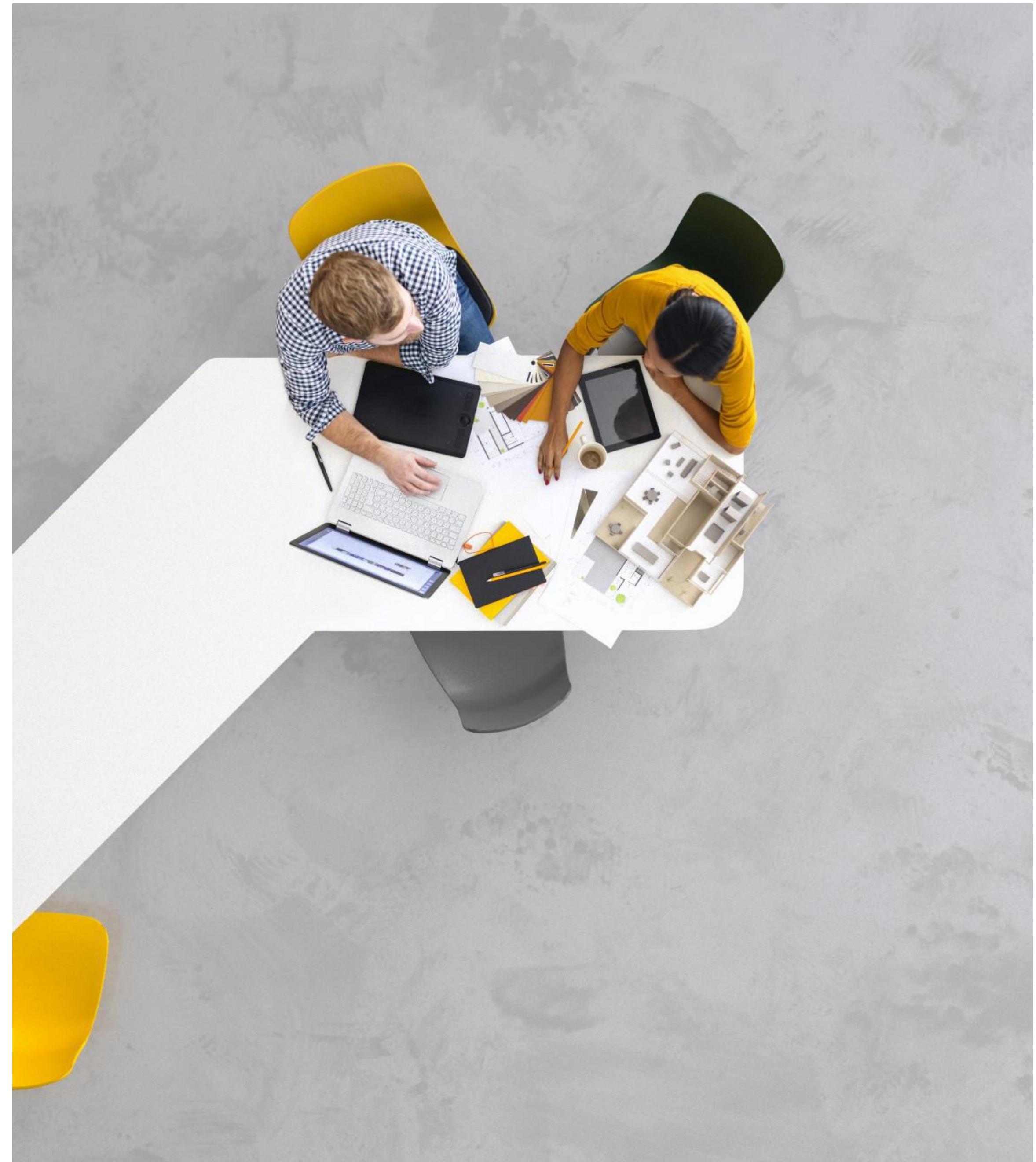


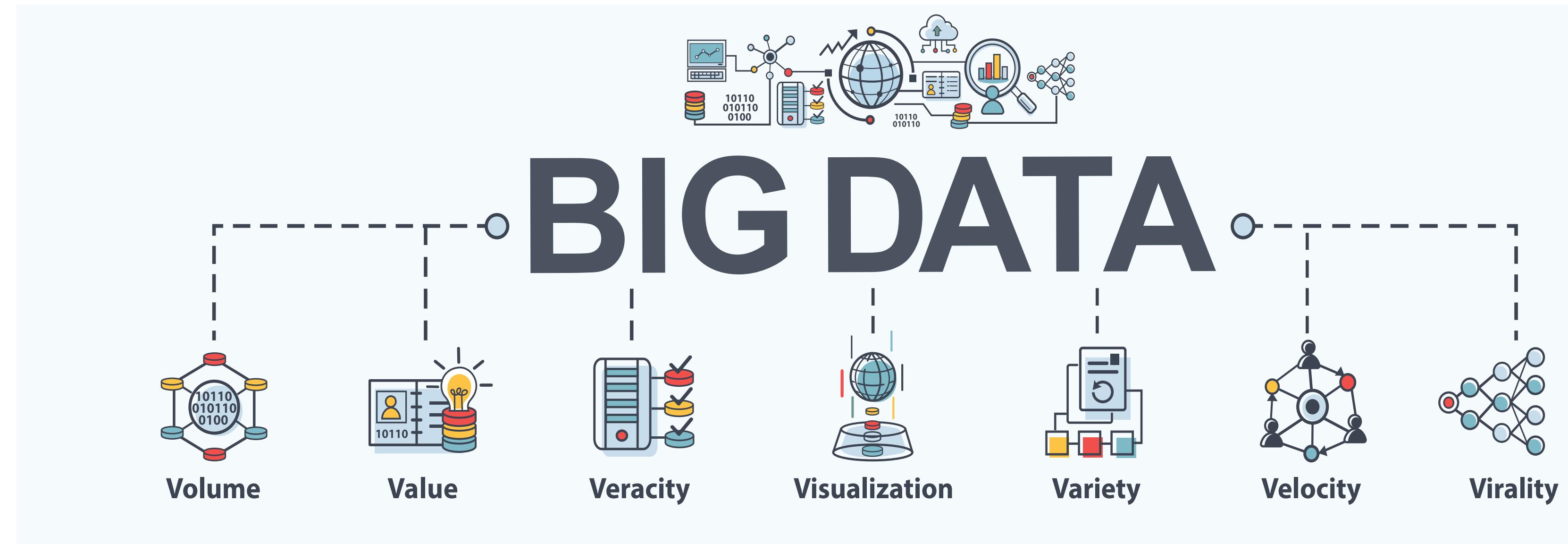
Usos da IA e Prompt Engineering

Parte IV

Glauco Reis



O que é BIG DATA?



Big Data é um termo adotado para representar atividades analíticas em um espectro mais amplo do dado, em sua maioria, para operações que envolvem um grande volume de informações.

Referimos sempre ao termo como uma agregação dos V's: **Volume**, **Veracidade**, **Variedade** e **Velocidade**. Recentemente, refere-se Big Data as atividades analíticas que englobam Volume, Valor, Veracidade, Visualização, Variedade, Velocidade e Viralidade

O que é Ciência de Dados?



 **Josh Wills**
@josh_wills Seguir

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Traducir del inglés

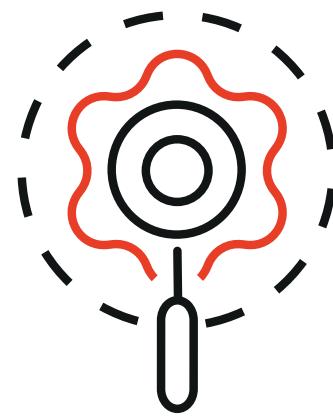
9:55 - 3 may. 2012

1.686 Retweets 1.417 Me gusta

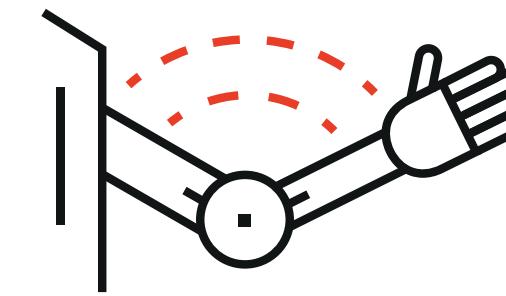
51 1,7K 1,4K

Ciência de dados é um área interdisciplinar que envolve o estudo dos dados e informações inerentes ao negócio, visando a extração de conhecimento, detecção de padrões, processamento, otimização, automação, transformação e análise de dados. É uma área que envolve as disciplinas de matemática, estatística, computação e conhecimento do negócio.

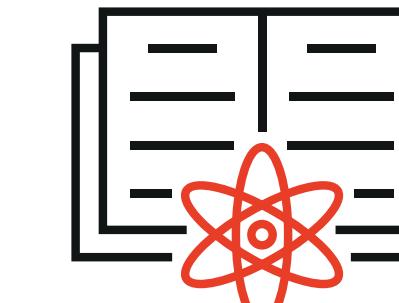
Quais problemas podem ser resolvidos
utilizando Ciência de Dados?



COGNITION



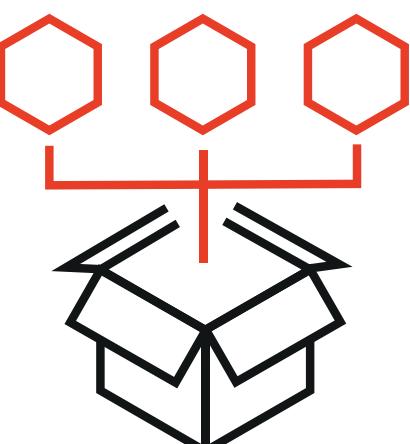
SENSORIMOTOR
SKILLS



AI
KNOWLEDGE



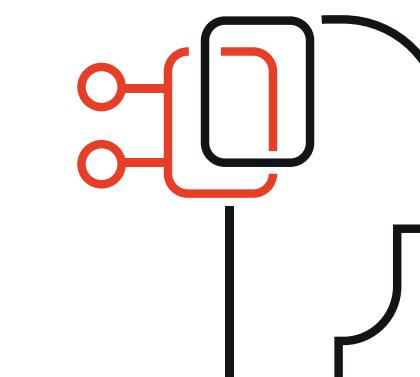
EXPERT SYSTEM



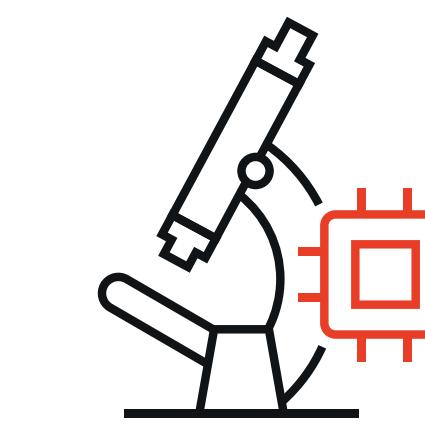
KNOWLEDGE
REPRESENTATION



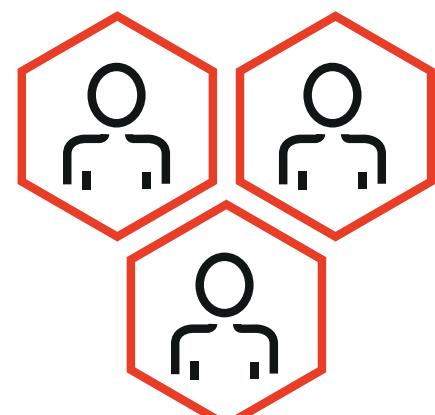
AUTOMATED
PLANNING



COMPUTATIONAL
INTELLIGENCE



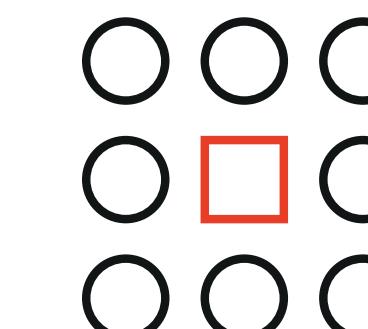
DEEP LEARNING



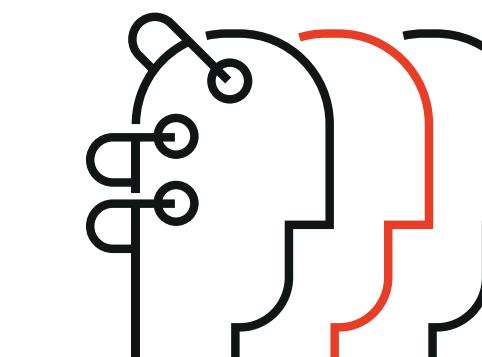
MULTI-AGENT
SYSTEM



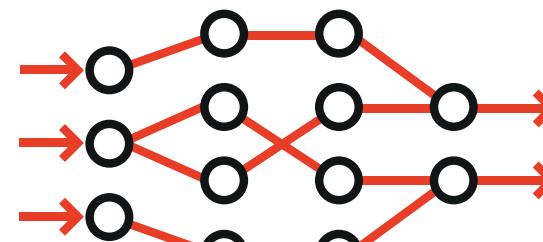
AI APPLICATIONS



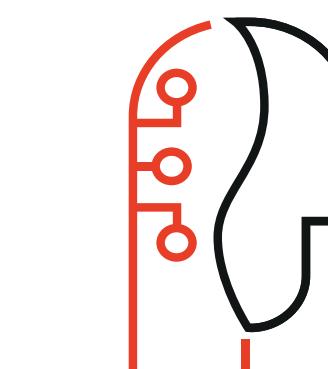
PATTERN
RECOGNITION



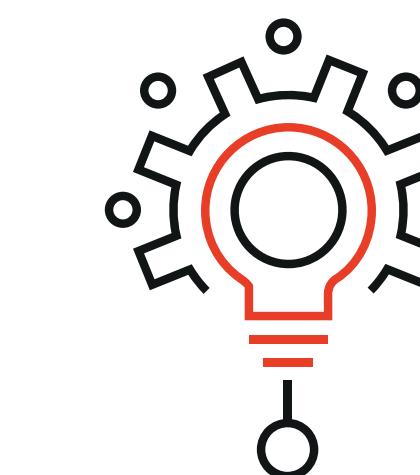
INTELLIGENT
AGENTS



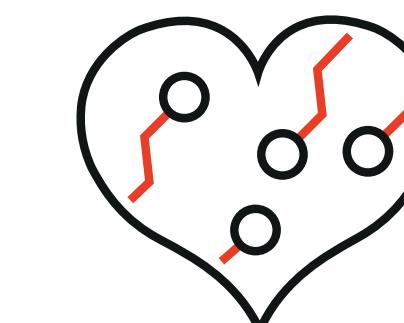
ARTIFICIAL
NEURAL
NETWORK



HUMAN-COMPUTER
INTERACTION



AI RESEARCH



EMERGENT
BEHAVIOR

Prever o que irá acontecer...

