Improving language models by retrieving from trillions of tokens

Borgeaud et. al.

By Ilya Lasy

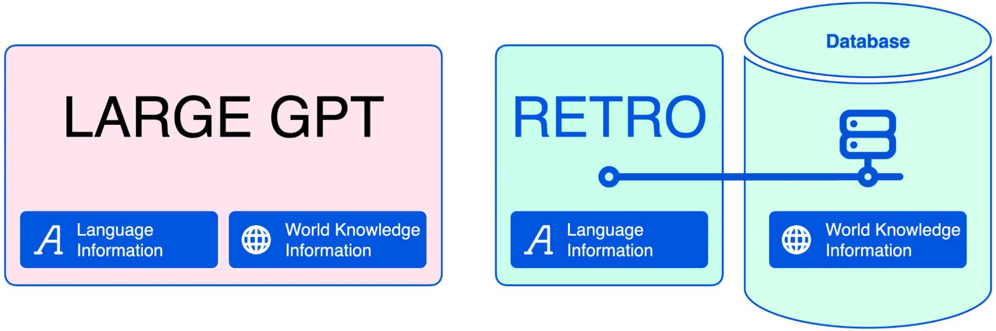
Outline

1. Motivation
2. Method
3. Related Work
4. Results
5. Limitations

Motivation

Decoupling “*Knowledge*” and “*Language information*” in Large Language Models.

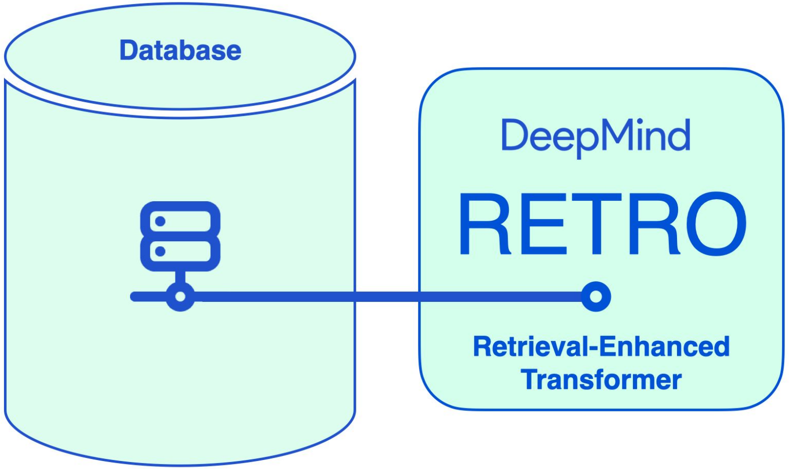
- Efficiency (smaller amount of parameters with same performance)
- Interpretability
- Controllable generation
- Domain adaptation



Method

RETRO (Retrieval-Enhanced Transformer)

1. Key-value database
2. Encoder-decoder
3. Chunked Cross-attention

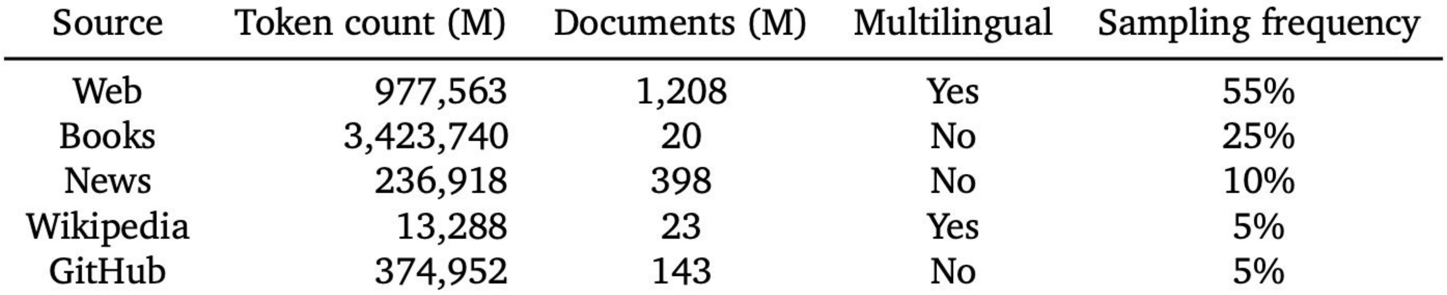


Dataset

MassiveText (Rae at. al.) - 5T tokens

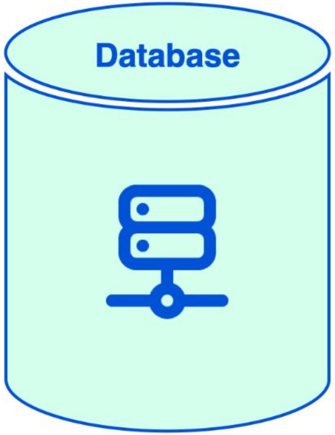
Retrieval during *training*: 600B tokens

