Part V AI 개발 연구 — 예측모형과 RAG

AI 산업보건 석사논문

12주차· 13주차· 14주차

이 Part 에서 답할 질문

- 예측모형 연구는 무엇이고 연관 연구(Part I)와 무엇이 다른가
- RAG 챗봇은 무엇이고, 만든 뒤 무엇을 보고해야 연구가 되는가
- 무엇이 가장 자주 틀리는가 — 좋은 성능, 정확도, 그리고 AI 가 만든 테스트

0. 문제로 시작합니다

두 가지 요청이 들어옵니다. 안전보건팀장: "특수건진과 측정 자료가 있으니 내년에 다칠 위험이 큰 사람을 미리 알려 주는 모델을 만들 수 있나요? AI 가 AUC 0.99 를 냈다는데요." 보건관리자: "KOSHA 지침이 너무 많아서 챗봇을 만들었습니다. 잘 답합니다. 논문이 되나요?"

첫 질문의 0.99 는 성과가 아니라 증상일 가능성이 높고, 둘째 질문의 "잘 답합니다"는 아직 아무것도 측정하지 않은 상태입니다.

1. 무엇인가 — 1.1 예측모형 연구

검진 시점의 정보로 미래의 사건(다음 해 산재)을 얼마나 잘 가려내는가를 묻는 연구입니다. Part I 의 연관 연구와 자료는 같지만 질문이 다릅니다.

	연관 (Part I)	예측 (Part V)
질문	교대근무는 산재 위험을 높이는가	누가 내년에 다칠 위험이 큰가
답	효과 추정치 + CI	판별·보정 지표 + 검증
변수 선택	DAG — 교란만, 매개는 뺌	결과 **이전**에 측정된 것이면 다 가능
해석	인과에 가까움	인과 아님. 중요도 ≠ 원인
보고 지침	STROBE	**TRIPOD+AI**

1. 무엇인가 — 1.2 RAG(검색증강생성) 챗봇

산업보건 규범(법·고시·KOSHA GUIDE)은 파편화되어 있고 개정이 잦습니다. 범용 챗봇은 개정을 모르고 고시 번호를 지어내며 근거 조문을 대지 못합니다. **RAG** 는 답을 지정된 원전 문서에서 검색된 근거에 묶고 인용을 붙이는 구조입니다. 만드는 것은 하루면 됩니다.

1. 무엇인가 — 1.2 RAG(검색증강생성) 챗봇

챗봇을 만든 것은 연구가 아닙니다

평가셋에 대한 성능을 보고해야 연구입니다

1. 무엇인가 — 1.3 언제 쓰는가

예측 — 개입 자원이 한정되어 우선순위가 필요할 때. RAG — 원전이 명확하고 현행 여부가 중요하며 인용이 필요할 때. 둘 다 **평가셋(정답)이 있어야** 연구가 됩니다. 컴퓨터 비전(보호구 탐지)·모션캡처(RULA 자동 채점)도 같은 논리이며 ·7.6.

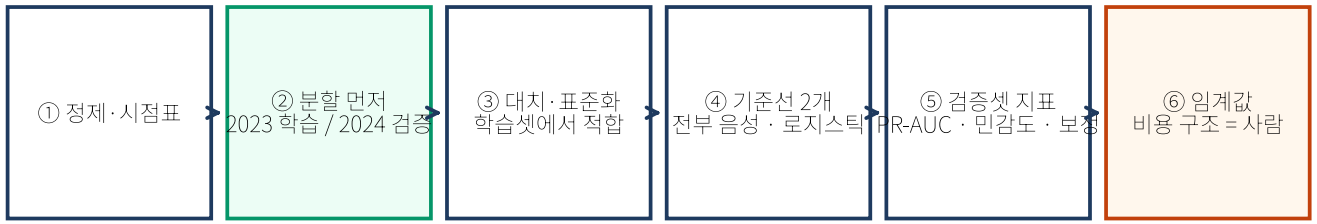
2. 기본 구조

요소	예측모형	RAG
입력	결과 이전 시점의 변수	질문
정답	다음 해 **신고** 산재 (라벨 불완전)	사람이 원전에서 만든 **평가셋** (질문·정답·근거 위치·거절 문항)
분할·검증	시간 분할(2023 → 2024) 또는 사업장 단위	평가셋은 구현에 노출하지 않음
지표	AUC · **PR-AUC** · 민감도·정밀도 · **보정**(Brier, 곡선)	근거 정확도 · **인용 정합률** · 부적절 답변률 · 거절 정확도
사람의 결정	임계값 (비용 구조)	원전 선정 · 정답 · 위험 답변 정의
산출물	자료 흐름+분할 · 성능표(기준선 포함) · 보정 곡선 · 임계값 표 · **누수 점검표**	명세 · 평가셋 · 지표 표 · **실패 사례 표**

