

AGENTIC RETRIEVAL-AUGMENTED GENERATION: A SURVEY ON AGENTIC RAG

AID 6기 강민석

기계공학부 4학년

Github: [@myeolinmalchi](https://github.com/myeolinmalchi)

Mail: rkd2274@pusan.ac.kr



부산대학교
PUSAN NATIONAL UNIVERSITY

Foundations of
Retrieval-Augmented Generation

Naïve RAG

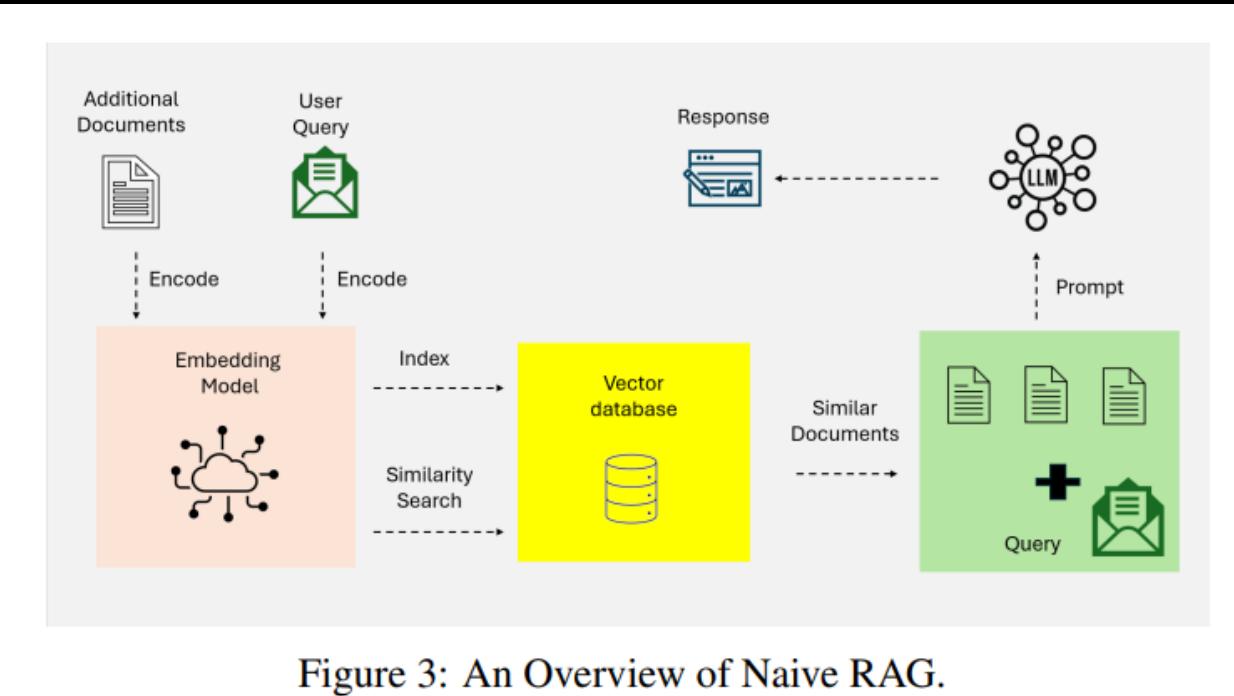
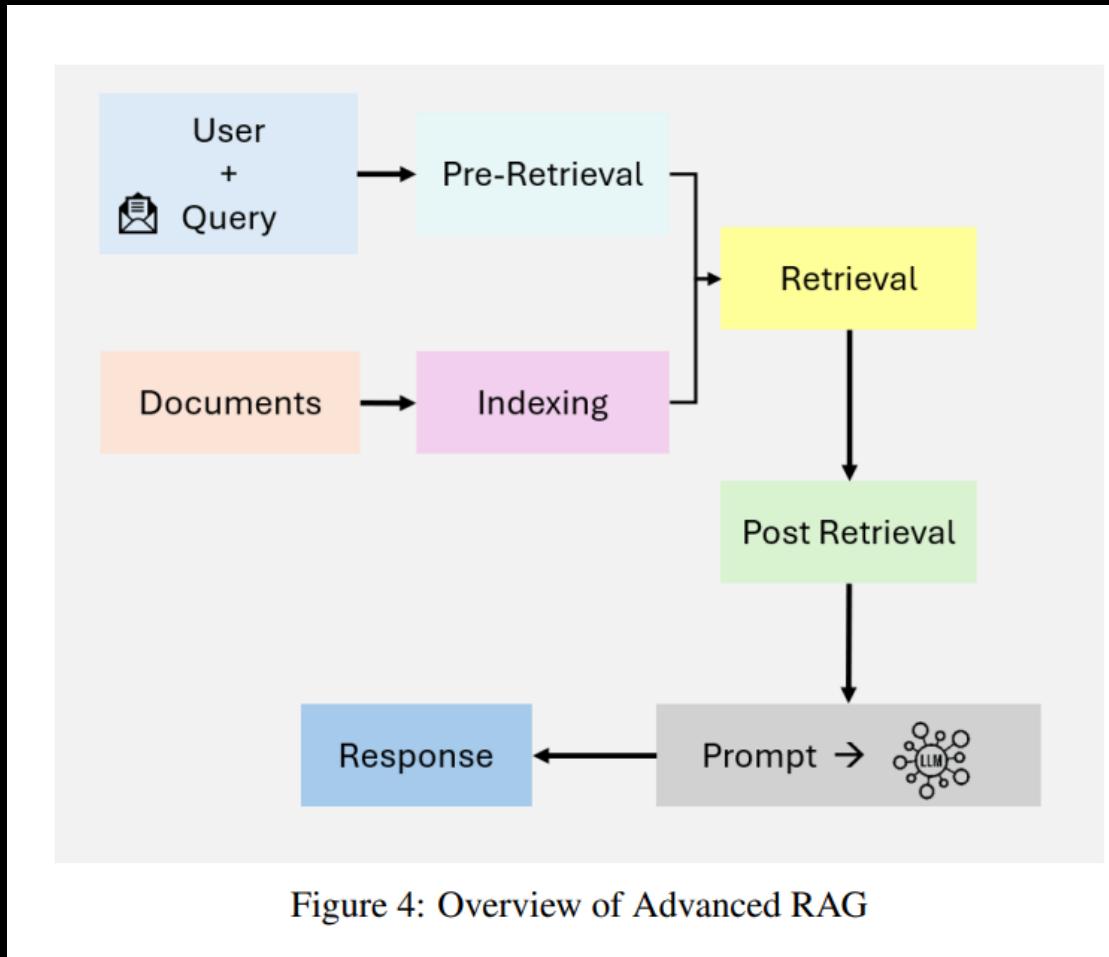


Figure 3: An Overview of Naïve RAG.

Limitations:

1. Lack of Contextual Awareness:
Limitation of Keyword based search
2. Fragmented Outputs
3. Scalability Issues

Advanced RAG



Key features:

1. Dense Vector Search: Semantic Search
2. Contextual Re-Ranking: Prioritize docs
3. Iterative Retrieval: Multi-hop retrieval

Challenges:

Computational overhead and limited scalability, particularly when dealing with large datasets or multi-step-queries

Modular RAG

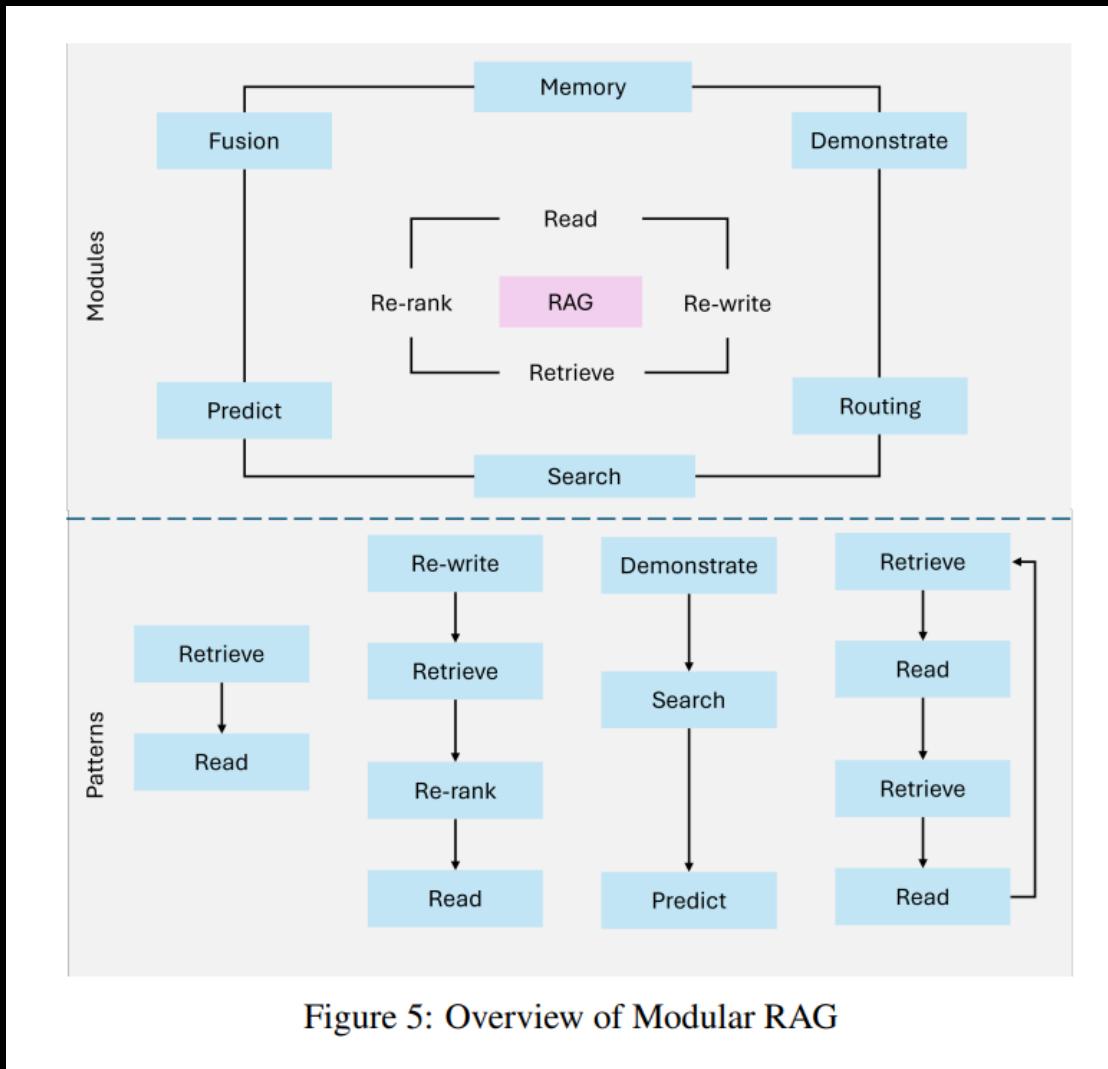


Figure 5: Overview of Modular RAG

Decompose the retrieval and generation pipeline into independent, reusable components.