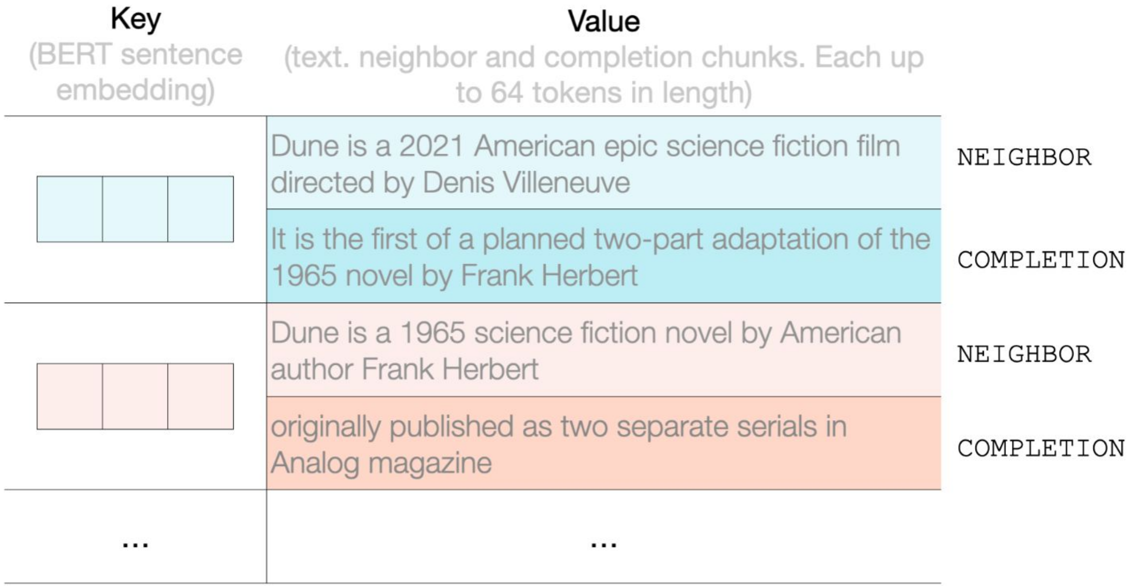
Retrieval during *evaluation*: up to 1.75T tokens



Source	Token count (M)	Documents (M)	Multilingual	Sampling frequency
Web	977,563	1,208	Yes	55%
Books	3,423,740	20	No	25%
News	236,918	398	No	10%
Wikipedia	13,288	23	Yes	5%
GitHub	374,952	143	No	5%