2. 기본 구조

요소	예측모형	RAG
보고 지침	TRIPOD+AI	CHART

3. 하는 법 단계별 — 3.1 예측모형



정확도는 보고하지 않는다 — 발생률 4% 에서 전부 '아니오'라 해도 96%

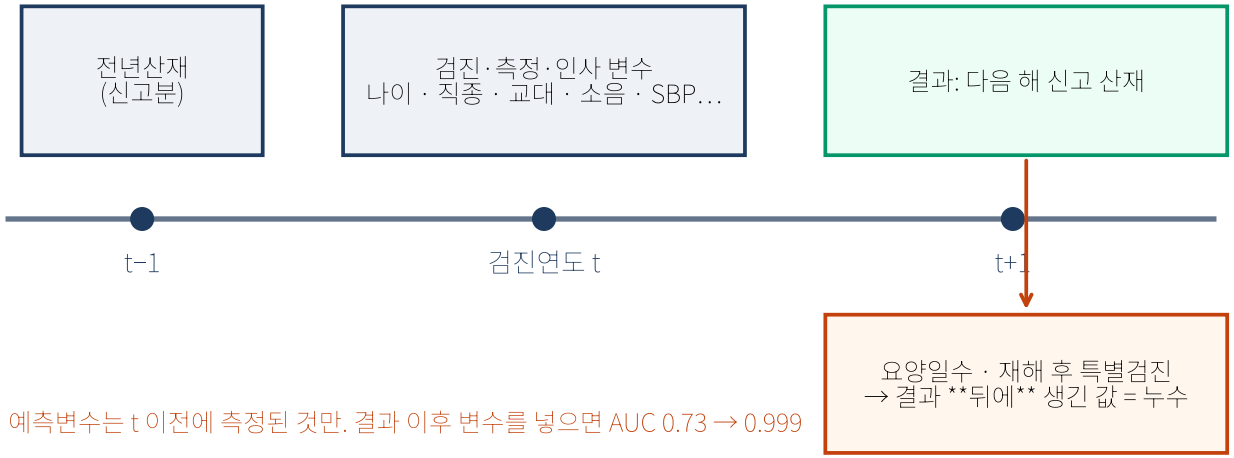
3. 하는 법 단계별 — 3.1 예측모형

단계	내용
① 시점표	모든 변수의 측정 시점. 결과 **이후** 변수(요양일수·재해 후 검진)는 제외
② 분할 먼저	학습 2023 / 검증 2024. 무작위 분할은 같은 사람·사업장이 양쪽에
③ 전처리	대치·표준화는 학습셋에서 적합해 검증셋에 적용
④ 기준선	전부 음성 · 로지스틱. 복잡한 모형은 기준선을 이긴 만큼만
⑤ 지표	발생률 4% → 정확도 무의미. PR-AUC·민감도·정밀도, 그리고 **보정**
⑥ 임계값	민감도–정밀도 표를 놓고 비용 구조로 사람이 결정

