

---

---

# 잼민이(Gemini)동생 잼마 (Gemma)에게 컨닝페이퍼를


ZeroPage 32기 김동우

---



# 목차

- 1. 잼민이와 잼마가 누구인가?
- 2. 잼마에게 컨닝페이퍼를 주는 법
- 3. 실제로 잼마에게 컨닝페이퍼를 제공해보자!
- 4. 더 생각해볼 거리



## 1. 잼민이와 잼마가 누구인가?

→ 잼민이(Gemini)

→ 잼마(Gemma)

# Gemini

동우 gemini가 원지 간단하게 소개해줘.

Show drafts ▾ 🔊

## ◆ 제미니 소개

제미니는 구글 AI에서 개발한 대화형 생성형 인공지능(AI) 챗봇입니다. 텍스트 생성, 번역, 코드 작성, 질문 답변 등 다양한 기능을 수행하며, 사용자와 자연스러운 대화를 이어갈 수 있습니다.

### 주요 특징

- **대화형:** 사용자와 자연스러운 대화를 이어갈 수 있습니다.
- **생성형:** 텍스트, 코드, 번역, 질문 답변 등 다양한 콘텐츠를 생성할 수 있습니다.
- **다국어:** 한국어를 포함한 여러 언어를 지원합니다.

### 사용 방법

- **클라우드 API:** 인터넷 연결만 있으면 누구나 사용할 수 있습니다.
- **자체 서버:** 더 많은 제어권과 맞춤화가 필요한 경우 자체 서버에 배포하여 사용할 수 있습니다.

### 용도

- **고객응대:** 고객 문의에 자동으로 응답하고, 24시간 고객 지원을 제공합니다.
- **콘텐츠 제작:** 텍스트, 코드, 번역, 질문 답변 등 다양한 콘텐츠를 자동으로 생성합니다.
- **교육:** 학생들에게 개인 맞춤형 학습 경험을 제공합니다.
- **엔터테인먼트:** 챗봇, 스토리텔링, 게임 등 다양한 엔터테인먼트 콘텐츠를 제공합니다.

제미니는 다양한 분야에서 활용될 수 있는 강력한 AI 도구입니다.

### 더 자세한 정보

- 제미니 공식 웹사이트: <https://gglirimi.com/desktop/>
- 제미니 블로그: <http://m.blog.naver.com/llionji8723/220983412219>
- 제미니 GitHub 저장소: <https://github.com/AutoMiny>

Gemini Advanced ▾



# Hello, 동우

## How can I help you today?

Generate an image in a specific style



Help me write HTML, CSS, and JS

```
<!DOCTYPE html>
<html>
<head>
<title>Image Slider</title>
<link rel="stylesheet"
href="style.css">
```

Role-play as an event planner

1. Type of event:  
What kind of event do you envision (e.g., a film screening, a gala evening, a networking event, a workshop)?
2. Target Audience

Walk me through solving a problem

- Step 1: Identify the chemical formula of cane sugar.  
Cane sugar, also known as sucrose, has the chemical formula  $C_{12}H_{22}O_{11}$ .
- Step 2: Calculate the molar mass of cane sugar.



Your conversations are processed by human reviewers to improve the technologies powering Gemini Apps. Don't enter anything you wouldn't want reviewed or used.

Enter a prompt here



# 구글의 최첨단 오픈 모델 ‘젬마(Gemma)’를 공개합니다

# Gemma

2월 21, 2024

젬마는 제미니AI 모델을 개발하기 위해 활용됐던 것과 동일한 연구와 기술을 바탕으로 책임감 있는 AI 개발을 위해 탄생하였습니다.



**Jeanine Banks**  
VP & GM, Developer X  
and DevRel



**Tris Warkentin**  
Director, Google  
DeepMind

공유하기



구글은 **모두에게 유용한 인공지능(AI)**을 만들 수 있다고 생각합니다. 이의 일환으로 오랫동안 구글은 오픈 커뮤니티에 [Transformers](#), [TensorFlow](#), [BERT](#), [T5](#), [JAX](#), [AlphaFold](#), and [AlphaCode](#)와 같은 혁신 기술들을 공유해 왔습니다. 그리고 오늘 구글은 다시 한번 책임감 있게 AI를 이용하는 개발자와 연구진들을 지원하기 위해 새로운 세대의 오픈 모델을 공개합니다.

## 오픈 모델 젬마

**젬마(Gemma)**는 제미니AI 개발의 핵심 기술과 연구를 기반으로 제작된 최첨단 경량 **오픈 모델군**입니다. 구글 딥마인드와 구글의 다양한 부서가 협력하여 개발한 젬마는 **제미니AI**를 비롯하여 '보석'을 뜻하는 라틴어 'gemma'에서 영감을 받은 이름입니다. 또한, 다양한 모델 규모와 함께 개발자들의 혁신을 지원하고, 협업을 촉진하며 젬마 모델을 더욱 책임감있게 사용할 수 있는 톨도 함께 공개됩니다.

출처 : [blog.google](https://blog.google)

# Gemma

CAPABILITY			Gemma		Llama-2	
BENCHMARK			7B		7B	13B
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities and others)	64.3		45.3	54.8
	5-shot, top-1					
Reasoning	BBH	Diverse set of challenging tasks requiring multi-step reasoning	55.1		32.6	39.4
	-					
	HellaSwag	Commonsense reasoning for everyday tasks	81.2		77.2	80.7
	0-shot					
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	46.4		14.6	28.7
	maj@1					
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	24.3		2.5	3.9
	4-shot					
Code	HumanEval	Python code generation	32.3		12.8	18.3
	pass@1					

출처 : [blog.google](https://blog.google/teams/ai/gemma/)

