TOTVS

GUIA DE IA TOTVS

2026

Outubro/2026

IA em evolução: do primeiro contato à operação em escala

Você está acessando a 3ª edição do Guia de IA da TOTVS. Os fundamentos continuam os mesmos, mas o cenário de mercado avançou de forma relevante, e este material foi totalmente revisado para acompanhar esse movimento.

O que era novidade e experimento nas edições anteriores agora é visto como infraestrutura e decisão de negócio, e o vocabulário que descrevia o futuro da IA nas empresas passou a descrever o presente. Atualmente, temos Agentes autônomos em produção, protocolos de interoperabilidade como padrão de mercado, custo por token como variável financeira de projeto e modelos especializados rodando em infraestrutura própria.

E esta edição acompanha esse movimento. Você vai encontrar conceitos das edições anteriores resumidos onde já são parte do repertório e aprofundados onde evoluíram, além de temas que simplesmente não tinham peso no vocabulário corporativo há pouco mais de 2 anos.

Para aproveitar melhor o conteúdo deste guia, sugerimos que acesse a edição de 2025, pois ela fornece toda a base conceitual que possibilita compreender na íntegra esta versão atualizada.

Sumário

GUIA DE IA TOTVS 2026

1 **IA Generativa:** novas abordagens e o que importa agora — **Pág. 4**

1.1 O que é? **4**
1.2 Modelos de IA Generativa **4**
1.2.1 LLMs, PLN e IA Generativa **4**
1.3 SLMs (Small Language Models) e a eficiência da IA **5**
1.4 A engenharia por trás das respostas da IA **6**
1.5 Alucinação e os modelos de raciocínio (Inference-time compute) **6**
1.6 Ética, risco e governança: o que a IA exige das empresas hoje **8**

2 Como o tema de **dados e IA evoluiu** na TOTVS? — **Pág. 9**

2.1 Os habilitadores de IA: o pré-requisito estrutural **9**
2.2 IA escalável, segura e com propósito específico **10**
2.3 Foundation LYNN **11**

3 **Agentes de IA:** arquitetura, integração e governança — **Pág. 12**

3.1 O que são Agentes de IA? **12**
3.2 Engenharia de Contexto **12**
3.3 Protocolos de interoperabilidade: MCP, Tool Use e Conectores **13**
3.4 Orquestração multiagente **13**
3.5 Segurança em sistemas agênticos: guardrails, prompt injection e riscos **13**
3.6 Human-in-the-loop **14**

4 **Checklists de governança** por tipo de IA — **Pág. 15**

- IA Generativa (assistentes e geração de conteúdo) **16**
- Modelos Preditivos (Machine Learning) **16**
- Agentes de IA **17**
- Soluções com Dados Sensíveis **17**

5 **Glossário** — **Pág. 18**

1

IA Generativa:

novas abordagens e o que importa agora

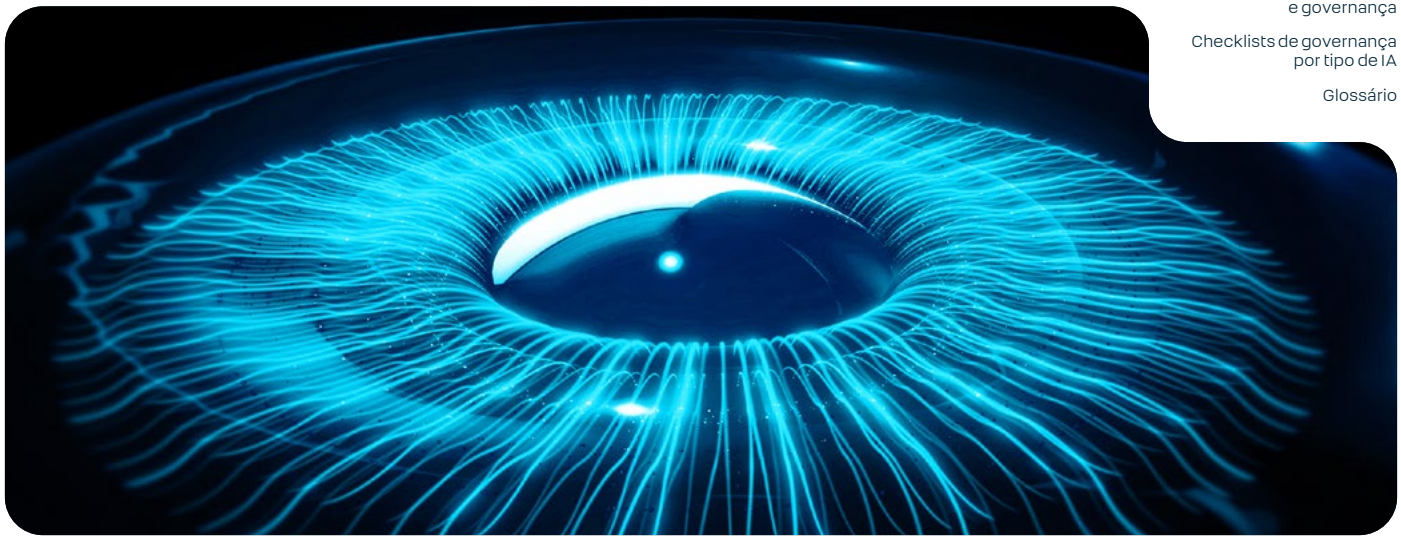
IA Generativa: novas abordagens e o que importa agora

Como o tema de dados e IA evoluiu na TOTVS?

Agentes de IA: arquitetura, integração e governança

Checklists de governança por tipo de IA

Glossário



1.1 O que é ?

IA Generativa é a capacidade de modelos de IA avaliarem e manipularem probabilidades para criarem novos conteúdos coerentes com um contexto. A popularidade do tema nos últimos anos se explica pela combinação de capacidade criativa inédita, avanços técnicos de modelos específicos, aplicações comerciais atrativas, acessibilidade crescente das plataformas e debate público sobre ética, autoria e criatividade.



