LLM - Fine-tuning

Luciano Barbosa

Centro de Informática UFPE

Universidade Federal de Pernambuco

Prompt Engineering: Limitações

- Tamanho da janela de contexto é limitado
- Requer esforço humano para construção e melhoria de prompts
- Um mesmo prompt pode entregar saídas diferentes
- Inconsistência na experiência com usuário

Por que utilizar modelos Open-Source?

- **Fine-tuning:** Não precisa treinar do zero ou depender apenas de prompt engineering
  - OpenAI disponibiliza fine-tuning apenas para alguns modelos através de sua API
- Não depende de terceiros
  - Perda de disponibilidade do serviço
  - Alteração nos custos do serviço
  - Latência
  - Sem controle sobre mudanças no modelo usado
  - Moderação e Segurança: dados sendo transmitidos para terceiros

[1] https://arxiv.org/abs/2302.13971
[2] https://arxiv.org/pdf/2203.15556.pdf

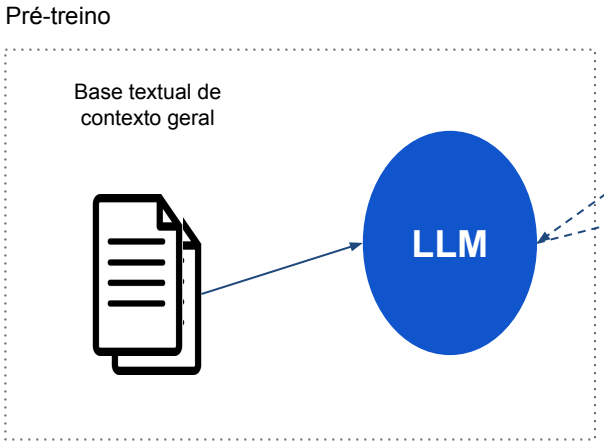
Por que utilizar modelos Open-Source?

- Mais customizáveis
- Privados
- Existem modelos abertos com menor número de parâmetros mas com performance competitiva

[1] https://arxiv.org/abs/2302.13971
[2] https://arxiv.org/pdf/2203.15556.pdf

Fine-tuning

- LLMs podem ser muito generalistas devido ao pré-treino em uma grande base de dados genérica



TAREFA: Rotulagem de Entidade Nomeada no Contexto Jurídico

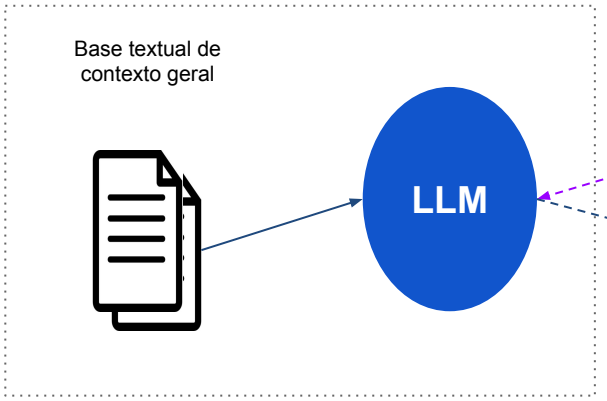
SAÍDA: Rotulagem de Entidade Nomeada no Contexto geral onde foi treinado o LLM

EXEMPLO: "conforme autoriza o art. 4.o, da Lei de Introdução às Normas do Direito Brasileiro, com redação determinada pela Lei 12.376/2010 (Dec.-lei 4.657/1942)." (Santana, 2019) *

* Exemplo obtido em https://www.jusbrasil.com.br/doutrina/secao/protecao-juridica-da-dignidade-do-consumidor-dano-moral-no-direito-do-consumidor/1207548435#a-177431376
** NER realizado através do modelo "text-da-vinci-002". Várias execuções de um mesmo prompt foram realizadas e observadas as entidades nomeadas retornadas com maior frequência.

Fine-tuning

- No fine-tuning, o LLM é especializado em uma tarefa específica e um domínio específico
- Uma nova etapa de treino é executada a partir dos parâmetros pré-treinados do LLM



fine-tuning

Base textual no contexto específico da tarefa

DADO ROTULADO

TAREFA: Rotulagem de Entidade Nomeada no Contexto Jurídico

Exemplo:

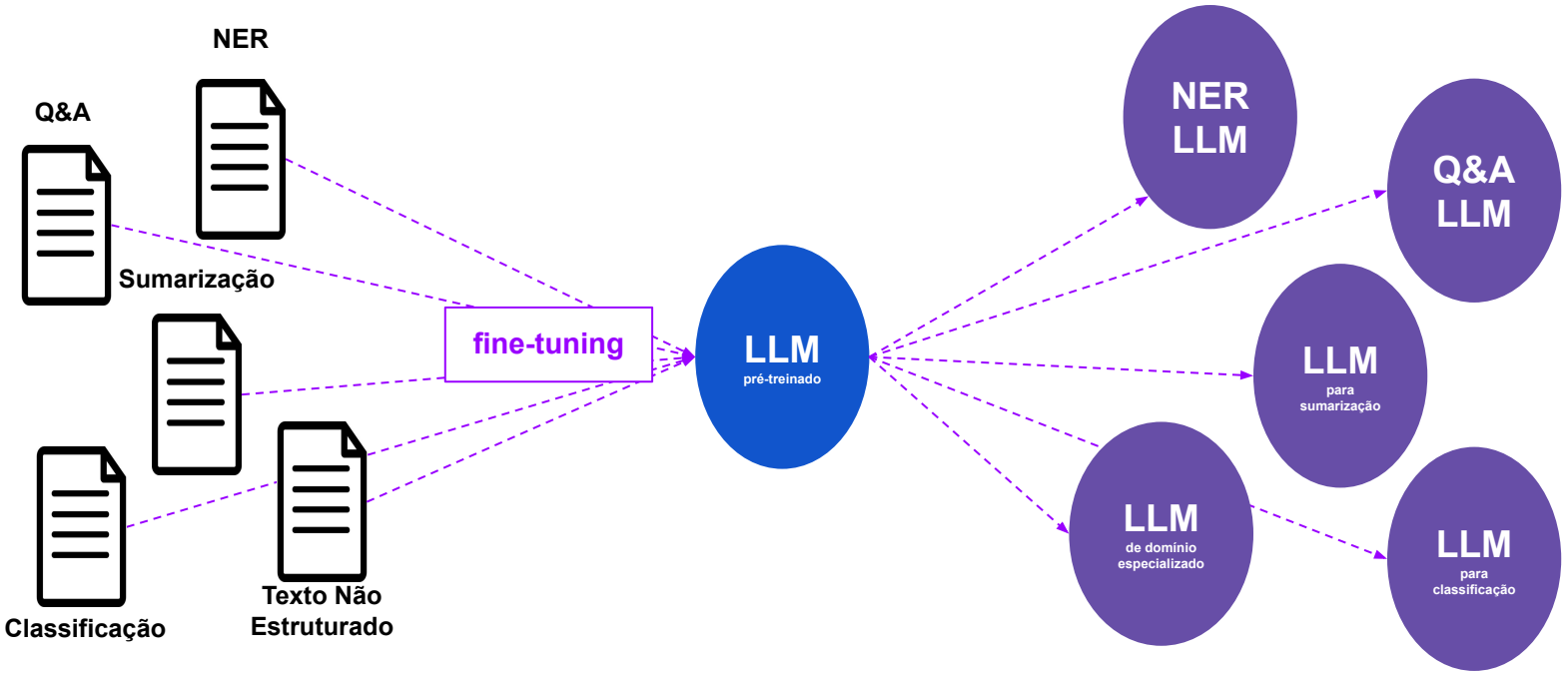
SAÍDA: Rotulagem de Entidade Nomeada no Contexto da tarefa

EXEMPLO: "conforme autoriza o art. 4.o, da Lei de Introdução às Normas do Direito Brasileiro, com redação determinada pela Lei 12.376/2010 (Dec.-lei 4.657/1942)." (Santana, 2019) *

* Exemplo obtido em https://www.jusbrasil.com.br/doutrina/secao/protecao-juridica-da-dignidade-do-consumidor-dano-moral-no-direito-do-consumidor/1207548435#a-177431376

Fine-tuning

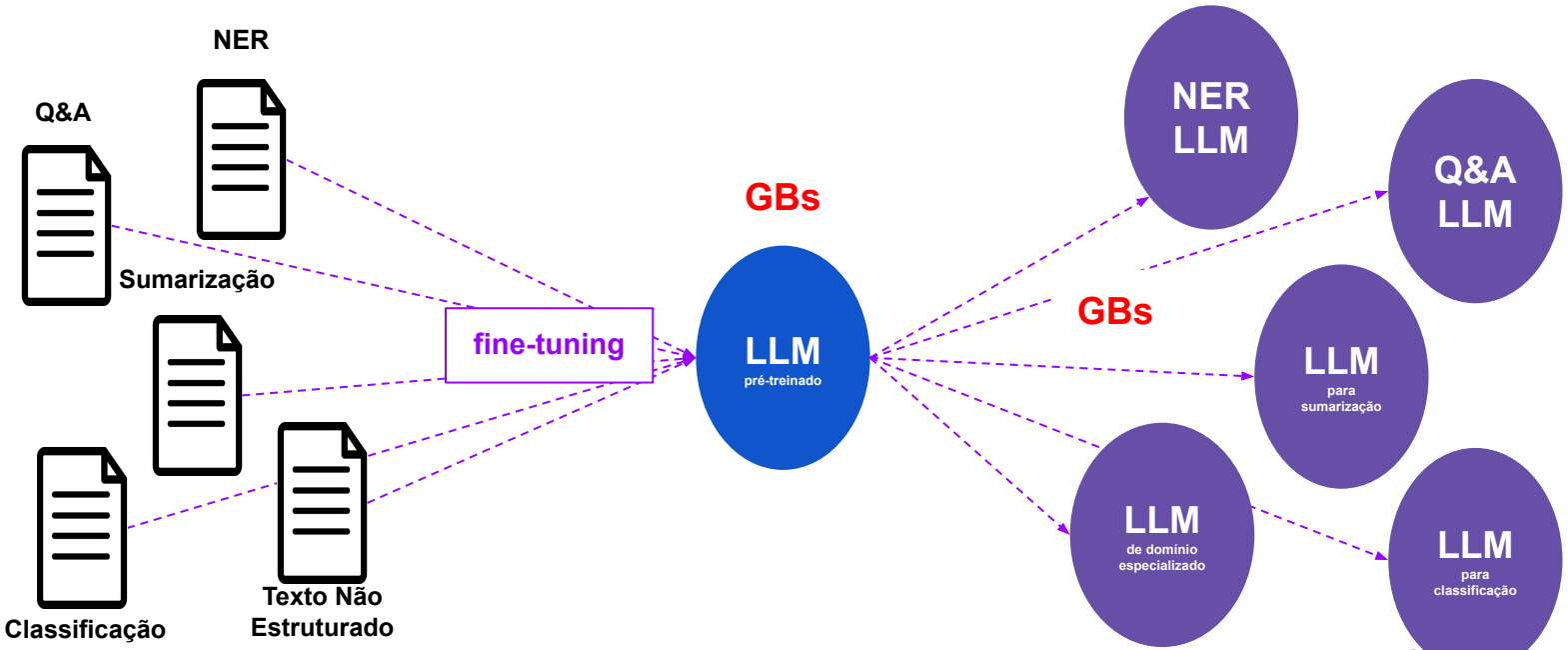
- No fine-tuning, o LLM é especializado em uma tarefa específica e um domínio específico
- Uma nova etapa de treino é executada a partir dos parâmetros pré-treinados do LLM



