Text Mining

Summer term 2024

Sandipan Sikdar

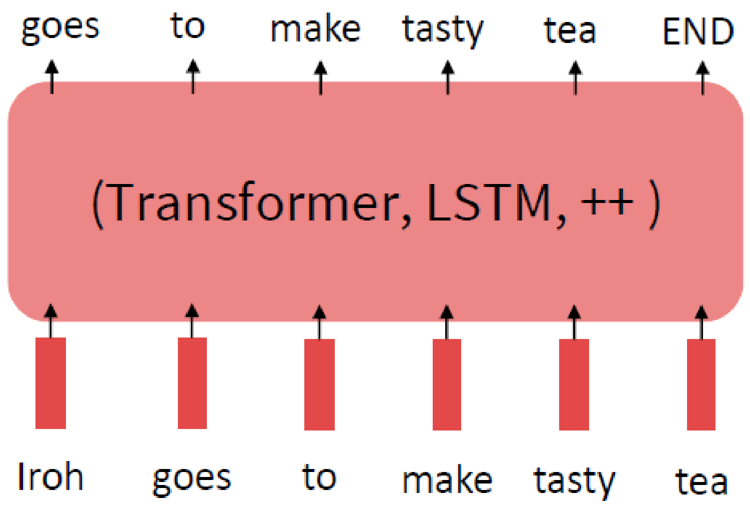
Leibniz Universität Hannover

Pretraining/Finetuning Paradigm

- Pretraining can improve NLP applications by serving as parameter initialization.

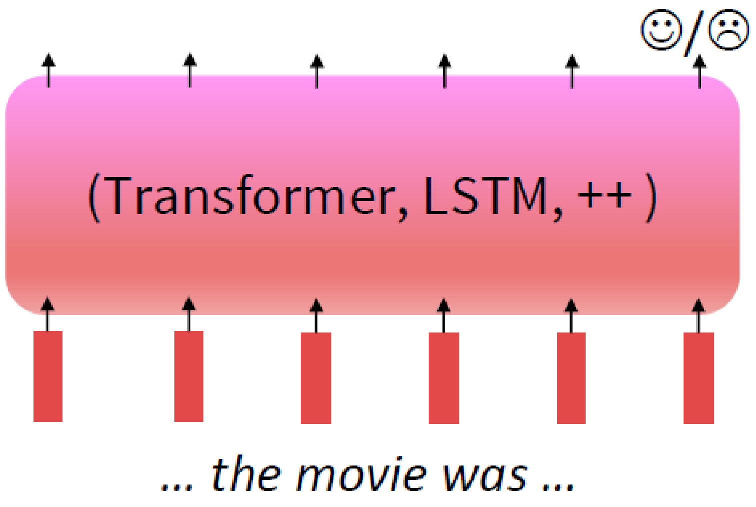
Step 1: Pretrain (on language modeling)

Lots of text; learn general things!

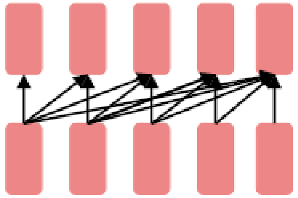


Step 2: Finetune (on your task)

Not many labels; adapt to the task!

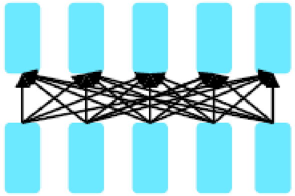


Model pretraining



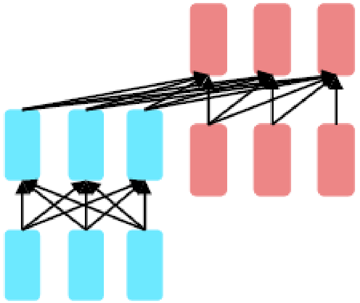
Decoders

- Language models!



Encoders

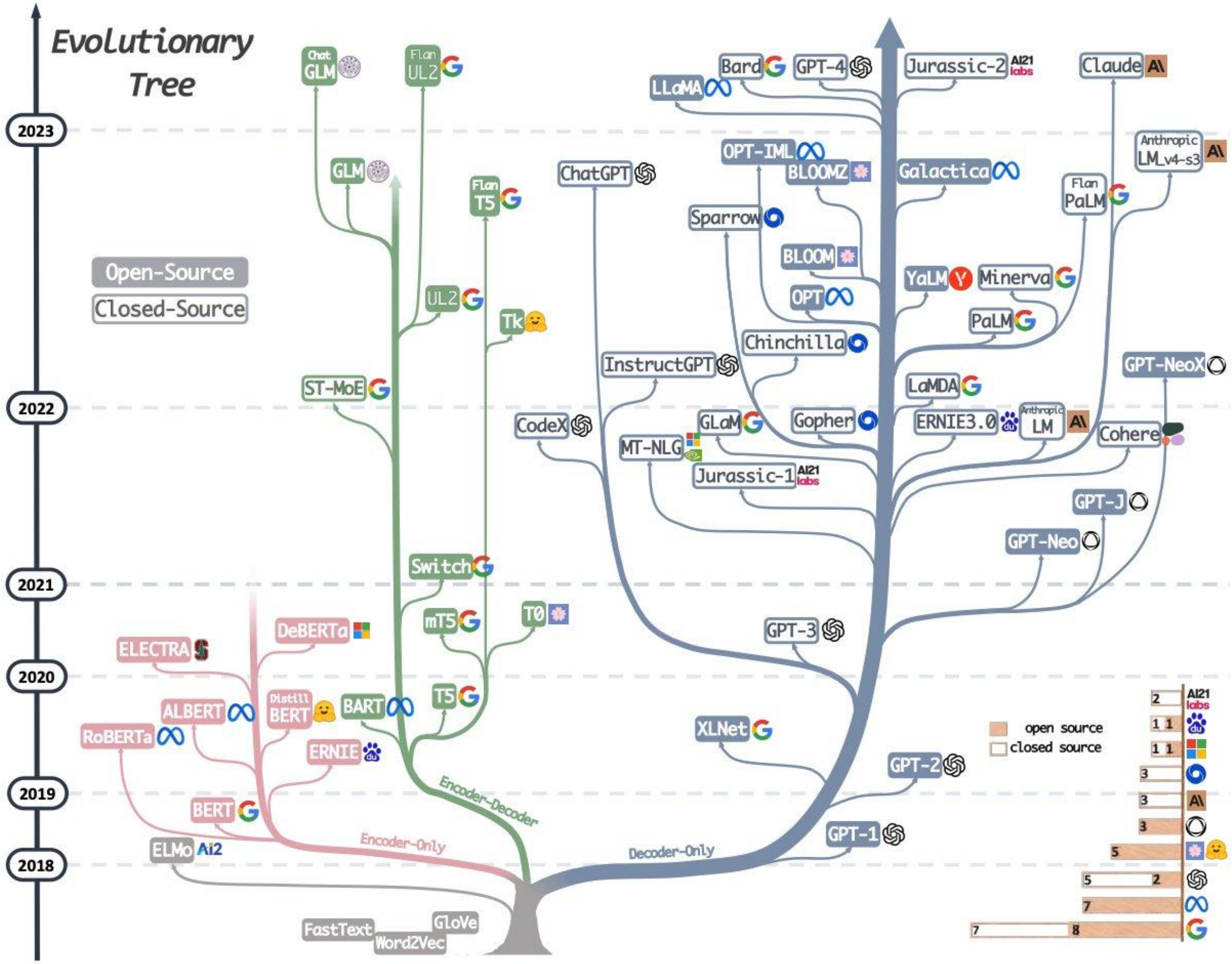
- Gets bidirectional context can condition on future!



