDonald's conseguirá não apenas gerar novos insights sobre o desempenho de equipamentos, mas também implementar soluções que reduzem interrupções nos negócios e simplificam as operações. Isso permitirá que as equipes do restaurante se concentrem em oferecer experiências excepcionais aos clientes.



Descoberta 07

A nuvem híbrida oferece flexibilidade para implantações da IA generativa

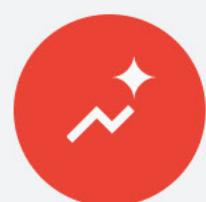
Uma abordagem híbrida envolve a combinação de infraestrutura local com recursos de nuvem pública.

01 02 03 04 05 06 07 08

Nossa pesquisa indica claramente uma forte preferência por uma abordagem de nuvem híbrida (74%).

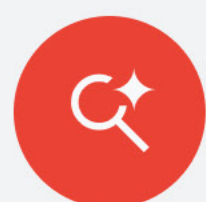
Essa abordagem é preferível às abordagens locais e de nuvem pública única.

A lógica é clara:



Maior flexibilidade

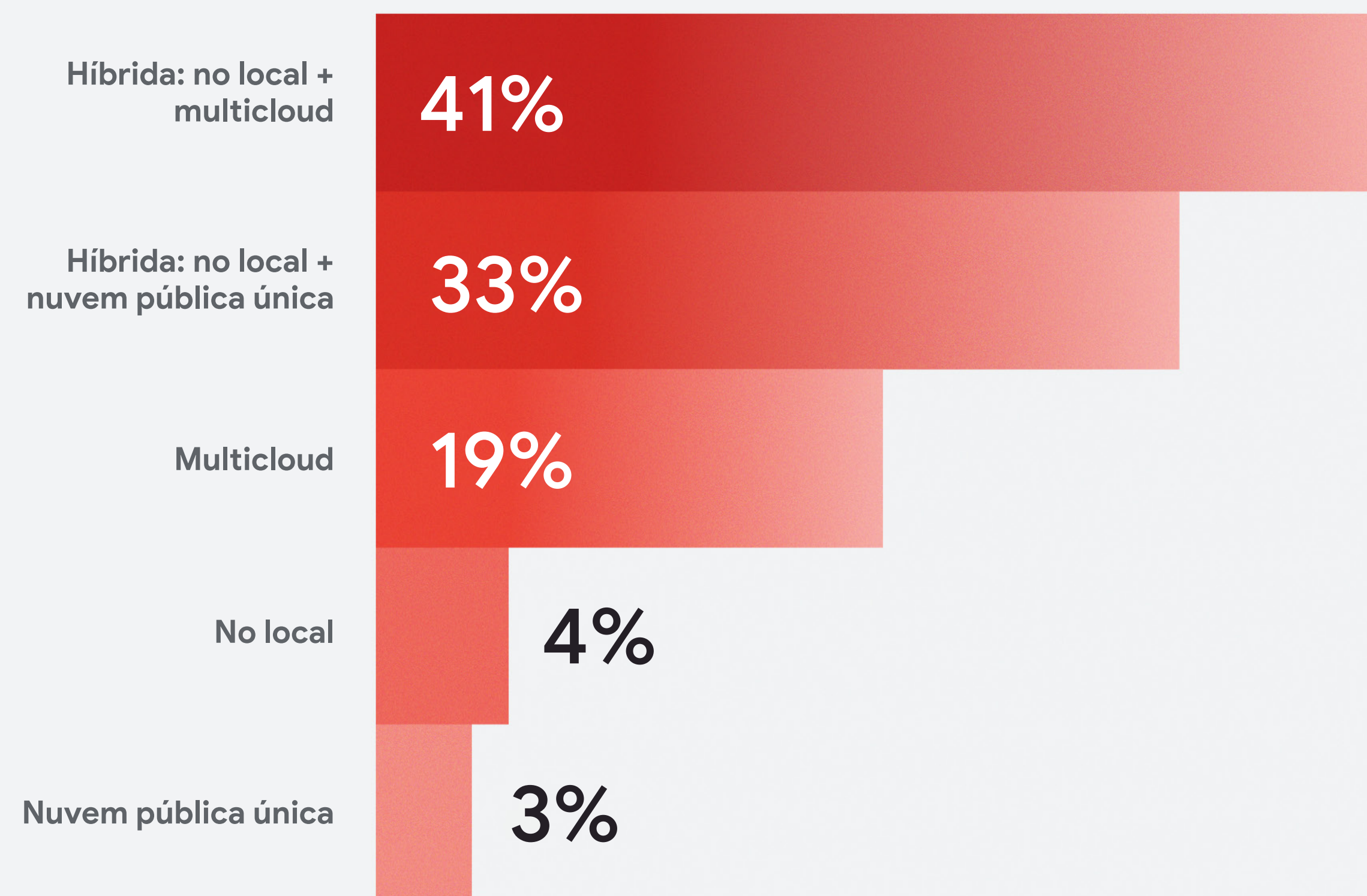
Uma abordagem híbrida permite que as organizações escolham o ambiente ideal para cada carga de trabalho, equilibrando desempenho, custo e ponderações de segurança.