https://arxiv.org/pdf/2004.10964.pdf

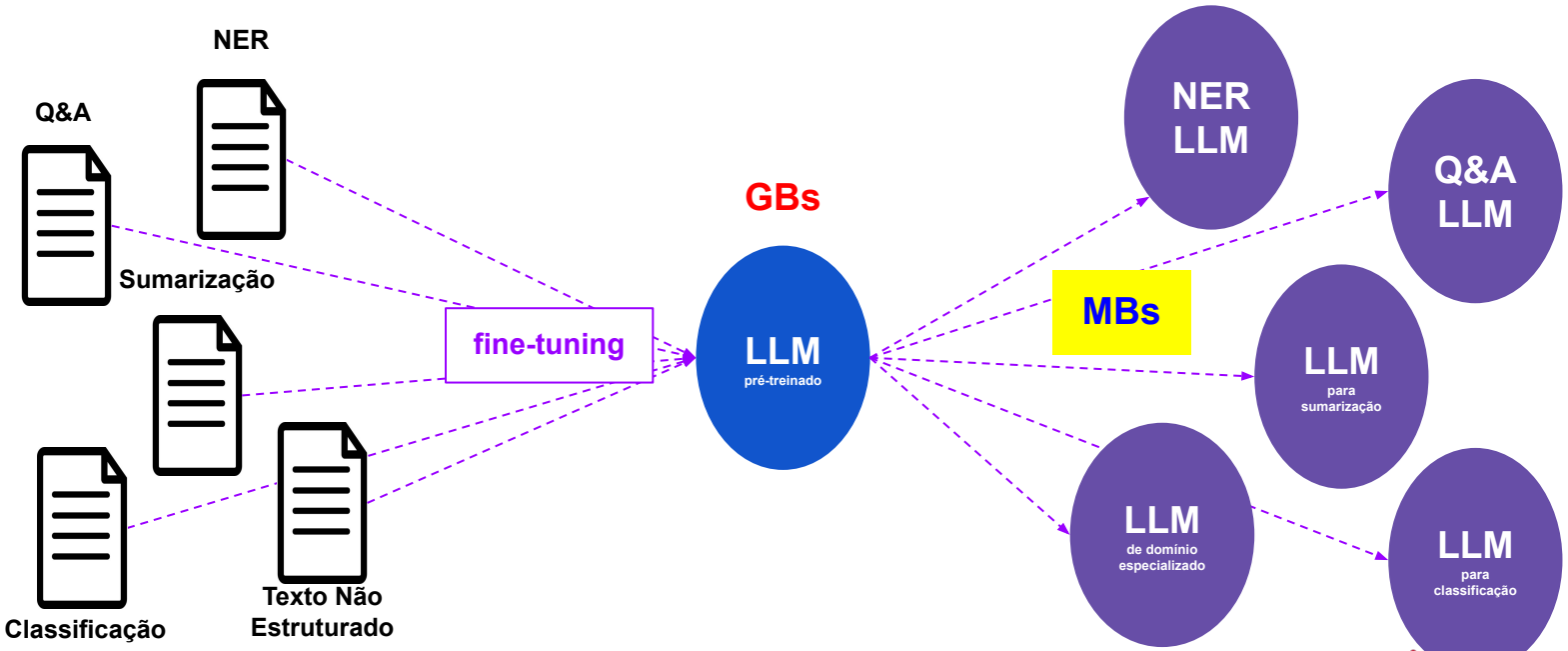
Fine-tuning

- É possível obter bons resultados a partir de um pequeno conjunto de dados
- Full fine-tuning:
  - Alto custo computacional
  - Catastrophic Forgetting



Fine-tuning: Diminuindo os Custos Computacionais

- Técnicas para redução do número de pesos “tunáveis” dos modelos
  - Parameter-Efficient Fine-Tuning (PEFT)
- Quantização: Técnica para “tunagem” de parâmetros com menor precisão numérica