Encoder-Decoders

- Fuse the good parts of both encoder and decoder

LLM zoo



Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
GLM
Flan T5
UL2
Tk
ST-MoE
Switch
T0
mT5
DeBERTa
ELECTRA
ALBERT
Distill BERT
BART
T5
RoBERTa
ERNIE
BERT
ELMo
Encoder-Decoder
Encoder-Only
Decoder-Only
FastText
GloVe
Word2Vec
Bard
GPT-4
Jurassic-2
Claude
LLaMA
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
ChatGPT
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source

Further advancements

Advancements

- Several advancements have been proposed to further improve performance and efficiency
- Encoder models
  - Sparse attention
- Decoder models
  - KV-cache
  - Multi-query, Group-Query and Latent query attention
  - Mixture of experts*

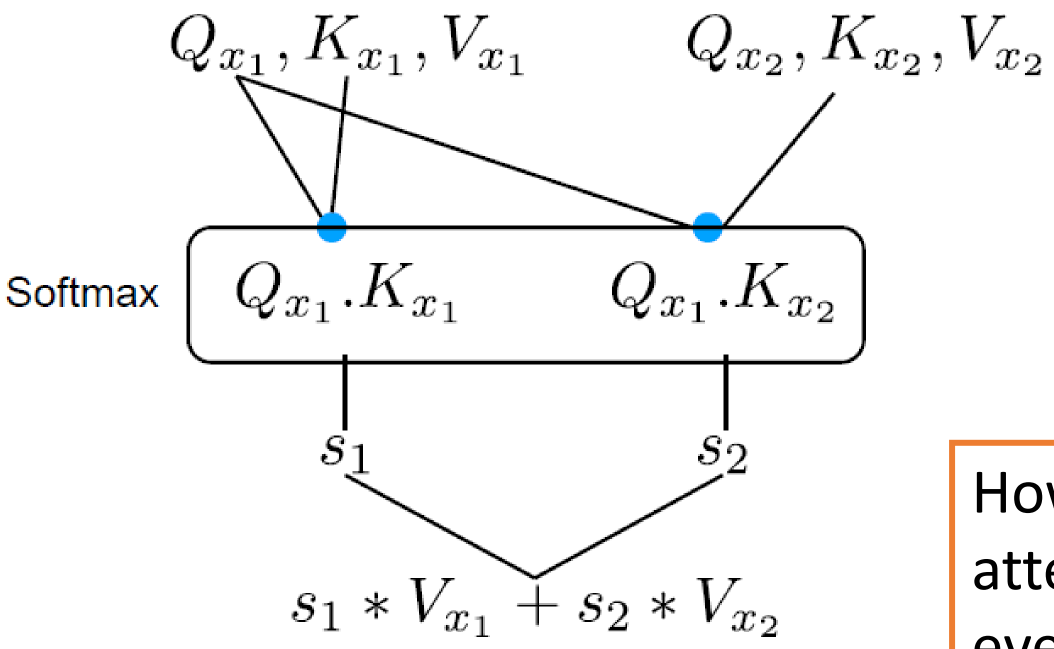
* Could also be deployed in encoder models

Sparse attention: Encoder Models

- BERT model can only process a sequence of length 512
- The self attention computation requires $O(n^2)$ inner product operations n is the length of the sequence
- Hence increasing the length of the sequence would lead to almost infeasible computational overhead
- Processing longer sequences is often required in many downstream tasks
- Models like BigBird are capable of processing longer sequences
- Instead of self attention, they deploy "sparse" attention

Sparse attention

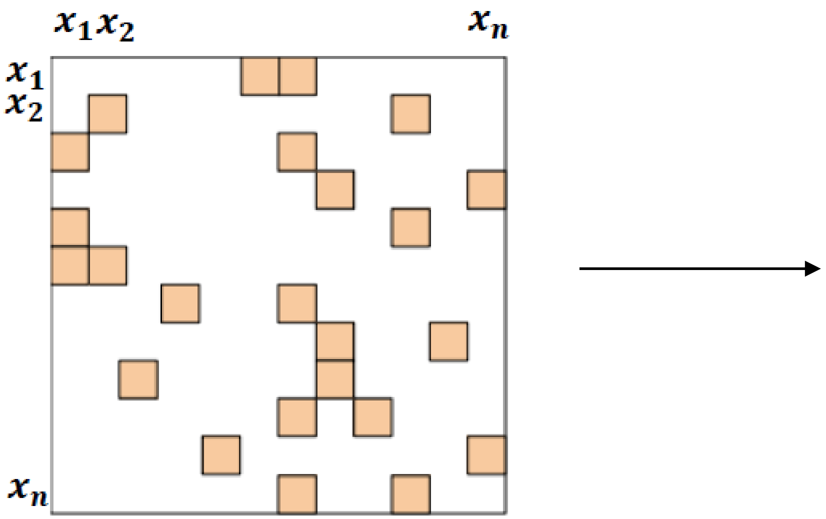
- Computing attention scores are exactly same as self-attention
- Computing attention of x_1 with respect to x_2



However, we won't compute attention of a token with respect to every other token

Sparse attention

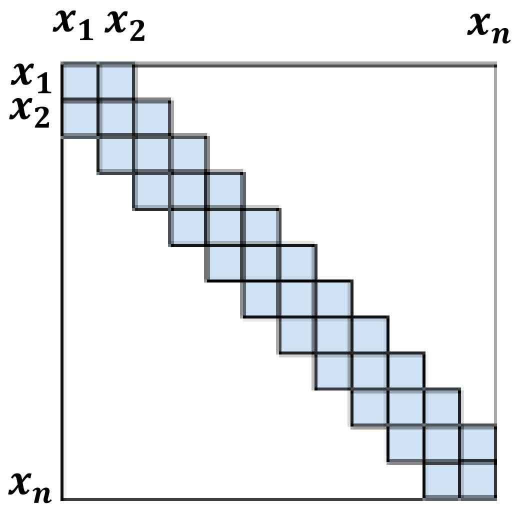
- Random attention



- White space implies absence of attention
- Randomly attend to r tokens in the sequence (here r=2)

Sparse attention

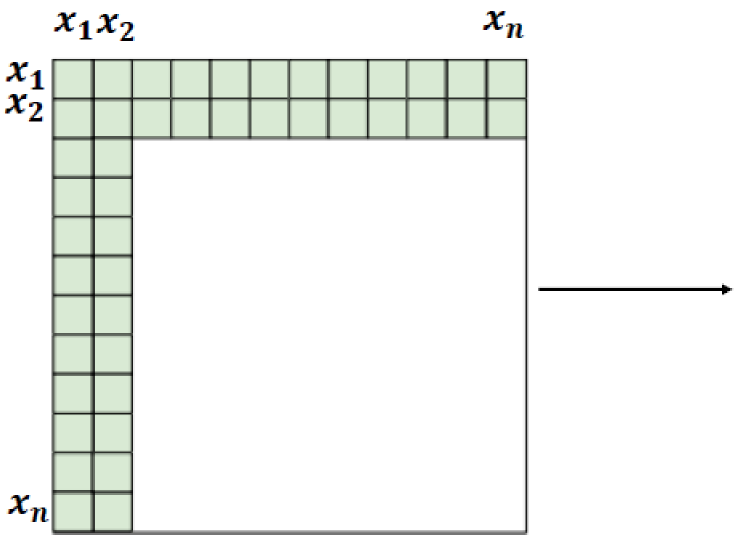
- Window attention



- White space implies absence of attention
- Attend to the local neighbors through sliding window (here w=3)
- $i + \left\lfloor \frac{w}{2} \right\rfloor, i - \left\lfloor \frac{w}{2} \right\rfloor$
- NLP tasks display “locality of reference”
- Self attention models in NLP tasks indicate that neighboring inner products are extremely important

