

Processamento de Linguagem Natural e Imagens

Aula 5

James Moraes de Almeida



Conteúdo do curso

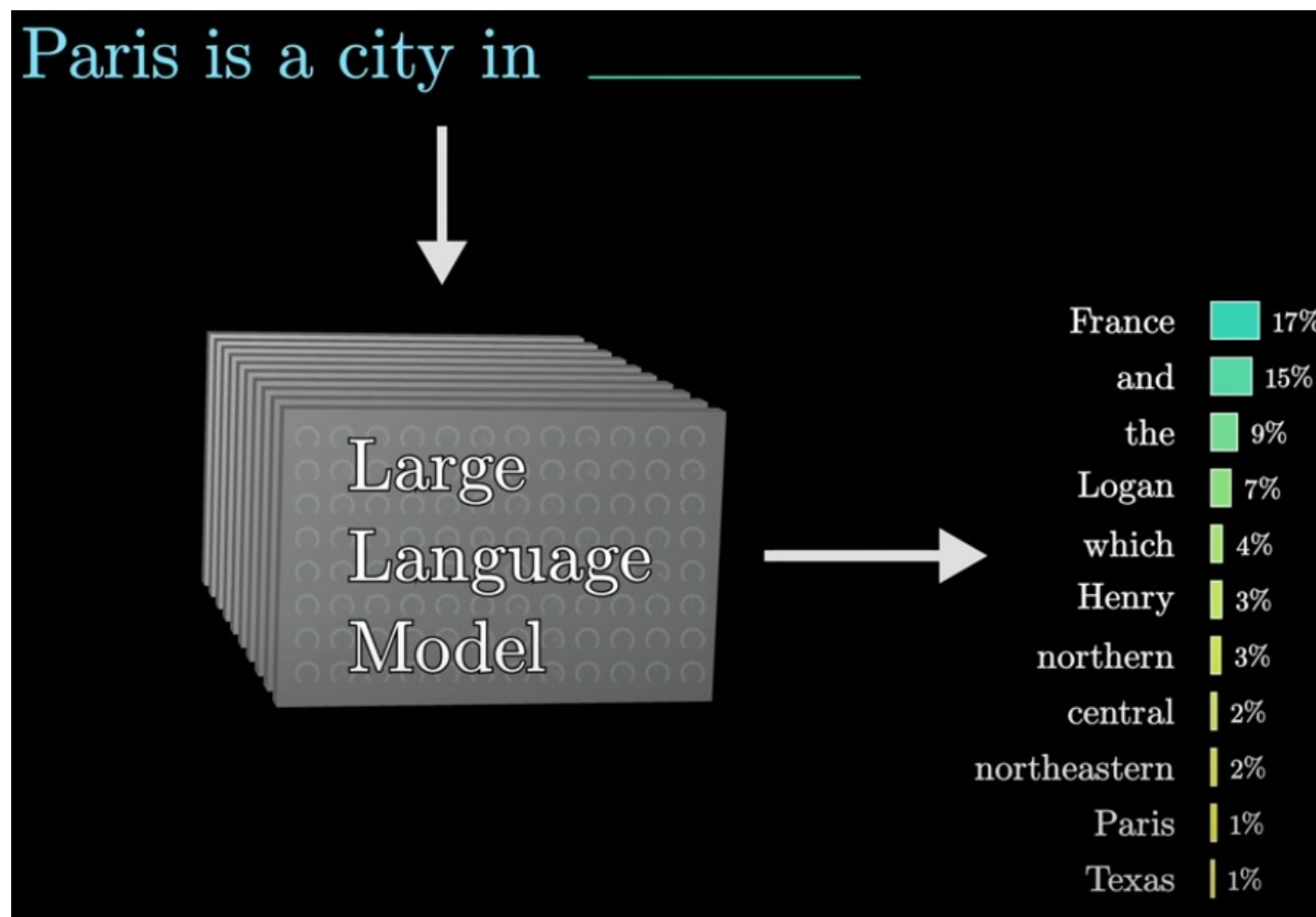
- Qual o plano?
 - Processamento de linguagem natural.
 - Mineração de dados e construção do corpus.
 - Pré-processamento, tokenização e lematização.
 - Bag-of-Words, TF-IDF e similaridade de cossenos.
 - Vetores de palavras e busca semântica.
 - Extração com expressões regulares.
 - Grandes Modelos de Linguagem para extração de dados.
 - Modelos de sequência e a arquitetura Transformer.
- Material principal: <https://www.youtube.com/@futuremojo>