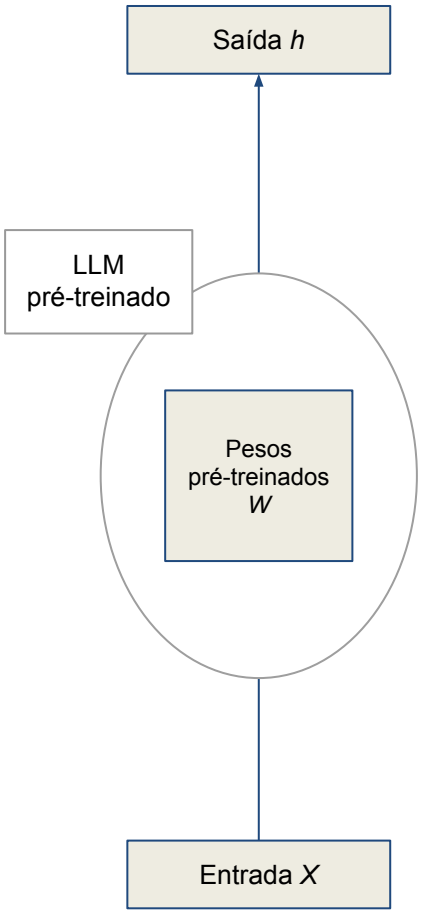
https://github.com/huggingface/peft

Parameter Efficient Fine-Tuning (PEFT)

- Vários métodos que permitem a adaptação do fine-tuning de um LLM pré-treinado
  - Congelam pesos ou camadas do modelo
  - Atualizam apenas alguns pesos ou camadas
  - Adicionam camadas ou parâmetros
- Performance comparável ao fine-tuning completo do LLM
- Menos propenso a esquecimento catastrófico durante fine-tuning completo

FONTE: https://arxiv.org/pdf/2303.15647v1.pdf

Low-Rank Adaptation (LoRA)



- Reparametriza os pesos do modelo utilizando uma representação de baixo nível

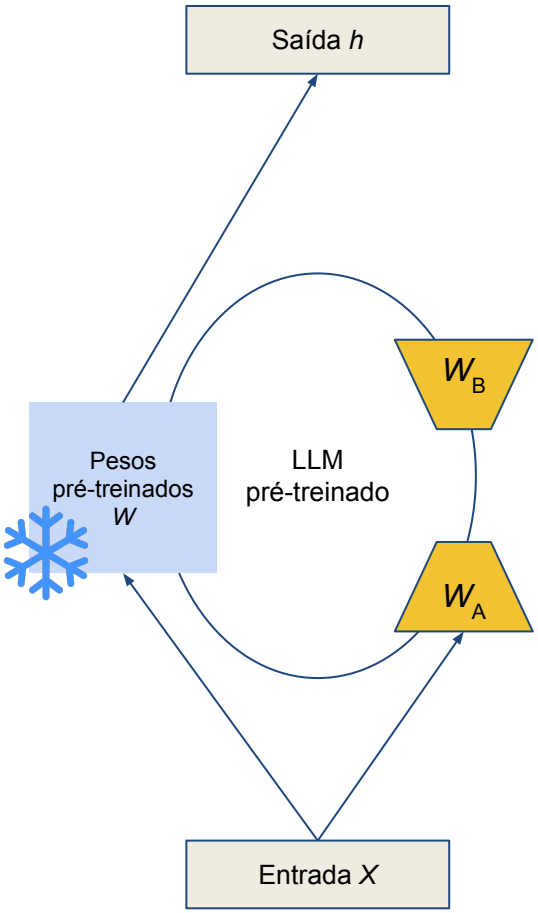
FONTE: https://arxiv.org/abs/2106.09685

Low-Rank Adaptation (LoRA)

- Reparametriza os pesos do modelo utilizando uma representação de baixo nível
- Congela os pesos originais do modelo

FONTE: https://arxiv.org/abs/2106.09685

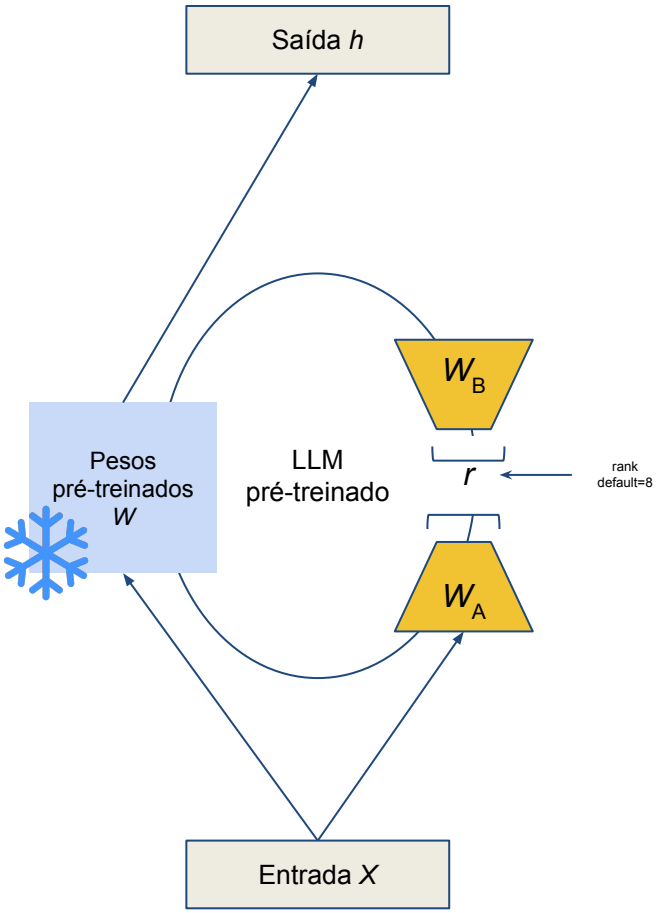
Low-Rank Adaptation (LoRA)