# Gemma

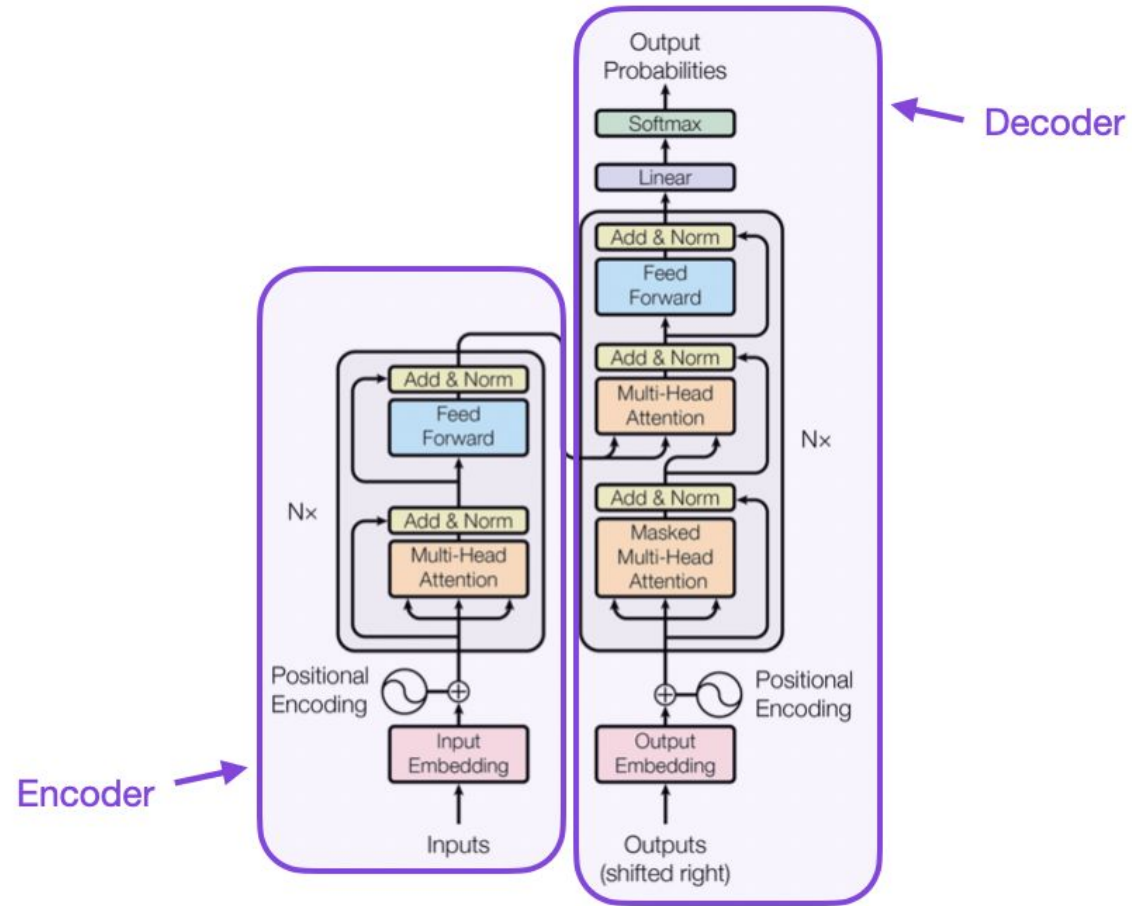


Figure 1: The Transformer - model architecture.

# Gemma

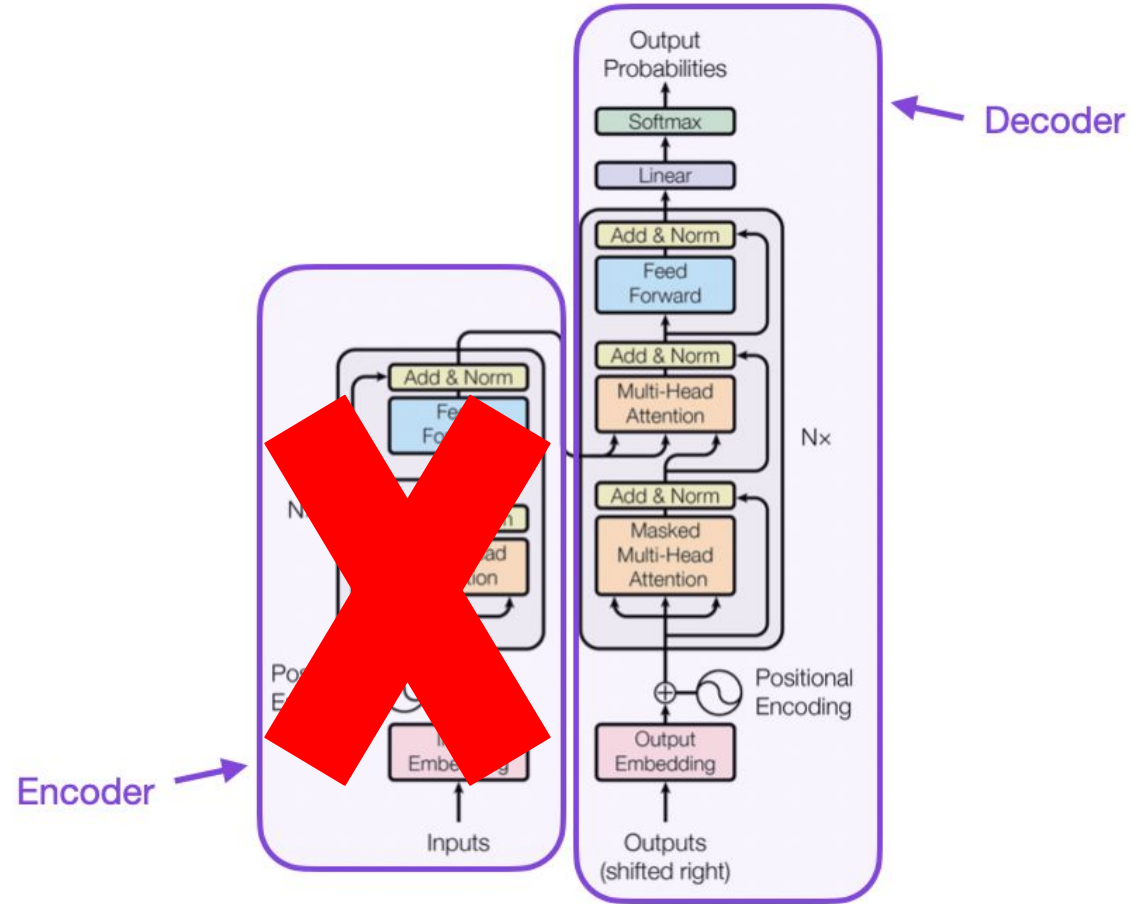



Figure 1: The Transformer - model architecture.





