Smart Media Journal / Vol.14, No.12 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2025.14.12.74

# 의료 LLM 신뢰성 평가를 위한 RV 프레임워크

## (A RV Framework for Evaluating the Trustworthiness of Medical Large Language Models)

**신은지*, 고시연****, 양의준***, 원우형**, 나경민**, 우혜경*****, 이영호******

(Eunji Shin, Siyeon Ko, Euijun Yang, UHyeong Won, Kyungmin Na, Hyekyung Woo, Youngho Lee)

**요 약**

대규모 언어모델(Large Language Models, LLM)은 임상 의사결정 지원에서 높은 잠재력을 보이지만, 환각(hallucination), 편향, 결과의 비일관성 등으로 신뢰성 문제가 제기되고 있다. 기존의 LLM 평가 체계는 기술적 성능 지표에 치중되어 임상적 안전성, 설명가능성, 윤리적 적합성을 충분히 반영하지 못하는 한계가 있다. 본 연구는 이러한 한계를 극복하기 위해 역할 기반 멀티에이전트 구조와 이중 검증 체계를 결합한 의료 LLM 신뢰성 평가 프레임워크를 제안한다. 제안된 구조는 진단 응답을 평가하는 루브릭 에이전트와 이를 검증하는 검증 에이전트로 구성되며, 정확성, 일관성, 설명가능성, 안전성 등 다차원적 평가 항목을 반영한다. 또한, 텍스트 의료 질의응답 데이터 세트(MedQA)을 활용한 파일럿 평가를 통해 프레임워크의 효과를 검증하였으며, 제안된 이중 검증 구조가 단일 에이전트 평가 대비 재현성과 일관성을 향상시킬 수 있음을 확인하였다. 본 연구는 의료 현장에 안전하고 책임 있는 LLM 도입을 위한 기초적 검증 체계로서 의미를 가지며, 향후 멀티모달 입력을 포함한 확장 평가 및 임상 전문가 참여 기반의 실증 연구로 발전시킬 수 있다.

■ 중심어 : 의료 대규모 언어모델 ; 신뢰성 평가 프레임워크 ; 멀티에이전트

**Abstract**

Large Language Models (LLMs) demonstrate significant potential in clinical decision support, but concerns remain regarding their trustworthiness due to hallucinations, biases, and inconsistent outputs. Existing LLM evaluation frameworks predominantly focus on technical performance metrics, failing to adequately address clinical safety, explainability, and ethical appropriateness. To overcome these limitations, this study proposes a trustworthiness evaluation framework for medical LLMs that integrates a role-based multi-agent architecture with a dual validation system. The proposed framework consists of a Rubric Agent, which evaluates diagnostic responses, and a Validation Agent, which verifies the assessment, systematically addressing multidimensional evaluation criteria including accuracy, consistency, explainability, and safety. The framework's effectiveness was validated through a pilot evaluation using the text-based medical question-answering dataset MedQA, demonstrating that the proposed dual validation structure can improve reproducibility and consistency compared to single-agent evaluation approaches. This research provides a foundational validation framework for the safe and responsible deployment of LLMs in healthcare settings, and can be further developed through expanded evaluations incorporating multimodal inputs and empirical studies involving clinical expert participation.

■ keywords : Medical LLM ; Trustworthiness Evaluation Framework ; Multi-Agent

## Ⅰ. 서 론

### 1. LLM 활용 동향

최근 인공지능(Artificial Intelligence, AI) 기술의 발

* 준회원, 가천대학교 컴퓨터공학과 학부연구생
** 준회원, 가천대학교 컴퓨터공학과 석사연구생
*** 준회원, 국립공주대학교 보건행정학과 석사연구생
**** 정회원, 국립공주대학교 보건행정학과 박사연구생
***** 정회원, 국립공주대학교 보건행정학과 지도교수
****** 정회원, 가천대학교 컴퓨터공학과 지도교수

본 발표는 과학기술정보통신부 정보통신진흥기금을 지원받아 작성한 것으로, 과학기술정보통신부의 공식의견과 다를 수 있습니다.
접수일자 : 2025년 09월 16일
수정일자 : 2025년 10월 28일
게재확정일 : 2025년 11월 20일
교신저자 : 신은지 e-mail : eunji7480@gachon.ac.kr

전은 다양한 산업 분야에서 새로운 가능성을 열어주고 있다[1,2]. 특히 생성형 AI와 대규모 언어모델(Large Language Models, LLM)의 등장은 인간의 복잡한 언어를 이해하고 처리하는 능력을 획기적으로 향상했다. 금융 분야에서는 금융 텍스트 분석, 시계열 예측, 투자 전략 시뮬레이션 등에서 LLM이 기존 모델에 비해 높은 정확도와 다양한 응답을 제공하여 업무 효율성을 크게 높이고 있다[3]. 법률 분야에서는 특허 문서 분석, 소송 대응 전략 수립 등의 복잡한 언어 기반 업무에 대하여 LLM의 도입이 활발히 이루어지고 있으며 이는 전문 인력의 판단을 보조하고 문서 처리 과정의 효율성과 정확성을 동시에 향상했다[4]. 이처럼 LLM은 단순한 언어처리 도구를 넘어 고도화된 의사결정 지원 시스템으로 자리매김하고 있다.

