

# 자기소개 및 지원동기

Black-box 평가를 넘어, **모델 내부 메커니즘을 진단**하는 연구로의 확장

DSBA LAB APPLICATION

Ph.D. Candidate



신정운

AI/ML Engineer

## CORE COMPETENCIES

LLM Evaluation

RAG Systems

Mechanistic Interpretability

✉ jungwoonshin@gmail.com



## 문제의식: 평가의 비효율성과 블랙박스의 한계

### 평가의 비효율성 (Efficiency Bottleneck)

현업 대규모 RAG 시스템 운영 시, 단순 평가를 위해 막대한 추론 자원과 데이터를 소모하는 방식의 한계를 체감

### 설명력의 부재 (Lack of Explainability)

LLM-as-a-Judge는 정답 여부는 판별하나, "왜 그런 판단을 내렸는지"에 대한 인과관계 설명 불가



## 지원동기 및 연구 비전

강필성 교수님의 **CheckEval** 연구가 지향하는 '평가의 체계화'에 깊이 공감하며, 이를 모델의 **Internal Mechanism** 영역으로 확장하고 싶습니다.

### RESEARCH APPROACH

#### 기계적 해석 (Mechanistic Interpretability)

결과값이 아닌 모델 내부의 정보 흐름과 사고 과정을 직접 관찰

### ULTIMATE GOAL

#### 자원 효율적 (Resource-Efficient) 평가


적은 연산 비용으로 신뢰성 검증이 가능한 프레임워크 제안

SELECTED PUBLICATIONS

1저자 2020  
**Bipartite Link Prediction by Intra-class Connection based Triadic Closure**  
IEEE Access (IF: 4.64)  
딥러닝 기반 이분 그래프 링크 예측 알고리즘 제안 및 최신 성능 달성

공동저자 2019  
**Kitchenette: Predicting and Recommending Food Ingredient Pairings**  
IJCAI 2019 (Top Conference)  
Siamese Neural Network를 활용한 식재료 조합 예측 및 추천 모델

EDUCATION

 **고려대학교 컴퓨터과학 석사**  
2018 - 2020 | GPA 4.00/4.50  
DMIS-Lab 연구원 (PyTorch 링크예측 연구)

 **보스턴대학교(BU) 컴퓨터과학 학사**  
2011 - 2016

PROFESSIONAL EXPERIENCE

 **Coxwave** 2025.06 - Current  
AI/ML 엔지니어

**통합 RAG 시스템 개발**  
LLaVA-Next 및 하이브리드 검색(BM25+BGE-M3) 활용 PDF QA 자동화

**임베딩 모델 파인튜닝 (Top-1 Acc 82.8%)**  
BGE-M3 학습 데이터 최적화(Summary+QA) 및 도메인 취약점 분석

 **CJ대한통운** 2023.11 - 2025.05  
데이터 사이언티스트

- VRP 최적화:** LP & Heuristic 기반 배송 경로 최적화 및 검증 시스템
- LLM 문서 관리:** 한국어 LLaMA + E5 임베딩 기반 사내 RAG 구축
- Transformer 기반 대규모 차량 운행 주문 추천 모델 개발

 **오버테이크 (Startup)** 2022.08 - 2023.09  
데이터 사이언티스트

- 금융 추천 시스템 상용화:** CTR 모델링으로 클릭 수 5-10배 증대
- 금융위 주최 D-Testbed 신용정보원장상 수상 (대출 증감 예측)

# 한국어 도메인 특화 RAG 시스템 최적화

교육 QA 시스템의 Document Retrieval Rate 최대화 (60% → 89.33%)

Path: similarity\_search/finetune/kywin3-finetuned-best-train\_test  
KEY PROJECT 01  
Performance Optimization

## PROJECT GOAL

Baseline

60.0%

+29.33%p

Target

89.3%

## EVALUATION PIPELINE

1

Question Gen

2

Retrieval

3

Verification

## FINE-TUNING RESULTS

Methodology	Top-1 Acc
QA Multi Pos Neg (1:N)	43.56%
Original (Baseline)	68.12%
Summary	74.69%
Summary + QA <span>Best</span>	83.75%

