SQUADRA digital | AI Full Cycle Platform genius Powered by SQUADRA

EBOOK

GESTÃO DE SISTEMAS AGÊNTICOS

Da checagem periódica ao monitoramento contínuo: uma disciplina de governança para orquestradores e agentes de inteligência artificial em produção na Plataforma Genius AI da Squadra.

QUALIDADE DE RESPOSTA | SAÚDE DE BASE SEMÂNTICA | EFICIÊNCIA DE CUSTO | ADEQUAÇÃO DE MODELO

Por Rafael Vieira

AI Product Manager na Squadra

Led by humans, ***powered by*** *Genius*

SUMÁRIO

A jornada da checagem periódica ao monitoramento contínuo

INTRODUÇÃO

Um sistema agêntico nunca está pronto.

A maioria das equipes trata um agente de inteligência artificial como um software convencional: publica, valida em produção e revisita apenas quando algo quebra. Essa lógica falha porque um sistema agêntico não é estático.

Ele responde a uma base de conhecimento que muda, a um comportamento de usuário que evolui e a um custo de inferência que se acumula silenciosamente a cada interação.

A qualidade de um sistema agêntico não se mede no dia do lançamento, e sim na disciplina com que ele é revisado depois.

Este documento reúne a metodologia que usamos para isso, organizada em torno de quatro pilares e sustentada por uma arquitetura de observabilidade construída em camadas independentes de qualquer fornecedor específico.

Essa abordagem é fundamentada nos princípios do Design Integral da Squadra, garantindo uma visão sistêmica, holística e equilibrada de toda a operação.

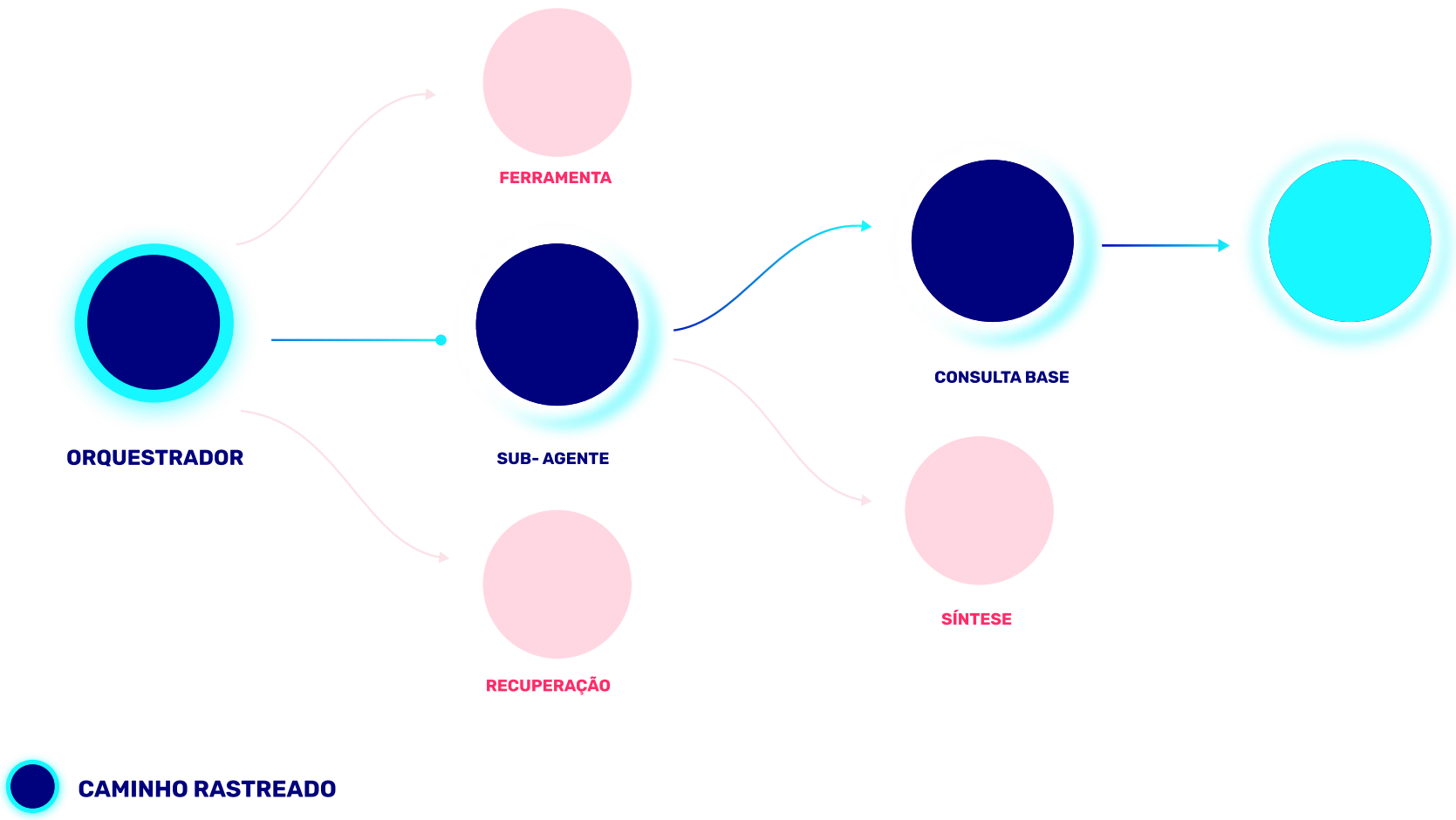
A rotina que descrevo começou como uma auditoria mensal e hoje é uma prática semanal apoiada por painéis em tempo real, uma mudança que os próprios dados exigiram: problemas de contexto, revisão de custo, análise e métricas sobre modelos que evidênciam riscos muito antes de virarem reclamação de usuário.

