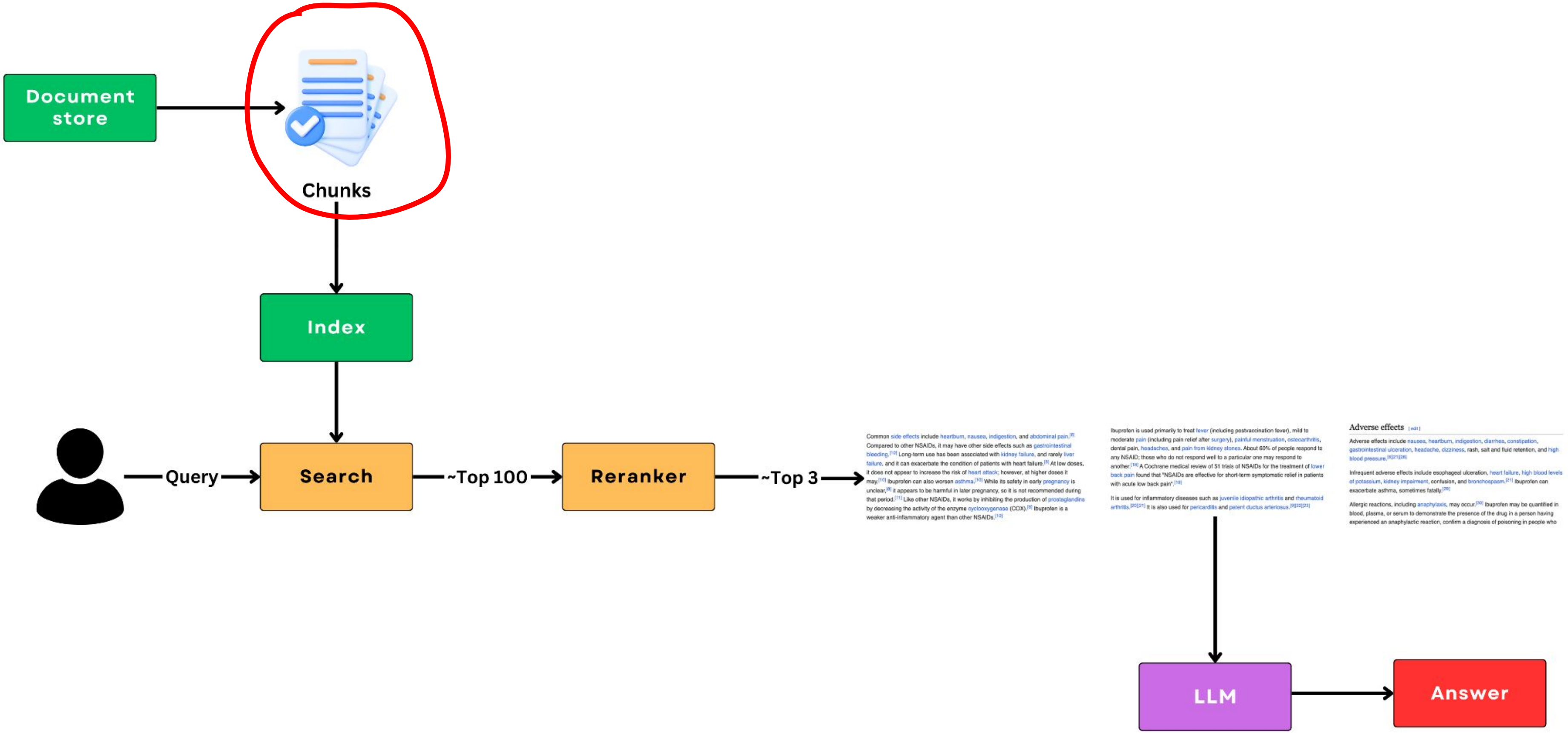
RAG



Common side effects include heartburn, nausea, indigestion, and abdominal pain.[8] Compared to other NSAIDs, it may have other side effects such as gastrointestinal bleeding.[10] Long-term use has been associated with kidney failure, and rarely liver failure, and it can exacerbate the condition of patients with heart failure.[8] At low doses, it does not appear to increase the risk of heart attack; however, at higher doses it may.[10] Ibuprofen can also worsen asthma.[10] While its safety in early pregnancy is unclear,[8] it appears to be harmful in later pregnancy, so it is not recommended during that period.[11] Like other NSAIDs, it works by inhibiting the production of prostaglandins by decreasing the activity of the enzyme cyclooxygenase (COX).[8] Ibuprofen is a weaker anti-inflammatory agent than other NSAIDs.[10]

Ibuprofen is used primarily to treat fever (including postvaccination fever), mild to moderate pain (including pain relief after surgery), painful menstruation, osteoarthritis, dental pain, headaches, and pain from kidney stones. About 60% of people respond to any NSAID; those who do not respond well to a particular one may respond to another.[18] A Cochrane medical review of 51 trials of NSAIDs for the treatment of lower back pain found that "NSAIDs are effective for short-term symptomatic relief in patients with acute low back pain".[19]

It is used for inflammatory diseases such as juvenile idiopathic arthritis and rheumatoid arthritis.[20][21] It is also used for pericarditis and patent ductus arteriosus.[9][22][23]

Adverse effects [edit]

Adverse effects include nausea, heartburn, indigestion, diarrhea, constipation, gastrointestinal ulceration, headache, dizziness, rash, salt and fluid retention, and high blood pressure.[8][21][28]

Infrequent adverse effects include esophageal ulceration, heart failure, high blood levels of potassium, kidney impairment, confusion, and bronchospasm.[21] Ibuprofen can exacerbate asthma, sometimes fatally.[29]