- ...baseado em **dados históricos**.



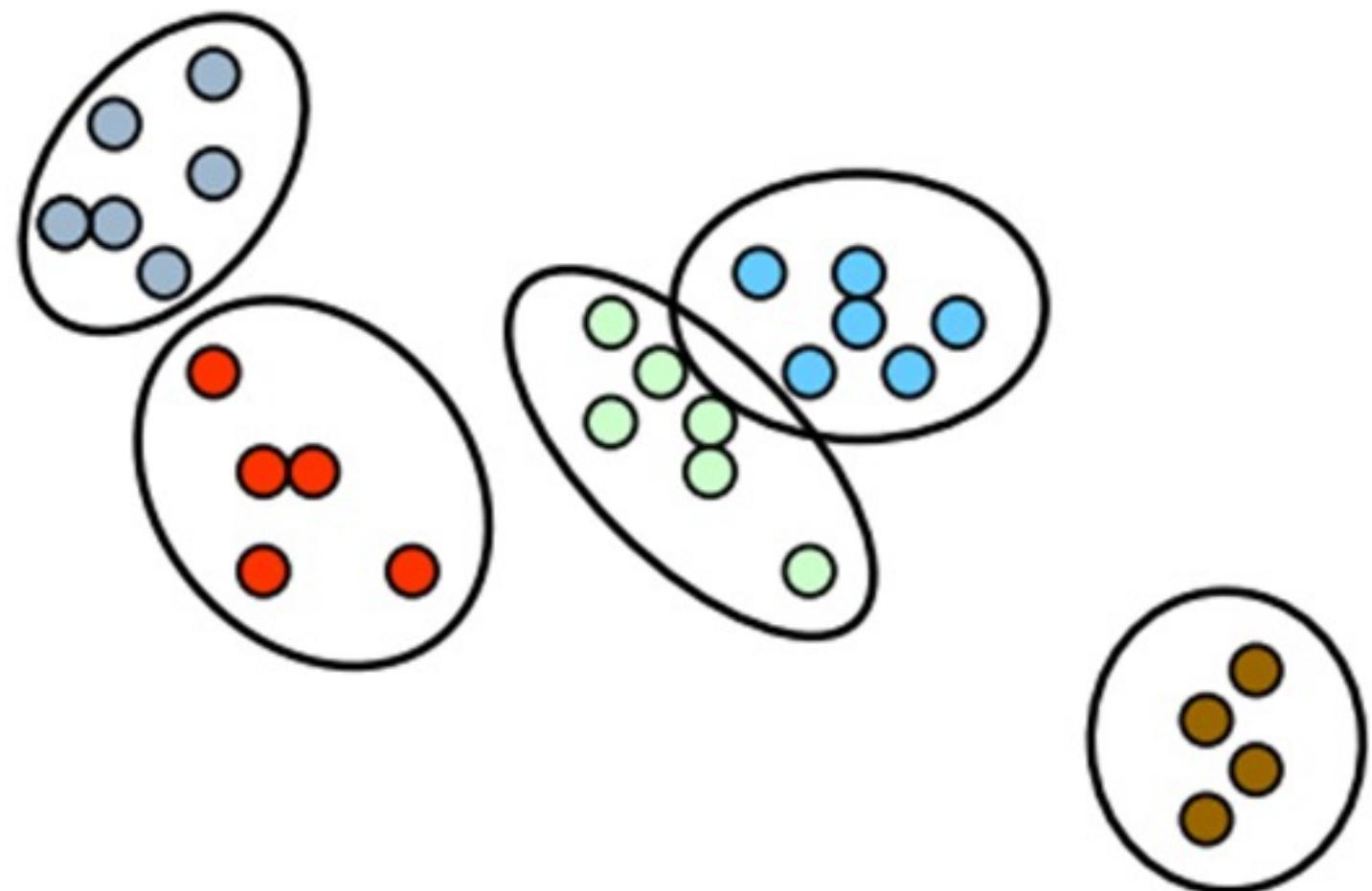
- ...baseado em **tendência** de dados antigos.

EXEMPLOS

- **Quantos** incidentes espera-se que sejam criados na **próxima segunda?**
- **Quanto tempo** leva-se geralmente para criar **esse tipo de incidente?** (Problemas de estimativa)
- Prever a necessidade computacional necessária em um futuro próximo.
- Previsão de falhas

Categorizar coisas automaticamente...

- ...através de **reconhecimento de padrões.**



EXEMPLOS

- **Agrupar itens similares**
 - Agrupamento de filas sendo trabalhadas em incidentes de problemas similares;
 - Classificação de clientes;
 - Reconhecimento de padrões.

Relacionamento entre itens...



- ...Através de análise de redes

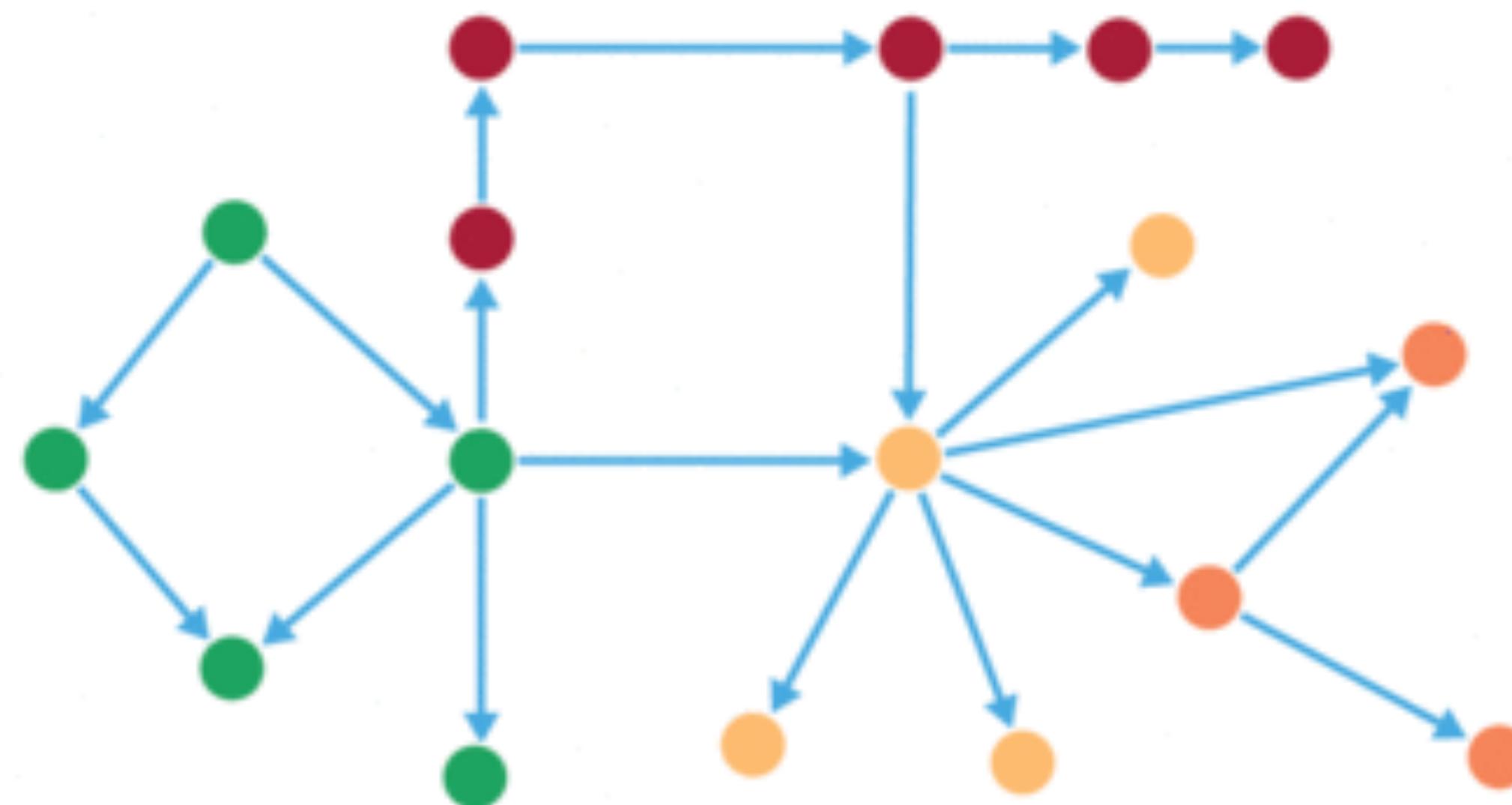


EXEMPLOS

- **Agrupamento de itens ou grupos que trabalham em conjunto**
 - Relacionamento de clientes ou times para recomendação
 - Entendimento do **impacto** entre um time ou outro na cadeia produtiva

Dependência entre coisas

- ...através de **redes Bayesianas**



EXEMPLOS

- Qual o **impacto** que um **ítem de configuração** pode ter? (Análise de suporte ao cliente)
- Quais **itens** pode ter **causado** uma falha em um determinado componente? (Análise de causa raiz).

Fornecer sugestões ou recomendações

- ...através de **sistemas de recomendação**



EXEMPLOS

- Recomendar soluções de resolução de problemas;
- Sugerir itens similares em um processo de vendas;

Detectar anormalidades

- ...através de algoritmos de **detecção de anomalias**



EXEMPLOS

- Detectar **anomalias** em **dados de monitoramento** ou de **desempenho**:
 - Auxilia na prevenção de falhas, manutenção preventiva, etc.

Tipos de algorítmos de aprendizado de máquina



Treinados com supervisão humana ou por rótulos/anotação nos dados

- Supervisionados
- Não-supervisionados
- Aprendizado por reforço

Podem aprender incrementalmente ou em tempo-real

**Online vs.
Batch Learning**

Como realiza o processo de generalização

- Baseado em instância
- Aprendizado baseado em modelo

Tipos de algorítmos de aprendizado de máquina



● **ALGORÍTMOS SUPERVISIONADOS**

- Há existência de uma variável alvo como rótulo ou sinal do que se deseja prever, classificar, recomendar, etc.
- Tem-se conhecimento prévio nos dados de treinamento sobre a “resposta”, sendo possível validar o resultado entre o valor observado e o previsto.

● **ALGORÍTMOS NÃO-SUPERVISIONADOS**

- Não há nenhum rótulo ou anotação sobre os dados.
- Nesse caso, procura-se por alguma estrutura ou padrão no dado de acordo com algum critério ou característica específica.

Classificação de câncer de pele

SPAM FILTERING

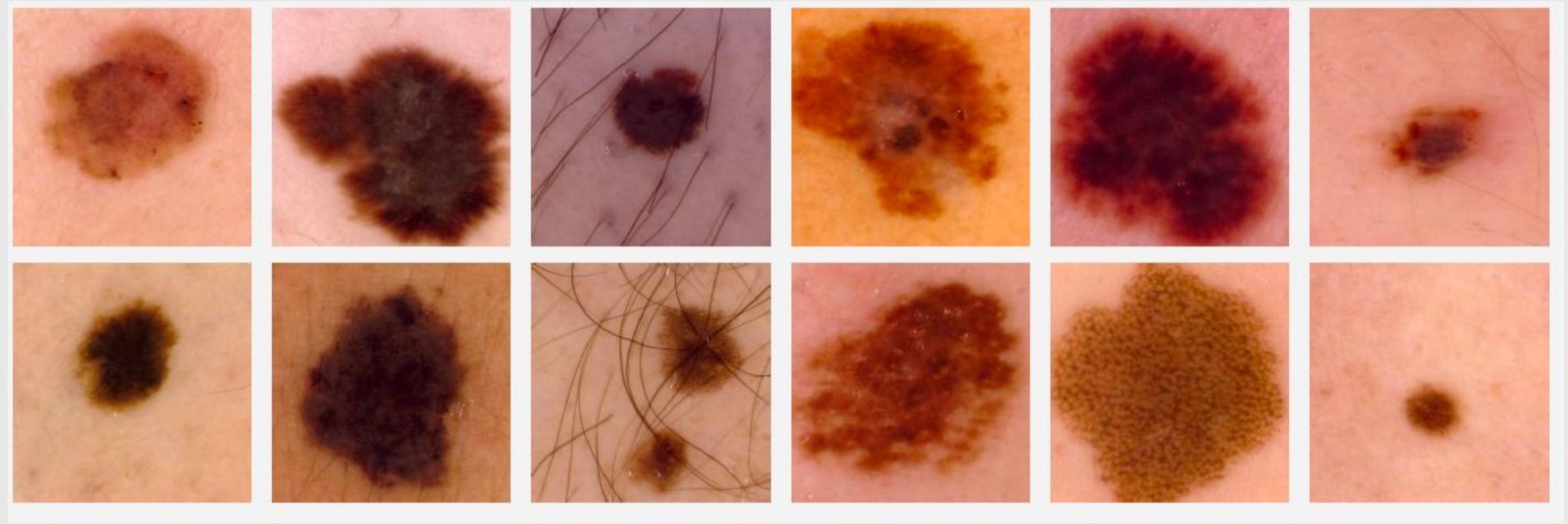


Bad Cures fast and effective! - Canadian *** Pharmacy #1 Internet
Inline Drugstore Viagra Cheap Our price \$1.99 ...

