2 Methods For Improving Retrieval in RAG

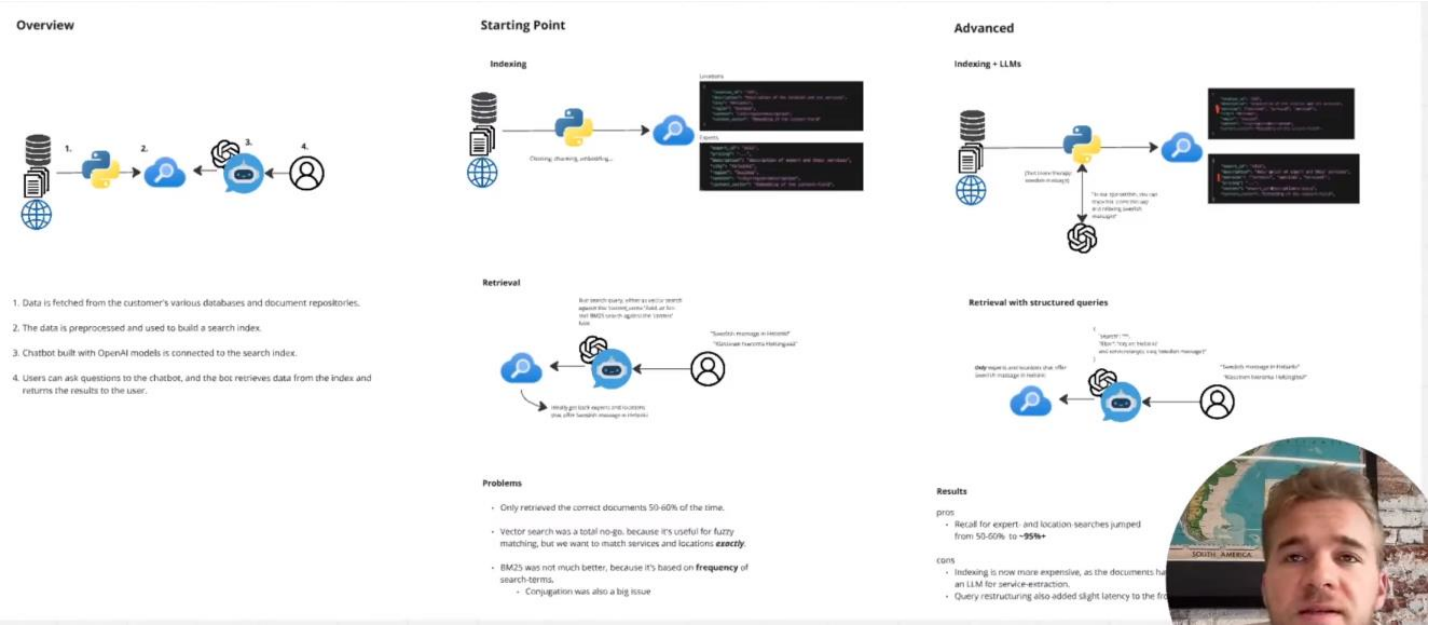
Johannes Jolkkonen | Funktio AI
7.52K subscribers

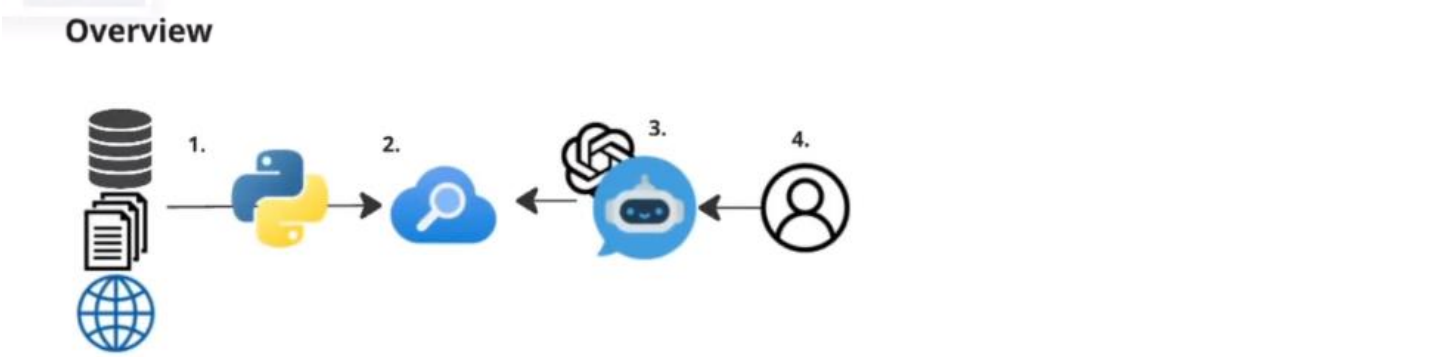
Subscribe

1.1K Share Ask

32,423 views Dec 19, 2024

Want to learn more about automating your business with AI?
https://cal.com/johannes-jolkkonen-xd...