## PROJECT OVERVIEW & ARCHITECTURE

2024 (6mo) | ML Engineer (E2E)

Domain	Education (교육)
Problem	Dense Model Limits on Tech Terms

### 1. SEMANTIC DOCUMENT PARSING

PDF → DocAI → Structure-Aware Chunking

### 2. HYBRID SEARCH OPTIMIZATION

SPARSE (BM25)

70%

KiwiPiePy Tokenizer

DENSE (BGE-M3)

30%

Fine-tuned

### TECH STACK

BGE-M3

BM25

KiwiPiePy

DocAI

Pinecone

## BEST MODEL DETAIL ANALYSIS

OVERALL ACC 65.01% TOP-K ACC 83.75% DOCUMENTS 11 QUESTIONS 1,126

Document Name	Acc	Top-K	Q's
[패스트캠퍼스] 클라우드 환경... 파트2_ch1	0.4314	0.7647	51
[비법노트] 9회 내부회계관리제도	0.8500	0.9167	60
[Ch3-2. 프로세스와 스프레드] 스프레드 다루기	0.4478	0.7313	67
02_01_개체명인식	0.7333	0.8667	15
1. PT1.CH1.CL1 _ 머신러닝 엔지니어란...	0.5303	0.7424	66
4. 흐름제어 (조건문, 반복문, 예외처리)	0.5159	0.7778	126

Document Name	Acc	Top-K	Q's
Part04. 배당_강의자료	0.4051	0.6051	195
ubion_23	0.8667	0.9481	135
강의자료_부동산 인허가_29강	0.7204	0.9247	93
01_데이터분석적 사고	0.6667	0.8889	18
ubion_24	0.8100	0.9733	300
Total 11 Documents / 1,126 Questions			

