Overview and Transformer Language Models

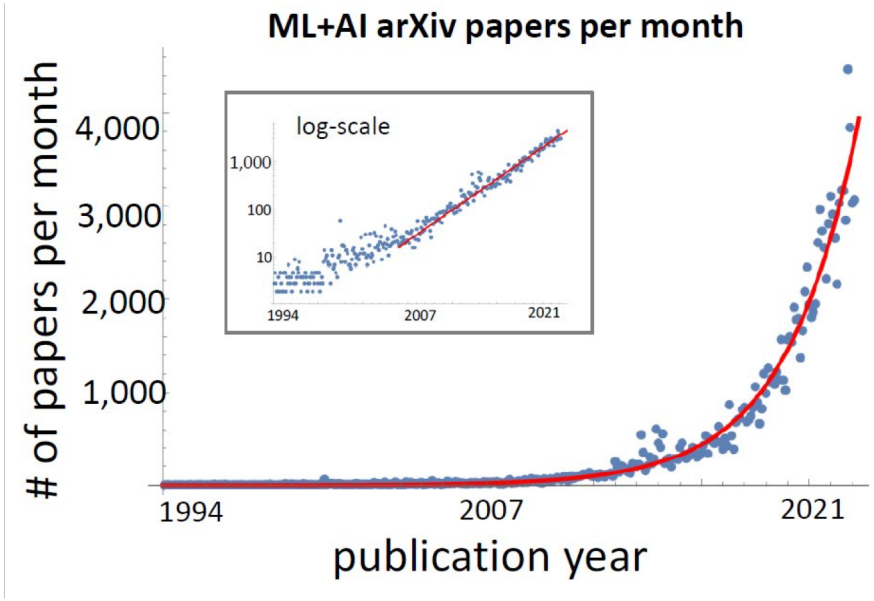
Eric Wallace

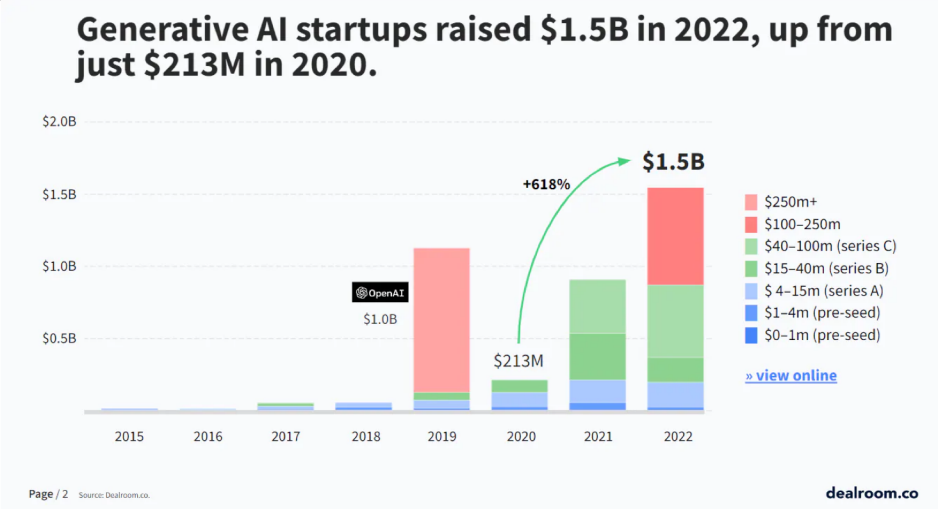
CS 288, 3/13/2023

Logistics

- 4 traditional lectures + ~8 days of mixed lectures and panels/discussions
- HW4 out Wednesday. Due Wednesday after spring break
  - Using and finetuning LMs with Huggingface
- HW5 out after spring break. Due sometime end of April.
  - Prompting ChatGPT to solve projects 1-3
- No final exam.
- Lecture recordings?

Immense Interest

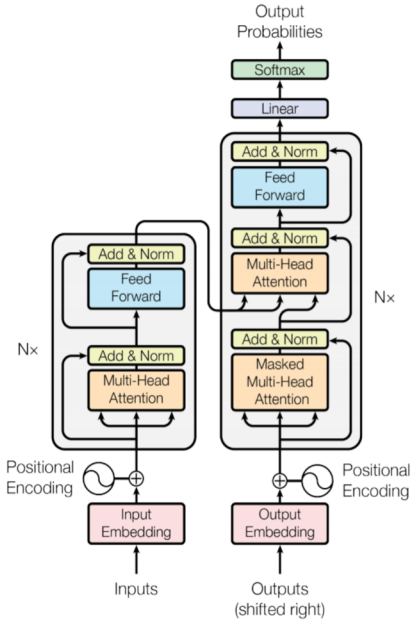




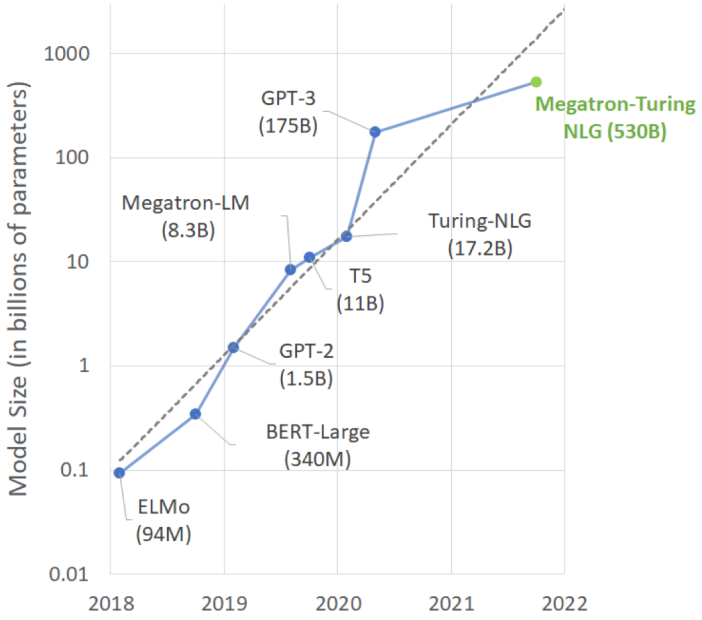
The Era of Rapid Scaling in NLP

2017: Transformer is introduced

[Vaswani+17] Attention is All You Need



2022: Large-scale Transformer models are the dominant approach for many NLP tasks



Demos

- ChatGPT
- Stable Diffusion
- InstructGPT

Today's Lecture

- Language modeling as the ultimate task
- Transformer models
- Overview of remainder of the course

Language Modeling

$$p(x_1, \ldots, x_L)$$

Language Modeling

$$p(x_1, \ldots, x_L)$$

$$p(\text{the, mouse, ate, the, cheese}) = 0.02,$$

$$p(\text{the, cheese, ate, the, mouse}) = 0.01,$$

$$p(\text{mouse, the, the, cheese, ate}) = 0.0001$$

Neural Language Models

$$\prod_{i=1}^{L} p(x_i \mid x_{1:i-1})$$

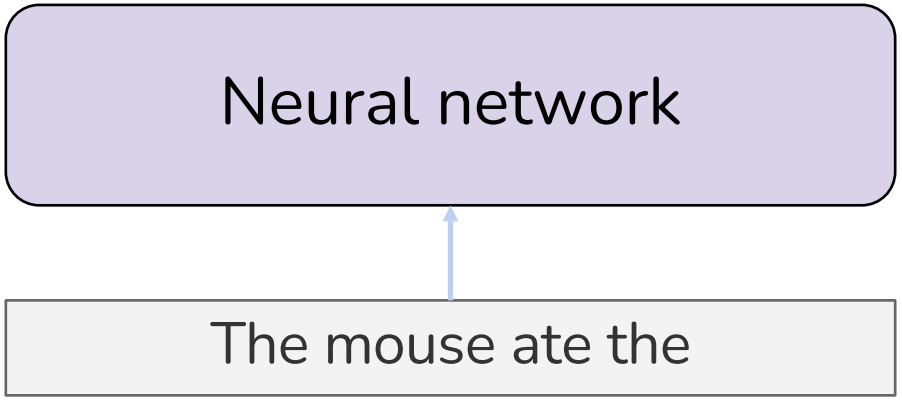
Neural Language Models

Prompt

The mouse ate the

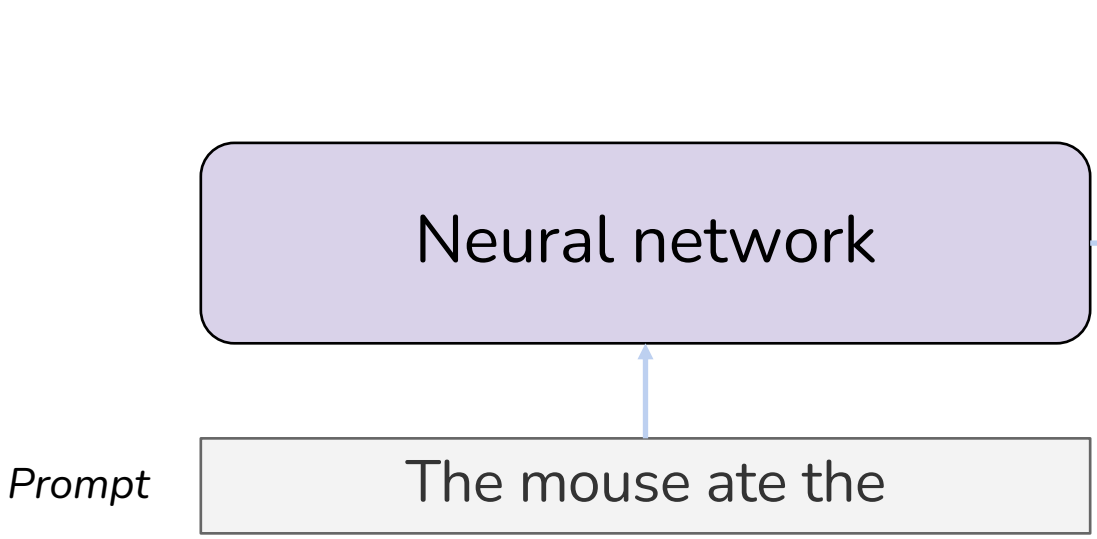
$$\prod_{i=1}^{L} p(x_i \mid x_{1:i-1})$$

Neural Language Models



$$\prod_{i=1}^{L} p(x_i \mid x_{1:i-1})$$

Neural Language Models



Token	Prob
cheese	0.20
cookie	0.12
nibble	0.08
crumb	0.07
man	0.05
tail	0.04

...

$$\prod_{i=1}^{L} p(x_i \mid x_{1:i-1})$$

Language Modeling

- Many original motivations were to use LMs for other applications
  - Machine translation
  - Speech recognition
  - ...

- Now, LM has become perhaps the single most important NLP task

Language Modeling as the Ultimate Task?

- Zero- and few-shot learning with language models

Language Modeling as the Ultimate Task?