1. Data is fetched from the customer's various databases and document repositories.
2. The data is preprocessed and used to build a search index.
3. Chatbot built with OpenAI models is connected to the search index.
4. Users can ask questions to the chatbot, and the bot retrieves data from the index and returns the results to the user.

Starting Point

Indexing

Locations

```
{
    "location_id": "123",
    "description": "Description of the location and its services",
    "city": "Helsinki",
    "region": "Uusimaa",
    "content": "city+region+description",
    "content_vector": "Embedding of the content-field"
}
```

Experts

```
    "expert_id": "e111",
    "pricing": "...",
    "description": "description of expert and their services",
    "city": "Helsinki",
    "region": "Uusimaa",
    "content": "city+region+description",
    "content_vector": "Embedding of the content-field",
```

Locations

```
{
    "location_id": "123",
    "description": "Description of the location and its services",
    "city": "Helsinki",
    "region": "Uusimaa",
    "content": "city+region+description",
    "content_vector": "Embedding of the content-field"
}
```

Experts

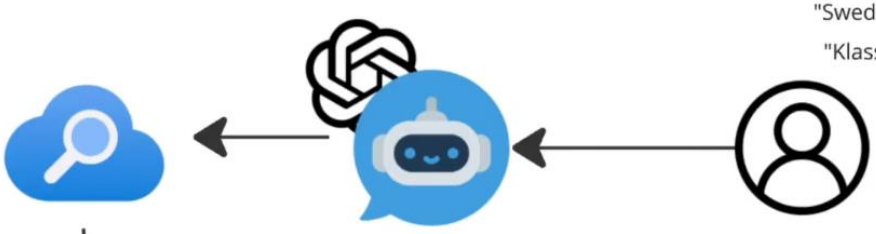
```
    "expert_id": "e111",
    "pricing": "...",
    "description": "description of expert and their services",
    "city": "Helsinki",
    "region": "Uusimaa",
    "content": "city+region+description",
    "content_vector": "Embedding of the content-field",
```

Retrieval

Run search query, either as vector search against the '*content_vector*' field, or full-text BM25 search against the '*content*' field.

"Swedish massage in Helsinki"

"Klassinen hieronta Helsingissä"