Retrieval Database

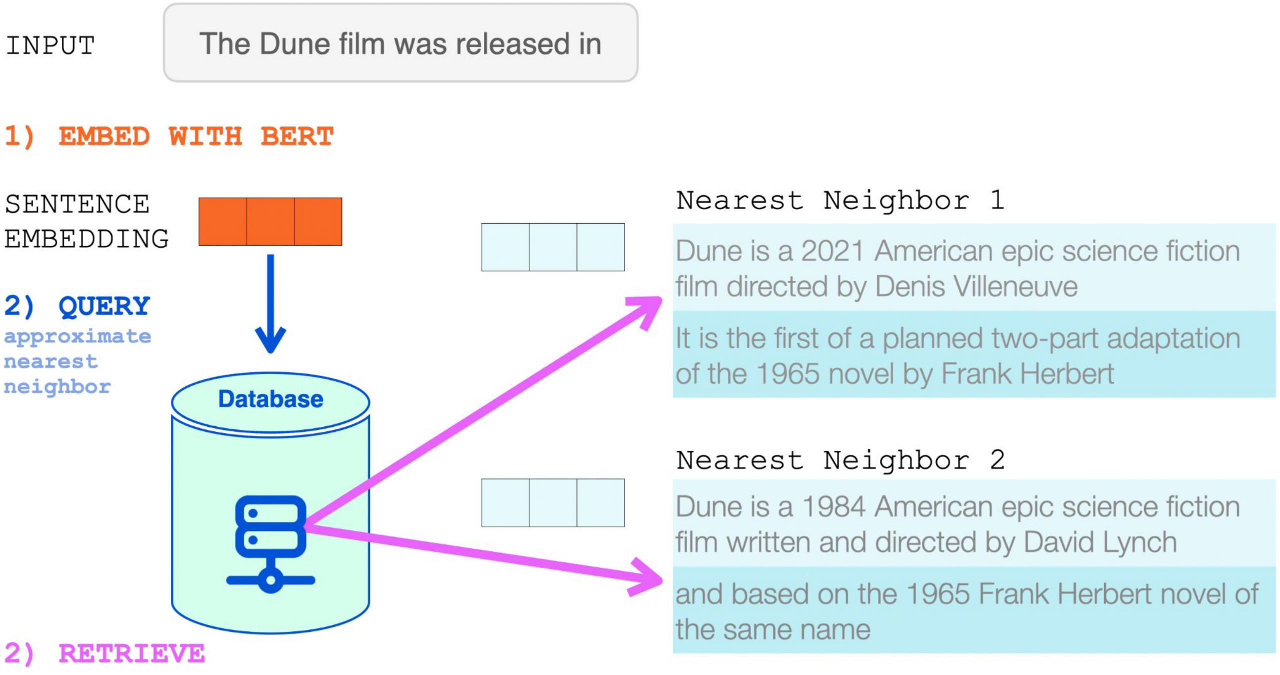




Key (BERT sentence embedding)	Value (text. neighbor and completion chunks. Each up to 64 tokens in length)	
□□□	Dune is a 2021 American epic science fiction film directed by Denis Villeneuve	NEIGHBOR
	It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert	COMPLETION
□□□	Dune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

Nearest neighbour retrieval

Faiss, SCaNN, etc.



Nearest Neighbor 1

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

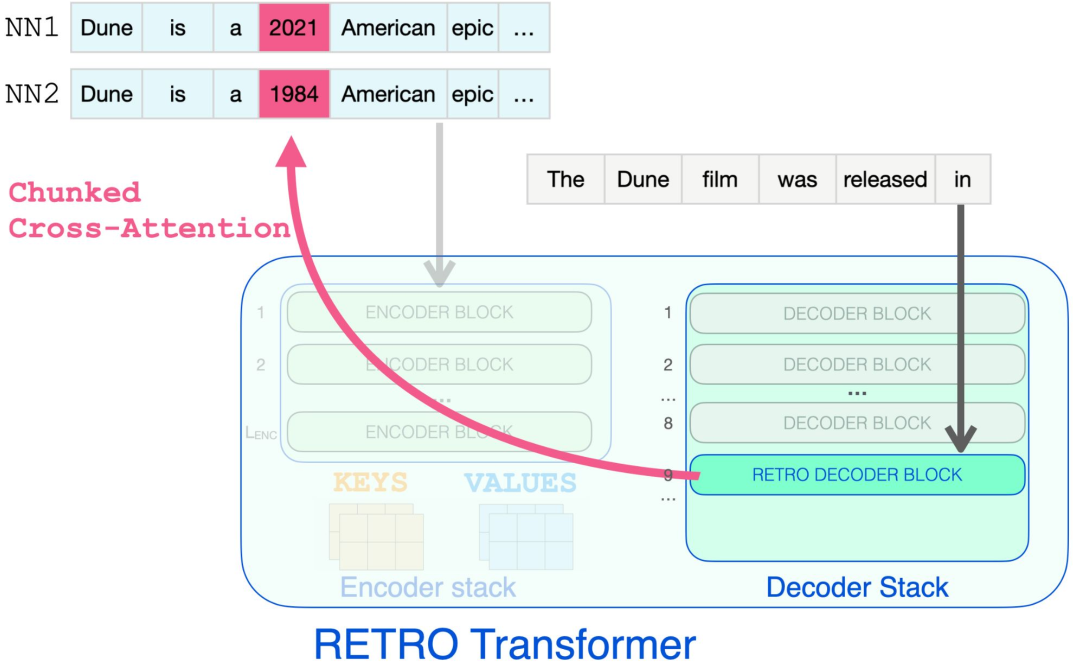
It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert

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

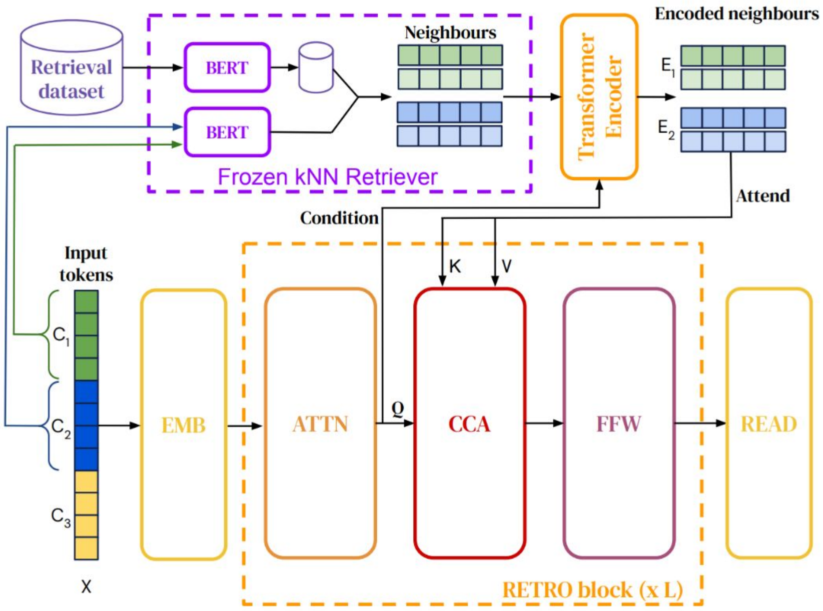
High-level architecture

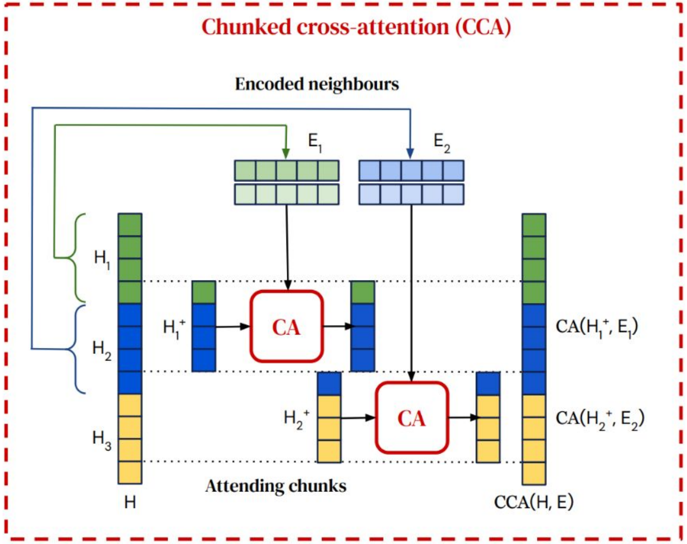


Encoder: non-autoregressive

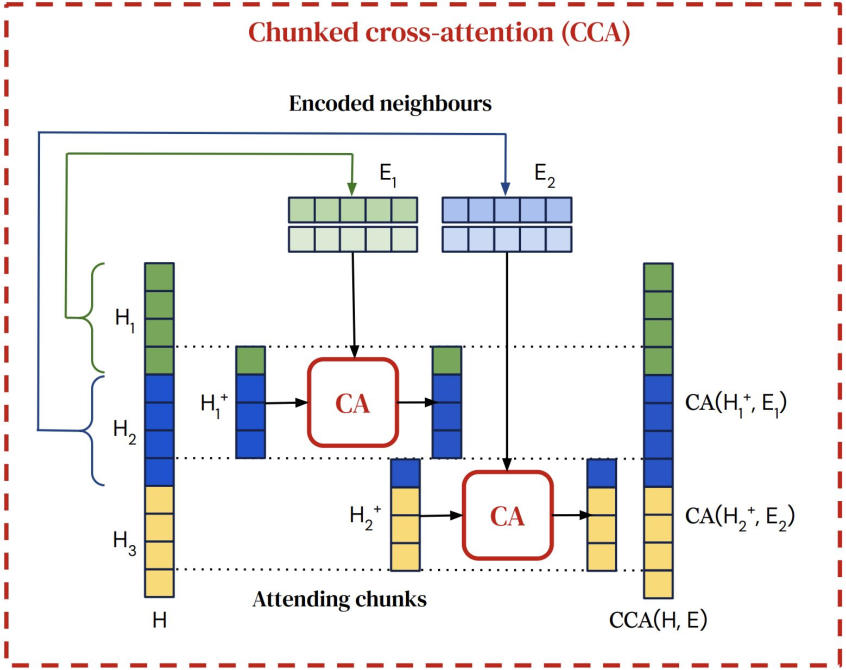
Decoder: autoregressive

Detailed architecture





Decoder block



H_i *- hidden activation of input tokens*
E_i *- encoded retrieved neighbors*
CA *- cross-attention*

