한국보건의료기술평가학회 2025년 후기 학술대회

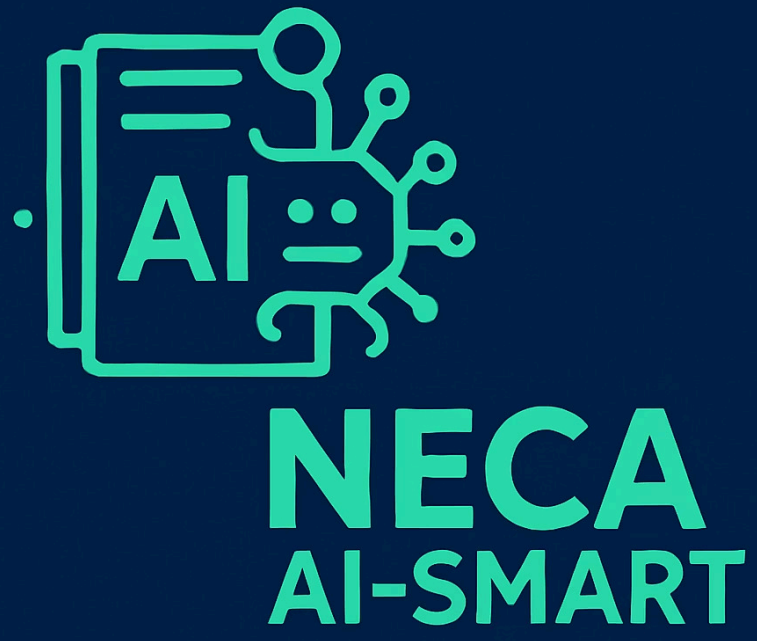
AI 시대의 HTA: Systematic review에 AI가 어떻게 활용되고 있나?(part 2)

NECA AI-SMART SR

: 체계적 문헌고찰 자동화를 향한 실무적 접근

김세희 (한국보건의료연구원 / 보건의료연구팀)

This study was supported by the National Evidence-based Healthcare Collaborating Agency (grant no. NECA-A-24-014 & Medical Advertisement Compliance Monitoring Project).



발표 개요

01

연구의 필요성

02

LLM 자동화 가능성 선행문헌고찰

03

LLM 자동화 가능성전문가 자문

04

NECA AI-SMART SR 소개

05

연구 워크플로우를 위한 AI 도구+ 실무적용 로드맵

미래 보건의료 연구의 필수 과제: 체계적 문헌고찰(SR) 혁신

- 급변하는 보건의료 환경: **근거 중심 의학(EBM) 중요성 증대**
- **체계적 문헌고찰(SR)**: 최고의 근거 종합 **핵심 방법론**
- **방대한 문헌 처리와 인력 소모**: 효율성 및 확장성 **한계 직면**

방대한 문헌 처리 비효율성

기하급수적 문헌량 증가로 고품질 SR 수행에 과도한 시간과 인적 자원 필요. 신속한 근거 생산 저해 요인

연구 일관성 및 신뢰도 저하

연구자 간 주관적 판단 개입으로 결과 일관성 결여. 인적 오류 발생 가능성 상존, SR 학술적 엄밀성 우려

정보 과부하 및 연구자 피로도 증대

정보 과부하로 연구자 인지적 부담 가중. 피로도 증가 및 집중력 저하, 연구 품질 영향

AI 기반 자동화 도입의 효율성

고품질 SR 신속·효율적 수행 및 연구 정확성·신뢰성 획기적 향상. AI 기반 혁신 솔루션 도입 필수

> "**연구자의 심층 전문성은 유지**하되, **반복적이고 단순한 작업은 AI가 담당**하는 **혁신 협업 패러다임** 요구."

체계적 문헌고찰(SR)의 당면 과제와 AI 혁신의 필요성

체계적 문헌고찰(SR) 수행의 현실적 한계 분석

시간적 비효율성 심화

- SR 195건 분석, 등록부터 출판까지 평균 **67.3주 소요**
- SR 수행에 **과도한 시간·인력 소요**. 증가하는 데이터 양 대비 **능동적 업데이트 어려움** (Borah et al., 2017)

정확성 및 포괄성 제약

- 인간 단독 스크리닝 시 관련 문헌 **누락 가능성 증대**
- 2인 검토자 합의에도 관련 논문 **누락 사례 발생** (Scherbakov et al., 2025)

일관성 유지의 어려움

- 평가자 판단 편차, 누적 피로, 주관적 개입으로 **결과 일관성 저하**
- **'결정 피로'**: 의사결정 신뢰도 저하, 판단 정확성 감소, 오류 증가 초래 (Grignoli et al., 2025)
- 연속 사례 처리 시 평가자 **판단 기준 일관성 결여** 또는 보수적 변화 (Maier M et al., 2025)

미래지향적 AI 도입의 당위성

- 거대언어모델(LLM), 방대한 텍스트에서 **핵심 의미 추출** 및 **포함/배제 기준 자동화**로 SR **효율성 혁신적 증대 가능**
- "SR의 근본적 병목 현상은 기술 문제 아닌, **인간 인지적 한계를 넘어선 전문성 요구**에 기인."

Part 1. LLM 자동화 가능성 선행문헌고찰

Systematic Review of LLM Automation Potential Across 8 SR Stages

체계적 문헌고찰 8 Stages (자동화 업무 단위별)

연구질문 정하기

PICO 구조 설정

검색 전략 수립

데이터베이스·MeSH 용어 설계

문헌 중복 정리

DOI·저자·연도 기반 중복 제거

제목·초록 선별

포함/배제 기준 적용

전문 선별

RAG 기반 전문 판독

자료 추출

표준화된 템플릿 적용

비뚤림 위험평가

RoB2 기준 평가

결과 합성 및 보고

메타분석·PRISMA 보고

체계적 문헌고찰 8 stages: AI/LLM 기여 가능성 (1)

1

1. 연구 질문 정하기

임상적·정책적 필요에 맞춰 **연구 질문 명확화**. 특히 **PICO (Patient/Population, Intervention, Comparison, Outcome) 구조 활용**

AI/LLM 기여: **관련 문헌 분석**을 통한 **질문 구체화 제안**

AI 예: **표준 용어 추천**, **유사 질문 사례 제시**, **PICO 구조화 초안 작성 가능**

2

2. 검색 전략 수립

데이터베이스 및 **검색식 설계**. **표준 용어** (예: **MeSH** (표준 의학 주제어 체계))와 **동의어 반영**

AI/LLM 기여: **키워드** 및 **검색식 자동 추천**, **MeSH 용어 제안** 지원

AI 예: **동의어/관련어 제안**, **검색식 초안 작성** 및 **오류 점검**

3

3. 문헌 중복 정리

다중 **데이터베이스**에서 수집된 문헌 **중복 제거**. 컨퍼런스 초록, 학회 초록, 저널 등 다양한 형태 포함

AI/LLM 기여: 검색 문헌의 **중복 자동 식별 및 제거**

AI/프로그램 예: **DOI/PMID, 제목, 저자 매칭 자동화**, 애매한 사례 후보 표시

4

4. 제목·초록 선별

포함/제외 기준에 따른 **1차 스크리닝 진행**

AI/LLM 기여: **포함/제외 기준 학습** 후 **1차 선별 자동화**

AI/프로그램 예: **DOI/PMID, 제목, 저자 매칭 자동화**, 애매한 사례 **후보 표시**하여 검토 지원

체계적 문헌고찰 8 stages: AI/LLM 기여 가능성 (2)

