

VALORIAN



Glossário Executivo de IA

O vocabulário que chegou ao conselho — e o que cada termo significa na prática

Um guia de referência para executivos, diretores e conselheiros

Sumário

·	Como usar este glossário	3
01	Fundamentos	4
02	Como um modelo aprende	7
03	Como um modelo responde	9
04	Agentes	12
05	Infraestrutura	15
06	Qualidade, avaliação e confiabilidade	17
07	Segurança e risco	19
08	Governança e regulação	21
09	Economia da IA e métricas de negócio	23
10	Termos que enganam o conselho	25
·	Doze perguntas para levar à próxima reunião	27
·	Nota de fontes e método	28

Como usar este glossário

Existe um custo silencioso em toda organização que adota inteligência artificial: o custo de decidir sobre algo que não se sabe nomear. Ele aparece em reuniões nas quais o conselho aprova um investimento sem conseguir formular a pergunta que revelaria o risco; em contratos assinados com cláusulas técnicas que ninguém do lado do negócio leu com rigor; em projetos que são declarados bem-sucedidos porque a métrica escolhida era a única que o fornecedor sabia produzir.

Este glossário não é um dicionário técnico. É um instrumento de conversa. Cada verbete responde a duas perguntas — **o que é** e **por que importa na mesa de decisão** — e a segunda é sempre mais importante que a primeira. O objetivo não é transformar executivos em engenheiros, mas dar a quem decide o vocabulário suficiente para separar substância de discurso comercial, e para fazer, na hora certa, a pergunta que o fornecedor não esperava.

Três orientações de leitura:

Não leia do início ao fim. As dez seções são independentes e seguem a lógica do stack de IA: fundamentos, treino, inferência, agentes, infraestrutura, avaliação, segurança, governança, economia e armadilhas semânticas. Vá à seção correspondente à decisão que está na sua mesa esta semana.

Leve as perguntas para a reunião. Cada verbete termina com uma linha marcada **No conselho**. São perguntas e critérios formulados para serem usados literalmente diante de um fornecedor, de um time interno ou de um comitê.

Termos envelhecem. Esta é a edição 2026.2. O vocabulário de IA se reorganiza em ciclos de seis a doze meses — três dos termos mais usados hoje em salas de diretoria não existiam com esse nome dois anos atrás. A seção 10 trata especificamente dos termos cujo significado se deslocou e hoje induzem erro.

Fundamentos

o vocabulário mínimo

Inteligência artificial (IA)

Campo que reúne técnicas para fazer sistemas executarem tarefas que exigiriam inteligência humana. É um guarda-chuva amplo: inclui desde regras codificadas manualmente até os modelos generativos atuais. Na prática corporativa de 2026, quando alguém diz "IA" quase sempre está falando de modelos de linguagem de grande porte.

NO CONSELHO

exija a especificidade. "Vamos usar IA" não é uma decisão; "vamos usar um modelo X para a tarefa Y com o dado Z" é.

Modelo de linguagem de grande porte (LLM)

Sistema treinado sobre volumes massivos de texto para prever a continuação mais provável de uma sequência. Dessa capacidade aparentemente simples emergem redação, tradução, raciocínio, programação e análise. É a tecnologia central por trás de praticamente todo produto de IA generativa em uso hoje.

NO CONSELHO

um LLM não consulta um banco de fatos — ele gera texto plausível. Toda arquitetura corporativa séria precisa dizer de onde vêm os fatos.

IA generativa

Categoria de modelos que produz conteúdo novo — texto, imagem, áudio, vídeo, código — em vez de apenas classificar ou prever valores. Distingue-se da IA preditiva tradicional, que responde "qual a probabilidade de churn deste cliente" em vez de "escreva o e-mail de retenção".

NO CONSELHO

as duas convivem. Boa parte do valor capturado em operação ainda vem de IA preditiva bem calibrada, que é mais barata, mais auditável e mais antiga na casa.

Parâmetros

Os valores numéricos ajustados durante o treinamento que codificam o que o modelo "sabe". Modelos de fronteira operam na casa das centenas de bilhões a trilhões de parâmetros. Mais parâmetros historicamente significaram mais capacidade, mas a relação deixou de ser linear.

