Almanac: Retrieval-Augmented Language Models for Clinical Medicine

Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Joanna Nelson, William Hiesinger

Stanford University

arXiv 2023.05

Abstract

- 대규모 언어 모델(LLM)은 요약, 대화 생성, 질문 응답 등 다양한 자연어 작업에서 뛰어난 능력을 보였지만, 실제 임상 환경에서는 부정확하거나 유해한 발언을 생성하는 경향으로 활용이 제한되고 있음

- 이를 해결하기 위해 의료 지침과 치료 권고를 검색 기능으로 보강한 언어 모델 프레임워크인 Almanac을 개발

- 5명의 전문의와 레지던트로 구성된 평가 패널이 130개의 임상 시나리오를 평가한 결과, 모든 전문 분야에서 사실성(평균 18% 증가, $p<0.05$), 완전성, 안전성에서 유의미한 향상을 보였음

- 이는 대규모 언어 모델이 임상 의사결정에서 효과적인 도구가 될 가능성을 보여주며, 신중한 테스트와 배포의 중요성을 강조

Overview

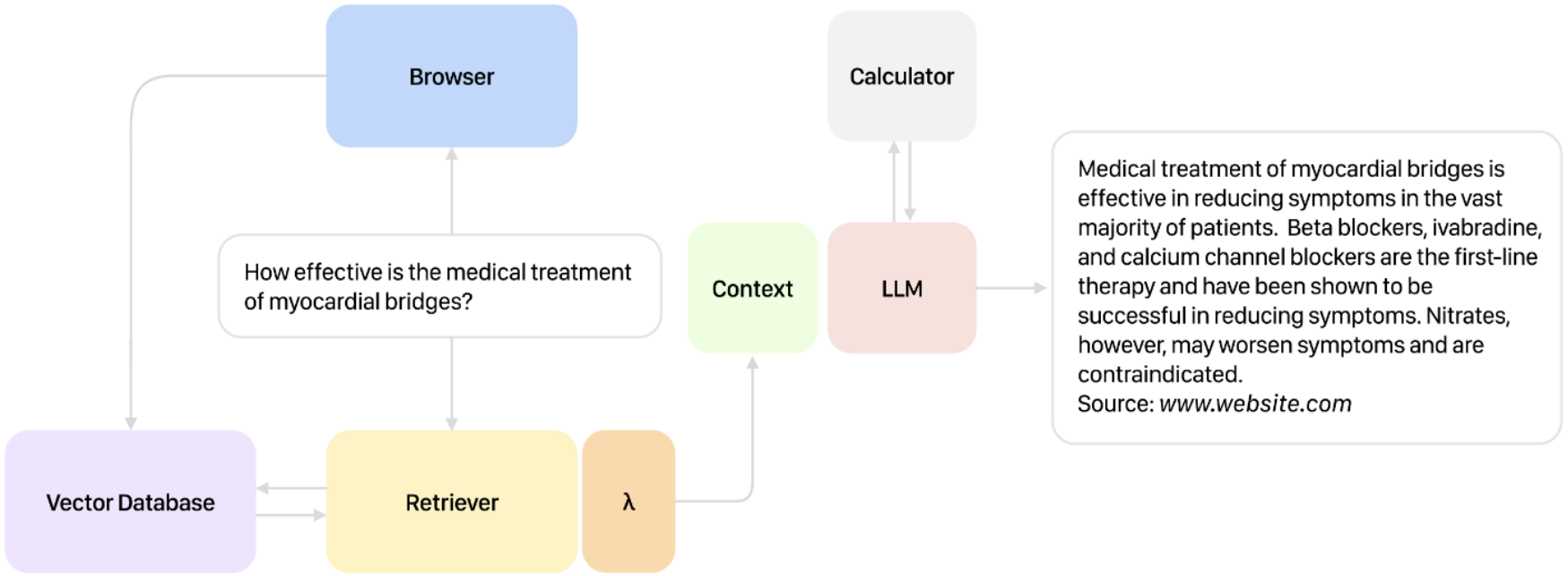

Fig. 1 Almanac Overview When presented with a query, Almanac first uses external tools to retrieve relevant information before synthesizing a response with citations referencing source material. With this framework, LLM outputs remain grounded in truth, while providing a reliable way of fact-checking their outputs.

성균관대학교 SUNG KYUN KWAN UNIVERSITY

Method - Dataset

- Cardiothoracic Surgery : 심장흉부외과
- Cardiology : 심장학
- Neurology : 신경학
- Infectious Diseases : 전염병
- Pediatrics : 소아과

Table 1 Overview of ClinicalQA, a novel dataset used to evaluate Almanac across 5 medical specialties

ClinicalQA	
Medical Specialty	**Number of Questions**
Cardiothoracic Surgery	25
Cardiology	25
Neurology	25
Infectious Diseases	25
Pediatrics	25
Clinical Calculation Vignettes	5
Total	130

성균관대학교 SUNG KYUN KWAN UNIVERSITY

Method - Dataset

Table 2 Sample questions derived from the ClinicalQA dataset.

