

# Improving language models by retrieving from trillions of tokens

Mikhail Domanin

# Retrieval Augmented Generation (RAG)

- We separate linguistic information and knowledge about the world
- Knowing the language is enough:

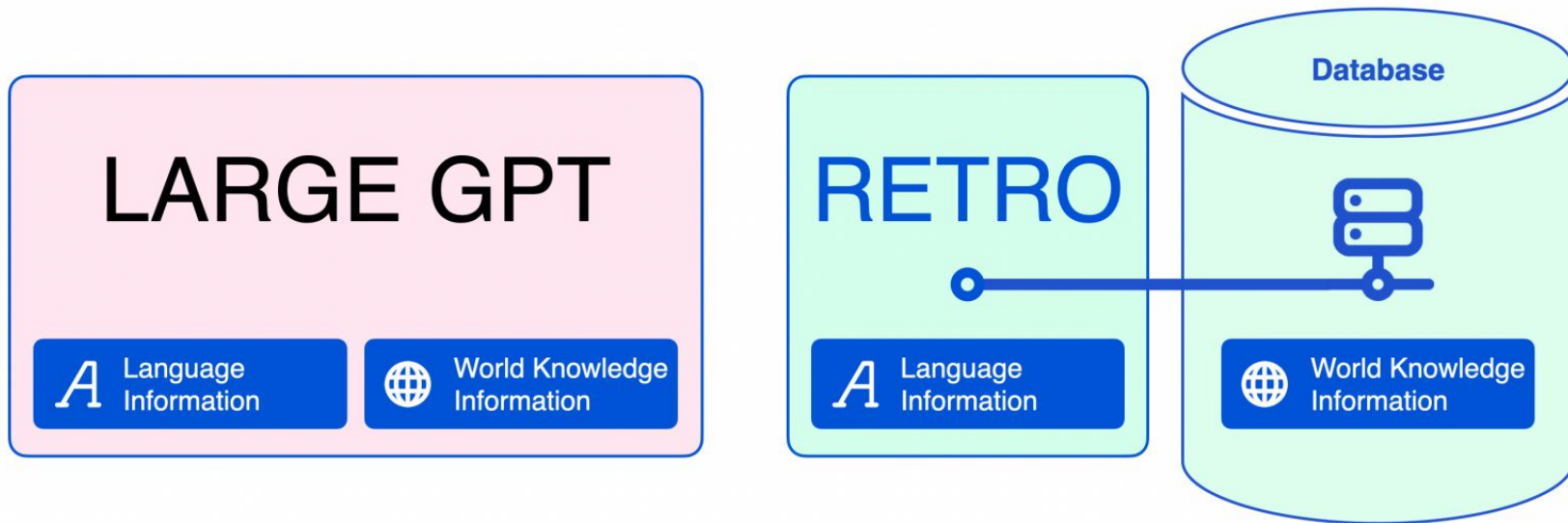
its popularity spread by word-of-mouth to  
allow Herbert to start working full \_\_\_\_\_

- Need information from additional sources:

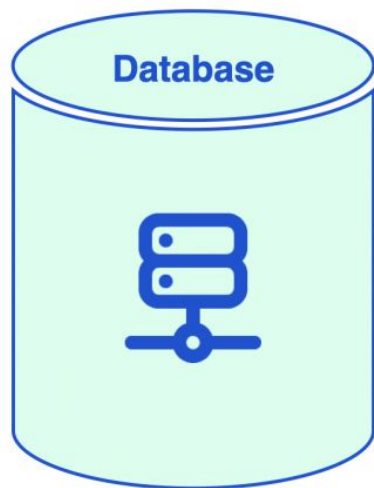
The Dune film was released in \_\_\_\_\_

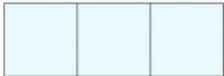
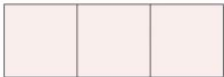
# Retrieval-Enhanced TRansfOrmer (RETRO)

- Text quality is comparable to GPT-3
- Uses 7.5 billion parameters, instead of 185 billion for GPT-3 Da Vinci (that is, 4%)

