XXXIII Congresso Brasileiro de Custos
17, 18 e 19 de novembro de 2026
- Porto de Galinhas / PE -

AI-SaaS e a reintrodução do custeio variável: o encontro do modelo de software com a contabilidade tradicional de custos

Pedro Martins de Oliveira (UFSC) - pmartins.oliveira@gmail.com

Resumo:

O modelo de Software como Serviço (SaaS) consolidou um regime de custo em que o custo marginal de atender um cliente adicional se aproxima de zero, afastando o setor de software de instrumentos clássicos da contabilidade de custos, como a margem de contribuição por cliente e a curva ABC de rentabilidade. A integração de modelos de linguagem de grande escala (LLMs) aos produtos SaaS, fenômeno definido neste artigo como AI-SaaS, reintroduz custo variável real, mensurado em tokens processados a cada interação do cliente com a plataforma. Este ensaio teórico, apoiado em revisão de literatura e análise documental de nove empresas de capital aberto, organiza sua contribuição em três eixos: a mudança factual do custo de entrega, de quase-fixo para variável por interação; o desafio de identificar o direcionador de valor do cliente e traduzi-lo em um direcionador de custo operacional, base para o custeio e a precificação; e a inadequação da precificação por assento diante dessa nova estrutura de custo, com migração para referenciais de precificação por consumo. A evidência documental mostra que essa mudança se manifesta no reconhecimento, pela administração dessas empresas, de que o custo de receita cresce por investimentos em infraestrutura de IA, na criação de métricas próprias de consumo de IA e no lançamento de modelos de precificação por consumo. O artigo propõe uma tipologia de estruturas de custo de software (local, SaaS e AI-SaaS) e uma agenda de pesquisa sobre o direcionador de custo e a adequação dos modelos de precificação ao consumo.

Palavras-chave: *Inteligência Artificial. Custeio Variável. Precificação. Software. Contabilidade de Custos*

Área temática: *Tecnologia e gestão de custos*

AI-SaaS e a reintrodução do custeio variável: o encontro do modelo de software com a contabilidade tradicional de custos

RESUMO

O modelo de Software como Serviço (SaaS) consolidou um regime de custo em que o custo marginal de atender um cliente adicional se aproxima de zero, afastando o setor de software de instrumentos clássicos da contabilidade de custos, como a margem de contribuição por cliente e a curva ABC de rentabilidade. A integração de modelos de linguagem de grande escala (LLMs) aos produtos SaaS, fenômeno definido neste artigo como AI-SaaS, reintroduz custo variável real, mensurado em tokens processados a cada interação do cliente com a plataforma. Este ensaio teórico, apoiado em revisão de literatura e análise documental de nove empresas de capital aberto, organiza sua contribuição em três eixos: a mudança factual do custo de entrega, de quase-fixo para variável por interação; o desafio de identificar o direcionador de valor do cliente e traduzi-lo em um direcionador de custo operacional, base para o custeio e a precificação; e a inadequação da precificação por assento diante dessa nova estrutura de custo, com migração para referenciais de precificação por consumo. A evidência documental mostra que essa mudança se manifesta no reconhecimento, pela administração dessas empresas, de que o custo de receita cresce por investimentos em infraestrutura de IA, na criação de métricas próprias de consumo de IA e no lançamento de modelos de precificação por consumo. O artigo propõe uma tipologia de estruturas de custo de software (local, SaaS e AI-SaaS) e uma agenda de pesquisa sobre o direcionador de custo e a adequação dos modelos de precificação ao consumo.

Palavras-chave: Inteligência Artificial. Custeio Variável. Precificação. Software. Contabilidade de Custos.

Área Temática: Tecnologia e gestão de custos.

1 INTRODUÇÃO

A contabilidade de custos sempre reconheceu que o custo variável acompanha o volume: quanto mais se produz ou entrega, mais se incorre em custo direto. Esse princípio, consolidado desde as formulações clássicas de custeio por absorção e custeio variável (Horngren, Datar, & Rajan, 2021; Martins, 2018), estrutura a análise de margem de contribuição, a curva ABC (Custeio Baseado em Atividades, do inglês Activity-Based Costing) de rentabilidade e o custeio por atividade. Trata-se de um dos fundamentos da disciplina.

O advento do modelo Software como Serviço (Software as a Service, SaaS), contudo, consolidou um regime de custo que afastou, na prática, o setor de software dessa lógica. Ao centralizar a entrega em plataformas na nuvem, o SaaS reduziu o custo marginal de atender cada usuário adicional a um valor próximo de zero: uma vez desenvolvida a plataforma, o custo de servir o milésimo cliente é essencialmente igual ao de servir o primeiro (Shapiro & Varian, 1999). Essa característica afastou o setor das análises clássicas de custeio, como a margem por cliente e a curva ABC de rentabilidade, em favor de métricas baseadas no número de clientes, como Custo

de Aquisição de Clientes (CAC), Valor do Cliente no Tempo (Lifetime Value, LTV) e Receita Recorrente Mensal (MRR), e sustentou um modelo de precificação por assento (seat-based pricing), no qual o preço é fixo por usuário e independe da intensidade de uso (Ojala, 2013).

A integração de modelos de linguagem de grande escala (LLMs) aos produtos SaaS, fenômeno que define o segmento AI-SaaS, reintroduz a variabilidade de custo nesse setor. Cada interação do usuário com um componente de inteligência artificial exige processamento computacional adicional, cobrado por token processado, a um custo diretamente proporcional ao volume e à complexidade do uso. O custo de servir deixa de ser integralmente quase-fixo e passa a incorporar um componente variável, não a totalidade do custo, mas parcela suficiente para alterar a lógica de custeio herdada do SaaS tradicional, atrelado não ao cliente, mas ao volume de interação do cliente com a plataforma.