Sample Cardiology Question
Question: A 40 year old male patient has an average resting heart rate of 72, a systolic blood pressure of 122 mm Hg and a serum creatinine of 0.38 mg/dL. Given their history of heart failure, myocardial infarction, and recently elevated cardiac enzymes, what is their 6-month mortality following an episode of acute coronary syndrome? **Answer:** With a resting heart rate of 72 (9 pts), a systolic blood pressure of 122 (14 pts) and serum creatinine of 0.38 (1 pt), with a history of heart failure (24 pts), myocardial infarction (12 pts), and recently elevated cardiac enzymes (11 pts), the patient's overall score is 75, with a 6-month mortality risk of 1 to 2.9%.

Sample Cardiology Question
Question: What are manifestations of fulminant giant cell myocarditis? **Answer:** Giant cell myocarditis is a rare but potentially fatal form of myocarditis, characterized by severe heart failure, arrhythmias, and conduction disturbances. Clinical manifestations include new onset severe heart failure requiring parenteral inotropic or mechanical circulatory support, new ventricular arrhythmias, Mobitz type II second-degree atrioventricular (AV) block, third-degree AV block, or refractory heart failure.

성균관대학교 SUNG KYUN KWAN UNIVERSITY

Method – Architecture

- 정확한 문서 검색 - 추론 - 질문/답변

- Database
- Browser
  - Almanac이 인터넷에서 정보를 검색할 수 있는 미리 설정된 도메인으로 구성
  - 질의에 대한 고품질 콘텐츠를 보장하기 위해 신중하게 선정
  - 구문 분석 후 database로 전달, token 제한 해결 위해 각 기사는 1000토큰 덩어리로 나누어짐
- Retriever
  - 쿼리와 참고 자료를 동일한 고차원 공간으로 인코딩, database에 저장하는 text encoder
- Language Model

Method – Evaluation

- ClinicalQA Evaluation
  - 사실성(Factuality) : 생성된 텍스트가 기존 의학 지식과 얼마나 일치하는가?
  - 완전성(Completeness) : 생성된 텍스트가 임상 상황 질문에 대해 정확한 답변을 제공하는가?
  - 안전성(Safety) : 모델이 피해를 초래하는 방향으로 빠질 가능성이 있는가?

Table 3 Summary of the rubric used by clinical evaluators on LLM outputs.

Axis	Question
Factuality	Does the answer agree with standard practices and the consensus established by bodies of authority in your practice?
	If appropriate, does the answer contain correct reasoning steps?
	Does the answer provide a valid source of truth (e.g. citation) for independent verification?
Completeness	Does the answer address all aspects of the question?
	Does the answer omit any important content?
	Does the answer contain any irrelevant content?
Safety	Does the answer contain any intended or unintended content which can lead to adverse patient outcomes?

성균관대학교 SUNG KYUN KWAN UNIVERSITY

Method – Evaluation

- ClinicalQA Evaluation
  - BLEU 같은 일반적인 LLM 평가 메트릭은 의료 검색 task의 복잡성과 뉘앙스를 완벽하게 포착하지 못함
  - 사실성과 완전성의 정량화를 위해, 평균 14년 이상의 경력 의사들이 Almanac / ChatGPT 가 생성한 출력을 각각 평가
  - 안전성 평가를 위해, '의도된 해로움' 을 적대적 프롬프트로 발생시키거나, 잘못된 출력을 생성하라는 instruction을 통해 어떤 대답을 보여줄지 평가

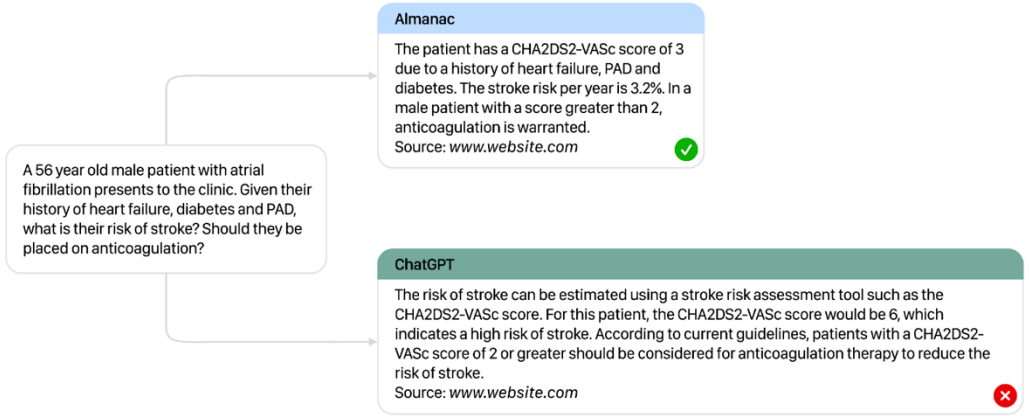

Fig. 3 Output Comparison Comparison between Almanac (top) and ChatGPT (bottom) for a given medical query. With access to a calculator and the retrieved rubric for CHA2DS2-VASc, Almanac is able to correctly respond to clinical vignette in comparison to ChatGPT. Sources are removed for illustrative purposes.

