DeepMind

RETRO: Improving language models by retrieving from trillions of tokens

Sebastian Borgeaud*, Arthur Mensch*, Jordan Hoffmann*, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae[†], Erich Elsen[†], **Laurent Sifre**[*,†]

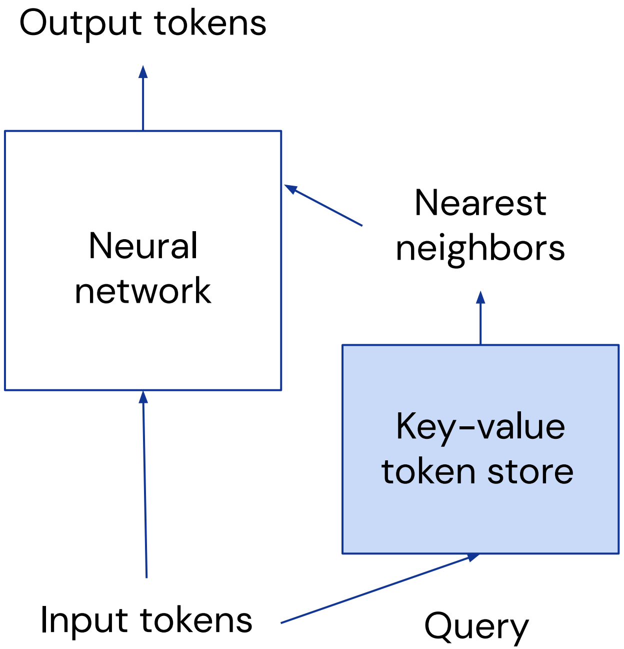
*Equal contributions, [†]Equal senior authorship

Adding an explicit memory to language models

- Increasing the model size has two effects:
  - Increase **memorisation** of training data
  - Increase **generalisation** performance

- Can we increase memorisation without the extra parameters ?
  - External memory mechanism
  - Associated to a trainable neural network

- Combine a **parametric model** (neural network) with a **non-parametric model** (data store providing nearest neighbours)



RETRO: adding a very large database of tokens

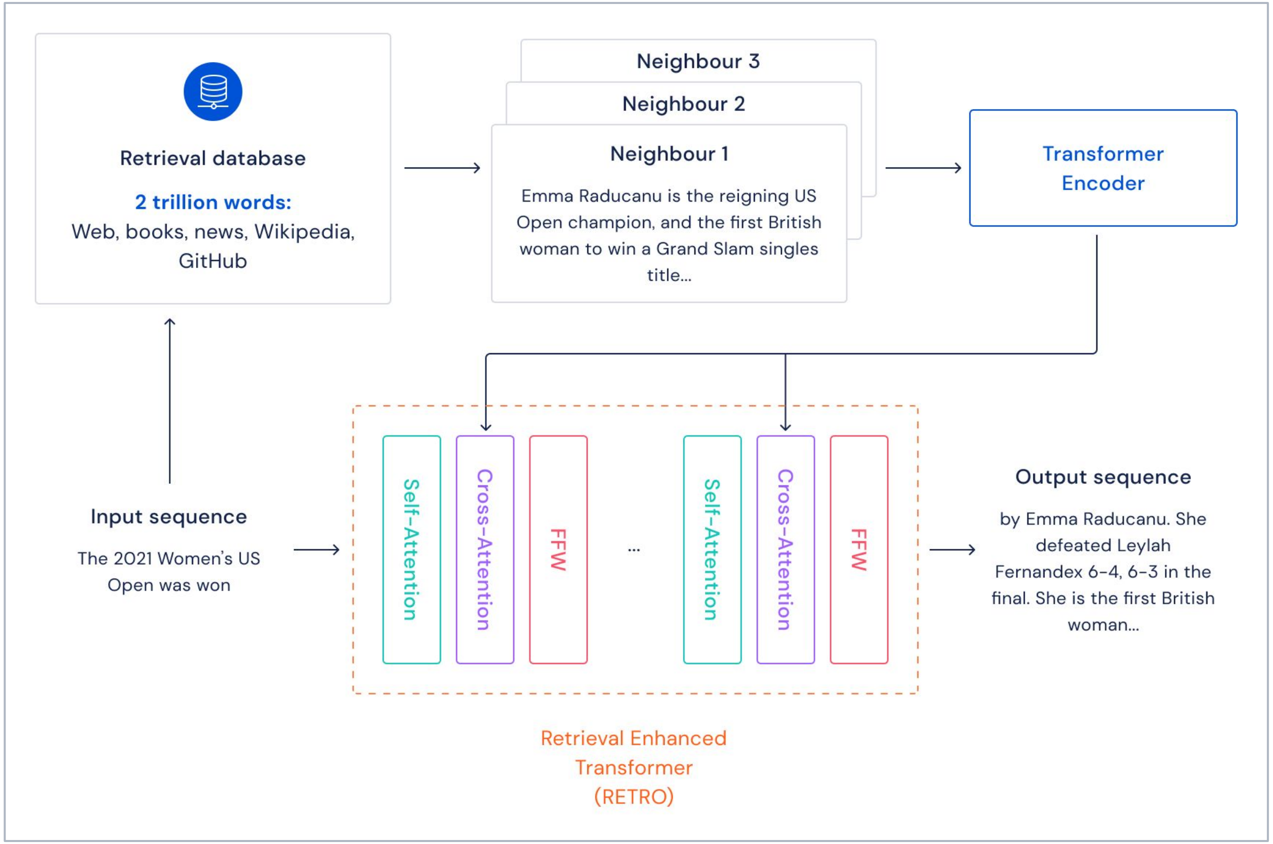
Previous work on retrieval have focused on retrieving from Wikipedia (**3B tokens**)