O PONTO DE PARTIDA

Um agente entrando em loop, um pico de custo ou uma queda de relevância não esperam pelo fim do mês para acontecer. A gestão precisa observar no mesmo ritmo em que o sistema opera.

INTRODUÇÃO

Um sistema agêntico nunca está pronto.



A execução real de um agente é uma árvore de decisões.

Governar significa enxergar qual caminho ele percorreu, e por quê.

Observar é manter conhecimento sob as ações por fluxo, qualidade das respostas, consumo de token, equivalência de modelos e escalabilidade.

OS QUATRO PILARES DA GESTÃO AGÊNTICA

Toda revisão de um sistema agêntico precisa cobrir quatro frentes ao mesmo tempo. Olhar apenas para uma delas produz uma falsa sensação de controle: um agente pode ter custo baixo e ainda assim alucinar, pode citar a fonte certa e ainda assim usar um modelo caro demais para uma tarefa simples.

1. QUALIDADE DA RESPOSTA

A resposta final resolve a intenção do usuário, mantém coerência com o histórico e evita alucinação, mesmo quando o agente encadeia múltiplas ferramentas ou etapas de raciocínio.

2. SAÚDE DA BASE SEMÂNTICA

O conteúdo recuperado por busca vetorial realmente sustenta a resposta gerada. Depende de como o conhecimento foi fragmentado, da sobreposição entre blocos e da qualidade do modelo de embedding, controle do overlap dos chunks, score de cada fragmento.

3. EFICIÊNCIA FINANCEIRA

O custo por interação é proporcional à complexidade real da tarefa, e não ao hábito de sempre acionar o modelo mais robusto disponível. A qualidade da resposta e o fluxo que leva a ela deve ser equilibrado. A previsão de escalar um agente em produção, sem medição e controle, pode comprometer todo o budget do mês em 2 dias.

4. ADEQUAÇÃO DO MODELO À ATIVIDADE

Cada etapa aciona o modelo de inteligência artificial certo para sua profundidade de raciocínio, reservando os mais avançados para as decisões que de fato precisam deles. Saber escalar o modelo certo para cada agente é fundamental para um fluxo agêntico.

Esses quatro pilares não competem entre si, eles se equilibram. Um agente ajustado para gastar pouco, mas que recupera contexto insuficiente, apenas transfere o custo do orçamento de infraestrutura para o custo de confiança do usuário. **O objetivo da governança agêntica é manter os quatro em equilíbrio simultâneo, e não otimizar um deles isoladamente.**

“

A checagem mensal é uma auditoria retrospectiva. O monitoramento contínuo é uma propriedade permanente do sistema.

Do Capítulo 02 • Da checagem periódica ao painel em tempo real

DA CHECAGEM PERIÓDICA AO PAINEL EM TEMPO REAL

A primeira versão dessa rotina era mensal: uma vez por mês, eu reunia amostras de conversas, revisava os documentos mais recuperados pela busca semântica, cruzava isso o modelo usado por cada agente do orquestrador no Genius, consumo de token, custos do período e decidia se algo precisava mudar em chunking, overlap, embedding, fluxo ou roteamento de modelo. Funcionava, mas tinha um problema estrutural, uma falha no meio do mês só aparecia na análise semanas depois de já ter afetado usuários reais e impactado diretamente a operação e o valor entregue x custo de token.

A rotina evoluiu para um **ciclo semanal apoiado por painéis em tempo real.** A diferença não é apenas de frequência, é de natureza. Os painéis mostram latência, custo, profundidade de interação e sinais de qualidade a cada momento, e a revisão semanal é o espaço em que essas leituras viram ação: fine-tuning nos agentes corretos, um ajuste fino de overlap, uma recalibração de roteamento entre modelos ou a criação de um novo teste de regressão. As vezes até a implementação de agentes observadores acionados periodicamente para monitorar os estágios.

