서로 다른 임상 시나리오에서 항생제 처방을 위한 대규모 언어 모델 비교: 어느 모델이 더 우수한가?

Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better?

Clin Microbiol Infect. 2025;31:1336-1342 (Original article)

Clinical Microbiology and Infection 31 (2025) 1336-1342

원저(Original article)

Andrea De Vito [1,*], Nicholas Geremia [2,3], Davide Fiore Bavaro [4,5], Susan K. Seo [6], Justin Laracy [6], Maria Mazzitelli [7], Andrea Marino [8], Alberto Enrico Maraolo [9], Antonio Russo [10], Agnese Colpani [1], Michele Bartoletti [4,5], Anna Maria Cattelan [7], Cristina Mussini [11], Saverio Giuseppe Parisi [12], Luigi Angelo Vaira [13], Giuseppe Nunnari [8], Giordano Madeddu [1]

1. Unit of Infectious Diseases, Department of Medicine, Surgery and Pharmacy, Sassari, Italy
2. Unit of Infectious Diseases, Department of Clinical Medicine, Ospedale dell'Angelo, Venice, Italy
3. Unit of Infectious Diseases, Department of Clinical Medicine, Ospedale Civile S.S. Giovanni e Paolo, Venice, Italy
4. Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy
5. Infectious Diseases Unit - Department of Biomedical Sciences - Istituti di Ricovero e Cura a Carattere Scientifico (IRCCS) Humanitas Research Hospital, Rozzano, Milan, Italy
6. Infectious Diseases Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
7. Infectious and Tropical Diseases Unit, Department of Molecular Medicine, Padua University Hospital, Padua, Italy
8. Unit of Infectious Diseases, Department of Clinical and Experimental Medicine, Azienda Ospedaliera di Rilievo Nazionale e di Alta Specializzazione (ARNAS), Garibaldi Hospital, University of Catania, Catania, Italy
9. Section of Infectious Diseases, Department of Clinical Medicine and Surgery, University of Naples "Federico II," Naples, Italy
10. Department of Mental Health and Public Medicine-Infectious Diseases Unit, University of Campania Luigi Vanvitelli, Naples, Italy
11. Infectious Diseases Unit, Department of Surgical, Medical, Dental and Morphological Sciences, Azienda Ospedaliera-Universitaria of Modena, University of Modena and Reggio Emilia, Modena, Italy
12. Department of Molecular Medicine, University of Padua, Padua, Italy

13. Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy

초록

논문 이력: 접수 2025년 1월 4일 · 수정본 접수 2025년 3월 6일 · 게재 확정 2025년 3월 11일 · 온라인 공개 2025년 3월 19일

목적(Objectives): 대규모 언어 모델(large language models, LLM)은 임상 의사결정에서 가능성을 보여주고 있으나, 항생제 처방 정확도에 대한 모델 간 비교 평가는 제한적입니다. 본 연구는 다양한 임상 시나리오에서 여러 LLM이 항생제 치료를 권고하는 성능을 평가합니다.

방법(Methods): ChatGPT, Claude, Copilot, Gemini, Le Chat, Grok, Perplexity, Pi.ai의 표준판과 프리미엄판을 포함한 14종의 LLM을, 10개 감염 유형을 아우르는 항생제 감수성 결과(antibiogram) 동반 60개 임상증례로 평가했습니다. 약제 선택, 용량, 치료기간에 초점을 둔 표준화된 프롬프트를 사용해 항생제 권고를 받았습니다. 응답은 익명화 후, 항생제 적절성·용량 정확성·치료기간 적절성을 평가하는 맹검 전문가 패널이 검토했습니다.

결과(Results): 총 840개 응답을 수집해 분석했습니다. ChatGPT-o1이 항생제 처방에서 가장 높은 정확도를 보여, 권고의 71.7%(43/60)가 정답으로 분류되었고 오답은 단 1건(1.7%)이었습니다. Gemini와 Claude 3 Opus가 가장 낮은 정확도를 기록했습니다. 용량 정확성은 ChatGPT-o1(96.7%, 58/60)이 가장 높았고, 이어 Perplexity Pro(90.0%, 54/60)와 Claude 3.5 Sonnet(91.7%, 55/60) 순이었습니다. 치료기간에서는 Gemini가 가장 적절한 권고(75.0%, 45/60)를 제공한 반면, Claude 3.5 Sonnet은 치료기간을 과다 처방하는 경향을 보였습니다. 증례 복잡도가 높아질수록, 특히 치료곤란(difficult-to-treat) 미생물에서 성능이 저하되었습니다.

고찰(Discussion): LLM 간에는 적절한 항생제, 용량, 치료기간을 처방하는 능력에 상당한 편차가 있습니다. ChatGPT-o1이 다른 모델보다 우수해, 첨단 LLM이 항생제 처방의 의사결정 지원 도구로서 가능성이 있음을 시사합니다. 그러나 복잡한 증례에서의 정확도 저하와 모델 간 불일치는 임상 활용 이전에 신중한 검증이 필요함을 부각합니다. Andrea De Vito, Clin Microbiol Infect 2025;31:1336

편집인(Editor): Andre Kalil

핵심어(Keywords): 항생제 치료(Antibiotic treatment); 항생제 감수성 검사(Antimicrobial susceptibility testing); ChatGPT-o1; 치료곤란 감염(Difficult-to-treat infection); 대규모 언어 모델(Large language models); LLMs

---

서론

적절한 항생제 선택은 효과적인 환자 진료의 초석입니다. 치료가 지연되거나 부정확하면 중증 합병증, 사망률 증가, 항균제 내성(antimicrobial resistance, AMR) 발생으로 이어질 수 있다는 점은 널리 인정되고 있습니다 [1-3]. 항균제 스튜어드십(antimicrobial stewardship, AMS) 프로그램은 환자 예후를 최적화하고 내성 위험을 줄이기 위해 정확하고 시의적절하며 환자 맞춤형 처방이 필요하다는 점을 강조합니다 [4,5]. 그러나 올바른 항생제, 정확한 용량, 적절한 치료기간을 선택하는 과정은 미묘하며, 특히 다제내성균(multidrug-resistant organisms, MDRO)이 흔하고 감염내과(infectious disease, ID)

자문이 부재한 환경에서 그러합니다. MDRO 관리에 대한 진료 지침이 있더라도, 복잡한 임상증례를 관리할 때는 임상적 판단이 결정적으로 중요합니다.