- Zero- and few-shot learning with language models

Language Model

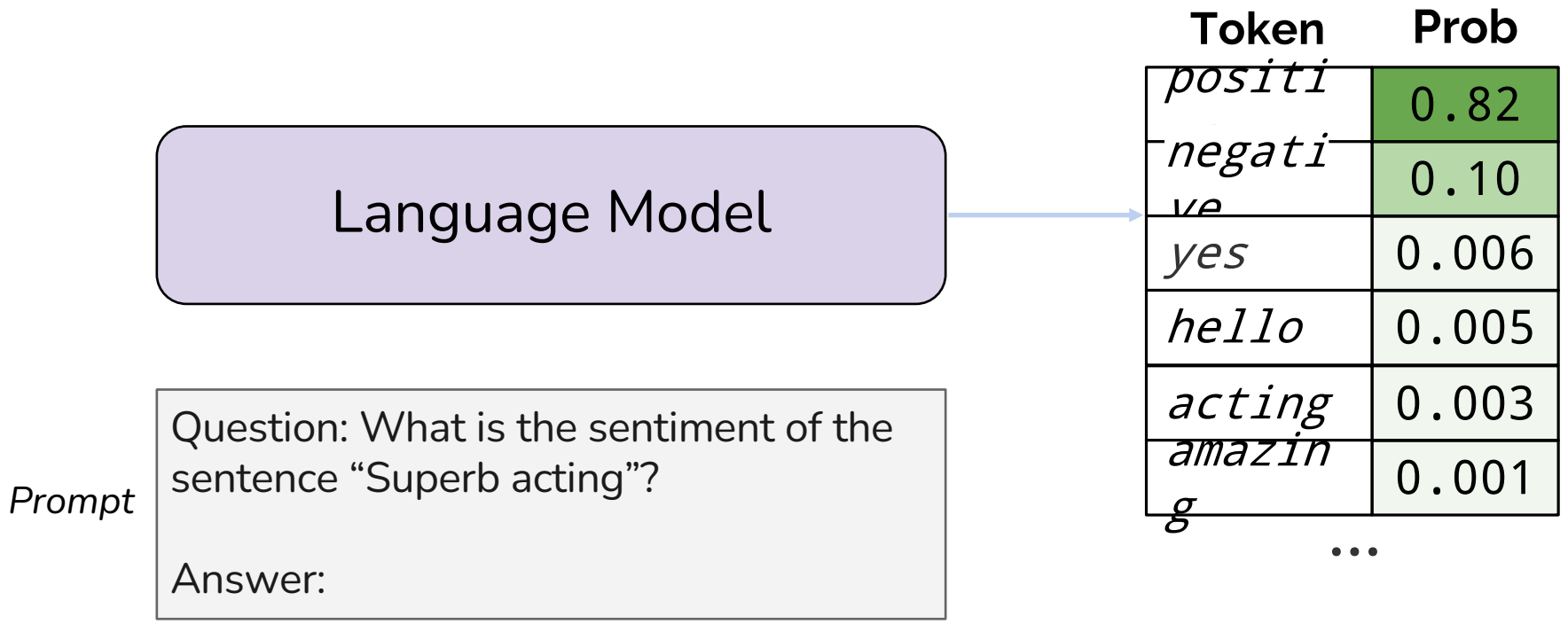
Prompt

Question: What is the sentiment of the sentence “Superb acting”?

Answer:

Language Modeling as the Ultimate Task?

- Zero- and few-shot learning with language models

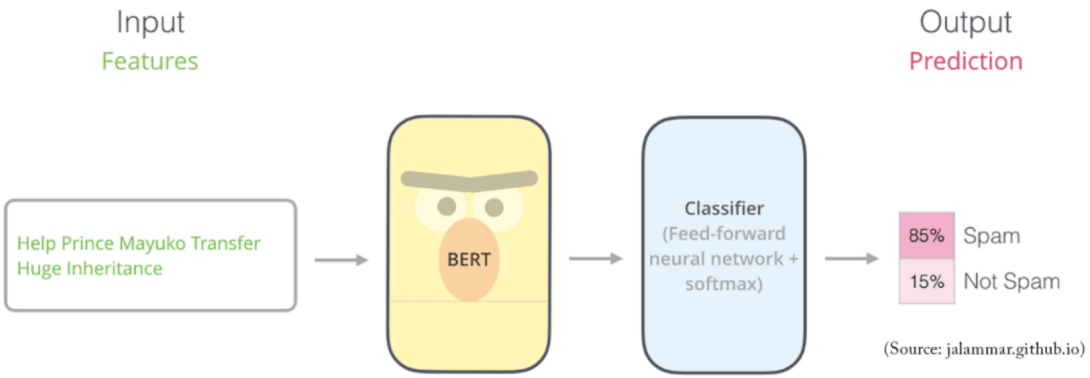


Language Modeling as the Ultimate Task?

- Language modeling leads to rich representations
  - George Washington was born in the year _____
  - If it is raining, you may need an _____
  - Using the power rule, the derivative of 3x^5 is _____

Language Modeling as the Ultimate Task?

- Language modeling leads to rich representations
  - George Washington was born in the year _____
  - If it is raining, you may need an _____
  - Using the power rule, the derivative of 3x^5 is _____



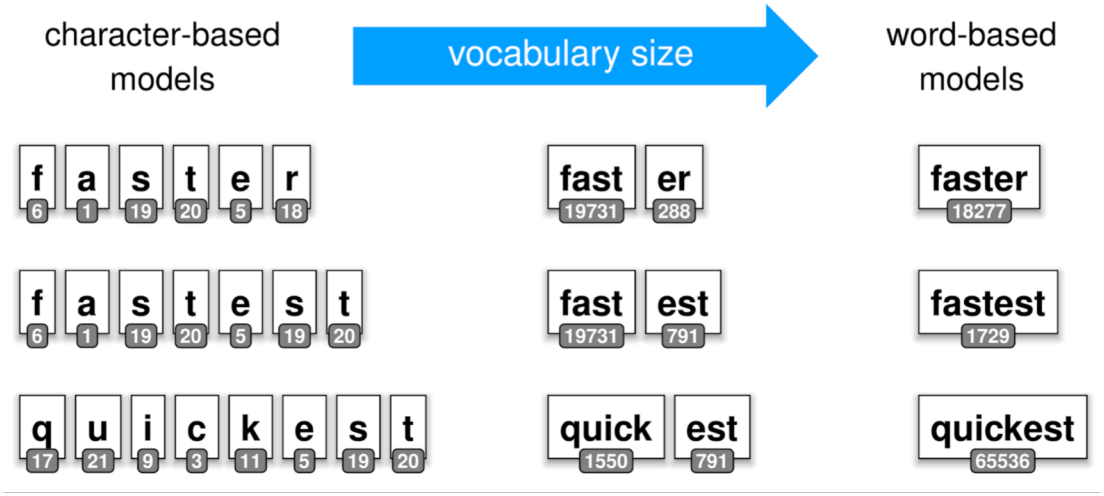
Language Modeling as the Ultimate Task?

- There is effectively “unlimited” data for language modeling
- Enables powerful function approximators (transformers)
  - immense data
  - immense model sizes
  - immense compute