NA PRÁTICA

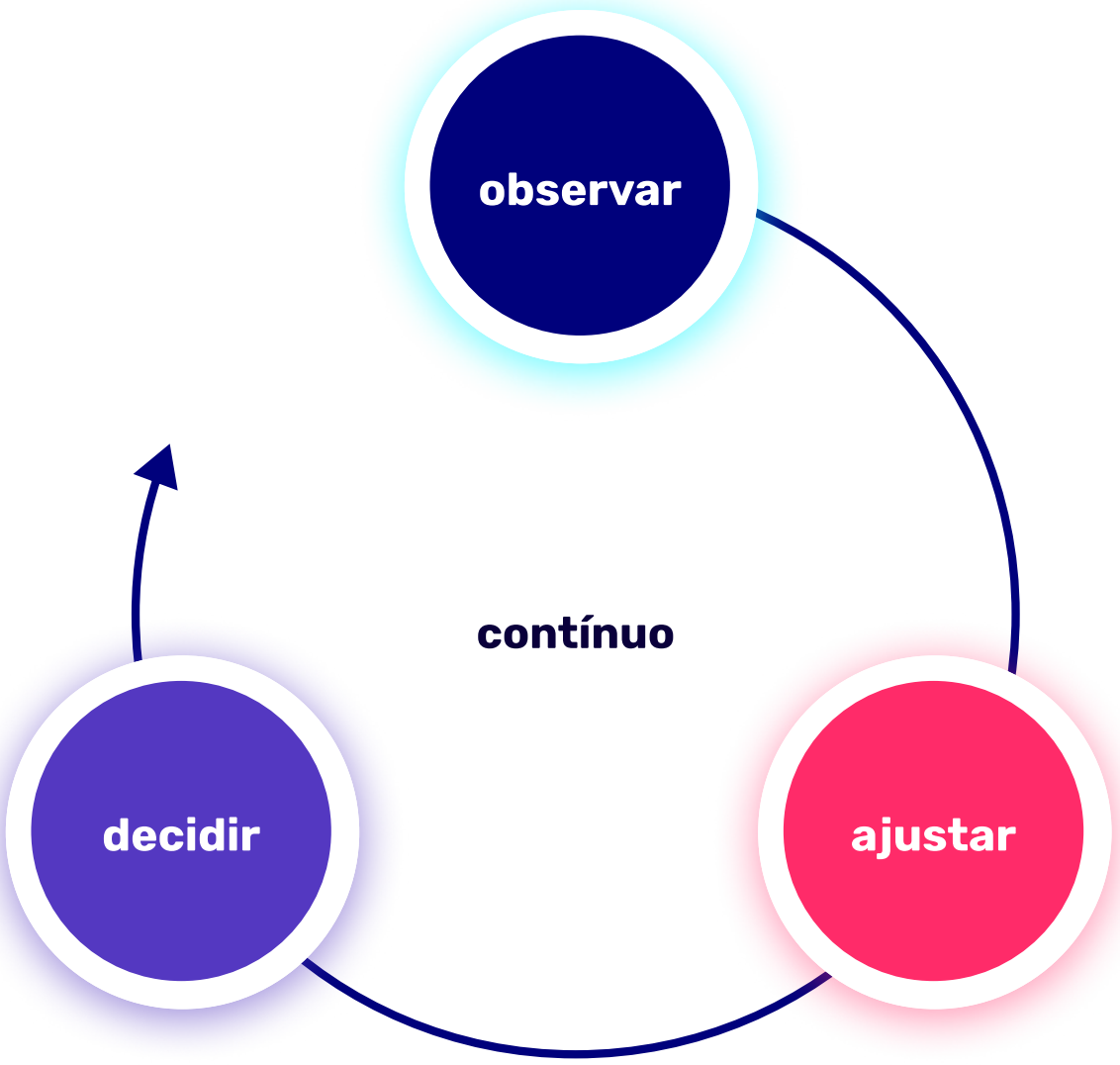
Nenhuma decisão de calibração espera pela reunião semanal para ser identificada. Ela espera apenas para ser validada e implementada com consistência, evitando ajustes reativos e isolados feitos no calor do momento.

Essa mudança também redefiniu o que vale a pena automatizar.

Erros críticos e picos de custo passaram a gerar alerta no instante em que ocorrem. A análise semanal ficou reservada para o que exige julgamento humano: interpretar tendência, decidir o trade-off entre custo e qualidade, e priorizar qual ajuste de base semântica traz mais retorno.

O acompanhamento da comunicação do usuário com a solução também é monitorado.

A estrutura, raciocínio, arquitetura e orquestração da ferramenta é padrão, o usuário é imprevisível.



Observar, decidir, ajustar: um ciclo que deixa de depender de uma data no calendário.