Residência e conformidade de dados aprimoradas

Para certas cargas de trabalho, requisitos regulatórios podem exigir a manutenção de dados no local. Uma abordagem híbrida permite isso e ainda aproveita os benefícios da nuvem pública.

Qual abordagem de infraestrutura em nuvem sua organização usa, primordialmente, para cargas de trabalho de IA generativa?



Por Isso é importante para você?

A preferência esmagadora pela nuvem híbrida significa que um líder de tecnologia precisa entender como arquitetar, implementar e gerenciar esse ambiente. Se ignorarem essa tendência, podem acabar correndo o risco de ficar para trás dos concorrentes que já estão aproveitando a flexibilidade e o controle que o ambiente híbrido oferece. Eles precisam avaliar as parceiras de nuvem e a infraestrutura atual para saber se implementar uma abordagem híbrida é possível.

A flexibilidade na escolha do ambiente certo para cada carga de trabalho é fundamental. Isso requer pensar estrategicamente sobre o posicionamento da carga de trabalho, considerando fatores como desempenho, custo e requisitos de segurança por aplicativo.



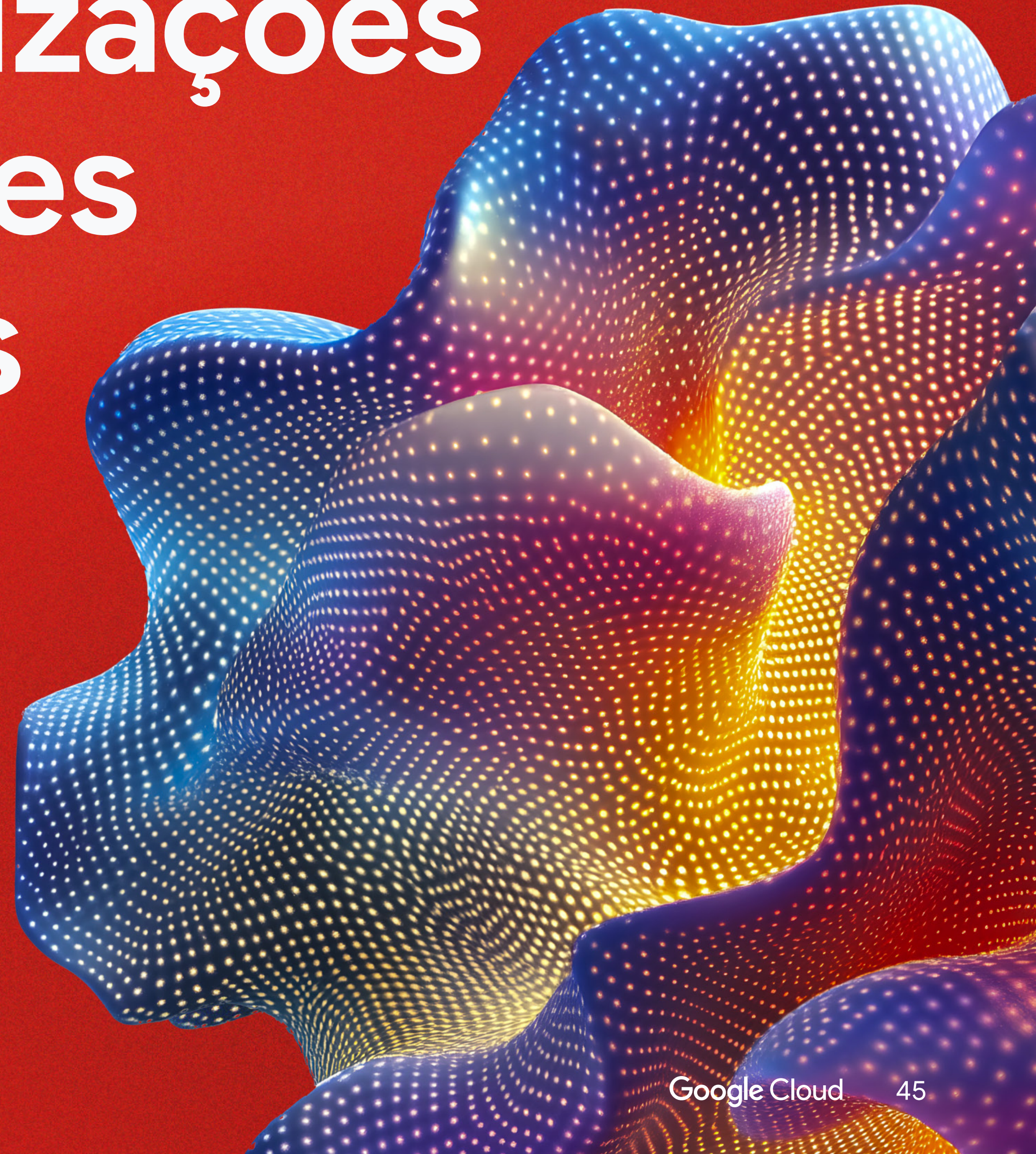


A Toyota utilizou o Hipercomputador de IA do Google Cloud para construir uma arquitetura de nuvem híbrida a fim de impulsionar sua inovadora plataforma de IA, que capacita os trabalhadores da fábrica a desenvolver e implantar modelos de IA em casos de uso importantes. A Toyota conseguiu uma redução de 20% no tempo de criação do modelo de aprendizagem, melhorando a experiência do usuário e aumentando a adoção. Com mais funcionários de chão de fábrica capazes de criar modelos de IA, a Toyota automatizou muitas tarefas manuais e que exigem mão de obra, poupando mais de 10.000 horas por ano por meio da eficiência de fabricação e otimização de processos.



Descoberta 08

A maioria das organizações depende das soluções de IA generativa dos provedores de nuvem



