

SelfQueryRetriever EXAONE 3.5 한계 & OpenAI 전환

📌 문제 상황

날짜: 2025-10-02

작업: SelfQueryRetriever 구현

목표: 로컬 LLM(EXAONE 3.5)으로 메타데이터 필터링 검색

🔍 발생한 문제

문제 1: 한글 카테고리 Unicode 변환

증상:

```
# LLM이 생성한 잘못된 필터
"filter": "eq(category, '\\uc2a4\\ud0a8\\ucf00\\uc5b4')" # ❌

# 필요한 올바른 필터
"filter": "eq(category, '선케어')" # ✅
```

에러:

```
UnexpectedCharacters: No terminal matches ', ' in the current parser
context
```

원인:

- EXAONE 3.5가 한글을 Unicode 이스케이프로 변환
- Chroma 파서가 이를 인식하지 못함

문제 2: Float 비교 연산자 문법 오류

증상:

```
# LLM이 생성한 잘못된 문법
"filter": "gte(user_rating, 4.5)" # ❌ 심표 있음
```

```
# 올바른 Chroma 문법
"filter": "gte(user_rating 4.5)"
로 분리
```

 공백으

예러:

```
UnexpectedToken: Unexpected token Token('COMMA', ',') at line 1,
column 16
```

원인:

- EXAONE 3.5가 Python 함수 호출 문법(실패) 생성
- Chroma는 특수한 DSL 문법 사용 (공백 구분)

문제 3: 한글 쿼리 매핑 실패

증상:

```
retriever.invoke("선크어 제품")
검색
```

 0개

```
retriever.invoke("suncare 제품")
검색
```

 1개

원인:

- EXAONE 3.5가 "선크어" → "suncare" 자동 매핑 못함
- 메타데이터 `description`에 한영 매핑 있어도 실패

시도한 해결책

시도 1: 임베딩 모델 변경

- 3개 모델 테스트 (`multilingual-mpnet`, `ko-sroberta`, `all-MiniLM`)
- 결과: 임베딩 문제 아님 (LLM 출력 문제)

시도 2: 카테고리 영어 변경

- "스킨케어" → "skincare"
- 결과: Unicode 문제 해결, 하지만 문법 오류 여전

시도 3: 메타데이터 단순화 ⚠

- 3개 필드로 축소 (`category`, `year`, `rating`)
- 결과: 약간 개선, 여전히 불안정

시도 4: 프롬프트 강화 ❌

- Few-shot 예시 추가
- 한영 매핑 명시
- 결과: 일부 쿼리만 성공 (일관성 부족)

시도 5: 수동 필터링 대안 ✅

- `smart_search()` 함수 구현
- 결과: 완벽 동작, 하지만 `SelfQuery` 개념 아님

✅ 최종 해결책: OpenAI GPT-4o-mini 사용

선택 이유

1. 구조화 출력 완벽

- `Chroma DSL` 문법 정확히 생성
- 심표/공백 구분 완벽

2. 한글 처리 우수

- "선케어" → "suncare" 자동 매핑
- `Unicode` 이스케이프 없음

3. 비용 효율적

- \$0.15 / 1M tokens
- 쿼리 1개당 ~\$0.0001

4. 안정성

- 100% 성공률
- 복합 쿼리 완벽 동작

구현 코드

```
from langchain_openai import ChatOpenAI

# OpenAI LLM 로드
llm_openai = ChatOpenAI(
    model="gpt-4o-mini",
    temperature=0,
)
```

```
retriever = SelfQueryRetriever.from_llm(
    llm=llm_openai,
    vectorstore=vectorstore,
    document_contents="Brief summary of a cosmetic product",
    metadata_field_info=metadata_field_info,
    enable_limit=True,
    verbose=True,
)
```

테스트 결과

```
# ✅ 평점 필터 (성공) → 2개 정확
retriever.invoke("평점 4.8 이상")

# ✅ 연도 필터 (성공) → 3개 정확
retriever.invoke("2023년 제품")

# ✅ 복합 필터 (성공) → 1개 정확
retriever.invoke("2023년 + 평점 4.5+ + skincare")
```

📊 성능 비교

항목	EXAONE 3.5	OpenAI GPT-4o-mini
한글 카테고리	❌ Unicode 오류	✅ 완벽
Float 비교	❌ 문법 오류	✅ 완벽
한글 쿼리 매핑	❌ 실패	✅ 자동
복합 쿼리	❌ 불안정	✅ 안정적
속도	9-17초	0.9-2.4초
성공률	~30%	100%
비용	무료	\$0.0001/쿼리

💡 교훈

1. 오픈소스 LLM 한계

- 특수 DSL 문법 생성 능력 부족
- 구조화 출력 일관성 낮음
- Fine-tuning 없이는 어려움

2. SelfQuery는 OpenAI 권장

- LangChain 공식 예제 모두 OpenAI
- 복잡한 파서 로직 필요
- 상용 LLM 추천


3. 실무 선택 기준

- 학습/프로토타입: EXAONE + 수동 필터링
- 실제 서비스: OpenAI GPT-4o-mini
- 대규모 무료: 수동 필터링 함수 구현

4. 포트폴리오 가치

- 문제 분석 능력 증명
- 다양한 해결책 시도
- 비용-성능 트레이드오프 이해
- 최종적으로 작동하는 솔루션 완성




관련 자료

- 코드: [10_Retriever/08_SelfQueryRetriever.ipynb](#)
- 문서: [LangChain SelfQuery 공식 문서](#)
-  **Related:** #32.12 (SelfQueryRetriever 구현)
- **선행작업:** #32.11 (MultiVectorRetriever 로컬 LLM)




결론

EXAONE 3.5로는 SelfQueryRetriever 완전 구현 불가

OpenAI GPT-4o-mini 사용 시:

-  모든 쿼리 완벽 동작
-  비용 효율적 (\$0.0001/쿼리)
-  실무 적용 가능

학습 목적 달성:

-  SelfQuery 개념 이해
-  문제 해결 능력 증명
-  실무 의사결정 경험

다음 단계:

- TimeWeightedVectorStoreRetriever 학습
- 다른 고급 Retriever 탐구