Agenda

UNIDADE 6: Large Laguage Models

6.1 Redes Neurais Generativas (RNG)

6.2 Conceitos de Processamento de Linguagem Natural

6.3 Principais modelos LLM

- 6.3.1 Open AI
- 6.3.2 Llama

Prática

Prática: Chatbot Genérico com LLM Local

1

Fine-Tuning Supervisionado do Llama3-8B (usando HuggingFace)

4

Prática: Reconhecimento de Entidades Nomeadas (Named Entity Recognition - NER)

Descrição: Encontrar as entidades nomeadas no processo

Número do processo: 123456-12.3456.7.89.1234
Classe judicial: PROCEDIMENTO COMUM CÍVEL
REQUERENTE: MINHA EMPRESA ADVOGADOS ASSOCIADOS S/S
REQUERIDO: PEDRO ALVARES CABRAL

DECISÃO INTERLOCUTÓRIA
(Emenda à Inicial)

Convido o autor a promover a emenda à inicial, no prazo de 15 dias, a fim cumprir as disposições constantes dos itens abaixo, sob pena de incidência do art. 321, parágrafo único, do Código de Processo Civil:

1. Juntar procuração atualizada, haja vista que o instrumento de ID 123456789 está datado do ano de 2022. 2)Esclarecer o ajuizamento da ação na jurisdição do Brasília/DF, haja vista que no contrato objeto da ação (ID 123456789, cláusula VIII, item 24), consta que as partes elegeram o foro da cidade de Goiânia/Goiás. Com efeito, após recente alteração, o §1º do art. 63 do CPC passou a dispor que "a eleição de foro somente produz efeito quando constar de instrumento escrito, aludir expressamente a determinado negócio jurídico e guardar pertinência com o domicílio ou a residência de uma das partes ou com o local da

DECISÃO INTERLOCUTÓRIA
(Emenda à Inicial MISC)

Convido o autor a promover a emenda à inicial, no prazo de 15 dias, a fim cumprir as disposições constantes dos itens abaixo, sob pena de incidência do art. 321, parágrafo único, do Código de Processo Civil MISC :
1) Juntar procuração atualizada, haja vista que o instrumento de ID 123456789 está datado do ano de 2022.
2)Esclarecer o ajuizamento da ação na jurisdição do Brasília LOC / DF LOC , haja vista que no contrato objeto da ação (ID 123456789, cláusula VIII, item 24), consta que as partes elegeram o foro da cidade de Goiânia LOC / Goiás LOC . Com efeito, após recente alteração, o §1º do art. 63 do CPC PER passou a dispor que "a eleição de foro somente produz efeito quando constar de instrumento escrito, aludir expressamente a determinado negócio jurídico e guardar pertinência com o domicílio ou a residência de uma das partes ou com o local da obrigação, ressalvada a pactuação consumerista, quando favorável ao consumidor". No caso dos autos, o foro eleito pelas partes guarda pertinência com o domicílio do requerente, ID 123456789.
3) Comprovar a efetiva atuação do causídico, ora requerente, na demanda objeto do contrato entabulado entre as partes. Para tanto, junte-se cópia integral do referido processo, até o momento em que efetivada a renúncia.
Sem prejuízo, à Secretaria ORG para retificação da autuação, alterando a classe processual para " Execução de Título Extrajudicial MISC ".