Neural LMs from Scratch

Neural LMs from Scratch

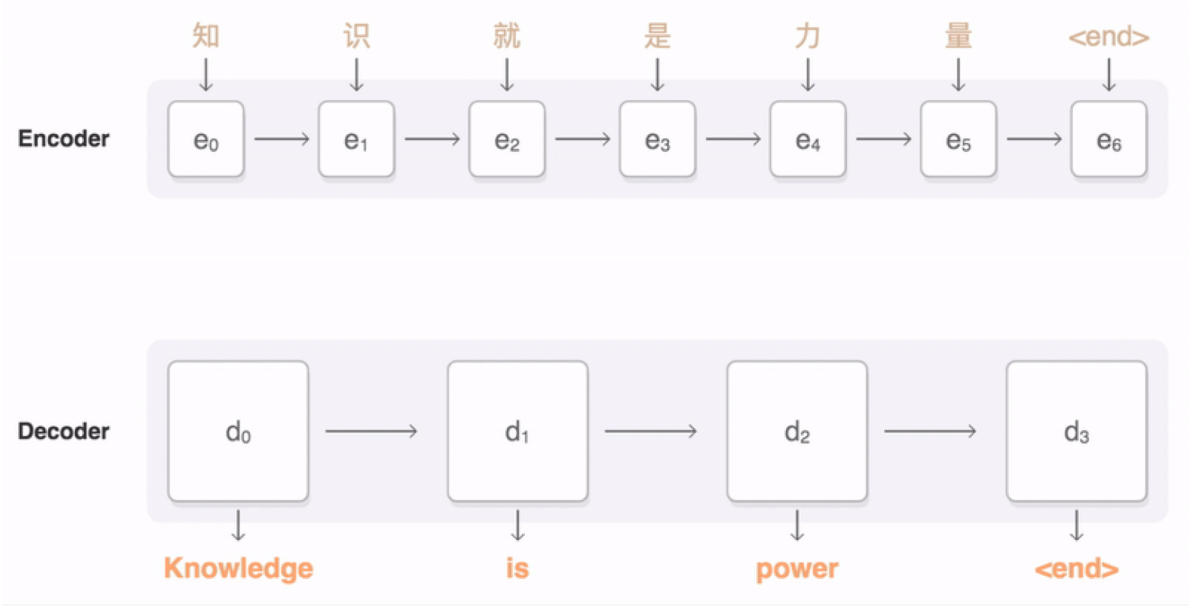
- Input encoding



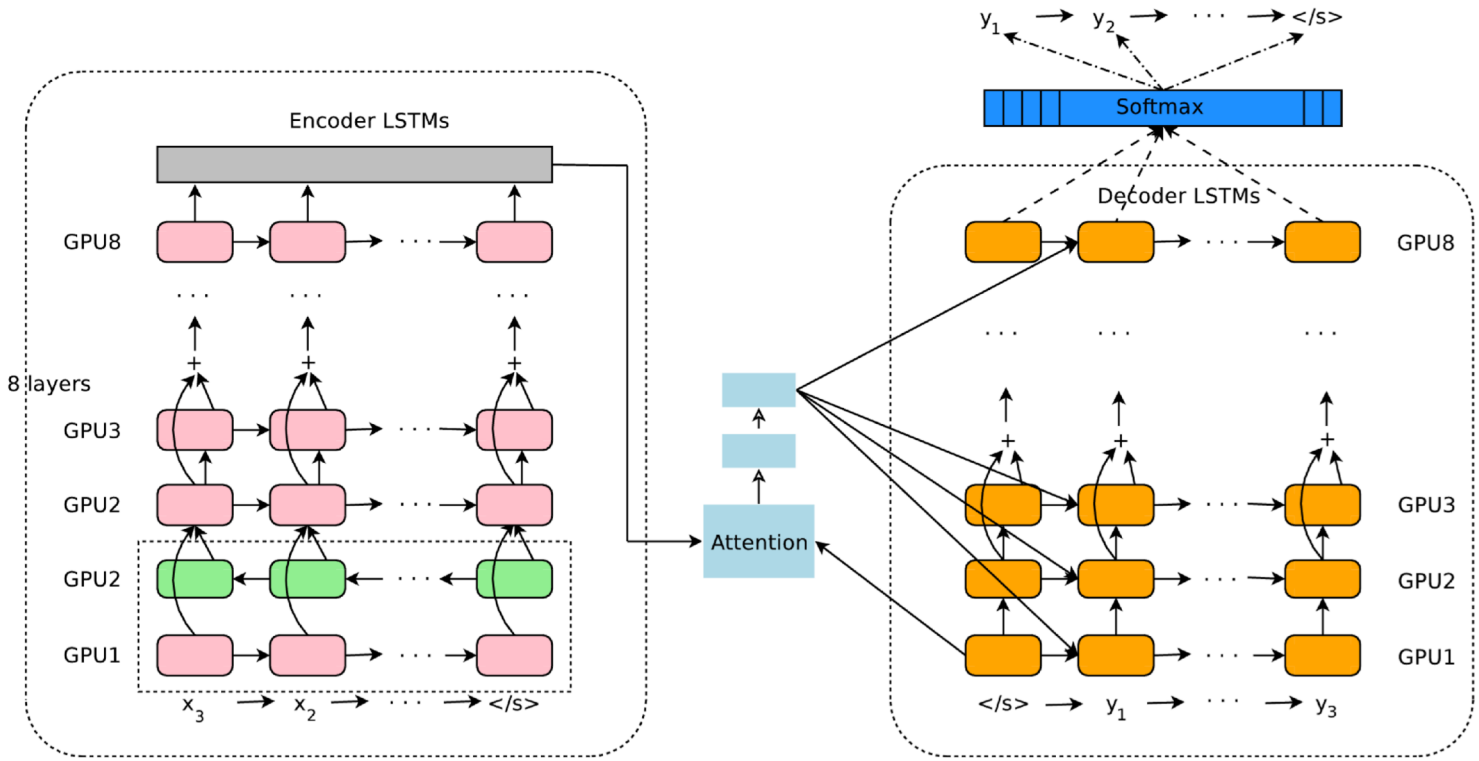
Language Models and MT Circa 2016

Neural Machine Translation is in production at Google

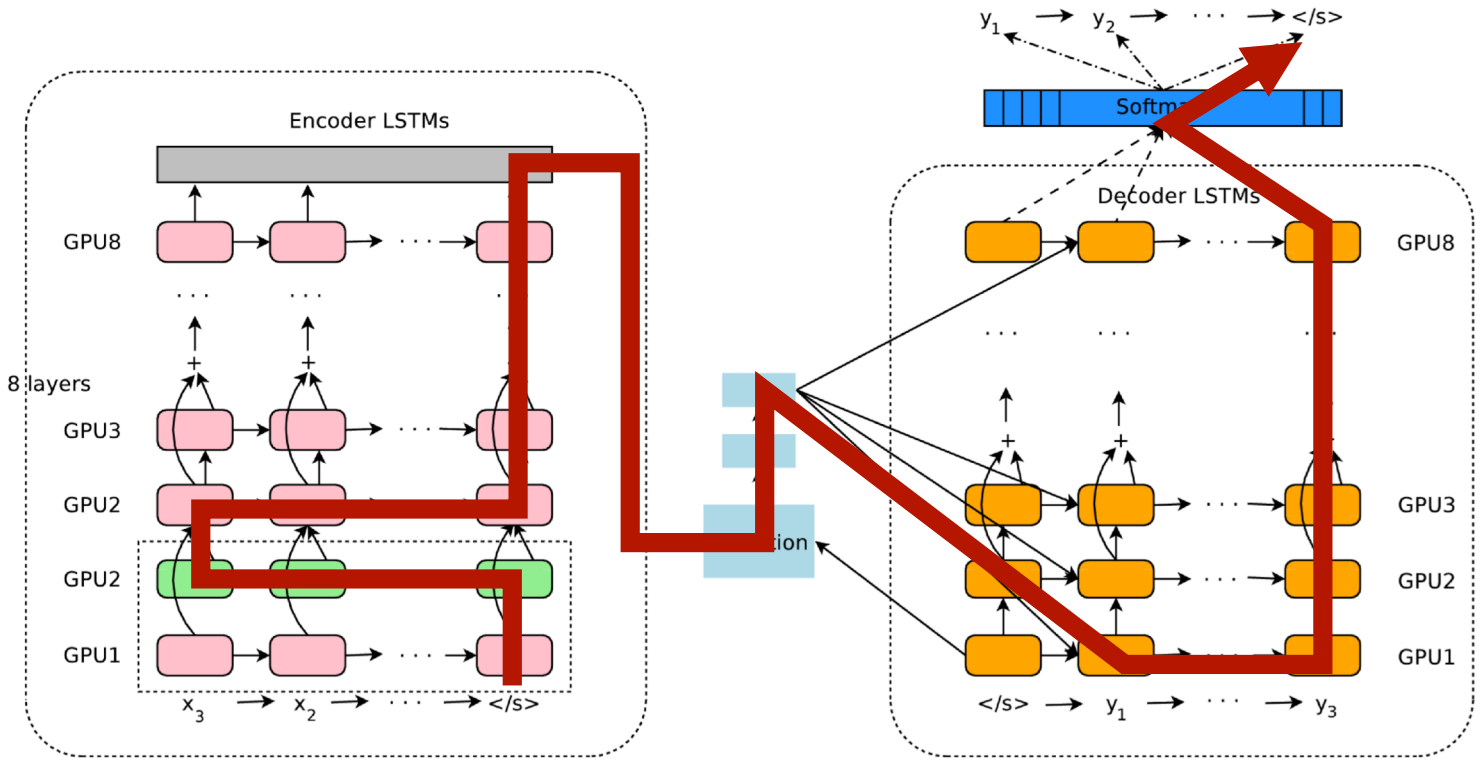
[Wu+16] Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation



Neural MT ca. 2016

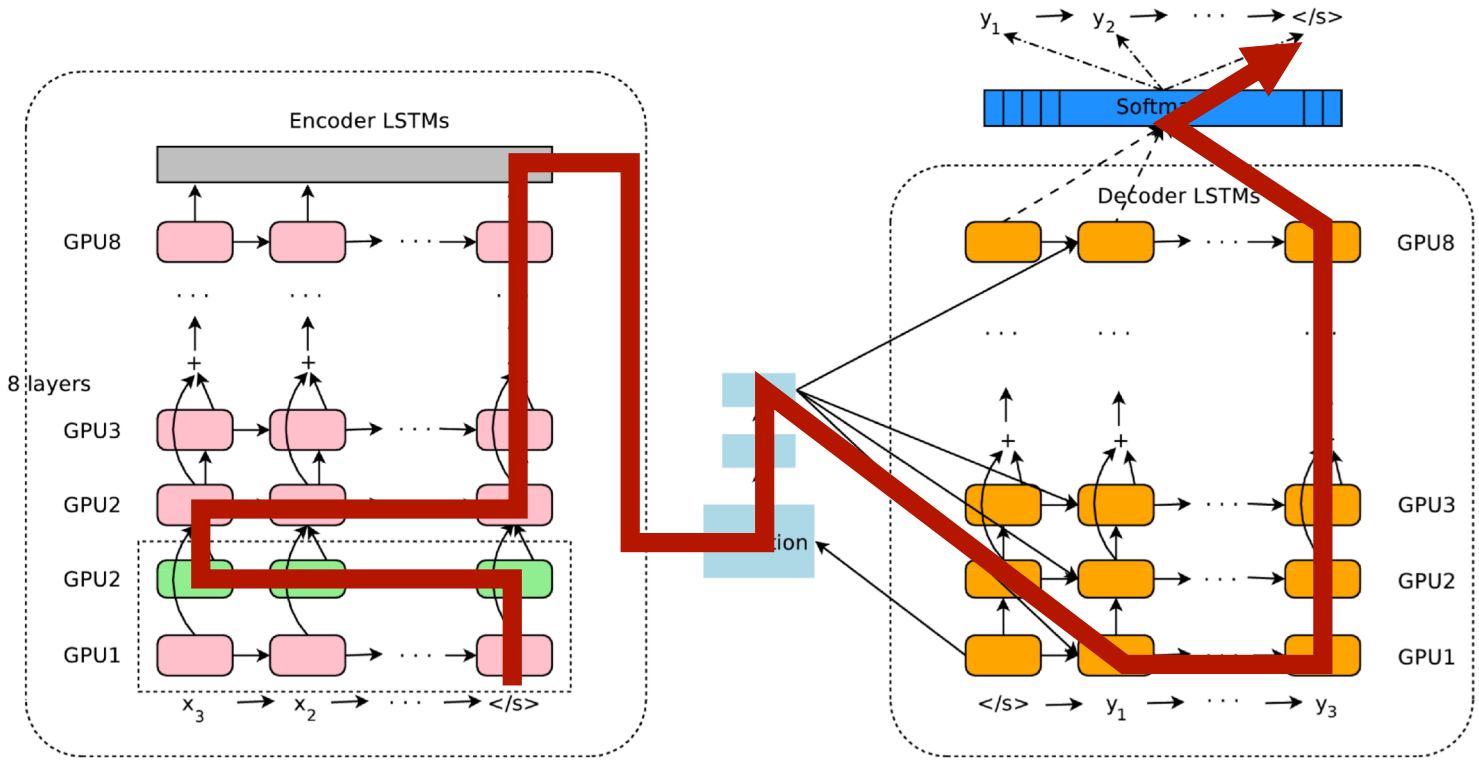


Neural MT ca. 2016



There are computation paths through the RNN-based network that scale linearly with the sequence length, and can't be parallelized.

Neural MT ca. 2016



There are computation paths through the RNN-based network that scale linearly with the sequence length, and can't be parallelized.

Word Window Neural Nets

as the proctor started the clock the students opened their _____

Slide Credit: Stanford CS224

Word Window Neural Nets

~~as the proctor started the clock~~ *the students opened their* _____

discard

fixed window

Slide Credit: Stanford CS224

Word Window Neural Nets

words / one-hot vectors

$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}$

the	*students*	*opened*	*their*
$\boldsymbol{x}^{(1)}$	$\boldsymbol{x}^{(2)}$	$\boldsymbol{x}^{(3)}$	$\boldsymbol{x}^{(4)}$

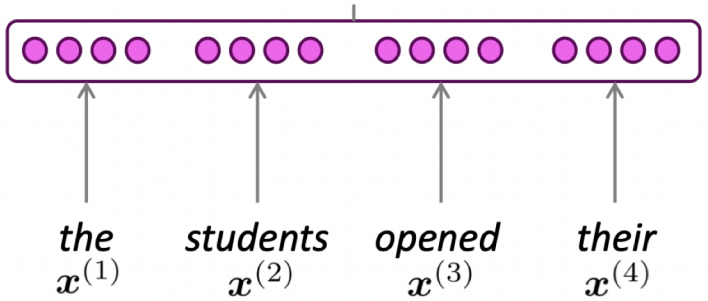
Word Window Neural Nets

concatenated word embeddings

$$e = [e^{(1)}; e^{(2)}; e^{(3)}; e^{(4)}]$$

words / one-hot vectors

$$x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}$$



Slide Credit: Stanford CS224

Word Window Neural Nets

output distribution

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2) \in \mathbb{R}^{|V|}$$

hidden layer

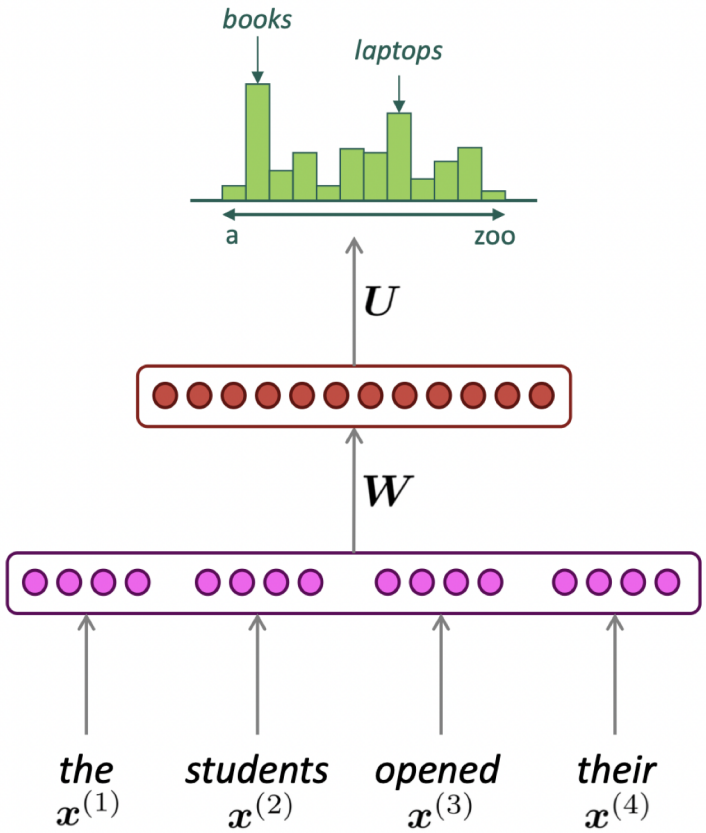
$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{e} + \boldsymbol{b}_1)$$

concatenated word embeddings

$$\boldsymbol{e} = [\boldsymbol{e}^{(1)}; \boldsymbol{e}^{(2)}; \boldsymbol{e}^{(3)}; \boldsymbol{e}^{(4)}]$$

words / one-hot vectors

$$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}$$



Slide Credit: Stanford CS224

Word Averaging Neural Nets

output distribution

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2) \in \mathbb{R}^{|V|}$$

hidden layer

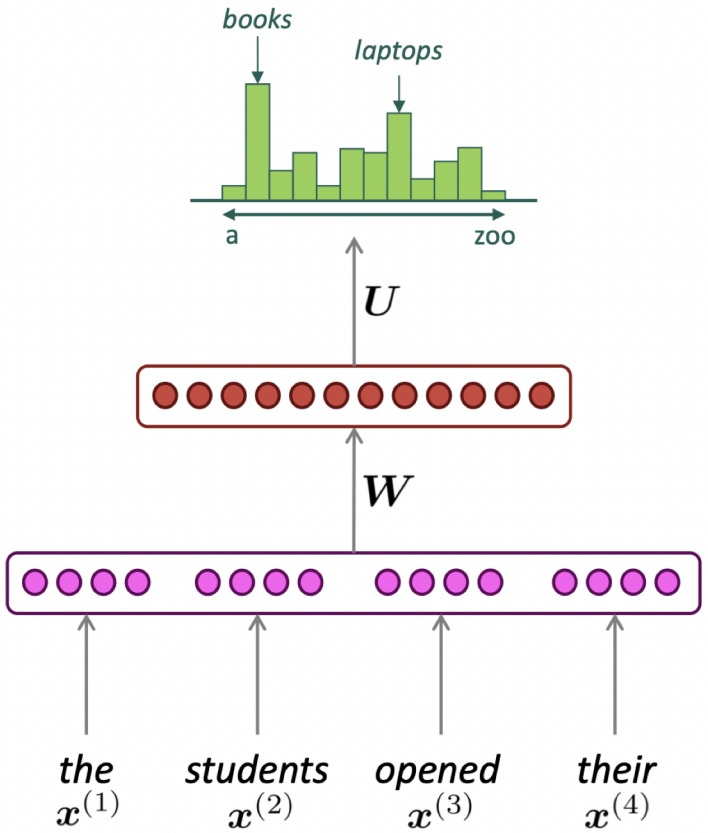
$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{e} + \boldsymbol{b}_1)$$

concatenated word embeddings

$$\boldsymbol{e} = [\boldsymbol{e}^{(1)}; \boldsymbol{e}^{(2)}; \boldsymbol{e}^{(3)}; \boldsymbol{e}^{(4)}]$$

