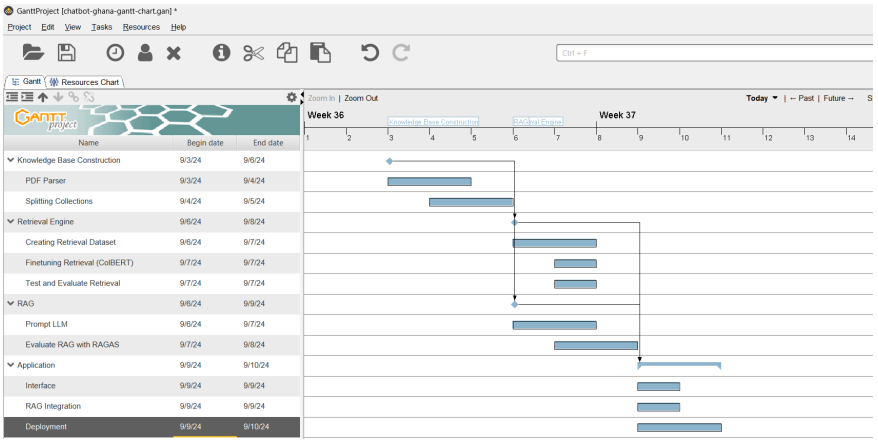
Assignee: Ghana Ahmada Yudistira

General Approach



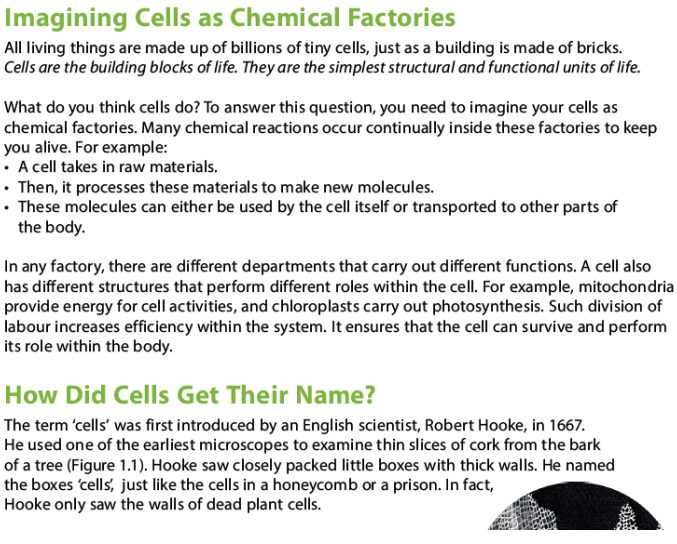
1. Knowledge Base Construction

a. PDF Parser

Parse raw text from the given PDF material related to biology using PyMuPDF library. Next, I segment the raw text into structured chunk text with format of Header and Content by utilizing the difference in font size.

b. Splitting Collections

The header and content on each page have different font sizes. Therefore, I extract the unique font sizes from each page, calculate the median, and use it as a threshold. If a block of text on a page has a font size larger than the threshold, it is classified as a header; otherwise, it is treated as content under the previous header. Using this method, the PDF material was transformed into 88 header-content pairs.



Imagining Cells as Chemical Factories

All living things are made up of billions of tiny cells, just as a building is made of bricks.
Cells are the building blocks of life. They are the simplest structural and functional units of life.

What do you think cells do? To answer this question, you need to imagine your cells as chemical factories. Many chemical reactions occur continually inside these factories to keep you alive. For example:

- A cell takes in raw materials.
- Then, it processes these materials to make new molecules.
- These molecules can either be used by the cell itself or transported to other parts of the body.

In any factory, there are different departments that carry out different functions. A cell also has different structures that perform different roles within the cell. For example, mitochondria provide energy for cell activities, and chloroplasts carry out photosynthesis. Such division of labour increases efficiency within the system. It ensures that the cell can survive and perform its role within the body.

How Did Cells Get Their Name?

The term 'cells' was first introduced by an English scientist, Robert Hooke, in 1667. He used one of the earliest microscopes to examine thin slices of cork from the bark of a tree (Figure 1.1). Hooke saw closely packed little boxes with thick walls. He named the boxes 'cells', just like the cells in a honeycomb or a prison. In fact, Hooke only saw the walls of dead plant cells.

Imagining Cells as Chemical Factories	All living things are made up of billions of tiny cells, just as a building is made of bricks.Cells are the building blocks of life. They are the simplest structural and functional units of life.What do you think cells do? To answer this question, you need to imagine your cells aschemical factories. Many chemical reactions occur continually inside these factories to keepyou alive. For example:â€¢ A cell takes in raw materials.â€¢ Then, it processes these materials to make new molecules.â€¢ These molecules can either be used by the cell itself or transported to other parts ofthe body.In any factory, there are different departments that carry out different functions. A cell alsohas different structures that perform different roles within the cell. For example, mitochondriaprovide energy for cell activities, and chloroplasts carry out photosynthesis. Such division oflabour increases efficiency within the system. It ensures that the cell can survive and performits role within the body.
How Did Cells Get Their Name?	The term â€˜cellsâ€™ was first introduced by an English scientist, Robert Hooke, in 1667.He used one of the earliest microscopes to examine thin slices of cork from the barkof a tree (Figure 1.1). Hooke saw closely packed little boxes with thick walls. He namedthe boxes â€˜cellsâ€™, just like the cells in a honeycomb or a prison. In fact,Hooke only saw the walls of dead plant cells.