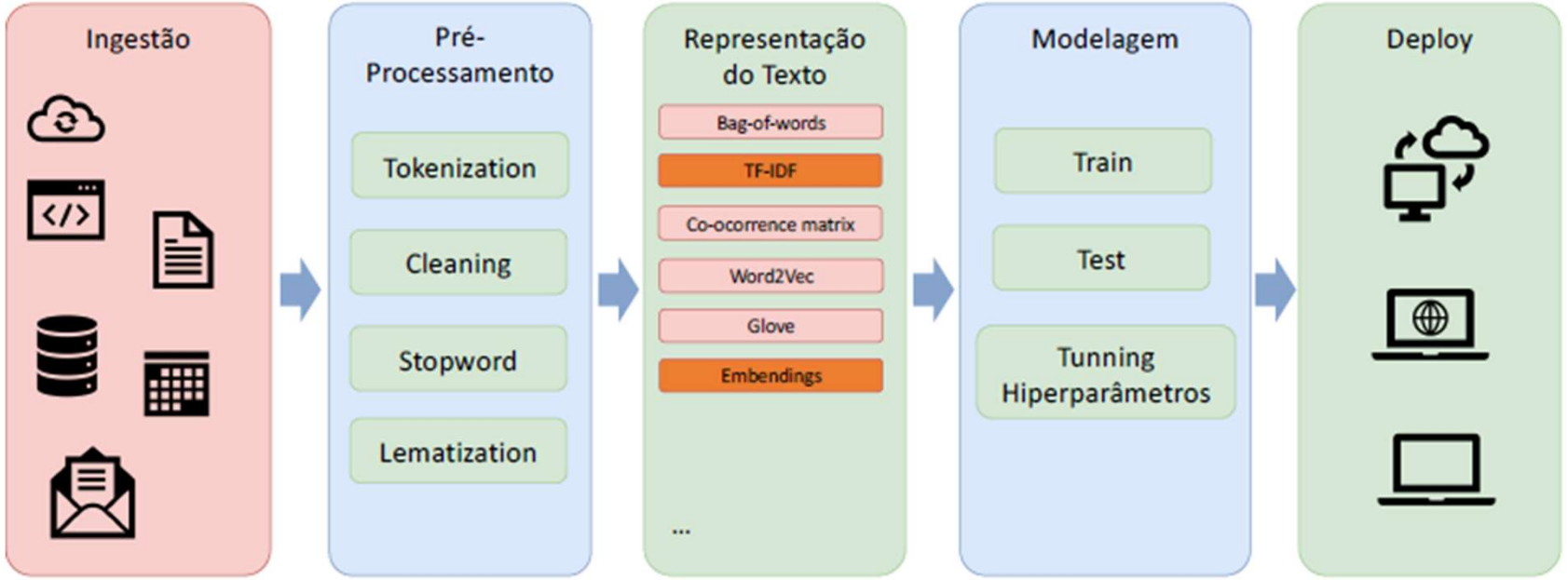
2 groq

3 spaCy

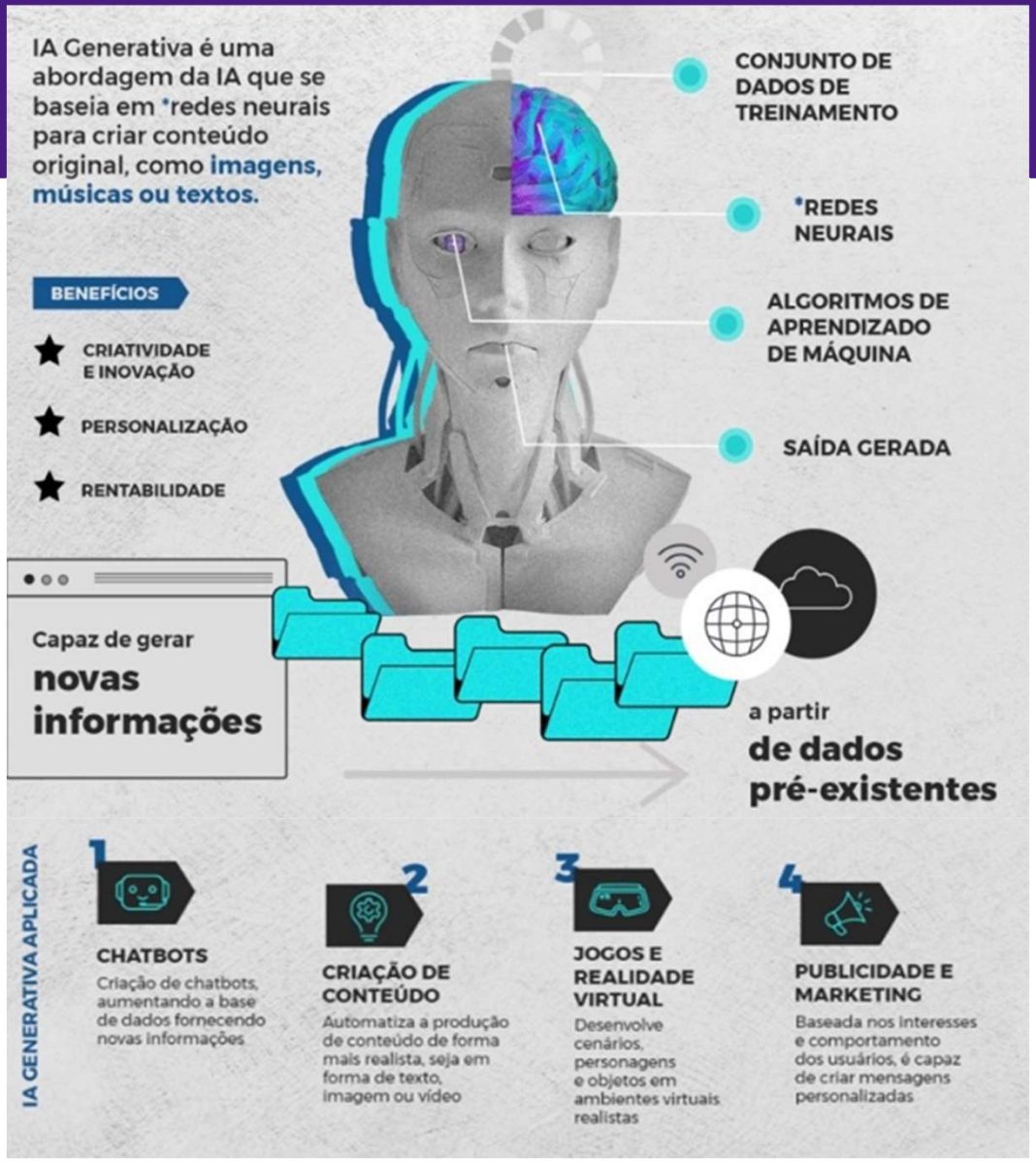
Processamento de Linguagem Natural (PLN)

- Combina linguística computacional, baseada em regras, modelos estatísticos e aprendizagem de máquina.
- Processar a linguagem humana na forma de texto ou voz, e “entender” todo o seu significado
  - intenção, conceitos, sutilezas e o sentimento do falante ou do escritor.

Um pouco de história



Processamento básico	Processamento avançado	Outros passos
Segmentação de sentenças	Extração e Recuperação de Informação	Machine Translation
Tokenização de palavras	Named Entity Recognition	Sumarização Automática
Stopwords	Extração de Relacionamento	Geração de Linguagem Natural
Remoção de digitos/pontuação, lowercasing	Análise de Discurso	Sistemas de Respostas
Stemming		Sistemas de Diálogo
Lemmatization		Image e Video Captioning
Part of Speech Tagging (POS Tagging)		Multimodel Tasks
		Reasoning over Knowledge Base



INTELIGÊNCIA ARTIFICIAL GENERATIVA:

CRIANDO NOVAS REALIDADES

https://aiotbrasil.com.br/noticias/o-que-voce-precisa-saber-sobre-ia-generativa

IA Generativa

Redes neurais tradicionais, que são frequentemente associadas à classificação e previsão

- Focada na criação de novos conteúdos a partir de dados existentes.
- Pode produzir **textos, imagens, músicas, vídeos e vozes humanas**.
- Utiliza modelos de machine learning avançados:
  - Redes Neurais Generativas Adversariais (GANs)
  - Modelos de Transformadores (Transformers, em inglês)

A principal característica dessa tecnologia é sua capacidade de aprender padrões complexos a partir de grandes volumes de dados e, em seguida, usar esse aprendizado para gerar novos conteúdos que se assemelham aos dados originais.

