

최신 대규모 언어 모델의 신경과 인서비스 시험 수행능력에 대한 예비 관찰: 멀티모달 추론과 전공의 성적 비교

고판우

경북대학교 의과대학 신경과학교실

A Preliminary Observation of Performance of Contemporary Large Language Models on Neurology In-service Examination: Multimodal Reasoning and Comparison with Resident Performance

Pan-Woo Ko, MD

Department of Neurology, School of Medicine, Kyungpook National University, Daegu, Korea

J Korean Neurol Assoc 44(3):234-236, 2026

Key Words: Artificial intelligence, Neurology, Clinical competence

Address for correspondence

Pan-Woo Ko, MD

Department of Neurology, School of Medicine, Kyungpook National University, 680 Gukchaebosang-ro, Jung-gu, Daegu 41944, Korea

Tel: +82-53-200-5765

Fax: +82-53-422-4265

E-mail: panwoo.ko@gmail.com

Received April 19, 2026

Revised June 12, 2026

Accepted June 15, 2026

대규모 언어모델(large language model, LLM)의 의료 분야 적용 가능성에 대한 관심이 빠르게 증가하고 있다. 2024년 국내 연구에서 ChatGPT 3.5는 신경과 인서비스 시험의 텍스트 기반 문항에서 45.1%의 정답률을 보여 전공의 1년차 수준에 해당하는 성능을 보였으며, 특히 이미지 해석 능력의 부재가 주요 한계로 지적되었다.¹ 이후 등장한 최신 LLM은 멀티모달 아키텍처와 향상된 추론 능력을 바탕으로 텍스트뿐 아니라 이미지 정보 처리에서도 성능 향상이 보고되고 있다.^{2,3} 본 연구에서는 2025년도 신경과 전공의 수련 평가(인서비스) 시험을 활용하여 최신 LLM의 성능을 평가하고, 특히 이미지 기반 문항에서의 멀티모달 추론 능력을 분석하고자 하였다.

2025년도 신경과 인서비스 시험 60문항을 분석 대상으로 하였으며 뇌혈관질환, 뇌전증, 신경근육질환, 신경퇴행질환, 두통, 수면장애, 자율신경계질환 등 다양한 세부 영역을 포함하였다. 이 중 26문항(43.3%)은 자기공명영상(magnetic resonance imaging)/컴퓨터단층촬영(computed tomography) 영상, 신경전도 검사, 병리 이미지 등 이미지 정보가 포함된 문항이었으며 34문항(56.7%)은 텍스트 기반 문항이었다. 출제 오류로 판단된 1문항은 모든 모델에서 정답 처리하였다.

비교 대상 모델은 ChatGPT 5.3 (OpenAI, San Francisco, CA, USA), Gemini 3 Pro (Google DeepMind, Mountain View, CA, USA), Claude Sonnet 4.6 (Anthropic, San Francisco, CA, USA)으로 선정하였다. 평가는 2026년 4월 7일에 웹 인

터페이스를 통해 수행하였다. 각 문항은 단일 시행으로 설정하였다. 모델에 입력한 프롬프트와 각 LLM 모델의 정답 표시 내용은 Supplementary Table로 제시하였으며, 채점은 수련평가 위원회에서 제공하는 인서비스 정답을 기반으로 일치율로 평가하였다. Claude의 경우 상위 모델인 Claude Opus 4.6 (Anthropic) 모델이 있으나 같은 입력 프롬프트에서 오류 가능성을 이유로 답변 생성을 거부하여 분석에서 제외하였다. 이미지 문항의 경우 기존 연구와 달리 텍스트 변환 없이 원본 이미지를 직접 입력하는 vision input 방식(수직 144 DPI, 수평 144 DPI 해상도)을 적용하였다. 분석 지표는 전체 정답률, 이미지 포함 여부에 따른 정답률, 전공의 연차별 평균 성적($n=240$)과의 비교로 설정하였다. 본 연구는 대한신경과학회 수련위원회로부터 인서비스 시험 문항과 개인 정보가 포함되지 않은 전공의 연차별 성적 자료를 제공받아 수행하였다. 전공의 성적은 연차별 평균값만을 분석에 사용하였다.

