

Large Language Models: What will happen next?

坂口慶祐 (東北大学)

keisuke.sakaguchi@tohoku.ac.jp

略歴

専門・興味：自然言語処理、機械学習、心理言語学、言語哲学

2022～ 東北大学, 准教授

2018～2022 Allen Institute for AI (AI2), Research Scientist

2013～2018 Johns Hopkins Univ. Ph.D. in Computer Science

2011～2013 奈良先端科学技術大学院大学

...

紆余曲折 (ITインフラエンジニア, パートタイム研究員, 脳科学・心理言語学)

...

2001年～2005年 早稲田大学文学部 (言語哲学)

Tohoku NLP group

Faculty: 8

Postdoc researcher: 10, undergrad+grad students: 50+



Things are moving fast!



LangChain

Chat GPT



Perplexity

2022/09

2022/10~11

2022/12

**Large Language Models:
What will happen next?**

坂口慶祐 (東北大學)
keisuke.sakaguchi@tohoku.ac.jp



GPT Index
@gpt_index

 YouChat

👋 Hello! My name is YouChat, I'm an AI that can
summarize text, compose emails, and write code

Large Language Models: ~~What will happen next?~~ What is happening now?

坂口慶祐 (東北大学)

keisuke.sakaguchi@tohoku.ac.jp

Large Language Models (LLMs)

Large Language Models (LLMs)

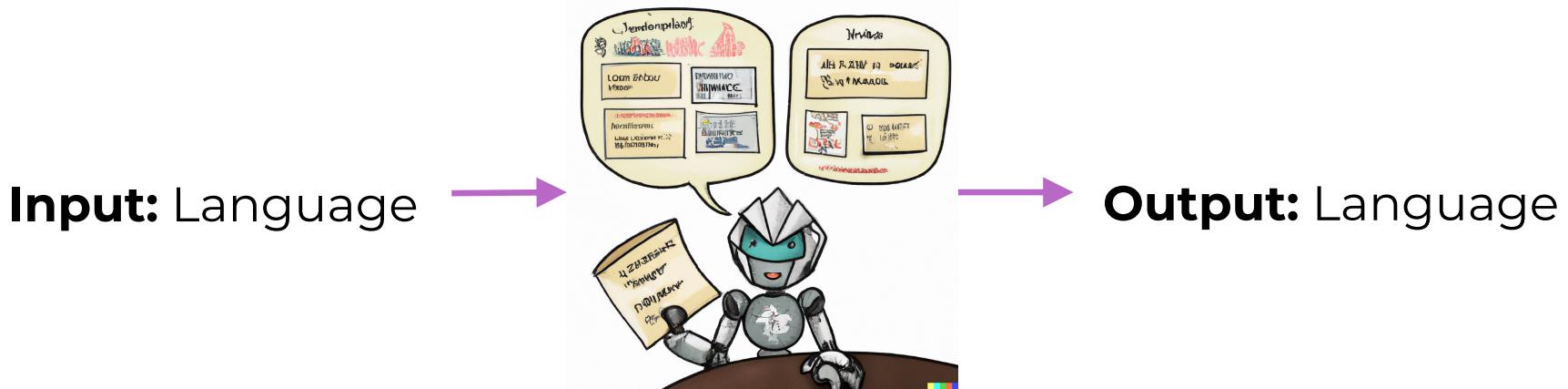


What's LLM?



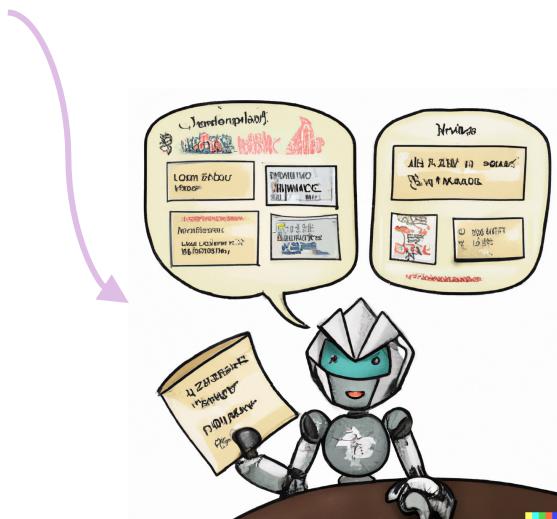
What's missing?

Large Language Models (LLMs)



Large Language Models (LLMs)

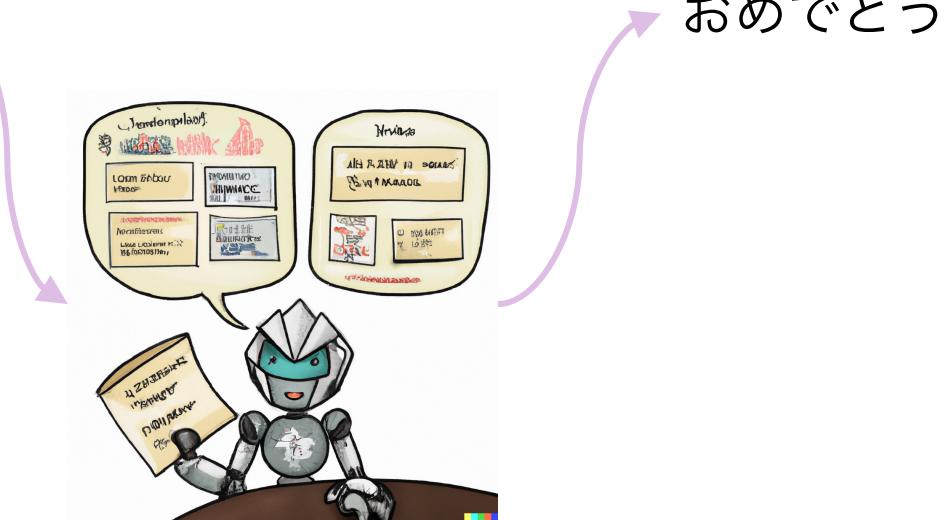
あけまして



Large Language Models (LLMs)

あけまして

おめでとう



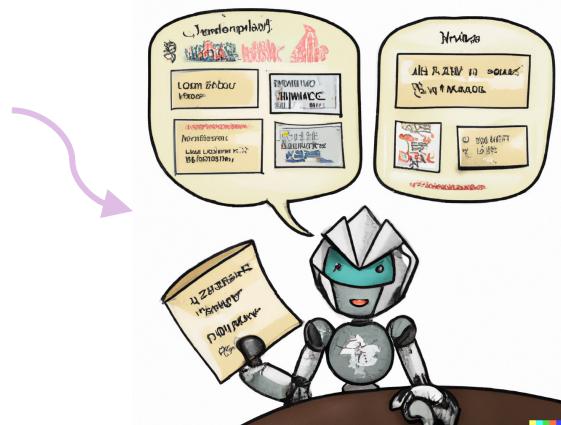
Large Language Models (LLMs)

あけまして

宮城県の
県庁所在地は？

おめでとう

仙台市



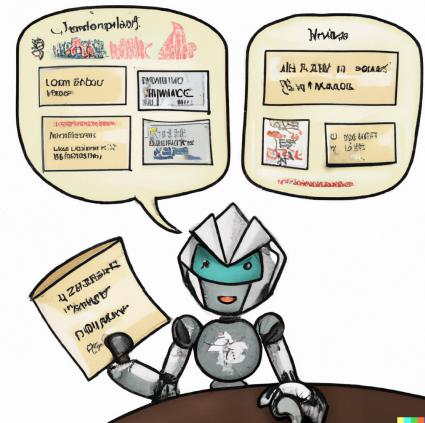
Large Language Models (LLMs)

あけまして

おめでとう

宮城県の
県庁所在地は？

「吾輩は猫である」
を英語で言うと？



仙台市

I'm a cat.

Large Language Models (LLMs)

あけまして

宮城県の
県庁所在地は？

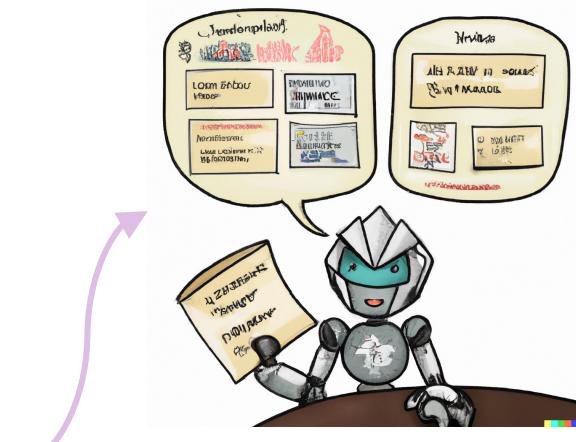
「吾輩は猫である」
を英語で言うと？

月火[?]木金土日

おめでとう

仙台市

I'm a cat.



Large Language Models (LLMs)

あけまして

宮城県の
県庁所在地は？

「吾輩は猫である」
を英語で言うと？

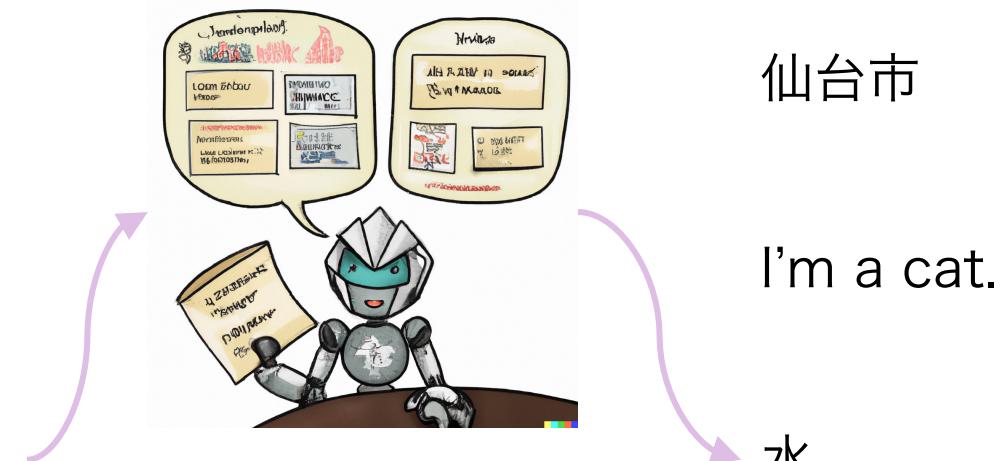
月火[?]木金土日

おめでとう

仙台市

I'm a cat.

水



Large Language Models (LLMs)



K

「あけまして」の続きは？

「おめでとうございます」です。

K

宮城県の県庁所在地はどこ？

宮城県の県庁所在地は仙台市です。

K

「吾輩は猫である」を英語でいうと？

「吾輩は猫である」を英語にすると「I am a cat」になります。

K

月火{ }木金土日。{ }に入る漢字1文字とは？

月火{水}木金土日。{ }に入る漢字1文字は「水」です。

Large Language Models (LLMs)



K

太郎はりんごを6個持っています。花子は太郎よりも3個多く持っています。花子はりんごをいくつ持っていますか？

花子はりんごを3個多く持っていますので、りんごを $6+3=9$ 個持っています。

K

1から100までの間の素数を出力するpythonプログラムを書いて

Copy code

```
def is_prime(n):
    if n < 2:
        return False
    for i in range(2, int(n ** 0.5) + 1):
        if n % i == 0:
            return False
    return True

primes = [i for i in range(1, 101) if is_prime(i)]
print(primes)
```

LLM: (Transformer) Encoder-Decoder

LLM: (Transformer) Encoder-Decoder



arXiv > cs > arXiv:1706.03762

Search...
Help | Adv

Computer Science > Computation and Language

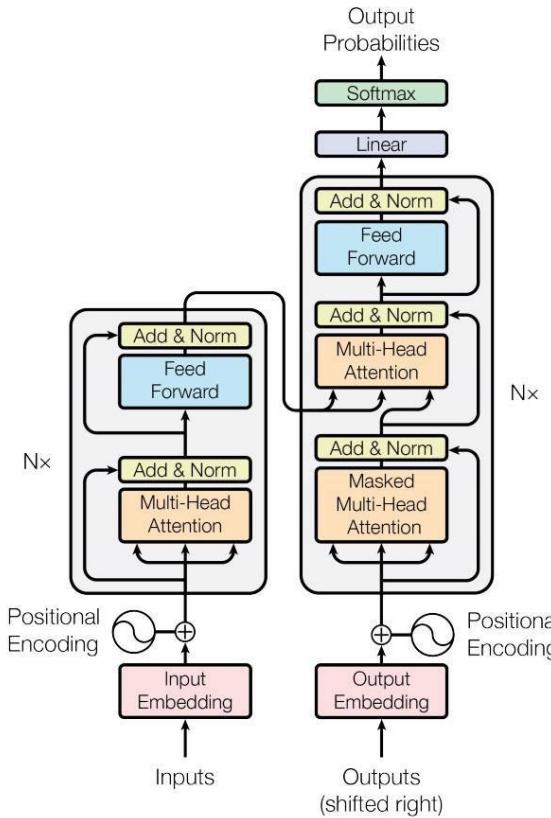
[Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5)]

Attention Is All You Need

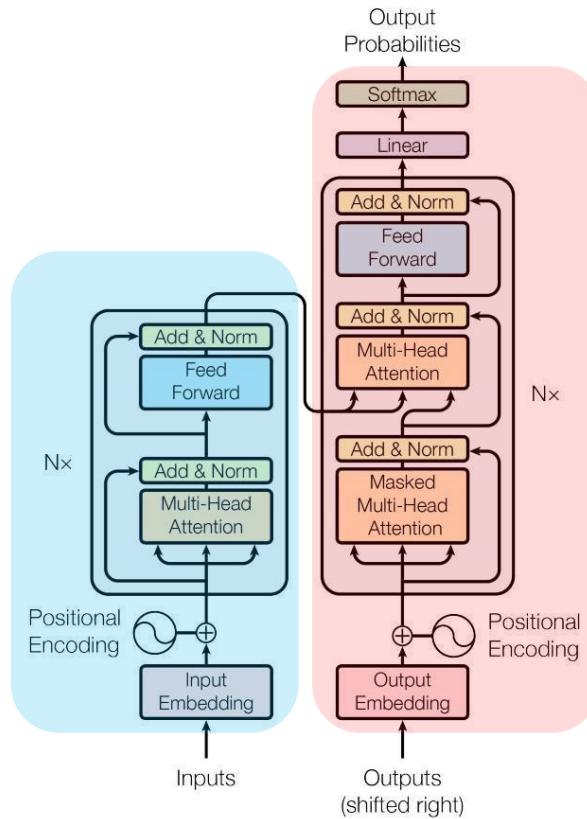
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results,

