Augment LLMs with Retrieval

Self-designing RAG Systems

Qitong Wang, qitong@seas.harvard.edu

CS265. February 3, 2026

DASlab
@ Harvard SEAS

- Why we need retrieval
- How RAG works
- Algorithm designs

This class

- System designs
  Storage, pipelining, tuning, resource allocation
- Self-designing RAG systems

Next class

Large Language Models (LLMs)

Icons are obtained from https://www.flaticon.com

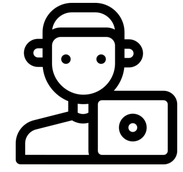
What is PyTorch?

PyTorch is an open-source machine learning and deep learning framework primarily used to build and train artificial intelligence models.

How many people are using PyTorch now?

As of early 2026, millions of developers and researchers are actively using PyTorch.

Write the code that uses torch.profile to profile the performance of a LLM pipeline. Here is the API document.



```
…
with profile(
    activities=[ProfilerActivity.CPU, ProfilerActivity.CUDA] if
torch.cuda.is_available() else [ProfilerActivity.CPU],
…
```

DASlab @ Harvard SEAS

Icons are obtained from https://www.flaticon.com

LLM input

LLM

LLM output

LLM input

LLM

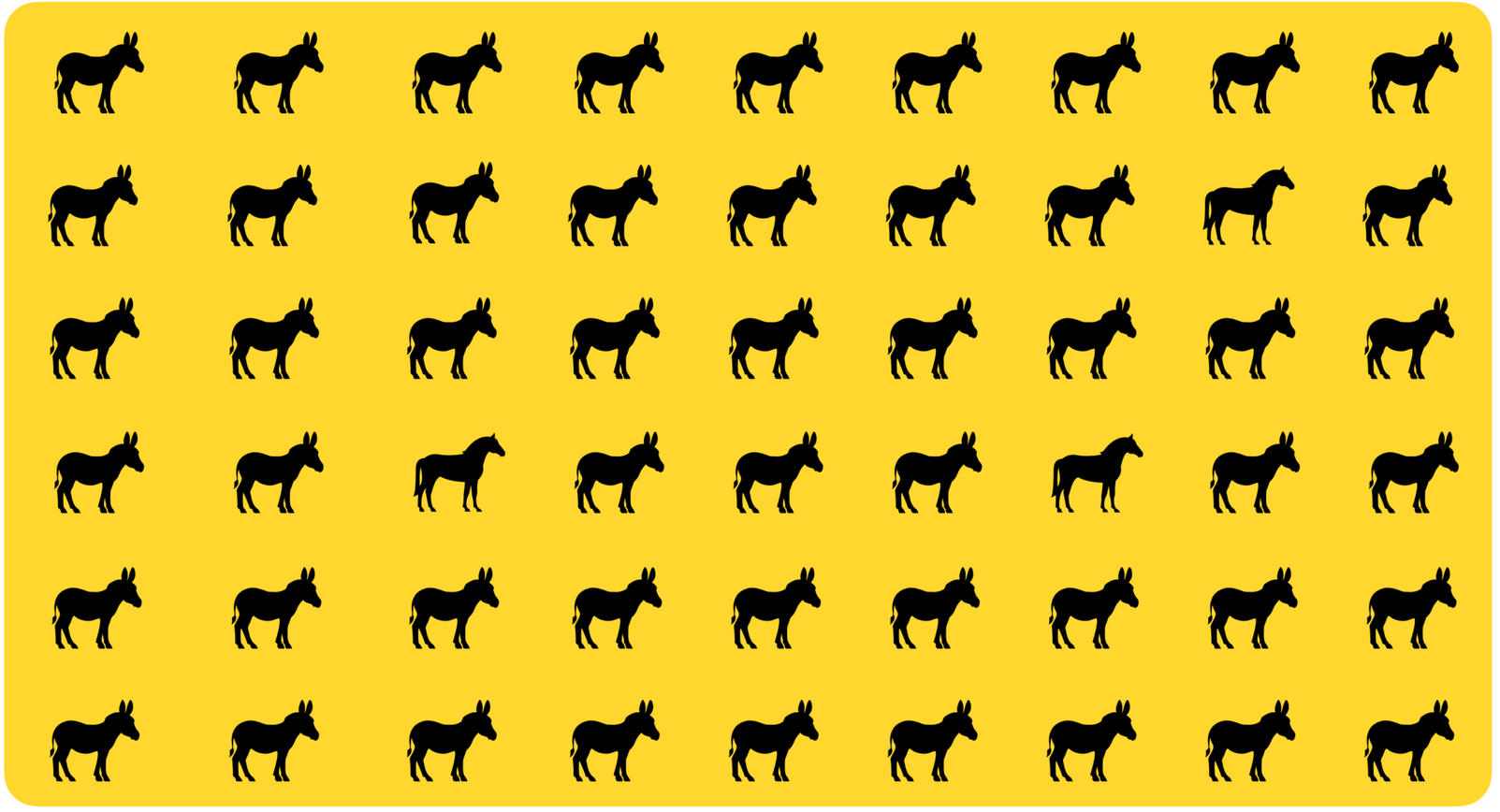
LLM output

LLM input

LLM

```
with profile(
    activities=[ProfilerActivity.CPU, ProfilerActivity.CUDA] if
torch.cuda.is_available() else [ProfilerActivity.CPU],
```

LLM output



Want to start a startup? Get funded by Y Combinator.

November 2009 I don't think Apple realizes how badly the App Store approval process is broken. Or rather, I don't think they realize how much it matters that it's broken. The way Apple runs the App Store has harmed their reputation with programmers more than anything else they've ever done. Their reputation with programmers used to be great. It used to be the most common complaint you heard about Apple was that their fans admired them too uncritically. The App Store has changed that. Now a lot of programmers have started to see Apple as evil.How much of the goodwill Apple once had with programmers have they lost over the App Store? A third? Half? And that's just so far. The App Store is an ongoing karma leak. How did Apple get into this mess? Their fundamental problem is that they don’t understand software.They treat iPhone apps the way they treat the music they sell through iTunes. Apple is the channel; they own the user; if you want to reach users, you do it on their terms. The record labels agreed, reluctantly. Want to start a startup? Get funded by Y Combinator. But this model doesn't work for software. It doesn’t work for an intermediary to own the user. The software business learned that in the early 1980s, when companies like VisiCorp showed that although the words "software" and "publisher" fit together, the underlying concepts don't.

Want to start a startup? Get funded by Y Combinator.

November 2009 I don't think Apple realizes how badly the App Store approval process is broken. Or rather, I don't think they realize how much it matters that it's broken. The way Apple runs the App Store has harmed their reputation with programmers more than anything else they've ever done. Their reputation with progra[illegible] you heard about Apple was that their fans admired them too uncritically. The App Store [illegible] as evil.How much of the goodwill Apple once had with programmers have they lost over the App Store? A third? Half? And that's just so far. The App Store is an ongoing karma leak. How did Apple get into this mess? Their fundamental problem is that they don't understand software.They treat iPhone apps the way they treat the music they sell through iTunes. Apple is the channel; they own the user; if you want to reach users, you do it on their terms. The record labels agreed, reluctantly. Want to start a startup? Get funded by Y Combinator. But this model doesn't work for software. It doesn't work for an intermediary to own the user. The software business learned that in the early 1980s, when companies like VisiCorp showed that although the words "software" and "publisher" fit together, the underlying concepts don't.

As data scale matters in data systems, data scale also matters in AI systems!

Greg Kamradt. Needle In A Haystack - Pressure Testing LLMs. 2023.

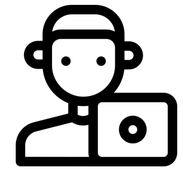
What is PyTorch?

Short and simple question
~10 words

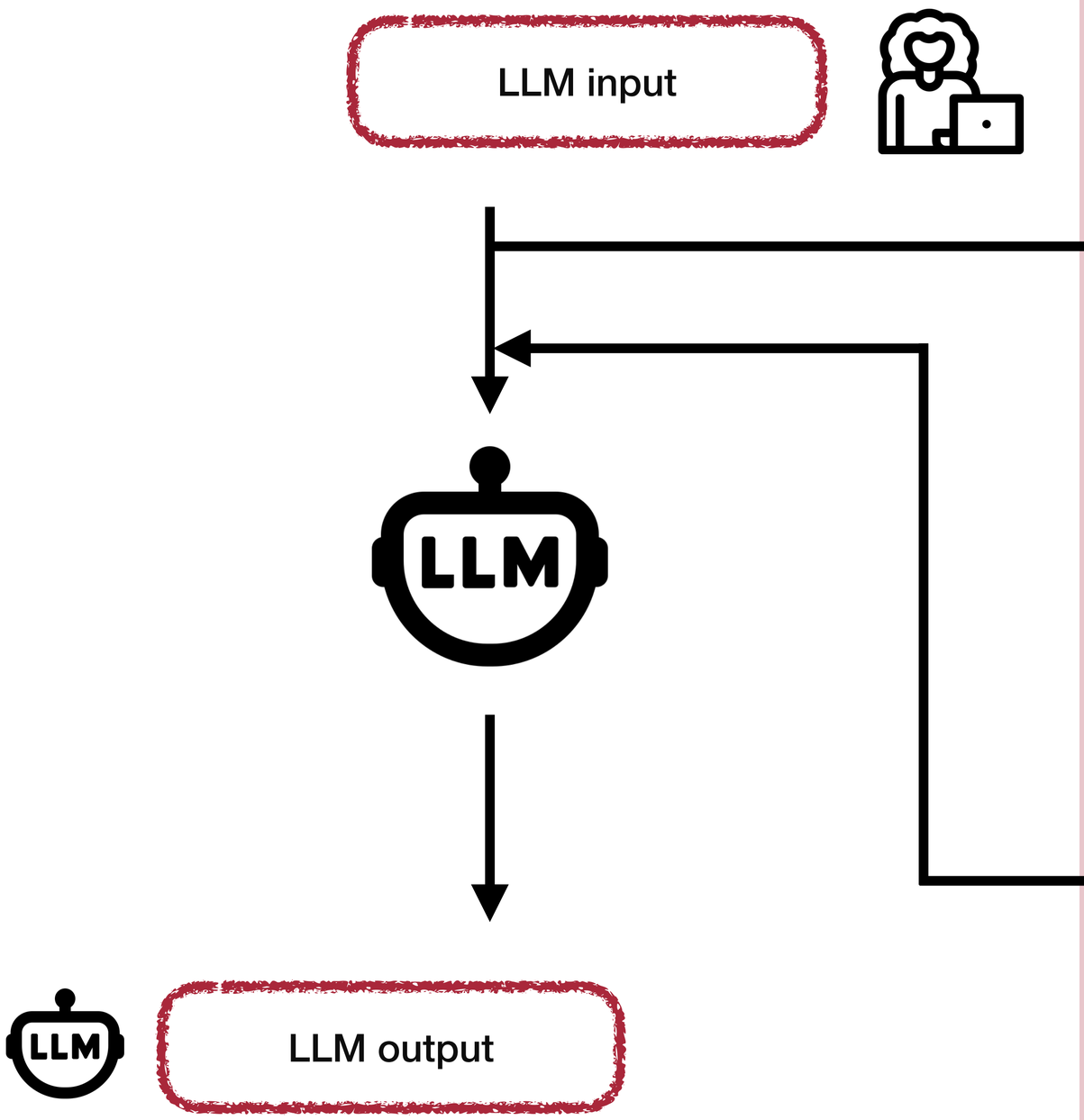
How many people are using PyTorch now?

Short question, but requires external information
~1k words combined

Write the code that uses torch.profile to profile the performance of a LLM pipeline. Here is the API document.



