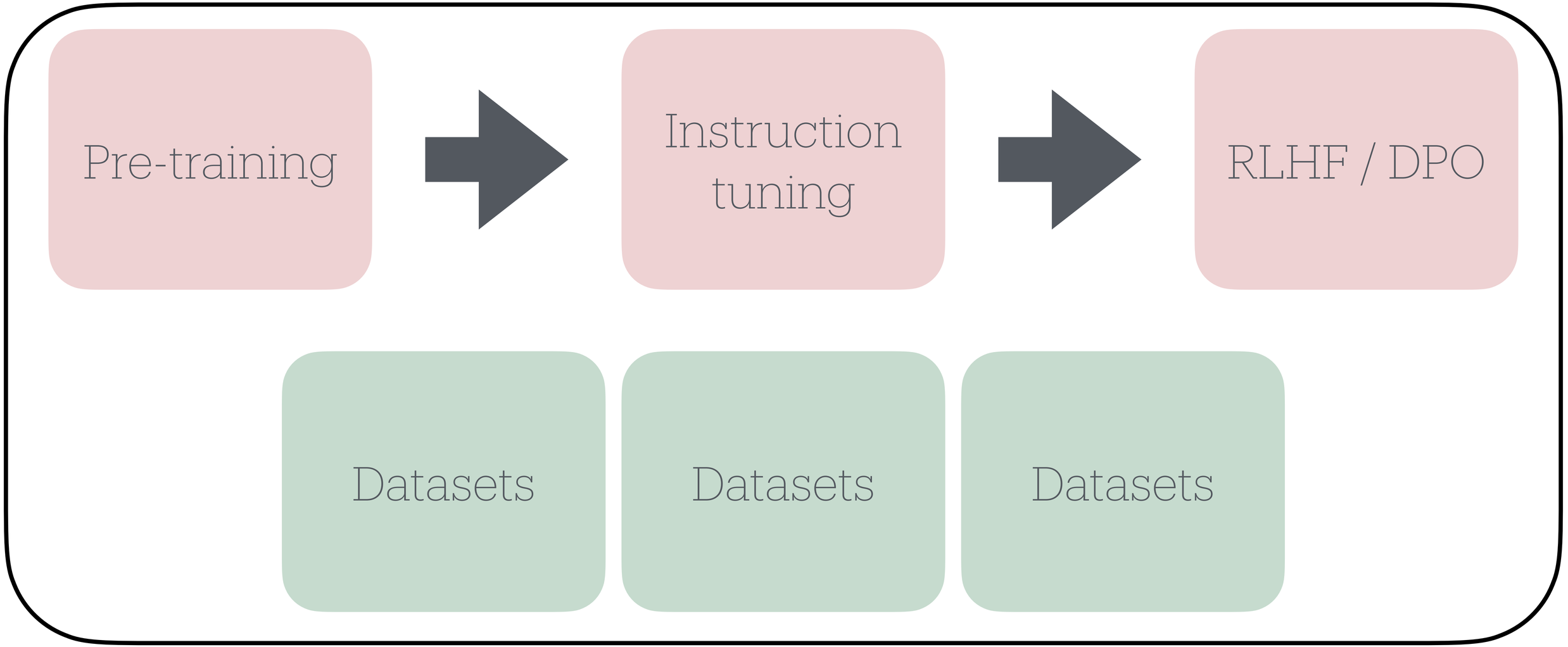
Retrieval Augmented Generation

Philipp Krähenbühl, UT Austin

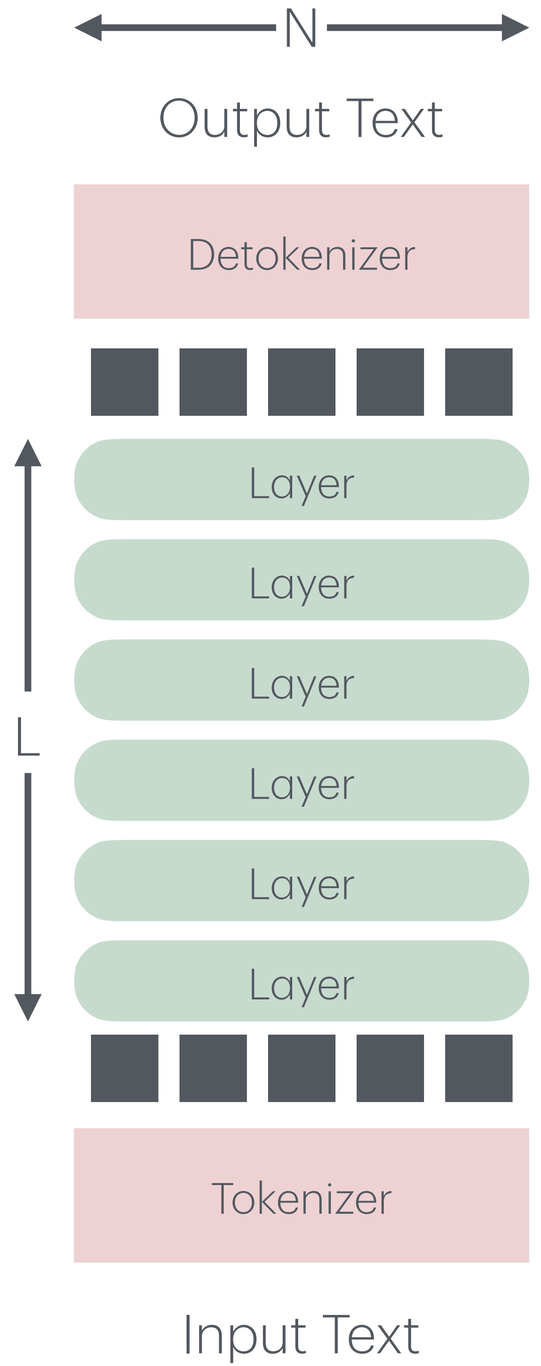
Full Picture

Basic LLM



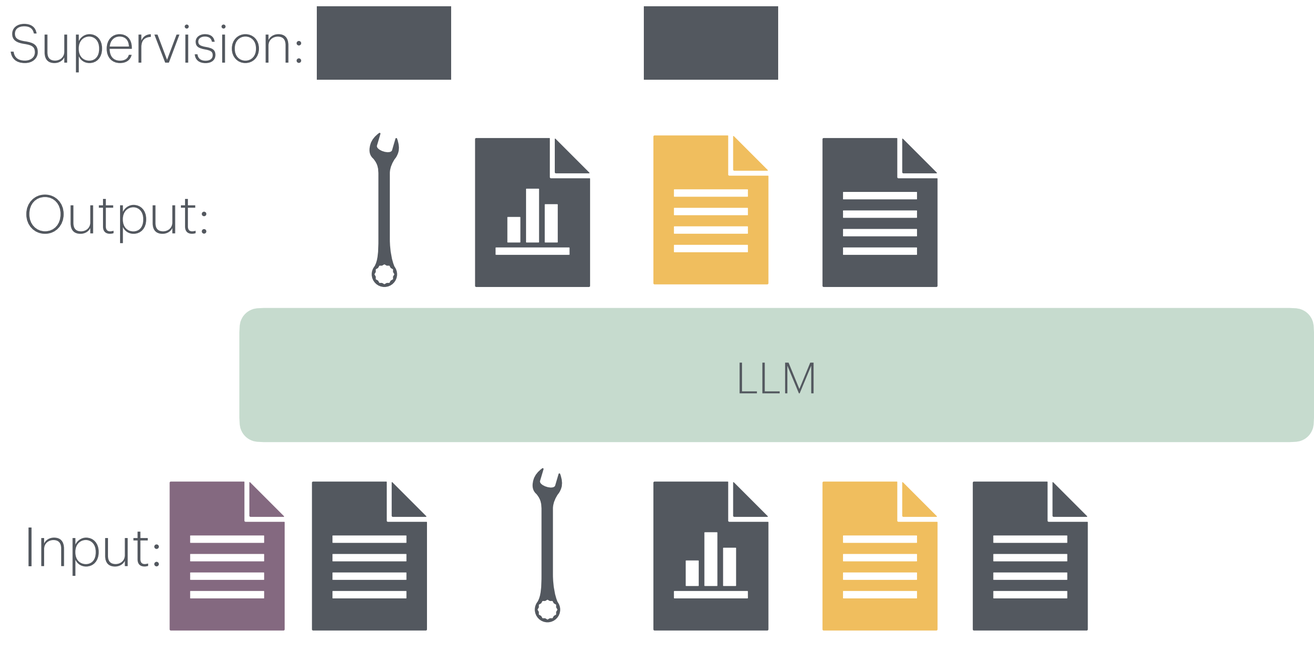
Training and Generation

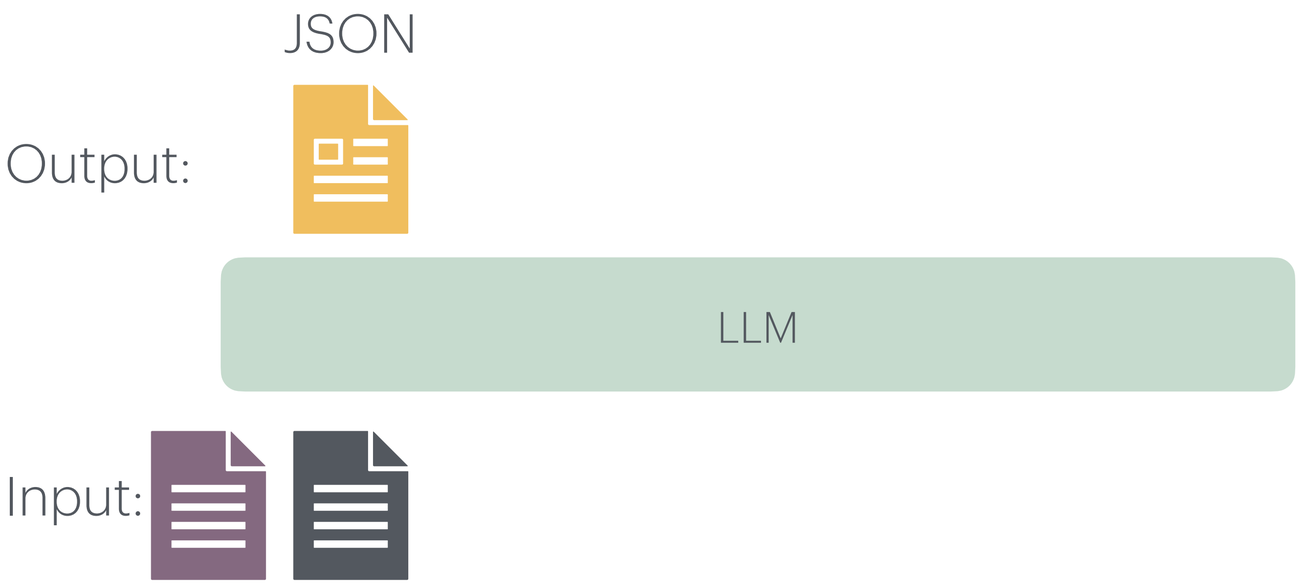