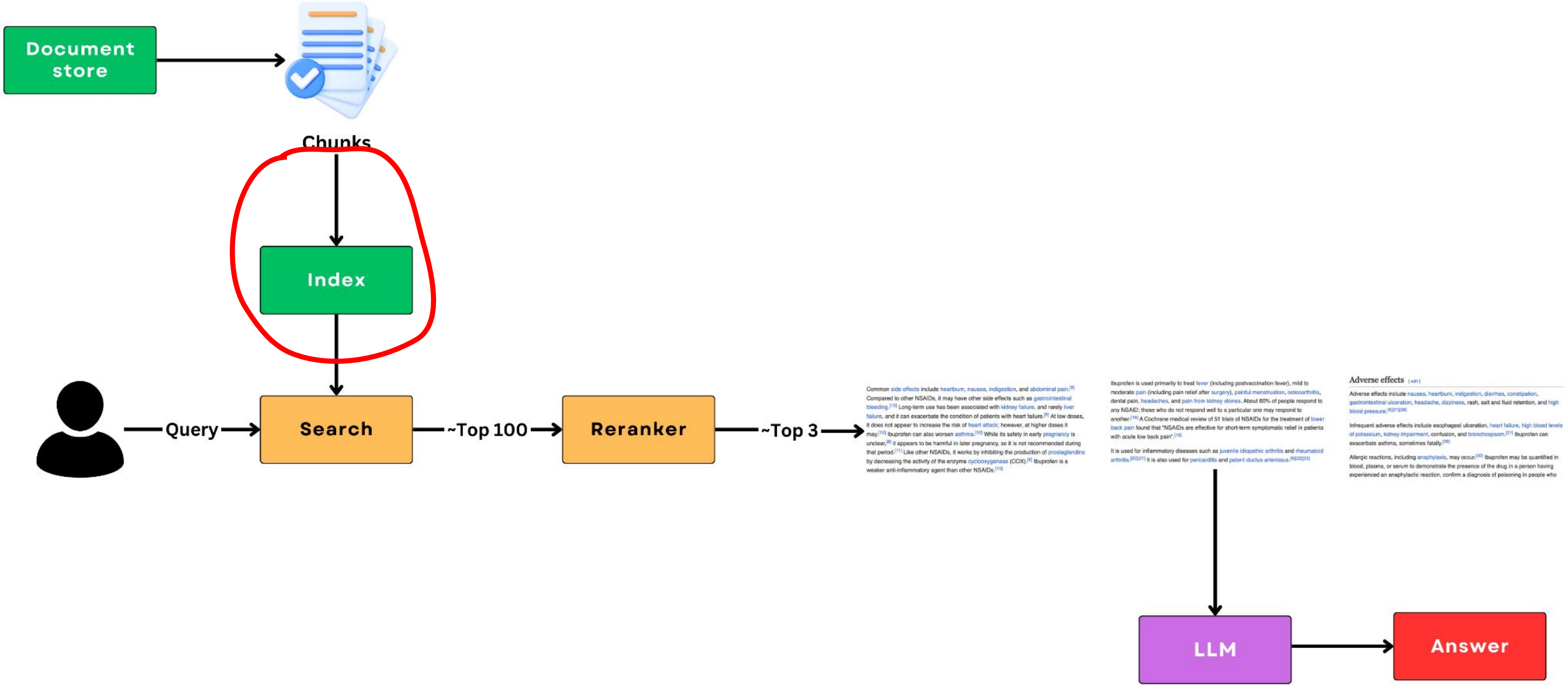
Allergic reactions, including anaphylaxis, may occur.[30] Ibuprofen may be quantified in blood, plasma, or serum to demonstrate the presence of the drug in a person having experienced an anaphylactic reaction, confirm a diagnosis of poisoning in people who

Chunking

- Usually the document store contains lots of long documents, e.g. wikipedia articles (we need to link the answer to these documents)
- But we need to split the documents into smaller pieces that we want to index
- Why?
    - ML models still can't process long documents well (Embeddings, Reranker, LLM, etc..)
    - Smaller size usually means better precision, but less context
    - Needs a good balance

Chunking

- What is the *correct* size? -> Hard to set, up for experimentation (most ML models still have around 512 token limit)
- How to split?
  - Based on structure (XML, JSON)
  - NLP methods (tokenizer, sentencizer, etc..)
  - Sliding window approach
  - Semantic chunking?
  - spaCy, NLTK, etc..



Common side effects include heartburn, nausea, indigestion, and abdominal pain.[8] Compared to other NSAIDs, it may have other side effects such as gastrointestinal bleeding.[10] Long-term use has been associated with kidney failure, and rarely liver failure, and it can exacerbate the condition of patients with heart failure.[8] At low doses, it does not appear to increase the risk of heart attack; however, at higher doses it may.[10] Ibuprofen can also worsen asthma.[10] While its safety in early pregnancy is unclear,[8] it appears to be harmful in later pregnancy, so it is not recommended during that period.[11] Like other NSAIDs, it works by inhibiting the production of prostaglandins by decreasing the activity of the enzyme cyclooxygenase (COX).[8] Ibuprofen is a weaker anti-inflammatory agent than other NSAIDs.[10]

Ibuprofen is used primarily to treat fever (including postvaccination fever), mild to moderate pain (including pain relief after surgery), painful menstruation, osteoarthritis, dental pain, headaches, and pain from kidney stones. About 60% of people respond to any NSAID; those who do not respond well to a particular one may respond to another.[18] A Cochrane medical review of 51 trials of NSAIDs for the treatment of lower back pain found that "NSAIDs are effective for short-term symptomatic relief in patients with acute low back pain".[19]

It is used for inflammatory diseases such as juvenile idiopathic arthritis and rheumatoid arthritis.[20][21] It is also used for pericarditis and patent ductus arteriosus.[9][22][23]

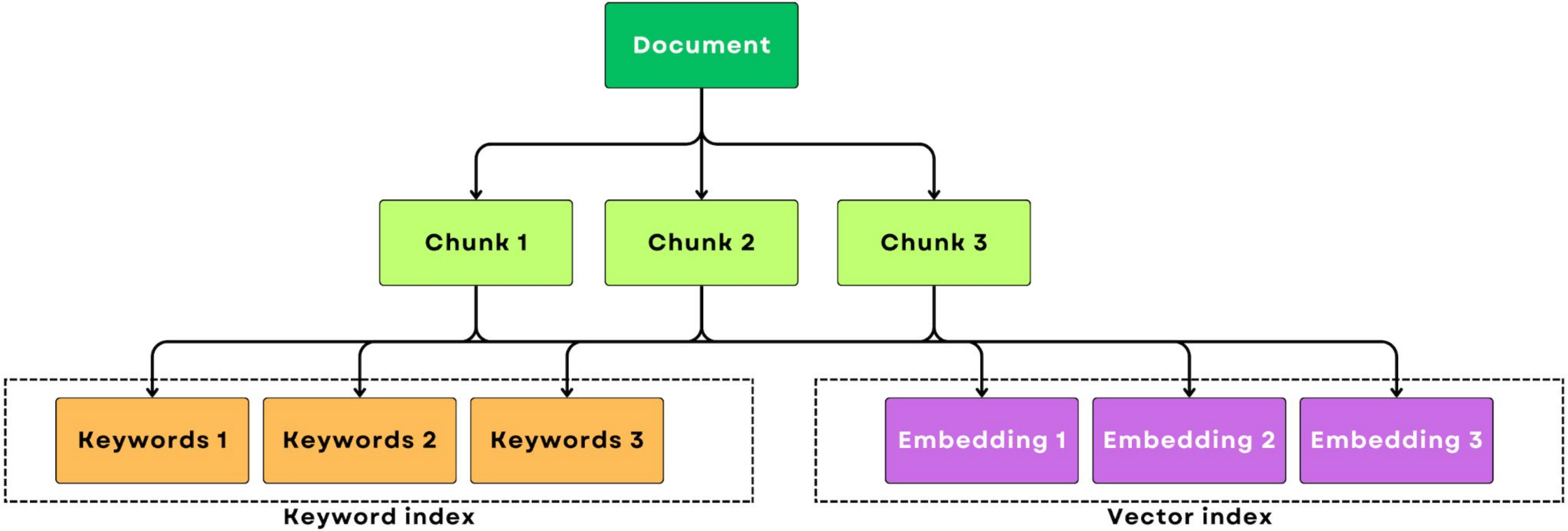
Adverse effects [edit]