Redes Neurais Generativas (RNGs)

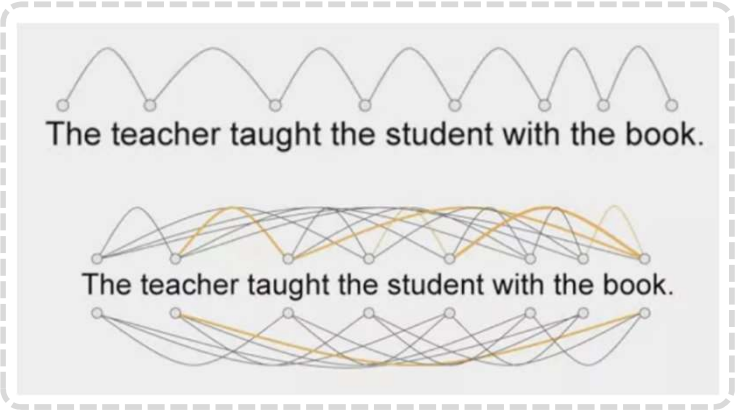
- Arquitetura complexa
- Consistem em duas partes principais:
  - gerador é responsável por criar dados,
  - discriminador avalia a autenticidade desses dados.
- Modelos de destaque
  - **Redes Neurais Generativas Adversariais (GANs):** as GANs consistem em duas redes competindo entre si – uma rede geradora e a discriminadora. .
  - **Variational Autoencoders (VAEs):** servem para gerar dados novos ao aprender uma representação compacta dos dados originais. Permitem uma geração mais controlada e interpretável.
  - **Modelos de difusão (stable diffusion):** são usados para gerar novas imagens ou outros dados de alta qualidade, começando com ruído e refinando gradualmente até que uma imagem clara seja produzida.

Modelo Transformers

- Arquitetura de rede neural projetada para lidar com sequências de dados e foram introduzidos no artigo "Attention is All You Need" em 2017.
- Principais características chave:
  - **Mecanismo de Atenção:** presta atenção a diferentes partes da entrada ao mesmo tempo, melhorando a captura de dependências de longo alcance em sequências de dados.
  - **Modelos de Transformadores (Modelo Transformer):** Modelos como GPT (Generative Pre-trained Transformer) utilizam mecanismos de atenção para gerar textos coerentes e contextualmente relevantes.

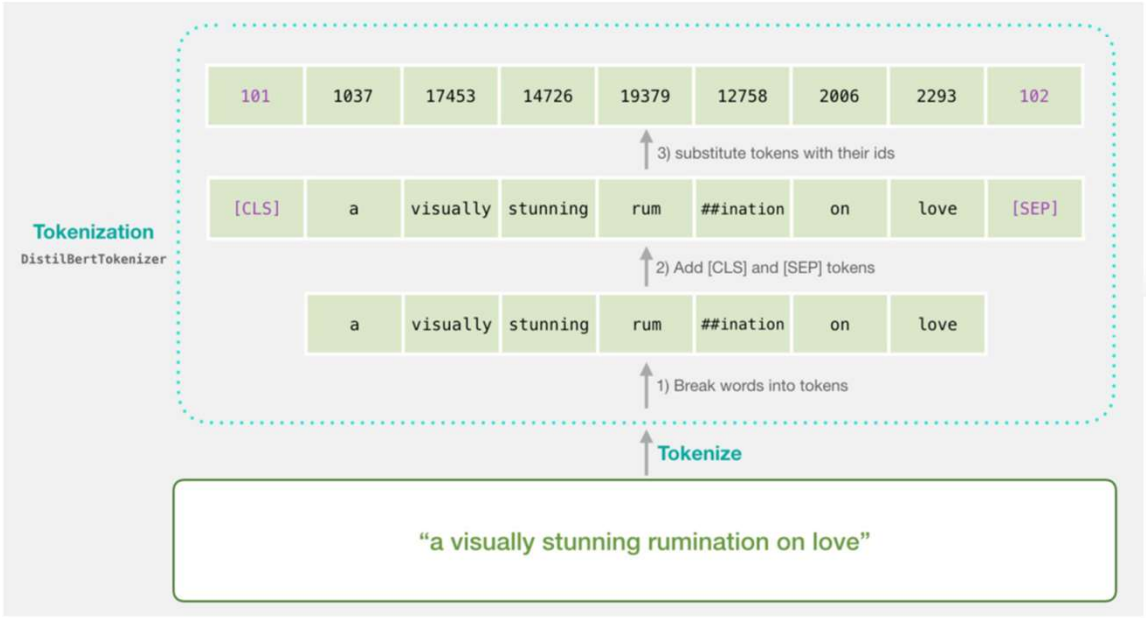
Modelo Transformers

- Utiliza um "mecanismo de auto-atenção" como camada adicional
  - Pondera o significado de cada parte dos dados



- Processam dados de entrada sequenciais em dados de saída também sequenciais
  - tarefas sequence-to-sequence
  - dados sequenciais não precisam ser processados em ordem
  - dependências de longo alcance
  - identifica o contexto para qualquer posição na sequência de entrada
- Aplicado a tarefas complexas que necessitam de memória longa

Tokenização



GPT-3 Codex

The OpenAI API can be applied to virtually any task that involves understanding or generating natural language or code. We offer a spectrum of models with different levels of power suitable for different tasks, as well as the ability to fine-tune your own custom models. These models can be used for everything from content generation to semantic search and classification.

Clear Show example

Tokens 68 Characters 373

The OpenAI API can be applied to virtually any task that involves understanding or generating natural language or code. We offer a spectrum of models with different levels of power suitable for different tasks, as well as the ability to fine-tune your own custom models. These models can be used for everything from content generation to semantic search and classification.