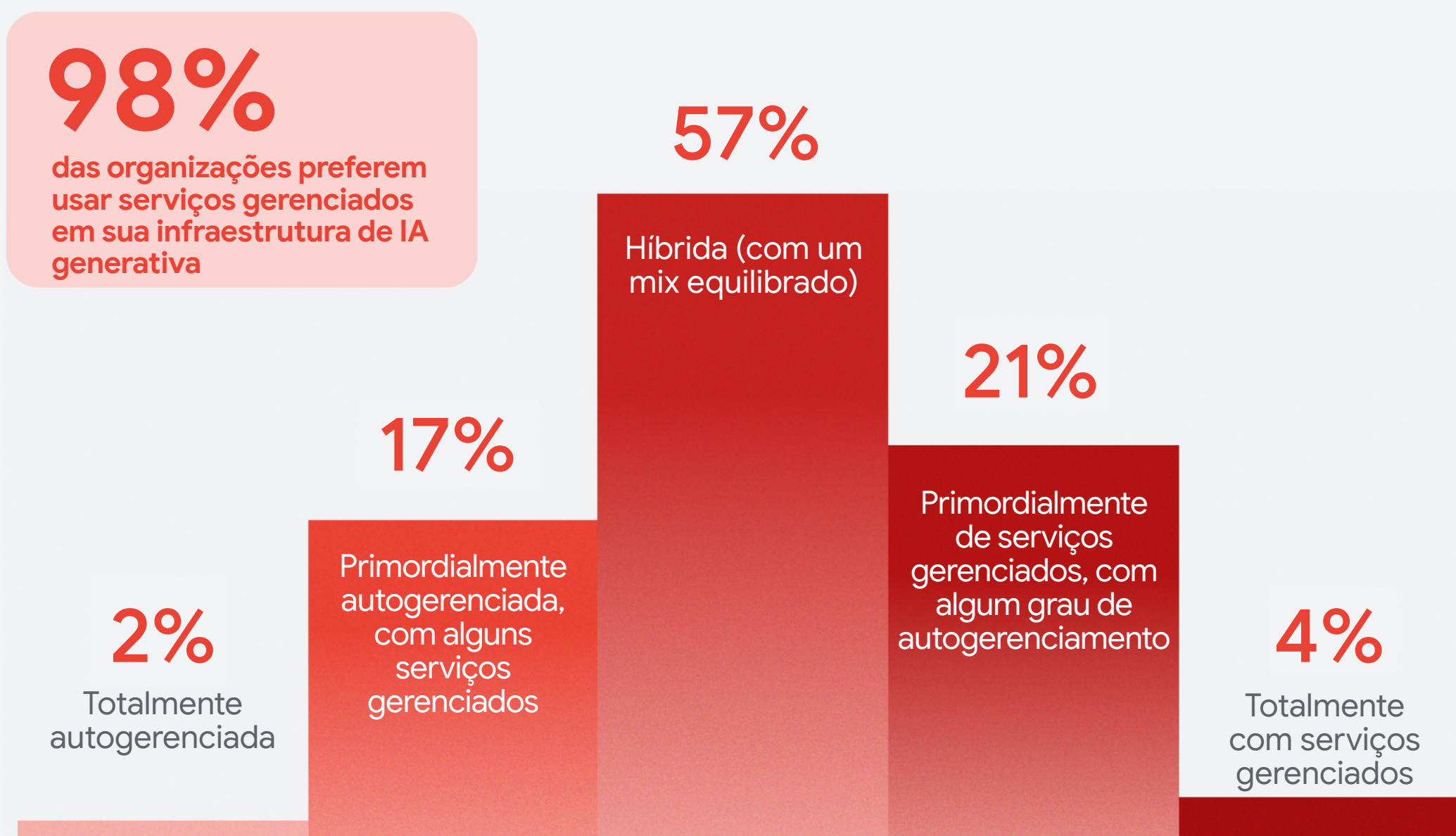
01 02 03 04 05 06 07 08

Criar e executar cargas de trabalho de IA é algo complexo.

