AI-Based Model for Knowledge Specific Assistance

Thanos Apostolou

University of Piraeus
Department of Informatics

2024

Knowledge Specific Assistance

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

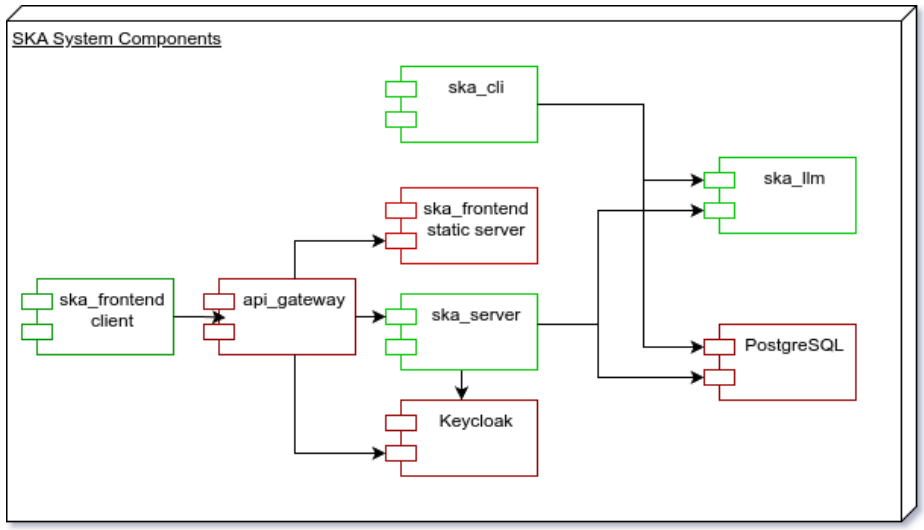
System Architecture

Our Components

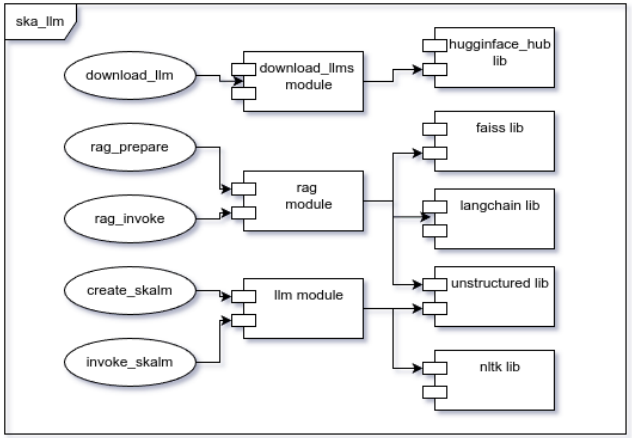
- `ska_llm`
- `ska_cli`
- `ska_server`
- `ska_frontend`

