

**ACL 2023 Tutorial:**

# **Retrieval-based Language Models and Applications**

Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen

University of Washington, Princeton University

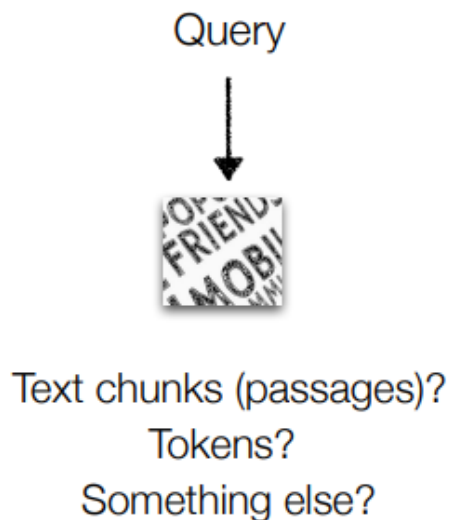
발표자: 송선영

2024/01/16

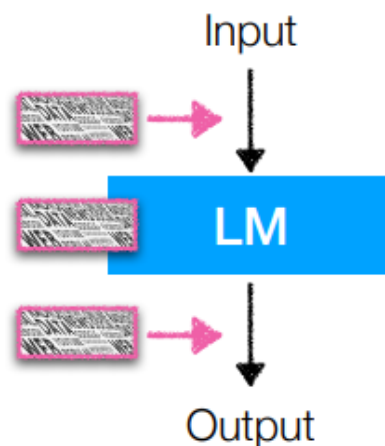
# Retrieval-based language model

- We can categorize retrieval-based language models based on **three key questions**
  - What do we retrieve from a datastore?
  - How do we incorporate retriever result to the language model?
  - When do we use retrieval?

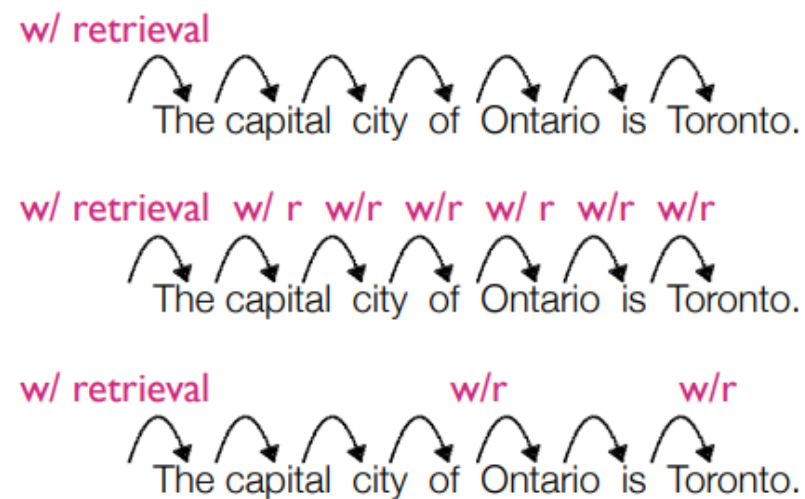
**What** to retrieve?