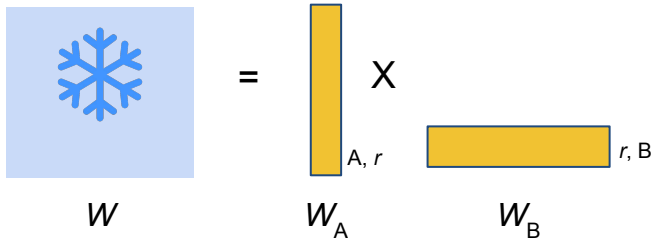
- Reparametriza os pesos do modelo utilizando uma representação de baixo nível
- Congela os pesos originais do modelo
- Injeta duas **matrizes menores de decomposição**

FONTE: https://arxiv.org/abs/2106.09685

Low-Rank Adaptation (LoRA)

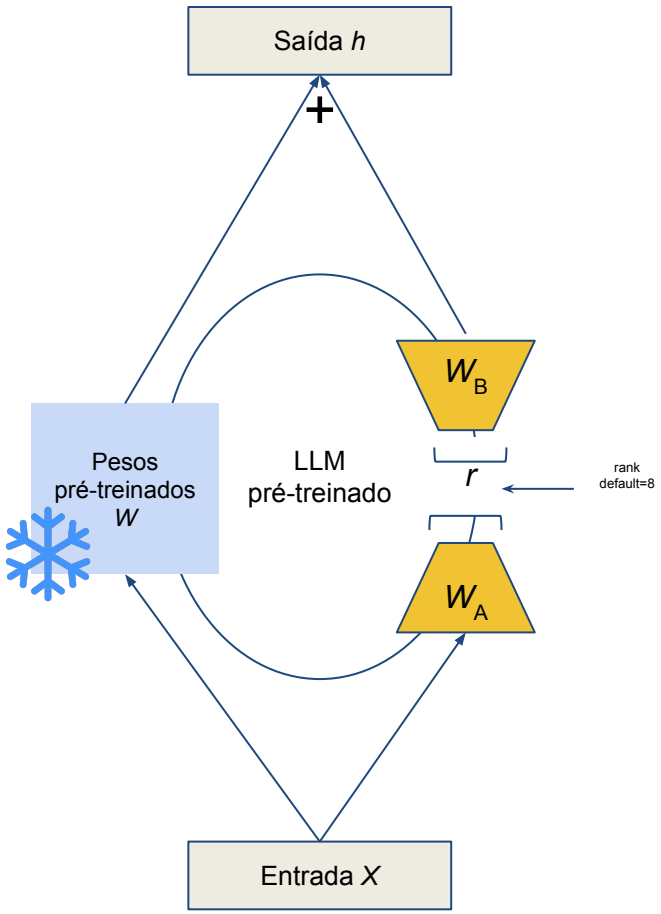


- Reparametriza os pesos do modelo utilizando uma representação de baixo nível
- Congela os pesos originais do modelo
- Injeta duas matrizes menores de decomposição
  - O produto das dimensões dessas duas matrizes deve ser igual às dimensões dos pesos originais

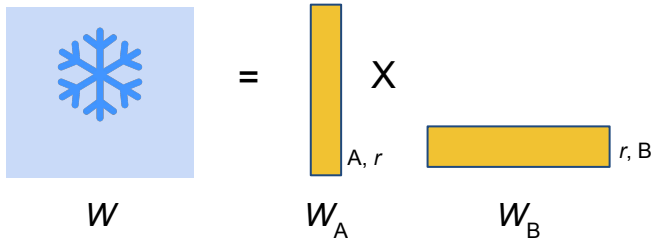


FONTE: https://arxiv.org/abs/2106.09685

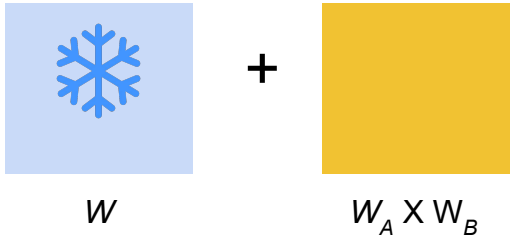
Low-Rank Adaptation (LoRA)



- Reparametriza os pesos do modelo utilizando uma representação de baixo nível
- Congela os pesos originais do modelo
- Injeta duas matrizes menores de decomposição
  - O produto das dimensões dessas duas matrizes deve ser igual às dimensões dos pesos originais



- Treina os pesos das matrizes de decomposição
- Adiciona os pesos aprendidos aos congelados



FONTE: https://arxiv.org/abs/2106.09685

Low-Rank Adaptation (LoRA)

- Reduz o número de parâmetros treináveis
- Possibilita a execução do fine-tuning com menos recurso computacional
- Possibilita Inferência mais customizada:
  - Adicionando as matrizes de decomposição, aprendidas anteriormente, aos pesos congelados do modelo em tempo de inferência

FONTE: https://arxiv.org/abs/2106.09685

Exemplo de Fine-tuning

Comentário Cliente	**Horrível! Restaurante vazio e o atendimento ainda demora. Pedi um frango e veio com gosto de de peixe. Nojo!**
Resposta Falcon 7B	Horrível! Restaurante vazio e o atendimento ainda demora. Pedi um frango e veio com gosto de de peixe. Nojo!
Resposta LLaMA 2	Obrigado pelo seu comentário. Nós estamos muito preocupados com a sua experiência.
Resposta Open-LLaMA	O cliente não gostou do frango.