전담 감염내과 팀을 갖춘 병원에서는 이러한 복잡한 결정을 안내받기 위해 전문의 자문을 자주 구합니다. 감염내과 의사는 항생제 감수성 검사(antibiotic susceptibility testing, AST)를 해석하고, 지역 내성 양상을 이해하며, 끊임없이 변화하는 항균제 치료 환경을 항해하는 데 매우 귀중한 경험을 제공합니다 [6-8]. 그러나 소규모 병원이나 지역사회 환경에서는 감염내과 전문의 접근성이 제한되는 경우가 많아, 치료 실패, 이상반응, AMR의 출현 및 확산 위험이 모두 증가합니다 [9].

이러한 문제들은 모든 의료 환경에서 항생제 의사결정 지원을 강화하는 것이 중요함을 부각합니다. 이 맥락에서 인공지능 시스템과 대규모 언어 모델(LLM)은 임상 진료를 혁신할 잠재력이 있습니다. 의학 분야에서 LLM의 활용은 빠르게 주목받아 왔으며, ChatGPT와 같은 첨단 시스템이 다양한 임상 영역에서 가능성을 입증해 왔습니다. LLM 기반 도구가 진단, 의사결정, 의학 교육을 지원하는 능력은 여러 전문 분야에 걸쳐 연구되어 왔습니다 [10-12]. 세균 감염에 대한 LLM 활용에 관해서는 소수의 연구만이 수행되었으나, 서로 다른 LLM 간 비교는 이루어진 바가 없습니다 [13-15]. 일부 연구만이 서로 다른 LLM의 성능을 평가했고 [16], 항생제 처방 맥락에서 ChatBot의 능력을 구체적으로 검토한 연구는 없었습니다. 따라서 본 논문은 선도적인 LLM들을 포괄적으로 평가하고 항생제 권고 제공 성능을 비교함으로써 이 공백을 다루는 것을 목표로 합니다.

방법

본 연구진은 여러 LLM이 임상 시나리오와 AST를 평가해 적절한 항생제 치료를 처방하는 능력을 평가하기 위해 비교 연구를 수행했습니다. 감염내과 전문의 3명이 10개 주제에 초점을 둔 AST 동반 60개 임상증례를 작성했습니다. 주제는 다음과 같습니다. (a) 급성 세균성 피부 및 피부구조 감염(acute bacterial skin and skin structure infections); (b) 혈류감염(bloodstream infection, BSI); (c) 골관절감염(bone and joint infection); (d) 중추신경계(central nervous system, CNS) 감염; (e) 귀 및 눈 감염; (f) 심내막염(endocarditis); (g) 복강내감염(intrabdominal infections, IAI); (h) 폐렴(pneumonia); (i) 이식 관련 감염(transplant-related infections); (j) 요로감염(urinary tract infections). 임상증례 목록은 보충 자료에서 확인할 수 있습니다.

ChatBot 선정

본 연구진은 표준판과 프리미엄판 양쪽을 포함해 8개 LLM, 총 14종의 LLM 성능을 조사했습니다(표 S1).

각 임상증례에 대해 2024년 9월 1일부터 30일 사이에 질의를 수동으로 입력하고, 응답을 인터페이스에서 직접 수집했습니다. 기억(memorization)이 결과에 영향을 미치는 것을 방지하기 위해 각 시나리오마다 새로운 작업(task)을 생성했고, 모든 LLM과 증례에 동일한 프롬프트를 사용했습니다(표 S2).

맹검 검토 및 평가

모든 응답을 익명화한 후, 맹검 전문가 패널이 2024년 10월 1일부터 11월 10일 사이에 답변을 검토했습니다. 패널은 세 가지 측면, 즉 적절성(appropriateness), 권고 용량의 정확성, 치료기간의 적절성을 평가하도록 요청받았습니다(표 S3).

각 임상증례는 AMS 및 임상 의사결정에 5년 이상의 경력을 가진 감염내과 전문의 9명으로 구성된 패널이 독립적으로 평가했습니다(표 S4). 각 전문가는 증례를 개별적으로 검토하고 사전에 정의된 기준에 따라 점수를 부여했습니다. 검토자 간 불일치가 발생한 경우, 선임 감염내과 교수들로 구성된 2차 판정 패널이 응답을 재평가해 합의에 도달했습니다. 전문가들은 단일한 진료 지침 하나에 따르지 않고, 가용한 가장 최신이며 견고한 근거에 기반해 평가했습니다. 특정 감염에 대해 유럽 지침과 미국 지침이 모두 존재하는 경우, 미국 전문가들은 주로 미국 지침을, 유럽 전문가들은 유럽 지침을 따랐습니다.

안과·이과 감염처럼 특정 국제 지침이 없는 임상 상황의 경우, 패널은 항생제 적절성·용량·치료기간을 평가하기 위해 자신들의 임상 전문성과 최신 동료심사(peer-reviewed) 문헌에 의존했습니다.

