# Improving language models by retrieving from trillions of tokens

Ilya German

# LLM

- Many weights
- Long-time learning
- Retraining or tuning with new data

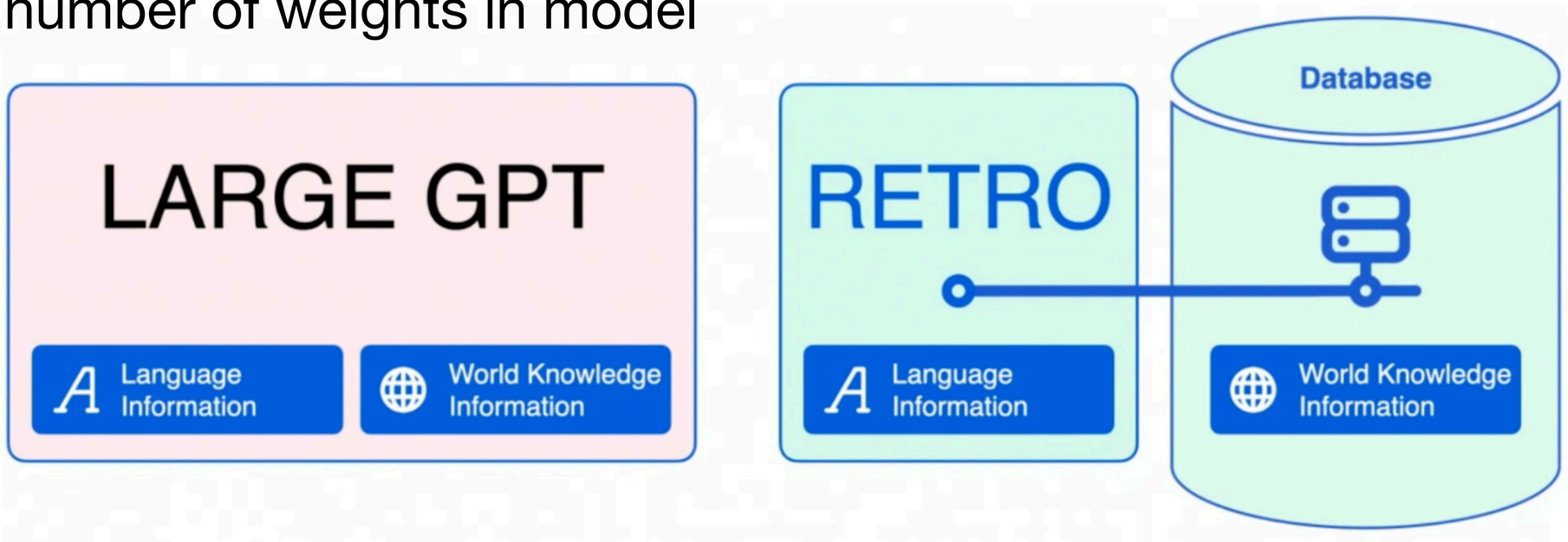
# How to generate next token

- London is a capital of Great _____ .
- You can go on a paid internship at Yandex and work there part-_____ .

- The Dune: Part Two film was released in _____ .
- Spring in 1985 was _____ . (early)