Key features:

1. Hybrid Retrieval Strategies:
Combine Sparse +Dense Retrieval
2. Tool Integration: Incorporate external APIs, databases or computational tools
3. Composable Pipelines: Enables retrievers, generators, and other components to be replaced

Modular RAG

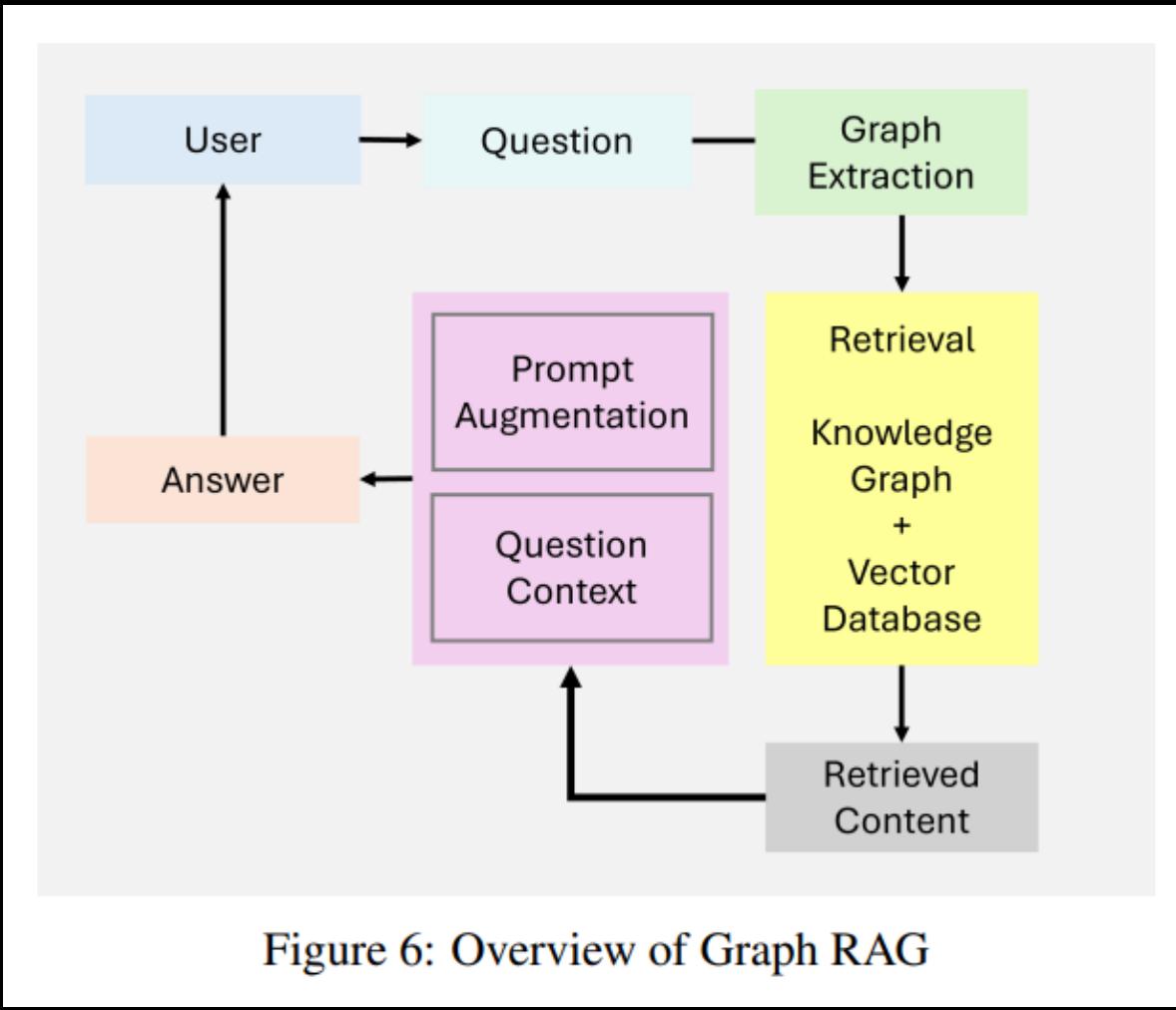
Strategies		
ver	Strategy	Description
V1	Dense-only	본문 Semantic Search
V2	Hybrid (Dense + Sparse)	본문 Lexical + Semantic Score 가중합
V3	Hybrid + RRF	본문 및 요약 Hybrid Score + RRF
V4	🚧 V3 + Reranker	Top-10 Retrieval 결과 Rerank

Results



ver	rrf k	w_{lex}	Recall@3 (FP32)	Recall@3 (Q8)
V1	-	-	76.90	62.30
V2	-	0.3	78.74	69.54
V2	-	0.5	77.36	70.57
V2	-	0.7	74.25	68.05
V3	40	0.3	77.36	68.85
V3	40	0.5	78.51	74.02
V3	40	0.7	77.01	76.09
V3	20	0.7	77.36	77.13
V3	10	0.7	77.82	77.82
V4	40	0.5	73.93	-

Graph RAG



Key features:

1. Node Connectivity: Captures over relationships between entities
2. Hierarchical Knowledge Management
3. Context Enrichment:
Adds relational understanding

Limitations:

1. Limited Scalability
2. Data Dependency:
Difficulties with preprocessing
3. Complexity of Integration:
Integrating with unstructured retrieval increases complexity

Agentic RAG

Key features:

1. Autonomous Decision-Making:
Agents manage retrieval strategies based on query complexity
2. Iterative Refinement:
Incorporates feedback loops
3. Workflow Optimization:
Dynamically orchestrates tasks

Challenges:

1. Coordination Complexity
2. Computational Overhead: The use of multiple agents increases resource requirements
3. Scalability Limitations:
Scaling is resource intensive.

Table 1: Comparative Analysis of RAG Paradigms

Paradigm	Key Features	Strengths
Naïve RAG	<ul style="list-style-type: none"> Keyword-based retrieval (e.g., TF-IDF, BM25) 	<ul style="list-style-type: none"> Simple and easy to implement Suitable for fact-based queries
Advanced RAG	<ul style="list-style-type: none"> Dense retrieval models (e.g., DPR) Neural ranking and re-ranking Multi-hop retrieval 	<ul style="list-style-type: none"> High precision retrieval Improved contextual relevance
Modular RAG	<ul style="list-style-type: none"> Hybrid retrieval (sparse and dense) Tool and API integration Composable, domain-specific pipelines 	<ul style="list-style-type: none"> High flexibility and customization Suitable for diverse applications Scalable
Graph RAG	<ul style="list-style-type: none"> Integration of graph-based structures Multi-hop reasoning Contextual enrichment via nodes 	<ul style="list-style-type: none"> Relational reasoning capabilities Mitigates hallucinations Ideal for structured data tasks
Agentic RAG	<ul style="list-style-type: none"> Autonomous agents Dynamic decision-making Iterative refinement and workflow optimization 	<ul style="list-style-type: none"> Adaptable to real-time changes Scalable for multi-domain tasks High accuracy

Core Principles and Background of
Agentic Intelligence

Components of an AI Agent

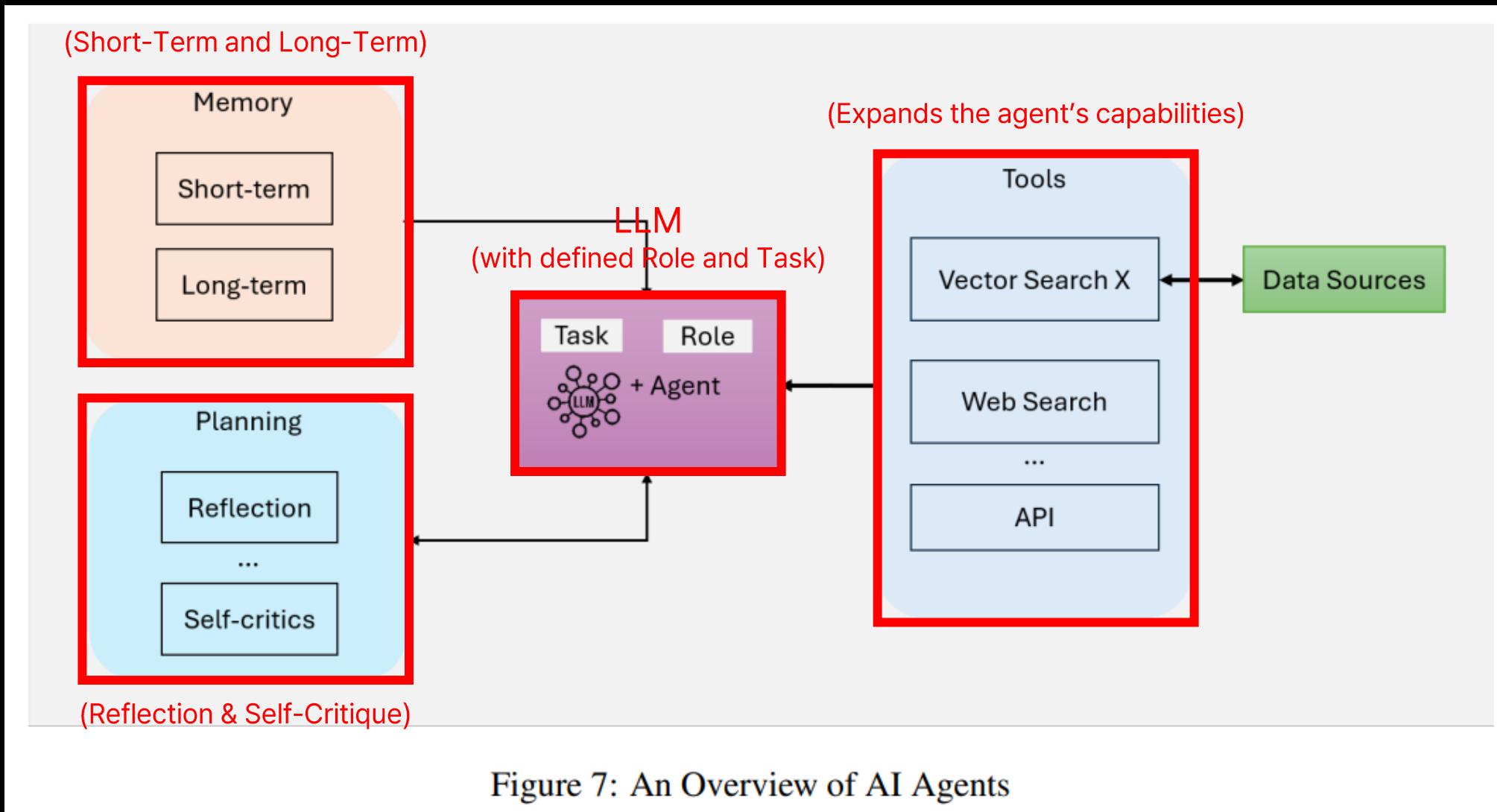
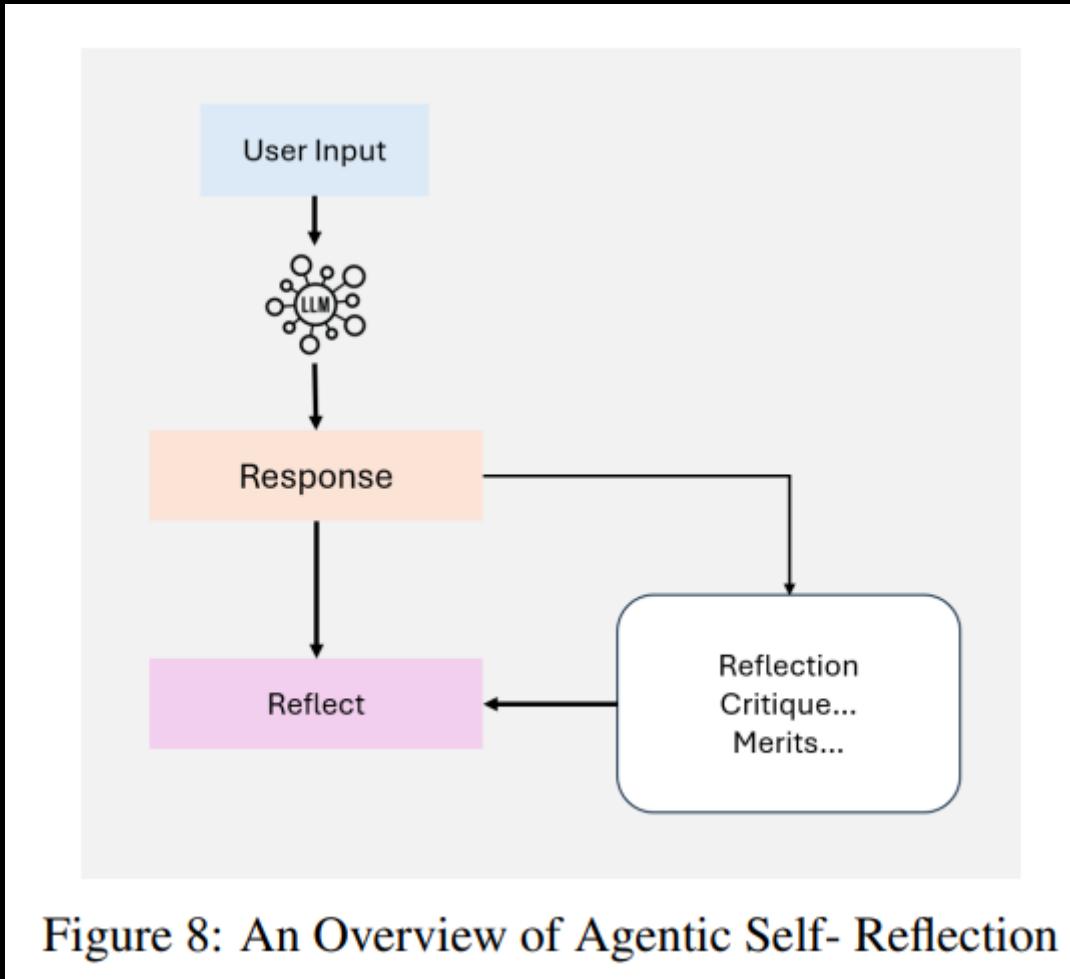


Figure 7: An Overview of AI Agents

Reflection



Reflection involves prompting an agent to critique its outputs for **correctness, style, and efficiency**, then incorporating this feedback into subsequent iterations.