5

5. 전문(full text) 선별

1차 선별 문헌의 **전문 검토, 최종 포함 여부 결정**. 이 과정에서 **포함/제외 기준 재적용**

AI/LLM 기여: **전문 검토** 통한 **최종 포함 문헌 결정 보조**

AI 예: **전문 내용 기반 포함/제외 기준 위배 여부 자동 판단**(일부), **애매 사례 후보 제시**

6

6. 자료 추출 (안전성 분석)

최종 포함 문헌에서 **연구 설계, 환자 특성, 중재 내용, 주요 결과** (특히 **안전성 정보**) **추출**

AI/LLM 기여: 문헌에서 **핵심 정보** 및 **결과 데이터 자동 추출**

AI 예: **관심 정보 자동 추출**(테이블/텍스트), **이종 서식 자료 표준화, 추출 가이드라인 기반 오류 점검**

7

7. 비뚤림 위험 평가

포함 연구의 **방법론적 질 평가, 결과 신뢰성 검토** (**RoB 2.0** 등 **도구 활용**)

AI/LLM 기여: **비뚤림 위험 평가 도구 사용 지원, 보고서 초안 작성**

AI 예: **RoB 도구별 항목 초안 작성, 근거 문장 제시, 일관성 없는 평가 사례 표시**

8

8. 결과 합성·보고

추출 자료 정성적/정량적(메타분석) 합성, 체계적 문헌고찰 보고서 작성 (**PRISMA** 등 **지침 준수**)

AI/LLM 기여: **결과 요약** 및 **보고서 초안 생성, PRISMA 체크리스트 자동 작성**

AI 예: **메타분석 결과 해석 요약, 초안 보고서 작성(PRISMA), 주요 결과 시각화**

Stage 2. 검색 전략 수립(1)
- De Cassai A, et al. (2025)

1

연구 배경

체계적 문헌고찰(SR) 검색 전략 수립에 LLM 활용 가능성 평가

2

연구 대상

마취과학 상위 10개 저널에 발표된 85개 체계적 문헌고찰(SR) 분석 (2024년 1월-8월)

3

연구 설계

- **세 가지 비교 그룹**:
  - **원 논문 검색어** (인간 전문가 벤치마크)
  - **ChatGPT 4o** (범용 LLM)
  - **Meta-Analysis Librarian** (PICO 기반 특화 모델)
- 각 SR당 4개 검색어 생성 (ChatGPT 3개 + Librarian 1개)
- PubMed 검색 후 원 참고문헌 대비 검색률 측정

4

연구 질문

"LLM이 체계적 문헌고찰 검색 전략 수립에서 인간 전문가를 대체할 수 있는가?"

Stage 2. 검색 전략 수립 (2)

- LLM별 검색전략 수립 방법론 비교 (비구조화 vs. 구조화)

ChatGPT 4o (비구조화 방식):

입력 프롬프트:

- "Provide a search string for the paper titled '[SR 제목]'"
- "Provide another string"
- "Provide another string"

방법론 특징:

- × **추가 지침 없는 대화형 접근**
- × **일반 AI 능력에만 의존**
- × **무작위적 키워드 변형**
- × **결과 일관성 낮음**
- × **전문 지식 미반영**

Meta-Analysis Librarian (구조화 방식):

체계적 프로세스:

1. **Step 1: PICO 프레임워크 구축**
2. **Step 2: 검색어 생성**
   - **PICO 요소 기반 키워드** 추출
   - **Boolean 연산자** 정밀 적용 (AND/OR/NOT)
   - **와일드카드(*)**, **절단 기호** 활용
3. **Step 3: 반복 개선**
   - **Cochrane Handbook 표준** 준수
   - **증거기반 검색 전략** 최적화

방법론 특징:

- ✓ **체계적이고 구조화된 접근**
- ✓ **국제 표준 가이드라인 준수**
- ✓ **전문 사서처럼 협력적 개선**
- ✓ **높은 일관성**과 **재현성**

Stage 2. 검색 전략 수립 (3) - 연구결과 및 시사점

원 검색어 (인간 전문가):
65% 중앙값 (IQR: 43-81%). 통계적 유의성 기반 최우수 (p=0.001).

Meta-Analysis Librarian:
24% 중앙값 (IQR: 13-38%). ChatGPT 4o 대비 **4배 높은 성능**, **7건(25%)에서 인간 전문가와 동등 또는 초과** 결과 달성.

ChatGPT 4o:
6% 중앙값 (IQR: 0-14%). **최저 성능** 및 높은 변동성.

결과 기반 **핵심 발견 사항 및 시사점**

구조화 방법론의 중요성

- ✓ **PICO 프레임워크 기반 접근** 핵심
- ✓ **도메인 특화 모델**, 범용 LLM보다 **4배 우수**
- ✓ **체계적 가이드라인 준수**, 성과 결정 요인

현재 실무 활용 전략

- **보조 도구로 활용** (완전 대체 불가)
- **초기 검색전략 개발의 출발점**
- 반드시 **인간 전문가 검토 병행 필요**
- 다중 데이터베이스 적용 시 주의 요망

미래 발전 가능성

- **모델 성능 지속 향상** (GPT-3.5→4o)
- **생의학 특화 LLM 개발** (BioGPT 등)
- 모델 미세조정 및 인간 피드백 학습
- 향후 **인간 전문가 수준 근접 전망**

결론:
"구조화된 프롬프트와 도메인 특화 모델 사용 시 SR 검색전략 자동화의 실용적 가능성 확인. 현재는 인간-AI 협력 모델이 최적"

Stage 4/5. 제목·초록, 전문 선별(1)

- Cao et al. (2025)

Stage 4: Abstract Screening

- **프롬프트:** Abstract ScreenPrompt
- **민감도:** 97.7% (86.7-100%)
- **특이도:** 85.2% (68.3-95.9%)
- **테스트:** 48,425개 인용문헌

Stage 5: Full-text Screening

- **프롬프트:** ISO-ScreenPrompt
- **민감도:** 96.5% (89.7-100%)
- **특이도:** 91.2% (80.7-100%)
- **테스트:** 12,690개 전문