3. 하는 법 단계별 — 3.1 예측모형

단계	내용
⑦ 설명	변수 중요도·SHAP 은 "모델이 의존한 변수"이지 원인이 아님

3. 하는 법 단계별 — 3.1 예측모형

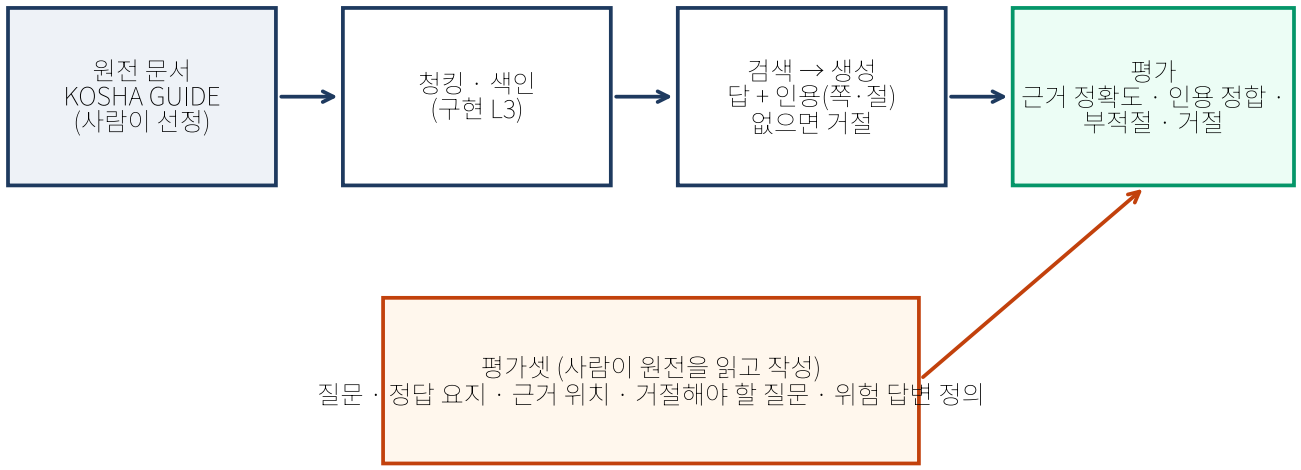


3. 하는 법 단계별 — 3.1 예측모형

보정(calibration) — 순위를 잘 매기는 모델(판별)이 확률을 맞게 말하는 것은 아닙니다. 클래스 가중을 쓰면 예측확률 평균이 0.38 로 부풀어 재보정이 필요합니다.

3. 하는 법 단계별 — 3.2 RAG

챗봇을 만든 것은 연구가 아니다 — 평가셋에 대한 지표와 실패 사례가 연구다



평가셋 파일은 구현 프롬프트에 붙이지 않는다 (테스트 유출)

3. 하는 법 단계별 — 3.2 RAG

단계	내용
① 원전 선정	어떤 문서를 권위 있는 원전으로 인정할지, 현행 개정본인지 — 사람
② 명세	입력·출력·인용 형식·거절 규칙("이 지침에서 답할 수 없습니다")
③ **평가셋**	사람이 원전을 읽고 질문·정답 요지·근거 위치·유형(사실/절차/수치/**거절해야 함**)·위험 답변 정의를 작성. 20문항 이상
④ 구현	청킹·색인·검색·생성 — 위임 L3~L4. **평가셋을 프롬프트에 붙이지 않음**
⑤ 평가	평가셋 전수 실행 → **사람이 판정** → 지표 집계 → 실패 사례 분류(검색 실패/생성 오류/인용 불일치/거절 실패)

3. 하는 법 단계별 — 3.3 명세와 테스트는 사람이

명세와 테스트를 둘 다 AI 에 맡기면 AI 의 오해가 테스트에 그대로 들어가 **통과하는데 틀린** 시스템이 됩니다. 이것이 검증 비대칭의 가장 구체적인 형태입니다.