Este artigo organiza sua contribuição em torno de três eixos complementares. O primeiro é a mudança factual na natureza do custo de entrega: no AI-SaaS, o custo passa a incorporar um componente variável, proporcional ao volume de tokens processados a cada interação; não é a totalidade do custo de entrega que se torna variável, mas essa parcela já é suficiente para alterar a lógica de custeio herdada do SaaS tradicional. O segundo é o desafio de identificar o direcionador de valor percebido pelo cliente, isto é, aquilo que a IA efetivamente entrega (uma tarefa executada, um relatório gerado, uma resposta obtida), e de traduzi-lo em um direcionador de custo operacional capaz de servir de base tanto ao custeio quanto à precificação. O terceiro é a inadequação da precificação por assento diante de uma estrutura de custo que varia por interação, e a consequente necessidade de um referencial de precificação baseado em consumo.

Esse retorno à lógica de custeio variável reabilita instrumentos que o setor desconsiderava, como a margem por cliente e a curva ABC de rentabilidade: um cliente que utiliza intensivamente a plataforma pode consumir dezenas de vezes mais tokens que um cliente de uso moderado, tornando seu custo real incompatível com o preço único por usuário que o modelo SaaS tradicional havia consagrado. A análise documental conduzida neste artigo, reunindo nove empresas de capital aberto detalhadas na Seção 3.1, mostra que essa mudança de natureza do custo é identificável no reconhecimento, pela própria administração dessas empresas, em suas divulgações regulatórias, a Discussão e Análise da Administração (Management Discussion and Analysis, MD&A), e comunicações públicas, de que o crescimento do custo de receita decorre de investimentos em infraestrutura de IA, na criação de métricas operacionais específicas de inferência (como volume de tokens e unidades proprietárias de consumo de IA) e no lançamento de modelos de precificação por consumo.

O custo de inferência corresponde ao custo computacional incorrido para executar um modelo de IA já treinado e gerar uma resposta a uma solicitação do usuário; é tipicamente calculado a partir do número de tokens de entrada e de saída processados em cada requisição, multiplicados pelo preço unitário cobrado por cada tipo de token (Sardana, Portes, Doubov, & Frankle, 2024). O desafio central, porém, não é apenas o retorno do custo variável que esse tipo de custo introduz. O custo de inferência é mensurado em tokens, unidade operacional que o cliente não conhece e com a qual não se relaciona. Na prática, o cliente adquire uma pesquisa, um relatório gerado, uma tarefa concluída, uma resposta: esse é o direcionador de valor que orienta sua percepção de preço justo. A empresa precisa identificar esse direcionador de valor e construir, a partir dele, um direcionador de custo operacional

que seja simultaneamente rastreável ao custo real em tokens e inteligível como base de precificação. Sem essa definição, não há mapeamento de custos diretos confiável, nem uma base coerente para migrar da precificação por assento para a precificação por consumo; é precisamente nessa interface entre valor percebido, direcionador de custo e precificação que reside a lacuna que este artigo se propõe a mapear.

Diante desse cenário, emergem as seguintes questões de pesquisa: considerando que o modelo AI-SaaS representa um retorno à lógica de custeio variável clássica no setor de software, (i) quais instrumentos da contabilidade de custos esse retorno reabilita; (ii) como identificar o direcionador de valor do cliente e traduzi-lo em um direcionador de custo que sirva de base ao custeio e à precificação; e (iii) por que a precificação por assento deixa de ser adequada e que referenciais de precificação por consumo a substituem?

Para respondê-las, o artigo tem como objetivo geral analisar, por meio de revisão de literatura e análise documental, como o modelo AI-SaaS reintroduz a lógica de custeio variável no setor de software, a partir dos três eixos descritos. A relevância acadêmica decorre da ausência de literatura revisada por pares que trate esses três eixos como um problema conjunto, lacuna que reflete, em boa medida, a novidade do próprio fenômeno: a criação de direcionadores de custo proprietários e o lançamento de modelos de precificação por consumo em produtos de IA são práticas de mercado concentradas nos últimos dois a três anos, e a literatura contábil ainda não teve tempo de processá-las sistematicamente. A relevância prática decorre das implicações diretas para gestores financeiros e controllers: parte deles precisa voltar a rastrear custo por interação, prática já conhecida da contabilidade de custos tradicional; outra parte, formada profissionalmente já sob a lógica das métricas de receita típicas do SaaS (CAC, LTV, MRR), precisa aprendê-la pela primeira vez, definindo um direcionador de custo operacional e redesenhando modelos de precificação que não dependam exclusivamente do assento como unidade de cobrança.

2 REFERENCIAL TEÓRICO

2.1 Fundamentos da contabilidade de custos

A análise de margem de contribuição por objeto de custo (cliente, produto ou canal) é um dos instrumentos mais consolidados da contabilidade gerencial (Horngren, Datar, & Rajan, 2021). Sua utilidade depende, contudo, da existência de custo variável identificável por objeto: quando o custo de servir se aproxima de zero e é homogêneo entre clientes, a margem de contribuição converge para a própria receita, esvaziando seu poder analítico. Fenômeno análogo ocorre com a curva ABC de rentabilidade, que ordena clientes por lucratividade decrescente e tipicamente revela concentração de lucro nos primeiros decis e destruição de valor nos últimos (Cooper & Kaplan, 1999); sem variabilidade de custo por cliente, a curva se achata e perde relevância como indicador.