1.2 Modelos de IA Generativa

1.2.1 LLMs, PLN e IA Generativa

Os modelos de LLM (Modelo de Linguagem de Grande Escala ou Large Language Models) são a intersecção entre o Processamento de Linguagem Natural (PLN) e a IA Generativa. O PLN é a camada do entendimento, pois interpreta o que foi dito, em que contexto e com qual intenção para viabilizar tarefas como reconhecimento de fala, análise de sentimentos, extração de informações, tradução e sumarização. Já a IA Generativa é a camada da criação.

A partir desse entendimento, produz conteúdos novos em diversos formatos, como texto, vídeo, imagem, código etc., gerando recombinações criativas do conhecimento adquirido no treinamento e coerentes com o contexto da solicitação.

Um LLM em funcionamento opera as duas camadas simultaneamente, e é essa combinação que torna a interação com esses modelos tão natural. Vale observar, porém, que LLMs são por definição modelos generalistas, treinados em volumes massivos de dados para cobrir praticamente qualquer assunto. Essa amplitude gera o custo de mais infraestrutura, mais processamento por requisição e, em escala, custos operacionais relevantes.

1.3 SLMs (Small Language Models) e a eficiência da IA

Boa parte da atenção do mercado recai sobre os LLMs, com bilhões de parâmetros e treinados para lidar com praticamente qualquer assunto. Mas nem todo caso de uso precisa de um modelo "generalista" gigantesco. É aqui que entram os SLMs (Small Language Models), ou modelos de linguagem de pequeno porte.

SLMs são modelos de linguagem mais compactos, treinados ou ajustados para domínios e tarefas mais específicas. Por terem menos parâmetros, exigem menos poder computacional para operar, o que traz três vantagens diretas para o ambiente corporativo:

- **Custo e eficiência:** um SLM consome significativamente menos recursos de processamento por requisição do que um LLM de grande porte. Em escala, com milhares de chamadas diárias, essa diferença tem impacto determinante no custo total de uma aplicação de IA.

- **Especialização:** quando o caso de uso é delimitado (por exemplo, classificar documentos fiscais ou navegar em um módulo específico do ERP), um modelo menor e especializado pode entregar precisão equivalente ou superior à de um modelo gigante e genérico, com uma fração do custo.



- **Inferência local e soberania de dados:** por serem mais leves, os SLMs podem ser executados em infraestrutura própria (on-premise) ou em nuvem dedicada, sem que os dados precisem trafegar para serviços externos de terceiros. Isso é especialmente relevante para empresas que lidam com informações sensíveis, sujeitas a requisitos de privacidade, compliance ou soberania de dados.

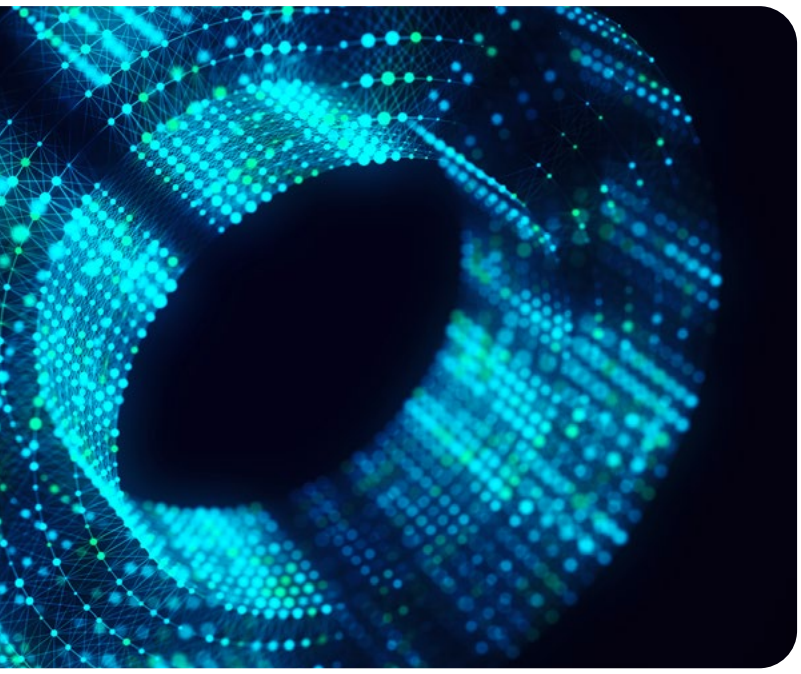
Na prática, o futuro da IA corporativa não é uma escolha entre "LLM ou SLM", mas a combinação inteligente dos dois: modelos grandes para tarefas abertas e complexas, e modelos pequenos e especializados para rotinas delimitadas, de alto volume e que exigem mais controle sobre dados e custos. Essa lógica de escolher o modelo certo para cada caso de uso, em vez de depender de um único fornecedor ou de um único modelo, o chamado agnosticismo de modelos, é justamente o que sustenta uma estratégia de IA escalável, soberana e economicamente sustentável. É esse princípio que a TOTVS pratica no LYNN, como veremos na seção 3.

1.4 A engenharia por trás das respostas da IA

Todo texto processado por um modelo de PLN passa primeiro pela ***Tokenização,*** onde o conteúdo é fragmentado em unidades menores chamadas tokens, que podem ser palavras inteiras, partes de palavras ou símbolos de pontuação. Em seguida, cada token é convertido em um vetor numérico, processo conhecido como ***Embedding.*** Esses vetores permitem que a rede neural processe linguagem. Em aplicações de busca semântica e RAG, embeddings também podem ser armazenados em bancos vetoriais.