Long question (including the document content)
~1m words combined



the entire file/database, offloaded

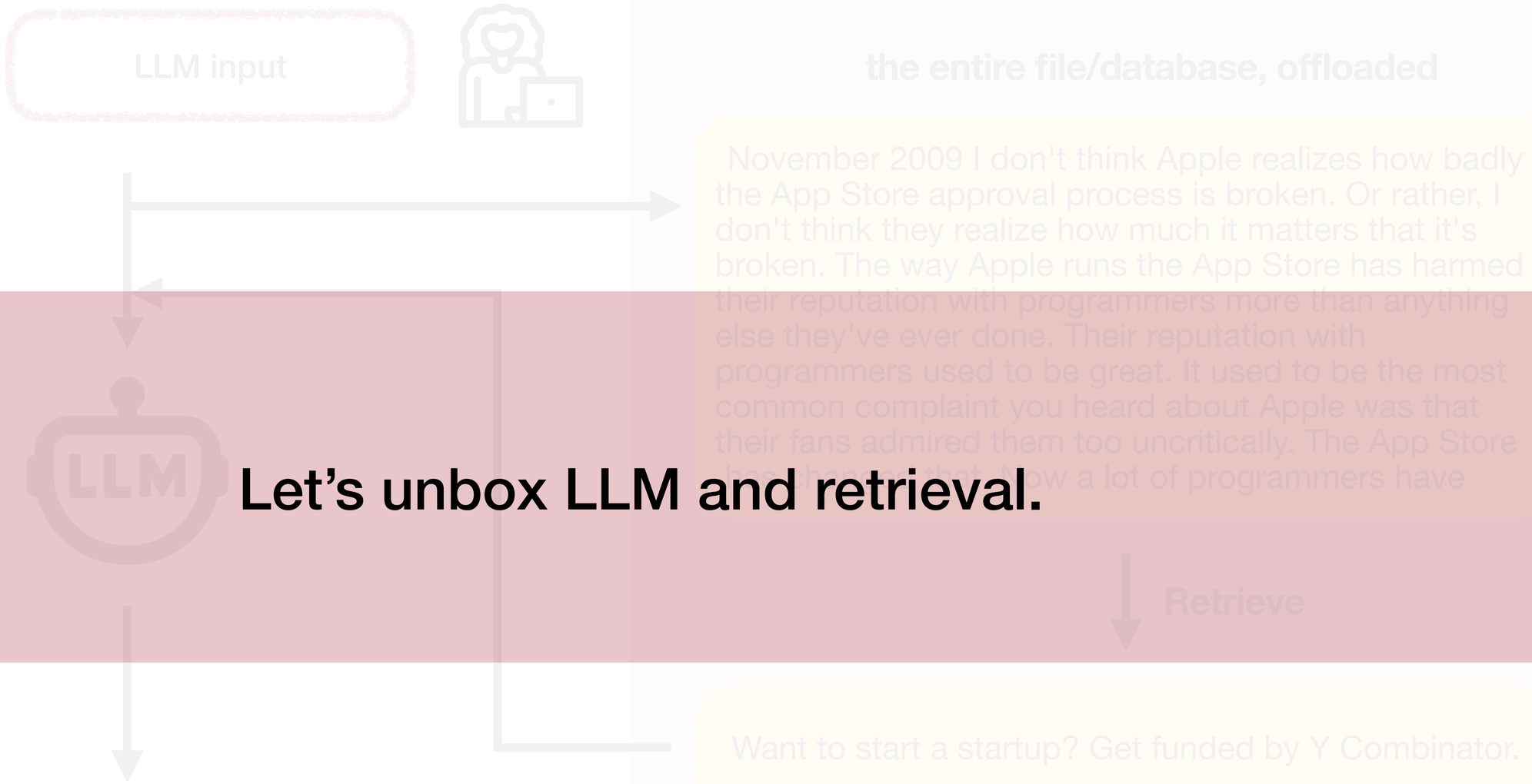
November 2009 I don't think Apple realizes how badly the App Store approval process is broken. Or rather, I don't think they realize how much it matters that it's broken. The way Apple runs the App Store has harmed their reputation with programmers more than anything else they've ever done. Their reputation with programmers used to be great. It used to be the most common complaint you heard about Apple was that their fans admired them too uncritically. The App Store has changed that. Now a lot of programmers have

Retrieve

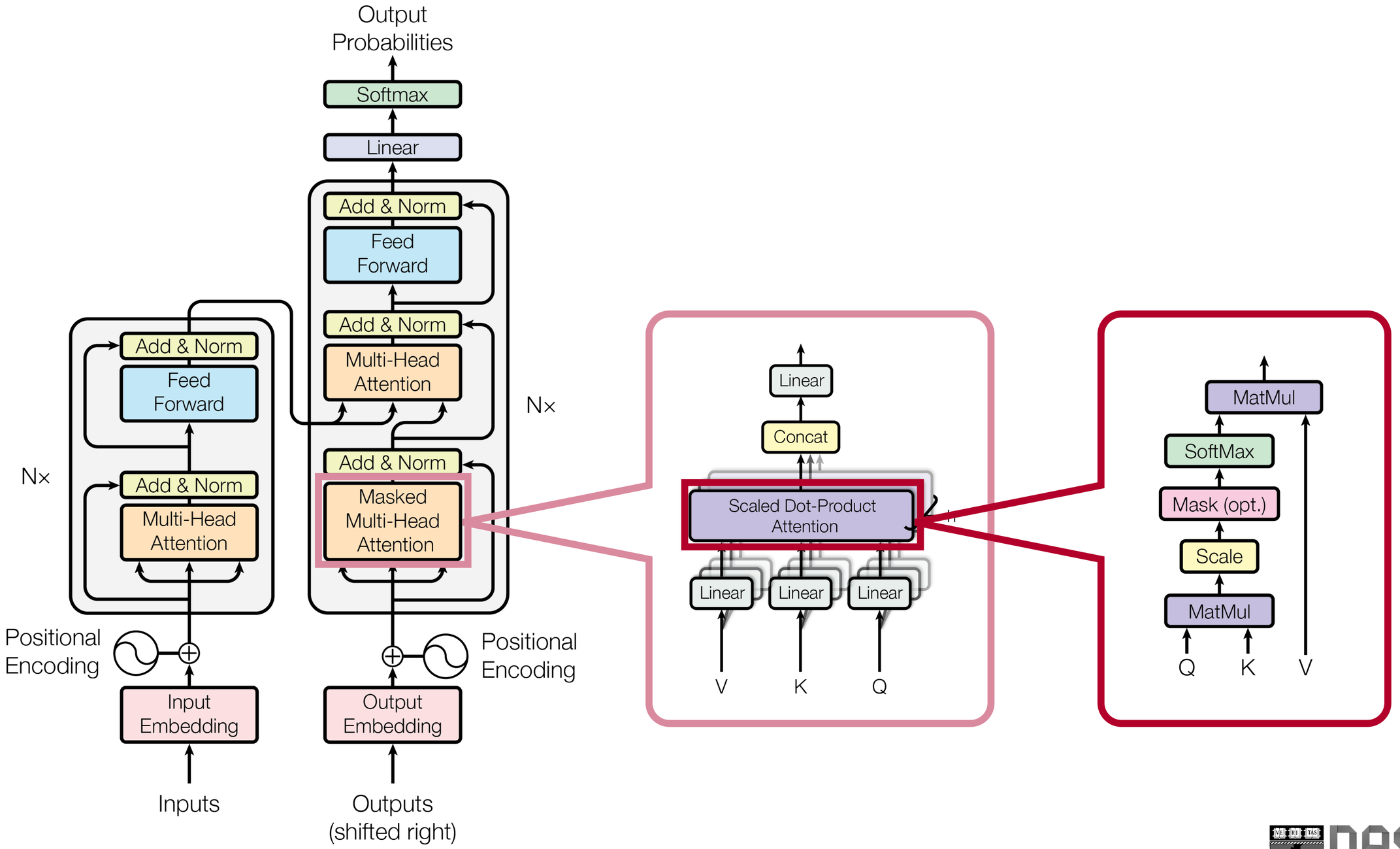
Want to start a startup? Get funded by Y Combinator.

the paragraphs relevant to the input

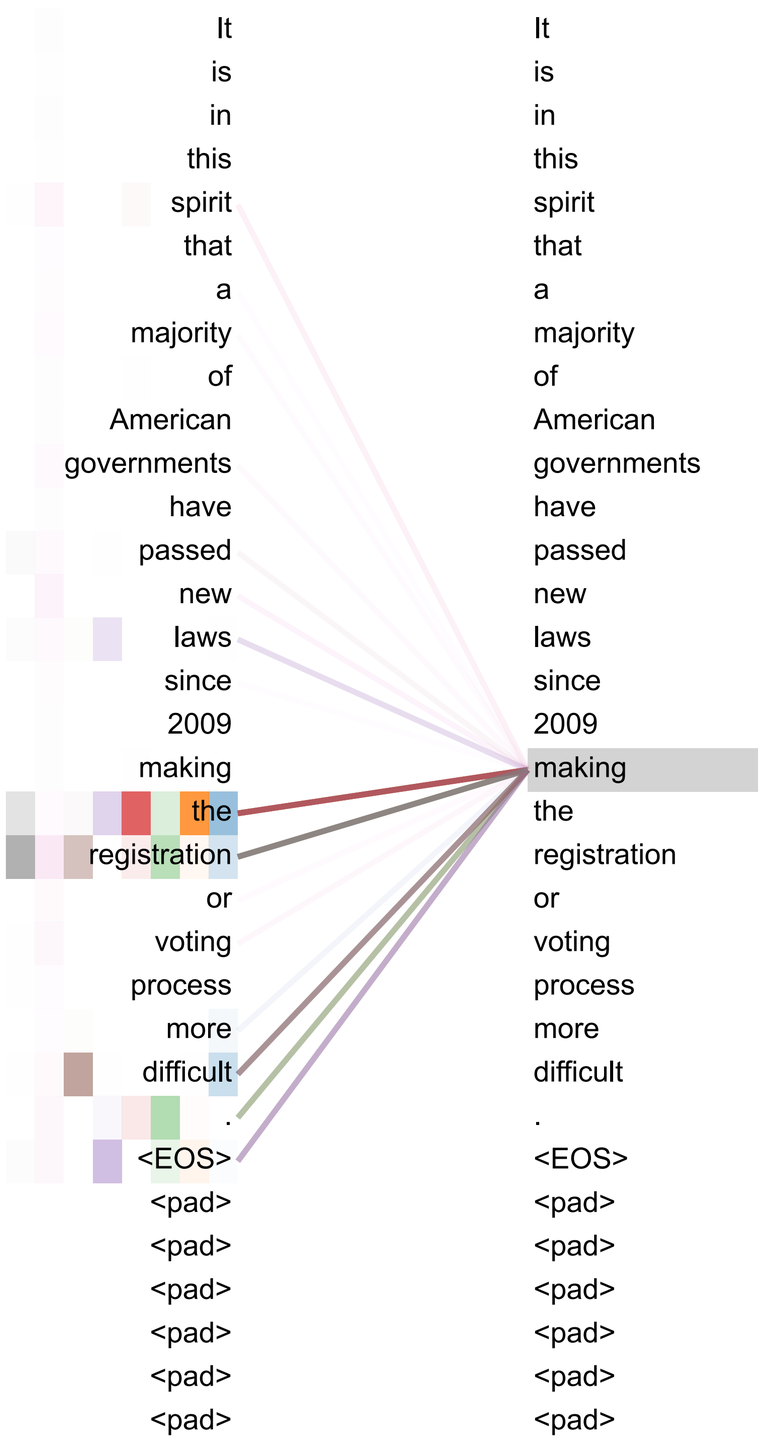
Retrieval-augmentation



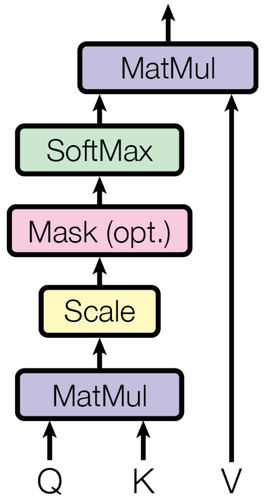
Let’s unbox LLM and retrieval.