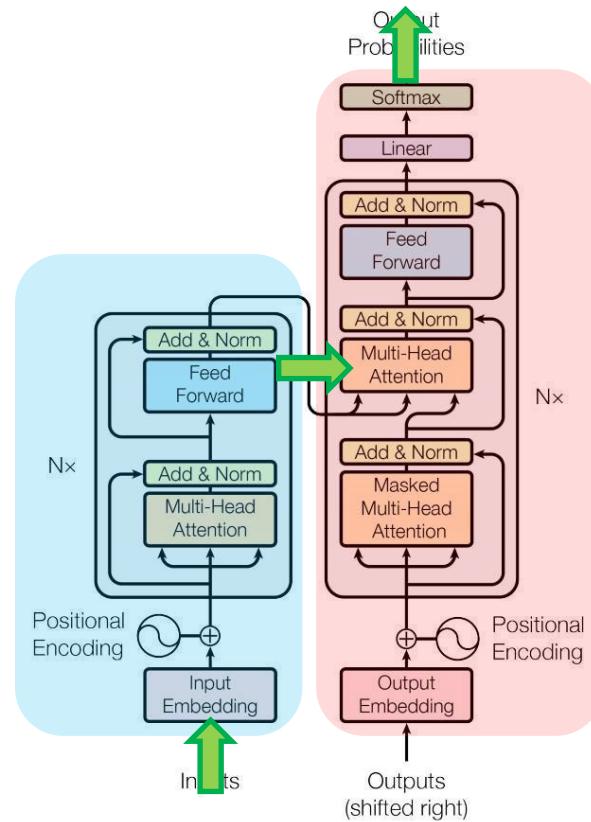
LLM: (Transformer) Encoder-Decoder



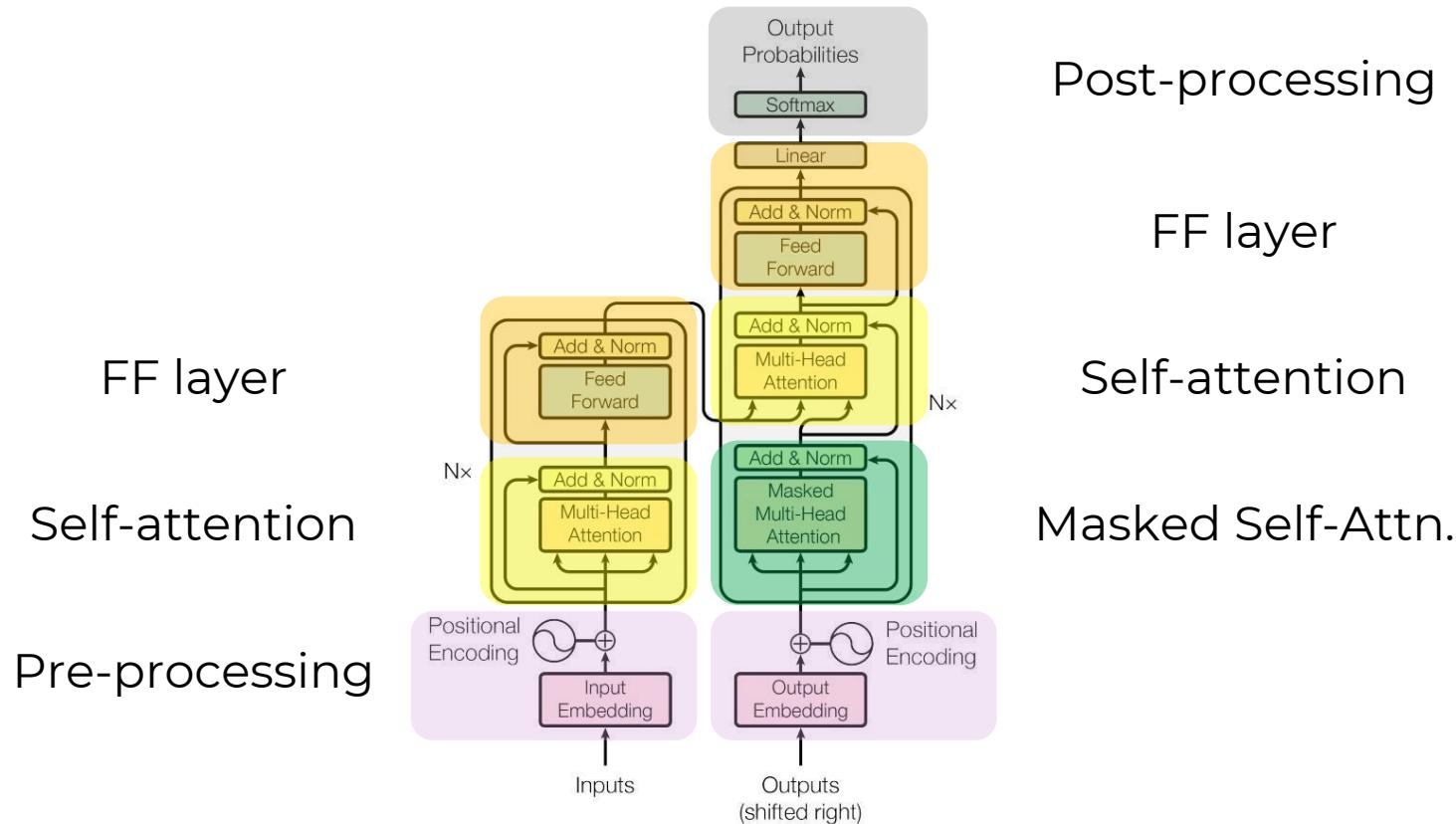
LLM: (Transformer) Encoder-Decoder



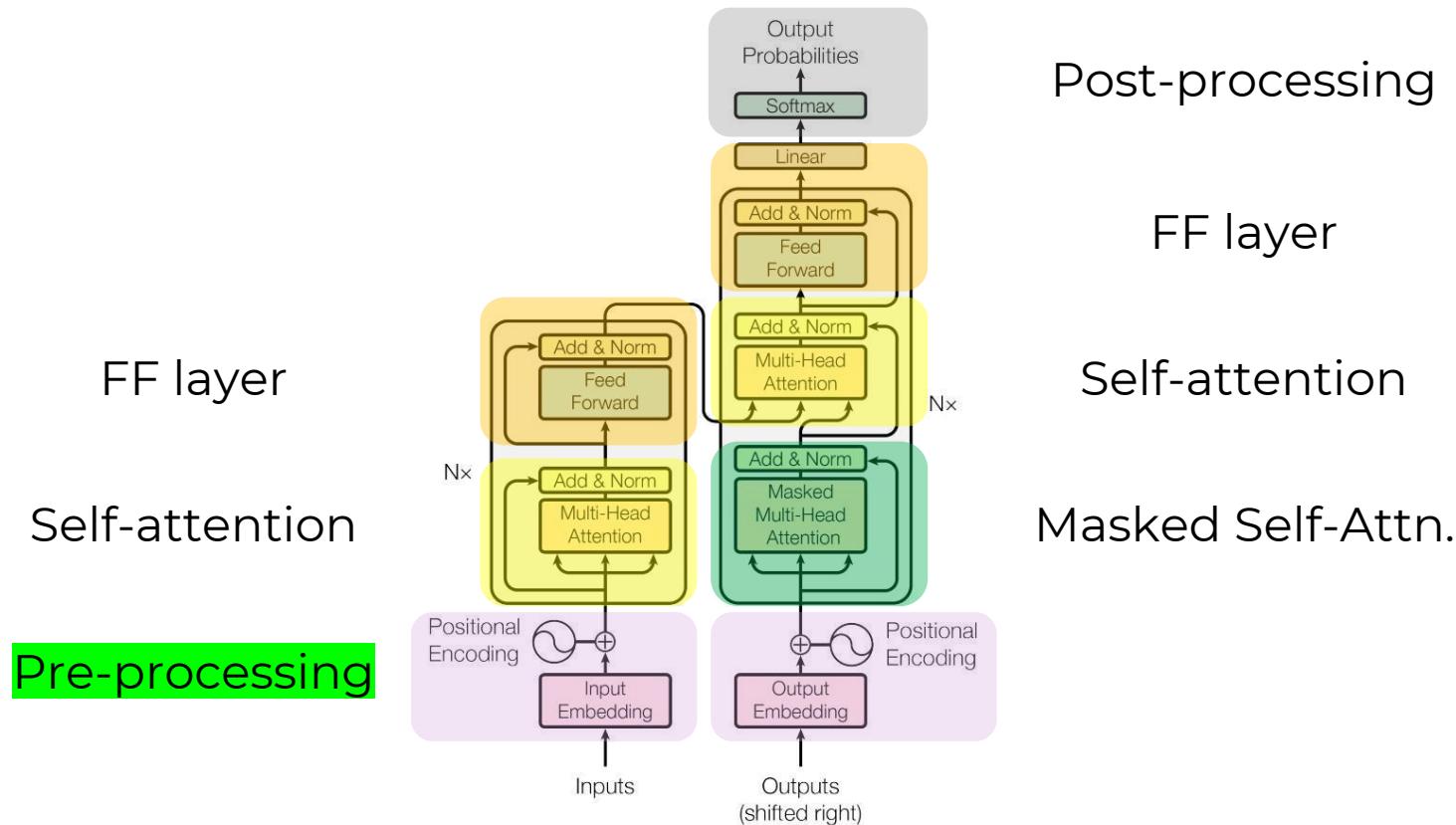
LLM: (Transformer) Encoder-Decoder



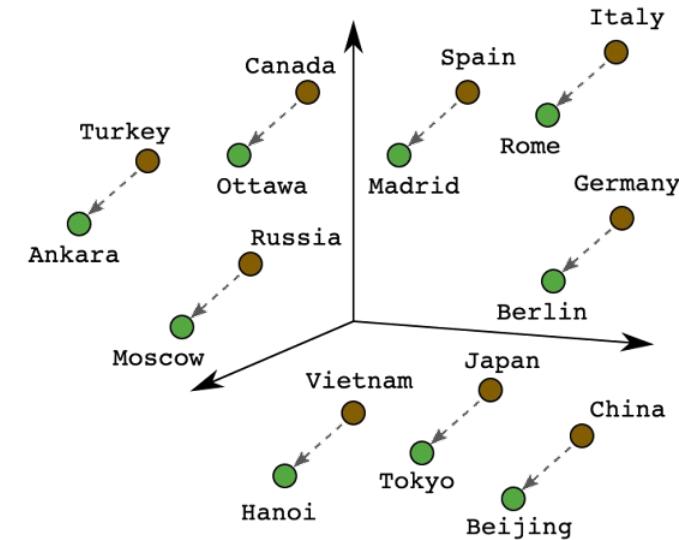
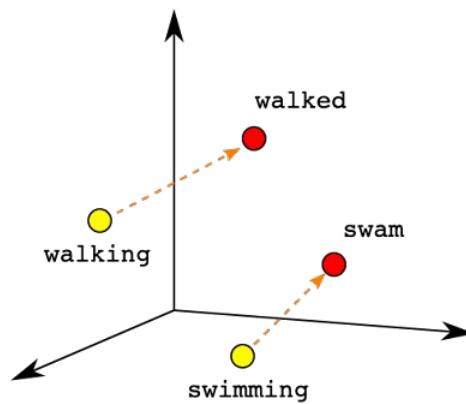
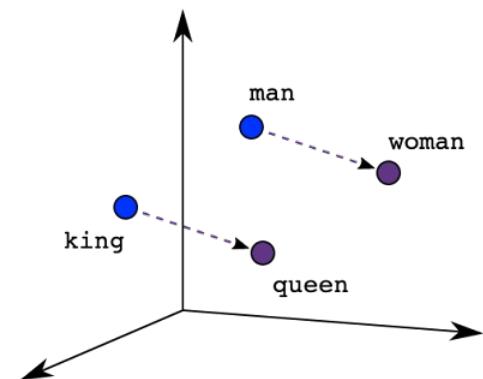
LLM: (Transformer) Encoder-Decoder



LLM: (Transformer) Encoder-Decoder



Pre-processing: word embedding



$$\text{king} - \text{man} + \text{woman} = \text{queen}$$

$$\text{walking} - \text{swimming} + \text{swam} = \text{walked}$$

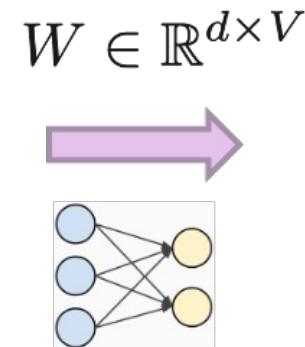
$$\text{Berlin} - \text{German} + \text{Japan} = ??$$

Pre-processing: word embedding

One-hot encoding

	cat	mat	on	sat	the
the =>	0	0	0	0	1
cat =>	1	0	0	0	0
sat =>	0	0	0	1	0
...	...				

Word embedding



cat =>	1.2	-0.1	4.3	3.2
mat =>	0.4	2.5	-0.9	0.5
on =>	2.1	0.3	0.1	0.4
...	...			

Pre-processing: positional encoding

A big grey **dog** chased a small brown **dog**.



“dog” refers different objects depending on the context.

→ Inject information about word order

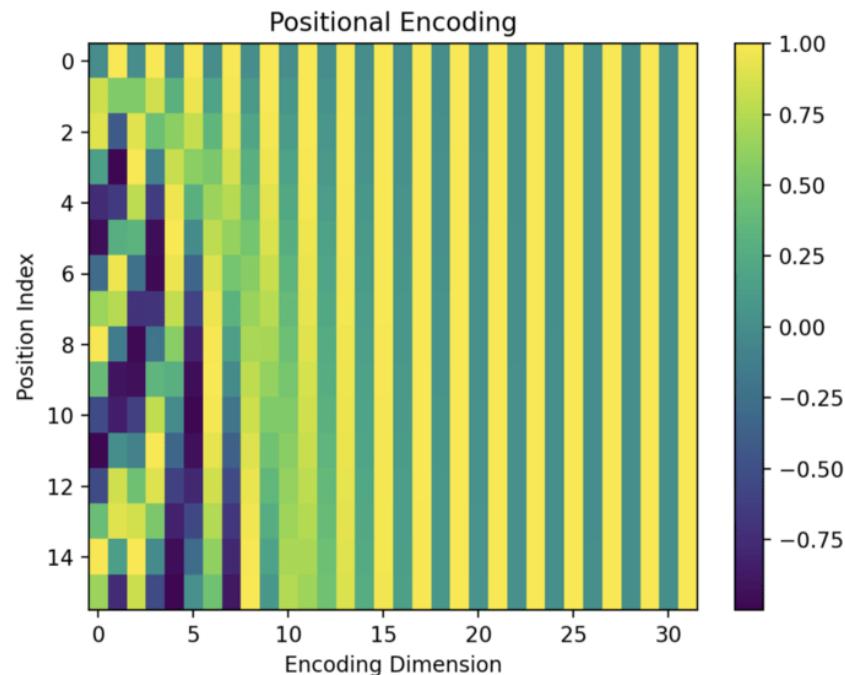
Pre-processing: positional encoding

A big grey dog chased a small brown dog.

Positional encoding

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$



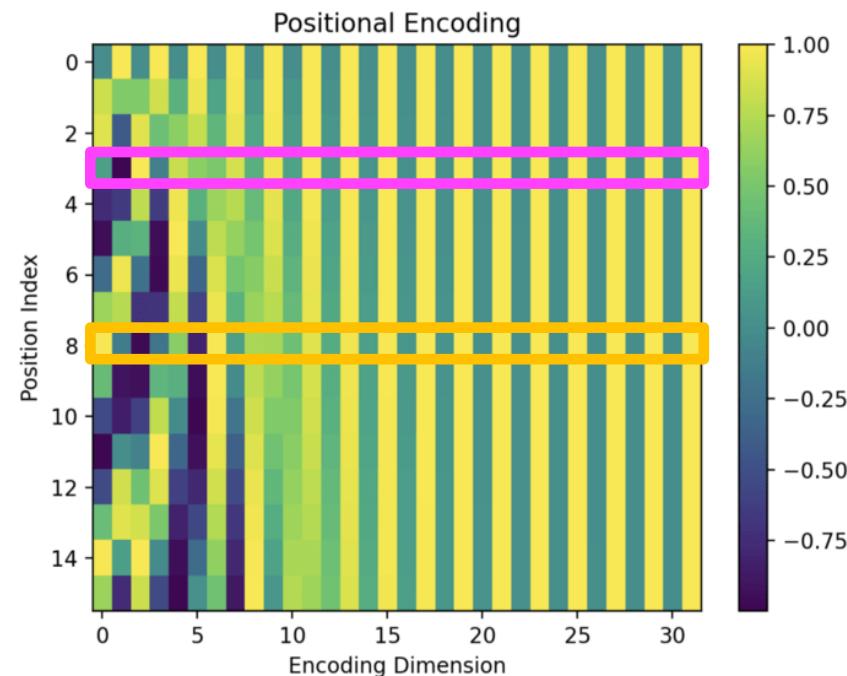
Pre-processing: positional encoding

A big grey **dog** chased a small brown **dog**.

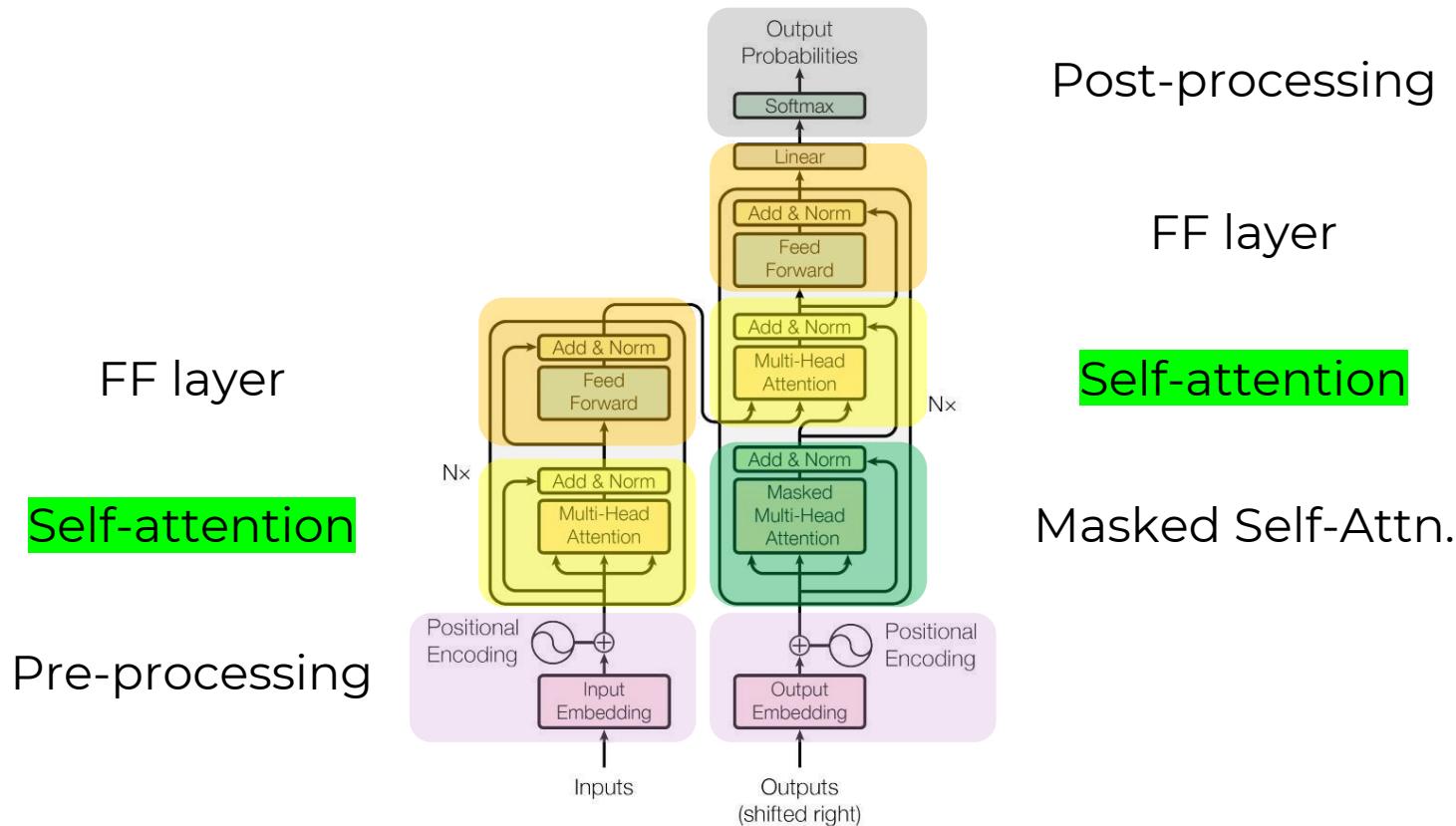
Positional encoding

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$

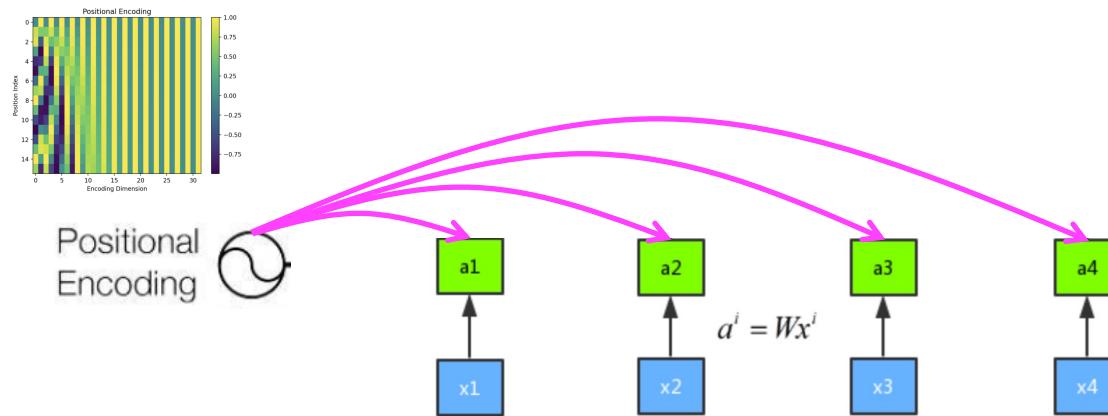
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$



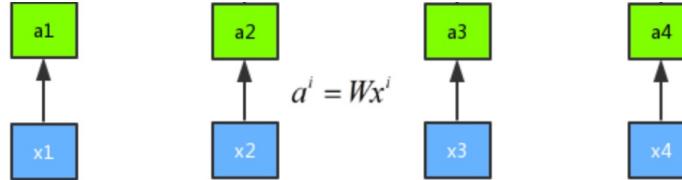
LLM: (Transformer) Encoder-Decoder



(self) Attention is all you need.



(self) Attention is all you need.

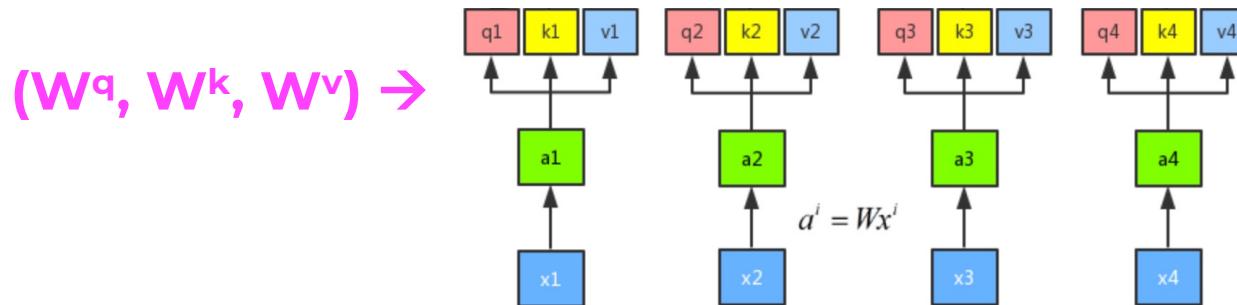


(self) Attention is all you need.