미생물은 AST에 따른 치료곤란도(difficulty to treat, DTR)에 기반해 분류했습니다. 다음 네 군으로 나누었습니다. (a) DTR 미생물(*Acinetobacter baumannii*, trimethoprim/sulfamethoxazole에 내성인 *Stenotrophomonas maltophilia*, DTR *Pseudomonas aeruginosa* [17]), 카바페넴 내성 장내세균목(carbapenem-resistant Enterobacterales), 반코마이신 내성 *Enterococcus*; (b) 치료 선택지가 제한된 미생물(메티실린 내성 포도상구균, 페니실린 내성 연쇄상구균, 3세대 세팔로스포린에 내성인 그람음성균 [확장스펙트럼 베타락타마제(extended-spectrum beta-lactamases) 또는 AmpC 베타락타마제]); (c) 야생형(wild-type) 미생물; (d) AST가 없는 미생물. 미생물 목록과 그람염색 및 내성 범주별 AST 분포는 표 S5와 표 S6에서 확인할 수 있습니다.

통계 분석

서로 다른 ChatBot 간 성능을 비교하기 위해 통계 방법을 적용했습니다. 검토자 간 신뢰도(inter-rater reliability)는 Fleiss' Kappa를 사용해 검토자 간 일치도를 측정했습니다. Kappa 값은 표준 기준에 따라 해석했습니다. 군 간 차이의 존재 여부를 평가하기 위해 카이제곱 검정(Chi-squared test)을 사용했습니다. 사전에 정의된 공변량 및 단변량 분석에서 통계적으로 유의한 공변량을 보정해 95% 신뢰구간(CI)과 함께 교차비(OR)를 추정하기 위해 이항 로지스틱 회귀(binomial logistic regression)를 수행했습니다. 통계적 유의성은 p 값 0.05 미만으로 설정했으며, 데이터 분석은 STATA(Version 16.1, StataCorp, College Station, TX)로 수행했습니다.

윤리적 고려사항

본 연구는 실제 세균 균주의 AST에 기반한 가상의 임상 시나리오를 다루되 실제 환자 데이터나 인구학적 특성을 포함하지 않는 성격이므로, 인간 대상 연구에 관한 기관 지침에 부합하게 윤리 심사 면제를 신청해 승인받았습니다.

결과

전체적으로 840개의 답변을 수집해 항생제 치료 전문가 맹검 패널이 평가했습니다.

처방 항생제의 정확도

항생제 처방과 관련해, 전문가 평가는 항생제 선택에서 상당한 일치도를 보였습니다($k = 0.799$, $p < 0.001$).

여러 LLM 가운데 ChatGPT-o1이 정확한 항생제 처방 측면에서 가장 우수해 응답의 71.7%가 정답으로 분류되었고, 이어 ChatGPT 4o(53.3%)와 Perplexity Pro(56.7%) 순이었습니다. 반면 Gemini와 Claude 3 Opus는 정답으로 인정된 응답이 각각 30.0%, 48.3%에 그쳐 가장 부정확했습니다(표 1). 오답 비율은 ChatGPT-o1과 ChatGPT 4o에서 가장 낮았던 반면, Gemini와 Claude 3 Opus가 가장 높은 부정확 처방률을 보였습니다.

무료 접근 LLM 가운데 Perplexity가 정답 비율이 가장 높았으나 부분 정답 비율도 가장 높았습니다. 정답과 과다치료(overtreatment) 답변을 합산하면, Copilot이 다른 무료 접근 LLM을 능가했고 ChatGPT-o1에만 뒤졌습니다.

그람염색별로 범주화하면(표 2, 표 3), 정답 비율은 그람양성 미생물에서 유의하게 높았습니다($p < 0.001$). ChatGPT-o1이 일관되게 다른 모델을 능가한 반면, Gemini는 특히 그람음성 시나리오에서 두드러지게 저조했습니다(정답 24.3%)(표 3).

DTR로 분류된 미생물이 오답 수가 가장 많았으며, DTR 미생물에서는 과다치료율이 유의하게 낮았습니다(표 4). ChatGPT-o1은 모든 난도 범주에서 탁월했던 반면, Gemini가 가장 저조했습니다(표 S7-S12).

감염 유형과 관련해서는, 복강내감염(IAI, 22.6%), 귀 및 눈 감염(19.1%), CNS 감염(13.1%), BSI(10.7%)에서 오답이 가장 빈번했습니다. 과다치료는 폐렴(60.7%)과 귀 및 눈 감염(39.4%)에서 가장 흔했습니다(표 S12).

표 1. 서로 다른 대규모 언어 모델별로 구분한, 항생제 선택 답변이 부정확·부분 정답·정답·과다치료로 평가된 비율

대규모 언어 모델	오답, n (%)	부분 정답, n (%)	정답, n (%)	과다치료, n (%)
ChatGPT	4 (6.7)	9 (15.0)	29 (48.3)	18 (30.0)
ChatGPT-o1	1 (1.7)	10 (16.7)	43 (71.7)	6 (10.0)
ChatGPT4o	2 (3.3)	11 (18.3)	32 (53.3)	15 (25.0)
Claude 3 Opus	9 (15.0)	13 (21.7)	29 (48.3)	9 (15.0)
Claude 3.5 Sonnet	6 (10.0)	10 (16.7)	29 (48.3)	15 (25.0)
Copilot	3 (5.0)	10 (16.7)	29 (48.3)	18 (30.0)
Copilot Pro	5 (8.3)	10 (16.7)	26 (43.3)	19 (31.7)
Gemini	14 (23.3)	8 (13.3)	19 (31.7)	19 (31.7)
Gemini Advance	7 (11.7)	12 (20.0)	24 (40.0)	17 (28.3)
Grok 2	5 (8.3)	12 (20.0)	24 (40.0)	19 (31.7)
Le Chat - Large 2	7 (11.7)	10 (16.7)	28 (46.7)	15 (25.0)
Perplexity	4 (6.7)	14 (23.3)	31 (51.7)	11 (18.3)
Perplexity Pro	4 (6.7)	11 (18.3)	34 (56.7)	11 (18.3)
pi.ai	4 (6.7)	12 (20.0)	26 (43.3)	18 (30.0)
전체(Total)	75 (8.9)	152 (18.1)	403 (48.0)	210 (25.0)