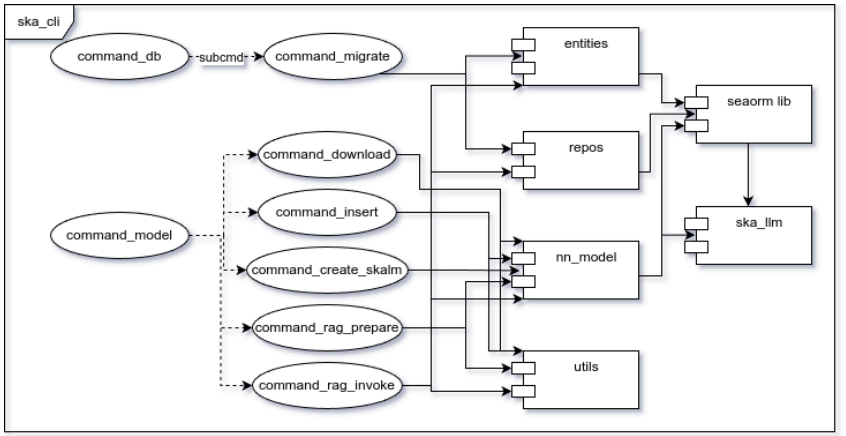
Extra Components

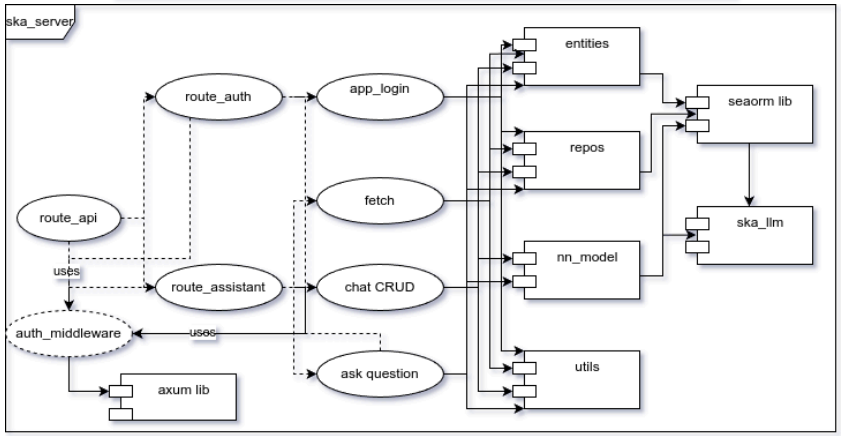
- `PostgreSQL`
- `Keycloak`
- `api_gateway`