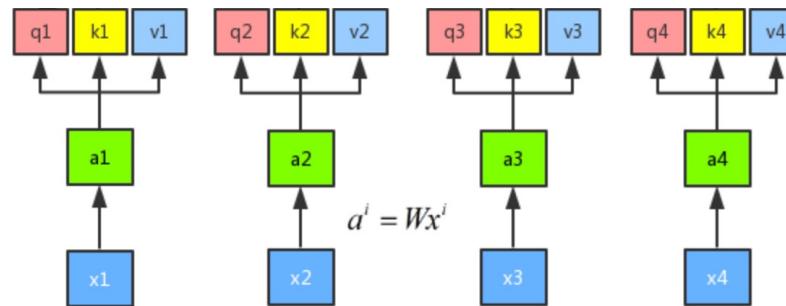
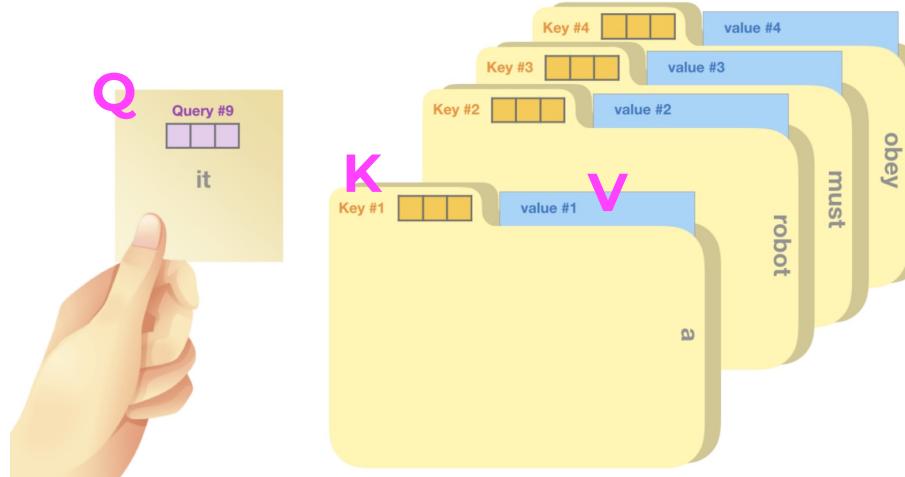
q: query ($q^i = W^q a^i$)

k: key ($k^i = W^k a^i$)

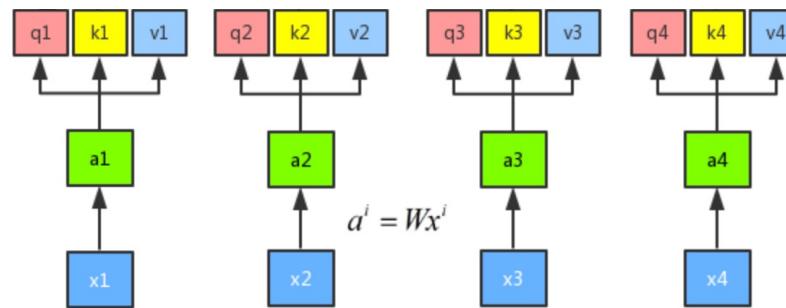
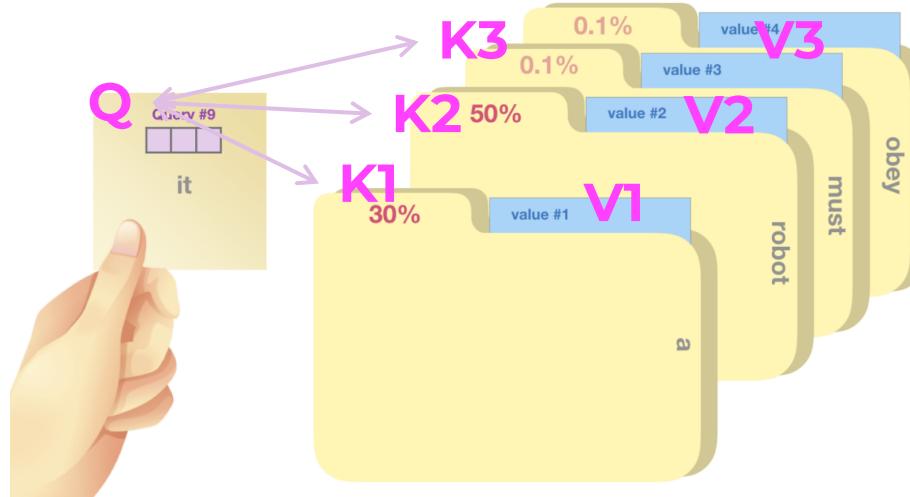
v: value ($v^i = W^v a^i$)



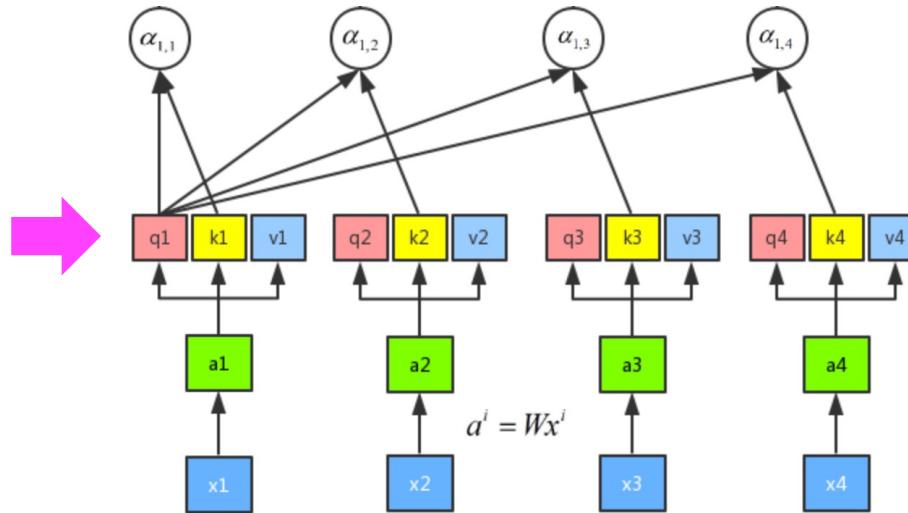
(self) Attention is all you need.



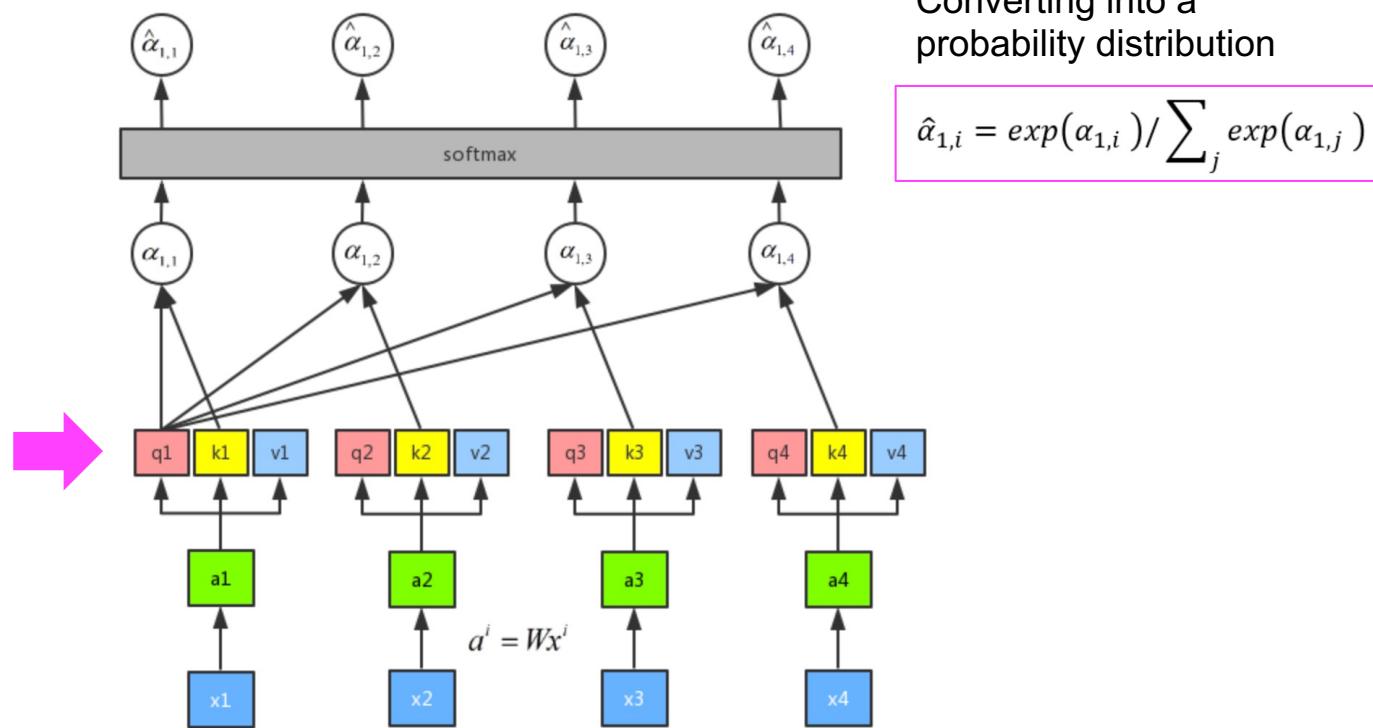
(self) Attention is all you need.



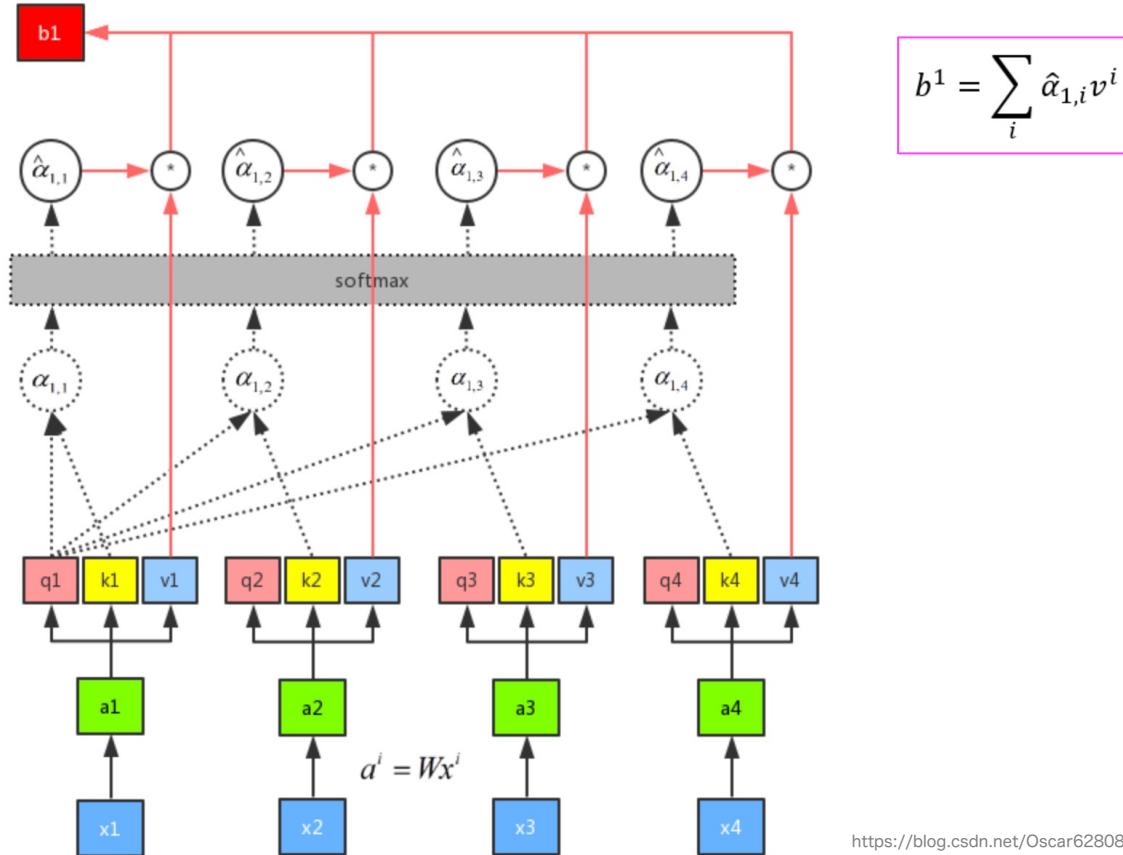
(self) Attention is all you need.



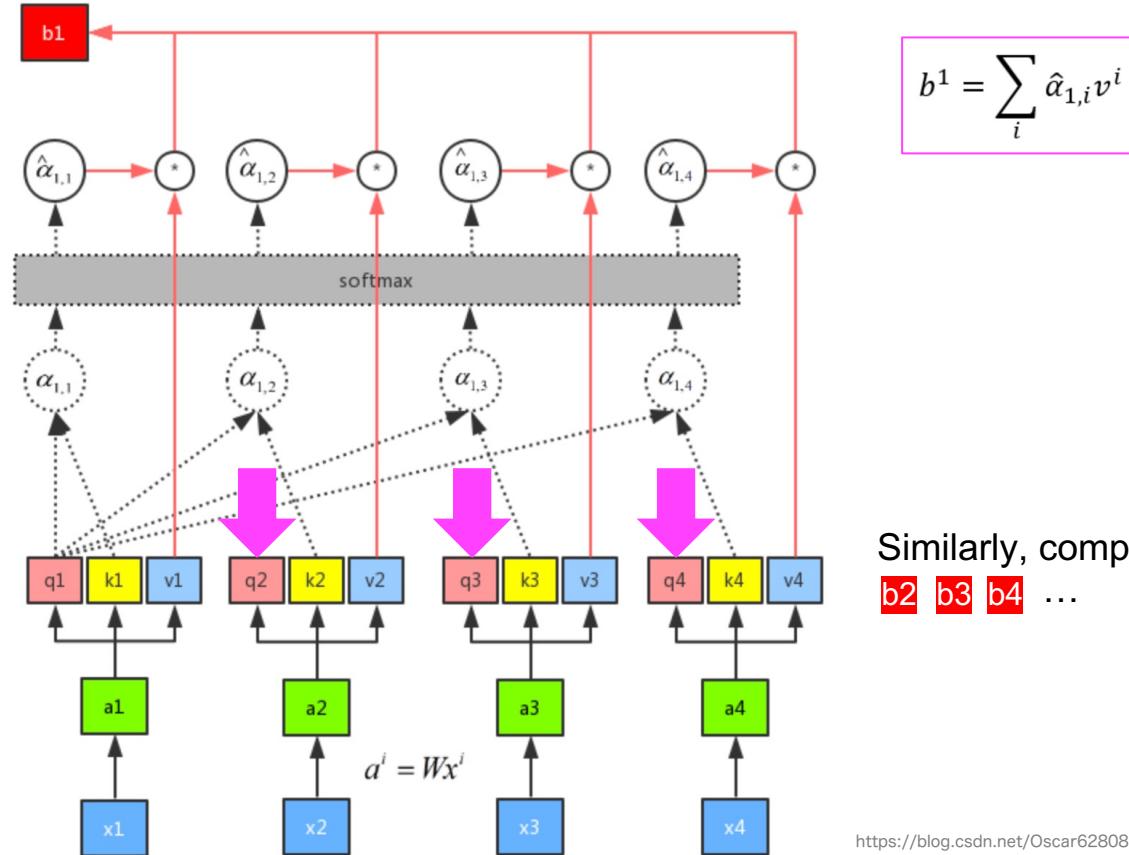
(self) Attention is all you need.



(self) Attention is all you need.



(self) Attention is all you need.



Visualizing attention

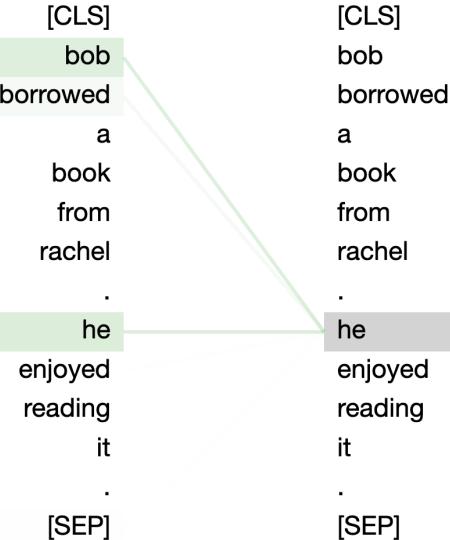
Bob borrowed a book from Rachel. He enjoyed reading it.

Visualizing attention

Bob borrowed a book from Rachel. He enjoyed reading it.

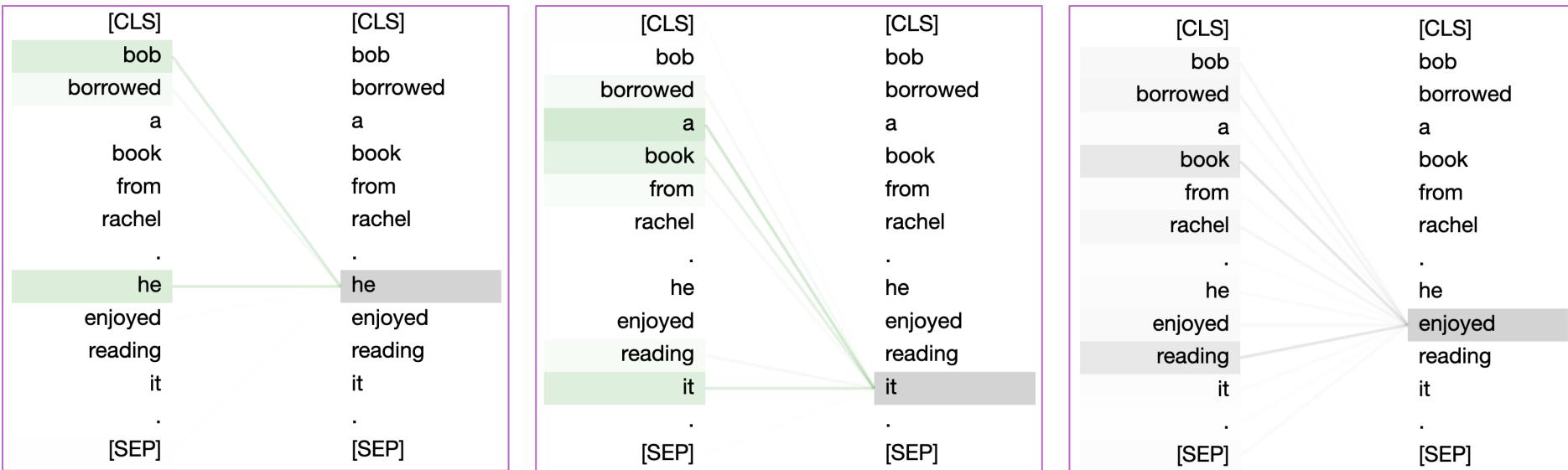
Attend from

to



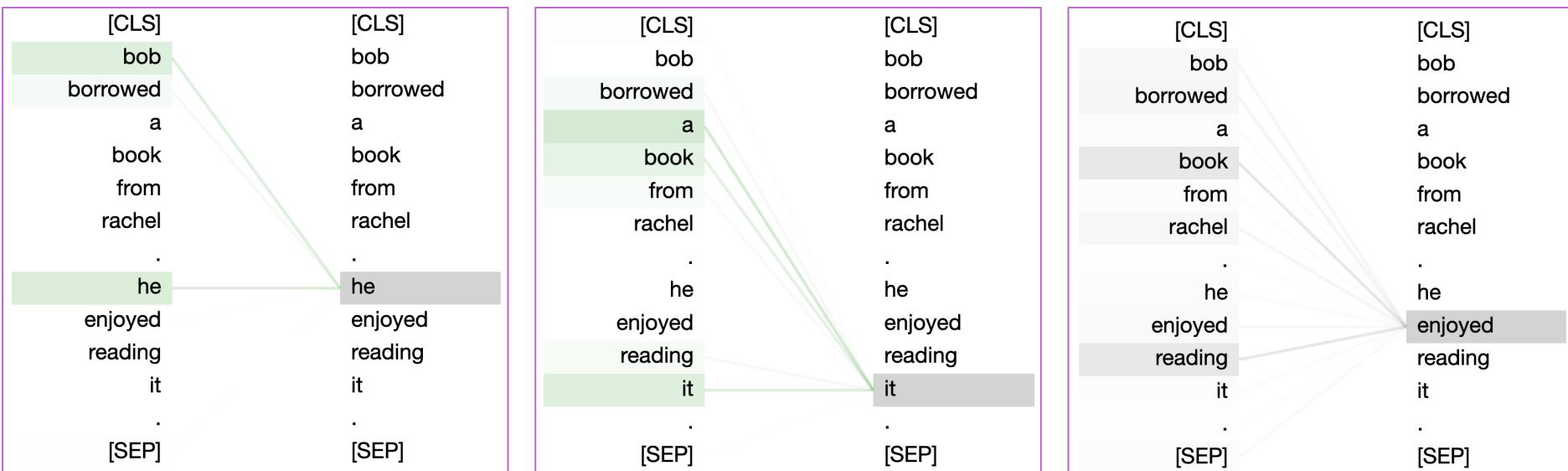
Visualizing attention

Bob borrowed a book from Rachel. He enjoyed reading it.