💰 **비용 절감: 초록 선별 90%↓, 전문 선별 98%↓**

⏱️ **시간 단축: 83시간 → 24시간 이내 단축**

📊 **잘못 제외율 감소:**

초록 선별: 12.5개 → 0.5개

전문 선별: 5.5개 → 0개

Stage 4/5. 제목·초록, 전문 선별(2)

통합 성능 평가 및 실무 적용 전략(초록 vs 전문 스크리닝 성능 비교)(Cao et al. (2025))

지표	Zero-shot	Abstract SP	Zero-shot	ISO-SP
민감도	49.0%	97.7% ⬆️	49.1%	96.5% ⬆️
특이도	97.9%	85.2%	97.5%	91.2%
잘못 제외	12.5개	0.5개	5.5개	0개
평가 범위	전체	전체	7.56%*	7.56%*

*무료 PMC 전문만 평가 가능

모델 성능 비교

🥇 **GPT-4 variants**: 93-95% (최고 성능)

🥈 **Claude-3.5-Sonnet**: 93-94% (**GPT-4와 유사**) ⭐

🥉 **Gemini Pro**: 67-77% (중간 성능)

📉 **GPT-3.5**: 69-78% (저성능)

---

stage별 실무 적용 전략

Strategy 1: 초록 스크리닝 (2가지 경로)

Path A: "독립 검토자"
인간 1명과 **협업** 또는 **완전 자동화**
작업 부담 50% 감소

Path B: "사전 스크리닝" (권장)
2명의 인간 검토 전 **AI 필터링**
스크리닝 양 **66-95% 감소**
누락 위험 **최소화 (0.5개)**

Strategy 2: 전문 스크리닝 (제한적 적용)

⚠️ **현실적 제약**:

무료 전문 접근 7.56%만 가능

페이월/저작권 문제 존재

초록 우회 **시기상조**

💡 **권장 접근**: "초록 사전스크리닝 → 인간 초록 검토 → 전문 검토"

1️⃣ 초록 스크리닝:
즉시 **실무 적용 가능**

97.7% 민감도와 **90% 비용 절감** 효과 기대

2️⃣ 전문 스크리닝:
기술적 우수하나, **접근성 제약** 존재

3️⃣ 단계별 전략:
리소스와 위험 허용도에 따라 **구현 경로 선택 필요**

Stage 4/5. 제목·초록, 전문 선별(3)

프롬프트 엔지니어링 6단계 반복 개발(Cao et al. (2025))

Stage 1) Zero-shot
30% 민감도 달성
기본 지시사항만 제공

Stage 2) Few-shot
라벨링 예시 추가
성능 개선 제한적

Stage 3) Zero-shot CoT
74.5% 민감도 달성
"단계별 사고" 지시 추가

Stage 4) Framework CoT
86.5% 민감도 달성
구조화된 추론 프레임워크 적용

Stage 5) ScreenPrompt
89% 민감도 달성
SR 목표 명시적 포함

Stage 6) Abstract ScreenPrompt
94.5% 민감도, 94% 특이도
초록 특화 지침 추가

핵심 요소:

- ✓ SR 목표 그대로 포함
- ✓ 포함/제외 기준별 단계별 추론
- ✓ 불확실시 포함(inclusive) 원칙
- ✓ 초록 간결성 고려한 맥락 제공

Stage 6. 자료 추출(1)

전통적 자료 추출의 문제점

- **가장 노동 집약적인 단계** (수개월 소요)
- **수동 추출 오류율**: 최대 50%
- **높은 비용** (연구당 수천 달러)
- **Living SR 유지의 주요 병목 구간**

단일 LLM 사용의 위험성

- **환각 현상 (Hallucination)**: 2-3% 발생
- **신뢰성 검증** 어려움
- **일관성 부족**

Khan et al. (2025) 혁신적 접근법

📄 **Title**: Collaborative LLMs for Automated Data Extraction

👥 **Authors**: Khan, Ayub, Riaz et al.

🏥 **Institution**: Mayo Clinic, ASU

📊 **Dataset**: 10개 임상시험, 22개 출판물, 23개 변수 분석

2-Reviewer 협업 LLM 워크플로우

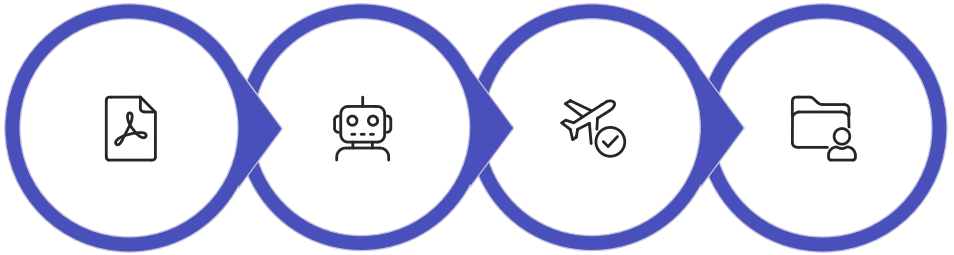
- **GPT-4-turbo** + **Claude-3-Opus** 활용
- **실제 체계적 문헌고찰** 프로세스 모방
- **일치도 평가** (Concordance) 및 **교차 비판** (Cross-critique) 도입

Stage 6. 자료 추출(2)

Khan et al. (2025)

협력적 자료 추출 워크플로우

Collaborative LLM Data Extraction Pipeline



📊 **Key Performance Metrics**

- **일치도 정확도: 0.94**
- **환각(Hallucination) 발생률: 0.25%**
- 논문당 비용: **$3.40**
- 처리 속도: **실시간**

"시스템 템플릿 활용 **Zero-shot 프롬프팅**"

"숫자 **정확히 일치**, 텍스트 **의미론적 일치**"

"LLM 간 **상호 답변 비판**"

Stage 6. 자료 추출(3)

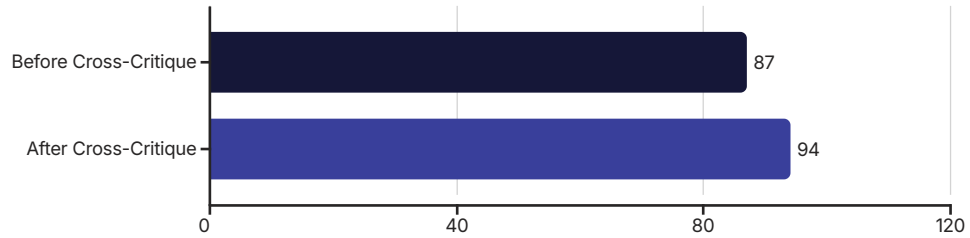
협력적 LLM 접근법의 성과와 향후 방향 (Khan et al. (2025))

단일 LLM과 상호협력형태 LLM 접근법 비교 (Test Set, n=391)

Approach	Accuracy	Precision	Recall	F1 Score	Hallucination
GPT-4 alone	89%	98%	87%	92%	2.29%
Claude-3 alone	90%	94%	91%	92%	2.76%
🏆 **Concordant**	**94%***	**100%***	**92%***	**96%***	**0.25%**
Discordant (GPT)	41%	54%	46%	50%	26.93%
Discordant (Claude)	50%	60%	72%	65%	41.00%