Planning & Tool Use

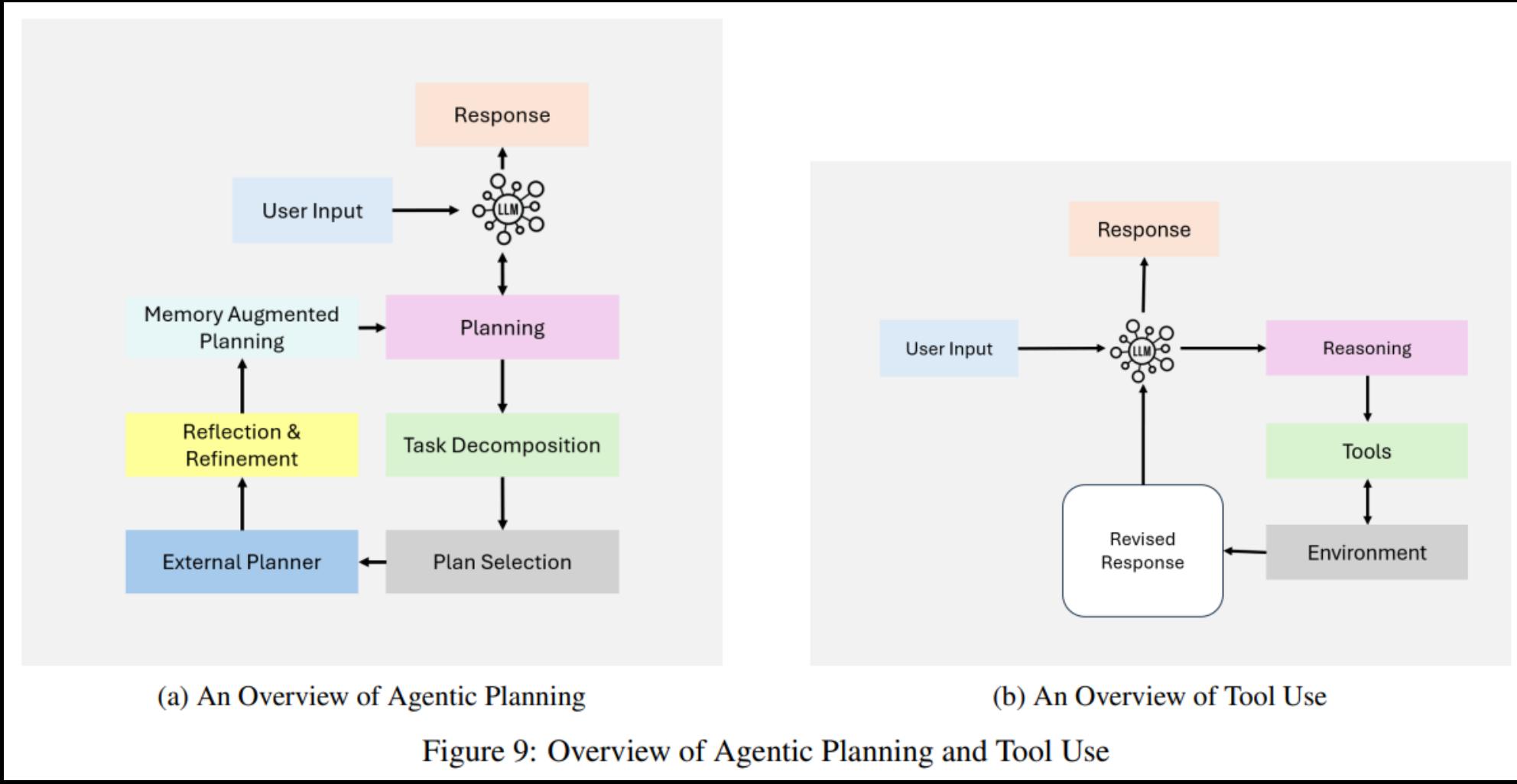


Figure 9: Overview of Agentic Planning and Tool Use

Multi-Agent RAG

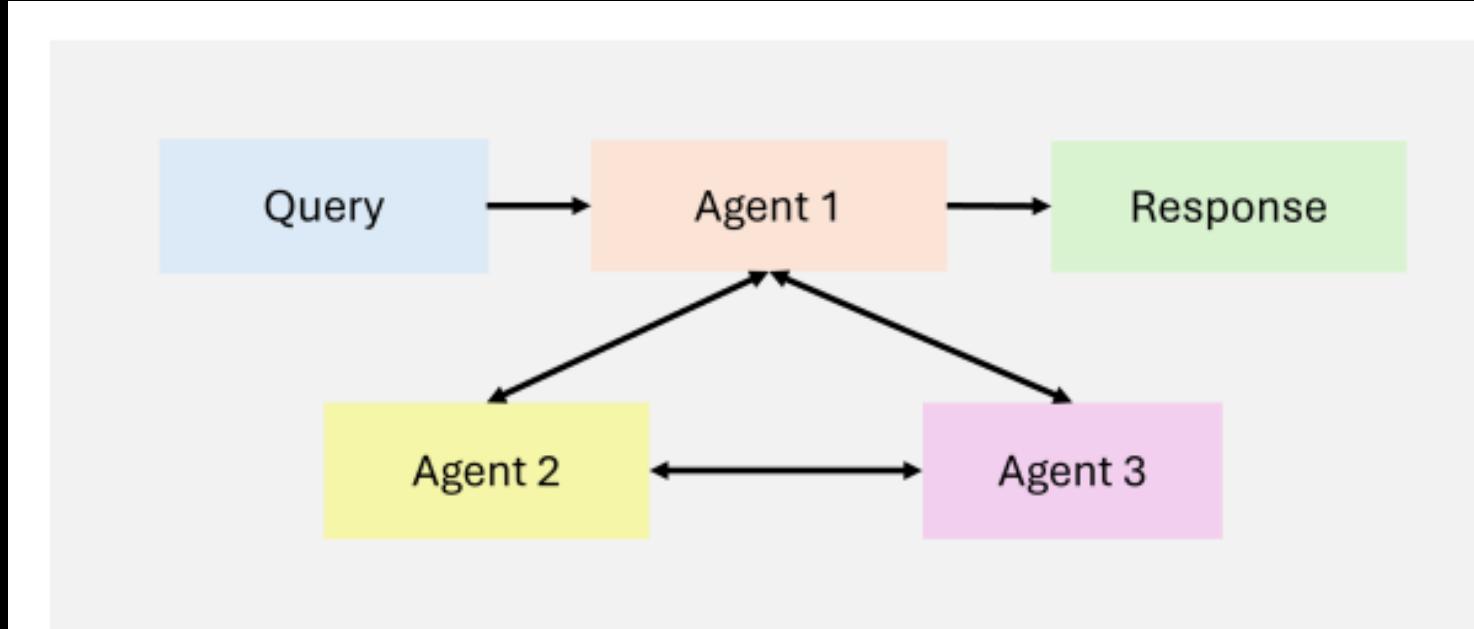


Figure 10: An Overview of MultiAgent

Agentic Workflow Patterns:
**Adaptive Strategies for
Dynamic Collaboration**

Prompt Chaining

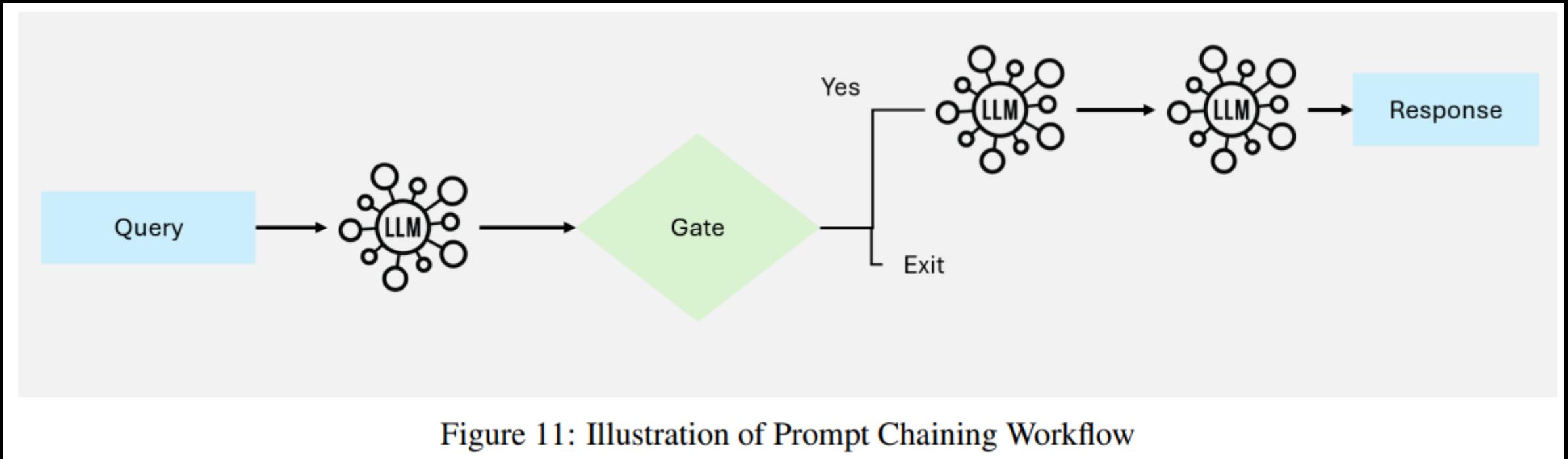
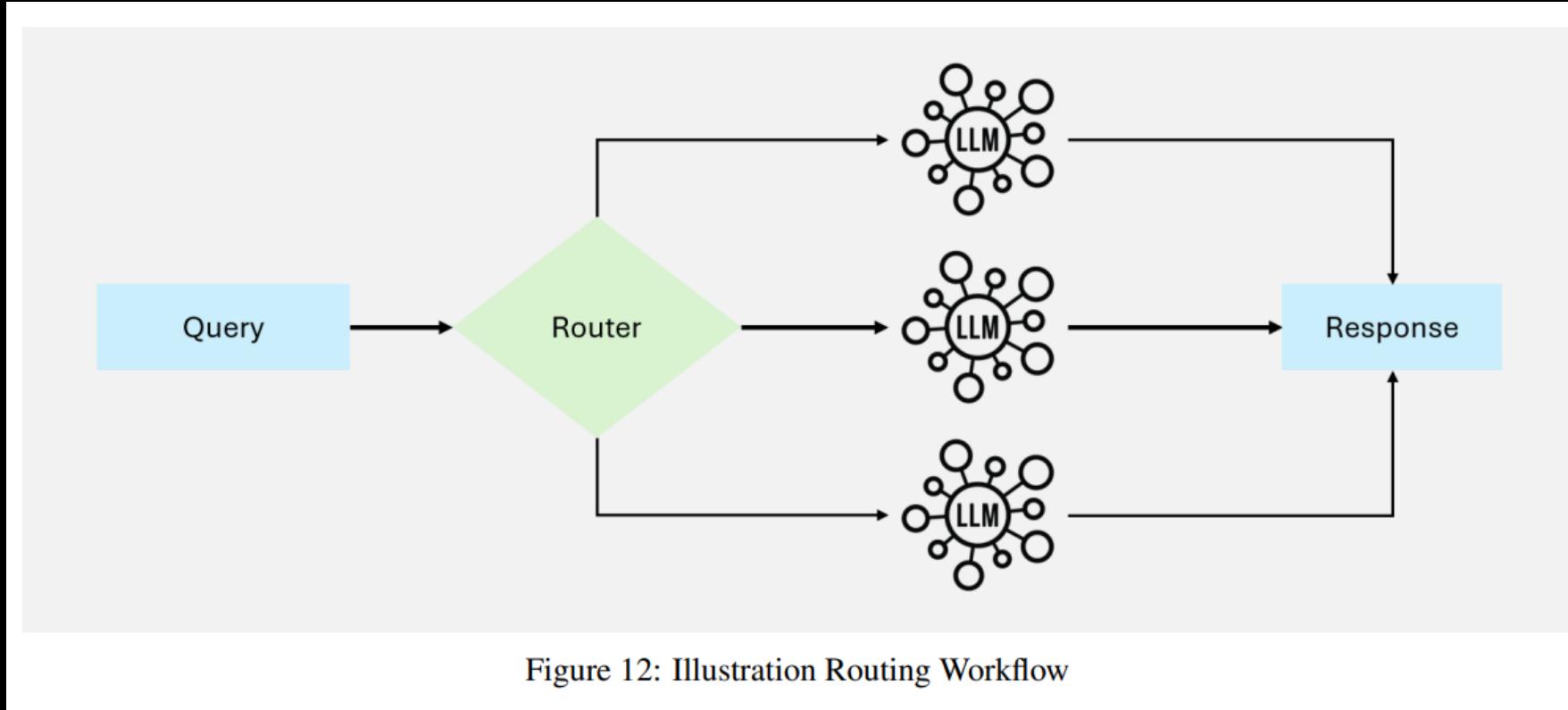