O Custeio Baseado em Atividade e Tempo (Time-Driven Activity-Based Costing, TDABC), proposto por Kaplan e Anderson (2007), oferece um paralelo conceitual direto ao problema do AI-SaaS. O TDABC estima o custo de uma atividade a partir de duas variáveis, o custo da capacidade fornecida e o tempo necessário para executar a atividade, convertendo uma unidade de processamento interna (o tempo) em custo alocado a um objeto de custo externo, como um pedido, um cliente ou um produto. A analogia com o AI-SaaS é direta: assim como o TDABC

usa o tempo como unidade interna de processamento a ser convertida em custo por transação, o AI-SaaS precisa usar o token como unidade interna de processamento a ser convertida em um direcionador de custo de interação, atrelado ao valor entregue ao cliente. Essa relação entre direcionador de custo (cost driver) e rastreabilidade de custo variável (Cooper & Kaplan, 1999) é o conceito central deste bloco e fundamenta a discussão apresentada na Seção 5.

2.2 Economia da informação e o custo marginal em software

Shapiro e Varian (1999) estabeleceram que bens de informação possuem estrutura de custos com elevados custos fixos de produção e custo marginal de reprodução próximo de zero, pressuposto que tornou os instrumentos da Seção 2.1 pouco relevantes, na prática, para o setor de software por duas décadas. Brynjolfsson e McAfee (2014) situam esse fenômeno no contexto mais amplo da economia da digitalização, em que a escala se dissocia do custo marginal. Buxmann, Diefenbach e Hess (2013) descrevem o modelo econômico da indústria de software resultante dessa estrutura de custos, no qual a receita recorrente se descola do custo de entrega marginal.

2.3 Modelos de negócio SaaS: estrutura de custos, métricas unitárias (unit economics) e precificação por assento

O regime de custo marginal próximo de zero gerou, no setor SaaS, um conjunto específico de métricas de gestão e um modelo de precificação característico. Ojala (2013) propõe uma tipologia de modelos de receita SaaS centrada na precificação por assento (seat-based pricing), na qual o cliente paga um valor fixo por usuário habilitado, independentemente da intensidade de uso, consistente com a premissa de que o custo de servir não varia de forma relevante com o consumo. Choudhary (2007) compara a estrutura econômica do modelo on-premise, o software instalado localmente, com o SaaS, demonstrando a transição de um modelo de custo fixo elevado por licença para um modelo de custo de entrega marginal decrescente. LTV, CAC, ARR (Receita Recorrente Anual) e churn (taxa de cancelamento de clientes) tornam-se as métricas centrais porque a variável determinante da rentabilidade é a aquisição e retenção do cliente, não o custo de servi-lo (Lehmann & Buxmann, 2009). Li e Kumar (2022), em agenda de pesquisa publicada na Production and Operations Management, identificam a escassez de pesquisa sobre gestão operacional de empresas SaaS, reforçando a pertinência de investigações sobre a estrutura de custos e de precificação do setor.

2.4 Custos de inferência como fenômeno emergente

A literatura técnica sobre custo computacional de modelos de IA generativa é recente e concentrada fora do campo contábil, mas já permite delimitar o objeto deste artigo. O custo de inferência distingue-se do custo de treinamento por sua natureza recorrente: enquanto o treinamento é incorrido uma única vez, o custo de inferência é incorrido a cada requisição atendida pelo modelo em produção, e cresce proporcionalmente ao volume de tokens processados e à extensão da resposta gerada (Sardana, Portes, Doubov, & Frankle, 2024). Patterson et al. (2021) quantificam o custo computacional e energético do treinamento de redes neurais de grande escala, estabelecendo a base de comparação com a fase de inferência. Samsi et al. (2023) realizam uma avaliação comparativa (benchmarking) do custo

energético por consulta (query) em LLMs em produção, demonstrando empiricamente a variabilidade do custo por interação. Wu et al. (2022) discutem os custos de operação de modelos de IA em ambiente produtivo sob a ótica da sustentabilidade. Sardana, Portes, Doubov e Frankle (2024) demonstram que decisões de dimensionamento de modelo devem considerar conjuntamente o custo de treinamento e o custo de inferência esperado ao longo da vida útil do modelo, tratando ambos como componentes de uma mesma função de custo total. Erdil (2025) modela a relação de compensação entre custo e velocidade na inferência de LLMs, relevante para decisões de roteamento entre modelos de diferentes portes e custos. Bergemann, Bonatti e Smolin (2025) analisam a alocação ótima de tokens e a precificação ótima de modelos de linguagem sob uma perspectiva de desenho de mercado, relevante para a discussão de precificação por consumo retomada na Seção 6. Zhu (2026) propõe o conceito de "tokenomics" de IA, formalizando o token como unidade que conecta processamento de informação, computação, consumo de energia, precificação e valor econômico, e distinguindo explicitamente o gasto em tokens do valor econômico gerado por eles, distinção que conecta diretamente com o problema do direcionador de custo e do direcionador de valor discutido na Seção 5. Em conjunto, essa literatura estabelece que o custo de inferência é variável por natureza, proporcional ao volume de tokens processados, mas, salvo por incursões pontuais e recentes (Zhu, 2026; Bergemann, Bonatti, & Smolin, 2025), não trata esse fenômeno como problema de contabilidade de custos gerencial, o que constitui a lacuna que este artigo busca situar dentro do campo contábil.

3 METODOLOGIA

O presente estudo classifica-se como pesquisa qualitativa exploratória, de natureza teórico-conceitual, combinando revisão de literatura e análise documental. Essa combinação é coerente com as diretrizes de Meneghetti (2011) para ensaios teóricos em contabilidade, particularmente quando o objetivo é estruturar e discutir um fenômeno emergente ainda não tratado sistematicamente na literatura acadêmica, a partir do confronto entre pressupostos teóricos e evidências empíricas.