결 과

세 모델 모두 2023년 인서비스 문항으로 평가한 ChatGPT 3.5의 정답률(45.1%)을 크게 상회하였다. ChatGPT 5.3은 93.3% (95% 신뢰구간, 84.1-97.4)로 가장 높은 정답률을 보

였으며 Gemini 3 Pro는 91.7% (95% 신뢰구간, 81.9-96.4), Claude Sonnet 4.6은 81.7% (95% 신뢰구간, 70.1-89.4)를 기록하였다(Table).

이미지 문항에서 ChatGPT 5.3은 88.5% (95% 신뢰구간, 71.0-96.0)로 가장 높은 정답률을 보였고 Gemini 3 Pro는 84.6% (95% 신뢰구간, 66.5-93.9), Claude Sonnet 4.6은 65.4% (95% 신뢰구간, 46.2-80.6)였다. 세 모델 모두 텍스트 문항 대비 이미지 문항에서 낮은 정답률을 보였으나 그 격차는 ChatGPT 5.3에서 8.6%p, Gemini 3 Pro에서 12.5%p, Claude Sonnet 4.6에서 28.7%p로 차이를 보였다. 다만 전체 및 이미지 문항 수가 적어 신뢰구간이 넓고 모델 간 신뢰구간이 상당 부분 겹치므로 상위 모델 간의 정답률 차이(예: ChatGPT 5.3와 Gemini 3 Pro 간 1.6%p)를 포함한 점추정치 기반의 모델 순위 해석은 신중할 필요가 있다. 본 연구에서는 모델 간 차이에 대한 별도의 통계적 유의성 검정을 수행하지 않았다.

전공의 성적과 비교하면 세 모델 모두 전공의 4년차 평균(65.1%)을 상회하였다. ChatGPT 5.3와 Gemini 3 Pro는 각각 28.2%p, 26.6%p의 높은 성적을 보였으며 Claude Sonnet 4.6 역시 16.6%p로 높은 정답률을 기록하였다. 4년차 전공의의 실제 최고 점수는 83.3점이었으며 ChatGPT 5.3 (93.3점)과 Gemini 3 Pro (91.7점)는 4년차 전공의 최고 득점을 초과하

Table. Performance comparison of large language models and neurology residents on the 2025 Korean Neurology In-Service Examination

Group	Mean score±SD ^a	Question type accuracy			Difference vs. R4 mean ^c (%p)
		Text-based (%)	Image-based (%)	Text-image difference ^b (%p)	
Neurology residents					
R1 (n=45)	49.0±7.4	NA	NA	NA	-16.1
R2 (n=66)	56.1±10.7	NA	NA	NA	-9.0
R3 (n=64)	59.9±9.9	NA	NA	NA	-5.2
R4 (n=65)	65.1±8.2	NA	NA	NA	NA
Total (n=240)	57.9±10.5	NA	NA	NA	NA
Large language models					
ChatGPT 5.3	93.3	97.1	88.5	-8.6	+28.2
Gemini 3	91.7	97.1	84.6	-12.5	+26.6
Claude Sonnet 4.6	81.7	94.1	65.4	-28.7	+16.6

SD; standard deviation, R1-4; residency year 1 to 4, NA; not applicable, LLM; large language model.

^aMean score for residents is presented as a converted 100-point scale score (mean $\pm$ SD) and mean score for LLMs represents the percentage of correctly answered items across all questions.

^bText-image difference indicates the percentage-point gap between text-based and image-based question accuracy (negative values indicate lower performance on image-based items).

^cDifference vs. R4 mean indicates the percentage-point difference between each LLM's overall accuracy and the R4 resident mean score (65.1 $\pm$ 8.2).

는 수준의 성적을 기록하였다. Claude Sonnet 4.6 (81.7점)은 4년차 최고 득점자(83.3점)에는 미치지 못하였으나 4년차 전공의 65명 중 상위 2위 수준에 해당하는 성적을 보였다. 세 모델 모두 4년차 전공의 평균(65.1 ± 8.2 점)보다 높은 점수를 보였으며 실제 점수 분포를 기준으로 할 때 최상위권에 해당하는 수준의 성적을 나타냈다.