NO CONSELHO

número de parâmetros virou métrica de marketing. Desde 2025, modelos menores e bem treinados superam modelos maiores e mal treinados em tarefas específicas.

Token

A unidade que o modelo efetivamente processa: um pedaço de palavra, uma palavra ou um símbolo. Regra prática em português: um token equivale a cerca de 0,7 palavra. É a unidade de cobrança — você paga por token de entrada e por token de saída — e a unidade em que se mede a janela de contexto.

NO CONSELHO

todo orçamento de IA é, no fundo, um orçamento de tokens. Peça a projeção em tokens/mês, não em "licenças".

Multimodal

Modelo capaz de receber e produzir mais de um tipo de dado — texto, imagem, áudio, vídeo — no mesmo fluxo. Um modelo multimodal lê a foto de um laudo, o áudio de uma consulta e o PDF do contrato sem pipelines separados.

NO CONSELHO

em setores documentais — saúde, jurídico, seguros, construção — a multimodalidade elimina uma camada inteira de integração e é frequentemente o maior ganho isolado.

Modelo de fronteira

Os sistemas mais capazes disponíveis em determinado momento, produzidos por um número pequeno de laboratórios. São mais caros por token, mais lentos e, para muitas tarefas corporativas, desnecessários.

NO CONSELHO

a pergunta não é "temos o melhor modelo?", e sim "temos o modelo adequado ao custo e ao risco de cada tarefa?".

Modelo aberto / peso aberto (open weights)

Modelo cujos parâmetros são publicados e podem ser baixados, executados e ajustados em infraestrutura própria. Não é sinônimo de código aberto no sentido clássico: os dados de treino e a licença podem ter restrições relevantes.

NO CONSELHO

é o principal caminho para soberania de dados e previsibilidade de custo, ao preço de assumir a operação. Leia a licença — "aberto" tem pelo menos cinco significados no mercado.

Modelo pequeno / SLM

Modelo compacto, executável em hardware modesto ou no próprio dispositivo. Perde em raciocínio aberto, ganha em latência, custo e privacidade.

NO CONSELHO

classificação, extração e roteamento — o volume real da operação — raramente exigem modelo de fronteira.

Destilação

Técnica em que um modelo grande "ensina" um modelo menor, transferindo boa parte da capacidade a uma fração do custo. Foi o mecanismo central da compressão de custos observada desde 2024.

NO CONSELHO

é a razão econômica pela qual capacidade de ponta chega ao seu orçamento com doze a dezoito meses de atraso — e por que esperar às vezes é a decisão financeira correta.

Como um modelo aprende

treino, ajuste e dados

Pré-treinamento

Fase inicial e mais cara, em que o modelo aprende padrões gerais de linguagem a partir de um corpus massivo. Custa dezenas a centenas de milhões de dólares e é feita por um punhado de laboratórios no mundo.

NO CONSELHO

nenhuma empresa brasileira não-laboratório deveria estar pré-treinando um modelo de propósito geral. Se um fornecedor propõe isso, peça a justificativa econômica linha a linha.

Fine-tuning (ajuste fino)

Continuação do treino de um modelo existente com dados específicos da sua operação, para fixar formato, tom, jargão ou padrão de decisão. Muda o comportamento do modelo de forma permanente.

NO CONSELHO

ajuste fino resolve **como** o modelo responde. Não resolve **o que** ele sabe sobre fatos que mudam — isso é RAG.

RAG (geração aumentada por recuperação)

Arquitetura em que o sistema busca documentos relevantes na sua base e os entrega ao modelo como contexto antes de gerar a resposta. É o padrão de mercado para respostas ancoradas em conhecimento corporativo.

NO CONSELHO

é a arquitetura que permite citar a fonte. Se um sistema corporativo responde sobre políticas internas sem apontar o documento de origem, há um problema de auditabilidade.

Embedding

Representação numérica de um texto, imagem ou registro que posiciona conteúdos semanticamente próximos em coordenadas próximas. É o mecanismo que faz a busca por significado funcionar.