Adverse effects include nausea, heartburn, indigestion, diarrhea, constipation, gastrointestinal ulceration, headache, dizziness, rash, salt and fluid retention, and high blood pressure.[8][21][28]

Infrequent adverse effects include esophageal ulceration, heart failure, high blood levels of potassium, kidney impairment, confusion, and bronchospasm.[21] Ibuprofen can exacerbate asthma, sometimes fatally.[29]

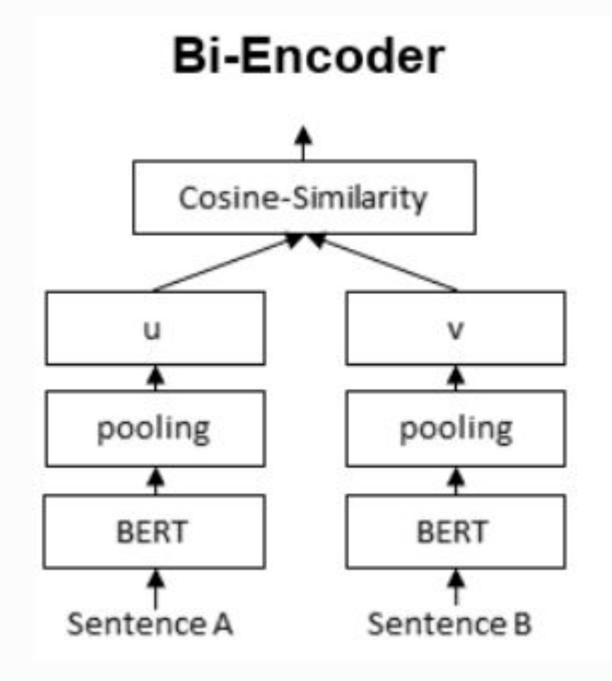
Allergic reactions, including anaphylaxis, may occur.[30] Ibuprofen may be quantified in blood, plasma, or serum to demonstrate the presence of the drug in a person having experienced an anaphylactic reaction, confirm a diagnosis of poisoning in people who

Indexing


Document
Chunk 1
Chunk 2
Chunk 3
Keywords 1
Keywords 2
Keywords 3
Embedding 1
Embedding 2
Embedding 3
Keyword index
Vector index

Indexing - Vectors

- Turn chunks into embeddings
  - Use sentence-embeddings (https://www.sbert.net/index.html)
  - Sentence-embeddings are trained as Bi-Encoders with a Cosine Similarity Loss
  - Sentence-embeddings for Information Retrieval are trained with query-passage pairs with positive and negative examples
  - Queries to relevant passages are “more” similar than to not-relevant passages



source

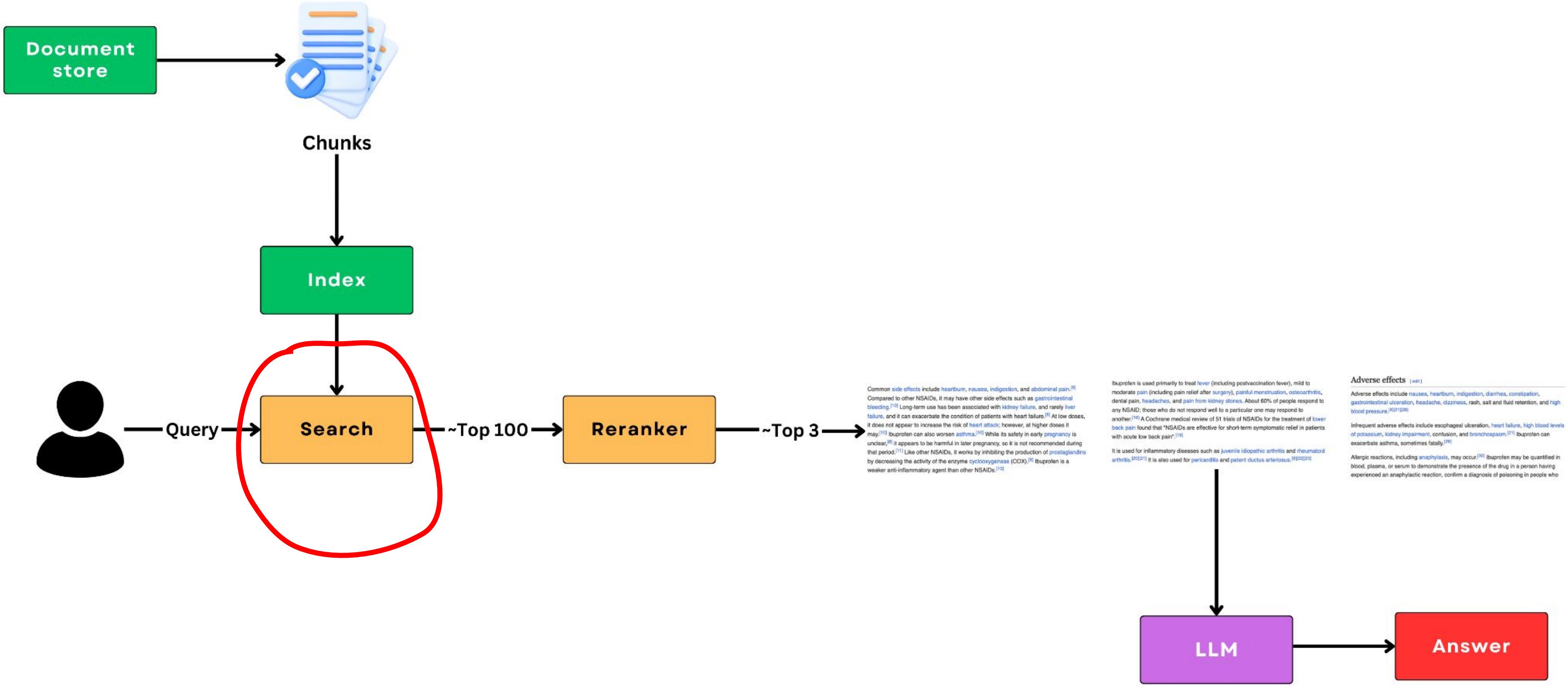
Indexing - Vectors

- The embeddings of the chunks are stored in a vector index
- In memory index vs. vector database
- The functionality we need: fast KNN with vectors
- Voyager (https://github.com/spotify/voyager) in memory vector store
- FAISS (https://github.com/facebookresearch/faiss) in memory vector store
- Weaviate (https://weaviate.io/) -> fully fledged vector database
- Elasticsearch (https://www.elastic.co/) -> also has vector database capability