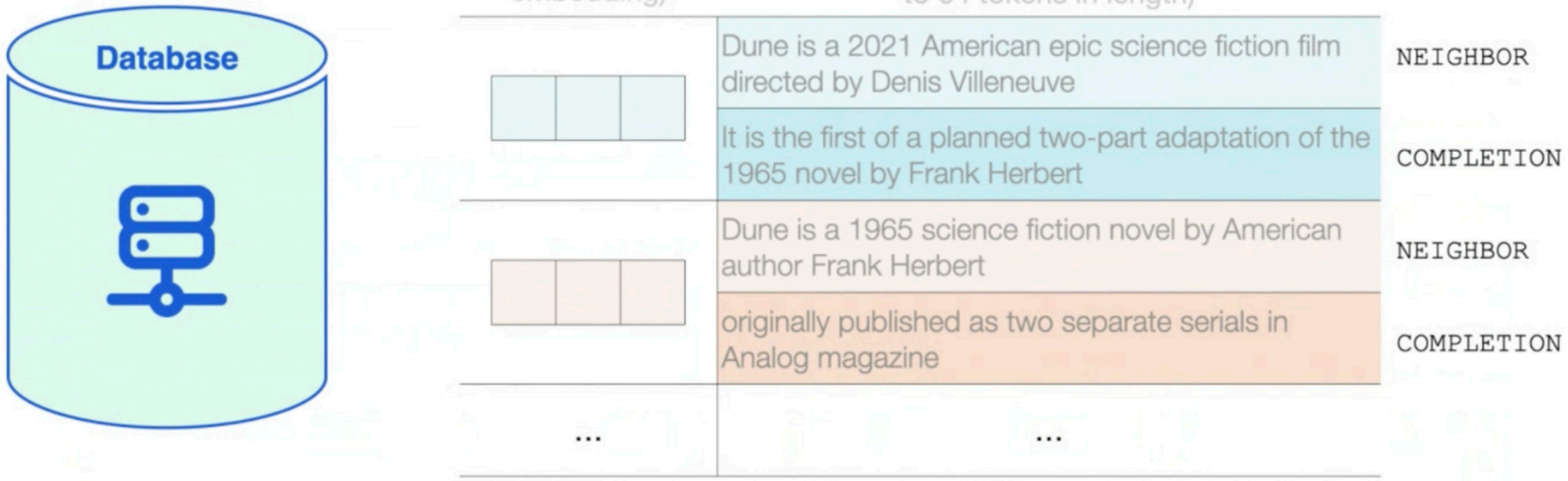
# RETRO (Retrieval-Enhanced TRansfOrmer)

- Working with external DB.
- Independent model training and database updates
- Lower number of weights in model



# DataBase

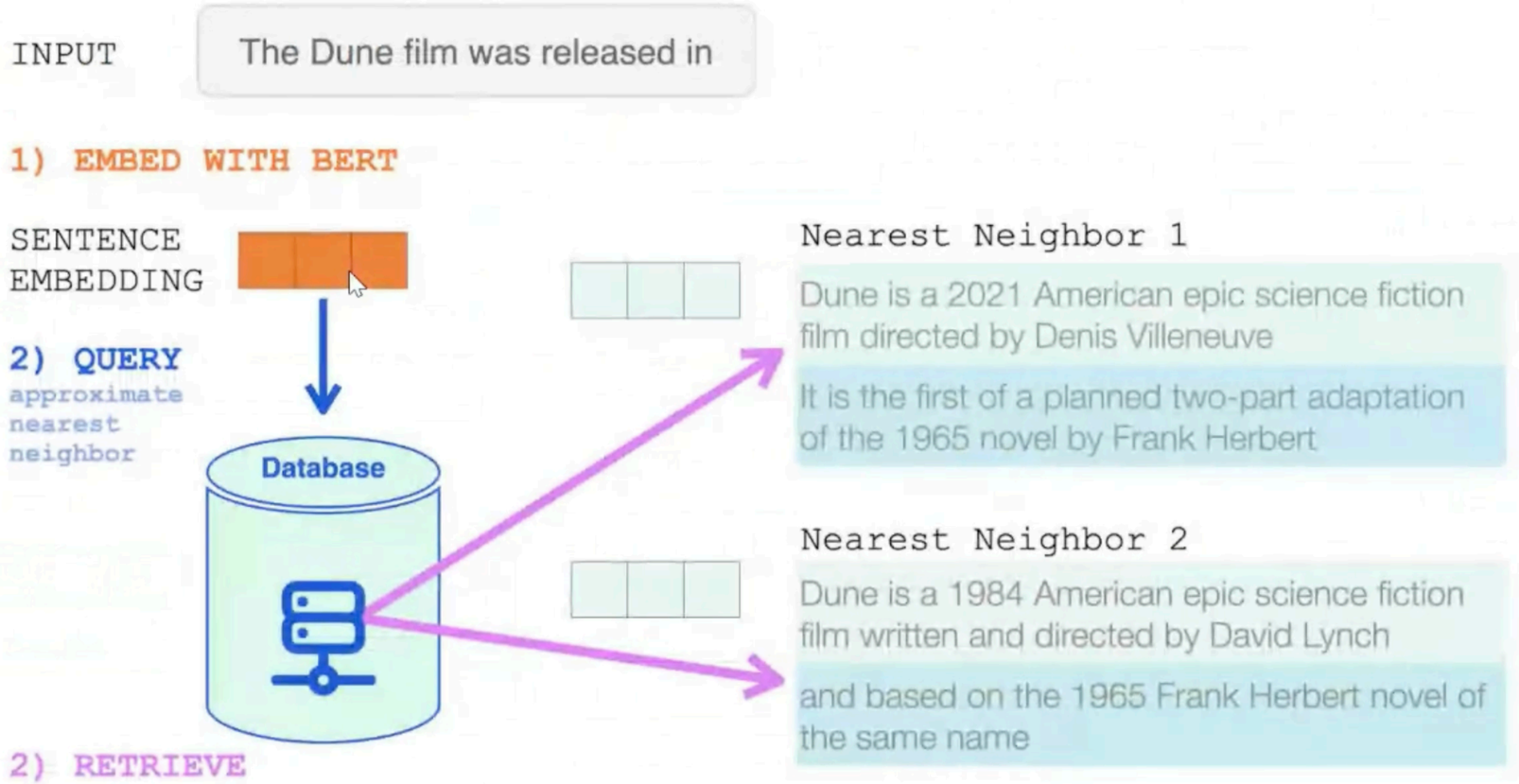
- Split text into chunks
- Download chunks in key-value database