Figure 11: Illustration of Prompt Chaining Workflow

Example Applications:

1. Generating content and translating it while preserving nuances.
2. First, create an outline, then verify its completeness, and finally write the full text.

Routing: Directing Inputs to Specialized Process



Example Applications:

1. Directing customer service queries into categories such as technical support, refund requests, or general inquiries.
2. Assigning simple queries to smaller models for cost efficiency, while complex requests go to advanced models

Parallelization

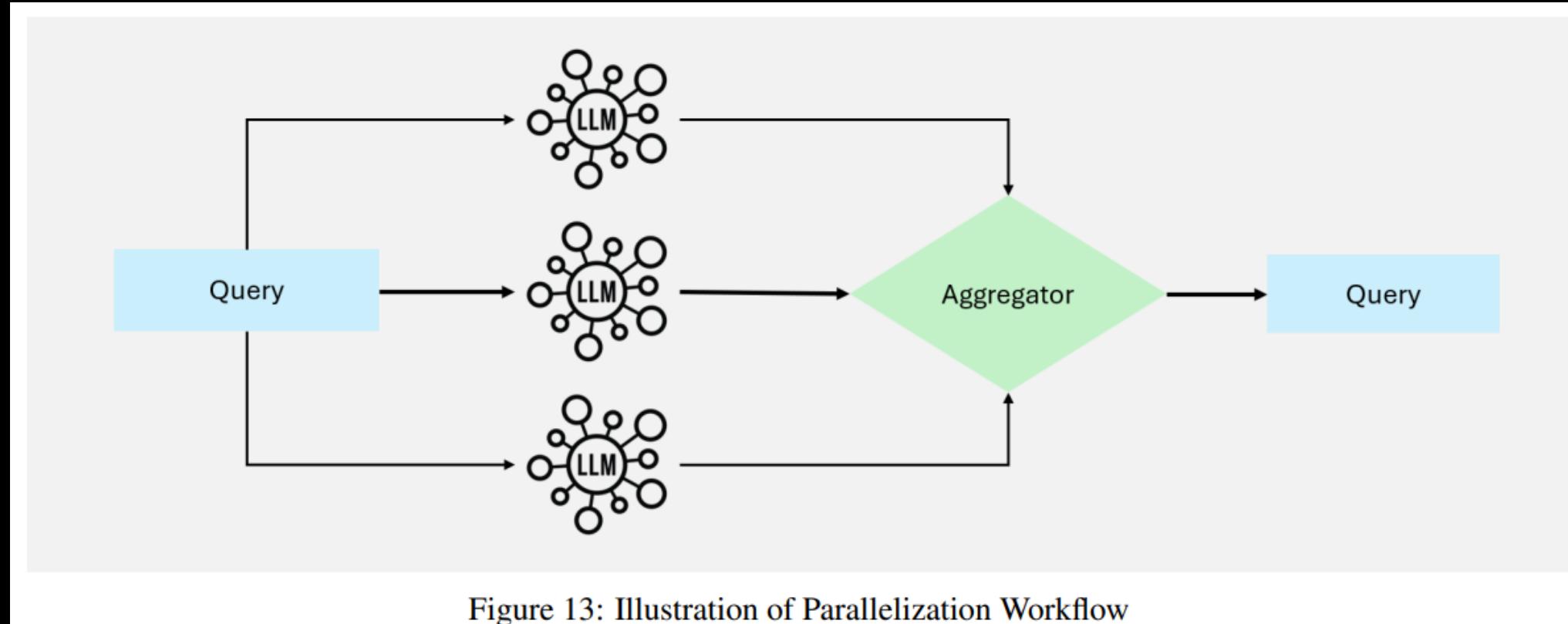


Figure 13: Illustration of Parallelization Workflow

Example Applications:

1. Sectioning: Splitting tasks like content moderation, where one model screens input while another generates a response
2. Voting: Using multiple models to cross-check code for vulnerabilities or analyze content moderation decisions

Orchestrator-Workers: Dynamic Task Delegation

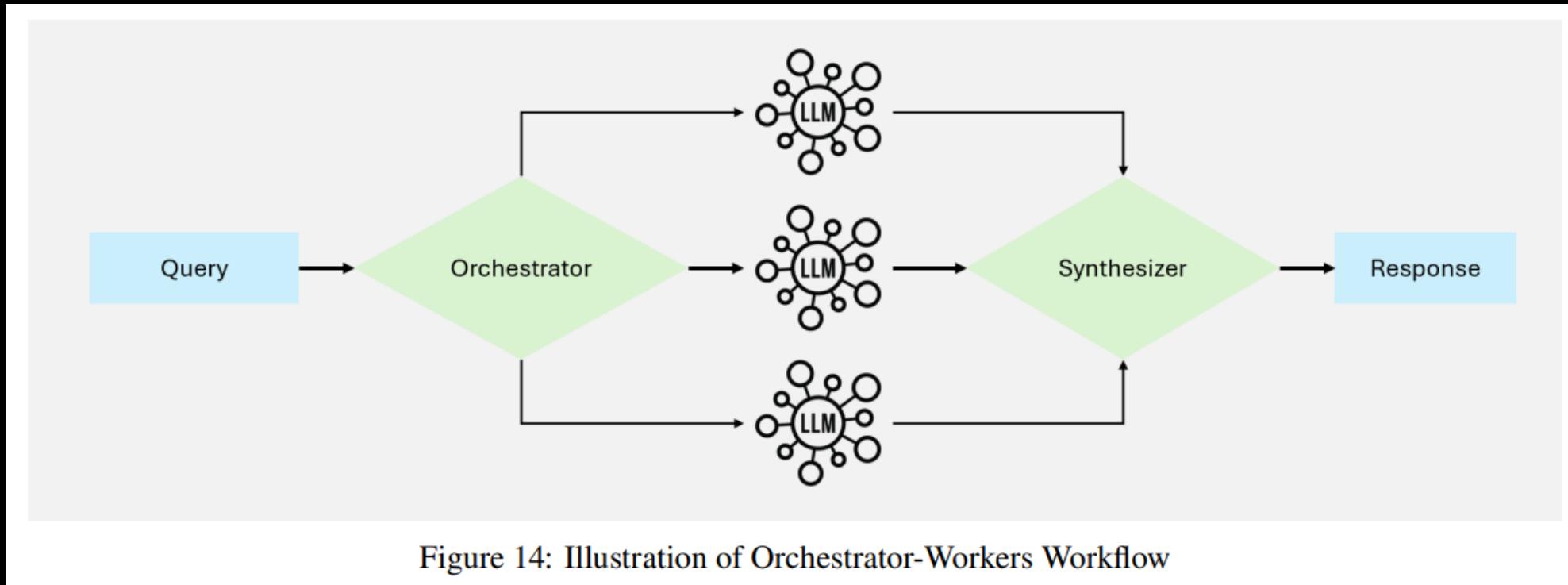
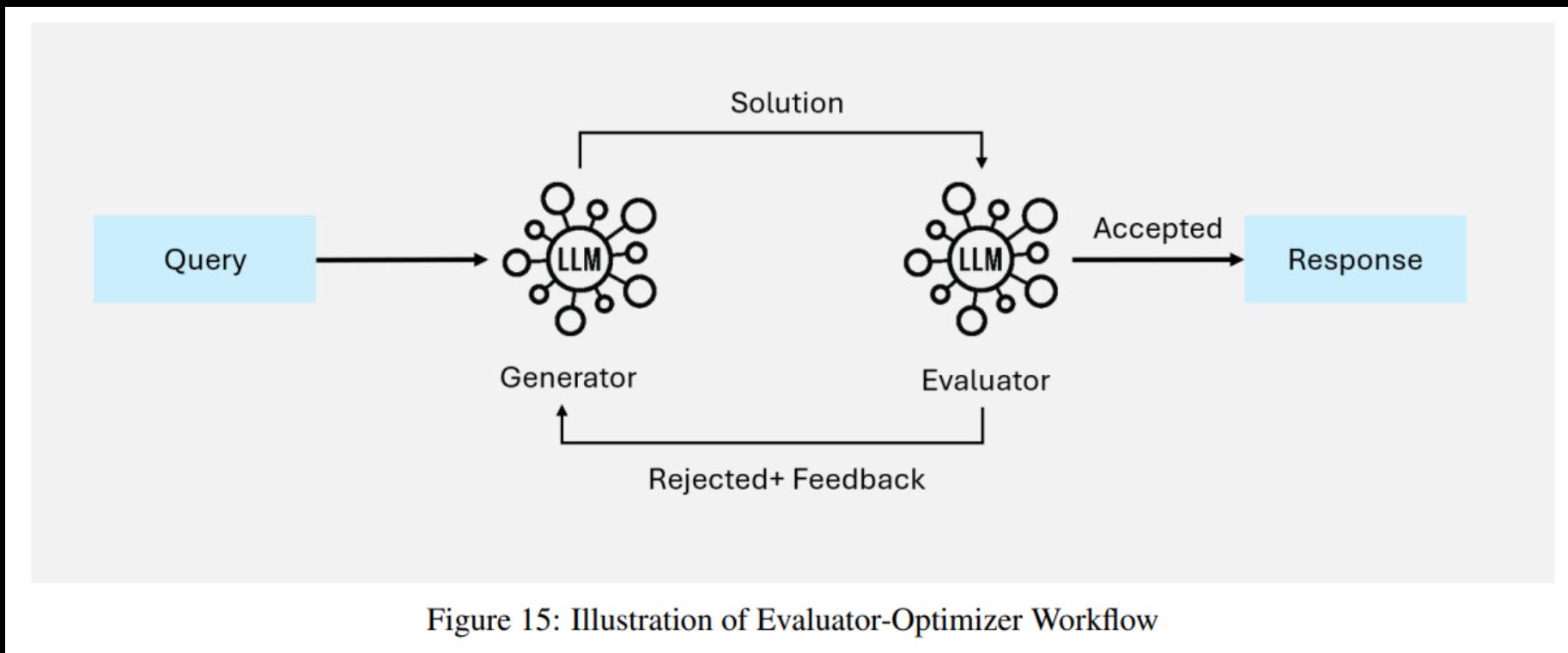


Figure 14: Illustration of Orchestrator-Workers Workflow

Example Applications:

1. Automatically modifying multiple files in a codebase based on the nature of requested changes.
2. Conducting real-time research by gathering and synthesizing relevant information from multiple sources.