O que torna os *embeddings* poderosos são as informações que os vetores carregam sobre o significado das palavras e suas relações entre si, aproximando ou distanciando os resultados dos temas e contextos relacionados.

Para processar esse volume de cálculos simultâneos, especialmente as multiplicações de matrizes envolvidas no processamento de tokens em modelos Transformer, entra o papel das GPUs. Originalmente desenvolvidas para processamento gráfico, as GPUs se tornaram o hardware central da IA Generativa por sua capacidade de executar milhares de operações em paralelo, reduzindo significativamente o tempo de treinamento e inferência dos modelos.



Janela de Contexto

É o volume máximo de tokens que um modelo consegue considerar em uma única interação, ou seja, a "memória de curto prazo" da IA. Quanto maior a janela, mais histórico, documentos ou instruções o modelo consegue manter ativos ao gerar uma resposta. Isso é determinante para Agentes que lidam com conversas longas, múltiplos documentos ou instruções complexas: uma janela insuficiente força o modelo a "esquecer" partes relevantes do contexto ao longo da interação.

Token Economics

Cada requisição a um modelo de IA tem um custo associado ao volume de tokens processados, tanto na entrada quanto na saída. Em aplicações de baixo volume, essa variável passa despercebida. Em escala corporativa, com milhares ou milhões de chamadas por mês, o custo por token se torna uma variável de decisão tão relevante quanto a escolha do próprio modelo. Projetar uma solução de IA atualmente exige estimar esse custo desde a etapa de concepção, não como detalhe técnico de implementação, mas como critério de viabilidade do projeto.

1.5 Alucinação e os modelos de raciocínio *(Inference-time compute)*

Alucinação é a geração de informações falsas, inventadas ou inconsistentes com o contexto ou as fontes relevantes, apresentadas como verdadeiras. O fato decorre da forma como o modelo opera (correlação estatística) e é influenciado por fatores como qualidade dos dados de treinamento, viés e complexidade da tarefa.

Este permanece como um dos principais desafios da IA Generativa. Mas o campo avançou, e hoje existem abordagens mais sofisticadas para reduzi-lo:

- **Ancoragem (*Grounding*):**
  vincular as respostas do modelo a fontes confiáveis e específicas, em vez de deixá-lo operar apenas com o conhecimento internalizado durante o treinamento. Se um Agente precisa responder sobre uma política interna da empresa, ele consulta o documento oficial, e não infere uma resposta a partir de padrões genéricos aprendidos.

- **RAG (*Retrieval-Augmented Generation*):**
  uma extensão do *grounding,* o qual o modelo, antes de responder, consulta ativamente uma base de dados vetorial ou documental para recuperar os trechos mais relevantes sobre o tema da pergunta e usa esse conteúdo como contexto para construir a resposta. O resultado é mais preciso e fundamentado do que uma resposta gerada "de memória".

- **Modelos de raciocínio (*reasoning models*) com computação em tempo de inferência (*inference-time compute*):**
  em vez de gerar a resposta de forma imediata, palavra após palavra, esses modelos alocam tempo e capacidade computacional adicionais no momento da resposta para percorrer etapas internas de raciocínio antes de entregar o resultado ao usuário.

  Para ilustrar este conceito, podemos pensar que um modelo convencional responde como alguém que fala sem parar. Já um modelo de raciocínio responde como alguém que para, pensa em voz alta, confere cada passo e só então responde.

  Essa cadeia de pensamento interna, chamada de ***chain of thought,*** valida a lógica de cada etapa antes de avançar para a próxima. O custo disso é um tempo de resposta um pouco maior e mais recursos computacionais por requisição.
  O benefício é uma redução relevante de alucinações em tarefas complexas como análise de contratos, validação de dados financeiros, raciocínio jurídico e diagnósticos técnicos.



Importante:

nenhuma dessas três abordagens elimina o problema completamente. A alucinação continua sendo um risco não resolvido, e por isso a supervisão humana em decisões de maior impacto segue sendo inegociável e diversas vezes exigido pelas normativas globais, que requerem que sistemas automatizados ofereçam explicações claras, acessíveis e não técnicas sobre os critérios utilizados permitindo a revisão humana de determinadas decisões.

1.6 Ética, risco e governança: o que a IA exige das empresas hoje

Atualmente, os dilemas éticos da IA são decisões que empresas, legisladores e desenvolvedores já estão sendo obrigados a tomar como parte da rotina de trabalho. Em especial, três temas ganharam relevância recentemente:

- **Responsabilidade em sistemas agênticos:** com Agentes de IA tomando decisões operacionais em tempo real ao aprovar transações, enviar comunicações institucionais e alterar registros em sistemas críticos, a pergunta "quem responde por uma ação errada?" se torna cada vez mais relevante.

  Definir previamente os limites de autonomia de cada Agente e os pontos de supervisão humana obrigatória, além de boa prática de governança, é exigência para uma gestão de risco responsável e em conformidade com os principais frameworks de mercado.

- **Direitos autorais e uso de dados para treinamento:** o uso de conteúdo protegido para treinar modelos de IA gerou processos judiciais relevantes contra grandes desenvolvedores nos últimos anos, trazendo mais clareza e mais risco sobre o que é permitido. Empresas que desenvolvem ou adaptam modelos precisam documentar a origem dos dados e garantir que o processo esteja em conformidade com a legislação aplicável.