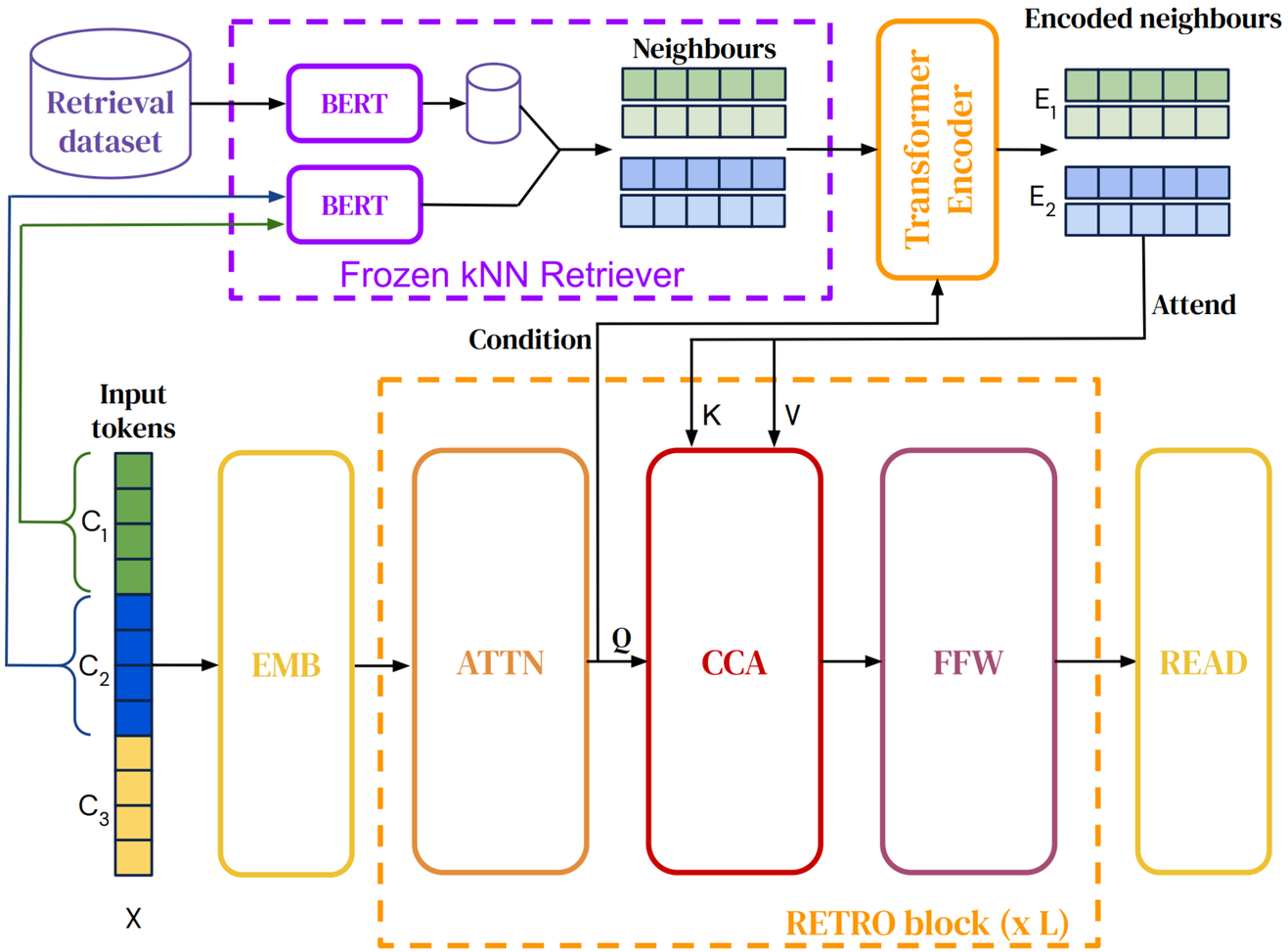
# Scann

- Almost binary tree
- Fast finding k-nearest keys in dictionary
- O(log T)-heuristics finding nearest key
- Rebuilding dataset
- No hash-map!!!