Evaluator-Optimizer: Refining Through Iteration



Example Applications:

1. Automatically modifying multiple files in a codebase based on the nature of requested changes.
2. Conducting real-time research by gathering and synthesizing relevant information from multiple sources.

팀원 소개



강민석

myeolinmalchi

프로젝트 및 개발 총괄
텍스트 임베딩 서버



박상훈

sanghunii

서버 엔드포인트 구현
데이터 수집 보조



박준혁

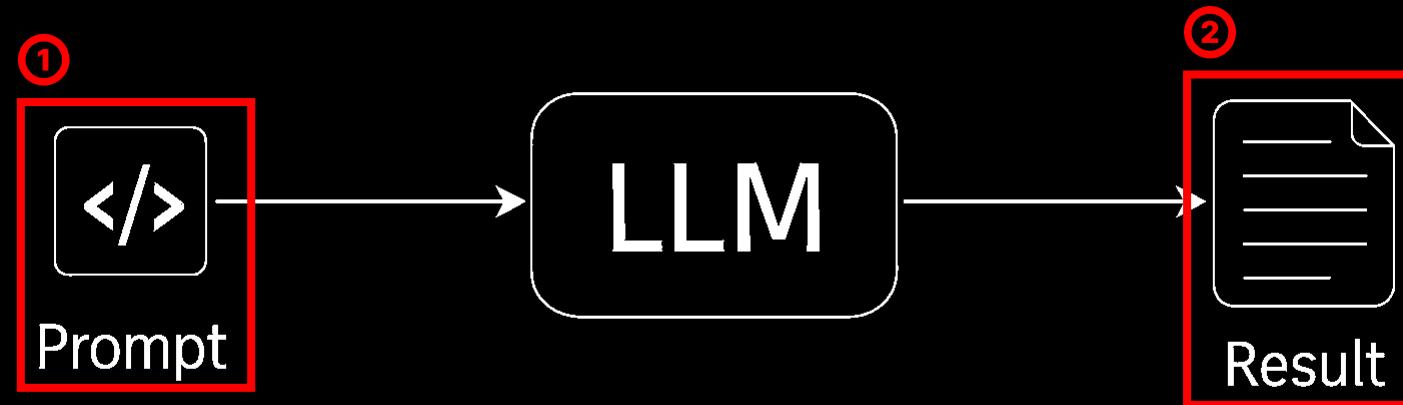
JakeFRCSE

첨부파일 파싱 서버
데이터 수집 보조

들어가기에 앞서:
What is RAG?

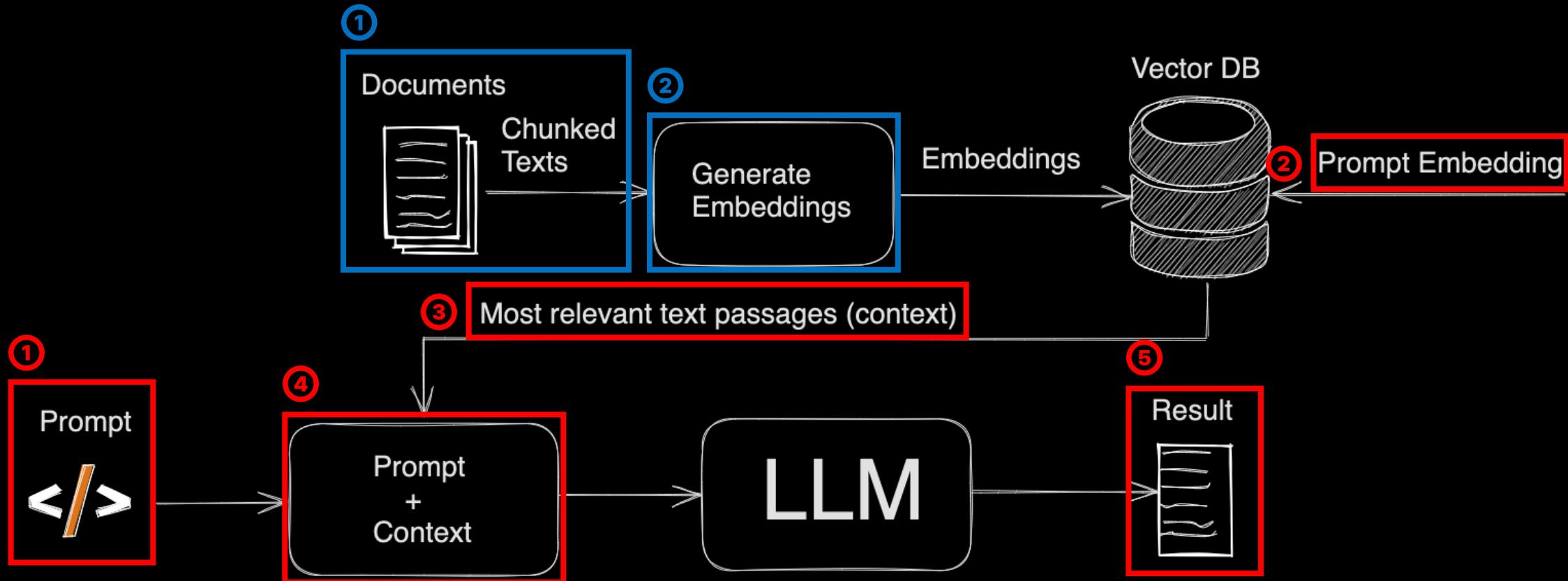
기본 QA 챗봇:

외부 정보 없이 LLM이 바로 답변



RAG 챗봇:

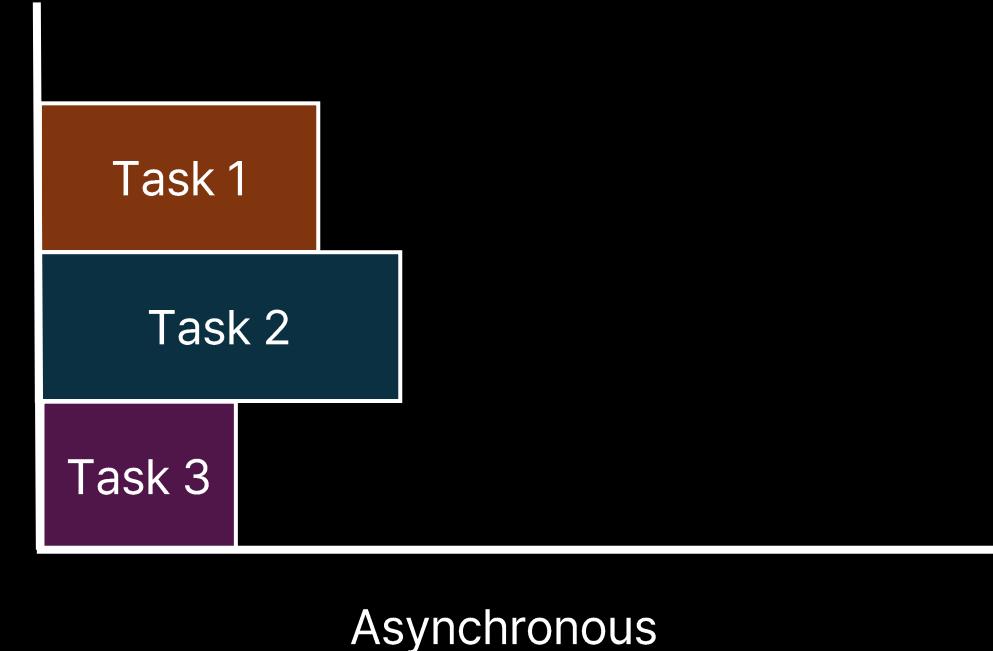
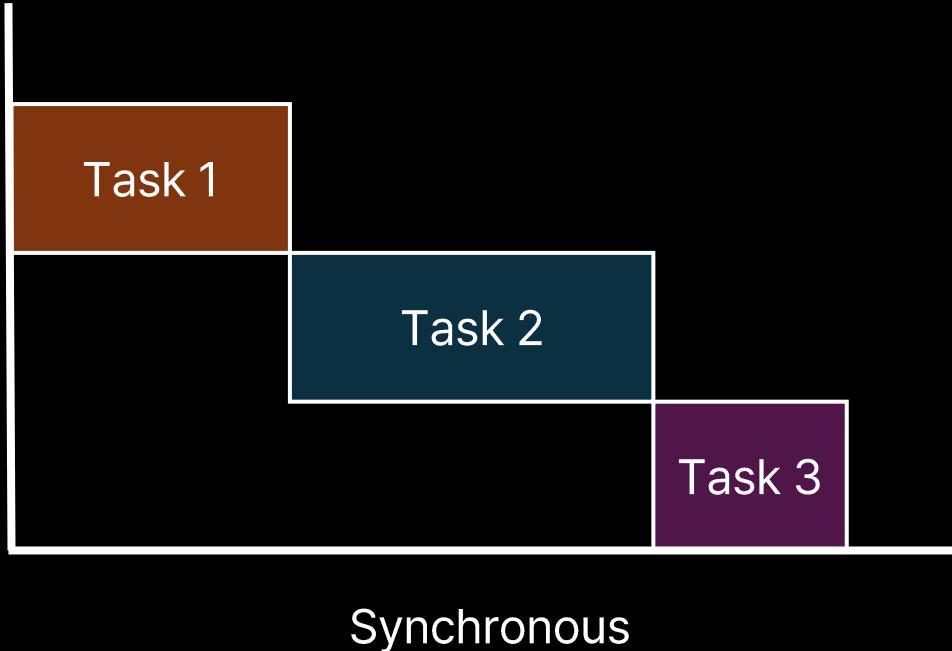
질문과 관련된 외부 문서를 참고하여 답변



1. 데이터 크롤링 및 전처리
2. 텍스트 임베딩 후 VectorDB 저장
3. 데이터 증강을 위한 파이프라인 설계
4. 답변 도출을 위한 Prompt Engineering

데이터 크롤링:
학교 서버야 뻔지 말아다오....

동기? 비동기?



시도 1: 될 때까지 재시도

```
def retry_async(
    times: int = 10,           # 최대 재시도 횟수
    delay: float = 5.0,        # 재시도 딜레이
    is_success=lambda _: True, # 성공 여부를 판단하는 람다 함수
):

    def decorator(func):

        @wraps(func)
        async def wrapped(*args, **kwargs):
            for _ in range(times):
                try:
                    result = await func(*args, **kwargs)
                    if not is_success(result):
                        continue
                    if isinstance(result, BaseException):
                        raise result
                    return result

                except Exception as e:
                    logger(f"요청을 재시도 합니다: {e}", level=logging.WARNING)
                    await asyncio.sleep(delay)

            raise TimeoutError(f'{times}번의 재시도에 실패했습니다. ')

        return wrapped

    return decorator
```