표 2. 서로 다른 대규모 언어 모델별로 구분한, 그람양성 미생물에 대한 항생제 선택 답변이 부정확·부분 정답·정답·과다치료로 평가된 비율

대규모 언어 모델	오답, n (%)	부분 정답, n (%)	정답, n (%)	과다치료, n (%)
ChatGPT	1 (4.3)	0	15 (65.2)	7 (30.4)
ChatGPT-o1	0	1 (4.3)	20 (87.0)	2 (8.7)
ChatGPT4o	1 (4.3)	1 (4.3)	18 (78.3)	3 (13.0)
Claude 3 Opus	3 (13.0)	4 (17.4)	15 (65.2)	1 (4.3)
Claude 3.5 Sonnet	2 (8.7)	1 (4.3)	16 (69.6)	4 (17.4)
Copilot	1 (4.4)	0	15 (65.2)	7 (30.4)
Copilot pro	1 (4.4)	1 (4.4)	14 (60.9)	7 (30.4)
Gemini	4 (17.4)	1 (4.4)	10 (43.5)	8 (34.8)
Gemini Advance	2 (8.7)	3 (13.0)	11 (47.8)	7 (30.4)
Grok 2	2 (8.7)	5 (21.7)	11 (47.8)	5 (21.7)

대규모 언어 모델	오답, n (%)	부분 정답, n (%)	정답, n (%)	과다치료, n (%)
Le Chat	1 (4.4)	3 (13.0)	17 (73.9)	2 (8.7)
Perplexity	1 (4.4)	3 (13.0)	15 (65.2)	4 (17.4)
Perplexity pro	2 (8.7)	3 (13.0)	16 (69.6)	2 (8.7)
pi.ai	0	4 (17.4)	14 (60.9)	5 (21.7)
전체(Total)	21 (6.5)	30 (9.3)	207 (64.3)	64 (19.9)

표 3. 서로 다른 대규모 언어 모델별로 구분한, 그람음성 미생물에 대한 항생제 선택 답변이 부정확·부분 정답·정답·과다치료로 평가된 비율

대규모 언어 모델	오답, n (%)	부분 정답, n (%)	정답, n (%)	과다치료, n (%)
ChatGPT	3 (8.1)	9 (24.3)	14 (37.8)	11 (29.7)
ChatGPT-o1	1 (2.7)	9 (24.3)	23 (62.2)	4 (10.8)
ChatGPT4o	1 (2.7)	10 (27.0)	14 (37.8)	12 (32.4)
Claude 3 Opus	6 (16.2)	9 (24.3)	14 (37.8)	8 (21.6)
Claude 3.5 Sonnet	4 (10.8)	9 (24.3)	13 (35.1)	11 (29.7)
Copilot	2 (5.4)	10 (27.0)	14 (37.8)	11 (29.7)
Copilot pro	4 (10.8)	9 (24.3)	12 (32.4)	12 (32.4)
Gemini	10 (27.0)	7 (18.9)	9 (24.3)	11 (29.7)
Gemini Advance	5 (13.5)	9 (24.3)	13 (35.1)	10 (27.0)
Grok 2	3 (8.1)	7 (18.9)	13 (35.1)	14 (37.8)
Le Chat	6 (16.2)	7 (18.9)	11 (29.7)	13 (35.1)
Perplexity	3 (8.1)	11 (29.7)	16 (43.2)	7 (18.9)
Perplexity pro	2 (5.4)	8 (21.6)	18 (48.7)	9 (24.3)
pi.ai	4 (10.8)	8 (21.6)	12 (32.4)	13 (35.1)
전체(Total)	54 (10.4)	122 (23.6)	196 (37.8)	146 (28.2)

표 4. 서로 다른 대규모 언어 모델별로 구분한, 치료곤란(difficult-to-treat) 미생물에 대한 항생제 선택 답변이 부정확·부분 정답·정답·과다치료로 평가된 비율

대규모 언어 모델	오답, n (%)	부분 정답, n (%)	정답, n (%)	과다치료, n (%)
ChatGPT	3 (14.3)	8 (38.1)	10 (47.6)	0
ChatGPT-o1	1 (4.8)	6 (28.6)	14 (66.7)	0
ChatGPT4o	1 (4.8)	10 (47.6)	9 (42.9)	1 (4.8)
Claude 3 Opus	7 (33.3)	6 (28.6)	8 (38.1)	0
Claude 3.5 Sonnet	3 (14.3)	8 (38.1)	10 (47.6)	0

