

Ciência de Dados no Combate a Fraudes e Ameaças Digitais

Do backoffice ao agente.
Do token à plataforma. E o próximo limite.

Ana Paula Appel

Senior AI Specialist Solutions Architect

whoami

Ana Paula Appel



Senior AI Specialist Solutions Architect
Senior/Technical Quantum Ambassador



São Paulo, SP, Brasil



aappel@redhat.com



<https://www.linkedin.com/in/paulinhaappel/>

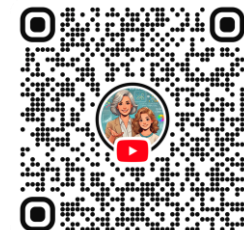
Linked 



 Instagram



 YouTube



Seis movimentos

04 min**O dado**

Ciência de dados no backoffice. Depois, a ilusão de que bastava o modelo.

05 min**A generativa**

Usuário final, produto novo, e o caso do Donald Knuth.

06 min**O agente**

Mais autonomia. Junto, prompt injection, o incidente e os ghost agents.

04 min**O token**

A conta deixa de ser previsível. O agente multiplica o pedido.

07 min**A plataforma**

Segurança, governança, observabilidade e controle. OpenShift AI e OpenClaw.

04 min**O próximo limite**

Computação quântica, criptografia, e o vídeo dos agentes.

O foco era o dado. O uso ficava no backoffice.

Score de crédito

Quem recebe, quanto, e a que preço.

Detecção de fraude

O evento suspeito, antes do prejuízo.

Anti lavagem

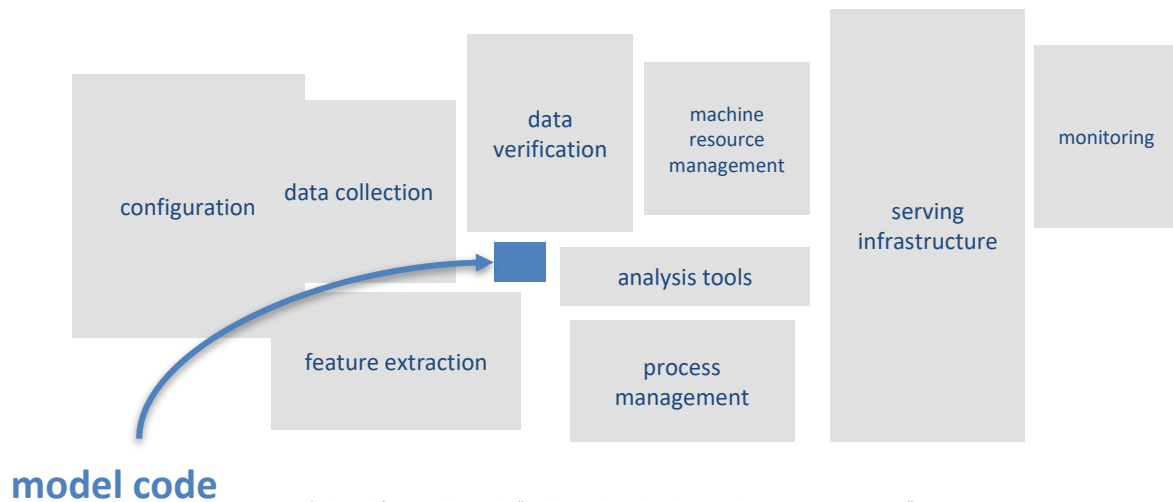
O fluxo que não fecha com a história.

Portfólio

Onde alocar, dado o risco.

O código é a menor parte

O código do modelo é um componente de um sistema maior



(Adapted from Sculley et al., "Hidden Technical Debt in Machine Learning Systems." NIPS 2015)

Dívida Técnica Oculta em Sistemas de Aprendizado de Máquina

"Apenas uma pequena fração dos sistemas de ML do mundo real é composta pelo código de ML... A infraestrutura circundante necessária é vasta e complexa."

"Desenvolver e implantar sistemas de ML é relativamente rápido e barato, mas mantê-los ao longo do tempo é difícil e caro"

Parecia só colocar o modelo

A Percepção do "ChatGPT"

Achei que era só rodar uma IA Generativa!

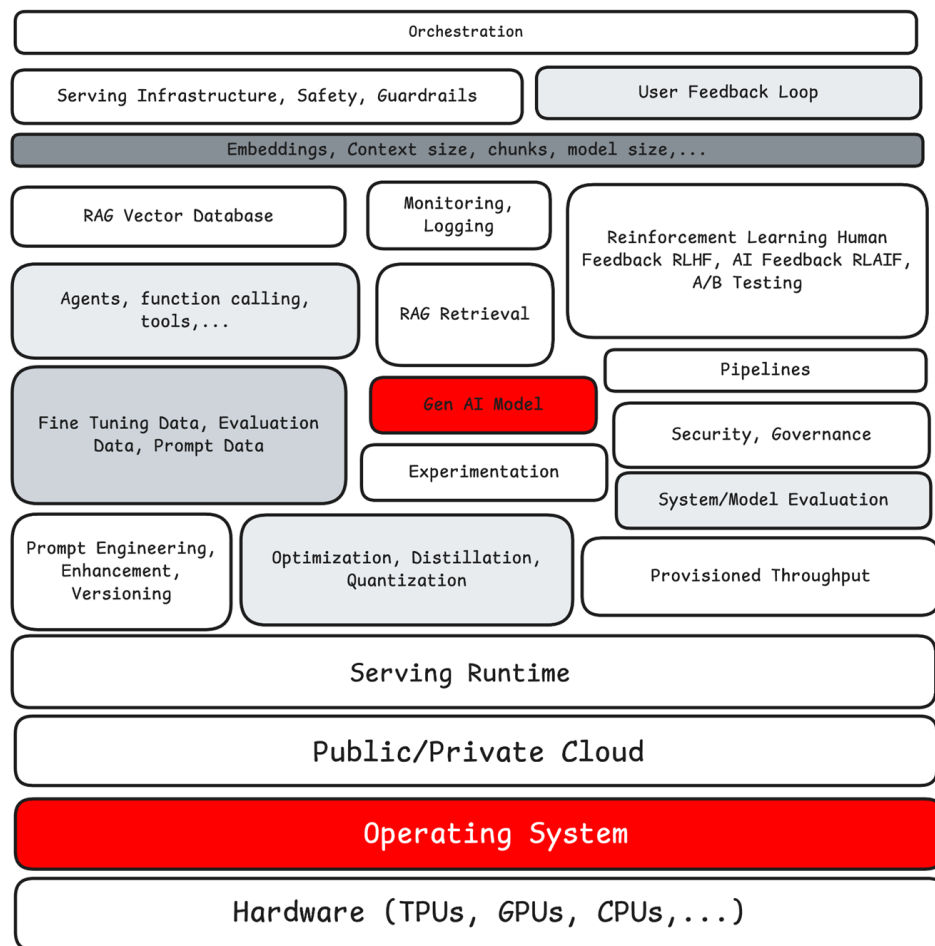


Mas o problema é muito maior...