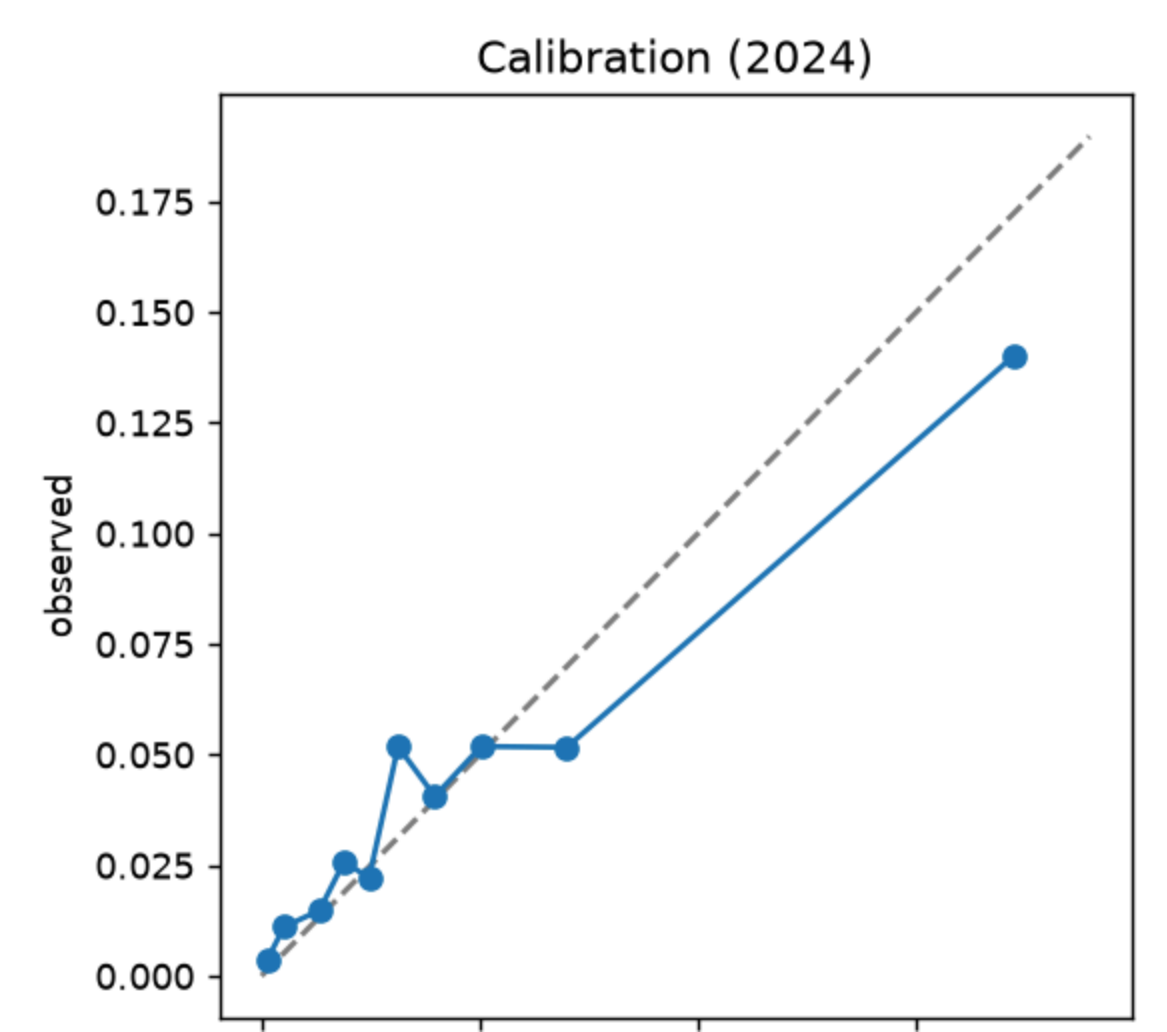
3. 하는 법 단계별 — 3.3 명세와 테스트는 사람이

골든 테스트와 평가셋은
반드시 사람이 만듭니다

4. 시범 — 끝까지 한 번 — 4.1 예측 (`예시/predict_example.py`, 합성 특수건진 연계자료)

모형	AUC	PR-AUC	Brier	예측평균 / 관측
기준선 0 전부 음성	0.50	0.04	—	정확도 0.96
기준선 1 로지스틱 (누수 없음)	**0.73**	**0.13**	0.039	0.043 / 0.041
+ 클래스 가중, 재보정 없음	0.72	0.13	0.192	**0.378** / 0.041
+ 가중 + isotonic 재보정	0.72	0.13	0.039	0.044 / 0.041
누수 실험 결과 이후 변수 포함	**0.999**	0.99	0.003	—

4. 시범 — 끝까지 한 번 — 4.1 예측 (`예시/predict_example.py`, 합성 특수건진 연계자료)



4. 시범 — 끝까지 한 번 — 4.1 예측 (`예시/predict_example.py`, 합성 특수건진 연계자료)

임계값 표 — 0.04 에서 고위험 933명 판정, 실제 재해 70 포착, 42 놓침(민감도 0.62, 정밀도 0.08). 0.12 면 157명 판정, 33 포착, 79 놓침. **어느 것을 고를지는 놓친 재해와 헛경보의 비용이 정합니다.**

4. 시범 — 끝까지 한 번 — 4.2 RAG (`예시/mini_rag.py` , KOSHA GUIDE G-12-2013 개인보호구 지침)

청크 42개, TF-IDF 검색 + 생성. 평가셋 예시 10문항(사실 7 · 절차 1 · 수치 1 · **거절해야 함 1**). "특정 유기용제의 TWA 는 몇 ppm 인가" — 이 지침에 없으므로 거절이 정답. 지어내 답하면 가장 위험한 실패입니다. 결과는 사람이 판정한 뒤 `score.py` 로 집계합니다.

5. 주의점과 흔한 오류

오류	무엇이 잘못인가	대신
AUC 0.99 를 성과로	결과 이후 변수(누수)	시점표 · 성능이 좋으면 의심
정확도 96%	전부 음성도 96%	PR-AUC·민감도·보정
무작위 분할	같은 사업장·사람이 양쪽	시간·사업장 단위 분할
대치를 분할 전에	전처리 누수	학습셋에서 적합
가중 후 확률 그대로	예측평균 0.38 vs 관측 0.04	재보정
"AI 가 최적 임계값 제시"	비용 구조는 사람의 판단	표를 놓고 사람이