Retro-fitting

~10% of weights, ~3% of dataset

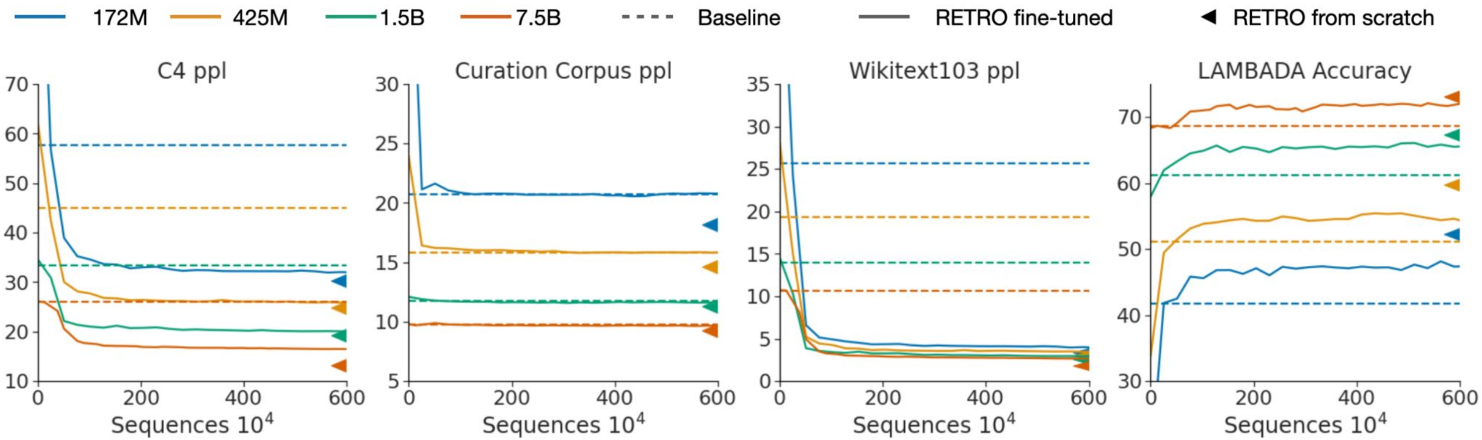
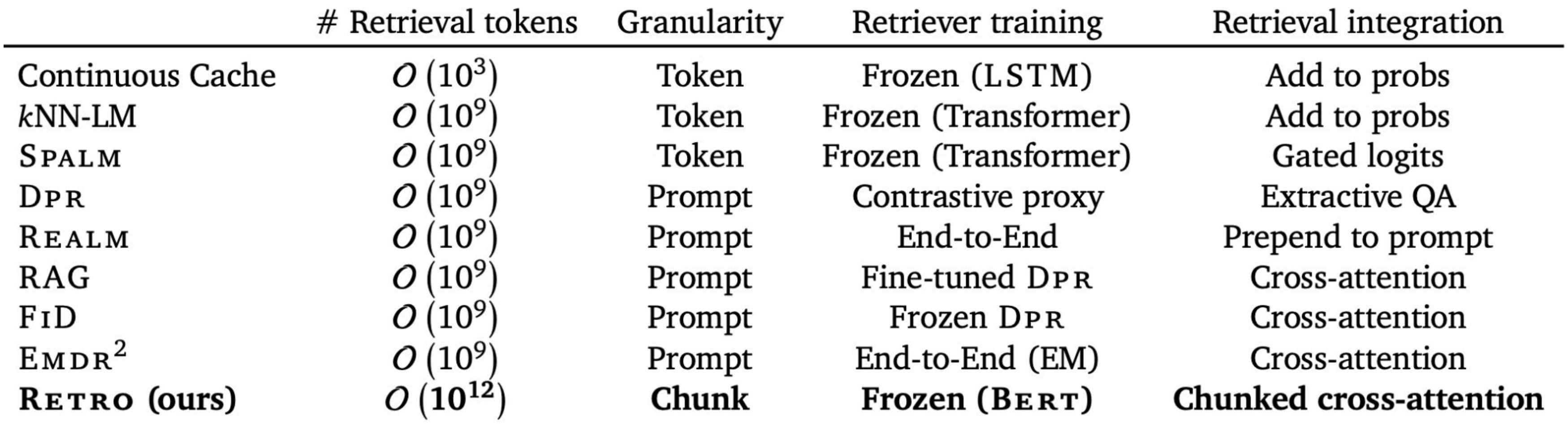


Figure: Retro-fitting a baseline transformer

Related work



	# Retrieval tokens	Granularity	Retriever training	Retrieval integration
Continuous Cache	$\mathcal{O}(10^3)$	Token	Frozen (LSTM)	Add to probs
*k*NN-LM	$\mathcal{O}(10^9)$	Token	Frozen (Transformer)	Add to probs
SPALM	$\mathcal{O}(10^9)$	Token	Frozen (Transformer)	Gated logits
DPR	$\mathcal{O}(10^9)$	Prompt	Contrastive proxy	Extractive QA
REALM	$\mathcal{O}(10^9)$	Prompt	End-to-End	Prepend to prompt
RAG	$\mathcal{O}(10^9)$	Prompt	Fine-tuned DPR	Cross-attention
FID	$\mathcal{O}(10^9)$	Prompt	Frozen DPR	Cross-attention
EMDR^2	$\mathcal{O}(10^9)$	Prompt	End-to-End (EM)	Cross-attention
RETRO (ours)	$\mathcal{O}(\mathbf{10^{12}})$	**Chunk**	**Frozen (BERT)**	**Chunked cross-attention**

Table: Comparison of Retro with existing retrieval approaches.

Evaluation metric: bits-per-bytes

Bits-per-byte (bpb) measures the average number of bits required to predict the next token in a sequence, which is usually a byte.