NO CONSELHO

embeddings dos seus documentos são derivados dos seus dados e podem vaziar informação. Trate-os como dado sensível no contrato e na arquitetura.

Banco vetorial

Base de dados otimizada para armazenar embeddings e recuperar os itens mais similares a uma consulta. Componente de infraestrutura de quase todo sistema RAG.

NO CONSELHO

é mais um sistema para operar, versionar e proteger. Pergunte quem faz o backup e quem responde pela reindexação.

Aprendizado por reforço com feedback humano (RLHF)

Método de ajuste em que avaliadores humanos classificam respostas e o modelo é otimizado para produzir o que foi melhor avaliado. É o que transforma um preditor de texto em assistente utilizável.

NO CONSELHO

as preferências de quem avaliou ficam impressas no modelo. Modelos treinados em outro contexto cultural e regulatório trazem esses valores embutidos.

Dados sintéticos

Dados gerados por modelos para treinar outros modelos. Contornam escassez e restrições de privacidade, com o risco de amplificar vieses e erros da geração anterior.

NO CONSELHO

útil e legítimo, mas exige rastreabilidade. Peça a proporção de dado sintético em qualquer modelo ajustado com seus casos.

Curadoria de dados

Trabalho de seleção, limpeza, rotulagem e versionamento dos dados que alimentam treino e recuperação. É a parte menos glamourosa e mais determinante da qualidade final.

NO CONSELHO

em quase todo projeto que falha, a causa está aqui — não no modelo. Orce curadoria como item próprio, não como rodapé do projeto.

Cutoff de conhecimento

Data limite dos dados usados no treino. Sobre eventos posteriores, o modelo não sabe — e, pior, pode responder com confiança a partir de informação desatualizada.

NO CONSELHO

para qualquer domínio que muda — preço, norma, tabela, regulação — só arquitetura com busca ou base própria é aceitável.

Como um modelo responde

contexto, inferência e custo

Inferência

A execução do modelo para produzir uma resposta. É onde está o custo recorrente e o desafio de latência da operação; o treino é investimento único, a inferência é conta de luz.

NO CONSELHO

a pergunta financeira certa é o custo de inferência por transação de negócio, não o preço por milhão de tokens da tabela do fornecedor.

Janela de contexto

Quantidade máxima de tokens que o modelo consegue considerar de uma vez — prompt, documentos, histórico e resultados de ferramentas somados. Modelos de ponta em 2026 operam entre 400 mil e mais de 1 milhão de tokens.

NO CONSELHO

janela grande não é licença para despejar tudo dentro. Quanto mais ruído no contexto, pior a precisão — e maior a conta.

Engenharia de contexto

Disciplina que organiza tudo o que o modelo enxerga no momento da resposta: instrução de sistema, prompt do usuário, memória, documentos recuperados, histórico e retornos de ferramenta. Consolidou-se como termo em 2026 e engloba a engenharia de prompts, que trata apenas do texto digitado.

NO CONSELHO

é a camada onde mora a qualidade em sistemas com agentes. Se o time descreve a solução só em termos de "prompt", o desenho provavelmente não escala.

Prompt de sistema

Instrução persistente que define papel, limites e formato de resposta do modelo, invisível ao usuário final. É onde normalmente se colocam as regras de negócio e as restrições de conduta.

NO CONSELHO

é um artefato de governança. Deve ser versionado, revisado e auditável como qualquer política — não editado ad hoc por quem tem acesso ao painel.

Compactação de contexto

Compressão automática do histórico quando a conversa excede a janela. Resumos sucessivos podem diluir instruções críticas: uma regra explícita de "peça confirmação antes de apagar" pode virar "trabalhe com cuidado" — e o agente segue achando que obedece.

NO CONSELHO

pergunte ao time onde estão os guardrails e como estão protegidos da compactação. Se a resposta for "no prompt do sistema", investigue mais.

Memória

Capacidade de reter informação entre sessões, em arquivo, banco ou serviço dedicado. Diferente da janela de contexto, que é volátil.

NO CONSELHO

memória é dado pessoal acumulado ao longo do tempo. Precisa de política de retenção, exclusão e resposta a titular sob a LGPD.