- Onde paramos
 - O regex extraiu 45 linhas de 48.719 sentenças — e 44 estavam certas.
 - Precisão altíssima. Abrangência de 0,09%.
 - E duas extrações só foram pegadas na revisão manual:
 - "...was about 500 MPa higher than..." — o número era uma diferença.
 - "...maximum compressive yield strength..." — outra grandeza.
 - Nos dois casos a sintaxe estava certa e o sentido, errado.
- Hoje vamos ver o que é a ferramenta que entende o sentido.

- Objetivos desta aula
 - Explicar o que um modelo de linguagem faz, em uma frase.
 - Entender por que prever a próxima palavra basta para tanta coisa.
 - Saber o que é um token, e por que ele decide o seu custo.
 - Reconhecer as três etapas de treino de um modelo moderno.
 - Passar da extração por sentença para o abstract inteiro.
 - Estimar custo, detectar alucinação e validar o resultado.

O que são grandes modelos de linguagem?

- Os modelos de linguagem são modelos generativos.
 - Geram a próxima palavra pela probabilidade dela — um super autocompletar.



O que são grandes modelos de linguagem?

- Os modelos de linguagem são modelos generativos.
- Geram a próxima palavra nela probabilidade dela — um super autocompletar

TRANSFORMER EXPLAINER

Examples ▾