## 2. 젤마에게 컨닝페이퍼를 주는 법

→ 젤마에게 컨닝페이퍼가 없다면?

→ RAG

# 젼마에게 컨닝페이퍼가 없다면?

```
In 6 1 print(llms.invoke("Who is DoYeop Kim?"))
```

Executed at 2024.03.11 22:48:02 in 33s 168ms

DoYeop Kim is a software engineer and entrepreneur specializing in artificial intelligence and machine learning. He is known for his work on AI-powered systems for image recognition, natural language processing, and data analysis. Here's a summary of his accomplishments:

**\*\*Key Contributions:\*\***

\* **\*\*Image Recognition:\*\*** Developed AI models for object detection and facial recognition with applications in self-driving cars, security systems, and medical diagnosis.

\* **\*\*Natural Language Processing:\*\*** Created AI systems for text summarization, sentiment analysis, and machine translation, improving communication and information extraction.

\* **\*\*Data Analysis:\*\*** Designed and implemented AI-powered data analysis platforms for cleaning insights from large datasets, optimizing business processes and predicting future

```
In 7 1 print(llms.invoke("Who is DoYeop Kim, undergraduate student of CAU?"))
```

Executed at 2024.03.11 22:49:54 in 8s 179ms

I do not have access to private information, therefore I cannot answer the question of "Who is DoYeop Kim, undergraduate student of CAU?"

# 젼마에게 컨닝페이퍼가 없다면?

```
In 10 1 print(llms.invoke("김동우가 누구야?"))
```

Executed at 2024.03.11 23:07:36 in 11s 788ms

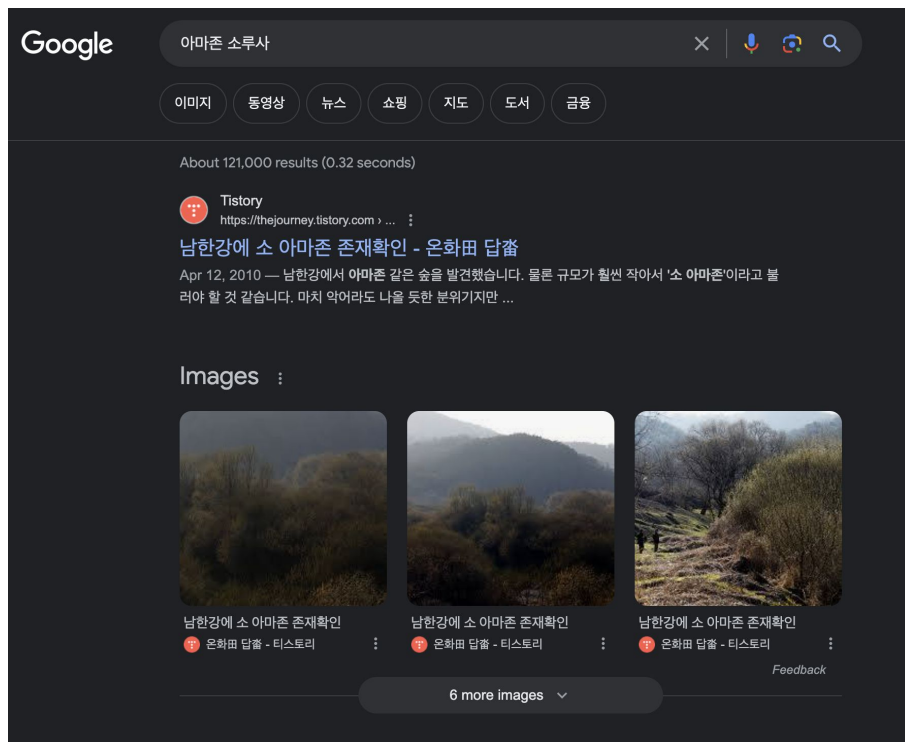
김동우는 한국의 배우입니다. 그는 다양한 역할을 맡아 많은 인정을 받은 배우입니다. 그의 가장 유명한 역할은 아마존 소루사의 역할입니다.

```
In 11 1 print(llms.invoke("중앙대학교 학부생인 김동우가 누구야?"))
```

Executed at 2024.03.11 23:07:41 in 3s 294ms

김동우는 중앙대학교 학부생으로, 현재는 미상 알려지지 않은 정보입니다.

# 젼마에게 컨닝페이퍼가 없다면?



## 아마존 소루사???

# 젼마에게 컨닝페이퍼가 없다면?

```
In 13 1 print(llms.invoke("김동우가 누구야?"))
```

Executed at 2024.03.11 23:16:46 in 3s 475ms

김동우는 대한민국의 정치인이자 국회의원입니다.

```
In 14 1 print(llms.invoke("중앙대학교 학부생인 김동우가 누구야?"))
```

Executed at 2024.03.11 23:16:47 in 1s 119ms

김동우는 중앙대학교 학부생입니다. 그의 학번은 20230101입니다.

# 젼마에게 컨닝페이퍼가 없다면?

=> 김동우가 23학번?



젼마에게 컨닝페이퍼가 없다면?

환각을  
줄일려면???