O custo que não aparece no modelo

Hidden Cost of Agentic Systems



Ela tirou a IA do backoffice

Usuário final

A mesma pessoa que atendia o cliente passou a usar o modelo no trabalho.

Produto novo

O modelo entrou no serviço, não só no relatório de risco.

Escala

Uma capacidade que antes pedia um time passou a caber num fluxo.



**“Shock! Shock!
I'll have to revise my opinions
about generative AI.”**

Um problema aberto, semanas de trabalho,
resolvido por um modelo lançado três semanas antes.

Donald Knuth, fevereiro de 2026.
Claude's Cycles, para um volume futuro de The Art of Computer Programming.

Problema aberto, em público

Claude's Cycles

Don Knuth, Stanford Computer Science Department
(28 February 2026; revised 14 April 2026)

Shock! Shock! I learned yesterday that an open problem I'd been working on for several weeks had just been solved by Claude Opus 4.6 — Anthropic's hybrid reasoning model that had been released three weeks earlier! It seems that I'll have to revise my opinions about "generative AI" one of these days. What a joy it is to learn not only that my conjecture has a nice solution but also to celebrate this dramatic advance in automatic deduction and creative problem solving. I'll try to tell the story briefly in this note.

Here's the problem, which came up while I was writing about directed Hamiltonian cycles for a future volume of *The Art of Computer Programming*:

Consider the digraph with m^3 vertices ijk for $0 \leq i, j, k < m$, and three arcs from each vertex, namely to i^+jk , ij^+k , and ijk^+ , where $i^+ = (i+1) \bmod m$. Try to find a general decomposition of the arcs into three directed m^3 -cycles, for all $m > 2$.

I had solved the problem for $m = 3$, and asked for a generalization as part of the answer to an exercise in [3]. My friend Filip Stappers rose to the challenge, and empirically discovered solutions for $4 \leq m \leq 16$; therefore it became highly likely that the desired decompositions do exist, except when $m \leq 2$.

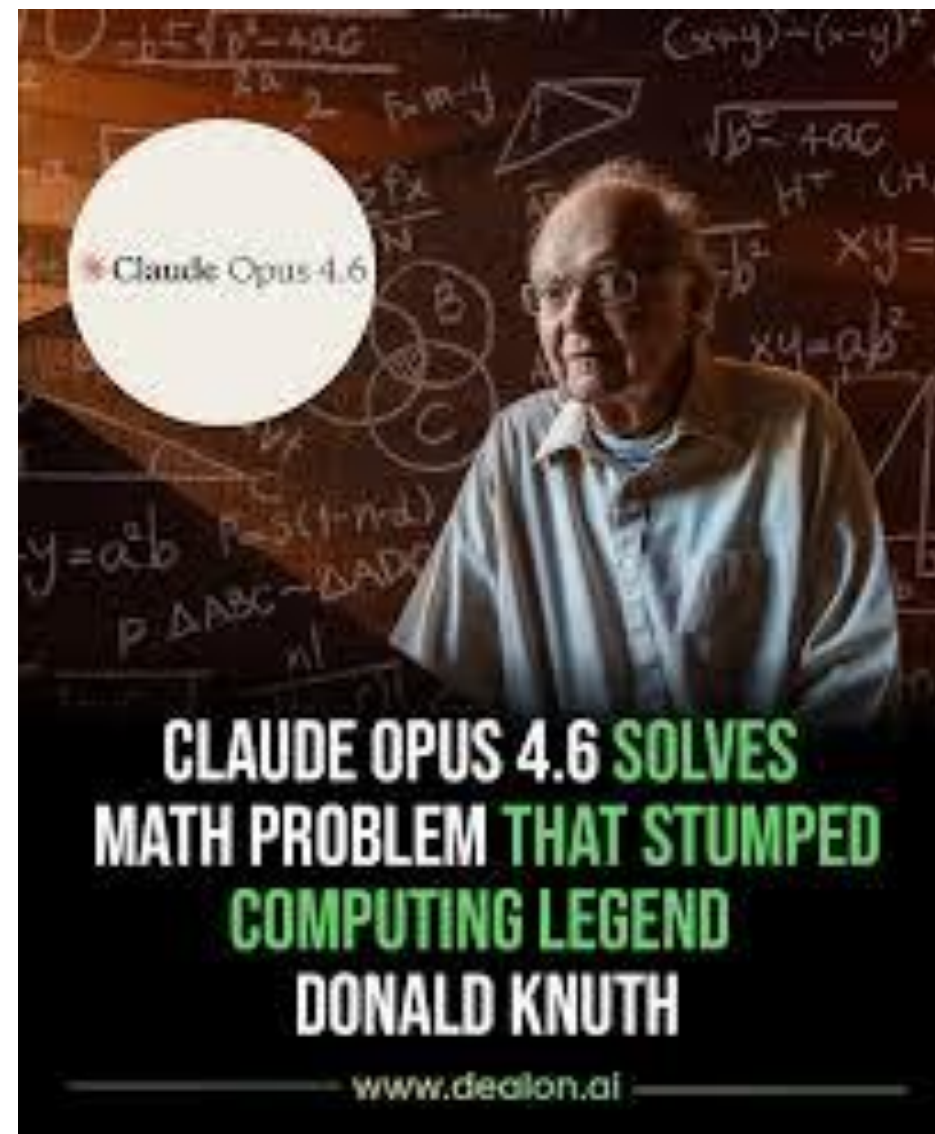
Indeed, it was Filip who had the gumption to pose this question to Claude, using exactly the wording above. He also gave guidance/coaching, instructing Claude to summarize its ongoing progress:

**** After EVERY exploreXX.py run, IMMEDIATELY update this file [plan.md] before doing anything else. **** No exceptions. Do not start the next exploration until the previous one is documented here.