***p<0.05 vs 단일 LLM 접근법**

교차 확인의 효과



49개 불일치 중 **25개(51%) 일치로 전환**

💡 핵심 발견

일치 응답: 94% 정확도

환각(Hallucination) 감소

(2.5% → 0.25%)

비용 효율성

($3.40/paper + $2.06/critique)

거의 실시간 처리

Stage 6. 자료 추출(4)

🚀향후 연구 방향 및 적용(Khan et al. (2025))

📚 확장성

- **다른 의학 분야** 검증
- **문헌 스크리닝** 효율화
- **품질 평가** 자동화
- **GRADE 평가** 자동화

🔧 기술 개선

- **3인 검토자** 도입
- **추론 과정** 분석
- **메타분석** 계획 지원
- **템플릿** 확장

🏥 실무 적용

- **살아있는 증거 프레임워크** 통합
- **임상 진료 지침** 업데이트
- **실제 적용** 사례 개발
- **교육 프로그램** 개발

"협력적 LLM, 추후 **'Living' Systematic Reviews** 가능"

→ **높은 정확도 + 낮은 비용 + 실시간 처리**

Stage 7. 비뚤림 위험평가

Huang et al. (2025), Lai et al. (2024) 등 연구의 **비뚤림 위험(Risk of Bias 2)을 자동 평가하는 모듈** 개발

LLM **세분화된 프롬프트(QA 방식)**로 Reviewer 수준의 **판단 정확도(57–70%)** 및 **일관성(Cohen kappa)** 달성.

🎲 무작위 배정

무작위 순서 생성 및 배정 은폐

✅ Low Risk

💉 중재 실행

중재 시행 및 참여자 눈가림

⚠️ **Some Concerns**

📉 결과 누락

탈락 및 결측 처리

✅ Low Risk

👁️ 결과 측정

평가자 눈가림 및 측정 오류

❌ High Risk

📑 결과 선택

선택적 보고 / 분석 편향

✅ Low Risk

> 💬 **자동 판정 근거 예시 (LLM 출력 요약)**
>
> **Reasoning:** Randomization clearly described (✓) / Double-blind design (✓) / Outcome assessor not blinded (×) / Missing data <5% (✓) → **Overall judgment: "Some Concerns"**

이러한 **평가 프로세스 자동화**, 기존 사람 방식 대비 **속도와 객관성 향상**

판정 기준 및 설명(예: 'Low risk?'의 이유) 자동화 기능도 보고

Stage 8. 결과 합성·보고(1)

Meta-Analysis에서 ChatGPT-4o의 정량적 합성 능력(Purewal et al. (2025))

연구 배경

- Systematic Review 자동화의 마지막 단계: **결과 합성 및 메타분석**
- 통증의학 분야 **발표 메타분석** 기반 **검증 연구** (Klasova et al., 2024)
- **ChatGPT-4o**의 **데이터 통합** 및 **forest plot 생성 능력** 최초 평가

평가 과제

- **통합 평균 차이(Pooled Mean Difference, MD)** 및 **95% 신뢰 구간(CI)** 계산
- **이질성 추정치(Heterogeneity estimate, I^2 score)** 산출
- **타우 제곱 값(Tau-squared value, τ^2)** 계산
- **Forest plot** 생성

데이터 출처

- **척수 자극기(Spinal Cord Stimulation)** 연구 **5가지 결과 지표** (outcome)
- 결과 지표: **불안(HADS-A)**, **우울(BDI)**, **전반적 기능(GAF)**, **정신 건강(SF-36 MCS)**, **통증 파국화(PCS)**
- 각 결과 지표별: **8-23개 연구/하위 그룹**
- 분석 방법: **랜덤 효과 모델(Random-effects model)**을 사용한 통합 효과 추정

비교 기준

- **사람이 직접 수행한 메타분석(Human-executed meta-analysis)**을 **골드 스탠다드(gold standard)**로 설정

Research Question

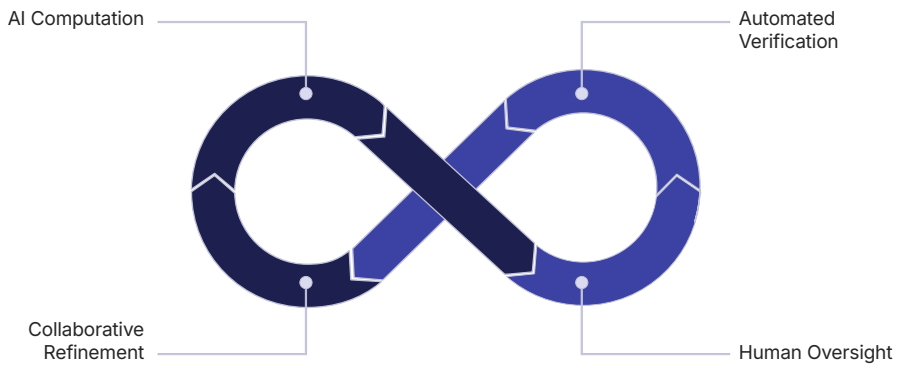
Can **ChatGPT-4o** accurately perform **quantitative synthesis tasks** in systematic reviews?

Stage 8. 결과 합성·보고(2)(Purewal et al. (2025))

📊 성능 결과 (5개 Forest Plot 기준)

지표	정확도	비고
Pooled MD	100% (4/5)	1개 outcome: **0.01 차이**
95% CI	100% (5/5)	**완벽한 일치**
I^2 Score	100% (5/5)	**완벽한 일치**
τ^2 Value	40% (2/5)	**0.01-0.05 범위 차이**
Forest Plot	100% (5/5)	**시각적 차이 없음**

💡 핵심 발견: AI + Human Collaboration



✅ 주요 강점

- Meta-analysis 계산에서 **거의 완벽한 정확도** 달성
- 계산 오류 **감소** 및 **시간 절약** 효과
- Forest plot의 **우수한 시각적 품질**

⚠️ 현재 한계점

- Tau-squared value의 **미세한 불일치**
- **학습 데이터 오염 가능성**
- 새로운 데이터셋에 대한 **추가 검증 필요**

향후 연구 방향

- 미발표된 메타분석 대상 **동시 평가 필요**
- 다양한 **의료 분야로 확대 검증**
- Stage 8 자동화를 위한 **표준 프로토콜 개발**