- **Desinformação, deepfakes e integridade institucional:** o que antes era descrito como risco genérico de "manipulação de opiniões" tem hoje forma definida. Vídeos sintéticos, clonagem de voz e conteúdo gerado em escala para influenciar decisões compõem um cenário desafiador, pois representam um risco que afeta diretamente empresas que precisam proteger a imagem de suas lideranças e a credibilidade de suas comunicações. Como consequência, torna-se indispensável a construção de políticas internas claras sobre o que pode e o que não pode ser produzido com IA generativa. Neste sentido, existem normas sendo desenvolvidas já exigindo que sistemas de IA que criam conteúdo sintético (como deepfakes) marquem suas saídas como geradas artificialmente, mesmo que estes sistemas e conteúdos não tenham intenção de enganar. Ou seja, o conteúdo que se pareça ou soe como uma pessoa real deve ser rotulado, mesmo sem indivíduo real retratado, no intuito de trazer uma maior transparência para o público.

  Na TOTVS, esse posicionamento já está em vigor. Desde 2024, a empresa conta com uma Política de Uso de Inteligência Artificial que circula internamente e orienta as ações de todos os colaboradores. A política está estruturada em quatro pilares: Compliance e Segurança, com compromisso de aderência às normas de segurança da informação e cibersegurança; Uso Responsável, garantindo respeito aos direitos fundamentais e à propriedade intelectual; Transparência, assegurando comunicação clara e em conformidade com a legislação; e Inovação e Estratégia, incentivando a exploração de novas possibilidades dentro de limites éticos e estratégicos definidos.

IA Generativa: novas abordagens e o que importa agora

Como o tema de dados e IA evoluiu na TOTVS?

Agentes de IA: arquitetura, integração e governança

Checklists de governança por tipo de IA

Glossário

2

Como o tema de Dados e IA evoluiu na TOTVS?



Quando as primeiras edições deste guia descreveram a abordagem da TOTVS em relação a dados e IA, o foco estava em estabelecer as bases culturais dessa jornada: porque colocamos o caso de uso em primeiro lugar, a importância de uma plataforma de dados sólida como pré-requisito e a necessidade de uma governança que funcionasse sem travar a velocidade do negócio.

Esses princípios continuam válidos e sendo praticados e evoluídos, mas o estágio mudou. A IA já se tornou parte concreta dos produtos e soluções TOTVS, portanto o direcionamento evoluiu de "como convencer a organização a usar IA" para "como escalar, governar e extrair ROI das iniciativas de IA que já estão em produção".

Essa evolução reflete como institucionalizamos o conceito de IH + IA (Inteligência Humana + Inteligência Artificial) como parte dos nossos valores organizacionais. Acreditamos que a tecnologia é sempre um meio a serviço das pessoas, e não o contrário. Ao combinarmos o julgamento crítico, a empatia e o aprendizado contínuo da nossa inteligência humana com o poder de escala e eficiência da inteligência artificial, impulsionamos decisões mais inteligentes, fortalecemos o relacionamento com nossos clientes e garantimos o uso responsável da tecnologia para transformar o mercado e crescermos juntos.

A seguir, confira nossas principais novidades:

2.1 Os habilitadores de IA: o pré-requisito estrutural

A adoção de IA em escala não começa no modelo, e sim na prontidão do ambiente do cliente. Por isso, a TOTVS estrutura sua estratégia sobre três **habilitadores de IA:** (i) a **universalização do T-Cloud,** que garante o ambiente seguro, soberano e padronizado onde a IA opera; (ii) o **domínio das APIs,** que permite que Agentes e aplicações acessem e executem ações nos sistemas de gestão; e (iii) a **organização das bases de dados** e a atualização dos sistemas, que asseguram dados íntegros, acessíveis e em prontidão.

Esses habilitadores são a tradução prática do princípio que orienta este guia desde suas primeiras edições: **sem dados íntegros, acessíveis e em prontidão, não tem IA.** É sobre essa fundação que o LYNN e os Agentes de IA da TOTVS entregam valor.

2.2 IA escalável, segura e com propósito específico

Facilitamos para todos os produtos e serviços TOTVS os melhores recursos de Inteligência Artificial disponíveis, de forma escalável, governada e com propósito de negócio. Essa orquestração se materializa no LYNN, o foundation de inteligência artificial da TOTVS, que é organizado em três pilares:



- **Agent Builder:**
  o ambiente para conceber, construir e orquestrar Agentes de IA conectados aos sistemas TOTVS (ERP, CRM e demais aplicações). É aqui que uma oportunidade identificada ou uma "dor" de negócio se transforma em um Agente funcional, capaz de raciocinar, acessar dados, executar ações e operar de forma autônoma ou semi-autônoma dentro de um fluxo real da operação.

- **Enterprise Layer:**
  a camada que conecta os Agentes ao contexto corporativo de forma confiável e segura. É o que assegura que a IA não opere "no vácuo", mas sim ancorada nos dados, nos sistemas e nas regras de negócio de cada cliente, viabilizando acesso aos dados certos, integração com as aplicações certas e o grounding necessário para respostas precisas e fundamentadas na realidade da empresa. Ao operar, essa camada também registra metadados de uso e de resultado de cada aplicação, um ativo proprietário que retroalimenta a evolução dos Agentes e a inteligência da plataforma.

- **Governance:**
  o pilar que garante que tudo opere com diligência, rastreabilidade e controle. Abrange a governança no uso dos modelos, o controle de consumo e custo das requisições, a observabilidade das ações executadas pelos Agentes, proteções para os MCPs e APIs, rate-limits para as aplicações e os guardrails que asseguram conformidade, segurança da informação e uso responsável da IA, sem sacrificar a agilidade do negócio.
  Os metadados de uso e resultado gerados pelas camadas do LYNN são proprietários e estratégicos: são eles que permitem medir eficácia, otimizar custo por tarefa e melhorar continuamente cada aplicação em produção.

2.3 Foundation LYNN

O LYNN é o primeiro foundation de Inteligência Artificial B2B do mercado brasileiro, lançado pela TOTVS em fevereiro de 2026. Ele viabiliza, de ponta a ponta, a jornada de uma aplicação de IA, da concepção do Agente à sua operação governada em escala. Em vez de cada área partir do zero, recorrendo a infraestrutura crua de mercado, o LYNN entrega um caminho padronizado, seguro e reaproveitável para colocar IA em produção.