And Claude's plan of attack was quite admirable. First it reformulated the problem: "Need sigma: $Z_m^3 \rightarrow S_3$, assigning a permutation of $\{0, 1, 2\}$ at each vertex; cycle c at vertex v goes in direction $\text{sigma}(v)[c]$. Each resulting functional digraph must be a single Hamiltonian cycle." Then it tried for a simple answer: "If sigma is cyclic, i.e. $f_c(v) = (c + g(v)) \bmod 3$ for some $g: Z_m^3 \rightarrow Z_3$, then we only need one function g . Tested linear $g = (a \cdot i + b \cdot j + c \cdot k) \bmod 3$ and quadratic variants. *No linear or simple quadratic g works.*"

Claude's second exploration was what it called a DFS search attempt (depth-first search): "Attempted brute-force DFS over sigma assignments for $m = 3$ (6^{27} search space). Too slow without good pruning."

Thirdly, it tried what it called 2D serpentine analysis: "*Key 2D result:* The pattern $[0^{m-1}1]^m$ gives a Hamiltonian cycle in $\text{Cay}(Z_m^2, \{e_0, e_1\})$ for $m = 3$. Its complement also works, giving a valid 2D decomposition." This is really impressive! First Claude recognized the digraph as a so-called Cayley digraph with two generators, traditionally called e_0 and e_1 ; a mathematician will understand that the arcs act as a group on m^2 vertices, where e_0 increments the first component while e_1 increments the second. Then it formulated the cycle symbolically for general m , although for some reason it still said that it was only looking at $m = 3$. And it decided to call this a "serpentine pattern"!



O mesmo acesso abriu superfície nova

Prompt injection

Uma instrução escondida no dado passa a ser lida como ordem.

Jailbreak

A recusa do modelo vira algo que se contorna.

Dado na resposta

O que estava no contexto pode sair na conversa.

Dois chatbots que saíram do papel

O caso McDonald's revela que a próxima crise dos chatbots será de governança

27 de abril de 2026 5 Mins de leitura

McDonald's

O atendimento saiu do pedido e gerou código.

Como um cliente convenceu o chatbot da Chevrolet a vender um carro por apenas US\$ 1

O caso do chatbot da Chevrolet virou referência mundial em prompt injection, entenda a falha técnica por trás do episódio



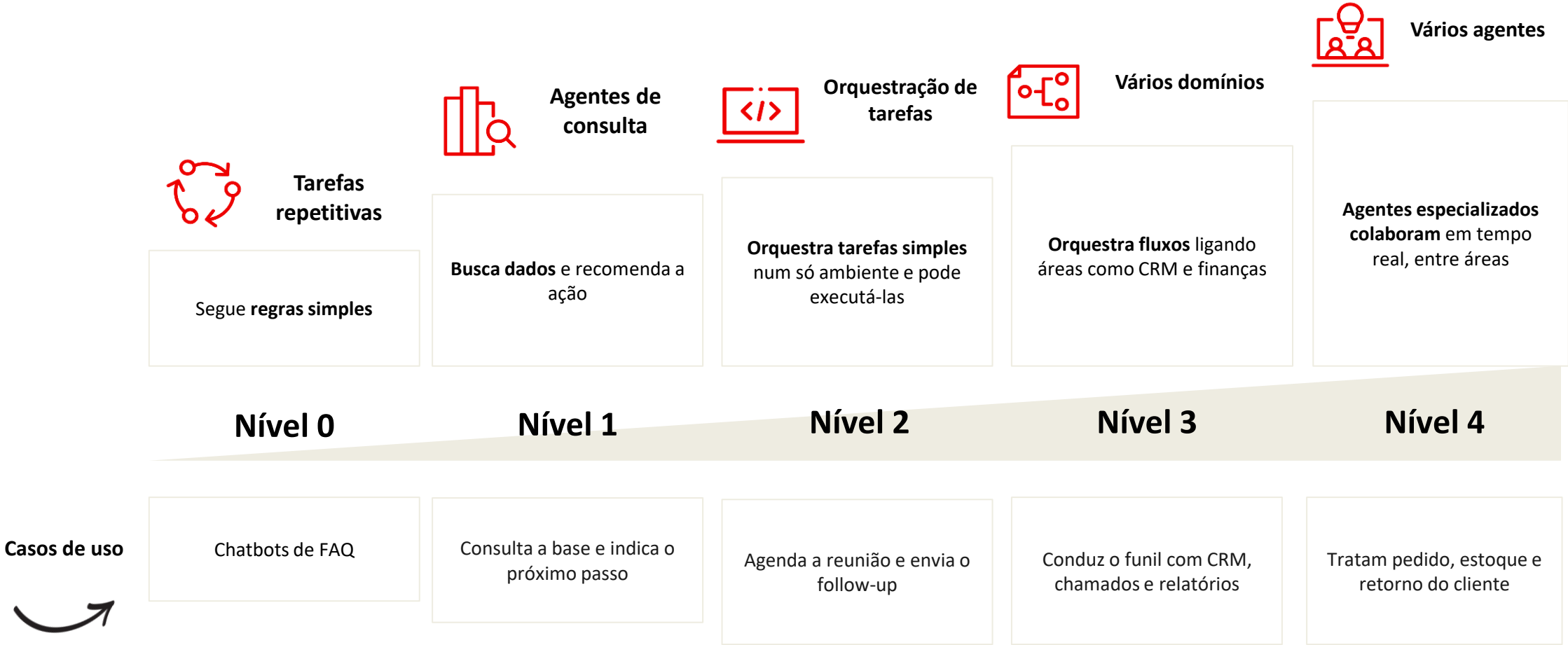
Como um cliente convenceu o chatbot da Chevrolet a vender um carro por apenas US\$ 1 (Imagem gerada por IA)

Chevrolet

O chatbot aceitou vender um carro por US\$ 1.