Part 2. 8 Stages별 LLM 자동화 가능성 전문가 의견

Systematic Review of LLM Automation Potential Across 8 SR Stages

Part 3. NECA AI-SMART Project

Towards AI-Integrated Systematic Reviews for Evidence Acceleration

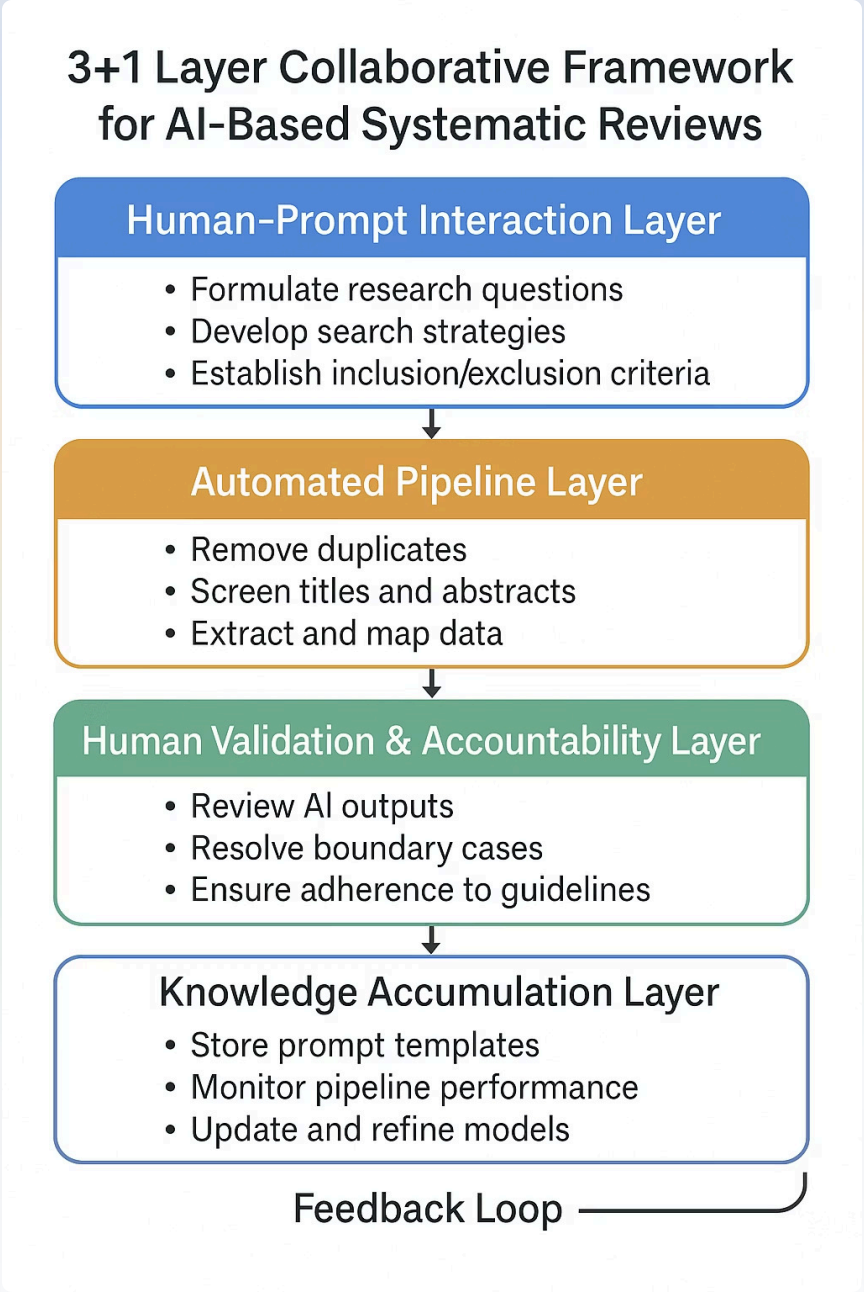
SMARTER FRAMEWORK는 NECA AI-SMART SR의 설계기반모델

SR 전주기 통합연계모델 구축를 위한 SMARTER FRAMEWORK

Strategy **M**apping → **A**utomated **R**eview → **T**ested Validation → **E**volutionary **R**efinement

핵심 구성

1. **인간-프롬프트 상호작용층 (Human-Prompt Interaction Layer)**: **AI와 효과적인 소통** 통한 전문성 구현
2. **자동화 파이프라인층 (Automated Pipeline Layer)**: **AI 기반 자동화**로 효율성 및 생산성 극대화
3. **인간 검증층 (Human Validation & Accountability Layer)**: **AI 결과의 학술적 엄밀성** 및 **신뢰성** 확보
4. **지식 축적층 (Knowledge Accumulation Layer)** (지속 학습): **자율적 학습** 통한 시스템의 **지속적 고도화 달성**



1. 사람-프롬프트 상호작용층

Human-Prompt Interaction Layer: 전략적 탐색·기획 중심층

핵심 특성

- **연구자의 인지 능력**과 **AI의 언어 분석 능력 만남**
- **인간**의 방향 제시, **AI**의 구체화 — **전략적 조율층**
- **지시-피드백 루프** 통한 지속적 **최적화**
- **근거**: Delgado-Chaves(2025) 연구서 **인간 감독의 필수성** 강조

주요 업무

- **PICO 질문 구조화**
  → 예: "**제2형 당뇨 환자**에서 **메트포민**의 **심혈관 효과**는?"
- **검색 전략 설계**
  → **MeSH terms**: "Diabetes Mellitus, Type 2"[MeSH]
  → **Boolean**: (diabetes AND metformin AND cardiovascular)
- **포함/제외 기준 초안 작성**
  → **RCT**만, 추적기간 **≥12개월**, **영어/한국어** 논문
- **피드백 루프 형성**
  → **연구자**: "stroke 포함, 관찰연구 제외"
  → **AI**: 검색식 수정 → **3-5회 반복 최적화**

"이를 기반으로 포괄적 검색식 생성 가능"

핵심: 인간은 **방향성**과 **판단** 제시. AI는 단순한 도구를 넘어, 연구자의 **전략적 사고 증강 파트너 역할**

2. 자동화 파이프라인층

Automated Pipeline Layer: 대량처리·효율성 중심층

핵심 특성

- 프롬프트 및 기준 시스템화로 **대규모 병렬 처리** 가능
- **LangChain, LlamaIndex, API, NLP, Vision AI** 등 최신 기술 활용
- 근거: Guo et al.(2024) 연구에서 **정확도 77.3%, 정밀도 83%, 재현율 86%** 달성.