시도 1: 될 때까지 재시도

```
async def scrape_async(
    url: str | List[str],
    session: aiohttp.ClientSession,
    post_process: Optional[Callable[[BeautifulSoup], T]] = None,
    retry_delay: float = 5.0,
    retry_times: int = 10,
    delay_range: Tuple[float, float] = (0, 1)
) → T | List[T]:

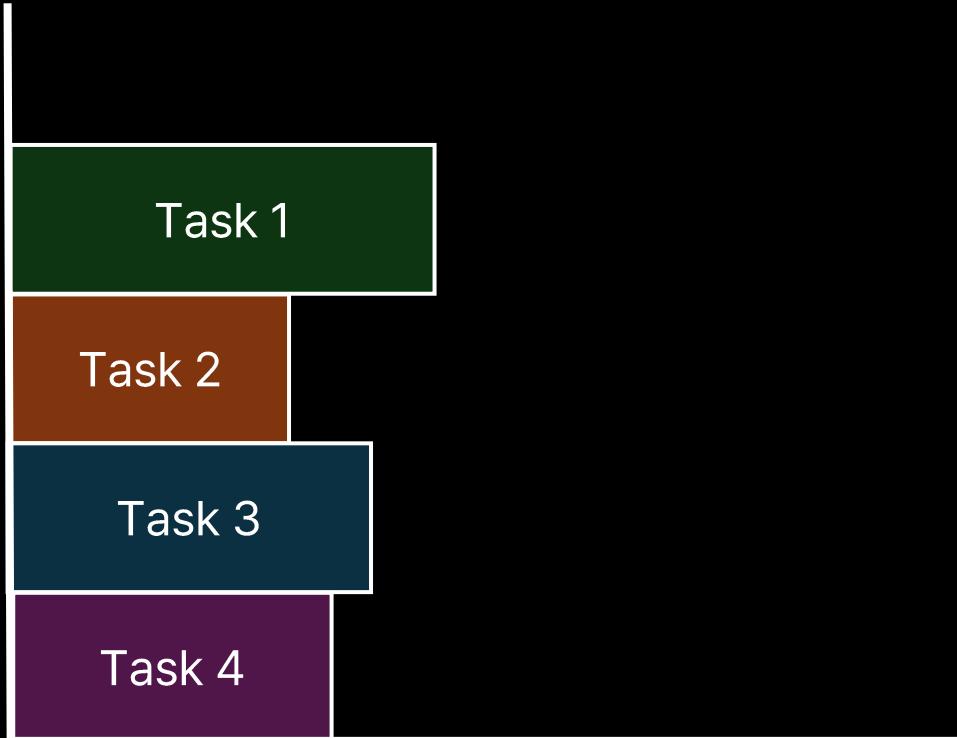
    @retry_async(delay=retry_delay, times=retry_times)
    async def help(_url: str) → Any:
        await asyncio.sleep(random.uniform(*delay_range))
        async with session.get(_url) as res:
            if res.ok:
                html = await res.text(errors="ignore")
                soup = BeautifulSoup(html, "html5lib")
                result = post_process(soup) if post_process else soup
                return result

            raise aiohttp.ClientError

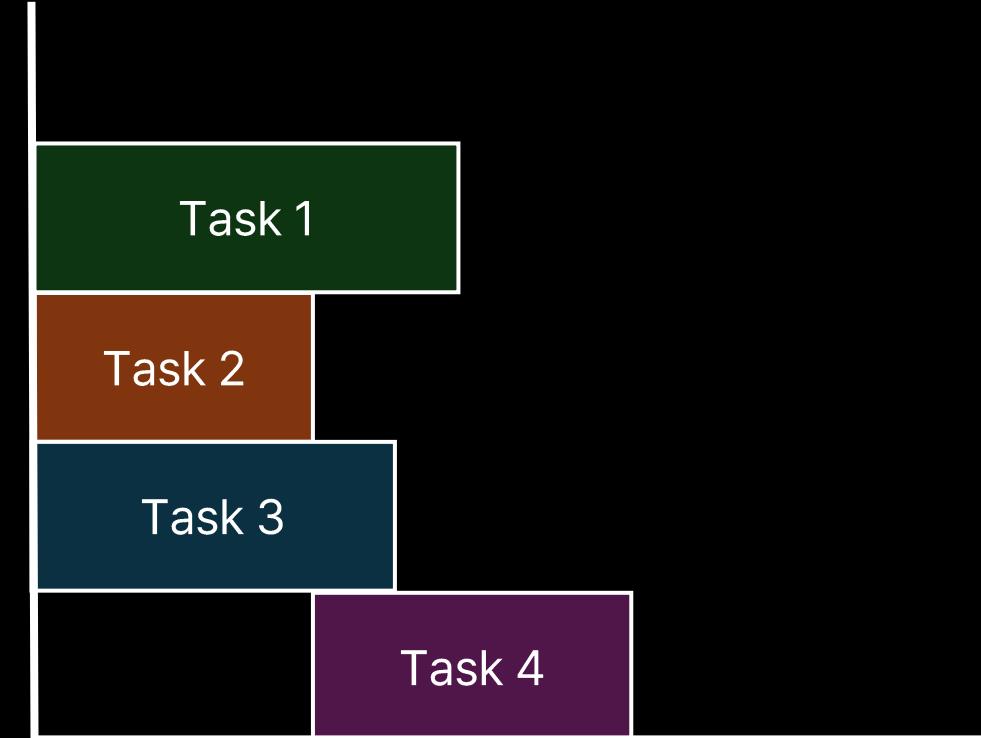
    if isinstance(url, str):
        return await help(url)

    return await asyncio.gather(*[help(url) for url in url])
```

시도 2: 최대 요청 수 제한하기



Asynchronous



Asynchronous(with semaphore)

시도 2: 최대 요청 수 제한하기

```
❶ @asynccontextmanager
❷     async def semaphore(limit: int = 10):
❸         global _Semaphore
❹
❺         if _Semaphore is None:
❻             _Semaphore = Semaphore(limit)
❽
❼         semaphore = _Semaphore
⫽
⫾         try:
⫿             async with semaphore:
⫻                 yield
⫽
⫾         except Exception as e:
⫿             logger(f"비동기 작업 중 오류가 발생했습니다. ({e})", level=logging.ERROR)
⫻             raise e
```

데이터 전처리:
기괴한 HTML 구조

[효원상담원] 찾아가는 심리검사 프로그램(집단심리검사) 신청 안내

1. 관련: 효원상담원-1458(2025.03.06.)
 2. 학생들의 학과에 대한 소속감 향상 및 자신과 타인에 대해 이해 증진을 위해 찾아가는 심리검사 프로그램(집단심리검사)을 운영하고 있으니, 많은 관심과 참여 부탁드립니다.
- 가. 목적: 학생들의 학과에 대한 소속감 향상 및 자기와 타인에 대한 이해 증진에 도움
캠퍼스 간 심리지원서비스 수혜 격차 완화
- 나. 신청기간: 2025년 3월 ~ 2026년 2월 상시 진행
- 다. 신청방법: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)
- 신청 최소 인원 : 10명 이상
 - 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청
 - 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호
 - 신청 메일 내용 : (1)신청 명단 학번 및 이름, (2)신청 일자, 장소
- 마. 신청 가능 검사

검사 종류	검사 내용	소요시간
마인드핏(적응역량검사)	대학생활 적응도 및 정신건강 수준, 스트레스 대응 능력 및 영역별 스트레스 확인(해석 특강 진행)	약 1시간
MBTI(성격유형검사)	자신의 성격 유형 및 타인과의 차이 이해, 공동체 유대감 증진 (해석 특강 및 소그룹 활동 진행)	약 2시간
TCI(기질 및 성격검사)	자신의 기질 및 성격 특성 이해, 공동체 유대감 증진 (해석 특강 및 소그룹 활동 진행)	약 2시간

바. 진행절차

- 1) 희망 학과 조교가 효원상담원 홈페이지에 접속하여 온라인 신청
- 2) 매달 말 신청내역 확인 후 선정된 학과 조교에게 개별 안내(밀양, 양산 소재 학과 우선 선정)
- 3) 선정된 학과 조교가 장소 및 참가자 확정
- 4) 참가하는 학생은 학생역량지원시스템 내 비교과 프로그램 신청
- 5) 효원상담원에서 강사를 파견하여 프로그램 진행

사. 담당 및 문의 : 김예주(☎ 3708)

난해한 구조의 Raw Data

```
<div class="artclView">
  <p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; letter-spacing: 0pt; font-size: 10pt;">…</p>
  <p style="line-height: 160%; margin: 5pt 5pt 0pt 16pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; letter-spacing: 0pt; font-size: 10pt; text-indent: -16pt;">…</p>
  <p style="line-height: 160%; margin: 5pt 5pt 0pt 80.3pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; letter-spacing: 0pt; font-size: 10pt; text-indent: -80.3pt;">…</p>
  <p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; letter-spacing: 0pt; font-size: 10pt;">…</p>
  <p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; letter-spacing: 0pt; font-size: 10pt;">…</p>
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;"> = $0
  <span style="letter-spacing: 0pt; font-family: 돋움; font-size: 11pt;">&ampnbsp&ampnbsp</span>
  <span style="letter-spacing: 0pt; font-family: 돋움; font-size: 11pt;">다</span>
  <span style="letter-spacing: 0pt; font-family: 돋움; font-size: 11pt;">.&ampnbsp</span>
  <span style="letter-spacing: 0pt; text-indent: 0pt; font-family: 돋움; font-size: 11pt;">신청방법</span>
  <span style="letter-spacing: 0pt; text-indent: 0pt; font-family: 돋움; font-size: 11pt;">: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)&ampnbsp</span>
</p>
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;">…</p>
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;">
  <span style="letter-spacing: 0pt; text-indent: 0pt; font-family: 돋움; font-size: 11pt;">&ampnbsp &ampnbsp - 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청&ampnbsp</span>
</p>
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;">
  <span style="font-family: 돋움; font-size: 11pt; letter-spacing: 0pt; text-indent: 0pt;">&ampnbsp &ampnbsp - 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호&ampnbsp&ampnbsp</span>
</p>
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;">
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;">
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;">
<div class="con-table on" style="overflow: hidden; outline: none;" tabindex="0">…</div>
<p>&ampnbsp</p>
<p style="line-height: 160%; margin: 0pt 5pt 0pt 0pt; text-indent: 0pt; text-align: justify; vertical-align: baseline; font-family: 한양신명조; font-size: 10pt;">
  <span style="font-family:돋움;font-width:95%;font-size:11.0pt;">&ampnbsp&ampnbsp</span>
  <span style="font-family:돋움;font-family:돋움;font-width:95%;font-size:11.0pt;">바</span>
  <span style="font-family:돋움;font-family:돋움;font-family:돋움;font-width:95%;letter-spacing:0pt;text-raise:0pt;font-size:11.0pt;"> 진행절차</span>
</p>
```

1. 태그 속성 대부분이 의미상 불필요

2. 태그간 과도한 중첩 및 분리

→ RAG 품질 저하

난해한 구조의 Raw Data

```
ALLOWED_ATTRS = ["colspan", "rowspan", "scope", "headers"]

def _clean_html_tag(soup: Tag | BeautifulSoup, element: Tag) → int:
    ...
    for attr in [*element.attrs.keys()]:
        if attr not in ALLOWED_ATTRS:
            del element[attr]
    ...

    match element:
        ...
        case Tag(
            name="a" | "span" | "p" | "u" | "b" | "strong",
            parent=Tag(name="span" | "p" | "li" | "td" | "th" | "b" | "u"),
        ):
            if element.name == "a":
                element.extract()
                return affected

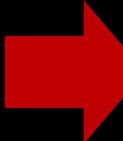
            if not only_string:
                return affected

            if element.next_sibling is not None and element.name == "p":
                element.append(NavigableString(" "))

            element.unwrap()
            return affected + 1
    ...

