



AI-Based Model for Knowledge Specific Assistance

Thanos Apostolou

University of Piraeus
Department of Informatics

2024

Knowledge Specific Assistant

Define the problem domain.

- Documents in different formats about a specific knowledge field.
- Investigation of Knowledge Assistance Approaches
- Specific Knowledge Assistant (SKA) Application

Specific Knowledge Assistance Approaches

Two different Methods

1. Custom Text Generation Model Method

Create custom text generation model from scratch

2. Retrieval Augmented Generation (RAG) Method

Use pre-trained LLMs together with RAG technique

1. Custom Text Generation Model Method

- Create custom text generation model based on LSTM
- Train this model with users' Documents
- Invoke model by predicting the next word each time

Advantages

- Independence
- Flexibility

Disadvantages

- Needs many sources
- Highly consuming
- Hard to implement
- Hard to return sources

2. Retrieval Augmented Generation (RAG) Method

- Split Documents in chunks and save in vector store.
- Use pre-trained LLMs
- Invoke LLM and instruct to answer only based on context from vector store.

Advantages

- Needs only relative sources
- Adaptable
- Easy to implement
- Able to return sources
- Prompting

Disadvantages

- Dependency on external LLMs:
- Inflexibility

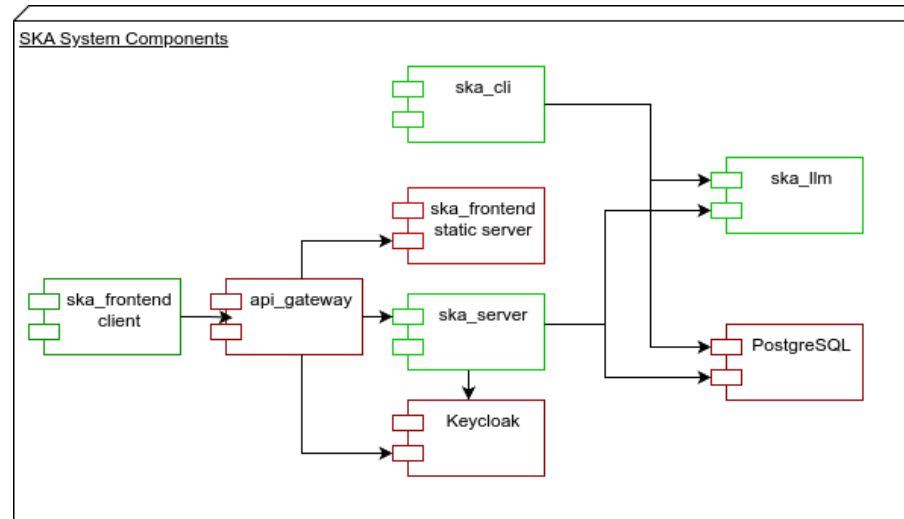
System Architecture

Our Components

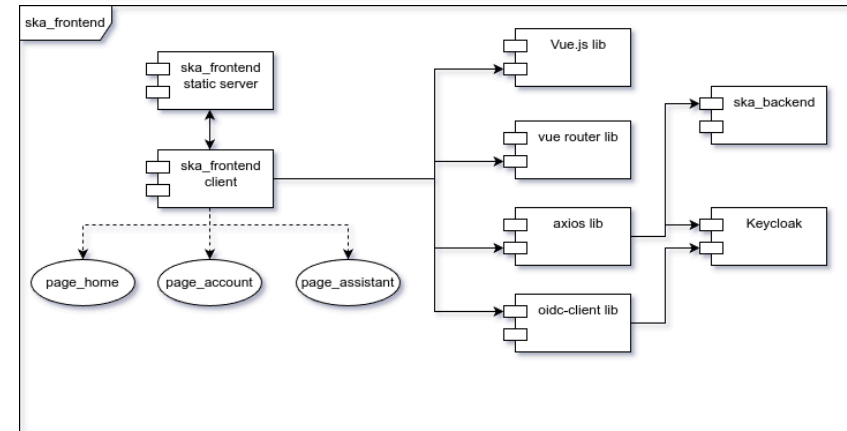
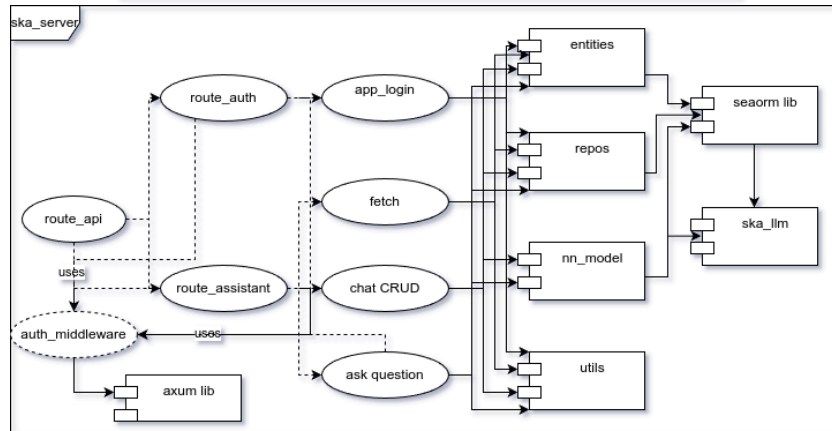
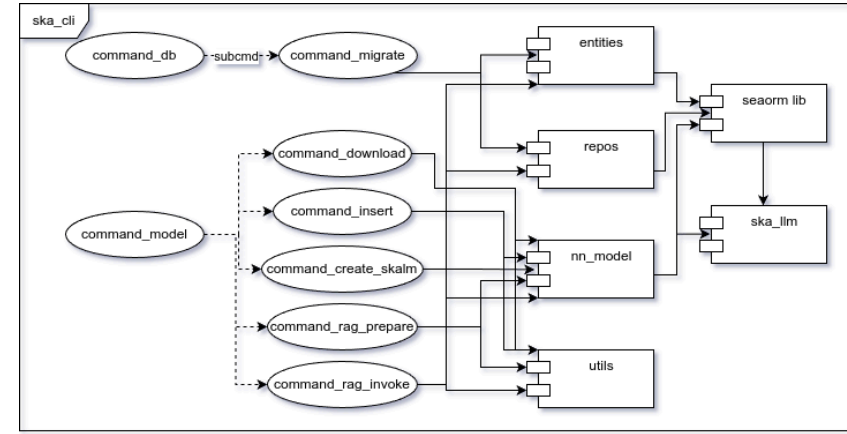
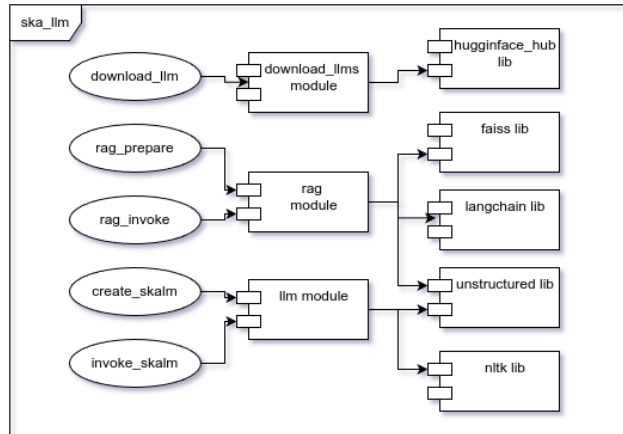
- ska_llm
- ska_cli
- ska_server
- ska_frontend

Extra Components

- PostgreSQL
- Keycloak
- api_gateway

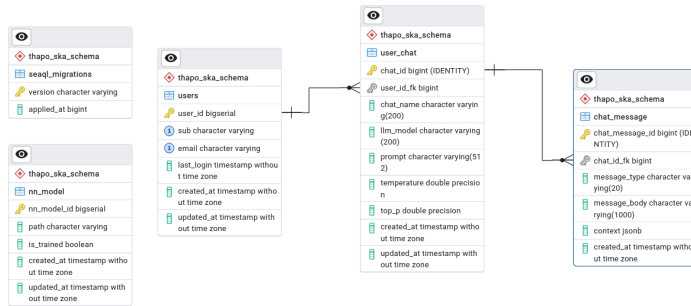


Our Components



Extra Components

PostgreSQL

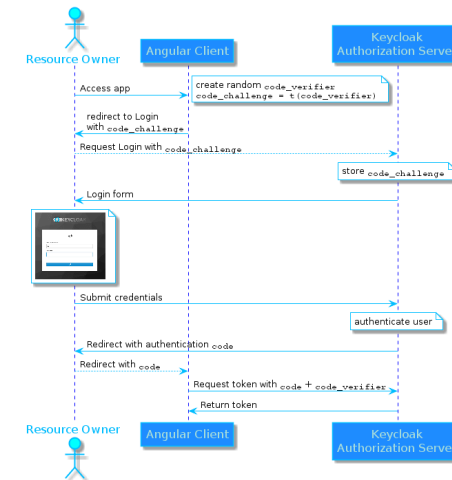


API Gateway

- “/app” to ska_frontend
- “/backend” to ska_server
- “/iam” to Keycloak

Keycloak Roles

- SKA_ADMIN
- SKA_USER
- SKA_GUEST



SKA Application

CLI

Usage: `app-cli <COMMAND>`

Commands:

`model`

`db`

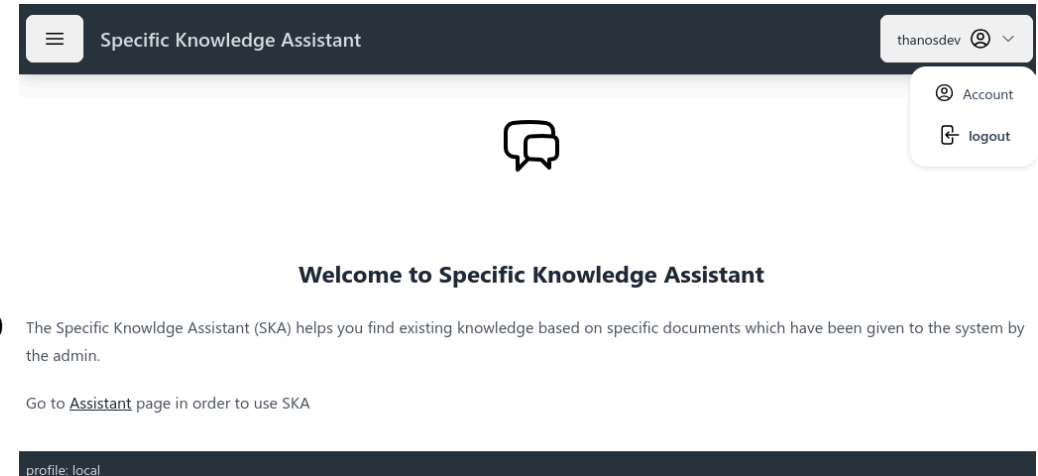
`help` Print this message or the help
of the given subcommand(s)

Options:

`-h, --help` Print help

`-V, --version` Print version

GUI



Usage skalm

Create skalm chat

Add Chat
Choose a LLM:

skalm

Chat Name

Edit Parameters?

Temperature

Top-P

Prompt

our_skalm

0

0

Prompt

Cancel

Add

Ask skalm question

+ Add Chat

our_skalm (skalm)

Please ask me anything related to this field

what is domain driven design?

a pattern is that it is not a solution to the model . the team can readily distinguish two models . the developer was able to respond to the model and the design . the team had encountered in the model , the team had given a new concept . the model is a set of concepts that can be integrated by the domain experts , the bones of the model is the same concept . the team may not be a very deep model for the domain .

what is domain driven design?

Usage Llama3

Create Llama3 chat

Add Chat
Choose a LLM:

llama3-8B ▾

Chat Name

our_llama3

Edit Parameters? ☒
Temperature

0 ▾

Top-P

0 ▾

Prompt

<|begin_of_text|> <|start_header_id|>system<|end_header_id|>
You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.<|eot_id|> <|start_header_id|>user<|end_header_id|>
Context: {context}. Question: {question}<|eot_id|> <|start_header_id|>assistant<|end_header_id|>

Ask Llama3 question

+ Add Chat

☐ our_skalm (skalm) ⋮

☒ our_llama3 (llama3-8B) ⋮

Please ask me anything related to this field

what is domain driven design?

Here's a concise answer: Domain-driven design is an approach to software development that focuses on modeling the domain, rather than the technology. It emphasizes understanding the problem domain, creating a model of that domain, and then using that model to drive the design of the software. ⋮

what is a database?

You're asking about databases! A database is a collection of data, which is stored in a specific format. It's like a big filing cabinet, where you can store and retrieve data as needed. In the context of object-oriented programming, a database is often used to store and retrieve objects. This is known as an object-relational database. I hope that helps clarify things! ⋮

Conclusions

- Custom text generation model was not able to perform well.
- Custom text generation model method is not suitable for this task.
- RAG method achieved the desired outcome.
- System and infrastructure can be self hosted.
- High resources are needed.

Future Work

- Deprecate support of the custom text generation model.
- Utilize a supported GPU.
- Support more generation formats like images, sounds and videos.
- Add integrations with other applications.

That was all...

Thank you for watching!