	Training	Training - Checkpointi	Generation	Paged Attention	Speculative decoding
Peak Memory	O(NL)	$O(NL^{½})$	O(N)	O(NL)	O(NL)
Runtime	$O(N^2L)$	$O(2\ N^2L)$	$O(N^3L)$	$O(N^2L)$	$O(N^2L)$
# forward	1	1	N	N	N / α



Tools and Structured outputs

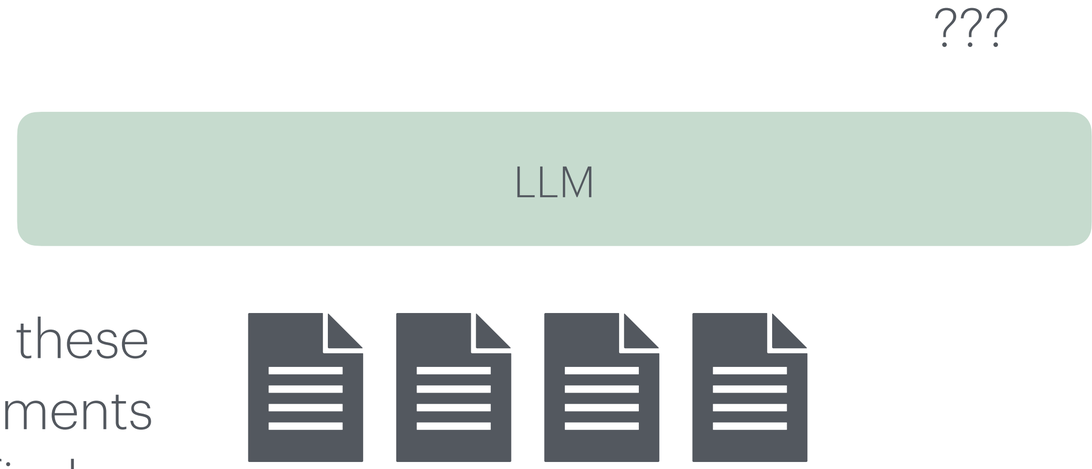
- Tools
  - Special tags, Special chat-template
- Structured output
  - Option 1.1: Write a robust parser (in python)
    - Let LLM know that you failed to parse
  - Option 1.2: Constrain output
  - Option 2: Use a tool, arguments = json fields