Sparse attention

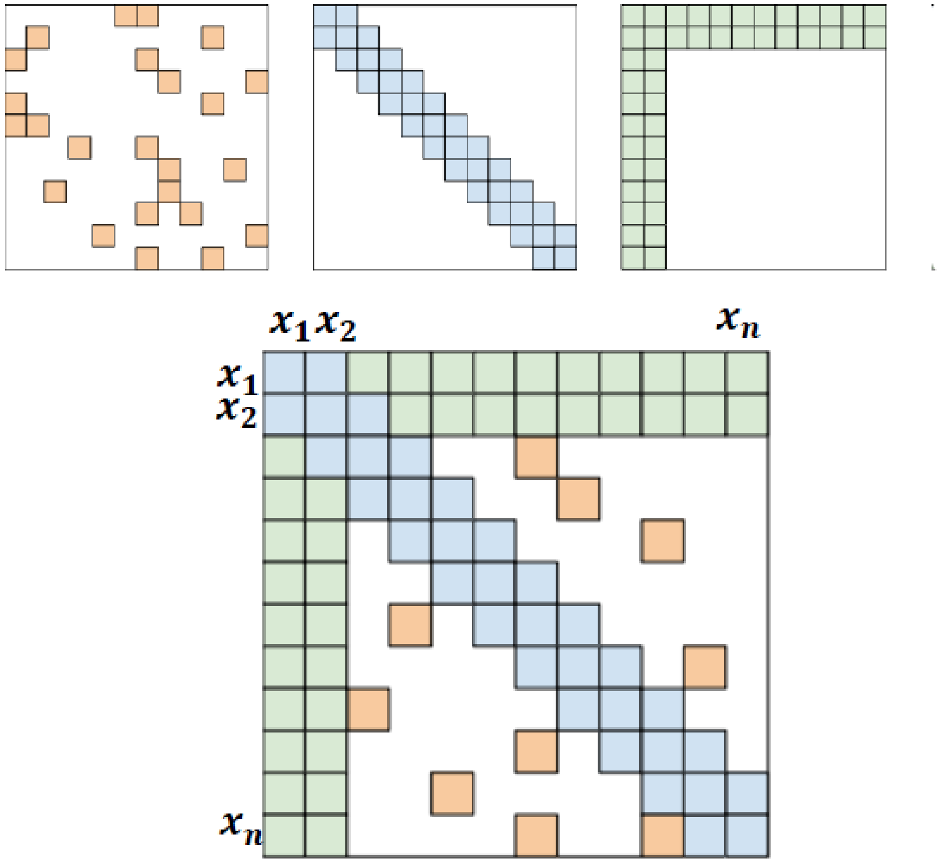
- Global attention



- White space implies absence of attention
- Select a few tokens as global tokens (g=2)
- Tokens that attend to all tokens in the sequence and to whom all tokens attend to
- The global tokens can be from existing tokens or extra added tokens

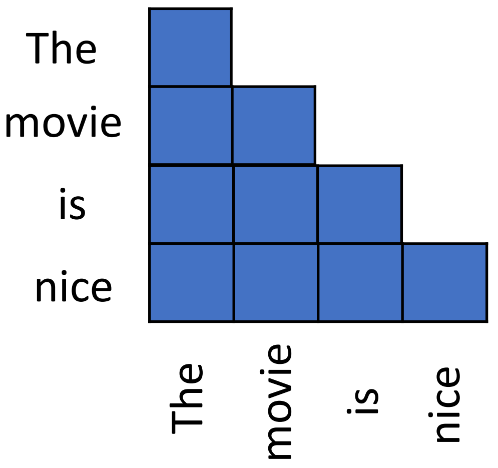
Sparse attention

- Putting it all together

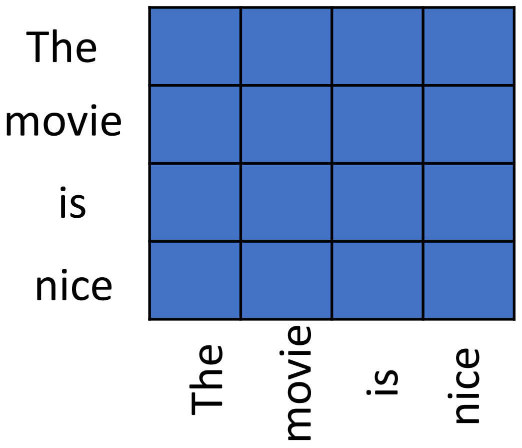


Masked Attention

- Decoder-only models deploy masked attention i.e., each token attends only to itself and previous tokens