The **lower** the bpb value - the more **efficient** the model in making predictions, i.e., it requires fewer bits to predict the next token.

$$L \times \log_2(e)$$

where ***L*** is the cross-entropy loss

Results

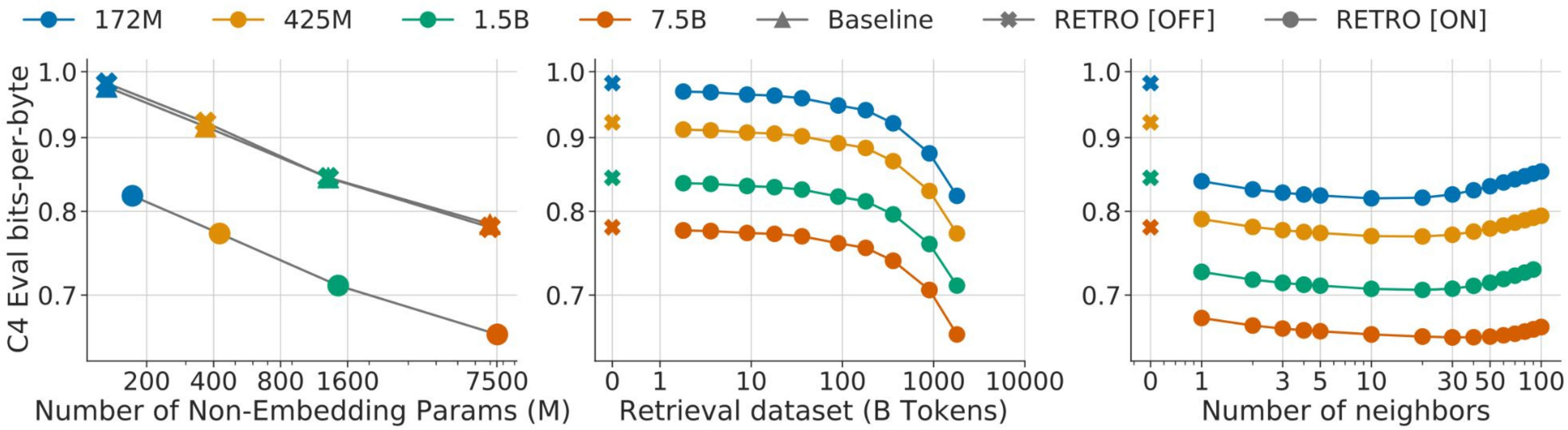


Figure: Scaling of Retro

Results

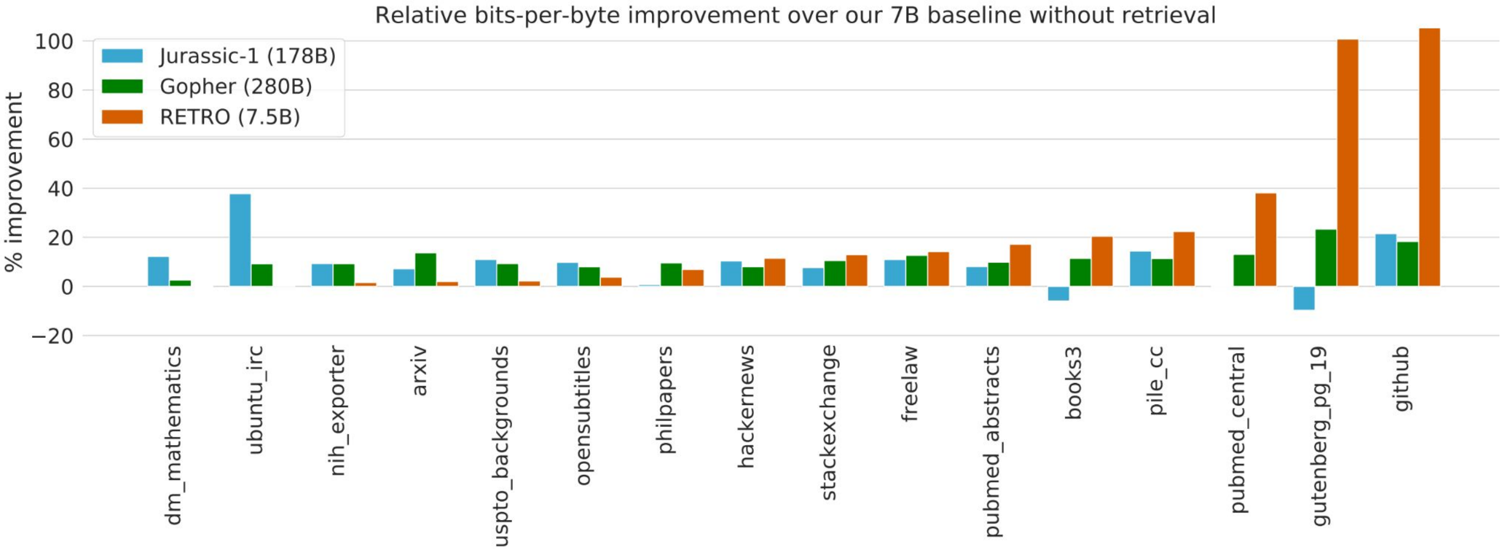


Figure: comparison to non-retrieval LLM

Results

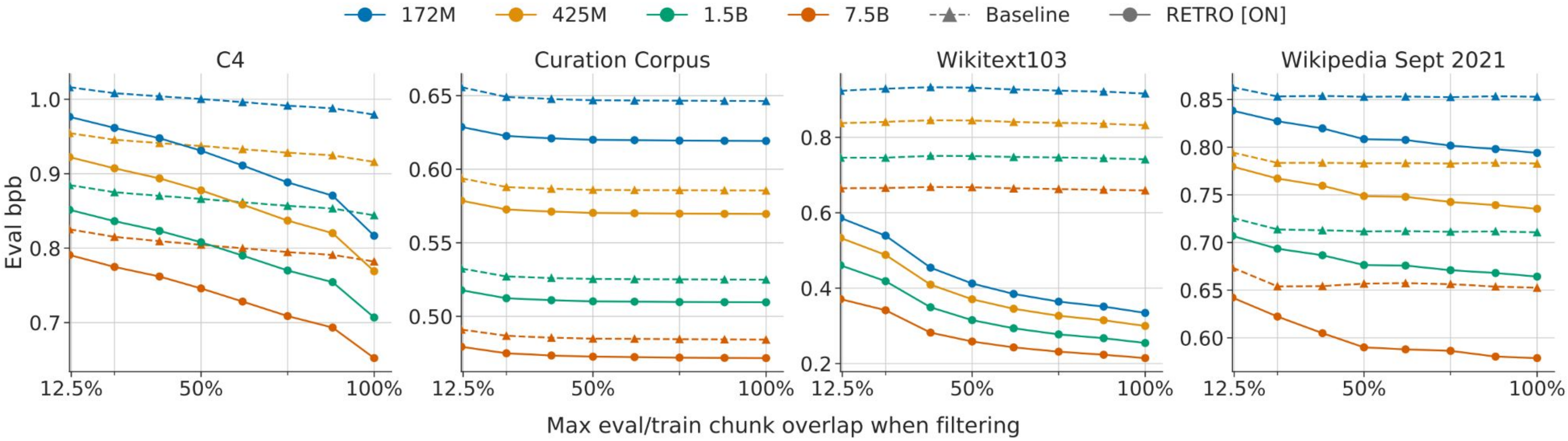
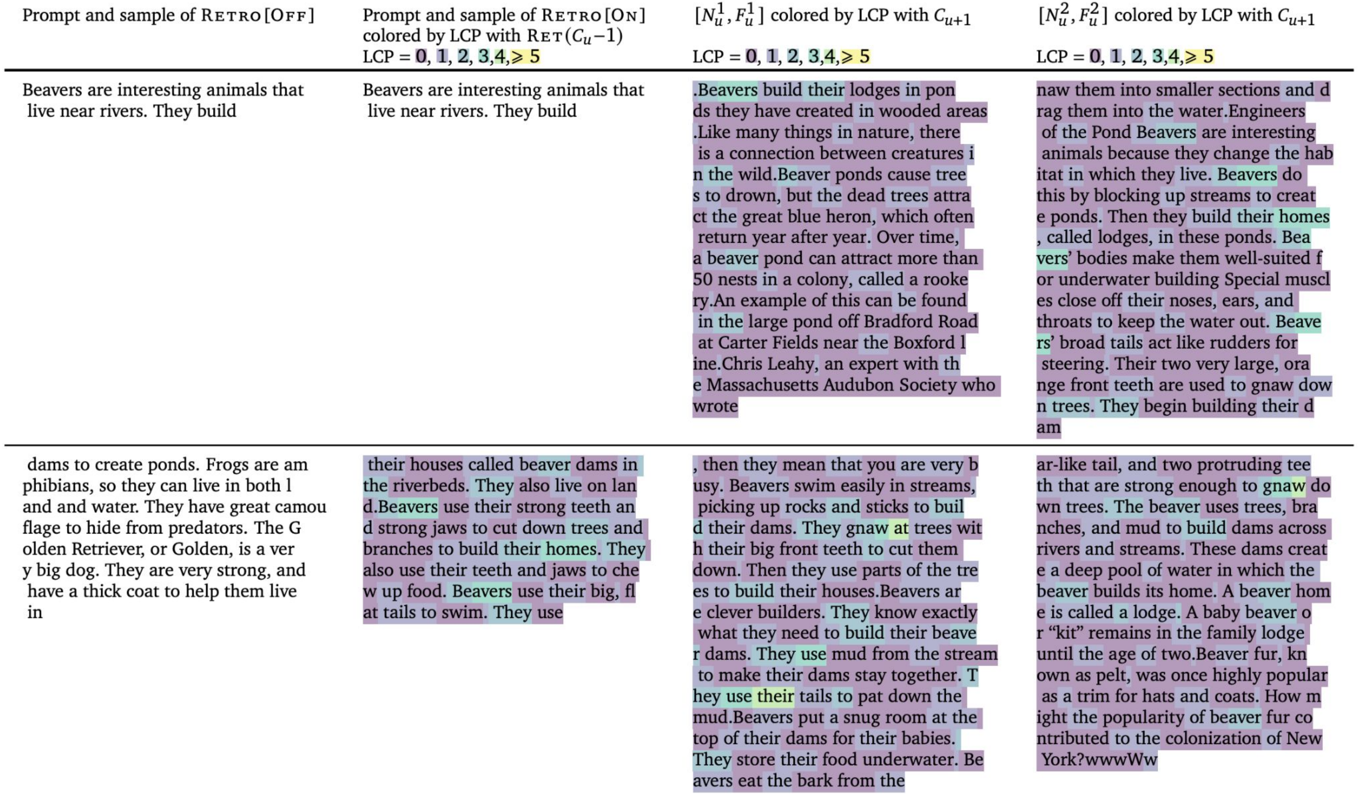