5. 주의점과 흔한 오류

오류	무엇이 잘못인가	대신
SHAP 상위 = 원인	모델이 의존한 변수	인과는 Part I
라벨을 진실로	신고된 산재	사업장별 신고율 제한점
"이 문서로 질문 20개 만들어 줘"	시스템이 잘 답하는 질문만	사람이 원전 읽고 작성
평가셋을 구현 프롬프트에	테스트 유출	분리
"잘 작동한다"	측정 없음	지표 표 + 실패 사례 표
인용이 있으니 맞다	인용 조문이 그 내용을 안 담음	인용 정합률 (9장 9.3 의 2차 오류)

6. AI 와 함께 — 위임 가능 과업과 검증

> 자율성 수준 L0(쓰지 않음) · L1(보조 — 사람이 전부 검토) · L2(초안 — 전면 재검토) · L3(위임 — 표본 검수) · L4(자동 — 테스트 통과 시 채택). 정의는 Part 0 §2.2, 약어는 `00_공통/용어집.md`.

과업	수준	무엇을 맡기나	틀리면 어떻게 알아채나
명세 문장화	L2	사람이 쓴 요구사항 → 명세 문서	요구사항에 없는 기능 추가 여부
평가셋·골든 테스트	**L0**	—	사람이 원전에서
파이프라인·RAG 구현	L3–L4	명세 + 인터페이스 (평가셋 미첨부)	골든 테스트 통과 · 합성자료 성능 범위 복원 · 누수 assert
누수 점검표 초안	L3	7.3 네 유형별 코드 줄 인용	판정은 사람. AI 는 자기 코드에 관대
보정 곡선·재보정	L3	10분위 표·isotonic	재보정 전후 예측평균
임계값 표 재배열	L2	"개입 가능 인원 N" 기준으로	추천 문장 삭제. 선택은 사람

6. AI 와 함께 — 위임 가능 과업과 검증

과업	수준	무엇을 맡기나	틀리면 어떻게 알아채나
실패 사례 분류	L2	사람 판정 결과를 유형화	판정을 바꾸지 않았는가
정답 판정 · 임계값 · 원전 선정	L0	—	사람

프롬프트 원문은 `AI실습/00_README.md` (P-E01~E07) 과 `../part1_2차자료연관분석/_attic/AI실습_특수건진/03_모형과_검증.md` 의 P-B20~B28.

7. 실습 — 예시 따라 하기

회차	하는 일	주는 것 (발판)
12주	특수건진 코드북으로 변수 시점표 작성 → 계획 커밋(분할·기준선·지표·임계값 정책) → `predict_example.py` 실행 → **누수 실험** → 자기 파이프라인(assert·누수 점검표·보정 곡선·임계값 표)	예시 코드·명세만
13주	KOSHA GUIDE 원문 읽고 **평가셋 10문항 이상 직접** → 명세 → `mini_rag.py` 구조 확인 → 자기 구현	예시 코드·평가셋 양식
14주	평가셋 전수 실행 → 사람 판정 → `score.py` → 실패 사례 분류 → 방법 사실 목록 → 문장화	—
제출	P5 결과 보고 V	루브릭 `30_실습.md`

8. 되돌아보기

- 오늘 잡은 오류 — 성능이 좋아서 의심한 순간이 있었는가
- 내가 만든 평가셋에서 시스템이 틀린 문항은 어떤 유형이었는가
- 다음 Part 로 가져갈 것: **정답을 정의하는 일의 값** — 계획서에서는 그것이 검증 계획입니다

참고문헌

- Lewis et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. (NeurIPS 2020 proceedings — Crossref 미등재. arXiv 2005.11401. DOI 없음)
- Collins et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. doi:10.1136/bmj.q902
- Huo et al. (2025). Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. doi:10.1136/bmjmed-2025-001632
- Saito et al. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. doi:10.1371/journal.pone.0118432
- Van Calster et al. (2019). Calibration: the Achilles heel of predictive analytics. doi:10.1186/s12916-019-1466-7
- Kaufman et al. (2012). Leakage in data mining: formulation, detection, and avoidance. doi:10.1145/2382577.2382579

(실존 확인: `reference/refs_verified.csv`)

가져갈 다섯 줄

1. 예측은 연관과 다른 질문이다. 결과 이후 변수는 어디에도 못 들어간다
2. **성능이 좋으면 누수를 의심한다.** 분할이 먼저, 시간·사업장 단위로
3. 정확도는 쓰지 않는다. PR-AUC·민감도·**보정**·비용 구조의 임계값
4. **명세와 골든 테스트·평가셋은 사람이 만든다.** 둘 다 맡기면 검증이 무력화된다
5. 만든 것은 연구가 아니다. **평가셋에 대한 지표와 실패 사례**가 연구다