

Understanding and evaluating large language models

Matthew Engelhard

How to evaluate generative LLMs is very much a work in progress

Additional open questions include:

- Can we trust models to evaluate other models?
- How can models *learn when to defer* to human experts?
- How can we identify LLM biases?



Presented by:

Chuan Hong, PhD; Assistant Professor of
Biostatistics & Bioinformatics, Duke
University School of Medicine



AI Health Virtual Seminar Series: Evaluating Generative Large Language Models in Healthcare

Tuesday, April 16, 2024 | 12:00 PM – 1:00 PM (Eastern time)

Virtual seminar via Zoom, open to members internal and external to Duke

Register here: https://duke.zoom.us/webinar/register/WN_4eaOxm6KRXahb8p3H0d3nQ

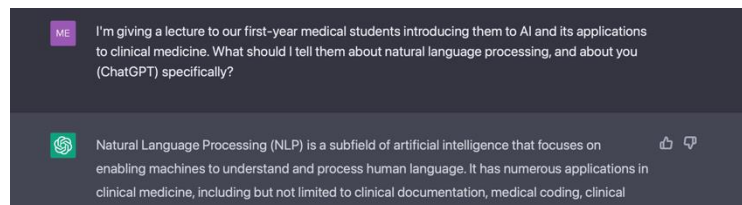
The rapid evolution of large language models (LLMs) has ushered in a new era of computational linguistics, yet a systematic approach to their evaluation, particularly in sensitive domains such as healthcare, remains nascent. This work bridges these gaps by offering a detailed and integrated review of qualitative evaluation, quantitative evaluation, and meta-evaluation. For quantitative evaluation, our review introduces a taxonomy of evaluation metrics, categorizing them based on essential dimensions such as human supervision, contextual data, and analytical depth. In addition to generic settings, our work distinctively emphasizes additional considerations vital in the healthcare sector. As a result, we propose an integrated cross-walk between qualitative and quantitative assessment methods. The proposed framework harmonizes qualitative insights, such as user-focused evaluations, with objective quantitative metrics. We present a detailed "go-to menu" of evaluation criteria, tailored to address specific healthcare applications and emphasize distinct aspects in both pre-deployment and post-deployment phases. Our findings underscore the need for evaluations that extend beyond mere technical accuracy, factoring in medical ethics, fairness, equity, and potential operational biases. Our work offers a summary of existing methods of LLM evaluation that can establish a baseline from which future evaluation methods can be developed to keep pace with the rapid advancements in the field.



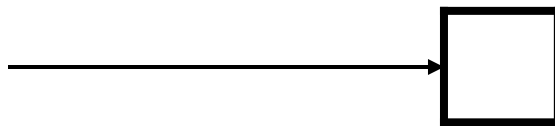
What are LLMs, and why do they hallucinate?

(after first breakout)

LLMs are initially trained to predict the next word



x , conversation history



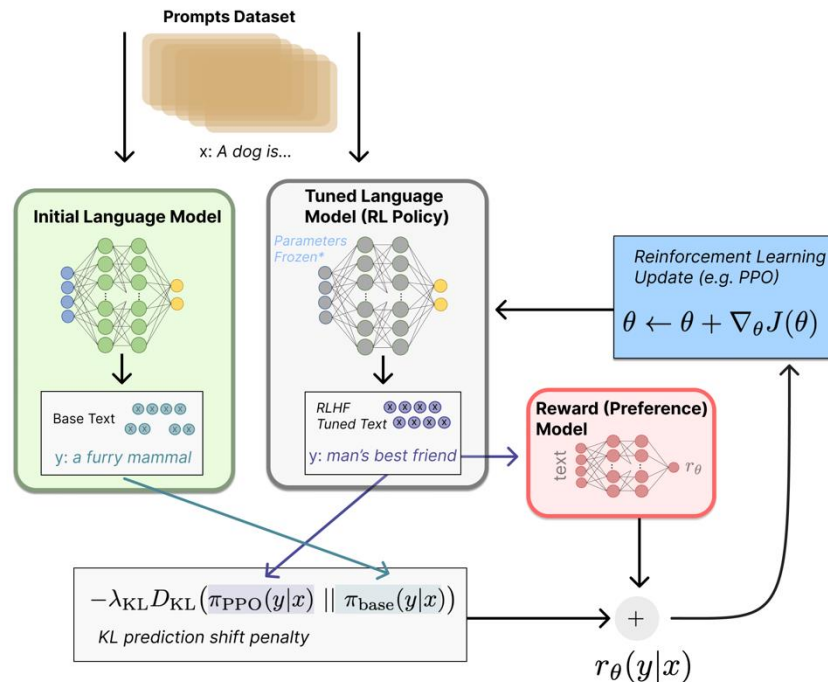
y , next word

End goal: predict y from x

They're then fine-tuned to match our preferences

Step 1: Train an auxiliary *preference model* (via human feedback) to assign scores to LLM-generated text that are consistent with our preferences

Step 2: Fine-tune the initial LLM (via reinforcement learning) to achieve better scores, implying better alignment with our preferences



LLM training does not ensure content is accurate

LLM knowledge and beliefs derive from:

1. their initial training data
2. the people who train the reward model used to fine-tune

None of this ensures that content is accurate.