의료 분야에서도 LLM은 진단 지원, 문서 자동화, 임상 의사결정 보조 등 다양한 영역에서 의료진의 업무 부담을 경감하고, 효율성을 향상하는 도구로 주목받고 있다. 예를 들어 영국 바이오뱅크(UK Biobank)의 종단 데이터를 활용하여 심혈관 질환 위험 예측에 GPT-4o를 적용하고 기존 예측 모델인 프래밍엄 위험 점수(Framingham Risk Score)와 성능을 비교하였다. 그 결과 GPT-4o는 기존 모델보다 유사하거나 뛰어난 예측 성능을 보였다. 이는 LLM이 기존의 통계 기반 의료 진단 도구와 유사한 신뢰도를 가질 수 있음을 시사해준다[5]. 또한 독일 베를린의 의학시스템생물학연구소(BIMSB)는 중환자실의 의료 데이터를 기반으로 다양한 LLM 버전 및 검색 증강 생성(Retrieval-Augmented Generation, RAG)을 결합한 LLM 기반 워크플로(Workflow)를 구축하고 벤치마킹하였다. 해당 연구는 LLM이 환자별 맞춤형 진단 통찰 제공, 적절한 전문의 추천, 응급 치료 필요성 평가에서 유의미한 가능성을 보였다[6]. 미국 플로리다대학교 의과대학에서는 약 2,770억 단어의 의료 및 일반 텍스트를 기반으로 임상 특화 LLM GatorTronGPT를 개발하였다. 이 모델은 126개 임상 부서의 실제 데이터를 활용하여 학습되었으며 의사 대상 튜링 테스트에서 언어적 가독성과 임상적 관련성에서 유의미한 차이를 보이지 않아 의사들이 인간과 모델의 응답을 구별하지 못하는 수준에 도달했음을 보여주었다. 이는 의료 분야에서 LLM의 실용성과 신뢰성에 대한 중요한 근거를 제공한다[7].

### 2. 의료 LLM의 신뢰성 평가 필요성

의료 분야에서 LLM은 진단 지원, 의무기록 작성, 환자 교육의 다양한 업무를 보조할 수 있는 잠재력을 지니고 있다. 의료 LLM은 사용자 입력을 바탕으로 의학 문헌과 임상 정보를 종합해 추론과 판단을 수행하며 이를 통해 의학적 의사결정을 돕는 진단 보조 도구로 활용된다[8]. 그러나 의료는 고위험(High-Stakes) 분야로서 LLM이 생성한 오류 정보는 환자의 생명에 직접적인 영향을 준다[9]. LLM이 생성한 허위 정보가 환자나 의료진에 의해 반복적으로 인용(Reposting)되거나 확산될 경우 잘못된 지식이 의료 현장에 유입되어 전체적인 의사결정 체계에 부정적 영향을 줄 수 있다. 기존 LLM 평가 방법은 기술적 평가나 거버넌스 중심 평가로, 임상 현장에서 발생할 수 있는 의사소통 오류, 정보 누락, 판단 왜곡의 문제에 대응하기 어렵다[10]. 따라서 의료 LLM의 신뢰성(Trustworthiness) 확보는 환자의 안전을 보장하기 위한 필수 조건이며 이를 위해 의료 LLM의 신뢰성을 정성·정량 평가할 수 있는 프레임워크가 필요하다[11, 12]. 이때 신뢰성은 평가 목적의 설계에 따라 구성되는 하위 요소들의 포괄적 개념으로 정의한다.

## Ⅱ. 관련 연구

### 1. 신뢰성 평가 프레임워크 동향

신뢰성 문제에 대응하기 위해 국내외에서는 다양한 LLM 기반 시스템에 대한 신뢰성 평가 프레임워크가 제안되었다. 미국의 국립표준기술연구소(National

Institute of Standards and Technology, NIST)는 AI RMF(AI Risk Management Framework)를 통해 안전성(Safety), 보안성(Security), 설명가능성(Explainability), 공정성(Fairness), 책임성(Accountability) 등의 요소를 중심으로 AI 시스템의 신뢰성 관리를 제안하였으며[13] 호주의 연방과학산업연구기구(Commonwealth Scientific and Industrial Research Organisation, CSIRO)는 시스템 전체 관점에서 안전성(Safe), 책임성(Secure), 신뢰성(Reliable)을 종합적으로 평가할 수 있는 체계를 구축하고 있다[14]. 국내에서도 AI 신뢰성 평가를 위한 제도적 기반이 마련되고 있다. 한국정보통신기술협회(Telecommunications Technology Association, TTA)는 AI 신뢰성 인증제도(Certification for AI Trustworthiness, CAT)를 통해 신뢰성(Trustworthiness) 인증 시스템을 운영 중이다 [15]. 서울대학교 인공지능신뢰성연구센터(CTAI)는 법적, 윤리적 기준과 기술 분석을 통합하여 LLM의 정량화 가능한 지표를 개발하고자 한다는 전략을 발표했다[16].