Na prática, ela facilita as principais etapas de viabilização de qualquer aplicação:

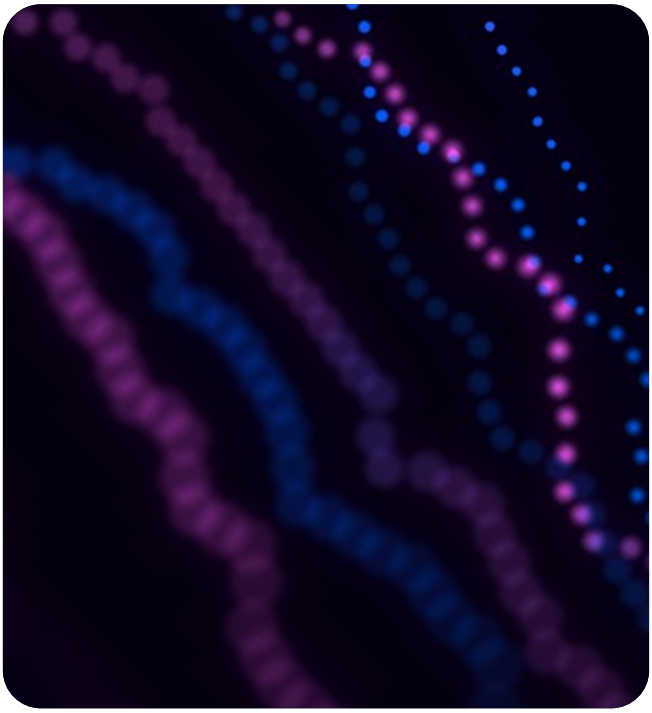
- **Construir**
  Agentes e soluções de IA conectados ao negócio, sobre o Agent Builder.
- **Acessar**
  Os principais modelos de IA disponíveis no mercado, de forma abstraída e intercambiável.
- **Conectar**
  Da IA aos dados e sistemas corporativos por meio da Enterprise Layer, garantindo contexto e precisão.
- **Governar**
  O uso dos modelos com rastreabilidade, segurança, governança e controle de consumo orçamentário a partir do custo das requisições dos diferentes modelos.



Com o LYNN, a TOTVS viabiliza a IA para aprimorar a eficiência operacional e ampliar a capacidade de tomada de decisão de nossos clientes, sempre fiéis ao nosso princípio: sem dados íntegros, acessíveis e em prontidão, não tem IA.

O LYNN, porém, não é apenas infraestrutura interna: é a base sobre a qual a TOTVS constrói sua oferta de Agentes de IA para os clientes. Esses Agentes executam tarefas diretamente ligadas aos processos de gestão, do back-office ao relacionamento com o cliente, e materializam o modelo TaaS (Task as a Service), em que a cobrança é baseada na volumetria de tarefas executadas pelos Agentes.

Na prática, isso significa que o cliente contrata resultado de trabalho realizado, e não apenas licenças de software. É uma expansão relevante do papel da tecnologia na operação das empresas.



IA Generativa: novas abordagens e o que importa agora

Como o tema de dados e IA evoluiu na TOTVS?

Agentes de IA: arquitetura, integração e governança

Checklists de governança por tipo de IA

Glossário

3

Agentes de IA:

arquitetura, integração e governança



3.1 O que são Agentes de IA?

Quando falamos em IA no dia a dia das empresas, muita gente já teve contato com a IA conversacional, na maioria das vezes representada por assistentes virtuais que respondem perguntas de forma natural. Os Agentes de IA representam uma evolução dessa dinâmica, pois são sistemas que atuam de forma autônoma ou semi-autônoma, tomando decisões e executando tarefas sem depender de um comando a cada passo.

Vale observar que nem todo Agente tem interface visível. Muitos operam nos bastidores, integrados diretamente a sistemas como ERP e CRM, conversando com outros Agentes e com as ferramentas da empresa sem que ninguém precise acioná-los a cada ciclo.

3.2 Engenharia de Contexto

Enquanto o *Prompt Engineering* trata de como formular a instrução certa para um modelo, a Engenharia de Contexto trata de como organizar, priorizar e entregar ao Agente exatamente as informações que ele precisa, no momento em que precisa. Um Agente moderno pode ter acesso a históricos de conversa, documentos internos, dados de sistemas conectados, resultados de ferramentas que já executou e instruções de governança.

Organizar esse volume de informação de forma que o modelo use o que é relevante, sem se perder ou saturar sua janela de contexto, é hoje uma disciplina própria dentro da construção de Agentes, e isso é tão importante quanto a escolha do modelo em si. Um Agente mal projetado nessa dimensão tende a ignorar informações relevantes ou, no extremo oposto, tratar dados irrelevantes como se fossem instruções, o que aumenta o risco de respostas equivocadas.

3.3 Protocolos de interoperabilidade: MCP, Tool Use e Conectores

Um dos avanços mais relevantes do último ano na área de Agentes de IA está em como os modelos se conectam ao mundo ao redor deles.

Tool Use é a capacidade de um Agente acionar ferramentas externas durante seu raciocínio (consultar uma API, rodar uma busca, executar um cálculo, atualizar um registro em um sistema) em vez de depender apenas do conhecimento que já tem internalizado.

Um dos maiores desafios do Tool Use sempre foi a integração. Cada sistema (CRM, ERP, planilha, banco de dados) exigia uma conexão sob medida entre o Agente e aquela ferramenta específica. Isso mudou com a chegada do **Model Context Protocol (MCP),** um padrão aberto que uniformiza a forma como aplicações expõem dados e ferramentas para Agentes de IA. Em vez de construir uma integração exclusiva para cada combinação de Agente e sistema, o MCP funciona como uma linguagem comum. Qualquer Agente compatível consegue se conectar a qualquer sistema que fale esse mesmo protocolo, o que reduz drasticamente o custo e o tempo de integração.