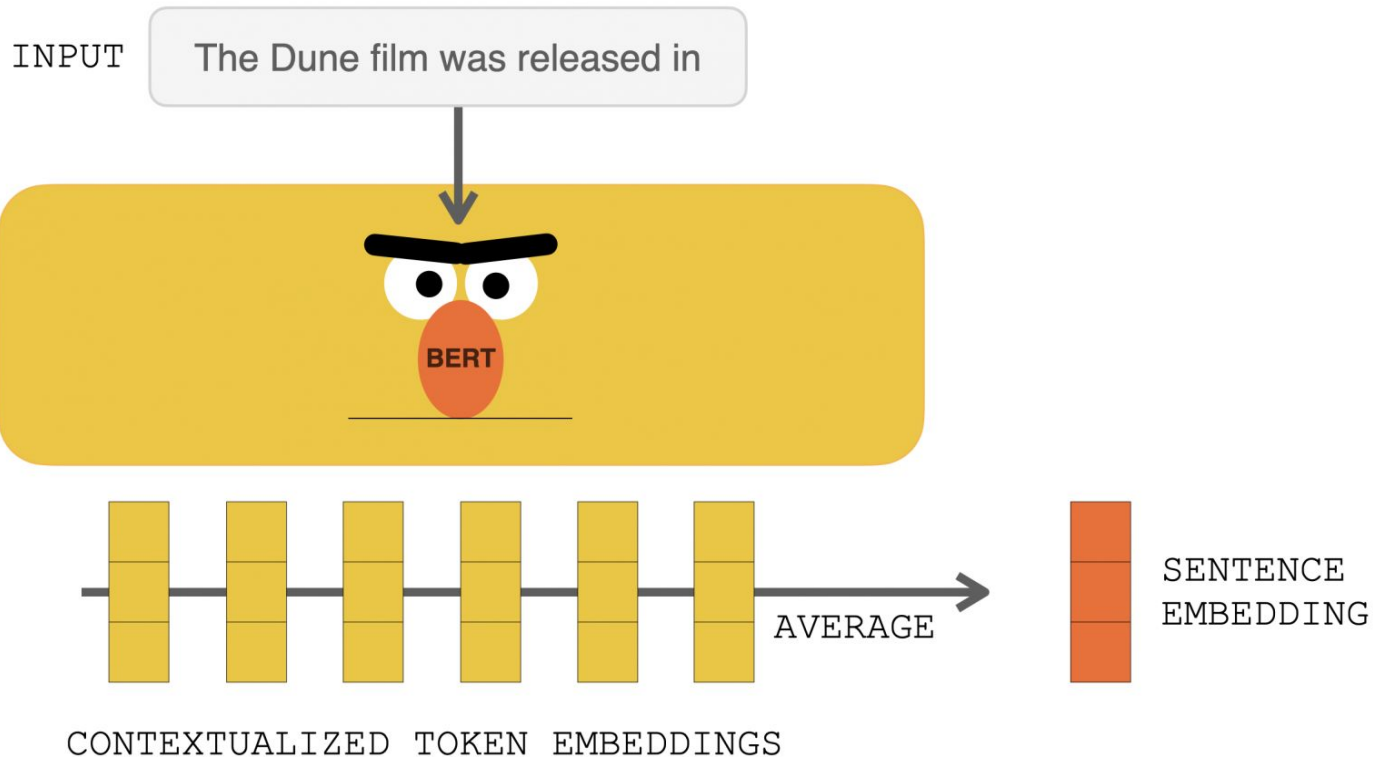


# How to build database?



Key (BERT sentence embedding)	Value (text. neighbor and completion chunks. Each up to 64 tokens in length)	
	Dune is a 2021 American epic science fiction film directed by Denis Villeneuve	NEIGHBOR
	It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert	COMPLETION
	Dune is a 1965 science fiction novel by American author Frank Herbert	NEIGHBOR
	originally published as two separate serials in Analog magazine	COMPLETION
...	...	

