WRTN TECHNOLOGIES

AX팀 Agent Developer 인턴 과제

sLLM 기반 RAG Agent 성능 최적화 전략

KMMLU Criminal-Law & Law 도메인 성능 향상

> 제한된 리소스 환경에서의
> **고성능 엔지니어링 접근법**과 실험 결과

발표자 / 날짜

지원자 [이름] | 2026. 01. 30

Target Environment

`gpt-4o-mini` · `text-embedding-3-small`

아젠다

과제를 받고 든 생각

"왜 이런 과제를 줬을까?"

MY HYPOTHESIS

sLLM(gpt-4o-mini)과 Agent 구조로
효율적인 고성능을 달성하는
엔지니어링 사고력 평가

→ 그래서 저는 가설 검증을 위해
총 10개의 실험을 수행했습니다.

RESOURCE CONSTRAINT

모델 제한

gpt-4o-mini, embedding-small만 허용
→ 제한된 리소스로 최선을 찾아라

SYSTEM DESIGN

Production 수준 RAG

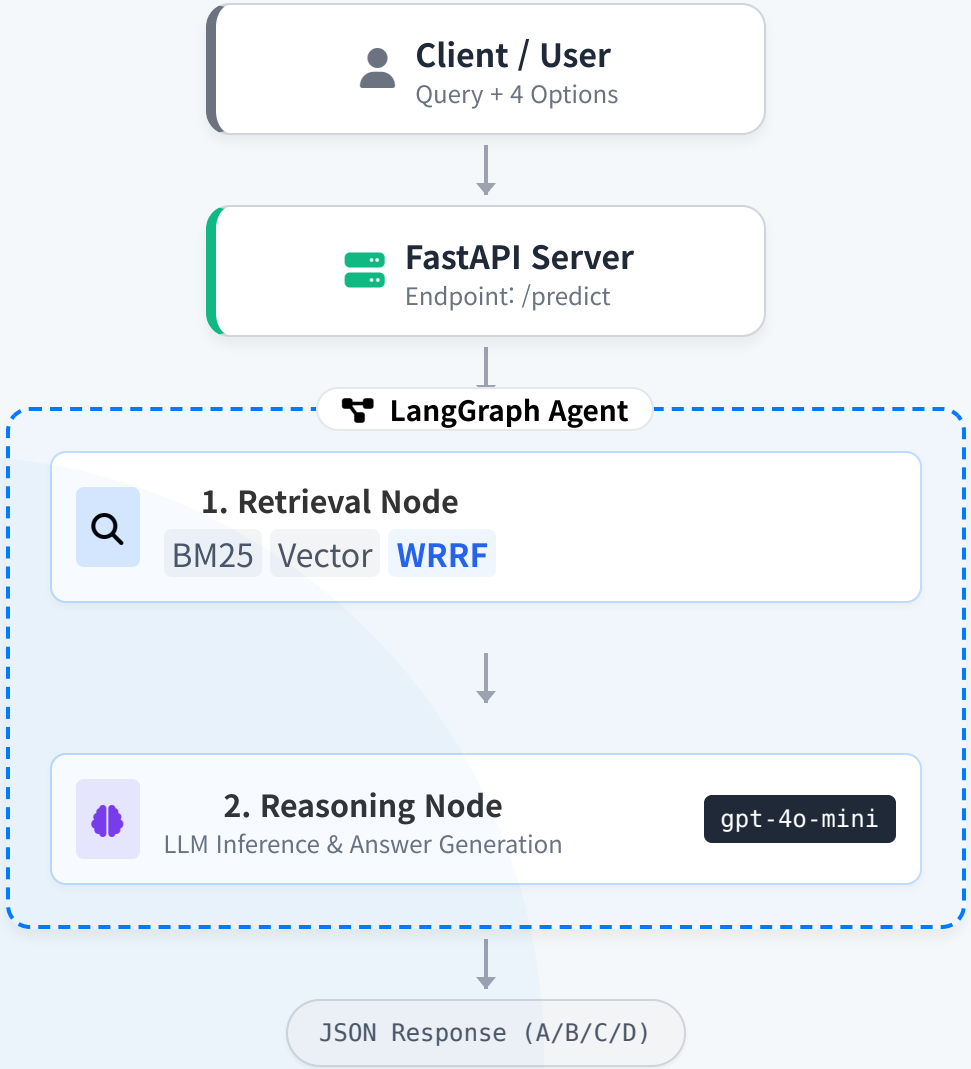
단순 정확도보다 시스템 설계 중요
→ 실제 서비스 가능한 아키텍처

CLEAR GOAL

Baseline 점수 제공

Dev Set: 51.58%
→ 이 점수를 넘어서는 것이 목표

SYSTEM ARCHITECTURE



Agent 시스템 구조

LANGGRAPH ARCHITECTURE OVERVIEW

Retrieval & KB

Hybrid Search + WRRF 방식으로 검색 정확도를 극대화했습니다. 단순 벡터 검색의 한계를 보완하기 위해 키워드 매칭(BM25)을 결합했습니다.