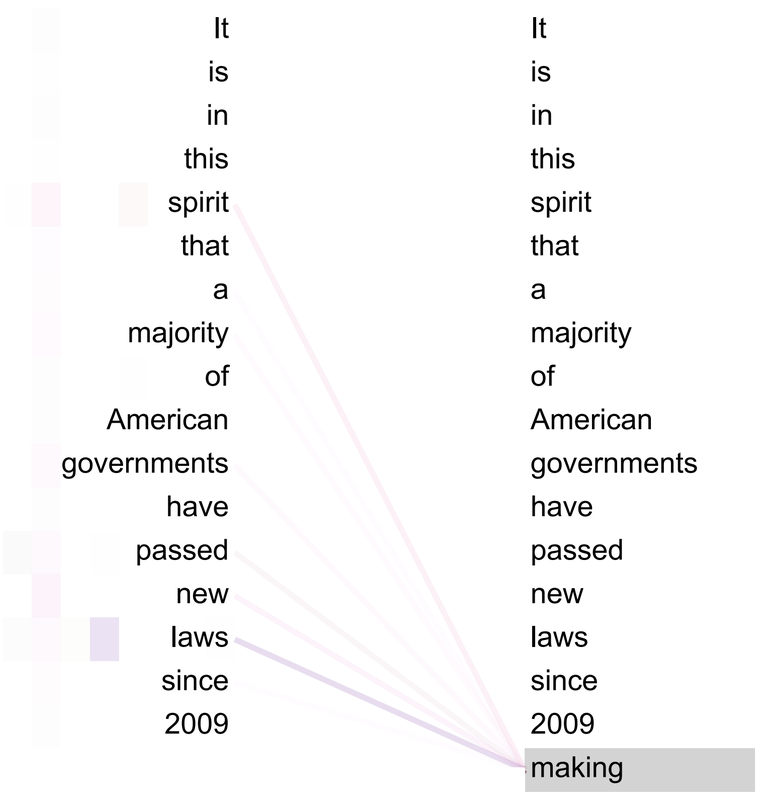


Ashish Vaswani, et al. Attention is All you Need. NIPS 2017

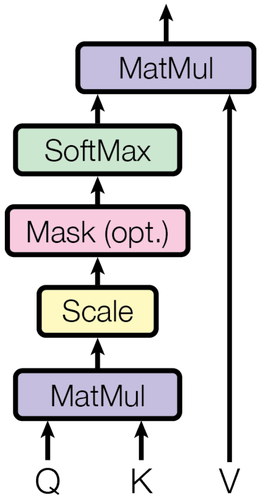


The meaning of a word is determined by its context

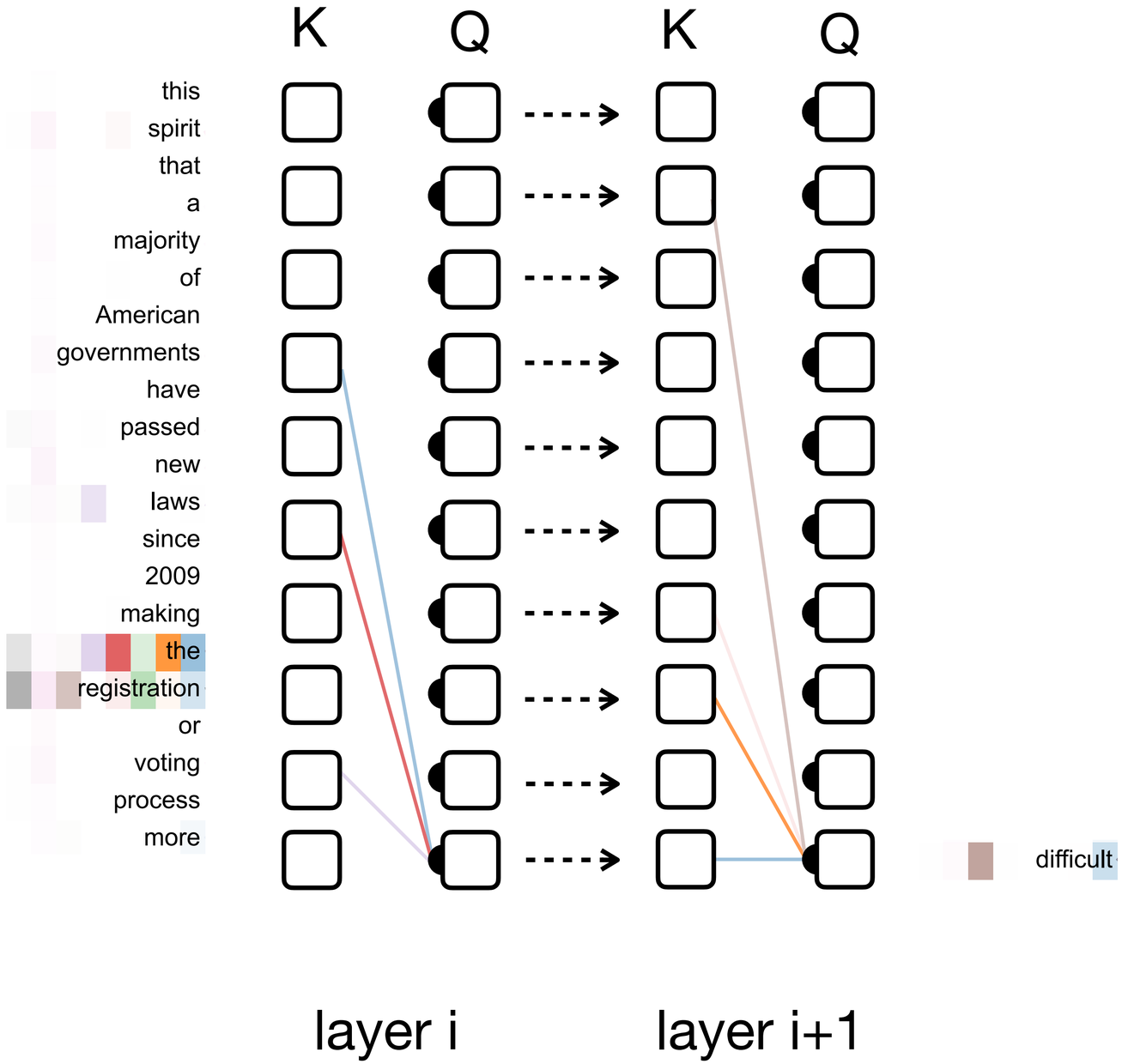




The meaning of a word is determined by its context

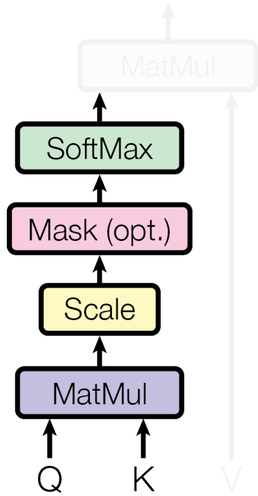


generate the next word: check the influence of all previous words

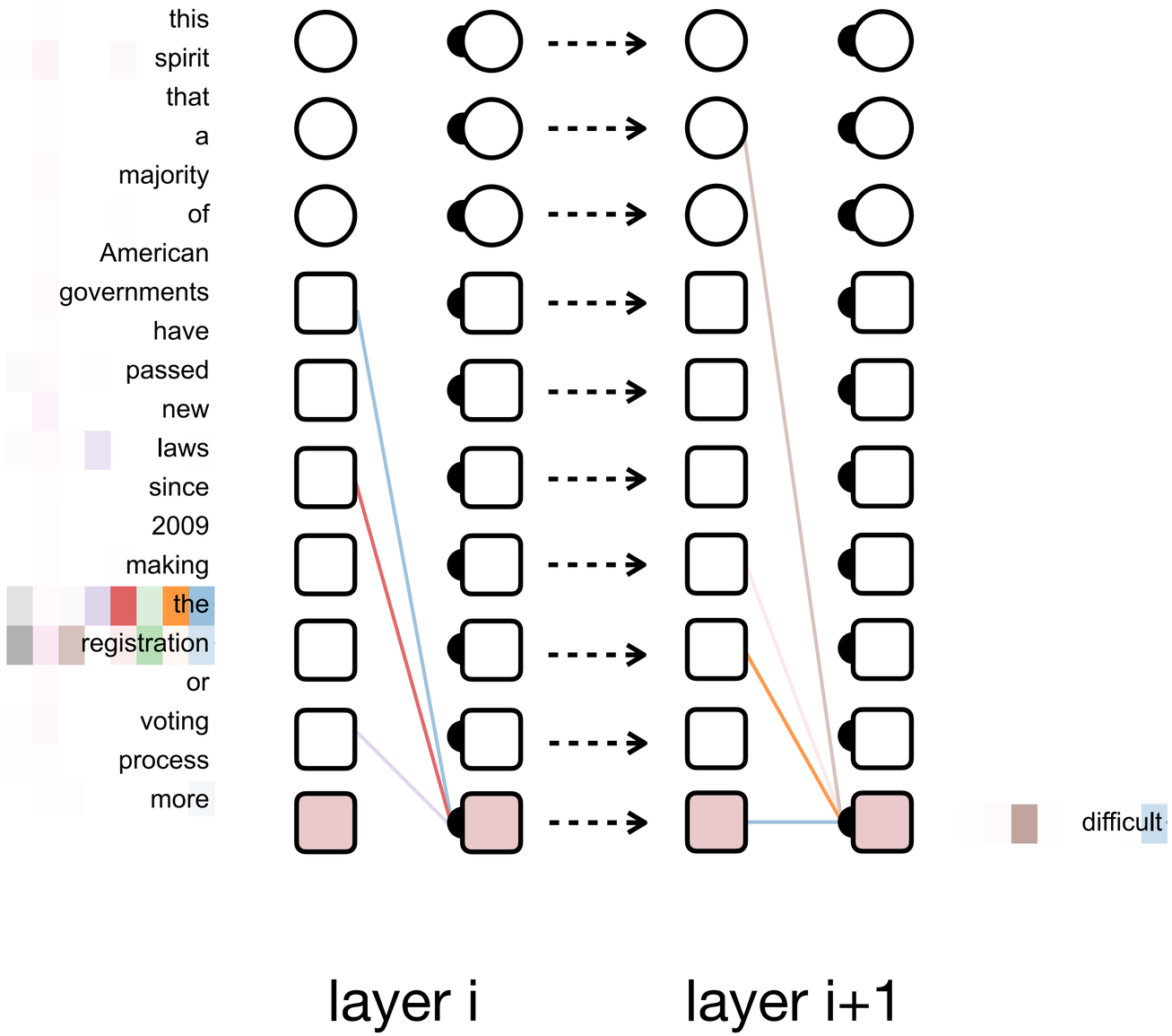


A semantic vector

[5.4, 1.7, 4.7, 3.4, …, 5.3, 5.1, 8.1, 6.9]