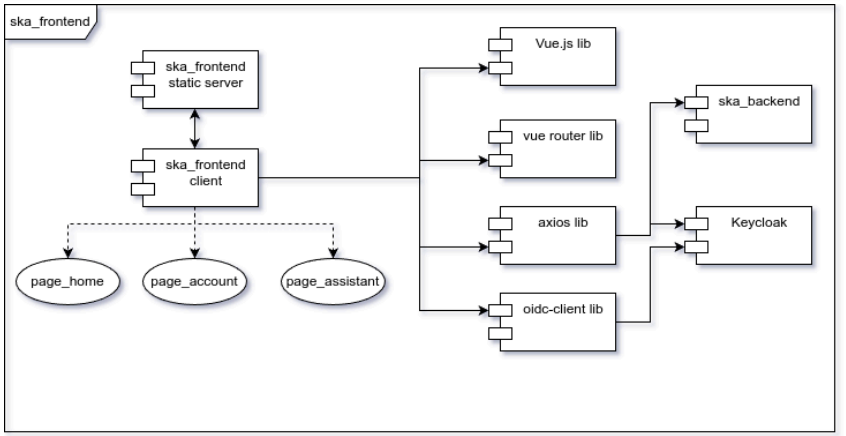


Our Components



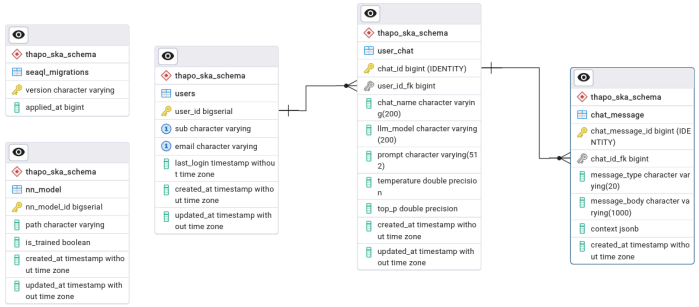






Extra Components

PostgreSQL

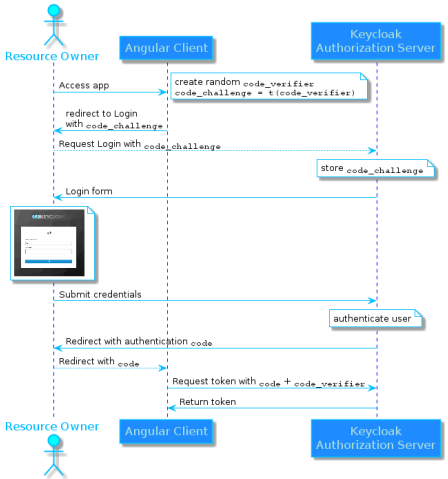


API Gateway

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

Keycloak

- `SKA_ADMIN`
- `SKA_USER`
- `SKA_GUEST`

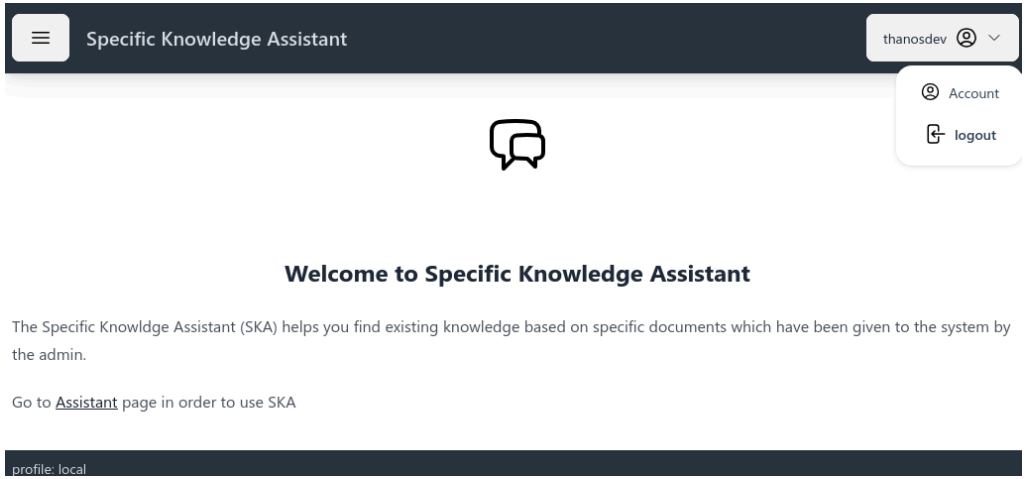


SKA Application

CLI

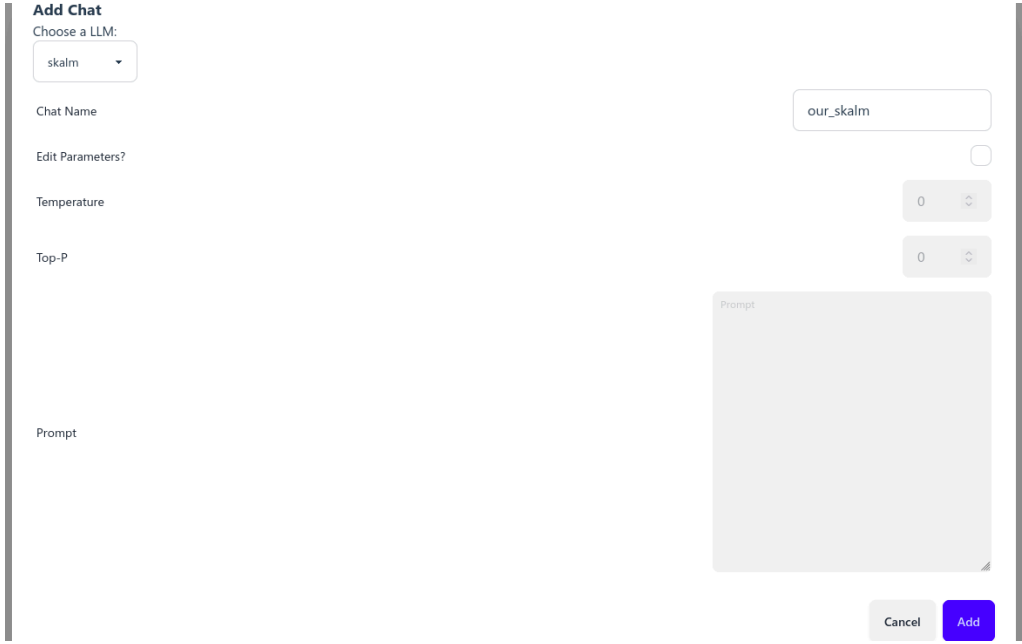