# RAG - Retrieval Augmented Generation

---

## Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

---

Patrick Lewis<sup>†‡</sup>, Ethan Perez<sup>\*</sup>,

Aleksandra Piktus<sup>†</sup>, Fabio Petroni<sup>†</sup>, Vladimir Karpukhin<sup>†</sup>, Naman Goyal<sup>†</sup>, Heinrich Küttler<sup>†</sup>,

Mike Lewis<sup>†</sup>, Wen-tau Yih<sup>†</sup>, Tim Rocktäschel<sup>†‡</sup>, Sebastian Riedel<sup>†‡</sup>, Douwe Kiela<sup>†</sup>

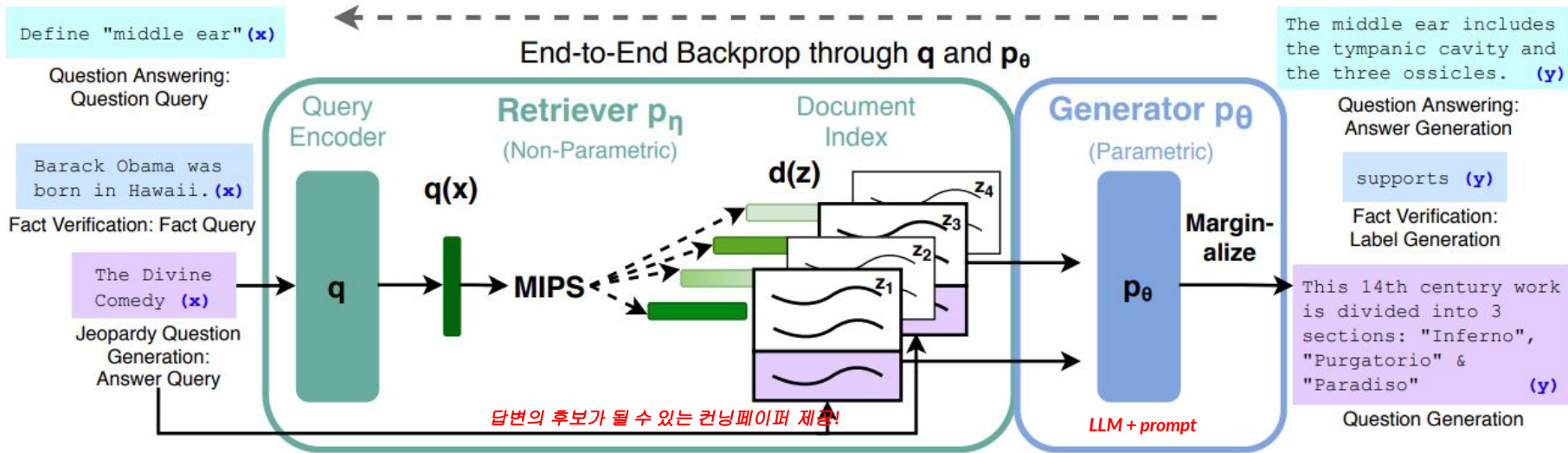
<sup>†</sup>Facebook AI Research; <sup>‡</sup>University College London; <sup>\*</sup>New York University;  
plewis@fb.com

### Abstract

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, and another which can use different passages

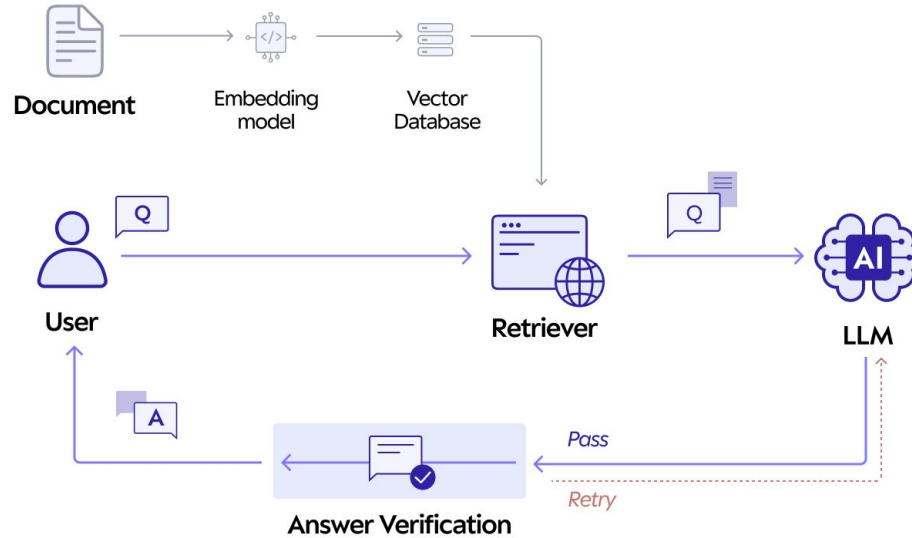


# RAG - Retrieval Augmented Generation



$x$ : 인풋 시퀀스,  $q(x)$ : query encoder 를 통해 나온 representattion,  $z$ : 문서 chunk,  $d(z)$ :document  $z$ 가 임베딩된 벡터,  $p(z|x)$ : input이  $x$ 일때  $z$ 가 검색될 확률