# How to find k-nearest



# How to find k-nearest

- Use Scann to find k-nearest
- Time complexity:  $O(\log T)$ ;  $T$  - dictionary size
- 10 ms to find k-nearest

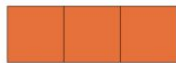
# Example

INPUT

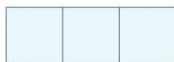
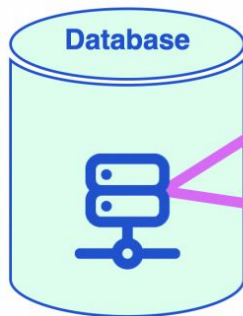
The Dune film was released in

## 1) EMBED WITH BERT

SENTENCE  
EMBEDDING



2) QUERY  
approximate  
nearest  
neighbor



Nearest Neighbor 1

Dune is a 2021 American epic science fiction film directed by Denis Villeneuve

It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert



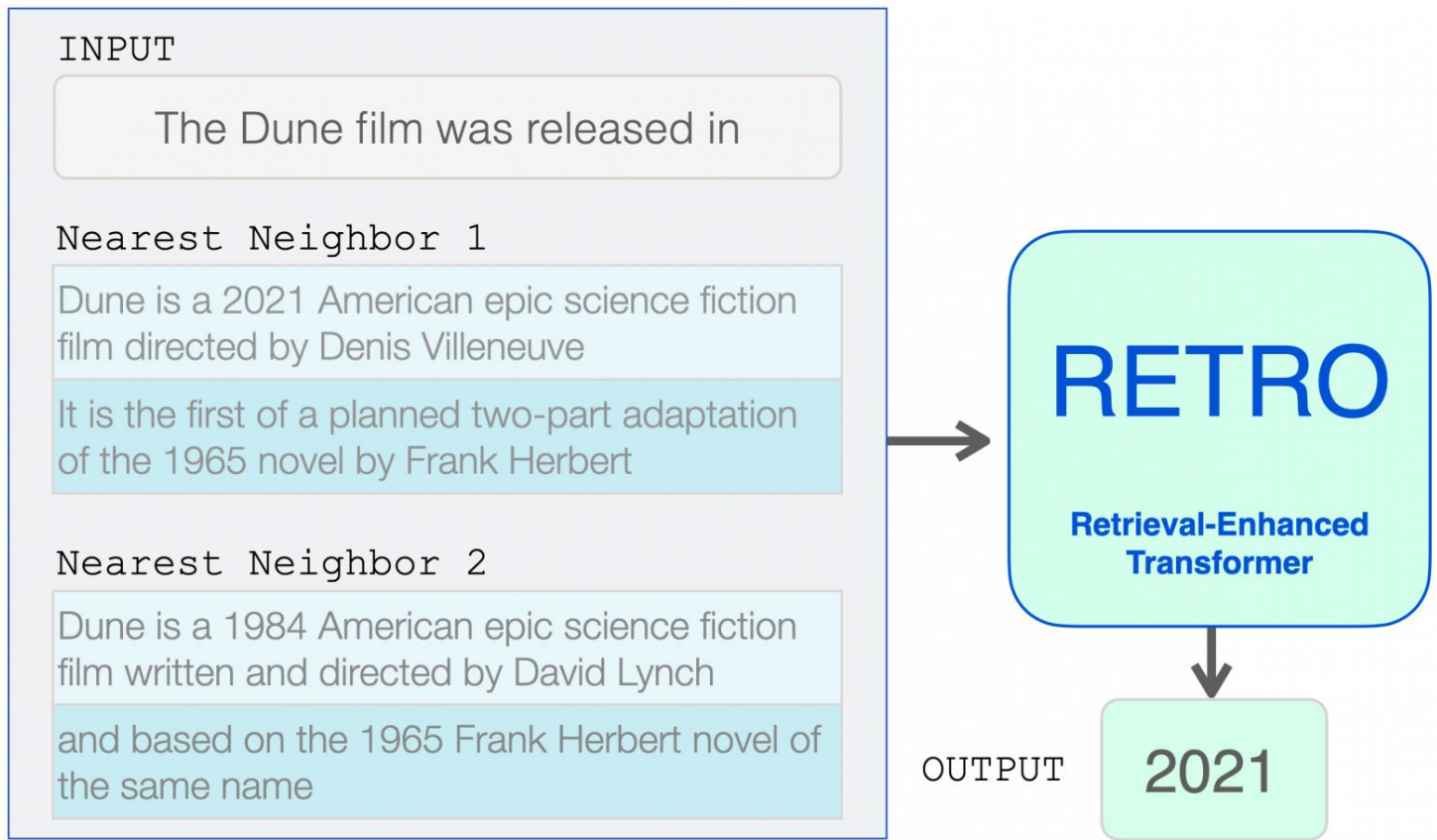
Nearest Neighbor 2

Dune is a 1984 American epic science fiction film written and directed by David Lynch