Good Interested in your research on graphical models - Dear Prof., I
have read some of your papers on probabilistic graphical
models. Because I ...

Exemplo de problemas: Supervisionado

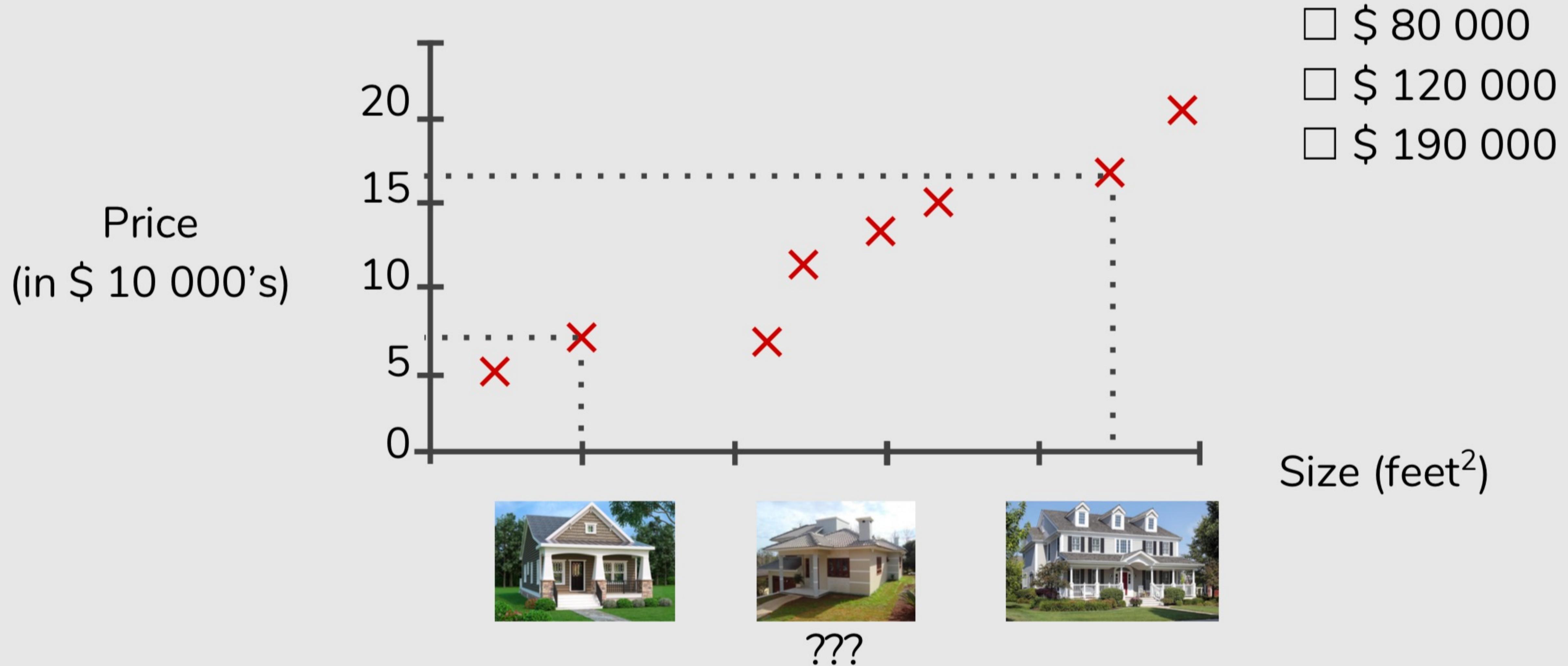
Classificação de câncer de pele



Melanomas (linha superior) e lesões **benignas** (linha inferior)