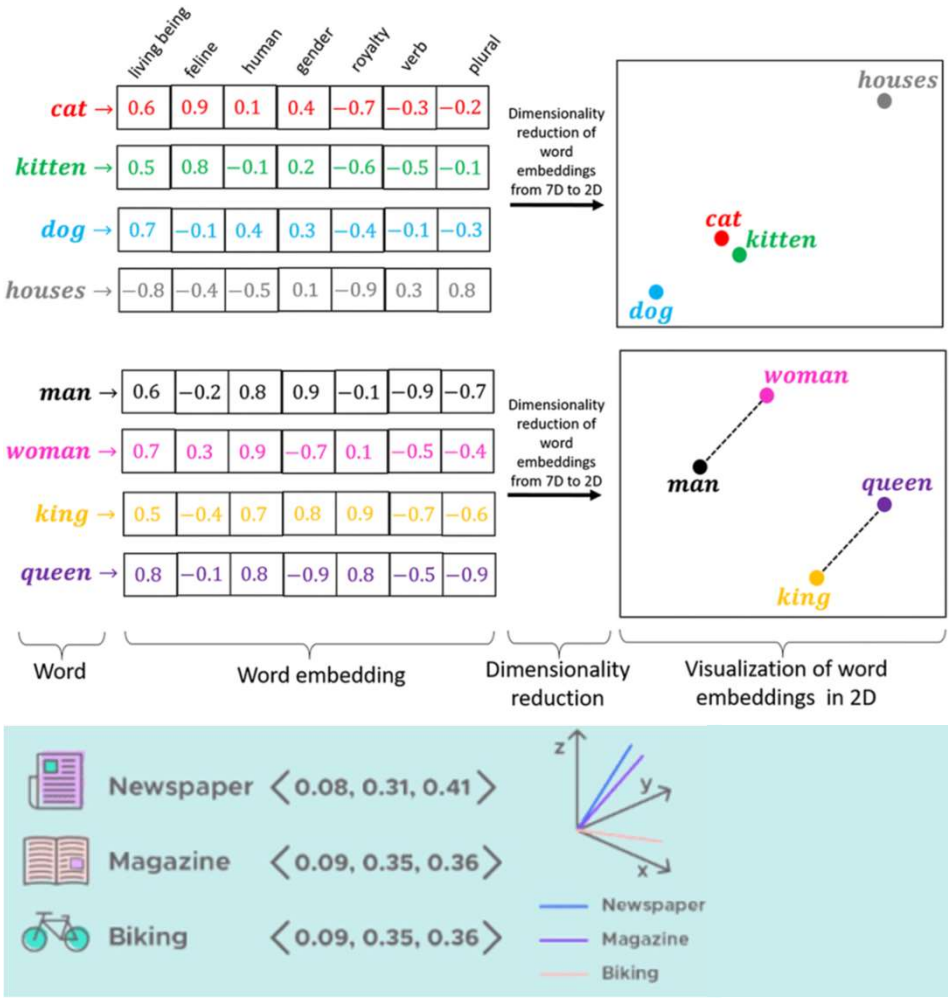
TEXT TOKEN IDS

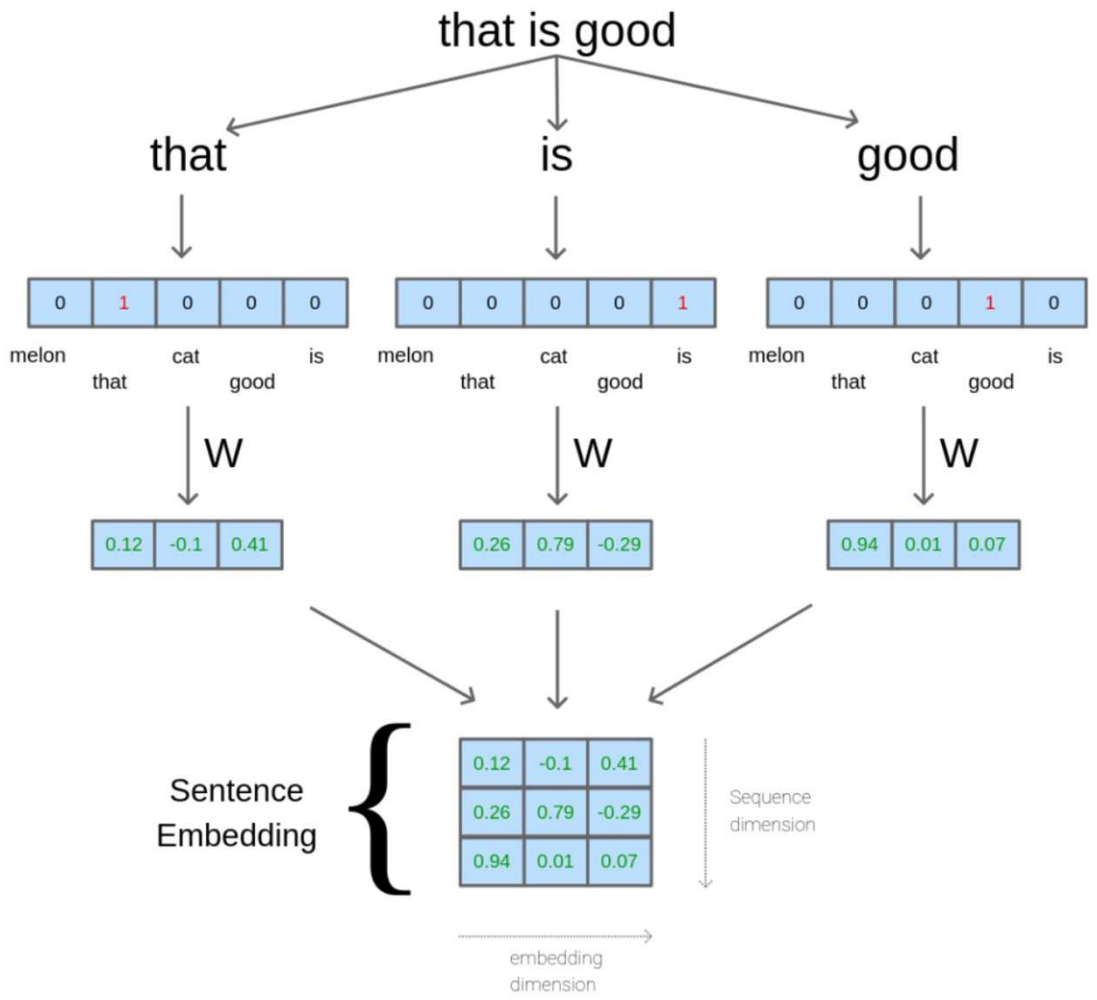
[464, 4946, 20185, 7824, 460, 307, 5625, 284, 9826, 597, 4876, 326, 9018, 4547, 393, 15453, 3288, 3303, 393, 2438, 13, 775, 2897, 257, 10958, 286, 4981, 351, 1180, 2974, 286, 1176, 11080, 329, 1180, 8861, 11, 355, 880, 355, 262, 2694, 284, 3734, 12, 83, 1726, 534, 898, 2183, 4981, 13, 2312, 4981, 460, 307, 973, 329, 2279, 422, 2695, 5270, 284, 37865, 2989, 290, 17923, 13]

TEXT TOKEN IDS

Embedding

https://tungmphung.com/the-transformer-neural-network-architecture/
http://jalammar.github.io/illustrated-transformer/
https://github.com/TranQuocTrinh/transformer
https://www.kaggle.com/code/alejopaullier/introduction-to-transformers





Camada de atenção

The **cat** sat on the **rug** and it was dry-cleaned.