3.1 Critérios de seleção da amostra documental

A amostra documental é composta por nove empresas de capital aberto nos Estados Unidos, organizadas em dois grupos. O primeiro grupo é composto por sete fornecedores SaaS estabelecidos que integraram IA generativa a produtos legados (HubSpot, monday.com, Salesforce, ServiceNow, Adobe, Atlassian e DocuSign). O segundo grupo é composto por duas empresas cujo núcleo de negócio é a própria inteligência artificial (Palantir Technologies e C3.ai), incluídas para verificar se o comportamento observado no primeiro grupo se repete, ou não, em empresas cuja proposta de valor já nasce baseada em IA.

A escolha dessas nove empresas decorre dos critérios explicados a seguir. Primeiro, o critério de disponibilidade da informação: apenas empresas de capital aberto sujeitas às exigências de divulgação da Security Exchange Comission (SEC) publicam, de forma auditada, o detalhamento de custo da receita e a discussão textual da Discussão e Análise da Administração (MD&A), que permite rastrear o comportamento do custo de entrega do serviço; isso exclui empresas privadas ou de capital fechado, ainda que fossem relevantes para o tema. Segundo, o critério de diversidade de segmento: a lista das empresas cobre diferentes segmentos de SaaS (CRM, gestão de trabalho, plataforma empresarial, criação de conteúdo,

colaboração e gestão de contratos), além de duas empresas nativas de IA. Terceiro, o critério de disponibilidade de evidência textual direta e específica aos três eixos deste artigo: cada uma das nove empresas foi selecionada por disponibilizar, em seus formulários regulatórios, press releases ou declarações públicas de executivos, evidência explícita sobre pelo menos um dos três eixos analisados (mudança para custeio variável, identificação do direcionador de valor do cliente, ou definição de um direcionador de custo para precificação por consumo). A bolsa brasileira tem uma amostra muito pequena de empresas SaaS e o movimento de integrações com Inteligência Artificial em produtos está ocorrendo há mais tempo nos Estados Unidos.

Uma pergunta metodológica pertinente é por que a amostra compreende nove empresas, e não um número maior ou menor. A resposta decorre da lógica de saturação teórica característica de pesquisas qualitativas exploratórias, e não de um número-alvo definido a priori. A coleta documental foi conduzida de forma incremental, iniciando pelas três empresas que motivaram o projeto original de pesquisa (HubSpot, monday.com e Salesforce) e expandindo-se, em rodadas sucessivas, a outras empresas de capital aberto identificadas como potencialmente relevantes aos três eixos do artigo. A cada rodada, empresas candidatas foram avaliadas quanto à existência de evidência documental direta a pelo menos um dos três eixos. A coleta foi encerrada quando empresas candidatas adicionais deixaram de acrescentar evidência nova aos três eixos, oferecendo apenas variações do mesmo tipo de evidência já documentado nas nove empresas incluídas. Esse critério de saturação, e não um número arbitrário, determinou o fechamento da amostra: um número menor comprometeria a diversidade de segmentos e deixaria de capturar evidência específica de pelo menos um eixo, enquanto um número maior exigiria incluir empresas cuja divulgação disponível não acrescentasse informação nova aos três eixos.

A combinação desses critérios caracteriza uma amostragem não probabilística por julgamento, adequada a estudos exploratórios de natureza qualitativa, mas que limita a generalização estatística dos achados, limitação reconhecida na Seção 8 e que orienta a agenda de pesquisa da Seção 7. Os documentos primários efetivamente consultados, com identificação completa de emissor, tipo de documento, período de referência e link de acesso, estão identificados na lista de Referências, permitindo a auditoria e a replicação da análise documental.

Empresa	**Segmento**	**Evidência documental**	**Eixo evidenciado**
HubSpot Inc.	CRM e marketing	Custo de receita atribuído textualmente ao crescimento da assistente de IA Breeze	Custeio variável
monday.com Ltd.	Gestão de trabalho	Administração sinaliza formalmente aumento futuro de custo de infraestrutura de IA	Custeio variável
Salesforce Inc.	CRM e nuvem empresarial	Divulga volume de tokens processados; criou as unidades Agentic Work Unit e Flex Credits	Custeio variável; direcionador de custo
ServiceNow, Inc.	ITSM e CRM	CFO distingue token	Direcionador de valor

	empresarial	(custo) de resultado resolvido (valor); quantifica a parcela de IA no custo de servir	
Adobe Inc.	Criação de conteúdo	Criou os "generative credits" como unidade de consumo do Firefly	Direcionador de custo
Atlassian Corporation	Colaboração e devops	Criou os "Rovo credits"; consumo de créditos de IA cresce mês a mês	Direcionador de custo
Docusign, Inc.	Gestão de contratos	Lançou precificação por consumo para sua linha de produtos de IA (IAM)	Precificação por consumo

Quadro 1. Amostra documental: fornecedores SaaS estabelecidos

Fonte: elaborado pelo autor a partir dos formulários 10-K/20-F e press releases das empresas (2025, 2026).

Empresa	**Núcleo de negócio**	**Evidência documental**	**Eixo evidenciado**
Palantir Technologies Inc.	Plataforma de dados e IA (AIP)	Custo de entrega dominado por serviços profissionais ("forward deployed engineers"), não por tokens	Custeio variável (delimitação)
C3.ai, Inc.	Aplicações de IA empresarial	Custo de receita varia conforme a fase do contrato (piloto ou produção); parte da receita já é precificada por consumo	Custeio variável; precificação por consumo

Quadro 2. Amostra documental: empresas de núcleo em inteligência artificial

Fonte: elaborado pelo autor a partir dos formulários 10-K e press releases das empresas (2025, 2026).