Figure 1. Example of Splitting Collection

2. Retrieval Engine

a. Creating Retrieval Dataset

Next, I used GPT-4 to generate three synthetic questions for each content section. This approach was motivated by the fact that the headers alone were not sufficient to capture potential user questions. In other words, the synthetic questions are designed to mimic user inquiries and guide them to the relevant content as a source of knowledge.

question	context
What are cells compared to in the text to help us understand their importance?	All living things are made up of billions of tiny cells, just as a building is made of bricks.Cells are the building blocks of life. They are the simplest structural and functional units of life.What do you think cells do? To answer this question, you need to imagine your cells aschemical factories. Many chemical reactions occur continually inside these factories to keepyou alive. For example:• A cell takes in raw materials.• Then, it processes these materials to make new molecules.• These molecules can either be used by the cell itself or transported to other parts ofthe body.In any factory, there are different departments that carry out different functions. A cell alsohas different structures that perform different roles within the cell. For example, mitochondriaprovide energy for cell activities, and chloroplasts carry out photosynthesis. Such division oflabour increases efficiency within the system. It ensures that the cell can survive and performits role within the body.
What do mitochondria do in a cell?	All living things are made up of billions of tiny cells, just as a building is made of bricks.Cells are the building blocks of life. They are the simplest structural and functional units of life.What do you think cells do? To answer this question, you need to imagine your cells aschemical factories. Many chemical reactions occur continually inside these factories to keepyou alive. For example:• A cell takes in raw materials.• Then, it processes these materials to make new molecules.• These molecules can either be used by the cell itself or transported to other parts ofthe body.In any factory, there are different departments that carry out different functions. A cell alsohas different structures that perform different roles within the cell. For example, mitochondriaprovide energy for cell activities, and chloroplasts carry out photosynthesis. Such division oflabour increases efficiency within the system. It ensures that the cell can survive and performits role within the body.
Why is it useful for a cell to have different structures that perform different roles?	All living things are made up of billions of tiny cells, just as a building is made of bricks.Cells are the building blocks of life. They are the simplest structural and functional units of life.What do you think cells do? To answer this question, you need to imagine your cells aschemical factories. Many chemical reactions occur continually inside these factories to keepyou alive. For example:• A cell takes in raw materials.• Then, it processes these materials to make new molecules.• These molecules can either be used by the cell itself or transported to other parts ofthe body.In any factory, there are different departments that carry out different functions. A cell alsohas different structures that perform different roles within the cell. For example, mitochondriaprovide energy for cell activities, and chloroplasts carry out photosynthesis. Such division oflabour increases efficiency within the system. It ensures that the cell can survive and performits role within the body.

b. Finetuning Retrieval

Then, I created 10 pairs of questions and content for each content section. Each pair includes one relevant question (labeled as a positive sample, 1) and nine irrelevant questions (labeled as negative samples, 0). In total, I generated 1,920 samples to fine-tune ColBERT as the retrieval engine using supervised learning.

c. Test and Evaluate Retrieval

The metric used to evaluate the retrieval performance is Mean Reciprocal Rank (MRR), as illustrated in the figure below.

After completing the fine-tuning process for ColBERT, I compared the fine-tuned model with the zero-shot performance of other backbone models used in ColBERT. The fine-tuned version achieved an MRR@10 of 0.846, while the zero-shot version scored 0.81.

3. RAG

a. Prompt LLM

The selected language model is LLaMa3-8b-Instruct, which I accessed via the Huggingface Inference API. While it is free to use, it has a limited rate of API calls. I experimented with several prompts, ranging from basic to advanced, ensuring they met the requirements for this technical test to guide the LLM effectively. After these experiments, I developed a prompt that is neither too short nor overly complex.

Original prompt

```
Jawab pertanyaan berikut sebagai guru sesuai dengan bahasa yang digunakan dalam pertanyaan. Jangan jawab bila pertanyaan berkaitan dengan isu kekerasan, sex, LGBT, dan hal buruk lainnya. Gunakan tone ramah dan friendly

<context>

{context}

</context>

Apabila pertanyaan tidak sesuai dengan konteks yang dibahas, maka jawab saja pertanyaan dengan pengetahuan yang ada.

Question: {input}

Note: Answer in the same language as the question.
```

Translated prompt

```
Answer the following questions as a teacher in the same language used in the question. Do not answer if the questions involve issues related to violence, sex, LGBT, or other harmful topics. Use a friendly and welcoming tone.

<context>

{context}

</context>
```

```
If the question does not fit the context discussed,
answer the question with the knowledge available.

Question: {input}

Note: Answer in the same language as the question.
```

b. Evaluate RAG with RAGAS

The dataset used to evaluate the RAG with RAGAS is the same as the retrieval dataset but selecting only the positive sample, consisting of 192 samples. Note that the content used as context for the RAG is retrieved by the fine-tuned ColBERT retrieval engine, rather than directly using the ground truth content.

```
context_precision: 0.8743455496507854
context_recall: 0.7972324346405228
faithfulness: 0.8300685425685425
answer_relevancy: 0.9247891386626608
```

RAGAS Evaluation Report

RAGAS Notebook Experiment

4. Application

For the application development, I used the Django framework, as I find it comfortable to work with. While developing the RAG for the chatbot system, I have not yet integrated previous dialogues into the upcoming ones for continual conversation. As a result, the chatbot is currently only capable of responding to direct instructions.