주요 성과

- **검색·중복제거:** 8시간 → **15분 (96% 단축)**
- **1차 스크리닝:** 40시간 → **2시간 (95% 단축)**
- **데이터 추출:** 60시간 → **4시간 (93% 단축)**

실제 적용 사례

- **4개 DB 동시 검색**으로 **총 2,847편 수집**
- **DOI/제목 매칭**을 통해 **1,923편**으로 정제
- **AI 스크리닝: Include 156편, Exclude 1,654편, Uncertain 113편**
- **샘플수, 중재용량, RR, p-value** 등 구조화된 **Excel 자동 추출 완료**

"이를 기반으로 검색–선별–추출 통합 모듈을 개발 가능"

핵심: 인간의 시간 및 인지 부담 해방 — "고속 병렬처리 엔진" 역할. 반복 노동 제거, 연구자는 해석 및 의사결정에 집중

3. 인간 검증층

Human Validation & Accountability Layer: 품질·책임 중심층

핵심 특성

- **AI 자동화 결과** 전문가 재검토 및 승인
- 오류 및 편향 탐지, **임상적 타당성 확보**
- 법적 책임 및 **투명한 보고 보장**
- 근거: Woelfle et al.(2024) 연구, PRISMA 89-96%, AMSTAR 91-95% 달성

검증 프로세스 예시

Uncertain 케이스 재검토

- AI 판단 애매한 **113건** 수동 검토
- 예: "cardiovascular events"에 MACE 포함 여부 확인

무작위 샘플 검증

- 포함 논문 **10% 샘플링** (총 16편)
- **독립 검증자 2인** 교차 검증
- 검증자 간 일치도: Cohen's Kappa = **0.91**

비뚤림 위험 평가

- **ROB 2.0 도구** 적용
- AI 초안 작성 후 **인간 검증 및 수정**
- Supplementary material (보충 자료) 확인

PRISMA 준수 점검

- PRISMA **27개 항목 체크**
- AI, 검색 전략, 플로우차트 등 구조적 부분 지원
- 인간, 제한점, 임상적 의의 등 해석적 부분 담당

"이를 기반으로 비뚤림 위험 평가 보조 도구 개발 가능"

핵심: AI 효율성-인간 판단력 결합, **"책임의 최종 방어선"** — 자동화 속도, **임상적 타당성 및 윤리적 책임**으로 균형

4. 지식 축적층

Knowledge Accumulation Layer: 순환적 학습 구조

핵심 특성

- **결과물과 피드백 축적으로 지속 개선**
- SR 경험 누적으로 다음 SR **성능 강화**
- **자가 개선 시스템** (Self-improving system) 구현

학습 사이클

1. **오류 패턴 분석**: False Negative 8건, False Positive 3건 기록
2. **프롬프트 자동 갱신**: 동의어 추가로 재현율 **4.3% 향상**
3. **템플릿 재사용**: 초기 설정 시간 3일 → 4시간 (**87% 단축**)
4. **성능 모니터링**: 정확도 카파(Kappa) 점수 0.78 → **0.91 상승**

피드백 루프

- **검증층 → 상호작용층**: 'elderly' 오류 발견으로 Uncertain 케이스 **21% 감소**
- **파이프라인층 → 축적층**: 'metformine' 철자 오류 발견으로 재현율 **3.1% 증가**
- **축적층 → 상호작용층**: 24개월 기준 제안 연구자 **채택**

누적 성과 예시

- 📚 **재사용 템플릿**: 47개
- 🗄️ **오류 학습 DB**: 326건
- ⏱️ **처리 시간**: 108시간 → 18시간 (**83% 단축**)
- 🎯 **정확도**: 0.78 → **0.91**

"이를 기반으로 순환적 지속 학습 모듈을 통하여 성능 개선"

핵심: 모든 연구가 다음 연구의 자산이 되는 "자가 진화 학습 생태계" — 경험을 데이터로, 오류를 개선의 씨앗으로 전환하는 순환적 지능

NECA AI-SMART SR project 소개

NECA AI-SMART의 핵심 비전

"AI + 연구자 = 차세대 SR 역량"

프로젝트의 목표

NECA AI-SMART SR은 2025–2027년 단계별로 **SMARTER Layer 구조를 적용한** 국내 표준 모델 개발을 추진할 예정

NECA AI-SMART 프로젝트의 핵심 가치

AI는 연구 효율을 극대화하는 **강력한 도구임**

하지만 **최종적 전문 판단**과 **학술적 책임**은 변함없이 인간 연구자에게 있음

NECA AI-SMART SR: 체계적 문헌고찰 자동화 프로그램

NECA AI-SMART(NECA Systematic Meta-Analysis and Review Toolkit)는 한국보건의료연구원(NECA) 주도 **혁신적 AI 기반 안전성 체계적 문헌고찰 자동화 시스템** 개발 프로젝트.

프로젝트 핵심 목표:

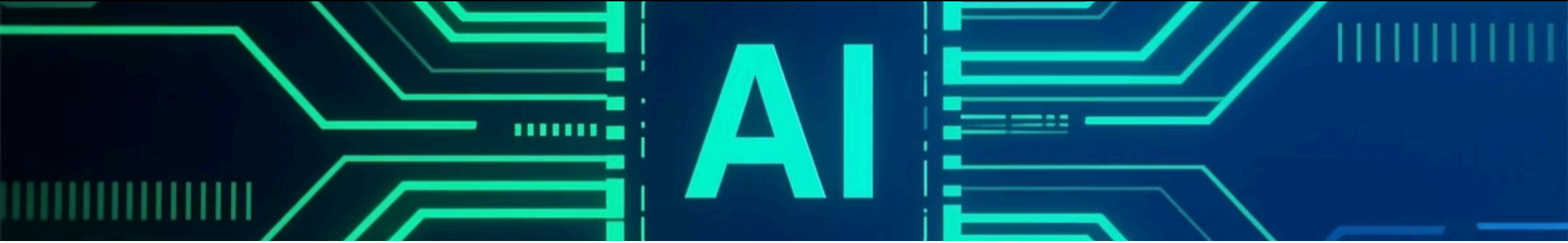
- **고도화된 지능형 문헌 스크리닝 자동화** 구현
- **Human-AI 협업 모델**을 통한 시너지 극대화
- **프롬프트 기반** 및 **프로그램 기반 하이브리드 접근법** 활용
- **한국 보건의료 환경에 최적화**된 설계

차별화된 기술 혁신:

- 최신 **대규모 언어모델(LLM)의 심층 활용**
- **정교한 맞춤형 프롬프트 엔지니어링 기술** 도입
- **고도화된 알고리즘 파이프라인** 구축 및 운영
- **실시간 성능 모니터링** 및 지속적 최적화

전략적 적용 분야:

- **체계적 문헌고찰** 및 **심층 메타분석** 지원
- 과학적 근거 기반 **임상진료지침 개발 지원**
- 정확하고 신뢰성 높은 **보건의료기술평가(HTA)** 수행
- **데이터 기반 근거중심 의료정책 수립** 강화



AI 기반 체계적 문헌고찰 고도화: 핵심 개발 로드맵

?