Paris is a city in which

Generate

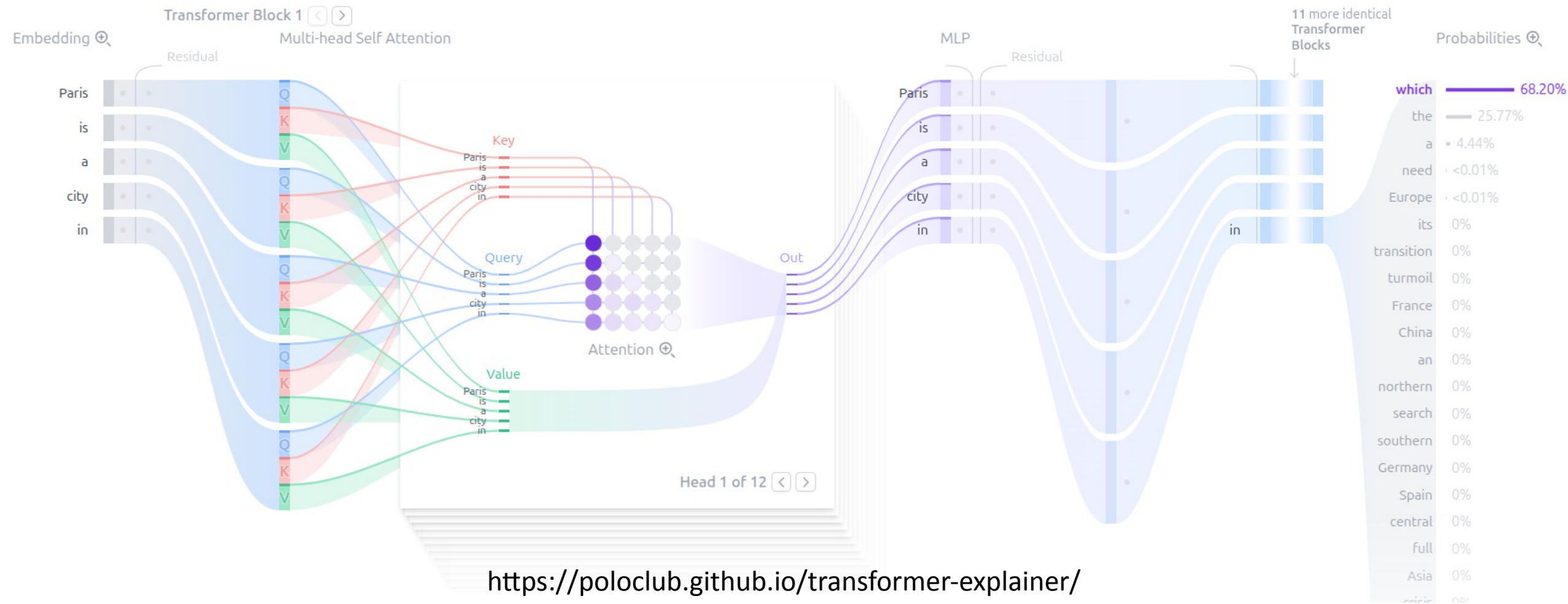
Temperature

0.6

Sampling

Top-k

k=5



<https://poloclub.github.io/transformer-explainer/>

O que são grandes modelos de linguagem?



- É só isso mesmo?
 - Prever a próxima palavra parece pouco. Mas repare no que exige:
 - "O limite de escoamento da liga foi de 1511 ____" → uma unidade.
 - "A capital da França é ____" → um fato.
 - "Traduza para o inglês: bom dia. → ____" → uma tradução.
 - "Extraia a composição desta frase: ... → ____" → o nosso trabalho.
 - Para prever bem, o modelo precisa ter aprendido gramática, fatos e tarefas.
- A tarefa é simples. O que ela obriga a aprender é que não é.

O que são grandes modelos de linguagem?



- Token, não palavra
 - O texto é fatiado em pedaços — os tokens — antes de entrar no modelo.
 - Palavras comuns viram um token; raras viram vários.
 - "AlCoCrFeNi" pode virar quatro ou cinco tokens.
- Por que isso importa para você:
 - A cobrança é por token, de entrada e de saída.
 - Português gasta mais tokens que inglês para dizer o mesmo.
 - É por isso que o modelo erra ao contar letras de uma palavra:
 - ele nunca viu as letras — viu os pedaços.

O que são grandes modelos de linguagem?