**How** to use retrieval?



**When** to retrieve?

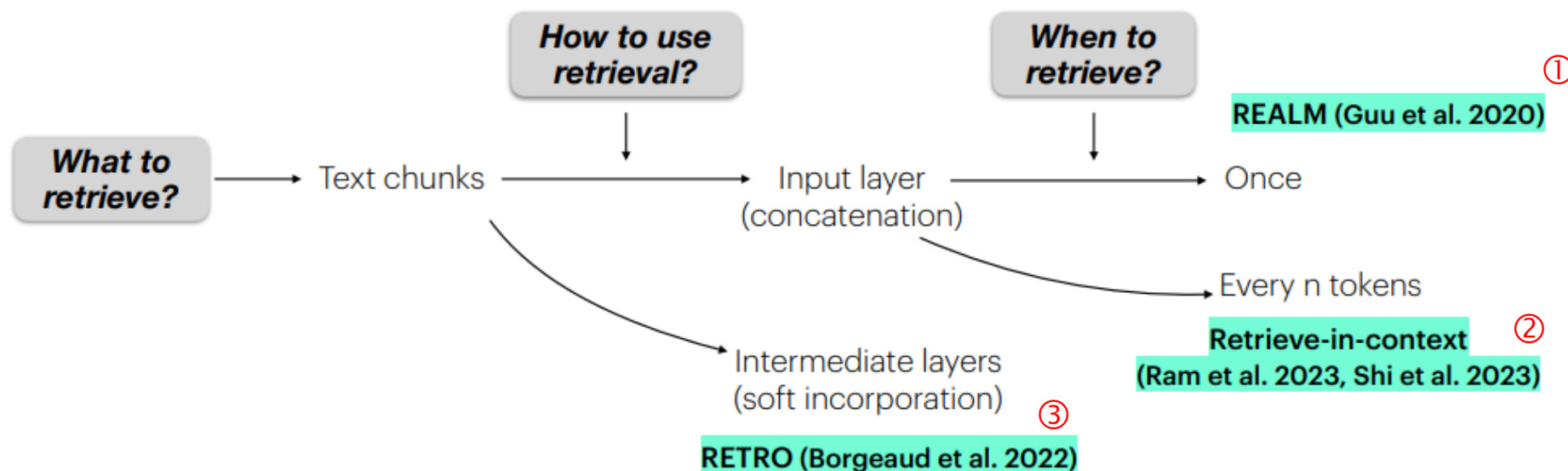


# Retrieval-based language model

- Road map of Key papers

- REALM
- Retrieve-in-context
- RETRO

	What do retrieve?	How to use retrieval?	When to retrieve?
REALM (Guu et al 2020)	Text chunks	Input layer	Once
Retrieve-in-context LM (Shi et al 2023, Ram et al 2023)	Text chunks	Input layer	Every n tokens
RETRO (Borgeaud et al. 2021)	Text chunks	Intermediate layers	Every n tokens