and based on the 1965 Frank Herbert novel of the same name

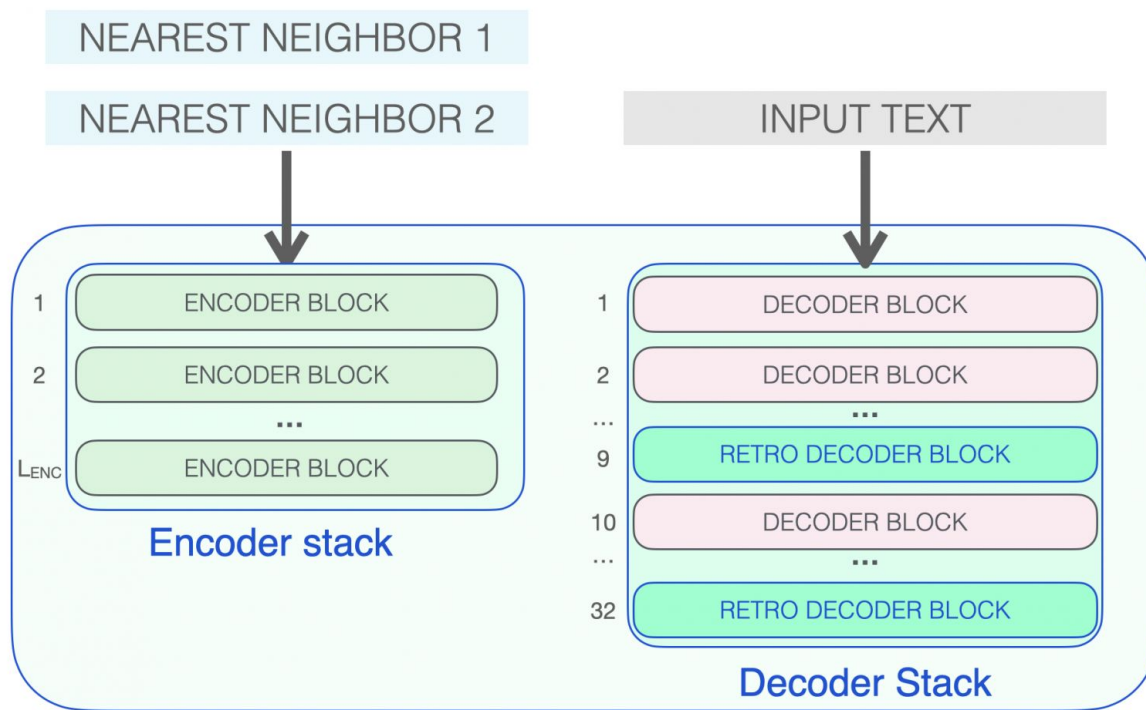
## 2) RETRIEVE

# Example





# RETRO Architecture



RETRO Transformer

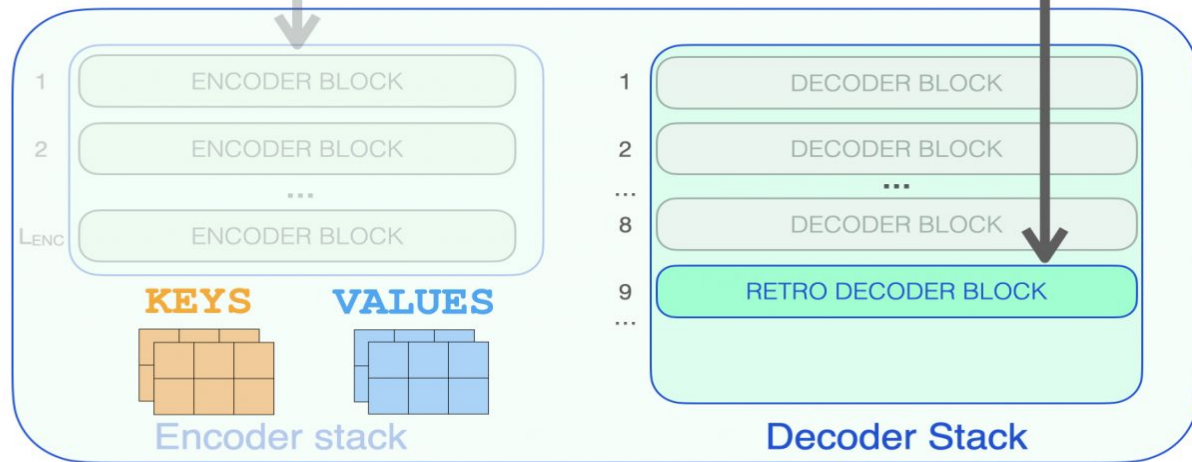
