



DESAFIOS PARA O USO DA IA GENERATIVA NO TRABALHO PARA AUMENTO DA PRODUTIVIDADE

Genaro Costa

EVIDEN



Nossa visão: Estamos apenas no início da era exponencial de dados/IA



Mapeando horizontes de HPC com tecnologia de IA

Eviden, líder mundial em HPC, IA e Quantum

#1 na Europa, #3 globalmente*

Entregando os principais centros de pesquisa europeus



1 na América do Sul e Índia

Entregando grandes supercomputadores

Aumentando nossa presença mundial na APAC, América do Sul, América do Norte e MEA



Fornecimento de grandes parceiros industriais
Com soluções complexas e otimizadas



Em 1950 ...

MIND
A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

O Jogo da Imitação

Proponho considerar a questão: “As máquinas podem pensar?”

Isso deveria começar com definições dos significados dos termos “máquina” e “pensar”. As definições poderiam ser elaboradas de forma a refletir, tanto quanto possível, o uso normal das palavras, mas essa postura é perigosa. (...)

I.—COMPUTING MACHINERY AND
INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, ‘Can machines think?’ This should begin with definitions of the meaning of the terms ‘machine’ and ‘think’. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ‘machine’ and ‘think’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ‘Can machines think?’ is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.



Geopública 2025



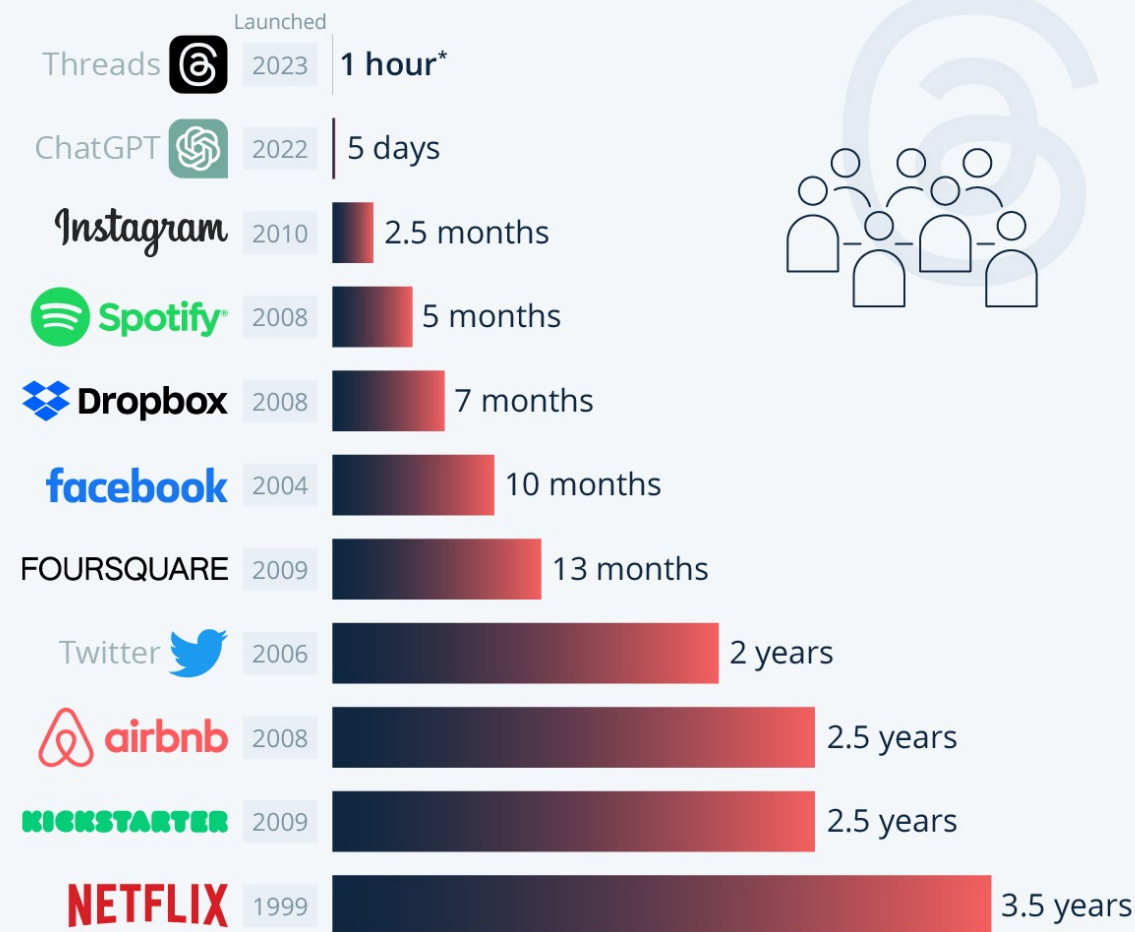
SotM BR 2025
STATE OF THE MAP BRASIL

Contexto

Crescimento acelerado da IA generativa

Threads Shoots Past One Million User Mark at Lightning Speed

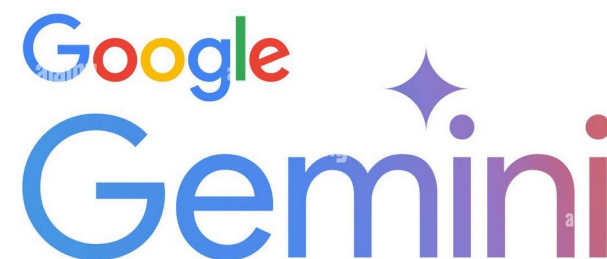
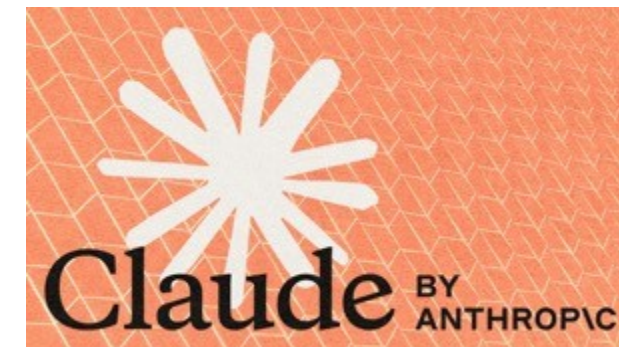
Time it took for selected online services to reach one million users



Refers to one million backers (Kickstarter), nights booked (Airbnb), downloads (Instagram/Foursquare)

* Two million signups in two hours

Source: Company announcements via Business Insider/LinkedIn



Contexto

Crescimento acelerado da IA generativa

COMO FUNCIONA UMA IA?