Vector search needs to compute KNN efficiently, comparing to every vector is costly -> ANN (approximate nearest neighbors) computes similarity to clusters instead

Indexing - Full text search

- Transform chunks into list of tokens (keywords)
- Store the keywords in an inverted index
- Text -> Keywords (tokenization, lemmatization, filtering stopwords, synonyms, etc..)
- Elasticsearch
  - Go to solution for full text search
  - Rich feature set with language analyzers, synonym lists, etc..
  - Scaling
  - Fuzzy search
- Tantivy (https://github.com/quickwit-oss/tantivy)
  - Fully open source
  - Works locally without setting up a server



Common side effects include heartburn, nausea, indigestion, and abdominal pain.[8] Compared to other NSAIDs, it may have other side effects such as gastrointestinal bleeding.[10] Long-term use has been associated with kidney failure, and rarely liver failure, and it can exacerbate the condition of patients with heart failure.[8] At low doses, it does not appear to increase the risk of heart attack; however, at higher doses it may.[10] Ibuprofen can also worsen asthma.[10] While its safety in early pregnancy is unclear,[8] it appears to be harmful in later pregnancy, so it is not recommended during that period.[11] Like other NSAIDs, it works by inhibiting the production of prostaglandins by decreasing the activity of the enzyme cyclooxygenase (COX).[8] Ibuprofen is a weaker anti-inflammatory agent than other NSAIDs.[10]

Ibuprofen is used primarily to treat fever (including postvaccination fever), mild to moderate pain (including pain relief after surgery), painful menstruation, osteoarthritis, dental pain, headaches, and pain from kidney stones. About 60% of people respond to any NSAID; those who do not respond well to a particular one may respond to another.[18] A Cochrane medical review of 51 trials of NSAIDs for the treatment of lower back pain found that "NSAIDs are effective for short-term symptomatic relief in patients with acute low back pain".[19]

It is used for inflammatory diseases such as juvenile idiopathic arthritis and rheumatoid arthritis.[20][21] It is also used for pericarditis and patent ductus arteriosus.[9][22][23]

Adverse effects [edit]

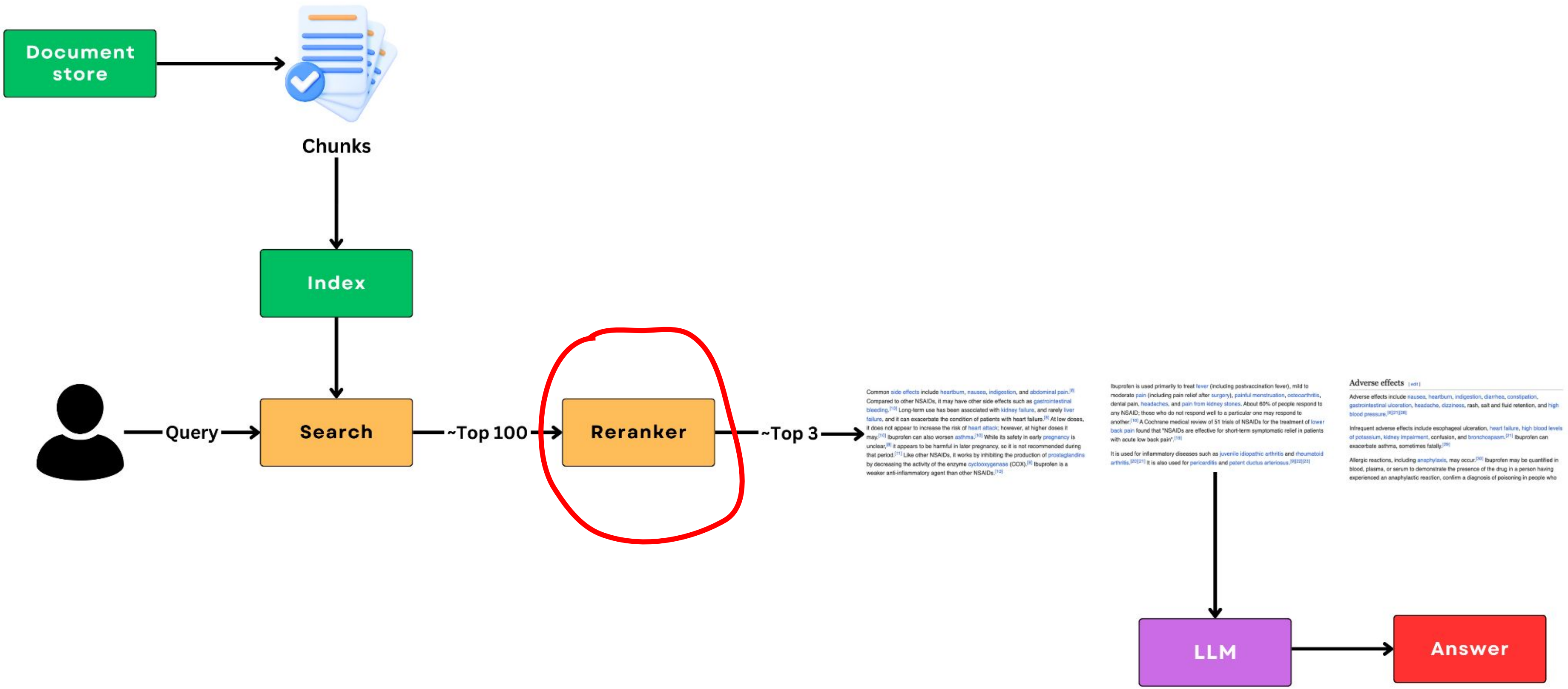
Adverse effects include nausea, heartburn, indigestion, diarrhea, constipation, gastrointestinal ulceration, headache, dizziness, rash, salt and fluid retention, and high blood pressure.[8][21][28]

Infrequent adverse effects include esophageal ulceration, heart failure, high blood levels of potassium, kidney impairment, confusion, and bronchospasm.[21] Ibuprofen can exacerbate asthma, sometimes fatally.[29]