# RETRO architecture

- Encoder block: self-attention + FFNN
- Decoder block: ATTN + FFNN
- RETRO decoder block: ATTN + Chunked Cross Attention (CCA) + FFNN

NN1 Dune is a 2021 American epic ...

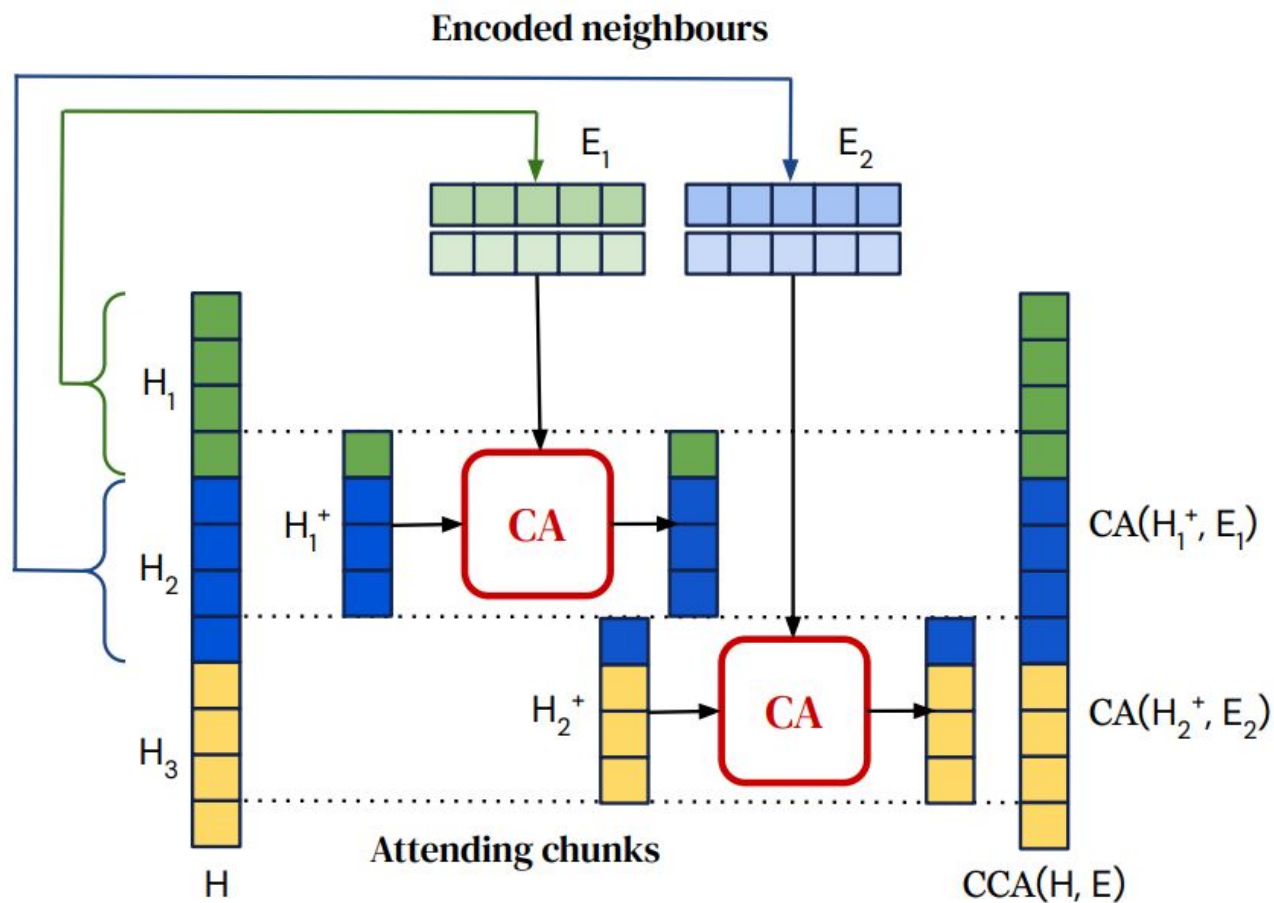
NN2 Dune is a 1984 American epic ...