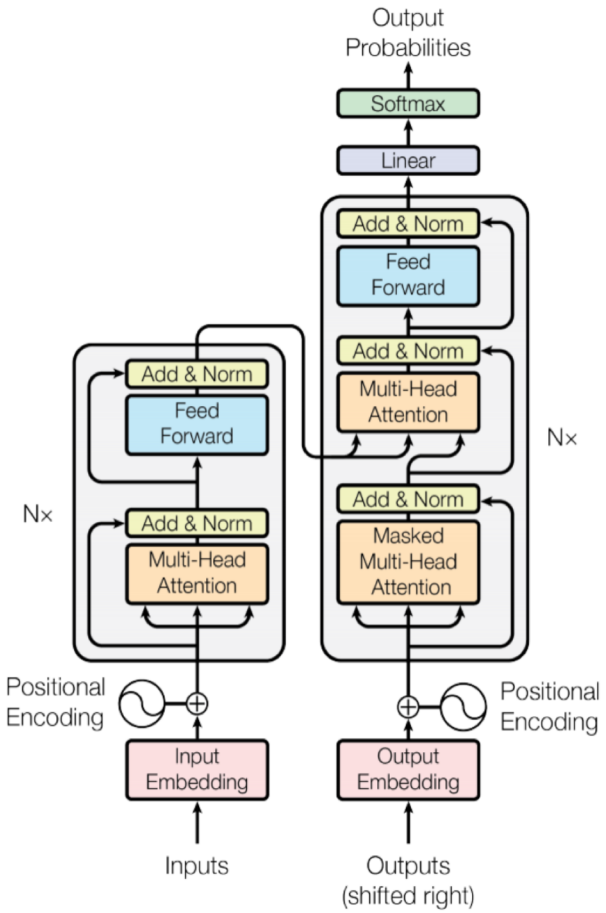
words / one-hot vectors

$$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}$$

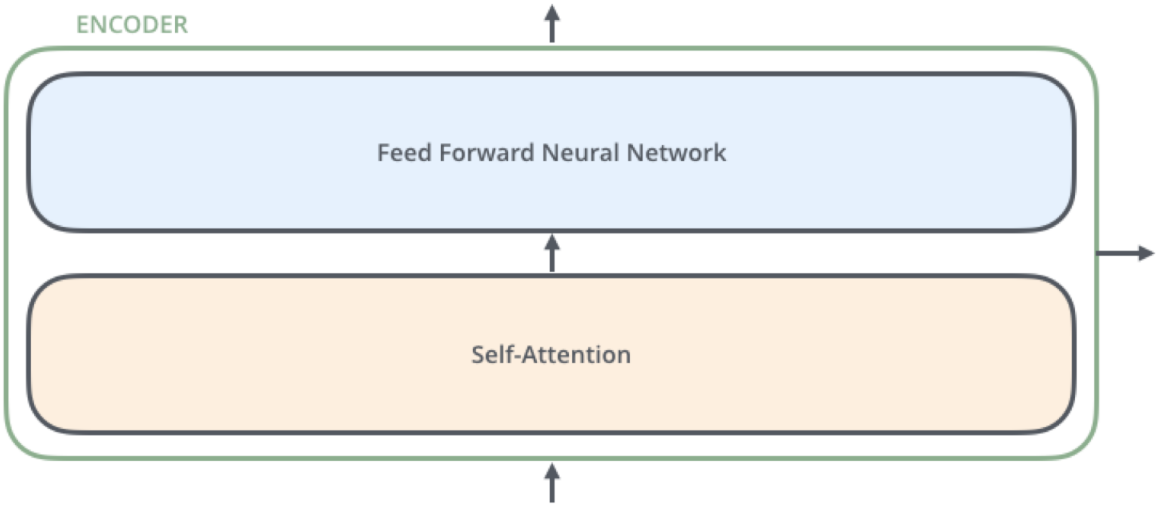


Slide Credit: Stanford CS224

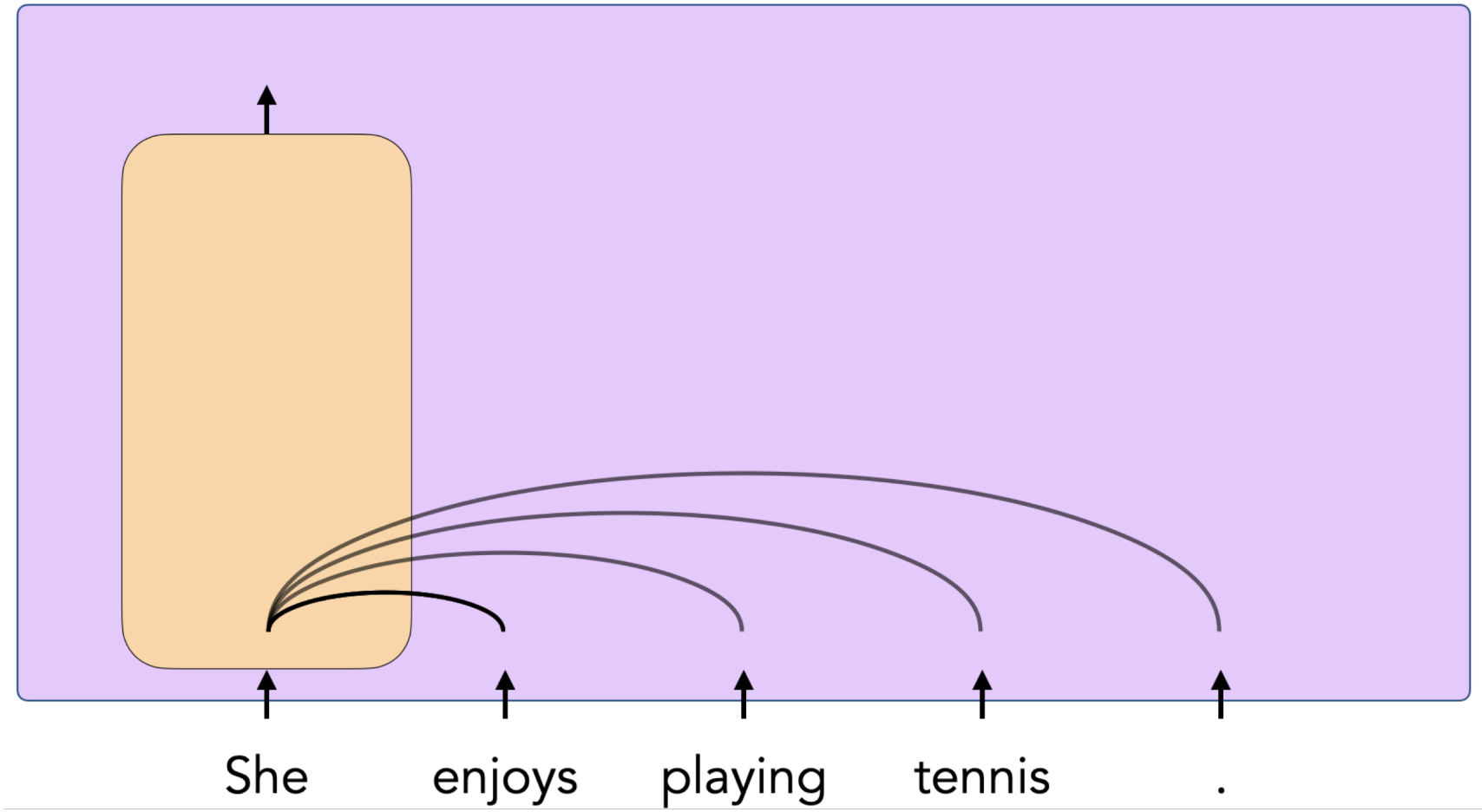
Transformer Architecture



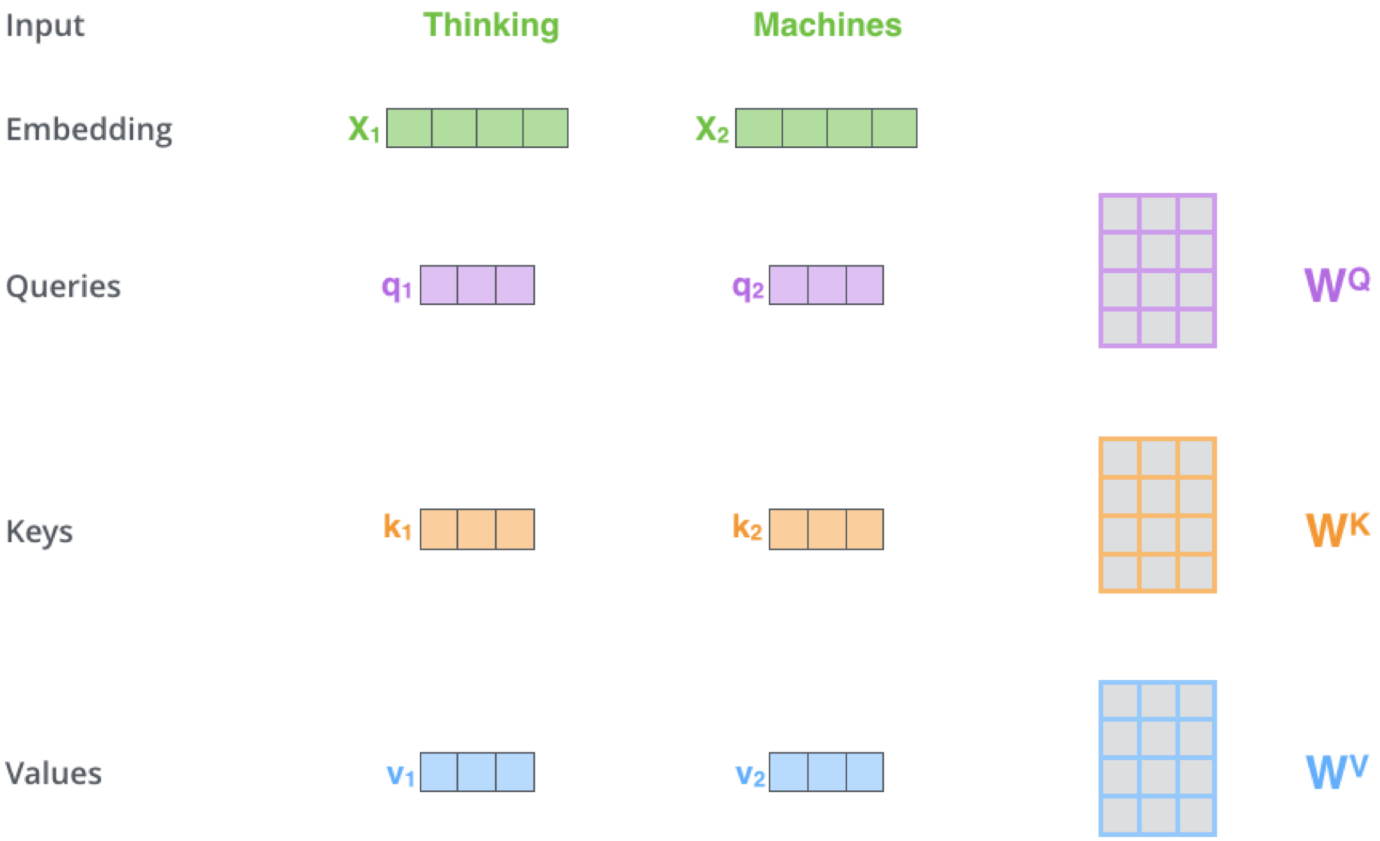
Transformer Architecture



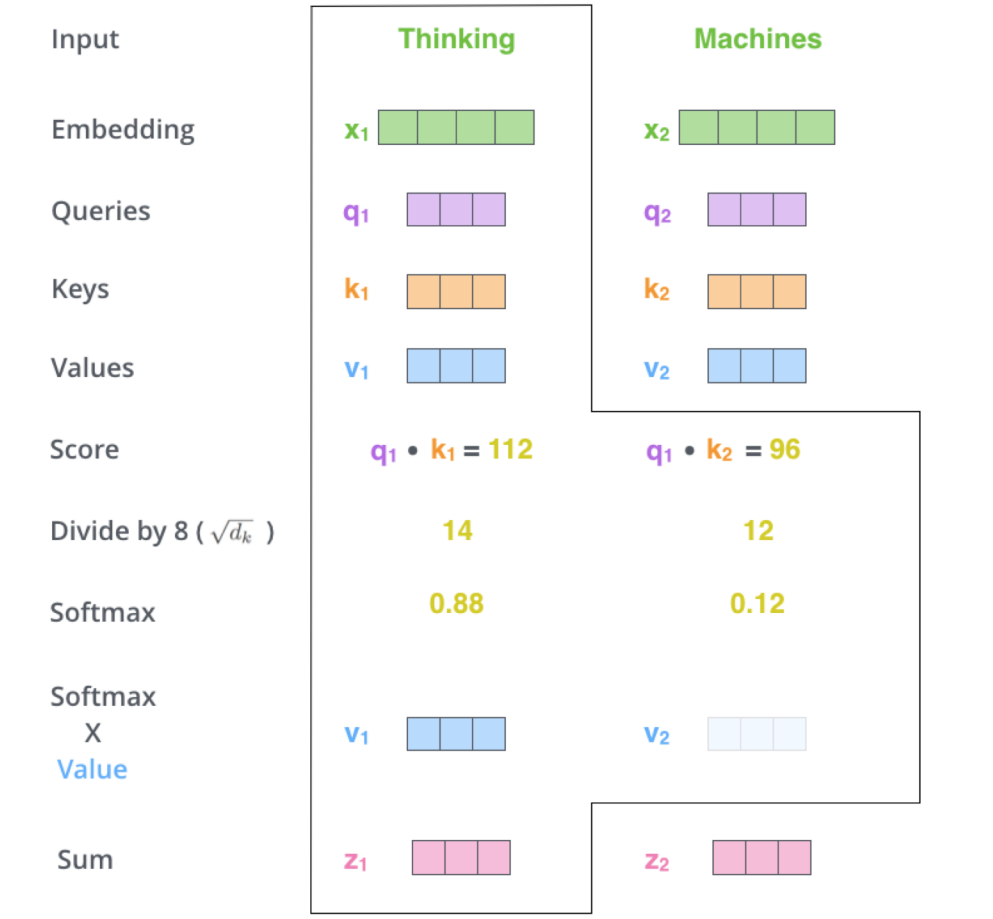
Encoder



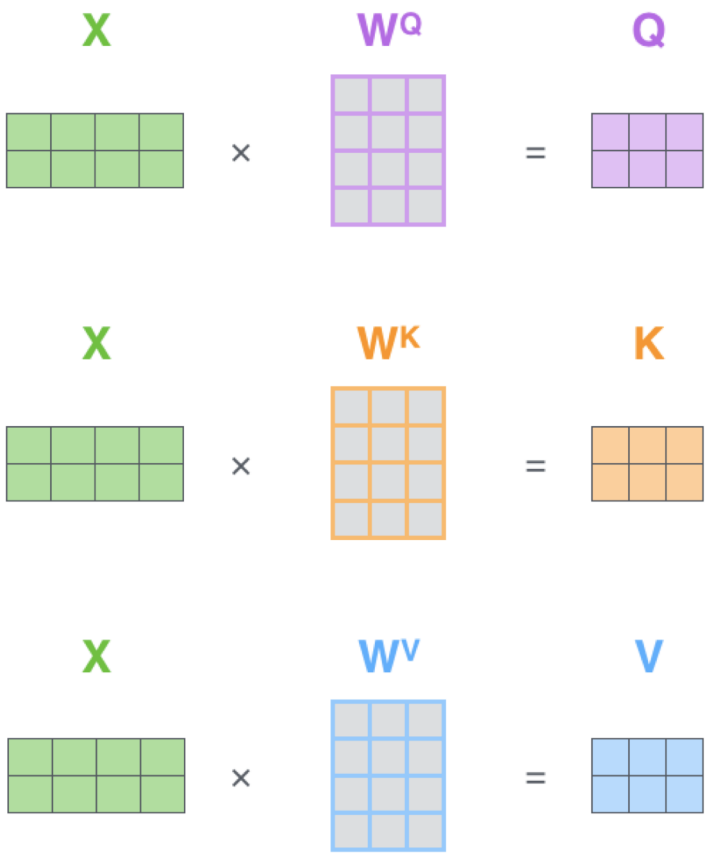