Visualizing attention

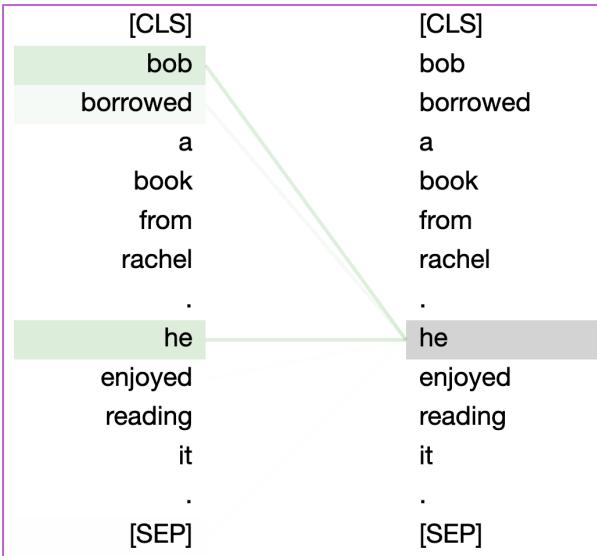
Bob borrowed a book from Rachel. He enjoyed reading it.



But wait, you may have a different point of view.

Visualizing multi-head attention

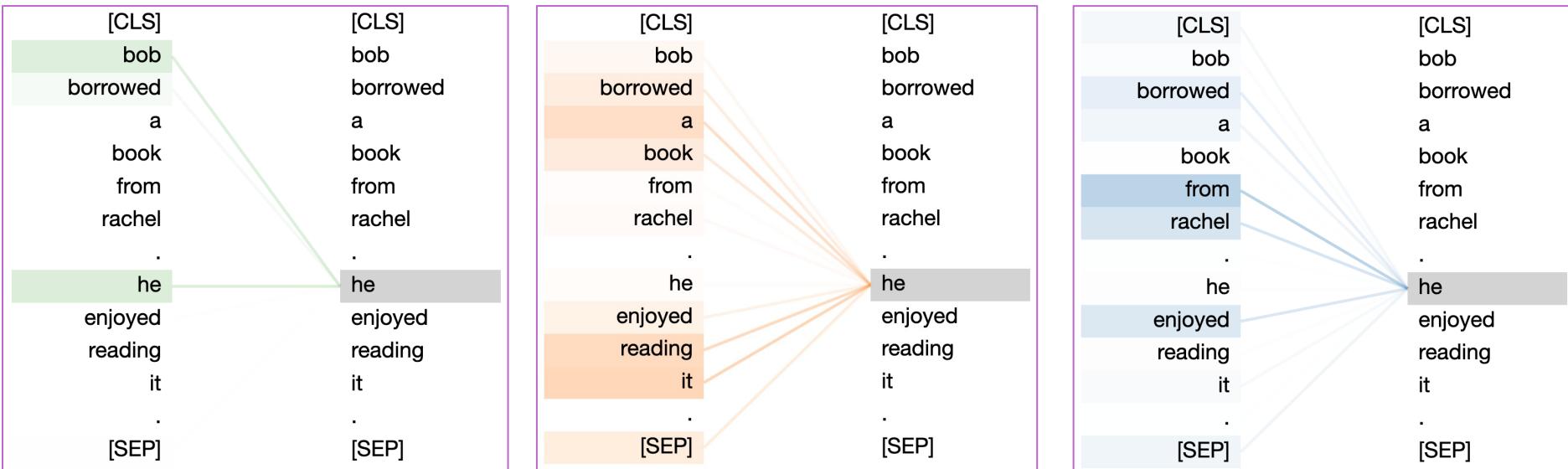
Bob borrowed a book from Rachel. He enjoyed reading it.



Multi-head attn. captures different perspectives/context!

Visualizing multi-head attention

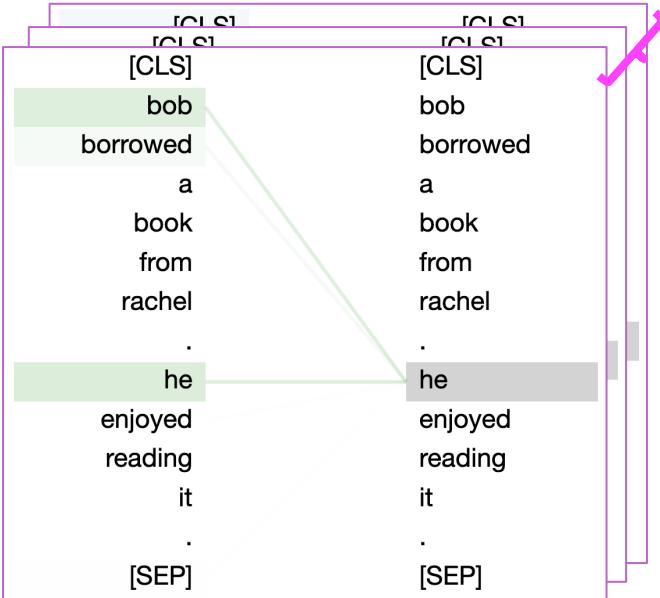
Bob borrowed a book from Rachel. He enjoyed reading it.



Multi-head attn. captures different perspectives/context!

Visualizing multi-head attention

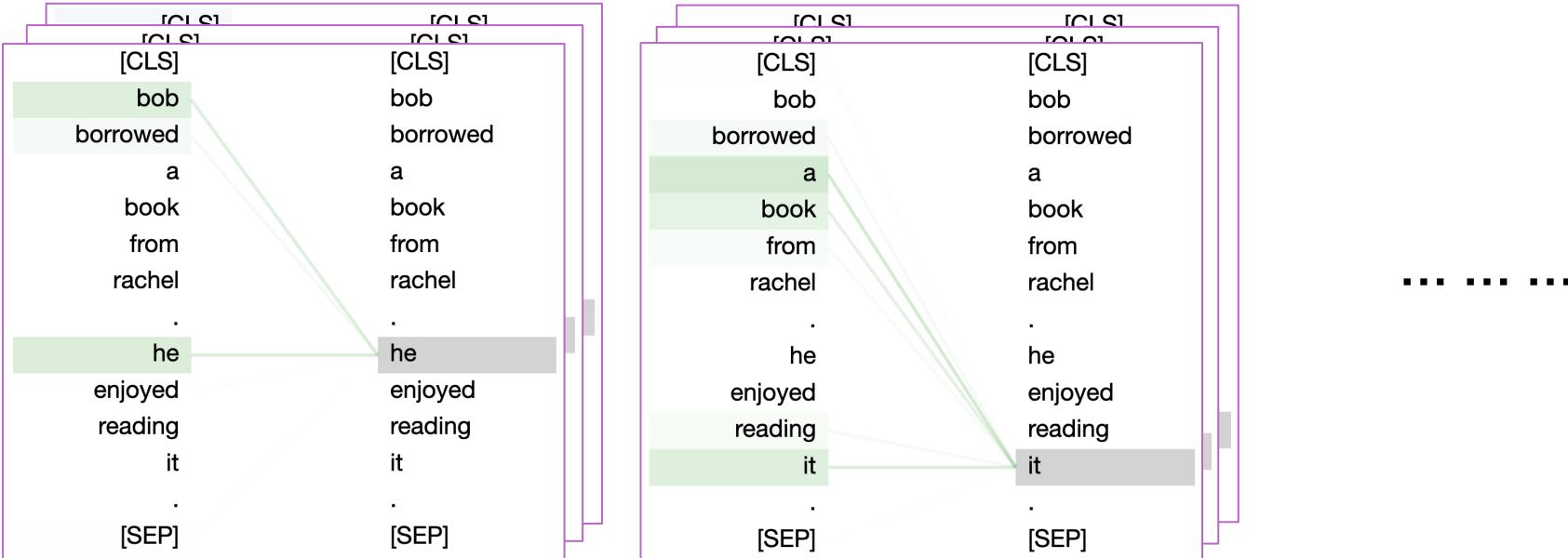
Bob borrowed a book from Rachel. He enjoyed reading it.



Multi-head attn. captures different perspectives/context!

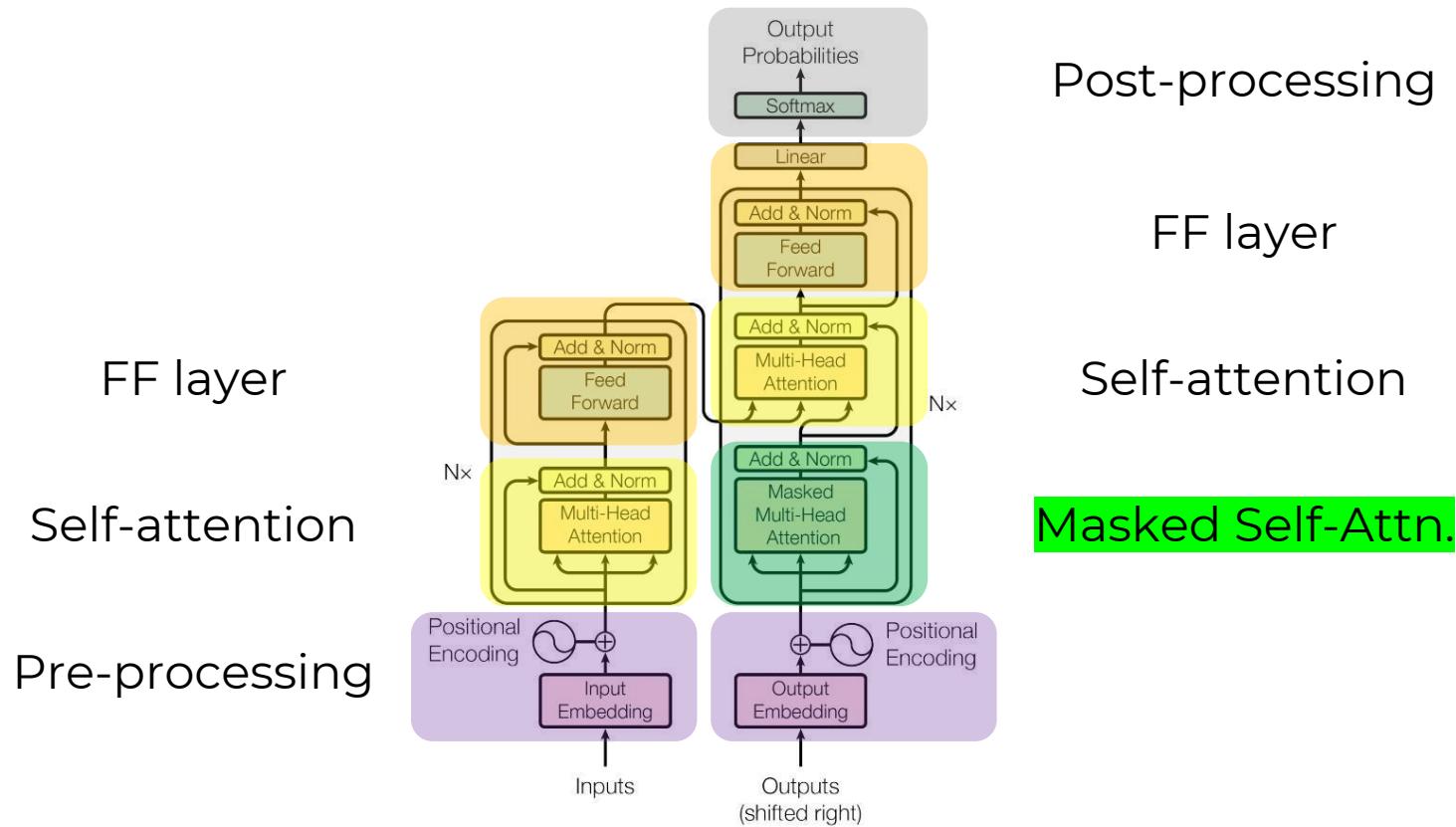
Visualizing multi-head attention

Bob borrowed a book from Rachel. He enjoyed reading it.

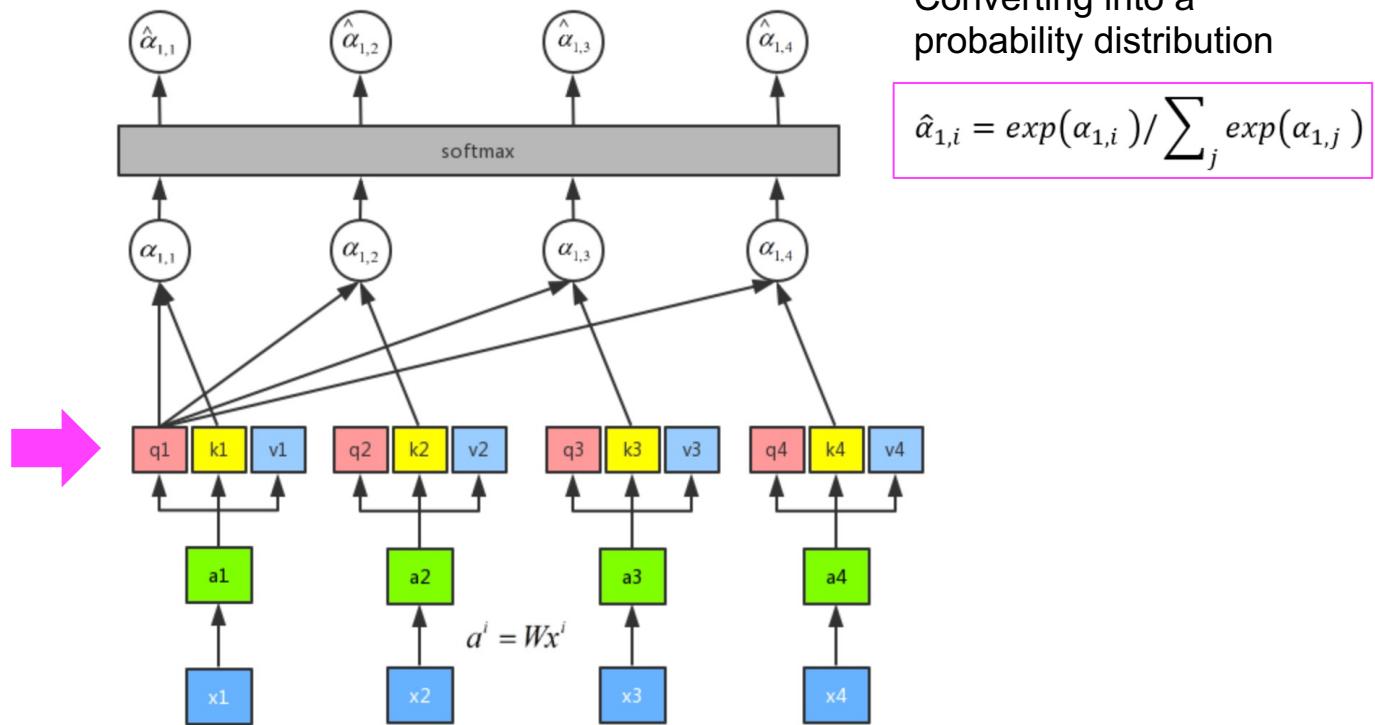


Multi-head attn. captures different perspectives/context!