최적화 연구 질문 설계 보조

PICO/PICOS 프레임워크 활용 **정교한 연구 질문** 체계화

지능형 검색 전략 구축

각 데이터베이스에 **최적화 검색식** 제안

혁신적 문헌 스크리닝

AI 기반 제목/초록 및 전문 **스크리닝 자동화 시스템** 구축

자동화 자료 추출

연구 특성 및 **핵심 결과 데이터**의 **정밀 자동 추출** 프로그램 개발

✳

객관적 비뚤림 위험 평가

AI 지원 연구 품질 평가 및 **편향 위험 진단 프로그램** 개발

고급 통계 분석 프로그램 연계

메타분석 및 이질성 평가 지원

NECA AI-SMART SR 설계 원칙

01

Layer 1: 사람-프롬프트 상호작용층 🧠

- 전략 조율: PICO 구조화, MeSH 설계
- 인간 방향 제시 → AI 구체화

02

Layer 2: 자동화 파이프라인층 ⚙️

- 대규모 병렬 처리: LangChain, API, NLP
- 중복제거, 자동 스크리닝, PICO 추출

03

Layer 3: 인간 검증층 ✔

- 경계 사례 판단, AI 결정 검증
- PRISMA·AMSTAR 준수 평가

04

Layer 4: 지식 축적층 🔄

- 순환적 학습: 피드백 → 개선을 통한 지속적 프롬프트 최적화
- 다음 SR 성능 강화

핵심 원칙 강조: "불확실시 포함 분류('YYY') - 민감도 우선"

> **"Given the nature of abstracts as concise summaries... If uncertainty persists, classify as includable ('YYY')."**

AI-SMART의 핵심 목표 : "AI의 진정한 성공은 정답 예측을 넘어, '어떻게, 그리고 왜 판단했는가'를 명확히 제시하는 설명가능성을 높이는 데 있음"

NECA AI-SMART SR 3개년 로드맵

📚 1차년도 ('25.7–'25.12)
준비 및 기반 구축 Stage

- 선행 문헌고찰 수행
- 전문가 자문 및 FGI
- HUMAN–AI 비교 데이터셋 구축

⚙️ 2차년도 ('26)
각 파트별 자동화 실현

- 체계적 문헌고찰 8 Stage별 자동화 순차 실현
- 개발용 데이터셋 추가 구축
- 검증용 테스트셋 마련 및 검증 실행

🌐 3차년도 ('27)
확장 및 통합 Stage

- 개별 프로세스 간 연결 및 전체 통합
- 질환·디바이스 단위 범용화
- NECA 내부 SR 플랫폼 연동

1 2 3

성과지향 구조: AI–Human 협업 기반 표준화·자동화 실현

중장기 목표: SMARTER framework 기반 전주기통합연계 모델 구축

핵심 키워드: Pipeline • Dashboard • Integration • Scalability

결과 비전: AI-SMART SR, NECA 근거생산 프로세스 내 핵심 역할 점진적 확대

NECA AI-SMART SR 기대효과

50–80%

⏱️ SR 수행시간 단축

검색·스크리닝·추출 자동화로

해당 단계프로세스 **대폭 단축**

20–30%

🤝 Reviewer 간 합의율 향상

AI 보조 판정 및 **근거 인용 자동화**로

일관된 판단체계 구축

2X

🔁 효율성과 신뢰성 증대

AI의 속도와 **인간의 전문성** 결합으로 SR 신뢰도

2배 이상 향상

최종적으로 Living SR 운영 가능

자동 업데이트 및 **재검증 기능**으로 항상 최신 근거 유지

새로운 문헌 발표 시 **자동 통합** 및 결과 갱신으로 **지속가능한 근거생산 체계 구축** 목표

지속가능한 근거생산

주기적인 피드백 루프를 통해 **자동 개선** 및 학습 가능한 시스템 제공

시간 경과에 따른 **정확도와 효율성 향상**을 통한 **자기진화형 플랫폼** 목표

핵심 슬로건: "연구자의 통찰과 **AI의 효율**로, **근거생산을 새롭게"**

Part 4. 연구 워크플로우를 위한 AI 도구

AI Tools for Research Workflows

추천 AI 도구 개요

도구	강점	주의점	최적 Stage
💬 ChatGPT Pro	**Structured Outputs**를 공식 제공하여 **데이터 추출·정형화에 유리**, PICO 프레임으로 질문을 구조화하도록 유도하기에 적합	생성 정보 **과신 주의, 검증 필수**	PICO 등 프레임워크에 맞춘 **질문 정교화, 스키마 기반 데이터 추출**
🧠 Claude	**긴 문맥 처리** 및 **문서 이해** 탁월	**프롬프트/예시로 형식 통제를 해야함**	**보고서 초안 작성**
📚 Consensus	**학술 근거 중심 요약, 근거와 함께 요약**	**학술 코퍼스 커버리지/갱신주기 의존으로 속보성 이슈에는 제한**	**초기 학술 질의 및 근거 탐색, 요약에 적합**
🔍 Perplexity	**출처 표기 명확(실시간 웹검색 + 답변 내 인라인 출처/링크**를 기본 제공), **연관/팔로업 질문 제시**로 탐색을 넓혀 주므로 **키워드 확장력 우수**	검색 **정확도 확인 필요**	초기 정보 지형 파악, **팔로업 중심 키워드/질문 확장**
📄 SciSpace	PDF 논문 **직접 질의응답, 개념 설명 우수**	**복잡한 통계 해석 한계, 정확도 재확인**	**논문 이해, 문헌 정리**
🎨 Gamma	**프레젠테이션 자동 생성, 시각화 탁월**	내용보다 **형식 중심, 내용의 전문성은 별도 검토**	**결과 시각화, 발표 자료 작성**

적재적소의 AI 선택으로 **속도와 신뢰성 확보**
SR 단계별 AI 도구 **활용 가이드 참고**하여 **최적 도구 선택**

AI 도구 활용 시나리오

💬 ChatGPT Pro

- PICO **프레임 기반 항목 추출** 및 스키마 정의
- **JSON 기반 데이터 표준화** 및 추출표 생성
- 포맷/라벨 **일관성 유지**와 초안 검수
- 규칙 기반 **요약·정리** 및 산출물 통합