Geopública 2025

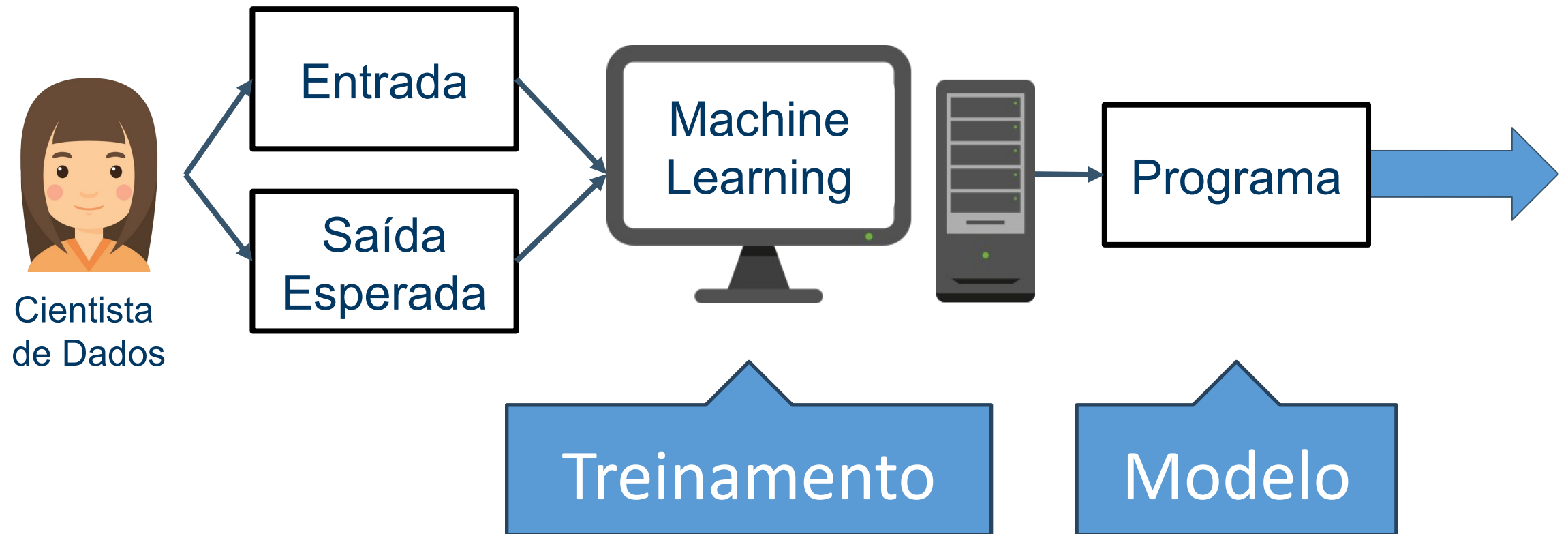


SotM BR 2025
STATE OF THE MAP BRASIL

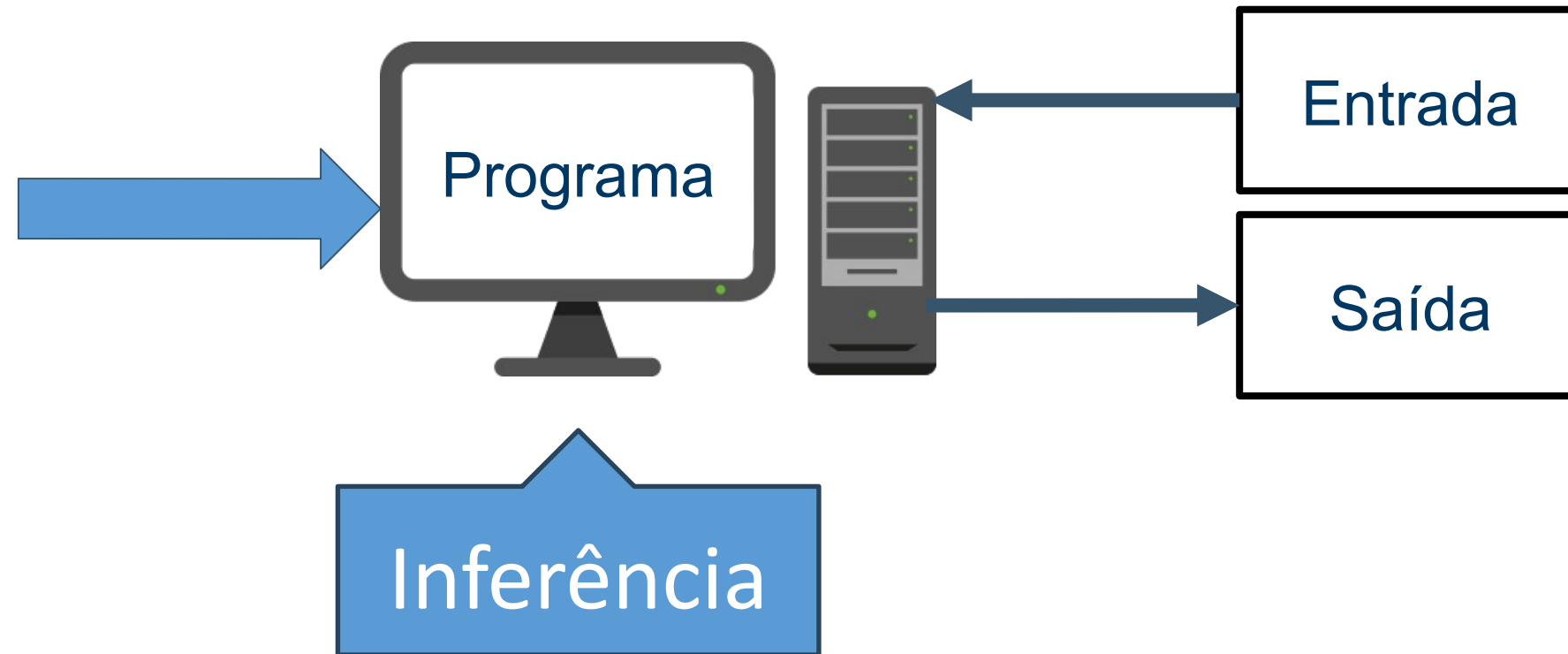
Programação



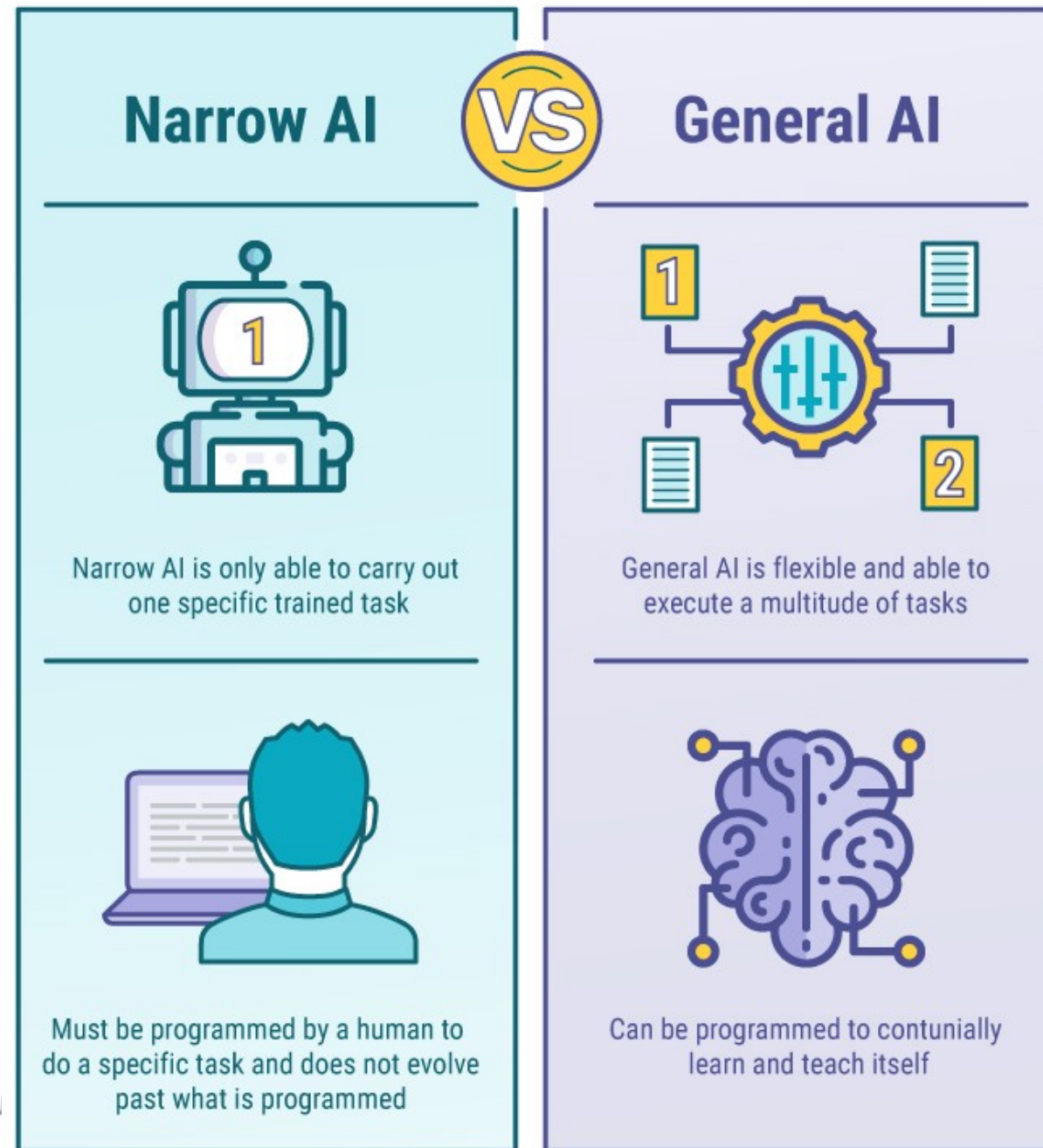
Aprendizagem de Máquina (ML)



Execução

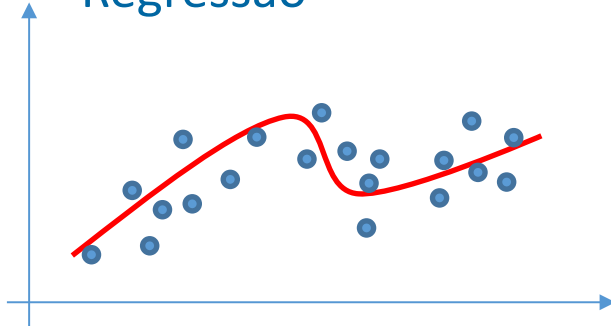


IA “estreita” vs. IA “generalista”



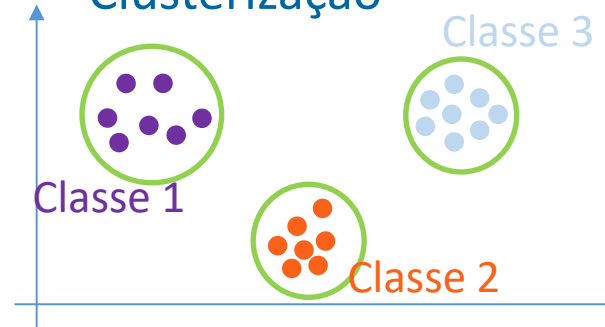
Alguns Algoritmos da AI “estreita”

Regressão



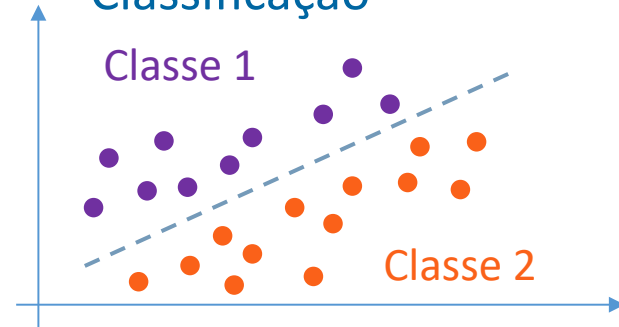
Previsão de resultados com base em registros históricos

Clusterização



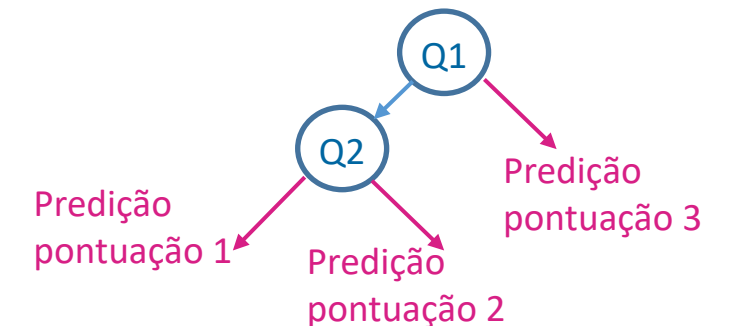
Encontrar itens semelhantes entre si

Classificação



Classificando itens individuais em um conjunto de classes predefinidas

Árvore de decisão



Ilustrando todos os resultados possíveis de uma decisão

A Jornada da IA Generativa

2010

Tradução quase perfeita da linguagem natural

2014

Dominando o significado das palavras

**2017 a
2022**

Modelos de Fundação de Linguagem de Grande Escala (LLMs)

2022

**Modelos de Fundação de Linguagem de Grande Escala para
Conversação**



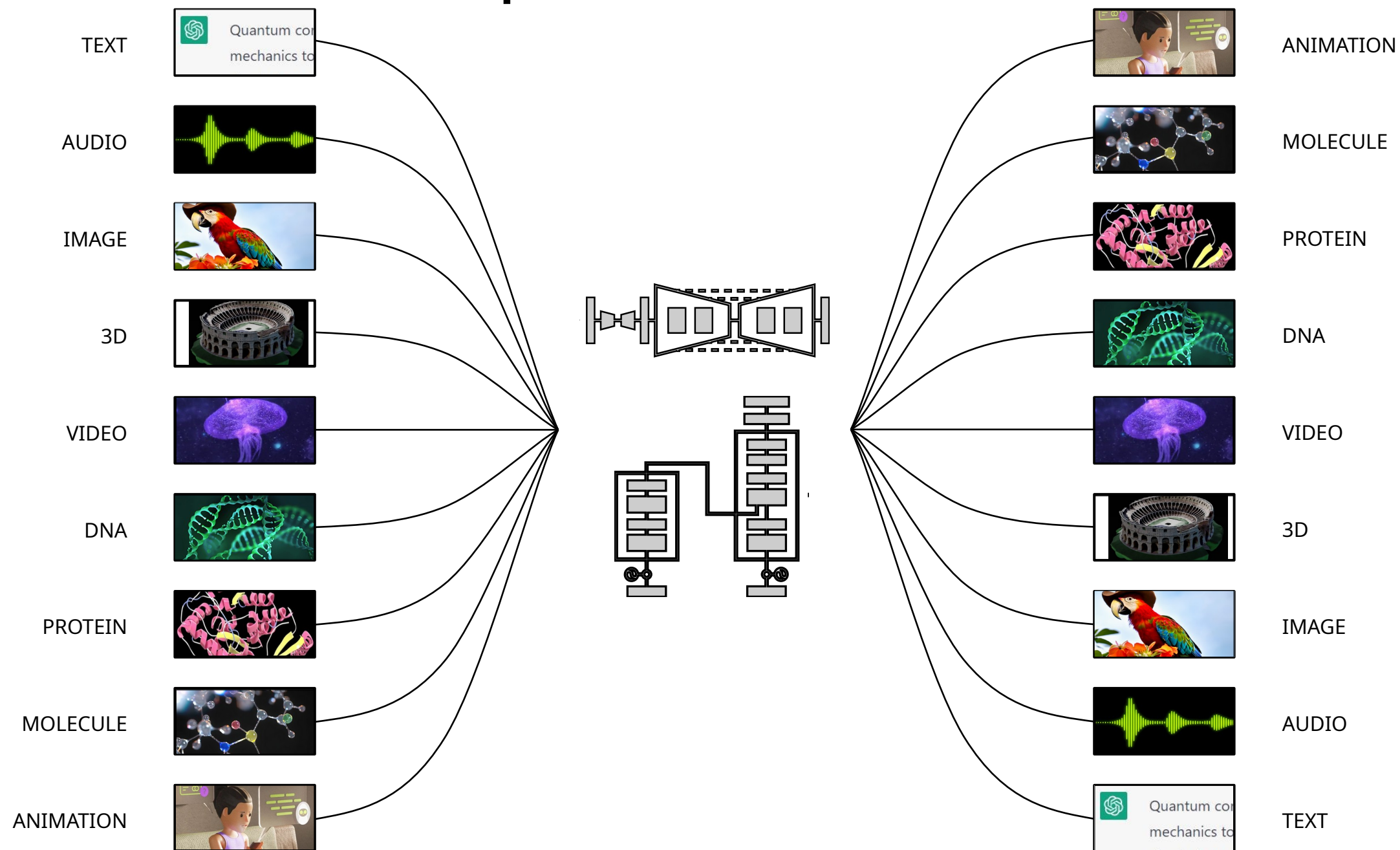
Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Fonte: Gartner (adaptado)

O que é IA GenerativaI?