OBSERVABILIDADE EM TRÊS CAMADAS

Independentemente de qual conjunto de ferramentas sustenta a operação, a observabilidade de um sistema agêntico, como a operacionalizada de forma contínua pela Plataforma Genius, organiza-se bem quando pensada em camadas funcionais, cada uma respondendo a uma pergunta diferente sobre o comportamento do agente.

RASTREAMENTO

CAMADA DE EXECUÇÃO

Registra a árvore completa de decisões: quais ferramentas foram acionadas, em que ordem e em qual profundidade de raciocínio. Revela quando um agente busca contexto além do necessário para uma pergunta simples, ou entra em ciclos repetitivos entre etapas.

CUSTO

CAMADA DE MÉTRICAS

Consolida latência, volume de chamadas e custo de inferência por etapa, por agente e por tipo de atividade, com granularidade suficiente para identificar onde um modelo avançado está sendo usado em uma tarefa que um modelo leve resolveria.

QUALIDADE

CAMADA DE AVALIAÇÃO SEMÂNTICA

Mede matematicamente se o conteúdo recuperado sustenta a resposta final, combinando relevância do contexto com fidelidade da resposta ao que foi encontrado, expondo alucinação mesmo quando o texto gerado soa convincente.

Base • repositório vetorial + registro de conversas, unindo dado operacional e dado de auditoria em uma única fonte de verdade

As três camadas conversam entre si.

Um pico de custo na camada de métricas frequentemente tem origem em um comportamento visível na camada de rastreamento. E uma queda na avaliação semântica quase sempre aponta para um ajuste na forma como o conteúdo foi fragmentado e indexado.

CALIBRAÇÃO DA RECUPERAÇÃO NA PRÁTICA

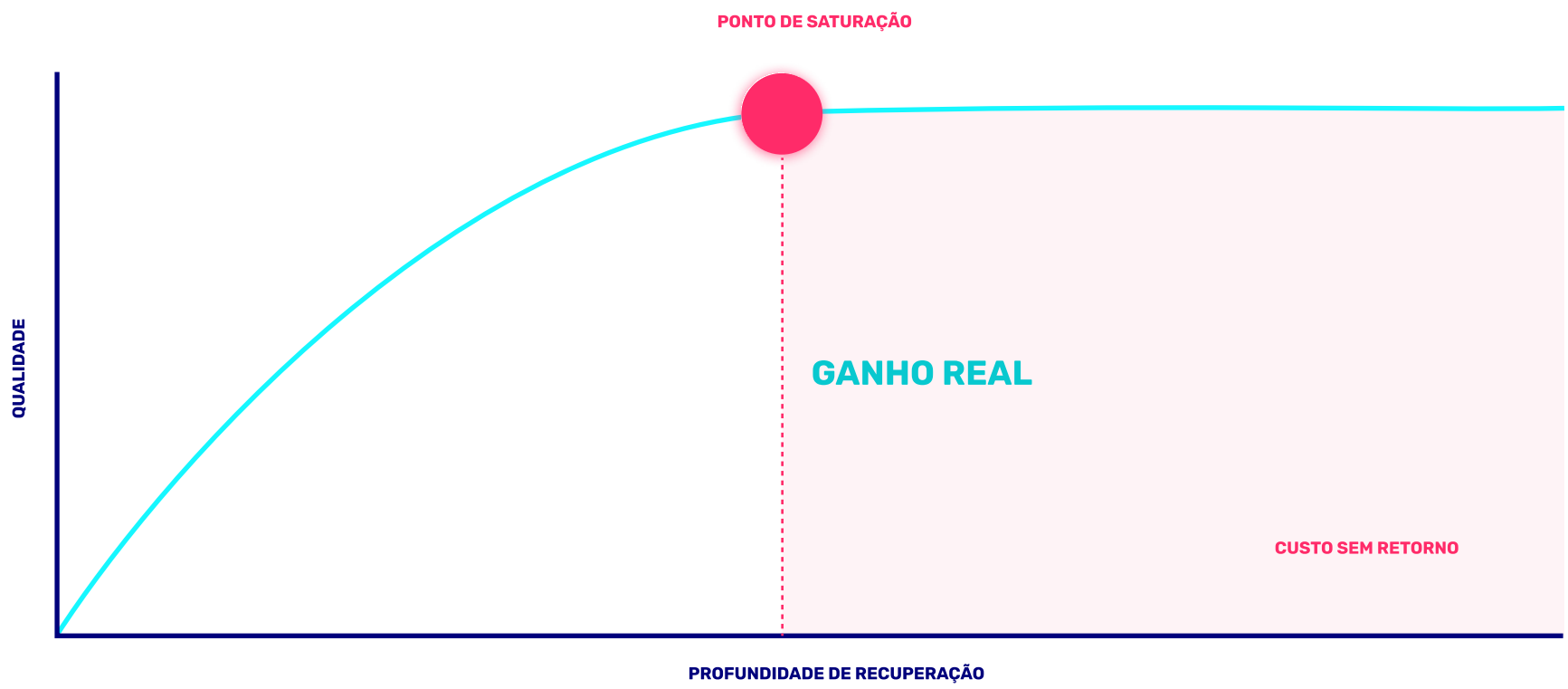
Um repositório vetorial informa qual bloco de conteúdo está mais próximo semanticamente da pergunta feita, mas nunca informa se esse bloco é suficiente para sustentar uma boa resposta. Essa lacuna se resolve em duas frentes.