Estimativa de preços de imóveis

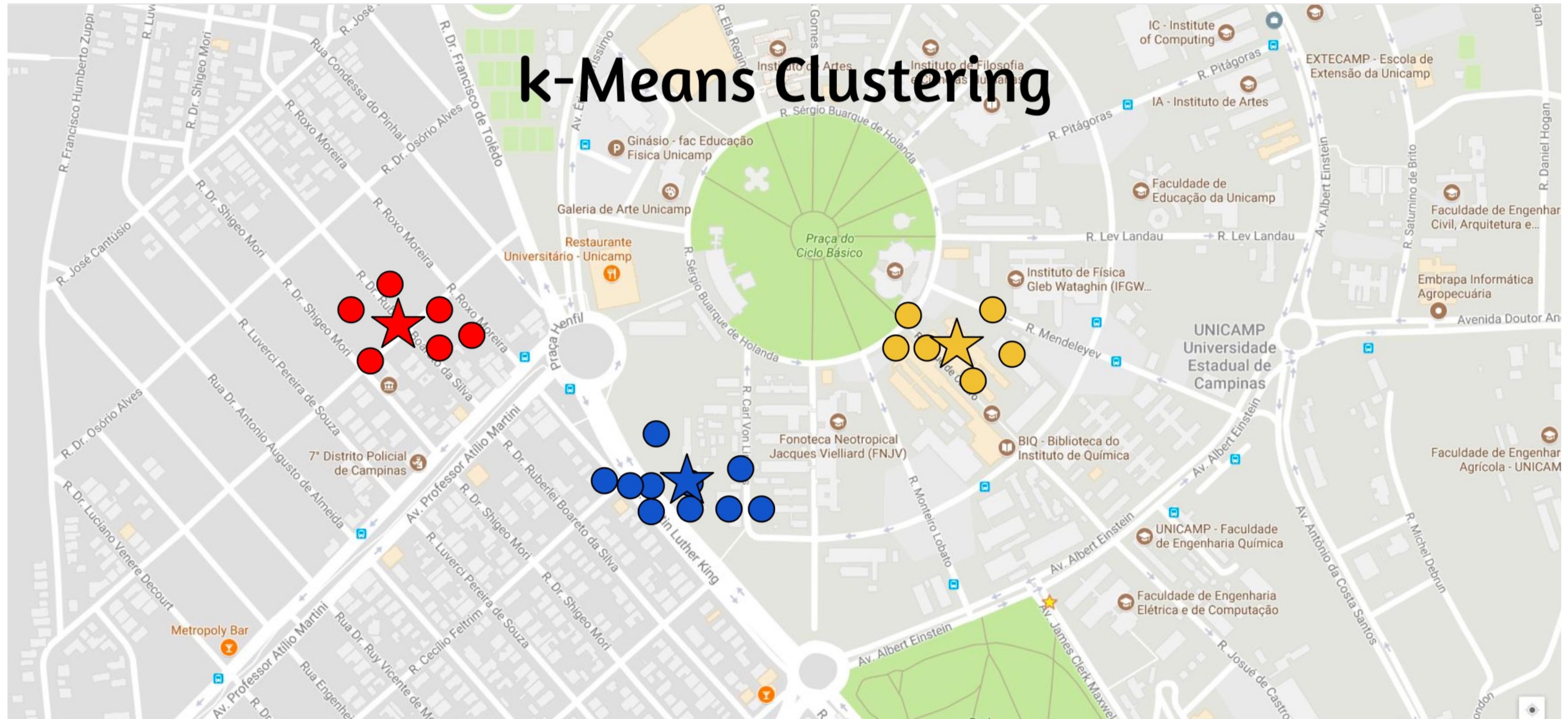
Problemas de estimativa: Regressão

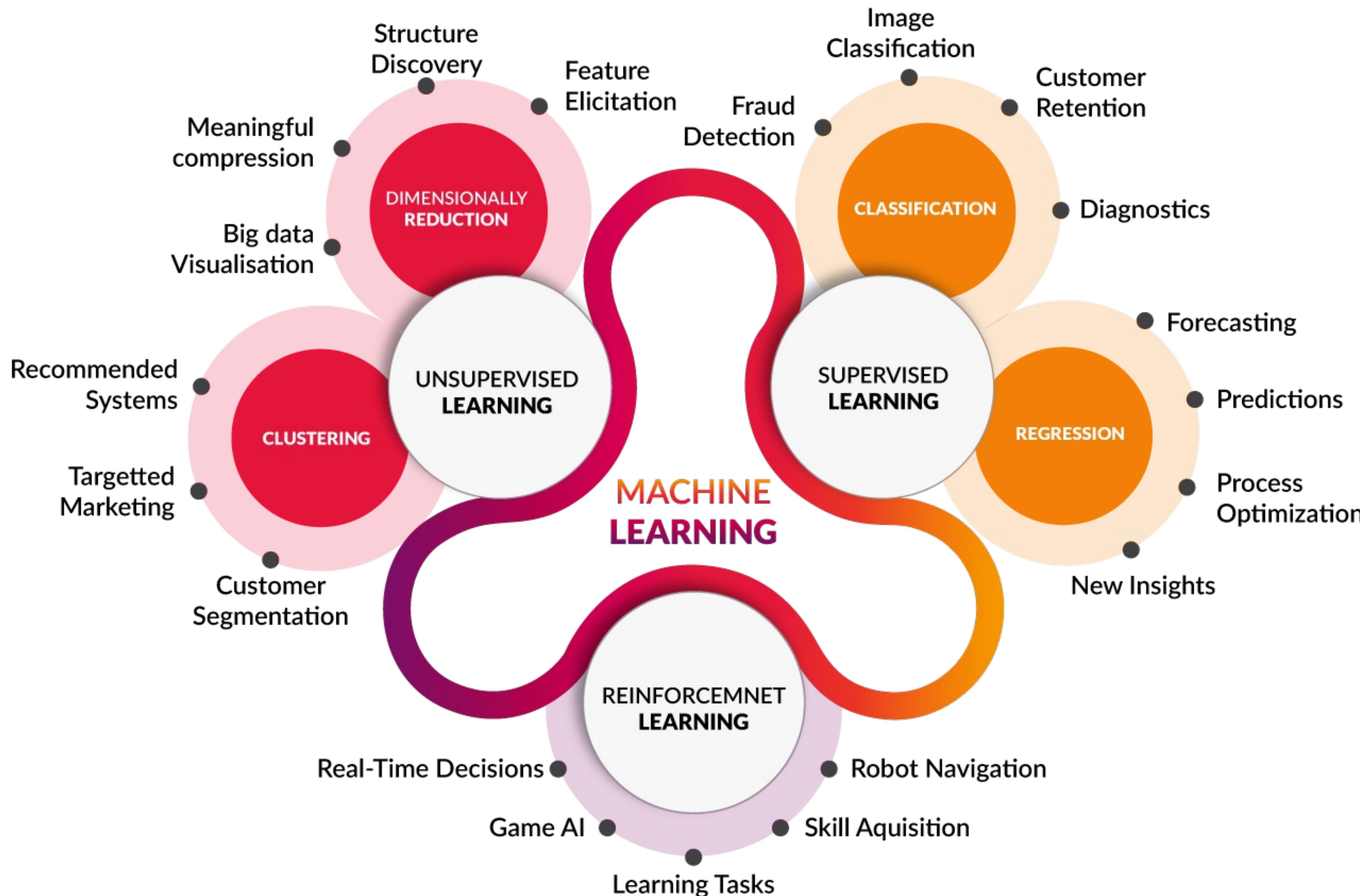


Exemplo de problemas: Não-Supervisionado

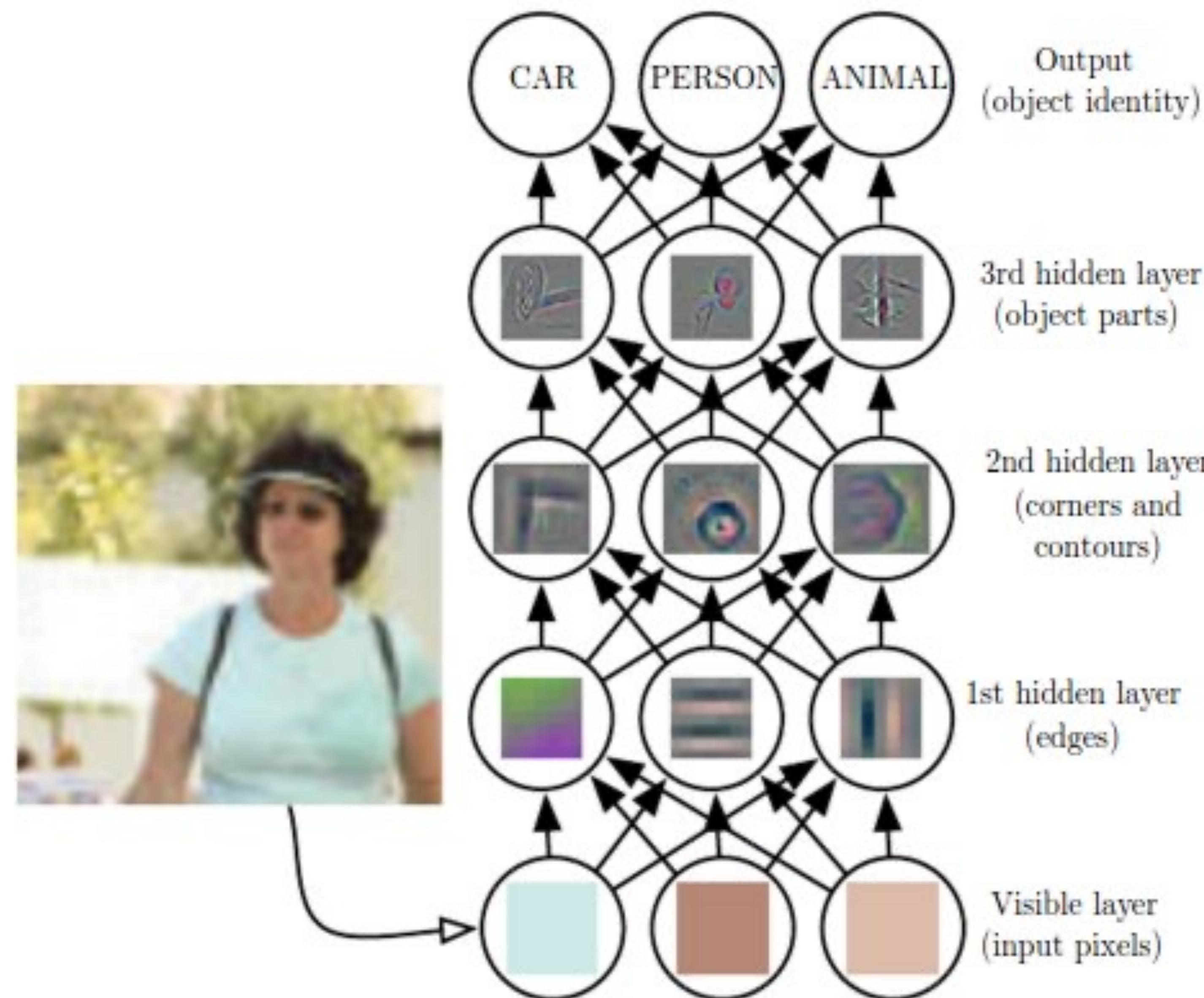
Agrupamento de restaurantes por tipo de cardápio

k-Means Clustering





Feature Extractions on a Deep Neural Network



Multi-class Classification



Cat



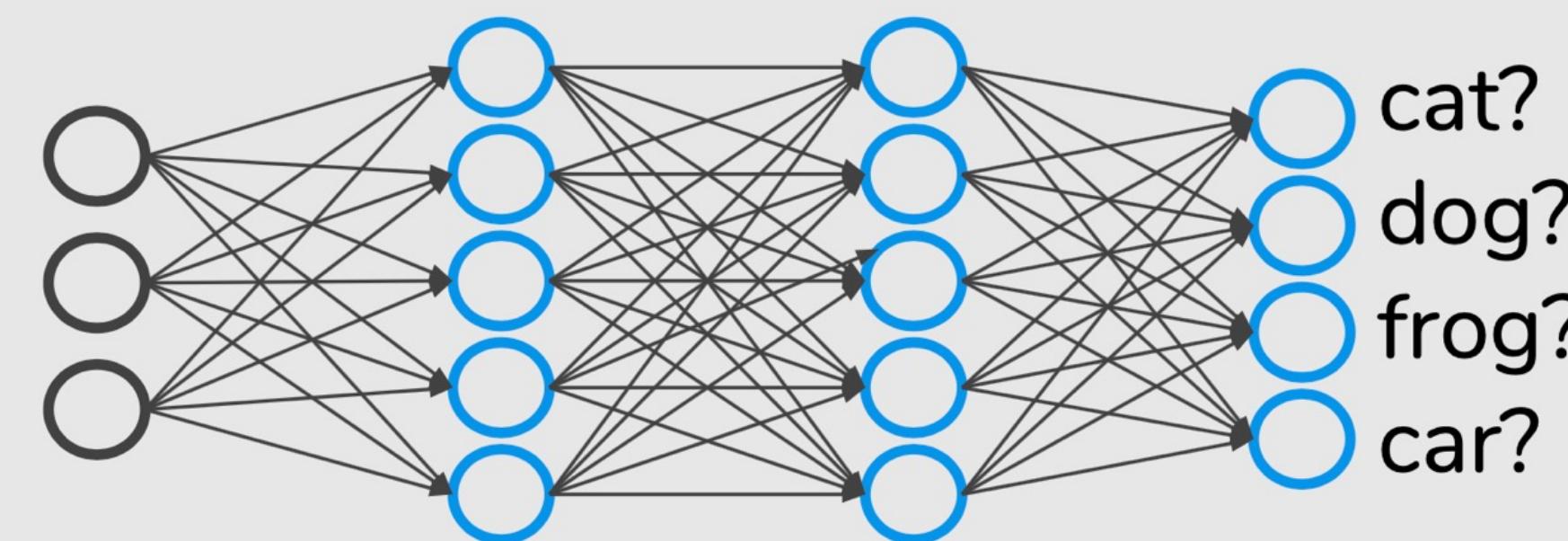
Dog



Frog



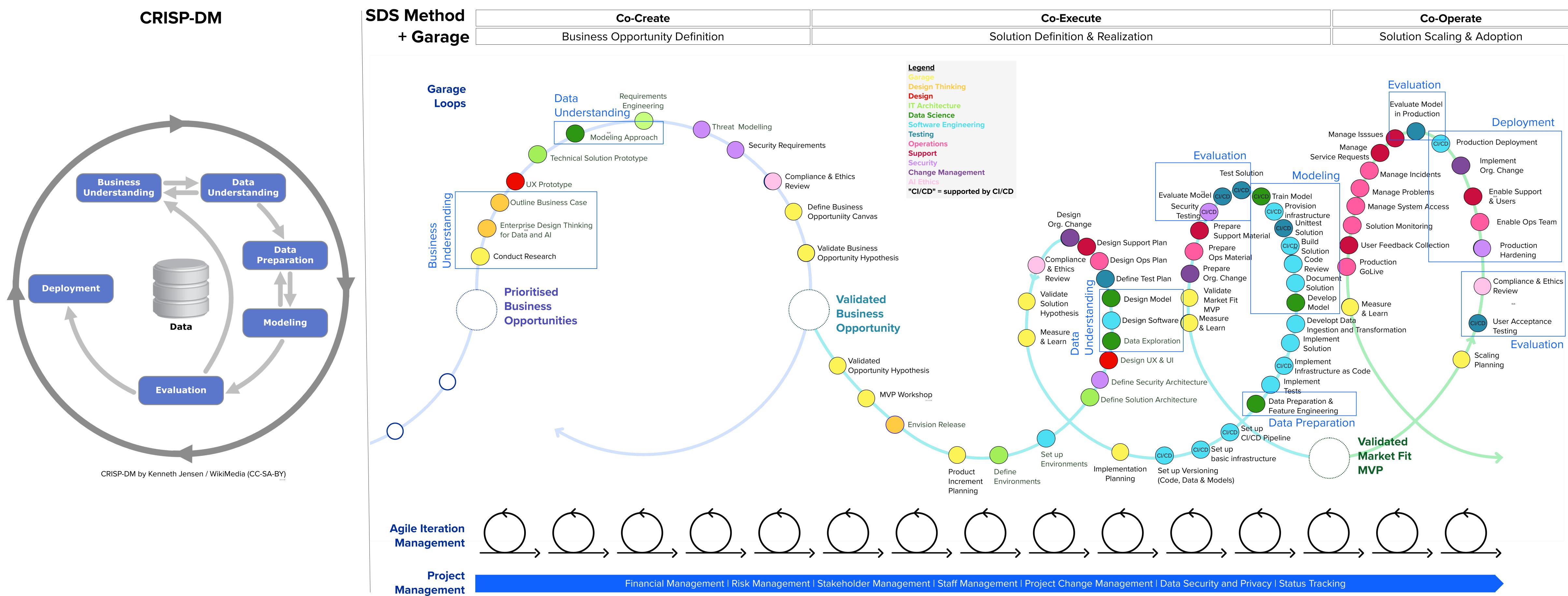
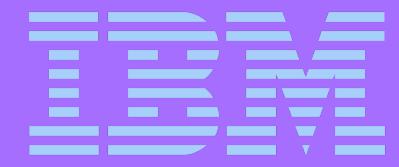
Car



Want $h_{\Theta}(x) \approx \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$, $h_{\Theta}(x) \approx \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$, $h_{\Theta}(x) \approx \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$, $h_{\Theta}(x) \approx \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$

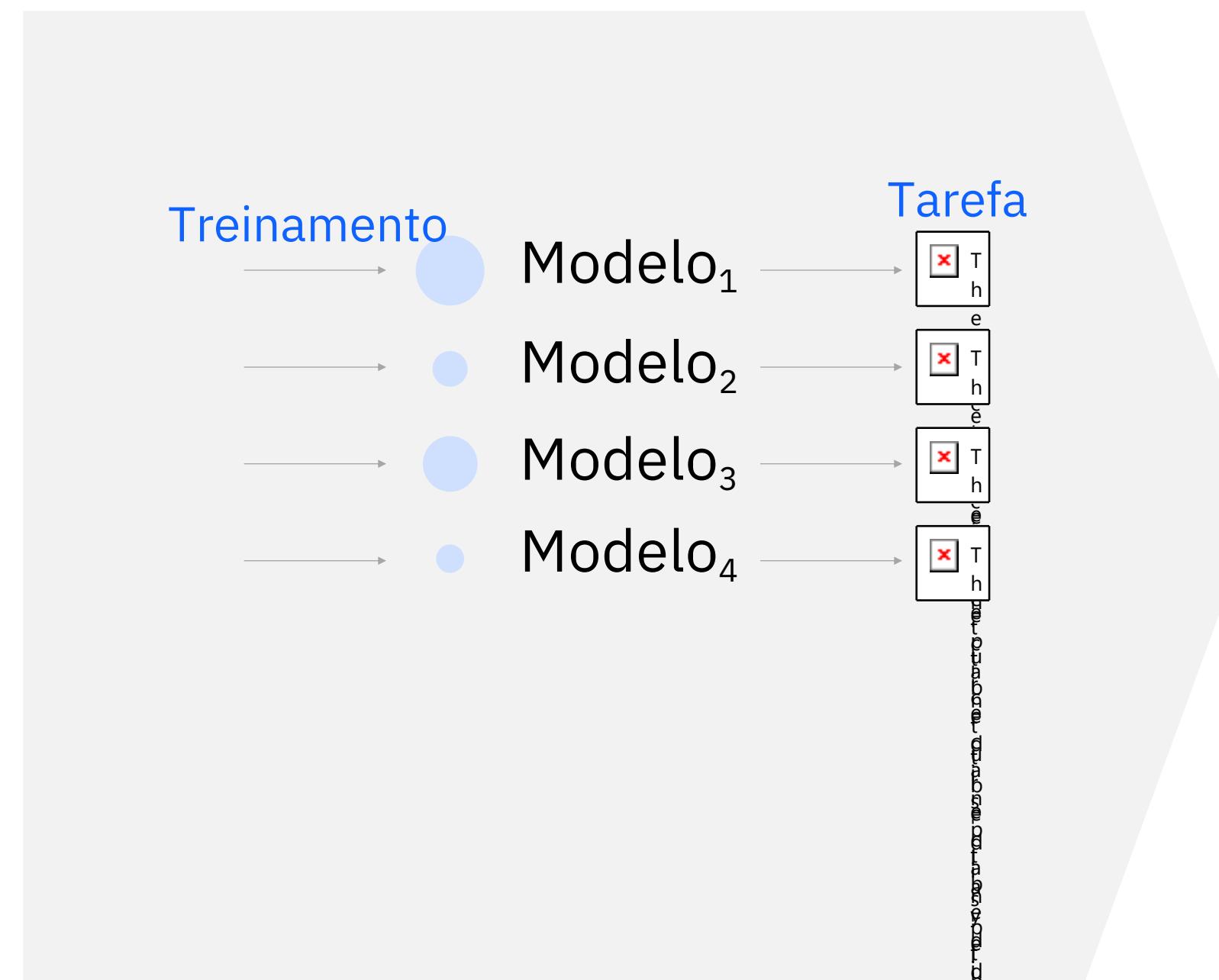
Crispier-DM

Scaled Data Science Method expands CRISP-DM with Software Engineering, Design Thinking, and many more