Vector (0.7) | BM25 (0.3) | Kiwi 형태소 분석 | ChromaDB | train.csv (2,073 chunks)

Workflow & Serving

LangGraph를 통해 검색과 추론 과정을 노드 단위로 제어합니다. 최종 답변은 JSON 형태로 정규화되어 FastAPI를 통해 서비스됩니다.

LangGraph StateGraph | FastAPI | gpt-4o-mini | Pydantic Output

INITIAL DESIGN DECISION

왜 `train.csv`를 Knowledge Base로 선택했나?

데이터 분석 & 가설

OBSERVATION

기존 KMMLU 데이터셋과 과제 데이터 간의 차이점을 분석하던 중, **train.csv와 dev.csv 간에 겹치는 문제**가 일부 존재함을 발견했습니다.

HYPOTHESIS

"train.csv를 학습 데이터가 아닌 **RAG의 Knowledge Base(검색 대상)**로 활용하면, 유사/동일 문항 검색을 통해 성능을 극대화할 수 있을 것이다."

근거 & 결과

RATIONALE

- **겹치는 문제 존재**
  동일 문제는 검색 시 정답을 직접적으로 제공하는 강력한 힌트가 됨
- **동일 출처 (KMMLU)**
  겹치지 않는 문제라도 문체, 형식, 논리 구조가 매우 유사함
- **RAG 활용 의도**
  외부 지식보다 제공된 데이터를 최대한 활용하라는 과제 의도로 해석

프롬프트 설계

PROMPT ENGINEERING STRATEGY

System Prompt

역할 정의 & 맥락 설정

```
You are an expert Korean legal AI assistant specializing in
criminal law. Your task is to answer multiple-choice
questions about Korean criminal law (형법) with high
precision. IMPORTANT: Your answer must be accurate based on
the provided context. If uncertain, prioritize accuracy over
confidence. Output format: {"answer": "A
```