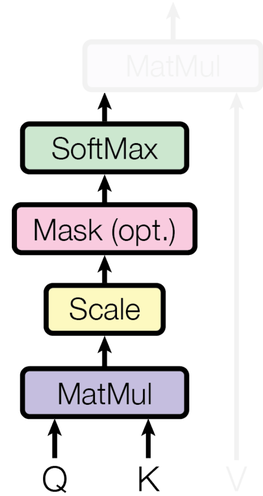


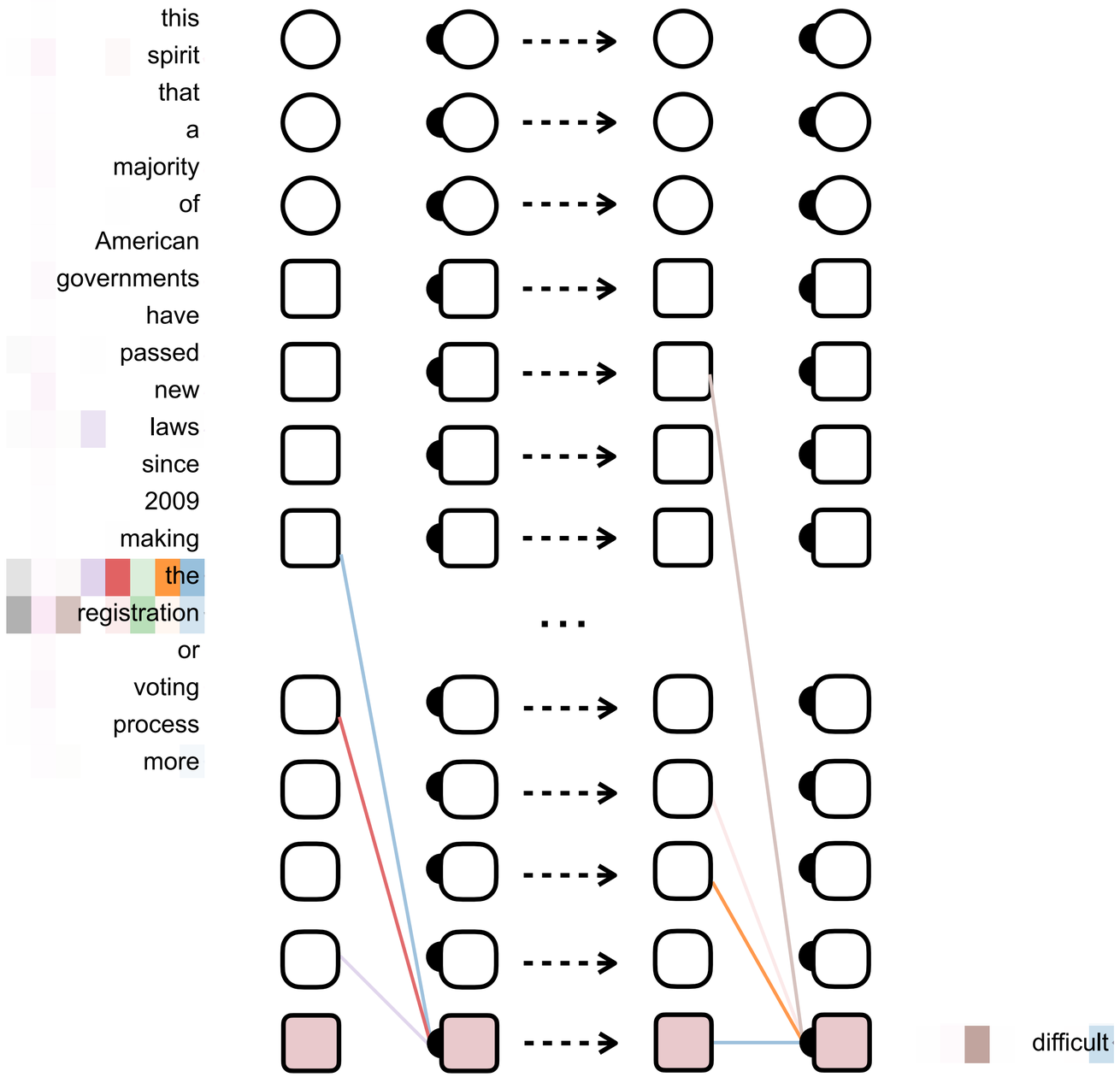
How LLM with retrieval works



A word in user request

A word in documents





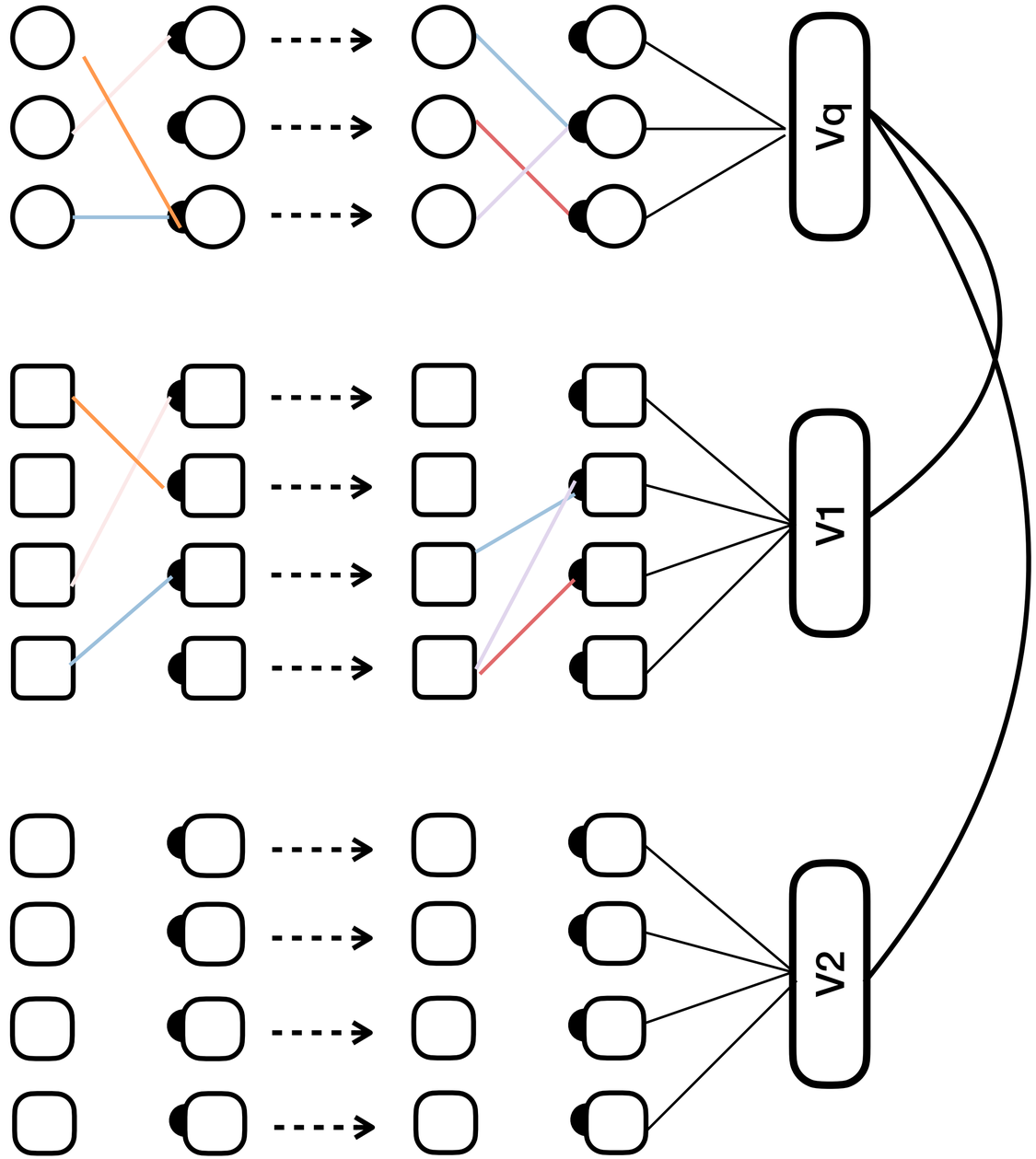
A word in user request

A word in document 1

A word in document 2

Problem:
Total words in documents grow (no. documents grows)

(Embedding) Vector search

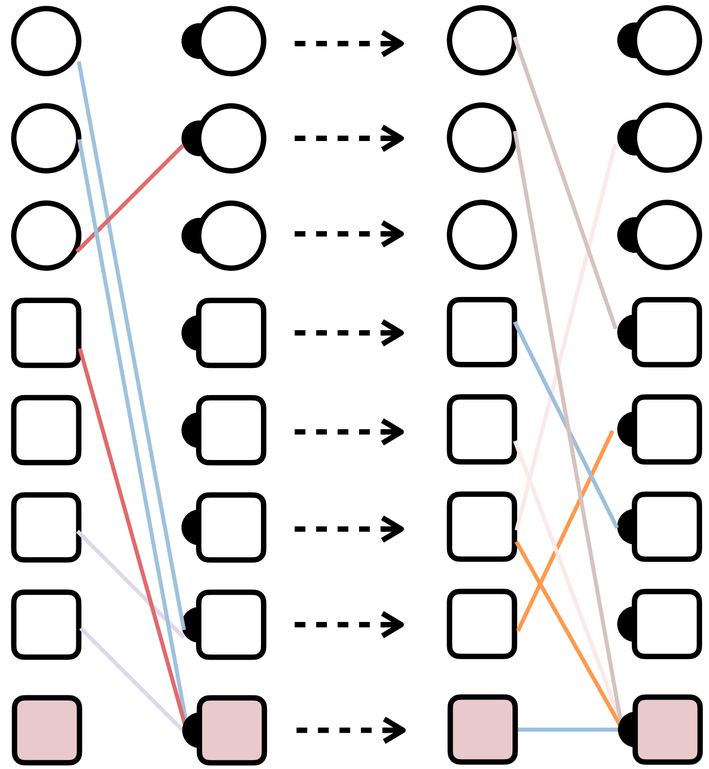


Solution (approximation) 1: Have one vector for the user request and each document

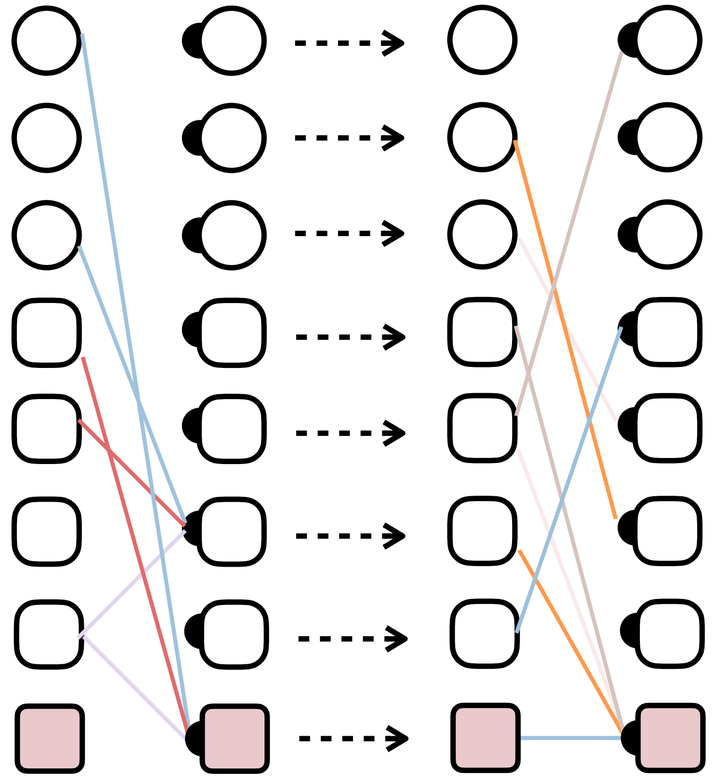
Compare (to score) document vectors with request vector