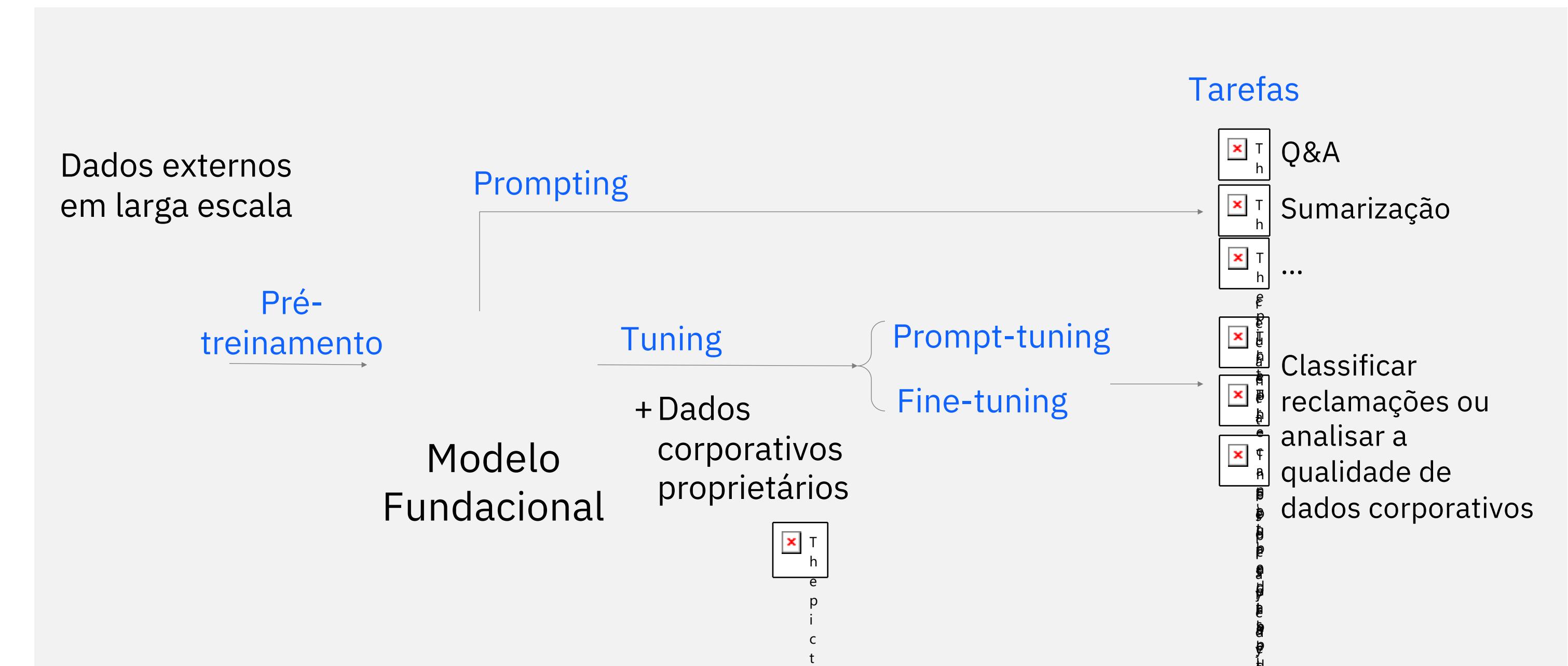
Os Modelos Fundacionais estão permitindo um novo paradigma no desenvolvimento eficiente de Inteligência Artificial em escala

Modelos de ML / DL convencionais

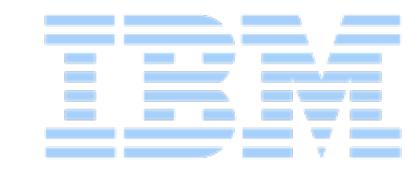


- Modelos de silos individuais
- Requer treinamento específico para cada tarefa
- Requer treinamento supervisionado

Modelos Fundacionais



- Adaptação rápida a várias tarefas com pequenas quantidades de dados específicos da tarefa
- Aprendizagem não supervisionada pré-treinada



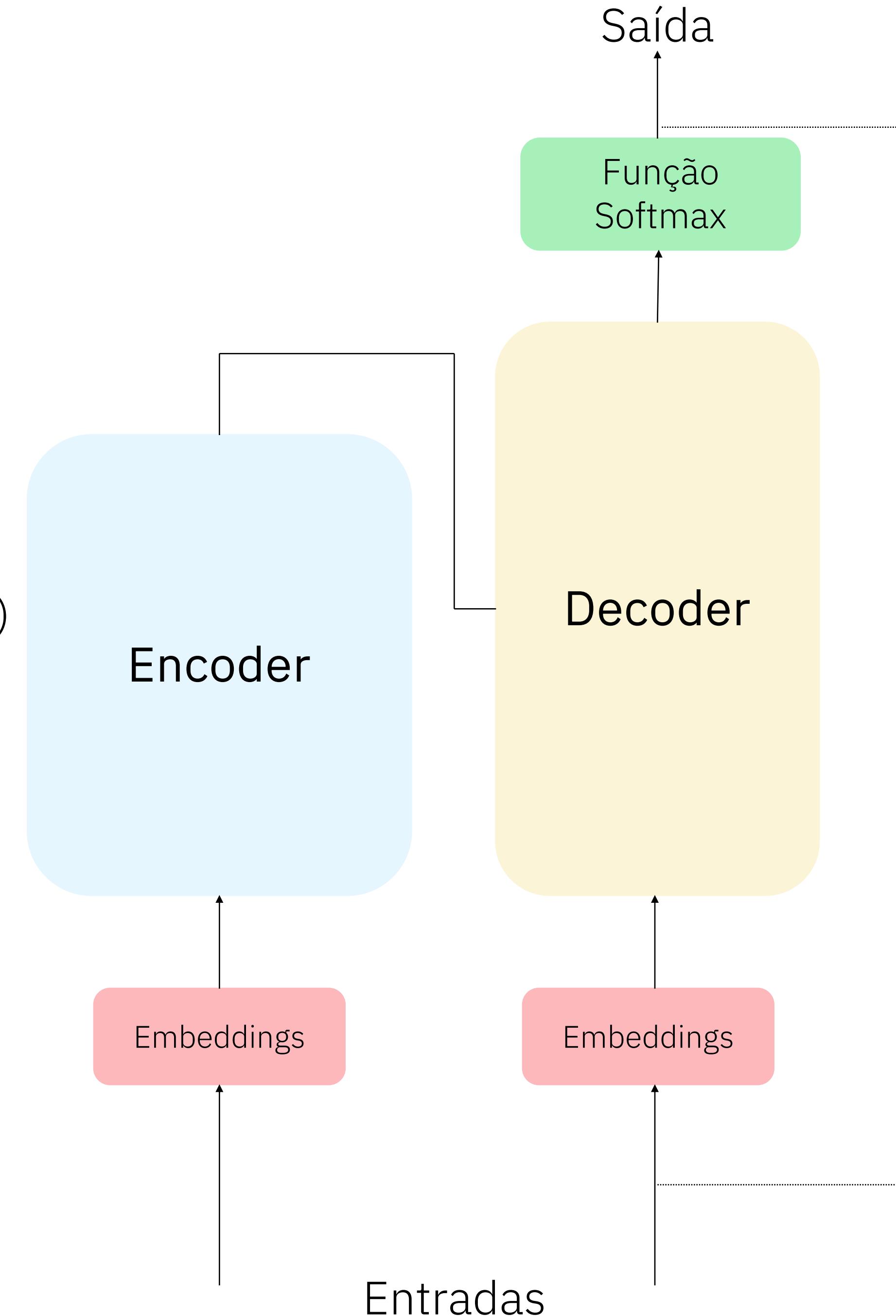
Prompts e Geração de Texto

Transformers

Arquitetura simplificada

Encoder

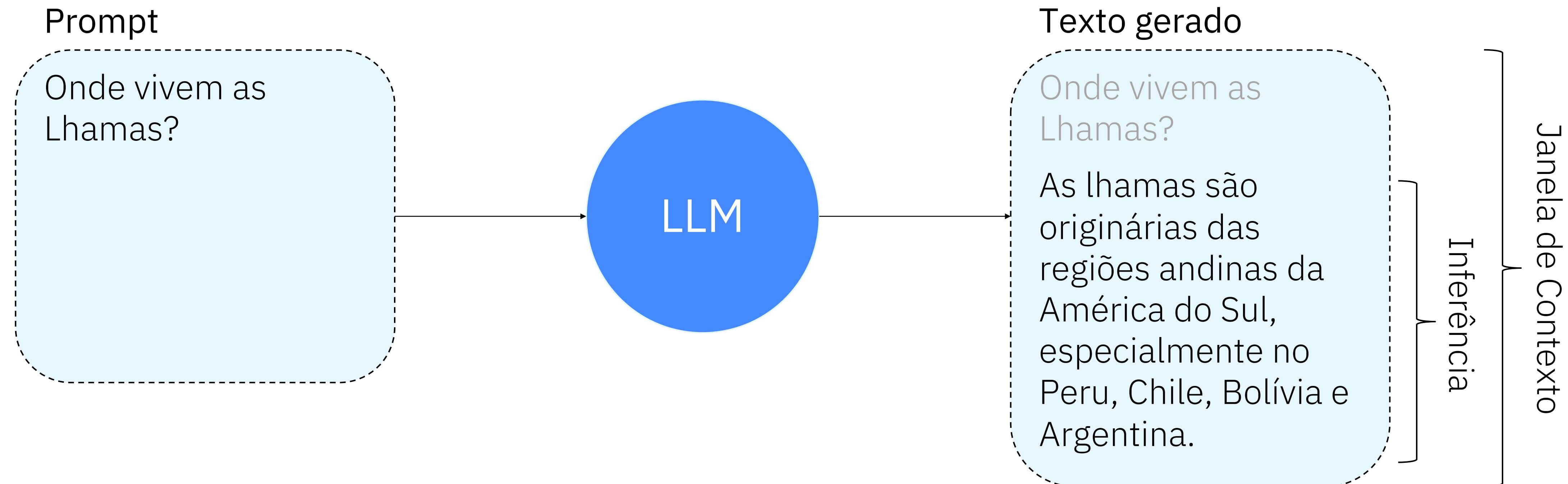
Codifica a entrada (“prompts”) com um entendimento contextual e produz um vetor para cada token de entrada.



Decoder

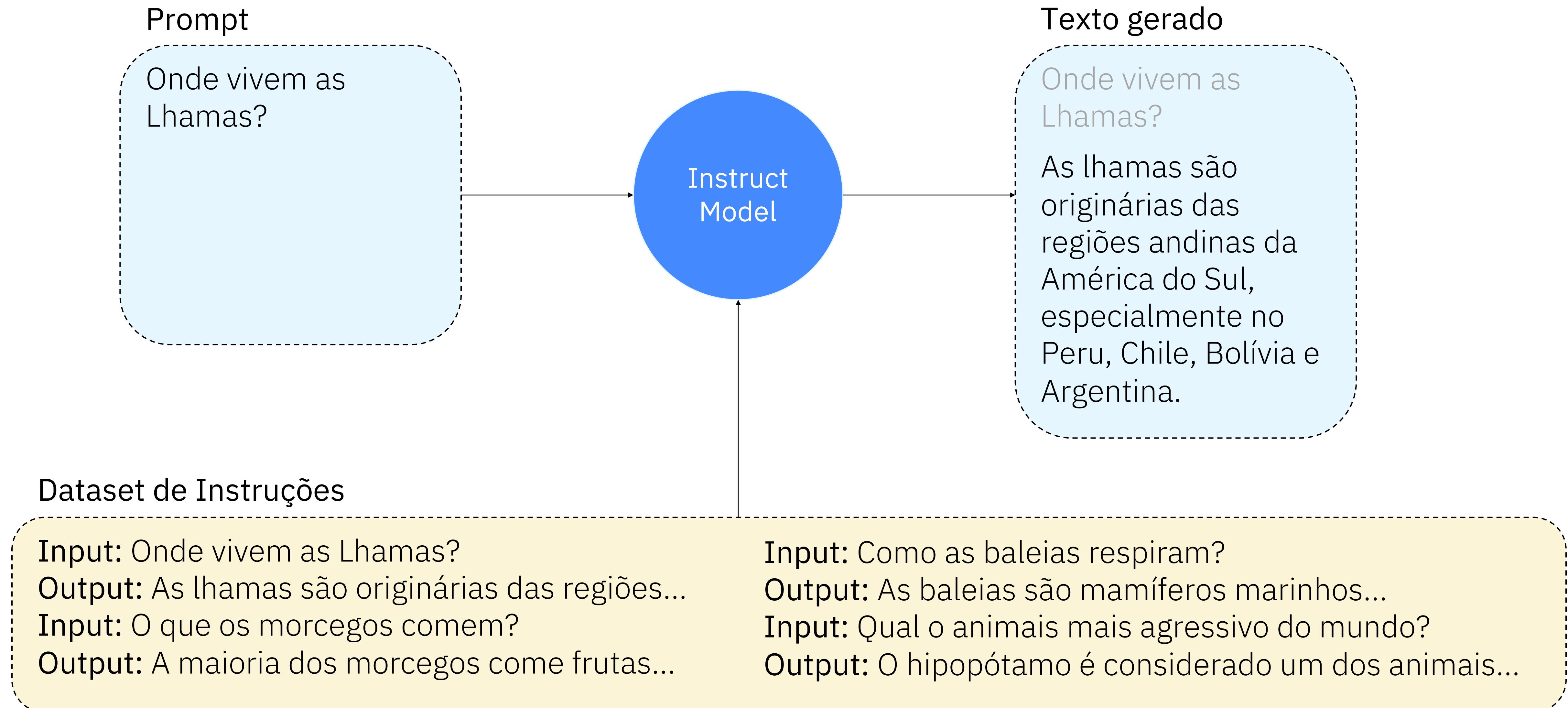
Aceita tokens de entrada e gera novos tokens.