We retrieve from our entire training set: **2T tokens (700x more)**

Three major ingredients:

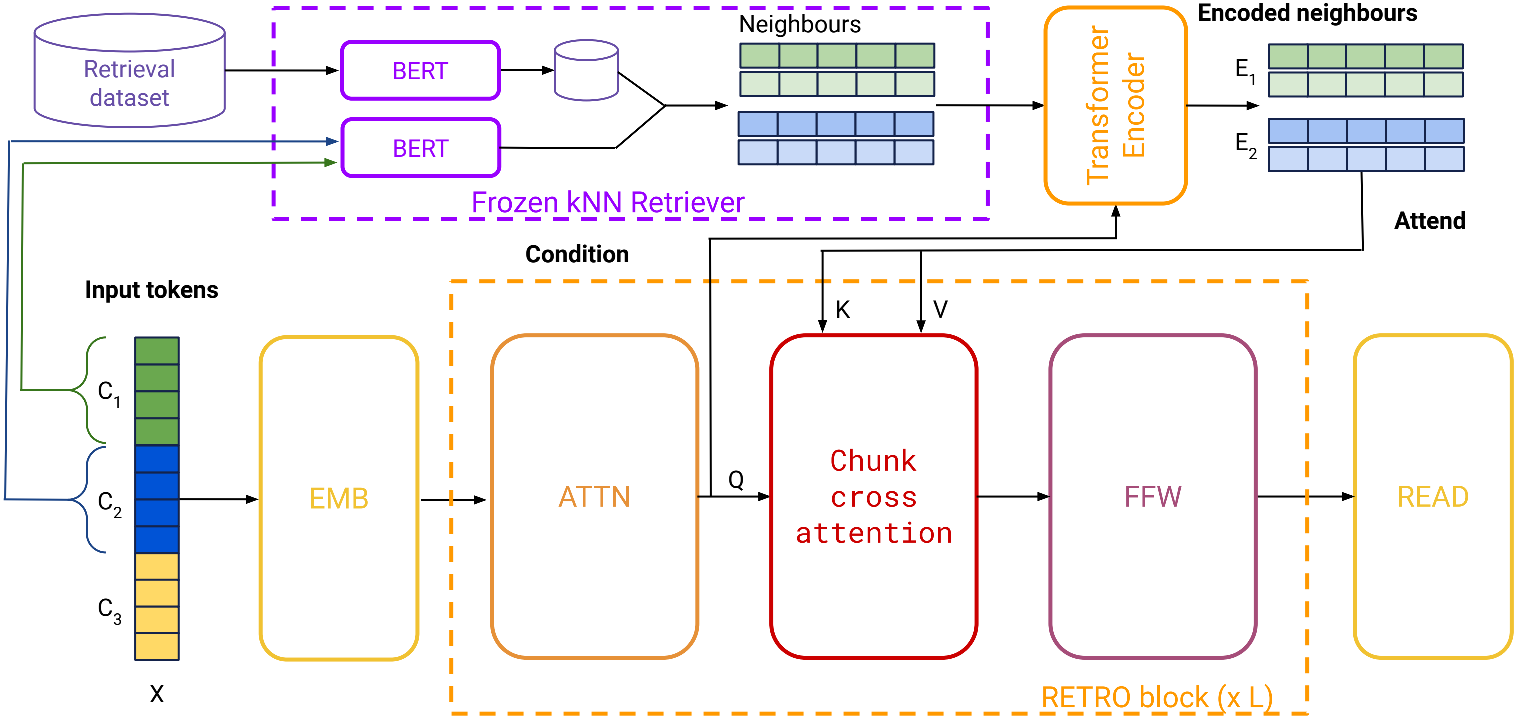
a. **Frozen dense retriever** (pooled BERT on small sequences)
b. **Approx k-nearest neighbors** run as a preprocessing step (quantification and tree search)
c. **DB element is a chunk and not a token:** 30B keys