# REALM (Guu et al. 2020)

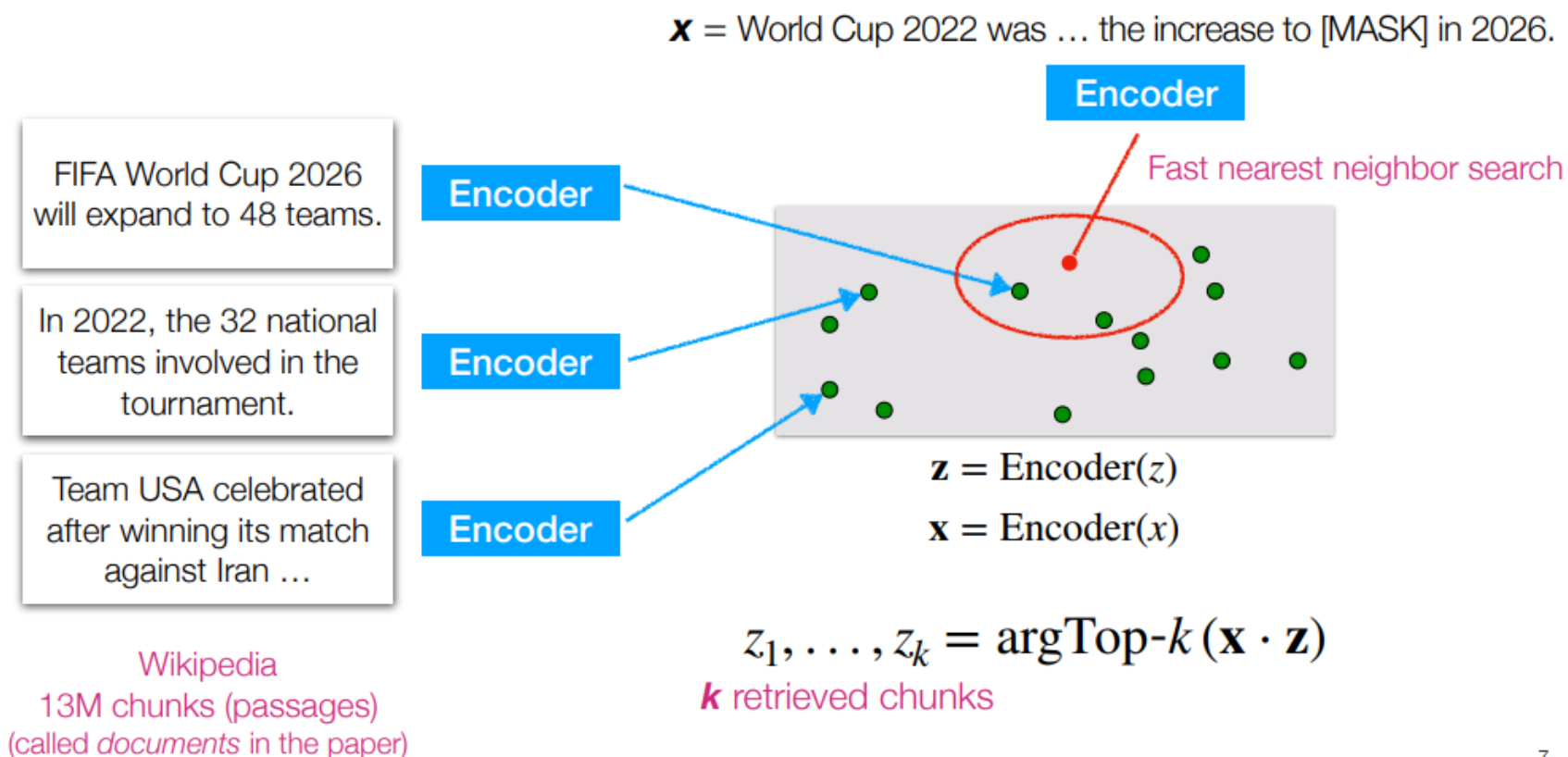
- A masked language model

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

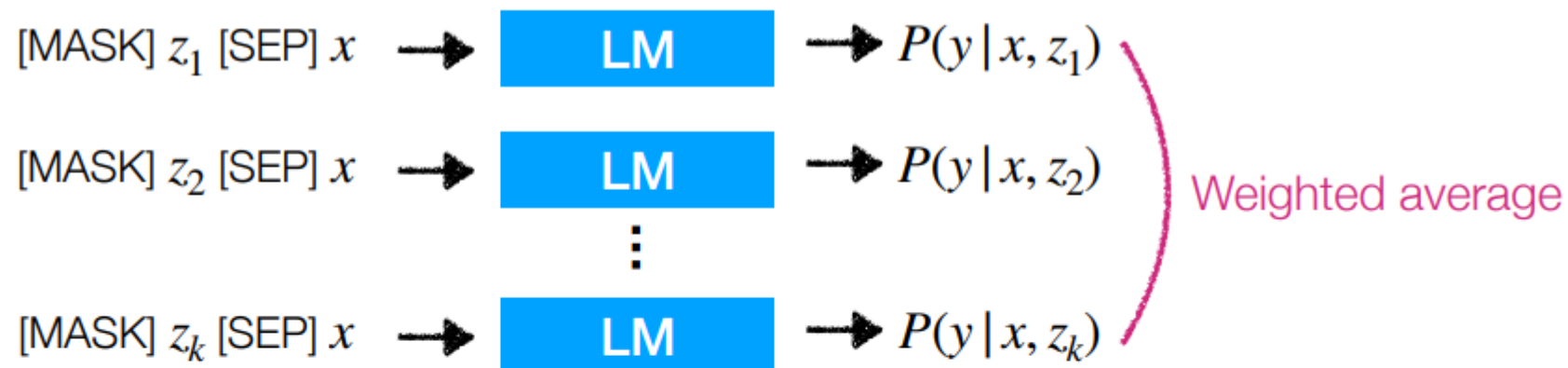


# REALM – Retrieve stage

- Datastore: 13 million text chunks of Wikipedia



# REALM – Read stage



Need to approximate  
 $\rightarrow$  Consider top  $k$  chunks only

$$\sum_{z \in \mathcal{D}} P(z | x) P(y | x, z)$$

$z \in \mathcal{D}$  from the retrieve stage

$P(y | x, z)$  from the read stage

0 if not one of top  $k$

# Retrieval-in-context LM (Ram et al. 2023, Shi et al. 2023)

- An auto-regressive language model

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

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

↓  
**Retrieval**  
↓

\* Can use multiple text blocks too (see the papers!)

FIFA World Cup 2026 will expand to 48 teams. World Cup 2022 was the last with 32 teams, before the increase to