Metadado rico em cada fragmento

Cada bloco indexado em um projeto no Genius carrega metadados sobre seu próprio tamanho, sua sobreposição com os vizinhos, sua origem documental e sua versão. Sem isso, quando a qualidade cai, não há como saber se o problema está no tamanho do chunk, na falta de sobreposição para preservar o fio da explicação, ou em uma fonte desatualizada. Com o metadado presente, a causa aparece nos próprios dados.

O ponto de saturação

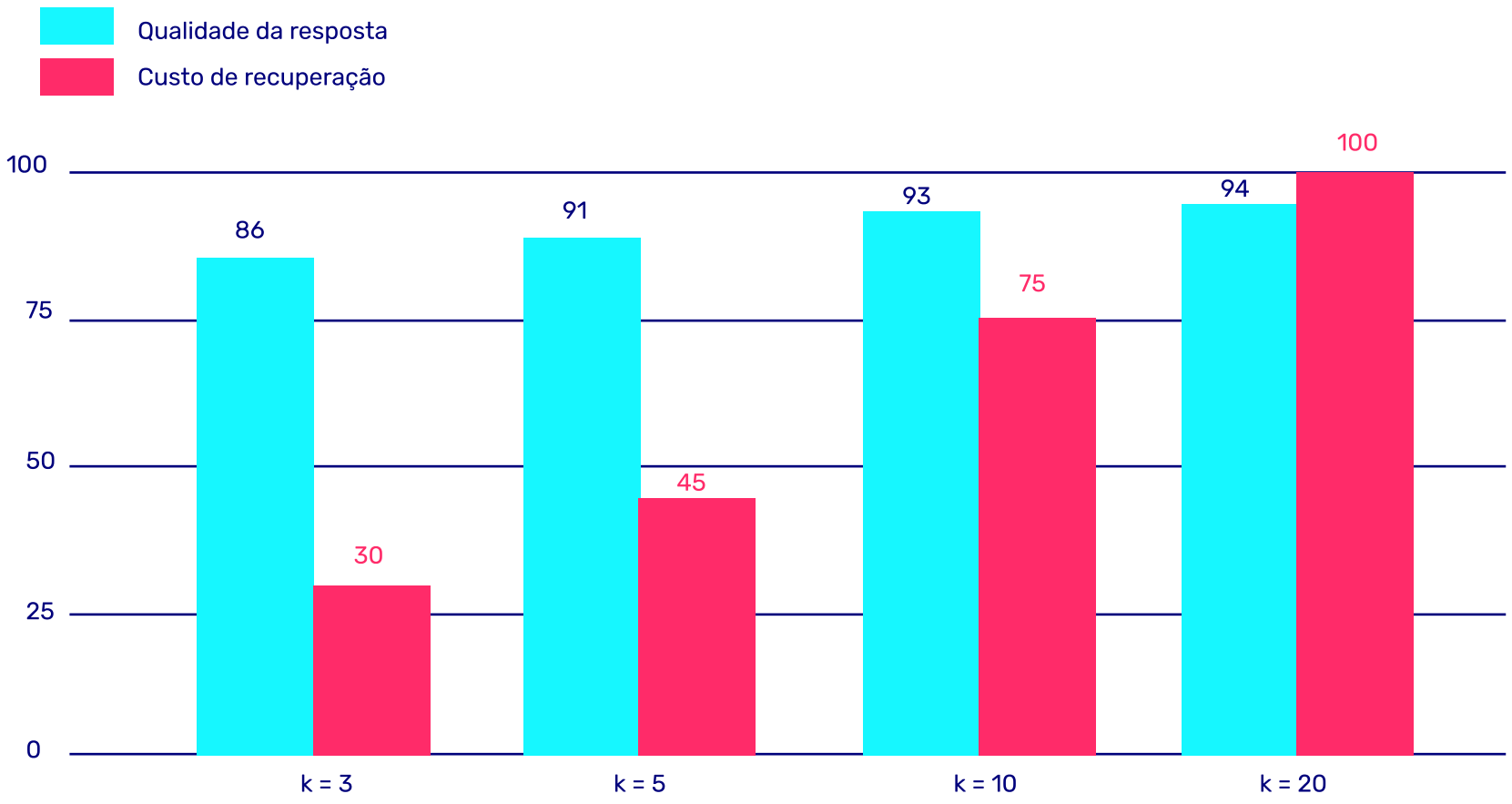
Existe um volume de contexto a partir do qual trazer mais blocos deixa de melhorar a resposta e passa apenas a aumentar custo. Encontrar esse ponto exige comparar, de forma sistemática, o ganho real de qualidade contra o custo de recuperação em diferentes profundidades de busca.