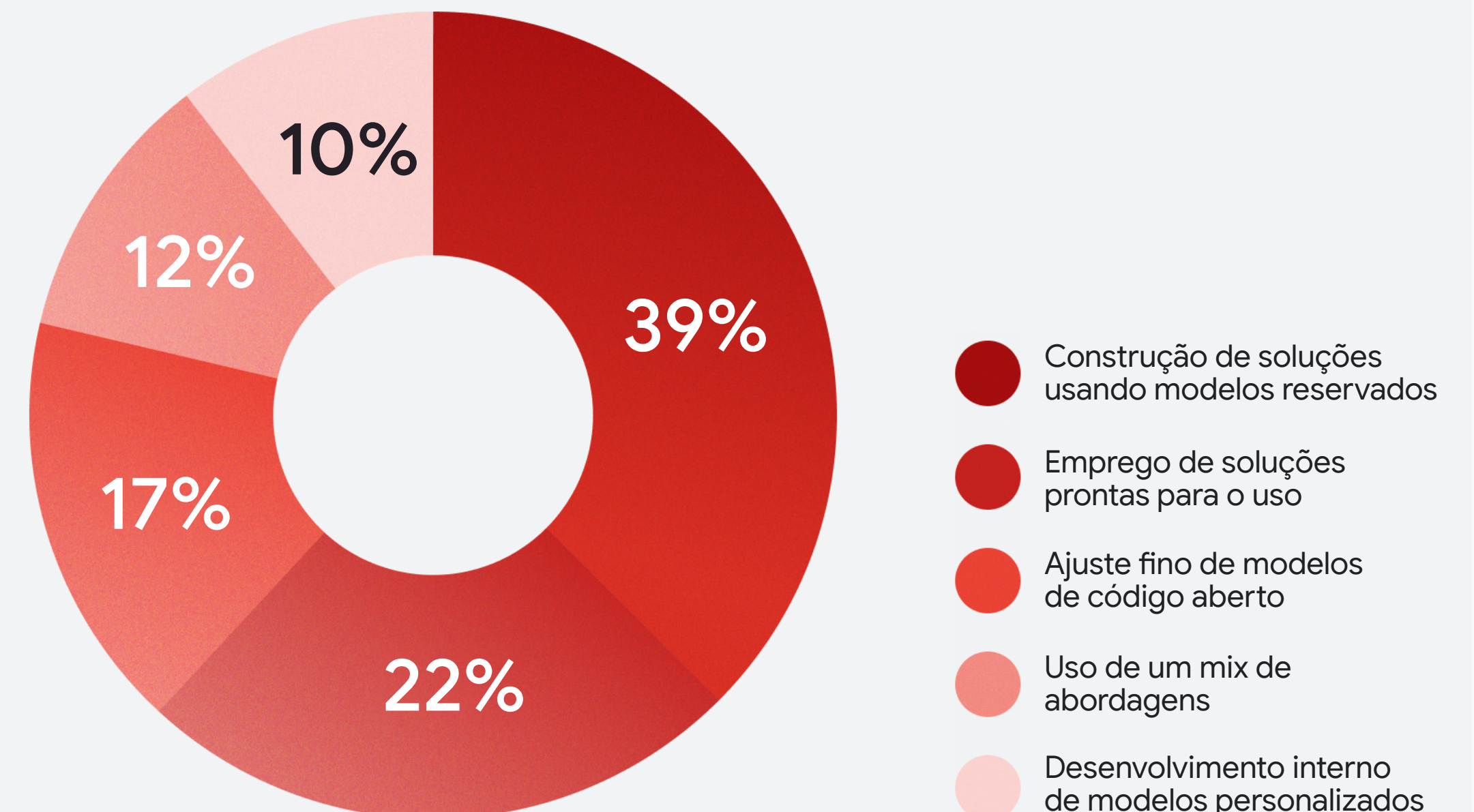
As organizações precisam de serviços gerenciados para otimizar as operações e aproveitar totalmente a nuvem.