Long Context

- Current model are **pre-trained** on **2-8k** token sequences
- Late stage pre-training **8k-128k**
  - RoPE Scaling
- Fine-tuned on variable length sequences



Longer Context

- What is we have even more inputs

Longer Context

- What is we have even more inputs
  - We have to manage context

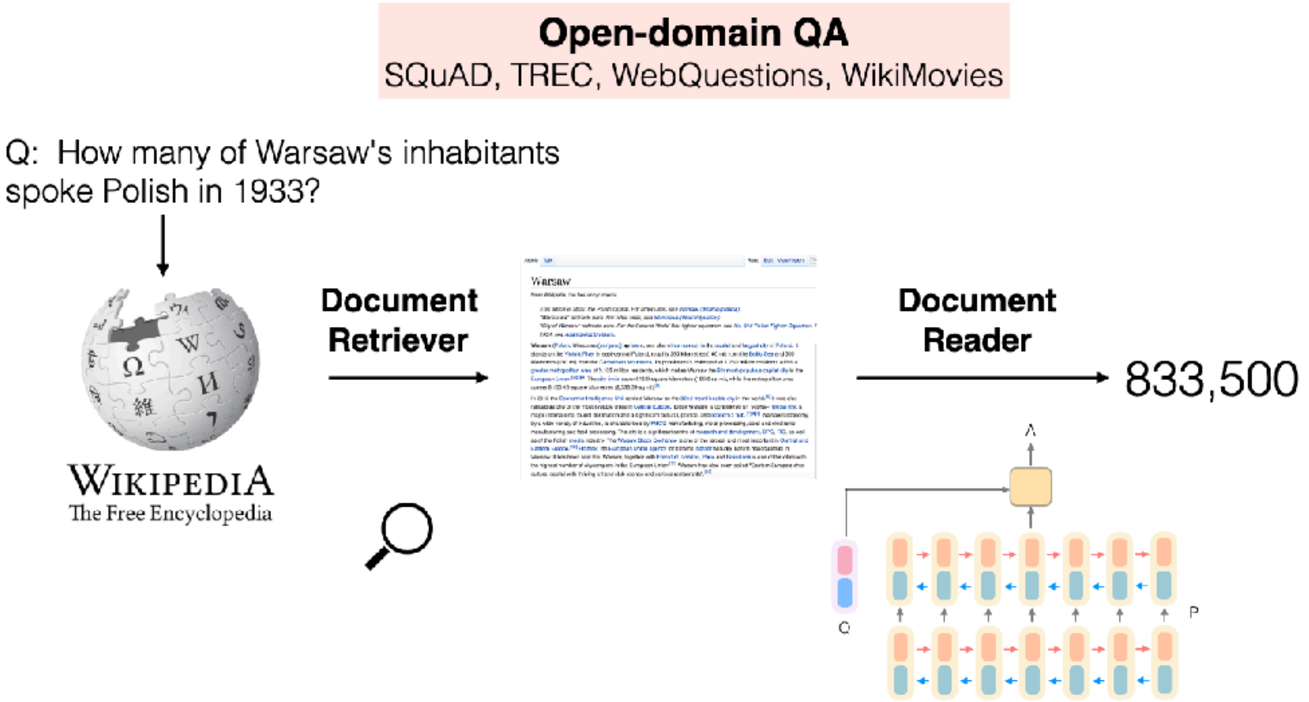
???