Depois do ponto de saturação, cada bloco extra recuperado adiciona custo sem retorno proporcional em qualidade.

Quando três blocos valem por dez

Num teste comparativo típico, o salto de qualidade entre recuperar três e cinco blocos é expressivo. Já o ganho entre dez e vinte é marginal, enquanto o custo praticamente dobra. Esse padrão de curva se repete em bases de conhecimento distintas.



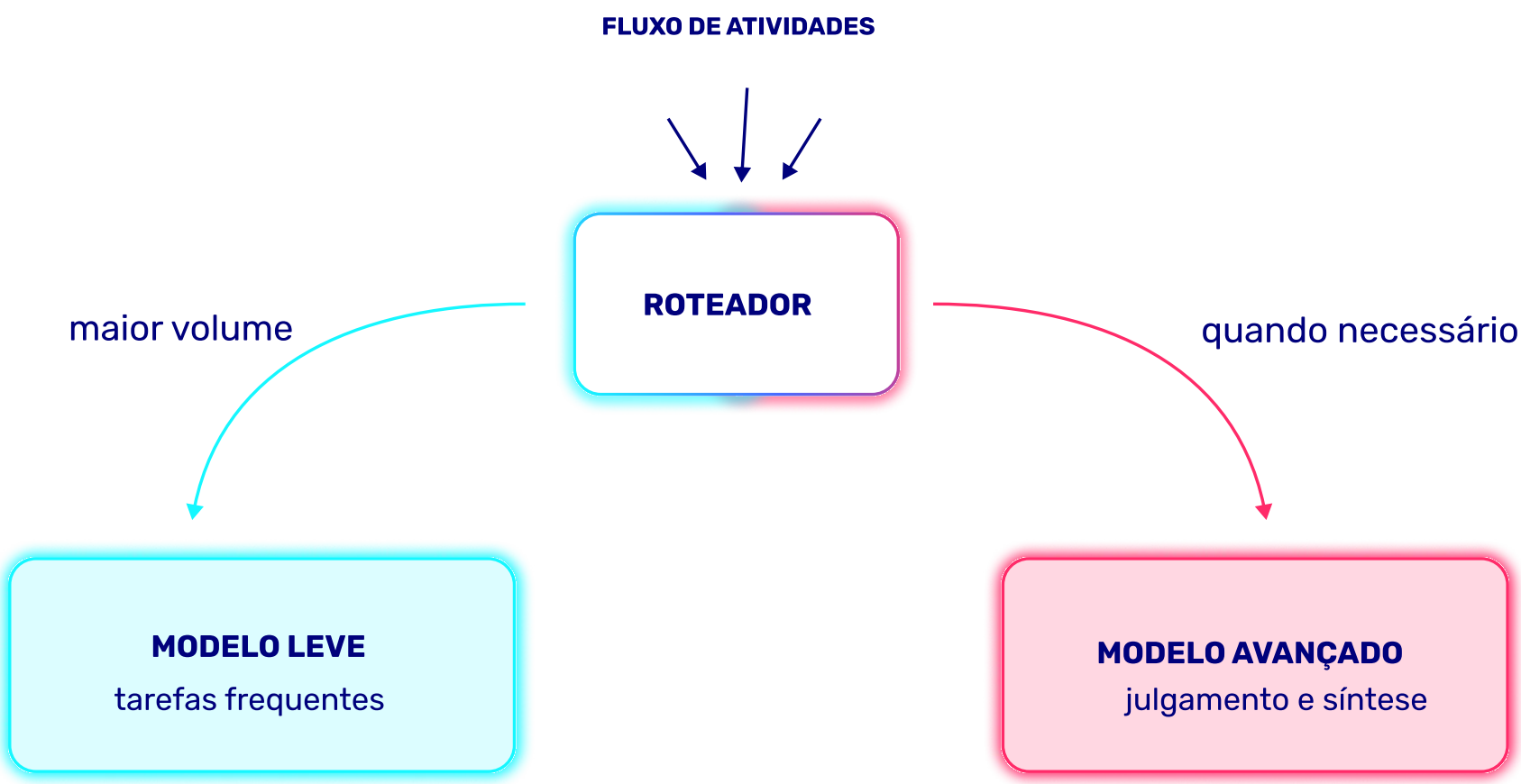
Qualidade da resposta e custo relativo por profundidade de recuperação (escala relativa, %).