A análise foi conduzida em duas etapas. Na primeira, a revisão de literatura mapeou o regime de custo marginal zero e os instrumentos clássicos que o setor de software, especialmente no modelo SaaS, havia relegado à irrelevância prática. Na segunda, a análise documental identificou, nas nove empresas, evidências da reintrodução do custeio variável, do desafio de definição do direcionador de custo a partir do valor percebido pelo cliente, e da transição da precificação por assento para a precificação por consumo. O confronto entre pressupostos teóricos e evidências documentais constitui a estrutura central, do qual foram derivados a tipologia de estruturas de Custo Variável (COGS) e a agenda de pesquisa apresentadas na Seção 7.

4 A MUDANÇA PARA O CUSTEIO VARIÁVEL: EVIDÊNCIAS DOCUMENTAIS

O custo direto de entrega das nove empresas analisadas passou a incorporar um componente variável, proporcional ao consumo de inferência. Essa evidência é, em sua maior parte, de natureza textual e operacional, e aparece de três formas recorrentes nas empresas do Quadro 1. Primeiro, a atribuição textual, pela própria

administração, do crescimento do custo de receita a investimentos em infraestrutura de IA: a HubSpot associa o crescimento do custo de receita à escala de sua assistente Breeze (HubSpot Inc., 2025; TIKR, 2026); a monday.com sinaliza formalmente a investidores que o crescimento do uso de IA deverá aumentar seus custos de infraestrutura e computação (monday.com Ltd., 2025); a Adobe reporta aumento de 8% no custo de receita no exercício de 2025, período de expansão do Firefly (Adobe Inc., 2026). Segundo, a criação de métricas operacionais de volume de inferência que não existiam no modelo SaaS tradicional: a Salesforce divulga 19 trilhões de tokens processados no exercício fiscal de 2026 (Salesforce Inc., 2026); a Atlassian reporta crescimento mensal superior a 20% no consumo de créditos de IA de sua plataforma Rovo (Futurum Group, 2026); a ServiceNow quantifica textualmente a parcela de seu custo de servir associada ao processamento de IA, o que a empresa denomina "reasoning" (raciocínio) (Fortune, 2026). Terceiro, o próprio lançamento de modelos de precificação por consumo, que pressupõe a existência de um custo variável a ser referenciado: a Docusign lançou precificação baseada em consumo para sua linha de produtos de IA, denominada IAM (Intelligent Agreement Management), no início de 2026 (Alpha-Sense, 2026).

O Quadro 2 mostra que esse mesmo tipo de evidência aparece, com contornos distintos, entre empresas cujo núcleo de negócio já é a inteligência artificial, o que reforça o argumento de que a mudança de natureza do custo é um fenômeno estrutural, e não um efeito específico da transição de incumbentes SaaS. A C3.ai atribui, em suas próprias divulgações, a variação de seu custo de receita à diferença entre projetos em fase de piloto e em fase de produção (C3 AI, 2026), evidência direta de custo variável associada à natureza do contrato, e já comercializa parte de sua receita sob precificação baseada em consumo (Datainsightsmarket.com, 2025). A Palantir, por sua vez, traz um componente para delimitar o argumento deste artigo: seu custo de entrega é dominado por serviços profissionais de implementação, os chamados "forward deployed engineers" (engenheiros de implantação em campo, em tradução livre), que respondem por cerca de 18% a 20% da receita, e não pelo custo de inferência de tokens propriamente dito, que a empresa não divulga separadamente (PitchGrade Research, 2026). O que interessa para neste artigo é especificamente a mudança no custo de processamento computacional relacionado à Inteligência Artificial, e os custos de implantação ou a oferta de serviços profissionais poderiam estar associados a qualquer produto de software.

A leitura conjunta dos Quadros 1 e 2 permite uma conclusão delimitada: nove empresas de capital aberto, cobrindo seis segmentos de software e dois estágios distintos de maturidade em IA (incumbentes em transição e nativas de IA), apresentam evidência textual convergente de que seu custo direto de entrega passou a incorporar um componente variável proporcional ao consumo de inferência. O que decorre diretamente dessa mudança de natureza do custo, e que é objeto das Seções 5 e 6, é a necessidade prática de identificar o que o cliente valoriza (o direcionador de valor) e de definir uma unidade operacional (o direcionador de custo) que sirva de base tanto ao custeio quanto à precificação.

5 O DIRECIONADOR DE VALOR DO CLIENTE E A DEFINIÇÃO DO DIRECIONADOR DE CUSTO PARA PRECIFICAÇÃO

Estabelecida a mudança para o custeio variável, o desafio seguinte é como transformar o custo de inferência, mensurado em tokens, em uma base confiável

para o custeio por cliente e para a precificação. Esse desafio tem dois componentes distintos que a literatura já trata separadamente em outros contextos, mas que o AI-SaaS obriga a tratar em conjunto: de um lado, a definição de um direcionador de custo operacional, tema central da literatura de custeio por atividade (Cooper & Kaplan, 1999; Kaplan & Anderson, 2007); de outro, a identificação do direcionador de valor percebido pelo cliente, tema tradicionalmente associado à literatura de precificação por valor (Nagle & Holden, 2002).