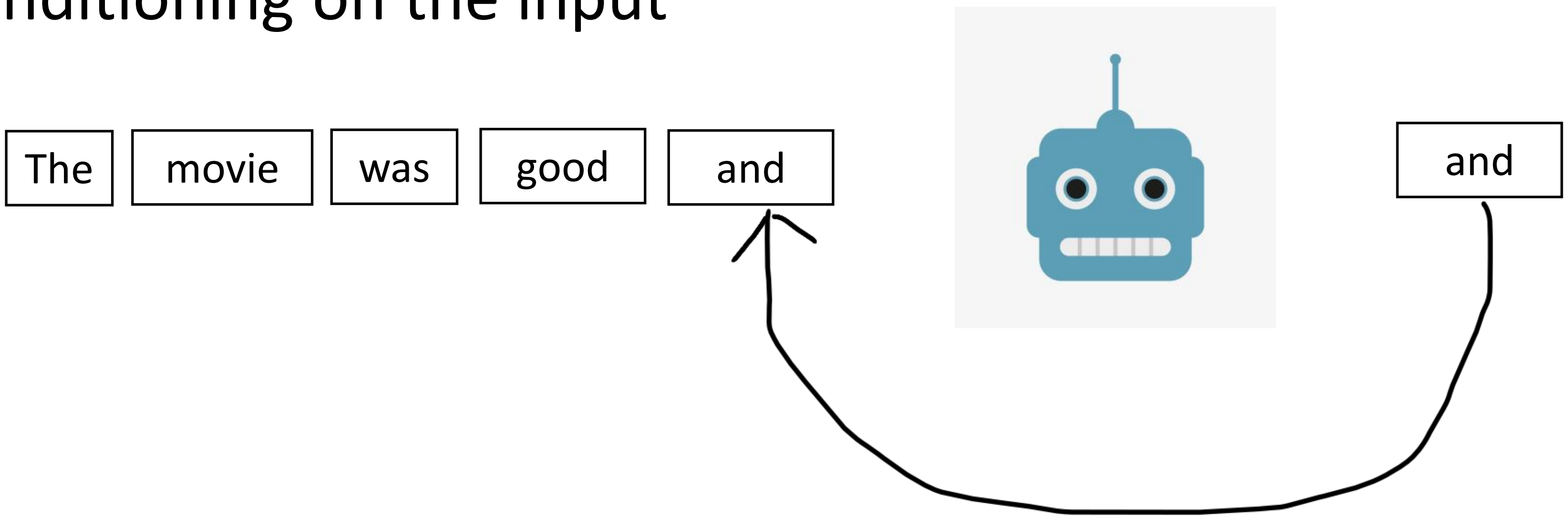
Masked attention



Self-attention

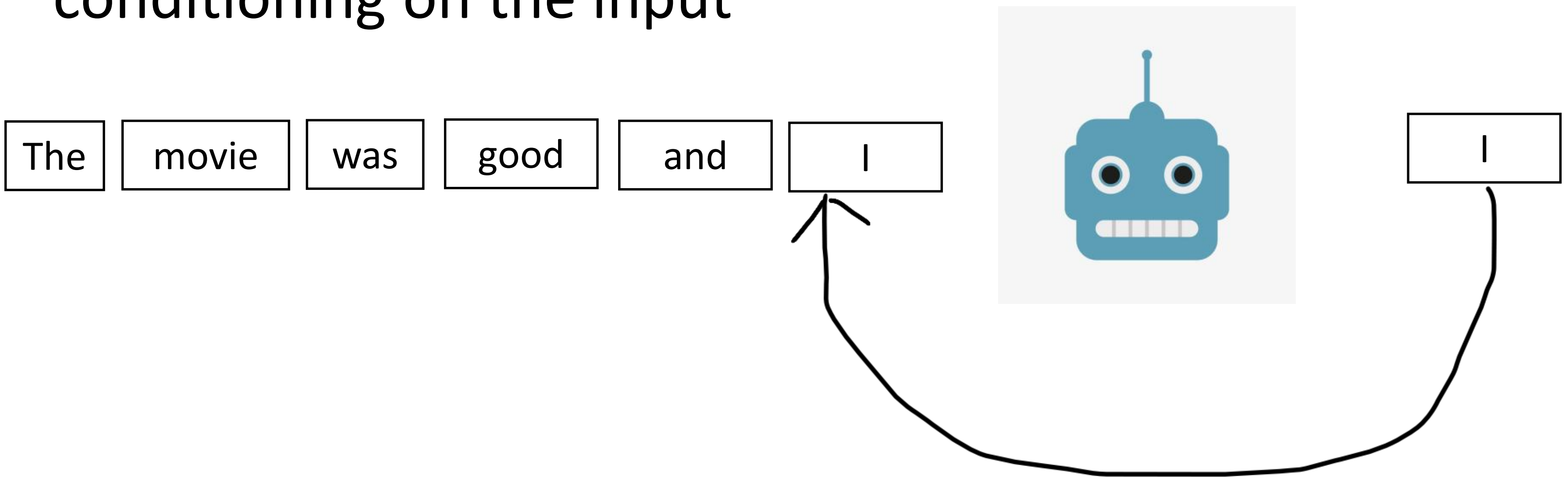
KV-Cache

- You can further speedup computation of mask attention
- Recall: For decoder-only models, we have an input text (also referred to as prompt) and the model generates one token at a time conditioning on the input



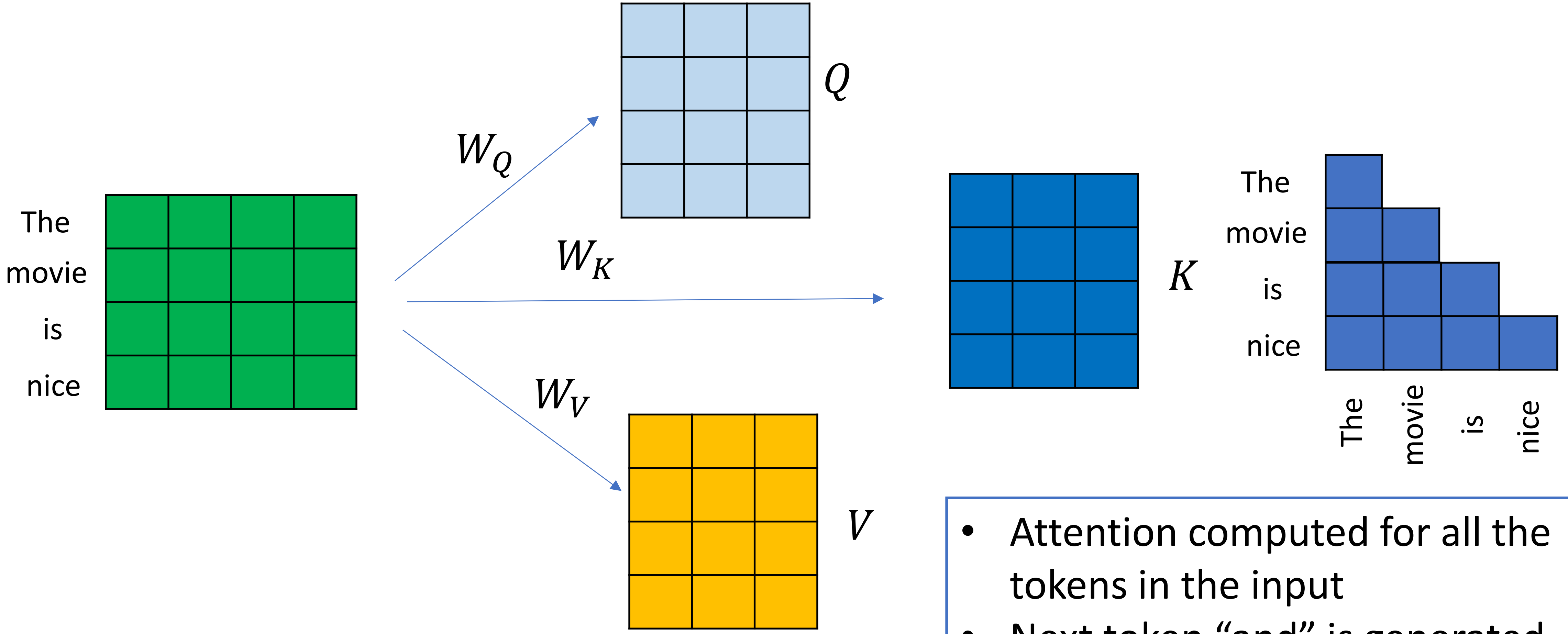
KV-Cache

- You can further speedup computation of mask attention
- Recall: For decoder-only models, we have an input text (also referred to as prompt) and the model generates one token at a time conditioning on the input



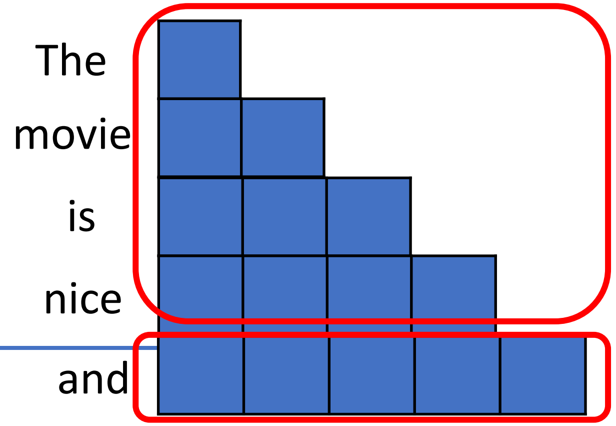
KV-Cache

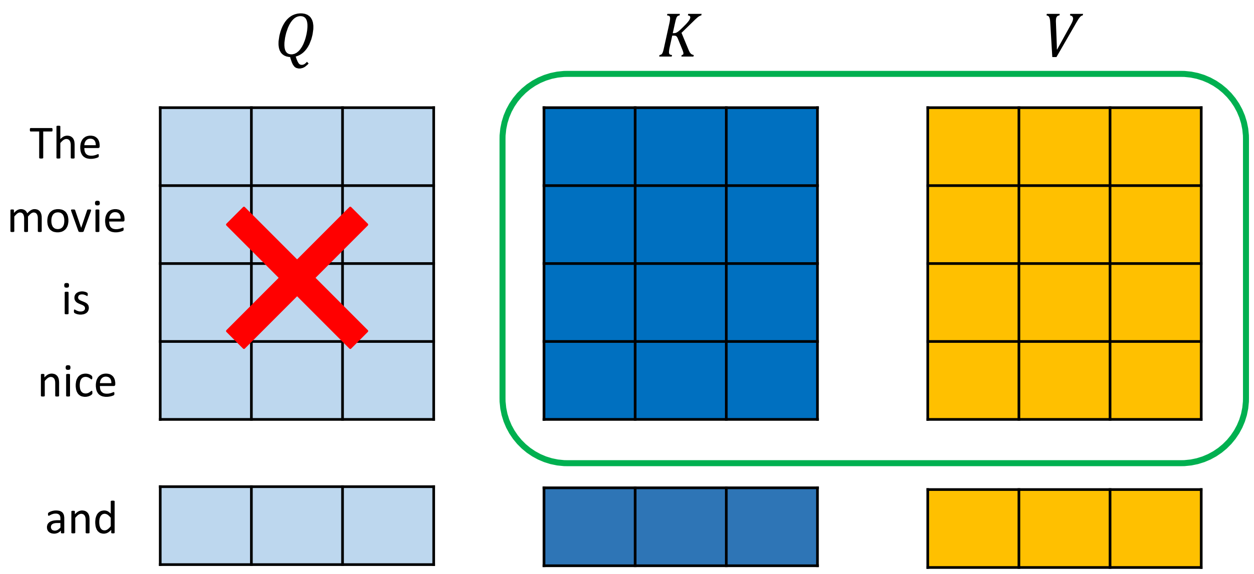
$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