LEITURA

Trazer três blocos bem estruturados, com bom overlap, frequentemente entrega a mesma qualidade que trazer dez, com uma fração do custo de recuperação. A calibração periódica existe para encontrar esse ponto antes que o custo suba sem retorno.

CUSTO POR MODELO, USO POR ATIVIDADE

O erro mais comum de alocação de custo não é usar um modelo caro, é usá-lo onde ele não precisava estar. Roteirizar a atividade para o modelo de inteligência artificial adequado à sua profundidade de raciocínio costuma gerar mais economia do que qualquer otimização de infraestrutura isolada. A capacidade de realizar esse roteamento inteligente e a calibração contínua é um dos pilares de eficiência integrados à Plataforma Genius. Cada agente tem uma identidade, papel e função bem definidos, isso permite medir modelo x desempenho x custo.



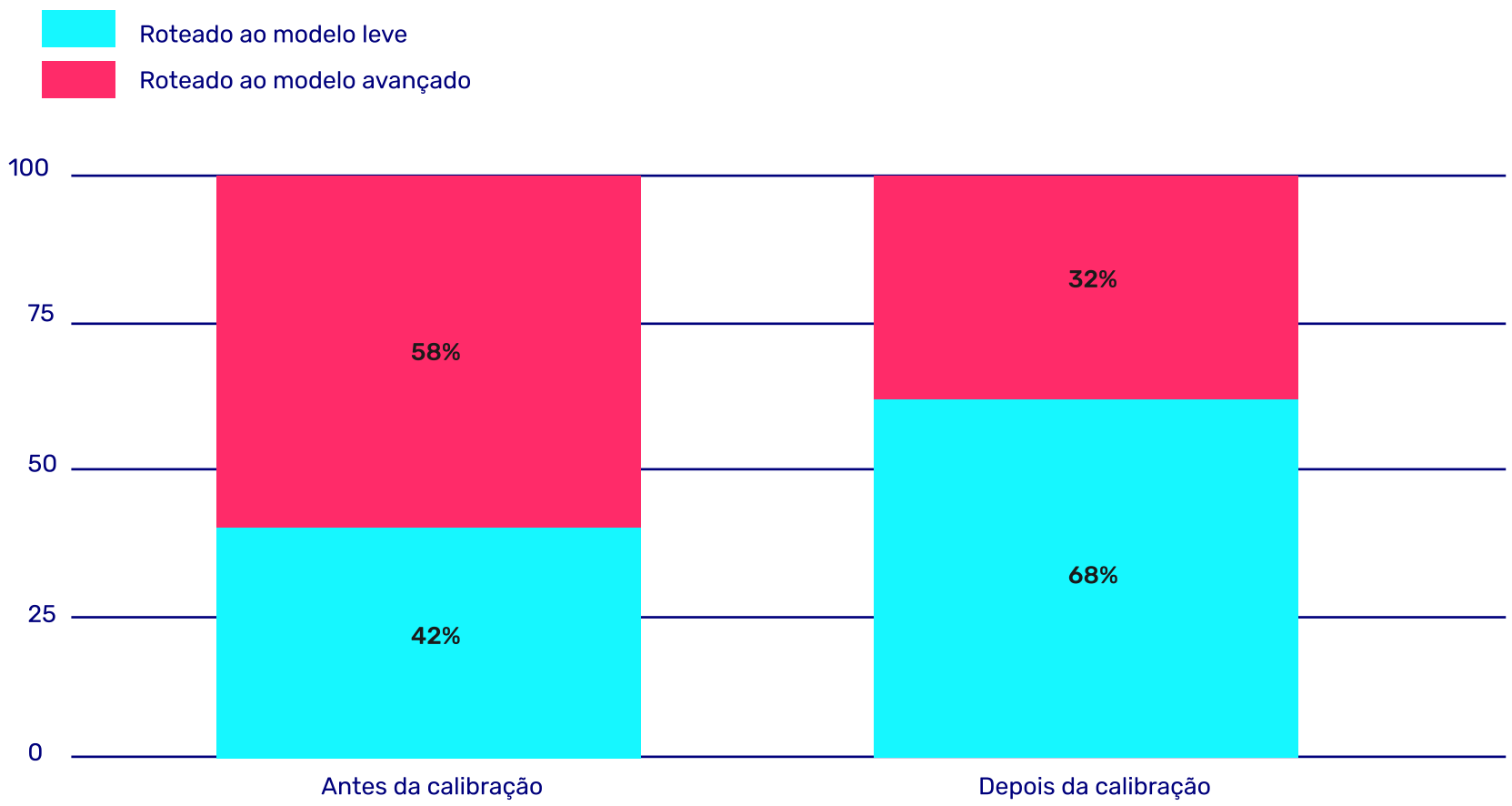
Tarefas frequentes e de baixa ambiguidade seguem ao modelo leve; o modelo avançado é reservado ao julgamento e à síntese.