Retrieving at chunk level, predicting at sequence level



Retrieval database
2 trillion words:
Web, books, news, Wikipedia, GitHub
Neighbour 3
Neighbour 2
Neighbour 1
Emma Raducanu is the reigning US Open champion, and the first British woman to win a Grand Slam singles title...
Transformer Encoder
Input sequence
The 2021 Women's US Open was won
Self-Attention
Cross-Attention
FFW
...
Self-Attention
Cross-Attention
FFW
Output sequence
by Emma Raducanu. She defeated Leylah Fernandex 6-4, 6-3 in the final. She is the first British woman...
Retrieval Enhanced Transformer (RETRO)

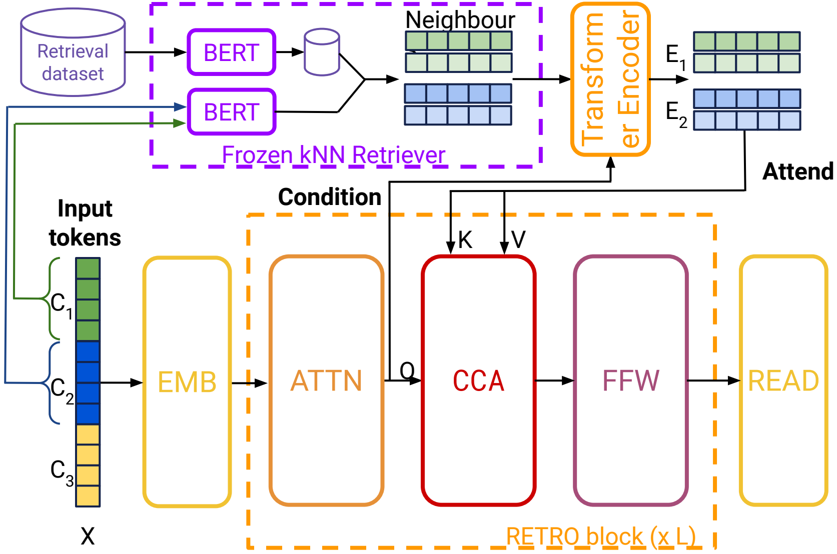
Condition new chunk generation on previous chunk neighbours