```

Clean logic



<div>

<p>1.관련:효원상담원-1458(2025.03.06.)</p>

<p>2.학생들의 학과에 대한 소속감 향상 및 자신과 타인에 대한 이해 증진을 위해...</p>

<p>가.목 적:학생들의 학과에 대한 소속감 향상 및 자기와 타인에 대한 이해 증진에 도움</p>

<p>캠퍼스 간 심리지원서비스 수혜 격차 완화</p>

<p>나.신청기간: 2025년3월~ 2026년2월 상시 진행</p>

<p>다.신청방법: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)</p>

<p>- 신청 최소 인원 : 10명 이상</p>

<p>- 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청</p>

<p>- 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호</p>

<p>- 신청 메일 내용 : (1)신청 명단 학번 및 이름, (2)신청 일자, 장소</p>

<p>마.신청 가능 검사</p>

검사 종류	검사 내용	소요시간
마인드핏(적응역량검사)	대학생활 적응도 및 정신건강 수준, 스트레스 대응 능력 및 영역별 스트레스 확인(해석 특강 진행)	약1시간
MBTI(성격유형검사)	자신의 성격 유형 및 타인과의 차이 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)	약2시간
TCI(기질 및 성격검사)	자신의 기질 및 성격 특성 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)	약2시간

</div>

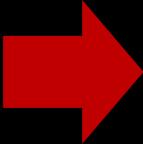
꼭 HTML이어야 하나?

```
<div>
  <p>1.관련:효원상담원-1458(2025.03.06.)</p>
  <p>2.학생들의 학과에 대한 소속감 향상 및 자신과 타인에 대해 이해 증진을 위해...</p>
  <p>가.목 적:학생들의 학과에 대한 소속감 향상 및 자기와 타인에 대한 이해 증진에 도움</p>
  <p>캠퍼스 간 심리지원서비스 수혜 격차 완화</p>
  <p>나.신청기간: 2025년3월~ 2026년2월 상시 진행</p>
  <p>다.신청방법: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)</p>
  <p> - 신청 최소 인원 : 10명 이상</p>
  <p> - 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청</p>
  <p> - 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호</p>
  <p> - 신청 메일 내용 : (1)신청 명단 학번 및 이름, (2)신청 일자, 장소</p>
  <p>마.신청 가능 검사</p>
  <table>
    <tbody>
      <tr>
        <td>검사 종류</td>
        <td>검사 내용</td>
        <td>소요시간</td>
      </tr>
      <tr>
        <td>마인드핏(적응역량검사)</td>
        <td>대학생활 적응도 및 정신건강 수준, 스트레스 대응 능력 및 영역별 스트레스 확인(해석 특강 진행)</td>
        <td>약1시간</td>
      </tr>
      <tr>
        <td>MBTI(성격유형검사)</td>
        <td>자신의 성격 유형 및 타인과의 차이 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)</td>
        <td>약2시간</td>
      </tr>
      <tr>
```

꼭 HTML이어야 하나?

```
<div>
<p>1.관련:효원상담원-1458(2025.03.06.)</p>
<p>2.학생들의 학과에 대한 소속감 향상 및 자신과 타인에 대한 이해 증진을 위해...</p>
<p>가.목 적:학생들의 학과에 대한 소속감 향상 및 자기와 타인에 대한 이해 증진에 도움</p>
<p>캠퍼스 간 심리지원서비스 수혜 격차 완화</p>
<p>나.신청기간: 2025년3월~ 2026년2월 상시 진행</p>
<p>다.신청방법: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)</p>
<p> - 신청 최소 인원 : 10명 이상</p>
<p> - 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청</p>
<p> - 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호</p>
<p> - 신청 메일 내용 : (1)신청 명단 학번 및 이름, (2)신청 일자, 장소</p>
<p>마.신청 가능 검사</p>
<table>
<tbody>
<tr>
<td>검사 종류</td>
<td>검사 내용</td>
<td>소요시간</td>
</tr>
<tr>
<td>마인드핏(적응역량검사)</td>
<td>대학생활 적응도 및 정신건강 수준, 스트레스 대응 능력 및 영역별 스트레스 확인(해석 특강 진행)</td>
<td>약1시간</td>
</tr>
<tr>
<td>MBTI(성격유형검사)</td>
<td>자신의 성격 유형 및 타인과의 차이 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)</td>
<td>약2시간</td>
</tr>
<tr>
<td>TCI(기질 및 성격검사)</td>
<td>자신의 기질 및 성격 특성 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)</td>
<td>약2시간</td>
</tr>
</tbody>
</table>
<p>바.진행절차</p>
<p>1)희망 학과 조교가 효원상담원 홈페이지에 접속하여 온라인 신청</p>
<p>2)매달 말 신청내역 확인 후 선정된 학과 조교에게 개별 안내(밀양, 양산 소재 학과 우선 선정)</p>
<p>3)선정된 학과 조교가 장소 및 참가자 확정</p>
<p>4)참가하는 학생은 학생역량지원시스템 내 비교과 프로그램 신청</p>
<p>5)효원상담원에서 강사를 파견하여 프로그램 진행</p>
<p>사.담당 및 문의:김예주(☎3708)</p>
</div>
```

markdownify



```
1.관련:효원상담원-1458(2025.03.06.)

2.학생들의 학과에 대한 소속감 향상 및 자신과 타인에 대한 이해 증진을 위해 찾아가는 심리검사 프로그램(집단심리검사)을 운영하고 있으니, 많은 관심과 참여 부탁드립니다.

가.목 적:학생들의 학과에 대한 소속감 향상 및 자기와 타인에 대한 이해 증진에 도움

캠퍼스 간 심리지원서비스 수혜 격차 완화

나.신청기간: 2025년3월~ 2026년2월 상시 진행

다.신청방법: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)

- 신청 최소 인원 : 10명 이상

- 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청

- 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호

- 신청 메일 내용 : (1)신청 명단 학번 및 이름, (2)신청 일자, 장소

마.신청 가능 검사

| 검사 종류 | 검사 내용 | 소요시간 |
| --- | --- | --- |
| 마인드핏(적응역량검사) | 대학생활 적응도 및 정신건강 수준, 스트레스 대응 능력 및 영역별 스트레스 확인(해석 특강 진행) | 약1시간 |
| MBTI(성격유형검사) | 자신의 성격 유형 및 타인과의 차이 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행) | 약2시간 |
| TCI(기질 및 성격검사) | 자신의 기질 및 성격 특성 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행) | 약2시간 |

바.진행절차

1)희망 학과 조교가 효원상담원 홈페이지에 접속하여 온라인 신청

2)매달 말 신청내역 확인 후 선정된 학과 조교에게 개별 안내(밀양, 양산 소재 학과 우선 선정)

3)선정된 학과 조교가 장소 및 참가자 확정

4)참가하는 학생은 학생역량지원시스템 내 비교과 프로그램 신청

5)효원상담원에서 강사를 파견하여 프로그램 진행

사.담당 및 문의:김예주(☎3708)
```

복잡한 테이블 구조

제원	장학명	장학금액	선발기준		선발시기	선발절차	지급기간	계속지급요건	
교내	Premier (특전) ?	등록금 전액 장려금 300만원	<ul style="list-style-type: none"> 해당 학년도 계열별 수능 지정영역 백분위 점수 충족 자 - 인문·사회계 : 국어와 수학(미적분, 기하, 확률과 통계) 영역의 합산 백분위 점수가 189점 이상 - 자연계: 국어와 수학(미적분, 기하) 영역의 합산 백분위 점수가 189점 이상 - 예술계 : 국어 영역 백분위 점수가 98점 이상이고 영어영역 1등급 - 체육계 : 국어 영역 백분위 점수가 98점 이상이고 영어영역 1등급 - 의·약학계* : 국어와 수학(미적분, 기하) 영역의 합산 백분위 점수가 196점 이상 중 모집단위별 모집인원의 3% 이내 고득점 순 선발 <p>* 의과대학 의예과, 치의학전문대학원 학·석사동합과정, 한의학전문대학원 학·석사동합과정, 약학대학 약학부 해당</p>		1~2월 1월/7월말	학생과 선발 학생과 선발	수업 연한 수업 연한	평점평균 3.5 ?	
			<ul style="list-style-type: none"> 해당학기 평점평균 3.5 이상 총족 시 해당학기 말 지급 * 학·석사동합과정 학생은 학사과정만 지급 						
교내	생명자원과학 대학특별	등록금 전액	A	<ul style="list-style-type: none"> 생명자원과학대학 학생 중 수능 전체영역의 평균등급이 2.2등급 이내 (탐구영역은 2개 과목 평균 등급으로 반영) 		1~2월	학생과 선발 학생과 선발	평점평균 3.5 ?	
		수업료 I 전액 + 수업료 II 반액	B	<ul style="list-style-type: none"> 생명자원과학대학 학생 중 수능 전체영역의 평균등급이 2.6등급 이내 (탐구영역은 2개 과목 평균 등급으로 반영) 					

복잡한 테이블 구조

```
class KeepTableConverter(MarkdownConverter):

    def convert_table(self, el, text, convert_as_inline):
        return '\n\n' + TABLE_PLACEHOLDER + '\n'

def md(html: str | BeautifulSoup, **options) → str:
    """Convert html to markdown string"""

    soup = BeautifulSoup(html, "html.parser") if isinstance(html, str) else html
    tables = soup.select("table")
    markdown = KeepTableConverter(**options).convert(html)

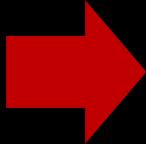
    for idx, table in enumerate(tables):
        table_str = str(table.prettify(formatter="html5"))
        markdown = markdown.replace(TABLE_PLACEHOLDER, table_str, idx + 1)

    return markdown
```

복잡한 테이블 구조

```
<div>
<p>1.관련:효원상담원-1458(2025.03.06.)</p>
<p>2. 학생들의 학과에 대한 소속감 항상 및 자신과 타인에 대한 이해 증진을 위해...</p>
<p>3.목 적:학생들의 학과에 대한 소속감 항상 및 자기와 타인에 대한 이해 증진에 도움</p>
<p>4)캠퍼스 간 심리지원서비스 수혜 격차 완화</p>
<p>나.신청기간: 2025년3월~ 2026년2월 상시 진행</p>
<p>다.신청방법: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)</p>
<p>- 신청 최소 인원 : 10명 이상</p>
<p>- 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청</p>
<p>- 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호</p>
<p>- 신청 내용 : (1)신청 명단 학번 및 이름, (2)신청 일자, 장소</p>
<p>마.신청 가능 검사</p>
<table>
<tbody>
<tr>
<td>검사 종류</td>
<td>검사 내용</td>
<td>소요시간</td>
</tr>
<tr>
<td>마인드핏(적응역량검사)</td>
<td>대학생활 적응도 및 정신건강 수준, 스트레스 대응 능력 및 영역별 스트레스 확인(해석 특강 진행)</td>
<td>약1시간</td>
</tr>
<tr>
<td>MBTI(성격유형검사)</td>
<td>자신의 성격 유형 및 타인과의 차이 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)</td>
<td>약2시간</td>
</tr>
<tr>
<td>TCI(기질 및 성격검사)</td>
<td>자신의 기질 및 성격 특성 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)</td>
<td>약2시간</td>
</tr>
</tbody>
</table>
<p>바.진행절차</p>
<p>1)희망 학과 조교가 효원상담원 홈페이지에 접속하여 온라인 신청</p>
<p>2)매달 말 신청내역 확인 후 선정된 학과 조교에게 개별 안내(밀양, 양산 소재 학과 우선 선정)</p>
<p>3)선정된 학과 조교가 장소 및 참가자 확정</p>
<p>4)참가하는 학생은 학생역량지원시스템 내 비교과 프로그램 신청</p>
<p>5)효원상담원에서 강사를 파견하여 프로그램 진행</p>
<p>사.담당 및 문의:김예주(☎3708)</p>
</div>
```

markdownify
(custom logic)



```
1.관련:효원상담원-1458(2025.03.06.)
2.학생들의 학과에 대한 소속감 항상 및 자신과 타인에 대한 이해 증진을 위해 찾아가는 심리검사 프로그램(집단심리검사)을 운영하고 있으니, 많은 관심과 참여 부탁드립니다.
가.목 적:학생들의 학과에 대한 소속감 항상 및 자기와 타인에 대한 이해 증진에 도움
캠퍼스 간 심리지원서비스 수혜 격차 완화
나.신청기간: 2025년3월~ 2026년2월 상시 진행
다.신청방법: 팀장이 학과사무실로 메일로 신청(yrjung@pusan.ac.kr)
- 신청 최소 인원 : 10명 이상
- 팀장은 시행 장소(학내)와 일시를 확정한 후 학과사무실로 신청
- 신청 메일 제목 : [찾아가는 심리검사 신청] 팀장 이름 / 핸드폰 번호
마.신청 가능 검사
<table>
<tbody>
<tr>
<td>검사 종류</td>
<td>검사 내용</td>
<td>소요시간</td>
</tr>
<tr>
<td>마인드핏(적응역량검사)</td>
<td>대학생활 적응도 및 정신건강 수준, 스트레스 대응 능력 및 영역별 스트레스 확인(해석 특강 진행)</td>
<td>약1시간</td>
</tr>
<tr>
<td>MBTI(성격유형검사)</td>
<td>자신의 성격 유형 및 타인과의 차이 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)</td>
<td>약2시간</td>
</tr>
<tr>
<td>TCI(기질 및 성격검사)</td>
<td>자신의 기질 및 성격 특성 이해, 공동체 유대감 증진(해석 특강 및 소그룹 활동 진행)</td>
<td>약2시간</td>
</tr>
</tbody>
</table>
바.진행절차
1)희망 학과 조교가 효원상담원 홈페이지에 접속하여 온라인 신청
2)매달 말 신청내역 확인 후 선정된 학과 조교에게 개별 안내(밀양, 양산 소재 학과 우선 선정)
3)선정된 학과 조교가 장소 및 참가자 확정
4)참가하는 학생은 학생역량지원시스템 내 비교과 프로그램 신청
5)효원상담원에서 강사를 파견하여 프로그램 진행
사.담당 및 문의:김예주(☎3708)
```