대규모 언어 모델	오답, n (%)	부분 정답, n (%)	정답, n (%)	과다치료, n (%)
Copilot	2 (9.5)	9 (42.9)	10 (47.6)	0
Copilot pro	4 (19.0)	8 (38.1)	8 (38.1)	1 (4.8)
Gemini	10 (47.6)	5 (23.8)	4 (19.0)	2 (9.5)
Gemini Advance	5 (23.8)	8 (38.1)	7 (33.3)	1 (4.8)
Grok 2	4 (19.0)	7 (33.3)	10 (47.6)	0
Le Chat	6 (28.6)	6 (28.6)	7 (33.3)	2 (9.5)
Perplexity	2 (9.5)	9 (42.9)	10 (47.6)	0
Perplexity pro	3 (14.3)	7 (33.3)	11 (52.4)	0
pi.ai	4 (19.0)	6 (28.6)	8 (38.1)	3 (14.3)
전체(Total)	55 (18.7)	103 (35.0)	126 (42.9)	10 (3.4)

그림 1. 서로 다른 대규모 언어 모델별로 구분한, 항생제 용량 답변이 저용량(low dosage)·정확 용량(correct dosage)·고용량(high dosage)으로 평가된 비율.

용량 권고

용량에 대한 검토자 간 일치도는 중등도에서 상당함(moderate to substantial) 수준이었습니다($k = 0.644$, $p < 0.001$). ChatGPT-o1이 용량 권고에서 가장 높은 정확도(96.7%)를 달성했으며, 저용량 사례가 전혀 없었고 과용량도 최소였습니다. Perplexity Pro와 Claude 3.5 Sonnet도 우수한 성능을 보였습니다(각각 90.0%, 91.7%). Gemini는 정확도가 가장 낮았고(60.0%) 저용량률이 가장 높았습니다(26.7%). 날조된(fabricated) 용량은 단 1건만 확인되었는데, Gemini가 *Escherichia coli*에 의한 심내막염에 대해 ceftazidime/avibactam의 비현실적인 용법을 제안한 사례였습니다(그림 1).

용량 정확도에서 그람양성 미생물과 그람음성 미생물 간에 유의한 차이는 관찰되지 않았습니다. 다만 정확한 용량 권고는 AST가 없는 미생물과 DTR 미생물에서 가장 높았던 반면, 저용량은 야생형 미생물에서, 과용량은 치료 선택지가 제한된 미생물에서 가장 흔했습니다(표 S13, 표 S14).

치료기간

치료기간에 대한 일치도는 중등도(moderate)였으며($k = 0.653$, $p < 0.001$), Gemini가 정확한 권고 비율이 가장 높아 75.0%가 적절한 것으로 분류되었고, 이어 Perplexity Pro 73.3%, ChatGPT-o1 70.0% 순이었습니다. 또한 Gemini는 지나치게 짧은 치료(overly short treatment) 비율이 가장 높았고(18.3%), 이어 Copilot Pro(13.3%)였습니다. 지나치게 긴 치료(overly long treatment) 권고와 관련해서는 Claude 3.5 Sonnet이 비율이 가장 높았으며(41.7%), 이어 Grok 2와 ChatGPT-4o 순이었습니다(그림 2).

그림 2. 서로 다른 대규모 언어 모델별로 구분한, 치료기간 답변이 저용량(low dosage)·정확 용량(correct dosage)·고용량(high dosage)으로 평가된 비율.

고찰

본 연구 결과는 LLM이 정확한 항생제와 용량을 처방하는 성능에 상당한 편차가 있음을 부각했습니다. 이러한 결과는 임상 진료에 LLM 도구를 통합할 잠재적 가능성에 중요한 함의를 가질 수 있습니다.

분석 결과 ChatGPT-o1이 항생제 처방에서 가장 높은 정확도를 달성해, 응답의 70.0%가 정답으로 분류되었고 부정확하다고 판단된 응답은 1.7%에 불과했습니다. 반면 Gemini와 Claude 3 Opus 같은 모델은 정확도가 상당히 낮아, 정답 처방률이 각각 30.0%, 48.3%였습니다. 용량 정확도를 고려하면 ChatGPT-o1이 96.7%의 정확 용량으로 선두였던 반면, Gemini가 가장 저조했습니다. 특히 부정확한 처방은 환자에게 잠재적으로 치명적인 결과를 포함해 심각한 결과를 초래할 수 있습니다. 따라서 본 연구진은 LLM이 보조 도구로서 가능성을 보이긴 하지만, 임상 환경에서 감독 없이(unsupervised) 사용하기에는 아직 준비되지 않았다고 봅니다. 사용자가 해당 의학 영역에 상당한 전문성을 갖추고 LLM을 주로 대안적 관점의 출처로 활용하는 것이 여전히 결정적으로 중요합니다. LLM이 의사결정에 도움을 줄 수는 있으나, 중요한 임상 판단을 LLM에만 의존해서는 안 됩니다.

ChatGPT-o1이 평가 대상 다른 LLM에 비해 우수한 성능을 보인 것은, 그 아키텍처 내에 사고연쇄(chain-of-thought) 추론이 구현되어 있고, 발전된 학습 데이터와 미세조정(fine-tuning) 기법이 적용된 데서 기인할 수 있습니다 [18].

그러나 서로 다른 LLM 간에 관찰된 편차는 주의의 필요성을 부각합니다. ChatGPT-o1은 높은 정확도를 입증했지만, Gemini와 Claude 3 Opus 같은 다른 모델은 그렇지 못했습니다. Gemini의 저조한 결과는 기존 문헌과 상충합니다. Pirkle 등 [19]은 흔한 소아 정형외과 질환에 대한 적절한 권고에서 ChatGPT와 Gemini 간 성능이 유사하다고 보고했고, Tong 등 [20]은 Gemini가 ChatGPT-3.5 및 ChatGPT-4보다 더 간결하고 직관적인 응답을 제공한다고 보고했습니다. 그러나 이들 연구와 본 연구의 핵심 차이는, 정형외과 평가는 이론적 질문에 기반했던 반면 본 연구는 AST를 동반한 임상 시나리오를 활용해 복잡한 추론과 개별화된 의사결정을 요구했다는 점입니다. LLM은 구조화되고 지침에 기반한 작업에서는 잘 수행할 수 있으나, 미묘한 항균제 처방은 숙주 상태, 감염 부위, 내성 양상 같은 요인을 포함하므로 더 까다로운 영역입니다.