- Attention computed for all the tokens in the input
- Next token “and” is generated

KV-Cache





- For the next token prediction, we need to compute attention as –

$$softmax\left(\frac{q_{\text{and}}K^T}{\sqrt{d_k}}\right)V$$

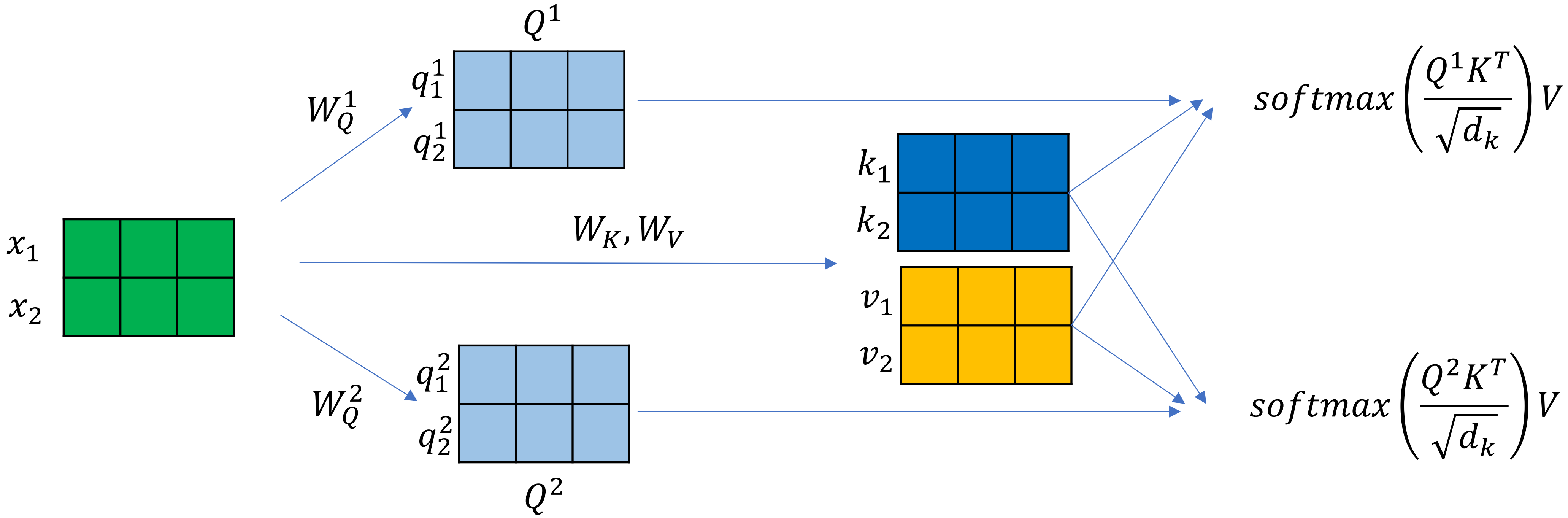
- Previous query vectors are no longer required, we can discard them to save memory
- The key and value vectors remain unchanged during inference. We don’t need to compute them every single time but can simply “cache” them

KV-Cache

- Faster computation at the cost of more memory requirement
- But?
  - For longer sequence, it has to cache larger key and value matrices
  - Might lead to out-of-memory error for longer sequences
- How to deal with this?
- Could we decrease the memory requirement?