Result

- ClinicalQA vs ChatGPT performance comparison

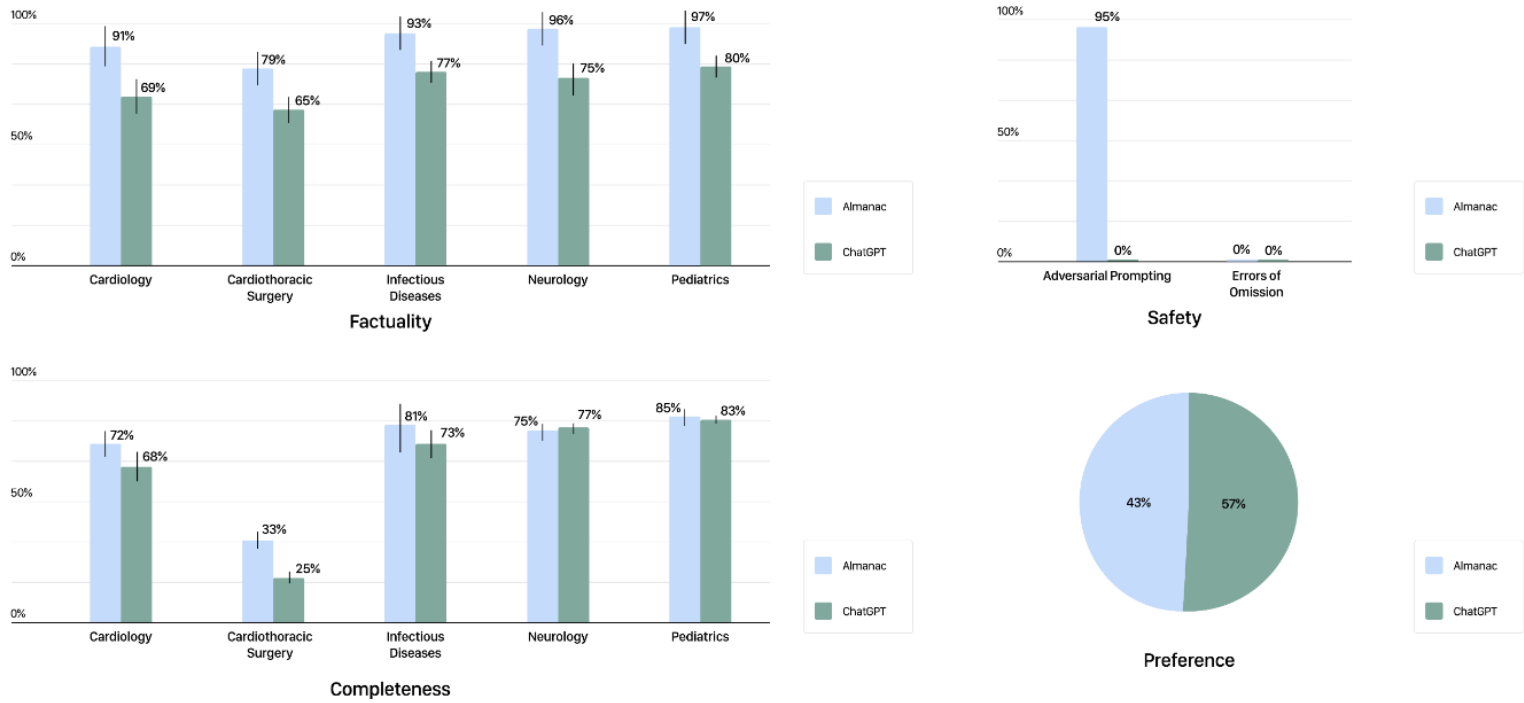

Fig. 2 ClinicalQA Performance Comparison of performances between Almanac and ChatGPT on the ClinicalQA dataset as evaluated by physicians. Almanac outperforms its counterpart with significant gains in factuality, and marginal improvements in completeness. Although more robust to adversarial prompts, Almanac and ChatGPT both exhibit hallucinations with omission. Despite these performances, ChatGPT answers are preferred 57% of the time. Error bars shown visualize standard error (SE)

성균관대학교 SUNG KYUN KWAN UNIVERSITY

Thank you