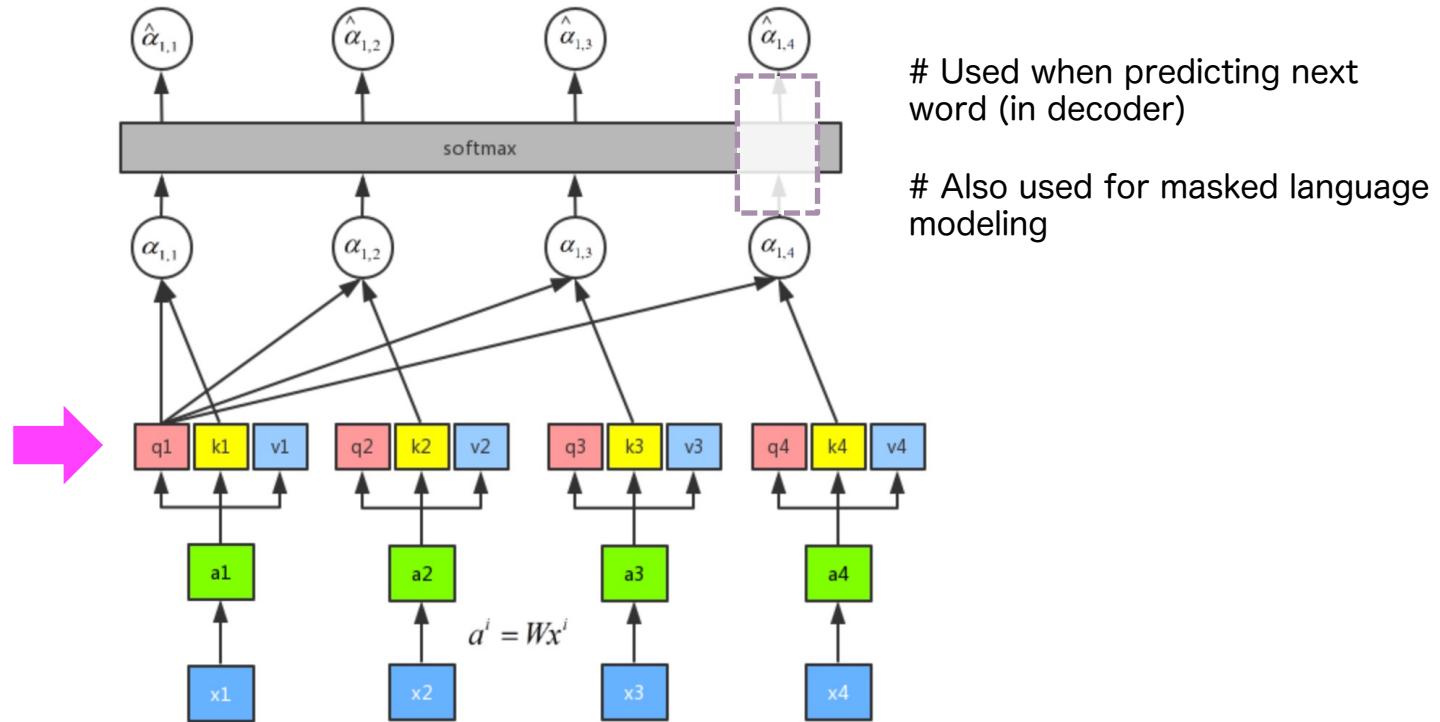
LLM: (Transformer) Encoder-Decoder



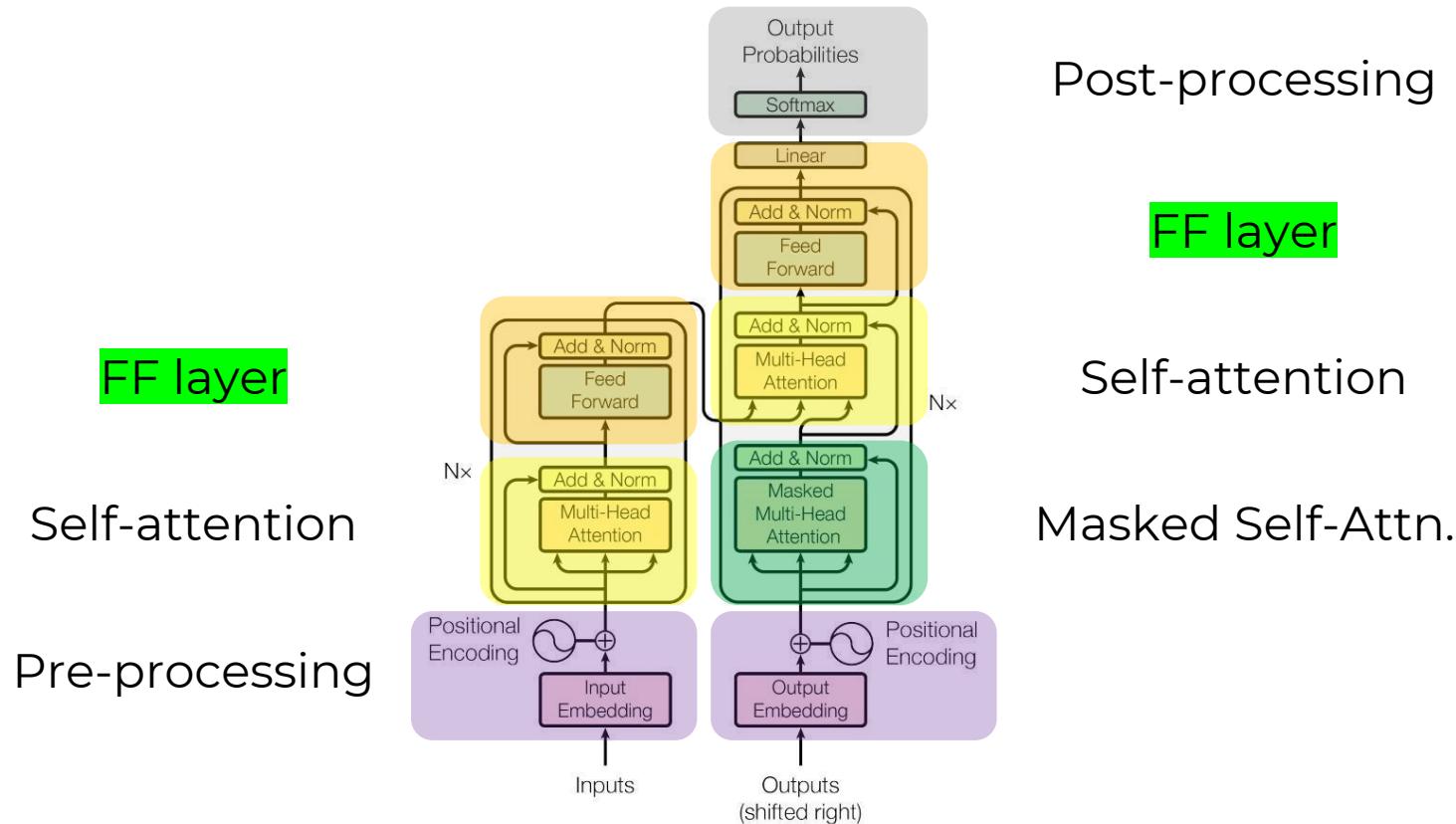
(self) Attention is all you need.



(self) Attention is all you need.

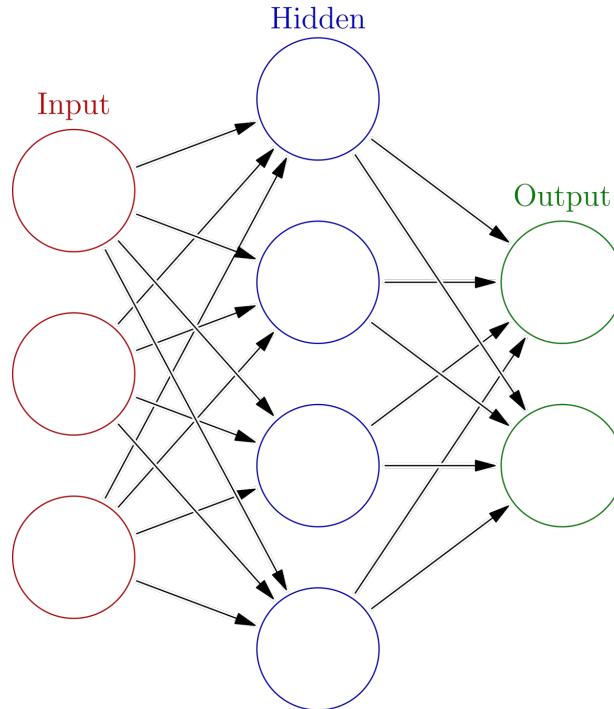


LLM: (Transformer) Encoder-Decoder

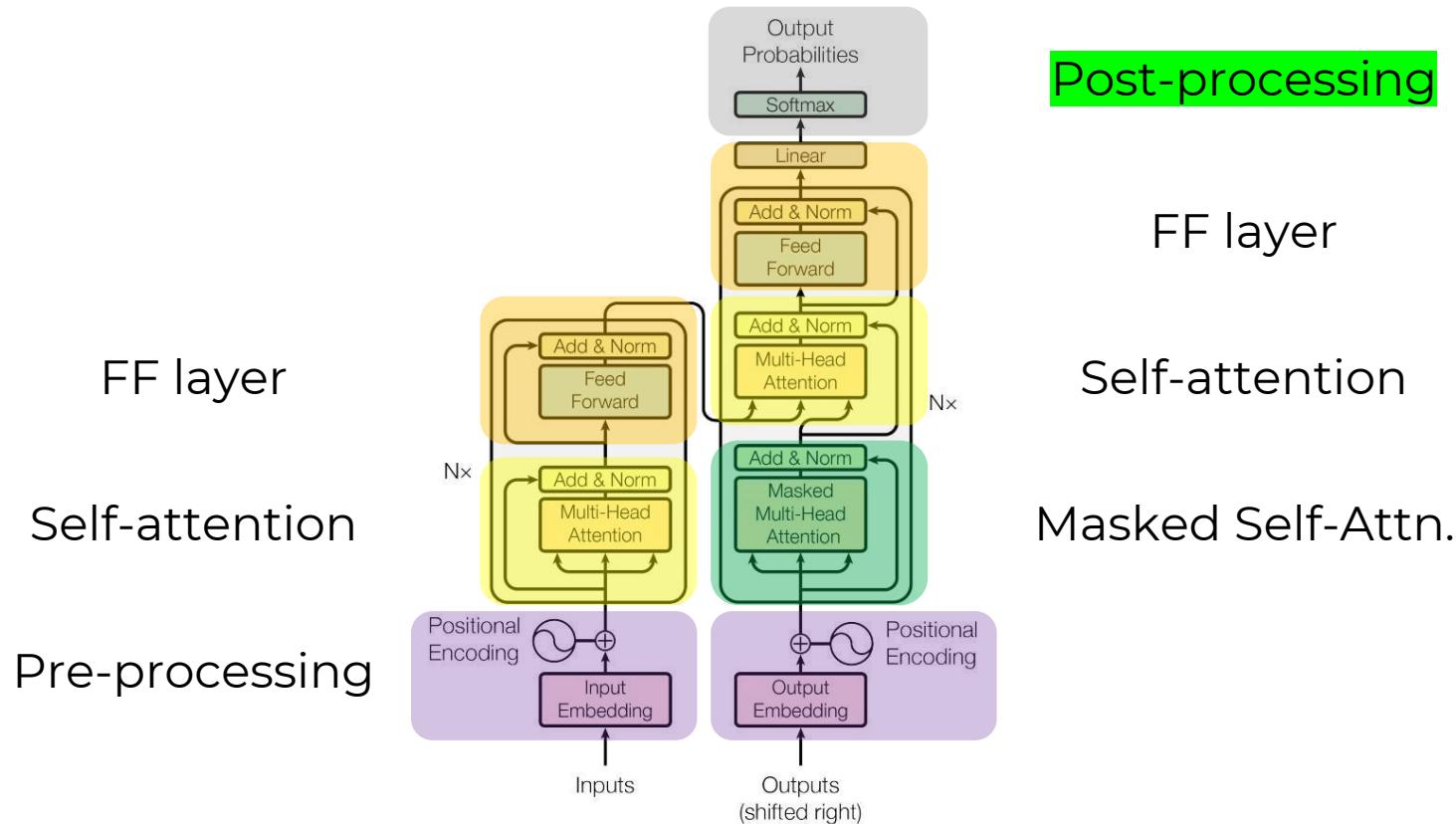


Feed-forward neural network

FFN enables both linear and nonlinear transformations for the input signals.



LLM: (Transformer) Encoder-Decoder



One-hot encoding → word embedding (pre-processing)

One-hot encoding

	cat	mat	on	sat	the
the =>	0	0	0	0	1
cat =>	1	0	0	0	0
sat =>	0	0	0	1	0
...

Word embedding



cat =>	1.2	-0.1	4.3	3.2
mat =>	0.4	2.5	-0.9	0.5
on =>	2.1	0.3	0.1	0.4
...

One-hot encoding ← word embedding (post-processing)

One-hot encoding

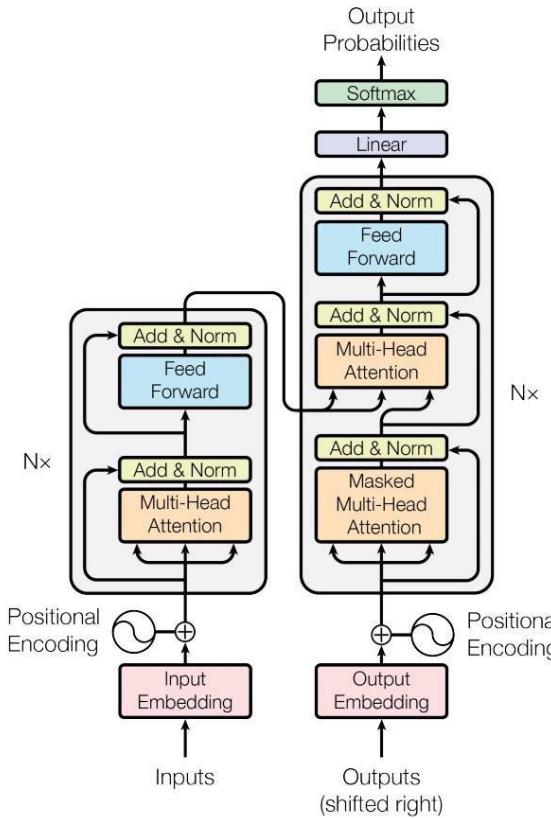
	cat	mat	on	sat	the
the =>	0	0	0	0	1
cat =>	1	0	0	0	0
sat =>	0	0	0	1	0
...

Word embedding



cat =>	1.2	-0.1	4.3	3.2
mat =>	0.4	2.5	-0.9	0.5
on =>	2.1	0.3	0.1	0.4
...

LLM: (Transformer) Encoder-Decoder



Self-supervised (pre)training:

Auto-regressive

Mask infilling



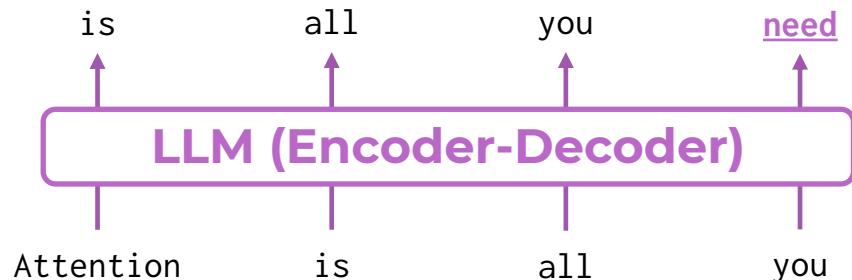
Large text corpora

(Wikipedia, Book Corpus, Common Crawl, Wayback machine, etc.)

Self-supervised (pre)training:

Auto-regressive

Mask infilling

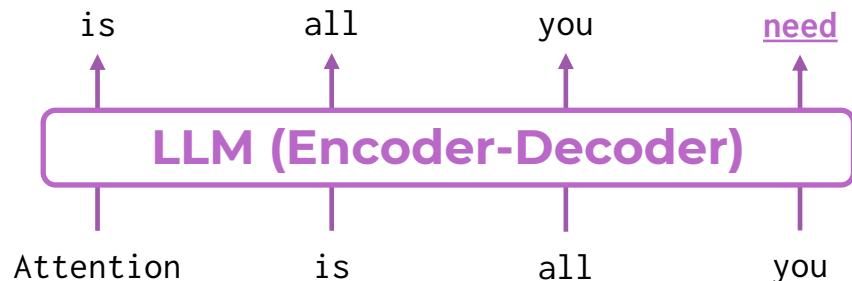


Large text corpora

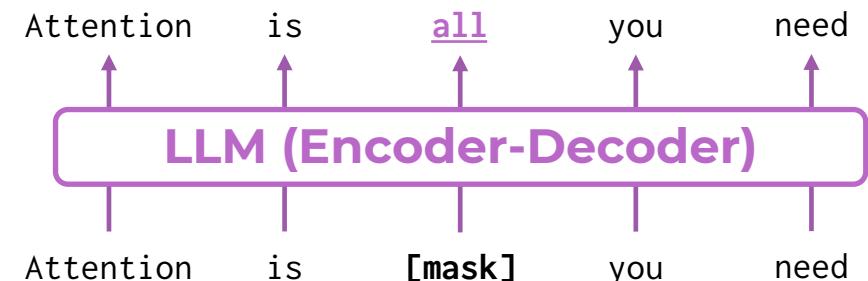
(Wikipedia, Book Corpus, Common Crawl, Wayback machine, etc.)

Self-supervised (pre)training:

Auto-regressive



Mask infilling



Large text corpora

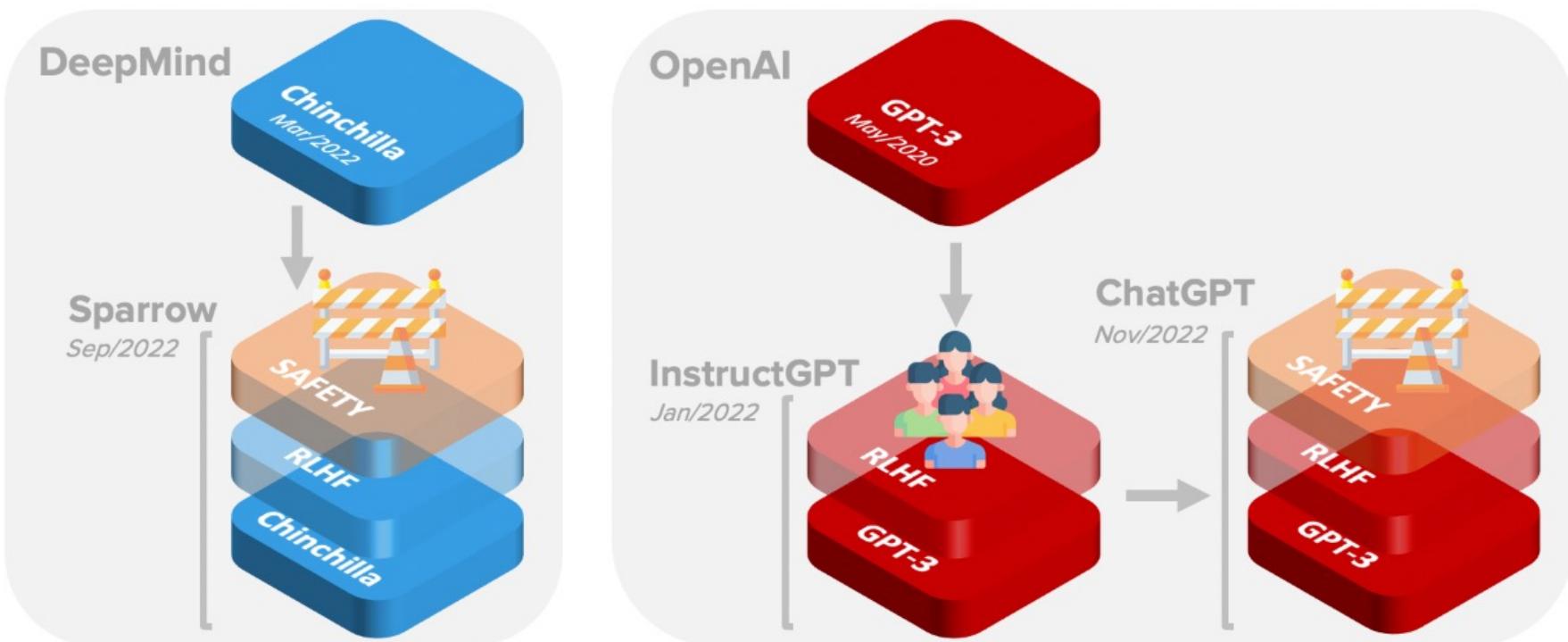
(Wikipedia, Book Corpus, Common Crawl, Wayback machine, etc.)

ARTIFICIAL INTELLIGENCE

Why Meta's latest large language model survived only three days online

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

AI alignment with human values



Reinforcement Learning from Human Feedback (RLHF)

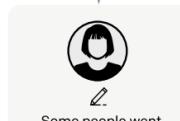
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



- A Explain gravity...
- B Explain war...
- C Moon is natural satellite of...
- D People went to the moon...