# 금융 상품 추천 시스템 개발 및 상용화

행동 로그 분석 기반 개인화 추천 시스템(CTR Prediction) 구축 및 성과

KEY PROJECT 02  
End-to-End Development

## OVERVIEW

### ROLE & PERIOD

단독 개발 (E2E) | 6 months

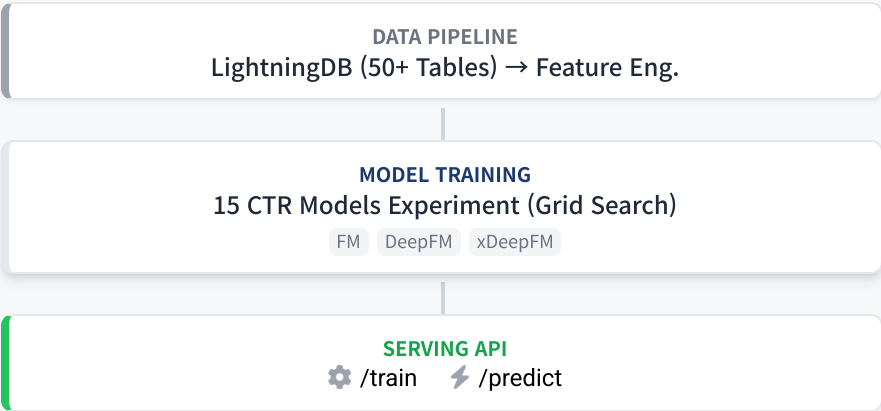
### OBJECTIVE

무작위 노출의 한계를 극복하고 잠재 고객을 사전 식별하여 추천

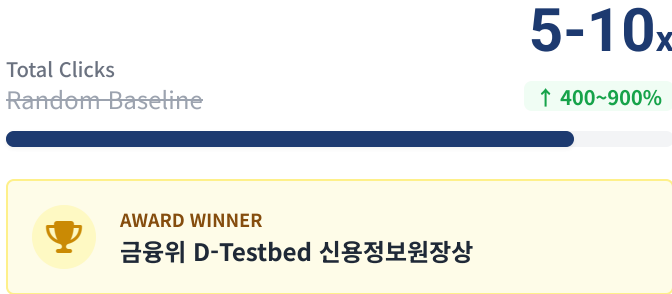
### KEY ACHIEVEMENT

클릭 수 5~10배 증가

## SYSTEM ARCHITECTURE



## BUSINESS IMPACT



## TECH STACK

### ML & Modeling

PyTorch torch\_fm DeepFM

### Data & Pipeline

Pandas LightningDB 50+ Tables

### Serving

RESTful API Scheduler

## CORE IMPLEMENTATION

- ✔ CTR 모델 체계적 실험  
torch\_fm 활용 15개 모델 실험, F1 Score 기반 최적 모델 선정 (Class Imbalance 고려)
- ✔ 대규모 로그 파이프라인  
수천 개의 고차원 카테고리 피처를 Embedding Layer로 저차원 학습
- ✔ Ray 기반 분산 실험  
Ray Tune 활용 15개 모델 병렬 튜닝(Grid/Random Search) 및 자동화된 모델 선택 파이프라인 구축

## TECHNICAL CHALLENGES

- ⚠ Class Imbalance  
Solution: Focal Loss 도입 및 Undersampling 전략, F1 Score 지표 채택
- ⚠ High Cardinality  
Solution: Feature Interaction 자동 학습을 위한 FM 계열 모델 및 Embedding 적용

# 향후 연구 계획: White-box LLM 평가 프레임워크

기계적 해석(Mechanistic Interpretability) 관점의 투명하고 효율적인 AI 평가 방법론

RESEARCH PLAN

Ph.D. Roadmap

## CONNECTION TO EXPERIENCE

"RAG 프로젝트에서 정확도 89.33%를 달성했으나, 모델이 정말로 검색된 문서를 보고 답했는지, 아니면 환각인지 구분할 수 없었다."

**핵심 문제의식:** 정답 여부(Output)만으로는 '과정의 타당성'을 검증할 수 없으며, 기존 Judge 모델은 고비용 문제와 사후 탐지의 한계가 존재함.

## PARADIGM SHIFT IN EVALUATION

Criteria	Current (Black-box)	Proposed (White-box)
분석 대상	Output (결과값)	Internal State (내부 상태)
평가 주체	External Judge (GPT-4)	Internal Signal (자체 신호)
평가 범위	Correctness (정답 여부)	Process Validity (과정 타당성)

## RESEARCH ROADMAP

### Exploratory Analysis

내부 활성화 패턴 ↔ 신뢰성 상관관계

❓ 모델이 정답/환각을 생성할 때, 내부 신경망(Attention, Hidden State)의 차이는 무엇인가?

**TARGET** Attention Map, Activation Pattern

**METHOD** 정보 흐름 시각화, Probing Classifier

**OUTPUT** 신뢰성 판단 Proxy Metric 후보군

### Causal Mechanism

RAG 정보 인용 메커니즘 인과성 분석

❓ 모델이 검색된 문서를 실제로 참고했는가? (Context Neglect 탐지)

**TARGET** Context → Generation 정보 흐름

**METHOD** Causal Intervention, Attention Attribution

**OUTPUT** Process-based Eval 프레임워크

### Framework Dev

자원 효율적 평가 방법론 정립

❓ 고비용 외부 모델 없이, 내부 신호만으로 답변 품질을 추정할 수 있는가?

**GOAL** 경량화된 Self-evaluation 정립

**VALUE** Explainability + Cost-efficiency

**OUTPUT** 범용적 White-box 평가 기준

## ACADEMIC CONTRIBUTION

Mechanistic Interpretability

RAG Evaluation

Trustworthy AI

## PRACTICAL IMPACT

💰 Cost-free Evaluation

🔍 Explainable "Why"

⚡ Real-time Detection