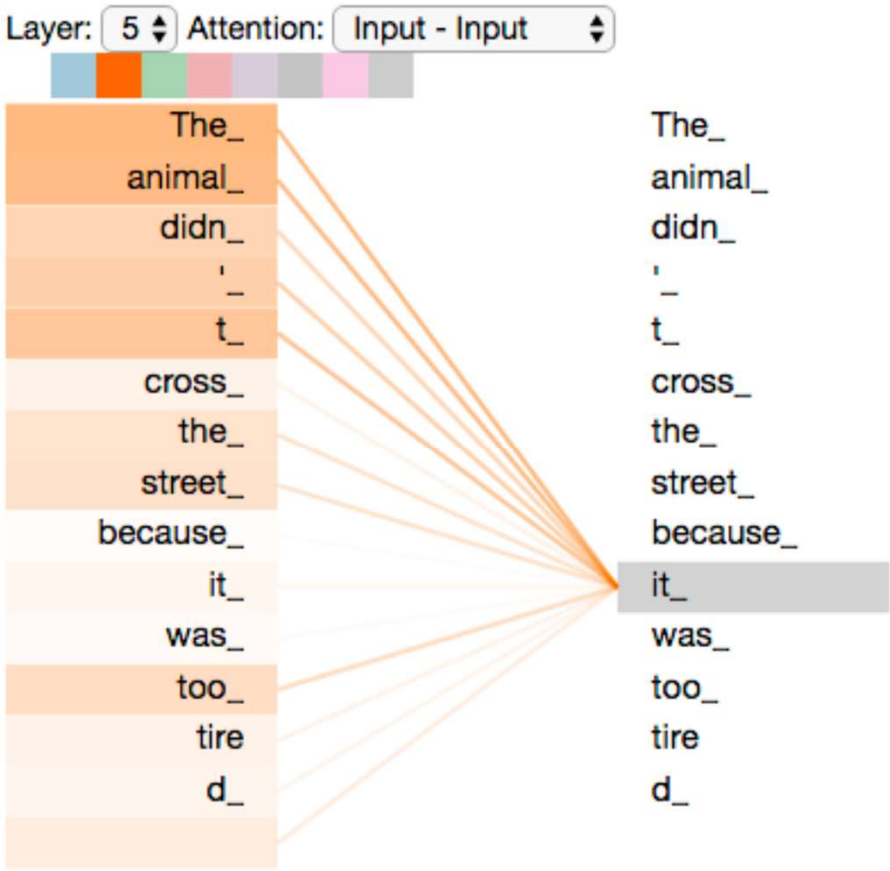
The cat sitting on the table is so cute. This one here is also …
The cat sitting on the table is so **cute**. This one here **is also** …
The cat sitting on the table is so **cute**. This one here **is also** …
adorable/cute/sweet

A camada de **atenção mede a** relevância das palavras em uma sequencia

- **Mecanismo de atenção:**

1. Descobre como uma palavra esta relacionada a outra
2. **Cada palavra é processada de acordo com as outras palavras**
3. **Concentra-se em parte** de um subconjunto das **informações** que recebem.

- A camada de atenção pode **acessar todos os estados anteriores**



Large Language Model (LLM)

- Componente chave dos sistemas de NLP
- Capacidade de gerar rapidamente texto legível.
- Treinados com enormes quantidades de dados.
- Os principais LLM possuem centenas de bilhões a mais de um trilhão de parâmetros.
  - GPT-3: 175 bilhões de parâmetros
  - Falcon: 180 bilhões de parâmetros
  - PaLM: 540 bilhões de parâmetros

Small Language Model (SLM)

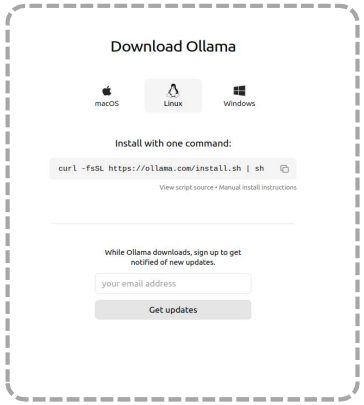
- Modelos pequenos contendo até 20 bilhões de parâmetros.
- Seu escopo e dados mais limitados os tornam mais adequados e personalizáveis para casos de uso empresarial focados.
  - DistilBERT
  - Orca 2
  - Phi 2
  - T5-Small

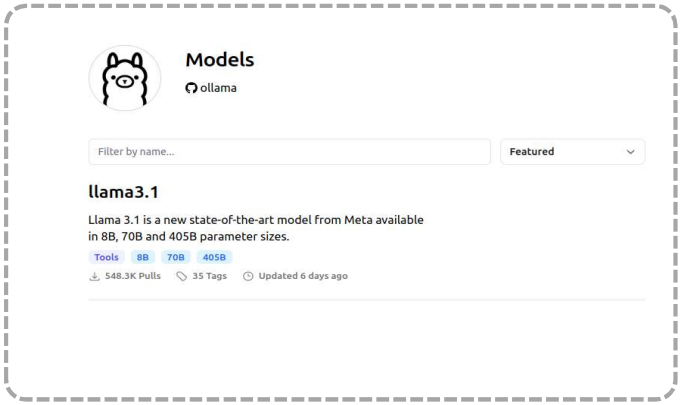
Small Language Model (SLM)