# RAG - Retrieval Augmented Generation



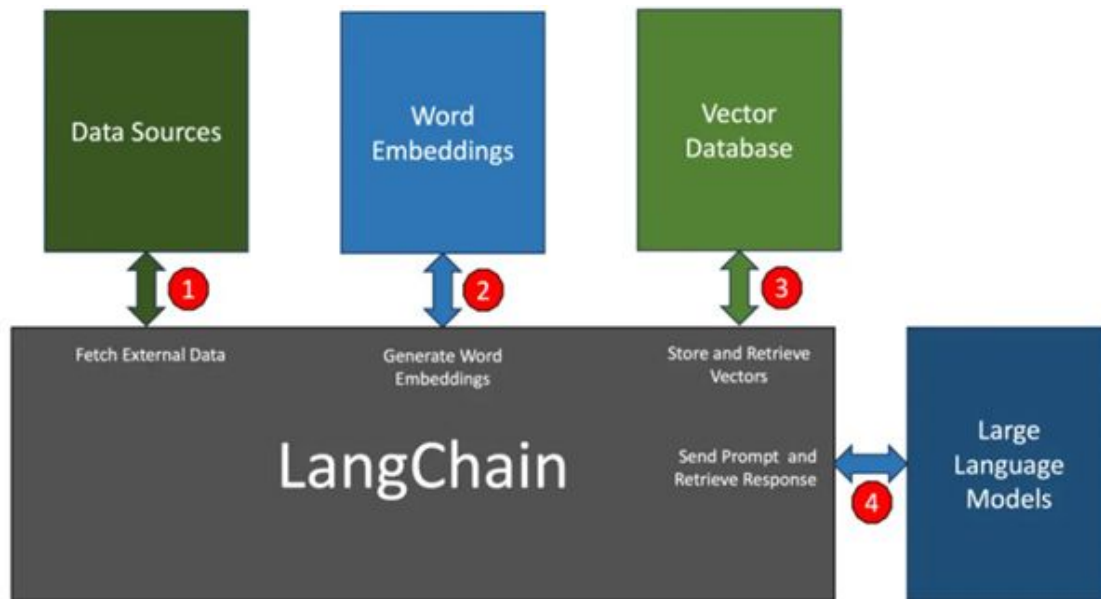


### 3. 실제로 줌마에게 컨닝페이퍼를     제공해보자!

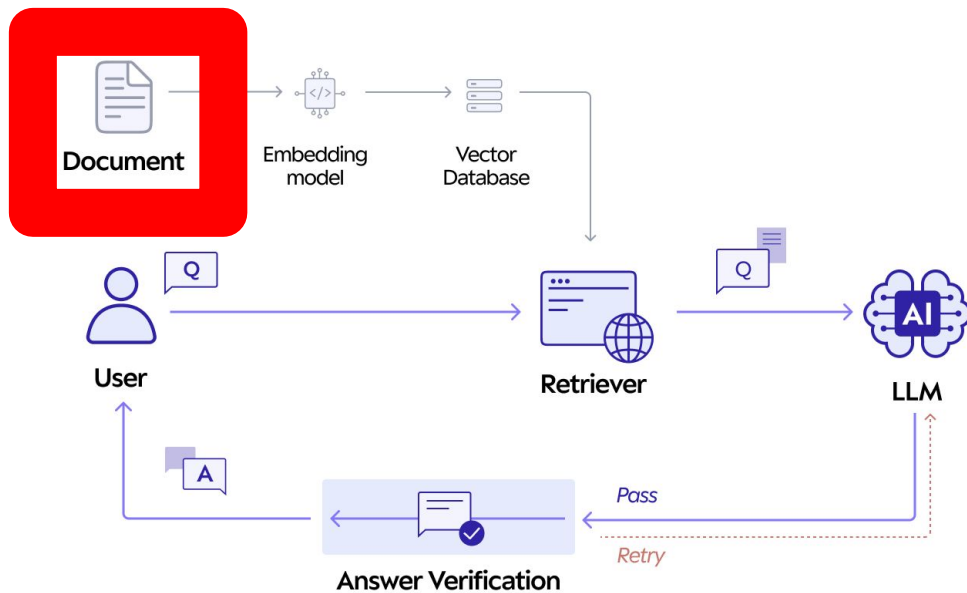
→ **LangChain**

→ 제로위키를 줌마에게 컨닝페이퍼로  
주기

# LangChain



# 제로위키를 쟈마에게 컨닝페이퍼로 주기



# 제로위키를 켄마에게 컨닝페이퍼로 주기

[FrontPage](#) [FindPage](#) [TitleIndex](#) [RecentChanges](#) [UserPreferences](#)

## Zero Page

중앙대학교 컴퓨터공학부 학술 연구회 **ZeroPage** 입니다.

[edit](#)

■ 위키 데이터가 남아가지 많은 부분에 공백이 있습니다. 특히 최근 05-07년 데이터가 업데이트 되지 않았습니다.

## ZeroPage 는

[edit](#)

[제로페이지](#)는 공부하고자하는 뜻이 있는 사람들이 모인 일종의 인력의 장입니다. 그 안에서 뜻이 같은 사람들을 만날수 있기를, 또는 자신이 아는 것에 대해 다른 사람들에게 전달해줄수 있기를, 또는 자신의 부족한 점을 다른 사람로부터 얻어갈 수 있었으면 합니다. 개인의 이익들이 모여서 집단의 이익을 만들어가며, 집단의 이익을 추구하는 것이 곧 개개인들에게 이익이 되는 경지가 되었으면 합니다.

여러가지 힘든 여건 속에서도 뜻이 있는 사람들의 수고와 노력으로 지금까지 이어져 온라인/오프라인 활동을 계속하고 있습니다. 지금까지의 **ZeroPage** 는 **ZeroPage**를 거쳐간 사람들이 만들어놓은, 혹은 현재 활동중인 사람들의 과정에 대한 결과이며, 다시 많은 사람들에 의해 재창조되고 새로운 모습을 만들어나가게 될 것입니다. 언제나 '현재 진행형' 중인 학회이기를. 이 정의 또한 사람들이 스스로 깨뜨리고, 더 좋은 방향으로 바꾸어나가기를.

## ZeroPage Wiki

[edit](#)

현재는 02년도부터 도입한 wiki 시스템을 통하여 각 프로젝트나 스터디를 진행할때마다 문서 산출물을 만들어 내어 양질의 정보들을 축적해 나가고 있다. 이 시스템은 스터디와 프로젝트를 팀으로 진행할때 공동 knowledge repository 의 역할을 함으로서 진행하는 회원 들에게 도움이 되고, 추후에 다른 회원이 비슷한 스터디나 프로젝트를 할때 그 wiki 페이지를 참고 함으로써 같은 곳에 쏟을 노력을 줄이고, 그 wiki 페이지를 다시 키워 나감으로써 지속적으로 양질의 정보를 축적하여왔다. 이로서 제로페이지의 wiki 시스템은 현재의 회원과 학교를 떠난 회원그리고 앞으로 제로페이지에 들어올 회원들 모두에게 도움이 되는 시스템으로서 자리매김하고 있다.