O direcionador de valor percebido pelo cliente é aquilo que o cliente efetivamente compra e reconhece como entrega, por exemplo, uma pesquisa concluída, um relatório gerado, uma tarefa executada por um agente de IA, e não o volume de tokens processados para produzi-la. A Chief Financial Officer da Servicenow declarou publicamente que os clientes não pagam pelos tokens processados, mas por um resultado resolvido (Fortune, 2026). Essa afirmação resume a distinção entre o direcionador de custo real da operação (o token, unidade de custo sem significado para o cliente) e o direcionador de valor percebido (o resultado entregue, unidade que orienta a disposição do cliente a pagar).

A tarefa da empresa é construir uma unidade intermediária que traduza o direcionador de custo real para o direcionador de valor percebido, de modo que essa unidade possa, simultaneamente, ser rastreada ao custo de inferência e servir de base à precificação. O Quadro 1 mostra que essa construção não é um caso isolado, mas um padrão que se repete, de forma independente, em pelo menos três das sete empresas do primeiro grupo. A Salesforce criou o Agentic Work Unit (AWU), definido textualmente como uma medida de tarefas executadas por agentes de IA em produção, e reporta essa métrica ao lado do volume bruto de tokens processados (Salesforce Inc., 2026); a empresa também introduziu "Flex Credits", precificados a US$ 0,10 por ação de agente executada (BuildMVPFast, 2026). A Adobe criou os "generative credits", unidade de consumo que baliza o uso do Firefly dentro dos planos do Creative Cloud (Adobe Inc., 2026). A Atlassian criou os "Rovo credits", unidade de consumo compartilhada entre os usuários de uma mesma conta, que baliza o uso de sua plataforma de agentes Rovo (Atlassian, 2026). Em nenhum dos três casos a unidade utilizada é o próprio token; em todos os três casos, ela é uma unidade proprietária, definida pela empresa, que se posiciona entre o custo real de inferência e a tarefa que o cliente reconhece como valor entregue.

Essa convergência entre empresas de segmentos distintos (CRM, criação de conteúdo, colaboração), que não coordenaram entre si a criação dessas unidades, sugere que a necessidade de um direcionador de custo de interação intermediário não é apenas uma escolha estratégica particular de uma empresa, mas uma exigência imposta pela própria mudança de natureza do custo discutida na Seção 4. Essa construção é, além disso, conceitualmente análoga à lógica do TDABC (Kaplan & Anderson, 2007): assim como o tempo de processamento de uma atividade é convertido em custo alocado a um pedido ou cliente, o token processado precisa ser convertido em um direcionador de custo de interação que seja, simultaneamente, rastreável ao custo de inferência e inteligível para fins de precificação e de análise de rentabilidade. Sem essa camada de tradução, a empresa mede corretamente seu custo total de inferência, mas não consegue alocá-lo de forma confiável a clientes, produtos ou funcionalidades individuais, o que inviabiliza tanto a visão de margem por cliente quanto da curva ABC de rentabilidade, mesmo quando os dados brutos de consumo de tokens estão disponíveis.

A definição desse direcionador de custo de interação é também o que permite reabilitar, na prática, os instrumentos discutidos na Seção 2.1. A análise de margem

por cliente volta a ser necessária porque o custo de servir deixa de ser homogêneo entre clientes: um cliente que consome grande volume de interações de IA gera custo variável superior ao de um cliente de uso moderado, ainda que ambos paguem hoje o mesmo preço por assento. A curva ABC de rentabilidade volta a ser instrumento de gestão relevante pelo mesmo motivo: uma vez definido o direcionador de custo de interação, ela deixa de convergir para a distribuição de receita e passa a revelar a concentração real de custo em conjuntos de clientes ou funcionalidades de IA, tal como o TDABC revela concentração de custo em atividades específicas (Kaplan & Anderson, 2007).

6 DA PRECIFICAÇÃO POR ASSENTO (SEAT-BASED PRICING) AO CONSUMO: A INADEQUAÇÃO DESSE MODELO EM AI-SAAS

A precificação por assento (seat-based pricing), consolidada no modelo SaaS tradicional (Ojala, 2013), pressupõe que o custo de servir cada usuário habilitado seja aproximadamente constante, independentemente da intensidade de uso. Essa premissa é coerente com o regime de custo marginal próximo de zero discutido na Seção 2.2, mas não se sustenta quando o custo de servir passa a variar por interação: dois usuários pagando o mesmo preço por assento podem gerar custos de inferência distintos, dependendo do volume e da complexidade das interações de IA que realizam.

A evidência documental deste artigo, reunida nos Quadros 1 e 2, confirma que essa dispersão de uso é tratada como um problema concreto de precificação pelas empresas. Ao menos quatro das nove empresas analisadas já adotaram, ou anunciaram, algum mecanismo de precificação por consumo atrelado a um direcionador de custo de interação: a Salesforce, com a AWU e os Flex Credits; a Adobe, com os generative credits; a Atlassian, com os Rovo credits; a Docusign, com a precificação por consumo de sua linha IAM; e a C3.ai, com parte de sua receita já precificada por consumo. Esse conjunto de casos, obtidos de forma independente entre segmentos distintos de software, sugere que a migração da precificação por assento para a precificação por consumo pode ser uma tendência geral do mercado de AI-SaaS como resposta à mudança de natureza do custo documentada na Seção 4.

O desenho mais comum observado na amostra é a adoção de um modelo híbrido, no qual um componente de assinatura ou de assento é somado a um componente variável atrelado ao direcionador de custo de interação. Esse desenho é viável porque a empresa definiu previamente um direcionador de custo de interação capaz de servir, simultaneamente, de referência de custo e de unidade de cobrança. Com a mudança no custo evidenciada pelas empresas, permanecer somente com o modelo de preço fixo colocaria em risco a margem bruta do negócio.