Allergic reactions, including anaphylaxis, may occur.[30] Ibuprofen may be quantified in blood, plasma, or serum to demonstrate the presence of the drug in a person having experienced an anaphylactic reaction, confirm a diagnosis of poisoning in people who

First stage search

- The goal is to pick around ~100 relevant candidates for the user query
- Hybrid model, use VectorDB and KeywordDB
- Question preprocessing
  - Query expansion, resolve abbreviations, etc..
- Needs to be a fast method that can pick relevant candidates from millions of documents -> cannot be deep learning model that classifies each pair
- Vector DB -> compute cosine similarity pairwise and rank them based on the distance
- Keyword DB -> compute BM25 and pick the top ones
- Merge the results -> Linear combination, Rank Fusion



Common side effects include heartburn, nausea, indigestion, and abdominal pain.[8] Compared to other NSAIDs, it may have other side effects such as gastrointestinal bleeding.[10] Long-term use has been associated with kidney failure, and rarely liver failure, and it can exacerbate the condition of patients with heart failure.[8] At low doses, it does not appear to increase the risk of heart attack; however, at higher doses it may.[10] Ibuprofen can also worsen asthma.[10] While its safety in early pregnancy is unclear,[8] it appears to be harmful in later pregnancy, so it is not recommended during that period.[11] Like other NSAIDs, it works by inhibiting the production of prostaglandins by decreasing the activity of the enzyme cyclooxygenase (COX).[8] Ibuprofen is a weaker anti-inflammatory agent than other NSAIDs.[10]

Ibuprofen is used primarily to treat fever (including postvaccination fever), mild to moderate pain (including pain relief after surgery), painful menstruation, osteoarthritis, dental pain, headaches, and pain from kidney stones. About 60% of people respond to any NSAID; those who do not respond well to a particular one may respond to another.[18] A Cochrane medical review of 51 trials of NSAIDs for the treatment of lower back pain found that "NSAIDs are effective for short-term symptomatic relief in patients with acute low back pain".[19]

It is used for inflammatory diseases such as juvenile idiopathic arthritis and rheumatoid arthritis.[20][21] It is also used for pericarditis and patent ductus arteriosus.[9][22][23]

Adverse effects [edit]

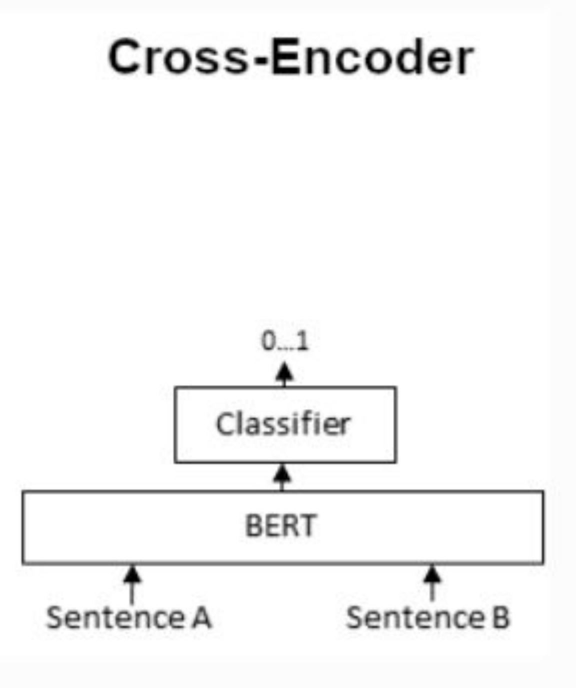
Adverse effects include nausea, heartburn, indigestion, diarrhea, constipation, gastrointestinal ulceration, headache, dizziness, rash, salt and fluid retention, and high blood pressure.[8][21][28]

Infrequent adverse effects include esophageal ulceration, heart failure, high blood levels of potassium, kidney impairment, confusion, and bronchospasm.[21] Ibuprofen can exacerbate asthma, sometimes fatally.[29]

Allergic reactions, including anaphylaxis, may occur.[30] Ibuprofen may be quantified in blood, plasma, or serum to demonstrate the presence of the drug in a person having experienced an anaphylactic reaction, confirm a diagnosis of poisoning in people who

Reranking - Second stage search

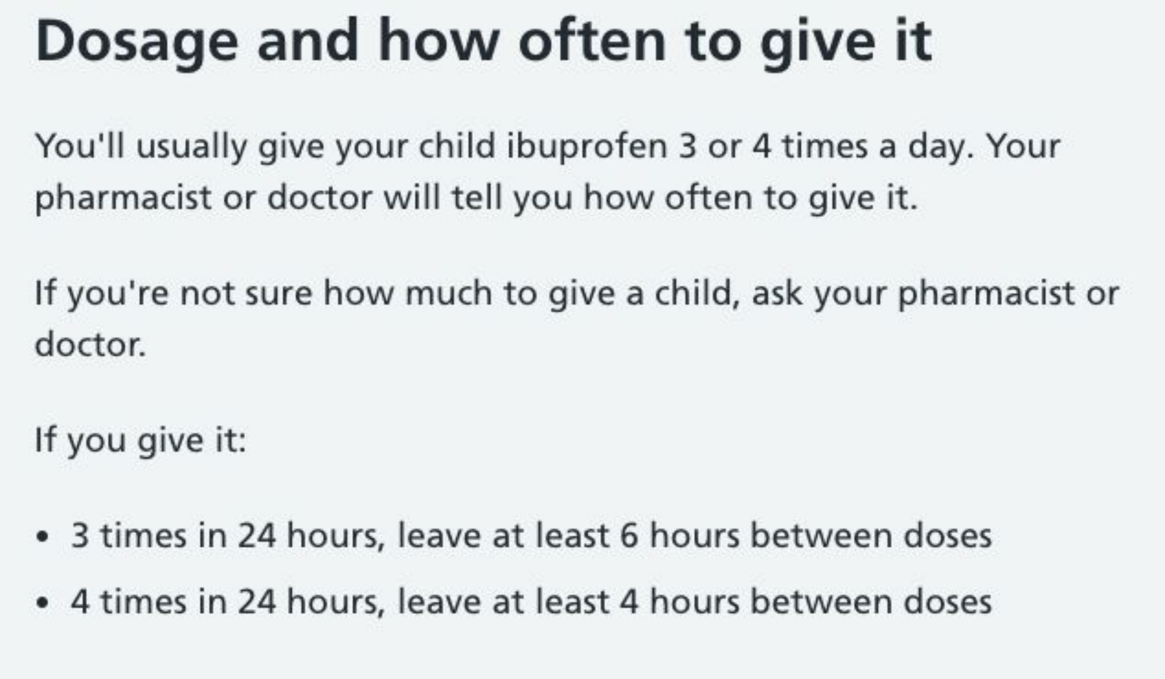
- The goal of the reranker is to go through the list of candidates picked by the first stage search and only pick the ones that answers the query
- Ideally, the output of this stage should be the final top 3-5 chunks
- The task is usually solved with CrossEncoders (Cross-Encoders — Sentence Transformers documentation)



Reranking

- There are big open source datasets of query-passage pairs, e.g. MSMARCO (https://microsoft.github.io/msmarco/)
- Pick a good model trained on these datasets, e.g. https://huggingface.co/cross-encoder/msmarco-MiniLM-L6-en-de-v1
- Collect training data on your field
- Query-Passage-Relevancy triplets
- Finetune on your domain as a binary classification problem
    - (there are also other way to model this task)

Query: ibuprofen dosage for kids



Dosage and how often to give it

You'll usually give your child ibuprofen 3 or 4 times a day. Your pharmacist or doctor will tell you how often to give it.

