



Custo da IA precisa incluir falhas e custo do tempo humano em revisão

Adilson Damasio (*)

As empresas costumam acompanhar a conta da inteligência artificial generativa pela unidade que o fornecedor usa para cobrar, o volume de tokens de entrada e de saída de cada chamada ao modelo, e esse número orienta a comparação entre modelos, a negociação de contratos e a projeção de orçamento.

A medida explica a fatura, mas diz pouco sobre o que a operação recebeu em troca, porque parte das chamadas gera respostas que alguém descarta, refaz ou corrige antes de usar. Com a passagem dos projetos-piloto para operações que atendem clientes e equipes internas em volume, a diferença entre o que a empresa paga e o que aproveita passou a pesar no orçamento de tecnologia.

O custo que interessa ao negócio é o de cada resultado aproveitável: a resposta, o documento, a classificação ou a ação que chega ao usuário e cumpre a função sem retrabalho. Para chegar a esse número, é preciso somar ao gasto com tokens o custo das tentativas descartadas, das tarefas que terminaram sem resultado, das ferramentas e bases de dados que o sistema consulta a cada etapa e das horas de revisão humana, e dividir o total pelo volume de resultados que a área usuária aceitou. A gestão financeira de IA depende, por isso, de instrumentação técnica e de uma definição de negócio sobre o que conta como resultado em cada caso de uso.

A distância entre medir consumo e medir resultado aparece na 6ª edição do State of FinOps, que a FinOps Foundation, divulgou em 2026 com respostas de 1.192 profissionais, cujas organizações somam mais de US\$ 83 bilhões em gasto anual com nuvem, com 15% das respostas vindas das Américas do Sul e Central.

Segundo o relatório, 98% dos respondentes incluem os gastos com IA entre as áreas que a equipe de FinOps gerencia hoje ou vai gerenciar no próximo ano, ante 63% na edição anterior e 31% dois anos antes. O estudo aponta a gestão de custos de IA como a principal competência que as equipes precisam desenvolver e reúne, entre as dificuldades mais citadas ao aplicar FinOps aos gastos com IA, a visibilidade sobre esses custos, a divisão do gasto entre as unidades de negócio e a medição do retorno.

Entre os recursos que os profissionais gostariam de encontrar nas ferramentas e que ainda não existem, o mais pedido é o monitoramento detalhado do gasto com IA por tokens, por requisições a modelos de linguagem e por uso de GPU.

Custo por resultado aproveitável começa na definição de cada caso de uso

A área que usa a entrega precisa definir, com apoio da equipe de tecnologia, o que conta como resultado aproveitável em cada processo. No service desk, o critério pode ser o chamado encerrado, e na análise de documentos, a extração de dados que o revisor aprova sem correção. Em assistentes de programação, vale a sugestão de código que passa pela revisão e chega à produção, enquanto no atendimento ao cliente a referência precisa ser a conversa que resolve a solicitação sem transferência para um atendente e sem novo contato sobre o mesmo assunto.

Entre o resultado aceito e o descartado ficam as respostas que o usuário aproveita depois de editar, e esse grupo precisa de registro próprio. As respostas editadas entram na conta com o tempo de edição, porque a economia que a empresa espera da IA depende de quanto trabalho humano a resposta ainda exige. A taxa de aceitação, que a equipe calcula sobre o total de respostas geradas, mostra que parcela da produção virou entrega, e a mesma conta, ponderada pelos tokens de cada resposta, mostra quanto do consumo ficou em tentativas que ninguém usou.

Depois que a área registra essa definição, a empresa consegue comparar o custo por resultado com o custo do mesmo processo sem IA, medido pelas horas que a equipe dedicava a cada entrega e pelo valor dessas horas. A comparação orienta a escolha dos casos de uso que merecem ampliação e evita que a decisão dependa de demonstrações que não reproduzem as condições da operação diária, com volume, exceções e dados incompletos.

Custo de cada resposta inclui tentativas, contexto e chamadas a ferramentas

O consumo de tokens de uma mesma tarefa varia mais do que a fatura sugere, porque o modelo recebe a cada chamada, além da pergunta, as instruções do sistema, o histórico da conversa e os trechos de documentos que a aplicação recupera da base da empresa.