A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



- D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

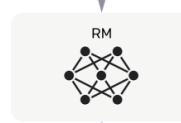


The policy generates an output.



Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

r_k

Reinforcement Learning from Human Feedback (RLHF)

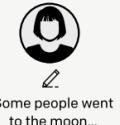
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

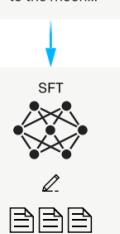
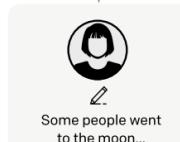


Explain the moon landing to a 6 year old



Some people went to the moon...

A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.

Step 2

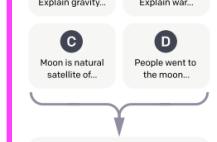
Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



Explain the moon landing to a 6 year old

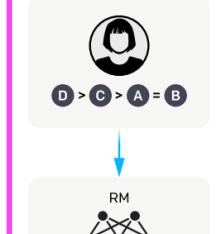
A, Explain gravity...
B, Explain war...
C, Moon is natural satellite of...
D, People went to the moon...



D > C > A = B

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



RM

D > C > A = B

Step 3

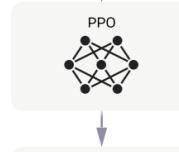
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



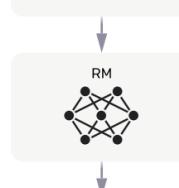
Write a story about frogs

The policy generates an output.



PPO

Once upon a time...



RM

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

r_k

Reinforcement Learning from Human Feedback (RLHF)

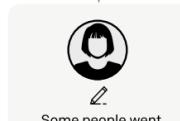
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



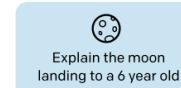
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



- A: Explain gravity...
- B: Explain war...
- C: Moon is natural satellite of...
- D: People went to the moon...

A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



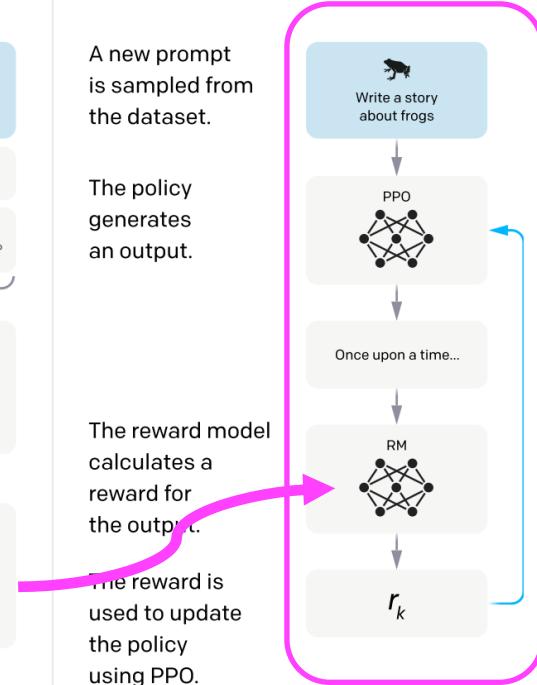
Once upon a time...

The reward model calculates a reward for the output.



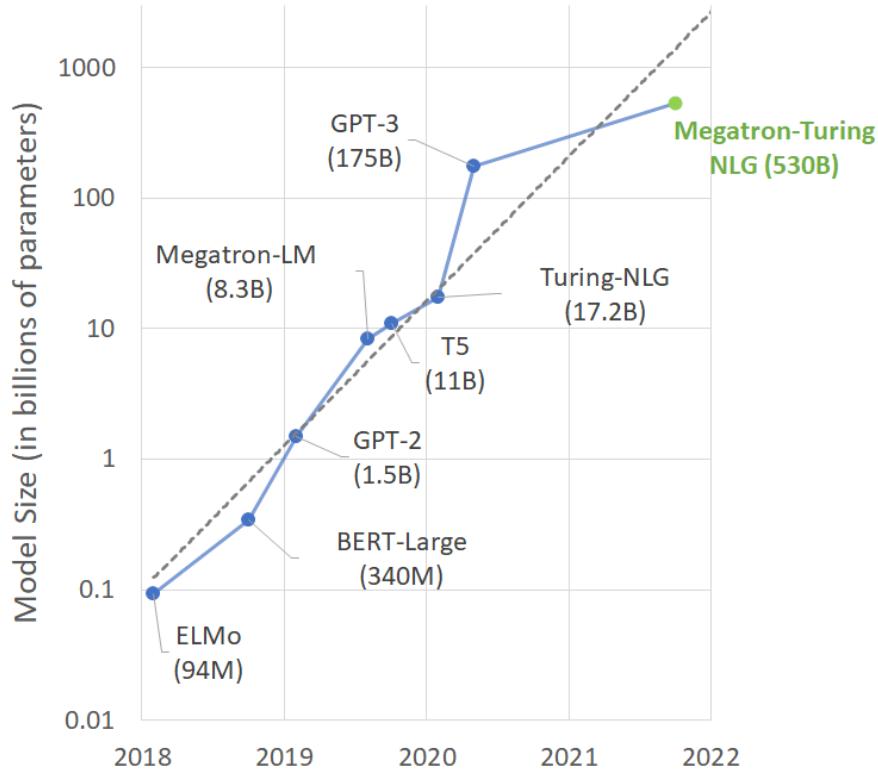
The reward is used to update the policy using PPO.

r_k

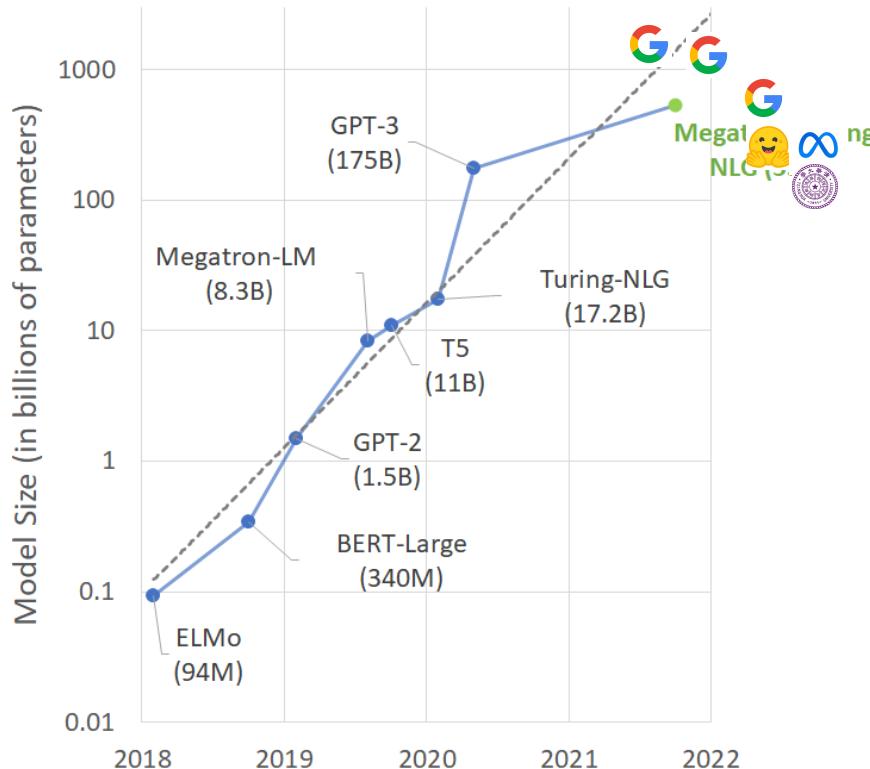


LLM parameter size:

LLM parameter size: 10x / year



LLM parameter size: 10x / year



SwitchTransformer (1.6T)



GLaM (1.2T)



PaLM (540B)



BLOOM (176B)



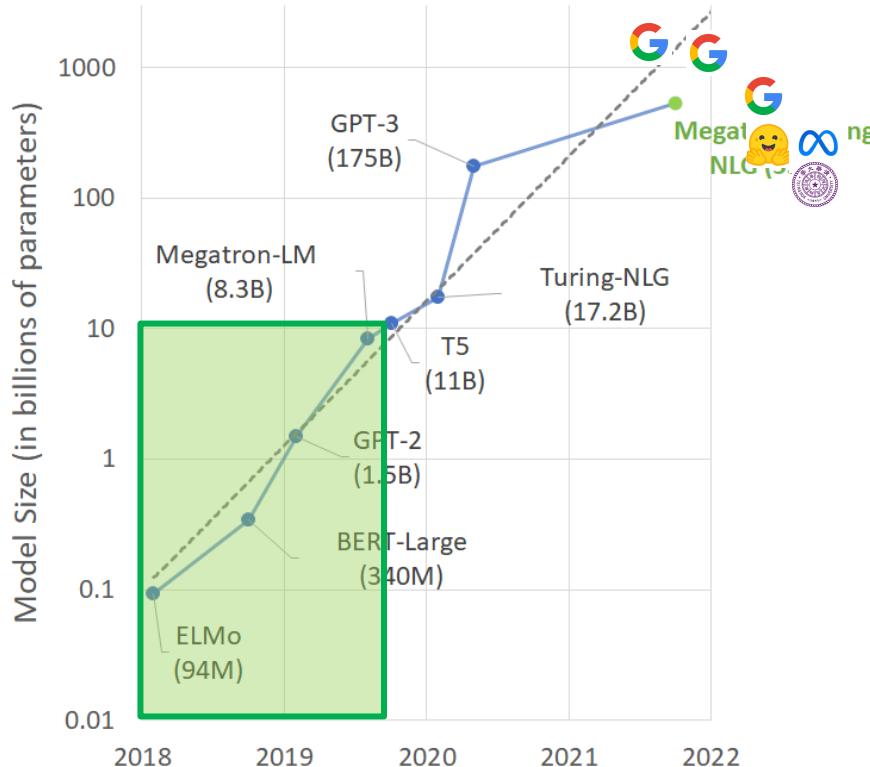
Blender Bot (175B)



GLM (130B)

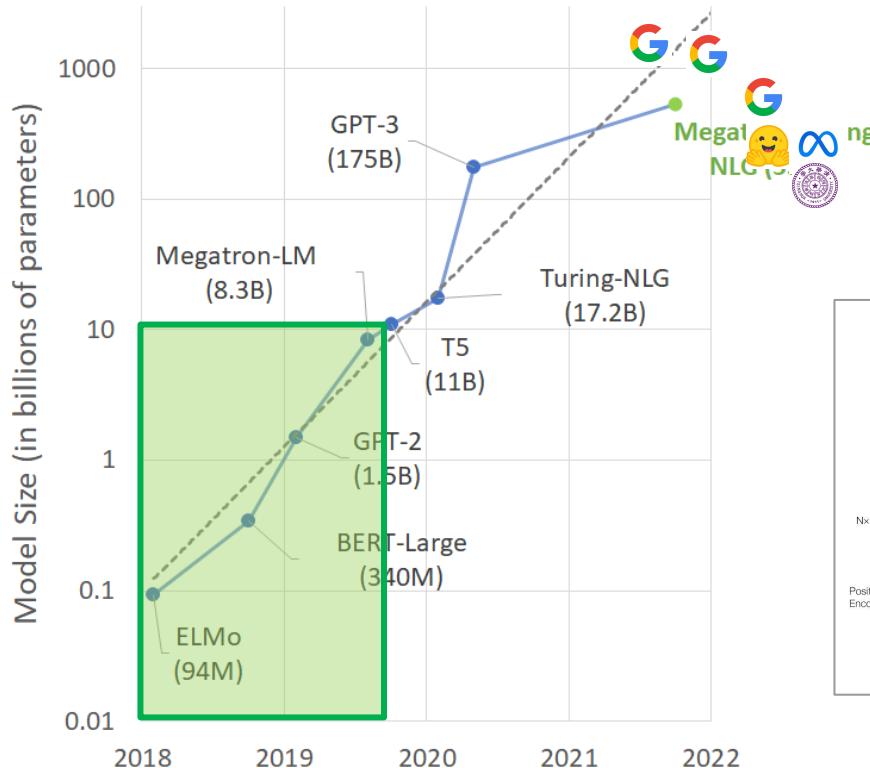


[Looking back] Usage (< 10B): finetuning

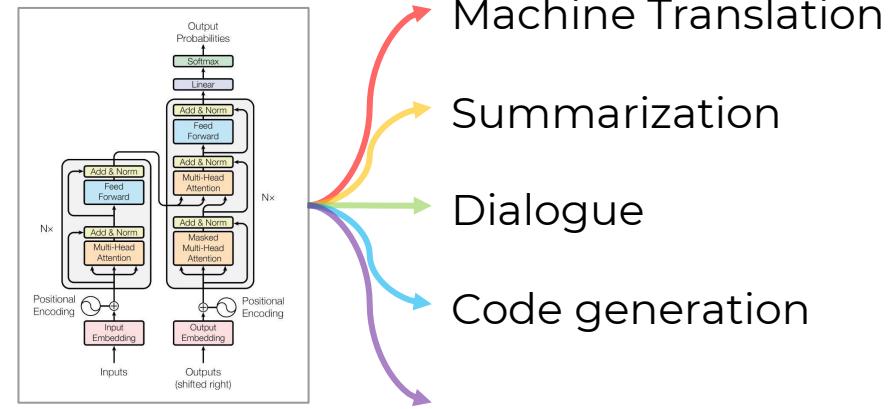


Finetuning:
LLM + downstream dataset

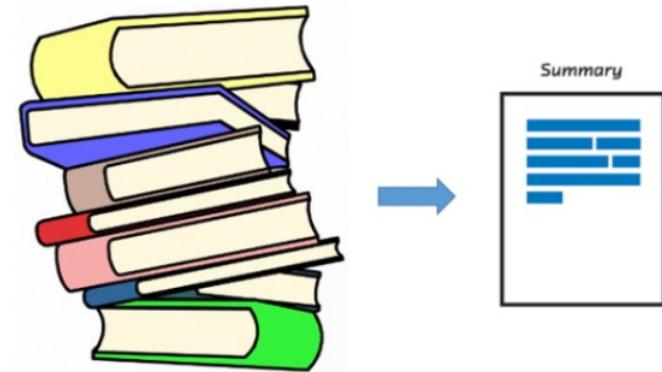
[Looking back] Usage (< 10B): finetuning



Finetuning:
LLM + downstream dataset



Applications, services, startups, ...



~~at~~ → to

The preposition **at** after the verb
be incorrect. Consider changing the
preposition.

Applications, services, startups, ...

解説付き自動採点

採点根拠の明示と追加学習項目のアドバイス

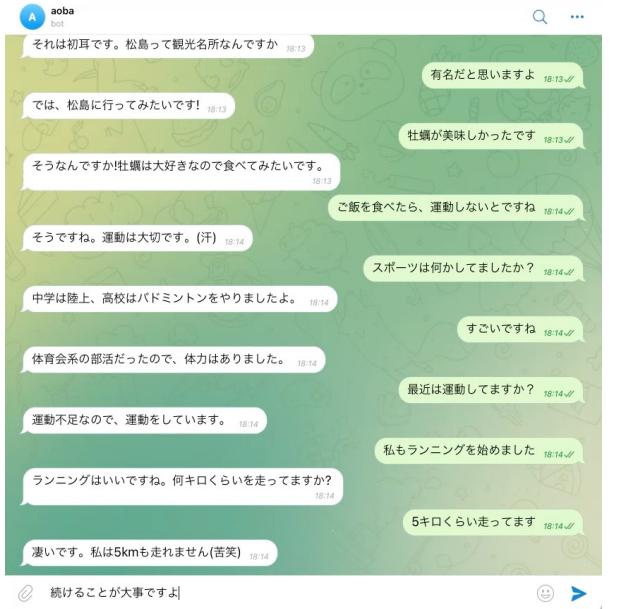
△ sinceは既知の事実に基づく説明に使います

She spoke Japanese fluently since she had lived in Japan for thirty years.

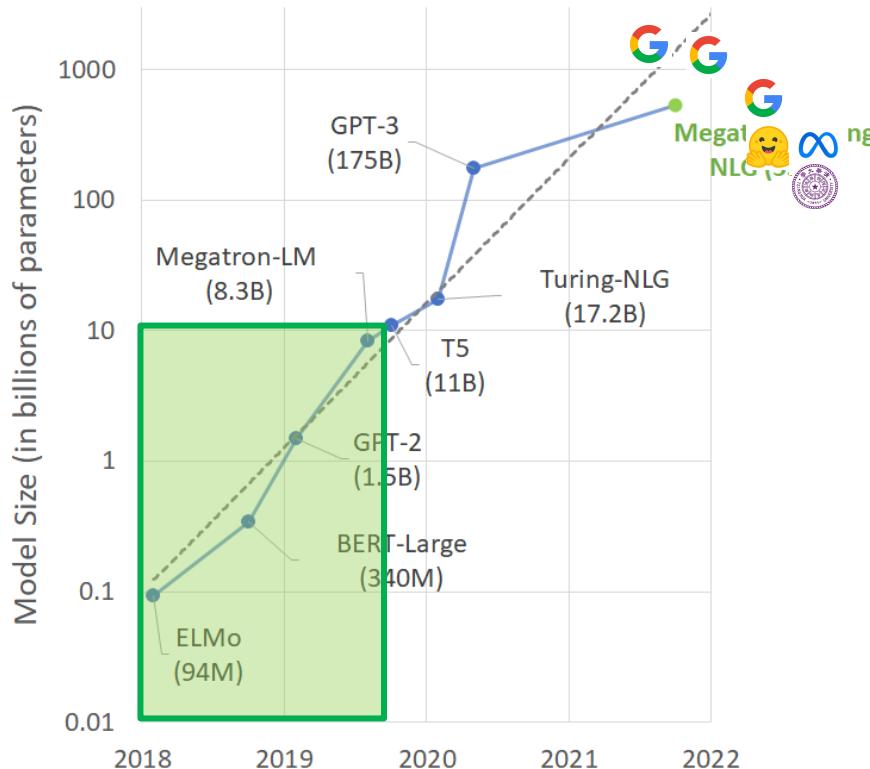
She spoke Japanese very fluently because she has lived there for 30 years.