Antes

Comentário Cliente	**Horrível! Restaurante vazio e o atendimento ainda demora. Pedi um frango e veio com gosto de de peixe. Nojo!**
Falcon 7B	Olá, agradecemos pelo seu feedback. Lamentamos muito pela experiência negativa que você teve em nosso restaurante. Pedimos desculpas pelo atendimento demorado e pelo frango com gosto de peixe. Vamos investigar o ocorrido e tomar as medidas necessárias para melhorar esses pontos negativos. Agradecemos por nos informar sobre sua experiência e esperamos ter a oportunidade de recebê-la novamente para uma experiência melhor.
LLaMA 2 7B	Caro cliente, lamentamos profundamente pela experiência negativa que você teve em nosso restaurante. Pedimos desculpas pelo atendimento lento e pela confusão com o prato que você pediu. Vamos investigar o ocorrido para garantir que isso não se repita. Agradecemos seu feedback e esperamos ter a oportunidade de oferecer uma experiência melhor no futuro.
Open LLaMA 7B	Caro cliente, lamentamos profundamente pela experiência negativa que você teve em nosso restaurante. Pedimos desculpas pelo atendimento demorado e pelo frango com gosto de peixe. Vamos investigar o ocorrido e tomar medidas para garantir que isso não aconteça novamente. Agradecemos seu feedback e esperamos ter a oportunidade de recebê-lo novamente para oferecer uma experiência melhor.

Depois

Quantização

- Técnica que reduz a quantidade de memória necessária para armazenar e treinar modelos
- Projeta os valores originais com precisão de 32 bits em espaços de precisão menor.
- Quantização + LoRA = QLoRA

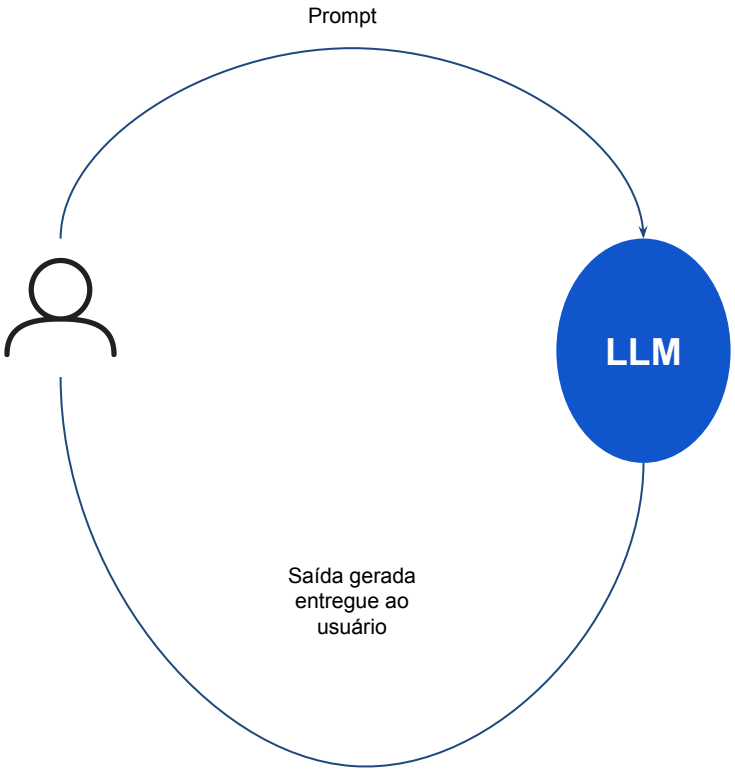
PRECISÃO	MEMORY	EXEMPLO
FP32	4 bytes	3.1415920257568359375
FP16	2 bytes	3.140625
INT8	1 Byte	3

https://huggingface.co/docs/transformers/perf_train_gpu_one#mixed-precision-training
https://huggingface.co/docs/transformers/model_memory_anatomy
https://cloud.google.com/tpu/docs/bfloat16
QLoRA: https://arxiv.org/abs/2305.14314