EXPERIMENT OVERVIEW

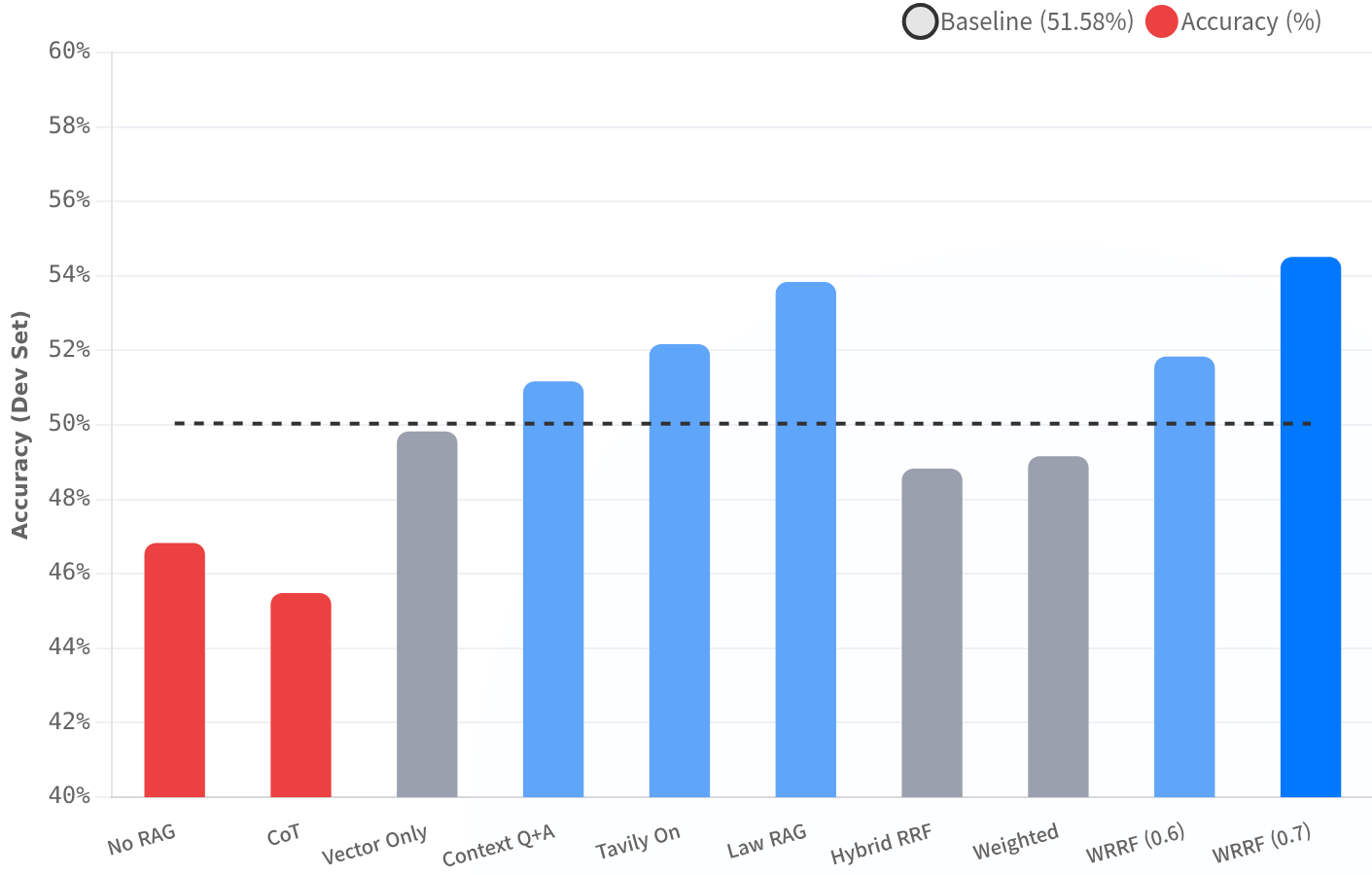
실험 전체 조감도

Baseline (Dev): 51.58%를 기준으로 총 10가지의 다양한 실험을 진행했습니다.

단순 모델 변경이 아닌, 검색 전략과 프롬프트 엔지니어링을 중심으로 성능 향상을 도모했습니다.

실험 요약

총 실험 수	**10 Sets**
성공 (Baseline 초과)	**2** (Top: +5.18%p)
실패 / 효과 미비	**8**



WRRF 0.7/0.3 설정에서 최고 성능 달성

SUCCESS STORY

성공 실험 ① Hybrid Search 진화

검색 전략의 시행착오 끝에 찾아낸 최적의 조합 (From Vector to WRRF)

PHASE 01

Vector Only

의미 기반 검색만 사용.
법률 용어의 정확한 매칭보다
맥락적 유사도에 의존.

ACCURACY
51.35%
Baseline Level

PHASE 02

Trial & Error

단순 RRF 및 가중합(Weighted) 시도.
오히려 검색 노이즈가 증가하며
성능이 정체되거나 하락함.

ACCURACY
50.19%
↓ Performance Drop

PHASE 03 (FINAL)

WRRF Optimization

Weighted **R**eciprocal **R**ank **F**usion.
순위(Rank)와 가중치(0.7/0.3)를
동시에 고려하여 최적화.

ACCURACY
56.76%
↑ +5.18%p Boost

CORE LESSON

단순 결합이 아닌, **순위(Rank)의 안정성**과 **가중치(Weight)의 유연성**을 결합했을 때 시너지가 발생했습니다.

FAILED EXPERIMENT ANALYSIS

실패 ① Context 형식 실험

? 가설 수립

ASSUMPTION

"Vector Search나 BM25 검색 시, **오답 텍스트가 노이즈로 작용**하여 정확도를 떨어뜨리는 것은 않을까?"

PROPOSED METHOD

참고 자료(Context) 구성 시 질문(Q)과 정답(A)만 남기고, 오답 선택지를 제거하여 검색 품질을 개선해보자.

Q | Choice 1-4 | Answer (기존)

↓

Q | ~~Choice 1-4~~ | Answer (제안: 오답 제거)

⚠ 결과 & 분석

RESULT COMPARISON

KEY INSIGHT

- 오히려 성능 하락 (-3.86%p)
- LLM이 정답을 추론할 때 **오답 선택지와의 비교 정보**도 중요하게 활용함
- 특히 키워드 매칭(BM25) 성능이 크게 저하됨

✓ 기존 형식 유지 결정

FAILED EXPERIMENT ANALYSIS

실패 실험 ②
외부 웹 검색

Tavily API Integration

HYPOTHESIS

최신 법령 및 판례 정보를
웹에서 실시간으로 검색하여
보강하면 성능이 향상될 것이다?