Prompts e Geração de Texto

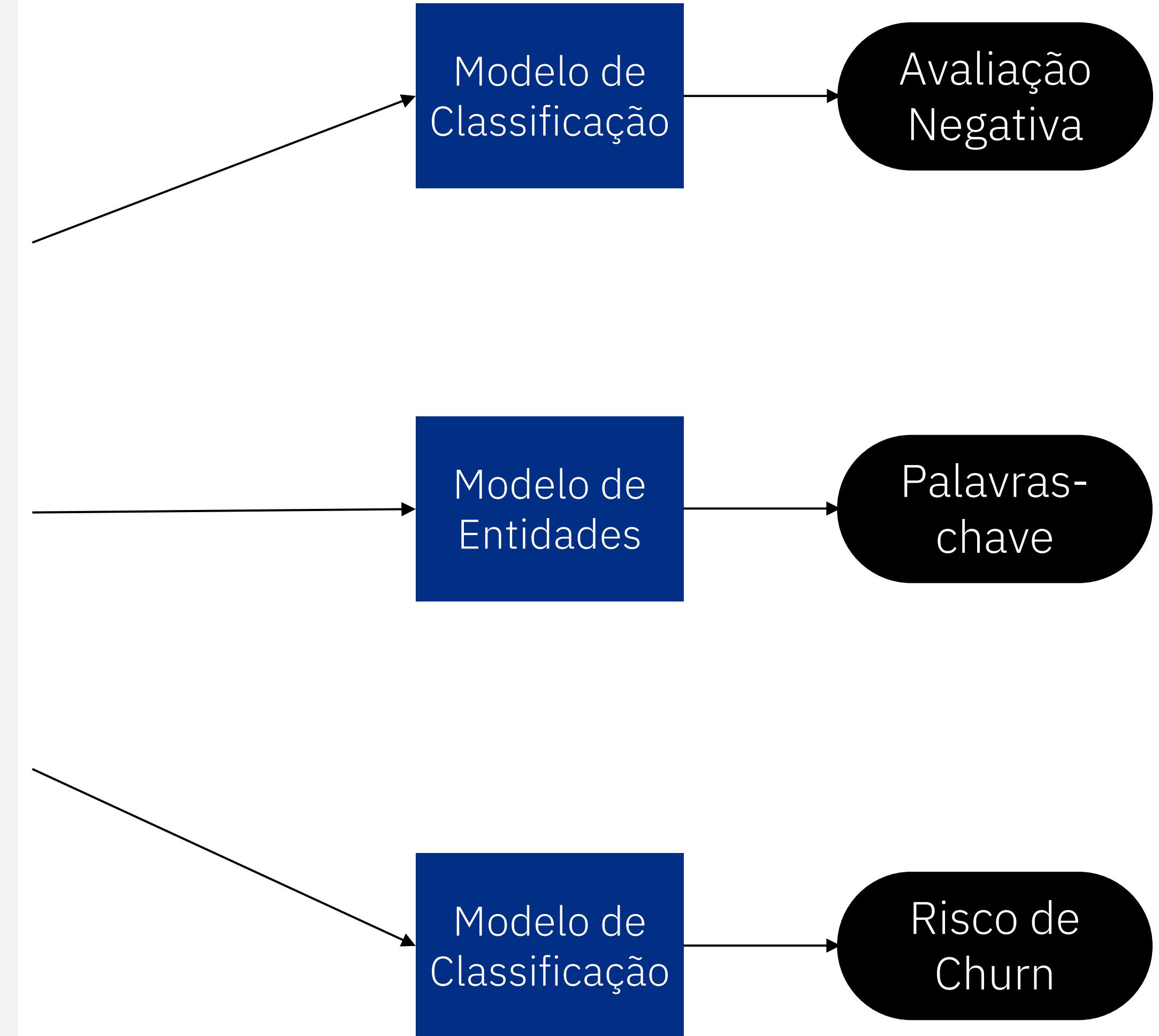


Diferentes modelos possuem diferentes tamanhos de *Context Window*:
2k ~ 16k tokens

Modelos *Instruct-tuned*



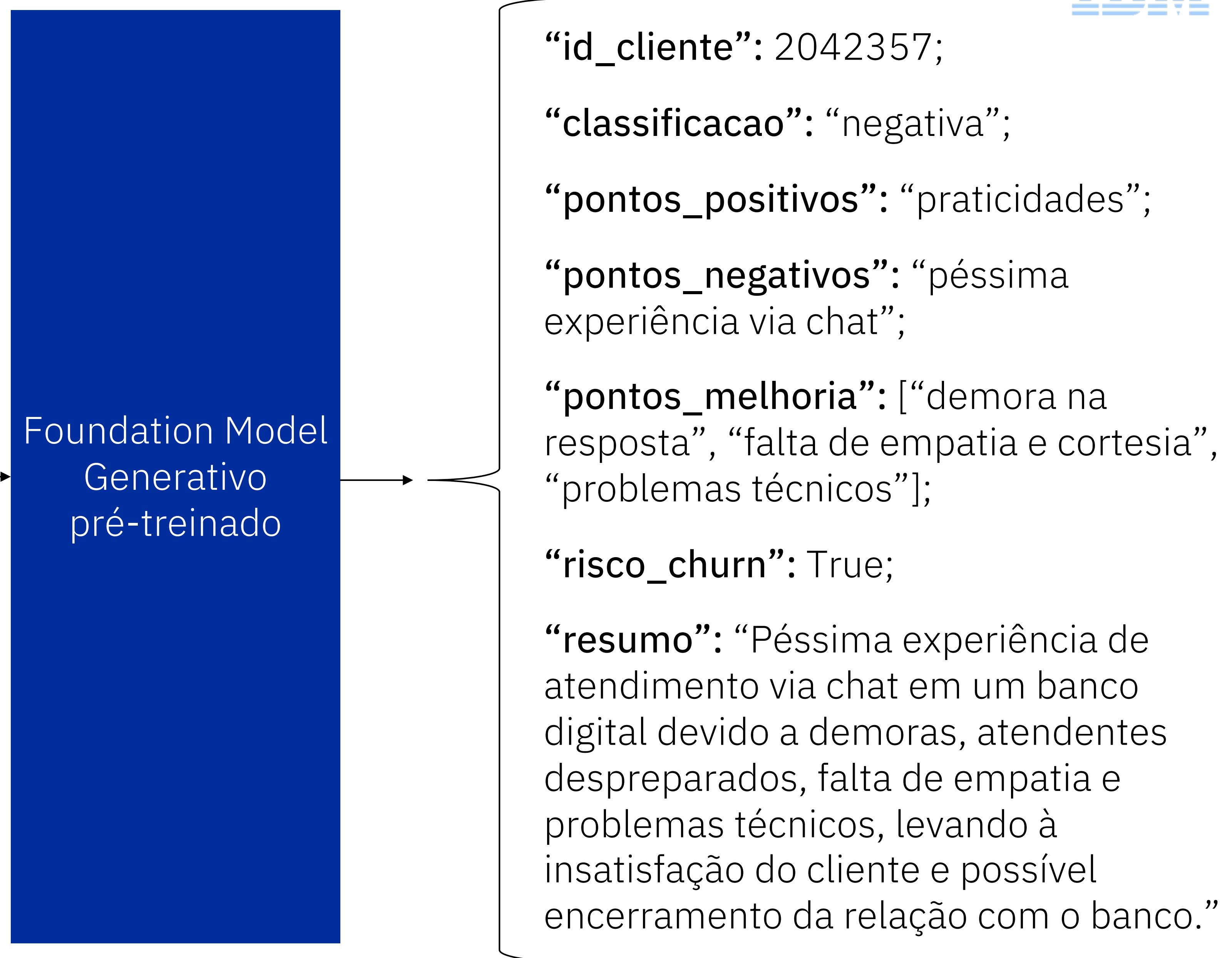
“
Apesar das praticidades do meu banco digital, tive uma péssima experiência via chat. Enfrentei demoras absurdas nas respostas, falta de empatia e cortesia, além de problemas técnicos.
(...) levando-me a considerar encerrar minha relação com o banco.”



“ Apesar das praticidades do meu banco digital, tive uma péssima experiência via chat. Enfrentei demoras absurdas nas respostas, falta de empatia e cortesia, além de problemas técnicos.

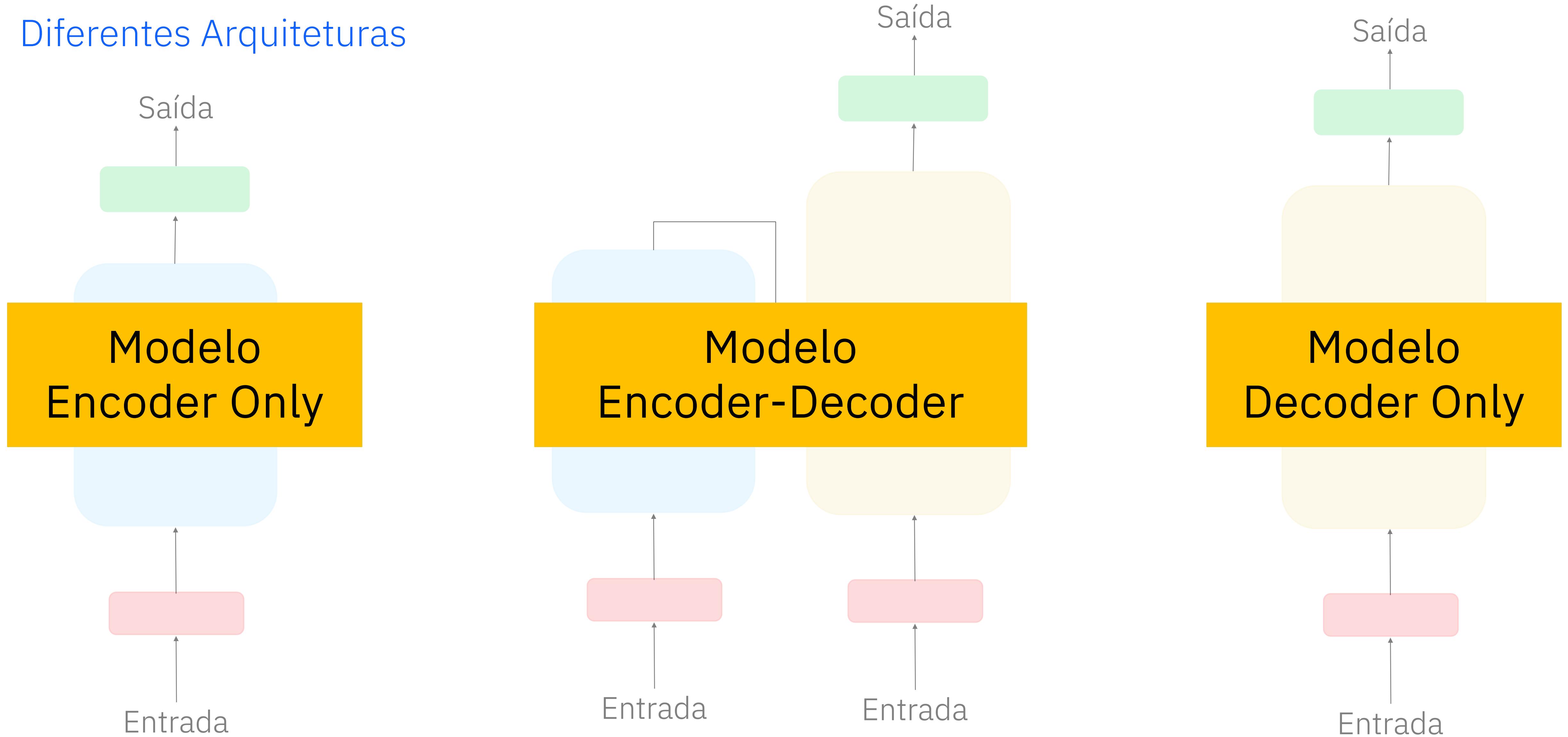
(...)

levando-me a considerar encerrar minha relação com o banco.



Transformers

Diferentes Arquiteturas

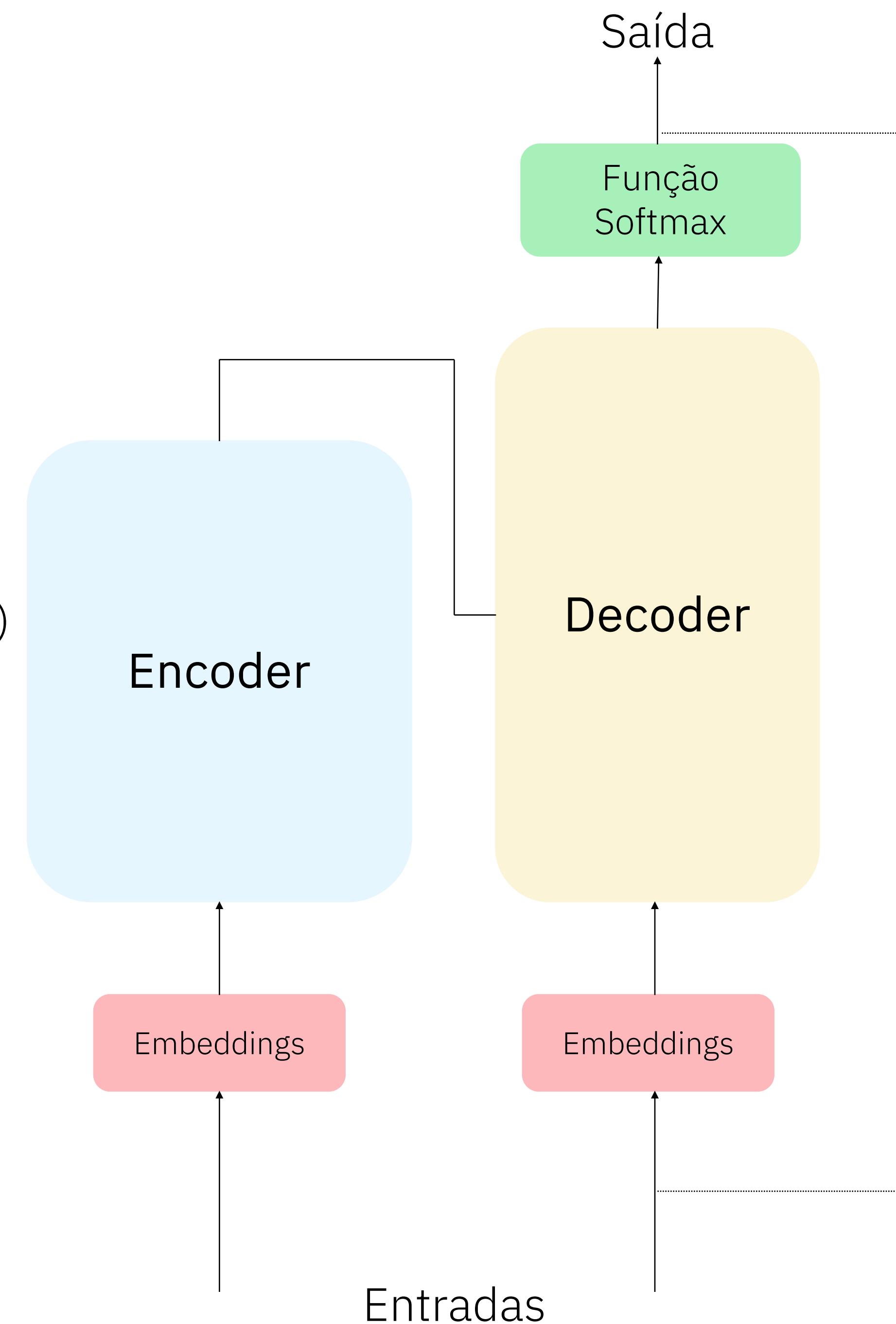


Transformers

Arquitetura simplificada

Encoder

Codifica a entrada (“prompts”) com um entendimento contextual e produz um vetor para cada token de entrada.

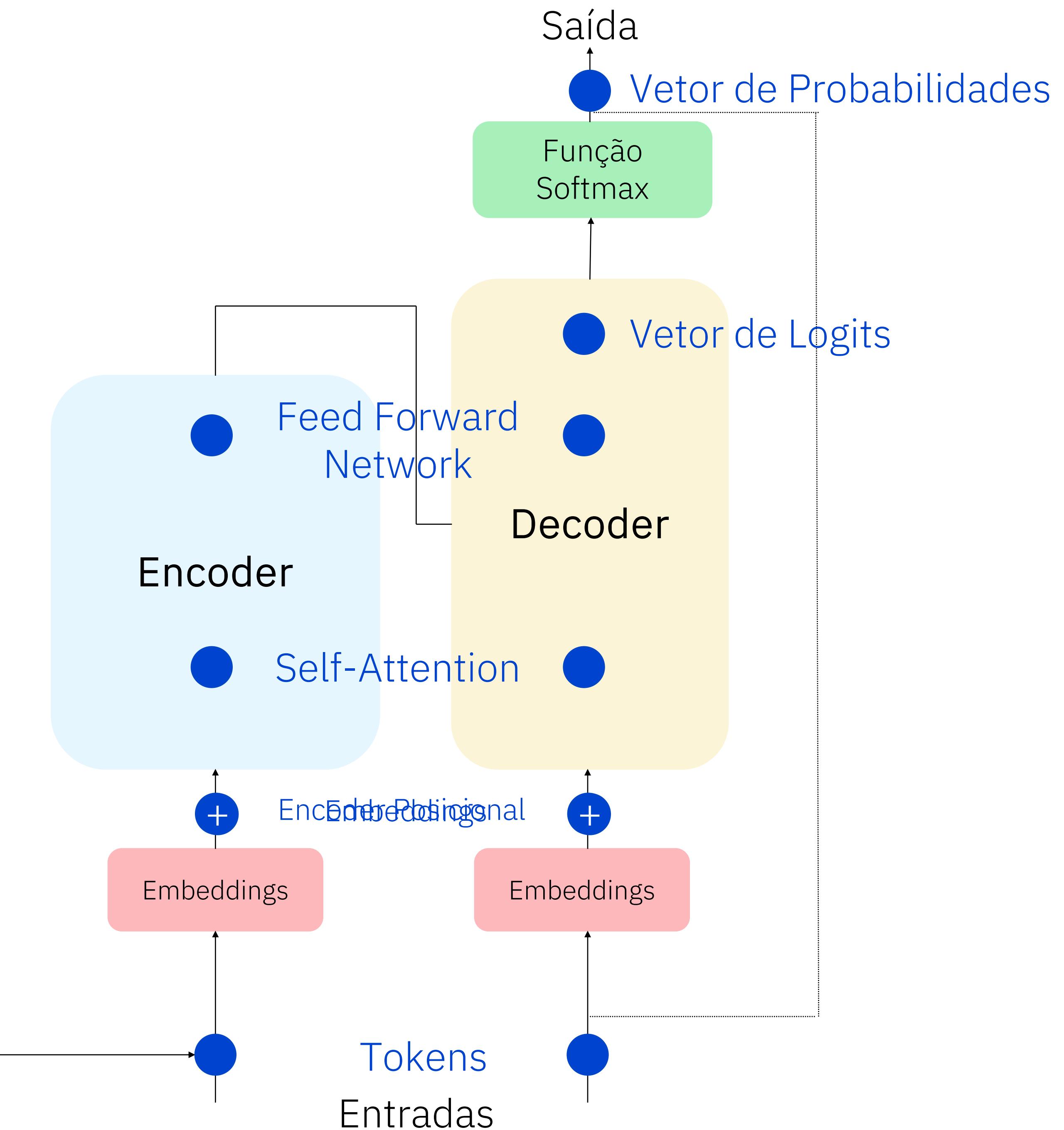


Decoder

Aceita tokens de entrada e gera novos tokens.

Transformers

Arquitetura simplificada



Parâmetros de Inferência

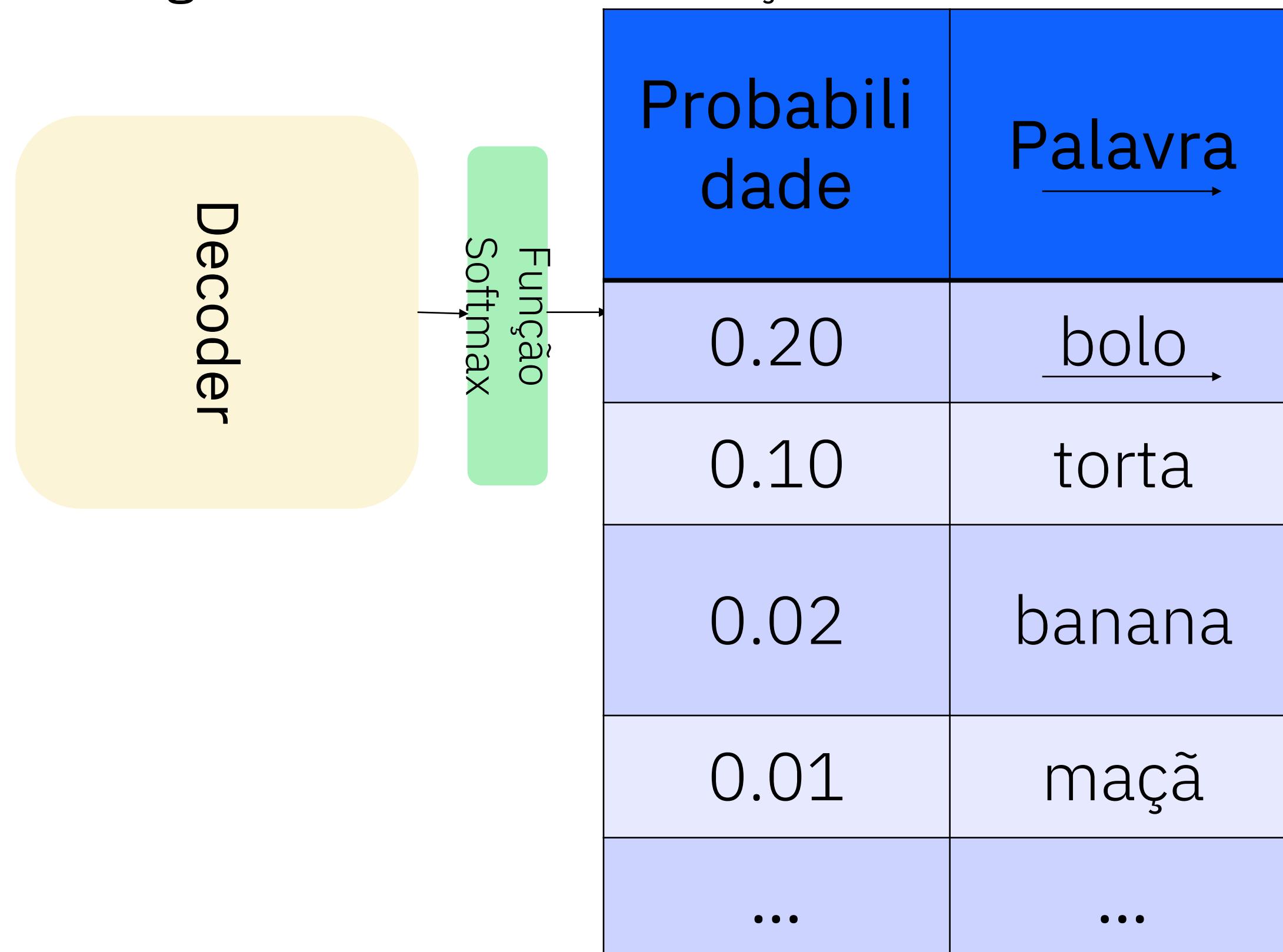
Definindo como os modelos irão responder

Min tokens: quantidade mínima de tokens a serem gerados.

Max tokens: quantidade máxima de tokens a serem gerados.

Stop sequences: caractere, ou sequência de caracteres, para delimitar o final da inferência.

Decoding: técnica de decodificação.



Greedy: a palavra/token mais provável é sempre selecionada.

Sampling: seleciona aleatoriamente um token usando diferentes estratégias para controlar a aleatoriedade.

Model parameters

Decoding

Greedy Sampling

Temperature 0,7

Top P (nucleus sampling) 1

Top K 100 50

Random seed

Repetition penalty 1

Stopping criteria

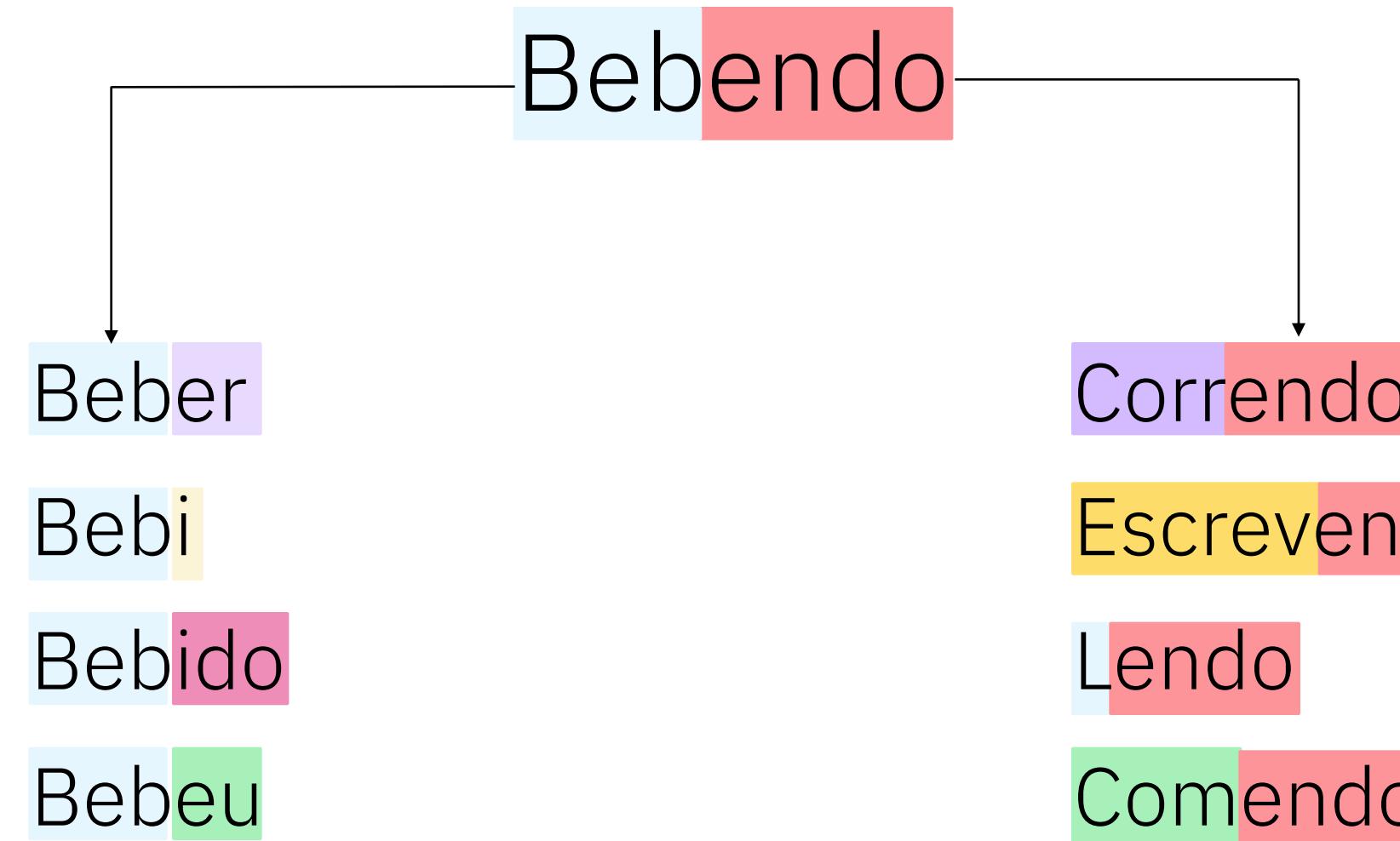
Stop sequences +

Min tokens Max tokens

Enter up to 6 sequences to stop output after the minimum number of tokens is reached.