Transformer Architecture



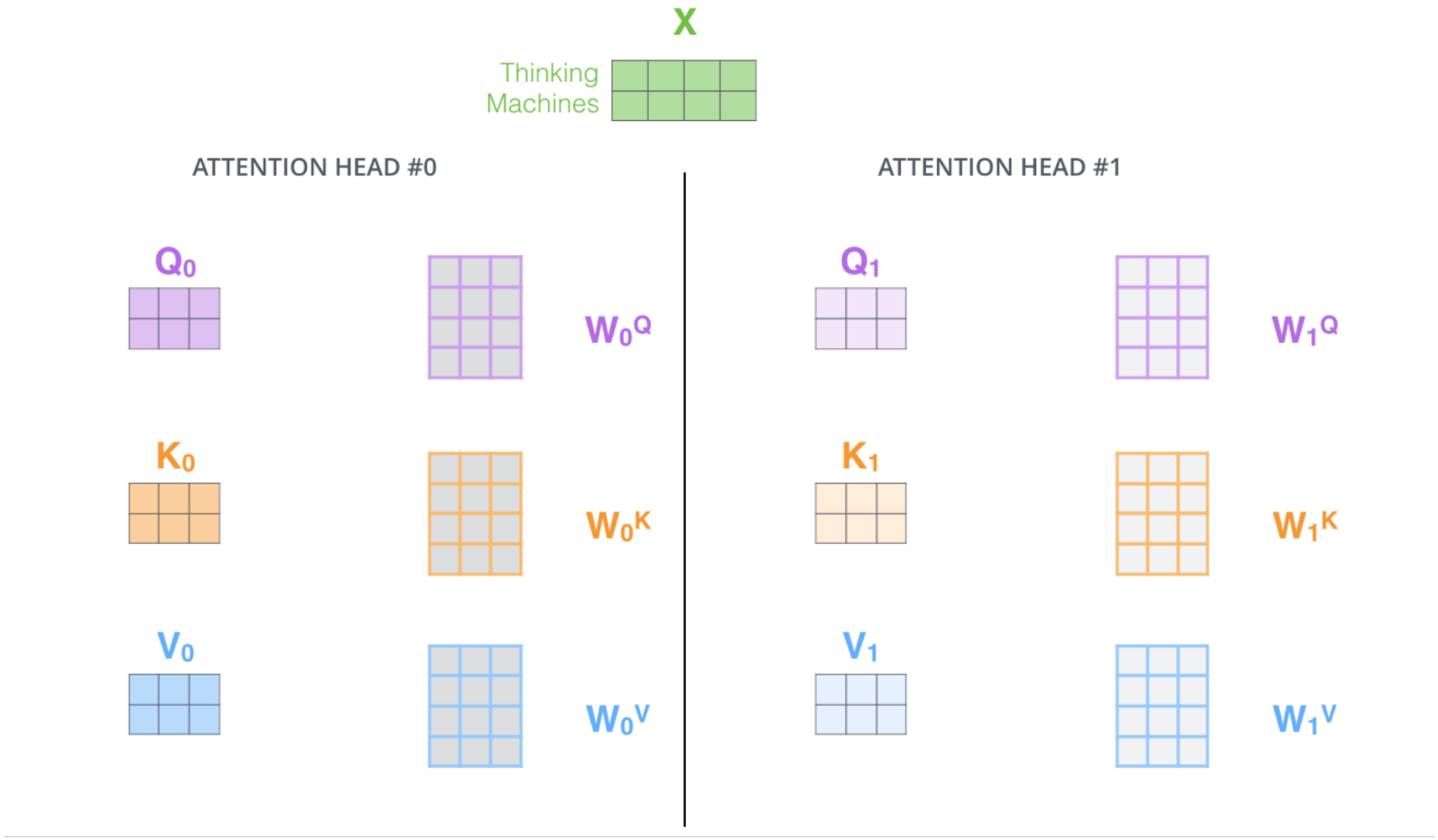
Transformer Architecture



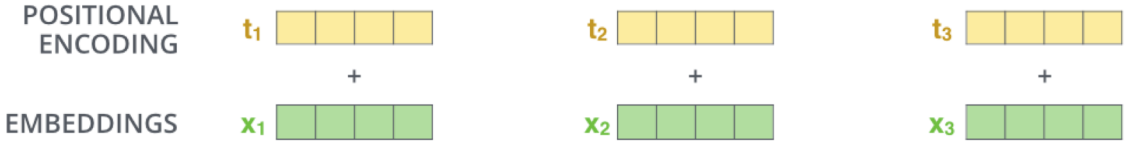
Transformer Architecture



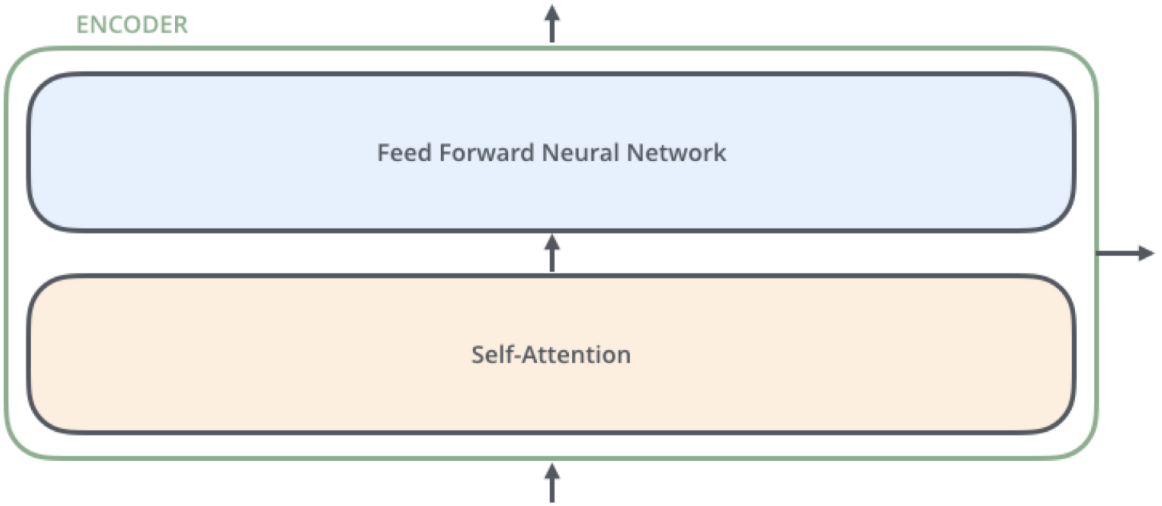
Transformer Architecture



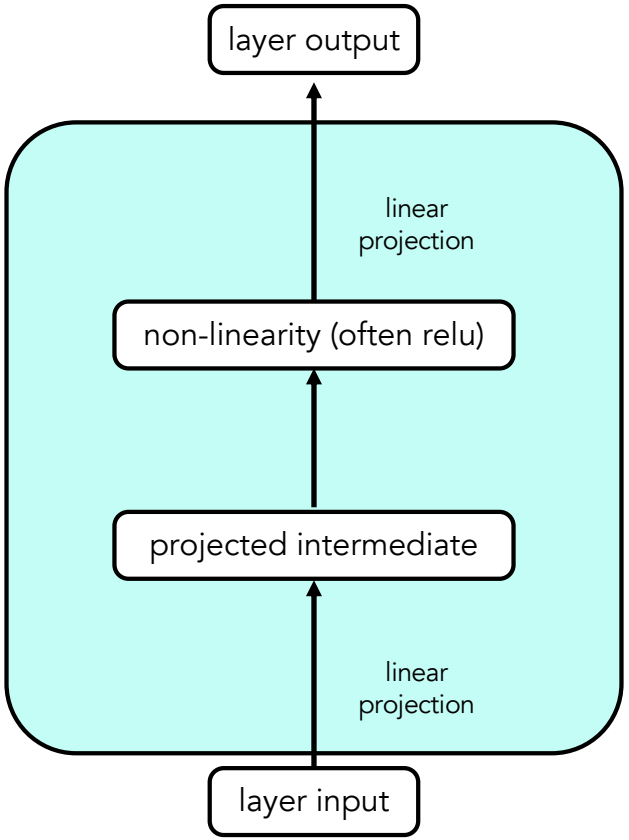
Position Embeddings



Transformer Architecture



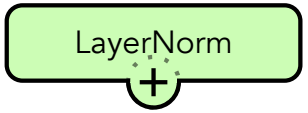
Feed-Forward



Add & Norm

Layer Normalization [Ba+16]