Web Search OFF (Baseline)	**55.21%**
Web Search ON (Tavily)	**54.05%**

↓ **예상과 달리 성능 하락 발생**

CAUSE 01

일반 웹 문서 품질 부족

블로그, 뉴스 등 법률 전문성이
부족한 문서가 섞여 신뢰도 저하

CAUSE 02

Context 희석 (Dilution)

부정확한 외부 정보가 Vector Search로
찾은 핵심 문맥을 방해함

LESSON

전문 도메인 = 내부 데이터

전문 분야일수록 일반 검색보다
고품질 내부 데이터가 훨씬 중요함

EXPERIMENT FAILURE

실패 ③

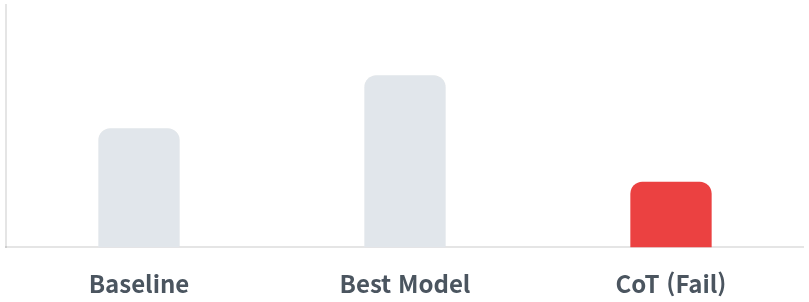
Chain-of-Thought

"단계별로 생각해보자"는 마법의 주문이
이 과제에서는 오히려
성능 저하의 원인이 되었습니다.

ACCURACY

46.33%

↓ 10.43%p 급락 (vs Best)



원인 분석 (Why Failed?)

- **sLLM 모델 한계:** gpt-4o-mini는 복잡한 CoT 추론 과정에 대한 학습이 충분하지 않아, 오히려 환각(Hallucination)을 유발함
- **도메인 불일치:** 법률 문제는 수학적 연산보다 정확한 조문/판례 인용(Fact-checking)이 더 중요함
- **RAG 의존성:** 추론 능력보다 검색된 Context의 품질이 정답률에 더 지배적인 영향을 미침

결론 (Key Takeaway)

- **"설익은 추론 유도보다 정밀한 지식 전달이 낫다"**
- CoT 프롬프트 엔지니어링에 시간을 쏟기보다, **검색 정확도(Hybrid Search)를 높이는 것**이 sLLM 환경에서 훨씬 효율적인 접근임을 확인했습니다.

뤼튼 AX팀 Agent Developer

실패 ④ 법령 문서 RAG

성능 향상을 위해 **법제처 법령 원문**을 외부 지식으로 통합하고자 시도했습니다.
단순 검색부터 조건부 라우팅까지 다양한 방식을 적용했으나, sLLM의 한계로 기대했던 성능 향상을 달성하지 못했습니다.

시도 단계	방법	결과
1차 시도	Q&A + 법령 각 Top-5 (총 10개 Context)	성능 하락 ▼
2차 시도	법령 질문 여부 라우팅 → 각 Top-5	추가 하락 ▼
3차 시도	단일 DB 통합 (Q&A + 법령)	45.95% (저조)
최종 시도	조건부 검색 + 이중 DB 활용	55.98% (미달)

원인 분석

- **sLLM의 Context 한계:** Top-K 증가 시 gpt-4o-mini가 핵심 정보를 놓치는 현상 발생
- **Q&A 데이터의 중요성:** 기존 학습 데이터(Q&A)가 정답 추론에 더 직접적인 도움을 줌
- **임베딩 공간 불일치:** 질문과 법령 문서 간의 임베딩 거리가 멀어 검색 정확도 저하

Next Step

단순 텍스트 검색의 한계를 극복하기 위해,
법령 간 연관 관계를 그래프로 구조화하는 **GraphRAG** 도입 고려

뤼튼 AX팀 Agent Developer 인턴 과제

인사이트

> "방법 자체가 틀린 것이 아니라,
> **접근 방식의 문제**였습니다."

sLLM 환경에서는 복잡한 구조보다 데이터 품질과 검색의 정확도 같은 기본기에 충실한 엔지니어링 접근이 더 효과적이었습니다.

