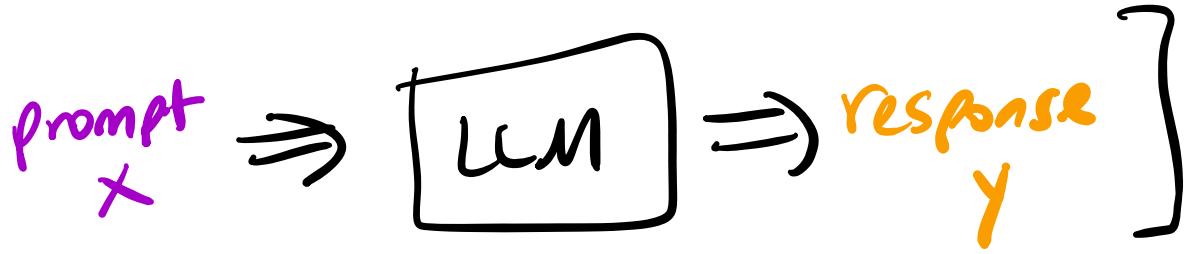
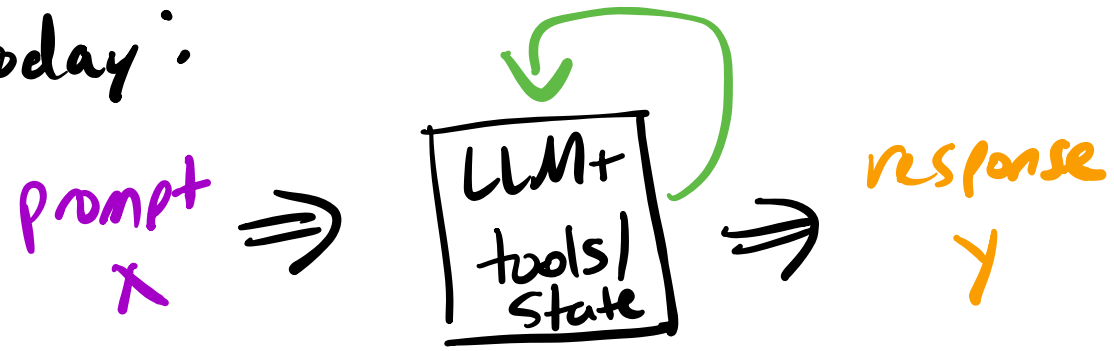
From LLMs ⇒ Systems

so far:



today:



---

retrieval augmented generation (RAG):

- ↳ assume frozen LLM P_θ
- ↳ create a datastore D of documents to retrieve from
- ↳ given a prompt x, retrieve top-k most relevant docs $z_1, z_2, \dots z_k$ from the datastore
- ↳ generate a response y from $P_\theta(y \mid x, z_1, z_2, \dots z_k)$

What date was Iribe opened to public?

↳ retrieve Iribe Wikipedia article / UMD CS website, ...

↳ concatenate these docs to my prompt

<doc 1> ... wikipedia ... </doc 1>

<doc 2> ... UMD CS page </doc 2>

⋮

↳ feed $x; z_{1...k}$ to p_θ to generate y

---

how do we perform retrieval?

↳ measure similarity between prompt x and docs z

↳ how?

↳ string matching

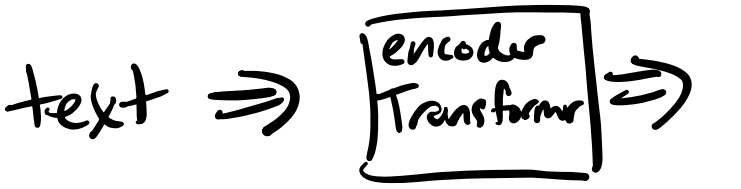
↳ BM-25

⇒ cheap, effective

⇒ not good at measuring semantic similarity

↳ vector database

↳ relies on a pretrained encoder model



↳ ave of final layer hidden states

↳ embed prompt x

embed each doc z_i in datastore

↳ identify k nearest neighbors of x

dot products:

$x \cdot z_1 = 5$

$x \cdot z_2 = -8$

$x \cdot z_3 = 0$

$x \cdot z_4 = 29$

} top 2 NN: z_1, z_4

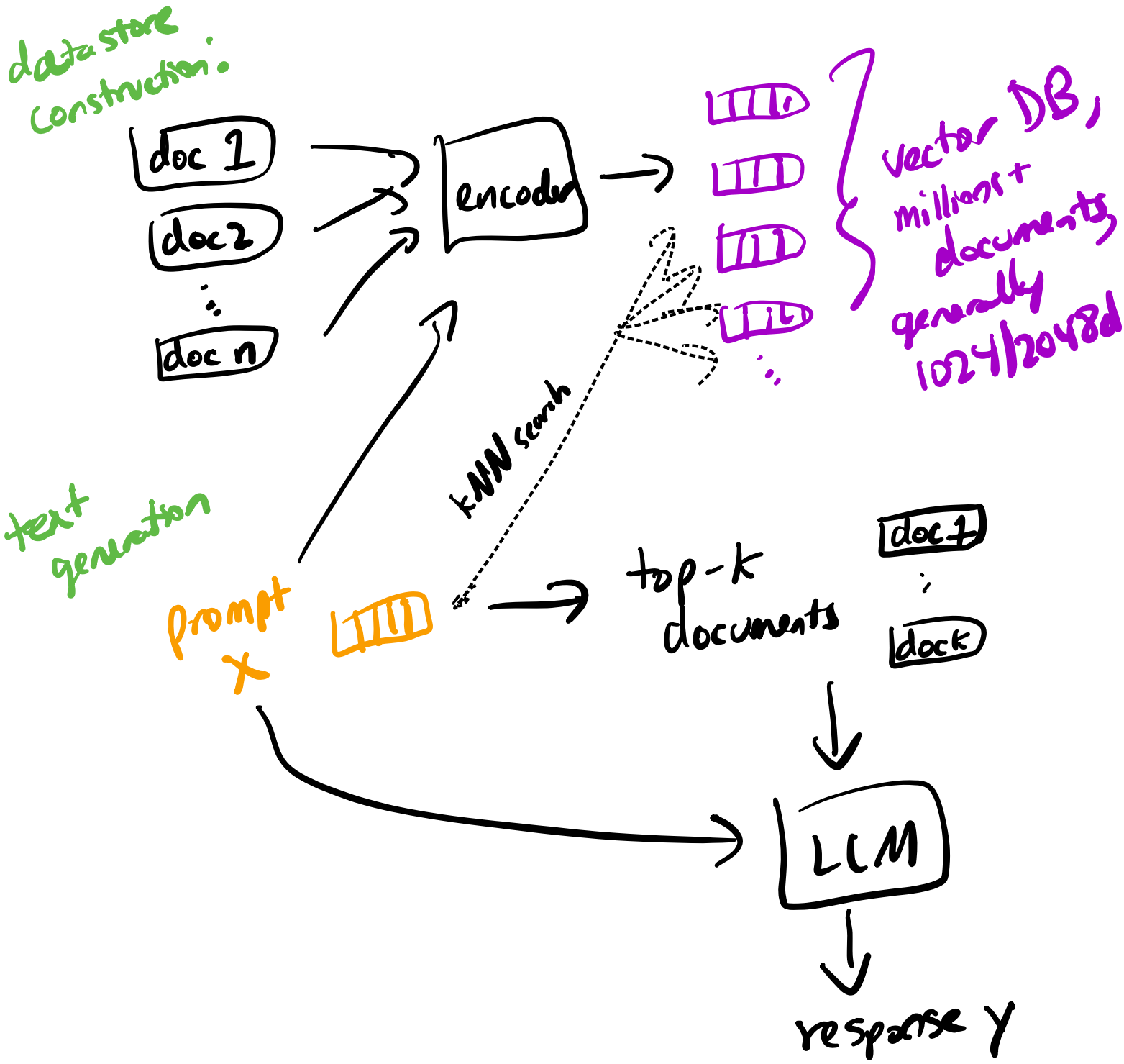
↳ exact nearest neighbors is intractable

↳ generally use approximate kNN

↳ FAISS

↳ encoder is fine-tuned on datasets of query/context pairs

↳ RAG w/ vector DB



RAG benefits:

- specific facts / domain knowledge may not be memorized by LLM during pretraining
- fresh knowledge
- reduce hallucinations, do attribution
  - connect generated text to source docs
    - not always reliable, LLM can misinterpret retrieved docs / ignore

RAG cons:

- retrieval is really hard!
- poor quality retrieval often leads to worse perf than no retrieval
  - parametric knowledge
    - knowledge encoded in the model's params

↳ non-parametric knowledge
↳ external docs / context

↳ information loss when embedding docs
↳ chunking → loses context

↳ increase inference cost as k increases
- k → # chunks retrieved
- k → chunk size

Tool use:

↳ retrieval is an example of a tool that the LLM can use

↳ in RAG, retrieval happens prior to response generation

↳ what if the model could call diff. tools during generation?

↳ Toolformer

Agents

↳ LLMs + tools + dynamic control flow

↳ some notion of state / memory