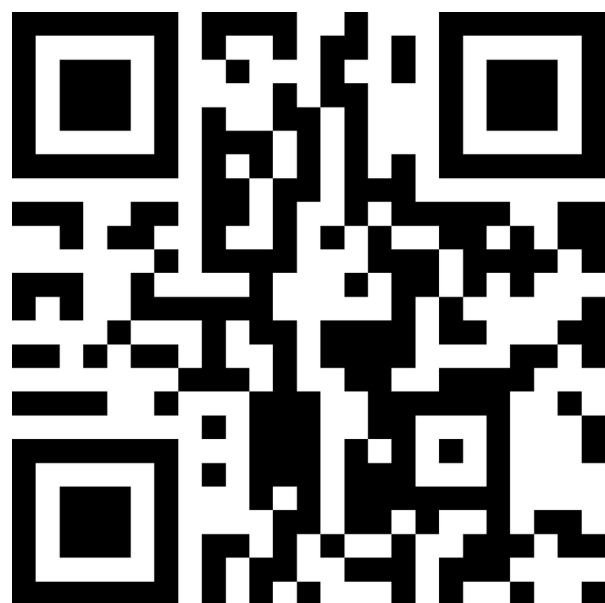


Tokenização de Textos

A tokenização é uma etapa fundamental no pré-processamento de texto para muitas tarefas de processamento de linguagem natural (NLP), inclusive para modelos de linguagem como GPT-4 e Llama-2. A tokenização envolve a divisão do texto em partes menores, ou "tokens", que podem ser tão curtos quanto um caractere ou tão longos quanto uma palavra (ou até mais longos em alguns casos). Esses tokens podem ser processados, analisados e usados como entrada para modelos de aprendizado de máquina.

<s> Bem vindo ao Geopública 2025. Minha terra tem palmeiras onde
canta o sabiá; As aves que aqui gorjeiam, não gorjeiam como lá.
Nosso céu tem mais estrelas, nossas várzeas tem mais flores, n
ossos bosques tem mais vida, nossa vida mais amores.

Tokenização
visualizada



IDs de token
resultantes

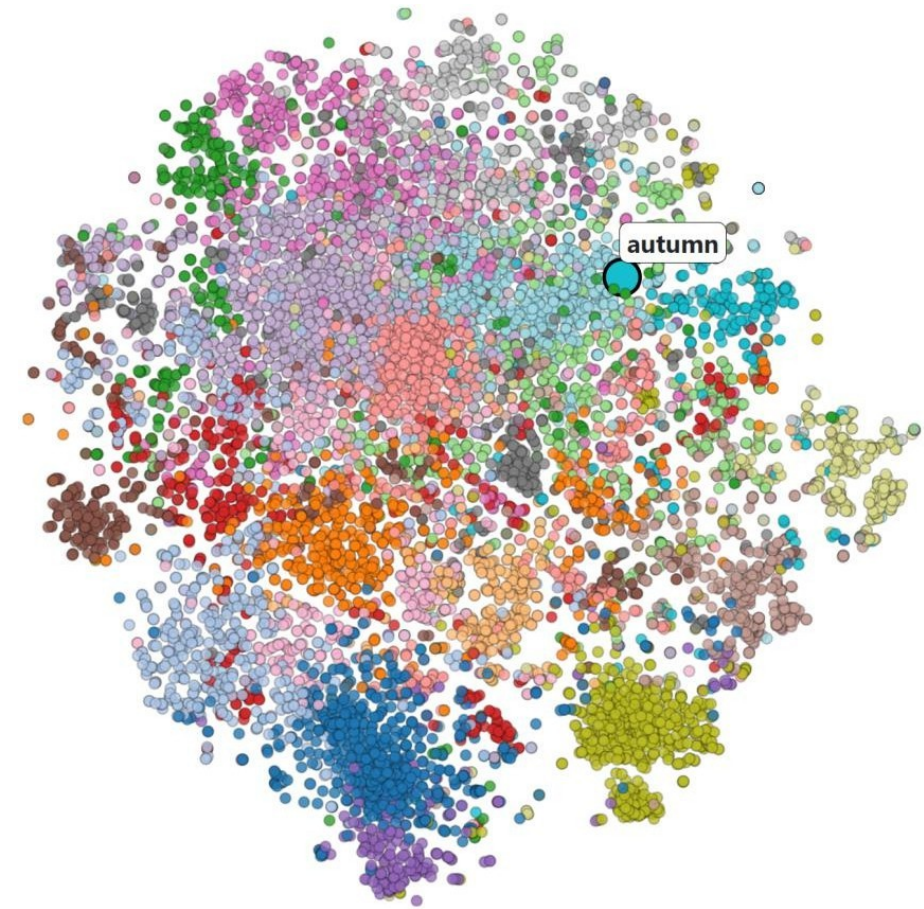
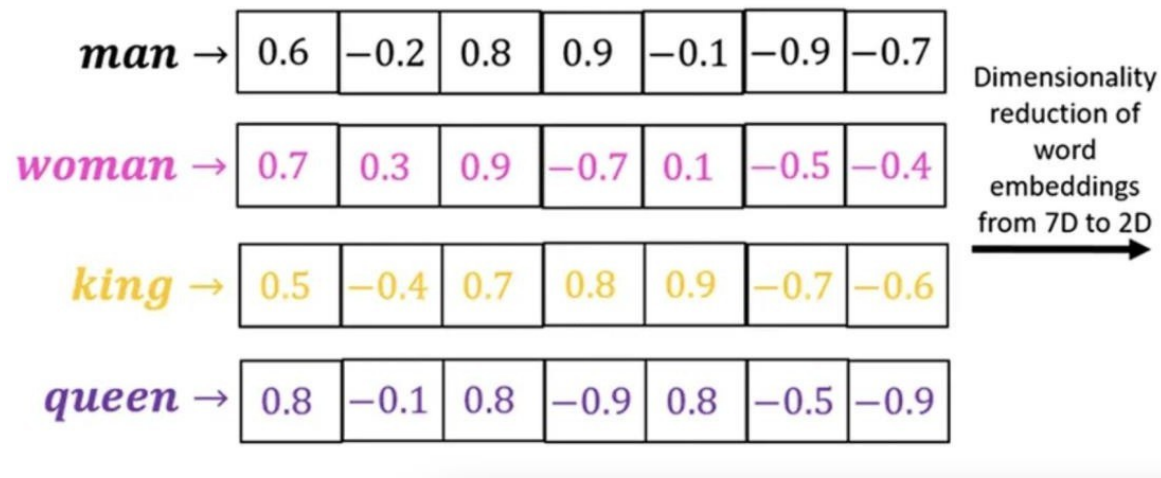
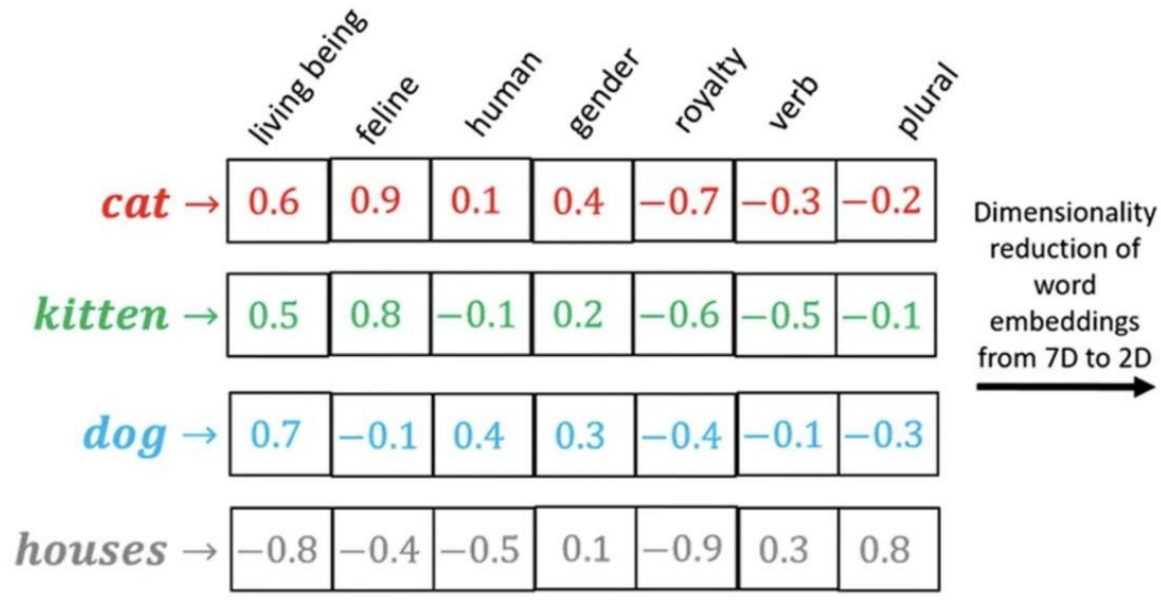


Word: **41** palavras → Mistral v3: **92** tokens

```
[30642, 1634, 318, 257, 43936, 2239, 287, 262, 662, 36948, 286, 2420,  
329, 867, 3288, 3303, 7587, 357, 45, 19930, 8, 8861, 11, 1390, 329, 3303,  
4981, 588, 402, 11571, 12, 19, 290, 18315, 1689, 12, 17, 13, 29130, 1634,  
9018, 7163, 866, 2420, 656, 4833, 22716, 11, 393, 366, 83, 482, 641,  
1600, 543, 460, 307, 355, 1790, 355, 530, 2095, 393, 355, 890, 355, 530,  
1573, 357, 273, 772, 2392, 287, 617, 2663, 737, 2312, 16326, 460, 788,  
307, 13686, 11, 15475, 11, 290, 973, 355, 5128, 329, 4572, 4673, 4981,  
13]
```


Vetor de Embeddings (incorporações)

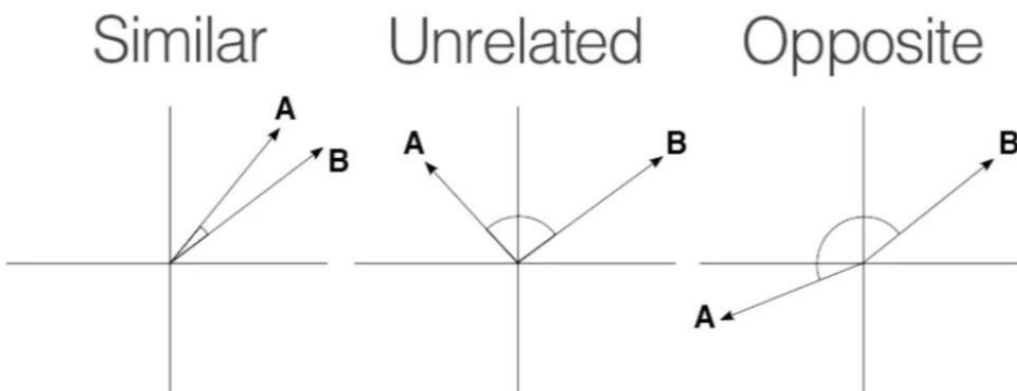
- Existem várias dezenas de modelos de incorporação
- Eles variam em complexidade de 384 a 1536 dimensões
- Eles variam em comprimento máximo de sequência de 512 a 8191 tokens
- Os modelos de incorporação geralmente não são compatíveis entre si



Interactive embedding explorer:
<https://github.com/echen/embedding-explorer>

Similaridade semântica de texto

- Mede o cosseno do ângulo entre dois vetores.
- Valor entre -1 e 1; onde 1 significa que os vetores são idênticos, 0 significa ortogonais e -1 significa diametralmente oposto (raro em incorporações de texto).