Select the most relevant documents

(Document) Rerank

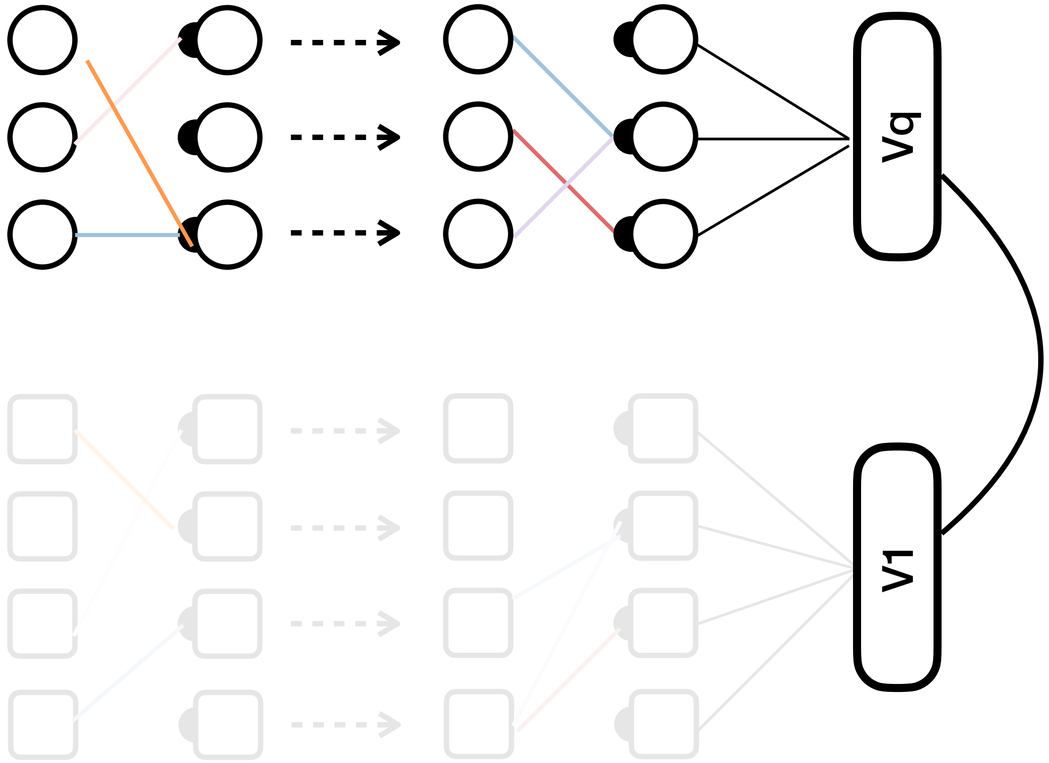


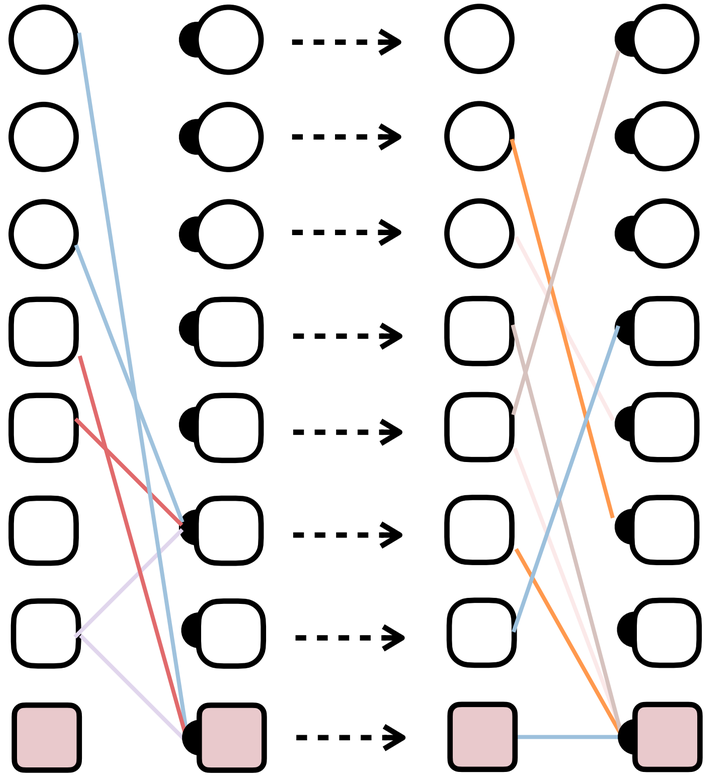
Relevant



Irrelevant

Solution (approximation) 2:
Interact one query-document at one time





Vector search

Speed +++
Accuracy +

?

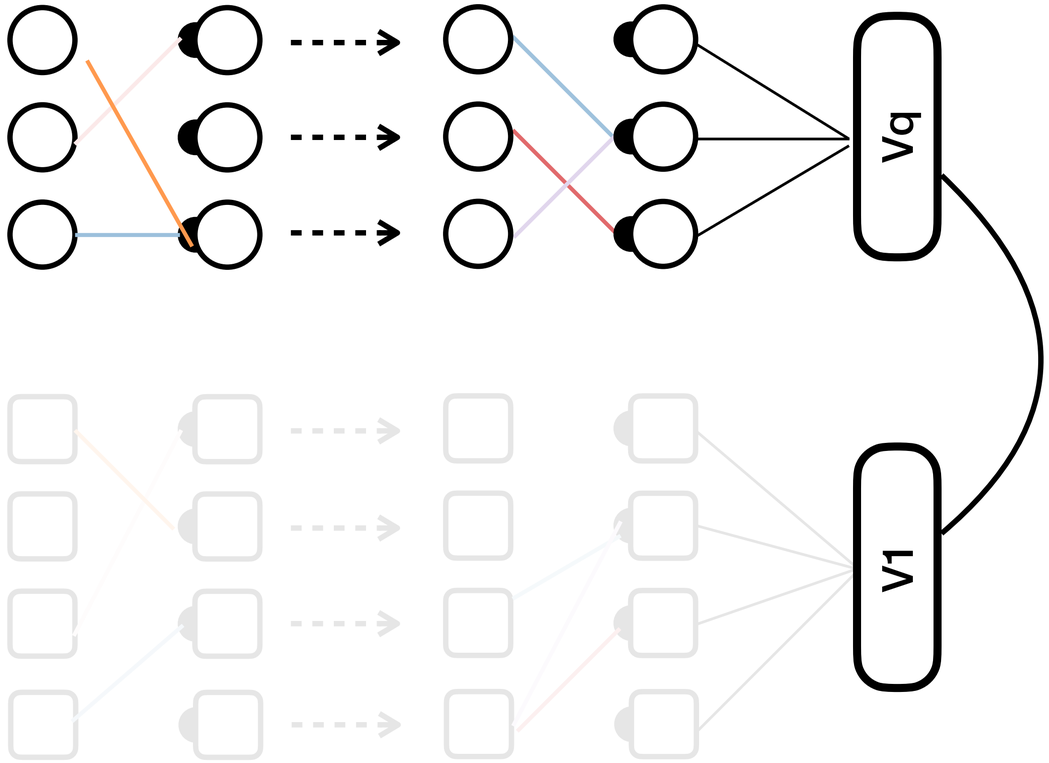
Reranking

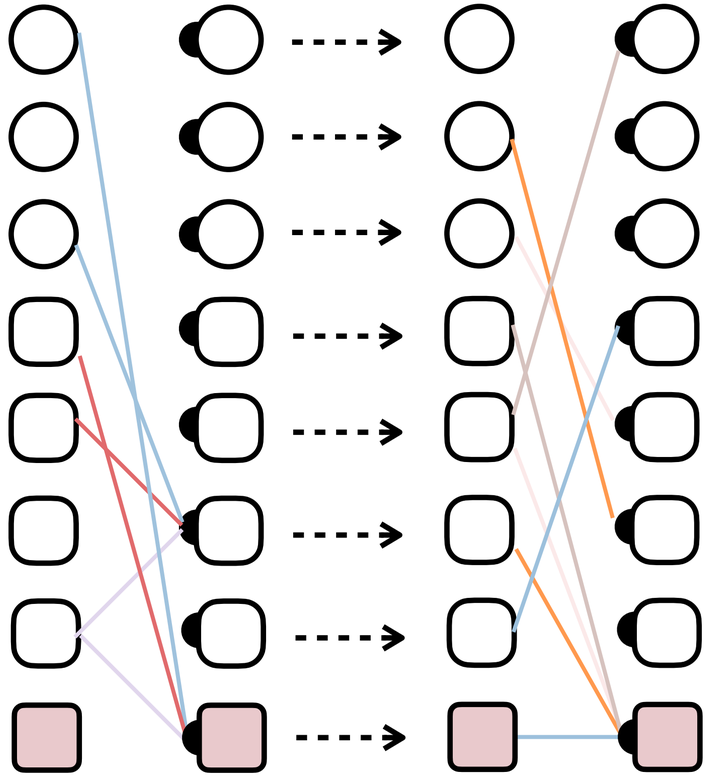
Speed ++
Accuracy ++

regular LLM

Speed ----
Accuracy +++

DASlab @ Harvard SEAS



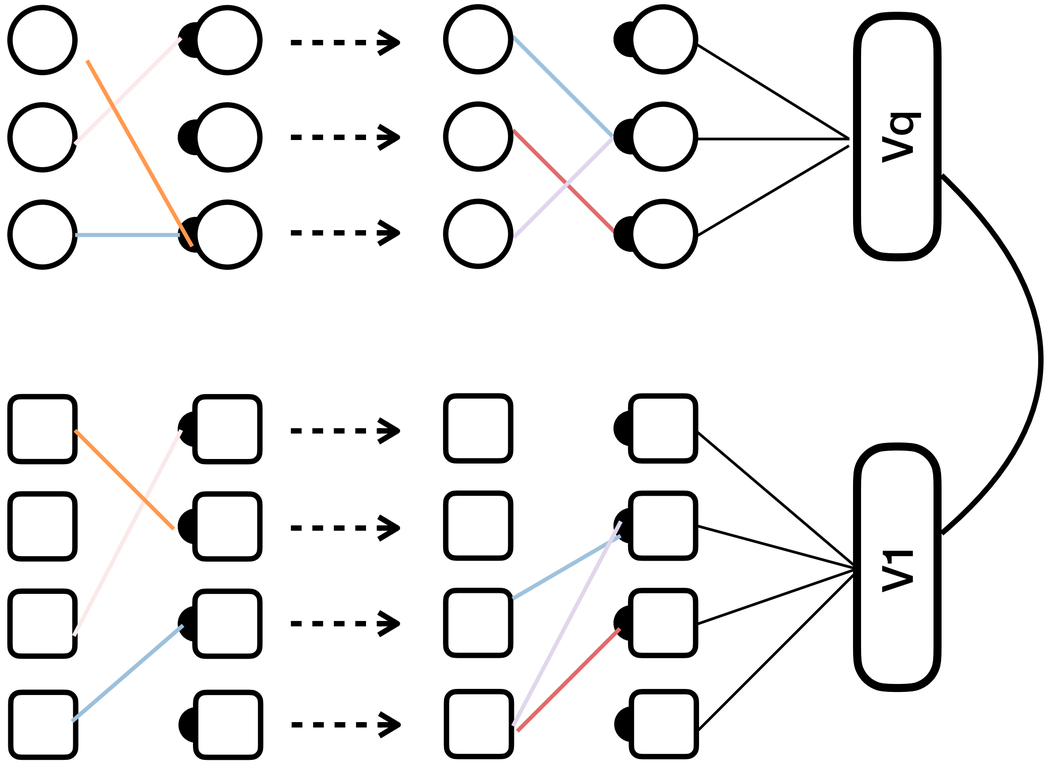


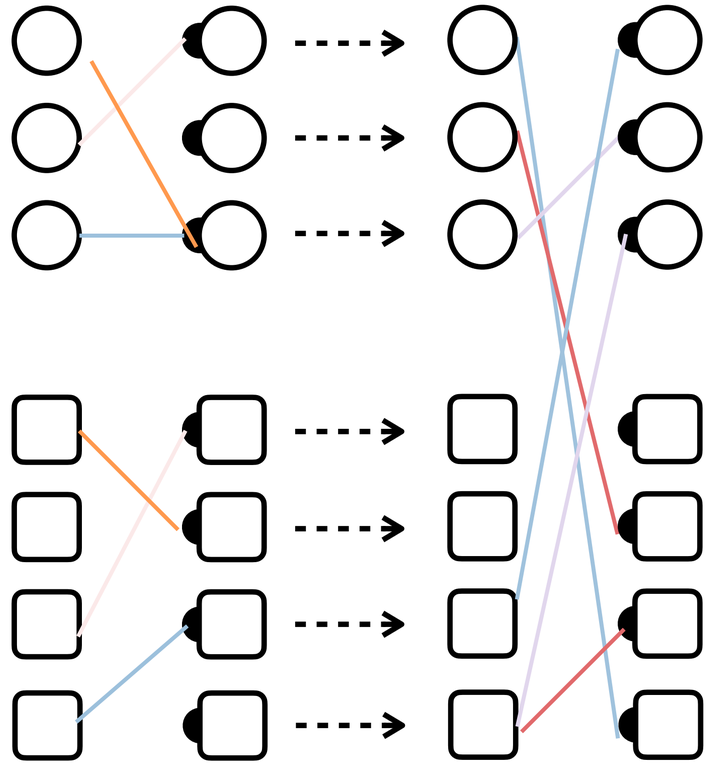
Vector search

?