Retrieval Augmented Generation

Desafios dos Modelos de Linguagem

- Não possuem informação atualizada
- Alucinações

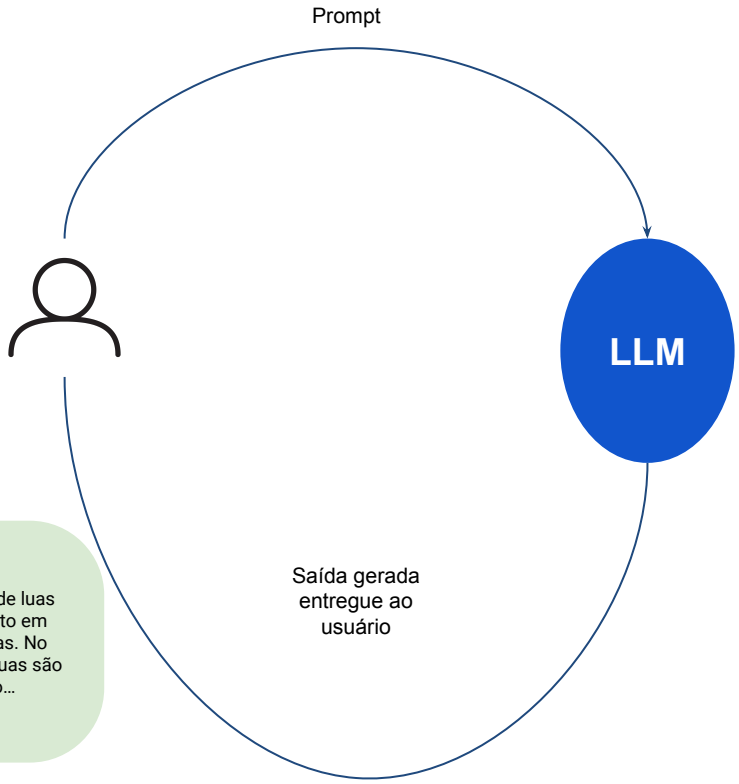


Retrieval Augmented Generation

Desafios dos Modelos de Linguagem

- Não possuem informação atualizada
- Alucinações

Qual o planeta com mais luas no sistema solar?



* **Júpiter** é o planeta do Sistema Solar com o maior número de luas conhecidas. Até a minha última atualização de conhecimento em setembro de 2021, Júpiter tinha mais de 80 luas confirmadas. No entanto, esse número pode ter mudado à medida que novas luas são descobertas por astrônomos e pesquisadores. Portanto...

* Resultado obtido com o ChatGPT

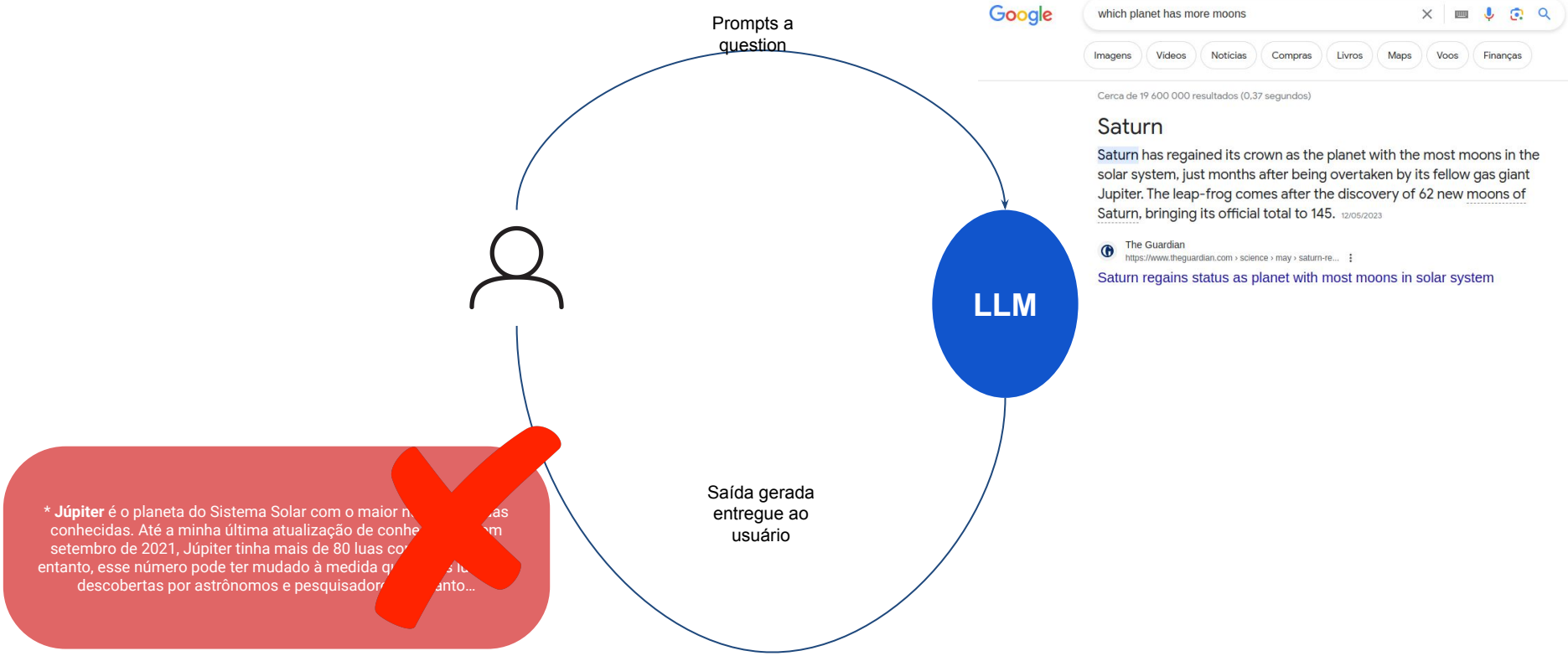
Meta Inc. (2020) https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/

Retrieval Augmented Generation

Desafios dos Modelos de Linguagem

- Não possuem informação atualizada
- Alucinações

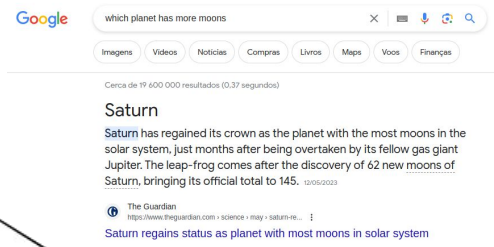
Qual o planeta com mais luas no sistema solar?

