

NLP and the Web – WS 2024/2025



Lecture 12 Neural Language Modeling 4

Dr. Thomas Arnold
Hovhannes Tamoyan
Kexin Wang

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt



Syllabus (tentative)

| <u>Nr.</u> | <u>Lecture</u> |
|------------|---|
| 01 | Introduction / NLP basics |
| 02 | Foundations of Text Classification |
| 03 | IR – Introduction, Evaluation |
| 04 | IR – Word Representation |
| 05 | IR – Transformer/BERT |
| 06 | IR – Dense Retrieval |
| 07 | IR – Neural Re-Ranking |
| 08 | LLM – Language Modeling Foundations, Tokenization |
| 09 | LLM – Neural LLM |
| 10 | LLM – Adaptation |
| 11 | LLM – Prompting, Alignment, Instruction Tuning |
| 12 | LLM – Long Contexts, RAG |
| 13 | LLM – Scaling, Computation Cost |
| 14 | Review & Preparation for the Exam |

Recap

Reinforcement Learning

Long Context

Retrieval-based LMs

In-Context Learning

```
1  Translate English to French:
2  sea otter => loutre de mer
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese => .....
```

← *task description*

← *examples*

← *prompt*

Language Modeling != Following Human Instructions

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

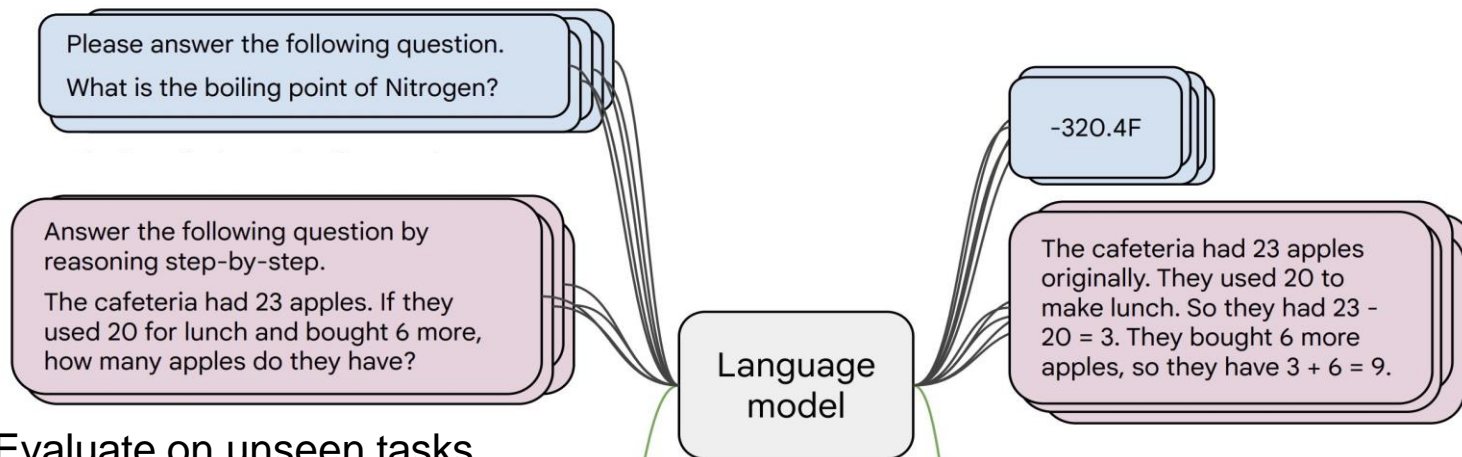
Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

There is a mismatch between LLM pre-training and **user intents**.

1. Collect examples of (instruction, output) pairs across many tasks and finetune an LM



2. Evaluate on unseen tasks

Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

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".

Recap

Reinforcement Learning

Long Context

Retrieval-based LMs

Why Reinforcement Learning?

- Remember the limits of Instruction-tuning?
 1. Difficult to collect diverse labeled data
 2. Rote learning (token by token) —
 - limited creativity
 3. Agnostic to model's knowledge —
 - may encourage hallucinations

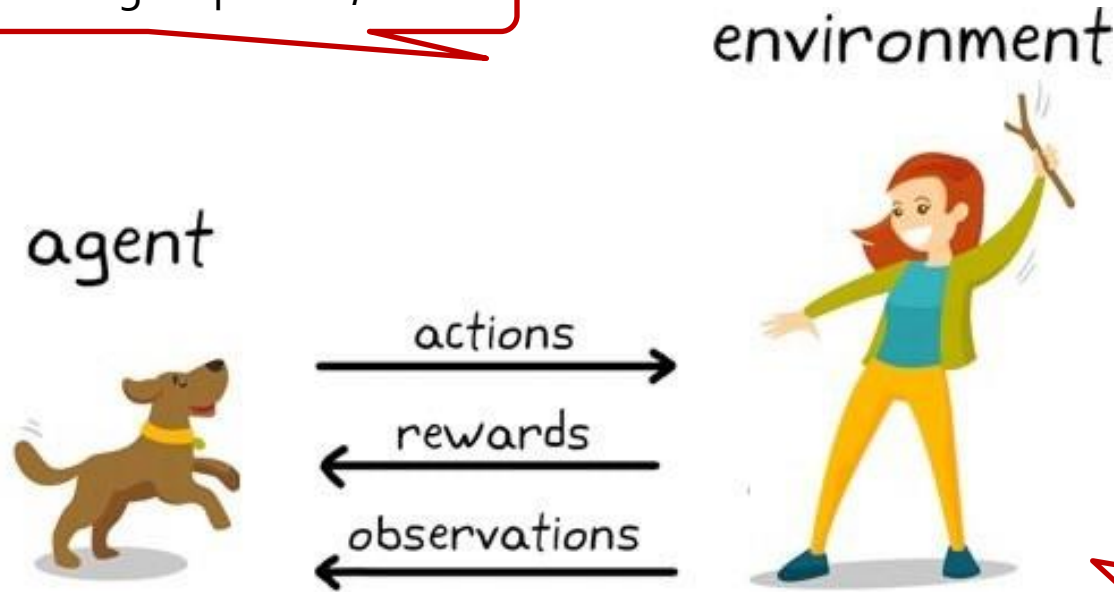
Limited/sparse feedback—usually considered a curse, but now a blessing.

“don't give a man fish rather teach him how to fish by himself”

The model itself should be involved in the alignment loop.

Reinforcement Learning: Intuition

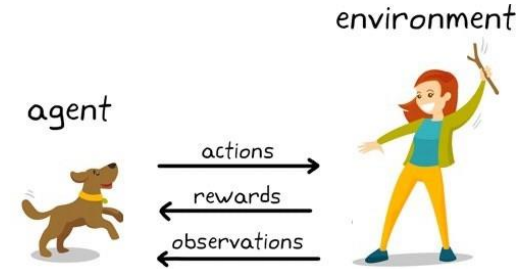
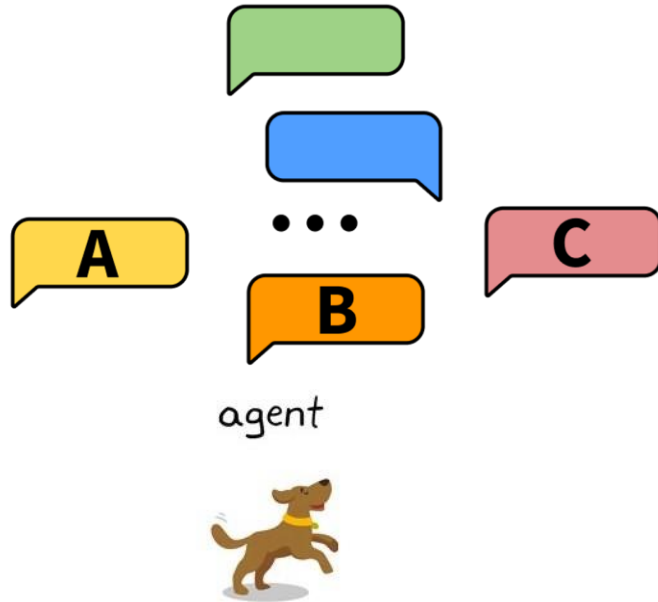
Action here: generating responses/token



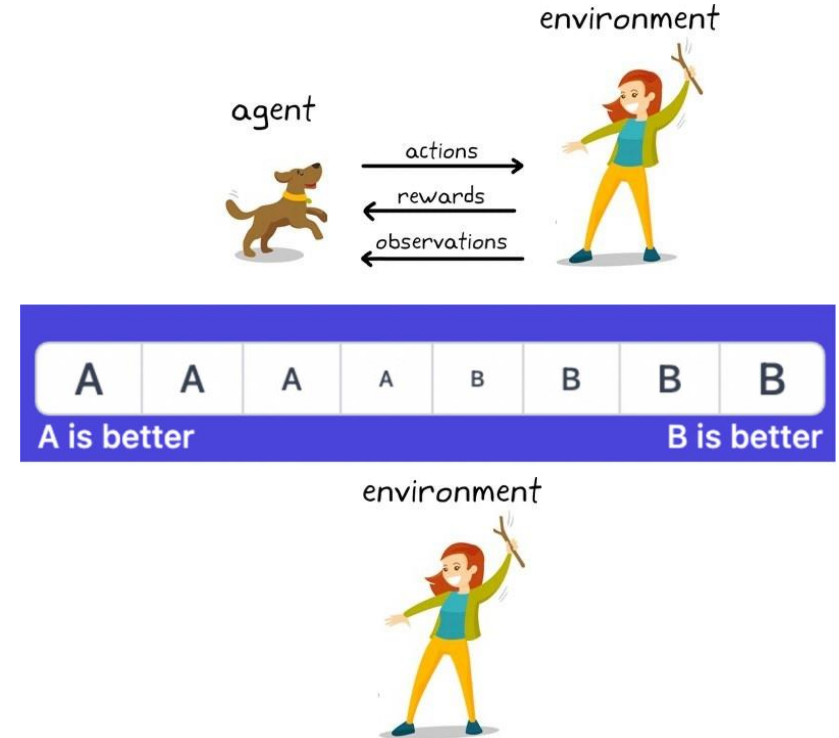
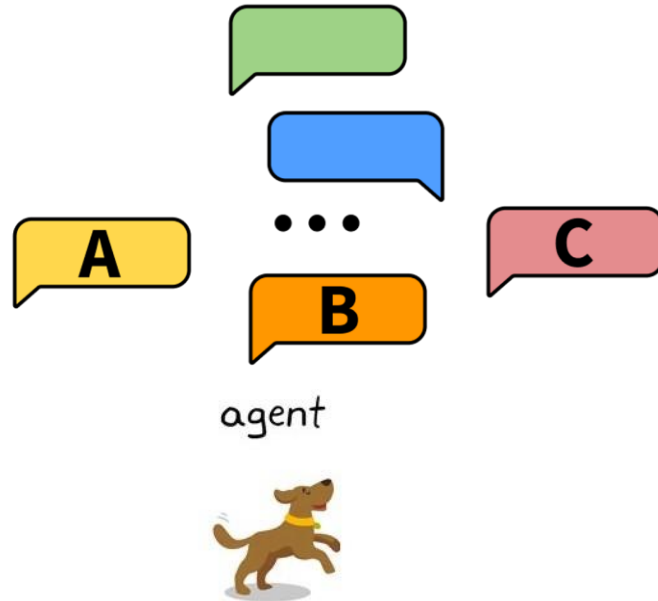
[figure credit]