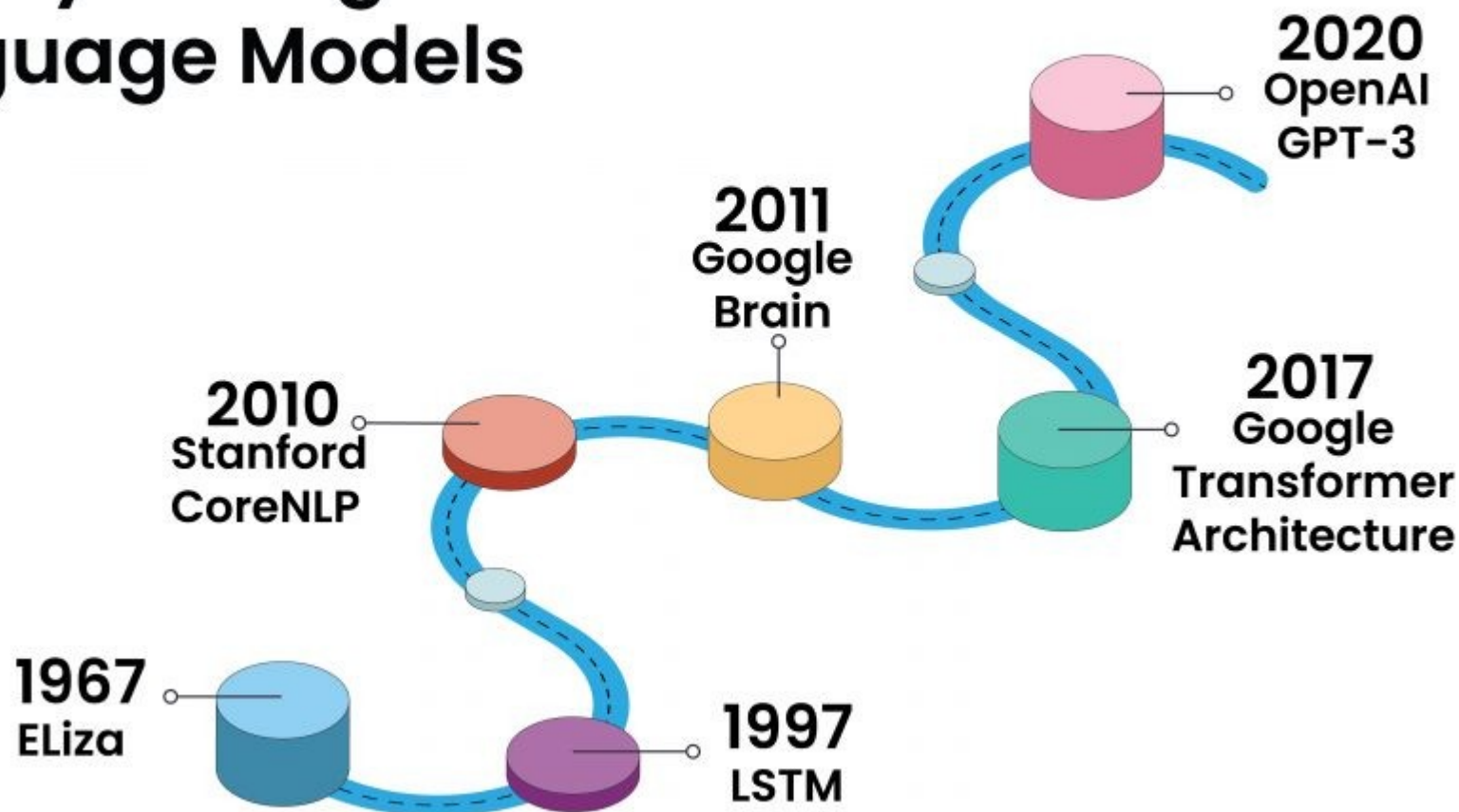


- A janela de contexto
 - É tudo que o modelo consegue enxergar de uma vez: o prompt inteiro.
 - System prompt + exemplos + a sua pergunta + raciocínio + a resposta que ele vai gerar.
 - O que está fora da janela simplesmente não existe para ele.
 - Não há memória entre chamadas — é a lista que você reenvia (Aula 04).
 - Na nossa extração isso decide quantos abstracts cabem por chamada.
- Janela maior custa mais. Mandar o corpus inteiro raramente é boa ideia.

O que são grandes modelos de linguagem?

- Quando começaram os modelos de linguagem?

History of Large Language Models



O que são grandes modelos de linguagem?



- O que aconteceu de 2020 para cá
 - A linha do tempo do slide anterior para no GPT-3. Falta o que veio depois:
 - 2022 — ajuste de instrução: o modelo passa a obedecer pedidos, não só completar.
 - 2022 — RLHF: humanos ordenam respostas e o modelo aprende a preferência.
 - É o que transformou um autocompletar num assistente.
 - 2023 — pesos abertos (Llama, Mistral, Qwen): dá para rodar na sua máquina.
 - É por isso que a Ilum pode ter o seu próprio servidor.
 - 2024 — multimodal de verdade: texto, imagem, áudio no mesmo modelo.
 - 2024 — modelos de raciocínio: o modelo "pensa" antes de responder.