A model made for sampling

- Fully autoregressive model

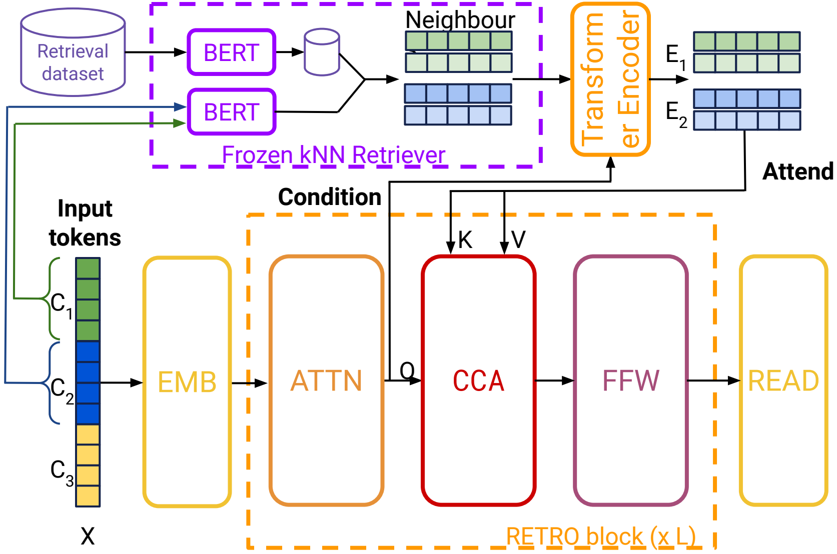
- Sampling with retrieval queries in the loop