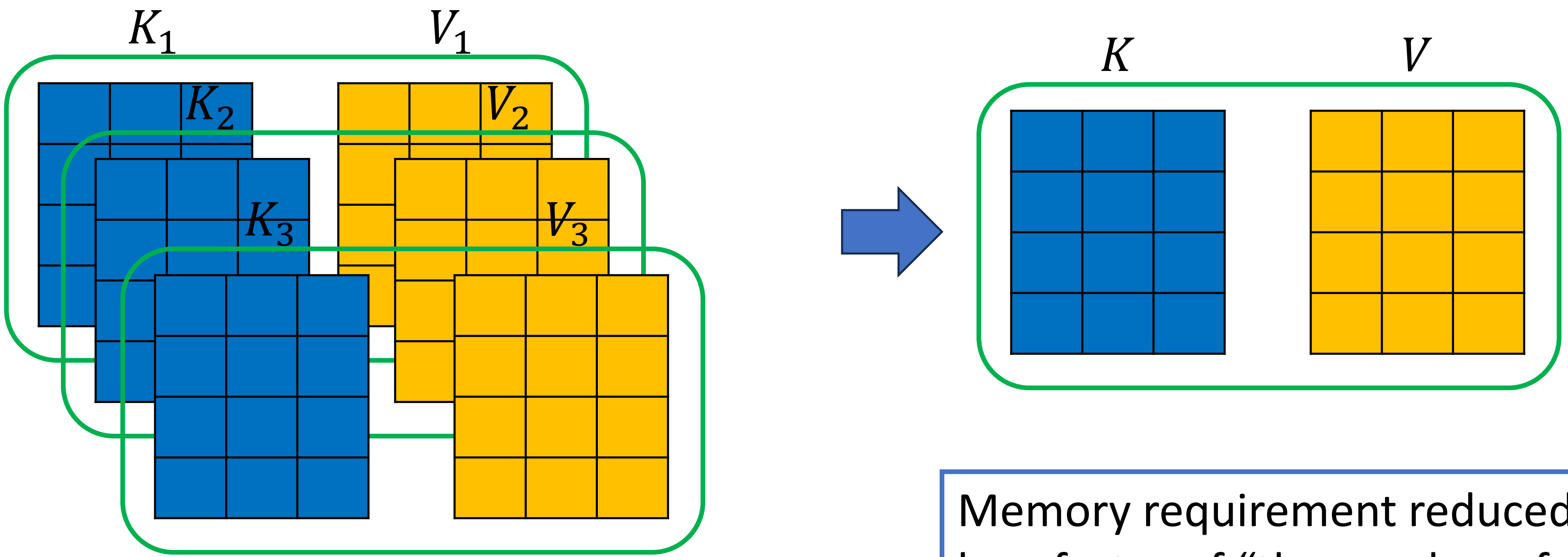
Multi Query Attention

- Share the key and value matrices across heads



Multi Query Attention

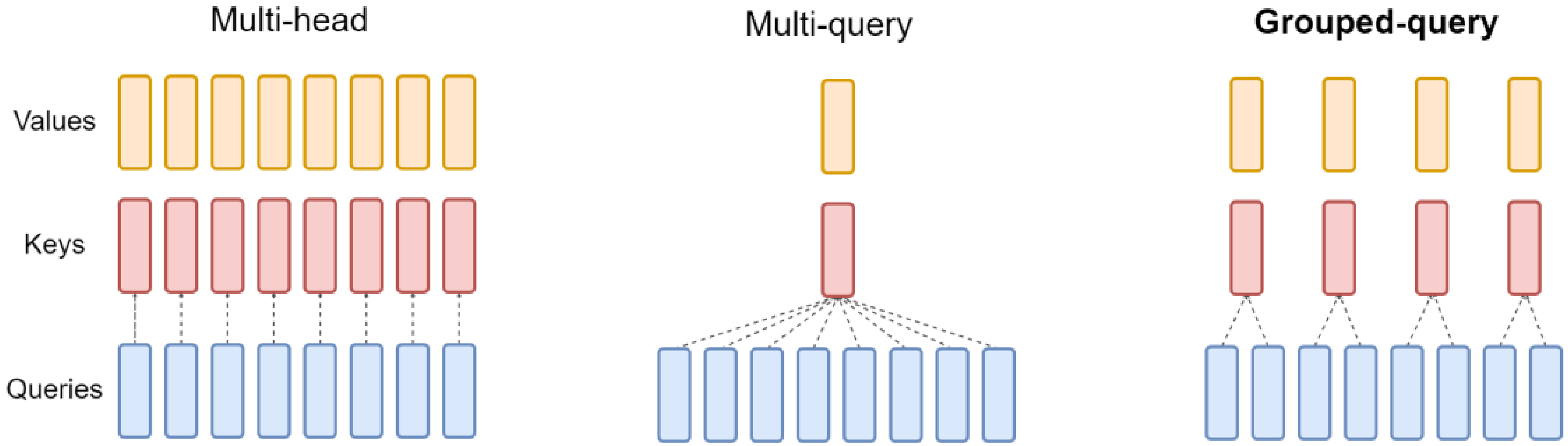
- What about Caching?



Memory requirement reduced by a factor of "the number of heads"

Grouped Query Attention

- Between multi-head and multi-query

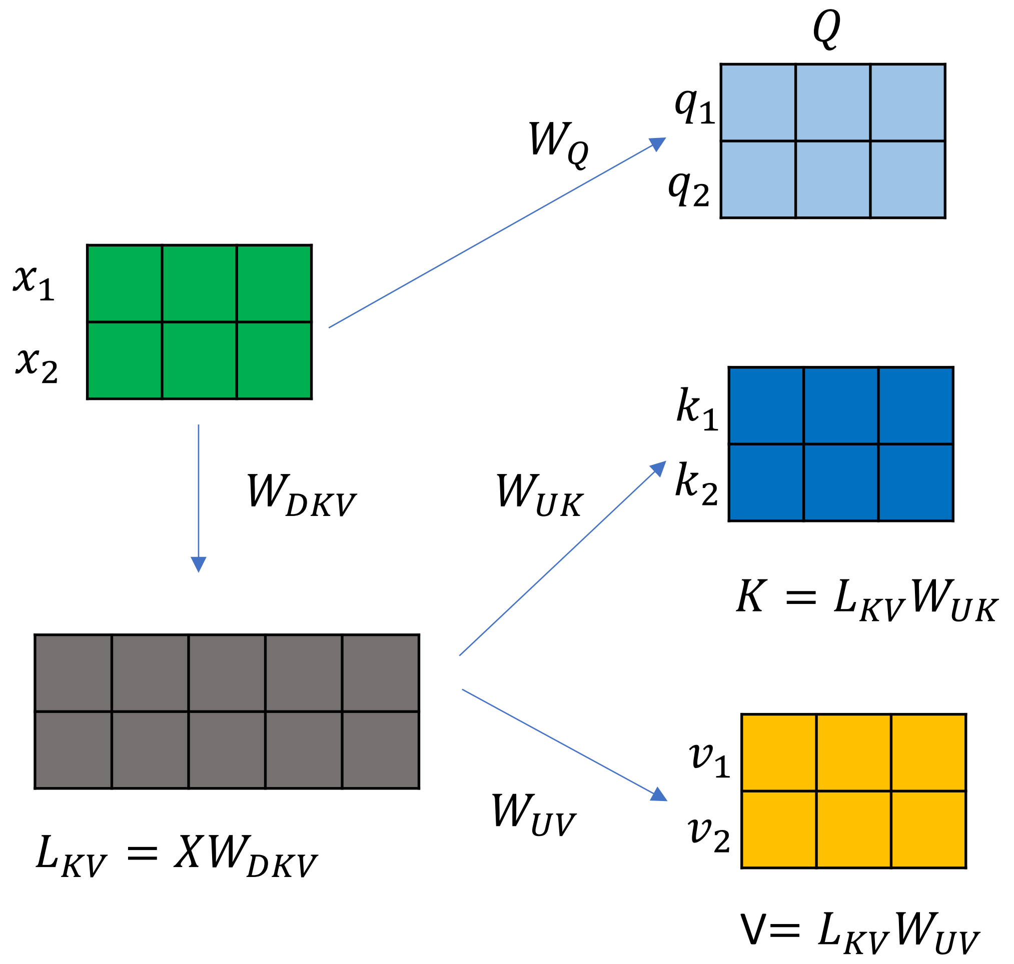


While more efficient, both multi and grouped query attention leads to degradation in performance

Latent Attention

- Introduced in DeepSeek models
- Can we develop a memory efficient attention comptation mechanism without degrading performance?
- What if we store the key values in a lower dimensional space?

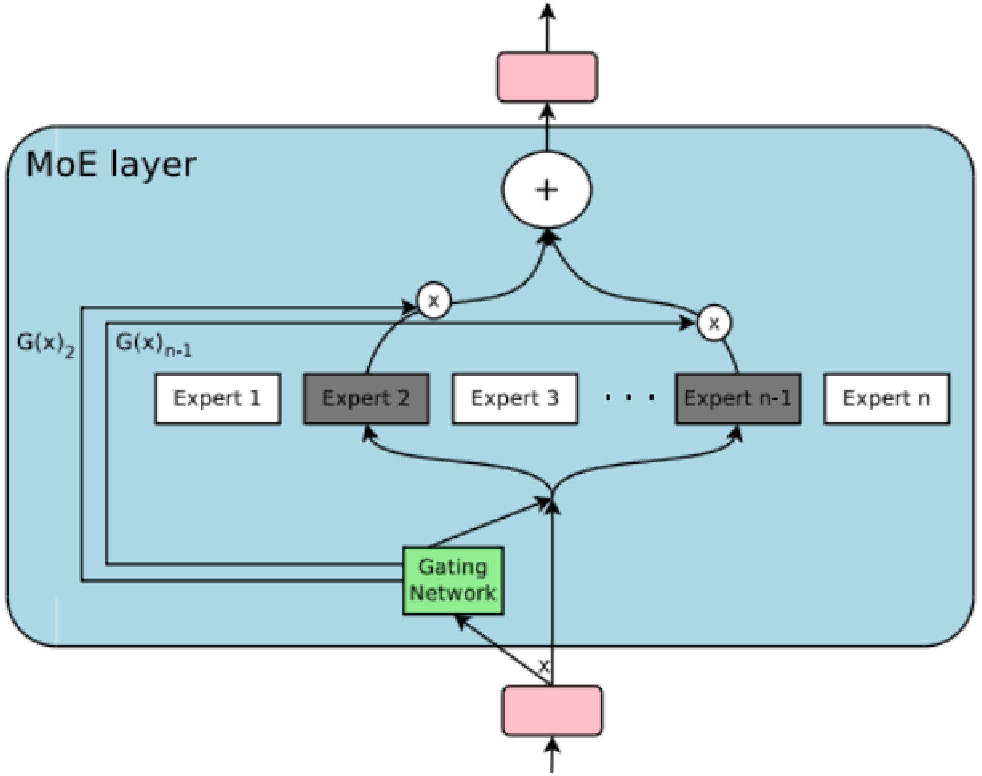
Latent Attention



- The latent KV (L_{KV}) is cached and shared across heads
- Additional parameters W_{DKV}, W_{UK} and W_{UV} are introduced

Mixture of Experts

- Feed forward module in the transformers are replaced by mixture of experts (MOE)
  - Consists of several experts $E_1, E_2, \dots, E_n$
  - Each expert is a neural network
  - Given an input x, the output for each expert is given by $E_i(x)$
  - The Gating network G, determines which expert to choose