Temperatura

Parâmetro que regula a variabilidade da saída. Baixa produz respostas estáveis e repetíveis; alta produz respostas criativas e menos previsíveis.

NO CONSELHO

processos regulados, contábeis ou clínicos exigem configuração determinística. Pergunte qual valor está em produção e quem pode alterá-lo.

Latência

Tempo entre a solicitação e a resposta. Determina quais casos de uso são viáveis: um copiloto de atendimento em tempo real e um processamento noturno de contratos têm exigências opostas.

NO CONSELHO

latência aceitável é decisão de negócio, não técnica. Defina-a no requisito, antes da escolha do modelo.

Cache de prompt

Reaproveitamento de partes fixas do contexto entre chamadas, reduzindo custo e latência de forma expressiva em aplicações com instruções longas e repetidas.

NO CONSELHO

é uma das poucas alavancas de economia que não exige troca de modelo. Vale perguntar se está ativada.

Roteamento de modelos

Estratégia de direcionar cada requisição ao modelo mais barato capaz de resolvê-la, escalando para modelos maiores só quando necessário.

NO CONSELHO

é a prática que separa operações de IA com custo sob controle das que estouram o orçamento no terceiro trimestre.

Agentes

a camada que virou produto em 2026

Agente

Sistema que usa um modelo para planejar e executar uma tarefa de várias etapas, chamando ferramentas, consultando sistemas e reagindo a resultados intermediários — em vez de apenas responder uma pergunta. É a mudança de comportamento que definiu 2026: de responder para agir.

NO CONSELHO

agente executa. Toda discussão sobre agentes é, portanto, uma discussão sobre permissões, reversibilidade e responsabilidade.

Tool use (uso de ferramentas)

Capacidade do modelo de solicitar a execução de uma função externa — consultar um ERP, emitir um boleto, abrir um chamado — e usar o retorno. O modelo decide qual chamar e com quais argumentos; quem executa é o seu código.

NO CONSELHO

a fronteira de segurança está no seu código, não no modelo. É lá que se implementam limites, validações e confirmações.

MCP (Model Context Protocol)

Protocolo aberto que padroniza a conexão entre modelos e sistemas externos — bases, APIs, ferramentas internas. Reduz integrações ponto a ponto e virou padrão de fato em arquiteturas corporativas de agentes.

NO CONSELHO

padronização de integração reduz aprisionamento a fornecedor. É um critério legítimo de arquitetura para constar em RFP.

Orquestração multiagente

Arquitetura em que vários agentes especializados dividem um fluxo, coordenados por um agente supervisor ou por um grafo de execução definido.

NO CONSELHO

aumenta capacidade e aumenta superfície de falha. Exija observabilidade por agente antes de aprovar produção.

Human-in-the-loop

Desenho em que um humano aprova, revisa ou intervém em pontos definidos do fluxo. Não é um controle genérico: é uma escolha explícita de onde colocar o ponto de parada.

NO CONSELHO

defina o gatilho por valor e por irreversibilidade — acima de tal valor, com efeito externo ou sem desfazer, passa por humano.

Autonomia graduada

Prática de escalonar a autonomia do agente conforme evidência acumulada de desempenho: começa sugerindo, passa a executar com aprovação, depois executa e notifica.

NO CONSELHO

é o mecanismo mais eficaz de gestão de risco em agentes. Peça a escada de autonomia documentada, com critérios de promoção.

Guardrails

Restrições explícitas sobre o que o agente pode ou não fazer, implementadas em código, política e permissão — não apenas em texto de instrução.

NO CONSELHO

guardrail que existe só no prompt é sugestão. Guardrail que existe na camada de permissão é controle.

Agente comercial / agentic commerce

Agentes que executam compra e negociação em nome de uma organização. Projeções de mercado apontam intermediação massiva de compras B2B por agentes até o fim da década.

NO CONSELHO

se o seu cliente passar a comprar por agente, seu processo comercial precisa ser legível por máquina. É uma questão de receita, não de TI.

Spec-driven development

Prática de desenvolvimento em que a especificação escrita é o artefato central e o código é gerado e regenerado a partir dela por agentes.