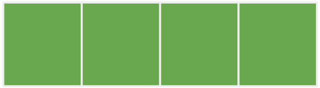


A model made for sampling

- Fully autoregressive model

- Sampling with retrieval queries in the loop

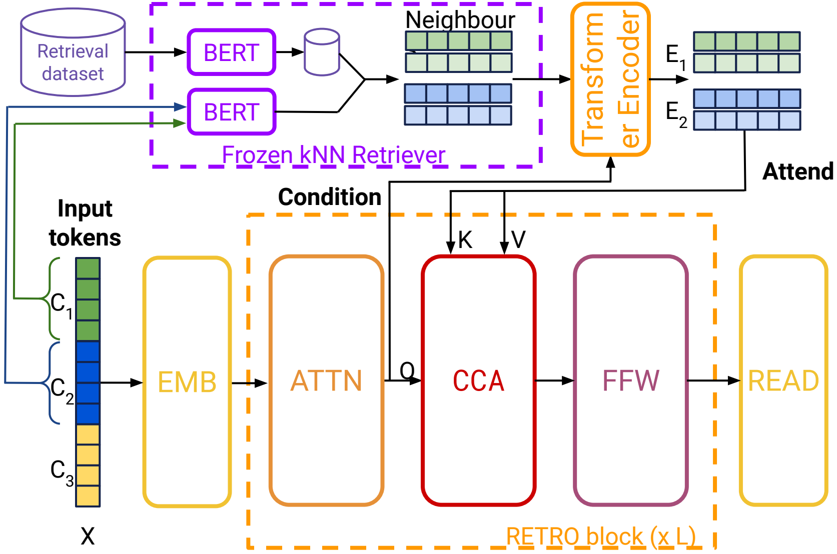


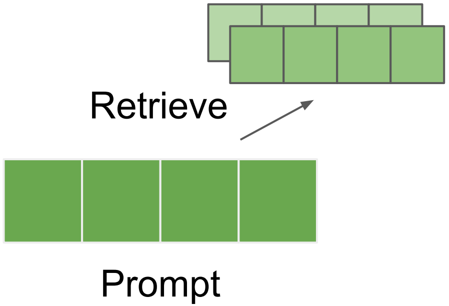


Prompt

A model made for sampling

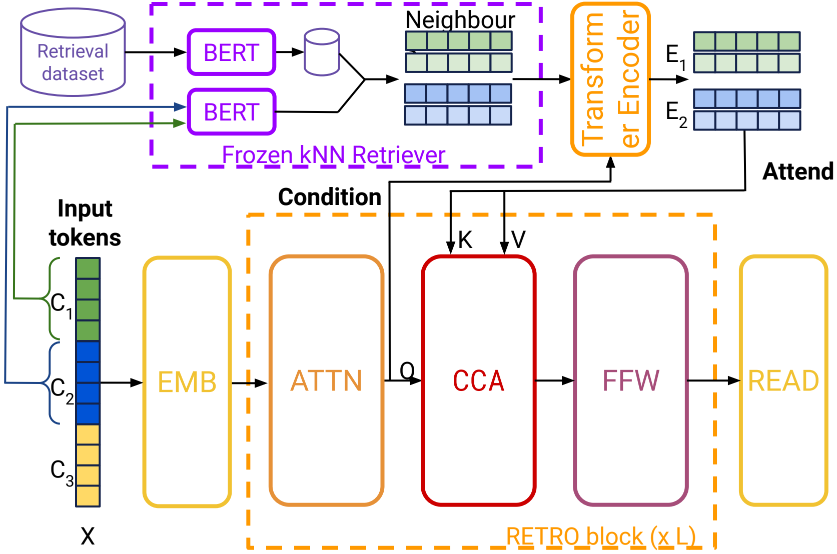
- Fully autoregressive model
- Sampling with retrieval queries in the loop

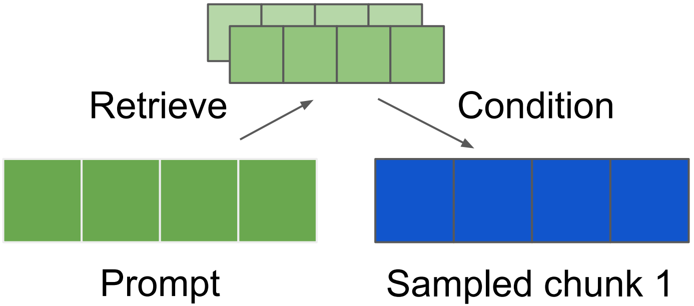




A model made for sampling

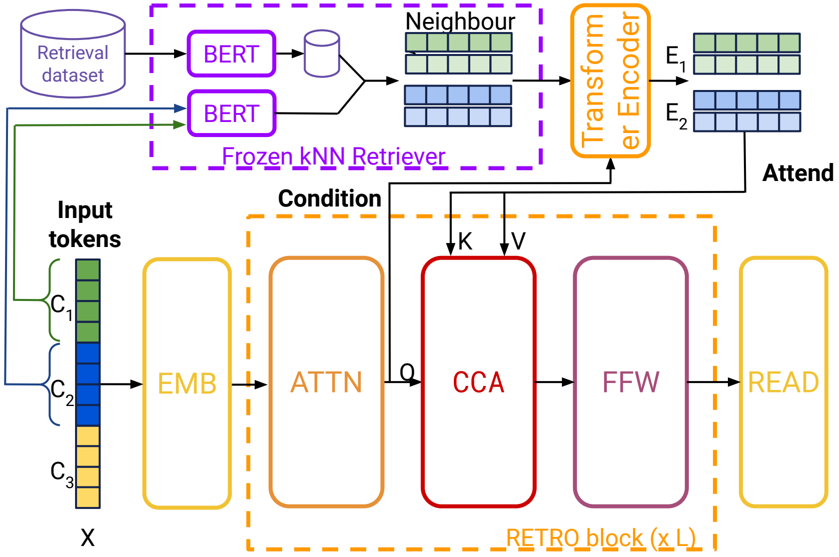
- Fully autoregressive model
- Sampling with retrieval queries in the loop