Sentence 1	Sentence 2	Cosine Similarity
The cat sits outside	The dog plays in the garden	0.2838
A man is playing guitar	A woman watches TV	-0.0327
The new movie is awesome	The new movie is so great	0.8939
Jim can run very fast	James is the fastest runner	0.6844
My goldfish is hungry	Pluto is a planet!	0.0454



```
from sentence_transformers import SentenceTransformer, util
model = SentenceTransformer('all-MiniLM-L6-v2')

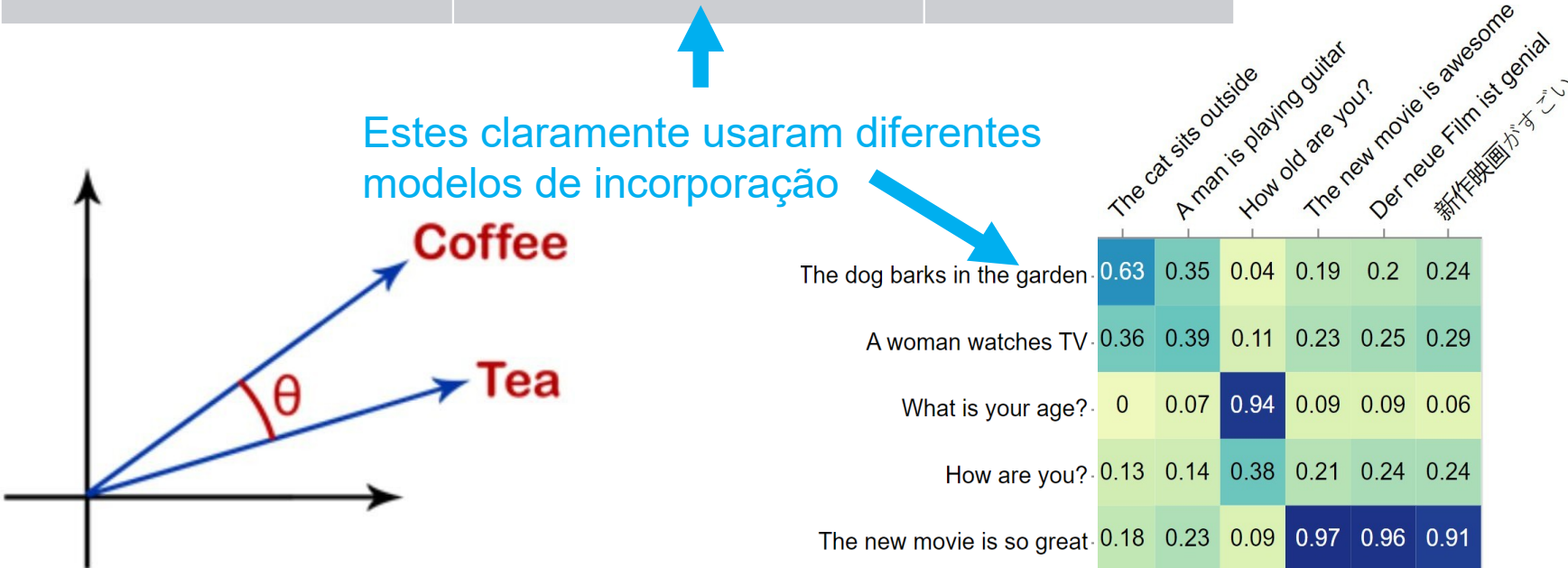
sentences1 = [
    'The cat sits outside',
    'A man is playing guitar',
    'The new movie is awesome',
    'Jim can run very fast',
    'My goldfish is hungry'
]

sentences2 = [
    'The dog plays in the garden',
    'A woman watches TV',
    'The new movie is so great',
    'Jim is the fastest runner',
    'Pluto is a planet!'
]

#Compute embedding for both lists
embeddings1 = model.encode(sentences1, convert_to_tensor=True)
embeddings2 = model.encode(sentences2, convert_to_tensor=True)

#Compute cosine-similarities
cosine_scores = util.cos_sim(embeddings1, embeddings2)

#Output the pairs with their score
for i in range(len(sentences1)):
    print("{:<28} {:<28} Score: {:.4f}".format(sentences1[i],
        sentences2[i], cosine_scores[i][i]))
```

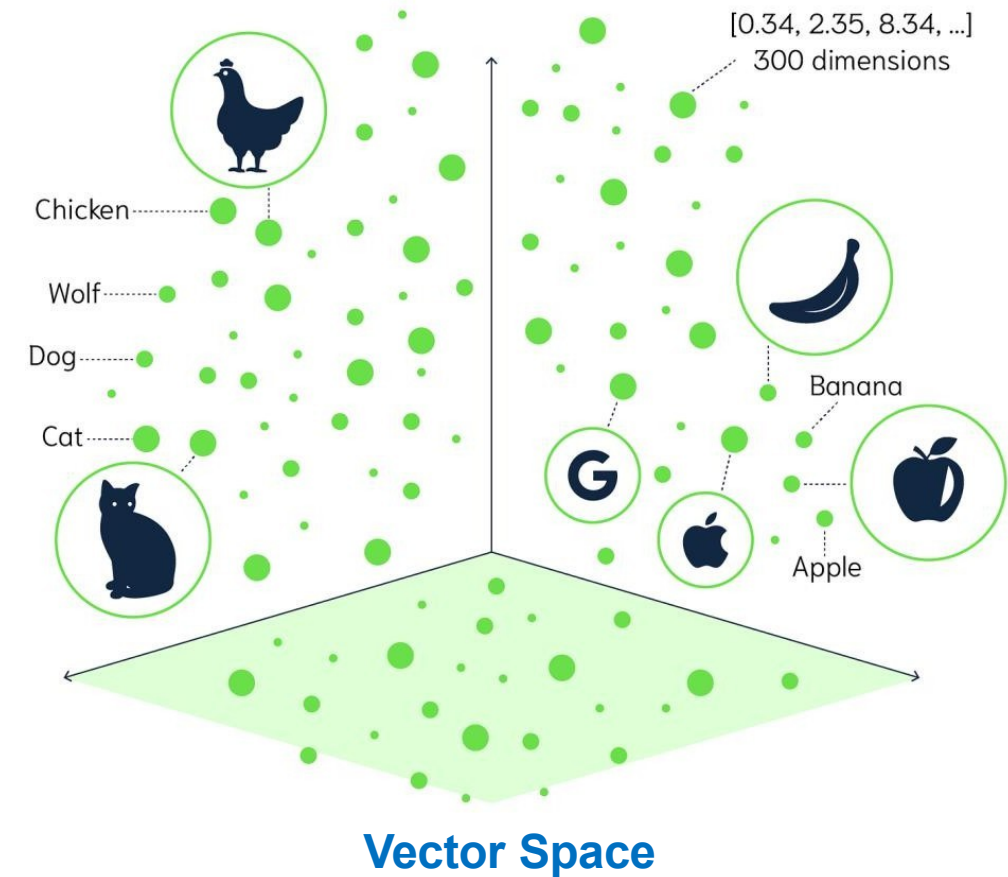
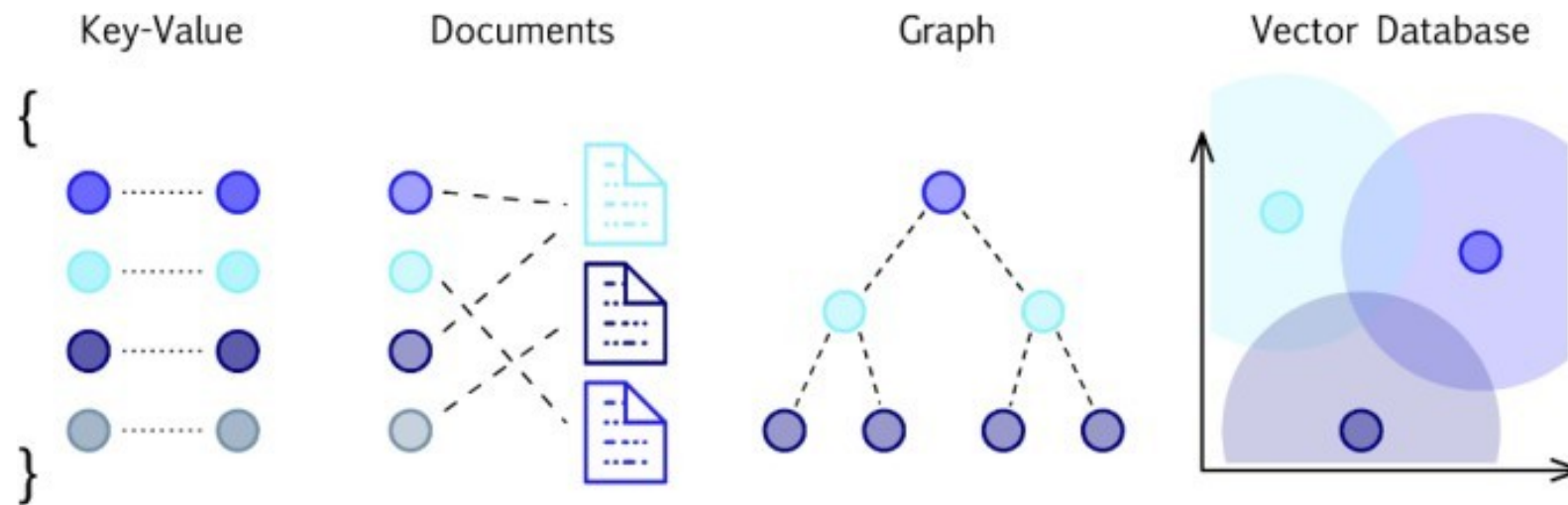


Bancos de dados vetoriais

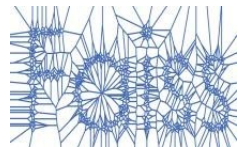
Os bancos de dados vetoriais, às vezes chamados de mecanismos de pesquisa de similaridade ou mecanismos de pesquisa do vizinho mais próximo aproximado (RNA), são bancos de dados especializados projetados para lidar com vetores de alta dimensão com eficiência.

Eles fornecem a capacidade de recuperar rapidamente os vetores mais semelhantes a um determinado vetor de consulta, o que é particularmente útil para tarefas que envolvem incorporações em grande escala.