$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$

Mixture of Experts

$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$

- If $G(x)_i$ is 0, no need to compute the respective expert operations and save compute
- Traditionally, G is a simple network with a softmax

$$G(x) = \mathrm{Softmax}(x\, W_g)$$

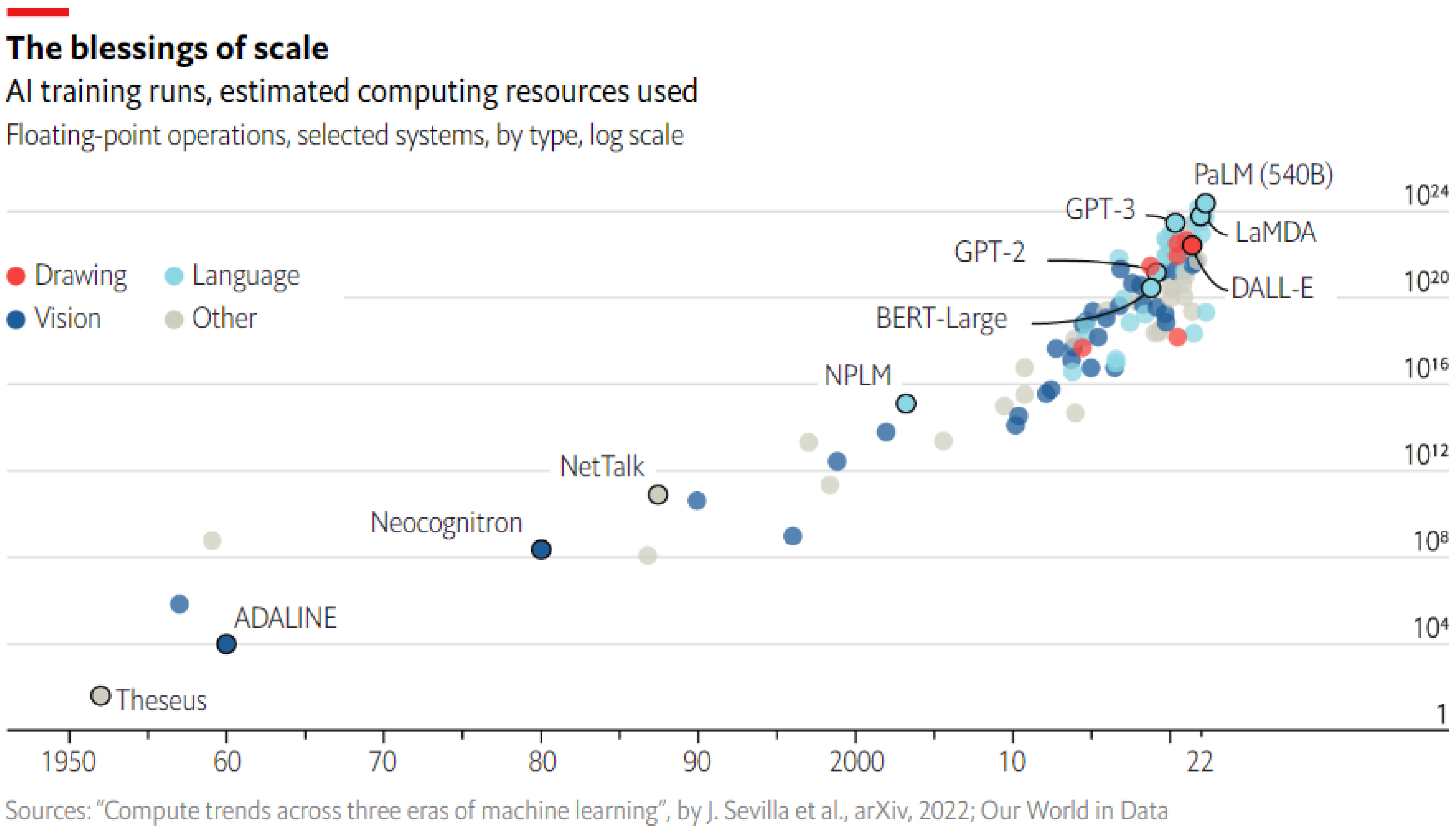
- Other sparsity approaches have also been explored

Mixture of Experts

- Although the parameter size increases, it is computationally much more efficient
- GLAM, language model from Google with 1.5 Trillion parameters deployed mixture of experts
  - 7x larger than GPT 3
  - 1/3 energy used during training
  - Required almost half computation during inference
  - Achieved better results across 29 NLP tasks