🧠 Claude

- **근거 문장 하이라이트** 및 **요약**
- **긴 문서 심층 분석**
- **보고서 및 논문 초안 작성**

📚 Consensus

- **배경지식 탐색** 및 **연구근거 요약**
- **학술적 합의 수준에 대한 개요 파악**
- 관련 문헌 **빠른 스캔**

🔍 Perplexity

- **핵심 키워드 확장** 및 **관련 연구 스캔**
- **최신 문헌** 및 **유사 주제 실시간 검색**
- **출처 명확 정보 제공**

📄 SciSpace

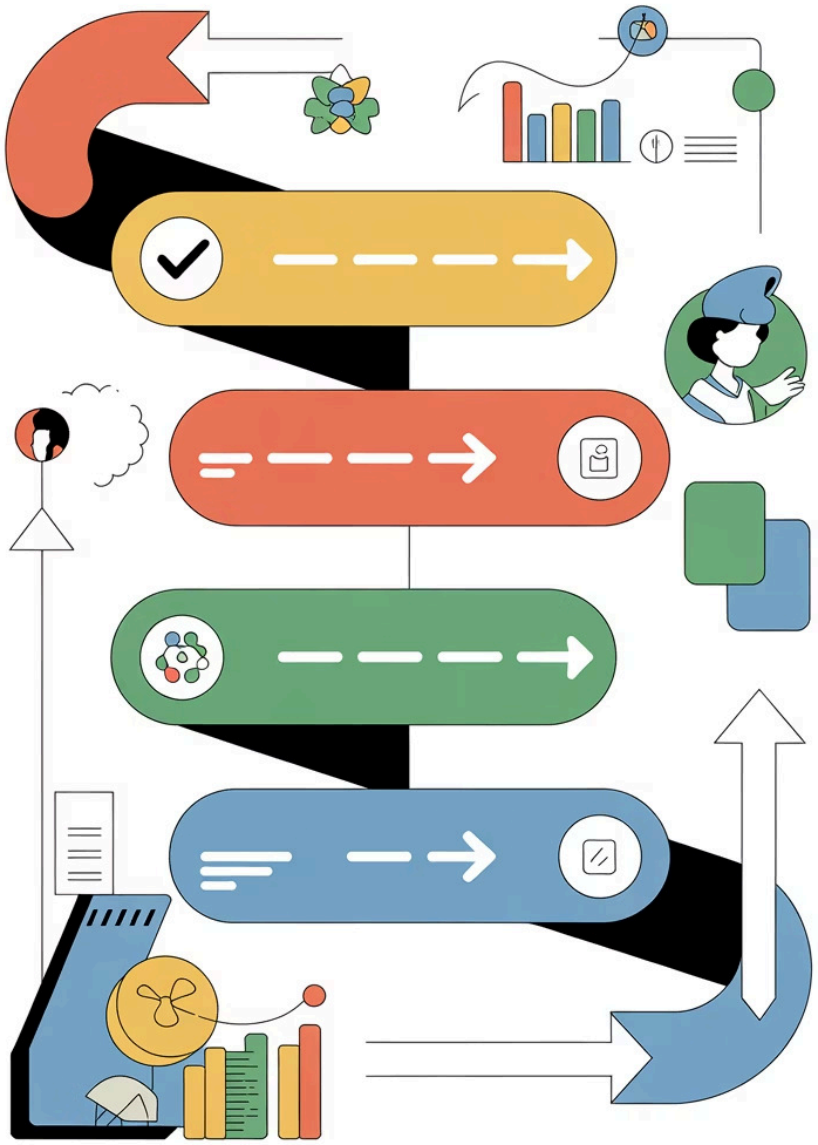
- **PDF 논문 업로드** 후 **직접 질의응답**
- 복잡한 **개념, 용어, 방법론** 즉시 설명
- **표·그래프 해석** 및 **통계 결과 이해** 지원

🎨 Gamma

- **연구결과**를 **프레젠테이션**으로 자동 변환
- **학회 발표용 슬라이드 초안** 빠르게 생성

Part 5. 실무 적용 로드맵 & 핵심 메시지

Practical Implementation Roadmap & Key Takeaways



AI-SMART SR 실전 가이드

: 성공적 도입을 위한 체크포인트

✅ 시작 전 준비사항

- **연구 질문** 및 **PICO** 명확화
- **포함/배제 기준** 상세 문서화
- **표준 프롬프트 템플릿** 사전 준비
- **Reviewer 역할** 명확 분담
- **검증 프로토콜** 사전 수립

⚠️ 주의사항

- **LLM 출력** 전문가 검증 필수
- **민감 데이터** 비식별화 처리
- 모든 **판단 근거** 상세 로그 기록
- **정기적 품질 점검** 철저 실시

✅ 진행 중 체크포인트

- **단계별 중간 검증** 주기적 수행
- **Reviewer 간 합의율** 지속 모니터링
- **오분류 사례** 발생 시 즉시 피드백
- **프롬프트 성능** 지속 개선
- **버전 관리** 및 상세 **로그** 꼼꼼 기록

📊 완료 후 평가

- **최종 결과 품질** 엄격 검증
- **시간·비용 절감 효과** 명확 측정
- **개선 사항** 체계적 문서화
- **지식베이스** 최신 정보 업데이트

Take-home Messages

AI-SMART SR: 근거생산의 속도와 품질을 동시에

1

LLM, SR의 속도·일관성·재현성 향상

검색, 선별, 추출 등 **반복 업무**의 **인력 부담 절감**

2

AI-SMARTER 프레임워크

SR **전 과정 체계적 통합**: **PromptOps, PipelineOps, Human Validation, Knowledge Loop**

3

LLM 자동화는 전문가 보완 및 능력 향상 목적

'전면 대체' 아닌 **'전문가 보완'. 인간 판단·검증** 통한 신뢰성·투명성 확보

4

표준화된 체계가 품질 결정

표준화된 **프롬프트·템플릿·검증 체계**로 **동일 입력 시 동일 결과**, **재현성**(reproducibility) 확보

5

NECA, AI-SMART SR 혁신 선도

AI-SMART SR 로드맵을 통한 근거생산 혁신 선도

(2025~2027: 시범 → 개발 → 범용화 Stage 추진)

참고문헌

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545.

2. De Cassai A, Dost B, Karapinar YE, et al. Evaluating the utility of large language models in generating search strings for systematic reviews in anesthesiology: a comparative analysis of top-ranked journals. Reg Anesth Pain Med. 2025;rapm-2024-106231.

3. Grignoli N, Manoni G, Gianini J, Schulz P, Gabutti L, Petrocchi S. Clinical decision fatigue: a systematic and scoping review with meta-synthesis. Fam Med Community Health. 2025;13(1):e003033.

4. Homiar A, Thomas J, Ostinelli EG, et al. Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. BMJ Ment Health. 2025;28(1):e301762.

5. Huang J, Lai H, Zhao W, et al. Large Language Model–Assisted Risk-of-Bias Assessment in Randomized Controlled Trials Using the Revised Risk-of-Bias Tool: Evaluation Study. J Med Internet Res. 2025;27:e70450.

6. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assoc. 2025;32(4):638-47.

7. Lai H, Ge L, Sun M, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open. 2024;7(5):e2412687.

8. Maier M, Powell D, Murchie P, Allan JL. Systematic review of the effects of decision fatigue in healthcare professionals on medical decision-making. Psychol Health. 2025:1-46.

9. Purewal A, Fautsch K, Klasova J, Hussain N, D'Souza RS. Human versus artificial intelligence: evaluating ChatGPT's performance in conducting published systematic reviews with meta-analysis in chronic pain research. Reg Anesth Pain Med. 2025;rapm-2024-106358.

10. Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. J Am Med Inform Assoc. 2025;32(6):1071-86.

감사합니다

질의응답

NECA AI-SMART SR, **LLM** 효율·품질 기반 근거 생산 새 가능성 제시

김세희

소속: NECA

이메일: sseng82@neca.re.kr