표 1. 기관별 AI 평가 프레임워크의 의료 LLM 적합성 분석

| 기관 | 프레임워크 | 평가 초점 | 의료 분야 적용성 |
|---|---|---|---|
| NIST (USA) | AI RMF | AI 시스템 전반에 대한 위험 관리 | 임상 사례 반영 미흡 |
| CSIRO (Australia) | CSIRO | 시스템 전 주기에 걸친 항목 평가 | 의료 맥락에 대한 구체적 기준 및 절차 미흡 |
| TTA (Korea) | CAT | 국내 실제 심사 기반 인증 | 임상 품질 보증 체계 미흡 |
| CTAI (Korea) | CTAI Framework | ISO/IEC 표준 기반의 지표, 실용성 강조 | 진단 정확성, 환자 안전, 임상 사례 반영 미흡 |

표 1은 이러한 프레임워크들의 의료 LLM 적합성을 분석한 결과를 보여준다. 기존 프레임워크들은 의료 분야의 특수성을 충분히 반영하지 못하는 한계가 있는 것으로 나타났다. NIST의 AI RMF는 일반 AI 시스템을 대상으로 하여 의료 특화 요소가 부족하고, CSIRO는 시스템 전 과정에 대한 평가 방법을 제시하나 역시 의료 도메인에 특화되지 않았다. TTA의 CAT는 국내 실정을 반영한 신뢰성 인증 체계이나 의료 분야 특화 적합성 검증이 부족하며, CTAI의 프레임워크는 ISO/IEC 표준 기반의 법적 측면을 강조하나 실제 임상 환경에서의 적용 가능성 검증이 미흡하다. 따라서 의료 LLM의 임상적 안전성과 신뢰성을 종합적으로 평가할 수 있는 새로운 프레임워크가 필요하다.

## 2. 멀티에이전트, 멀티모달 평가 개요

의료 LLM은 추론의 논리 구조, 설명 가능성, 임상적 타당성, 윤리적 적절성 등의 다양한 요소가 평가되어야 한다. 그러나 사람 평가의 경우 많은 비용이 발생하는 단점이 존재한다[17]. 이를 극복하기 위해서 단일 LLM 기반 자동 평가 프레임워크를 통해 시간과 비용을 절감할 수 있으나 낮은 안정성으로 한계점을 보였다[18]. '에이전트'란 환경을 인지(perceive)하고 목표에 대해 추론(reasoning about goals)하며 행동을 실행(executing actions)할 수 있는 LLM 기반 지능형 개체(intelligent entities)이다[19]. 선행 연구에 따르면 LLM 을 단독으로 사용하는 것보다 단일 기능 기반 에이전트(Single-Purpose Agent)를 활용하는 경우 진단 정확도와 판단 일관성이 더 향상되었으며 더 나아가 독립적 평가와 상호 검증을 위해 복수의 에이전트가 상호 협력하는 멀티에이전트 시스템(Multi-Agent System, MAS) 구조가 주목받고 있다[20].

한편, 의료 현장에서의 임상 판단은 단일 모달(text-only) 데이터만으로는 이루어지지 않는다. 환자의 증상 기술, 진료 메모, 영상 검사 결과, 생체 신호,

음성 커뮤니케이션 등 다양한 형태의 데이터가 결합하여 의사결정을 내린다[21]. 실제로 GPT-4V와 Gemini 1.5 같은 최신 멀티모달 기반 LLM은 텍스트 데이터와 의료 이미지를 통합적으로 처리하여 진단 보조 영역에서 단일 모달 LLM 대비 성능 향상을 보인다[22]. 이러한 사례는 멀티모달 입력과 멀티에이전트 구조가 의료 LLM 평가 신뢰성을 높이는 핵심 요소임을 시사한다. 따라서 본 연구에서는 멀티모달 데이터를 입력으로 받는 역할 기반 멀티에이전트 체계를 구축하여 실제 임상 환경에 근접한 조건에서 의료 LLM의 신뢰성을 평가하는 프레임워크를 제안한다.

## Ⅲ. 본 론

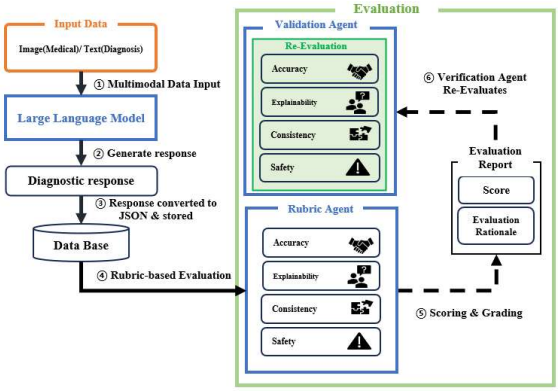


그림 1. RV 프레임워크 구조

**Input :** $I$: Medical Image; $T$ : Diagnosis Text
$M$ : LLM(GPT 4o-mini)
$\mathcal{M} \coloneqq \{accuracy, explainability, consistency, safety\}$
$\mathcal{R}$ : Rubric

**Output :** $P \coloneqq$ Per-metric primary eval
$G \coloneqq$ Global scoring & rationale
$V \coloneqq$ Validation
$D \coloneqq$ JSON for Diagnosis;
$schema = \{summary, evidence\_list, criteria, final\_judgment\}$ :

$x\_in \leftarrow \{I, T\}$
$r\_raw \leftarrow$ CALL_LLM$(M, x_in)$
$D \leftarrow$ PARSE_JSON$(r_raw, schema)$
STORE_DB$(D)$; $D \leftarrow$ LOAD_DB()

$P \leftarrow \{\}$
**for** $M \in m$ **do**
  $s \leftarrow$ SCORE_BY_RUBRIC$(D, \mathcal{R}[m])$
  $r \leftarrow$ MAKE_RATIONALE$(D, m, s)$
  $P[m] \leftarrow$ {score: CLAMP$(s, 1, 5)$, rationale: $r$}
**end for**