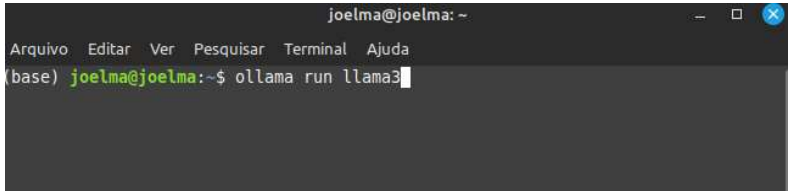
- Por exemplo, você pode ajustar esses modelos em domínios como:
  - **Médica**: podem ser **treinados em textos médicos, como artigos de pesquisa, ensaios clínicos e prontuários de pacientes**, para traduzir termos e conceitos médicos com precisão.
  - **Jurídica**: podem ser **treinados em textos jurídicos, como contratos, patentes e decisões judiciais**, para traduzir termos e conceitos jurídicos com precisão.
  - **Técnica**: podem ser **treinados em textos técnicos, como manuais, especificações e códigos**, para traduzir termos e conceitos técnicos com precisão.

Onde encontrar um LLM

- Ollama
  - Baixa uma versão otimizada para execução local (não tem a mesma precisão)
  - Instalar o Ollama (https://ollama.com/)
  - Depende da performance da sua máquina







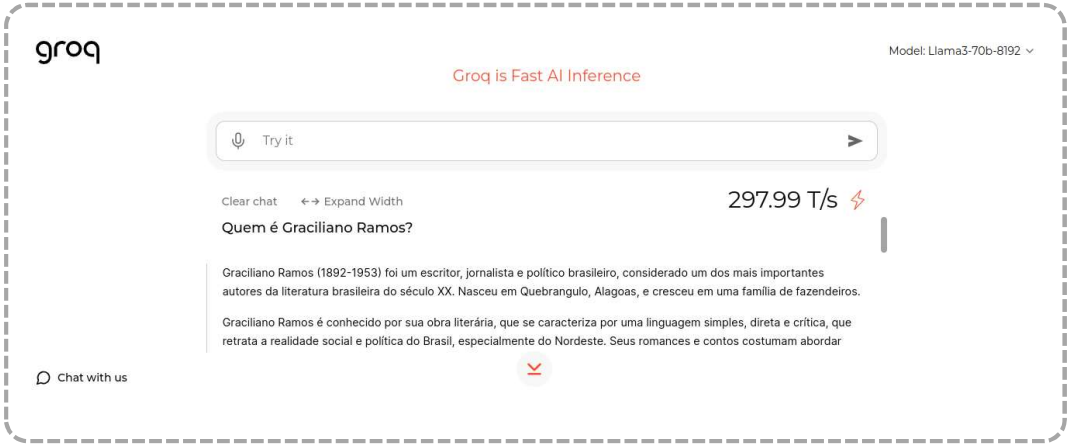
- Depois de baixar e instalar o Ollama execute
- A primeira vez o LLM será baixado

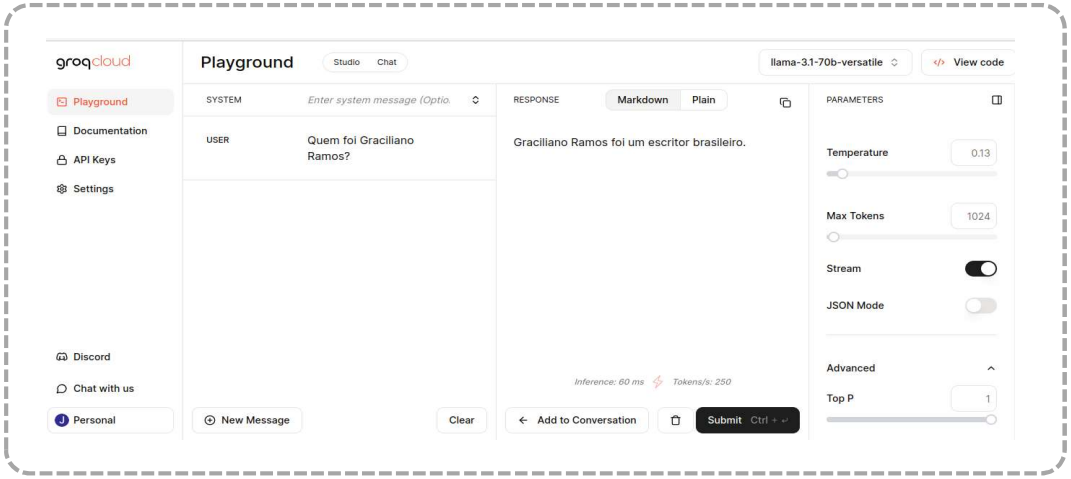

TEORIA
+
PRÁTICA

Onde encontrar um LLM

- Grog
  - A empresa Groq Inc., criou seu próprio chip ASIC, chamado de Unidade de Processamento de Linguagem (LPU), para executar modelos LLMs sem depender de GPUs
  - Plataforma para os desenvolvedores acessarem as LPUs como mecanismos de inferência para LLMs, especialmente os de código aberto, como Llama, Mixtral e Gemma
  - O playground do Groq oferece acesso gratuito ao Gemma 7B, ao Llama 3 70B e 8B e ao Mixtral 8x7b.

https://console.groq.com/playground





Prática: Reconhecimento de Entidade Nomeada

- Processar dados estruturados e não estruturados e classifique essas entidades nomeadas em categorias predefinidas.
  - Algumas categorias comuns: nome, local, empresa, horário, valores monetários, eventos e muito mais.

- Apple: é rotulado como ORG (Organização) e destacado em vermelho.
- Hoje: é rotulado como DATA e destacado em rosa.
- Segundo: é rotulado como QUANTIDADE e destacado em verde.
- iPhoneSE: é rotulado como COMM (produto comercial) e destacado em azul.