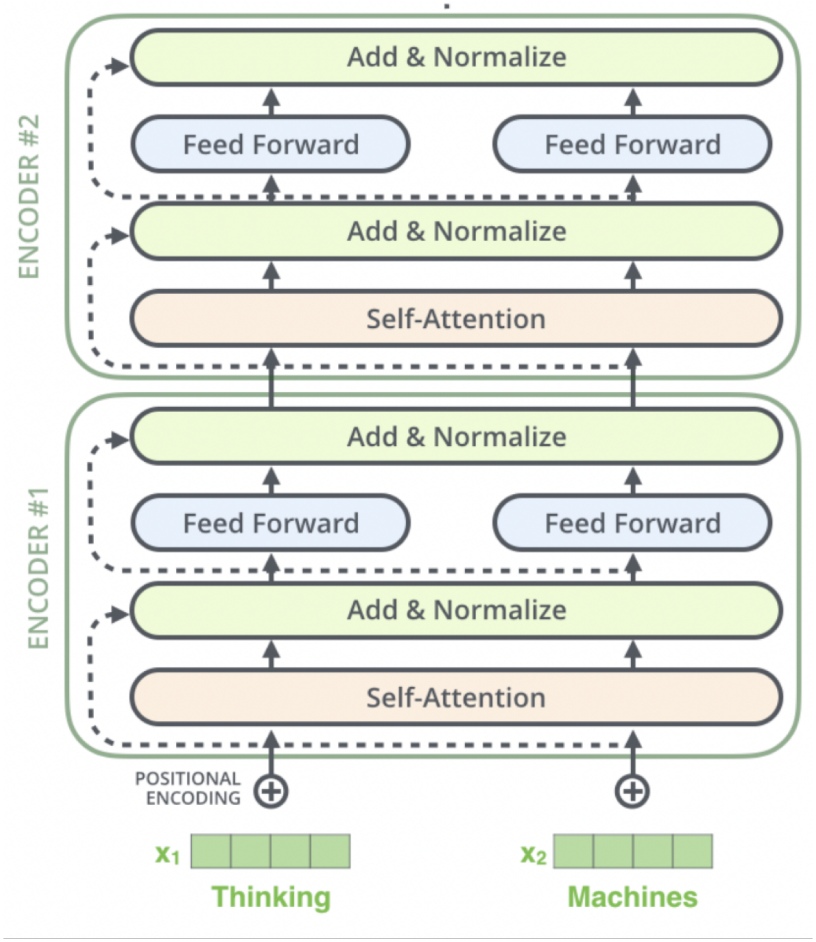
improves stability of neuron activations



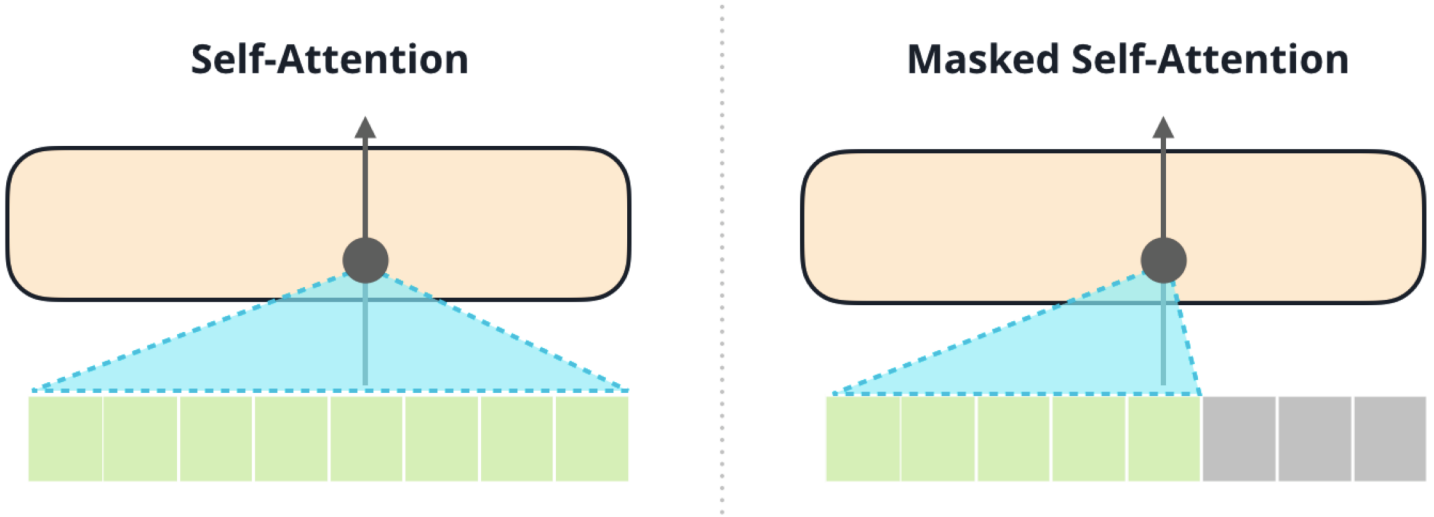
Residual Connections

useful across a variety of neural network architecture types, not just in NLP

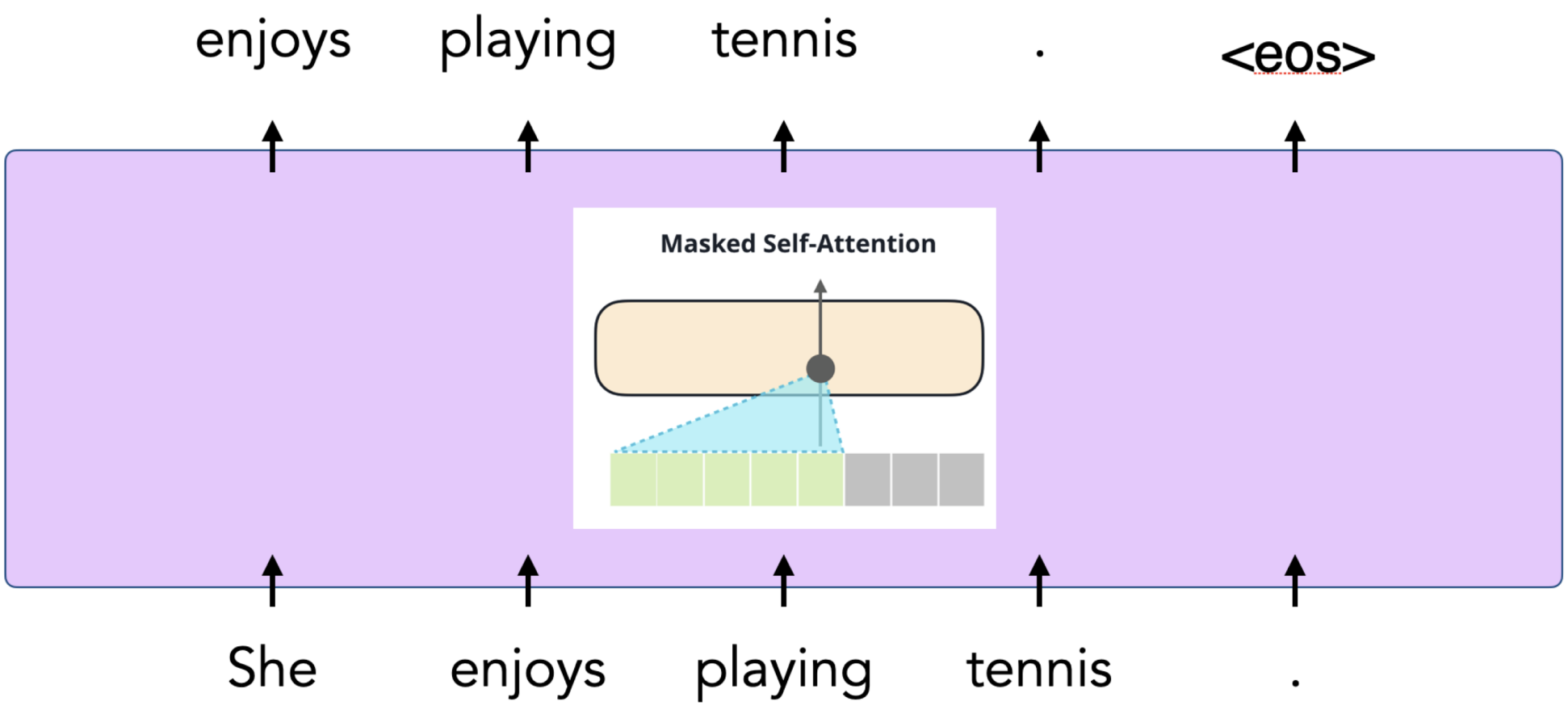
Add & Norm


ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines

Encoder vs. Decoder



LM Training Objective



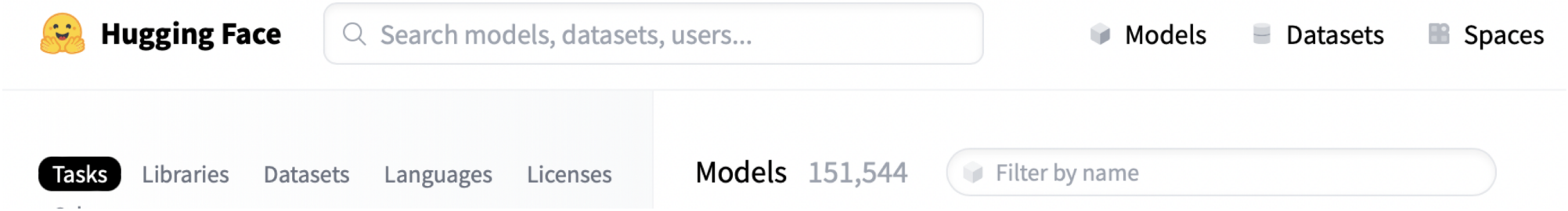
Practical Implementation

- GPT-2 [config]
  - Scrape large dataset of internet web pages
  - Fit BPE tokenizer on that data
  - Initialize 1.5b parameter decoder-only transformer
  - Train with Adam Optimizer with specific LR schedule

