RAG

(Retrieval-augmented generation)

Erling og Peder

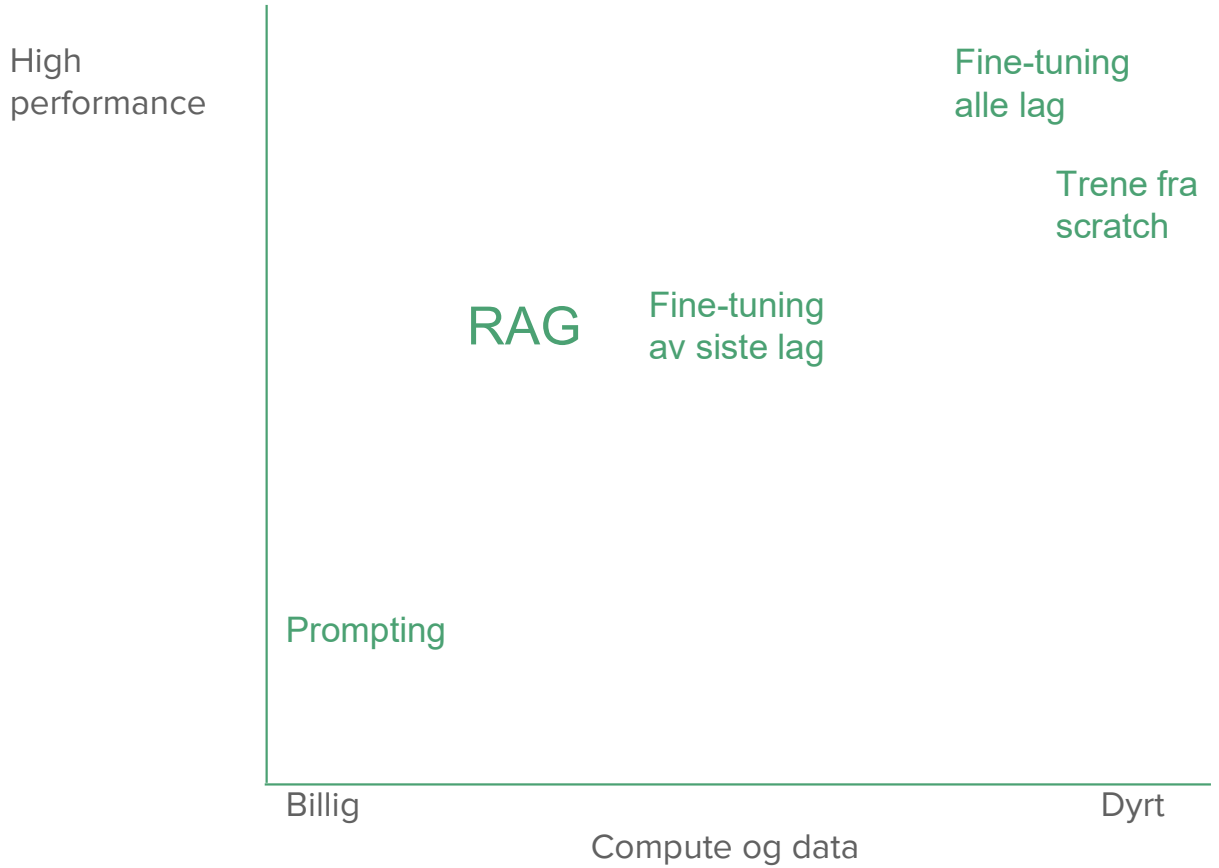
Problem:
Du har et sett med data hvor du vil bruke en LLM som grensesnitt til den informasjonen.

Underproblemer:

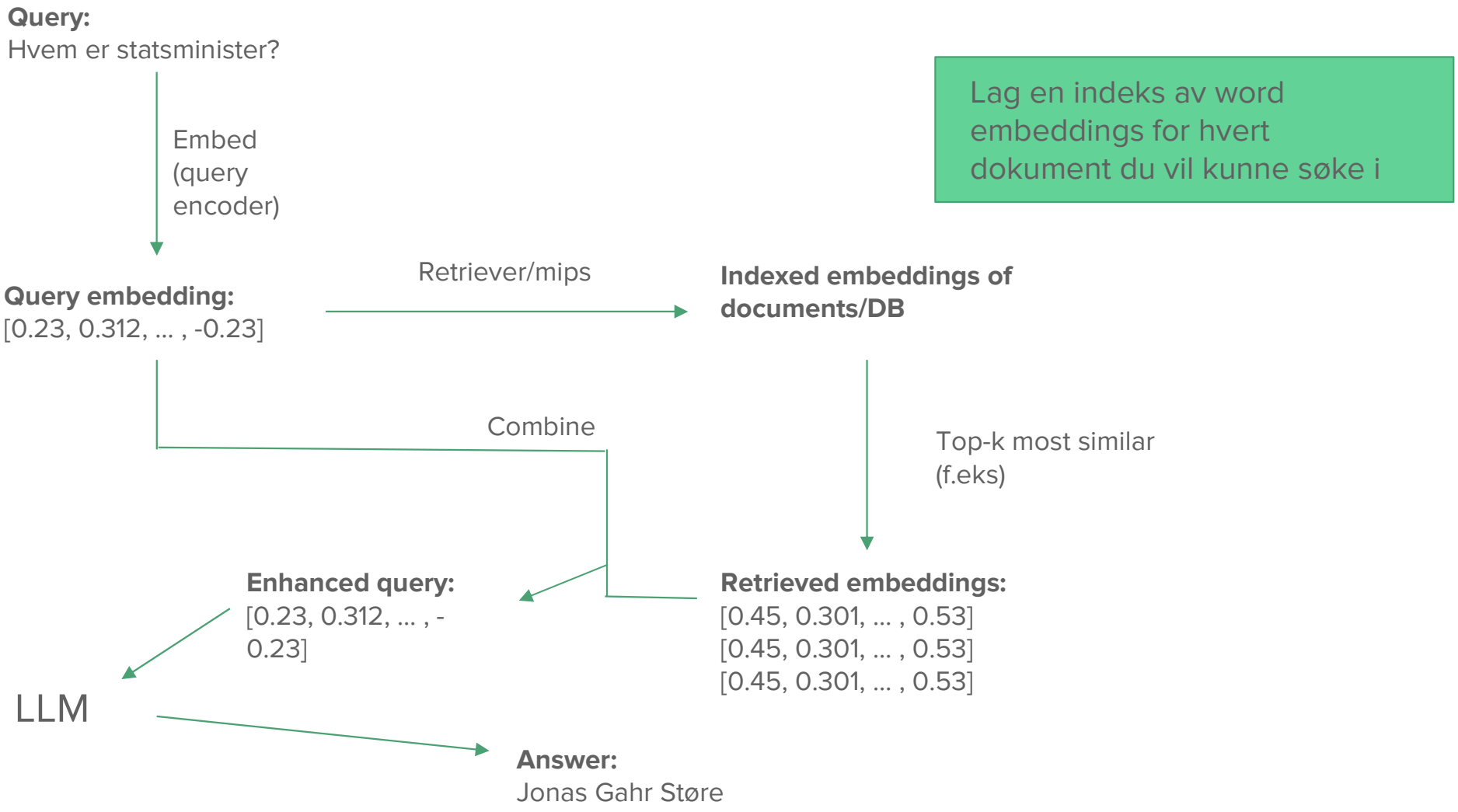
- Permission control
- Riktig data til riktig tid
- Begrenset med data
- Begrenset med compute
- Mindre hallusinasjoner
- Provenansinformasjon

Notat: Ikke-parametrisk minne er mer effektivt (kan gi flere garantier) og plassbesparende enn parametrisk minne som vi finner i LLMer

- RAG lar oss gi riktig informasjon under query-time
- gir oss direkte kontroll over hvilke brukere som har tilgang på hvilken data
- gir flere garantier for at modellen ikke hallusinerer
- Kan fortelle oss akkurat hvor den fikk informasjonen fra



* Ikke i skala


Query:
Hvem er statsminister?
Embed (query encoder)
Lag en indeks av word embeddings for hvert dokument du vil kunne søke i
Query embedding:
[0.23, 0.312, ... , -0.23]
Retriever/mips
Indexed embeddings of documents/DB
Combine
Top-k most similar (f.eks)
Enhanced query:
[0.23, 0.312, ... , -0.23]
Retrieved embeddings:
[0.45, 0.301, ... , 0.53]
[0.45, 0.301, ... , 0.53]
[0.45, 0.301, ... , 0.53]
LLM
Answer:
Jonas Gahr Støre

Generere output

- Rag-Token:
  for hvert (top-k) dokumenter:
  - Bruk en LLM til å generere en output gitt: query+dokument
  - Velg output med høyest sannsynlighet - returner til bruker

- Rag-Sequence:
  for hver token i query
  - Bruk en LLM til å generere en output gitt: query + tidligere genererte tokens + top-k dokumenter

Fungerer RAG?

Lewis et. al (2021) *Retrieval Augmented Generation for Knowledge-Intensive NLP Tasks*

4 tester:
- Open domain question answering
- Abstractive question answering
- Jeopardy question generation
- Fact verification

Gjennomgående bedre på fakta enn LLM-er uten RAG, men det varierte i hvor stor grad.

Table 3: Examples from generation tasks. RAG models generate more specific and factually accurate responses. '?' indicates factually incorrect responses, * indicates partially correct responses.

Task	Input	Model	Generation
MS-MARCO	define middle ear	BART	?The middle ear is the part of the ear between the middle ear and the nose.
		RAG-T	The middle ear is the portion of the ear internal to the eardrum.
		RAG-S	The middle ear includes the tympanic cavity and the three ossicles.
	what currency needed in scotland	BART	The currency needed in Scotland is Pound sterling.
		RAG-T	Pound is the currency needed in Scotland.
		RAG-S	The currency needed in Scotland is the pound sterling.
Jeopardy Question Gener -ation	Washington	BART	?This state has the largest number of counties in the U.S.
		RAG-T	It's the only U.S. state named for a U.S. president
		RAG-S	It's the state where you'll find Mount Rainier National Park
	The Divine Comedy	BART	*This epic poem by Dante is divided into 3 parts: the Inferno, the Purgatorio & the Purgatorio
		RAG-T	Dante's "Inferno" is the first part of this epic poem
		RAG-S	This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso"