RAG API 명세서

목차

1. /rag/insert - 문서 삽입
2. /rag/search - 문서 검색
3. /rag/delete - 문서 삭제
4. /rag/document - 문서 조회
5. /rag/data/show - 컬렉션 정보 조회
6. /reranker/enhanced-search - 통합 검색
7. /reranker/rerank - 순위 재조정
8. /reranker/batch_rerank - 배치 순위 재조정
9. /prompt/summarize - 문서 요약
10. /prompt/chat - 챗봇 응답
11. /prompt/models - 모델 목록 조회

1. /rag/insert - 문서 삽입

기본 정보

항목	내용
API 명	Milvus_data_삽입
Method	POST
URL	http://localhost/rag/insert
설명	문서를 벡터DB에 삽입
Content-Type	application/json

Request

Body Parameters

필드	필수	Type	길이	설명
domain	Y	String	32	데이터의 도메인 (예: news, description 등)
title	Y	String	128	문서 제목
author	Y	String	128	작성자 (기업 또는 특정인물)
text	Y	String	-	문서 본문 내용
info	N	Object	-	문서 추가 정보
ㄴ press_num	N	String	-	언론사 정보
ㄴ url	N	String	-	원문 URL

필드	필수	Type	길이	설명
tags	Y	Object	-	문서 태그 정보
└ date	Y	String	8	문서 작성일 (YYYYMMDD 형식)
└ user	N	String	-	사용자 정보

Response

성공 응답

필드	필수	Type	설명
status	Y	String	처리 상태 ("received")

실패 응답

HTTP 상태 코드	설명
400	잘못된 요청 (필수 필드 누락, 잘못된 형식)
500	서버 내부 오류

응답 예시

```
{
    "status": "received"
}
```

curl 예시

```
curl -X POST http://localhost/rag/insert \
  -H "Content-Type: application/json" \
  -d '{
    "domain": "news",
    "title": "메타버스 뉴스",
    "author": "삼성전자",
    "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
    "info": {
      "press_num": "비즈니스 워치",
      "url": "http://example.com/news/1"
    },
    "tags": {
      "date": "20240315",
      "user": "admin"
    }
  }'
```

2. /rag/search - 문서 검색

기본 정보

항목	내용
API 명	Milvus_데이터_검색
Method	GET
URL	http://localhost/rag/search
설명	벡터DB에서 유사 문서 검색

Request

Query Parameters

이름	필수/선택	Type	설명
query_text	Y	String	검색어
top_k	N	Integer	검색 결과 개수 (기본값: 5)
domain	N	String	검색할 도메인
author	N	String	작성자 필터
start_date	N	String	시작 날짜 (YYYYMMDD 형식)
end_date	N	String	종료 날짜 (YYYYMMDD 형식)
title	N	String	제목 검색
info_filter	N	String	info 필드 필터 (JSON 문자열)
tags_filter	N	String	tags 필드 필터 (JSON 문자열)

curl 예시

```
curl -X GET "http://localhost/rag/search?query_text=메타버스
&top_k=5&domain=news&start_date=20240301&end_date=20240315"
```

Response

성공 응답

필드	필수	Type	설명
result_code	Y	String	결과 코드 ("F000000": 성공)
message	Y	String	결과 메시지

필드	필수	Type	설명
search_params	Y	Object	검색 파라미터 정보
└ query_text	Y	String	검색어
└ top_k	Y	Integer	검색 결과 수
└ filters	Y	Object	적용된 필터 조건
search_result	Y	Array	검색 결과 목록
└ doc_id	Y	String	문서 ID
└ title	Y	String	문서 제목
└ author	Y	String	작성자
└ text	Y	String	문서 내용
└ score	Y	Float	유사도 점수
└ info	N	Object	문서 추가 정보
└ press_num	N	String	언론사 정보
└ url	N	String	원문 URL
└ tags	Y	Object	문서 태그 정보
└ date	Y	String	작성일
└ user	N	String	사용자 정보

실패 응답

HTTP 상태 코드	result_code	설명
400	F000001	검색어 누락
400	F000002	top_k 값 오류
400	F000003	info_filter JSON 형식 오류
400	F000004	tags_filter JSON 형식 오류
400	F000006	날짜 형식 오류
500	F000005	서버 내부 오류

응답 예시

```
{
    "result_code": "F000000",
    "message": "검색이 완료되었습니다.",
    "search_params": {
        "query_text": "메타버스",
```

```
            "top_k": 5,
            "filters": {
                "domain": "news",
                "date_range": {
                    "start": "20240301",
                    "end": "20240315"
                }
            }
        },
        "search_result": [
            {
                "doc_id": "20240315-메타버스-뉴스",
                "title": "메타버스 뉴스",
                "author": "삼성전자",
                "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
                "score": 0.95,
                "info": {
                    "press_num": "비즈니스 워치",
                    "url": "http://example.com/news/1"
                },
                "tags": {
                    "date": "20240315",
                    "user": "admin"
                }
            }
        ]
    }
```


3. /rag/delete - 문서 삭제

기본 정보

항목	내용
API 명	Milvus_데이터_삭제
Method	DELETE
URL	http://localhost/rag/delete
설명	벡터DB에서 문서 삭제

Request

Query Parameters

이름	필수/선택	Type	설명
date	Y	String	문서 날짜 (YYYYMMDD 형식)
title	Y	String	문서 제목
author	Y	String	작성자

이름	필수/선택	Type	설명
domain	Y	String	문서 도메인

curl 예시

```
curl -X DELETE "http://localhost/rag/delete?date=20240315&title=메타버스%20뉴스
&author=삼성전자&domain=news"
```

Response

성공 응답

필드	필수	Type	설명
status	Y	String	처리 상태 ("received")

실패 응답

HTTP 상태 코드	error	message	설명
400	date is required	날짜(date)는 필수 입력값입니다.	날짜 누락
400	title is required	제목(title)은 필수 입력값입니다.	제목 누락
400	author is required	작성자(author)는 필수 입력값입니다.	작성자 누락
400	domain is required	도메인(domain)은 필수 입력값입니다.	도메인 누락

응답 예시

```
{
    "status": "received"
}
```

4. /rag/document - 문서 조회

기본 정보

항목	내용
API 명	Milvus_문서_조회
Method	GET
URL	http://localhost/rag/document
설명	특정 문서 또는 패시지 조회

Request

Query Parameters

이름	필수/선택	Type	길이	설명
doc_id	Y	String	-	문서 ID
passage_id	N	String/Integer	-	패시지 ID

Response

성공 응답

필드	필수	Type	설명
doc_id	Y	String	문서 ID
domain	Y	String	데이터의 도메인
title	Y	String	문서 제목
info	N	Object	문서 추가 정보
└ press_num	N	String	언론사 정보
└ url	N	String	원문 URL
tags	Y	Object	문서 태그 정보
└ date	Y	String	작성일
└ user	N	String	사용자 정보
passages	Y	Array	패시지 목록
└ passage_id	Y	Integer	패시지 ID
└ text	Y	String	패시지 본문
└ position	Y	Integer	패시지 순서

실패 응답

HTTP 상태 코드	error	message	설명
400	doc_id is required	문서 ID는 필수 입력값입니다.	문서 ID 누락
404	Document not found	요청하신 문서를 찾을 수 없습니다.	문서 없음
404	Passage not found	요청하신 패시지를 찾을 수 없습니다.	패시지 없음
500	Internal server error	문서 조회 중 오류가 발생했습니다.	서버 오류

응답 예시

```
{
    "doc_id": "20240315-메타버스-뉴스",
    "domain": "news",
    "title": "메타버스 뉴스",
    "info": {
        "press_num": "비즈니스 워치",
        "url": "http://example.com/news/1"
    },
    "tags": {
        "date": "20240315",
        "user": "admin"
    },
    "passages": [
        {
            "passage_id": 1,
            "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
            "position": 1
        },
        {
            "passage_id": 2,
            "text": "특히 코로나19 이후 메타버스 시장이 급성장하고 있다...",
            "position": 2
        }
    ]
}
```

5. /rag/data/show - 컬렉션 정보 조회

기본 정보

항목	내용
API 명	Milvus_컬렉션_정보_조회
Method	GET
URL	http://localhost/rag/data/show
설명	컬렉션 정보 조회

Request

Query Parameters

이름	필수/선택	Type	설명
collection_name	Y	String	컬렉션 이름 (예: news)

curl 예시

```
curl -X GET "http://localhost/rag/data/show?collection_name=news"
```

Response

성공 응답

필드	필수	Type	설명
schema	Y	Object	컬렉션 스키마 정보
ㄴ fields	Y	Array	필드 정보 목록
ㄴ name	Y	String	필드 이름
ㄴ type	Y	String	필드 타입
ㄴ params	N	Object	필드 파라미터
partition_names	Y	Array	파티션 이름 목록
partition_nums	Y	Object	파티션별 엔티티 수

실패 응답

HTTP 상태 코드	error	message	설명
400	collection_name is required	collection_name은 필수 입력값입니다.	컬렉션 이름 누락
400	Invalid Collection	유효한 Collection 이름을 입력해야 합니다.	잘못된 컬렉션 이름

응답 예시

```
{
    "schema": {
        "fields": [
            {
                "name": "doc_id",
                "type": "VARCHAR",
                "params": {"max_length": 1024}
            },
            {
                "name": "passage_id",
                "type": "INT64"
            },
            {
                "name": "text",
                "type": "VARCHAR",
```

```
                "params": {"max_length": 512}
            }
        ]
    },
    "partition_names": ["p1", "p2"],
    "partition_nums": {
        "p1": 100,
        "p2": 150
    }
}
```

6. /reranker/enhanced-search - 통합 검색

기본 정보

항목	내용
API 명	Reranker_통합_검색
Method	GET
URL	http://localhost/reranker/enhanced-search
설명	RAG 검색 결과를 Reranker로 순위 재조정

Request

Query Parameters

이름	필수/선택	Type	설명
query_text	Y	String	검색어
top_k	N	Integer	최종 반환할 결과 개수 (기본값: 5)
raw_results	N	Integer	RAG에서 가져올 초기 결과 개수 (기본값: 20)
domain	N	String	검색할 도메인
author	N	String	작성자 필터
start_date	N	String	시작 날짜 (YYYYMMDD 형식)
end_date	N	String	종료 날짜 (YYYYMMDD 형식)
title	N	String	제목 검색

curl 예시

```
curl -X GET "http://localhost/reranker/enhanced-search?query_text=메타버스%20최신%20동향&top_k=5&raw_results=20&domain=news"
```

Response

성공 응답

필드	필수	Type	설명
result_code	Y	String	결과 코드 ("F000000": 성공)
message	Y	String	결과 메시지
search_params	Y	Object	검색 파라미터 정보
└ query_text	Y	String	검색어
└ top_k	Y	Integer	검색 결과 수
└ raw_results	Y	Integer	RAG 초기 결과 수
└ filters	Y	Object	적용된 필터 조건
search_result	Y	Array	재순위화된 검색 결과
└ doc_id	Y	String	문서 ID
└ text	Y	String	문서 내용
└ score	Y	Float	재순위화 점수
└ title	Y	String	문서 제목
└ author	Y	String	작성자
└ info	N	Object	문서 추가 정보
└ press_num	N	String	언론사 정보
└ url	N	String	원문 URL
└ tags	Y	Object	문서 태그 정보
└ date	Y	String	작성일
└ user	N	String	사용자 정보

실패 응답

HTTP 상태 코드	result_code	message	설명
400	F000001	검색어(query_text)는 필수 입력값입니다.	검색어 누락
500	F000002	검색 서비스 오류	RAG 서비스 오류
200	F000003	검색 결과가 없습니다.	검색 결과 없음
500	F000004	재랭킹 처리 중 오류가 발생했습니다.	재순위화 오류
500	F000005	통합 검색 중 오류가 발생했습니다.	기타 오류

응답 예시


```
{
    "result_code": "F000000",
    "message": "검색이 완료되었습니다.",
    "search_params": {
        "query_text": "메타버스 최신 동향",
        "top_k": 5,
        "raw_results": 20,
        "filters": {
            "domain": "news"
        }
    },
    "search_result": [
        {
            "doc_id": "20240315-메타버스-뉴스",
            "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
            "score": 0.95,
            "title": "메타버스 뉴스",
            "author": "삼성전자",
            "info": {
                "press_num": "비즈니스 워치",
                "url": "http://example.com/news/1"
            },
            "tags": {
                "date": "20240315",
                "user": "admin"
            }
        }
    ]
}
```


7. /reranker/rerank - 순위 재조정

기본 정보

항목	내용
API 명	Reranker_순위_재조정
Method	POST
URL	http://localhost/reranker/rerank
설명	문서 목록의 순위를 재조정
Content-Type	application/json

Request

Query Parameters

이름	필수/선택	Type	설명
top_k	N	Integer	최종 반환할 결과 개수

Body Parameters

이름	필수/선택	Type	설명
query	Y	String	쿼리 텍스트
results	Y	Array	순위를 재조정할 문서 목록
└ passage_id	N	Any	패시지 ID
└ doc_id	N	String	문서 ID
└ text	Y	String	문서 텍스트
└ score	N	Float	원본 점수
└ metadata	N	Object	메타데이터
└ title	N	String	문서 제목
└ author	N	String	작성자
└ info	N	Object	추가 정보
└ tags	N	Object	태그 정보

curl 예시

```
curl -X POST http://localhost/reranker/rerank \
  -H "Content-Type: application/json" \
  -d '{
    "query": "메타버스 최신 동향",
    "results": [
      {
        "passage_id": "1",
        "doc_id": "20240315-메타버스-뉴스",
        "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
        "score": 0.95,
        "metadata": {
          "title": "메타버스 뉴스",
          "author": "삼성전자"
        }
      }
    ]
  }'
```

Response

성공 응답

필드	필수	Type	설명
query	Y	String	쿼리 텍스트
results	Y	Array	재순위화된 문서 목록
└ passage_id	N	Any	패시지 ID
└ doc_id	N	String	문서 ID
└ text	Y	String	문서 텍스트
└ score	Y	Float	재순위화 점수
└ metadata	N	Object	메타데이터
total	Y	Integer	총 결과 수
reranked	Y	Boolean	재순위화 여부 (항상 true)

실패 응답

HTTP 상태 코드	error	설명
400	No JSON data provided	요청 본문 누락
400	Invalid input format	잘못된 입력 형식
500	Reranking failed	재순위화 실패

응답 예시

```
{
    "query": "메타버스 최신 동향",
    "results": [
        {
            "passage_id": "1",
            "doc_id": "20240315-메타버스-뉴스",
            "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
            "score": 0.98,
            "metadata": {
                "title": "메타버스 뉴스",
                "author": "삼성전자"
            }
        }
    ],
    "total": 1,
    "reranked": true
}
```

8. /reranker/batch_rerank - 배치 순위 재조정

기본 정보

항목	내용
API 명	Reranker_배치_순위_재조정
Method	POST
URL	http://localhost/reranker/batch_rerank
설명	여러 쿼리에 대한 문서 순위를 일괄 재조정
Content-Type	application/json

Request

Query Parameters

이름	필수/선택	Type	설명
top_k	N	Integer	최종 반환할 결과 개수

Body Parameters

이름	필수/선택	Type	설명
[배열]	Y	Array	각 쿼리별 순위 재조정 요청 목록
└ query	Y	String	쿼리 텍스트
└ results	Y	Array	순위를 재조정할 문서 목록
└ passage_id	N	Any	패시지 ID
└ doc_id	N	String	문서 ID
└ text	Y	String	문서 텍스트
└ score	N	Float	원본 점수
└ metadata	N	Object	메타데이터

curl 예시

```
curl -X POST http://localhost/reranker/batch_rerank \
  -H "Content-Type: application/json" \
  -d '[
    {
      "query": "메타버스 최신 동향",
      "results": [
        {
          "passage_id": "1",
          "doc_id": "20240315-메타버스-뉴스",
          "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
          "score": 0.95,
          "metadata": {
```

```
            "title": "메타버스 뉴스",
            "author": "삼성전자"
          }
        }
      ]
    }
  ]'
```

Response

성공 응답

필드	필수	Type	설명
[배열]	Y	Array	각 쿼리별 재순위화된 결과 목록
└ query	Y	String	쿼리 텍스트
└ results	Y	Array	재순위화된 문서 목록
└ passage_id	N	Any	패시지 ID
└ doc_id	N	String	문서 ID
└ text	Y	String	문서 텍스트
└ score	Y	Float	재순위화 점수
└ metadata	N	Object	메타데이터
└ total	Y	Integer	총 결과 수
└ reranked	Y	Boolean	재순위화 여부 (항상 true)

실패 응답

HTTP 상태 코드	error	설명
500	Batch reranking failed	배치 재순위화 실패

응답 예시

```
[
    {
        "query": "메타버스 최신 동향",
        "results": [
            {
                "passage_id": "1",
                "doc_id": "20240315-메타버스-뉴스",
                "text": "메타버스는 비대면 시대 뜨거운 화두로 떠올랐다...",
                "score": 0.98,
                "metadata": {
```

```
                    "title": "메타버스 뉴스",
                    "author": "삼성전자"
                }
            }
        ],
        "total": 1,
        "reranked": true
    }
]
```

9. /prompt/summarize - 문서 요약

기본 정보

항목	내용
API 명	Prompt_문서_요약
Method	POST
URL	http://localhost/prompt/summarize
설명	쿼리에 대한 문서를 검색하고 요약
Content-Type	application/json

Request

Body Parameters

이름	필수/선택	Type	설명
query	Y	String	검색 및 요약할 쿼리
domain	N	String	검색할 도메인
author	N	String	작성자 필터
start_date	N	String	시작 날짜 (YYYYMMDD 형식)
end_date	N	String	종료 날짜 (YYYYMMDD 형식)

curl 예시

```
curl -X POST http://localhost/prompt/summarize \
  -H "Content-Type: application/json" \
  -d '{
    "query": "메타버스 최신 동향",
    "domain": "news",
    "start_date": "20240301",
    "end_date": "20240315"
  }'
```

Response

성공 응답

필드	필수	Type	설명
query	Y	String	요청한 쿼리
summary	Y	String	문서들의 요약 결과
documents_count	Y	Integer	요약에 사용된 문서 수
prompt_length	Y	Integer	프롬프트 길이

실패 응답

HTTP 상태 코드	error	설명
400	쿼리가 필요합니다	쿼리 누락
500	문서 검색 중 오류가 발생했습니다	검색 오류
500	LLM 요청 중 오류가 발생했습니다	요약 생성 오류

응답 예시

```
{
    "query": "메타버스 최신 동향",
    "summary": "최근 메타버스 시장은 급격한 성장세를 보이고 있습니다. 특히 비대면 시대를 맞아 다양한 산업 분야에서 메타버스 기술을 도입하고 있으며, 교육, 엔터테인먼트, 업무 협업 등에서 활발히 활용되고 있습니다...",
    "documents_count": 5,
    "prompt_length": 1024
}
```

10. /prompt/chat - 챗봇 응답

기본 정보

항목	내용
API 명	Prompt_챗봇_응답
Method	POST
URL	http://localhost/prompt/chat
설명	단순 질의응답 챗봇 API
Content-Type	application/json

Request

Body Parameters

이름	필수/선택	Type	설명
query	Y	String	질문 내용
model	N	String	사용할 LLM 모델명 (기본값: 설정된 기본 모델)

curl 예시

```
curl -X POST http://localhost/prompt/chat \
  -H "Content-Type: application/json" \
  -d '{
    "query": "메타버스란 무엇인가요?",
    "model": "llama2"
  }'
```

Response

성공 응답

필드	필수	Type	설명
query	Y	String	요청한 질문
model	Y	String	사용된 모델명
response	Y	String	챗봇 응답 내용

실패 응답

HTTP 상태 코드	error	설명
400	질문이 필요합니다	질문 누락
500	LLM 요청 중 오류가 발생했습니다	응답 생성 오류
503	Ollama 서비스에 연결할 수 없습니다	서비스 연결 오류

응답 예시

```
{
    "query": "메타버스란 무엇인가요?",
    "model": "llama2",
    "response": "메타버스(Metaverse)는 가상과 현실이 융합된 초연결 디지털 세계를 의미
합니다. 이는 '초월'을 의미하는 'Meta'와 '우주'를 의미하는 'Universe'의 합성어로, 현실의
```

```
  물리적 한계를 넘어 가상공간에서 다양한 활동과 경험이 가능한 확장된 세계를 말합니다..."
}
```

11. /prompt/models - 모델 목록 조회

기본 정보

항목	내용
API 명	Prompt_모델_목록
Method	GET
URL	http://localhost/prompt/models
설명	사용 가능한 LLM 모델 목록 조회

Request

요청 파라미터 없음

curl 예시

```
curl -X GET http://localhost/prompt/models
```

Response

성공 응답

필드	필수	Type	설명
models	Y	Array	사용 가능한 모델 이름 목록
default_model	Y	String	기본 모델 이름
total	Y	Integer	총 모델 개수

실패 응답

HTTP 상태 코드	error	설명
500	모델 목록을 가져오는 중 오류가 발생했습니다	목록 조회 오류
503	Ollama 서비스에 연결할 수 없습니다	서비스 연결 오류

응답 예시

```
{
    "models": ["llama2", "mistral", "gemma"],
```

```
    "default_model": "llama2",
    "total": 3
}
```