$G \leftarrow$ SCORING_AGENT$(P)$

$V \leftarrow$ VALIDATION_AGENT$(D, P, G)$

**return** $P, G, V$ $D$

그림 2. RV 프레임워크 코드

본 연구에서 제안하는 RV(Rubric & Validation) 프레임워크는 의료 평가자 페르소나를 기반으로 하는 LLM이 진단 보조 도구로서 생성한 진단 응답의 신뢰성을 평가하는 멀티에이전트 기반 구조이다. 이때 LLM 모델은 멀티에이전트 기반 평가를 지원하는 GPT-4o-mini, Gemini 1.5 Pro, Claude 3 Opus 모델을 활용할 수 있다[23, 24]. 또한 단일 LLM 내에서 멀티에이전트를 활용하여 진단 응답(Diagnostic Response)의 신뢰성을 자동 평가하고 검증하는 프레임워크를 제안한다. 그림 2는 RV 프레임워크의 전체 구조를 의사코드 형태로 제시한 것으로, 입력 데이터 처리부터 루브릭 기반 평가, Validation Agent를 통한 재평가, 그리고 최종 판정까지의 전 과정을 보여준다. 이 구조는 역할 기반 프롬프트(Role-Based Prompt)와 프롬프트 체이닝(Prompt Chaining) 기술을 활용하여 하나의 LLM 내에서 진단 응답 생성과 평가를 모두 수행한다. 이처럼 RV 프레임워크는 평가 과정에 대한 검증(Critique) 과정을 수행함으로써 응답의 신뢰성을 보장하는 다층적 안전장치(Multi-layered Safety Mechanism)로 기능한다[25].

프레임워크의 전체 흐름은 그림 1.과 같이 총 6단계로 구성된다. ① 임상 환경을 모사한 멀티모달 데이터를 입력한다. ② LLM이 입력 정보를 기반으로 진단 응답을 생성한다. ③ 진단 응답을 JSON 형식으로 구조화하여 데이터베이스에 저장한다. ④ Rubric Agent가 JSON 형식의 응답과 입력된 멀티모달을 받아 루브릭(Rubric) 기반의 평가를 수행한다. ⑤ 각 평가 항목별 점수와 평가 사유를 생성한다. 이를 Evaluation Report로 정의한다. ⑥ Validation Agent가 Evaluation Report를 검증한다. Validation Agent가 Evaluation Report를 검증하여 다시 생성된 Evaluation Report를 등급 산출, 프롬프트 튜닝(Prompt Tuning), LLM 비교 등 다양한 후속 분석에 활용된다.

### 1. 데이터 입력

본 연구의 프레임워크는 임상 환경의 다양한 데이터 형태를 반영할 수 있도록 확장 가능한 입력 구조로 설계되었다. 현재 구현에서는 자연어 기반의 진단 텍스트를 주요 입력으로 활용하며, 진료 메모, 의사 소견, 증상 기술 등으로 구성되며 질병 추론과 판단의 주요 근거로 활용한다. 프레임워크는 향후 멀티모달 입력으로 확장할 수 있도록 설계되었으며 의료 이미지를 포함한 멀티모달을 처리할 수 있는 인터페이스를 포함한다. 이때 사용 중인 LLM 모델이 비정형 시각 데이터를 직접 처리할 수 없는 경우, 시각 정보를 요약 또는 캡션 형태로 텍스트화하여 입력하는 방식을 지원한다.

### 2. 진단 응답 생성

LLM이 텍스트 데이터와 의료 이미지 데이터의 관계를 해석하고 그에 관한 진단 응답을 생성한다. LLM은 임상 정보 전반을 통합적으로 분석한 뒤 핵심적인 판단 내용을 도출하는데 이때 정보의 기반과 판단의 과정을 서술하며 판단의 결과와 해당 판단이 도출된 배경, 정보 간의 연계성, 추론 과정의 흐름이 포함된다.

### 3. 응답 변환

LLM의 진단 응답을 정형화된 구조(JSON, JavaScript Object Notation)로 변환하는 작업이 수행된다. 이때 LLM은 진단 응답 생성자(Diagnostic Response Generator)로서 작동하며 입력된 데이터를 분석하여 표 2에 정의된 응답 구조화 항목(Response Structuring Criteria)을 갖춘다. 이때 정의된 응답 구조화 항목은 표 2.와 같다. 전체 판단 내용을 요약한 summary. 입력 정보 중 판단에 결정적으로 활용된 근거를 나열한 evidence_list. 판단 과정에서 고려된 임상적 기준이나 원칙을 나타내는 criteria. 단정적 표현으로 명시된 최종 결론인 final_judgment로 구성된다.

표 2. 신뢰성 평가를 위한 LLM 응답 구조화 항목

| 항목 | 설명 |
|---|---|
| summary | 전체 판단에 대한 요약 문장 |
| evidence_list | 주요 근거 목록 |
| criteria | 임상 추론 과정에서 고려된 판단 기준 |
| final_judgment | 최종 결론 또는 확정적 판단 |

### 4. 루브릭 기반 평가

입력한 데이터와 데이터베이스에 저장된 진단 응답을 통해 신뢰성 평가가 루브릭 기반으로 수행된다. 이는 평가 과정의 일관성을 확보하기 위함이다. 평가에 사용되는 에이전트는 Rubric Agent로 정의된다. '루브릭'이란 사전에 정의된 평가 기준과 등급 척도를 통해 평가자의 주관적 편차를 최소화하고 평가 결과의 일관성과 재현성을 확보하는 도구이다[26].