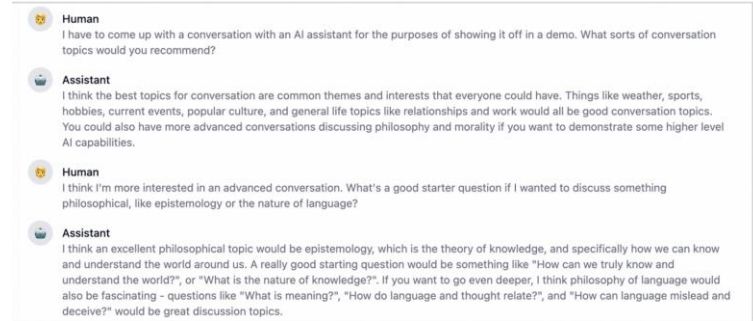
Reward here: whether humans liked the generation (sequence of actions=tokens)

Task: choose the better next message in a conversation



Scoring interface: Likert scale or rankings





Human
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

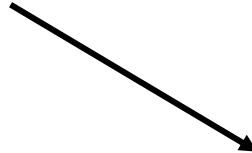
Assistant
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

Human
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

Assistant
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

human has conversation with the LLM

LLM provides two options for
next responses



Human
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

Assistant
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

Human
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

Assistant
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

Choose the most helpful and honest response

A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B
I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A A A A B B B B
A is better B is better

Intuition

- Human**
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?
- Assistant**
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.
- Human**
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?
- Assistant**
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

human rates better response

Choose the most helpful and honest response

A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B
I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A A A A B B B B
A is better B is better

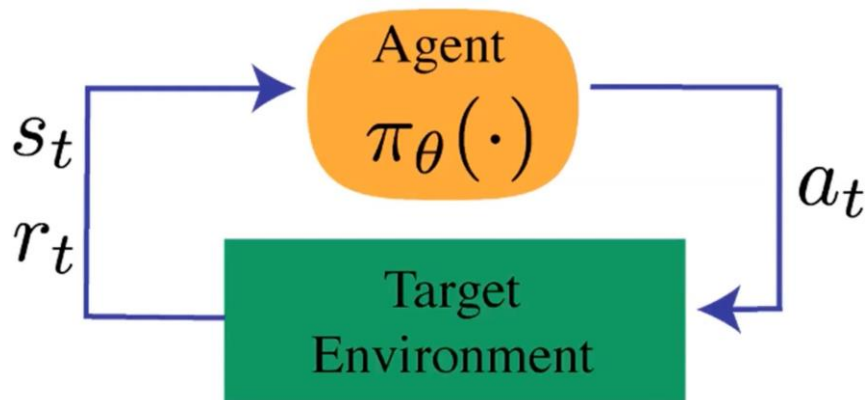
Reinforcement Learning: Abridged History

- The field of reinforcement learning (RL) has studied these (and related) problems for many years now [[Williams, 1992](#); [Sutton and Barto, 1998](#)]
- Circa 2013: resurgence of interest in RL applied to deep learning, game-playing [[Mnih et al., 2013](#)]
- But there is a renewed interest in applying RL. Why?
 - RL w/ LMs has commonly been viewed as very hard to get right (still is!)
 - We have found successful RL variants that work for language models (e.g., PPO; [[Schulman et al., 2017](#)])



Reinforcement Learning: Formalism

- An agent **interacts** with an environment by taking **actions**
- The environment returns a **reward** for the **action** and a **new state** (representation of the world at that moment).
- Agent uses a **policy function** to choose an action at a given **state**.
- We need to figure out: (1) reward function and (2) the policy function



Some notation:

s_t : state

r_t : reward

a_t : action

$a_t \sim \pi_{\theta}(s_t)$: policy

Reinforcement Learning from Human Feedback

- Imagine a reward function: $R(s; \text{prompt}) \in \mathbb{R}$ for any output s to a prompt.
- The reward is higher when humans prefer the output.
- Good generation is equivalent to finding reward-maximizing outputs:

Expected reward over the course of sampling from our policy (generative model)

$$\mathbb{E}_{\hat{s} \sim p_{\theta}} [R(\hat{s}; \text{prompt})]$$

$p_{\theta}(s)$ is a pre-trained model with params θ we would like to optimize (policy function)

Reinforcement Learning from Human Feedback

- Imagine a reward function: $R(s; \text{prompt}) \in \mathbb{R}$ for any output s to a prompt.
- The reward is higher when humans prefer the output.
- Good generation is equivalent to finding reward-maximizing outputs:

Expected reward over the course of sampling from our policy (generative model)

$$\mathbb{E}_{\hat{s} \sim p_{\theta}} [R(\hat{s}; \text{prompt})]$$

$p_{\theta}(s)$ is a pre-trained model with params θ we would like to optimize (policy function)

- On the notation:
 - “ \mathbb{E} ” here is an empirical expectation (i.e., average).
 - “ \sim ” indicates sampling from a given distribution.

Reinforcement Learning from Human Feedback

- Imagine a reward function: $R(s; \text{prompt}) \in \mathbb{R}$ for any output s to a prompt.
- The reward is higher when humans prefer the output
- Good generation is equivalent to finding reward-maximizing outputs:

$$\mathbb{E}_{\hat{s} \sim p_{\theta}} [R(\hat{s}; \text{prompt})]$$

- What we need to do:
 - (1) Estimate the reward function $R(s; \text{prompt})$.
 - (2) Find the best generative model p_{θ} that maximizes the expected reward:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}} [R(\hat{s}; \text{prompt})]$$

Estimating the Reward R



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Obviously, we don't want to use human feedback directly since that could be 💰💰💰
- Alternatively, we can build a model to mimic their preferences [[Knox and Stone, 2009](#)]

Estimating the Reward R

- Obviously, we don't want to use human feedback directly since that could be 💰💰💰
- Alternatively, we can build a model to mimic their preferences [[Knox and Stone, 2009](#)]
- Approach 1: get humans to provide absolute **scores for each output**

Challenge: human judgments on different instances and by different people can be noisy and mis-calibrated!

prompt

Explain "space elevators" to a 6-year-old.



LM

p_θ



s_1

It is like any typical elevator, but it goes to space. ...



→ 0.8

s_2

Explain gravity to a 6-year-old. ...

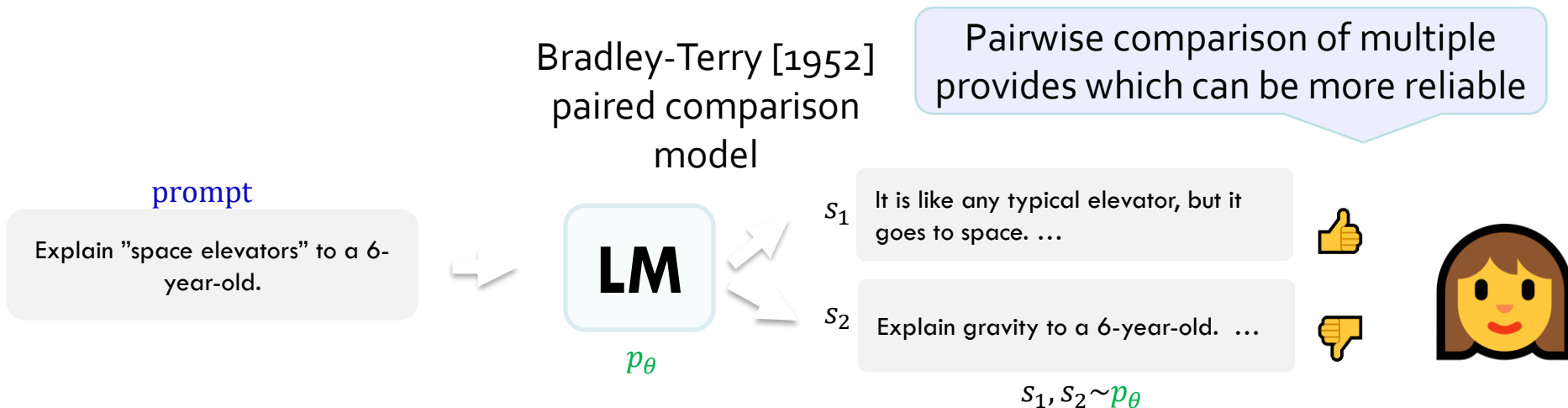


→ 1.2

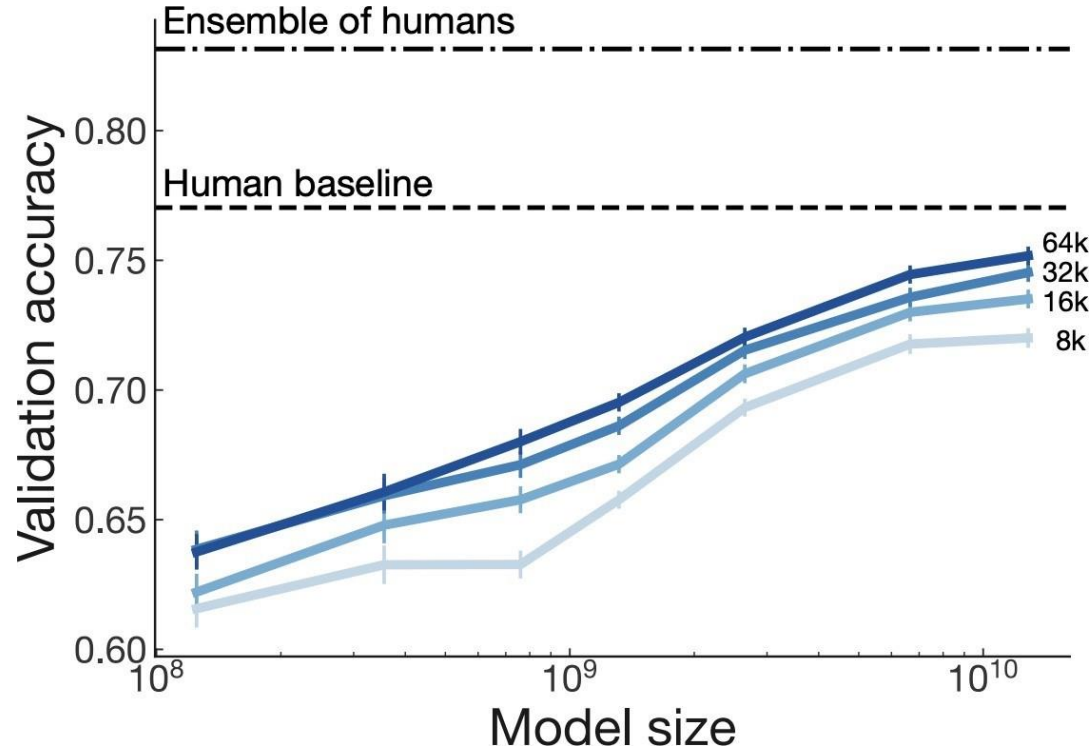
$s_1, s_2 \sim p_\theta$

Estimating the Reward R

- Obviously, we don't want to use human feedback directly since that could be 💰💰💰
- Alternatively, we can build a model to mimic their preferences [[Knox and Stone, 2009](#)]
- Approach 2: ask for **pairwise comparisons** [Phelps et al. 2015; Clark et al. 2018]