Em fluxos com agentes, cada etapa em geral reenvia ao modelo o contexto acumulado, e o consumo de entrada cresce mais depressa que o número de passos, a menos que a aplicação resuma o histórico ou reaproveite o contexto em cache. Por isso, o custo médio por chamada esconde a distribuição do gasto, na qual uma parcela pequena de tarefas longas pode concentrar boa parte da conta.

A nova chamada que a aplicação dispara depois de um erro ou de tempo esgotado, a resposta que a validação da própria aplicação rejeita por formato incorreto ou por violar regras de conteúdo e a resposta que o usuário pede de novo quando a primeira não atende geram cobrança como qualquer outra chamada, embora a tarefa termine com um único resultado. As falhas pesam da mesma forma quando um agente percorre várias etapas, consome tokens em cada uma e encerra a execução sem concluir a tarefa.

A revisão humana não aparece na fatura de tecnologia, mas esse custo também precisa ser contabilizado. Quando associa tokens, tentativas, ferramentas e revisão à mesma tarefa, a equipe de FinOps calcula o custo por resultado aproveitável de cada caso de uso e acompanha a evolução desse número.

Com essa avaliação, a empresa deixa de escolher o modelo pelo preço do token, já que um modelo mais barato pode sair mais caro por resultado quando exige mais tentativas, mais contexto e mais revisão. Antes de ampliar um caso de uso, passa a comparar o custo de cada resposta que chega pronta ao cliente ou à equipe interna com o custo do processo que a IA substitui, e essa conta oferece a medida de retorno que a fatura de tokens, sozinha, não mostra.

(*) Diretor de Operações na Faiston.

Quando o ecossistema de tecnologia cresce, quem assume o risco dos parceiros?

Durante anos, a homologação de fornecedores foi encarada como uma etapa burocrática do processo de compras

Keila Janaina Israel (*)

Essa percepção começou a mudar à medida que o compliance ganhou relevância no ambiente corporativo. No Brasil, a Lei Anticorrupção e as investigações da Lava Jato ampliaram a discussão sobre due diligence de terceiros. Posteriormente, casos envolvendo violações trabalhistas e trabalho análogo à escravidão reforçaram uma preocupação: os riscos de uma empresa não terminam nos limites de sua própria operação.



Keila Janaina Israel

Agora, uma nova transformação torna essa discussão especialmente importante para o mercado de tecnologia. No passado, grandes empresas do setor ampliaram seus portfólios por meio da aquisição de outras companhias. Incorporar uma empresa significava agregar tecnologias, competências e conhecimento para atender diferentes necessidades dos clientes. Esse movimento continua existindo, mas a velocidade atual da inovação tornou cada vez mais difícil concentrar internamente todas as capacidades necessárias.

Inteligência artificial, cloud, cibersegurança, dados e automação avançam rapidamente. Paralelamente, startups e empresas especializadas desenvolvem soluções capazes de complementar grandes projetos com conhecimento específico, inovação e velocidade. Em vez de necessariamente adquirir essas competências, tornou-se estratégico estabelecer parcerias e construir ecossistemas.

Essa tendência deve se intensificar. A pesquisa global de gestão de riscos de terceiros da KPMG de 2026, realizada com 851 organizações, aponta que 83% dos executivos pretendem ampliar suas redes de parceiros nos próximos três anos. O estudo mostra ainda que compliance regulatório e riscos cibernéticos estão entre os principais fatores que orientam as estratégias de gestão de terceiros.

Existe, porém, um choque entre dois modelos de operação. Startups foram estruturadas para responder rapidamente ao mercado e, em geral, trabalham com processos mais enxutos e ciclos mais curtos de decisão, inclusive na formação de parcerias. Grandes organizações, por sua vez, carregam estruturas de governança, controles e níveis de aprovação mais complexos. Em um projeto que reúne empresas com culturas e velocidades diferentes, o desafio passa a ser conciliar a agilidade necessária para aproveitar uma oportunidade comer-

cial com os controles indispensáveis para saber quem está sendo colocado dentro da entrega.

É justamente nesse momento que surge uma questão: quem assume o risco quando um projeto depende de várias empresas para ser entregue?