The Dune film was released in



RETRO Transformer

## Chunked cross-attention (CCA)



# Comparable architectures

## Baseline Transformer

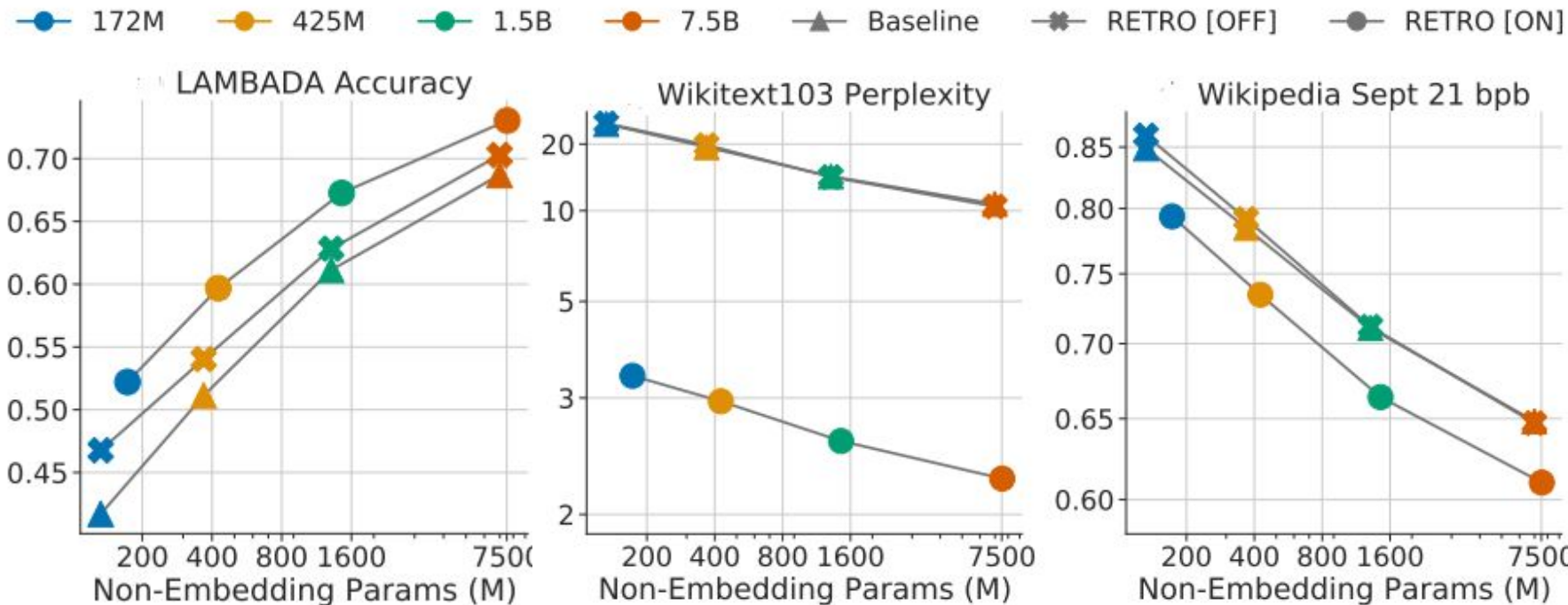
- Replace LayerNorm with RMSNorm
- Relative position encodings

## RETRO [Off]

- Without retrieval data

## RETRO [On]

# Some results (Top 1 result)



# Advantages

- Easily add new information
- It's possible to use internet as a database
- Much fewer parameters => easier to fit

# Literature

- <https://arxiv.org/abs/2112.04426> - The main article
- <https://habr.com/ru/articles/648705/> - Short russian adoption of the main article
- <https://www.promptingguide.ai/techniques/rag> - An article about RAG