NO CONSELHO

muda a natureza da dívida técnica — a especificação passa a ser o ativo a proteger e versionar.

Loop de falha

Situação em que o agente repete uma ação sem sucesso, consumindo tokens e tempo, às vezes com efeitos colaterais em sistemas reais.

NO CONSELHO

limites de tentativa, teto de gasto por execução e interruptor de emergência são requisitos mínimos de produção.

Infraestrutura

chips, nuvem e energia

GPU

Processador originalmente gráfico que se tornou o hardware padrão de treino e inferência de IA. Sua disponibilidade e preço são hoje uma variável macroeconômica, não apenas uma linha de compras.

NO CONSELHO

contratos de nuvem com IA embarcada carregam exposição a essa cadeia. Cláusulas de reajuste e de disponibilidade merecem leitura jurídica específica.

TPU e aceleradores dedicados

Chips desenhados especificamente para cargas de IA, alternativos às GPUs de propósito geral. Grandes provedores desenvolvem os seus para reduzir dependência e custo.

NO CONSELHO

diversificação de silício no seu fornecedor é um indicador de resiliência de preço no médio prazo.

Data center de IA

Instalação com densidade energética e refrigeração muito superiores às de um data center tradicional. A restrição real de expansão deixou de ser capital e passou a ser conexão à rede elétrica.

NO CONSELHO

para o Brasil, matriz renovável e disponibilidade hídrica são vantagem competitiva real na atração dessa infraestrutura — e uma tese de investimento defensável.

Soberania de dados

Princípio de que dados permanecem sob jurisdição e infraestrutura de um país ou bloco determinado. Vai além de "servidor no Brasil": alcança quem controla a chave, quem pode ser compelido judicialmente e sob qual lei.

NO CONSELHO

pergunte a qual jurisdição o fornecedor responde em caso de ordem judicial estrangeira. A resposta muda a análise de risco.

Residência de dados

Localização física em que os dados são armazenados e processados. Necessária para conformidade, mas insuficiente para soberania.

NO CONSELHO

exija a distinção no contrato. São promessas diferentes com consequências jurídicas diferentes.

Inferência local / on-premise

Execução do modelo em infraestrutura própria, sem tráfego externo. Elimina classes inteiras de risco de vazamento ao custo de operação e capex.

NO CONSELHO

faz sentido quando o dado é de altíssima sensibilidade ou o volume torna a API mais cara que o hardware. Peça o ponto de equilíbrio calculado.

Aprisionamento tecnológico (vendor lock-in)

Grau de dificuldade e custo para trocar de fornecedor de modelo, nuvem ou plataforma. Em IA, o aprisionamento se dá menos pelo modelo e mais pelos dados, embeddings, integrações e prompts acoplados.

NO CONSELHO

peça a estimativa de custo e prazo de migração para outro fornecedor. Se ninguém souber responder, o lock-in já é alto.

Qualidade, avaliação e confiabilidade

Alucinação

Produção de informação incorreta com aparência de confiança — citações inexistentes, números inventados, cláusulas que não estão no contrato. É uma característica estrutural de modelos generativos, mitigável mas não eliminável.

NO CONSELHO

a pergunta certa não é "o modelo alucina?", e sim "qual o custo do erro neste processo e qual a barreira de verificação antes dele produzir efeito?".

Benchmark

Teste padronizado de capacidade. Útil para comparar modelos entre si, fraco para prever desempenho no seu caso específico, e sujeito a contaminação quando o teste vaza para os dados de treino.

NO CONSELHO

benchmark público é folder. Decisão se toma com avaliação sobre os seus dados.

Eval (avaliação)

Conjunto de casos representativos do seu processo, com respostas esperadas, usado para medir e monitorar o desempenho real do sistema ao longo do tempo.

NO CONSELHO

é o artefato mais valioso e mais negligenciado de um projeto de IA. Nenhum sistema deveria entrar em produção sem um, e ele deve pertencer à empresa, não ao fornecedor.