Figure: Relating retrieval performance to dataset leakage

Results - sample



Prompt and sample of RETRO[OFF]	Prompt and sample of RETRO[ON] colored by LCP with RET(C_u−1) LCP = 0, 1, 2, 3,4,⩾ 5	$[N_u^1, F_u^1]$ colored by LCP with C_{u+1} LCP = 0, 1, 2, 3,4,⩾ 5	$[N_u^2, F_u^2]$ colored by LCP with C_{u+1} LCP = 0, 1, 2, 3,4,⩾ 5
Beavers are interesting animals that live near rivers. They build	Beavers are interesting animals that live near rivers. They build	.Beavers build their lodges in ponds they have created in wooded areas.Like many things in nature, there is a connection between creatures in the wild.Beaver ponds cause trees to drown, but the dead trees attract the great blue heron, which often return year after year. Over time, a beaver pond can attract more than 50 nests in a colony, called a rookery.An example of this can be found in the large pond off Bradford Road at Carter Fields near the Boxford line.Chris Leahy, an expert with the Massachusetts Audubon Society who wrote	naw them into smaller sections and drag them into the water.Engineers of the Pond Beavers are interesting animals because they change the habitat in which they live. Beavers do this by blocking up streams to create ponds. Then they build their homes, called lodges, in these ponds. Beavers' bodies make them well-suited for underwater building Special muscles close off their noses, ears, and throats to keep the water out. Beavers' broad tails act like rudders for steering. Their two very large, orange front teeth are used to gnaw down trees. They begin building their dam
dams to create ponds. Frogs are amphibians, so they can live in both land and water. They have great camouflage to hide from predators. The Golden Retriever, or Golden, is a very big dog. They are very strong, and have a thick coat to help them live in	their houses called beaver dams in the riverbeds. They also live on land.Beavers use their strong teeth and strong jaws to cut down trees and branches to build their homes. They also use their teeth and jaws to chew up food. Beavers use their big, flat tails to swim. They use	, then they mean that you are very busy. Beavers swim easily in streams, picking up rocks and sticks to build their dams. They gnaw at trees with their big front teeth to cut them down. Then they use parts of the trees to build their houses.Beavers are clever builders. They know exactly what they need to build their beaver dams. They use mud from the stream to make their dams stay together. They use their tails to pat down the mud.Beavers put a snug room at the top of their dams for their babies. They store their food underwater. Beavers eat the bark from the	ar-like tail, and two protruding teeth that are strong enough to gnaw down trees. The beaver uses trees, branches, and mud to build dams across rivers and streams. These dams create a deep pool of water in which the beaver builds its home. A beaver home is called a lodge. A baby beaver or “kit” remains in the family lodge until the age of two.Beaver fur, known as pelt, was once highly popular as a trim for hats and coats. How might the popularity of beaver fur contributed to the colonization of New York?wwwWw

Limitations

- Still Transformer - 100% decoupling is not achieved.
- Retrieved content length is still limited.
- Knowledge base is just unstructured text.
- What is “world knowledge”?
- What is “language information”?

References

- *Borgeaud et.al.* 2022. Improving language models by retrieving from trillions of tokens
- *Wang et.al.* 2023. Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study
- *Wang et.al.* 2023. InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining
- The Illustrated Retrieval Transformer