Complementar ao MCP, o protocolo **A2A (Agent-to-Agent)** resolve o problema de como Agentes de fornecedores e plataformas diferentes conversam entre si e coordenam tarefas em conjunto, em vez de operar como ilhas isoladas.

Esses protocolos são a infraestrutura que viabiliza os Agentes trabalharem em conjunto.



3.4 Orquestração multiagente

Nos primeiros passos da adoção de Agentes de IA nas empresas, o padrão era um Agente por tarefa. Um Agente de atendimento, um Agente de estoque, um Agente de relatórios, cada um operando isoladamente. Porém, esse modelo mudou. Hoje, a arquitetura predominante nas implementações corporativas mais maduras é a de **múltiplos Agentes especializados** que coordenam entre si para resolver um processo de ponta a ponta.

Um exemplo prático: um Agente de estoque identifica um padrão de baixa reposição, aciona um Agente de compras, que negocia com Agentes de fornecedores e formaliza o pedido, que por sua vez aciona um Agente de logística para agendar a entrega. Nenhum Agente sozinho executa o processo inteiro, ou seja, cada um contribui com sua especialização para um resultado compartilhado, coordenado por um Agente orquestrador ou por um fluxo de regras definido.

Essa mudança de arquitetura é hoje o padrão de referência para quem constrói soluções de IA em escala corporativa, e a razão pela qual protocolos como MCP e A2A se tornaram infraestrutura essencial.

Na TOTVS, essa arquitetura é praticada por meio do **Agent Builder** do LYNN, onde Agentes especializados são concebidos, conectados aos sistemas de gestão e orquestrados dentro de fluxos reais de operação.

3.5 Segurança em sistemas agênticos: guardrails, prompt injection e riscos

À medida que Agentes de IA passam a executar ações reais, a superfície de risco muda de natureza. Um erro de um chatbot gera uma resposta errada; já um erro de um Agente com permissão de ação pode gerar uma transação incorreta, um dado alterado indevidamente ou uma comunicação enviada por engano.

Conhecer os principais riscos é pré-requisito para qualquer projeto de Agente autônomo:

- **Prompt Injection (direto e indireto):** técnica em que instruções maliciosas são inseridas no fluxo de trabalho do Agente para manipular seu comportamento. No modo direto, a instrução vem do próprio usuário. No modo indireto, mais difícil de detectar, a instrução maliciosa está escondida em um conteúdo externo que o Agente consulta, como um documento, e-mail ou página web, e é executada sem que o usuário tenha ciência.

- **Goal Hijacking (Sequestro de objetivo):** quando um Agente é induzido, por meio de instruções manipuladas, a desviar do objetivo original para o qual foi projetado.

- **Contaminação de memória:** quando informações incorretas ou maliciosas são inseridas na memória de longo prazo de um Agente, influenciando decisões futuras de forma persistente.

Para mitigar esses riscos, a prática corporativa recomendada envolve **guardrails de Agentes,** ou seja, regras, limites de permissão e mecanismos de verificação que restringem o que um Agente pode fazer, mesmo que seu raciocínio interno seja manipulado. Por exemplo, exigir confirmação para ações irreversíveis, limitar o escopo de sistemas que o Agente pode acessar e monitorar continuamente as ações executadas.

Também é importante que todos os desenvolvedores e times de produto estejam atentos às recomendações técnicas da OWASP, tais como a **OWASP Top 10 for LLM Applications (2025)** que identifica os principais riscos de segurança em sistemas de IA e a **OWASP Top 10 for Agentic Applications,** que cobre riscos específicos de Agentes, tais como: tomada de decisão autônoma, memória persistente, acesso a ferramentas/APIs e coordenação multiagente etc.