LLM

Read these documents and find references to efficient long-context LLMs

Longer Context

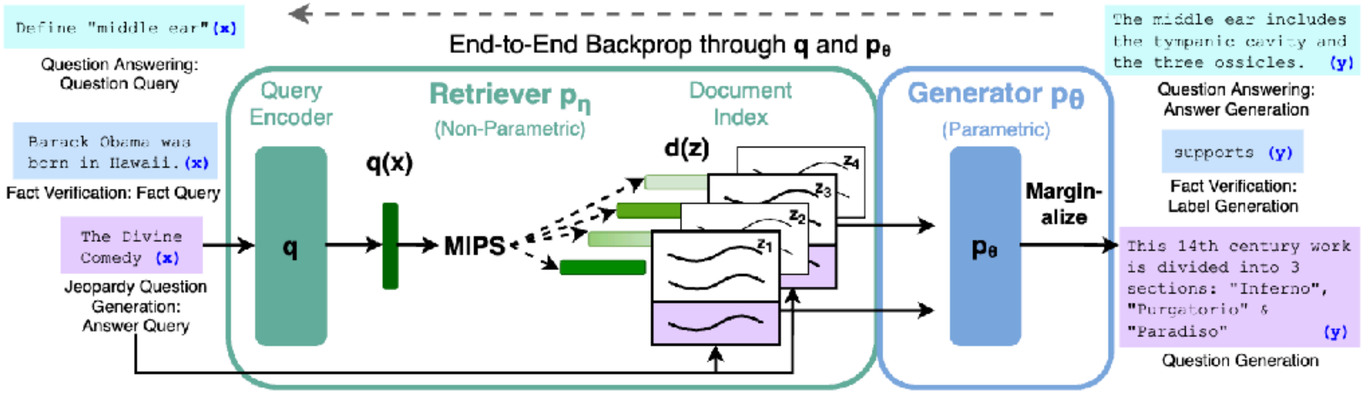
- Solution: Build a “system”
- Option 1
  - Document Retriever: LLM to retrieve most relevant document
  - Document Reader: LLM to answer request

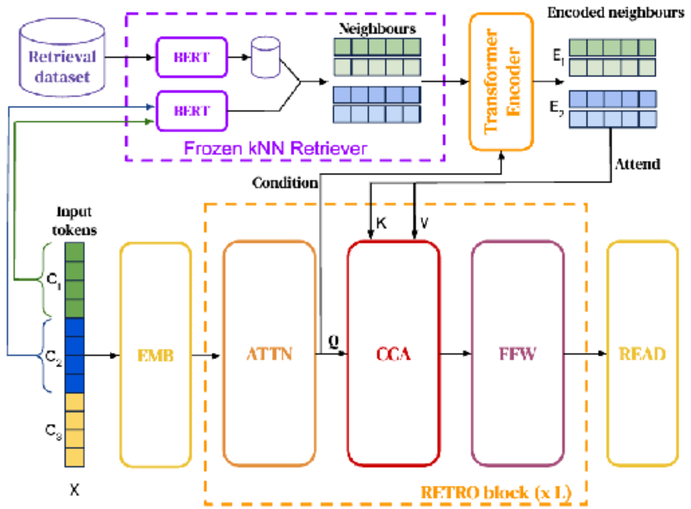


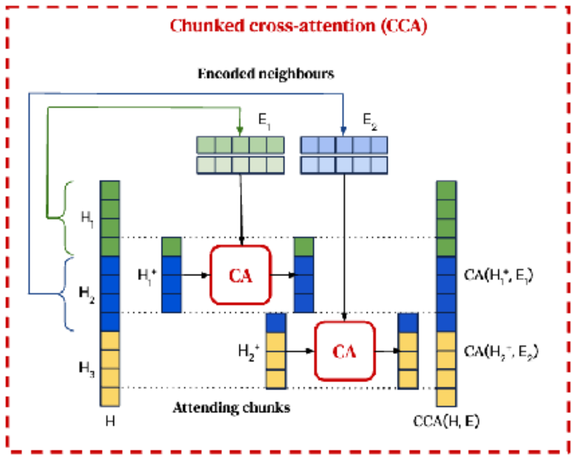
Reading Wikipedia to Answer Open-Domain Questions, Chen etal 2017

Longer Context

- Solution: Build a “system”
- Option 2
  - Document Retriever: LLM to retrieve all relevant documents
  - LLM to answer request with documents in context
  - Fine-tuned for task







Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Lewis etal 2020
REALM: Retrieval-Augmented Language Model Pre-Training, Guu etal 2020
Improving language models by retrieving from trillions of tokens, Borgeaud 2021

Longer Context

- Solution: Build a “system”
- Option 3
  - Document Retriever: LLM to retrieve all relevant documents
  - LLM to answer request with documents in context
  - ~~Fine-tuned for task~~ Model is prompted instead

In-Context Retrieval-Augmented Language Models, Ram etal 2023

Retrieval Augmented Generation

RAG

- A series of methods to manage the LLMs context
  - Some are trained
  - Some are just prompted

???

LLM

Read these documents and find references to efficient long-context LLMs

References

- [1] Reading Wikipedia to Answer Open-Domain Questions, Chen etal 2017 (link)
- [2] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Lewis etal 2020 (link)
- [3] REALM: Retrieval-Augmented Language Model Pre-Training, Guu etal 2020 (link)
- [4] Improving language models by retrieving from trillions of tokens, Borgeaud 2021 (link)
- [5] In-Context Retrieval-Augmented Language Models, Ram etal 2023 (link)