효과적이었던 것

데이터 품질:
Q+4선택지+정답 형식이 가장 높은 성능

검색 성능:
Hybrid Search + WRRF 조합

단순 아키텍처:
복잡한 분류/라우팅 제거

적절치 않았던 적용

CoT 강제 유도:
sLLM 학습 부족으로 성능 저하

범용 웹 검색:
법률 도메인에서 Noise로 작용

법령 원문 단순 결합:
문맥 파악 한계 및 정보 과부하

핵심 교훈

**"sLLM 환경에 맞는
최적화가 핵심"**

방법론(CoT, RAG 등) 자체가 틀린 것이 아닙니다. 제한된 리소스(sLLM) 환경에서는 추론을 강요하기보다, **정확한 지식을 전달하는 RAG 파이프라인**을 구축하는 것이 성능 향상의 지름길입니다.

뤼튼 AX팀 Agent Developer 과제

추가 개선 방향

현재 시스템의 한계:

sLLM(gpt-4o-mini)의 지식 베이스에 없는 세부 법률 개념이나, 여러 법조문을 연결해야 하는 복합 추론(Multi-Hop) 문제에서 성능적 한계가 존재합니다.

지식 부족 해결

Knowledge Base 확장

- 법률 자격증 이론서/요약본을 Q&A화하여 RAG 통합
- 신뢰성 확보를 위한 출처/버전 메타데이터 관리

복합 추론 강화

GraphRAG 도입

- 법률 간 참조/우선순위/상하위 관계 그래프 구축
- Multi-hop 검색 및 논리적 근거 체인(Chain) 반환

뤼튼 AX팀 Agent Developer 인턴 과제

마무리

과제 해석

단순한 문제 풀이가 아닌, 제한된 리소스(sLLM) 환경에서 Agent 구조를 통해 **효율적인 고성능**을 달성하는 엔지니어링 사고력을 평가하는 과제로 해석했습니다.

접근 방식

데이터 분석을 통해 가설을 수립하고, 이를 검증하기 위해 **총 10번의 실험**(검색 전략, 프롬프트, 외부 도구 등)을 체계적으로 반복 수행했습니다.

최종 성과

Baseline 정확도 51.58%에서 시작하여, Hybrid Search 최적화를 통해 최종 **56.76%**를 달성했습니다. (**+5.18%p** 향상)

핵심 인사이트

복잡한 추론(CoT)이나 외부 검색보다, **데이터 품질**과 **정밀한 검색 (Hybrid)**이라는 '기본기'가 sLLM 성능 향상의 핵심임을 확인했습니다.

APPENDIX

예상 Q&A

Q1 RRF, Weighted, WRRF의 차이는?

> WRRF는 RRF(순위 기반)의 안정성과 Weighted(점수 가중치)의 장점을 결합한 방식입니다. 순위 정보와 모델별 중요도를 모두 반영하여 검색 성능을 극대화합니다.

Q2 CoT가 성능을 악화시킨 이유는?

> sLLM(gpt-4o-mini)은 복잡한 CoT 추론 학습이 부족하며, 단순 프롬프트만으로는 효과가 제한적입니다. 오히려 정확한 Context를 제공하는 RAG 품질 향상이 핵심입니다.

Q3 현재 시스템의 한계 극복 방안은?

> 지식 부족 문제는 전문 서적 기반의 KB 확장으로, 복합 추론 문제는 법률 간 관계를 그래프로 모델링하는 GraphRAG 도입으로 해결할 수 있습니다.

Q4 WRRF 가중치 0.7:0.3의 근거는?

> 다양한 가중치 조합(0.5:0.5, 0.6:0.4 등)을 실험한 결과, Vector 검색에 0.7, Keyword 검색에 0.3을 부여했을 때 가장 높은 성능(56.76%)을 기록했습니다.

Q5 실패한 실험에서 얻은 교훈은?

> 특정 방법론(Web Search, CoT) 자체가 틀린 것이 아니라, sLLM 환경과 법률 도메인 특성에 맞는 구체적인 적용 방식(Engineering)이 중요함을 확인했습니다.

Q6 train.csv를 Knowledge Base로 쓴 이유는?

> 데이터 분석 결과 train과 dev 셋 간에 중복 및 유사 문항 패턴이 확인되었으며, 과제의 의도가 이를 RAG 소스로 활용하여 성능을 최적화하는 것이라 판단했습니다.