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on


TEORIA
+
PRÁTICA

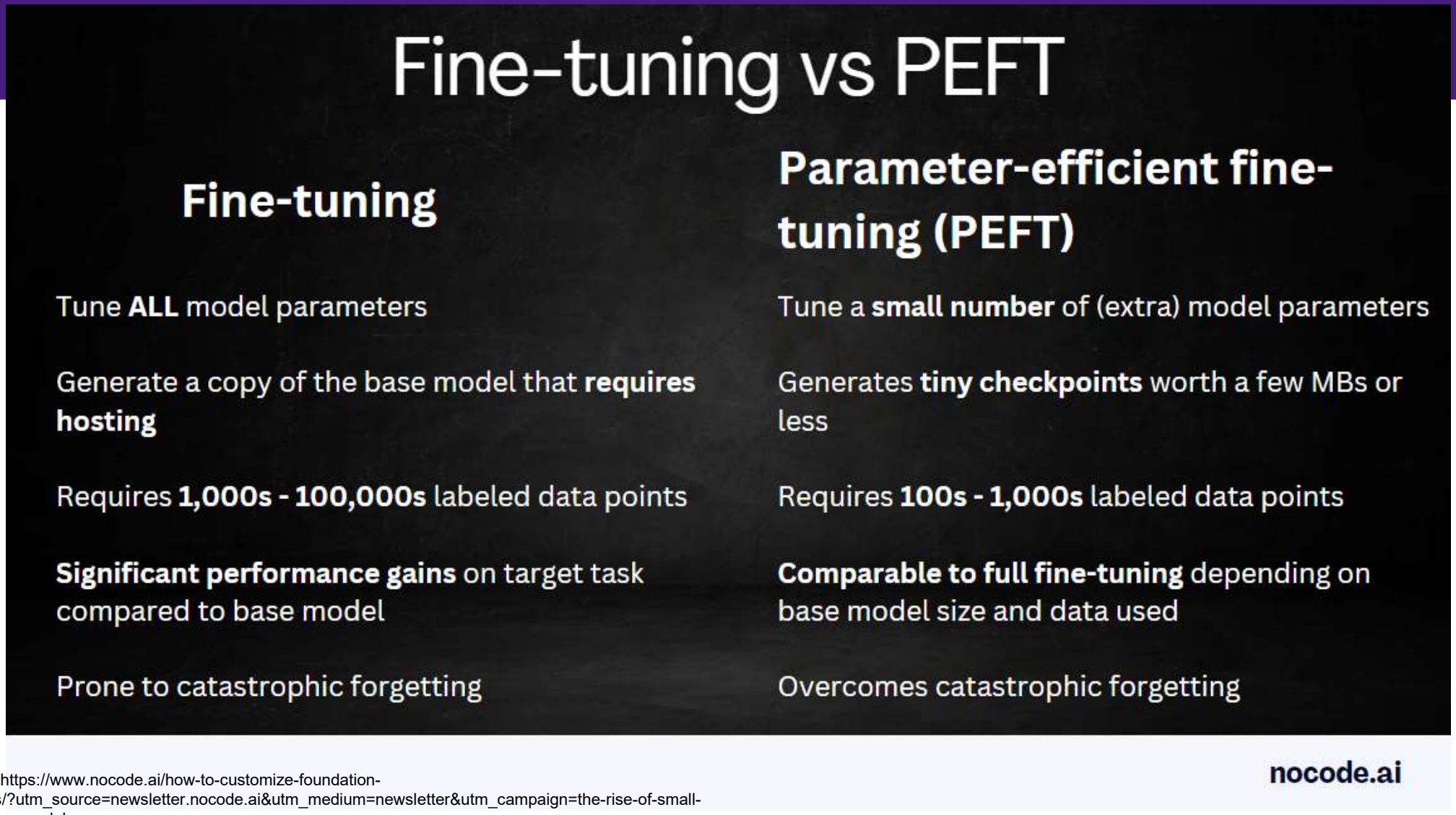
Fine-tuning

- Dado um modelo pré-treinado e treina-o ainda mais em um conjunto de dados específico de um domínio.
- Geralmente envolve:
  - congelar as camadas iniciais do modelo pré-treinado, que são responsáveis por aprender recursos gerais, como bordas, texturas e formas básicas.
  - camadas finais são descongeladas e treinadas em um novo conjunto de dados.

Fine-tuning Supervisionado (Supervised Fine-Tuning - SFT)

- Abordagem padrão para o fine-tuning.
- O modelo é treinado em um dataset rotulado, adaptado à tarefa específica.

Full parameter fine-tuning: fine-tuning de todo o modelo.
Parameter-efficient fine-tuning (PEFT): fine-tuning em um conjunto específico de parâmetros.
Instruction fine-tuning: fine-tuning baseado em um instruction-format dataset.



Fonte: https://www.nocode.ai/how-to-customize-foundation-models/?utm_source=newsletter.nocode.ai&utm_medium=newsletter&utm_campaign=the-rise-of-small-language-models

