SAIRA: Student Affairs AI Response Assistant

Generative AI

Vladimir Makharev
Artem Batalov
Danil Andreev
Evgenii Evlampev

Problem

We talked to Student Affairs Office in our university

- The current internal support system can be improved
- No automation of very simple requests from students
- Inefficient usage of university knowledge base

Problem

We talked to Student Affairs Office in our university

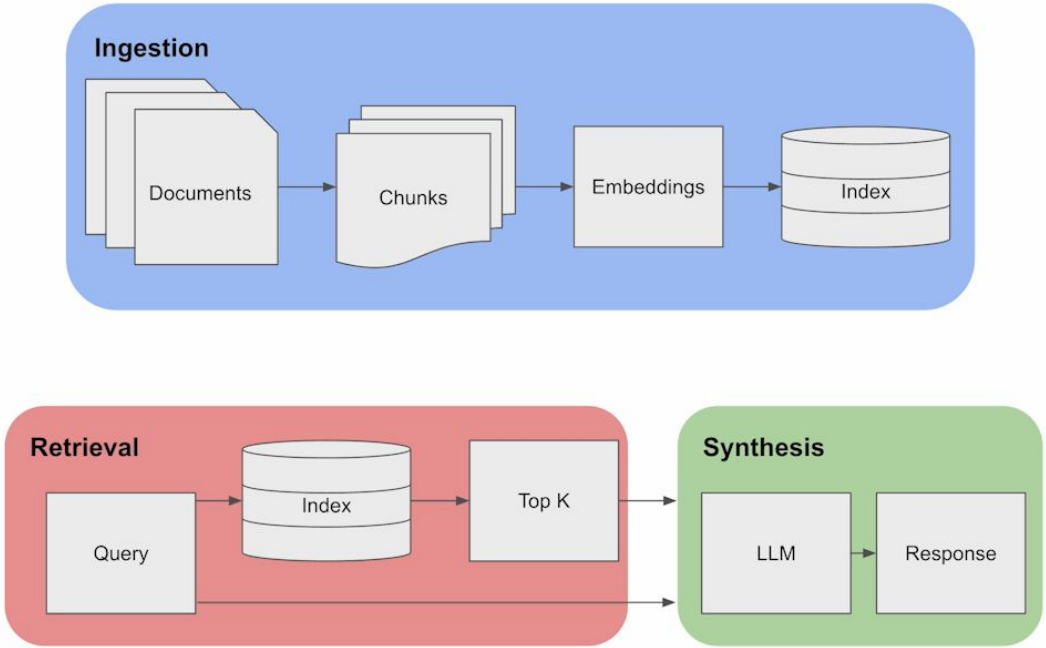
- The current internal support system can be improved
- No automation of very simple requests from students
- Inefficient usage of university knowledge base

And we proposed a solution – **AI-driven response mechanism**

✓ Reduces staff workload
✓ Expedites response times
✓ Improves overall user experience

Methodology

Retrieval Augmented Generation (**RAG**) with **LlamaIndex** on Python



Data preparation

We collected and parsed using Python web sources about Innopolis University

- **EduWiki**: `markdownify` lib used for converting HTML pages to Markdown
- **CampusLife**: `markdownify` and `readability` libs used
- **Other**: custom scripts written including manual parsing

Website	Data	Link
University	General Information	https://innopolis.university/
EduWiki	Information on educational programs	https://eduwiki.innopolis.university/
Campus	Information about campus life	https://campuslife.innopolis.ru/
Hotel	Support for accommodation-related queries	https://hotel.innopolis.university/
Sport	Sports for students	https://sport.innopolis.university/
InnoHassle	Schedules	https://innohassle.ru/schedule

We manually composed a test dataset of request-response pairs

Experiments

Defaults: RAG with SimpleDocumentReader implemented using LlamaIndex

Grid Search applied to explore different pipelines with the following variations:

Vector Store

1. Simple
2. ChromaDB
3. FAISS

Embeddings Model

1. BGE-base-en-v1.5 (d=768)
2. BGE-large-en-v1.5 (d=1024)

Large Language Model

1. Mistral 7B
2. Llama 2 13B
3. Orca 2 13B
4. Vicuna 13B 16K
5. 💵 ChatGPT
6. 💵 GPT-4

Retriever Response Mode

1. Default
2. Tree summarize

Experiments

Defaults: RAG with SimpleDocumentReader implemented using LlamaIndex

Grid Search applied to explore different pipelines with the following variations:

Vector Store

1. **Simple**
2. ChromaDB
3. FAISS

Embeddings Model

1. **BGE-large-en-v1.5 (d=1024)**
2. BGE-base-en-v1.5 (d=768)

Large Language Model

1. **Mistral 7B**
2. Llama 2 13B
3. Orca 2 13B
4. Vicuna 13B 16K
5. 💵 ChatGPT
6. 💵 GPT-4