Ao criar soluções de IA generativa, a abordagem mais comum é aproveitar modelos reservados como Gemini, Claude e AI21, priorizando velocidade, facilidade de uso, custo-benefício e escalabilidade. Criar modelos internos personalizados é a abordagem menos comum, porque exige consideravelmente mais tempo e recursos. Isso indica que, para muitas organizações, modelos reservados são mais do que eficientes para suas necessidades.

Qual abordagem sua organização prefere para gerenciar a infraestrutura de IA generativa?



Qual das opções a seguir descreve melhor a abordagem atual da sua organização para utilizar a IA generativa?



Os provedores de nuvem desempenham um papel fundamental na implementação da IA generativa.

Projetar para o sucesso em longo prazo requer um equilíbrio delicado: utilizar a nova funcionalidade de IA generativa para produtividade, otimizar os principais aplicativos para os clientes e modernizar a infraestrutura atual para acompanhar o ritmo. Estamos nos aproximando rapidamente de um ponto crítico de ruptura em que a infraestrutura e a computação em nuvem tradicionais não serão mais capazes de atender às demandas atuais de capacidade e desempenho.

Dada a complexidade e a intensidade de recursos para criar e gerenciar uma infraestrutura pronta para a IA, as organizações estão cada vez mais recorrendo a provedores de serviços de nuvem (48%) e, em menor escala, a fornecedores independentes de software (ISVs) (36%). Essa mudança envolve, em grande parte, a formação de parcerias estratégicas para alavancar conhecimentos e recursos especializados.

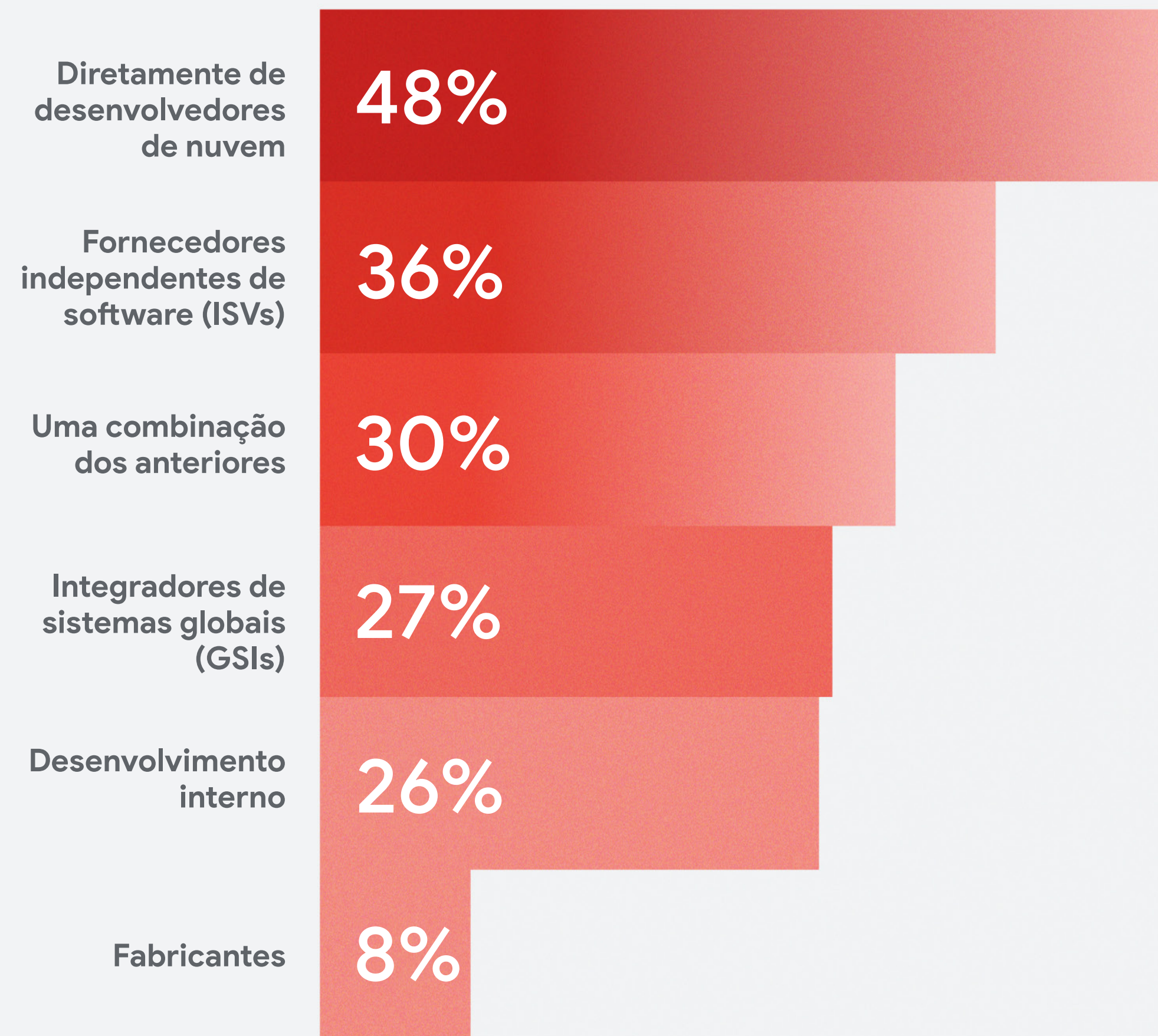
Isso está gerando um aumento considerável no tamanho do mercado de infraestrutura corporativa. Conforme descrito no relatório Tendências de IA para os negócios do Google Cloud de 2025, a adoção na infraestrutura corporativa deve aumentar em mais de 30% até 2026.¹ Essa rápida taxa de adoção criou uma demanda global por espaço em data centers capaz de lidar com o alto poder computacional e a alta densidade de energia necessários para cargas de trabalho de IA. A demanda por data centers prontos para a IA deve crescer a uma taxa média de 33% ao ano até 2030,² e os gastos com data centers devem dobrar nos próximos cinco anos.³