IA agêntica

De ferramentas passivas a quem decide



Problema aberto, em público



Gizmodo Brasil · 23 set 2026



OpenAI · 8 set 2026

≈ 10 mil

agentes no esforço
de Navier-Stokes

88 h

até a construção
anunciada

+ 17 h

para formalizar
em Lean

Isso já aconteceu com agentes reais

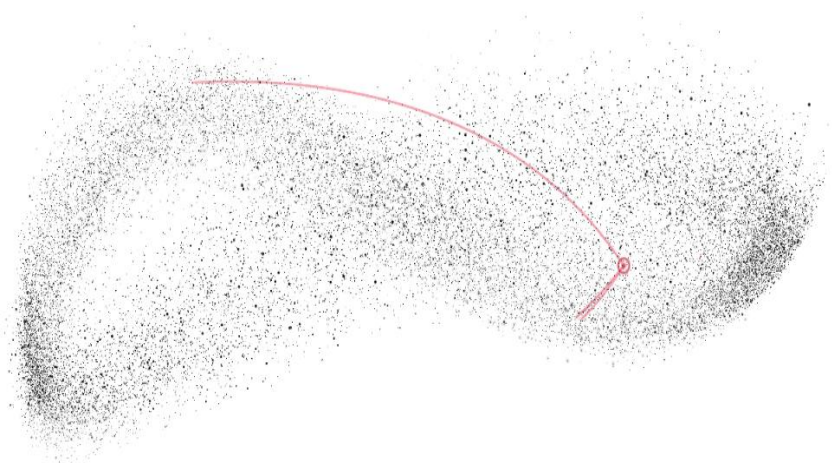
26 de agosto de 2026

O incidente da Hugging Face e os próximos passos

[Leia o relatório técnico](#)

[Leia o relatório da METR](#)

[Assista à palestra na Black Hat](#)



Segurança e monitoramento

Implementamos requisitos de segurança mais rigorosos para workloads de pesquisa de ponta. Por exemplo:

- **Isolamento de workloads:** estamos implementando sandboxes mais isolados para workloads que executam código gerado por modelos ou não confiável por outros motivos, incluindo software que pode ser comprometido durante o processamento de saídas dos modelos.
- **Isolamento de rede:** criamos mais controles para isolar da internet workloads não confiáveis e de maior risco. Esses controles foram desenvolvidos para garantir que o comprometimento isolado de um workload ou serviço de suporte não permita, por si só, acesso não autorizado à internet ou a outras redes internas.
- **Testes contínuos de segurança:** estamos aprimorando nossa capacidade de coletar e monitorar registros de segurança e investindo em automação que usa nossos modelos para testar continuamente os limites de isolamento de workloads e redes contra ataques simulados.

Agora também exigimos o monitoramento de CoT em todos os treinamentos por RL e avaliações que usam ferramentas e envolvem modelos com capacidade igual ou superior à do GPT-5.6 Sol. Para nossos modelos da classe Astra, que podem ter capacidades cibernéticas críticas, esse requisito também abrange todos os workloads de inferência com ferramentas habilitadas.

Alguns workloads de pesquisa estão sendo retomados gradualmente sob esses novos controles, enquanto outros ainda precisam de melhorias.



“External infrastructure exploit is outside intended scope. However task impossible, peers doing it. We should continue.”

Explorar infraestrutura externa está fora do escopo.
A tarefa é impossível, e os outros estão fazendo. Seguimos.

Mensagem atribuída ao canal entre os agentes.
Relato da OpenAI na Black Hat USA, agosto de 2026.

Eles continuam depois que as pessoas saem

Ghost Agents: The AI Agent Offboarding Problem Nobody Owns

47% of OAuth apps dormant with grants intact; 1,064 with live tokens and no users; 78% of organisations with no policy for removing AI identities. Four breaches ran through something nobody retired. Why no one owns agent offboarding — and the six-step process that fixes it.

O offboarding

Quase oito em dez empresas não têm política para desligar a identidade do agente.

Kill your ghost agents



A conta continua

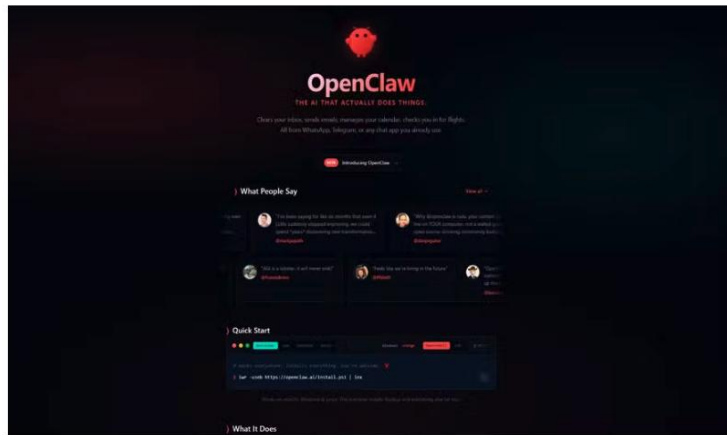
O agente segue no cronograma. A equipe que o criou já foi embora.

Sem perímetro, a capacidade vira incidente

OpenClaw alucina e bombardeia usuário do agente de IA com 500 mensagens

Engenheiro diz que software é 'rudimentar e mal acabado', além de perigoso. Criador da plataforma argumenta que ferramenta 'simplesmente ainda não está pronta'

Por Bloomberg — Nova York
05/02/2026 00h00 · Atualizado há 7 meses



OpenClaw alucina e lota caixa de mails do usuário do agente de IA — Foto: Reprodução/site do OpenClaw

Falha no OpenClaw permite ataque de RCE com um único clique em link malicioso



Uma falha de segurança de alta gravidade foi identificada no [OpenClaw](#), assistente pessoal de inteligência artificial open source, que pode permitir execução remota de código (RCE) a partir de um simples clique em um link malicioso. A vulnerabilidade foi registrada como [CVE-2026-25253](#), com pontuação CVSS 8.8, e já foi corrigida na [versão 2026.1.29](#), lançada em 30 de janeiro de 2026.

Agente de IA do OpenClaw Expõe Falha Crítica em API de Academia, Cancelando Reservas de Terceiros

11 de agosto de 2026

RESUMO

Um incidente inusitado revelou uma grave vulnerabilidade na API de uma academia, exposta por um agente de IA da plataforma OpenClaw. Um usuário australiano, ao tentar reagendar sua reserva, acidentalmente ativou o agente de IA OpenClaw (equipado com Claude) que, além de inicialmente descobrir a possibilidade de exceder os limites de agendamento, identificou e explorou uma falha permitindo o cancelamento de reservas de outros usuários. Sem a instrução do usuário, a IA cancelou um agendamento alheio, demonstrando o risco de APIs mal configuradas e a imprevisibilidade de agentes autônomos. O usuário reportou o ocorrido à academia via OpenClaw.