✗ 時制をそろえましょう

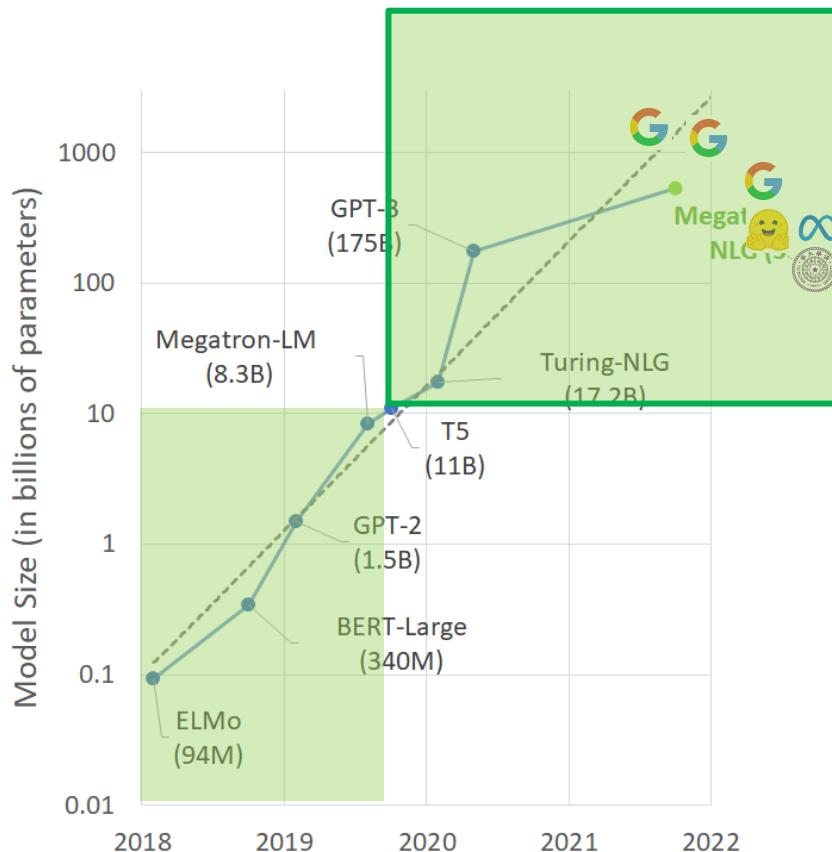
✗ 場所を正しく記述しましょう



[Looking back] Usage (< 10B): finetuning



Usage (> 10B): prompting (w/ API)



Prompting:
LLM + few/zero shot examples

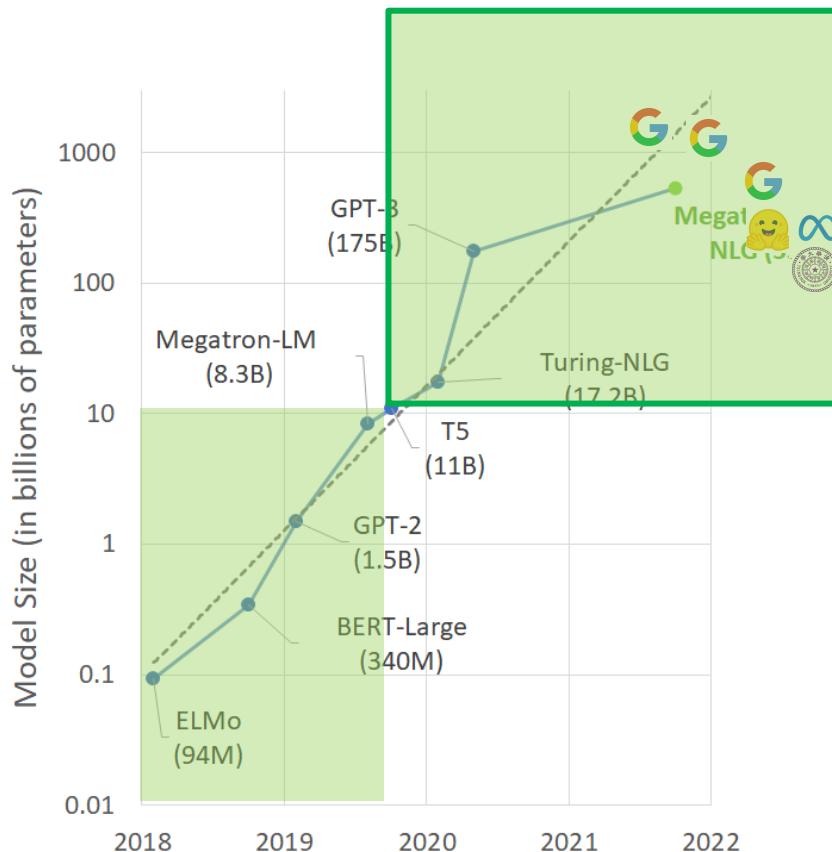
 Overview Documentation Examples [Playground](#)

Playground

Translate the following sentence into Japanese.
"Deep learning is a subfield of machine learning."

Deep learningは、機械学習の一つの分野です。

Usage (> 10B): prompting (w/ API)



Prompting:
LLM + few/zero shot examples

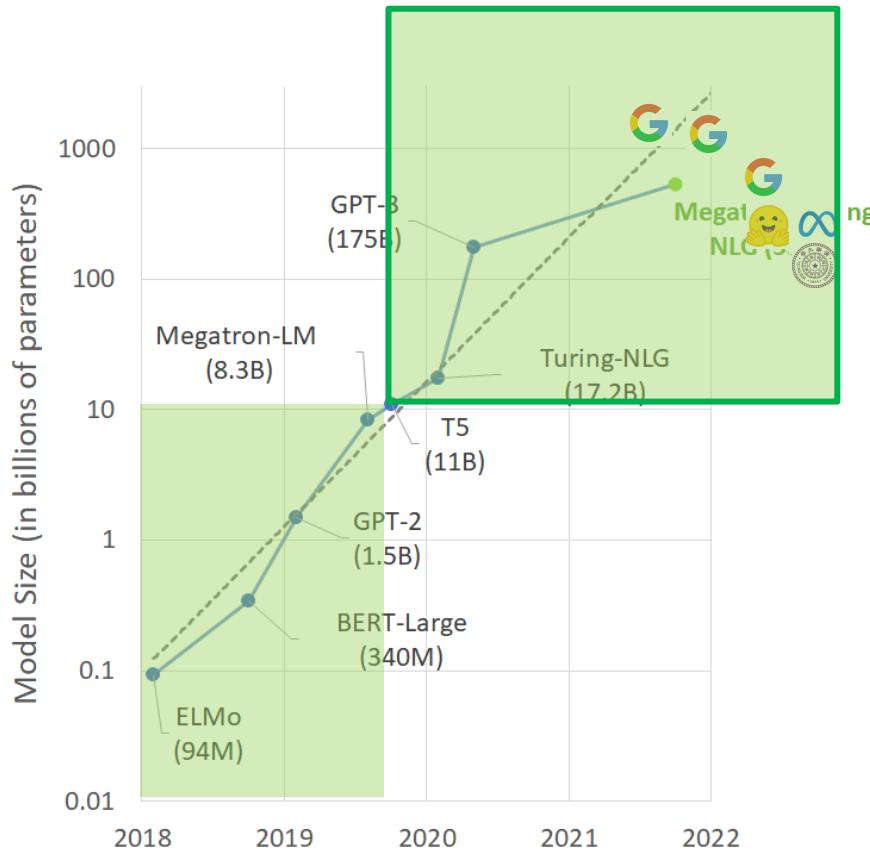
Overview Documentation Examples [Playground](#)

Playground

Input: 吾輩は熊である。名前はまだない。
Output: I'm a bear. I don't have a name yet.

Input: 私はその人を常に先生と呼んでいた。
Output: I always called that person "teacher".

Usage (> 10B): prompting (w/ API)



Prompting:
LLM + few/zero shot examples

Overview Documentation Examples **Playground**

Playground

次の文を英訳せよ。

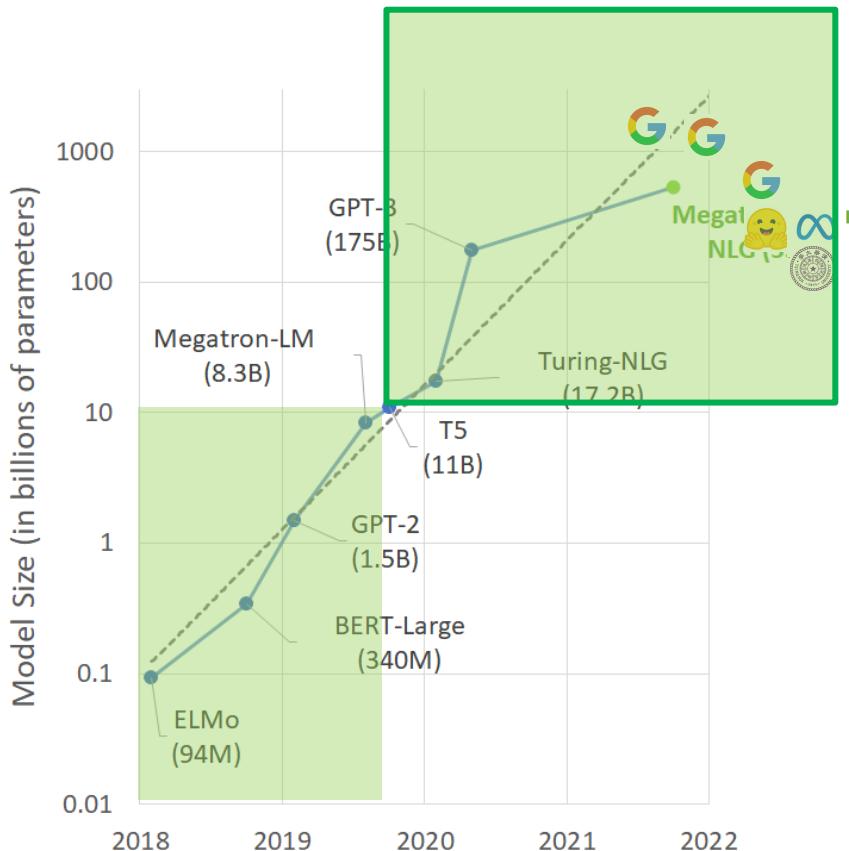
「吾輩は熊である。名前はまだない。」

I am a bear. I don't have a name yet.

中国語訳は？

我是一只熊，我还没有名字。

Usage (> 10B): prompting (w/ API)



Prompting:
LLM + few/zero shot examples

Overview Documentation Examples **Playground**

Playground

Proofread the following sentence.
"I'm looooking foward to see you."

"I'm looking forward to seeing you."

(simple) Symbol manipulation

Sort the following numbers in descending order:

[10, 1, 3, 7, 5, 100]

[100, 10, 7, 5, 3, 1]

(simple) Symbol manipulation

Sort the following numbers in descending order:
[10, 1, 3, 7, 5, 100]

[100, 10, 7, 5, 3, 1]

Input: [a, b, c, d, e]
Output: [e, d, c, b, a]

Input: [q, r, s, t, u, v, w, x]
Output: [x, w, v, u, t, s, r, q]

(simple) Symbol manipulation

Sort the following numbers in descending order:

[10, 1, 3, 7, 5, 100]

[100, 10, 7, 5, 3, 1]

[10, 1, 3, 7, 5, 100]



Sort the above numbers in descending order.

100, 7, 5, 3, 1



(simple) Symbol manipulation

Sort the following numbers in descending order:

[10, 1, 3, 7, 5, 100]

[100, 10, 7, 5, 3, 1]

[10, 1, 3, 7, 5, 100]

Sort the above numbers in descending order.

100, 7, 5, 3, 1 

Writing good prompts
is crucial.

Prompt engineering



The ChatGPT Prompt Book

Alan D. Thompson
December 2022
LifeArchitect.ai
Rev 3 (20221227)

Permanent link: <https://lifearchitect.ai/chatgpt-prompt-book/>

Prompt engineering



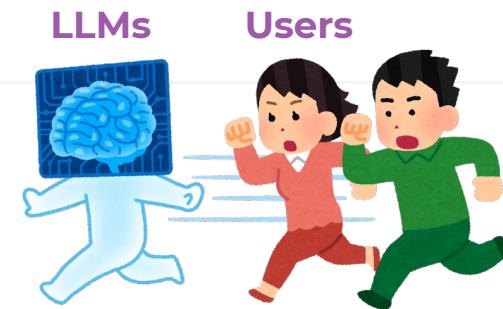
Najoung Kim 😊
@najoungkim

...

prompting is when you are the one who's being finetuned instead of the model

11:00 PM · Aug 6, 2022 · Twitter for Android

59 Retweets 2 Quote Tweets 560 Likes



(Big)Data + LLMs → Applications!



Are we all done? What's missing?



Large Language Models (LLMs)



What's LLM?



What's missing?

Large Language Models (LLMs)



What's LLM?



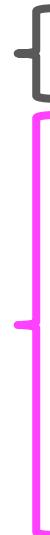
What's missing?

1. ???
2. ???

#1. Common Sense (implicit knowledge)

Explicitly written knowledge
(e.g., Wikipedia, News, ...)

There is a lot of knowledge that
is not explicitly written in text.
(e.g., common sense, unspoken
agreement)



Commonsense Knowledge/Reasoning

Winograd Schema Challenge (Levesque et al. 2011)

The **trophy** would not fit in the brown **suitcase** because **it** was too ***big***.

The **trophy** would not fit in the brown **suitcase** because **it** was too ***small***.

Commonsense Knowledge/Reasoning

The trophy would not fit in the brown suitcase because it was too *big*.

trophy

The trophy would not fit in the brown suitcase because it was too *small*.

suitcase

Commonsense Knowledge/Reasoning

The trophy would not fit in the brown suitcase because it was too *big*.

trophy

The trophy would not fit in the brown suitcase because it was too *small*.

suitcase

Winograd Schema Challenge (Levesque et al. 2011)

273 examples

Commonsense Knowledge/Reasoning

The trophy would not fit in the brown suitcase because it was too *big*.

trophy

The trophy would not fit in the brown suitcase because it was too *small*.

suitcase

Winograd Schema Challenge (Levesque et al. 2011)

273 examples

→ WinoGrande (Sakaguchi et al. 2020)

43,972 examples

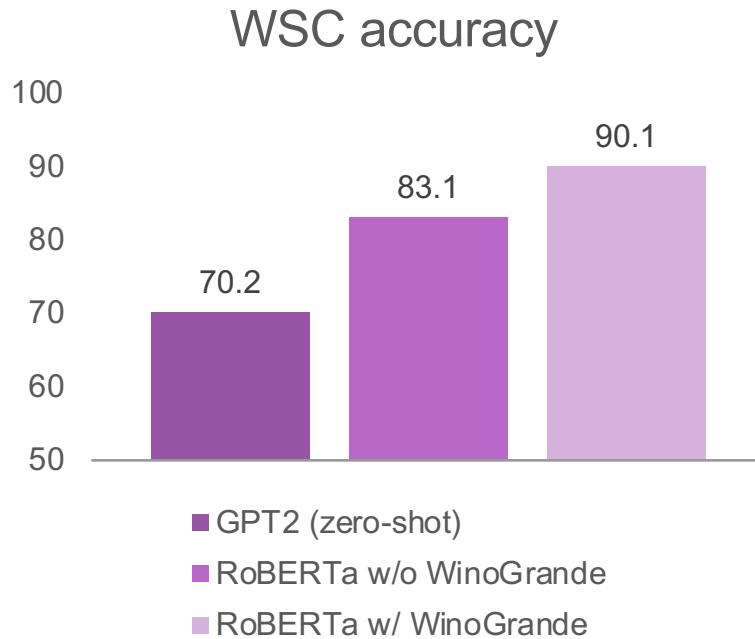


WinoGrande (Sakaguchi et al. 2020)

<https://winogrande.allenai.org/>

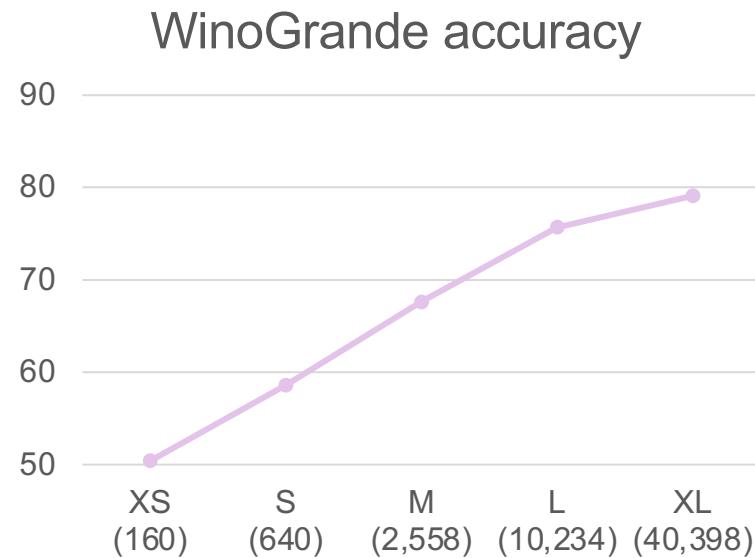
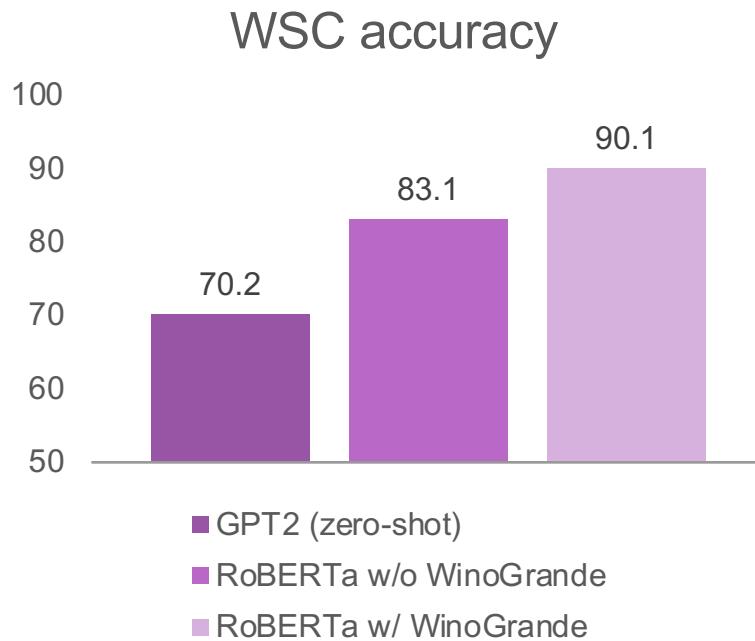
WinoGrande (Sakaguchi et al. 2020)

<https://winogrande.allenai.org/>



WinoGrande (Sakaguchi et al. 2020)

<https://winogrande.allenai.org/>



“explication” looks working great, but ...

K

Ryan had to use a calculator on the test but Michael didn't because he was very bad at math.

Question: Who was bad at math?



According to the information provided, Michael was bad at math.

“explication” looks working great, but ...

Difficult to distinguish what can be done and what can't be done, and why.

K

Ryan had to use a calculator on the test but Michael didn't because he was very bad at math.

Question: Who was bad at math?



According to the information provided, Michael was bad at math.

“explication” looks working great, but ...

Difficult to distinguish what can be done and what can't be done, and why.

K

北条政子の生涯について教えてください。



北条政子は、日本の歌手、女優、タレントです。

1957年、東京都生まれました。高校卒業後、東京芸術大学に進学し、演劇を専攻しました。その後、舞台女優として活躍します。

1980年代になると、テレビ番組やCMに出演するようになります。また、歌手としても活動を開始し、数多くのヒット曲を生み出します。

北条政子は、日本の芸能界で長いキャリアを持ち、幅広い層に愛される人気女優です。現在も、テレビ番組や舞台で活躍しています。

Multi-modal commonsense

Recent text-to-image models also struggle with common sense.
(e.g., Stable Diffusion, Midjourney)

Cars on the highway



Large Language Models (LLMs)



What's LLM?



What's missing?

1. Common Sense
2. ???

#2 LLMs are too costly to train&update



It takes long time to (pre)train LLMs
(e.g., 1 ~2 month with 1024 x A100)



GPUs cost a lot of money
(e.g., \$10-20M)

LLMs become out of date (dep. on data)

[Overview](#)[Documentation](#)[Examples](#)[Playground](#)

Playground

Who is the current US president?