3.6 Human-in-the-loop

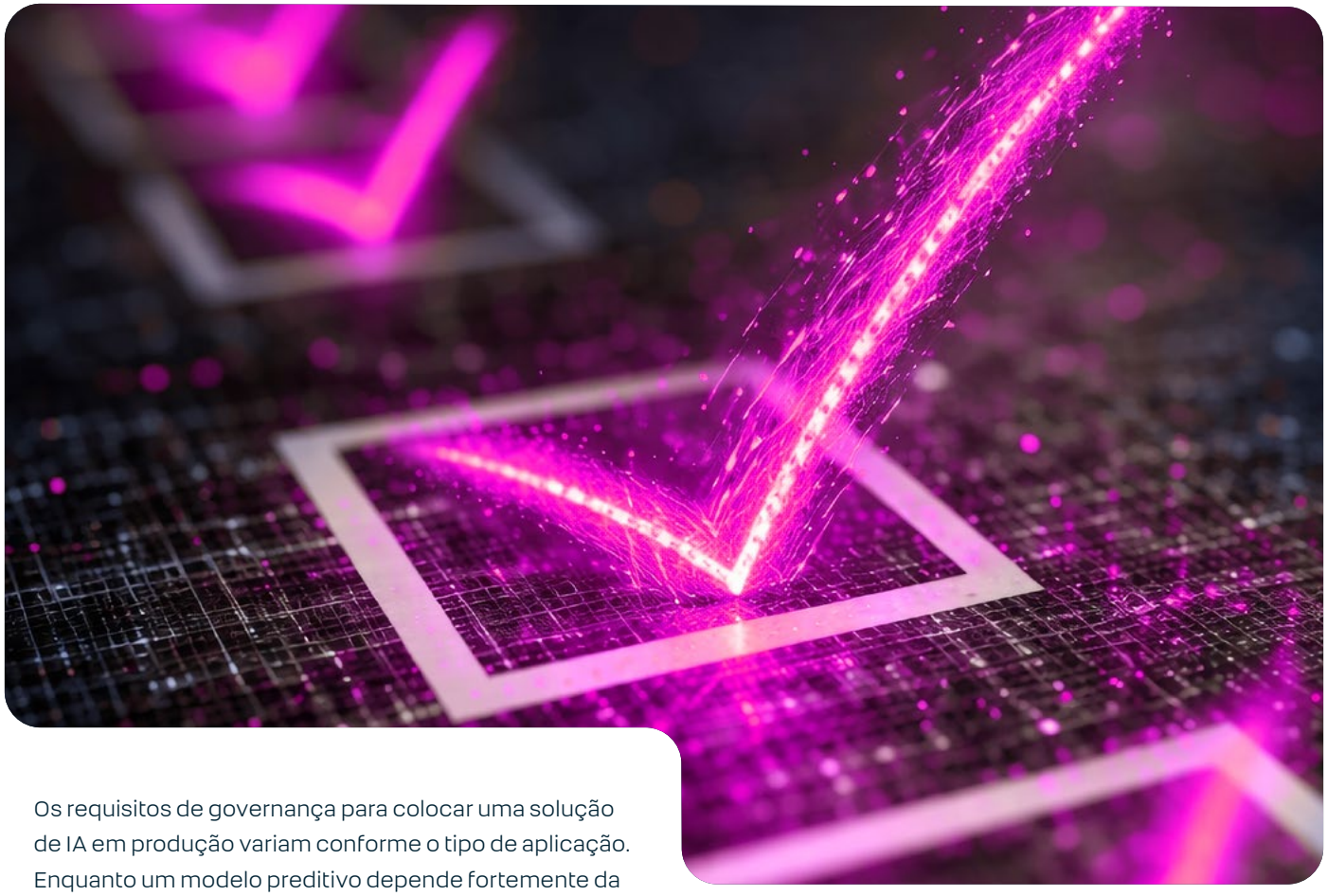
Human-in-the-loop (HITL) é o princípio de manter supervisão humana ativa em pontos estratégicos da atuação de um Agente de IA, em vez de conceder autonomia irrestrita. Na prática, isso significa definir com clareza em que momentos um Agente pode agir sozinho e em que momentos ele precisa de validação humana antes de prosseguir. Um Agente pode, por exemplo, consultar e sugerir de forma totalmente autônoma, mas precisará de aprovação explícita antes de enviar um e-mail em nome da empresa ou de aprovar uma transação financeira.

Definir esse ponto de corte é o que torna possível conceder autonomia real para atividades de baixo risco, com segurança, enquanto mantém controle humano onde o risco de um erro é mais alto. Um Human-in-the-loop bem projetado é o que permite escalar o uso de Agentes de IA com confiança, em vez de restringir seu uso por precaução alarmista.



4

Checklists de governança por tipo de IA



Os requisitos de governança para colocar uma solução de IA em produção variam conforme o tipo de aplicação. Enquanto um modelo preditivo depende fortemente da qualidade dos dados de treinamento.

Os checklists a seguir reúnem os principais elementos que devem ser considerados durante o desenvolvimento de cada categoria de solução.

Importante:
um Agente de IA exige mecanismos adicionais de controle, autorização e monitoramento.

IA Generativa (assistentes e geração de conteúdo)

Antes da disponibilização da solução, verifique se:

- ◯ O objetivo de negócio está claramente definido;
- ◯ A finalidade e as condições de uso foram estabelecidas e divulgadas;
- ◯ Existe política sobre quais informações podem ou não ser utilizadas nos prompts;
- ◯ Dados confidenciais ou pessoais possuem tratamento adequado antes do envio ao modelo;
- ◯ As respostas são ancoradas em fontes confiáveis quando necessário (Grounding ou RAG);
- ◯ Existe revisão humana para conteúdos de maior impacto;
- ◯ Há registro das interações (logs) para auditoria;
- ◯ Foram definidos indicadores de qualidade das respostas;
- ◯ Existe plano para tratamento de alucinações e respostas incorretas;
- ◯ O consumo de tokens foi estimado e monitorado;
- ◯ O modelo utilizado é compatível com os requisitos de privacidade, custo e desempenho do projeto.

Modelos Preditivos (Machine Learning)

Antes da implantação do modelo, verifique se:

- ◯ O problema do negócio está claramente definido;
- ◯ Existe documentação da origem dos dados utilizados;
- ◯ Os dados passaram por validação de qualidade;
- ◯ Foram avaliados possíveis vieses da base de treinamento;
- ◯ Existem métricas objetivas de desempenho;
- ◯ O modelo possui critérios de revalidação periódica;
- ◯ Há monitoramento para identificar degradação (*model drift*);
- ◯ Existe estratégia de versionamento do modelo;
- ◯ Os responsáveis pela manutenção estão definidos;
- ◯ Há documentação das limitações conhecidas do modelo.

Agentes de IA

Antes de permitir atuação autônoma do Agente, verifique se:

- ◯ O escopo de atuação do Agente está claramente delimitado;
- ◯ As ferramentas (Tool Use) disponíveis ao Agente foram autorizadas;
- ◯ Existem limites de permissão para execução de ações;
- ◯ Foram definidos limites de taxa de requisição (Rate Limiting) para controle de custos e prevenção de execuções descontroladas;
- ◯ Foram implementados guardrails;
- ◯ Existe proteção contra Prompt Injection;
- ◯ O Agente utiliza informações atualizadas por meio de Grounding ou RAG quando necessário;
- ◯ Foram definidos pontos obrigatórios de Human-in-the-loop;
- ◯ Existe rastreabilidade completa das ações executadas;
- ◯ Há monitoramento contínuo do comportamento do Agente;
- ◯ Existe mecanismo para interromper imediatamente sua execução em caso de comportamento inesperado.