A generalização desse padrão para o setor de AI-SaaS como um todo tem implicações diretas para a contabilidade gerencial: o direcionador de custo de interação deixa de ser apenas um instrumento de custeio interno e passa a ser, também, a unidade de cobrança externa, aproximando, na prática, a lógica de custeio por atividade (Cooper & Kaplan, 1999; Kaplan & Anderson, 2007) da lógica de precificação, algo que o modelo de assento fixo do SaaS tradicional nunca exigiu.

7 TIPOLOGIA DE ESTRUTURAS DE CUSTO DE BENS E SERVIÇOS VENDIDOS (COGS) EM SOFTWARE E AGENDA DE PESQUISA

A partir do referencial teórico e da evidência documental apresentados, este artigo propõe uma tipologia de três estruturas de custo de bens/serviços vendidos (COGS) no setor de software, organizadas por relação entre custo, direcionador de custo e modelo de precificação. No modelo on-premise, o custo dominante é fixo e concentrado na venda da licença, com custo marginal de distribuição praticamente nulo após o desenvolvimento do produto, e a precificação é predominantemente perpétua ou por licença (Choudhary, 2007). No modelo SaaS tradicional, o custo de entrega migra para hospedagem e suporte, mantendo-se, ainda assim, próximo de zero por usuário adicional em escala (Shapiro & Varian, 1999), e a precificação consolida-se por assento, sem necessidade de um direcionador de custo pelo uso do software. No modelo AI-SaaS, o custo de entrega passa a incorporar um componente variável e proporcional ao volume de interação, mensurado em tokens processados, exigindo a definição de um direcionador de custo de interação que traduza esse custo para o direcionador de valor percebido pelo cliente, e a precificação migra, ao menos parcialmente, para modelos de consumo referenciados a esse direcionador de custo.

Dessa tipologia e da análise documental decorre a seguinte agenda de pesquisa. Primeiro, a ausência de um modelo conceitual para a definição do direcionador de custo de interação a partir do direcionador de valor do cliente constitui a lacuna teórica mais imediata, sugerindo estudos futuros orientados a desenvolver e testar modelos de definição de direcionador de custo para fins de custeio e precificação em AI-SaaS. Segundo, a variabilidade de custo por intensidade de uso abre espaço para propostas de métricas unitárias (unit economics) adaptadas ao custo variável por interação. Terceiro, a transição de modelos de precificação por assento para modelos de consumo carece de investigação sistemática sobre seus determinantes e sobre os critérios de desenho de modelos híbridos, incluindo o papel do direcionador de custo de interação como unidade de cobrança. Quarto, a prática de evidenciação voluntária de custos de inferência como componente do custo de produtos vendidos ainda é incipiente e heterogênea entre as empresas analisadas, variando de menções textuais nos Relatórios da Administração à criação de métricas operacionais específicas, como volume de tokens e direcionadores de custo proprietários. Isso sugere pesquisa sobre os determinantes dessa prática de divulgação e sobre critérios de materialidade aplicáveis a esse tipo de custo, tema distinto da definição conceitual de custo direto já estabelecida pelos pronunciamentos contábeis vigentes. Quinto, a observação para confirmação do mesmo fenômeno em empresas brasileiras.

8 CONSIDERAÇÕES FINAIS

Este artigo argumentou que o AI-SaaS representa um retorno histórico à lógica de custeio variável tradicional no setor de software, a partir de três eixos complementares: a mudança do custo de entrega de quase-fixo para variável por interação do usuário; o desafio de identificar o direcionador de valor percebido pelo cliente e traduzi-lo em um direcionador de custo operacional; e a inadequação da precificação por assento diante dessa nova estrutura de custo, com a consequente migração para modelos de precificação por consumo. A partir da revisão de literatura e da análise documental de relatórios regulatórios e setoriais, identificaram-se os instrumentos clássicos da contabilidade de custos (margem por cliente, curva ABC

de rentabilidade e custeio por transação) que esse retorno reabilita, e caracterizou-se o desafio de definição do direcionador de custo de interação e sua relação com a migração para a precificação por consumo.

A principal contribuição do artigo é de natureza teórico-conceitual: articular, com apoio em evidência documental pública reunida a partir de nove empresas de capital aberto, um argumento estruturado em três camadas que se sustentam mutuamente. A primeira camada, a mudança de natureza do custo, é evidenciada textualmente em sete fornecedoras SaaS estabelecidas e em duas empresas de núcleo em inteligência artificial. A segunda e a terceira camadas, a definição do direcionador de custo de interação e a migração para a precificação por consumo, encontram evidência convergente em pelo menos três casos independentes (Salesforce, Adobe e Atlassian).

As principais limitações do estudo decorrem de sua natureza exploratória e do escopo documental: a amostra de nove empresas, selecionada por critério de propósito e justificada na Seção 3.1, não é probabilística e não permite generalização estatística. Estudos futuros, conforme a agenda proposta na Seção 7, poderão testar empiricamente o modelo conceitual de definição de direcionador de custo aqui discutido e ampliar a base documental para uma amostra mais representativa e numerosa de empresas SaaS e de empresas nativas de IA, avançando a reabilitação teórica proposta neste artigo em direção à validação empírica sistemática.