O que são grandes modelos de linguagem?



- 2017 — "Attention Is All You Need"
 - Introduziu o Transformer. A ideia: cada palavra olha para todas as outras.

Attention Is All You Need

Attention is all you need

[A Vaswani, N Shazeer, N Parmar...](#) - Advances in neural ..., 2017 - proceedings.neurips.cc

... to attend to **all** positions in the decoder up to and including that position. **We need** to prevent ... **We** implement this inside of scaled dot-product **attention** by masking out (setting to $-\infty$) ...

☆ Save 🔖 Cite Cited by 267078 Related articles All 26 versions 🔗

Ashish Vaswani*

Google Brain

avaswani@google.com

Noam Shazeer*

Google Brain

noam@google.com

Niki Parmar*

Google Research

nikip@google.com

Jakob Uszkoreit*

Google Research

usz@google.com

Llion Jones*

Google Research

llion@google.com

Aidan N. Gomez* †

University of Toronto

aidan@cs.toronto.edu

Łukasz Kaiser*

Google Brain

lukaszkaizer@google.com

Illia Polosukhin* ‡

illia.polosukhin@gmail.com

O que são grandes modelos de linguagem?

- Qual é a arquitetura?
 - Hoje fica só o mapa.

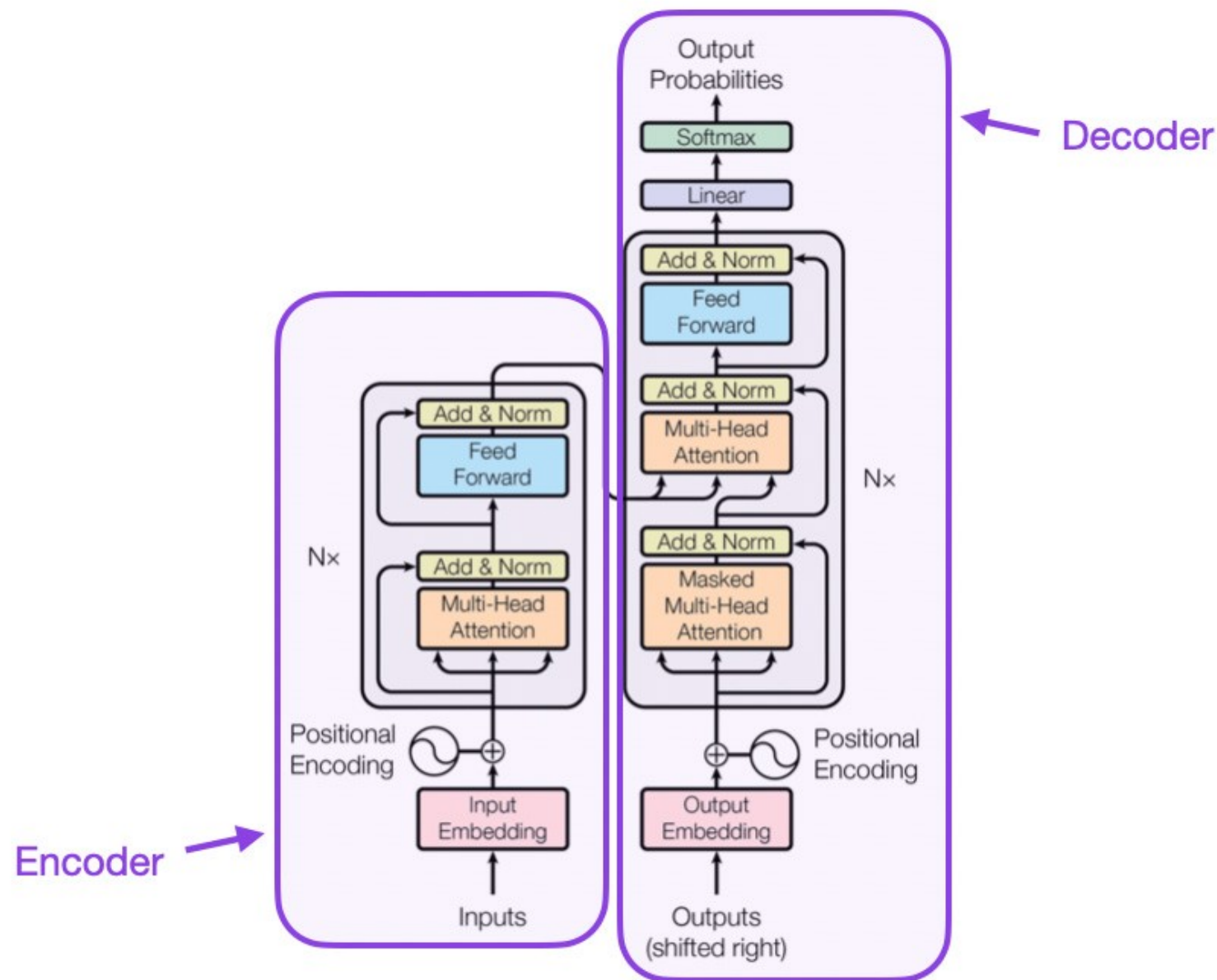


Figure 1: The Transformer - model architecture.

O que são grandes modelos de linguagem?



- Como são treinados — três etapas
 - 1. Pré-treino: prever a próxima palavra em bilhões de textos.
 - É onde entra a gramática, os fatos e a maior parte do custo.
 - 2. Ajuste de instrução: exemplos de pedido e resposta bem feita.
 - É o que faz ele responder em vez de continuar o seu texto.
 - 3. Alinhamento: humanos comparam respostas e o modelo aprende a preferida.
 - É de onde vem o tom, a recusa e boa parte do comportamento.
- Quando você escreve um system prompt, está conversando com a etapa 2.

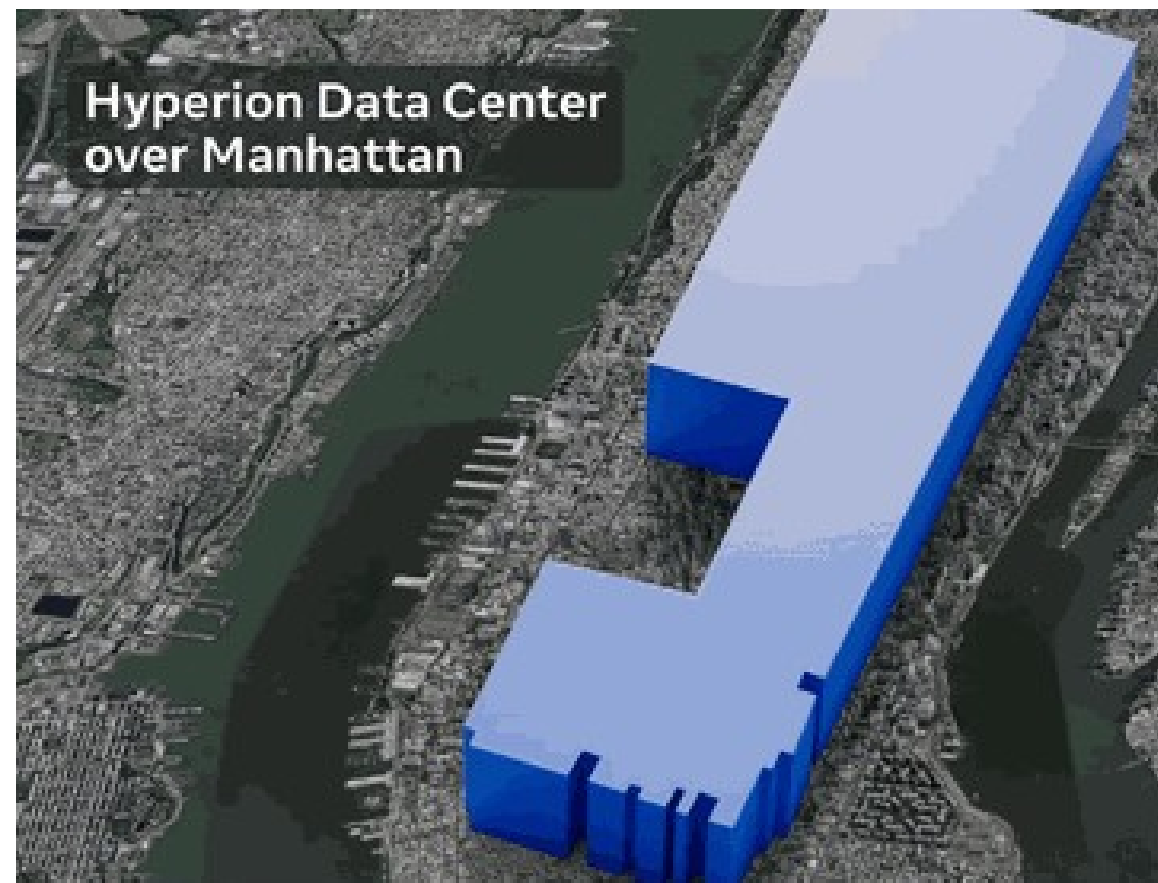
O que são grandes modelos de linguagem?