Reranking

Other derivatives?





only last layer

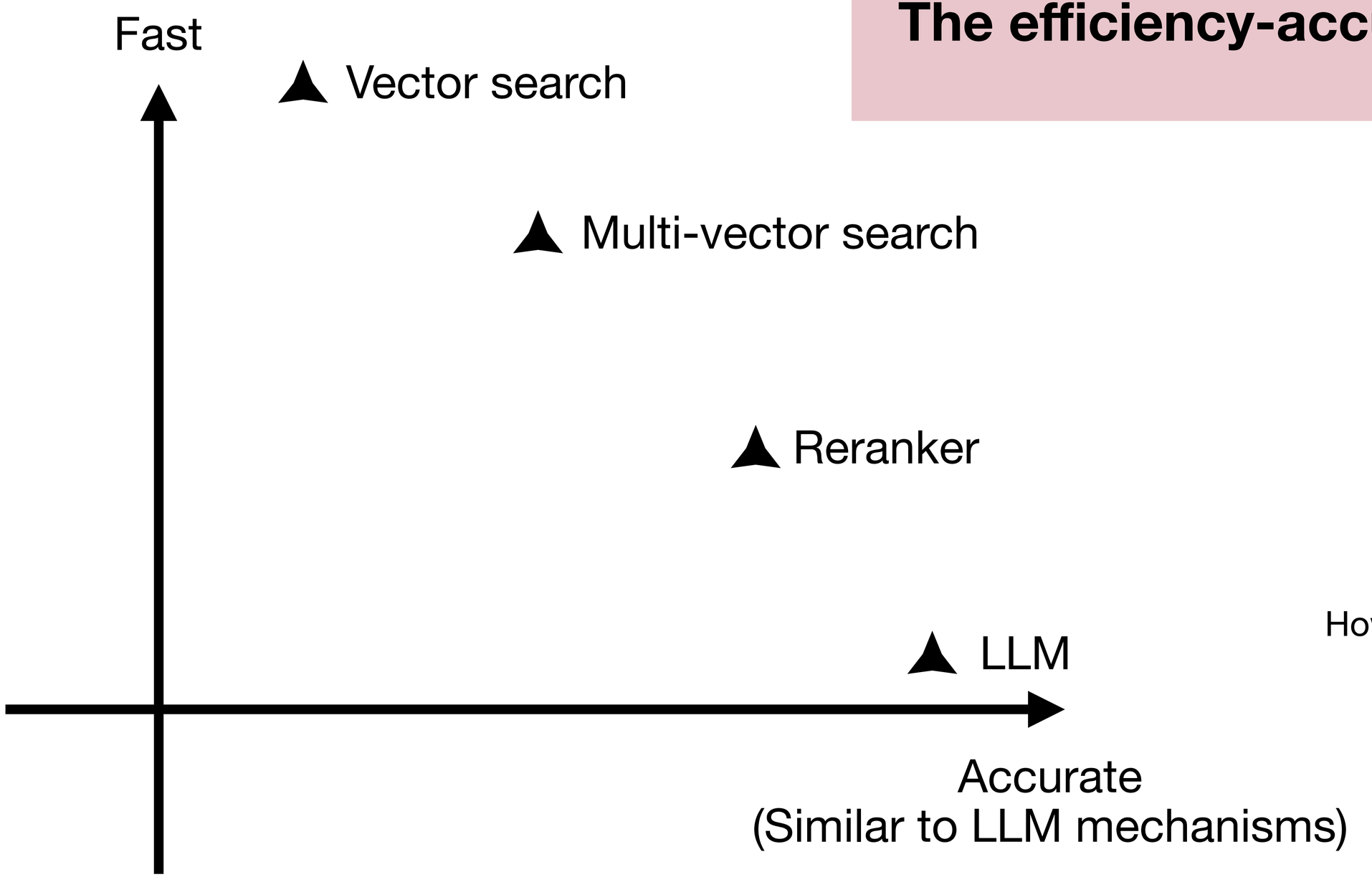
Vector search only one vector

Multi-vector search (ColBERT)

Reranking All layers, all vectors
but only one query-document pair at one time

LLM All layers, all vectors, all documents

The efficiency-accuracy tradeoff

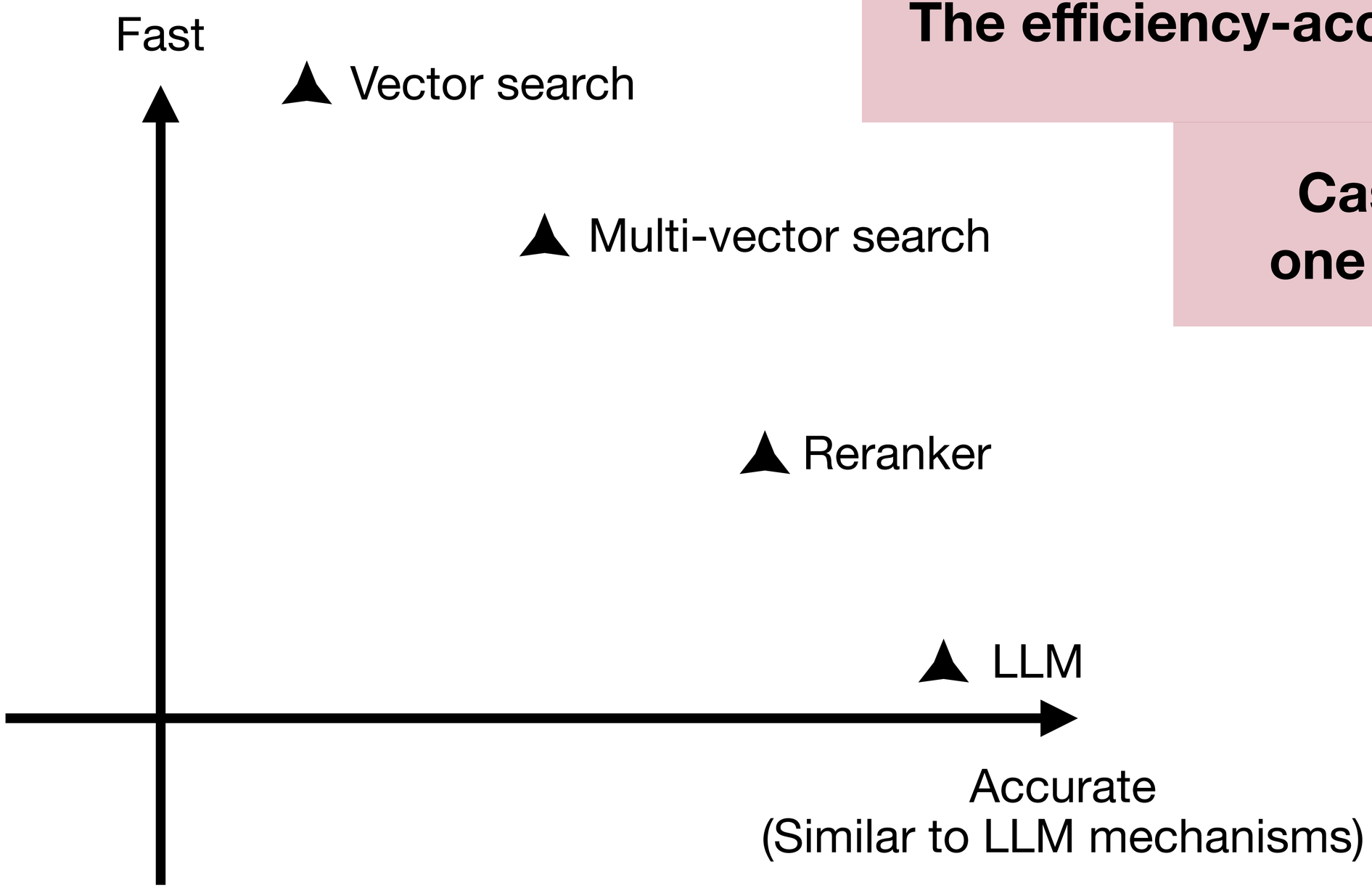


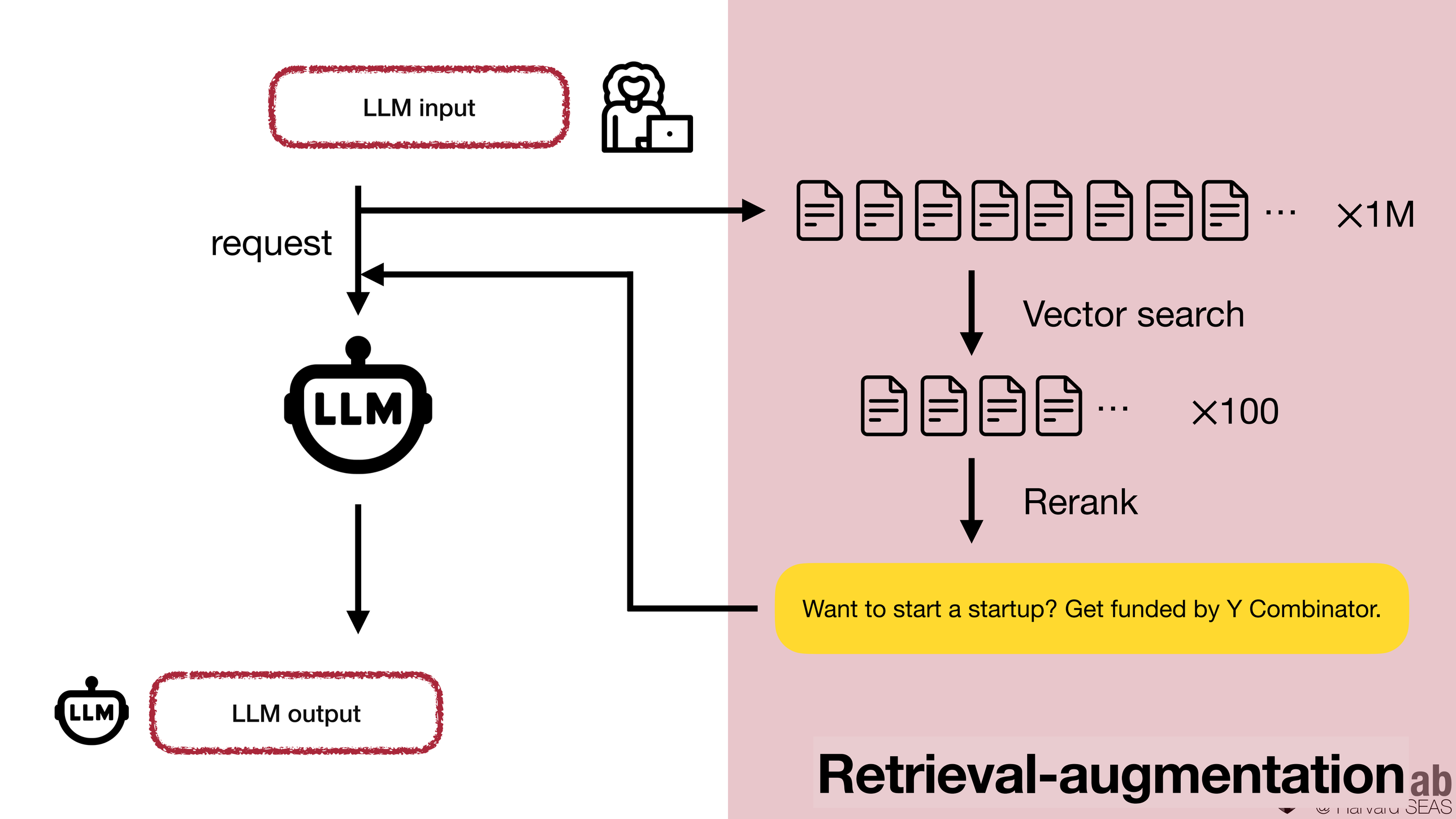
?

How to use them?