Types of Databases



Bancos de dados vetoriais dedicados



drant



Pinecone



Chroma



Weaviate



marqo

vespa



LanceDB

Treinamento LLM

Aprendizado Semi-Supervisionado

Etapa 1: Pré-treinamento

- Não é necessária rotulagem de dados (A próxima palavra é o rótulo)
- Alta quantidade de dados genéricos (internet)
- Modelo base com habilidades linguísticas genéricas

Estágio 2: Ajuste fino

- Rotulagem de dados por seres humanos
- Precisa de dados de qualidade, normalmente QnA
- Torne-se um "Assistente" em um determinado campo

Exemplos:

- Classificação de texto
- QnA
- Resumo
- Geração de texto
- Codificação

Estágio 3 (opcional): RLHF

- Reinforcement Learning from Human Feedback
- Use o modelo para gerar respostas diferentes
- O ser humano seleciona o melhor para o ajuste fino



COMO APLICAR ISSO?



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Produtividade como eixo central

“70% dos servidores públicos brasileiros antecipam um impacto significativo da IA na produtividade das suas organizações

Fonte: tiinside

“A implementação da IA em processos governamentais pode aumentar a eficiência operacional em até 40%

Fonte: <https://flexa.cloud/ia-generativa-gestao-publica/>



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Oportunidades

- **Automatização de tarefas repetitivas**

Sequência de operação de sistemas usando dados semi-estruturados

- Apoio à criatividade e inovação

- Democratização do acesso ao conhecimento

- Exemplos:

atendimento, jurídico, saúde, educação



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Oportunidades

- Automação de tarefas repetitivas
- **Apoio à criatividade e inovação**
- Democratização do acesso ao conhecimento
- Exemplos:
atendimento, jurídico, saúde, educação

Sequência de operação de sistemas usando dados semi-estruturados

Criação rápida de alternativas a textos, imagens ancoradas em conceitos



Oportunidades

- Automação de tarefas repetitivas
- Apoio à criatividade e inovação
- **Democratização do acesso ao conhecimento**
- Exemplos:

Sequência de operação de sistemas usando dados semi-estruturados

Criação rápida de alternativas a textos, imagens ancoradas em conceitos

Vencer o paradigma de informação em excesso

atendimento, jurídico, saúde, educação



Benefícios para Clientes	Benefícios para Empresas
Atendimento 24/7 via chatbots e voicebots	Redução de custos operacionais
Respostas rápidas a dúvidas comuns	Maior eficiência no atendimento
Personalização de respostas conforme perfil	Aumento da satisfação e fidelização
Recomendações de produtos/serviços sob medida	Incremento nas vendas cruzadas e upsell
Triagem inteligente para reduzir tempo de espera	Melhor alocação de agentes humanos
Acompanhamento pós-atendimento e feedback	Monitoramento contínuo da qualidade
Clareza com resumos automáticos do histórico	Agilidade na resolução de requisições complexas
Atendimento mais humano e empático (análise de sentimentos)	Insights estratégicos sobre necessidades do cliente
Segurança em interações digitais	Redução de fraudes e riscos

atendimento, jurídico, saúde, educação



Oportunidades

- Automatização de processos repetitivos
- Apoio à tomada de decisão
- Democratização do conhecimento
- Exemplos:

Areas onde o fluxo de dados é não estruturado e o comportamento esperado nas interações pode ser copiado de exemplos de interações anteriores.

atendimento, jurídico, saúde, educação



**POR QUE AINDA NÃO ESTAMOS
USAMOS IA EM TUDO?**



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Desafios

“ Relatório do MIT:
95% dos pilotos de IA
Generativa nas
empresas estão
falhando



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL





Saídas Viesados

Vieses discriminatórios ou
resultados imprecisos

RLHF, Reinforcement
Learning com Human
Feedback

Desafio

Potencial
Solução



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Desafio

Potencial
Solução



Outputs Viesados

Vieses discriminatórios ou
resultados imprecisos



Vazamento de Dados

Dados sensíveis podem ser
extraídos via prompt ou
outras técnicas sofisticadas

RLHF, Reinforcement
Learning com Human
Feedback

Anonimização e
criptografia de dados.
Modelos encadeados



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Desafio



Outputs Viesados

Vieses discriminatórios ou resultados imprecisos



Vazamento de Dados

Dados sensíveis podem ser extraídos via prompt ou outras técnicas sofisticadas



ALucinação

Argumentos convincentes, mas errados

Potencial
Solução

RLHF, Reinforcement
Learning com Human
Feedback

Anonimização e
criptografia de dados.
Modelos encadeados

Fine-tuning e
especialização dos
modelos. Ajuste de prompt



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Desafio



Outputs Viesados

Vieses discriminatórios ou resultados imprecisos



Vazamento de Dados

Dados sensíveis podem ser extraídos via prompt ou outras técnicas sofisticadas



ALucinação

Argumentos convincentes, mas errados



Shadow AI

IA operacionalizada de maneiras inapropriadas e desconhecidas

Potencial

Solução

RLHF, Reinforcement Learning com Human Feedback

Anonimização e criptografia de dados.
Modelos encadeados

Fine-tuning e especialização dos modelos. Ajuste de prompt

Guidelines claros, políticas reforçadas e mudança cultural



Geopública 2025

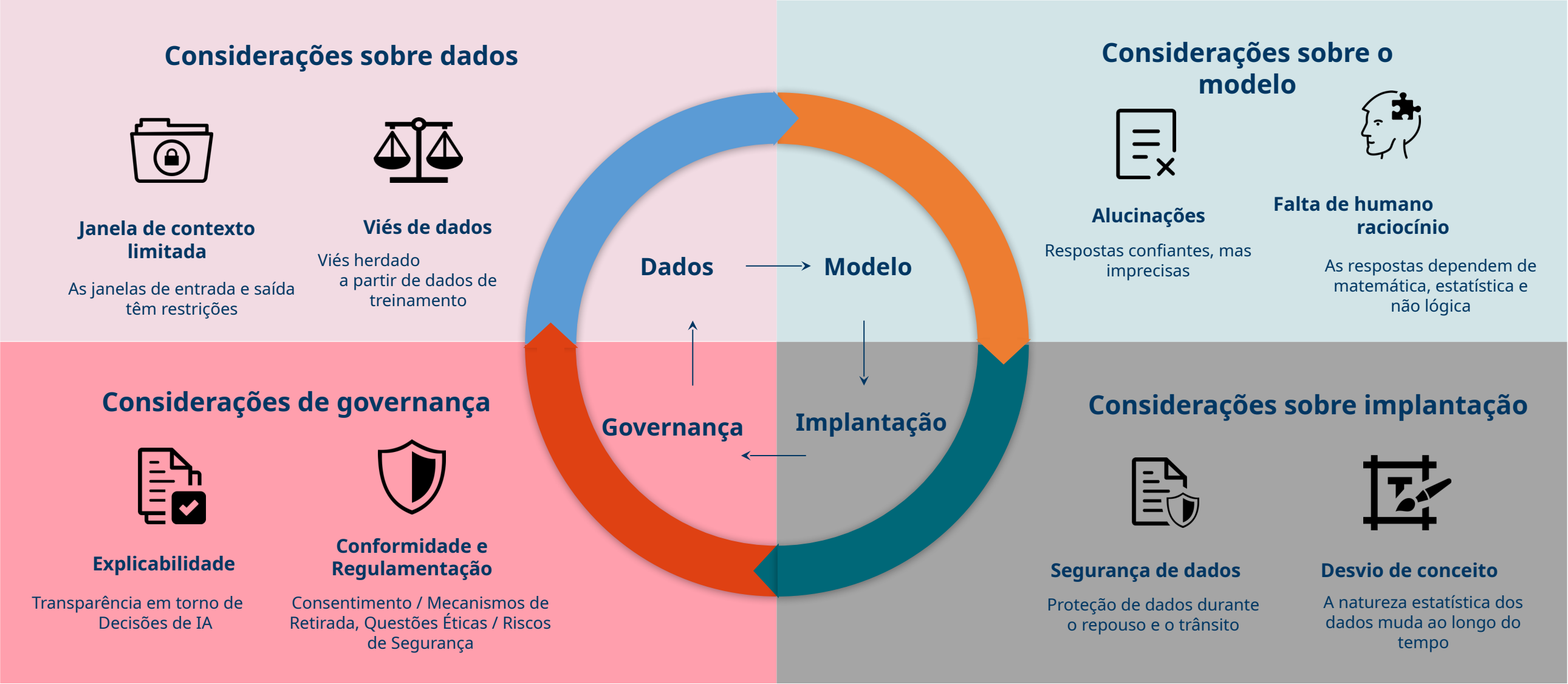


SotM BR 2025
STATE OF THE MAP BRASIL

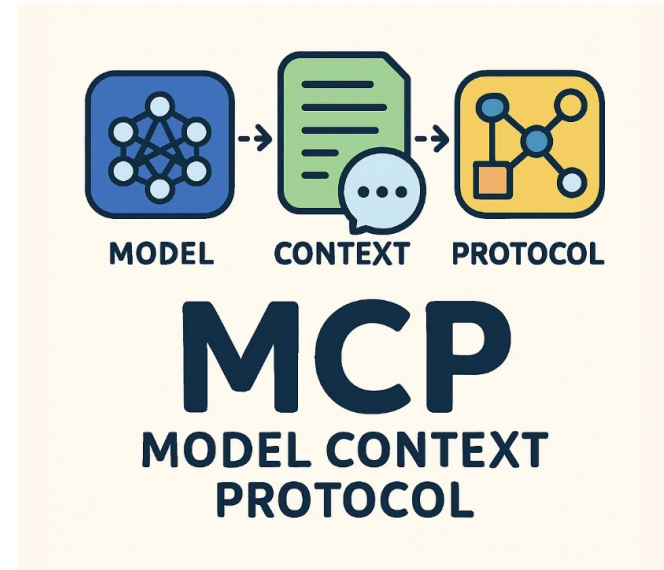
Desafio	 Outputs Viesados Vieses discriminatórios ou resultados imprecisos	 Vazamento de Dados Dados sensíveis podem ser extraídos via prompt ou outras técnicas sofisticadas	 ALucinação Argumentos convincentes, mas errados	 Shadow AI IA operacionalizada de maneiras inapropriadas e desconhecidas
	RLHF, Reinforcement Learning com Human Feedback	Anonimização e criptografia de dados. Modelos encadeados	Fine-tuning e especialização dos modelos. Ajuste de prompt	Guidelines claros, políticas reforçadas e mudança cultural

Um alerta

Desafios atuais com a IA Generativa



Outros Desafios Técnicos



- Integração em sistemas já existentes
 - Corrida para tornar sistemas existentes “compatíveis com IA – MCP”
 - Segurança – novas fronteiras de acesso.
 - Acesso similar ao RPA (Robotic Process Automation)
- Escalabilidade e custo computacional
 - Modelos + precisos → necessidade computacional \$\$\$



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Desafios Organizacionais



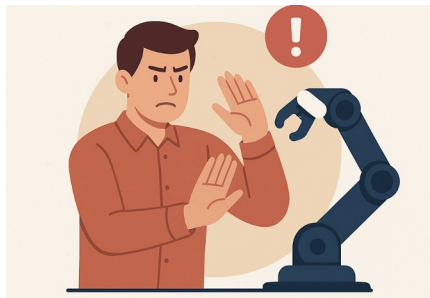
Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

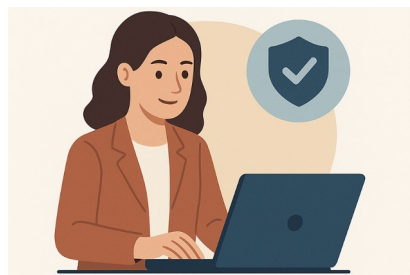
Fonte: MIT state of AI in Business, 2025

Desafios Organizacionais



Resistência **cultural** e
receio de substituição

Capacitação e novas
Habilidades Digitais



Governança interna
para uso **Responsável**

Transparência e
Responsabilização



Desafios Regulatórios e Éticos

- **Lei de Inteligência Artificial da União Europeia**

Baseada em risco: Classifica sistemas de IA em quatro níveis: risco inaceitável, alto risco, risco limitado e risco mínimo/nulo.