- A escala
 - Treinar é caro de um jeito difícil de imaginar:
 - Máquinas com mais de 100 mil GPUs.
 - Cada GPU custa ~30 mil dólares — 3 bilhões de dólares só em GPUs.
 - E os modelos são enormes:
 - Os de fronteira estimam-se estar em 10 trilhões de parâmetros.
 - Os abertos que rodam em clusters de gpu: por volta de 2 trilhões de parâmetros.
 - Os abertos para uma GPU: ~100 bilhões de parâmetros.
 - Os que rodam num notebook: de 1 a 8 bilhões.
- Usar é barato. Treinar do zero está fora do alcance de quase todos.

O que são grandes modelos de linguagem?

- A escala
 - Treinar é caro de um jeito difícil de imaginar:
 - Datacenter de 50 bilhões de dólares (Meta AI).
 - Investimento INICIAL!
 - 2 GW de potência.



O que são grandes modelos de linguagem?



- O cenário em 2026
 - Modelos de peso fechado — API de terceiros.
 - GPT, Claude, Gemini. Os maiores, e você não vê os pesos.
 - Modelos de peso aberto — você baixa e roda.
 - GPT-OSS, Qwen, DeepSeek, Gemma, Kimi, GLM, ...
 - Especializados por tamanho e por tarefa:
 - Pequenos (1-8 B) rodam num notebook; grandes exigem um servidor.
 - De raciocínio, de código, de embeddings, multimodais.

O que são grandes modelos de linguagem?



- Por que a Ilum tem o seu próprio servidor
 - Com peso aberto, o modelo roda dentro da instituição.
 - O texto que você manda não sai para uma empresa de fora.
 - Importa quando o dado é de pesquisa não publicada ou dado sensível/sigiloso.
 - O custo vira infraestrutura, não assinatura.
 - E dá para escolher e trocar o modelo.
- A contrapartida: o modelo local costuma ser menor que os de fronteira.

Extração de dados em escala

- Da sentença para o abstract inteiro
 - Na Aula 04 o modelo lia uma sentença por vez — como o regex fazia.
 - Agora ele lê o resumo inteiro. O que muda:
 - Ele resolve referências entre frases ("a liga acima", "esta amostra").
 - Ele vê a condição do ensaio, que costuma estar noutra frase.
 - Some a etapa de segmentação com spaCy.
 - E o que piora: mais tokens por chamada, e mais espaço para inventar.
- É a troca da aula: mais contexto, mais custo, mais risco.