↓  
**LM**  
↓

48 in the 2026 tournament.

# Retrieval-in-context LM - experiments

1. Is it necessary to use the entire prefix as an input to the retrieval system?

$x$  = Team USA celebrates after winning its match against Iran at Al Thumama Stadium in Group B play of the FIFA World Cup 2022 on Nov. 29, 2022. (..) World Cup 2022 was the last with 32 teams, before the increase to

Team USA celebrates after winning its match against Iran at Al Thumama Stadium in Group B play of the FIFA World Cup 2022 on Nov. 29, 2022. (..) World Cup 2022 was the last with 32 teams, before the increase to



Retrieval

The U.S. national team defeated Iran 1-0.

Does not cover "tokens that will come next"

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



Retrieval

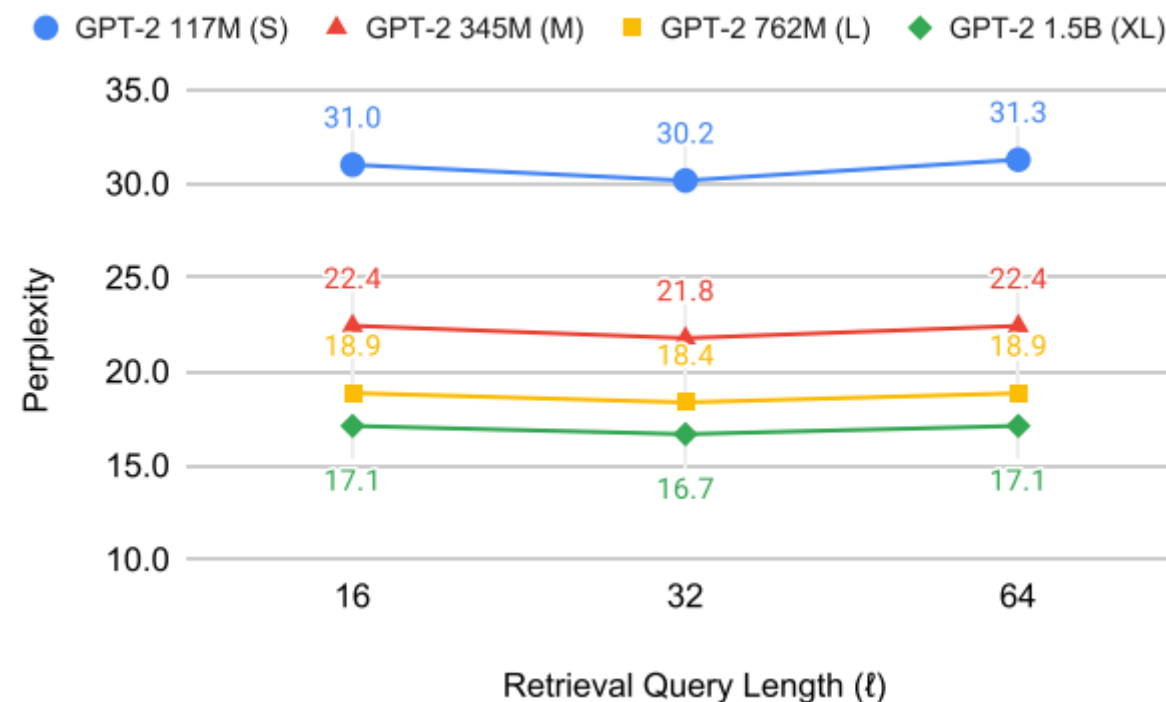
FIFA World Cup 2026 will expand to 48 teams.