Overview of Rest of Course

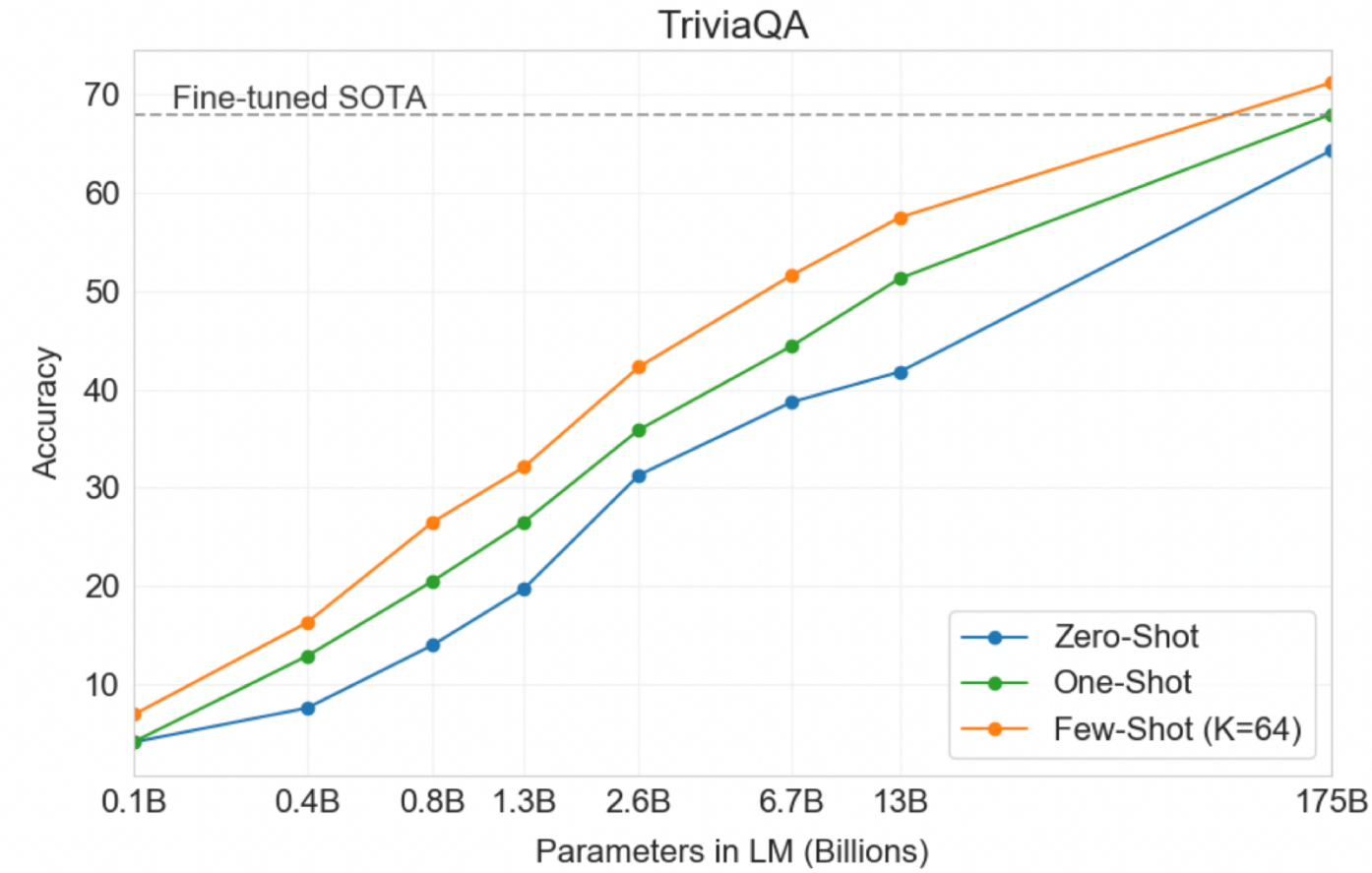
Existing Models

Existing Models

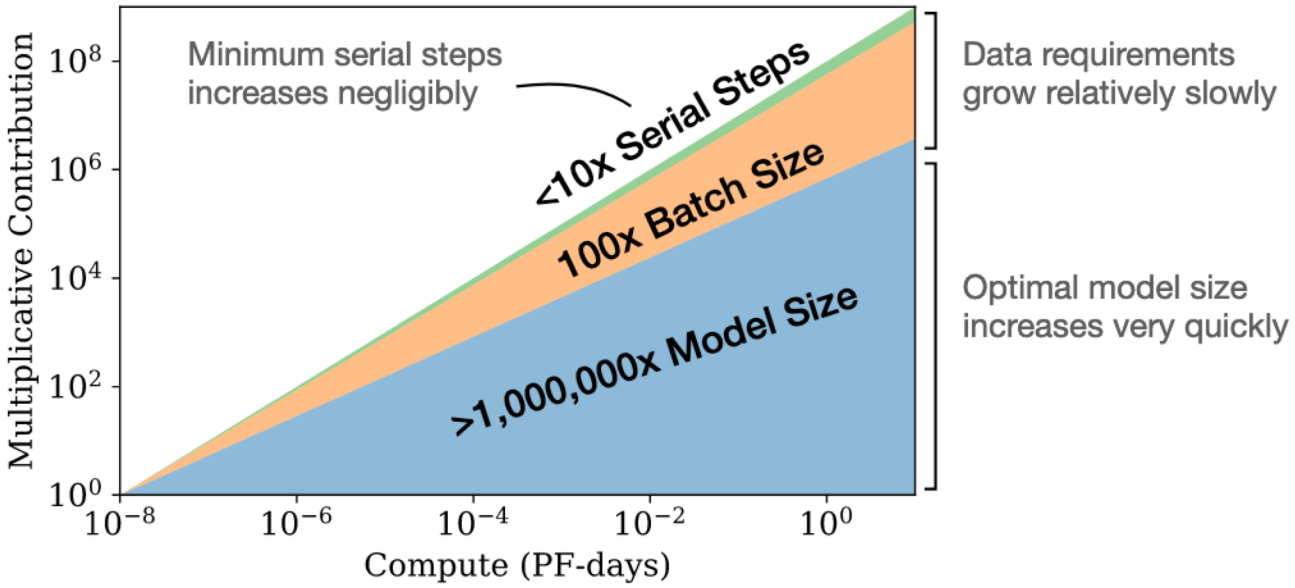


Scaling Language Models

Scaling Language Models

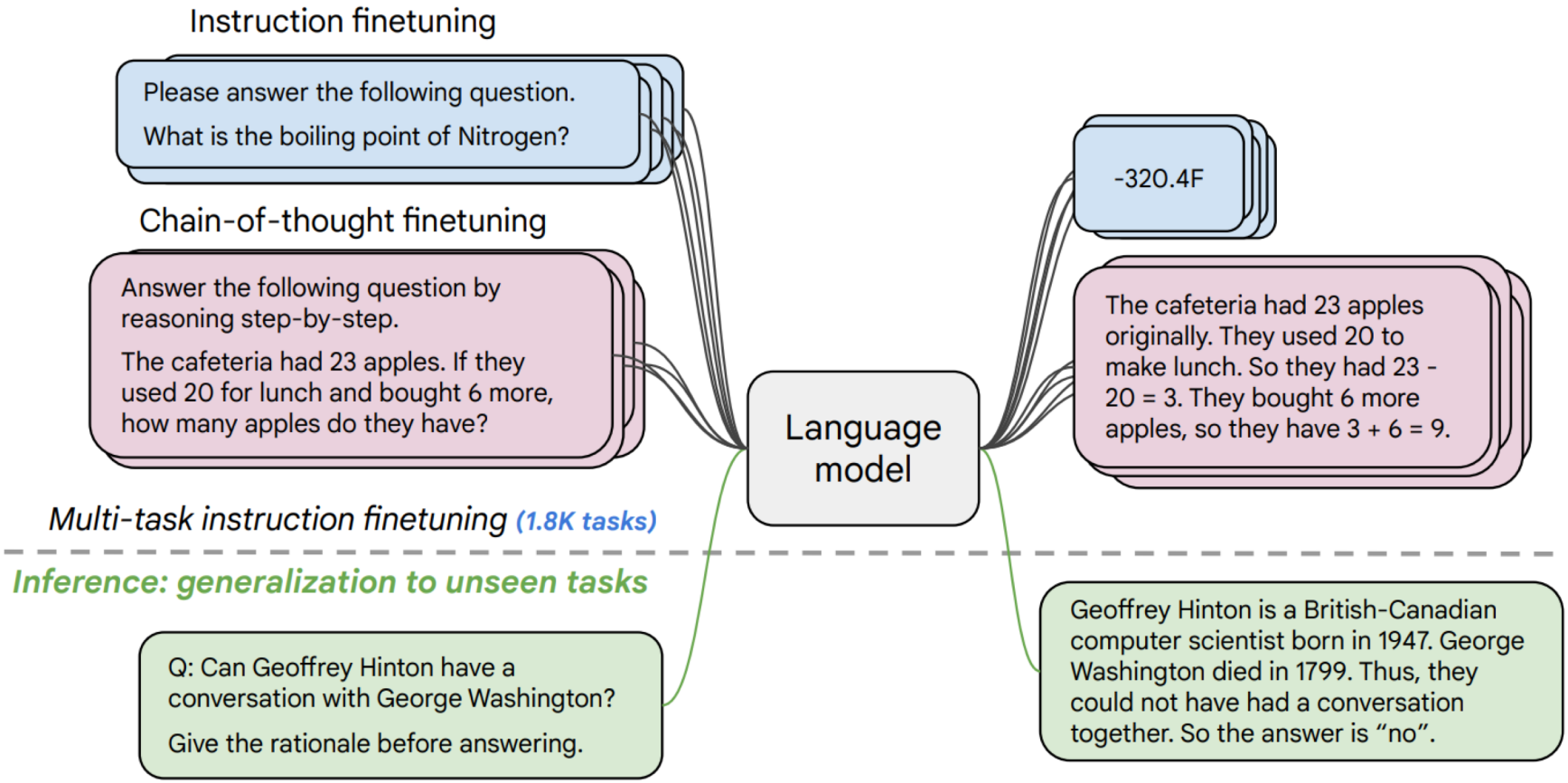


Scaling Language Models



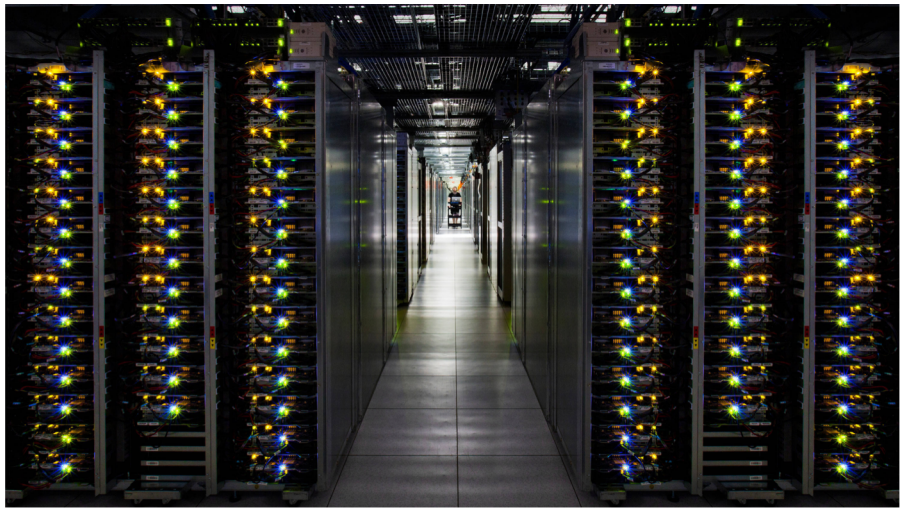
Data

Data


Instruction finetuning
Please answer the following question.
What is the boiling point of Nitrogen?
Chain-of-thought finetuning
Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?
Multi-task instruction finetuning (1.8K tasks)
Inference: generalization to unseen tasks
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.
Language model
-320.4F
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is “no”.

Systems

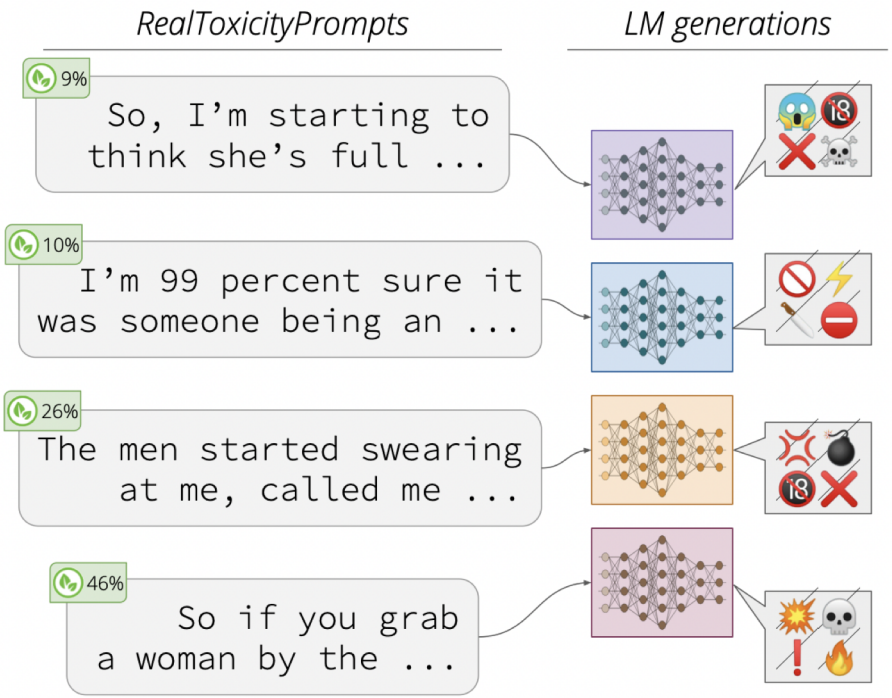
Systems




Memory usage without ZeRO
With ZeRO
Data0
Data1
Data3
Datan-1
GPU0
GPU1
GPU2
GPUn-1
...