Para o cliente, geralmente existe uma companhia responsável pela solução contratada. Se um dos parceiros apresentar problemas financeiros, jurídicos, fiscais, reputacionais, operacionais ou de segurança, as consequências podem atingir todo o projeto e a empresa que colocou seu nome à frente daquela entrega. Isso significa que a gestão de terceiros precisa começar antes do parceiro entrar efetivamente no projeto. Quanto mais estratégica for sua participação, maior deve ser a clareza sobre o que ele terá acesso, quais atividades executará, quais dados poderá tratar e qual impacto sua eventual indisponibilidade teria sobre a entrega. A homologação, nesse sentido, também ajuda a dimensionar o nível de exposição que cada parceiro representa.

O Global Cybersecurity Outlook 2026, do World Economic Forum, ajuda a dimensionar essa exposição. Entre as grandes empresas pesquisadas, 65% apontam vulnerabilidades relacionadas a terceiros e à cadeia de suprimentos como seu maior desafio para a resiliência cibernética, ante 54% em 2025. Embora 66% avaliem a maturidade de segurança de seus fornecedores, somente 33% afirmam mapear de forma abrangente seus ecossistemas para compreender as interdependências e a exposição a ameaças.

É nesse contexto que a homologação deixa de ser apenas uma prática de compras e compliance e passa a ser também um instrumento para construir ecossistemas tecnológicos mais seguros. E homologar é diferente de cadastrar. Enquanto o cadastro

reúne informações administrativas para formalizar um relacionamento, a homologação avalia se aquela empresa reúne condições para participar de uma entrega. Isso envolve regularidade jurídica e fiscal, saúde financeira, capacidade operacional, compliance, segurança da informação e outros fatores relacionados ao risco.

O desafio, portanto, não é escolher entre velocidade e controle, mas tornar a homologação compatível com a velocidade dos novos ecossistemas e com o nível de risco de cada parceria. Processos excessivamente longos podem fazer uma empresa perder uma oportunidade ou dificultar a entrada de parceiros inovadores. Por outro lado, acelerar uma parceria sem conhecer adequadamente o terceiro transfere para o projeto riscos que podem aparecer somente depois da contratação.

Essa questão é particularmente relevante para startups. Sua especialização, estruturas mais enxutas e velocidade podem ser fundamentais para incorporar inovação aos projetos. Homologá-las não deve significar submeter-las automaticamente à mesma burocracia de processos desenhados para grandes fornecedores, mas estabelecer critérios proporcionais ao risco, capazes de preservar a agilidade sem abrir mão da segurança.

Também não basta avaliar o parceiro apenas no início. Sua condição financeira, jurídica, fiscal ou operacional pode mudar. Por isso, a homologação precisa evoluir para um processo de monitoramento contínuo durante o relacionamento.

Em um mercado no qual nenhuma organização precisa dominar sozinha todas as tecnologias, construir ecossistemas tornou-se uma vantagem competitiva. Mas essa estratégia exige uma governança capaz de acompanhar a complexidade das relações. Quanto maior a dependência de parceiros, maior também a necessidade de conhecer, avaliar e monitorar quem participa da entrega.

O diferencial estará em conseguir fazer isso sem transformar governança em lentidão ou velocidade em exposição desnecessária. Afinal, quando uma empresa coloca seu nome à frente de um projeto, não está oferecendo apenas sua própria competência, mas também a responsabilidade pelos parceiros que escolheu para entregar junto com ela.

(*) Supervisora de Contratos de Fornecedores da SONDA do Brasil, líder regional em Transformação Digital.

Globalweb expande no Centro-Oeste e fatura R\$ 300 Mi

Uma das principais empresas de TI do Brasil, a Globalweb anuncia o fortalecimento de suas operações na região Centro-Oeste, com foco estratégico no Distrito Federal e no Estado de Goiás. O movimento integra o novo ciclo de expansão da companhia nos setores público e privado, impulsionado pela conquista de 11 novos contratos na região entre 2025 e 2026 que, juntos, representam mais de R\$ 300 milhões, nas áreas de fábrica de software e infraestrutura.

O crescimento no Centro-Oeste tem sido tracionado pela alta demanda por serviços de sustentação tecnológica, com destaque para o fornecimento de soluções de infraestrutura crítica. A empresa projeta como próximos passos a abertura de um escritório em Goiânia.

Para a CEO da Globalweb, Bruna Boner, a expansão regional é re-

flexo da maturidade da empresa em não apenas fornecer tecnologia, mas garantir a estabilidade das operações dos clientes. “A infraestrutura de TI não é uma commodity, ela é a fundação do negócio. O cliente não compra apenas software ou servidores, ele compra tranquilidade”, destaca.