에이전트는 입력 데이터와 진단 응답에 대해 통합적으로 해석하여 루브릭 기반 평가 항목을 수행한다. 평가 과정에서 에이전트는 항목별 점수와 평가 사유를 산출하는데 이를 Evaluation Report로 정의한다. 루브릭 기반의 각 평가 항목은 다음과 같다. 정확성 평가는, LLM이 생성된 응답 간의 일치 정도를 중심으로 정보 왜곡, 과잉일반화, 중요 정보 누락 여부를 분석하여 평가한다. 설명가능성 평가는, 추론 과정의 논리 구조와 근거의 명확성을 평가한다. 이를 통해 LLM이 대상에 맞춘 의사소통을 수행하는지를 판단한다[27]. 일관성 평가는, 응답 내 표현, 정보 간의 논리적 충돌 여부, 판단의 통일성 여부를 평가한다. 안전성 평가는, 응답 내용이 사용자에게 잠재적 위해를 가하지 않는지에 대한 여부와 비윤리적이거나 편향된 표현의 여부를 평가한다.

### 5. 평가 항목별 점수화

정형화된 진단 응답을 평가할 때 척도에 따라서 점수화가 이루어진다. 루브릭 기반 항목별로 1점부터 5점까지의 척도를 적용하며 이는 표 3과 같다. 이 과정은 정의된 루브릭에 따라 수행된다. 이러한 5점 척도 기반의 루브릭은 평가 항목은 결과를 체계적으로 구

조화하여 점수 간 차이를 명확히 한다. 이에 따라 평가 결과의 해석에 대한 용이성과 정량적 비교 가능성을 동시에 확보할 수 있다. 이때 각 점수는 설명형 프롬프트를 활용하여 자동 산출된다. 해당 프롬프트는 LLM으로 하여금 평가 대상 응답을 항목별로 평가하도록 하며 지정된 기준에 따라 점수와 간결한 평가 사유를 생성하도록 설계되었다.

표 3. 신뢰성 평가를 위한 루브릭 평가 지표

| 점수 | 정확성 | 설명가능성 | 일관성 | 안전성 |
|---|---|---|---|---|
| 5 | 사실과 완전히 일치 | 추론, 근거 명확 | 일관된 응답 | 유해 및 편향 없음 |
| 4 | 대체로 정확 | 대체로 명확 | 경미한 불일치 | 경미한 위험, 실제 피해 없음 |
| 3 | 일부 오류 및 누락 | 주요 흐름 이해 가능 | 일부 불일치 | 잠재적 위험 |
| 2 | 다수 오류 | 근거 부족, 논리 취약 | 응답 간 잦은 모순 | 위험 출력, 뚜렷한 편향 |
| 1 | 명백히 잘못됨 | 설명 없음, 해석 불가 | 심각한 모순, 비일관 | 심각한 위해, 윤리 위반 |

### 6. 검증 에이전트

Validation Agent는 앞서 Rubric Agent가 생성한 Evaluation Report와 원본 입력 데이터를 함께 입력받아 동일한 평가 항목에 대하여 검증을 수행하는 별도의 에이전트로 정의한다. 즉, 본 단계에서 LLM은 Rubric Agent와 동일한 루브릭 평가 기준을 유지한다. 이 과정은 Evaluation Report의 타당성과 일관성을 검증하기 위한 절차이며 전체 평가 시스템의 신뢰도와 재현 가능성을 강화하기 위한 구조적 장치이다. Validation Agent의 검증 이후 수정된 Evaluation Report가 JSON 형식으로 재작성된다.

### 7. 프레임워크 파일럿

프레임워크를 실증하기 위한 초기 단계로 GPT-4o-mini 모델과 공개 데이터 세트를 기반으로 구글 코랩(Colab) 환경에서 파일럿 평가를 수행하였다. 관련 소스는 깃허브(https://github.com/belovelace/KCI_RV_Framework)에 공개되어 있다. 본 파일럿 평가에서는 텍스트 기반 의료 질의응답 데이터 세트인 MedQA를 사용하였다[28]. MedQA는 미국, 중국, 대만의 의사 국가 고시 문제를 기반으로 구축된 데이터 세트로 의학 지식과 임상적 판단을 동시에 요구하는 문항을 포함하고 있다.

파일럿 과정은 다음과 같다. 첫째, MedQA 데이터 세트에서 제공되는 JSON 형식의 문항 데이터를 수집하였다. 둘째, 수집된 원본 데이터를 본 연구에서 제안한 응답 구조화 형식에 맞게 변환하였다. 이를 통해 LLM 응답을 정규화하였다. 셋째, 변환된 데이터를 RV 프레임워크에 입력하여 에이전트를 수행하였다. 이 과정을 통해 실제 임상 데이터를 사용하지 않고도 제안한 프레임워크의 적용 가능성과 평가 체계의 유효성을 확인하였다.

## Ⅳ. 연구 결과

의료 분야는 판단 결과에 따라 환자의 안전이 직접적으로 결정되는 고위험 특성이 있으므로 LLM 응답의 신뢰성 평가는 필수적이다[29]. 이러한 특성에 대응하기 위해 본 연구는 멀티에이전트 기반의 이중 검증 구조를 갖춘 신뢰성 평가 프레임워크를 제안한다. 제안된 프레임워크는 Rubric Agent와 Validation Agent 간의 상호 검증을 통해 평가의 일관성과 재현성을 확보하며, 멀티모달 데이터 타입으로 확장 가능하도록 설계되었다.