Fronteira irregular

Fenômeno em que o modelo executa brilhantemente tarefas difíceis e falha em tarefas aparentemente triviais, sem que a fronteira entre as duas seja intuitiva.

NO CONSELHO

é a razão pela qual a intuição executiva sobre "o que a IA consegue fazer" erra sistematicamente. Só teste empírico resolve.

Determinismo

Propriedade de produzir sempre a mesma saída para a mesma entrada. Modelos generativos são probabilísticos por natureza, o que colide com exigências de auditoria e reprodutibilidade.

NO CONSELHO

em processos regulados, registre entrada, versão do modelo, configuração e saída. Reprodutibilidade se constrói no log, não no modelo.

Drift

Degradação silenciosa do desempenho ao longo do tempo, por mudança nos dados de entrada, no comportamento dos usuários ou por atualização do modelo pelo fornecedor.

NO CONSELHO

exija monitoramento contínuo e notificação prévia de troca de versão. Um upgrade do fornecedor pode quebrar seu processo sem aviso.

Observabilidade

Capacidade de inspecionar o que o sistema fez: qual contexto recebeu, quais ferramentas chamou, quanto custou, onde falhou.

NO CONSELHO

sem observabilidade não há gestão, não há auditoria e não há defesa em caso de incidente. É requisito, não melhoria futura.

Explicabilidade

Capacidade de justificar como uma decisão foi produzida. Em modelos generativos, o que se obtém na prática é rastreabilidade — fontes, contexto e trilha de execução — mais do que explicação causal.

NO CONSELHO

o PL 2338 e o regramento europeu caminham para exigir explicação a quem é afetado por decisões automatizadas. Rastreabilidade é o que torna isso viável.

Segurança e risco

Prompt injection

Ataque em que instruções maliciosas escondidas em um conteúdo processado — um e-mail, um site, um PDF, um currículo — são interpretadas pelo agente como ordens legítimas.

NO CONSELHO

é a vulnerabilidade estrutural da era dos agentes. Todo conteúdo externo deve ser tratado como dado, nunca como comando.

Exfiltração de dados

Vazamento de informação sensível por meio do sistema de IA — em prompts enviados a terceiros, em respostas a usuários não autorizados ou em contexto recuperado indevidamente.

NO CONSELHO

controle de acesso precisa valer na camada de recuperação. Se o RAG ignora permissões, o chat vira canal de vazamento interno.

Shadow AI

Uso não autorizado de ferramentas de IA por colaboradores, geralmente com dados corporativos em contas pessoais.

NO CONSELHO

proibir aumenta o problema; o caminho eficaz é oferecer alternativa aprovada, boa e rápida, e comunicar com clareza o que pode e o que não pode.

Superfície de permissões

Conjunto de acessos, credenciais e sistemas que um agente pode alcançar. Cresce silenciosamente a cada nova integração.

NO CONSELHO

revise permissões de agentes com a mesma disciplina com que se revisa acesso de colaborador que muda de área.

Deepfake e fraude sintética

Uso de IA para imitar voz, imagem ou escrita de pessoas reais, com aplicação crescente em fraudes de autorização financeira e engenharia social contra executivos.

NO CONSELHO

processos de autorização de pagamento baseados em reconhecimento de voz ou vídeo estão obsoletos. Verificação por canal independente é o controle mínimo.

Red teaming

Teste adversarial deliberado do sistema, buscando fazê-lo falhar, vazar ou agir fora dos limites antes que alguém de má-fé o faça.

NO CONSELHO

deve ser recorrente e independente do time que construiu. Um relatório por ciclo de release é um pedido razoável.

Alinhamento

Grau em que o comportamento do sistema corresponde à intenção de quem o opera, inclusive em situações não previstas.

NO CONSELHO

o risco corporativo relevante não é ficção científica — é o agente que cumpre a instrução literal e produz um resultado que ninguém queria.

Governança e regulação

EU AI Act

Regulamento europeu que classifica sistemas de IA por nível de risco e impõe obrigações proporcionais. Alcança empresas de fora da UE cujos sistemas produzem efeitos no bloco, e o pacote de simplificação em discussão adiou parte das obrigações de alto risco para dezembro de 2027.