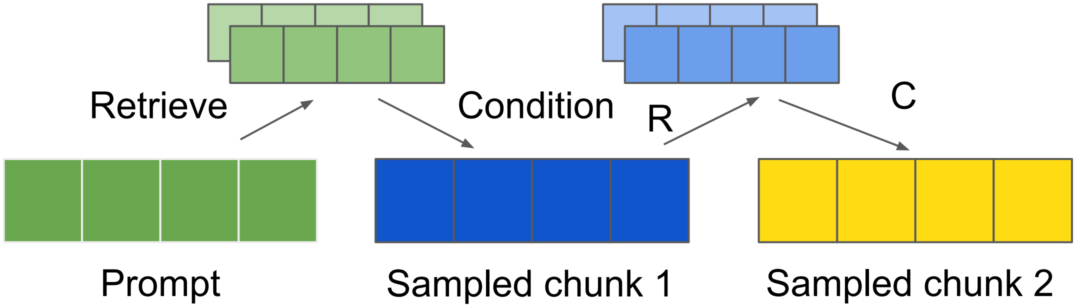




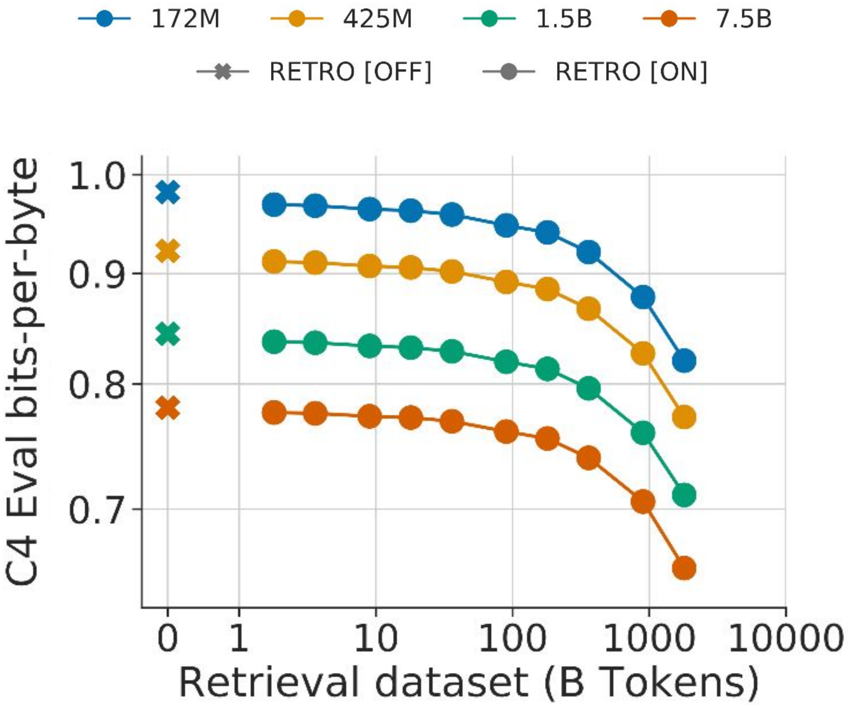
A model made for sampling

- Fully autoregressive model
- Sampling with retrieval queries in the loop





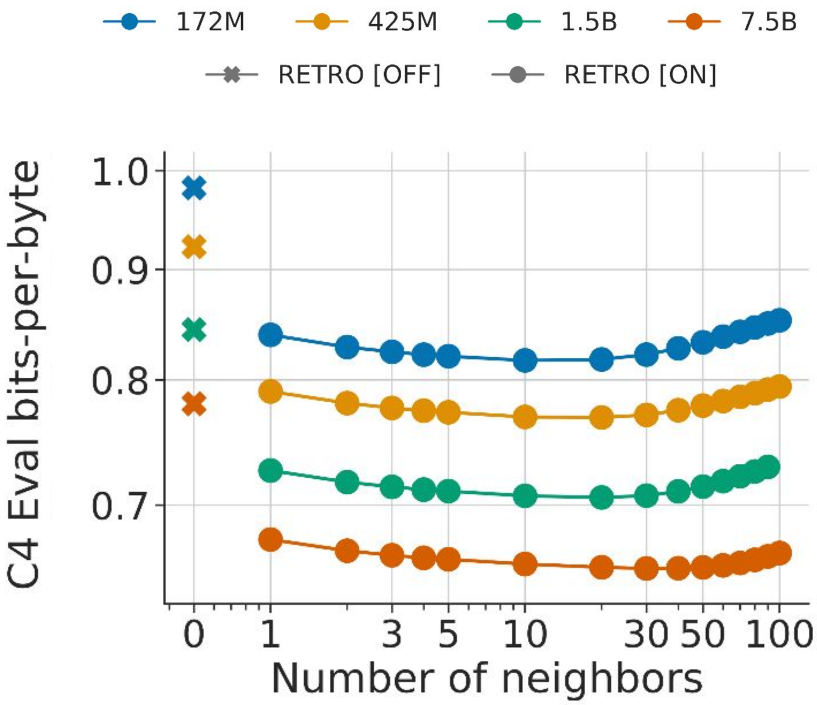
RETRO improves strongly with database size