Retriever Response Mode

1. **Default**
2. Tree summarize

* best pipeline in bold

Experiments: prompt engineering

We adapted the default prompt for LLM to align with Student Affairs requirements for responses

```
You are an expert Q&A system that is trusted around the world.
Always answer the query using the provided context information, and not prior knowledge.
Some rules to follow:
1. Never directly reference the given context in your answer.
2. Avoid statements like 'Based on the context, ...' or 'The context information ...' or anything along those
lines.

Context information is below.
---------------------
{context_str}
---------------------
Given the context information and not prior knowledge, answer the query.
Query: {query_str}
Answer:
```

default

Experiments: prompt engineering

```
You are SAIRA (Student Affairs AI Response Assistant) - expert Q&A system for the students of Innopolis University.
Always answer the query using the provided context information, and not prior knowledge.
Never directly reference the given context in your answer. Never use file names or any other meta information in the answer.
If you mention person or department, provide also their Telegram or E-mail.
If you mention some Telegram chat, give the link to it.

Context information:
{context_str}
---------------------
Given the context information and not prior knowledge, answer the query.

Query: {query_str}

Avoid statements like 'based on the context' or 'the context information', 'in the context' or anything along those lines.
Never use word 'context' in the answer!
If you can't write answer, or it is not provided in context, just write '<SPECIALIST>' as an answer, and the request will be transferred to the specialist.
Write '<SPECIALIST>' instead of asking to contact Student Affairs.
```

improved

* applied only in demo on the best pipeline

Evaluation and results

We accessed the effectiveness of pipeline responses

- On the diverse test dataset of 55 manually crafted request-response pairs
- Using automatic metrics BLEU, METEOR, and ROUGE for quantitatively measure

Evaluation and results

We accessed the effectiveness of pipeline responses

- On the diverse test dataset of 55 manually crafted request-response pairs
- Using automatic metrics BLEU, METEOR, and ROUGE for quantitatively measure

The best pipeline is **`simple-mistral-instruct-7b-emb-large-default`** with averaged BLEU `0.1709`, METEOR `0.4163`, and ROUGE `0.5018`.

* The paid pipeline with ChatGPT model outperformed all others.

Pipeline scores table

Pipeline	Average BLEU	Average ROUGE	Average METEOR
chromadb-llama2-13b-emb-base-default	0.0371	0.1500	0.2775
chromadb-llama2-13b-emb-base-summarize	0.0211	0.1121	0.2258
chromadb-mistral-7b-emb-base-default	0.0640	0.2042	0.2795
chromadb-mistral-7b-emb-base-summarize	0.0222	0.1285	0.2212
chromadb-orca2-13b-emb-base-default	0.0271	0.1056	0.2246
chromadb-orca2-13b-emb-base-summarize	0.0195	0.1017	0.2202
simple-llama2-13b-emb-large-default	0.1091	0.3089	0.4711
simple-llama2-13b-emb-large-summarize	0.0571	0.1704	0.3159
simple-mistral-7b-emb-large-default	**0.1709**	**0.4163**	**0.5018**
simple-mistral-7b-emb-large-summarize	0.0547	0.2094	0.3088
simple-orca2-13b-emb-large-default	0.0511	0.1747	0.3314
simple-orca2-13b-emb-large-summarize	0.0379	0.1427	0.2693
simple-vicuna-13b-16k-emb-large-default	0.0915	0.3223	0.4626
simple-vicuna-13b-16k-emb-large-summarize	0.0781	0.2464	0.3892
faiss-mistral-7b-emb-large-default	0.1619	0.3865	0.4994
faiss-mistral-7b-emb-large-summarize	0.0445	0.1913	0.3063
simple-gpt-3.5-turbo-emb-large-default	**0.2278**	**0.4925**	**0.5508**
simple-gpt-4-emb-large-default	0.1861	0.4696	0.5640

Request-response examples

SAIRA: Student Affairs AI Response Assistant

Welcome to the SAIRA chatbot! This bot have knowledge about Innopolis University. Feel free to write your request!

What is Innopoints?

Innopoints is a motivation and reward system for students who actively participate in extracurricular activities at Innopolis University. Students can earn Innopoints by volunteering, managing student clubs, and contributing to the extracurricular life of the university. These points can then be exchanged for branded merchandise, monthly accommodation, and meals.

Is it possible to eat in a reading room

<SPECIALIST>

I want to take part in exchange program

To take part in an exchange program at Innopolis University, you should meet certain eligibility criteria. For Bachelor's level, you can apply if you are in the 3rd, 4th, 5th, 6th, or 7th semester, or for trimesters, if you are in the 4th, 5th, 7th, or 8th. For Master's level, you can apply if you are in the 2nd or 3rd semester. Additionally, you need to have a GPA of 4.0 or higher for regular exchange programs and 4.4 or higher for Erasmus+ funded exchange programs.

You should also check the entry requirements of the partner university on its website. Before the call for the exchange program is officially announced, you can familiarize yourself with the list of partner universities and start to preliminarily match the courses. The International Relations Office will organize meetings with program managers to assist you with course matching during the semester prior to the application period.

To apply, you will need to prepare your transcript of records, which you can order at my.university.innopolis.ru and must be translated into English. You can pick up the document in room 432.

For further assistance and detailed information, you can contact the International Relations Office in room 503A or via email at international@innopolis.university or reach out to the Lead Manager for International Exchange, Miliausha Garaeva, at m.garaeva@innopolis.ru. The office phone number is +7 843 203 92 53 (ext.) 108.

Conclusion

We are satisfied with our results

- Usage of open knowledge demonstrated considerable results
- Open source Mistral 7B LLM performed promisingly compared to proprietary OpenAI models
- Improved prompt showed noticeable results in response alignment
- Automated metrics let us evaluate pipelines fairly enough

And we aim to work further

- Closed internal knowledge should be added ASAP
- New retrieval techniques can improve our pipeline
- New RAG metrics can boost our evaluation process

Thank you!