NO CONSELHO

se a empresa opera, vende ou processa dados na Europa, o enquadramento de risco de cada sistema precisa estar documentado.

PL 2338

Projeto brasileiro de marco legal da IA, aprovado no Senado em dezembro de 2024 e em tramitação na Câmara. Estrutura-se por níveis de risco, com deveres de transparência, governança e direitos de contestação para quem é afetado.

NO CONSELHO

empresas que já mantêm inventário de sistemas e trilha de decisão chegarão à vigência com custo marginal. As demais farão um projeto sob prazo.

LGPD aplicada a IA

A Lei Geral de Proteção de Dados já regula boa parte do uso de IA: base legal para tratamento, minimização, direito a revisão de decisão automatizada e transferência internacional.

NO CONSELHO

a pergunta prática é qual base legal sustenta o envio de dado pessoal a um modelo de terceiro — e se isso está no registro de operações.

ISO/IEC 42001

Norma internacional para sistema de gestão de IA, análoga em espírito à 27001 para segurança da informação. É certificável e vem sendo usada como evidência de diligência.

NO CONSELHO

é o caminho mais direto para transformar princípios de governança em processo auditável, e um diferencial concreto em licitações e contratos com grandes clientes.

Inventário de sistemas de IA

Registro de todos os sistemas de IA em uso, com finalidade, dados, fornecedor, responsável e classificação de risco.

NO CONSELHO

é o primeiro entregável de qualquer programa de governança e o primeiro documento que um regulador pedirá. Quase nenhuma empresa tem, e ele custa pouco.

Comitê de IA

Instância de governança com representação de negócio, jurídico, segurança, dados e tecnologia, responsável por aprovar casos de uso e definir limites de autonomia.

NO CONSELHO

sem foro de decisão, a governança vira documento. Com foro, vira prática — desde que tenha poder de barrar projeto.

Propriedade intelectual de saídas

Questão sobre quem detém direitos sobre conteúdo gerado por IA e qual o risco de infringir direitos de terceiros com base em material de treino.

NO CONSELHO

verifique no contrato se há indenização do fornecedor para reivindicações de terceiros sobre a saída. É uma cláusula que existe e se negocia.

Transparência ao usuário

Dever de informar quando uma pessoa interage com sistema automatizado ou é afetada por decisão automatizada.

NO CONSELHO

é obrigação de baixo custo e alto retorno reputacional. Implemente antes de ser exigido.

Economia da IA e métricas de negócio

Custo por transação

Custo total de IA — inferência, recuperação, orquestração, verificação humana — dividido pelo volume de transações de negócio efetivamente concluídas.

NO CONSELHO

é a única métrica de custo que se compara ao processo atual. Peça-a em toda apresentação de resultado.

Taxa de conclusão autônoma

Proporção de tarefas concluídas pelo sistema sem intervenção humana e com qualidade aceita. Métrica central da maturidade de agentes.

NO CONSELHO

olhe-a junto com a taxa de erro que escapou da verificação. As duas isoladas enganam.

Ganho de produtividade realizado

Diferença entre a economia de tempo medida e a economia de tempo efetivamente reconvertida em capacidade ou resultado.

NO CONSELHO

economizar duas horas por pessoa não é ganho até que essas horas tenham destino definido. Boa parte do ROI declarado de IA evapora neste ponto.

Piloto perpétuo

Padrão de falha em que a organização acumula provas de conceito bem-sucedidas que nunca entram em produção, por ausência de dono, integração ou modelo operacional.

NO CONSELHO

limite o número de pilotos simultâneos e exija critério de promoção definido antes do início. Piloto sem critério de saída é despesa recorrente.

Dívida de contexto

Acúmulo de prompts, integrações e configurações não documentadas que tornam o sistema frágil e dependente de indivíduos.

NO CONSELHO

pergunte quem consegue reconstruir a solução se a pessoa que a montou sair na semana que vem.

Buy, build ou compose

Decisão entre contratar solução pronta, construir internamente ou compor camadas próprias sobre modelos e componentes de terceiros. A terceira via domina os casos corporativos por equilibrar diferenciação e custo.