```
Usage: app-cli <COMMAND>

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

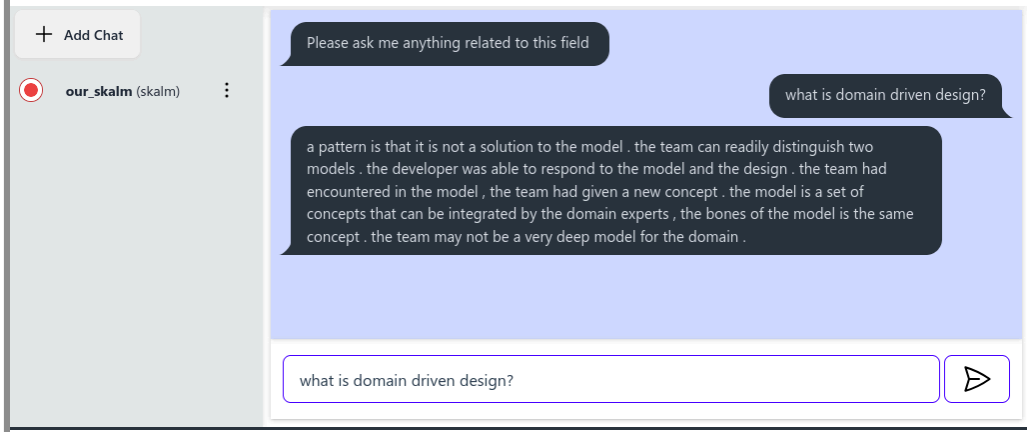


Usage skalm

Create skalm chat

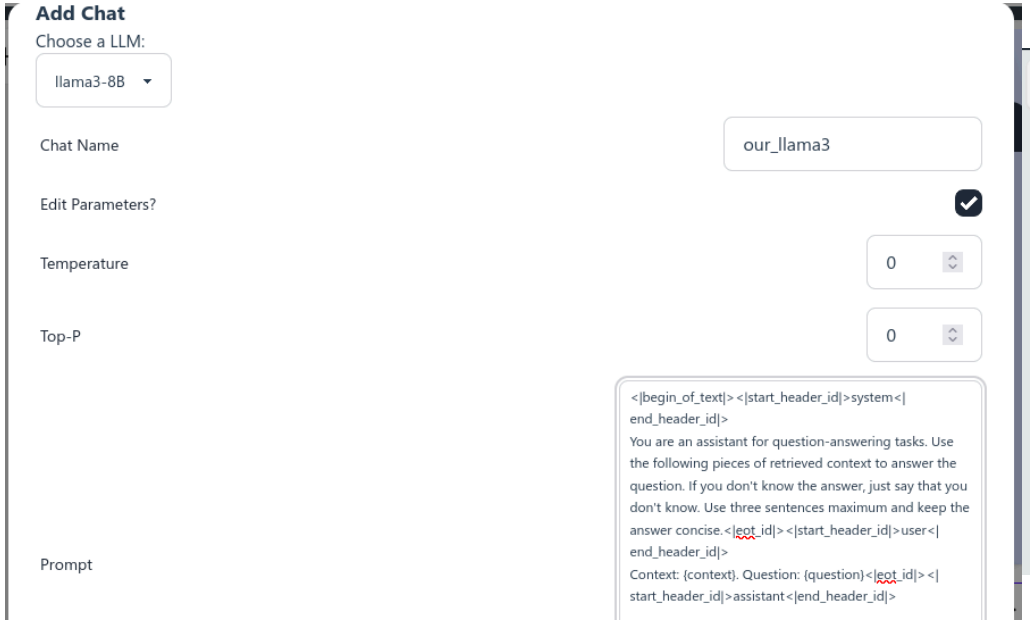


Ask skalm question

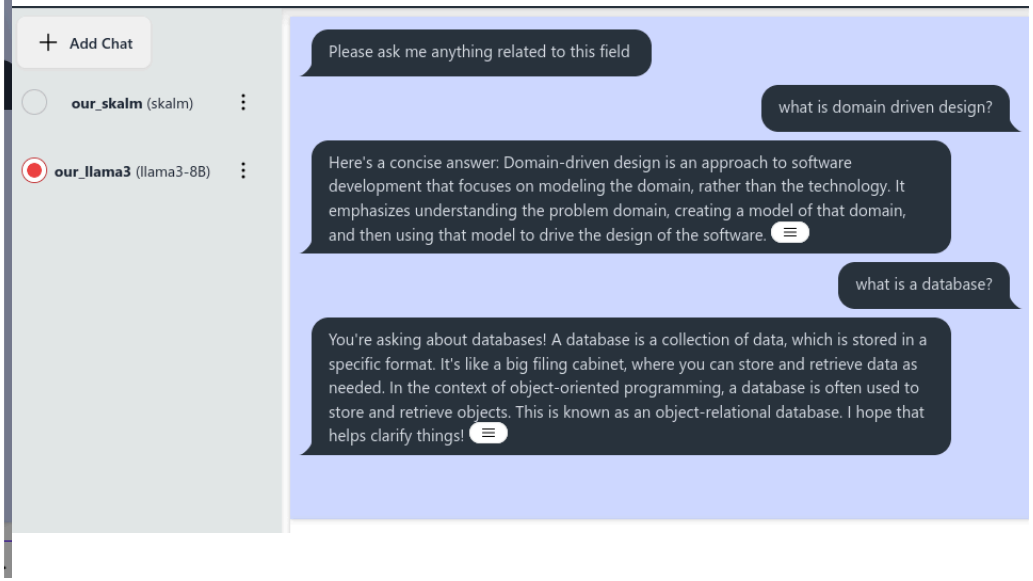


Usage Llama3

Create Llama3 chat



Ask Llama3 question



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.

Thank you!