본 연구는 또한 LLM의 정확도가 복잡도 증가에 따라, 특히 DTR 미생물에서 감소함을 발견했습니다. 오답은 MDRO가 관여한 증례에서 더 빈번했던 반면, 과다치료는 덜 흔했습니다. 이러한 결과는 LLM이 표준적 증례에서는 적절히 수행할 수 있으나, 더 복잡한 시나리오에서는 신뢰도가 떨어진다는 점을 시사합니다.

LLM 성능에서 관찰된 이러한 한계는 비(非)감염내과 전문의의 항생제 처방에서 잘 기록된 문제들과 부합합니다. 여러 연구에 따르면 입원환자 항생제 처방의 30-50%가 지침 준수 미흡, 광범위 항생제 남용, 또는 AST 오해석으로 인해 부적절할 수 있습니다 [21-23]. 반면 감염내과 전문의와 AMS 팀은 처방 정확도, 환자 예후를 개선하고 AMR을 줄입니다 [24]. 본 연구가 임상 의사결정에서 LLM의 한계를 부각하긴 하지만, 더 넓은 문제, 즉 인간 처방자조차 특히 감염내과 자문이 부재할 때 복잡한 항균제 의사결정에 어려움을 겪는다는 점을 재확인합니다. 이는 기성품(off-the-shelf) LLM에 의존하기보다 검증되고 전문가가 주도하는 의사결정 지원 도구가 필요함을 부각합니다.

본 연구 결과는 Maillard 등 [14]의 발견과 부합하는데, 이들은 ChatGPT-4가 상세하고 잘 구조화된 의학적 응답을 제공하긴 했으나 더 복잡한 임상 상황에서는 성능이 완벽과는 거리가 멀었음을 입증했습니다. 이들은 ChatGPT-4가 일부 증례에서는 만족스러운 관리 계획을 생성할 수 있었으나, 진단, 항생제 처방, 감염원 통제(source control) 조치 처리에서 상당한 오류가 발생했음을 발견했습니다.

본 연구에서 오답 비율이 가장 높았던 것은 IAI(22.6%), 귀 및 눈 감염(19.1%), CNS 감염(13.1%), BSI(10.7%) 증례였습니다. 과다치료 비율이 높았던 것은 폐렴(60.7%)과 귀 및 눈 감염(39.4%)이었습니다. 이러한 결과는 일부 영역에서 강력한 성능을 보임에도, LLM이 복잡한 증례, 특히 MDRO가 관여하거나 복잡한 임상 의사결정을 수반하는 증례에서 명백한 한계를 가진다는 점을 재확인합니다.

LLM이 임상 진료를 보완할 수는 있으나 인간의 전문성을 대체할 수는 없으며, 특히 표준 지침이 실제 증례의 복잡성을 충분히 다루지 못하는 상황에서 그러합니다. 또한 출판된 논문에 비추어 LLM을 갱신하는 데 따르는 한계도 고려해야 하는데, 특히

페이월(paywall) 뒤에 있는 가용 정보 전체에 접근할 수 없기 때문입니다. 이 한계는, 연구자나 임상의가 적절히 학습시키지 않으면 LLM이 문헌의 상당 부분을 반영하지 못해 지식의 공백과 덜 충분히 정보에 기반한 임상 안내로 이어질 수 있음을 의미합니다. 더불어 LLM은 최신 임상 발전, 지침, 출판 연구를 포함한 대규모 데이터셋으로 학습하는 데 방대한 연산 자원과 시간이 필요하므로, 새로운 의학 연구의 출판과 LLM 학습 데이터에 그것이 포함되는 시점 사이에 상당한 지연이 발생합니다.

이는 지침을 해석하고 각 환자의 특정 요구에 맞게 관리를 맞춤화하는 데 있어 인간 개입이 필수적인 역할을 한다는 점을 부각하는데, 임상 현실은 프로토콜의 단순한 적용을 넘어서는 경우가 많기 때문입니다.

본 연구의 몇 가지 한계를 인정해야 합니다. 첫째, 임상 시나리오는 항생제 알레르기, 약물-약물 상호작용 등 실제 증례의 전체 복잡성을 포착하지 못할 수 있습니다. 둘째, LLM은 단일 시점에 평가되었으나, 이 모델들은 끊임없이 진화하고 있어 갱신에 따라 성능이 달라질 수 있습니다. 셋째, 모든 LLM에 균일한 프롬프트를 사용했는데, 이는 각 모델의 강점에 최적화되지 않았을 수 있습니다. 또한 본 연구가 글로벌 표준에 기반한 최적의 항생제 치료를 평가하긴 하지만, 용량 지침과 약물 가용성은 특히 저·중소득 국가에서 지역별로 크게 다를 수 있다는 점을 고려하는 것이 중요합니다 [25,26]. 따라서 본 연구에서 차선으로 판단되어 '부분 정답'으로 분류된 치료가, 이러한 환경에서는 최선이거나 유일하게 가용한 선택지일 수 있습니다. 이러한 접근성 격차는 본 연구의 중대한 한계를 부각하는데, 즉 본 연구는 여러 지역에 걸친 항생제 가용성을 충분히 고려하지 못했습니다. 마지막으로, LLM이 AST에서 검사되지 않은 치료를 제안한 경우는 소수에 불과했음을 관찰했습니다. 이는 흔히 과다치료나 '부분 정답'으로 분류된 답변으로 귀결되었습니다. 따라서 더 새로운 항생제의 도입은 추가적인 도전 과제가 될 수 있습니다.