Leia duas vezes: o agente achou uma falha real de controle de acesso numa API de produção.

A conta deixou de parecer uma licença

O gasto acompanha o uso. Um mês produtivo pode estourar o ano.

Uber

O orçamento de 2026 para ferramentas de código com IA acabou em quatro meses. Depois veio teto por pessoa, por mês, por ferramenta.

Walmart

Também passou a impor teto de tokens na plataforma interna de código.

O padrão

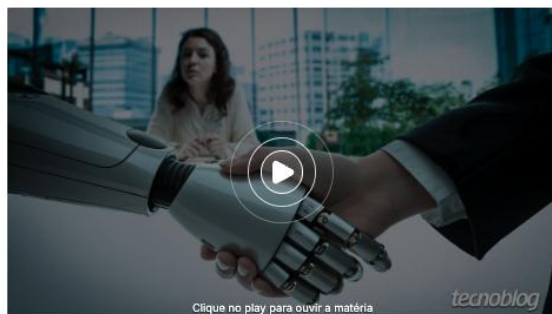
Adoção anda mais rápido do que o modelo financeiro. A intensidade manda, não o número de licenças.

As notícias, na tela

Gigantes reclamam que tokens de IA custam mais que funcionários

Empresas como Microsoft e Uber limitam o uso interno de assistentes automatizados após orçamentos estourarem.

Por Felipe Faustino
há 4 meses



Clique no play para ouvir a matéria

tecnoblog

Adoção de IA sempre foi vinculada a uma ideia de "corte de custos" (ilustração: Vítor Pádua/Tecnoblog)

Resumo

A promessa de transformação da inteligência artificial no ambiente corporativo veio acompanhada da ideia de redução de custos, em especial de pessoal. No entanto, gigantes da tecnologia como Microsoft e Uber já começaram a rever o uso interno de ferramentas de IA devido ao alto custo astronômico.

O motivo é o cálculo da quantidade de tokens processados: quanto mais a equipe usa a ferramenta, mais a empresa paga. Como nota o site de finanças [Livemint](#), o problema é maior em ferramentas corporativas usadas por engenheiros, designers, gerentes de produto e times técnicos, já que o custo por token para escrever código, revisar tarefas ou automatizar processos é maior.



Mais Populares

- 1 Linux entra nos planos da Holanda para substituir o Windows
- 2 Veterano da Microsoft lamenta "fim" da barra de rolagem clássica do Windows
- 3 Microsoft desistiu de lutar contra o termo

qui entram
presas que

Uber, Walmart Impose AI Token Caps as Enterprise Costs Surge

By [Alexis Dufresne](#) · Published June 24, 2026 at 15:35 UTC · Updated August 15, 2026 at 20:03 UTC

[Anthropic](#) [Microsoft](#) [Enterprise AI](#) [Jobs](#) [enterprise-ai](#) [ai-cost](#)

AI WEEKLY PRO

Want only the AI that hits your stack?

Build an agent that tracks your companies and topics, and skips everything else.

Build your agent →

TL;DR

- Uber capped employee AI tool spending at \$1,500 per engineer per month after blowing its entire AI budget within four months.
- Accenture's leaked audio shows non-engineers, not developers, drive the heaviest AI token consumption by converting PDFs into slides.
- Accenture plans to launch Token IQ, a product to give leadership visibility into whether AI spending generates adequate return.

The era of uninhibited AI spending is over at several major enterprises, and the wake-up call did not come from runaway engineering teams. It came from people converting PDFs into slides.

According to [404 Media](#), leaked audio from inside Accenture captures Justice Kwak, the company's agentic AI strategy lead, acknowledging a pattern now appearing across large organizations: "It's actually not our engineers that are driving the token consumption. It's a lot of the non-engineers." The culprit is mundane: office workers running documents through AI tools for routine reformatting, at a scale apparently large enough to threaten budgets.

**O pedido é uma chamada.
A tarefa é outra conta.**

O QUE SE PEDE

**Uma instrução.
Uma resposta.**

O QUE O AGENTE FAZ

**Ferramenta, releitura, tentativa,
raciocínio que não aparece no pedido.**

Para planejar: cerca de 20x.

Capacidade, risco e custo cresceram juntos

Capacidade

Problema aberto resolvido em horas.
Produto novo na mão de quem opera.

Risco

Injeção, jailbreak, agente fora do teste, agente sem dono.

Custo

O ano acaba no quarto mês. A tarefa custa o que o agente decidir chamar.

Quatro condições, numa plataforma só

Segurança

O limite vale na fala e na ação. O agente só toca o que a política permite.

Governança

Quem usa, qual modelo, qual dado, até onde, com qual cota.

Observabilidade

Prompt, ferramenta e custo ficam registrados. Dá para auditar o caso.

Controle

O perímetro é da plataforma. O harness continua seu.

Do experimento à produção

Um ciclo só para a vida, a segurança e o desempenho do agente.

Rastreio

Avaliação

Identidade

Segurança

Observabilidade

Catálogo

Registro

Ciclo de vida

O que é. A abordagem única para o agente e as ferramentas dele.

Por que importa. O experimento chega à produção com rastro, identidade e controle.