¹ Credence Research (2024). Tamanho, participação e previsão do mercado de infraestrutura multifuncional para 2032

² McKinsey & Company (2024). O poder da IA: expandindo a capacidade do data center para atender à crescente demanda.

³ National Bureau of Economic Research (2024). A rápida adoção da IA generativa.

Quais os principais canais que sua organização usa para adquirir e implementar soluções de IA generativa?



Por Isso é importante para você?

Construir uma infraestrutura de IA sólida é fundamental para o sucesso. Isso significa que, daqui para frente, o sucesso de uma organização vai depender da existência de uma infraestrutura capaz de fornecer escala, confiabilidade, desempenho computacional, eficiência, segurança, facilidade de uso e custo-benefício para permitir a combinação de cargas de trabalho de IA em larga escala, estados de aplicativos corporativos e desenvolvimento de novos aplicativos.

Muitas empresas e governos não têm as habilidades, o tempo ou os recursos necessários para desenvolver essas tecnologias e ferramentas de modo independente.

Não é surpresa, então, que eles estejam confiando em provedores de serviços gerenciados para controlar, dimensionar e proteger sua infraestrutura. Encontrar a melhor solução de IA para sua empresa ajuda você a se manter à frente da concorrência.

O ecossistema global de parceiros do Google Cloud especialmente treinado e certificado para fornecer soluções de nuvem que atendem às necessidades específicas do setor.





Os provedores de nuvem são parceiros estratégicos na era da IA

Essa mudança em direção aos provedores de nuvem não é apenas uma tendência. É um imperativo estratégico para líderes que estão lidando com as complexidades do desenvolvimento, implantação e gerenciamento de IA, ao mesmo tempo em que enfrentam restrições orçamentárias e pressões competitivas. A pesquisa confirma que a IA generativa é uma realidade imediata para muitas organizações. Cada vez mais, as empresas buscam soluções prontas para o uso e que sejam capazes de incorporar

tecnologias de IA e dados para simplificar a tomada de decisões ao operar, dimensionar, otimizar e proteger a infraestrutura.

À medida que os casos de uso de IA generativa se proliferam, as organizações precisarão de uma ampla gama de modelos para satisfazer diferentes necessidades. Pode ser complicado adotar os melhores modelos para seu caso de uso sem se afogar na complexidade.