Risco Inaceitável: manipulação cognitivo-comportamental.

Risco Alto: uso em áreas críticas (infraestruturas essenciais, educação, emprego, serviços públicos, aplicação da lei, migração e justiça) → gestão de riscos, supervisão humana, qualidade dos dados, transparência e segurança.



ESTRATÉGIAS E FERRAMENTAS



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

Estratégias para Superar os Desafios

Alucinações e Respostas Inconsistentes

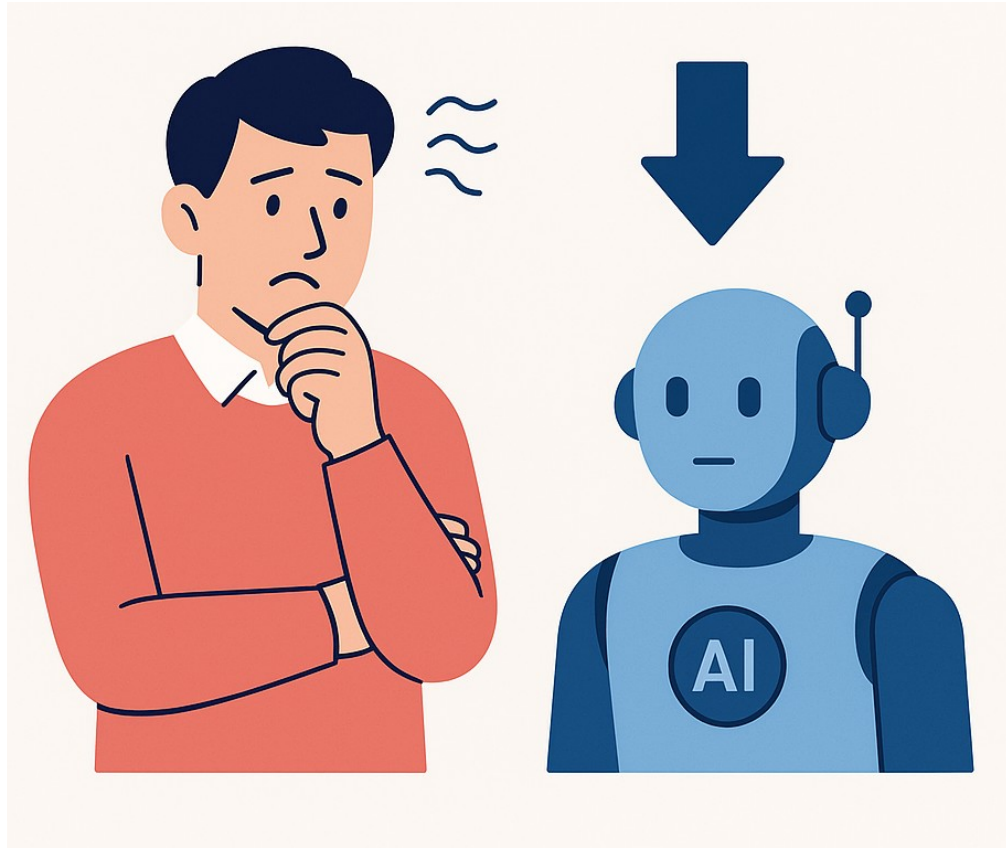
Problema: Modelos podem gerar informações incorretas ou inventadas.

Estratégias:

- Uso de human-in-the-loop (revisão humana em decisões críticas).
- Integração com bases de dados confiáveis (RAG – Retrieval-Augmented Generation).
- Adoção de métricas de acurácia e confiabilidade.



Estratégias para Superar os Desafios



Resistência Cultural e Medo de Substituição

Problema: Equipes podem rejeitar o uso da IA por receio de perda de emprego.

Estratégias:

- Comunicação clara: IA como copiloto, não substituto.
- Programas de reskilling/upskilling.
- Incluir trabalhadores na definição das regras de uso.

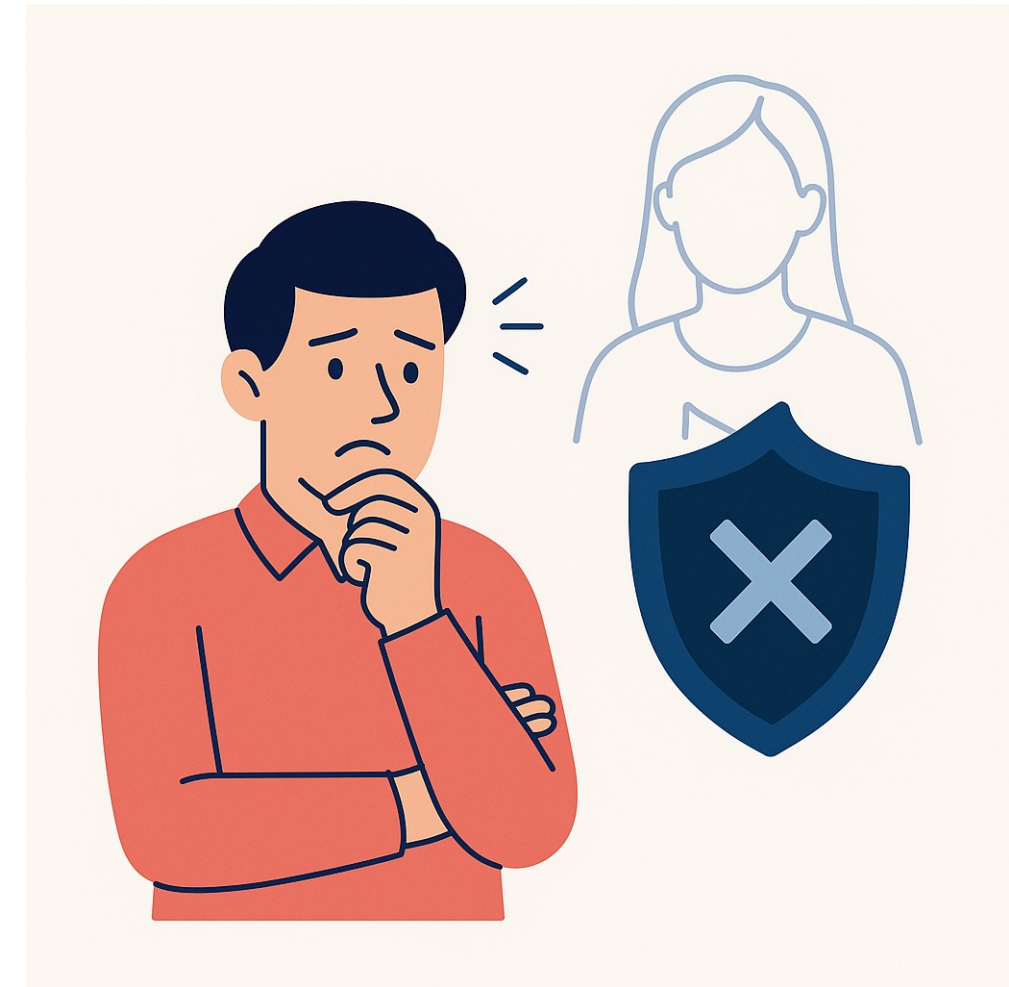
Estratégias para Superar os Desafios

Falta de Governança e Transparência

Problema: Uso indiscriminado sem diretrizes pode gerar riscos éticos e reputacionais.

Estratégias:

- Criação de comitês de ética e governança da IA.
- Estabelecimento de políticas internas claras.
- Ferramentas de auditoria e rastreabilidade dos outputs.



Estratégias para Superar os Desafios



Dependência Tecnológica e Custos

Problema: Soluções de IA podem ser caras e criar dependência de poucos fornecedores.

Estratégias:

- Avaliar soluções open source (LLaMA, Mistral, Falcon, qwen, gemma).
- Uso de infraestrutura híbrida (nuvem + local).
- Monitorar ROI e eficiência operacional continuamente.

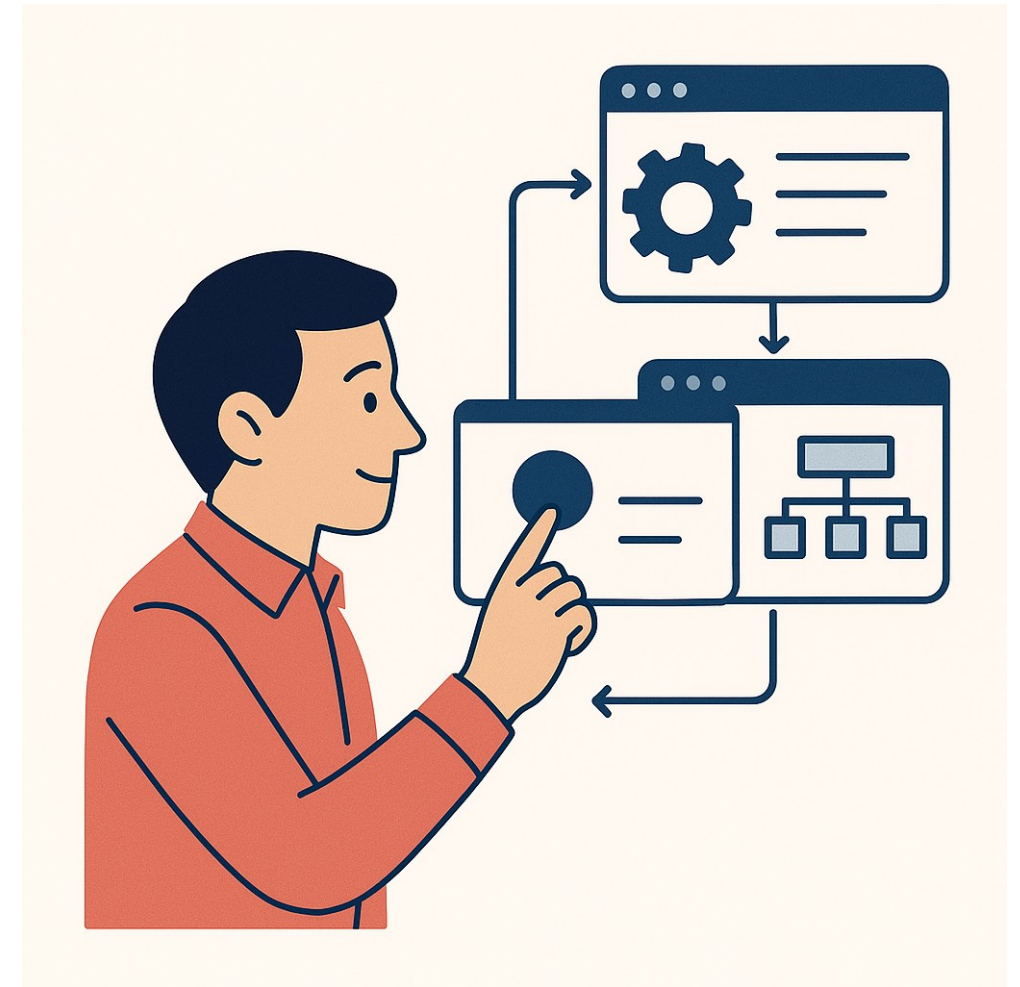
Estratégias para Superar os Desafios

Integração com Sistemas Existentes

Problema: Dificuldade de integrar IA a sistemas legados (ERP, CRM, plataformas públicas).

