Por Jefferson Souza · Edição Nº 07 · 10 de agosto de 2026 · As opiniões expressas são pessoais e não representam as de qualquer empregador.

ABSTRACT

The multi-agent complexity of large banks is not a prerequisite for AI in credit and collections. The article argues that deterministic decision graphs, well-structured retrieval-augmented generation (RAG) and secure access to models via API deliver the same analytical rigor with a fraction of the infrastructure, and that maturity is measured by the safety of logical transitions, not by the number of agents.

ARTIGO · ARQUITETURA DE IA EM CRÉDITO & COBRANÇA

IA e o mito da complexidade

Reproduzir a arquitetura multiagente de grandes bancos e empresas de tecnologia em uma operação de crédito ou cobrança média é caro e desnecessário. Grafos de decisão determinísticos, RAG bem estruturado e acesso a modelos via API com blindagem de dados entregam o mesmo rigor analítico com uma fração da complexidade.

Resian

Todos os direitos reservados 2026

O mercado financeiro vive uma corrida acelerada para colocar agentes de inteligência artificial em produção. Fintechs, credenciadoras e instituições de crédito integram modelos de linguagem diretamente em suas réguas de comunicação e motores de aprovação. O problema está no fato de que o desenho de infraestrutura dessas iniciativas frequentemente repete um erro clássico de engenharia: acoplar a lógica de negócio diretamente às APIs públicas dos provedores de modelos, sem camadas intermediárias de controle.

Ao olhar para bancos e empresas de tecnologia que operam IA em escala massiva, como o Itaú com sua plataforma proprietária Iara e a arquitetura multiagente da Inteligência Itaú, o iFood com a GenPlat, ou os motores de decisão automatizados do Nubank, vemos estruturas gigantescas processando dezenas de milhões de chamadas e trilhões de tokens por mês. Tentar reproduzir essa mesma complexidade e volume de infraestrutura em operações médias de crédito ou boutiques de investimento é um erro estratégico elementar: o resultado é aumento expressivo no custo de crédito e cobrança, além de latência impraticável.

A maturidade de uma operação não se mede pelo número de agentes que conversam livremente entre si, mas pela segurança de suas transições lógicas e pelo controle estrito de seus custos operacionais. A vantagem analítica está no equilíbrio: em vez de multiagentes redundantes, desenha-se uma arquitetura com poucos agentes altamente focados em temas específicos nos quais o julgamento e o tratamento especializado da informação são de fato necessários, como conduzir uma quebra de objeções em uma mesa de negociação ou validar automaticamente a conformidade documental de uma cessão.

01

O conceito de grafos de decisão e o poder dos dados históricos

Diferente de um assistente de conversação comum, que gera respostas de forma livre e probabilística, um agente voltado para o mercado de risco e recuperação de ativos deve operar sob uma estrutura determinística de estados. É aí que entra a engenharia de grafos de decisão. O comportamento do agente é representado por nós e arestas: cada nó representa um estado lógico rígido, como verificar a elegibilidade a um desconto ou validar o canal de cobrança, e as arestas representam as transições possíveis sob regras explícitas de negócio.

Essa dinâmica impede as chamadas alucinações de modelo. O agente cognitivo não tem autonomia para decidir os limites de uma oferta de desconto ou a concessão de novas faixas de crédito. Ele consulta dinamicamente o motor de regras de negócio (BRE) e transita o tomador apenas pelos caminhos pré-autorizados no grafo de estados. A IA atua de forma criativa apenas na interface, na empatia da redação, no tom de voz, na quebra de objeções e na personalização do atendimento, mantendo os parâmetros financeiros blindados.

O motor de transição de um grafo de alta performance é alimentado por dados históricos. Ao analisar o comportamento passado de carteiras reais, como a análise de safra, a velocidade de pagamento, as taxas de roll rate e o comportamento do FPD (First Payment Default), o sistema calibra as arestas do grafo. Se os dados históricos mostram que tomadores de determinado perfil cadastral têm taxa de perda severa ao ultrapassar trinta dias de atraso, o grafo direciona esse subgrupo para uma esteira de negociação especializada com maior urgência, reduzindo o custo total por conta trabalhada (cost of collection).



O grafo acima ilustra o princípio central da arquitetura: o nó cognitivo (em dourado) é isolado entre dois pontos de controle determinístico, a consulta ao motor de regras e a validação da proposta. A IA nunca decide sozinha o desfecho financeiro de uma interação, ela decide apenas como conduzir a conversa dentro dos limites que o BRE já autorizou.