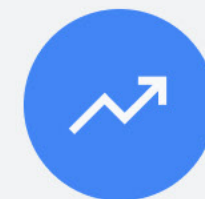


As grandes organizações estão aproveitando a experiência de provedores de nuvem para desenvolver uma solução abrangente. Por isso mesmo, estão abordando considerações importantes como:



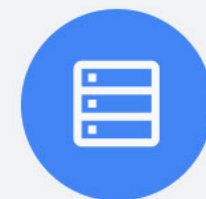
Segurança

Implementando medidas de segurança robustas para proteger dados sensíveis e garantir a conformidade



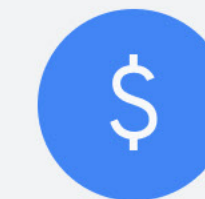
Escalonabilidade

Projetando uma infraestrutura que possa ser dimensionada para atender às crescentes demandas das cargas de trabalho de IA



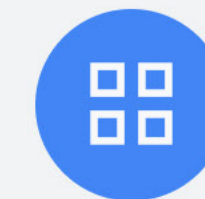
Qualidade dos dados

Garantindo acesso a dados de alta qualidade para treinar e implantar modelos eficazes de IA



Custo

Limitando as despesas associadas aos recursos computacionais significativos e à manutenção que a IA requer



Flexibilidade

Escolhendo o ambiente certo para cada tipo de carga de trabalho, por aplicativo

Ao selecionar cuidadosamente os parceiros, ferramentas e estruturas certas, as organizações podem se posicionar para o sucesso na era da IA.

O guia [Reprojetando sua infraestrutura para a IA generativa](#) ajuda líderes técnicos a projetar sistemas de IA generativa robustos, escalonáveis e econômicos.



O Google Cloud oferece uma plataforma abrangente para o sucesso da IA generativa

O imperativo para os líderes de tecnologia é claro: construir uma infraestrutura pronta para IA que possa atender às necessidades de negócio e, ao mesmo tempo, respaldar as demandas dessa tecnologia transformadora.



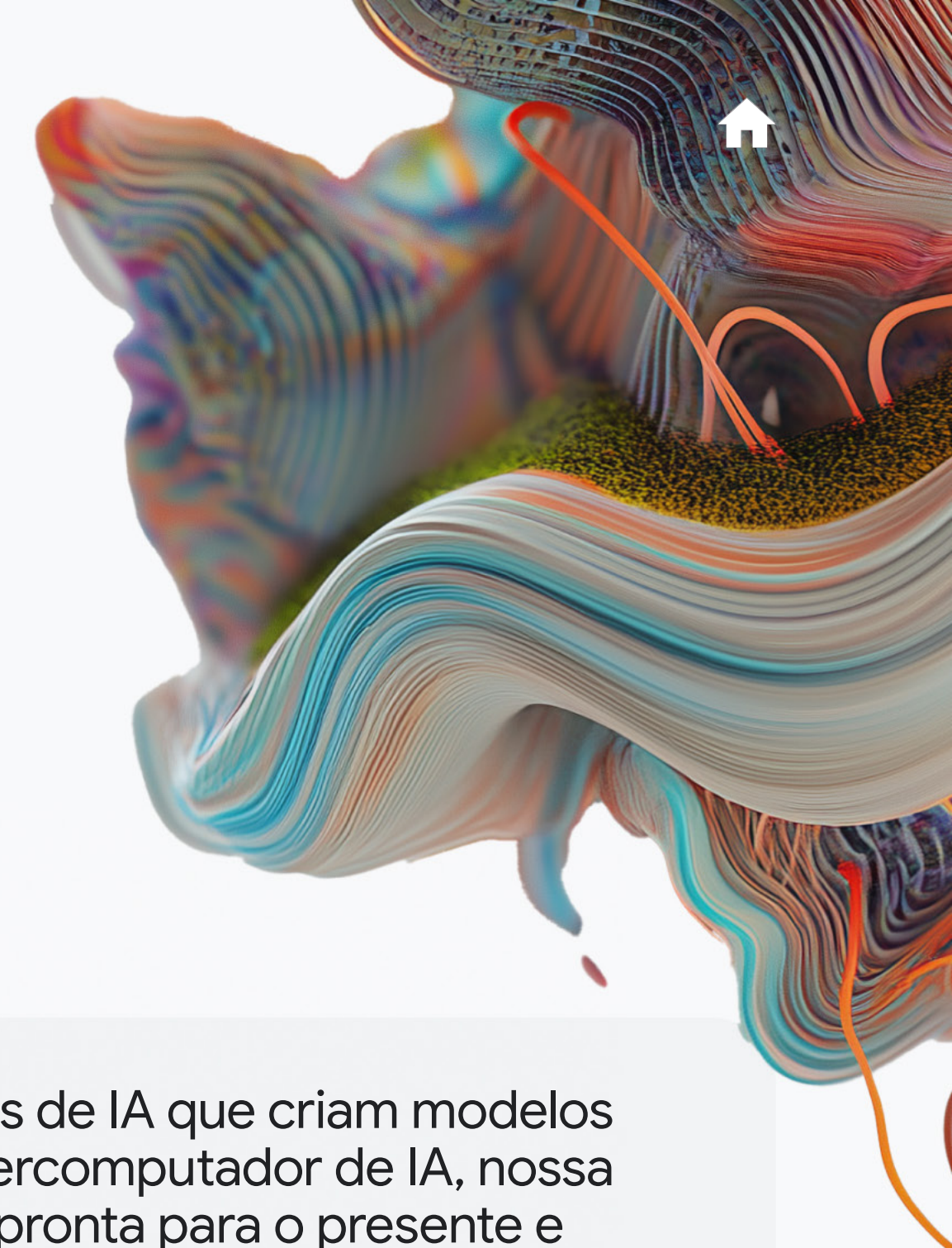
O **Hipercomputador de IA** é uma arquitetura de supercomputação totalmente integrada, projetada explicitamente para cargas de trabalho de IA.