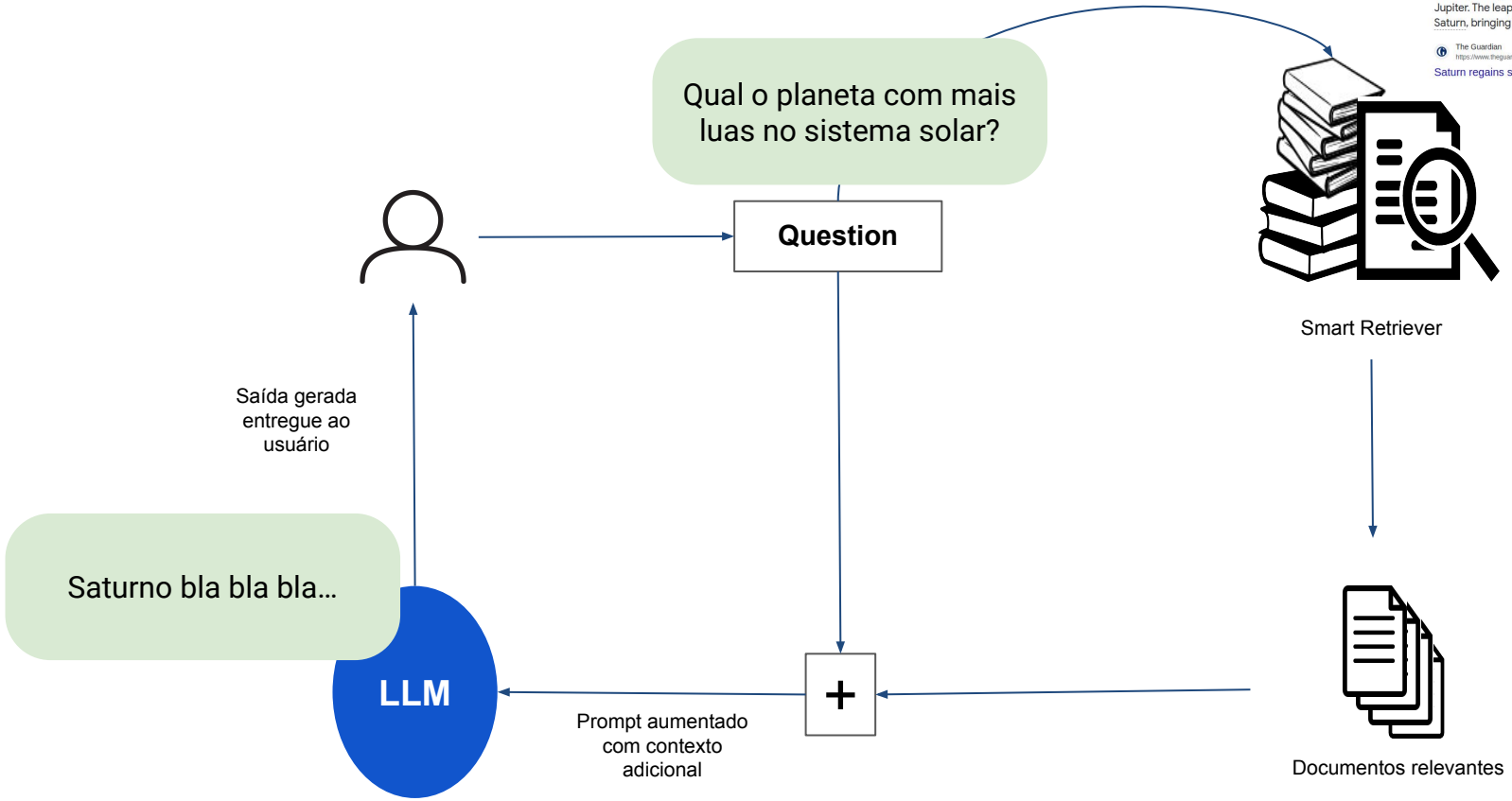


* Resultado obtido com o ChatGPT

Meta Inc. (2020) https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/

Retrieval Augmented Generation





Meta Inc. (2020)
https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/

Desafios: Deployment de Aplicações baseadas em LLMs

- Treino e Inferência
  - Treino do zero:
    - Muitos dados
    - Requer muita memória
  - Demanda poder computacional
  - Inferência com CPU: Técnicas de conversão para C++ [1]

[1]
https://github.com/ggerganov/llama.cpp
https://pub.towardsai.net/high-speed-inference-with-llama-cpp-and-vicuna-on-cpu-136d28e7887b