more relevant to what will come next



# Retrieval-in-context LM - experiments

1. Is it necessary to use the entire prefix as an input to the retrieval system?



→ Shorter prefix (more recent tokens) as a query helps

# Retrieval-in-context LM - experiments

---

## 2. How frequent should we use retrieval?



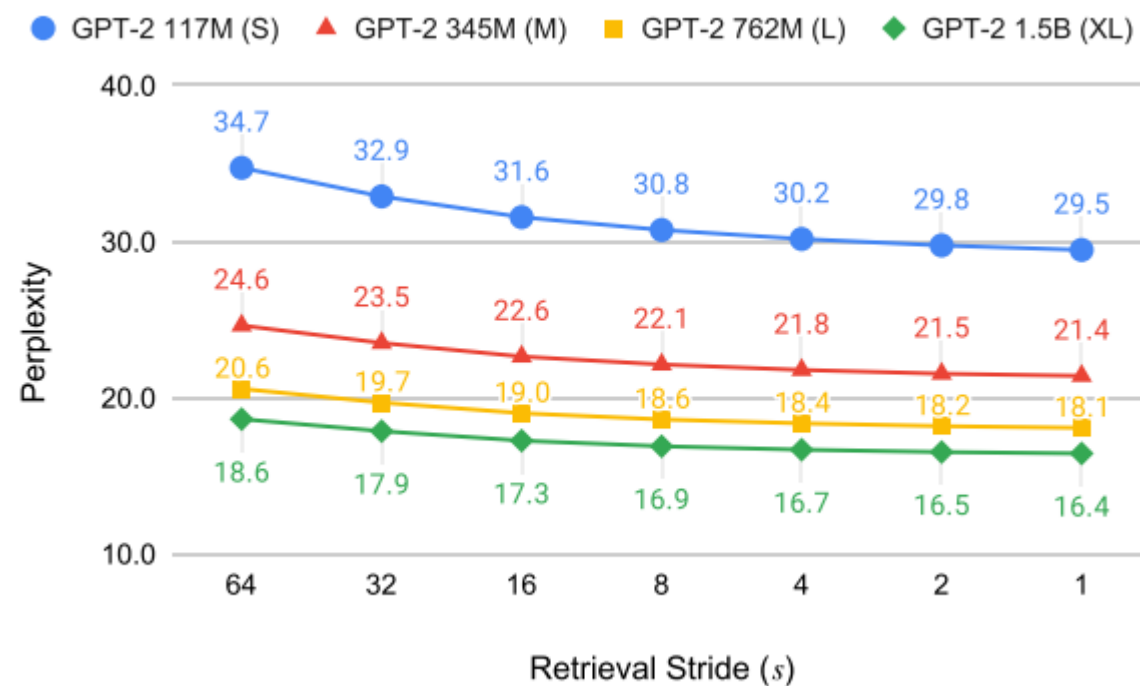
# Retrieval-in-context LM - experiments

## 2. How frequent should we use retrieval?



# Retrieval-in-context LM - experiments

2. How frequent should we use retrieval?