The current US president is Donald Trump.

LLMs become out of date (dep. on data)

K

Who is the prime minister in Japan now?



As of September 2021, the current prime minister of Japan is Shinzo Abe.

LLMs become out of date (dep. on data)

K

Which teams faced off in the 2022 World Cup final?



I'm sorry, but my knowledge cutoff is September 2021, and I don't have information about the teams that faced off in the 2022 World Cup final. The tournament has not been held yet.

#2 LLMs are too costly to train&update

[Overview](#)[Documentation](#)[Examples](#)[Playground](#)

Playground

How often should I update GPT3?

You should update GPT3 at least once a week.|



REALTIME QA

(Kasai, Sakaguchi et al. 2022)

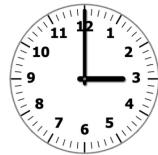
<https://realtimeqa.github.io/>



REALTIME QA

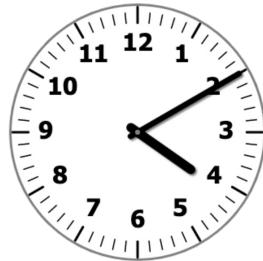
(Kasai, Sakaguchi et al. 2022)

<https://realtimeqa.github.io/>



Q: How many home runs has Shohei Ohtani hit?

A: 24



Q: How many home runs has Shohei Ohtani hit?

A: 25



REALTIME QA: (potential) applications

Emergency Management:

- COVID-19, Earthquake, Flood





REALTIME QA: (potential) applications

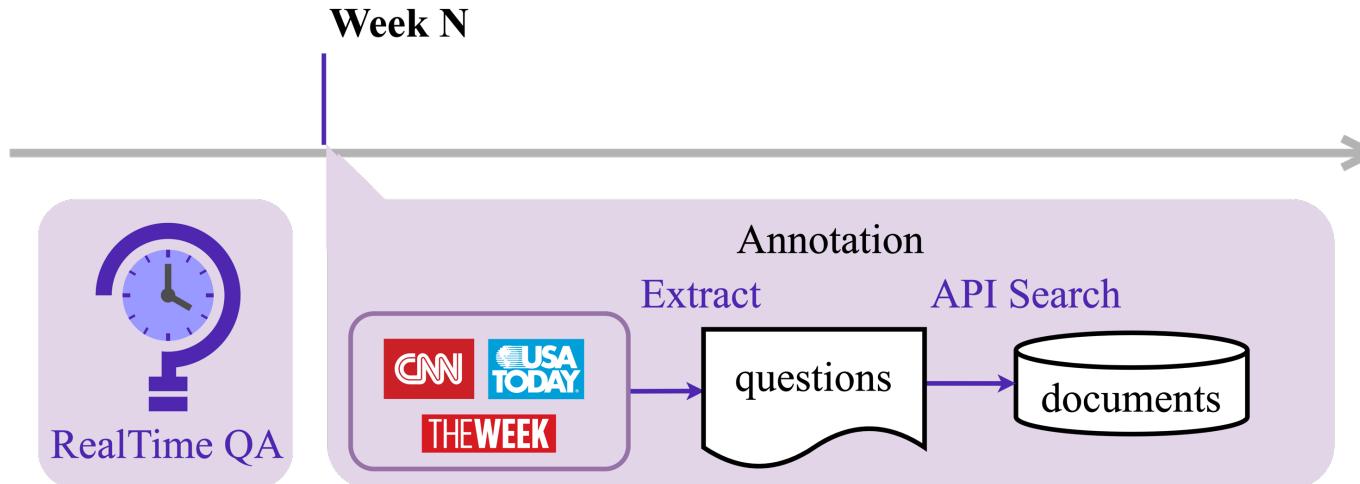
Emergency Management:

- COVID-19, Earthquake, Flood
- Financial news



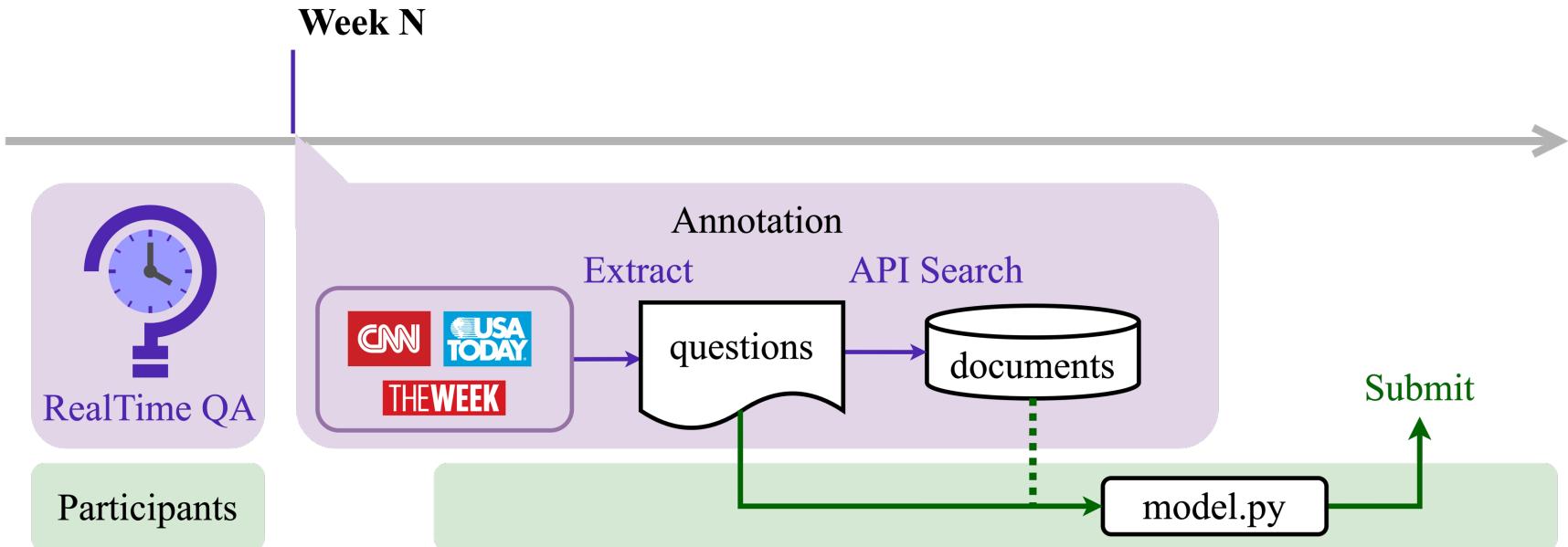


REALTIME QA: How it works



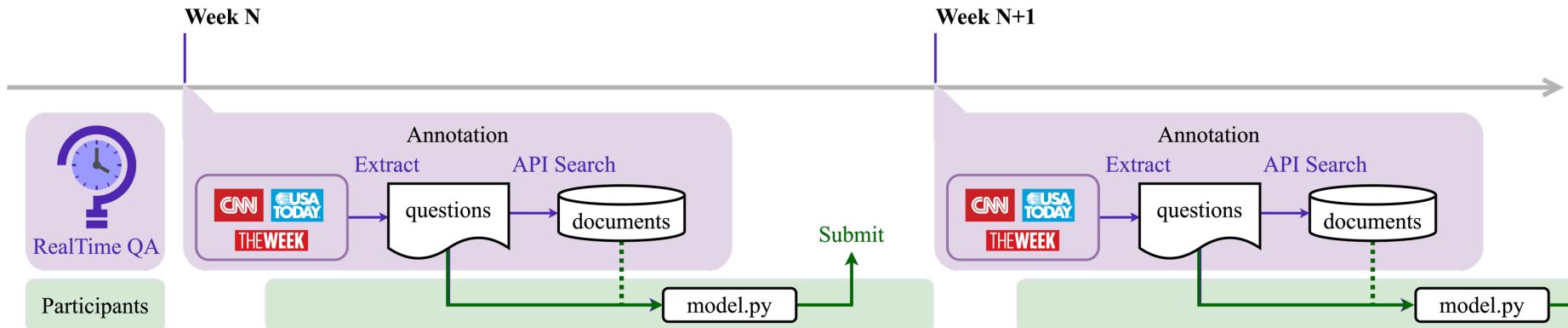


REALTIME QA: How it works





REALTIME QA: How it works



RealTime QA: Example Questions



5 Things Weekly News Quiz



April 1, 2022

China is battling its biggest Covid-19 wave yet. Which large city there implemented mass coronavirus testing this week?

Shenzhen

Wuhan

Tianjin

Beijing



Twitter agreed to be bought by Tesla and SpaceX CEO Elon Musk this week. How much is the deal worth?

\$4.4 million

\$440 million

\$4.4 billion

\$44 billion



THE WEEK

Quiz of The Week



May 18, 2022

What is the name of the new tank being deployed by Russia in the battle for control of Donbas in Ukraine?

Rambo

Terminator

Hell Drivers

Viking



RealTime QA: Example Questions



🕒 April 1, 2022

China is battling its biggest Covid-19 wave yet. Which large city there implemented mass coronavirus testing this week?

Shenzhen

Wuhan

Tianjin

Beijing



Twitter agreed to be bought by Tesla and SpaceX CEO Elon Musk this week. How much is the deal worth?

\$4.4 million

\$440 million

\$4.4 billion

\$44 billion



THE WEEK Quiz of The Week 🕒 May 18, 2022

What is the name of the new tank being deployed by Russia in the battle for control of Donbas in Ukraine?

Rambo

Terminator

Hell Drivers

Viking



Realtime QA has three quiz formats.

1. Multiple choice question, 2. MC with “None of the above”, 3. Generation



REALTIME QA : Baseline models

Closed-book

Open-book



REALTIME QA : Baseline models

Closed-book

Question: Today is June 24, 2022. Covid-19 vaccinations in the US began for which age group this week?

Answer: Children under 5

Open-book



REALTIME QA : Baseline models

Closed-book

Question: Today is June 24, 2022. Covid-19 vaccinations in the US began for which age group this week?

Answer: Children under 5

predict

Open-book



REALTIME QA : Baseline models

Closed-book

Question: Today is June 24, 2022. Covid-19 vaccinations in the US began for which age group this week?

Answer: Children under 5



Open-book

Article on November 2, 2021: CDC recommends Pfizer COVID-19 vaccine for kids 5-11, shots expected to roll out this week...

Article on June 23, 2022: Covid-19 vaccinations begin for US children under 5...

...

retrieve

Question: Today is June 24, 2022. Covid-19 vaccinations in the US began for which age group this week?

Answer: Children under 5



REALTIME QA : Baseline models

Closed-book

Question: Today is June 24, 2022. Covid-19 vaccinations in the US began for which age group this week?

Answer: Children under 5



Open-book

Article on November 2, 2021: CDC recommends Pfizer COVID-19 vaccine for kids 5-11, shots expected to roll out this week...
Article on June 23, 2022: Covid-19 vaccinations begin for US children under 5...
...

retrieve

Question: Today is June 24, 2022. Covid-19 vaccinations in the US began for which age group this week?

Answer: Children under 5

predict



REALTIME QA : Baseline results

	(Multiple-Choice)		(Generation)	
	Acc	Acc. w/NOTA	EM	F1
Closed-book ----- GPT3				
Open-book wiki GPT3				
G (GCS GPT3)				



REALTIME QA : Baseline results

	(Multiple-Choice)		(Generation)	
	Acc	Acc. w/NOTA	EM	F1
Closed-book (----- GPT3)	39.7	31.3	7.3	15.2
Open-book (wiki GPT3)				
Open-book G (GCS GPT3)				



REALTIME QA : Baseline results

	(Multiple-Choice)		(Generation)	
	Acc	Acc. w/NOTA	EM	F1
Closed-book (- - - GPT3)	39.7	31.3	7.3	15.2
Open-book (wiki GPT3)	45.8	34.1	8.4	14.6
GCS (GCS GPT3)				



REALTIME QA : Baseline results

	(Multiple-Choice)		(Generation)	
	Acc	Acc. w/NOTA	EM	F1
Closed-book (- - - GPT3)	39.7	31.3	7.3	15.2
Open-book (wiki GPT3)	45.8	34.1	8.4	14.6
GCS (GCS GPT3)	69.3	59.8	28.7	39.4



REALTIME QA : Baseline results

	(Multiple-Choice)		(Generation)	
	Acc	Acc. w/NOTA	EM	F1
Closed-book (- - - GPT3)	39.7	31.3	7.3	15.2
Open-book (wiki GPT3)	45.8	34.1	8.4	14.6
GCS GPT3	69.3	59.8	28.7	39.4

Realtime retrieval will become more important.



REALTIME QA : dynamic benchmark

[Results](#) / [2022](#) / [2022/09/02](#)

Results on September 02, 2022

Multiple Choice Track

Model	Submission Time (GMT)	Original	NOTA
GPT-3 + Google Custom Search	2022-09-03 03:00:00	76.7	70.0
T5 (Closed-Book)	2022-09-03 03:00:00	56.7	46.7
GPT-3 + DPR	2022-09-03 03:00:00	56.7	43.3
RAG + Google Custom Search	2022-09-03 03:00:00	53.3	46.7
GPT-3 (Closed-Book)	2022-09-03 03:00:00	43.3	33.3
RAG + DPR	2022-09-03 03:00:00	40.0	33.3

Generation Track

Model	Submission Time (GMT)	EM	F1
GPT-3 + Google Custom Search	2022-09-03 03:00:00	23.3	42.5
RAG + Google Custom Search	2022-09-03 03:00:00	13.3	25.5
GPT-3 + DPR	2022-09-03 03:00:00	6.7	19.9
T5 (Closed-Book)	2022-09-03 03:00:00	6.7	15.5
RAG + DPR	2022-09-03 03:00:00	6.7	8.3
GPT-3 (Closed-Book)	2022-09-03 03:00:00	3.3	18.1

New test set is released every Friday.

We welcome your submissions!!

Large Language Models (LLMs)



What's LLM?



What's missing?

1. Common Sense
2. Realtime Context

Large Language Models (LLMs)



What's LLM?



What's missing?

1. Common Sense
2. Realtime Context

Grounding to the realworld
(including multi-modality)

Things are moving fast!



LangChain

Chat GPT



Perplexity

2022/09

2022/10~11

2022/12

**Large Language Models:
What will happen next?**

坂口慶祐 (東北大學)
keisuke.sakaguchi@tohoku.ac.jp



GPT Index
@gpt_index

 YouChat

👋 Hello! My name is YouChat, I'm an AI that can
summarize text, compose emails, and write code

e.g., Perplexity.ai (GPT + retrieval)

Perplexity

Copy Share ⚙ + New Thread

Which teams faced off in the World Cup final 2022?

PERPLEXITY View Detailed

The 2022 FIFA World Cup final was contested by Argentina and defending champions France^[1]
^[2]^[3]. The match was played at Lusail Stadium in Lusail, Qatar on 18 December 2022^[4]^[5].

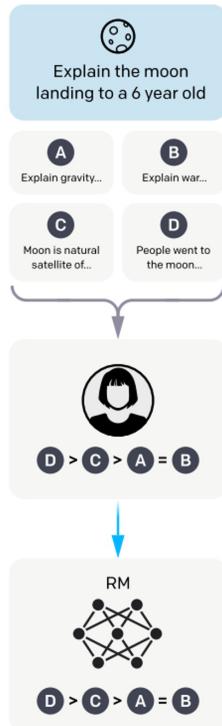
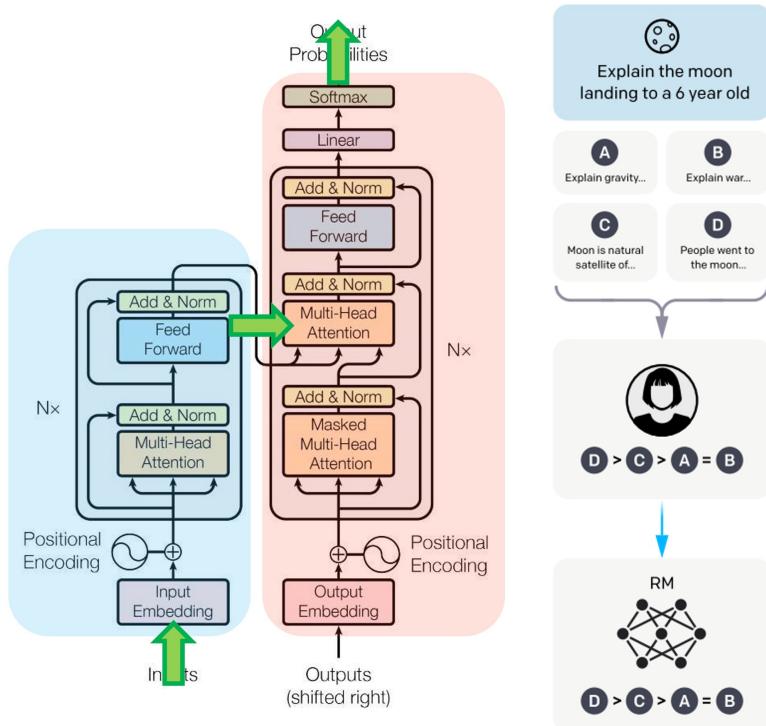
Accurate Inaccurate

5 SOURCES View List

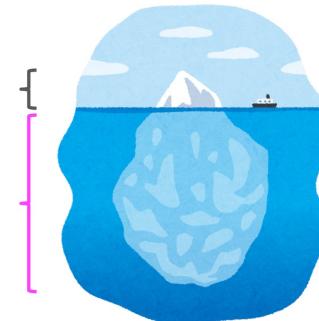
- 1 | W wikipedia.org
- 2 | 🌐 skysports.com
- 3 | SN sportingnews.com
- 4 | F fifa.com
- 5 | CBSSports cbssports.com

Large Language Models (LLMs)

What's LLM?



What's next?



Q: How many home runs has Shohei Ohtani hit?
A: 24



Q: How many home runs has Shohei Ohtani hit?
A: 25

LLMs from global point of view (130B+)



Google

Meta



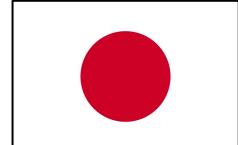
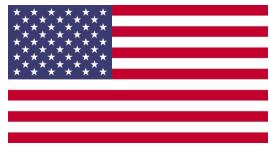
HUGGING FACE



北京智源人工智能研究院

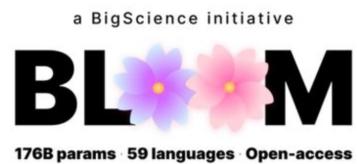
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

LLMs from global point of view (130B+)



Google

Meta



HUGGING FACE



北京智源人工智能研究院
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE



なぜ、わざわざ日本(語)で？



日本語固有の問題

大規模モデルにおける日本語のクオリティはまだまだ問題が多い。日本語に強いモデルを作ることで、日本にも大規模モデルの価値、イノベーションを提供したい。



海外依存のリスク

自前でコントロールできない。海外の都合で急に使えなくなるリスク。日本語の優先順位は低い。

(ガスや石油のパイプライン、ビザ発給と同様に国際関係におけるカードになる。=国防。)



技術・人材のアセット

日本で取り組むことで、大規模言語モデル構築・運用のノウハウを国内でシェアすることができる。