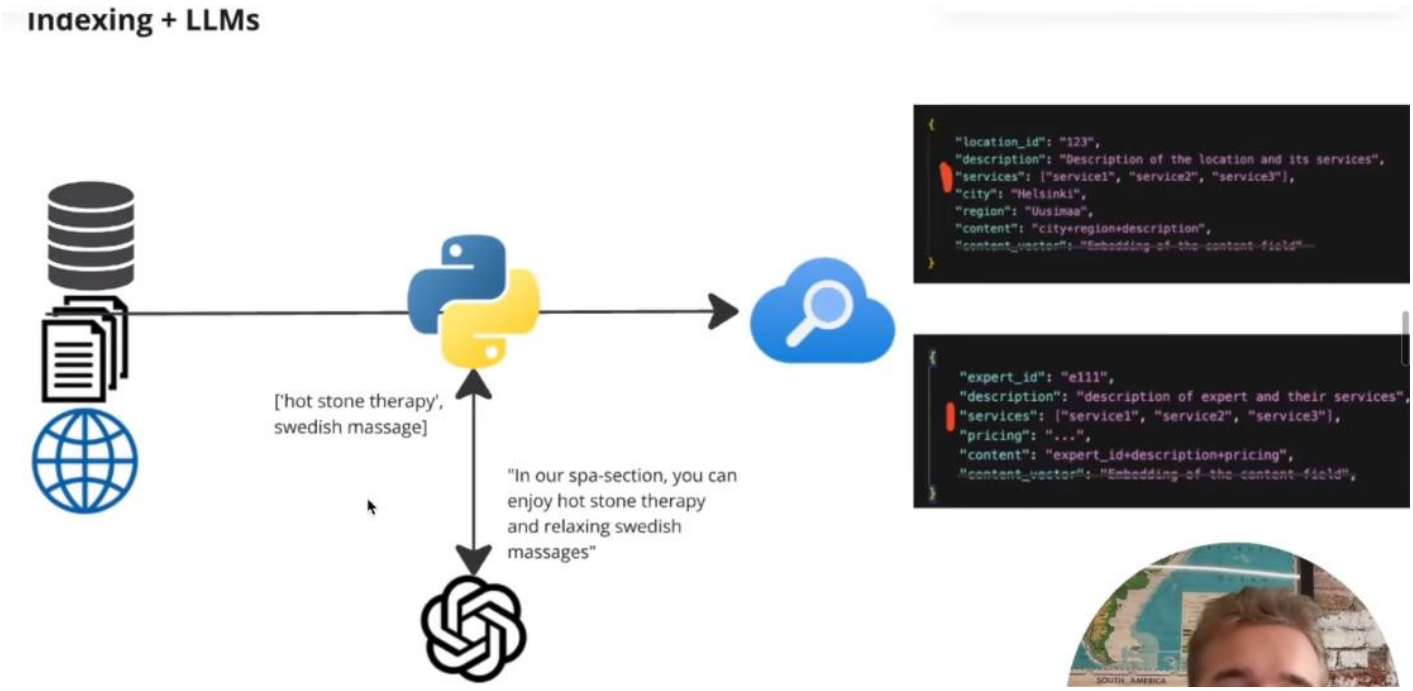


Ideally get back experts and locations that offer Swedish massage in Helsinki

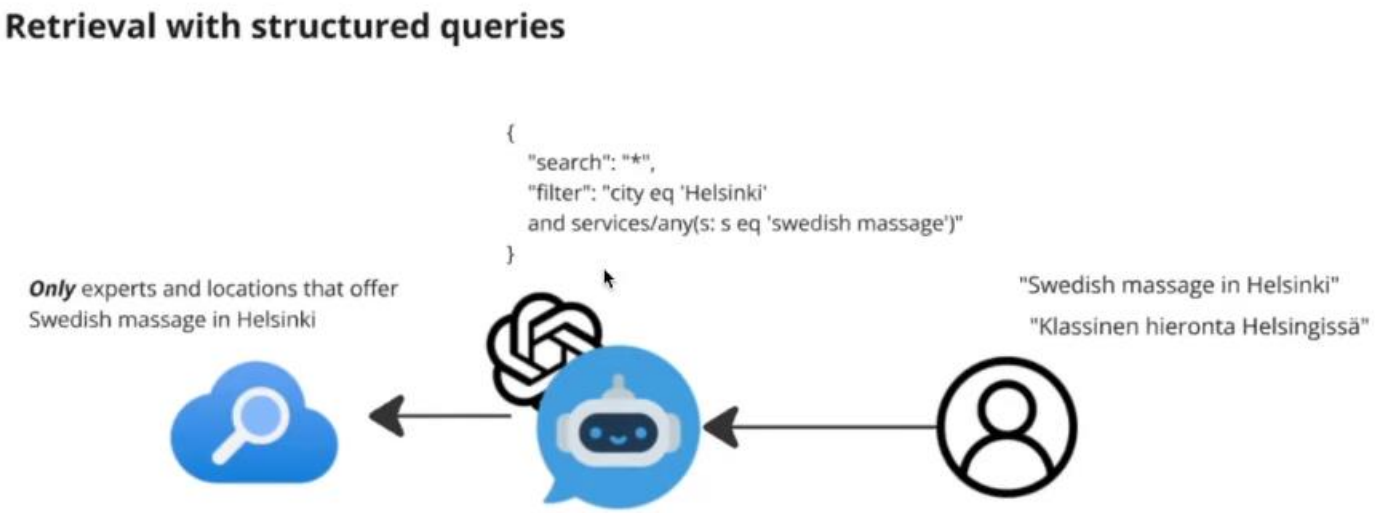
Problems

- Only retrieved the correct documents 50-60% of the time.
- Vector search was a total no-go, because it's useful for fuzzy matching, but we want to match services and locations ***exactly***.
- BM25 was not much better, because it's based on **frequency** of search-terms.
  - Conjugation was also a big issue

Next, let us see how we handled these problems and increased the accuracy of the RAG pipeline



We used an LLM to extract and add a services list node to the data



We now use an LLM to convert the raw query into a structured query to use that recognizes filters for cities and services.

Results

pros

- Recall for expert- and location-searches jumped from 50-60% to **~95%+**

cons

- Indexing is now more expensive, as the documents have to be run through an LLM for service-extraction.
- Query restructuring also added slight latency to the front-end.

Don't sleep on GAR

- Not only can retrieval support LLMs, but the reverse is also true