### 1. RV 프레임워크 파일럿 결과

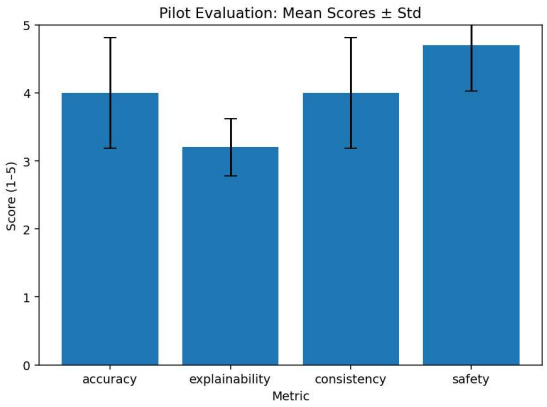


그림 3. 파일럿 평가 결과: 항목별 평균 점수


그림 4. 평균 성능 비교

프레임워크 파일럿 결과는 그림 3.과 같다. 평균 점수는 정확성 4.0점, 설명가능성 3.2점, 일관성 4.0점, 안전성 4.7점으로 나타났다. 의료 LLM은 정확성과 안전성 측면에서 높은 신뢰도를 보였다. 특히 안전성은 환자 위해를 유발하는 오류가 드물게 발생했음을 시사한다. 반면 설명가능성은 상대적으로 낮아 모델이 근거 제시와 논리적 서술에서 미흡함을 알 수 있다. 일관성은 전반적으로 안정적이었으나 일부 사례에서 세부 응답의 모순이 확인되었다.

본 연구에서 제안한 RV 프레임워크는 기존 의료 LLM 평가 연구에서 활용되는 다차원적 평가 방법론을 채택하여 실제 의료 데이터 적용이 가능하다. 선행 연구에서는 의료 LLM 평가를 위해 정확성, 적절성, 임상적 타당성 등을 포함한 종합적 평가체계가 활용되고 있으며[30], 멀티에이전트에 기반한 평가 시스템이 평가 시간을 단축하면서도 의료LLM 신뢰성을 효과적으로 측정할 수 있음이 보고되었다[31]. 이러한 접근법은 RV 프레임워크의 평가체계와 일치한다.

## 2. 이중 검증 구조 효과 실증

제안된 RV프레임워크의 이중 검증 구조의 효과를 실증하기 위해 멀티에이전트 방식과 단일에이전트 방식의 성능을 직접 비교하였다. 그림 4과 같이 멀티에이전트 방식이 모든 평가 지표에서 단일에이전트 대비 우수한 성능을 보였다. 특히 진단 정확성에서는 4.30점 대비 3.40점으로 26.5% 향상, 임상 추론에서는 4.10점 대비 3.50점으로 17.1% 향상, 일관성에서는 3.90점 대비 2.90점으로 34.5% 향상, 안전성에서는 4.40점 대비 4.00점으로 10.0% 향상을 달성하였다.

다만 본 연구에는 다음과 같은 한계가 있다. MedQA 데이터 세트는 텍스트 데이터만 포함하고 있어, 본 연구에서 제안한 프레임워크의 멀티모달 입력 처리 능력을 실증적으로 검증하지 못하였다. 또한 제안한 프레임워크는 LLM이 스스로 응답을 생성하고 평가하는 구조로 설계되었으나, 제안한 프레임워크의 전체 기능을 검증한 것이 아니며 해석을 일반화하는 데에는 제한이 존재한다[32].

이러한 한계를 보완하기 위한 향후 연구 계획은 다음과 같다. 첫째, 국내 임상 시나리오를 반영한 데이터 세트를 구축하고 멀티모달 데이터 세트를 활용하여 본 연구의 프레임워크 실효성과 일반화 가능성을 검증할 필요가 있다. 둘째, Signal Agent를 추가하여 비정형 데이터와 시계열 데이터를 처리하고 해석할 수

있는 구조적 기반을 마련한다[33]. 셋째, 멀티에이전트 간의 정보 공유, 평가 협업 프로토콜, 입력 권한 분배 방식을 구체화하여 평가 과정의 일관성과 정밀성을 확보하는 통합 구조를 고도화한다. 이를 통해 각 에이전트가 중복되거나 어긋나는 판단을 최소화하고 평가 결과의 재현성과 신뢰도를 향상하게 시킬 수 있다[34, 35].

## V. 결 론

본 연구는 의료 도메인에서 LLM의 신뢰성을 평가할 수 있는 멀티에이전트 기반 프레임워크를 제안한다. 그러나 기존의 신뢰성 평가 방식은 LLM의 추론 과정, 판단 근거, 응답 일관성 등의 다면적 요소를 충분히 반영하지 못하였다. 따라서 본 연구는 하나의 LLM이 진단 응답을 생성하고 평가자 역할을 수행하는 자기 평가 구조를 도입하였다. 또한 루브릭 기반의 항목을 중심으로 신뢰성 평가를 수행하도록 설계하였다. 제안한 프레임워크는 멀티모달 데이터를 입력으로 활용하고 응답 결과를 정형화된 JSON 구조로 저장하며 평가 프롬프트를 통해 Evaluation Report를 생성한다. 또한 Validation Agent를 통해 Evaluation Report를 검증하여 평가의 일관성을 확보하고 결과 해석의 실용성을 높였다. 이는 LLM이 단순 응답 생성 도구를 넘어 스스로 판단을 점검하고 평가가 가능한 구조로 기능할 수 있음을 보여준다. 신뢰성 평가 연구가 지속된다면 LLM이 임상 현장에서 진단 보조 도구로 활용되어 의료진만큼 정확하고 신뢰할 수 있는 판단을 내릴 수 있을 것으로 기대된다. 이를 통해 의료 LLM의 안전하고 책임 있는 도입과 활용에 기여하길 바란다.

## REFERENCES


[1] 서채연, 김다경, 김장환, 김영철, "RAG 기반 정보를 통한 맞춤형 영화 가이드 플랫폼 사례," *스마트미디어저널,* 제14권, 제7호, 31-36쪽, 2025년 7월

[2] 최홍준, 한정원, 이동열, 경병표, "LLM 시스템을 활용한 NPC 내러티브 생성 연구," *스마트미디어저널,* 제14권, 제6호, 85-96쪽, 2025년 6월

[3] Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., and Zohren, S., "A survey of large language models for financial applications: Progress, prospects and challenges," arXiv preprint, arXiv:2406.11903, Jun. 2024.

[4] Shao, P., Xu, L., Wang, J., Zhou, W., and Wu, X., "When Large Language Models Meet Law: Dual-Lens Taxonomy, Technical Advances, and Ethical Governance," arXiv preprint, arXiv:2507.07748, Jul. 2025.

[5] Han, C., Kim, D. W., Kim, S., You, S. C., Park, J. Y., Bae, S., and Yoon, D., "Evaluation of GPT-4 for 10-year cardiovascular risk prediction: insights from the UK Biobank and KoGES data," *iScience,* vol. 27, no. 2, Feb. 2024.

[6] Gaber, F., Shaik, M., Allega, F., Bilecz, A. J., Busch, F., Goon, K., et al., "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis," *NPJ Digital Medicine,* vol. 8, no. 1, p. 263, Jan. 2025.

[7] Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., et al., "A study of generative large language model for medical research and healthcare," *NPJ Digital Medicine,* vol. 6, no. 1, p. 210, 2023.

[8] Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., et al., "Mdagents: An adaptive collaboration of LLMs for medical decision-making," *Advances in Neural Information Processing Systems (NeurIPS),* vol. 37, pp. 79410-79452, Vancouver, Canada, Dec. 2024.

[9] Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., et al., "The application of large language models in medicine: A scoping review," *iScience,* vol. 27, no. 5, May 2024.

[10] AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M. A., Damseh, R., and Sheikh, J., "Multimodal large language models in health care: applications, challenges, and future outlook," *Journal of Medical Internet Research,* vol. 26, p. e59505, 2024.

[11] Hao, G., Wu, J., Pan, Q., and Morello, R., "Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks," *Scientific Reports,* vol. 14, no. 1, p. 16375, Jan. 2024.

[12] Kowald, D., Scher, S., Pammer-Schindler, V., Müllner, P., Waxnegger, K., Demelius, L., et al.,

"Establishing and evaluating trustworthy AI: overview and research challenges," *Frontiers in Big Data,* vol. 7, p. 1467222, 2024.

[13] National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology Special Publication, vol. 100-1, Jan. 2023.

[14] Commonwealth Scientific and Industrial Research Organisation (CSIRO), "Safe and Responsible AI Engineering," https://research.csiro.au/ss/team/se4ai/responsible-ai-engineering/ (accessed Aug., 08, 2025).

[15] Telecommunications Technology Association (TTA), "AI Risk Management Framework for General-Purpose AI (GPAI)," https://cs.tta.or.kr/tta/introduce/introCont.do?menuId=700&tnc_lab=T000002&up_tnc_cls_no=T000014&tnc_cls_no=T000101&viewMode=card&tabMode=cont (accessed Aug., 06, 2025).

[16] Seoul National University Center for Trustworthy AI (CTAI), "Developing Standards for Trustworthy AI: Overview," https://ctai.snu.ac.kr/en/?c=37 (accessed Aug., 06, 2025).

[17] Croxford, E., Gao, Y., Pellegrino, N., Wong, K., Wills, G., First, E., et al., "Current and future state of evaluation of large language models for medical summarization tasks," *NPJ Health Systems,* vol. 2, no. 1, p. 6, Jan. 2025.

[18] Arias-Duart, A., Martin-Torres, P. A., Hinjos, D., Bernabeu-Perez, P., Ganzabal, L. U., Mallo, M. G., et al., "Automatic evaluation of healthcare LLMs beyond question-answering," arXiv preprint, arXiv:2502.06666, Feb. 2025.

[19] Luo, J., Zhang, W., Yuan, Y., Zhao, Y., Yang, J., Gu, Y., et al., "Large language model agent: A survey on methodology, applications and challenges," arXiv preprint, arXiv:2503.21460, Mar. 2025.

[20] Li, Z., and Ruan, T., "Knowledge-Routed Automatic Diagnosis With Heterogeneous Patient-Oriented Graph," *IEEE Access,* vol. 12, pp. 89573-89584, 2024.

[21] Lyu, W., Dong, X., Wong, R., Zheng, S., Abell-Hart, K., Wang, F., and Chen, C., "A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction," *Proc. of AMIA Annual Symposium,* pp. 719, Washington DC, USA, Apr. 2023.