Scaling Reward Models

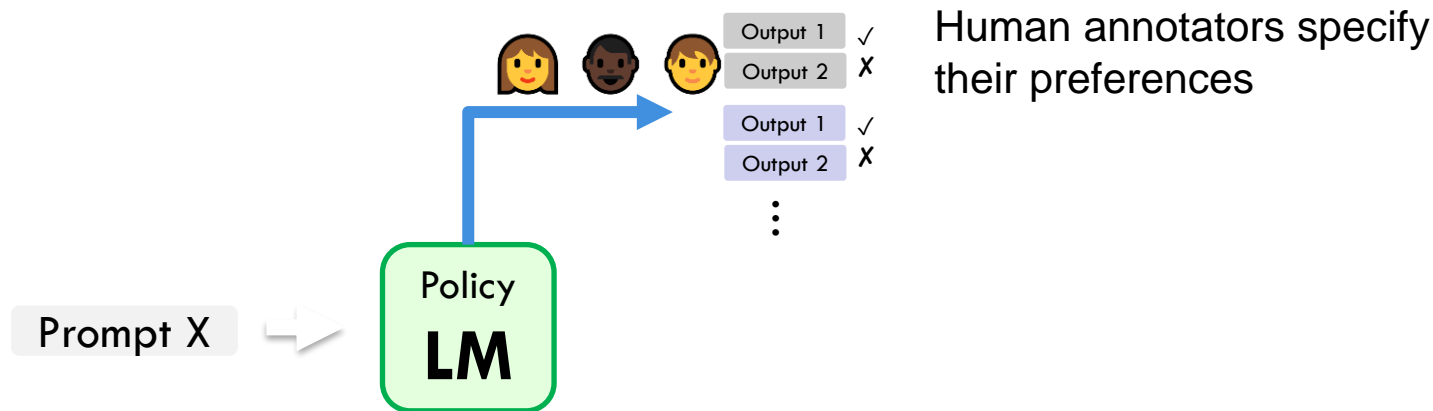


Large enough reward
trained on large enough
data approaching
human performance.

[[Stiennon et al., 2020](#)]

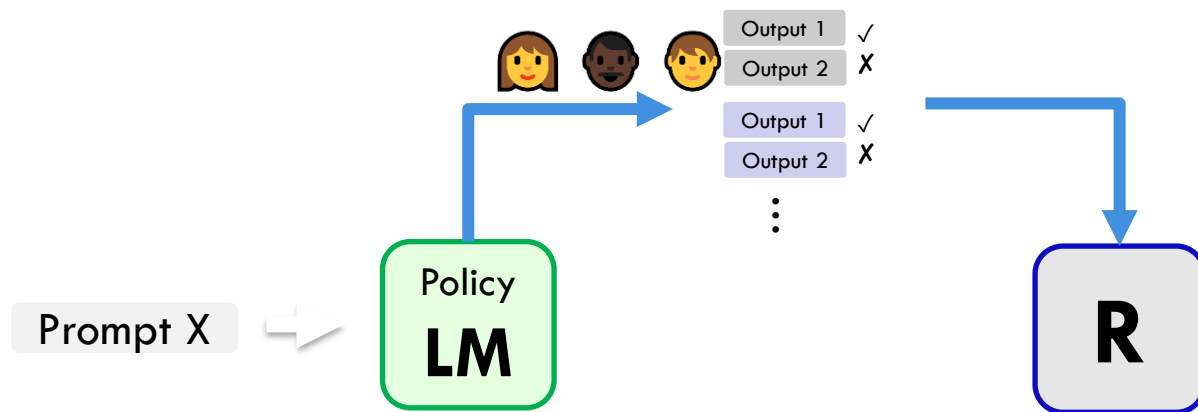
Putting it Together

- First collect a dataset of human preferences
 - Present multiple outputs to human annotators and ask them to rank the output based on preferability



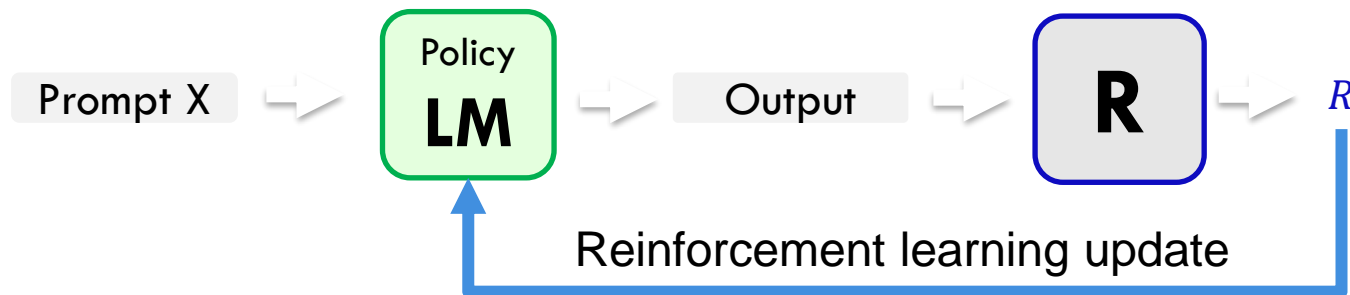
Putting it Together (2)

- Using this data, we can train a reward model
 - The reward model returns a scalar reward which should numerically represent the human preference.



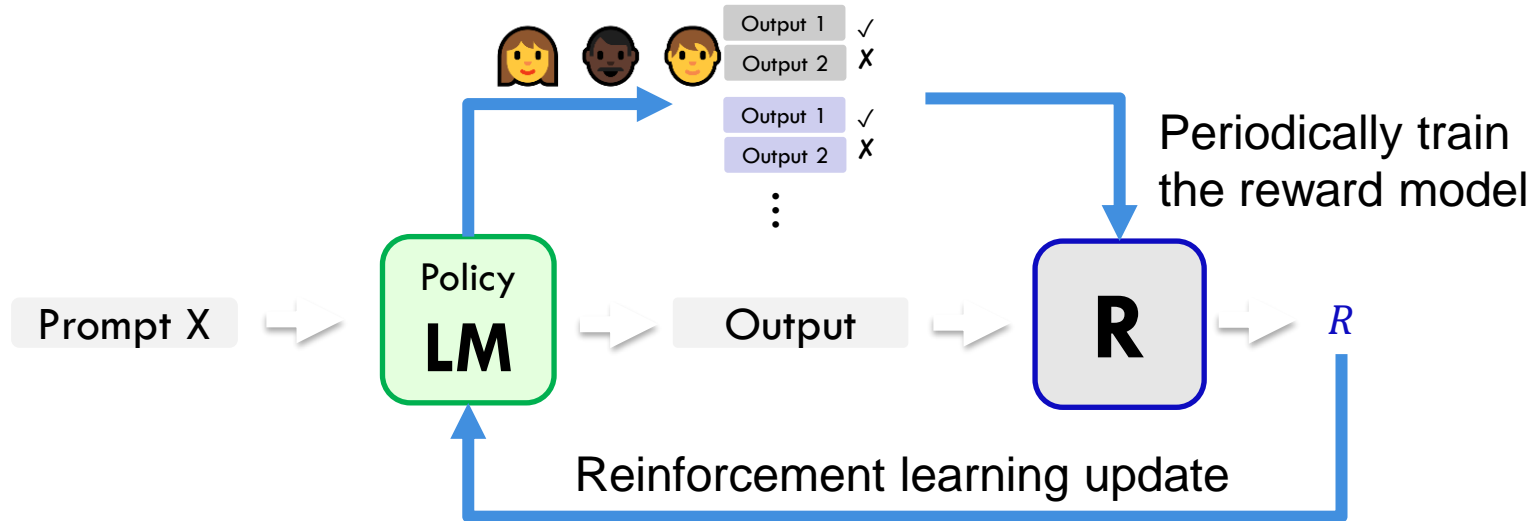
Putting it Together (3)

- We want to learn a policy (a Language Model) that optimizes against the reward model



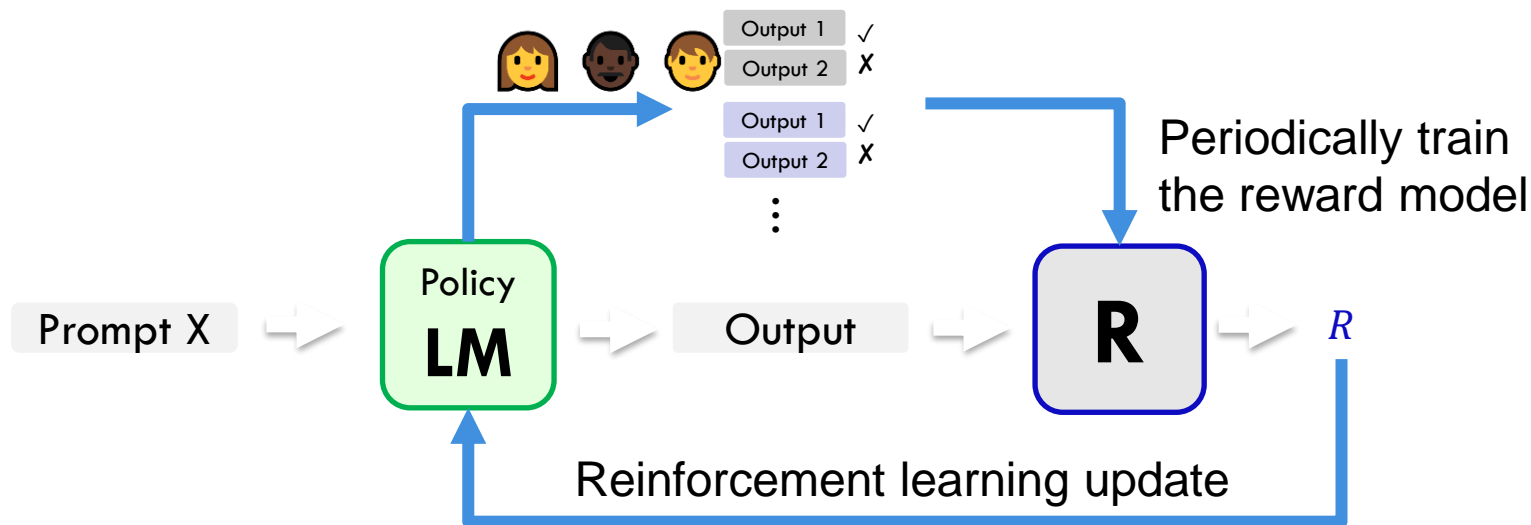
Putting it Together (4)

- Periodically train the reward model with more samples and human feedback



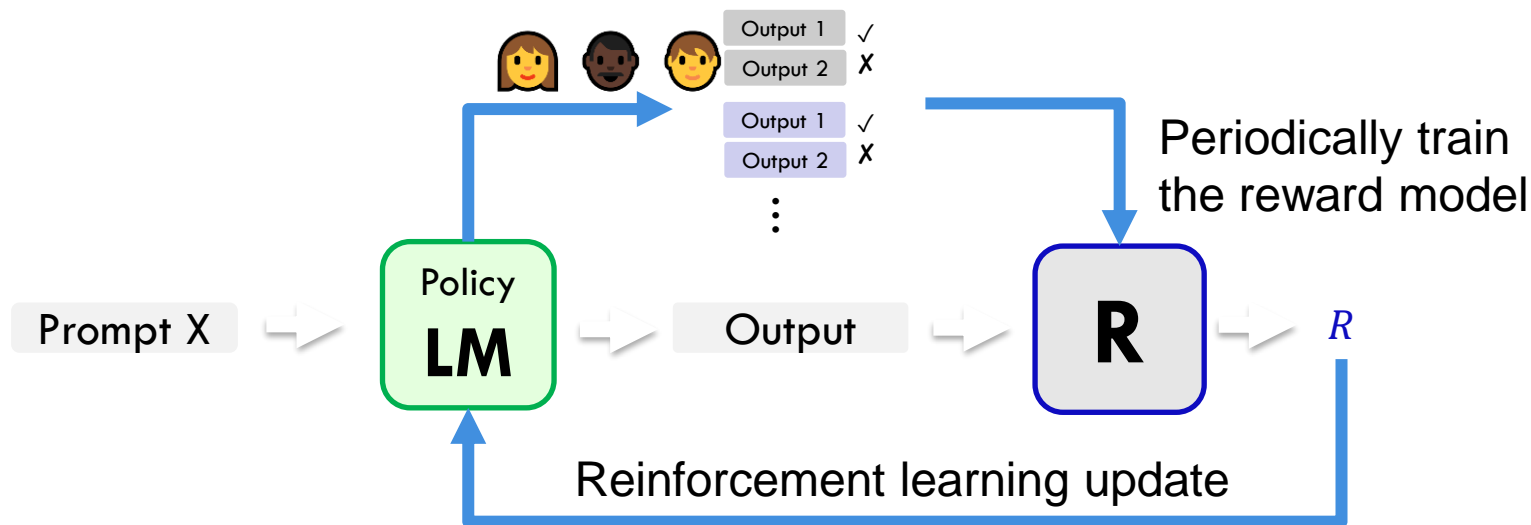
One missing ingredient

- It turns out that this approach doesn't quite work. The policy will learn to “cheat”.



One missing ingredient

- Will learn to produce an output that would get a **high** reward but might be **gibberish** or **irrelevant** to the prompt.
- Note, since R is trained on natural inputs, it may not generalize to unnatural inputs.



Regularizing with Pre-trained Model

- **Solution:** add a penalty term that penalizes too much deviations from the distribution of the pre-trained LM.

$$\hat{R}(s; p) := R(s; p) - \beta \log \left(\frac{p^{RL}(s)}{p^{PT}(s)} \right)$$

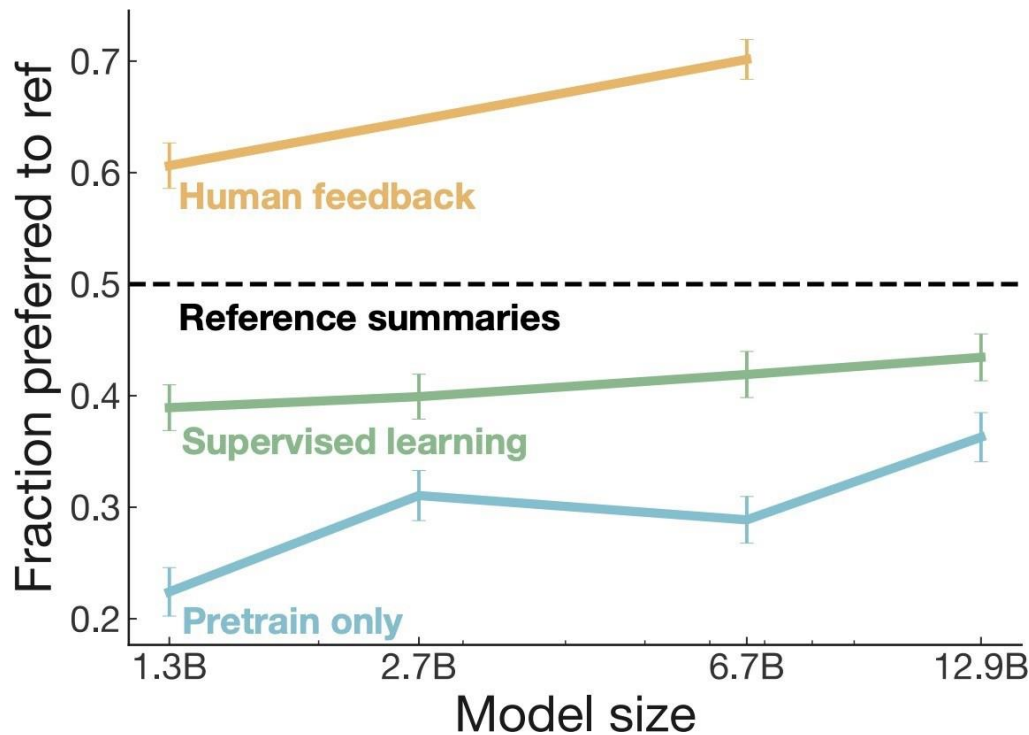
pay a price when
 $p^{RL}(s) < p^{PT}(s)$

- This prevents the policy model from diverging too far from the pretrained model.

The overall recipe



RLHF Gains over Instruction-Tuning



[[Stiennon et al., 2020](#)]

GPT3 vs. InstructGPT3 (RLHF-ed)



TECHNISCHE
UNIVERSITÄT
DARMSTADT

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

GPT3 vs. InstructGPT3 (RLHF-ed)



TECHNISCHE
UNIVERSITÄT
DARMSTADT

PROMPT *Write a short poem about a wise frog.*

COMPLETION GPT-3

Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

Write a short story in which a character has a bad dream.

InstructGPT

The frog is a symbol of wisdom
He knows all the secrets of the world
He is a master of disguise
And a great teacher of life
He is a symbol of transformation
And the bringer of change
He is the frog who has seen it all
And knows the meaning of it all

Can Help with Toxicity and Truthfulness

- Note, reward model can be used to induce any desired behavior as needed:
 - Avoiding bias
 - Avoiding responses outside its scope
 - Avoiding toxicity
 - ...

Lower is
better

| Dataset | |
|------------------------|--------------|
| RealToxicity | |
| GPT | 0.233 |
| Supervised Fine-Tuning | 0.199 |
| InstructGPT | 0.196 |

| Dataset | |
|------------------------|--------------|
| TruthfulQA | |
| GPT | 0.224 |
| Supervised Fine-Tuning | 0.206 |
| InstructGPT | 0.413 |

Higher is
better

Summary Thus Far

- Reinforcement learning can help mitigate some of the problems with supervised instruction tuning
- Reward model is trained via ranking feedback of humans.
- Regularization to restraint the model from deviating too far from the pre-trained policy
- Limitations:
 - RL can be tricky to get right
 - Training a good reward may require a lot of annotations

Recap

Reinforcement Learning

Long Context

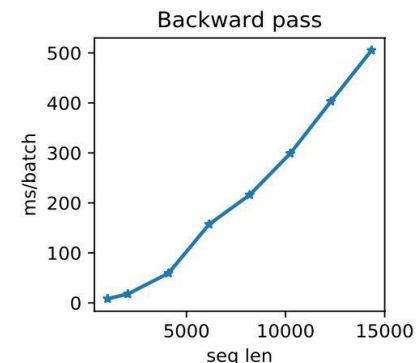
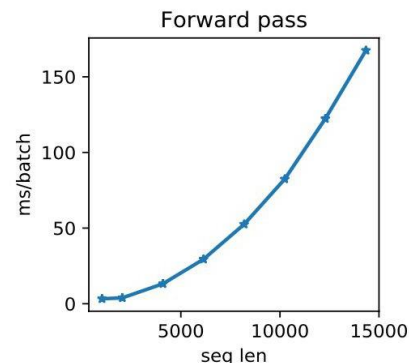
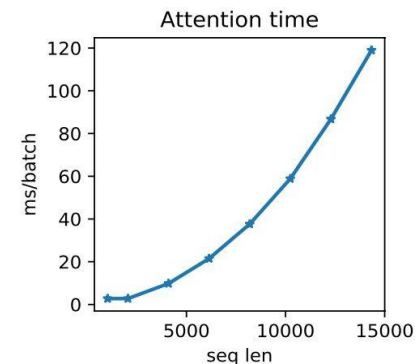
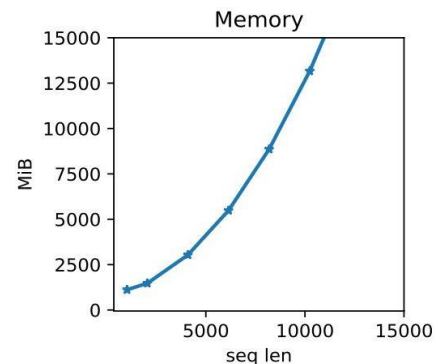
Retrieval-based LMs

Feeding Lots of Things to LM

- Books, scientific articles, government reports, videos, your daily experience, etc. they all are much longer than 2k tokens!!
- How do you enable language models process massive amounts of data?
- One approach: just scale up your model—train it on a much longer context window size.
 - The bottleneck: memory usage and number of operations in Self-Attention increases quadratically.