Estratégias:

- Uso de APIs padronizadas e middleware.
- Projetos-piloto de integração gradual.
- Definição de arquitetura de referência para IA corporativa.



COMO TER UMA IA PARA MEUS DADOS?



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

O que é RAG?

*Retrieval **A**ugmented **G**eneration (Geração Aumentada por Busca)*

>95% dos dados do mundo são privados!

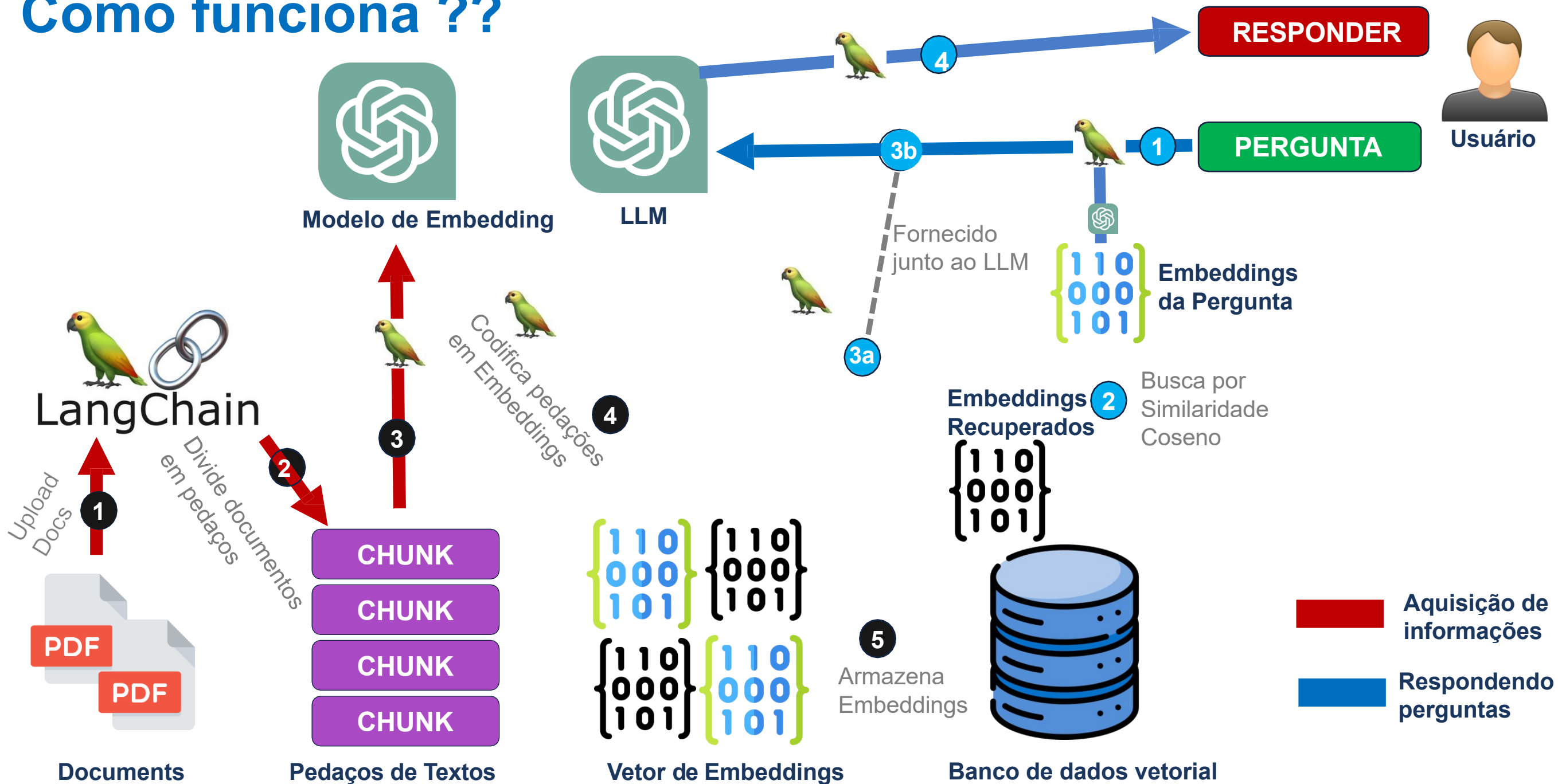
Capacidade de capacitar o LLM (que está congelado no tempo!) com:

- Informações proprietárias – por exemplo, normas internas
- Dados recentes não conhecidos quando o LLM foi treinado
- Fatos relevantes conhecidos → reduzir as alucinações

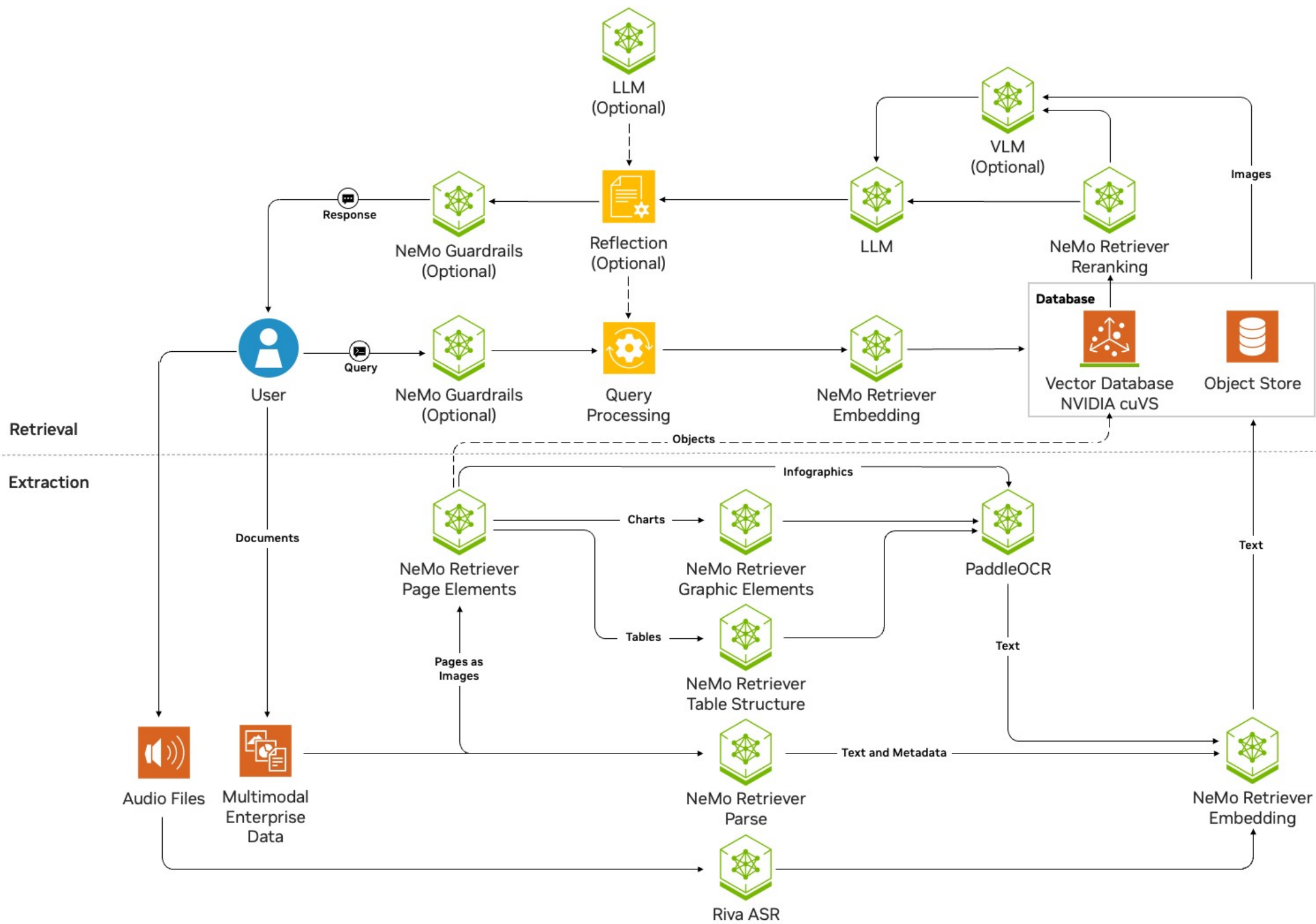
Possivelmente equipar o LLM com memória de longo prazo



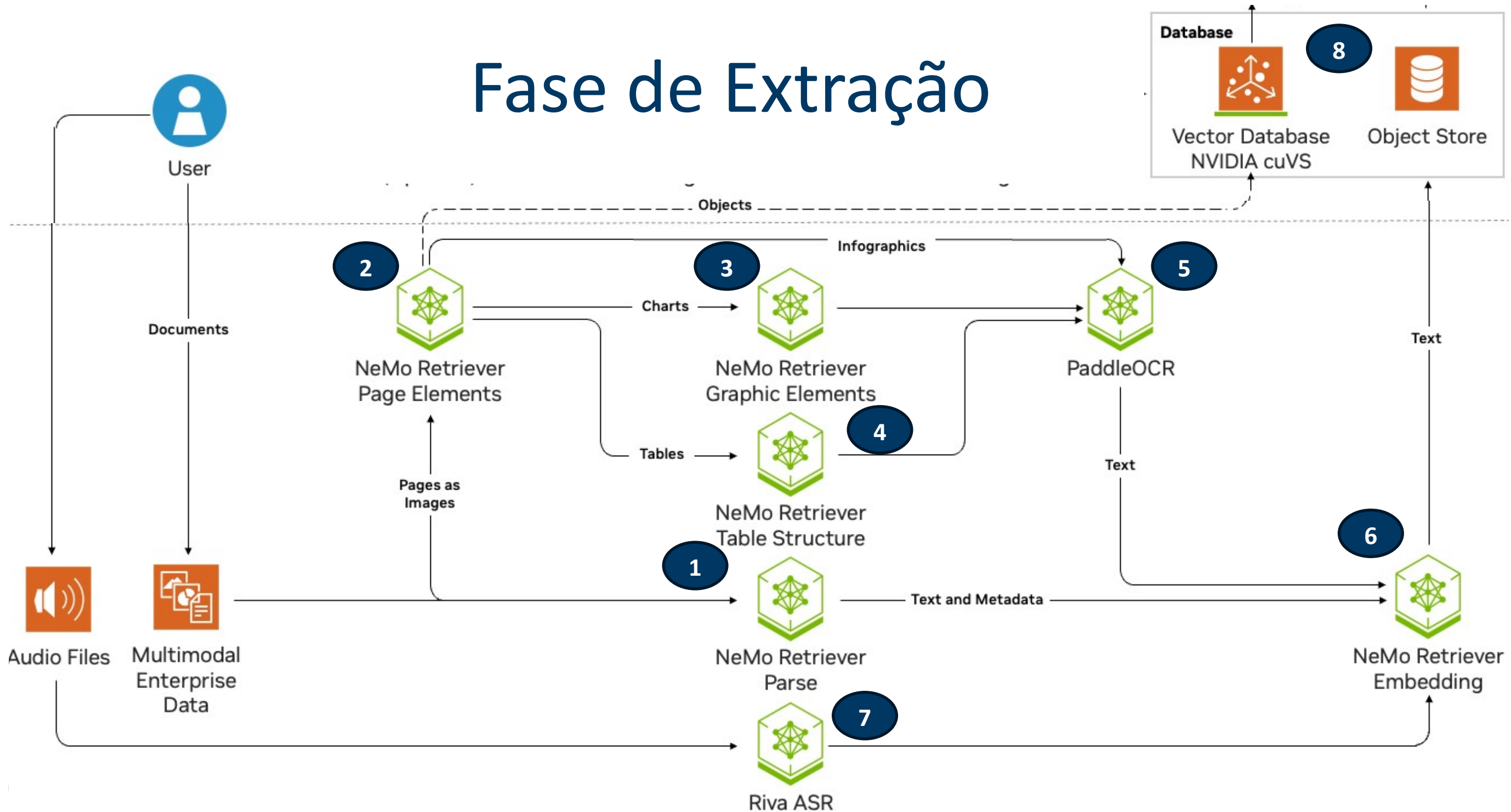
Como funciona ??