02

A aplicabilidade prática do RAG (Retrieval-Augmented Generation)

Em cenários operacionais complexos, os agentes de IA precisam de informação que vai além do conhecimento geral do modelo. O RAG (Retrieval-Augmented Generation) é a tecnologia que conecta o poder cognitivo do LLM ao repositório documental e ao histórico de transações da operação, de forma ágil e sem a necessidade de treinamentos custosos de novos modelos. Na concessão de crédito e na cobrança especializada, o RAG atua principalmente em duas frentes.

01

Compliance e políticas de crédito

Ao receber dúvidas ou objeções sobre taxas, cláusulas ou regulamento do FIDC, o agente usa o RAG para buscar e citar trechos específicos da política de crédito ou do regulamento do fundo. A resposta é formulada de forma segura e embasada, mitigando risco de contencioso jurídico ou passivo de auditoria.

02

Memória de relacionamento do cliente

O RAG recupera logs históricos de conversas anteriores, promessas de pagamento quebradas e renegociações passadas para entender a jornada e os hábitos do cliente. Quando o agente entra em contato, ele traz um nível de contexto que humaniza a abordagem e aumenta consideravelmente a taxa de conversão das propostas.

03

Soberania de dados e segurança no acesso a modelos via API

A segurança da informação e a conformidade com o sigilo bancário e a LGPD são barreiras severas na implementação de projetos de inteligência artificial em crédito e cobrança. Enviar dados sensíveis, como CPF, saldo devedor e histórico transacional, para servidores de terceiros sem controle pode gerar exposições regulatórias graves.

Parte dessa soberania se resolve na camada de orquestração e memória: frameworks de código aberto como o LangGraph e bases de dados vetoriais como o PostgreSQL com a extensão pgvector permitem que a lógica do grafo, o histórico de conversas e os embeddings documentais fiquem hospedados integralmente em infraestrutura de cloud proprietária, seja on-premise virtual ou em nuvens privadas como AWS, Oracle Cloud ou Azure. A operação mantém o controle total sobre onde e como esses dados circulam antes de qualquer chamada externa.

> **O acesso a modelos de linguagem via API, hoje oferecido pelos principais provedores comerciais, é perfeitamente viável para as etapas cognitivas do grafo, desde que a arquitetura trate a chamada externa como uma fronteira de risco explícita, não como uma extensão transparente do sistema interno.**

Isso significa nunca enviar à API o CPF, o saldo devedor ou qualquer dado que permita reidentificação direta do tomador: o nó cognitivo recebe apenas o contexto anonimizado ou tokenizado necessário para gerar a resposta, e a reidentificação acontece depois, dentro do perímetro controlado pela operação. Contratos com cláusula de retenção zero, ausência de uso dos dados para treinamento do provedor e auditoria completa de

logs de chamada fecham essa blindagem. O ganho é direto: a operação evita o custo de infraestrutura de GPU e o trabalho contínuo de fine-tuning e versionamento de modelo próprio, mantendo o controle real sobre o que sai da sua rede, que é, na prática, onde mora o risco regulatório, não no fato de o modelo estar hospedado em nuvem de terceiros.

04

Pragmatismo como alavanca de valor

No mercado de crédito estruturado e recuperação de ativos, a inovação tecnológica só se justifica se vier acompanhada de resiliência analítica e rigor de custos (ROI sempre!). Reduzir a complexidade de múltiplos agentes conversando entre si em favor de um conjunto mais enxuto de poucos agentes dedicados e bem calibrados por grafos de decisão de base histórica é o caminho mais rápido para capturar resultados reais.

FONTES E REFERÊNCIAS

1. TI Inside. *Itaú amplia em 88% uso de IA generativa, acelera digitalização e reduz custo de processamento em 44% com avanço da nuvem.* Sobre a plataforma Iara e a arquitetura multiagente do banco. Maio de 2026. Disponível em: tiinside.com.br
2. Itaú Tech. *Desenvolvendo a Inteligência Itaú: arquitetura multiagentes com foco no cliente.* Medium. Outubro de 2025. Disponível em: medium.com/itautech
3. iFood Tech. *GenPlat: A Plataforma de IA Generativa do iFood.* Sobre volume de tokens e arquitetura de acesso a modelos. Maio de 2025. Disponível em: iFood Tech (Medium)
4. Mobile Time. *iFood quer chegar a 8 mil agentes de IA internos até março de 2026.* Outubro de 2025. Disponível em: mobiletime.com.br
5. LangChain. *LangGraph overview.* Documentação oficial do framework de orquestração de agentes com estado. Disponível em: docs.langchain.com
6. pgvector. *Extensão de banco de dados vetorial de código aberto para PostgreSQL.* Disponível em: github.com/pgvector

"Resiliência analítica forjada nas maiores instituições financeiras do Brasil."