Transformer LMs and Long Inputs

- **Length generalization:** Do Transformers work accurately on long inputs?
- **Efficiency considerations:** How efficient are LMs on long inputs?



Transformer LMs and Long Inputs

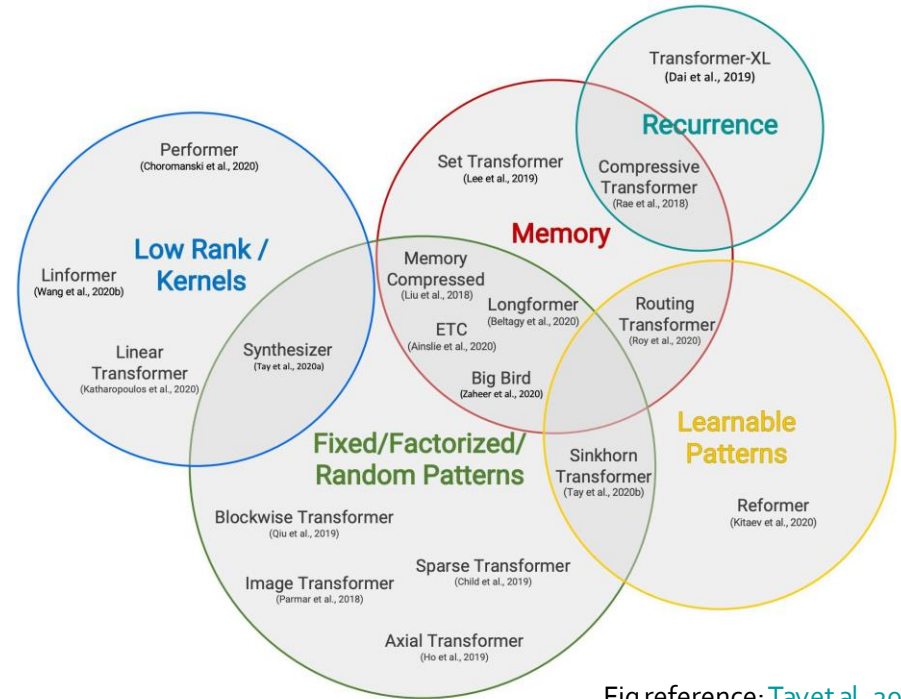
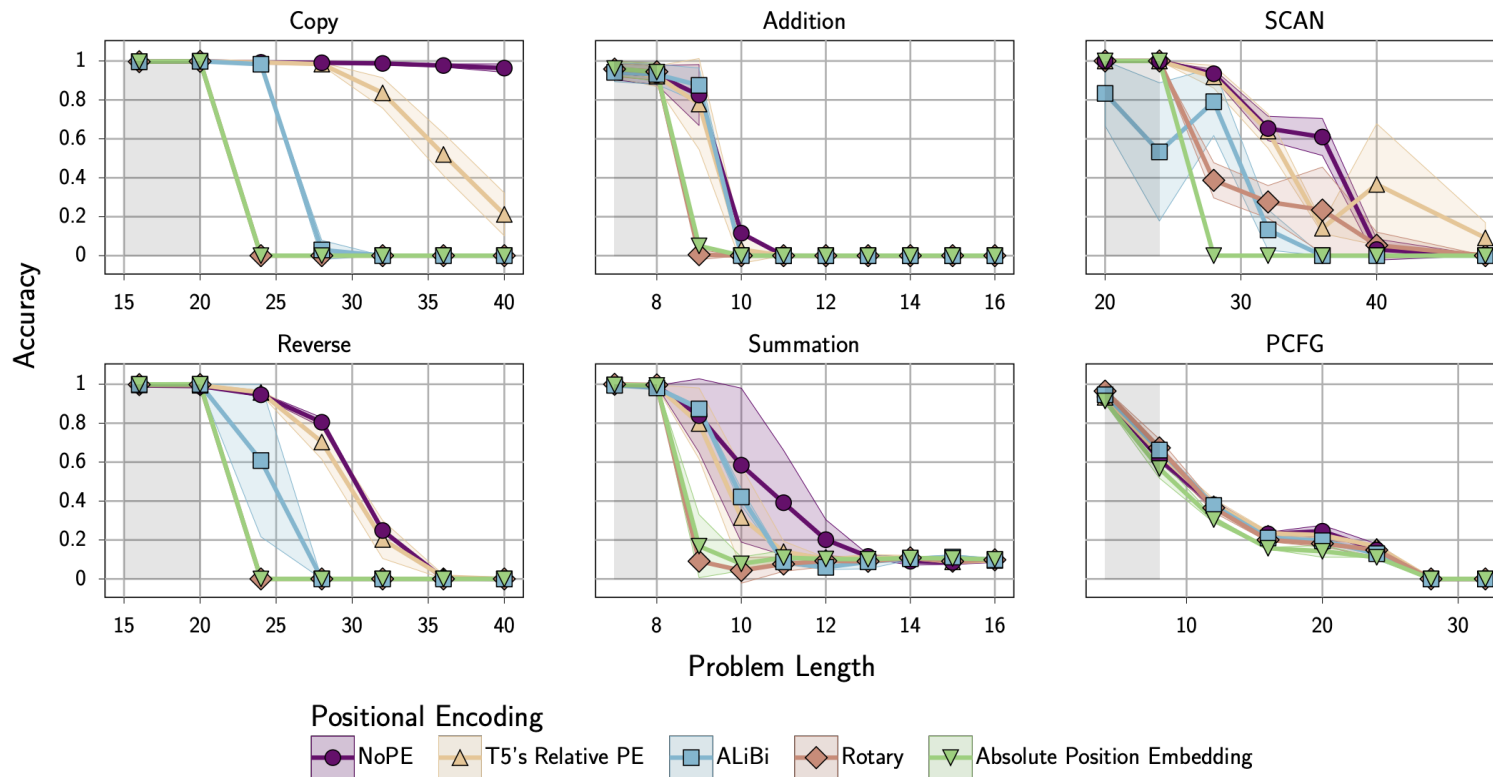


Fig reference: [Tay et al., 2020](#)

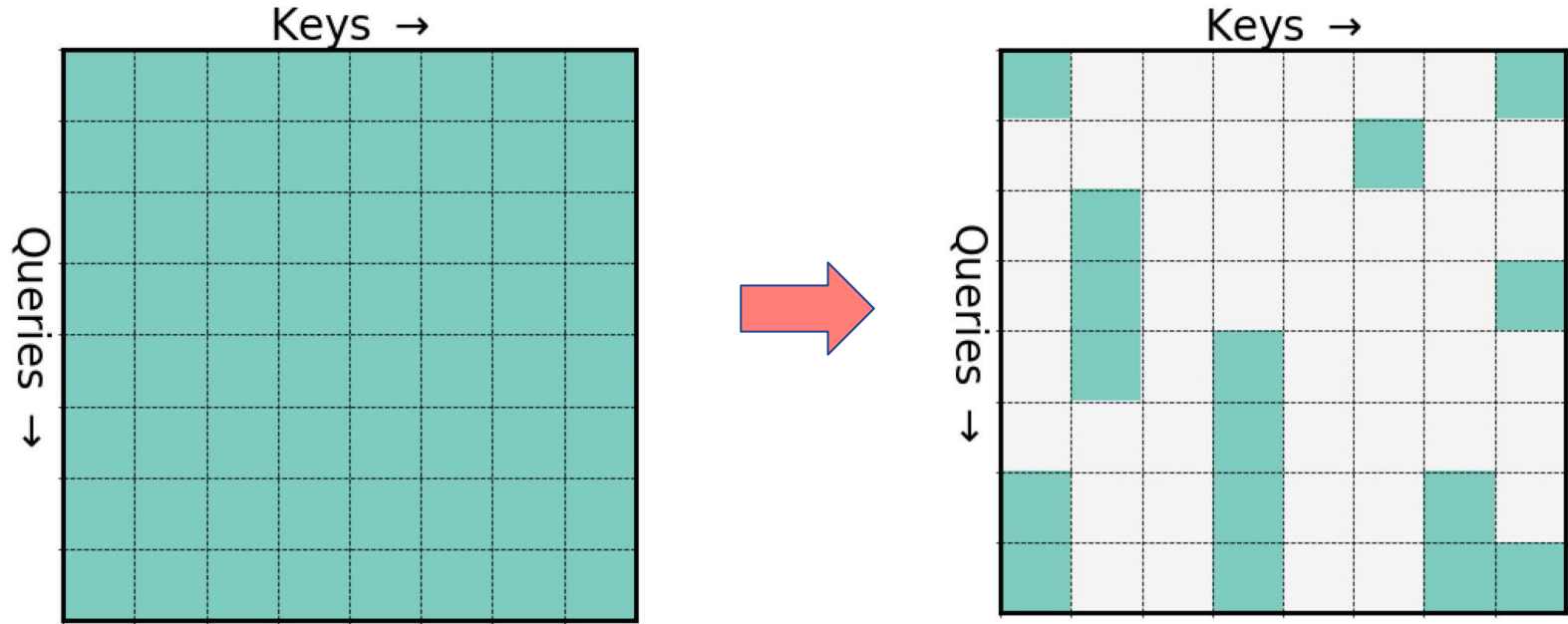
Length Generalization



Efficiency consideration: Sparse Attention Patterns

Sparse Attention Patterns

- The idea is to make the attention operation sparse

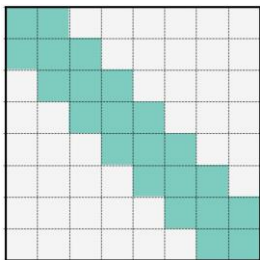


[NAACL 2021 Tutorial Beltagy, Cohan, Hajishirzi, Min, and Peters]

Pre-specified Sparsity Patterns

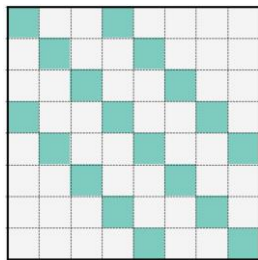
- A variety of patterns has been explored in the past work
 - Longformer ([Beltagy et al., 2020](#)), Sparse Transformer ([Child et al., 2019](#)), ...

Slidingwindow



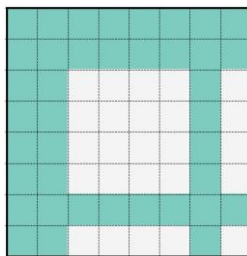
Sparse Transformer
Longformer

Dilated



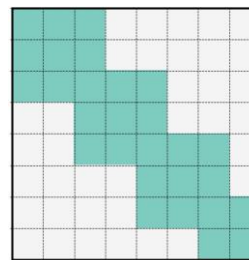
Longformer

Global



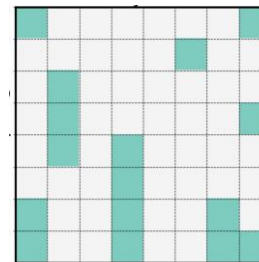
Big Bird

Blocked



Big Bird
Sinkhorn

Random

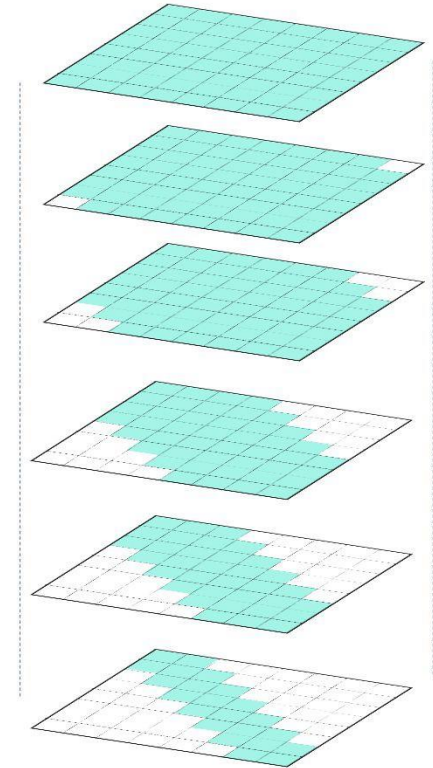


Big Bird

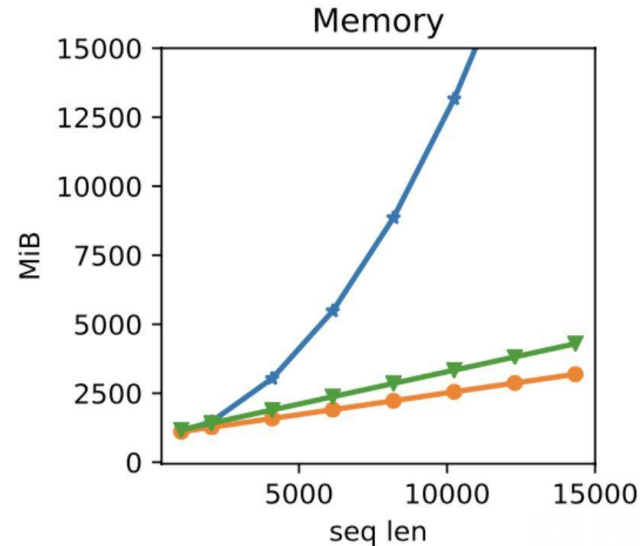
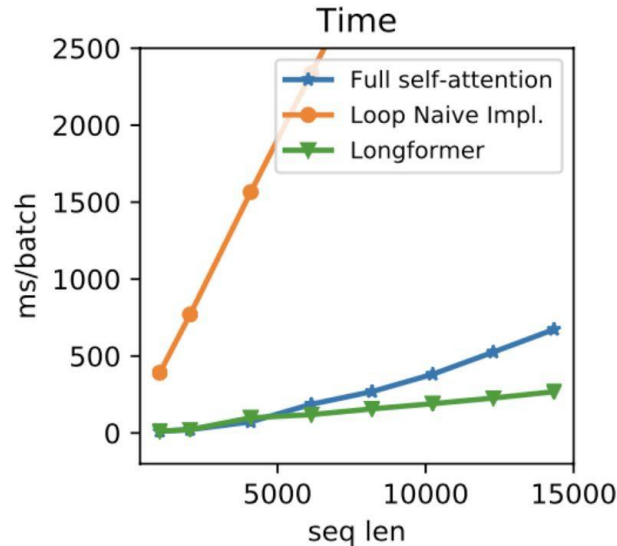
[[NAACL 2021 Tutorial Beltagy, Cohan, Hajishirzi, Min, and Peters](#)]

Pre-specified Sparsity Patterns

- Different layers and attention heads can follow different patterns
- A common setup is to have earlier layers with sparser attention pattern.
 - Longformer ([Beltagy et al., 2020](#))



Pre-specified Sparsity Patterns: Computations



[Longformer ([Beltagy et al., 2020](#))]

A Notable Adoption: GPT-3

- Sparse patterns also used in GPT-3 ([Brown et al., 2020](#))

2.1 Model and Architectures

We use the same model and architecture as GPT-2 [RWC⁺19], including the modified initialization, pre-normalization, and reversible tokenization described therein, with the exception that we use alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer [CGRS19]. To study the dependence of ML performance on model size, we train 8 different sizes of model, ranging over three orders of magnitude from 125 million parameters to 175 billion parameters, with the last being the model we call GPT-3. Previous work [KMH⁺20]

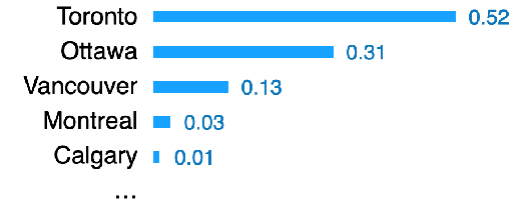
Recap

Long Context

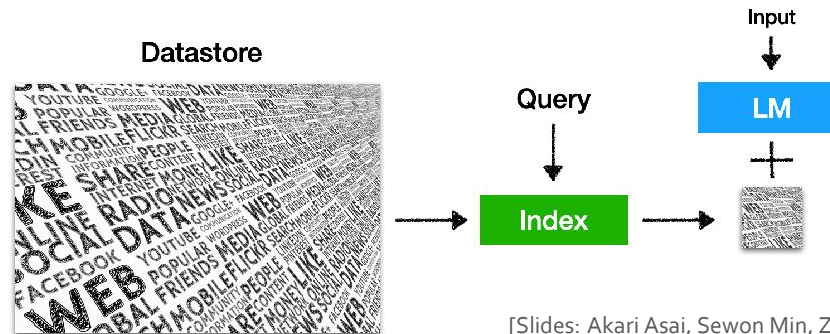
Retrieval-based LMs

- It is a **language model** $P(x_n | x_1, x_2, \dots, x_{n-1})$

The capital city of Ontario is ____



- It retrieves from an **external datastore** (at least during inference time)



[Slides: Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen]

Why Retrieval-based LMs?



LLMs can't memorize all (long-tail) knowledge in their parameters



List 5 important papers authored by Geoffrey Hinton



Geoffrey Hinton is a renowned computer scientist ...

Here are five important papers authored by him:

1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, .. ✓
2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, .. ✓
...
4. "**Deep Learning**" (with Y. Bengio and A. Courville) - Published as a book in **2016**,... ✗
5. "**Attention Is All You Need**" (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model,... ✗

Why Retrieval-based LMs?

LLMs' knowledge is easily outdated and hard to update



Who is the CEO of Twitter?



ChatGPT

As of my **knowledge cutoff in September 2021**, the CEO of Twitter is **Jack Dorsey**....



Who is the CEO of Twitter?



About 1,090,000,000 results (0.45 seconds)

Twitter / CEO

Linda Yaccarino

Jun 5, 2023–



Why Retrieval-based LMs?

LLMs' output is challenging to interpret and verify



Create an itinerary for exploring Toronto over a span of two days.



Toronto is a vibrant and diverse city with many attractions and activities to suit different tastes and preferences. Here are some suggestions for creating an itinerary for exploring Toronto over two days:

- **Day 1:** Start your day with a visit to the **CN Tower**, the iconic landmark that offers panoramic views of the city and beyond ¹. You can also experience the thrill of walking on the glass floor or the EdgeWalk, the world's highest hands-free walk ². Next, head to the nearby **St. Lawrence Market**, one of the world's best food markets, where you can sample a variety of cuisines and local specialties ³. After lunch, take a stroll along **Queen West**, a trendy neighborhood with eclectic shops, galleries, cafes and street art ⁴. In the evening, enjoy a show at one of the many theaters or comedy clubs in the **Entertainment District**, or catch a game at the **Scotiabank Arena** if you're a sports fan.

- **Day 2:** Explore the history and culture of Toronto at the **Royal Ontario Museum**, one of the largest and most comprehensive museums in North America, featuring exhibits on art, natural history, world cultures and more ⁴. Then, hop on a ferry to the **Toronto Islands**, a group of islands that offer a relaxing escape from the city, with beaches, parks, trails and amusement rides ³ ⁵. You can also rent a bike or kayak to explore the islands at your own pace. For dinner, head to **Chinatown**, one of the largest and most vibrant in North America, where you can find a variety of Asian cuisines and shops ³.

I hope this helps you plan your trip to Toronto. Have fun! 😊

Learn more:

1. [cntower.ca](https://www.cntower.ca)

2. [travel.usnews.com](https://www.travel.usnews.com)

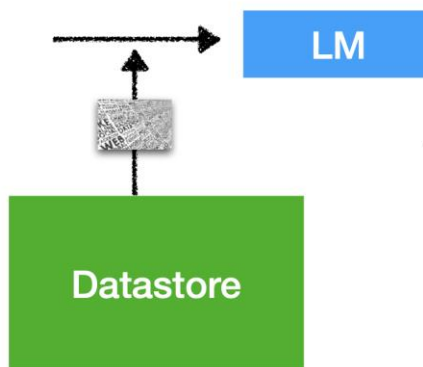
3. [bing.com](https://www.bing.com)

4. [rom.on.ca](https://www.rom.on.ca)

5. [tripadvisor.com](https://www.tripadvisor.com)

Why Retrieval-based LMs?

LLMs are **large** and expensive to train and run



vs.



Long-term goal: can we possibly reduce the **training** and **inference costs**, and scale down the size of LLMs?

e.g., RETRO (Borgeaud et al., 2021): “obtains comparable performance to GPT-3 on the Pile, despite using **25x fewer parameters**”

What are the Key Design Questions?