향후 연구에서는 이러한 지역적 불일치를 고려하는 것이 중요할 것인데, 이는 의료 맥락에 따라 '최적 치료'의 정의를 바꿀 수 있기 때문입니다. 또한 실제 환경에서 LLM의 통합을 평가하기 위한 추가 연구가 필요합니다. 연구는 알고리즘의 갱신과 개선을 고려해 시간 경과에 따른 모델 성능도 평가해야 합니다.

결론

본 연구는 항생제 처방에서 현행 LLM의 잠재력과 상당한 한계를 모두 부각합니다. ChatGPT-o1 같은 모델이 일부 측면에서 더 높은 정확도를 입증했지만, LLM 간의 상당한 편차와 복잡한 증례에서의 일관되지 않은 성능은 임상 의사결정에서의 신뢰 불가능성을 부각합니다. 어느 모델도 안전한 항생제 선택, 용량, 또는 치료기간에 요구되는 기준을 일관되게 충족하지 못했습니다.

이러한 결과는 기성품 LLM을 실제 진료의 항생제 처방에 의존해서는 안 된다는 점을 재확인합니다. 부정확하거나 부적절한 권고의 위험은 극도의 주의가 필요함을 부각합니다. 인공지능 기반 도구가 향후 AMS 프로그램에 통합될 가능성을 가질 수는 있으나, 독립적인(standalone) 임상 사용에는 여전히 불충분합니다. 임상의는 LLM을 직접적인 의학적 결정에 사용하는 것을 피해야 하는데, 이는 환자 안전을 저해하고 AMR에 기여할 수 있기 때문입니다. LLM이 임상 사용을 고려받을 수 있으려면 추가 연구, 모델 정교화, 무작위 임상시험(randomized clinical trials)을 통한 검증이 필수적입니다. 임상의, 개발자, 연구자 간의 협력은 위험을 완화하면서 LLM의 이점을 활용해 궁극적으로 환자 진료를 개선하고 AMS를 지원하는 데 결정적으로 중요합니다.

저자 기여

A.D.V.와 N.G.는 개념화(conceptualization)와 시각화(visualization)를 담당했습니다. A.D.V., N.G., G.M.은 방법론(methodology)을 담당했습니다. A.D.V.는 정형 분석(formal analysis)을 담당했습니다. D.F.B., S.K.S., J.L., M.M., A.M., A.E.M., A.R., A.C., M.B., A.M.C., C.M., S.G.P., L.A.V., G.N.은 조사(investigation)를 담당했습니다. D.F.B., S.K.S., J.L., M.M., A.M., A.E.M., A.R., A.C., M.B., A.M.C., C.M., S.G.P., L.A.V., G.N.은 데이터 큐레이션(data curation)을 담당했습니다.

A.D.V., N.G., A.C.는 원고 초안 작성(writing—original draft preparation)을 담당했습니다. D.F.B., S.K.S., J.L., M.M., A.M., A.E.M., A.R., M.B., A.M.C., C.M., S.G.P., L.A.V., G.N., G.M.은 원고 검토 및 편집(writing—review and editing)을 담당했습니다. G.M.은 감독(supervision)을 담당했습니다. A.D.V.와 N.G.는 본 연구에 동등하게 기여했습니다.

투명성 선언

잠재적 이해상충

저자들은 이해상충이 없음을 선언합니다.

재정 보고

본 원고에 대해 어떠한 연구비도 받지 않았습니다.

윤리 선언

본 연구는 인간 또는 동물에 대한 어떠한 분석도 포함하지 않았으므로 윤리 심사 및 승인 요건이 면제되었습니다.

데이터 가용성

본 연구를 위해 수집된 데이터는 타인에게 제공될 것이며, 본 논문의 출판 시점에 가용해지고 원고에 첨부된 보충 자료(Supplementary Materials)에서 접근할 수 있습니다. 데이터는 어떠한 제한도 없이 누구에게나 자유롭게 가용합니다.

감사의 글

본 연구의 그래픽 표현을 제작하고 다듬는 데 매우 귀중한 도움을 준 Stefano Zugno Brunetti에게 감사드립니다.

부록 A. 보충 자료

본 논문의 보충 자료는 https://doi.org/10.1016/j.cmi.2025.03.002 에서 온라인으로 확인할 수 있습니다.

참고문헌 (원문)

1. Worldwide Antimicrobial Resistance National/International Network Group (WARNING) Collaborators. Ten golden rules for optimal antibiotic use in hospital settings: the WARNING call to action. World J Emerg Surg 2023;18:50. https://doi.org/10.1186/S13017-023-00518-3.
2. López Romo A, Quirós R. Appropriate use of antibiotics: an unmet need. Ther Adv Urol 2019;11:1756287219832174. https://doi.org/10.1177/1756287219832174.
3. Patel K, Bunachita S, Agarwal AA, Bhamidipati A, Patel UK. A comprehensive overview of antibiotic selection and the factors affecting it. Cureus 2021;13:e13925. https://doi.org/10.7759/CUREUS.13925.
4. Zay Ya K, Win PTN, Bielicki J, Lambiris M, Fink G. Association between antimicrobial stewardship programs and antibiotic use globally: a systematic review and meta-analysis. JAMA Netw Open 2023;6:e2253806. https://doi.org/10.1001/JAMANETWORKOPEN.2022.53806.
5. Abdel Hadi H, Eltayeb F, Al Balushi S, Daghfal J, Ahmed F, Mateus C. Evaluation of hospital antimicrobial stewardship programs: implementation, process, impact, and outcomes, review of systematic reviews. Antibiotics (Basel) 2024;13:253. https://doi.org/10.3390/ANTIBIOTICS13030253.

