Norway’s National Data Portal – improving data discovery using AI

19 – 20 MARCH 2025

Introduction

- Data.norge.no is the national data portal of Norway
- The database consists of RDF-descriptions compatible to specifications and standards (i.e. dcat-ap-no)
- In this presentation I will demonstrate how and why we turned to using Artificial intelligence (AI) and Large Language Models (LLMs)

Background – what were the user needs?

- Datasets are often described using a field specific jargon.
- Users looking for datasets did not always know the terms used in the descriptions of the datasets they were interested in.
- Users did not know what data they needed but often knew what they wanted to use the data for.

What alternatives did we consider?

- Make the data owners improve their descriptions
- Tune relevance of elastic search engine
- Improve documentation and information about the search
- Experiment with Artificial intelligence

Why AI?

- We wanted more experience with AI
- The Minister of Digitalisation and Public Governance in Norway expressed that she wanted more of the public sector to use AI.

Why LLM search?

- LLM technology works well to roughly navigate and extract information from large text hubs using natural language.
- Short way from idea to market.
- Possible to alter as we gathered more insight and knowledge

Worries before we started

Cost

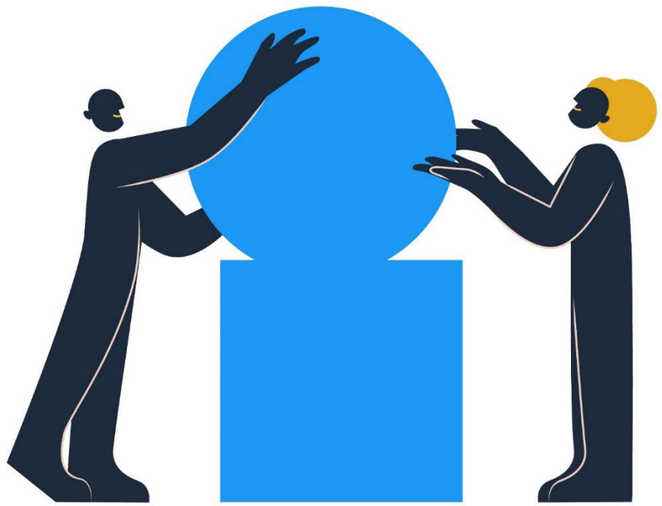
Quality

Lack of AI-competence

Security and privacy

What made it possible to start experimenting?

- Cross sector cooperation and initiative
- We had knowledge about our database.
- We had well structured metadata
- Others had time and resources to experiment with AI.



The prototype - «oracle search»

- Elapsed time: 3 weeks.
- Intentionally made with the same tech stack we already use.
- Cost: ~170 EUR/month of which 90% is cost related to Postgres database.
- Limitations: Based on one single extraction from the dataset catalogue. No dynamic update of data.

Findings after prototype

- Existing procedures for extracting metadata from the data catalogue made the work easier.
- The challenging part was choosing what information it was important to include.
- No automatic testing, but we manually tested a set of known user queries, to compare with existing solution
- User’s tests were positive.
- Positive feedback from other parts of the organization.
- The costs were lower than expected.

Result: We decided to put the AI search in production as part of national data portal - data.norge.no.

Taking the prototype to production in busy 3 months

June 2024:

- Competence transfer
- Mapping and architecture
- Make architecture proposal
- Approve architecture proposal
- Documentation

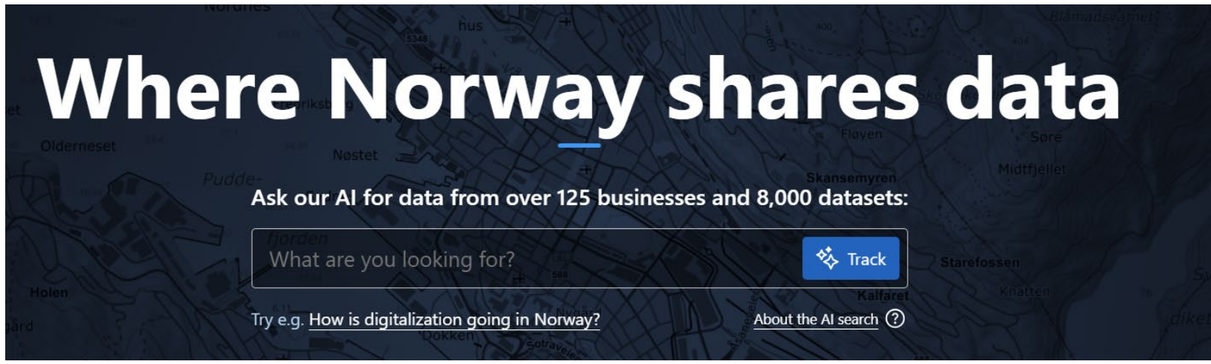
July 2024:

- Implementation (~2 weeks)
- Risk and vulnerability analysis

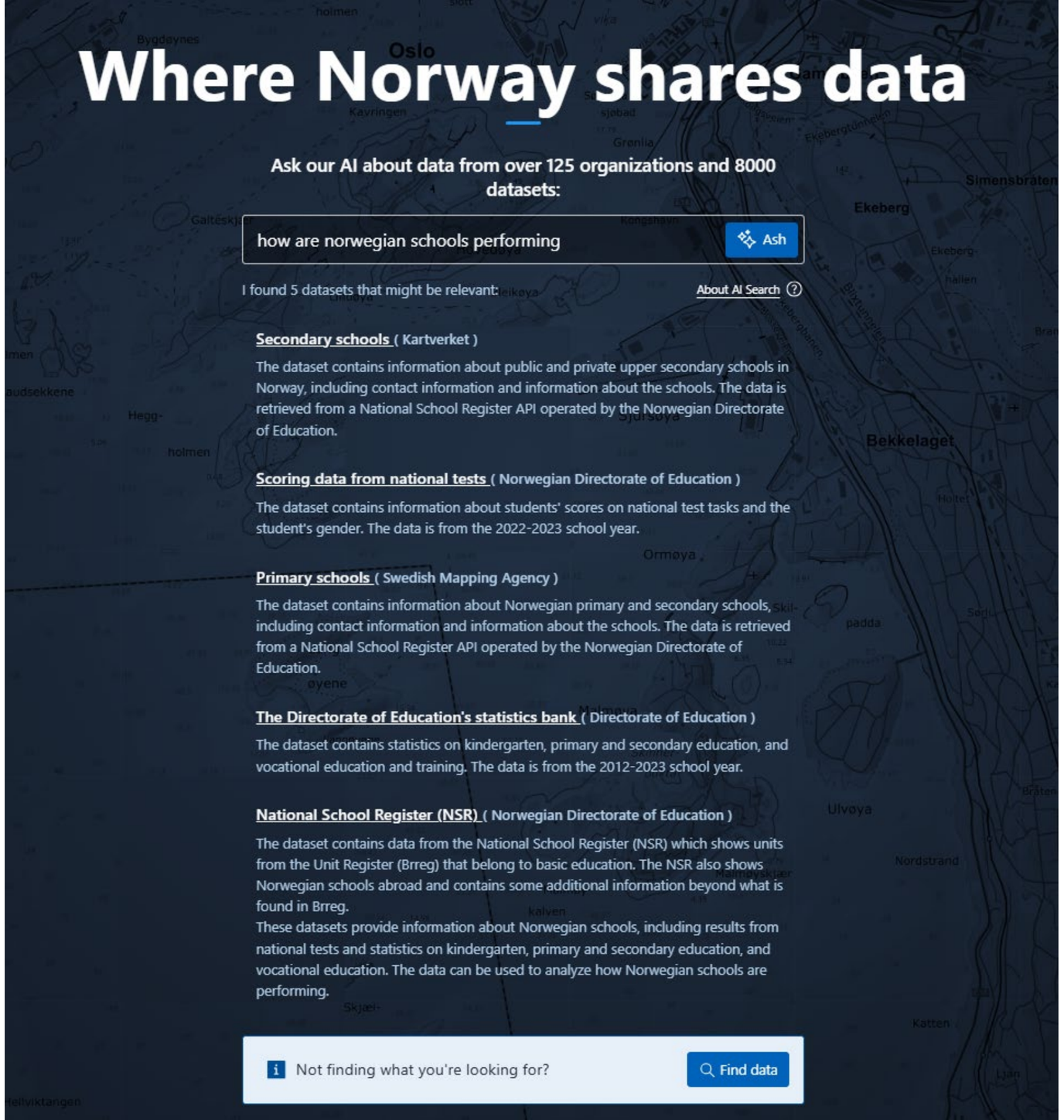
August 2024:

- GDPR and privacy
- Testing and security verification

August 23rd
Production



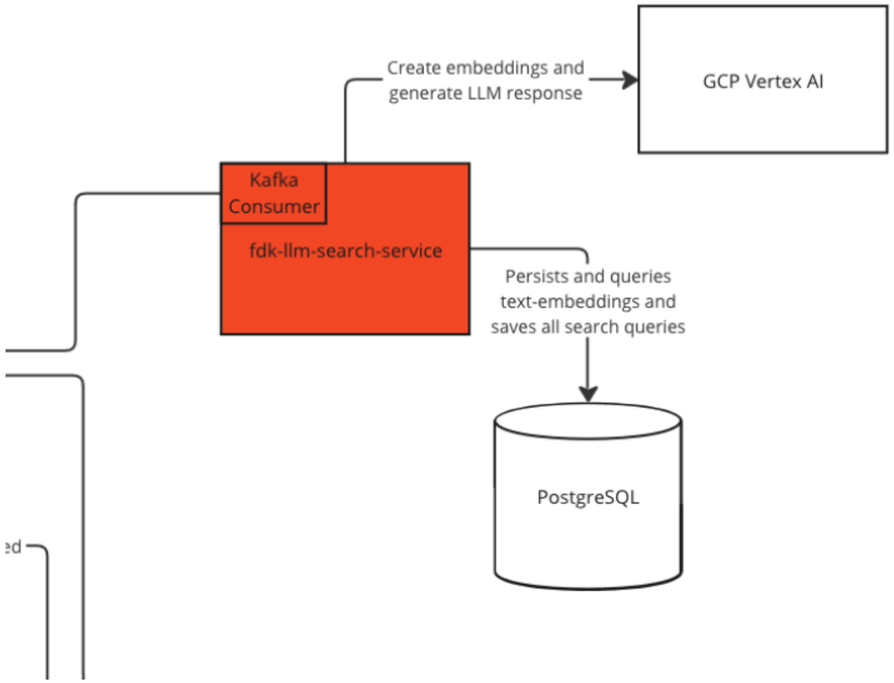
EU OPEN DATA DAYS



EU OPEN DATA DAYS

How does it work?

- Information on datasets from the dataset catalog are gathered into a Postgres database, saved as textual values.
- Uses Pgvector and Vertex AI to vectorize the datasets. Max 7 datasets are chosen.
- Prompt is created using LangChain to make Vertex AI filtrate datasets that are not considered relevant for the query.
- Vertex AI returns answers with an explanation of why these are considered as relevant datasets for the query.



Pros and cons with our AI search

Pros:

- By using an LLM as is, it will be easier for us to replace it with newer and better LLMs.
- Low power use and cost.
- We control what information we share with the LLM, and can therefore avoid sharing personal data.
- Our part of the AI search is open source.

Cons:

- The search is as good as the LLM, and we do not train the model.
- The LLM we use is not trained on all terms, i.e. dcat-ap-no is not found. This could be improved by adding different ranking functions, like BM25.
- Limitations on how much text we can include in the prompt limits the number of hits we can present to the user.
- Better and better models are emerging, and makes us feel like we are lagging behind.
- Vertex AI vector search is not open source.

What are the production costs?

- Vertex AI ~5-7 EUR/month
- Postgres database ~7 EUR/month
- Memory usage/CPU ~10 EUR/month

Total: ~25 EUR/month

- 11612 queries has been prosessed since we launched the AI search.
- ~2000 queries per month.

Impact on the environment

- With the use of on-demand AI models we avoid buying unnecessary hardware or idle capacity.
- By using Kafka events, the system only indexes changes.
- We consider the energy saved by finding and using a useful dataset as much larger than the energy that is used through the AI search.

Takeaways

- Don’t hesitate - Start using AI
- You do not necessarily have to train your own model
- Make sure you have control over security and privacy.
- After making an AI solution, keep on developing it
- AI/LLM solution will bridge the strict language of describing a dataset and the need for the consumer to describe what he or she is searching for

Useful links:

- https://data.norge.no/en
- https://data.norge.no/en/docs/finding-data/ai-search
- https://informasjonsforvaltning.github.io/data.norge.no/search/llm/
- https://github.com/Informasjonsforvaltning/fdk-llm-search-service

Contact us:

fellesdatakatalog@digdir.no