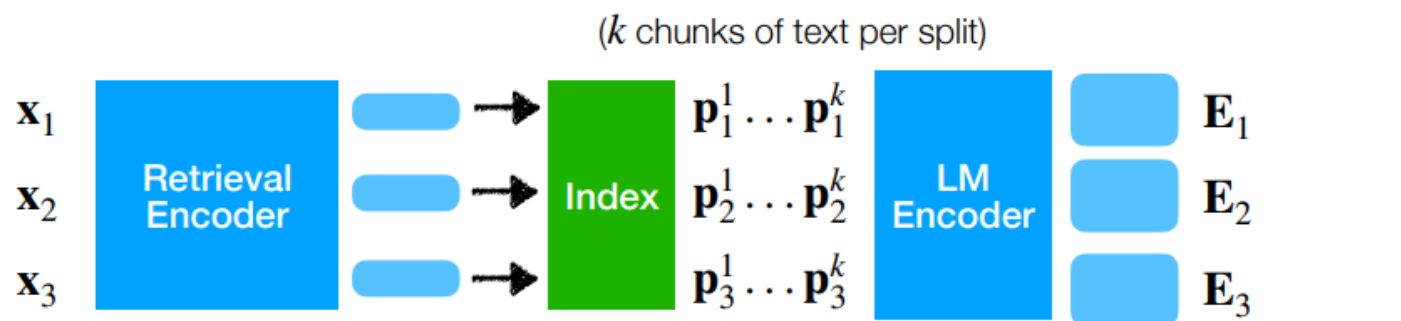


# RETRO (Borgeaud et al. 2022)

- An auto-regressive language model
- Designed for many chunks, frequently, more efficiently
- Scale the datastore (1.8T tokens)

$\mathbf{x}$  = World Cup 2022 was the last with 32 teams, before the increase to

$\mathbf{x}_1$   $\mathbf{x}_2$   $\mathbf{x}_3$



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

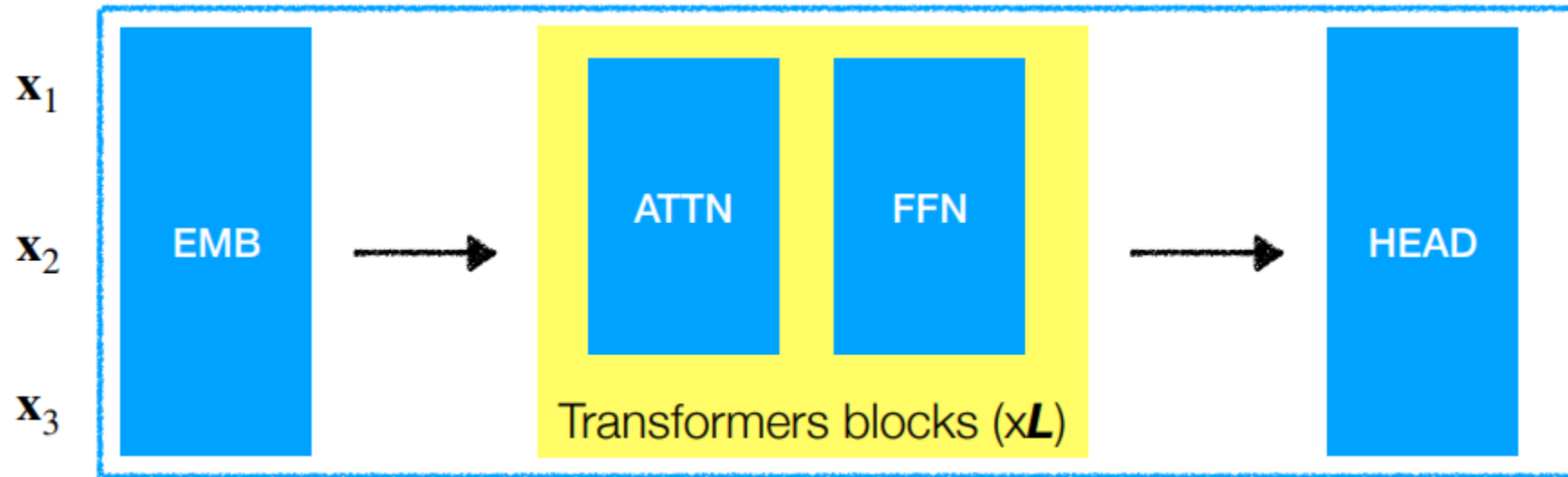
( $r$  = # tokens per text chunk)