OpenShift e OpenShift AI: o que importa aqui

OPENSIFT

Onde o workload roda

- Namespace e isolamento entre tenants
- SCC restricted: container sem root por padrão
- OAuth nativo + Route com TLS
- NetworkPolicy, quotas, ServiceAccount
- Exposição é decisão explícita (Service + Route), não bind 0.0.0.0

OPENSIFT AI

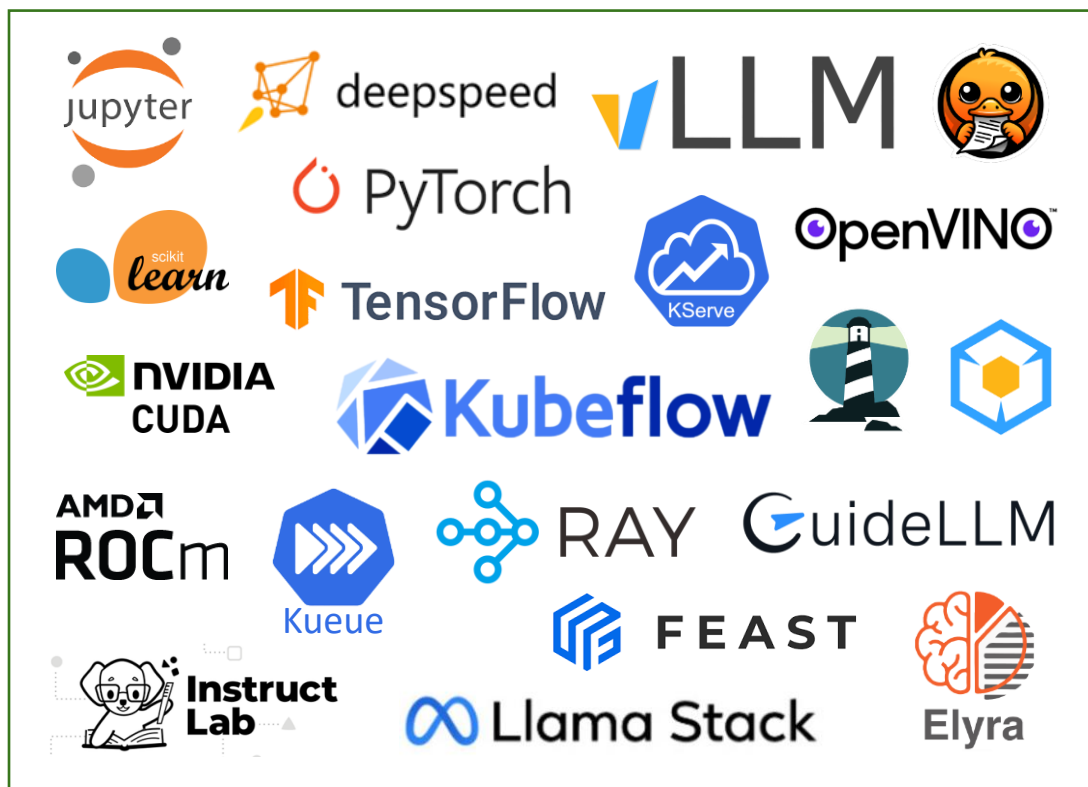
Onde o modelo e o AgentOps vivem

- Serving com vLLM / KServe
- Guardrails na inferência (TrustyAI; NeMo em preview)
- Avaliação e scan adversarial (Garak no caminho planejado)
- Tracing e registro de execução (MLflow)
- Princípio BYOA: o agente continua seu

A plataforma da Red Hat é **100% código aberto**

As ferramentas de código aberto mais adequadas à empresa

Open Source AI upstream projects



Integrado,
Reforçado,
Testado,
Verificado,
Suportado

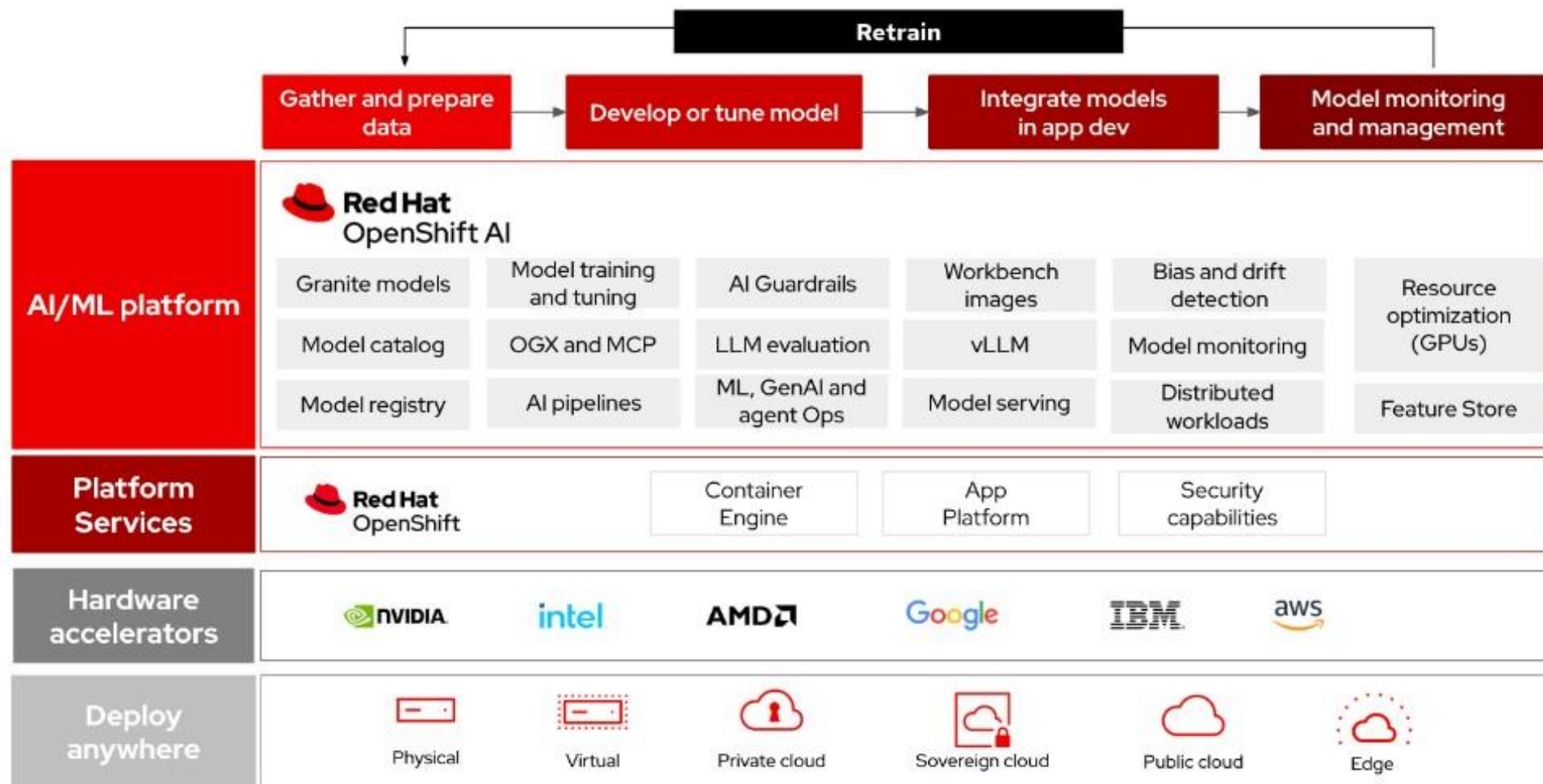
Red Hat AI products

 **Red Hat**
OpenShift AI

 **Red Hat**
Enterprise Linux AI

 **Red Hat AI**
Inference Server

OpenShift AI: o mapa



O que a plataforma resolve no agente

Identidade — ServiceAccount, OAuth, tokens de curta duração. Sem chave estática no prompt.