# 제로위키를 켜마에게 컨닝페이퍼로 주기

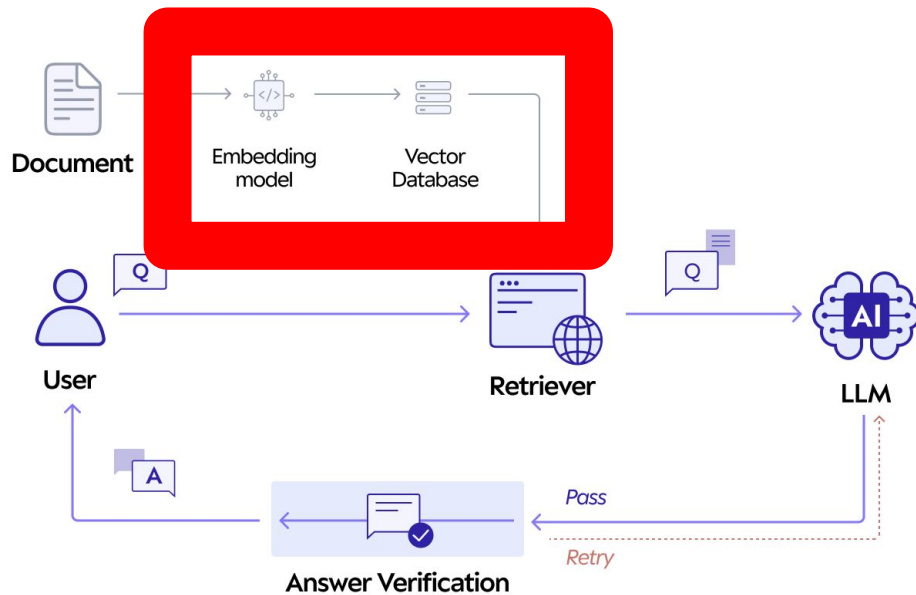
```
In [100]: loader = WebBaseLoader(
1         web_paths=("https://wiki.zeropage.org/wiki.php/ZeroPage"),
2         bs_kwargs=dict(
3             parse_only=bs4.SoupStrainer(
4                 "div",
5                 attrs={"class": ["section"]},
6             )
7         ),
8     )
9
10 docs = loader.load()
11 print(f"문서의 수: {len(docs)}")
12 print(docs)
13
```

Executed at 2024.03.13 00:42:25 in 219ms

▼ 문서의 수: 1

[Document(page\_content='[edit]\n중양대학교 컴퓨터공학부 학술 연구회 ZeroPage 입니다. ¶\n위키 데이터가 날아가서 많은 부분에 공백이 있습니다. 특히 최근 05-07년 데이터가 업데이트 되지 않았습니다.\n\n\n\n\n[edit]\nZeroPage는 ¶\n제로페이는 공부하고자하는 뜻이 있는 사람들이 모인 일종의 인력의 장입니다. 그 안에서 뜻이 같은 사람들을 만날수 있기를, 또는 자신이 아는 것에 대해 다른 사람들에게 전달해줄수 있기를, 또는 자신의 부족한 점을 다른 사람들로부터 얻어갈 수 있었으면 합니다. 개인의 이익들이 모여서 집단의 이익을 만들어가며, 집단의 이익을 추구하는 것이 곧 개개인들에게 이익이 되는 것이다'면 해피엔딩!')]

# 제로위키를 젤마에게 컨닝페이퍼로 주기





# 제로위키를 켄마에게 컨닝페이퍼로 주기

```
In 101 1 text_splitter = RecursiveCharacterTextSplitter(  
2         chunk_size=700, chunk_overlap=10)  
3  
4 splits = text_splitter.split_documents(docs)  
5 len(splits)
```

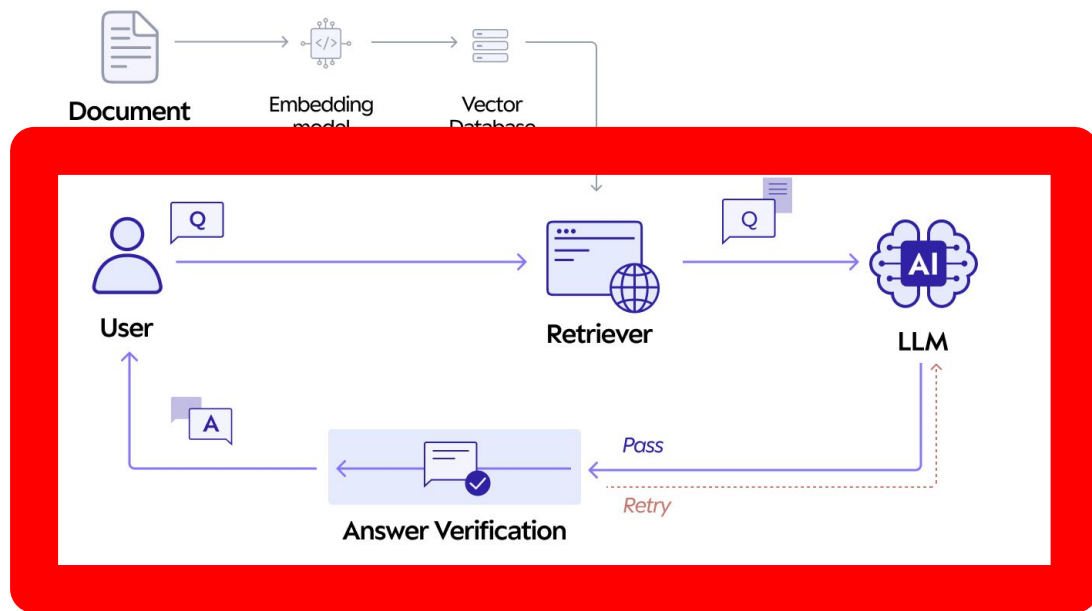
Executed at 2024.03.13 00:42:26 in 4ms

Out 101 8

```
In 102 1 vectorstore = FAISS.from_documents(  
2         documents=splits, embedding=openAIEmbedding  
3     )  
4  
5 retriever = vectorstore.as_retriever()
```

Executed at 2024.03.13 00:42:30 in 3s 386ms

# 제로위키를 켄마에게 컨닝페이퍼로 주기



# 제로위키를 젤마에게 컨닝페이퍼로 주기

```
In 107 1 llm = ChatOllama(model="gemma")
        2
        3 def format_docs(docs):
        4     return "\n\n".join(doc.page_content for doc in docs)
        5
        6 rag_chain = (
        7     {"context": retriever | format_docs, "question": RunnablePassthrough()}
        8     | prompt
        9     | llm
       10     | StrOutputParser()
       11 )
```

Executed at 2024.03.13 00:43:09 in 4ms

# 제로위키를 켄마에게 컨닝페이퍼로 주기

```
In 108 1 print(rag_chain.invoke(  
2     "ZeroPage 라는 명칭을 제시하신 분은 누구이신가요?"  
3 ))
```

Executed at 2024.03.13 00:43:42 in 31s 801ms

ZeroPage 라는 명칭을 제시하신 분은 91학번 이찬근님(현재 중앙대학교 컴퓨터공학부 교수.)입니다. ZeroPage 는 공부하고자 하는 뜻이 있는 사람들이 모인 일종의 인력의 장입니다.