Soluções com Dados Sensíveis

Independentemente da tecnologia utilizada, verifique se:

- ◯ Foi identificado se há tratamento de dados pessoais;
- ◯ Apenas os dados estritamente necessários são utilizados;
- ◯ Existe base legal ou autorização adequada para utilização dos dados;
- ◯ Dados sensíveis possuem controles adicionais de acesso;
- ◯ As informações armazenadas seguem as políticas de retenção da organização;
- ◯ Existe registro dos sistemas que acessam esses dados;
- ◯ O compartilhamento entre aplicações está documentado;
- ◯ Os riscos à privacidade foram avaliados;
- ◯ Existem mecanismos para atendimento às solicitações dos titulares quando aplicável;
- ◯ A solução está aderente às políticas internas de segurança e privacidade.

5

Glossário

A

A2A (Agent-to-Agent)
Protocolo que permite que Agentes de IA de diferentes plataformas e fornecedores se comuniquem e coordenem tarefas entre si.

AGI (Artificial General Intelligence)
Inteligência Artificial Geral – Conceito de IA capaz de aprender e realizar uma ampla variedade de tarefas cognitivas em nível comparável ou superior ao humano. Seus critérios de caracterização não têm definição única; atuar em múltiplos contextos, isoladamente, não caracteriza AGI.

Alucinação
Fenômeno em que um modelo de IA gera informações sem base real nos dados de entrada. Decorre do funcionamento estatístico dos modelos e é mitigada, mas não eliminada, por técnicas como Grounding, RAG e modelos de raciocínio.

ANI (Artificial Narrow Intelligence)
Inteligência artificial Estreita – Categoria de sistemas com capacidades restritas a uma tarefa ou conjunto delimitado de tarefas. O termo descreve o escopo das capacidades, e não uma técnica específica de contextualização.

API (Application Programming Interface)
Conjunto de regras, protocolos e rotinas que permite a comunicação e a troca de dados segura entre diferentes sistemas e softwares de forma automatizada. No contexto de IA, as APIs funcionam como as "pontes" que conectam os modelos de IA (ou Agentes) aos bancos de dados e sistemas de gestão, permitindo que a IA consulte informações, execute comandos e realize tarefas operacionais em tempo real na infraestrutura da empresa.

ASI (Artificial Superintelligence)
Superinteligência artificial – Conceito hipotético de IA com capacidades cognitivas amplamente superiores às humanas em uma vasta gama de tarefas.

C

Chain of Thought
Processo interno utilizado por modelos de raciocínio para decompor um problema em etapas antes de produzir uma resposta.

E

Embedding
Representação numérica de palavras, frases ou documentos que preserva relações de significado e permite que modelos de IA processem e comparem informações.

E

Engenharia de Contexto
Disciplina de organizar, priorizar e entregar a um Agente de IA exatamente as informações de que ele precisa, no momento em que precisa, sem saturar sua janela de contexto.

G

Grounding
Técnica que vincula as respostas da IA a fontes específicas e confiáveis, reduzindo o risco de informações incorretas.

Guardrails
Conjunto de regras, restrições e mecanismos de controle que limitam o comportamento de Agentes de IA para garantir segurança, conformidade e uso adequado.

H

Habilitadores de IA
Os três pré-requisitos estruturais da estratégia de IA da TOTVS: universalização do T-Cloud, domínio das APIs e organização das bases de dados com atualização dos sistemas.

Human-in-the-loop (HITL)
Abordagem que mantém a supervisão humana em etapas críticas da atuação de sistemas ou Agentes de IA.

I

Inference-time Compute
Capacidade de um modelo utilizar recursos computacionais adicionais durante a geração da resposta para realizar um raciocínio mais elaborado.

J

Janela de Contexto (Context Window)
Quantidade máxima de informações que um modelo consegue considerar simultaneamente durante uma interação.

L

LYNN
Primeiro foundation de Inteligência Artificial B2B do mercado brasileiro, lançado pela TOTVS. Viabiliza a construção, integração e governança de aplicações e Agentes de IA, organizados nos pilares Agent Builder, Enterprise Layer e Governance.

M

MCP (Model Context Protocol)
Protocolo aberto que padroniza a forma como Agentes de IA acessam ferramentas, sistemas e fontes de dados externas.

O

Orquestração Multiagente
Arquitetura em que diversos Agentes especializados trabalham de forma coordenada para executar um processo completo, compartilhando informações e responsabilidades.

P

Prompt Injection
Técnica utilizada para manipular o comportamento de um Agente de IA por meio da inserção de instruções maliciosas em prompts ou conteúdos consultados pelo próprio Agente.

R

RAG (Retrieval-Augmented Generation)
Técnica em que o modelo, antes de responder, consulta uma base de dados vetorial ou documental para recuperar os trechos mais relevantes e usá-los como contexto, produzindo respostas mais precisas e fundamentadas.

Reasoning Models (Modelos de Raciocínio)
Modelos de IA projetados para realizar etapas adicionais de raciocínio antes de produzir uma resposta, aumentando a precisão em tarefas complexas.

S

SLM (Small Language Model)
Modelo de linguagem de menor porte, geralmente especializado em domínios específicos, que demanda em geral menos recursos computacionais e apresenta menor custo operacional.

T

TaaS (Task as a Service)
Tarefa como serviço. Modelo de cobrança baseado na volumetria de tarefas executadas por meio de Agentes de IA. É o modelo comercial dos Agentes construídos sobre o LYNN.

Token Economics
Análise do impacto financeiro do consumo de tokens em aplicações de IA, considerando custos de processamento, escalabilidade e operação.

Tool Use
Capacidade de um Agente de IA utilizar ferramentas externas, como APIs, bancos de dados e sistemas corporativos, durante sua execução.