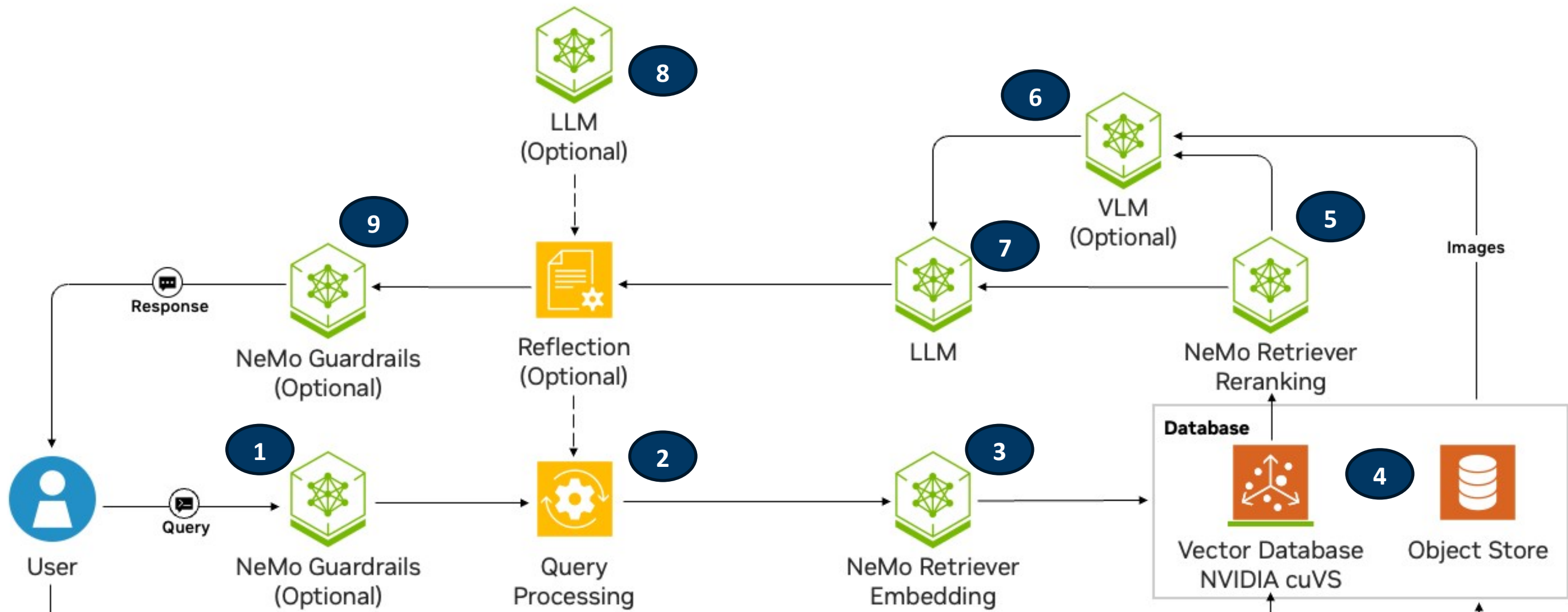
NVIDIA RAG Blueprint



Fase de Extração

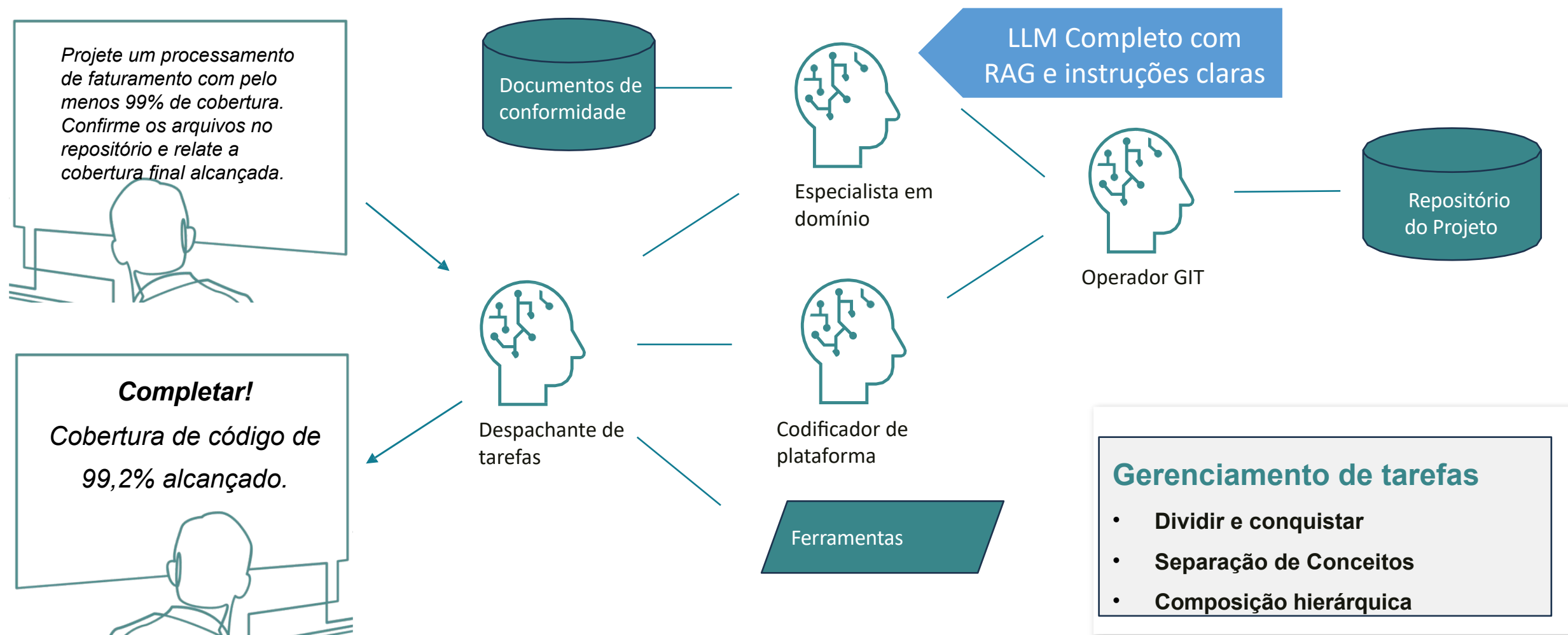


Fase de Recuperação



Futuro da IA Generativa no Trabalho

Multi Agent Workflow Automation – Agentic System



Futuro da IA Generativa no Trabalho

//

É a transformação do trabalho. Os empregos humanos serão substituídos – mas serão substituídos por outros humanos usando IA. A ideia é o humano usar a IA, enquanto ela aprende e faz acontecer.

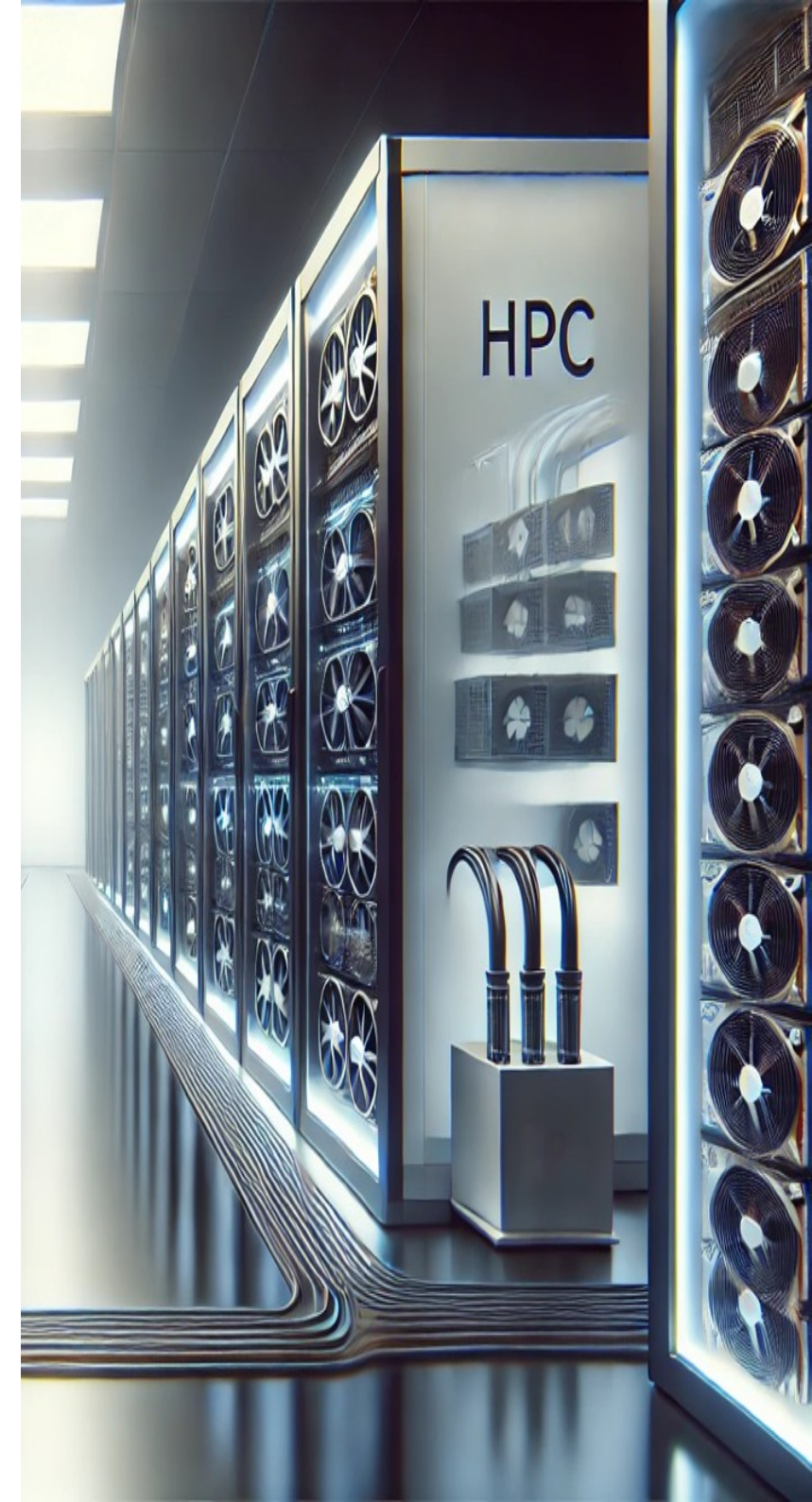
John Graham

//



Pontos Chave

- IA generativa é **oportunidade + desafio**
- Adoção exige equilíbrio entre **técnica, organização e regulação**
- Preparar **pessoas, processos e políticas** é essencial



EVIDEN

Obrigado

Perguntas?

Genaro Costa
genaro.costa@eviden.com
M +71 987 907 877



Geopública 2025



SotM BR 2025
STATE OF THE MAP BRASIL