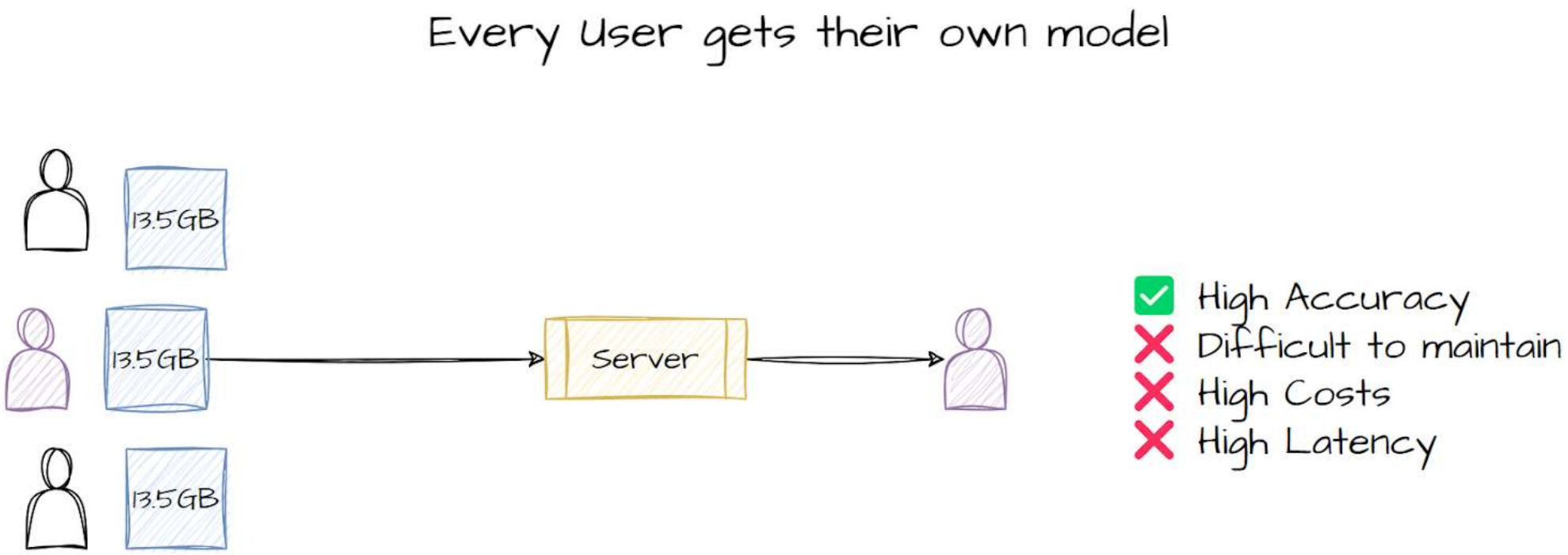
Fine-tuning Supervisionado (Supervised Fine-Tuning - SFT)

- **Few-Shot Learning:** fornece ao modelo um alguns exemplos (ou shots) da tarefa desejada no início dos prompts de entrada.

- **Full Transfer Learning:** um modelo pré-treinado é utilizado como ponto de partida para uma nova tarefa, mas todas as camadas do modelo são ajustadas durante o treinamento. Isso significa que o modelo pré-treinado é usado como uma espécie de "rede inicial" e, em seguida, todas as camadas são treinadas em conjunto com o novo conjunto de dados.

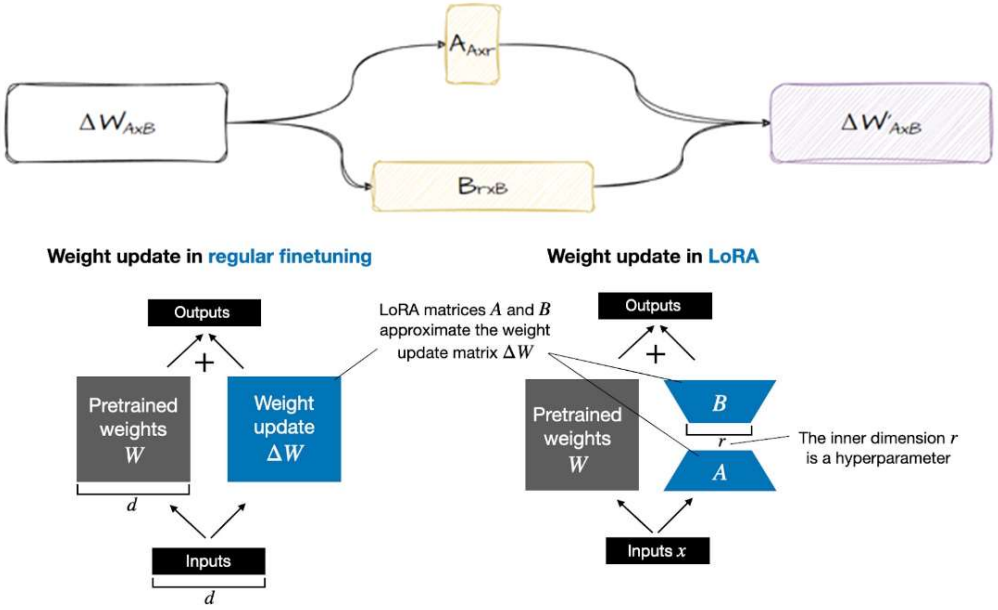
- **Fine-Tuning Específico de Domínio:** Esta variante de fine-tuning visa aclimatar o modelo para compreender e gerar texto pertinente a um domínio ou indústria específica. Por exemplo, para desenvolver um chatbot para uma aplicação jurídica, o modelo seria treinado em textos jurídicos para refinar suas habilidades de compreensão de linguagem no contexto.

Adaptação e Quantização

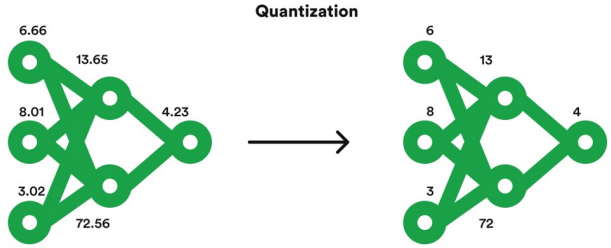


Adaptação e Quantização

- A adaptação tem como objetivo ajustar modelos pré-treinados para novas tarefas ou melhorar sua performance em tarefas existentes com eficiência, utilizando menos parâmetros
- LoRA é uma técnica que adapta modelos de aprendizado profundo usando matrizes de baixa ordem.



- Quantização é o processo de converter os pesos (e ativações) de um modelo para uma precisão mais baixa. Por exemplo, pesos armazenados usando 16 bits podem ser convertidos para uma representação de 4 bits.
- QLoRA (Quantized Low-Rank Adaptation) combina a quantização com a adaptação de baixa ordem. Isso significa que os pesos do modelo são quantizados para uma precisão mais baixa (como 4 bits) e, em seguida, ajustados usando a técnica LoRA.



https://www.mercity.ai/blog-post/guide-to-fine-tuning-llms-with-lora-and-qlora

https://dev.to/jackrover/understanding-quantization-in-ai-a-comprehensive-guide-including-lora-and-qlora-4dl1

https://magazine.sebastianraschka.com/p/practical-tips-for-finetuning-llms


TEORIA
+
PRÁTICA