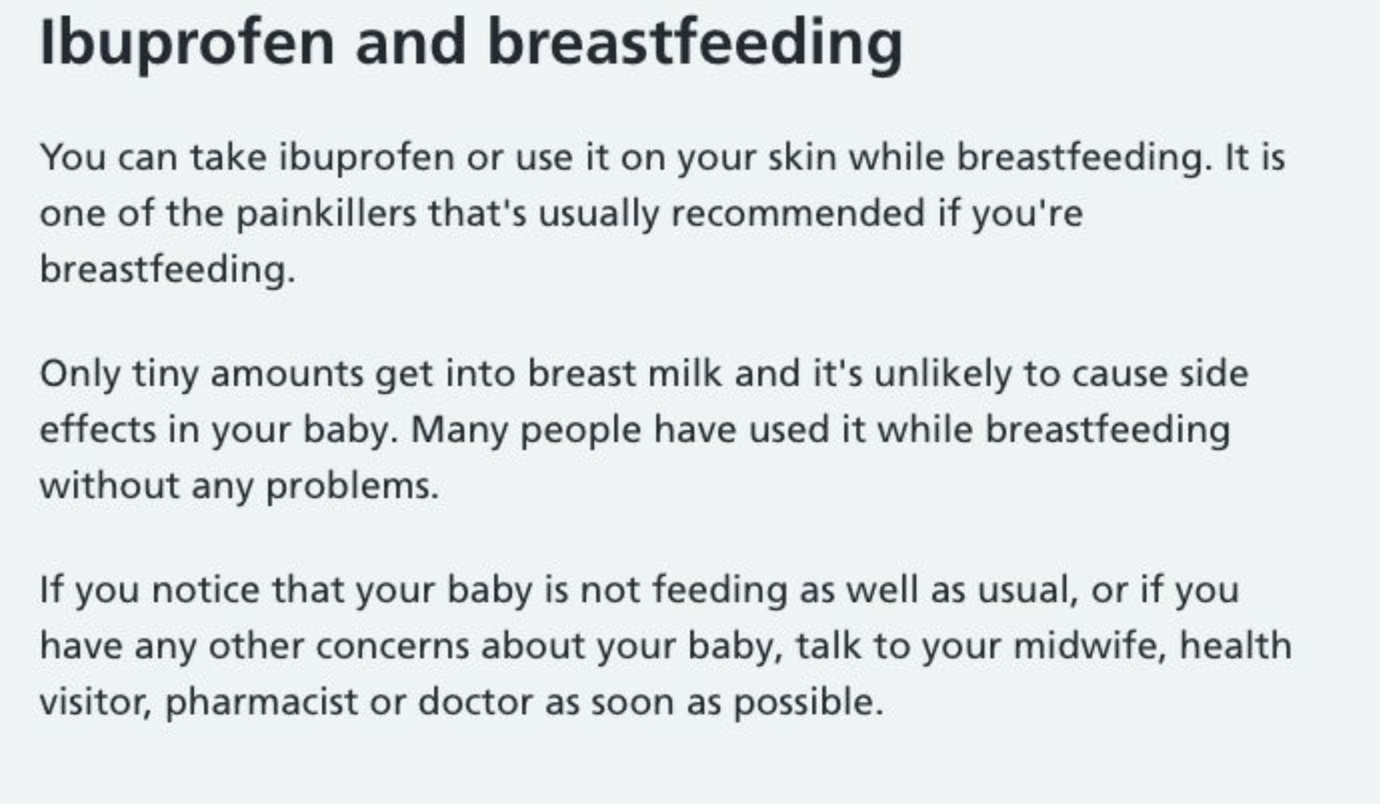
If you're not sure how much to give a child, ask your pharmacist or doctor.

If you give it:

- 3 times in 24 hours, leave at least 6 hours between doses
- 4 times in 24 hours, leave at least 4 hours between doses