REFERÊNCIAS

Adobe Inc. (2026). *Annual report on Form 10-K: Fiscal year ended November 28, 2025*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/796343/000079634326000003/adbe-20251128.htm

Alpha-Sense. (2026). *Docusign Inc earnings: Analysis and highlights for Q4 2025*. https://www.alpha-sense.com/earnings/docu/

Atlassian Corporation. (2025). *Annual report on Form 10-K: Fiscal year ended June 30, 2025*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1650372/000165037225000036/team-20250630.htm

Bergemann, D., Bonatti, A., & Smolin, A. (2025). *The economics of large language models: Token allocation, fine-tuning, and optimal pricing* (arXiv:2502.07736). https://arxiv.org/abs/2502.07736

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

BuildMVPFast. (2026). *AI agent pricing: Why subscriptions fail and what works in 2026*. https://www.buildmvpfast.com/blog/ai-agent-pricing-subscription-usage-outcome-2026

Buxmann, P., Diefenbach, H., & Hess, T. (2013). *The software industry: Economic principles, strategies, perspectives*. Springer.

C3 AI. (2026). *C3 AI announces fiscal third quarter 2026 results* [Press release]. https://c3.ai/c3-ai-fiscal-third-quarter-2026-results/

Choudhary, V. (2007). Comparison of software quality under perpetual licensing and software as a service. *Journal of Management Information Systems*, 24(2), 141–165.

Cooper, R., & Kaplan, R. S. (1999). *The design of cost management systems* (2nd ed.). Prentice Hall.

Datainsightsmarket.com. (2025). *C3.ai, Inc. (AI) stock price, market cap, segmented revenue & earnings*. https://www.datainsightsmarket.com/companies/AI

Docusign, Inc. (2026). *Annual report on Form 10-K: Fiscal year ended January 31, 2026*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1261333/000126133326000021/docu-20260131.htm

Erdil, E. (2025). *Inference economics of language models* (arXiv:2506.04645). https://arxiv.org/abs/2506.04645

Fortune. (2026). *ServiceNow just told Wall Street it's going to double again*. https://fortune.com/2026/05/06/servicenow-30-billion-revenue-not-crazy-why/

Futurum Group. (2026). *Atlassian Q3 FY 2026 earnings show continued cloud and AI-led expansion*. https://futurumgroup.com/insights/atlassian-q3-fy-2026-earnings-show-continued-cloud-and-ai-led-expansion/

Horngren, C. T., Datar, S. M., & Rajan, M. V. (2021). *Cost accounting: A managerial emphasis* (16th ed.). Pearson.

HubSpot Inc. (2025). *Annual report on Form 10-K: Fiscal year ended December 31, 2024*. U.S. Securities and Exchange Commission. https://www.sec.gov

HubSpot Inc. (2026). *HubSpot reports strong Q4 and full year 2025 results* [Press release]. https://ir.hubspot.com

Kaplan, R. S., & Anderson, S. R. (2007). *Time-driven activity-based costing: A simpler and more powerful path to higher profits*. Harvard Business School Press.

Lehmann, S., & Buxmann, P. (2009). Pricing strategies of software vendors. *Business & Information Systems Engineering*, 1(6), 452–462.

Li, B., & Kumar, S. (2022). Managing software-as-a-service: Pricing and operations. *Production and Operations Management*, 31(6), 2588–2608.

Martins, E. (2018). *Contabilidade de custos* (11a ed.). Atlas.

Meneghetti, F. K. (2011). O que é um ensaio-teórico? *Revista de Administração Contemporânea*, 15(2), 320–332.

monday.com Ltd. (2025). *Annual report on Form 20-F: Fiscal year ended December 31, 2024*. U.S. Securities and Exchange Commission. https://www.sec.gov

monday.com Ltd. (2026). *monday.com announces fourth quarter and fiscal year 2025 results* [Press release]. https://ir.monday.com

Nagle, T. T., & Holden, R. K. (2002). *The strategy and tactics of pricing: A guide to profitable decision making* (3rd ed.). Prentice Hall.

Ojala, A. (2013). Software-as-a-service revenue models. *IEEE Software*, 30(3), 54–60.

Palantir Technologies Inc. (2026). *Annual report on Form 10-K: Fiscal year ended December 31, 2025*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1321655/000132165526000011/0001321655-26-000011-index.htm

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon emissions and large neural network training* (arXiv:2104.10350). https://arxiv.org/abs/2104.10350

PitchGrade Research. (2026). *Palantir's business model: Government anchor, commercial optionality, and the AI pivot*. https://pitchgrade.com/research/palantir-business-model

Salesforce Inc. (2024). *Annual report on Form 10-K: Fiscal year ended January 31, 2024*. U.S. Securities and Exchange Commission. https://www.sec.gov

Salesforce Inc. (2026). *Salesforce delivers record fourth quarter fiscal 2026 results* [Press release]. https://investor.salesforce.com

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., & Gadepally, V. (2023). From words to watts: Benchmarking the energy costs of large language model inference. *2023 IEEE High Performance Extreme Computing Conference (HPEC)*.

Sardana, N., Portes, J., Doubov, S., & Frankle, J. (2024). Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. *Proceedings of the 41st International Conference on Machine Learning* (arXiv:2401.00448). https://arxiv.org/abs/2401.00448

ServiceNow, Inc. (2026). *Annual report on Form 10-K: Fiscal year ended December 31, 2025*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1373715/000137371526000007/now-20251231.htm

Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy*. Harvard Business School Press.

TIKR. (2026). *HubSpot Q1 2026 earnings: Margin expansion hits 2027 target a year early*. https://www.tikr.com/blog/hubspot-q1-2026-earnings-margin-expansion-hits-2027-target-a-year-early

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H. S., … Hazelwood, K. (2022). Sustainable AI: Environmental implications, challenges, and opportunities. *Proceedings of Machine Learning and Systems*, 4.

Zhu, Q. (2026). *AI tokenomics: The economics of tokens, computation, and pricing in foundation models* (arXiv:2606.24616). https://arxiv.org/abs/2606.24616