Extração de dados em escala

- Few-shot: mostrar em vez de descrever
 - A lista de mensagens da nossa extração tem cinco itens:
 - system — o papel e as regras.
 - user — um abstract de exemplo.
 - assistant — a resposta ideal para aquele abstract.
 - user — o abstract de verdade.
 - O par do meio nunca aconteceu: foi escrito por você (Aula 04).
- Um exemplo bem escolhido resolve o formato inteiro de uma vez.

Extração de dados em escala

- O que muda quando são 6.228
 - Tempo: uma chamada por resumo, em série, leva horas.
 - Custo: cobrado por token — estime ANTES de disparar o laço.
 - Rode 20, veja o consumo em Console → Usage Logs, multiplique.
 - Falhas: a rede cai, a chamada estoura o tempo, a resposta vem quebrada.
 - try/except por item, e salve o resultado a cada passo.
 - Guarde o índice: você vai querer retomar de onde parou.
- Um laço sem tratamento de erro perde três horas de trabalho no item 4.000.

Extração de dados em escala



- Alucinação
 - Um regex nunca devolve um número que não está no texto.
 - Um modelo pode. E devolve com a mesma confiança do acerto.
 - Frases com muitos números são o terreno mais perigoso.
 - Pedir um campo que não existe no texto é um convite ao invento.
 - Defesas práticas:
 - Permita explicitamente a resposta vazia: "se não houver, devolva []".
 - Confira se o valor extraído aparece literalmente no texto de origem.
- Essa checagem é barata e pega a maior parte dos casos.

Extração de dados em escala

- Como saber se funcionou
 - A pergunta não é "rodou?", é "quanto está certo?".
 - O procedimento é o mesmo da Aula 04:
 - Sorteie 30 extrações com uma semente fixa.
 - Confira à mão, contra a frase de origem.
 - Calcule a precisão — e, se der, a abrangência.
 - Compare com o placar do regex: 45 linhas, 44 corretas.
- Sem esse número, você tem uma impressão. Com ele, tem um resultado.

Extração de dados em escala



- Limites e cuidados
 - Corte de conhecimento: o modelo não sabe o que saiu depois do treino.
 - Viés: ele aprendeu do texto humano, com tudo que vem junto.
 - Não determinismo: mesmo com temperatura baixa, pode variar.
 - Privacidade: o seu texto vai para um servidor.
 - Aqui é o da Ilum — mas o hábito de perguntar "para onde isso vai?" é o que importa.
- Em trabalho publicado, diga qual modelo usou, com que prompt e que versão.

Extração de dados em escala

- Regex ou LLM? — revisitado
 - Regex, para o que tem forma fixa: unidades, DOIs, códigos, datas.
 - Instantâneo, de graça, determinístico, e nunca inventa.
 - LLM, para o que depende do sentido.
 - Diferença, condição do ensaio, referência entre frases.
 - Custa, demora e pode alucinar.
 - Na prática, os dois: o regex filtra, o modelo interpreta.
- Filtrar com regex antes de chamar o modelo corta a conta pela metade.

Extração de dados em escala



- Mão na massa
 - O receituário da IlumA: padrões prontos para copiar no seu projeto.
 - Conexão, abstract inteiro com few-shot, JSON confiável.
 - Lote com retomada, estimativa de custo, checagem de alucinação.
 - O notebook da Aula 04 continua valendo para o passo a passo.
- E agora: o que cada grupo está extraíndo?

O que são grandes modelos de linguagem?



- Para aprofundar
 - 3Blue1Brown, com áudio em português — a melhor explicação visual:
 - <https://www.youtube.com/watch?v=wjZofJX0v4M>
 - "Attention Is All You Need" (Vaswani et al., 2017).
- Então vamos aprender a usar um modelo.

OBRIGADO

James Moraes de Almeida
james.almeida@ilum.cnpem.br

ilum.cnpem.br