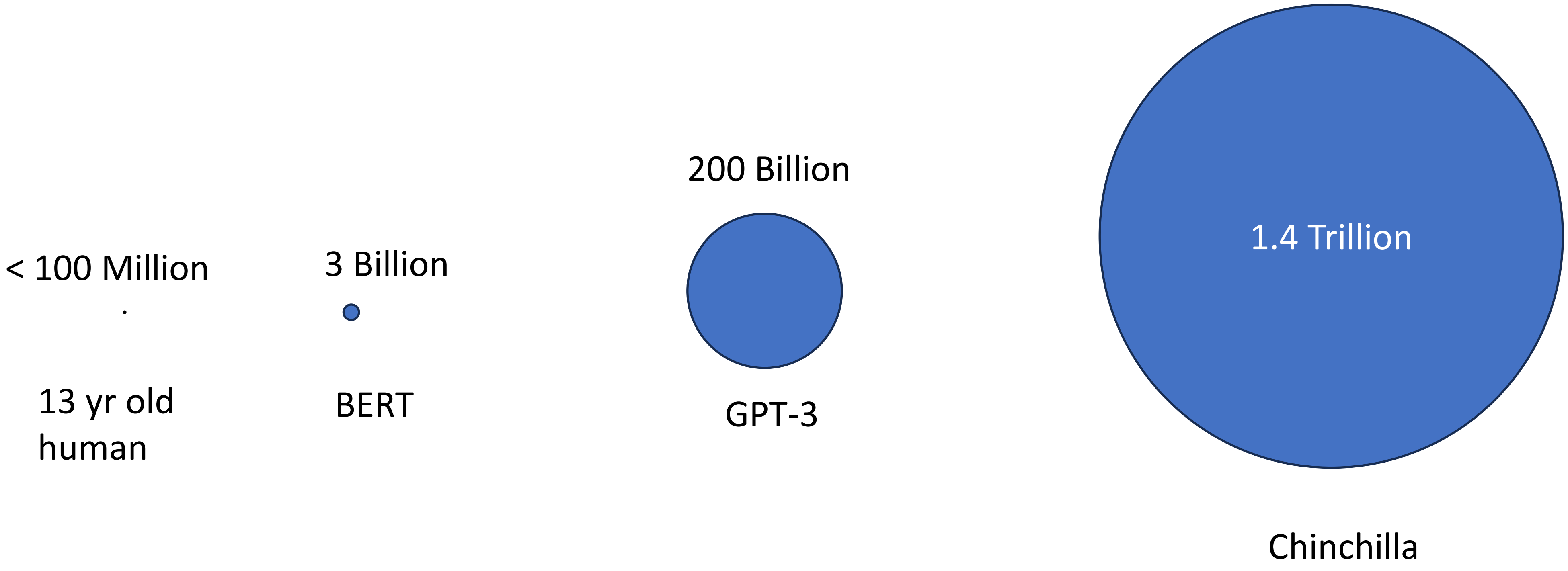

LARGE LANGUAGE MODELS
LIGHTYEAR
LARGE LANGUAGE MODELS EVERYWHERE
imgflip.com

Larger and larger models



Trained on more and more data

Llama4 and DeepseekV3 models have been trained on ~14 Trillion tokens



Language models as world models

- Natural language -> code

sentiments.ts write_sql.go parse_expenses.py addresses.rb

```
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```

Copilot

Language models as world models

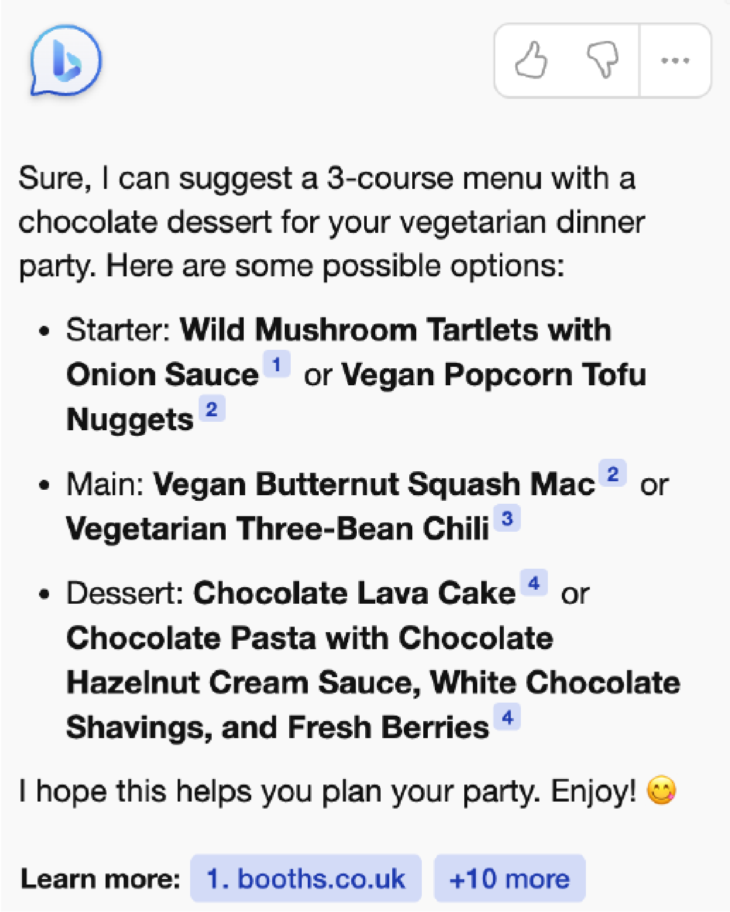
- Medicine

Following endaerectomy on the right common carotid, a patient is found to be blind in the right eye. It is appears that a small thrombus embolized during surgery and lodged in the aery supplying the optic nerve. Which aery would be blocked?

Central aery of the retina

Language models as multitask assistants

I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?

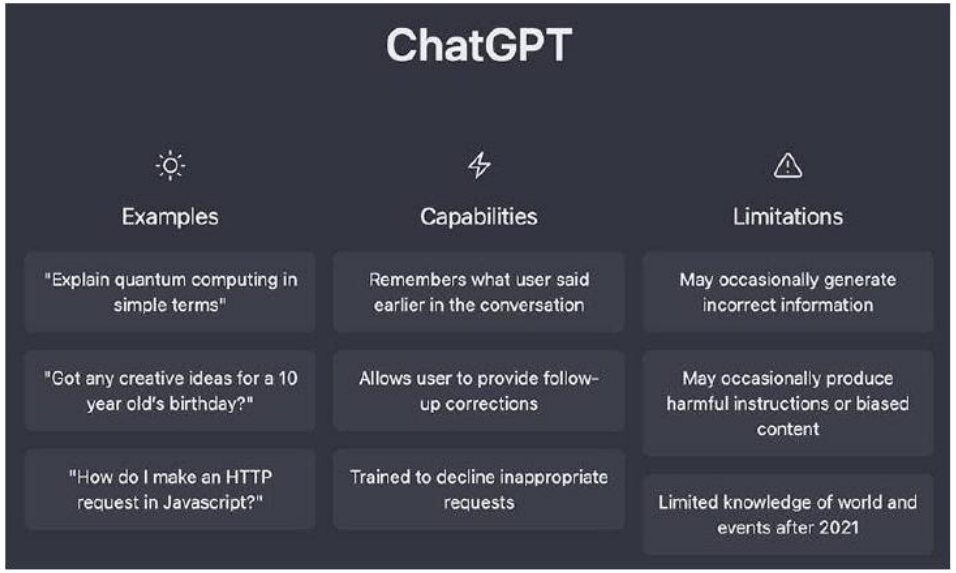


Language models as multitask assistants

- We started with –

The students opened their ____________

- And we have ended up -

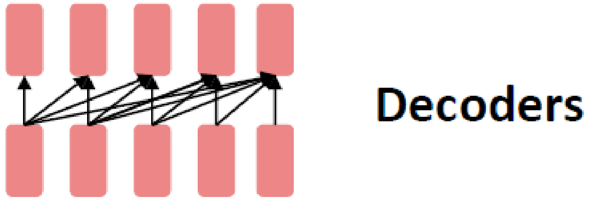


Language models to assistants

- Zero-shot and few-shot in-context learning (prompting)
- Instruction fine-tuning
- Reinforcement learning from human feedback (RLHF)

Emergent abilities of LLMs

- Let’s revisit the Generative Pretrained Transformer (GPT) models from OpenAI as an example:
- GPT (117M parameters)
- Transformer decoder with 12 layers.
- Trained on BooksCorpus : over 7000 unique books (4.6GB)
- Showed that language modeling at scale can be an effective pretraining technique for downstream tasks like natural language inference.



entailment

[START] *The man is in the doorway* [DELIM] *The person is near the door* [EXTRACT]

Emergent abilities of LLMs

- GPT 2 (1.5B parameters)
- Same architecture as GPT, just bigger (117M --> 1.5B)
- But trained on much more data : 4GB --> 40GB of internet text data WebText
- Scrape links posted on Reddit w/ at least 3 upvotes (rough proxy of human quality)

Emergent zero-shot learning

- One key emergent ability in GPT 2 is zero shot learning : the ability to do many tasks with no examples, and no gradient updates

- Specifying the right sequence prediction problem (e.g., question answering):

```
Passage: Tom Brady... Q: Where was Tom Brady born? A: ...
```

Emergent zero-shot learning

- One key emergent ability in GPT 2 is zero shot learning : the ability to do many tasks with no examples, and no gradient updates

The cat couldn't fit into the hat because it was too big.
Does it = the cat or the hat?

Emergent zero-shot learning

- GPT 2 beats SoTA on language modeling benchmarks with no task specific fine tuning

Context: "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel. "He was a great craftsman," said Heather. "That he was," said Flannery.
Target sentence: "And Polish, to boot," said ______.
Target word: Gabriel

LAMBADA (language modeling w/ long discourse dependencies) [Paperno et al., 2016]

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**

Emergent zero-shot learning

- You can get interesting zero shot behavior if you're creative enough with how you specify your task!
- Summarization:

```
Prehistoric man sketched an
incredible array of
prehistoric beasts on the
rough limestone walls of a
cave in modern day France
36,000 years ago… TL;DR:
```

→

The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man…

Language models to assistants

- Zero-shot and few-shot in-context learning (prompting)
- Instruction fine-tuning
- Reinforcement learning from human feedback (RLHF)

Reference and further reading

- CS224n: Chris Manning’s course at Stanford (slides are adopted from here)
- https://arxiv.org/pdf/2007.14062.pdf (Big Bird)
- https://arxiv.org/pdf/1910.13461.pdf (BART)
- https://arxiv.org/pdf/1910.10683.pdf (T5)
- https://arxiv.org/pdf/2005.14165.pdf (GPT-3)
- https://arxiv.org/pdf/1911.02150 (Multi-query attention)
- https://arxiv.org/pdf/2305.13245 (Grouped-query attention)