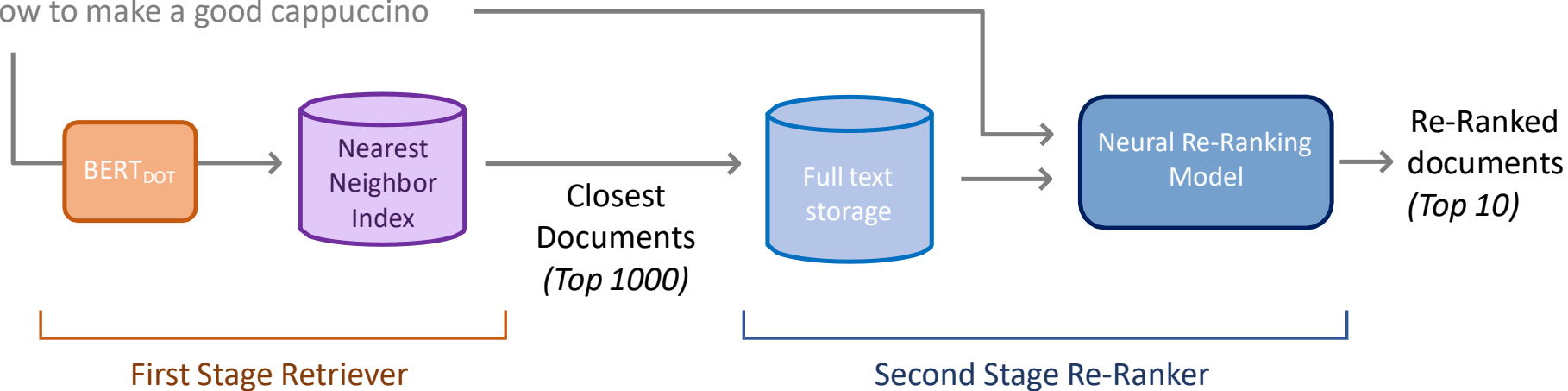
- What are your memories?
 - Documents, database records, training examples, etc.
- How to retrieve memories?
 - Use an off-the-shelf search engine (e.g. Google, StackOverflow).
 - How to train your own memory retriever.
- How to use retrieved memories?
 - "Text fusion"
 - Common failure modes:
 - Underutilization: model ignores retrieved memories.
 - Overreliance: model depends too much on memories!

Anatomy of a Neural Retreiver

Remember our IR lectures? (especially neural retrieval and approximate NN)

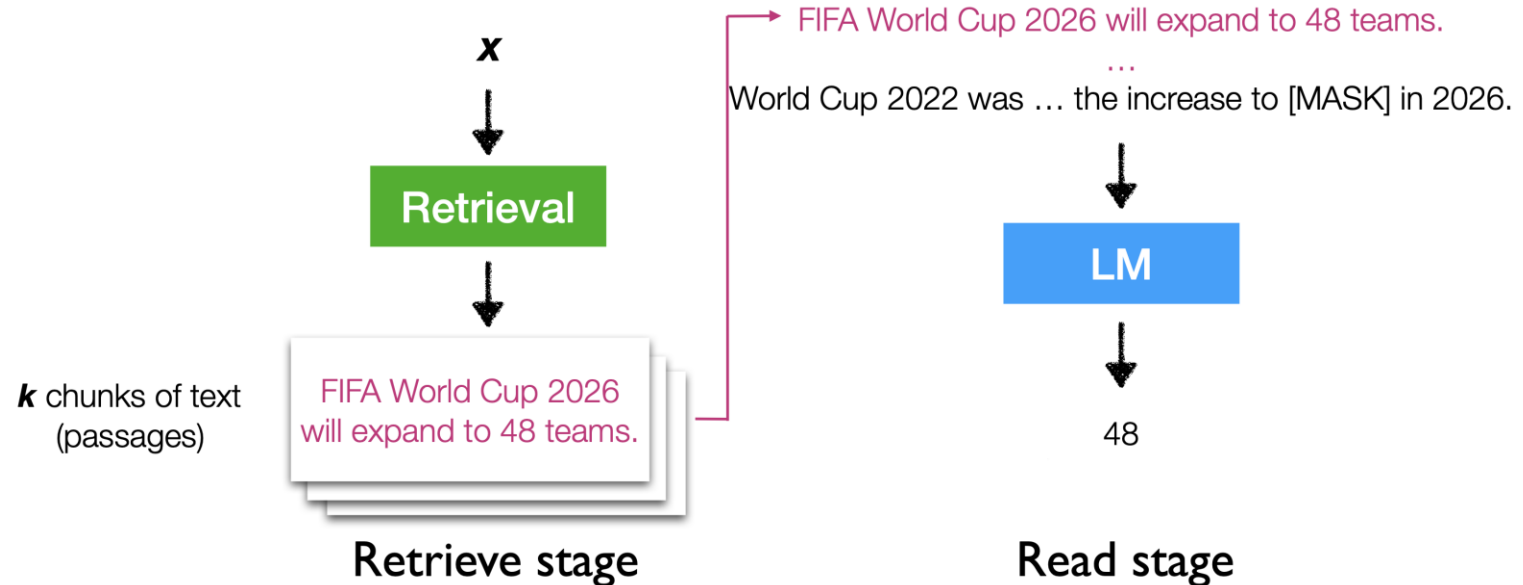
Prompt

How to make a good cappuccino



Retrieval-Augmented LM

- x = World Cup 2022 was the last before the increase to [MASK] in the 2026 tournament.



Retrieval-Augmented LM

FIFA World Cup 2026
will expand to 48 teams.

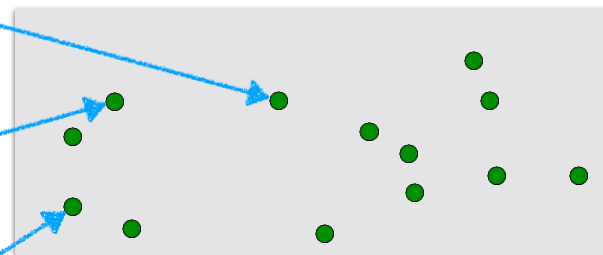
In 2022, the 32
national teams involved
in the tournament.

Team USA celebrated
after winning its match
against Iran ...

Encoder

Encoder

Encoder



$$\mathbf{z} = \text{Encoder}(z)$$

Wikipedia

13M chunks (passages)
(called *documents* in the paper)

Retrieval-Augmented LM

\mathbf{x} = World Cup 2022 was ... the increase to [MASK] in 2026.

FIFA World Cup 2026
will expand to 48 teams.

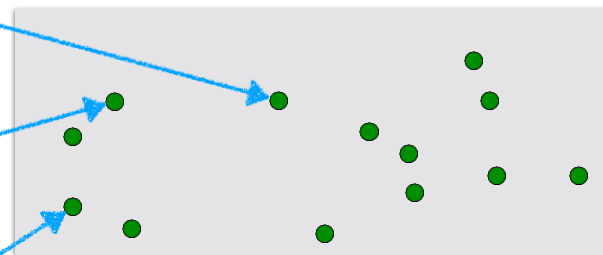
In 2022, the 32
national teams involved
in the tournament.

Team USA celebrated
after winning its match
against Iran ...

Encoder

Encoder

Encoder



$\mathbf{z} = \text{Encoder}(z)$

Wikipedia

13M chunks (passages)
(called *documents* in the paper)

Retrieval-Augmented LM

\mathbf{x} = World Cup 2022 was ... the increase to [MASK] in 2026.

Encoder

FIFA World Cup 2026
will expand to 48 teams.

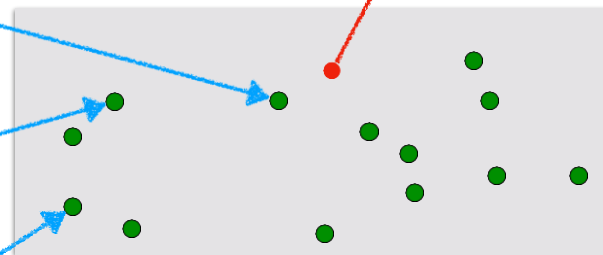
Encoder

In 2022, the 32
national teams involved
in the tournament.

Encoder

Team USA celebrated
after winning its match
against Iran ...

Encoder



$\mathbf{z} = \text{Encoder}(z)$

$\mathbf{x} = \text{Encoder}(x)$

Wikipedia

13M chunks (passages)
(called *documents* in the paper)

Retrieval-Augmented LM

\mathbf{x} = World Cup 2022 was ... the increase to [MASK] in 2026.

Encoder

FIFA World Cup 2026
will expand to 48 teams.

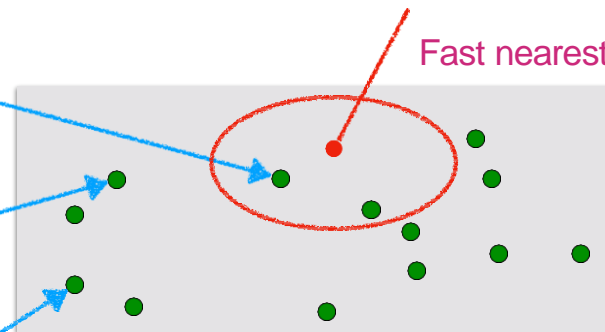
Encoder

In 2022, the 32
national teams involved
in the tournament.

Encoder

Team USA celebrated
after winning its match
against Iran ...

Encoder



Fast nearest neighbor search

$\mathbf{z} = \text{Encoder}(z)$

$\mathbf{x} = \text{Encoder}(x)$

Wikipedia

13M chunks (passages)
(called *documents* in the paper)

Retrieval-Augmented LM

\mathbf{x} = World Cup 2022 was ... the increase to [MASK] in 2026.

Encoder

FIFA World Cup 2026
will expand to 48 teams.

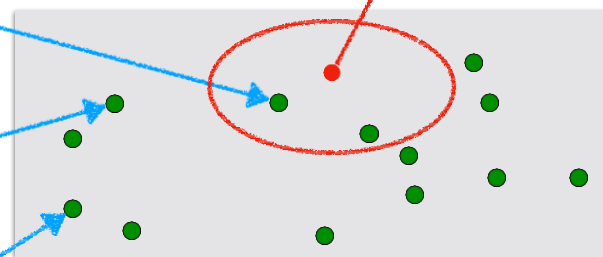
Encoder

In 2022, the 32
national teams involved
in the tournament.

Encoder

Team USA celebrated
after winning its match
against Iran ...

Encoder



Fast nearest neighbor search

$$\mathbf{z} = \text{Encoder}(z)$$

$$\mathbf{x} = \text{Encoder}(x)$$

$$z_1, \dots, z_k = \text{argTop-}k(\mathbf{x} \cdot \mathbf{z})$$

k retrieved chunks

Wikipedia

13M chunks (passages)
(called *documents* in the paper)

Retrieval-Augmented LM: Common Variant

What to retrieve?

- **Chunks**
- Tokens
- Others

How to use retrieval?

- **Input layer**
- Intermediate layers
- Output layer

When to retrieve?

- **Once**
- Every n tokens ($n > 1$)
- Every token

Retrieval-Augmented LM: Example Variation

What to retrieve?

- **Chunks**
- Tokens
- Others

How to use retrieval?

- Input layer
- **Intermediate layers**
- Output layer

When to retrieve?