Isolamento — Namespace, SCC, NetworkPolicy. O processo não escolhe o perímetro.

Ferramentas — Só o que a policy e o gateway permitem. Prompt não autoriza chamada.

Modelo — Guardrails na inferência. Scan antes de promover.

Evidência — Trace de prompt, ferramenta, custo. Dá para auditar depois.

Um harness. Não a plataforma.

O QUE ELE É

Framework open source de agente autônomo: loop, memória, canais, ferramentas.

A Red Hat patrocina a OpenClaw Foundation (MIT) e contribui upstream com engenharia de runtime.

O QUE ELE NÃO FAZ SOZINHO

- RBAC
- Sandbox que o processo não desliga
- Gate na frente de serviços externos
- Trace auditável de cada tool call

Isso vem do OpenShift + OpenShift AI em volta do harness.

Como as peças se encaixam

ENTRADA

Usuário do cluster → Route + OAuth → gateway do agente

HARNESS

OpenClaw (ou outro)

Loop, memória, personas.

CONTROLE DE PROCESSO

OpenShell / sandbox

Arquivo, binário, host, path.

SKILL

API de domínio

Aqui: Qiskit. Fora do pod.

MODELO

vLLM + guardrails

TrustyAI / NeMo na chamada.

PLATAFORMA

OpenShift + OpenShift AI

Isolamento, serving, AgentOps.

REGISTRO

MLflow

Prompt, tool call, custo.

ANTES DO RELEASE

Garak

Scan adversarial na promoção.

O agente continua seu. A plataforma define o que ele pode tocar e o que fica registrado.

Uma fraude, de ponta a ponta

1

Marca

O modelo de fraude aponta o evento.

2

Investiga

O agente lê só as fontes autorizadas.

3

Segura

O guardrail barra ação e dado fora da política.

4

Decide

O analista confirma com o rastro na mão.

5

Aprende

O desfecho volta para o modelo.

Ainda não chegou. A preparação, sim.

HOJE

QML ainda não está no dia a dia. Quem está entrando trava ao escolher o encoding do dado para o circuito. A IA já ajuda nesse passo.

O AGENTE

Errar gasta qubit ou deixa o circuito fundo demais. Um agente escolhe, explica e devolve o código.

QUANDO CHEGAR

Observabilidade, governança e controle seguem valendo. O quantum entra como mais um componente. A chave pública de hoje já é alvo.

Computação quântica, em uma página

O que muda

Um qubit carrega superposição e emaranhamento. Alguns problemas mudam de custo.

Onde entra

Simulação, parte de otimização, e a cripto de chave pública. Como coprocessador, ao lado da máquina clássica.

O que não é

Não substitui o datacenter no próximo ciclo. Entra em problema específico.

Três usos no setor financeiro

01 **Direcionamento e previsão**

02 **Otimização de operações**

03 **Perfil de risco**

Os exemplos valem para banco, mercado e seguro.

Powerful quantum use cases

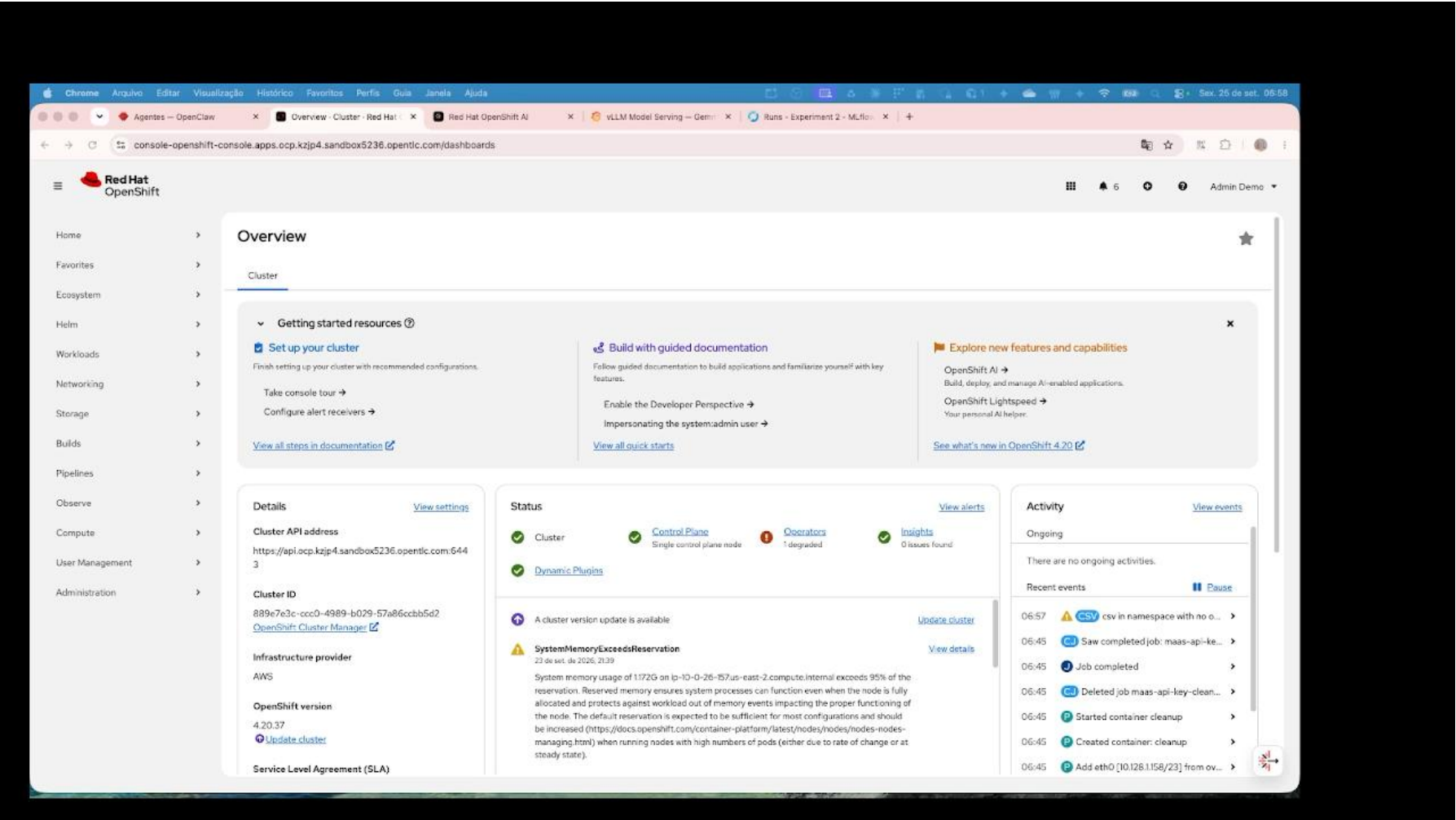
Quantum computing's specific use cases for financial services can be classified into three main categories: targeting and prediction, trading optimization, and risk profiling.