6. Hollingshead CM, Khazan AE, Franco JH, Ciricillo JA, Haddad MN, Berry JT, et al. A needs assessment for infectious diseases consultation in community hospitals. Infect Dis Ther 2023;12:1725. https://doi.org/10.1007/S40121-023-00810-4.
7. Vaughn VM, Greene MT, Ratz D, Fowler KE, Krein SL, Flanders SA, et al. Antibiotic stewardship teams and Clostridioides difficile practices in United States hospitals: a national survey in the joint commission antibiotic stewardship standard era. Infect Control Hosp Epidemiol 2020;41:143e8. https://doi.org/10.1017/ICE.2019.313.
8. Sunenshine RH, Liedtke LA, Jernigan DB, Strausbaugh LJ, Infectious Diseases Society of America Emerging Infections Network. Role of infectious diseases consultants in management of antimicrobial use in hospitals. Clin Infect Dis 2004;38:934e8. https://doi.org/10.1086/382358.
9. Pulcini C, Botelho-Nevers E, Dyar OJ, Harbarth S. The impact of infectious disease specialists on antibiotic prescribing in hospitals. Clin Microbiol Infec 2014;20:963e72. https://doi.org/10.1111/1469-0691.12751.
10. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol 2024;281:2159e65. https://doi.org/10.1007/S00405-023-08441-8.
11. De Vito A, Colpani A, Moi G, Babudieri S, Calcagno A, Calvino V, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav 2024;28:2746e54. https://doi.org/10.1007/S10461-024-04391-2.
12. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, et al. Gemini AI vs. ChatGPT: a comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol 2025;263:527e36. https://doi.org/10.1007/S00417-024-06625-4.
13. De Vito A, Geremia N, Marino A, Bavaro DF, Caruana G, Meschiari M, et al. Assessing ChatGPT's theoretical knowledge and prescriptive accuracy in bacterial infections: a comparative study with infectious diseases residents and specialists. Infection 2024. https://doi.org/10.1007/s15010-024-02350-6.
14. Maillard A, Micheli G, Lefevre L, Guyonnet C, Poyart C, Canoui Ë, et al. Can Chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. Clin Infect Dis 2024;78:825e32. https://doi.org/10.1093/CID/CIAD632.
15. Kaneda Y. ChatGPT in infectious diseases: a practical evaluation and future considerations. New Microbe. New Infect 2023;54:101166. https://doi.org/10.1016/J.NMNI.2023.101166.
16. Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med 2024;62:2425e34. https://doi.org/10.1515/CCLM-2024-0246.
17. Cosentino F, Viale P, Giannella M. MDR/XDR/PDR or DTR? Which definition best fits the resistance profile of Pseudomonas aeruginosa? Curr Opin Infect Dis 2023;36:564e71. https://doi.org/10.1097/QCO.0000000000000966.
18. OpenAI. OpenAI O1 system card. https://openai.com/index/openai-o1-system-card/. [Accessed 1 November 2024].
19. Pirkle S, Yang JW, Blumberg TJ. Do ChatGPT and Gemini provide appropriate recommendations for pediatric orthopaedic conditions? J Pediatr Orthop 2025;45:e66e71. https://doi.org/10.1097/BPO.0000000000002797.
20. Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. J Orthop Surg Res 2024;19:574. https://doi.org/10.1186/S13018-024-04996-2.
21. Durkin MJ, Keller M, Butler AM, Kwon JH, Dubberke ER, Miller AC, et al. An assessment of inappropriate antibiotic use and guideline adherence for uncomplicated urinary tract infections. Open Forum Infect Dis 2018;5:ofy198. https://doi.org/10.1093/OFID/OFY198.
22. Fleming-Dutra KE, Hersh AL, Shapiro DJ, Bartoces M, Enns EA, File TM, et al. Prevalence of inappropriate antibiotic prescriptions among US ambulatory care visits, 2010-2011. JAMA 2016;315:1864e73. https://doi.org/10.1001/JAMA.2016.4151.
23. Vazquez Deida AA, Bizune DJ, Kim C, Sahrmann JM, Sanchez GV, Hersh AL, et al. Opportunities to improve antibiotic prescribing for adults with acute sinusitis, United States, 2016e2020. Open Forum Infect Dis 2024;11:ofae420. https://doi.org/10.1093/OFID/OFAE420.
24. Davey P, Marwick CA, Scott CL, Charani E, Mcneil K, Brown E, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. Cochrane Database Syst Rev 2017;2:CD003543. https://doi.org/10.1002/14651858.CD003543.pub4.
25. Ren M, So AD, Chandy SJ, Mpundu M, Peralta AQ, Åkerfeldt K, et al. Equitable access to antibiotics: a core element and shared global responsibility for pandemic preparedness and response. J Law Med Ethics 2022;50:34e9. https://doi.org/10.1017/JME.2022.77.
26. Wasan H, Reeta KH, Gupta YK. Strategies to improve antibiotic access and a way forward for lower middle-income countries. J Antimicrob Chemother 2024;79:1e10. https://doi.org/10.1093/JAC/DKAD291.