- Once
- **Every n tokens ($n \geq 1$)**
- Every token

IR in the Middle of LM

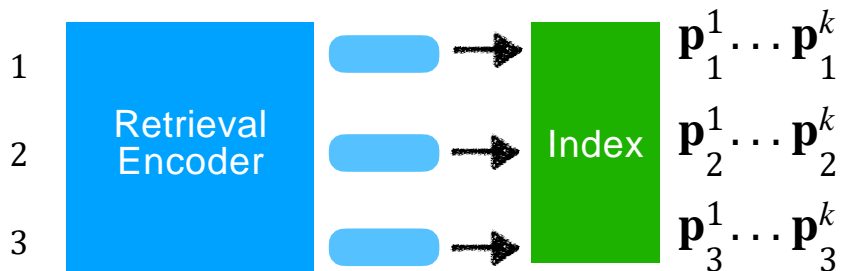
x = World Cup 2022 was the last with 32 teams, before the increase to

1

2

3

(k chunks of text per split)



Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

IR in the Middle of LM

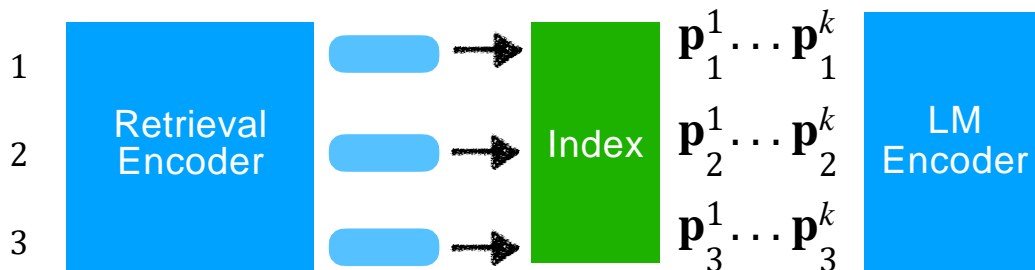
x = World Cup 2022 was the last with 32 teams, before the increase to

1

2

3

(k chunks of text per split)



IR in the Middle of LM



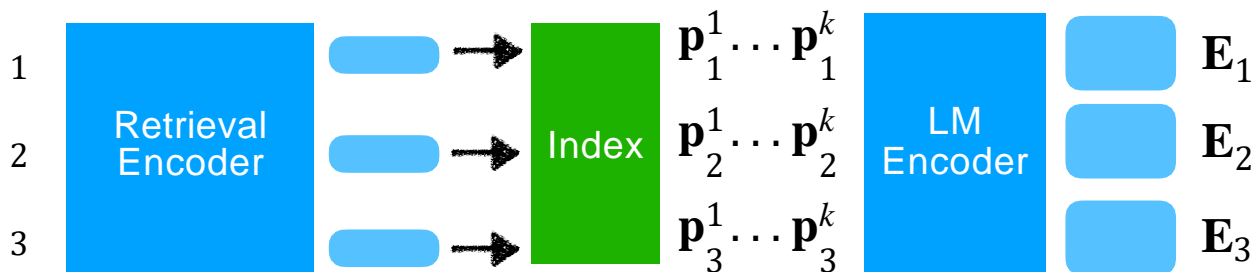
x = World Cup 2022 was the last with 32 teams, before the increase to

1

2

3

(k chunks of text per split)



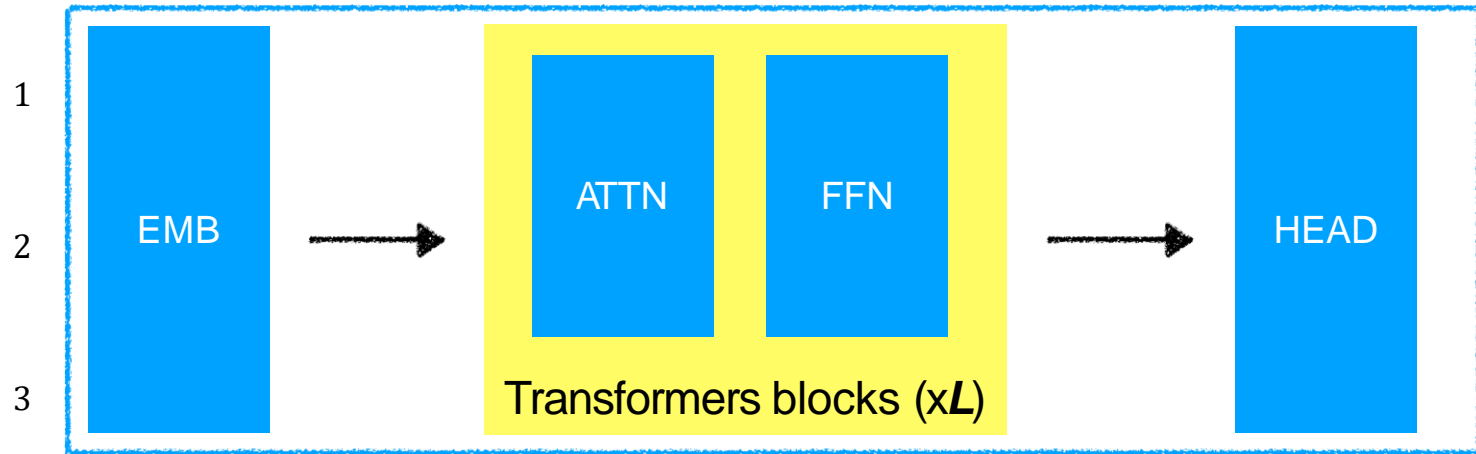
(A $r \times k \times d$ matrix)

(r = # tokens per text chunk)

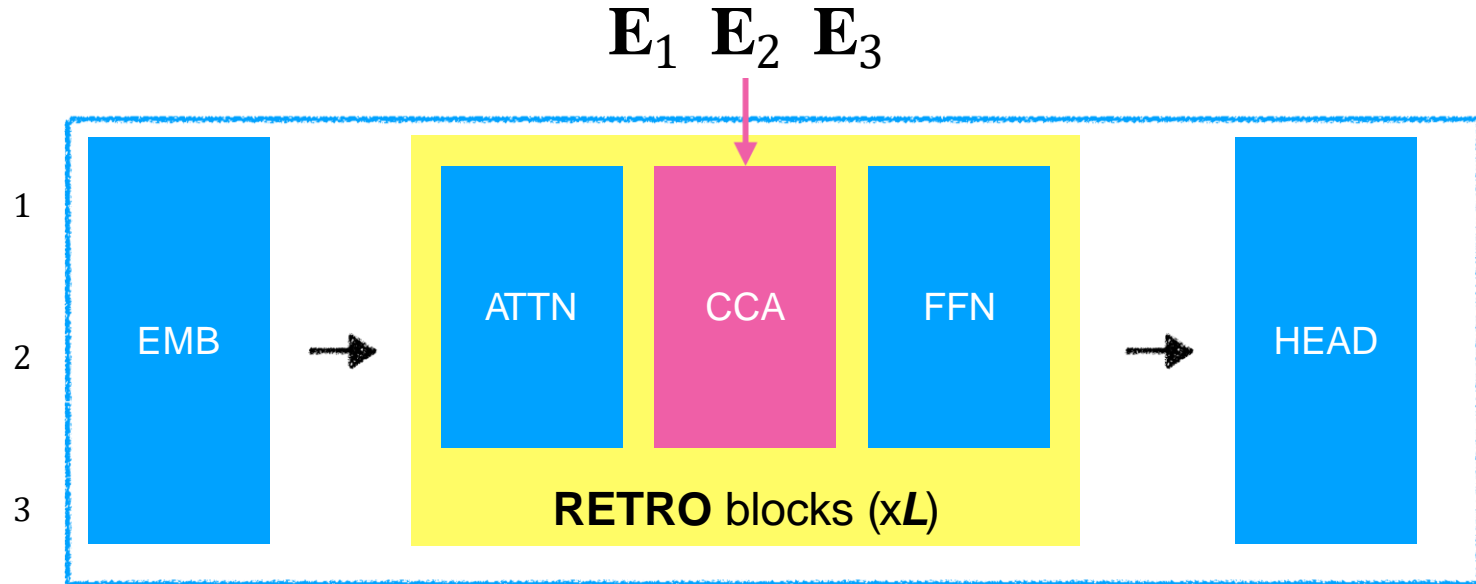
(d = hidden dimension)

(k = # retrieved chunks per split)

Regular Decoder

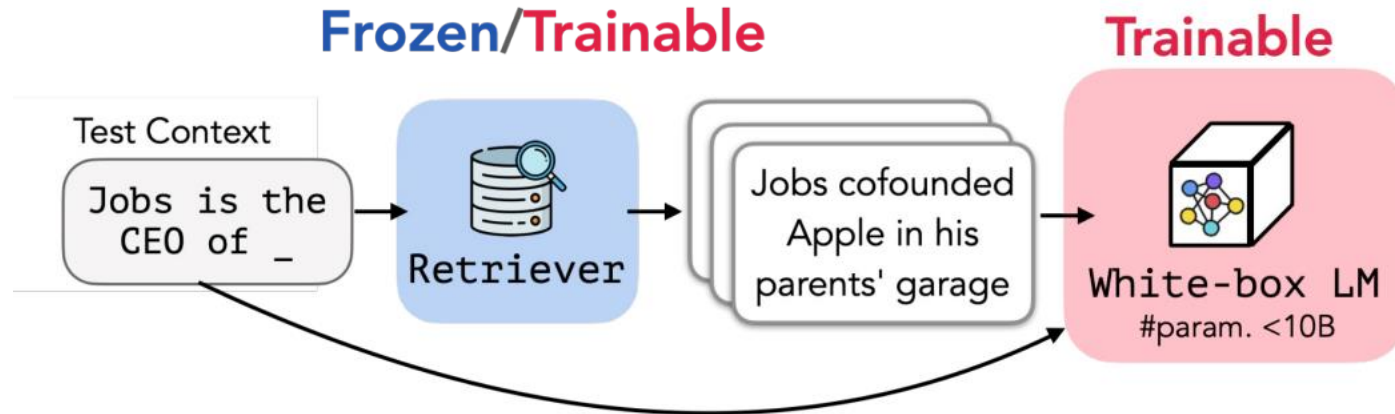


Regular Decoder with IR Embeddings



Chunked CrossAttention (CCA)

- There are various ideas in the literature for how to train these models efficiently and in an end-to-end fashion.



Main Takeaways

- How do we enable LMs to utilize external knowledge?
 - Retrieval-augmented language models
- A retriever is a function, $f(\text{input}, \text{memory}) \rightarrow \text{score}$
- What we did not discuss:
 - Attribution: Tracing decisions to the source knowledge
 - How to modify the knowledge
 - Conflicting knowledge
 - Editing knowledge
 - More efficient scaling
 -