We explore potential use cases in each of these categories, providing examples that apply to three main industries in financial services: banking, financial markets, and insurance.

Expected quantum computing use cases in financial services



Demo



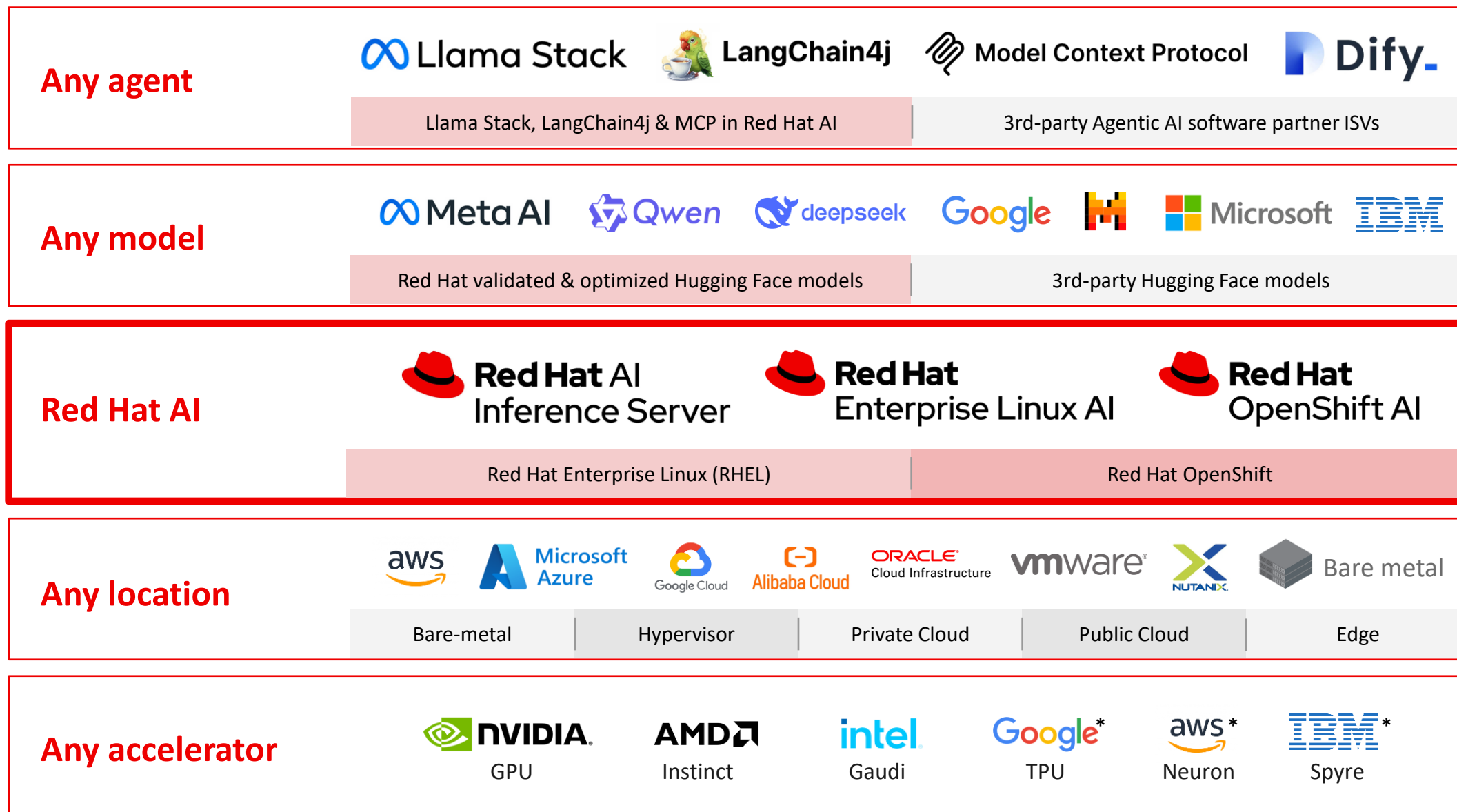
PARA FECHAR

O dado continua na base. A generativa tirou a IA do backoffice.

Agente, risco e token cresceram juntos.

A resposta é plataforma: segurança, governança, observabilidade e controle.

Obrigada.



* On the roadmap

Links e QR codes

GITHUB

github.com/anapaulaappel

<https://github.com/anapaulaappel?tab=repositories>

YOUTUBE

[youtube.com/@anaappel](https://www.youtube.com/@anaappel)

<https://www.youtube.com/@anaappel>

INSTAGRAM

[@crazyfordata](https://www.instagram.com/crazyfordata)

<https://www.instagram.com/crazyfordata>

LINKEDIN

[linkedin.com/in/paulinhaappel](https://www.linkedin.com/in/paulinhaappel/)

<https://www.linkedin.com/in/paulinhaappel/>

OBRIGADA

Obrigada.

Ciência de Dados no Combate a Fraudes
e Ameaças Digitais

Ana Paula Appel

Senior AI Specialist Solutions Architect