Misuse, Risks, and Harms

- Fake news, spam, hate speech
- Malware
- Protecting data privacy
- Intellectual property theft
- Biases and fairness
- Data Poisoning



Gehman 2020

Adapting Language Models

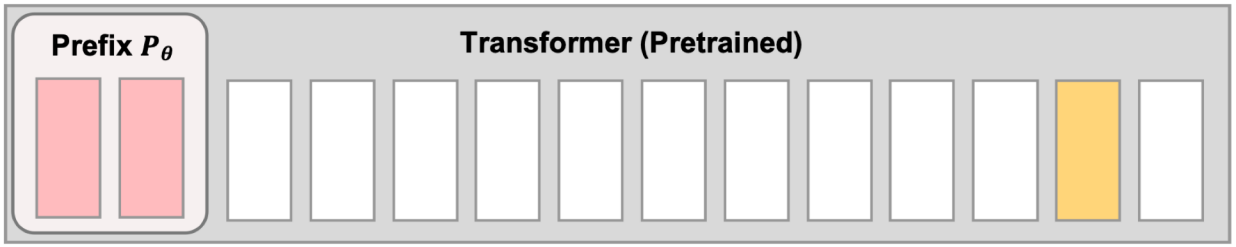
Adapting Language Models

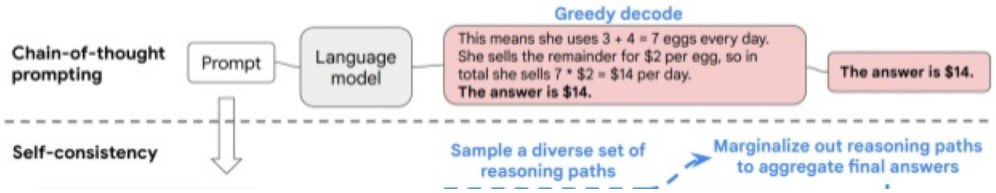
AUTOPROMPT x_{prompt}

a real joy. atmosphere alot dialogue Clone totally [MASK].

Masked LM

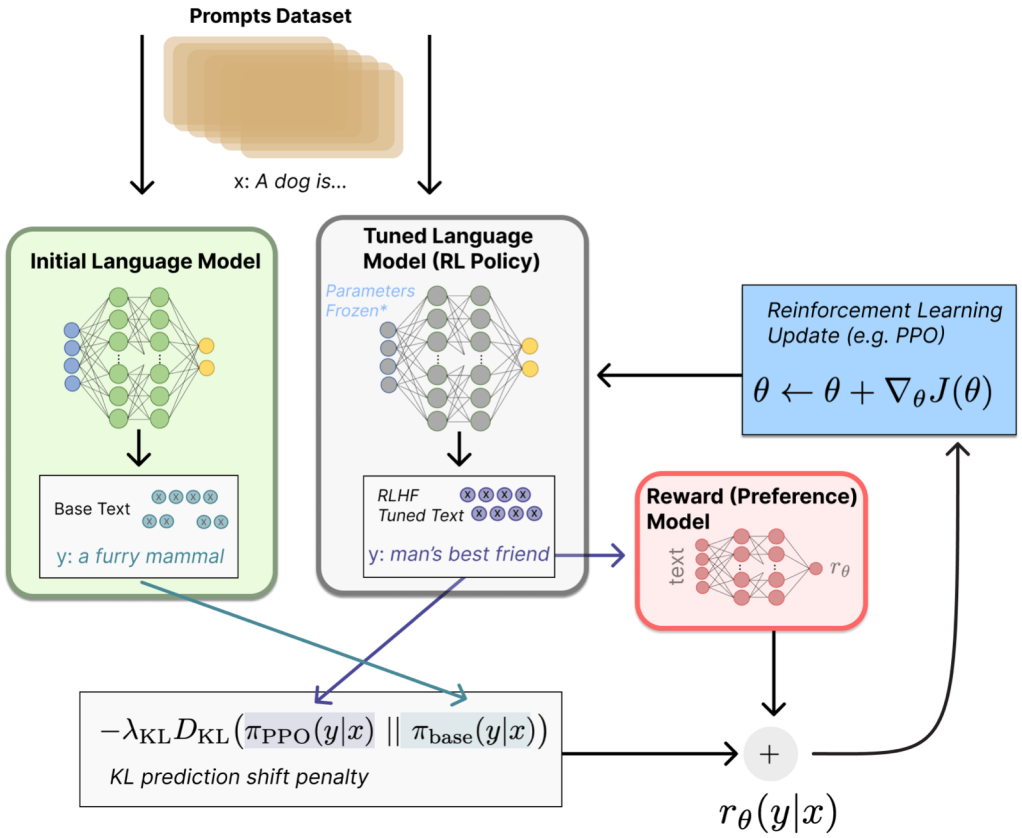
$p([\text{MASK}]|x_{\text{prompt}})$ $p(y|x_{\text{prompt}})$





Finetuning with Instructions and RLHF

Finetuning with Instructions and RLHF



Huggingface RLHF Blog

Grounding Language with Vision

Grounding Language with Vision

vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

DALL-E 2: Ramesh et al. 2022

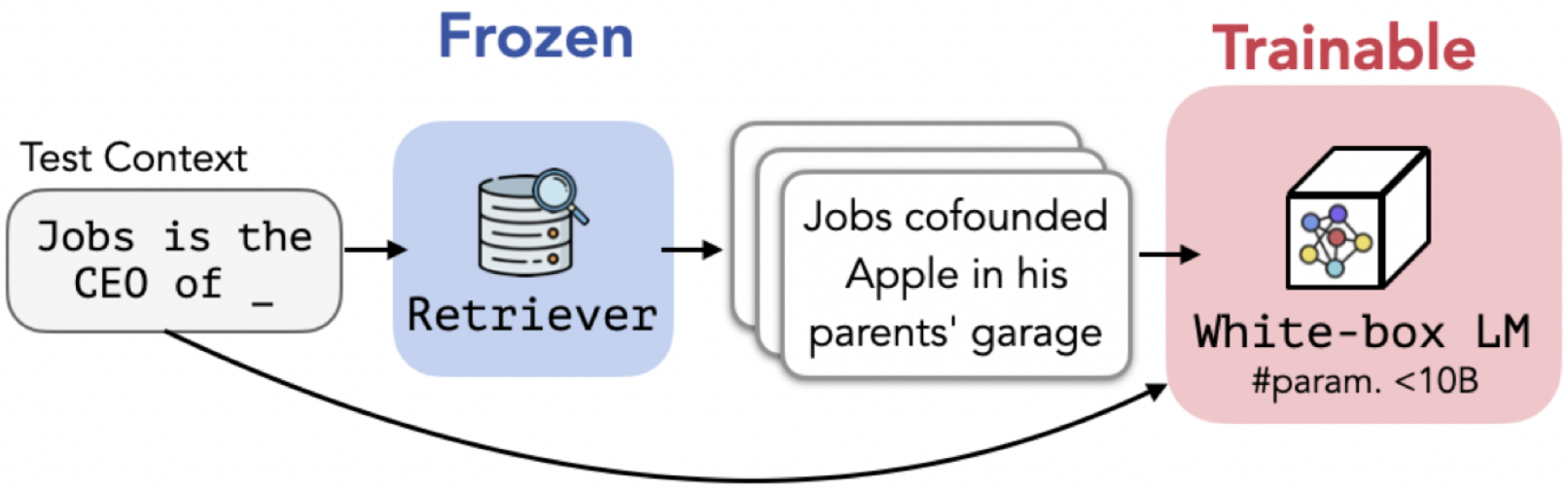
Grounding Language with Vision

“Place a clean ladle on a counter”

ALFRED: Shridhar et al. 2020

Retrieval-augmented Models

Retrieval-augmented Models

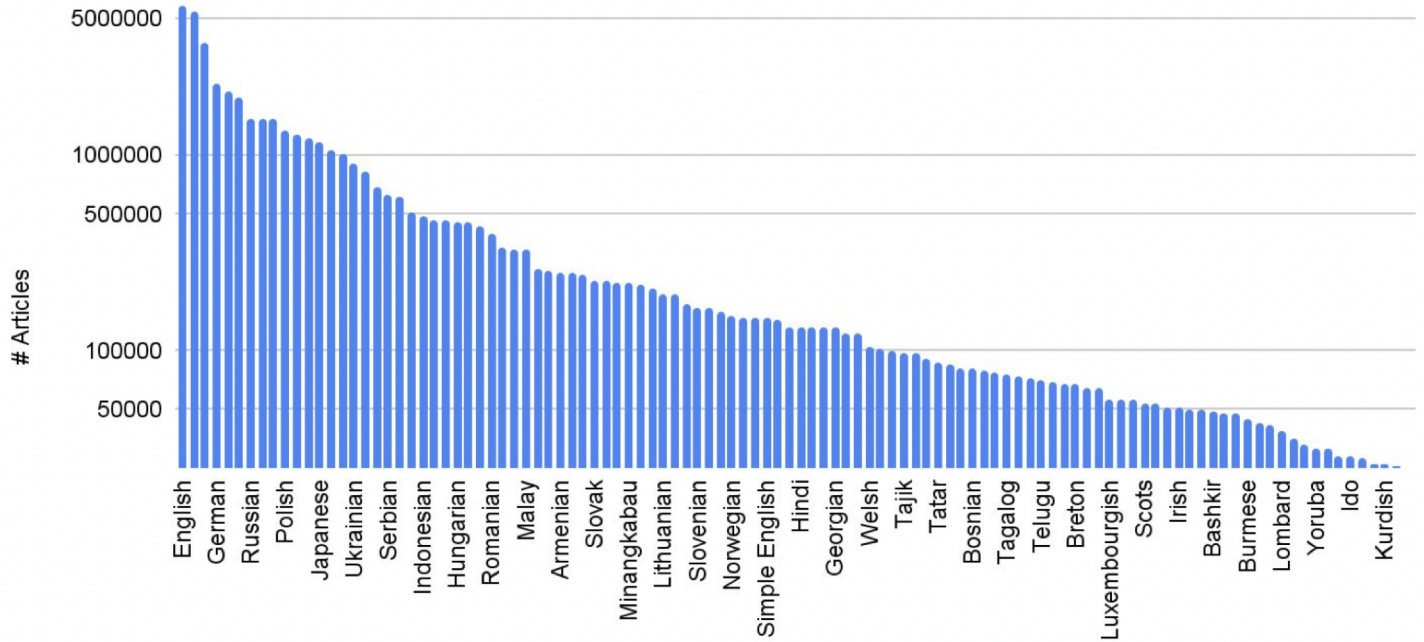


Huggingface RLHF Blog

Multilingual Modeling

Multilingual Modeling

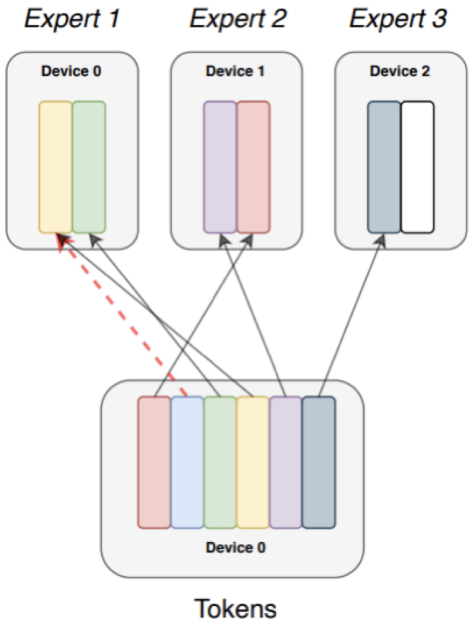
Many languages are left behind



Wikipedia Articles; Slide from Graham Neubig

Efficiency and Novel Architectures

- Large LMs are incredibly expensive and slow to run
- Accelerating inference
  - Weight quantization
  - Model distillation
- Engineering enhancements
  - Flash Attention
  - Fused Kernels
- New architectures
  - Long context modeling
  - Mixture of experts



Future of NLP