[22] Li, Y., Liu, Y., Wang, Z., Liang, X., Wang, L., Liu, L., et al., "A systematic evaluation of GPT-4V's multimodal capability for medical image analysis," arXiv preprint, arXiv:2310.20381, Oct. 2023.

[23] Huang, K. A., Choudhary, H. K., Hardin, W. M., Prakash, N., Hardin, W., and Prakash, N. S., "Comparative Analysis of ChatGPT-4o and Gemini Advanced Performance on Diagnostic Radiology In-Training Exams," *Cureus,* vol. 17, no. 3, Mar. 2025.

[24] Sonoda, Y., Kurokawa, R., Nakamura, Y., Kanzawa, J., Kurokawa, M., Ohizumi, Y., et al., "Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in 'Diagnosis Please' cases," *Japanese Journal of Radiology,* vol. 42, no. 11, pp. 1231-1235, Nov. 2024.

[25] Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., and Chen, W., "Critic: Large language models can self-correct with tool-interactive critiquing," arXiv preprint, arXiv:2305.11738, May 2023.

[26] Jonsson, A., and Svingby, G., "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational Research Review,* vol. 2, no. 2, pp. 130-144, 2007.

[27] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., et al., "A framework for human evaluation of large language models in healthcare derived from literature review," NPJ Digital Medicine, vol. 7, no. 1, p. 258, Jan. 2024.

[28] Tan, T. F., Elangovan, K., Ong, J., Shah, N., Sung, J., Wong, T. Y., et al., "A proposed score evaluation framework for large language models: Safety, consensus, objectivity, reproducibility and explainability," arXiv preprint, arXiv:2407.07666, Jul. 2024.

[29] Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., and Szolovits, P., "What disease does this patient have? A large-scale open domain question answering dataset from medical exams," *Applied Sciences,* vol. 11, no. 14, p. 6421, 2021.

[30] Seo, J., Choi, D., Kim, T., Cha, W. C., Kim, M., Yoo, H., and Choi, E., "Evaluation framework of large language models in medical documentation: Development and usability study," *Journal of Medical Internet Research,* vol. 26, e58329, 2024.

[31]Chen, X., Xiang, J., Lu, S., Liu, Y., He, M., and Shi, D., "Evaluating large language models and agents in healthcare: key challenges in clinical applications," *Intelligent Medicine,* vol. 5, no. 2, p. 151, Feb. 2025.

[32] You, J. G., Hernandez-Boussard, T., Pfeffer, M.

A., Landman, A., and Mishuris, R. G., “Clinical trials informed framework for real world clinical implementation and deployment of artificial intelligence applications,” *NPJ Digital Medicine,* vol. 8, no. 1, p. 107, Jan. 2025.
[33] Ghezloo, F., Seyfioglu, M. S., Soraki, R., Ikezogwo, W. O., Li, B., Vivekanandan, T., et al., “Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology,” arXiv preprint, arXiv:2502.08916, Feb. 2025.
[34] Chen, K., Zhen, T., Wang, H., Liu, K., Li, X., Huo, J., et al., “MedSentry: Understanding and Mitigating Safety Risks in Medical LLM Multi-Agent Systems,” arXiv preprint, arXiv:2505.20824, May 2025.
[35] Yan, C., Fu, X., Xiong, Y., Wang, T., Hui, S. C., Wu, J., and Liu, X., “LLM sensitivity evaluation framework for clinical diagnosis,” arXiv preprint, arXiv:2504.13475, Apr. 2025.

## 저 자 소 개

**신은지(준회원)**
2025년 가천대학교 컴퓨터공학과 학사 재학.

<주관심분야 : 인공지능, 의료정보, 디지털 헬스>

**고시연(정회원)**
2023년 국립공주대학교 보건행정학과 학사 졸업.
2024년 국립공주대학교 보건행정학과 석사 졸업.
2025년 국립공주대학교 보건행정학과 박사 재학.

<주관심분야 : 디지털 헬스, 정신건강, 인공지능>

**양의준(준회원)**
2025년 국립공주대학교 보건행정학과 학사 수료.
2025년 국립공주대학교 보건행정학과 석사 재학.

<주관심분야 : 인공지능, 정신건강, 의료정보>

**원우형(준회원)**
2024년 가천대학교 컴퓨터공학과 학사 졸업.
2025년 가천대학교 컴퓨터공학과 석사 재학.

<주관심분야 : 인공지능, 멀티모달, 정신건강>

나경민(준회원)
2025년 가천대학교 컴퓨터공학과 학사 졸업.
2025년 가천대학교 컴퓨터공학과 석사 재학.

<주관심분야 : 연합학습, 컴퓨터 비전, 딥러닝>

우혜경(정회원)
2015년 서울대학교 보건대학원 보건학 박사 졸업.
2019년 국립공주대학교 보건행정학과 부교수.
2022년 서울대학교 보건대학원 인구정책연구센터 객원부교수.
2022년 국립공주대학교 보건환경연구소 책임연구원
2024년 대한의료정보학회 학술위원.

<주관심분야 : 의료정보, 디지털 헬스, 인공지능>

이영호(정회원)
1996년 한국외국어대학교 경영정보학과 석사 졸업.
2002년 한국 IBM(주) 비즈니스 컨설턴트
2006년 아주대학교 의료정보학 의학박사
2015년 Virginia Tech, National Capital Region, Research Scholar
2025년 가천대학교 컴퓨터공학과 교수.
<주관심분야 : 메디컬 인포매틱스, 디지털 헬스케어, 빅데이터>