( $d$  = hidden dimension)

( $k$  = # retrieved chunks per split)

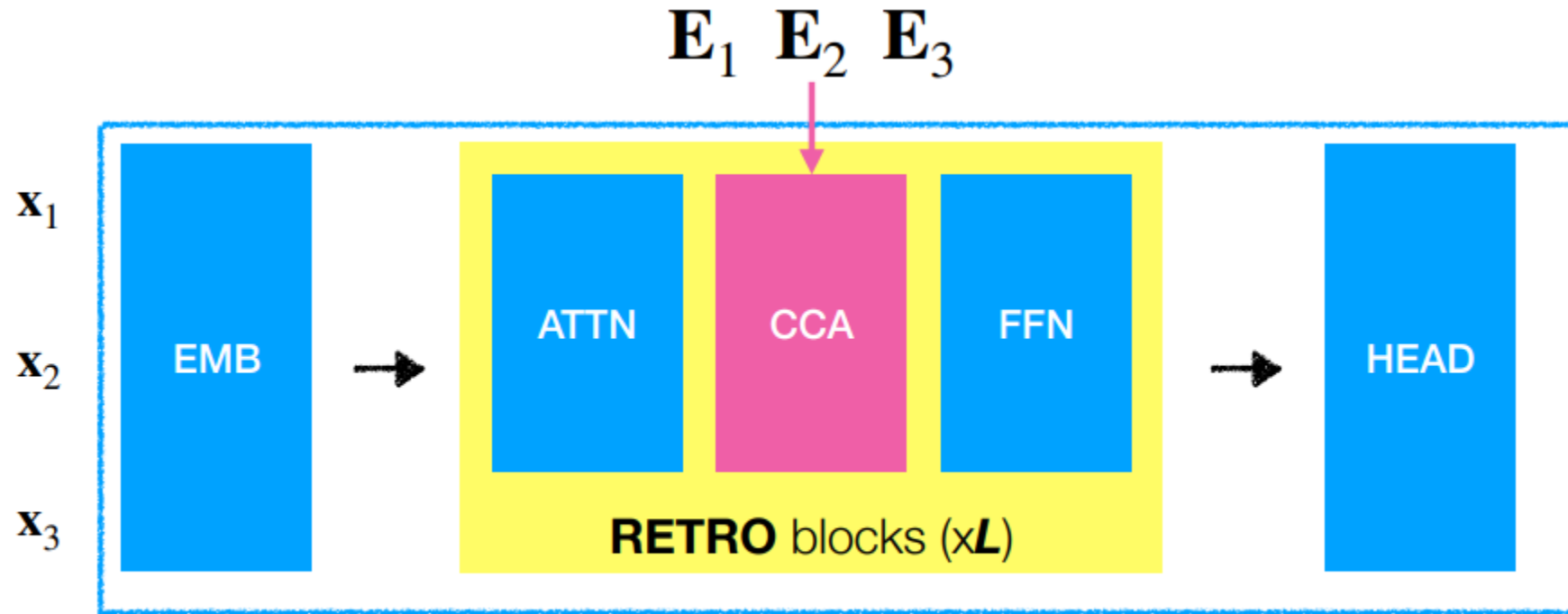
# Decoder in RETRO

- Regular decoder



# Decoder in RETRO

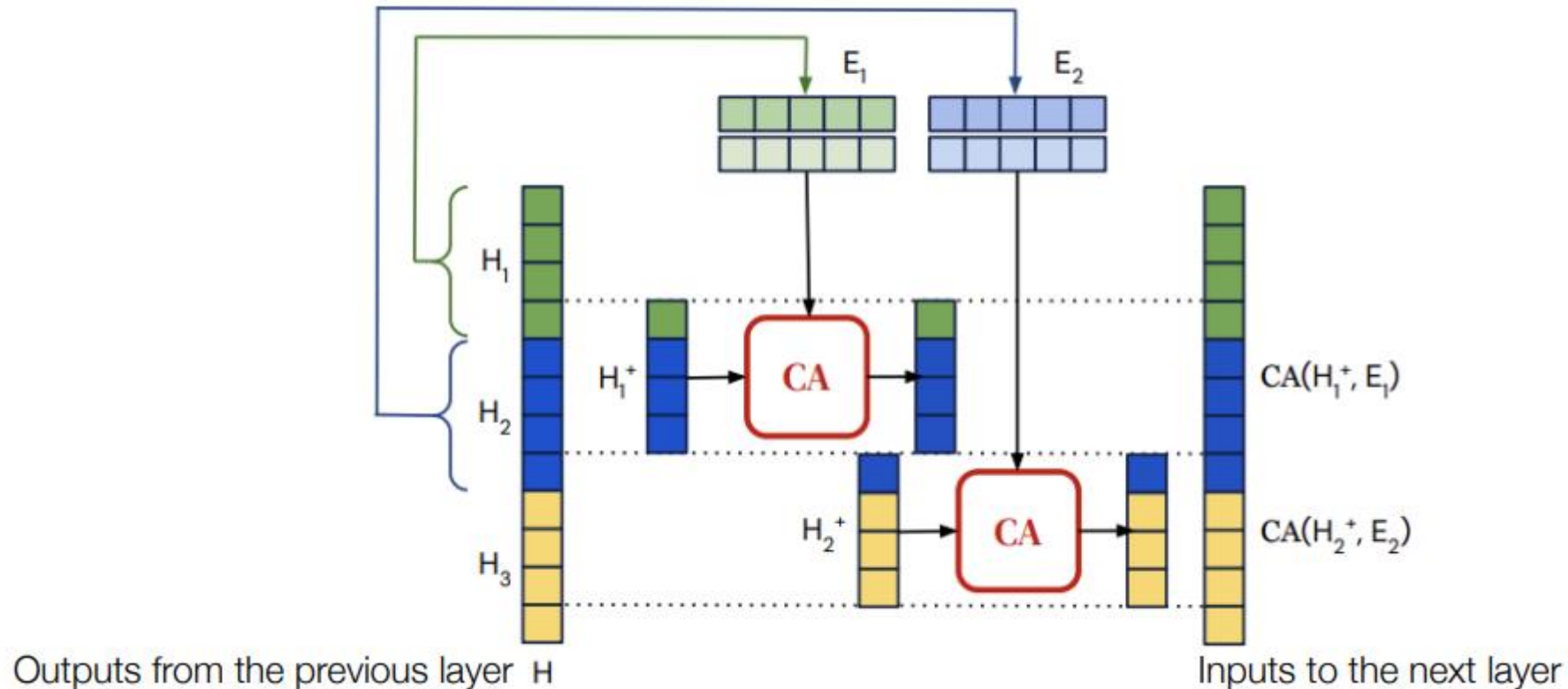
- RETRO decoder



Chunked Cross Attention (CCA)

# Decoder in RETRO

- What is the CAA(Chunked Cross Attention) ?
  - To compute a cross-attention between the retriever result and the representation from the previous layer





# Thank You

---

감사합니다.