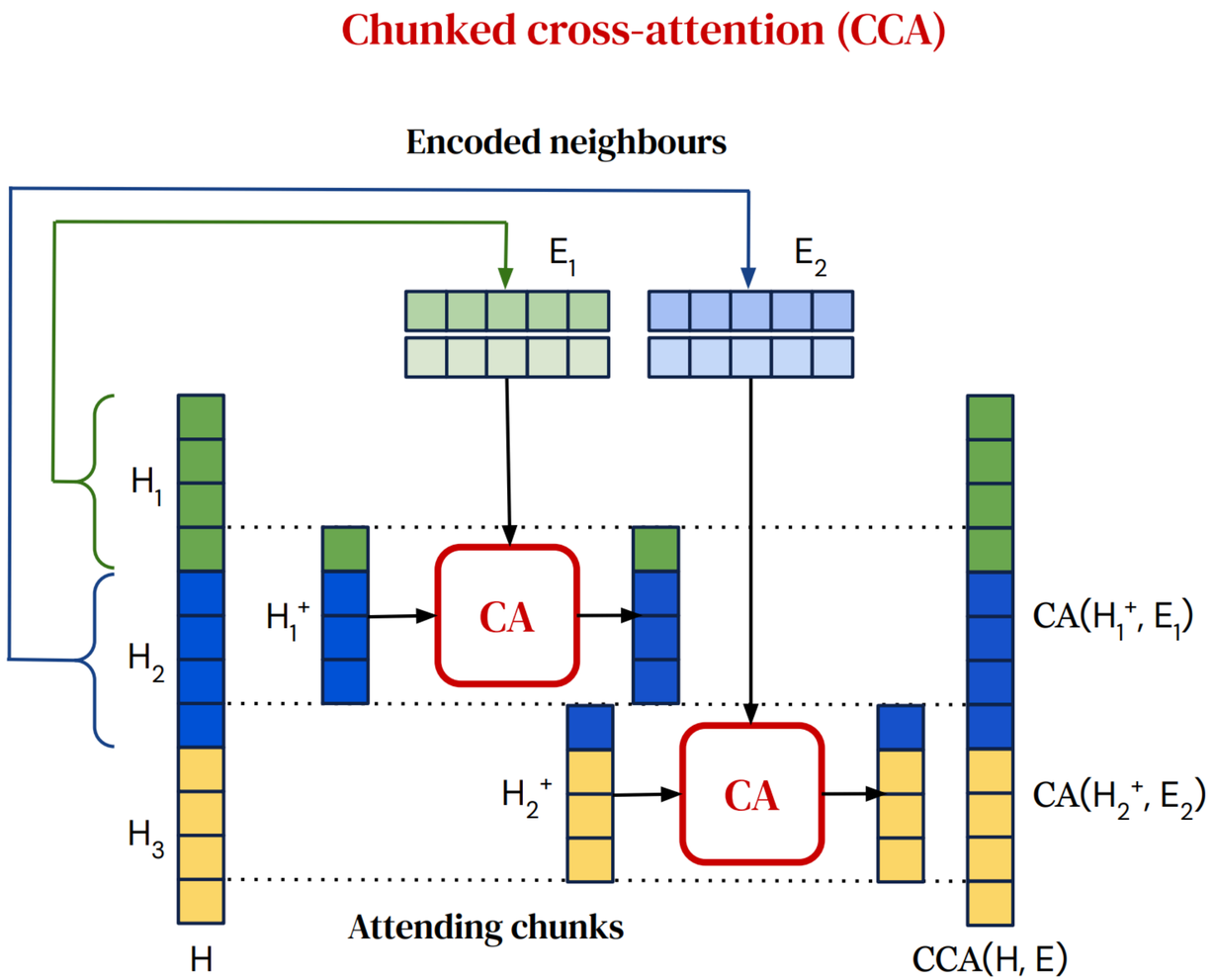
# Summing-up


INPUT
The Dune film was released in
1) EMBED WITH BERT
SENTENCE EMBEDDING
2) QUERY
approximate nearest neighbor
Database
2) RETRIEVE
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

# Transformer architecture



# Transformer architecture



# Architecture comparison

## Baseline Transformer

## RETRO

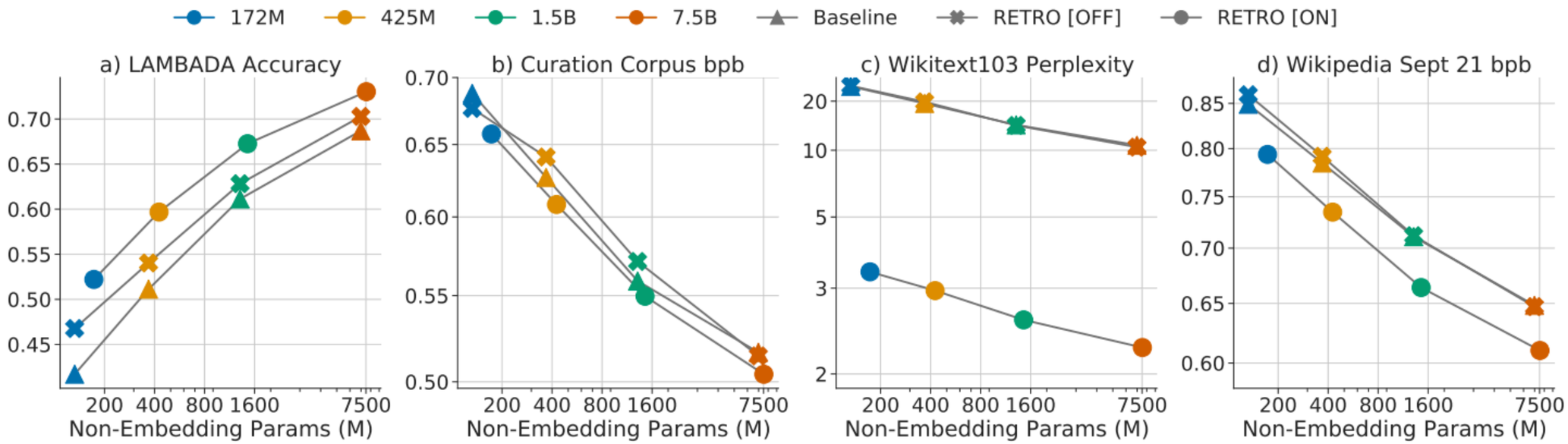



Figure 3 | **Scaling with respect to model size.** (a) LAMBADA top-1 accuracy. (b) Evaluation loss on curation corpus. (c) Perplexity on Wikitext103 valid. (d) Bits-per-byte on selected Wikipedia articles from September 2021.

# Results

- Results comparable with GPT-3 Da Vinci
- 7.5 billions parameter (vs 185b in GPT-3)
- Independent model learning and database updating
- Ability to add internet

# References


- “*Improving language models by retrieving from trillions of tokens.”* DeepMind, Sebastian Borgeaud , Arthur Mensch , Jordan Hoffmann. Link

- https://habr.com/ru/articles/648705/