✓

Query: ibuprofen dosage for kids



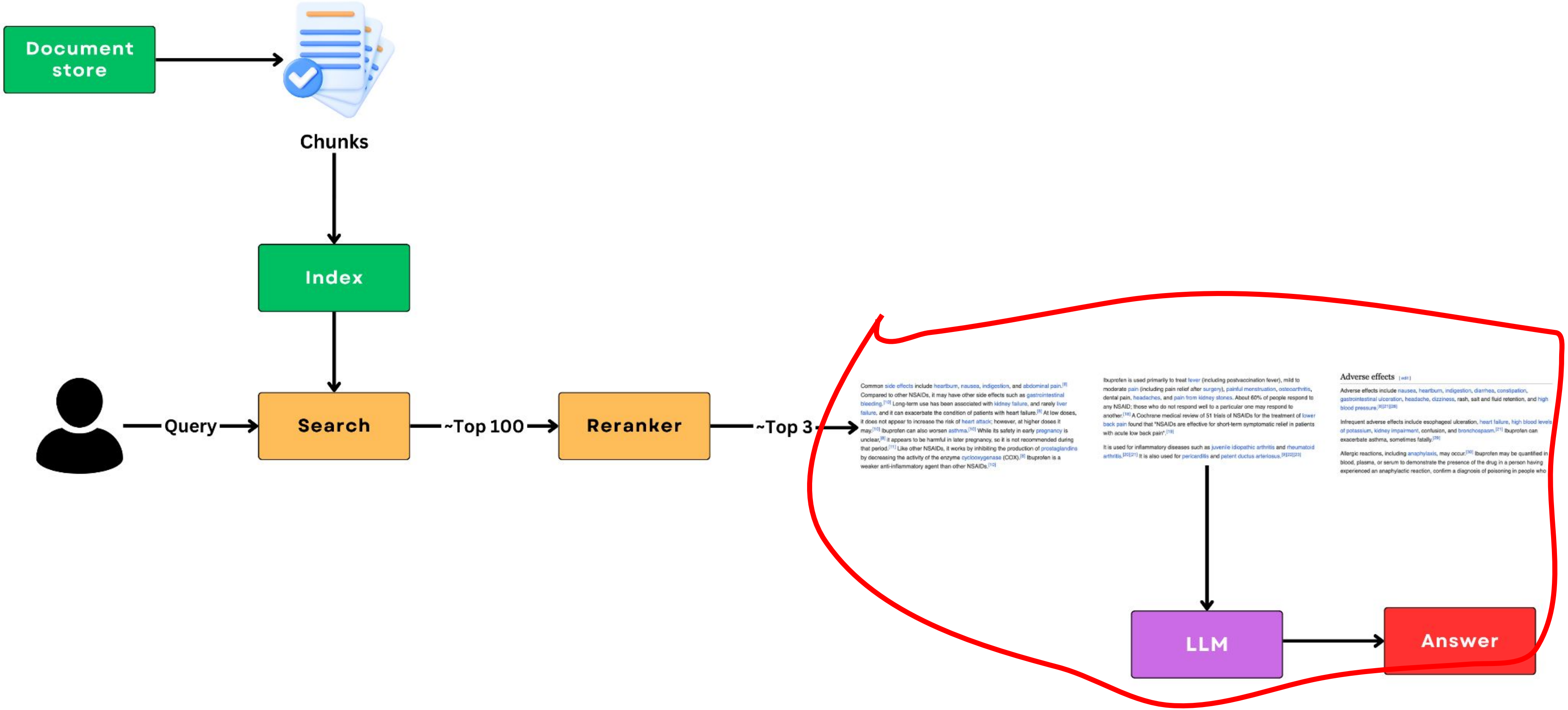
Ibuprofen and breastfeeding

You can take ibuprofen or use it on your skin while breastfeeding. It is one of the painkillers that's usually recommended if you're breastfeeding.

Only tiny amounts get into breast milk and it's unlikely to cause side effects in your baby. Many people have used it while breastfeeding without any problems.

If you notice that your baby is not feeding as well as usual, or if you have any other concerns about your baby, talk to your midwife, health visitor, pharmacist or doctor as soon as possible.

✗



Common side effects include heartburn, nausea, indigestion, and abdominal pain.[8] Compared to other NSAIDs, it may have other side effects such as gastrointestinal bleeding.[10] Long-term use has been associated with kidney failure, and rarely liver failure, and it can exacerbate the condition of patients with heart failure.[8] At low doses, it does not appear to increase the risk of heart attack; however, at higher doses it may.[10] Ibuprofen can also worsen asthma.[10] While its safety in early pregnancy is unclear,[8] it appears to be harmful in later pregnancy, so it is not recommended during that period.[11] Like other NSAIDs, it works by inhibiting the production of prostaglandins by decreasing the activity of the enzyme cyclooxygenase (COX).[8] Ibuprofen is a weaker anti-inflammatory agent than other NSAIDs.[10]

Ibuprofen is used primarily to treat fever (including postvaccination fever), mild to moderate pain (including pain relief after surgery), painful menstruation, osteoarthritis, dental pain, headaches, and pain from kidney stones. About 60% of people respond to any NSAID; those who do not respond well to a particular one may respond to another.[18] A Cochrane medical review of 51 trials of NSAIDs for the treatment of lower back pain found that "NSAIDs are effective for short-term symptomatic relief in patients with acute low back pain".[19]

It is used for inflammatory diseases such as juvenile idiopathic arthritis and rheumatoid arthritis.[20][21] It is also used for pericarditis and patent ductus arteriosus.[9][22][23]

Adverse effects [edit]

Adverse effects include nausea, heartburn, indigestion, diarrhea, constipation, gastrointestinal ulceration, headache, dizziness, rash, salt and fluid retention, and high blood pressure.[8][21][28]

Infrequent adverse effects include esophageal ulceration, heart failure, high blood levels of potassium, kidney impairment, confusion, and bronchospasm.[21] Ibuprofen can exacerbate asthma, sometimes fatally.[29]

Allergic reactions, including anaphylaxis, may occur.[30] Ibuprofen may be quantified in blood, plasma, or serum to demonstrate the presence of the drug in a person having experienced an anaphylactic reaction, confirm a diagnosis of poisoning in people who

The final step: LLM

- The last step is to take the output of the reranker (~top 3 chunks), form it as a *context* for the LLM and prompt it to answer based on the documents
- Pros:
  - We can make the LLM to answer based on our own data without actually training its internal knowledge
  - We can fact-check the given answers if the context serves as something that contains the gold answer for the query
  - We can reduce hallucinations
-

Forming the prompt

Prompt:
Answer questions based on medical documents.

Question: What are common symptoms of a peanut allergy?

First Document:
Peanut Allergy
Symptoms
Peanut allergies can cause mild to severe reactions. Common symptoms include itching, hives, swelling, and difficulty breathing. In severe cases, anaphylaxis can occur, which requires immediate medical attention.

Second Document:
Allergies and Immune Response
When someone with a peanut allergy consumes peanuts, the body's immune system reacts abnormally by releasing chemicals like histamine. This can result in symptoms such as skin rashes, gastrointestinal issues (nausea, vomiting), or respiratory distress.

Your answer:

Common …

Forming the prompt

Prompt:
Answer questions based on medical documents.

Question: What are common symptoms of a peanut allergy?

First Document:
Peanut Allergy
Symptoms
Peanut allergies can cause mild to severe reactions. Common symptoms include itching, hives, swelling, and difficulty breathing. In severe cases, anaphylaxis can occur, which requires immediate medical attention.

Second Document:
Allergies and Immune Response
When someone with a peanut allergy consumes peanuts, the body’s immune system reacts abnormally by releasing chemicals like histamine. This can result in symptoms such as skin rashes, gastrointestinal issues (nausea, vomiting), or respiratory distress.

Your answer:

Common symptoms…

Forming the prompt

Prompt:
Answer questions based on medical documents.

Question: What are common symptoms of a peanut allergy?

First Document:
Peanut Allergy
Symptoms
Peanut allergies can cause mild to severe reactions. Common symptoms include itching, hives, swelling, and difficulty breathing. In severe cases, anaphylaxis can occur, which requires immediate medical attention.

Second Document:
Allergies and Immune Response
When someone with a peanut allergy consumes peanuts, the body’s immune system reacts abnormally by releasing chemicals like histamine. This can result in symptoms such as skin rashes, gastrointestinal issues (nausea, vomiting), or respiratory distress.

Your answer:

Common symptoms include …

Forming the prompt

Prompt:
Answer questions based on medical documents.

Question: What are common symptoms of a peanut allergy?

