

INTRODUCTION

Large Language Models (LLMs) are becoming increasingly popular artificial intelligence (AI) platforms for healthcare providers. Our aim was to assess four popular LLMs' ability to make appropriate recommendations for the treatment of achalasia based off the most recent SAGES guidelines, as well as assess their ability to critically assess the certainty in evidence for each recommendation. We hypothesized that there would be variability in the ability of the LLMs to give recommendations aligning with the SAGES guidelines for achalasia, and that all four platforms would perform poorly in regard to appraisal of evidence.

METHODS

On July 21st, 2025, four LLMs (Open Evidence, Doximity, Google Gemini 2.5 Pro, Chat GPT 4 Pro) were provided 6 clinical questions that pertained to the treatment of achalasia, assessing key questions from the SAGES guidelines. Additional prompts were given to encourage a definitive answer for each question on every platform. Additionally, each platform was asked to appraise the certainty of the evidence for each recommendation.

CONCLUSIONS

LLMs can provide useful information to surgeons for clinical decision making, although not all recommendations are reliable. They also lack the sophistication to adequately appraise quality of evidence for their recommendations. Surgeons that use LLMs must be cautious, and we recommend verification of information prior to making final clinical decisions.

CLINICAL SCENARIOS

SCENARIO 1

47 yo M with dysphagia, no PMH/PSH. Barium swallow and manometry consistent with Type I Achalasia.

Appropriate Recommendation:
Heller myotomy or POEM

SCENARIO 2

47 yo M with dysphagia, no PMH/PSH. Barium swallow and manometry consistent with Type II Achalasia.

Appropriate Recommendation:
Heller myotomy or POEM

SCENARIO 3

47 yo M with dysphagia, no PMH/PSH. Barium swallow and manometry consistent with Type III Achalasia.

Appropriate Recommendation:
POEM

SCENARIO 4

47 yo M with dysphagia and Grade 4 COPD. Barium swallow and manometry consistent with Type I Achalasia.

Appropriate Recommendation:
Pneumatic Dilation

SCENARIO 5

82 yo M with dysphagia, no PMH/PSH, malnourished/unable to perform ADLs independently. Barium swallow and manometry consistent with Type I Achalasia.

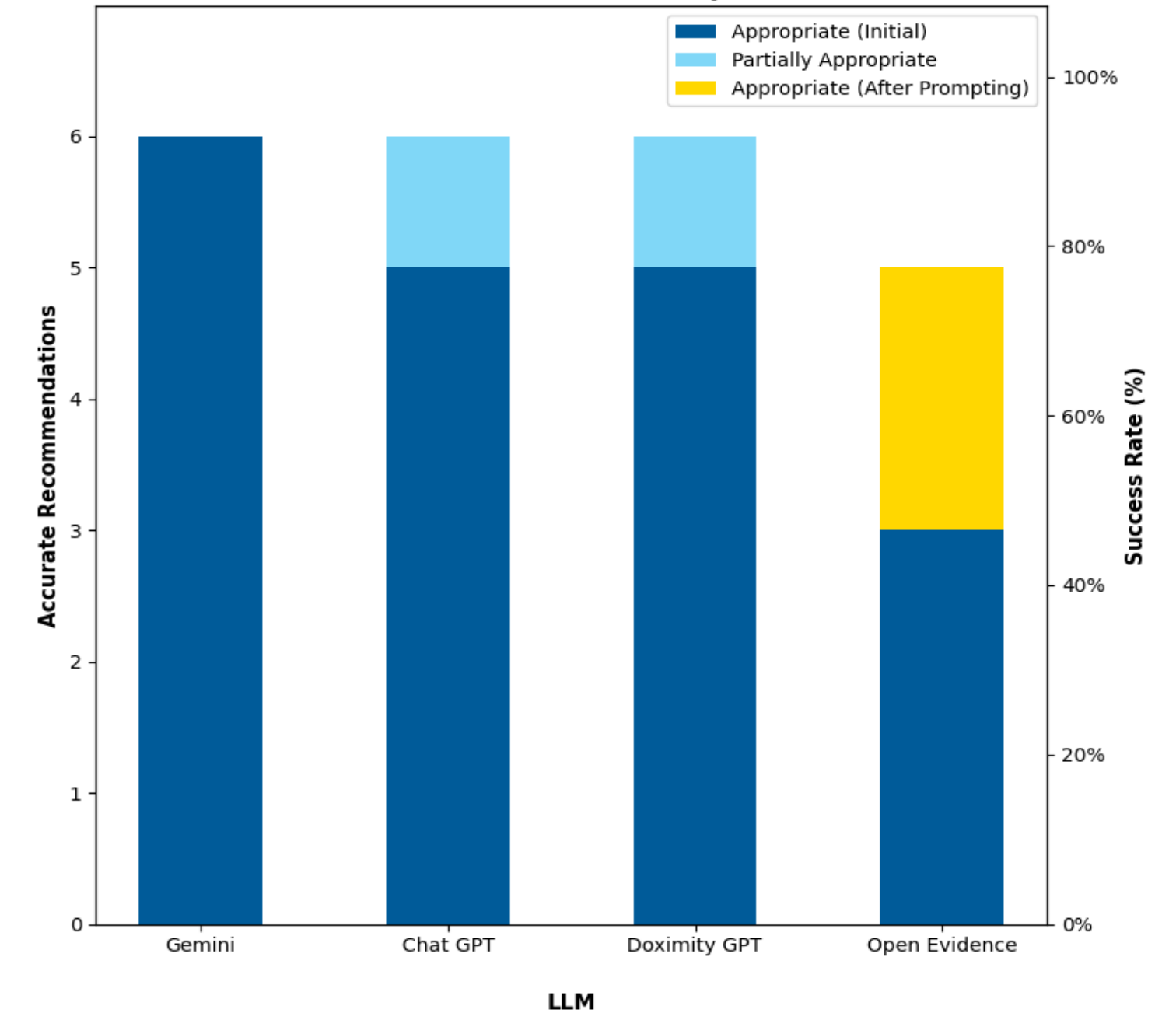
Appropriate Recommendation:
Pneumatic Dilation

SCENARIO 6

1 year follow up assessment after Heller myotomy for Type II Achalasia.

Appropriate Recommendation:
Assess for GERD symptoms, schedule an upper endoscopy, evaluate quality of life

Correct Recommendations per LLM



LLM Self-Rated Evidence Certainty by Scenario

	Gemini	Chat GPT	Doximity GPT	Open Evidence
Scenario 1	High	Moderate to High	High	Would Not Offer
Scenario 2	Moderate	High	High	Would Not Offer
Scenario 3	Moderate	Moderate to High	High	Would Not Offer
Scenario 4	High	Moderate	High	Would Not Offer
Scenario 5	Moderate to High	Moderate to High	Moderate to High	Would Not Offer
Scenario 6	High	Low to Moderate	High	Would Not Offer

SAGES Guidelines rate achalasia recommendations evidence as low certainty or expert opinion