Add Code Cell

Add Markdown Cell

```
In 16 1 print(rag_chain.invoke(  
2     "ZeroPage 라는 명칭을 최초로 제시하신 이찬근님은 몇 학번이신가요?"  
3 ))
```

Executed at 2024.03.13 12:59:20 in 24s 374ms

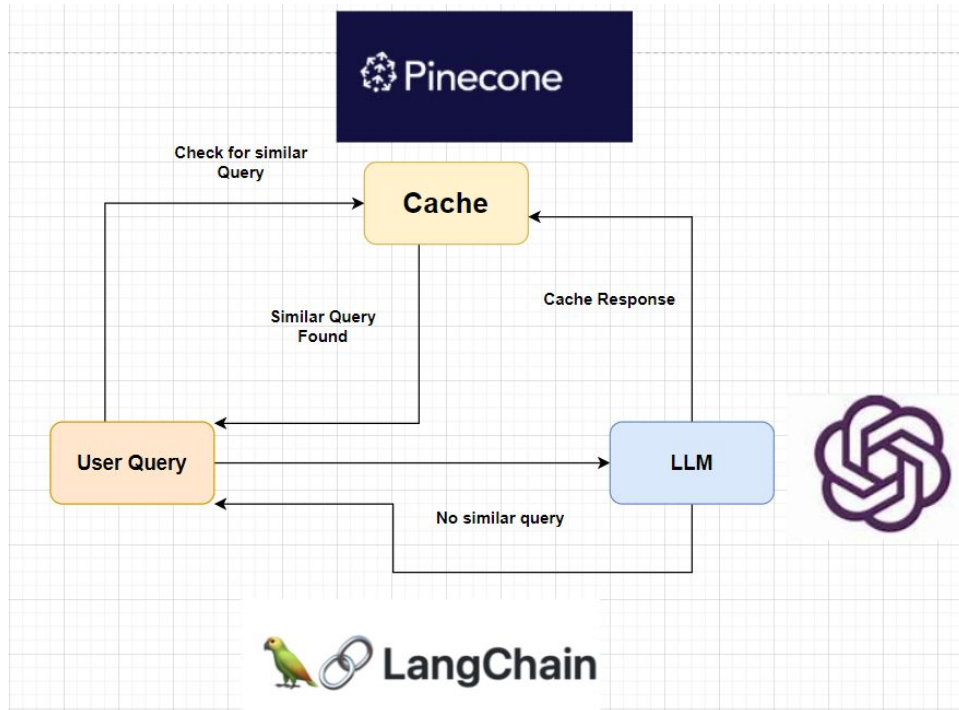
ZeroPage 라는 명칭을 최초로 제시하신 이찬근님은 91학번이었습니다. ZeroPage 는 공부하고자 하는 뜻이 있는 사람들이 모인 일종의 인력의 장입니다.



#### 4. 더 생각해볼 거리

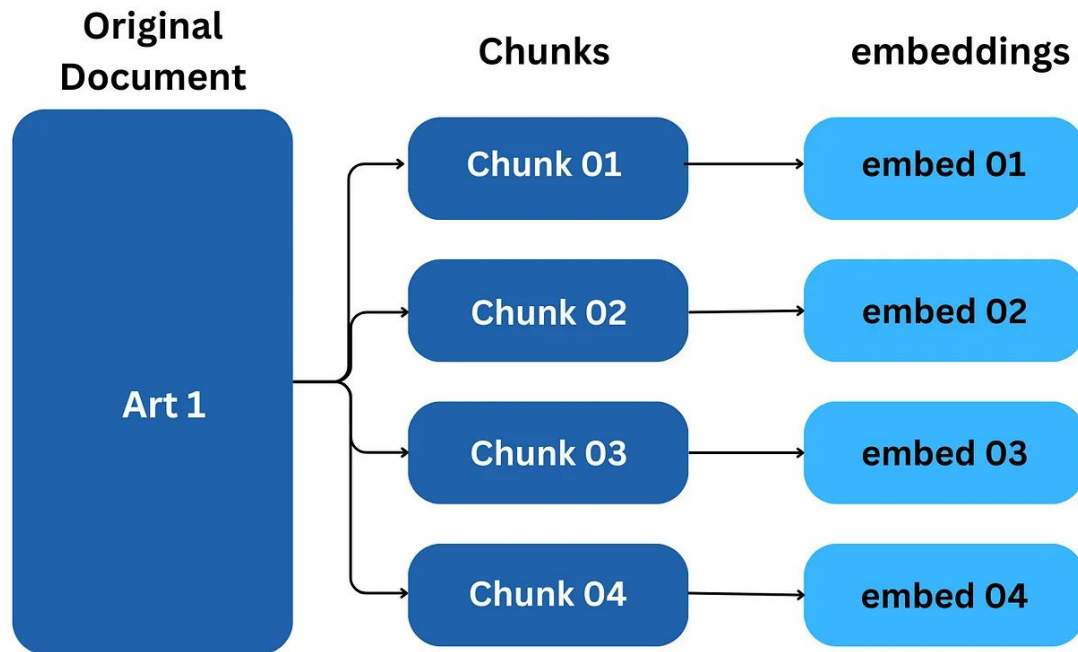
- Caching
- Chunking
- Query

# Caching



출처 : <https://medium.aiplanet.com/slash-openai-api-costs-using-caching-265e2638ac2d>