NO CONSELHO

construa onde há diferenciação competitiva e dado proprietário; compre onde é commodity. O erro caro é o inverso.

Termos que enganam o conselho

Seis expressões de uso corrente cujo significado se deslocou o suficiente para induzir decisões erradas.

"O modelo aprende com o nosso uso."

Na maioria dos produtos corporativos, não aprende. O modelo é estático; o que muda é o contexto recuperado ou a memória armazenada. Aprendizado real exige novo treino ou ajuste, com custo e ciclo próprios.

NO CONSELHO

pergunte se há treino, ajuste, memória ou nada disso — são quatro arquiteturas diferentes com quatro perfis de risco.

"É IA proprietária nossa."

Frequentemente significa uma camada de interface sobre um modelo de terceiro. Não há demérito nisso, mas muda inteiramente a análise de dependência e de valuation.

NO CONSELHO

pergunte qual modelo está por baixo e o que acontece se o preço dele dobrar ou o acesso for restringido.

"Precisão de 95%."

Número sem significado fora do contexto: precisão em qual tarefa, medida sobre qual conjunto, com qual distribuição de erro? Noventa e cinco por cento pode ser excelente ou inaceitável dependendo do custo do 5% restante.

NO CONSELHO

peça a matriz de erro e o custo unitário do falso positivo e do falso negativo.

"Copiloto."

Sugere assistência supervisionada, mas é usado hoje para descrever desde autocompletar até agentes com permissão de escrita em sistemas de produção.

NO CONSELHO

ignore o rótulo, pergunte pelas permissões efetivas.

"Automação de ponta a ponta."

Quase sempre significa automação da parte fácil, com exceções voltando ao humano — o que é legítimo, desde que dimensionado.

NO CONSELHO

peça a proporção de casos que caem na exceção e quem os absorve. É onde o ganho prometido costuma se dissolver.

"Vamos esperar a tecnologia amadurecer."

Postura defensável para escolha de fornecedor, indefensável para preparo organizacional. Inventário de dados, governança, curadoria e capacitação levam mais tempo que a adoção de qualquer ferramenta.

NO CONSELHO

separe as duas decisões. Adiar a compra é estratégia; adiar o preparo é atraso composto.

Doze perguntas para levar à próxima reunião

- 01 Qual problema de negócio este sistema resolve, e como saberemos que resolveu?
- 02 Qual modelo está por baixo, de quem, sob qual jurisdição?
- 03 Que dados nossos saem da nossa infraestrutura, e sob qual base legal?
- 04 Existe um conjunto de avaliação sobre os nossos casos? Quem é o dono dele?
- 05 Qual o custo por transação de negócio, comparado ao processo atual?
- 06 Onde está o ponto de parada humano, e qual gatilho o aciona?
- 07 Se o agente errar, o que é reversível e o que não é?
- 08 Quem é notificado quando o fornecedor troca a versão do modelo?
- 09 Qual o custo e o prazo para migrar para outro fornecedor?
- 10 Este sistema está no nosso inventário de IA, com classificação de risco?
- 11 Quem consegue reconstruir esta solução se a pessoa que a montou sair?
- 12 O que decidimos deliberadamente **não** automatizar, e por quê?

Nota de fontes e método

Os verbetes foram construídos a partir de três camadas: a literatura técnica e os relatórios setoriais de referência do período 2025–2026; o acompanhamento do processo regulatório brasileiro e europeu, incluindo o PL 2338 e o EU AI Act com seu pacote de simplificação; e a observação direta de projetos de IA em operação em saúde, jurídico, construção e varejo.

Números de mercado citados são projeções de terceiros e devem ser tratados como ordem de grandeza, não como dado auditado. A seção 8 descreve o estado regulatório vigente na data desta edição e é a que envelhece mais rápido — confirme a situação de tramitação antes de usar qualquer verbete como base para decisão jurídica. Este documento não constitui aconselhamento jurídico.

O vocabulário de IA se reorganiza em ciclos curtos. Esta é a edição 2026.2; correções, discordâncias e sugestões de novos verbetes são bem-vindas e alimentam a edição seguinte.