텍스트 임베딩:
GPU 서버는 너무 비싸

	인스턴스 크기	GPU	vCPU	메모리 (GiB)	인스턴스 스토리지 (GB)	네트워크 대역폭 (Gbps)	EBS 대역폭 (Gbps)	온디맨드 요금/시간*	1년 예약 인스턴스 실질 시간당*(Linux)	3년 예약 인스턴스 실질 시간당*(Linux)
G4dn										
단일 GPU VM	g4dn.xlarge	1	4	16	1 x 125 NVMe SSD	최대 25	최대 3.5	0.526 USD	0.316 USD	0.210 USD
	g4dn.2xlarge	1	8	32	1 x 225 NVMe SSD	최대 25	최대 3.5	0.752 USD	0.452 USD	0.300 USD
	g4dn.4xlarge	1	16	64	1 x 225 NVMe SSD	최대 25	4.75	1.204 USD	0.722 USD	0.482 USD
	g4dn.8xlarge	1	32	128	1 x 900 NVMe SSD	50	9.5	2.176 USD	1.306 USD	0.870 USD
	g4dn.16xlarge	1	64	256	1 x 900 NVMe SSD	50	9.5	4.352 USD	2.612 USD	1.740 USD
다중 GPU VM										
	g4dn.12xlarge	4	48	192	1 x 900 NVMe SSD	50	9.5	3.912 USD	2.348 USD	1.564 USD
	g4dn.metal	8	96	384	2 x 900 NVMe SSD	100	19	7.824 USD	4.694 USD	3.130 USD

text-embeddings-inference

text-embeddings-inference Public

Watch 38 Fork 235 Star 3.4k

main 9 Branches 20 Tags

Go to file Add file Code About

A blazing fast inference solution for text embeddings models

[huggingface.co/docs/text-embeddings-i...](#)

ai ml embeddings huggingface llm

Readme

kaixuanliu Optimize the performance of FlashBert on HPU by using fast mode s... b06b752 · 14 hours ago 212 Commits

.cargo fix: add cors_allow_origin to cli (#162) last year

.github Upgrade candle2 (#543) last week

assets v0.1.0 2 years ago

backends Optimize the performance of FlashBert on HPU by using fast...

candle-extensions Fixing the static-linking. (#547)

core feat: support multiple backends at the same time (#440)

docs add CLI flag disable-spans to toggle span trace logging (#48...)

load_tests fix(ort): fix mean pooling (#332)

proto feat: add default prompts (#312) 10 months ago

235 forks Report repository

1. 모델 양자화 미지원

2. bge-m3 모델 미지원
(sparse, colbert)

llama-cpp-python

 **llama-cpp-python** Public

[Watch 77](#) [Fork 1.1k](#) [Star 8.9k](#)

[main](#) [17 Branches](#) [304 Tags](#) [Go to file](#) [Add file](#) [Code](#)

Commit	Message	Time
 abetlen chore: Bump version ✓	37eb5f0 · 3 weeks ago	1,981 Commits
 .github	fix(ci): Fix the CUDA workflow (#1894)	3 months ago
 docker	fix(docker): Update Dockerfile BLAS options (#1632)	9 months ago
 docs	docs: Add project icon courtesy of 😊	8 months ago
 examples	fix(examples): Refactor Batching notebook to use new sampl...	4 months ago
 llama_cpp	chore: Bump version	3 weeks ago
 scripts	fix: pull all gh releases for self-hosted python index	8 months ago
 tests	fix: Fix memory allocation of ndarray (#1704)	7 months ago
 vendor	feat: Update llama.cpp	3 weeks ago
 .dockerignore	Add dockerignore	2 years ago

About

Python bindings for llama.cpp

[llama-cpp-python.readthedocs.io](#)

[Readme](#) [MIT license](#) [Activity](#) [8.9k stars](#) [77 watching](#) [1.1k forks](#)

[Report repository](#)

Releases 284

 [v0.3.5-metal](#) Latest
on Dec 10, 2024

[+ 283 releases](#)

llama-cpp-python(GGUF)

gpustack/bge-m3-GGUF like 4

Sentence Similarity sentence-transformers GGUF feature-extraction text-embeddings-inference arxiv:5 papers License: mit

Model card Files and versions Community ⋮ Train Deploy Use this model

main bge-m3-GGUF Go to file 2 contributors History: 2 commits

gpustack	Update README.md	2d48f17	VERIFIED	5 months ago	
imgs			feat: first commit	5 months ago	
.gitattributes	Safe		43 Bytes	feat: first commit	5 months ago
README.md	Safe		16.1 kB	Update README.md	5 months ago
bge-m3-FP16.gguf	Safe	1.16 GB	LFS	feat: first commit	5 months ago
bge-m3-Q2_K.gguf	Safe	366 MB	LFS	feat: first commit	5 months ago
bge-m3-Q3_K.gguf	Safe	402 MB	LFS	feat: first commit	5 months ago
bge-m3-Q4_0.gguf	Safe	422 MB	LFS	feat: first commit	5 months ago
bge-m3-Q4_K_M.gguf	Safe	438 MB	LFS	feat: first commit	5 months ago
bge-m3-Q5_0.gguf	Safe	459 MB	LFS	feat: first commit	5 months ago
bge-m3-Q5_K_M.gguf	Safe	468 MB	LFS	feat: first commit	5 months ago
bge-m3-Q6_K.gguf	Safe	499 MB	LFS	feat: first commit	5 months ago
bge-m3-Q8_0.gguf	Safe	635 MB	LFS	feat: first commit	5 months ago

Custom Logic

BAAI/FlagEmbedding

```
class M3Embedder(AbsEmbedder):
    """
    Embedder class for BGE-M3.

    Args:
        model_name_or_path (str): If it's a path to a local model, it loads the model from the path. Otherwise tries
            to load a model from HuggingFace Hub with the name.
        normalize_embeddings (bool, optional): If True, normalize the dense embedding vector. Defaults to :data:`True`.
        use_fp16 (bool, optional): If true, use half-precision floating-point to speed up computation with a slight
            degradation. Defaults to :data:`True`.
        query_instruction_for_retrieval: (Optional[str], optional): Query instruction for retrieval tasks, which will
            be used with :attr:`query_instruction_format`. Defaults to :data:`None`.
        query_instruction_format: (str, optional): The template for :attr:`query_instruction_for_retrieval`. Default:
        devices (Optional[Union[str, int, List[str], List[int]]], optional): Devices to use for model inference. Default:
        pooling_method (str, optional): Pooling method to get embedding vector from the last hidden state. Default:
        trust_remote_code (bool, optional): trust_remote_code for HF datasets or models. Defaults to :data:`False`.
        cache_dir (Optional[str], optional): Cache directory for the model. Defaults to :data:`None`.
        colbert_dim (int, optional): Dimension of colbert linear. Return the hidden_size if -1. Defaults to :data:`-1`.
        batch_size (int, optional): Batch size for inference. Defaults to :data:`256`.
        query_max_length (int, optional): Maximum length for query. Defaults to :data:`512`.
        passage_max_length (int, optional): Maximum length for passage. Defaults to :data:`512`.
        return_dense (bool, optional): If true, will return the dense embedding. Defaults to :data:`True`.
        return_sparse (bool, optional): If true, will return the sparse embedding. Defaults to :data:`False`.
        return_colbert_vecs (bool, optional): If true, will return the colbert vectors. Defaults to :data:`False`.

    Attributes:
        DEFAULT_POOLING_METHOD: The default pooling method when running the model.
    """
    DEFAULT_POOLING_METHOD = "cls"

    def __init__(
        self,
        model_name_or_path: str,
        normalize_embeddings: bool = True,
        use_fp16: bool = True,
        query_instruction_for_retrieval: Optional[str] = None,
        query_instruction_format: str = "{}{}", # specify the format of query_instruction_for_retrieval
        ...)
```

llama-cpp-python

```
def embed(
    self,
    input: Union[str, List[str]],
    normalize: bool = False,
    truncate: bool = True,
    return_count: bool = False,
):
    """Embed a string.

    Args:
        input: The utf-8 encoded string to embed.

    Returns:
        A list of embeddings
    """
    n_embd = self.n_embd()
    n_batch = self.n_batch

    # get pooling information
    pooling_type = self.pooling_type()
    logits_all = pooling_type == llama_cpp.LLAMA_POOLING_TYPE_NONE

    if self.context_params.embeddings is False:
        raise RuntimeError(
            "Llama model must be created with embedding=True to call this method"
        )

    if self.verbose:
        llama_cpp.llama_perf_context_reset(self._ctx.ctx)

    if isinstance(input, str):
        inputs = [input]
    else:
        inputs = input

    # reset batch
    self._batch.reset()

    # decode and fetch embeddings
    data: Union[List[List[float]], List[List[List[float]]]] = []

    def decode_batch(seq_sizes: List[int]):
        llama_cpp.llama_kv_cache_clear(self._ctx.ctx)
        self._ctx.decode(self._batch)
        self._batch.reset()

    # store embeddings
```

Custom Logic

```
class LlamaCppSession(llama_cpp.Llama):

    ...

    def _sparse_embedding(self, hidden_state: torch.Tensor, input_ids: List[int]):

        ...

        sparse_result_tensor = torch.nan_to_num(sparse_result_tensor, 0)
        token_weights_tensor = torch.nan_to_num(token_weights_tensor, 0)

        return sparse_result_tensor, token_weights_tensor

    def embed_(
        self,
        input: Union[str, List[str]],
        truncate: bool = True,
    ) → List[EmbedResult]:
        """Embed a string with lexical weights

        Args:
            input: The utf-8 encoded string to embed.

        Returns:
            A list of embeddings
        """
        ...

        return outputs
```

텍스트 임베딩:
CPU 인스턴스도 사실은 비싸다(?)

AWS EC2 T3 인스턴스

저렴한 비용으로 버스트 가능한 CPU 성능

T3 인스턴스는 훨씬 저렴한 비용으로 대부분의 범용 워크로드를 실행하도록 설계되었습니다. T3 인스턴스는 많은 일반 워크로드를 처리할 수 있는 기준 수준의 CPU 성능을 제공하면서 더 많은 성능이 필요할 때를 대비해 기준 이상으로 버스트할 수 있는 기능을 제공합니다. T3 인스턴스는 크레딧을 사용하여 CPU 사용량을 추적합니다. T3 인스턴스는 워크로드가 기준 임계값 미만에서 작동할 때 CPU 크레딧을 누적하고 기준 임계값 이상으로 실행될 때 크레딧을 사용합니다. T3 인스턴스는 고객이 필요할 때마다 얼마든지 오랫동안 높은 CPU 성능을 유지할 수 있다는 점에서 현재 시장에 출시된 다른 버스트 가능 인스턴스와는 다릅니다.

AWS EC2 T3 인스턴스

이름	vCPU(Virtual CPU)	메모리 (GiB)	기준 성능/vCPU	획득한 CPU 크레딧/시간	네트워크 버스트 대역폭(Gbps)	EBS 버스트 대역폭(Mbps)	온디맨드 요금/시간*	1년 약정 예약 인스턴스 실질 시간당*	3년 약정 예약 인스턴스 실질 시간당*
t3.nano	2	0.5	5%	6	5	최대 2,085	0.0052 USD	0.003 USD	0.002 USD
t3.micro	2	1.0	10%	12	5	최대 2,085	0.0104 USD	0.006 USD	0.005 USD
t3.small	2	2.0	20%	24	5	최대 2,085	0.0209 USD	0.012 USD	0.008 USD
t3.medium	2	4.0	20%	24	5	최대 2,085	0.0418 USD	0.025 USD	0.017 USD
t3.large	2	8.0	30%	36	5	최대 2,780	0.0835 USD	0.05 USD	0.036 USD
t3.xlarge	4	16.0	40%	96	5	최대 2,780	0.1670 USD	0.099 USD	0.067 USD
t3.2xlarge	8	32.0	40%	192	5	최대 2,780	0.3341 USD	0.199 USD	0.133 USD

2 x 0.05 = 0.1vCPU

QnA