# Chunking



# Chunking

## ZeroPage란 이름의 기원은?

---

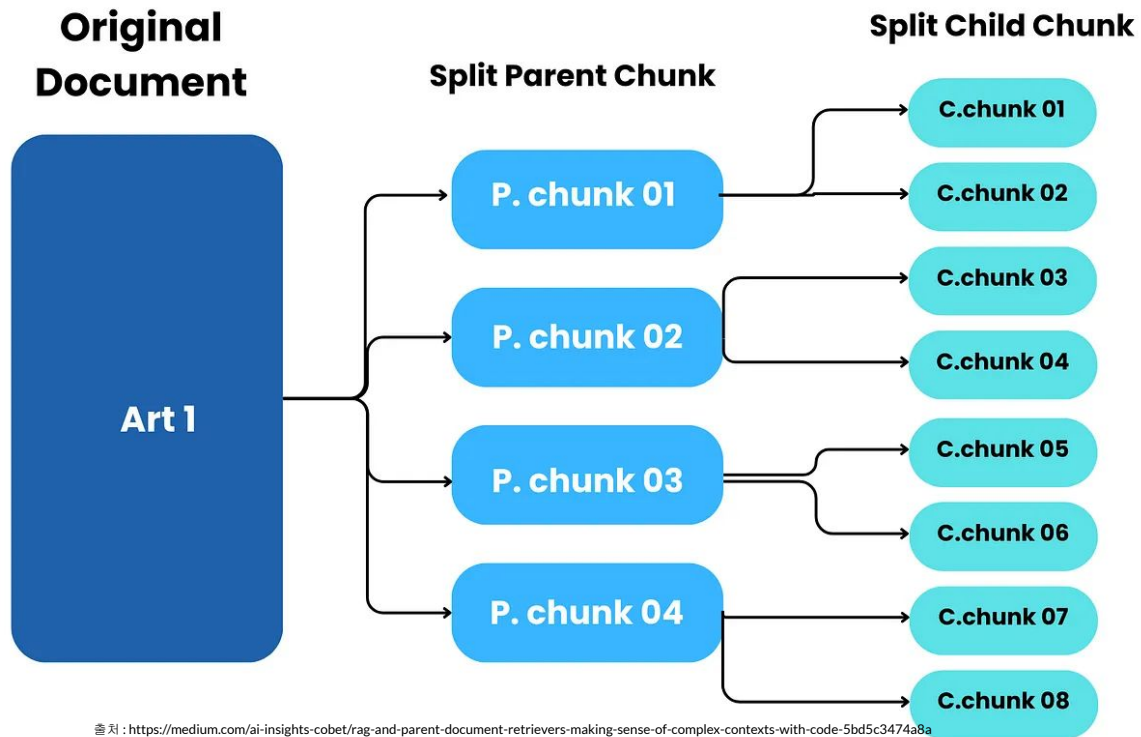
**ZeroPage** 라는 명칭에 대해서는 91학번 이찬근님(현재 중앙대학교 컴퓨터공학부 교수.)이 최초로 제시한 것으로 알려져있음. (from 테크니카(김승태) )

Chunk가 여기서 나뉘면

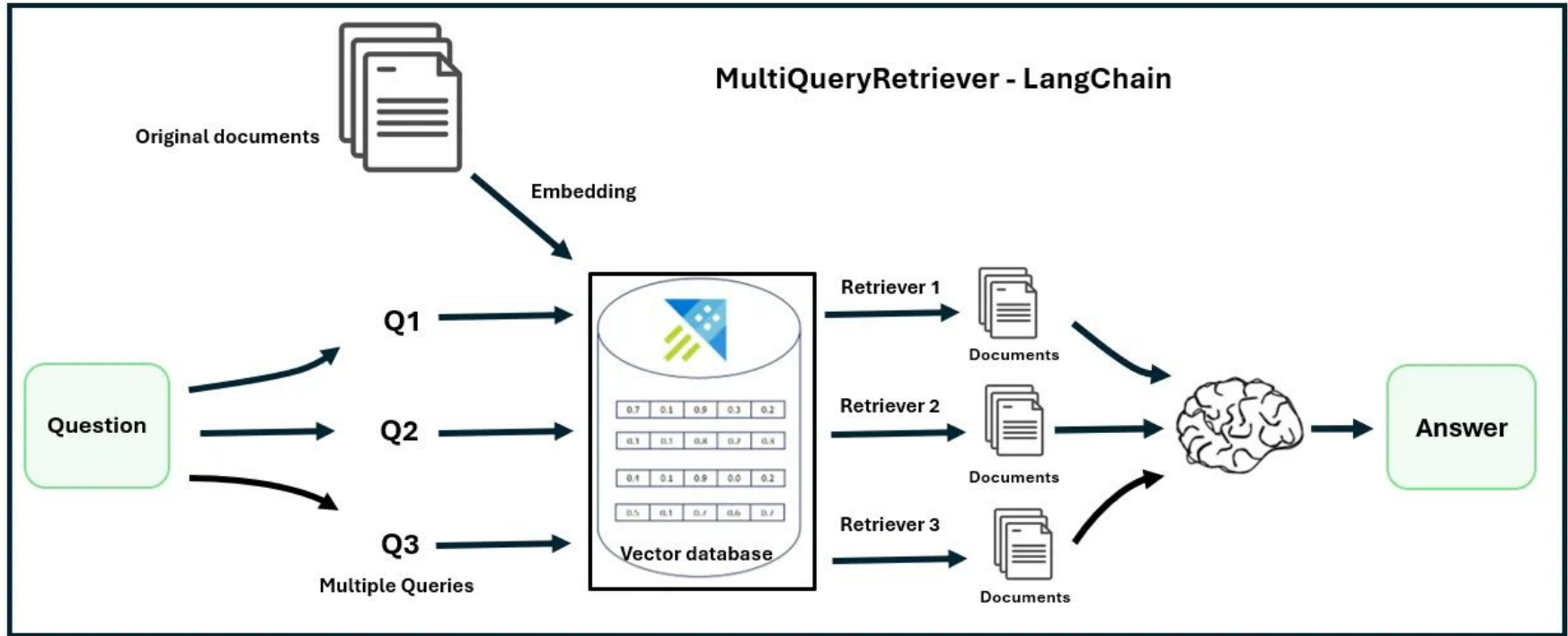
젼마는 이찬근 교수님이 몇 학번이신지 모를 것!



# Chunking



# Query



# Query

ZeroPage어때?

=> ZeroPage는 어떤 분야에서 활동하고 있어?

=> ZeroPage에서 참여할 수 있는 활동은 무엇이 있어?

=> ZeroPage의 분위기는 어때?



마무리