- Continuous improvement with database size

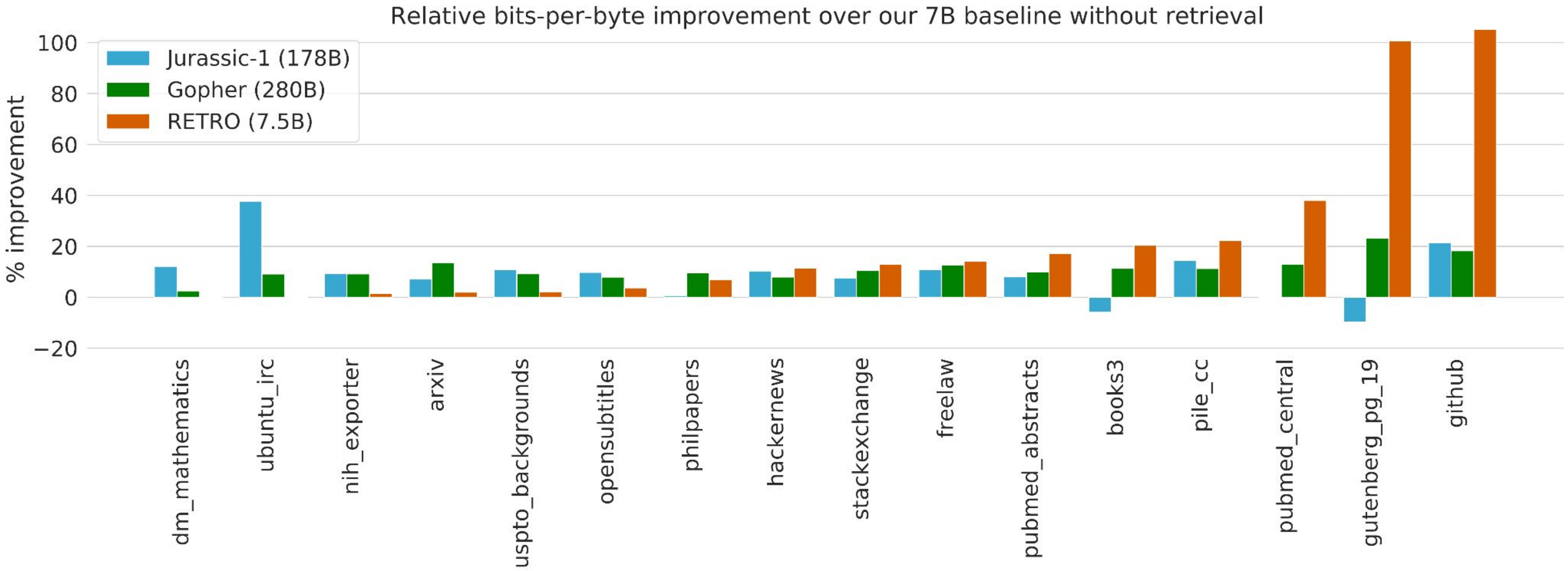
- Improvements for all models

RETRO improves with additional retrieved neighbours



- Trained with 2 neighbours
- Improvements up to 40 neighbours at evaluation with 7.5B
- Larger models can better utilise extra neighbours

Language modelling: The Pile



Gao et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Conclusion

- **RETRO** is a **general** architecture, that is fully **autoregressive** and enables **large scale retrieval**

- Adding a **2T token database** yields a performance improvement that's constant with model size:
  - Similar performance to models with 10x more parameters on the Pile

- Consistent performance across benchmarks
  - Retrieval does exploit train-test leakage more than standard language models
  - But performance also improves on held-out tokens

- Future work on few-shot evaluation