Nossa pesquisa mostra que o sucesso não depende de soluções pontuais, mas da implantação de uma plataforma de IA robusta e integrada.

O foco deve ser integrar a IA aos negócios — e o Google Cloud pode fazer o trabalho pesado com uma infraestrutura de IA que oferece o desempenho, a escala e a eficiência que já impulsionam os produtos de IA mais avançados do mundo. Estamos sintonizados com o mercado, já executando modelos de raciocínio de ponta, oferecendo IA na borda e ajudando os clientes a implantar IA com segurança nos seus próprios data centers. Desde cientistas de dados corporativos que usam o Vertex AI até desenvolvedores que incorporam a IA com endpoints do CloudRun e do Gemini,

bem como empresas de IA que criam modelos de fundação no Hipercomputador de IA, nossa pilha completa está pronta para o presente e o futuro em rápida evolução. O Google Cloud não é apenas um provedor de nuvem, mas um parceiro estratégico para a transformação da IA. Nosso compromisso com os princípios de código aberto, nossa inovação contínua impulsionada pelo Google DeepMind e nossa oferta de tecnologias de IA a bilhões de usuários diariamente, somados ao foco em segurança e governança de nível corporativo, nos posicionam como a escolha preferida para organizações que buscam alcançar a liderança na era da IA generativa.

Vamos descobrir como o Google Cloud pode ajudar você a garantir seu futuro e liderar o caminho na jornada da IA.



**Crie hoje mesmo
uma infraestrutura
pronta para IA.**

Entre em contato



Metodologia

Este relatório é baseado em pesquisas quantitativas conduzidas em agosto e setembro de 2024.

Os entrevistados estão em suas organizações há pelo menos um ano e têm autoridade (ou influência significativa) sobre os processos de tomada de decisões de compra ou uso de soluções tecnológicas/seleção de fornecedores de tecnologia. As empresas representadas nesta pesquisa já têm uma carga de trabalho que usa a IA ou querem aproveitá-la nos próximos 12 meses.

Além disso, os líderes de tecnologia

deste estudo estão familiarizados com diversas aplicações de IA generativa e testaram ou implementaram algumas em suas respectivas empresas. Eles acreditam que a IA generativa é pelo menos um pouco importante para as operações comerciais atuais e futuras de sua organização e têm pelo menos alguma influência nas decisões relacionadas à adoção de tecnologia de IA e aprendizado de máquina (ML) nas organizações.

Os entrevistados não sabiam que o Google era o patrocinador da pesquisa e a identidade dos participantes não foi revelada ao Google.

Líderes de tecnologia das empresas em:

América do Norte

Estados Unidos e Canadá

Europa

Reino Unido
França
Alemanha

Ásia-Pacífico

Singapura
Austrália
Japão

Empresas da América do Norte e da Europa têm pelo menos

1.000 funcionários

As empresas da Ásia-Pacífico têm pelo menos

500 funcionários

513

líderes de tecnologia foram entrevistados para nossa pesquisa nos Estados Unidos, Canadá, Reino Unido, França, Alemanha, Singapura, Austrália e Japão, sendo que os EUA representaram mais de um quarto da amostra (29%).

Região		Porte da empresa		Nível do cargo	
América do Norte	40%	500-999	5%	Diretor executivo	19%
Europa	30%	1.000 a 4.999	54%	Vice-Presidente/Gerente Geral	14%
Ásia-Pacífico	30%	5.000 ou mais	41%	Diretor sênior/Diretor	33%
Nível de gerência					35%

Tempo no cargo		Idade		Familiaridade com IA	
De 1 a 2 anos	9%	18 a 34	32%	Intermediário	21%
3 a 5 anos	41%	35 a 44	42%	Avançado	55%
6 a 9 anos	38%	45 a 54	21%	Especialista	24%
10 anos ou mais	12%	55-64	6%		