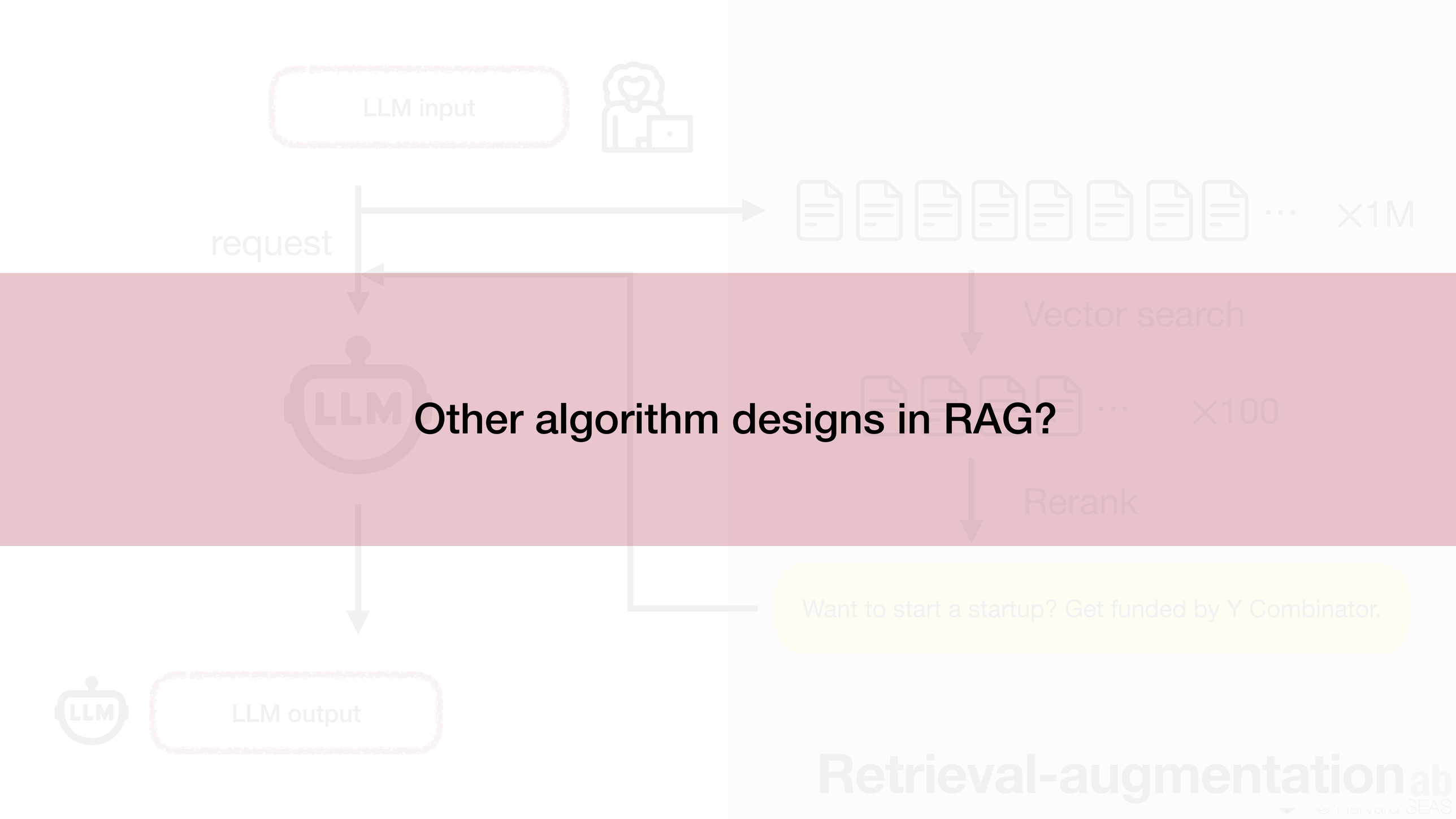
The efficiency-accuracy tradeoff

Cascaded into one RAG system

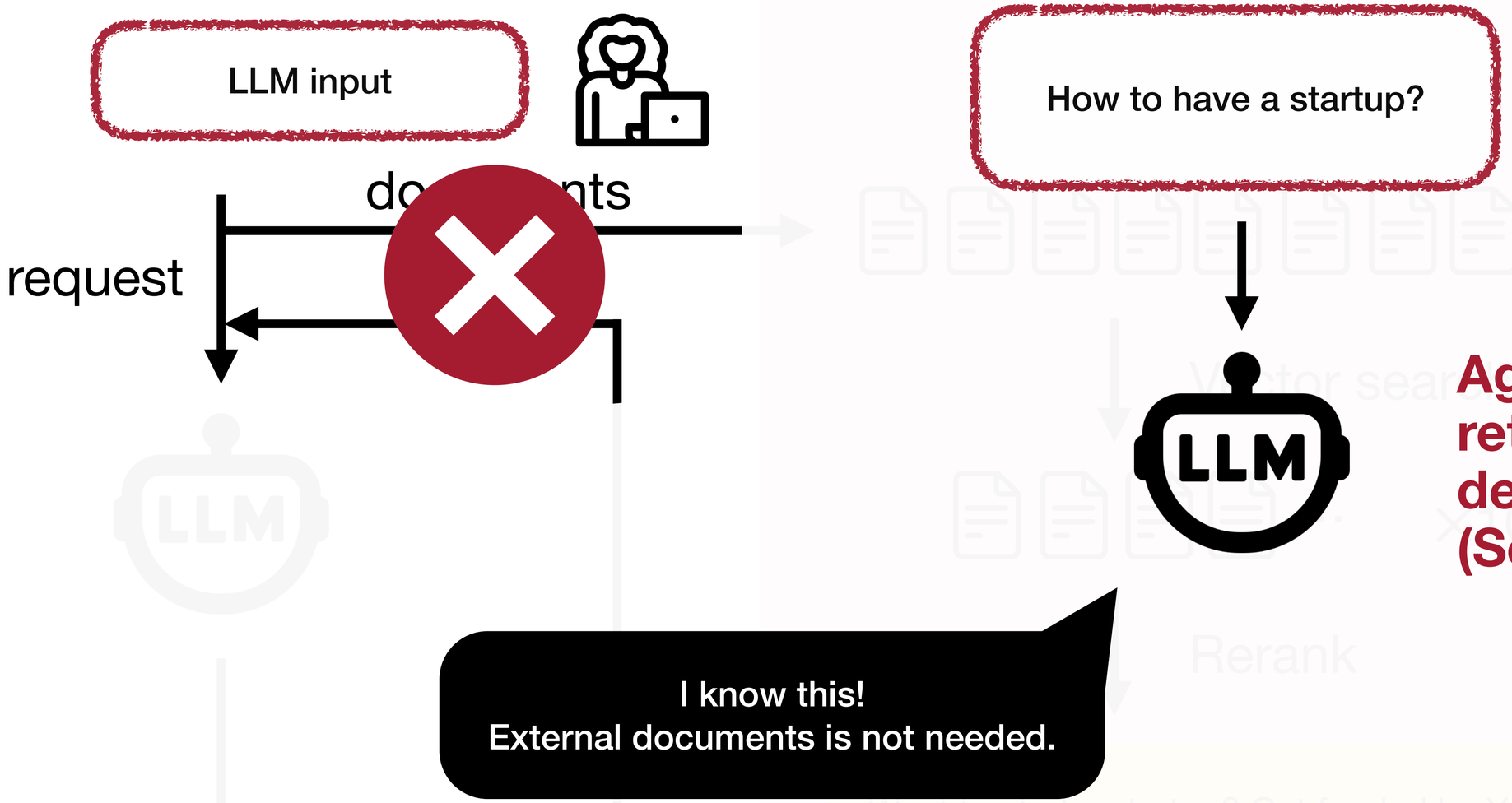


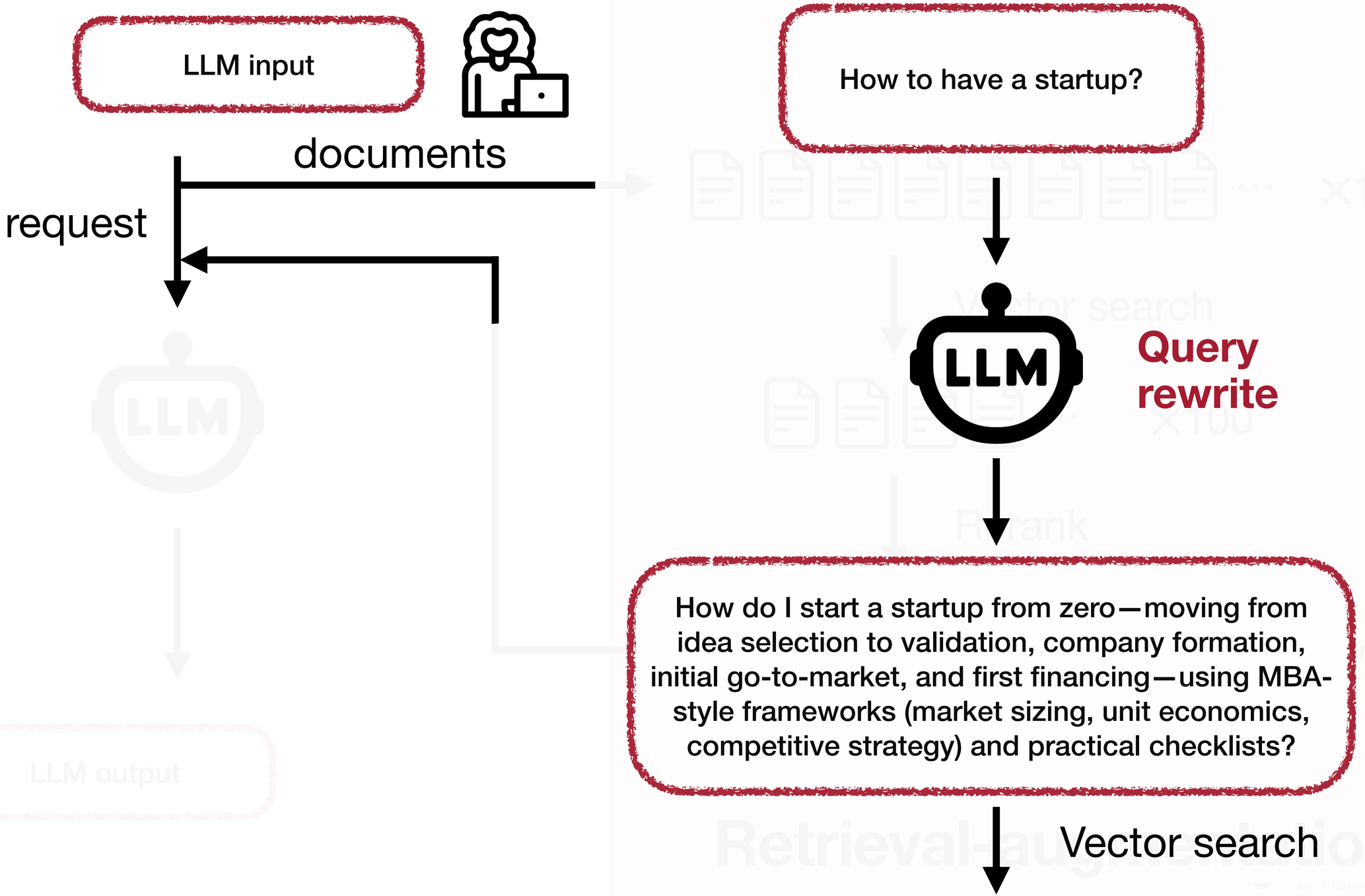


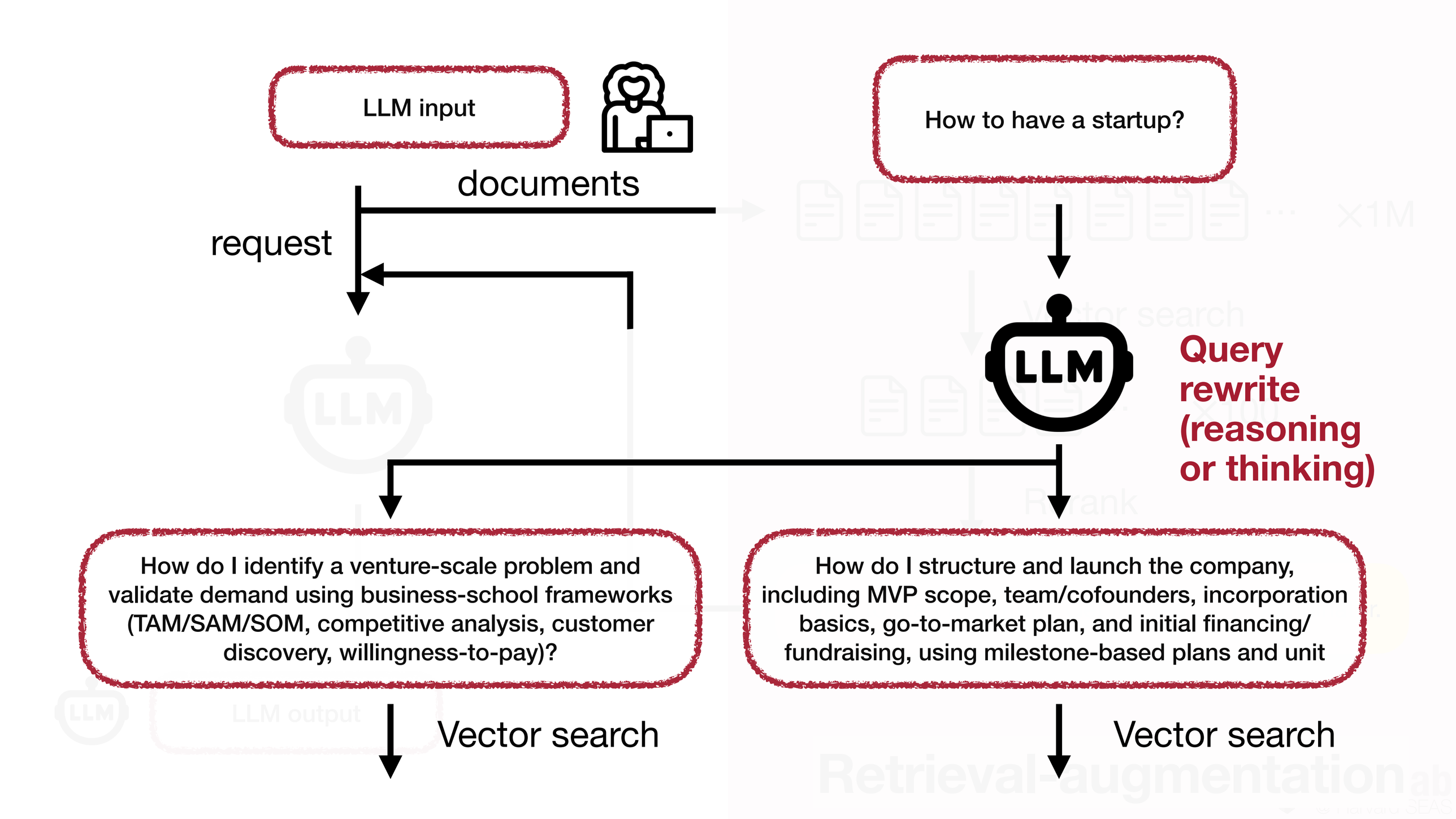
Retrieval-augmentation



Other algorithm designs in RAG?


LLM input
How to have a startup?
request
LLM
I know this!
External documents is not needed.
Agentic retrieval decision (Self-RAG)


LLM input
documents
request
How to have a startup?
LLM
Query rewrite
How do I start a startup from zero—moving from idea selection to validation, company formation, initial go-to-market, and first financing—using MBA-style frameworks (market sizing, unit economics, competitive strategy) and practical checklists?
Vector search


LLM input
How to have a startup?
documents
request
LLM
Query rewrite (reasoning or thinking)
How do I identify a venture-scale problem and validate demand using business-school frameworks (TAM/SAM/SOM, competitive analysis, customer discovery, willingness-to-pay)?
How do I structure and launch the company, including MVP scope, team/cofounders, incorporation basics, go-to-market plan, and initial financing/ fundraising, using milestone-based plans and unit