A forma de enxergar isso é rotular cada chamada pelo tipo de atividade, não apenas pelo agente que a originou, e observar a distribuição de custo por esse rótulo. Classificação, extração estruturada e perguntas frequentes raramente exigem o modelo mais avançado. Decisões com ambiguidade real, síntese entre fontes ou julgamento de negócio, sim.

O custo cai, a qualidade permanece

Uma rodada de calibração de roteamento realoca as tarefas simples para o modelo leve e reserva o modelo avançado apenas para o que exige profundidade real. O efeito no volume de chamadas é direto.

Diferente do cenário de 2 anos atrás, hoje temos muitos modelos disponíveis, cada um com uma capacidade própria de execução, velocidade de análise, qualidade de entrega e custos por token. Isso é uma oportunidade de gerar mais, com menos.



Redistribuição do volume de chamadas após uma rodada de calibração de roteamento.

O TESTE DE UMA BOA CALIBRAÇÃO

A mudança não reduzia qualidade percebida pelo usuário porque as tarefas realocadas eram exatamente aquelas em que a profundidade adicional do modelo avançado não se traduzia em resposta melhor. O custo cai, e nenhum indicador de qualidade se move junto.

DO REATIVO AO CONTÍNUO

Painéis em tempo real implementados nos agentes do Genius resolvem a velocidade da observação, mas não substituem dois elementos que tornam a gestão de um sistema agêntico verdadeiramente madura sob a ótica do Design Integral.

Um conjunto fixo de perguntas de regressão.

Antes de qualquer mudança em chunking, overlap, embedding ou roteamento, esse conjunto de referência roda de novo e ele existe para responder as perguntas:

O ajuste que resolveu um problema não criou outro em outro ponto da base de conhecimento?

O "ambiente" sofrerá impacto?

Os "serviços" continuarão com seus propósitos de entrega?

Os "processos" precisarão de alterações?

A "arquitetura" agêntica suportará as mudanças?

Um canal simples de feedback do usuário final.

Nenhum painel técnico substitui o sinal mais direto de qualidade, que é o usuário indicando se aquela resposta resolveu o que ele precisava. Cruzado com rastreamento e avaliação semântica, é o que expõe a divergência entre o que a métrica considera boa resposta e o que o usuário considera útil.

Além do feedback direto, existe o feedback de observação, conforme citado anteriormente, onde usamos métricas para medir interação para uma resposta, reprocessamento para chegar na qualidade esperada, palavras reusadas e frequencia, entre outros.

O CICLO FECHADO

Alerta em tempo real, regressão automatizada e feedback do usuário fecham o ciclo. A observação deixa de depender de uma data no calendário e passa a ser uma propriedade permanente do sistema, com a revisão periódica reservada para decidir com contexto, não apenas detectar.

CONCLUSÃO

Um sistema agêntico não falha porque foi mal projetado no primeiro dia. Ele falha porque ninguém continuou perguntando, semana após semana, se ele ainda faz o que deveria.

Aqui na Squadra, os sistemas agênticos maduros não se diferenciam pelo modelo de inteligência artificial que usam, todos têm acesso às mesmas opções na plataforma Genius. O que temos acompanhado no mercado é que a diferença está em quem constrói uma disciplina de revisão capaz de enxergar, ao mesmo tempo, qualidade, saúde da base semântica, custo e adequação de modelo, e em quem transforma essa disciplina em rotina permanente em vez de checklist ocasional.

Com a Plataforma Genius, provamos na Squadra que a gestão técnica, operacional, semântica e financeira de agentes de inteligência artificial não precisa ser uma caixa-preta. No entanto, o grande desafio para o mercado permanece:

Você tem clareza sobre o comportamento real, o nível de controle e o retorno sobre o investimento dos seus agentes em produção neste exato momento?

Criar e publicar um agente inteligente tornou-se uma tarefa simples; mantê-lo seguro, calibrado e lucrativo semana após semana é onde o jogo real se decide.

A pergunta que fica para sua operação é: você está governando os seus sistemas agênticos, ou apenas financiando as interações invisíveis deles?

CONHEÇA A
SQUADRA

É uma das maiores consultorias de tecnologia do Brasil, parceira digital de marcas como Inter, Vale, VLI, Unimed, Claro e MRV. Combina design integral, IA aplicada e excelência técnica em desenvolvimento, operação e dados para construir plataformas digitais modernas, inteligentes e escaláveis.

Reúne mais de 700 talentos em tecnologia, atuando em todo o Brasil e em projetos que extrapolam fronteiras.

Com sua plataforma proprietária de IA, Genius, potencializa resultados e acelera jornadas digitais com impacto real.