First Document:
Peanut Allergy
Symptoms
Peanut allergies can cause mild to severe reactions. Common symptoms include itching, hives, swelling, and difficulty breathing. In severe cases, anaphylaxis can occur, which requires immediate medical attention.

Second Document:
Allergies and Immune Response
When someone with a peanut allergy consumes peanuts, the body's immune system reacts abnormally by releasing chemicals like histamine. This can result in symptoms such as skin rashes, gastrointestinal issues (nausea, vomiting), or respiratory distress.

Your answer:

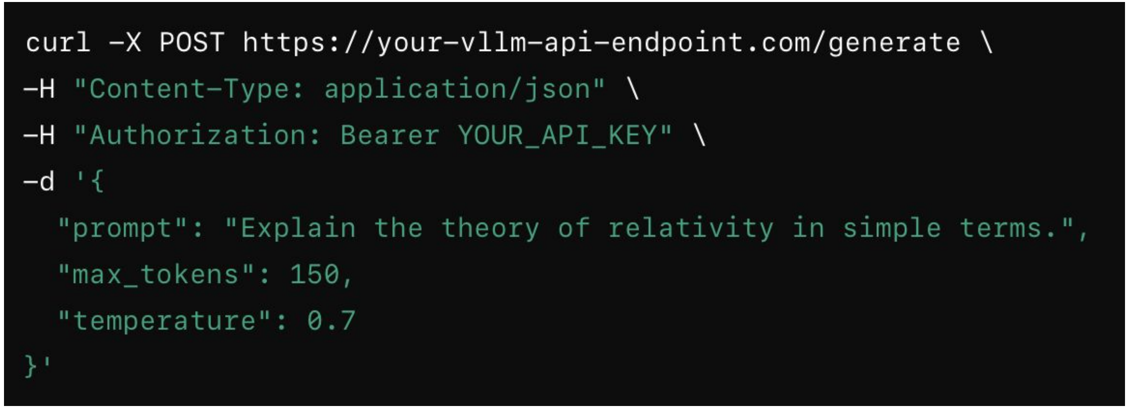
Common symptoms include itching…

How?

- LLMs are just ML models that take sequence of words as an input and generate the next one in an autoregressive manner
- To handle efficient requests, usually there is a need for an optimized LLM server architecture
    - Something that handles concurrent requests, does efficient KV caching with an optimized attention mechanism like FlashAttention
    - This can be an API, like OpenAI
    - Or you can run your own LLM servers
- Suggestion, use vllm GitHub - vllm-project/vllm: A high-throughput and memory-efficient inference and serving engine for LLMs as your local server
- VLLM runs a REST API with the same interface as OpenAI

How?

- Then, you need a client to call the REST API, this you can construct yourself, or use the ***openai*** client (https://pypi.org/project/openai/)



```
curl -X POST https://your-vllm-api-endpoint.com/generate \
-H "Content-Type: application/json" \
-H "Authorization: Bearer YOUR_API_KEY" \
-d '{
  "prompt": "Explain the theory of relativity in simple terms.",
  "max_tokens": 150,
  "temperature": 0.7
}'
```

How?

- There are good libraries to orchestrate LLM apps, e.g.
    - langchain (https://www.langchain.com/)
    - llamaindex (https://www.llamaindex.ai/)
    - haystack (https://haystack.deepset.ai/)
- These libraries let you build quick LLM solutions, but hide lots of details from you
- The suggestion is to try to construct your own prompts and only use LLM REST APIs to handle the rest, this way you will have control over your solution
- If you don’t have the resources to run your own LLMs, try https://www.together.ai/ (for a free 20$ credit)

Chat templates

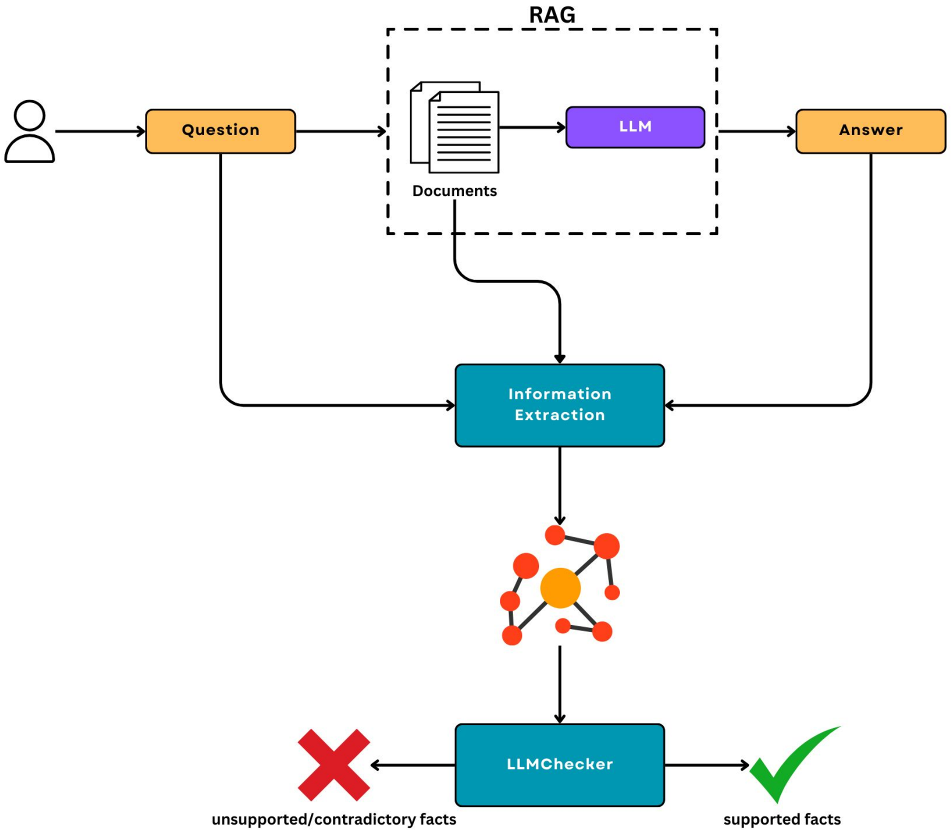
- The chat like manner of LLMs are achieved by training them with *chat templates*, these templates are also applied in the background when you are using an off-the-shelf solution like ChatGPT

A chat template looks like this:

```
<|system|>

You are a friendly chatbot who always responds in the style of a pirate</s>

<|user|>

How many helicopters can a human eat in one sitting?</s>

<|assistant|
```

A good overview of chat templates: https://huggingface.co/docs/transformers/main/en/chat_templating

LLMChecker


RAG
Question
Documents
LLM
Answer
Information Extraction
LLMChecker
unsupported/contradictory facts
supported facts