Vector search
Vector search

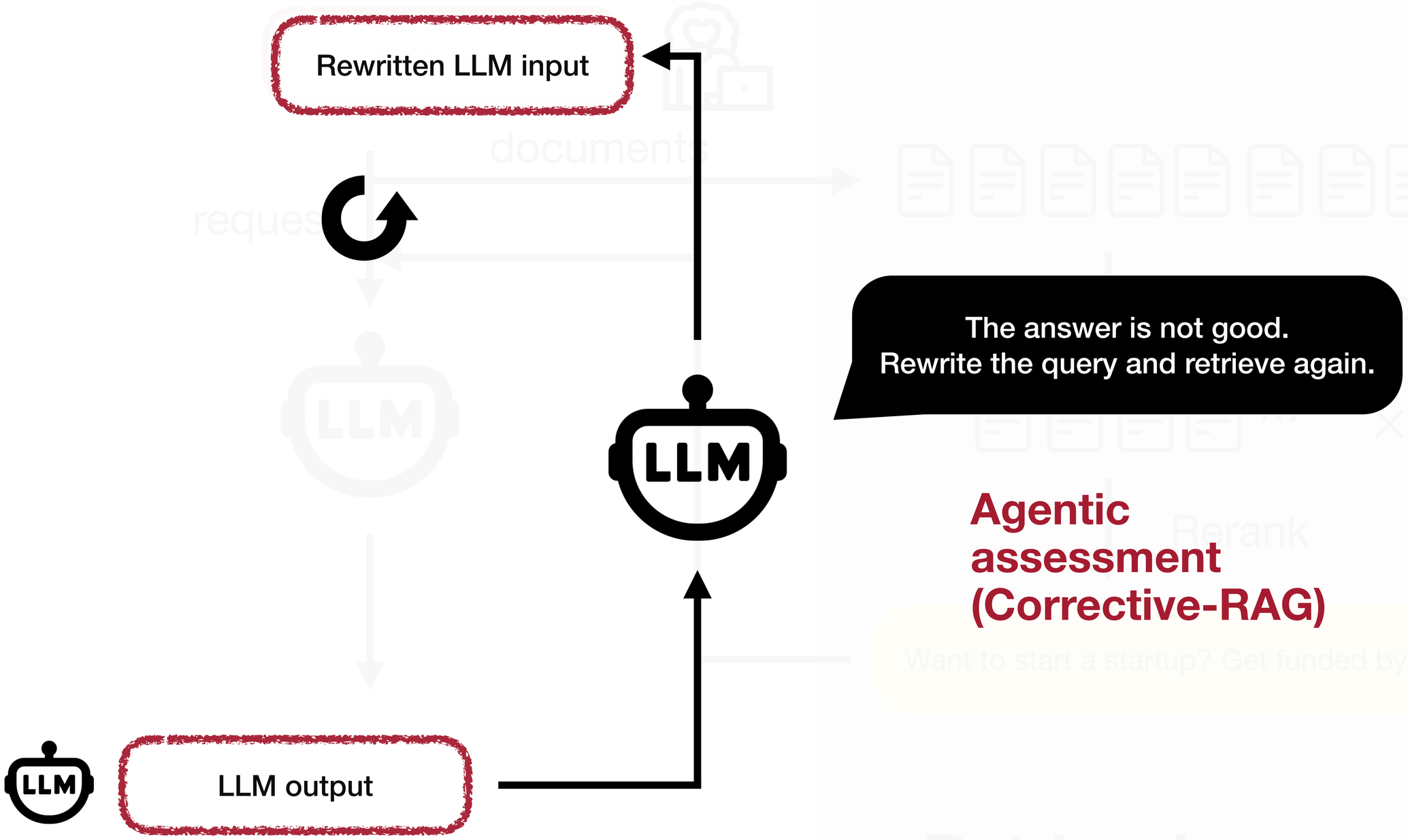
Want to start a startup? Get funded by Y Combinator.

Document rewrite

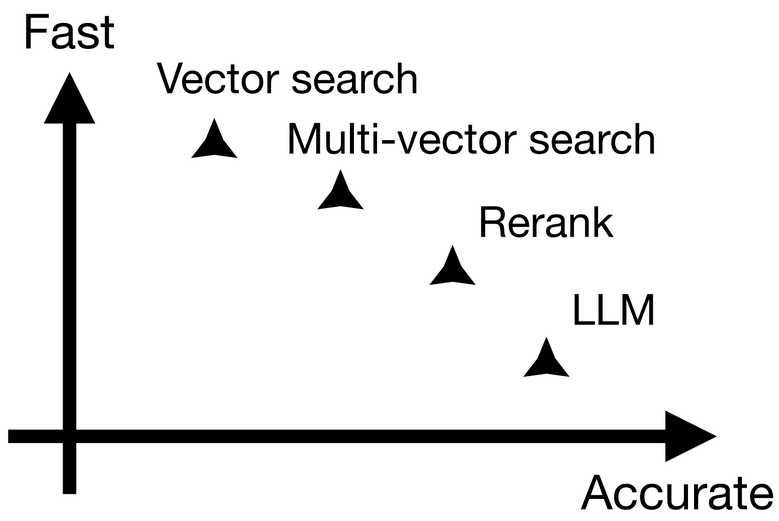
Want to start a startup? Y Combinator (YC) is a well-known early-stage accelerator that provides seed funding, mentorship, and a strong founder network in exchange for equity, helping teams quickly validate a problem, build an MVP, and gain early traction. The program ends with Demo Day, where startups pitch to many investors, which can improve follow-on fundraising and credibility, but it’s optional and comes with dilution and a bias toward fast growth, so apply when you can clearly explain the customer, the problem, why your team, and what milestones you’ll hit with the funding.

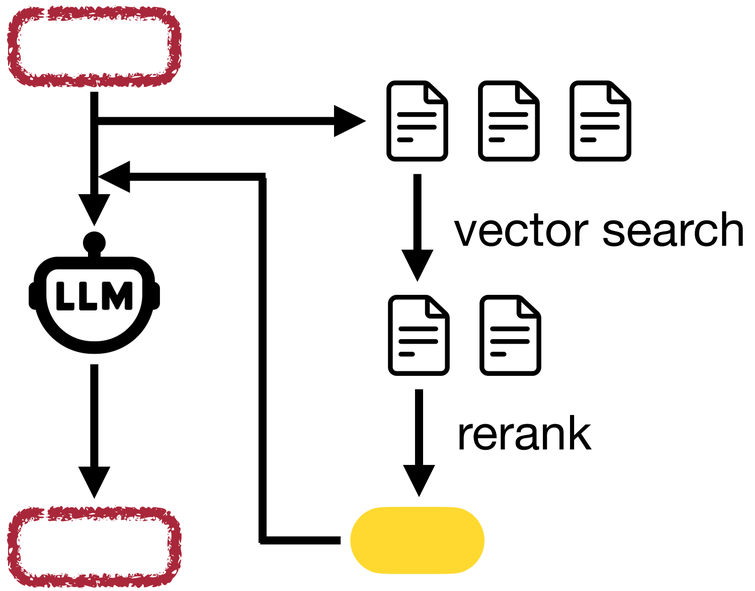
Want to start a startup? Get funded by Y Combinator.

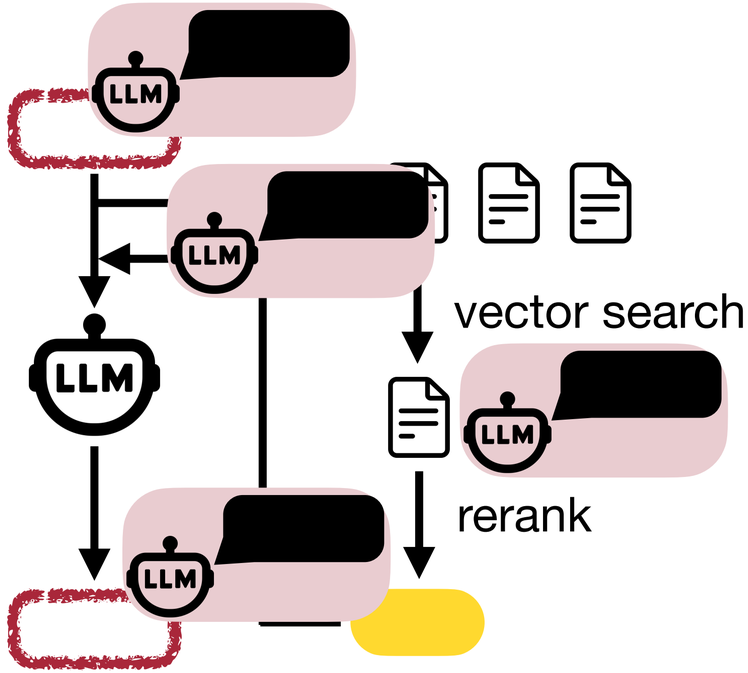
the paragraphs relevant to the input



Agentic assessment (Corrective-RAG)







- System designs
  Storage, pipelining, tuning, resource allocation
- Self-designing RAG systems

Next class