고 찰

본 연구는 최신 LLM이 한국어 기반 신경과 전문 시험에서 전공의 상위 수준에 해당하는 성능을 보임을 관찰하였다. 신경과 인서비스 시험 문항은 특성상 단순 사실 확인을 넘어 수련 과정을 점검하기 위한 복합적인 신경계 증상, 영상 소견, 검사 결과를 통합적으로 해석하는 고차원적 사고를 요한다. 세 모델이 모두 80% 이상의 정답률을 보인 것은 이들이 신경과 전문 지식을 요구하는 다양한 문항에서 높은 정답률을 도출할 수 있음을 보여 준다. 특히 이미지 문항에서의 성능 향상은 과거 LLM의 주요 한계였던 이미지 해석 능력이 상당 부분 개선되었음을 시사한다.⁴

다만 몇 가지 제한점이 존재한다. 첫째, 시험 기반 정답률은 실제 임상 환경에서의 의사결정 능력을 완전히 반영하지 않는다. 실제 진료에서는 불완전한 정보, 환자와의 상호작용, 윤리적 판단 등이 중요한 요소로 작용한다. 문제 기반의 평가는 증례의 핵심 요소를 일차적으로 선별한 것이므로 실제 임상에서의 적용 가능성을 평가하는 도구로는 한계가 있다.⁵ 둘째, 모델 간 성능 차이가 존재하며, 특히 Claude Sonnet 4.6에서 관찰된 바와 같이 이미지 기반 추론 능력은 아직 균질하지 않다. 각 모델마다 강점을 보이는 분야나 영역이 다를 수 있어 전체적인 성능의 차이로 해석하는 것에는 무리가 있다. 셋째, LLM의 환각(hallucination) 가능성은 여전히 임상 적용에 있어 중요한 위험 요소로 남아 있다.⁶ 넷째, 신경과 인서비스 시험 문항이 웹에 공개되었거나 유사한 문항이 모델의 학습 데이터에 포함되었을 가능성을 완전히 배제하기는 어렵다. 그 가능성이 높지 않으나 이러한 데이터 노출은 정답률을 실제 수행능력보다 높게 추정하게 할 수 있어 결과 해석 시 고려가 필요하다.

따라서 LLM은 임상의를 대체하기보다는 진단 및 의사결정을 보조하는 도구로 활용되어야 하며, 실제 임상 환경에서의 검증과 오류 분석이 지속적으로 이루어질 필요가 있다.

최신 LLM은 신경과 인서비스 시험에서 전공의 상위 수준의 성능을 보였으며, 특히 멀티모달 추론 능력에서 뚜렷한 향상을 나타냈다. 다만 분석 문항 수가 적고 평가 환경이 안정화되기 이전 시기의 시험을 대상으로 하였으므로 본 결과는 확증적 결론이라기보다 예비 관찰로 해석되어야 하고, 보다 많은 문항을 포함한 후속 연구를 통한 검증이 필요하다. 향후 추론능력과 기술이 발전하면서 임상 의사결정 지원 도구로의 활용 가능성이 더욱 높아질 것으로 기대되므로 임상에서의 활용 방안에 대한 지속적인 논의가 필요할 것이다.⁷ 아울러 LLM이 전공의 평균을 상회하는 수행능력을 보인 점은 향후 신경과 전공의 수련 과정에서 LLM이 학습 보조 및 문항 검토 도구로 활용될 수 있는 가능성과 함께, 단순 지식의 암기를 넘어 임상 추론과 통합적 판단을 평가하는 방향으로 수련 평가 체계가 보완될 필요가 있음을 시사한다.

REFERENCES

1. Ko PW. Will ChatGPT be useful for Korean neurologists in clinical practice? *J Korean Neurol Assoc* 2024;42:233-240.
2. Hirano Y, Miki S, Yamagishi Y, Hanaoka S, Nakao T, Kikuchi T, et al. Assessing accuracy and legitimacy of multimodal large language models on Japan Diagnostic Radiology Board Examination. *Jpn J Radiol* 2026;44:209-217.
3. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Sci Rep* 2024;14:9330.
4. Tomita K, Nishida T, Kitaguchi Y, Kitazawa K, Miyake M. Image recognition performance of GPT-4V(ision) and GPT-4o in ophthalmology: use of images in clinical questions. *Clin Ophthalmol* 2025;19:1557-1564.
5. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Sci Rep* 2025;15:39426.
6. Asgari E, Montana-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarization. *NPJ Digit Med* 2025;8:274.
7. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. *Nat Med* 2025;31:943-950.