Tokens

Mais eficientes



Tokens **Characters**
23 68

0 arquiteto não reiniciou o servidor porque ele estava com preguiça.

TEXT TOKEN IDS

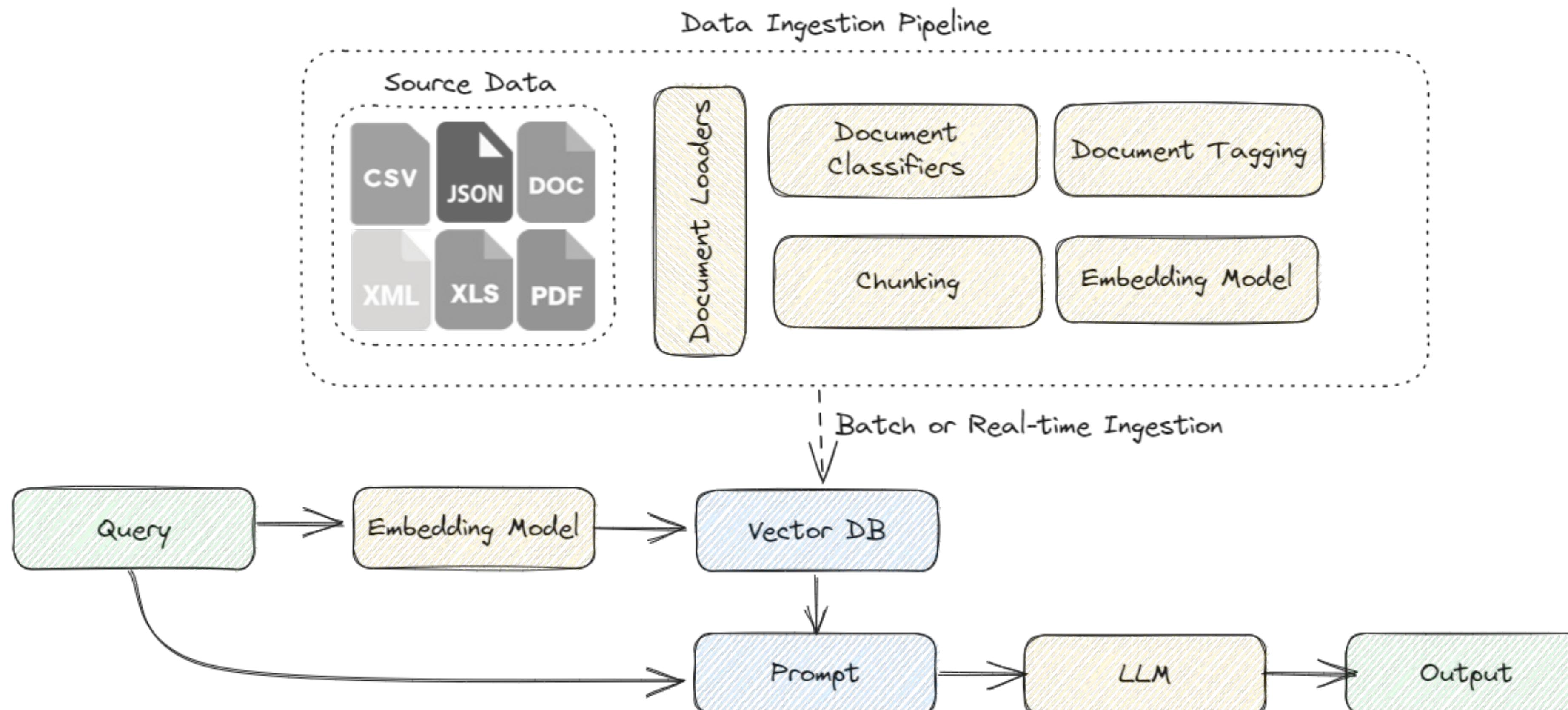
[46, 610, 47391, 27206, 299, 28749, 6865, 44070, 280, 267, 1113, 312, 273, 16964, 4188, 9766, 1556, 4170, 401, 662, 48317, 50041, 13]

TEXT TOKEN IDS

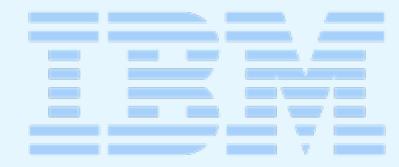
O que é RAG?

Retrieval-Augmented Generation

RAG combina processos de recuperação e geração para aprimorar as capacidades dos LLMs



O que é RAG?

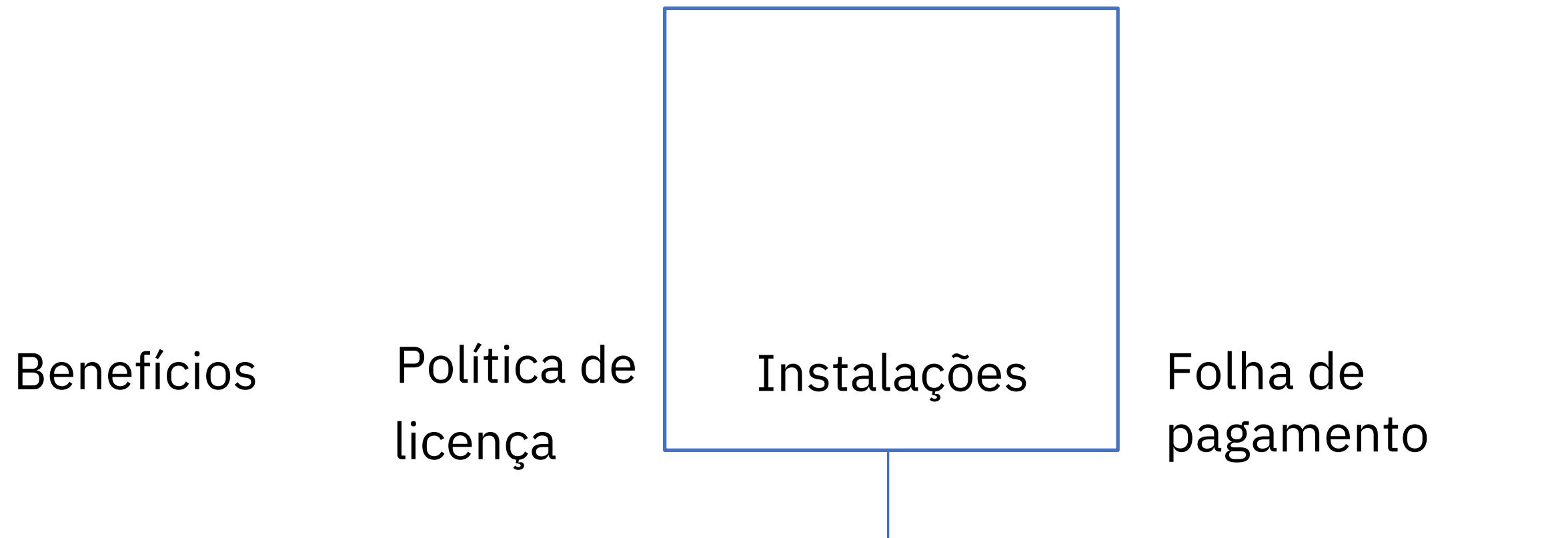


Exemplo muito simples...

1. Dada uma pergunta, pesquise documentos relevantes para obter resposta

Há estacionamento para funcionários?

Documentos da empresa



2. Incorpore o texto recuperado em um prompt atualizado

Use as seguintes partes do contexto para responder à pergunta no final:

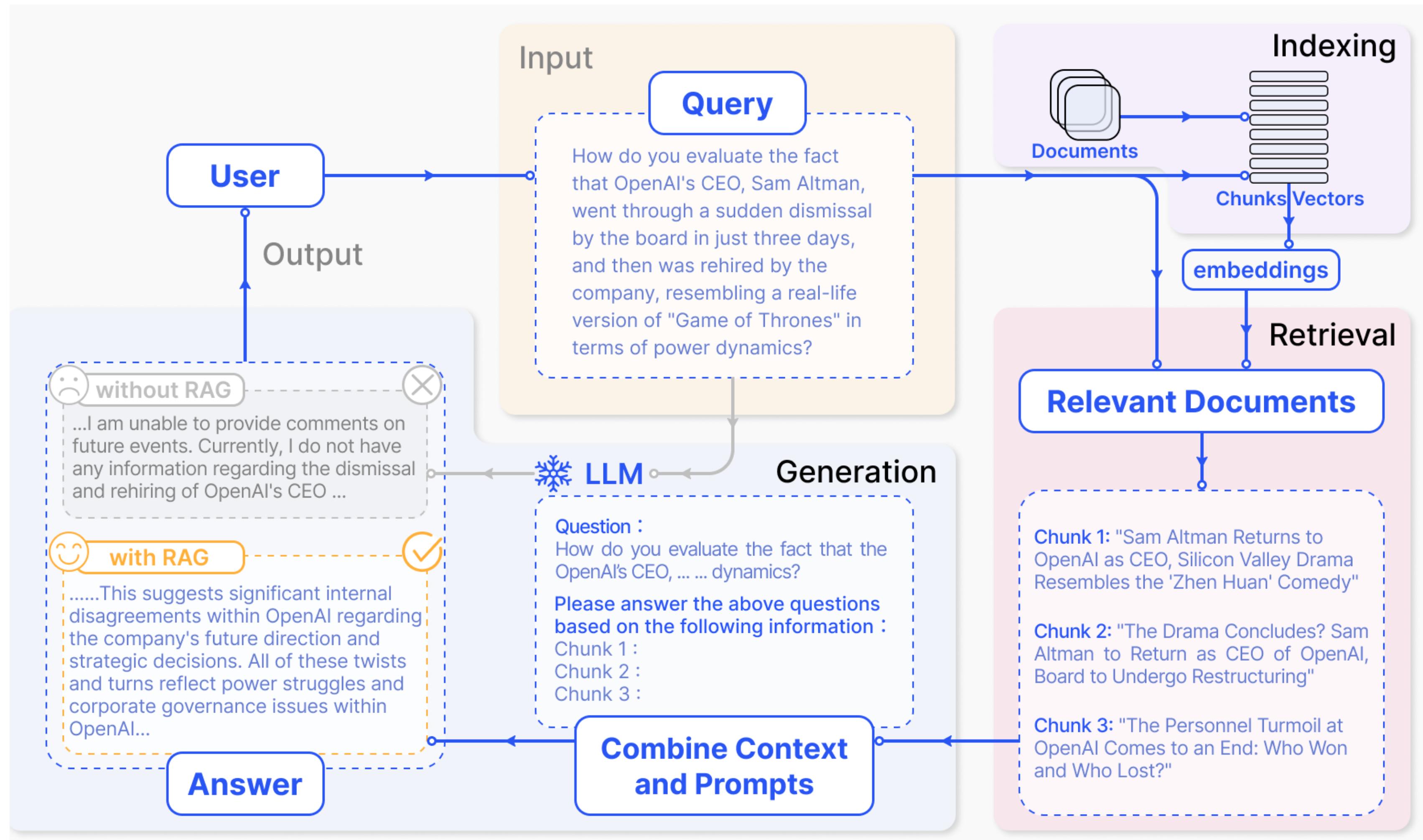
Política de Estacionamento: Todos os funcionários podem estacionar nos níveis 1 e 2 do escritório.

Use a entrada na Front St [...]

Há estacionamento para funcionários?

O que é RAG?

Exemplo de ponta a ponta...



Obrigado ! Perguntas