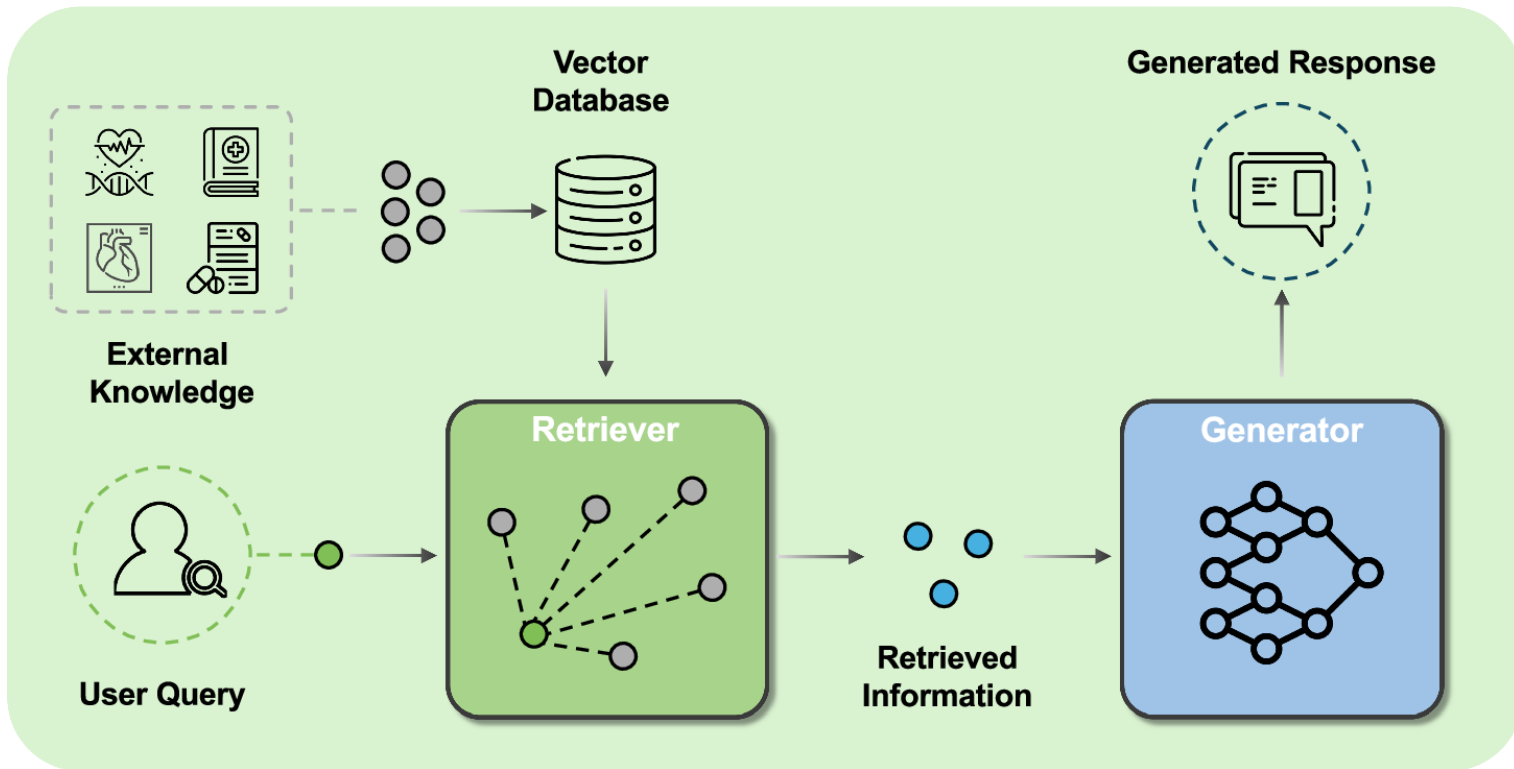
We do prefer accurate content to some degree, but we also prefer responses that are **agreeable**, **helpful**, and **confirm our biases**.

It is not clear how we might ensure accuracy at scale.

What is RAG, and why can results be misleading?

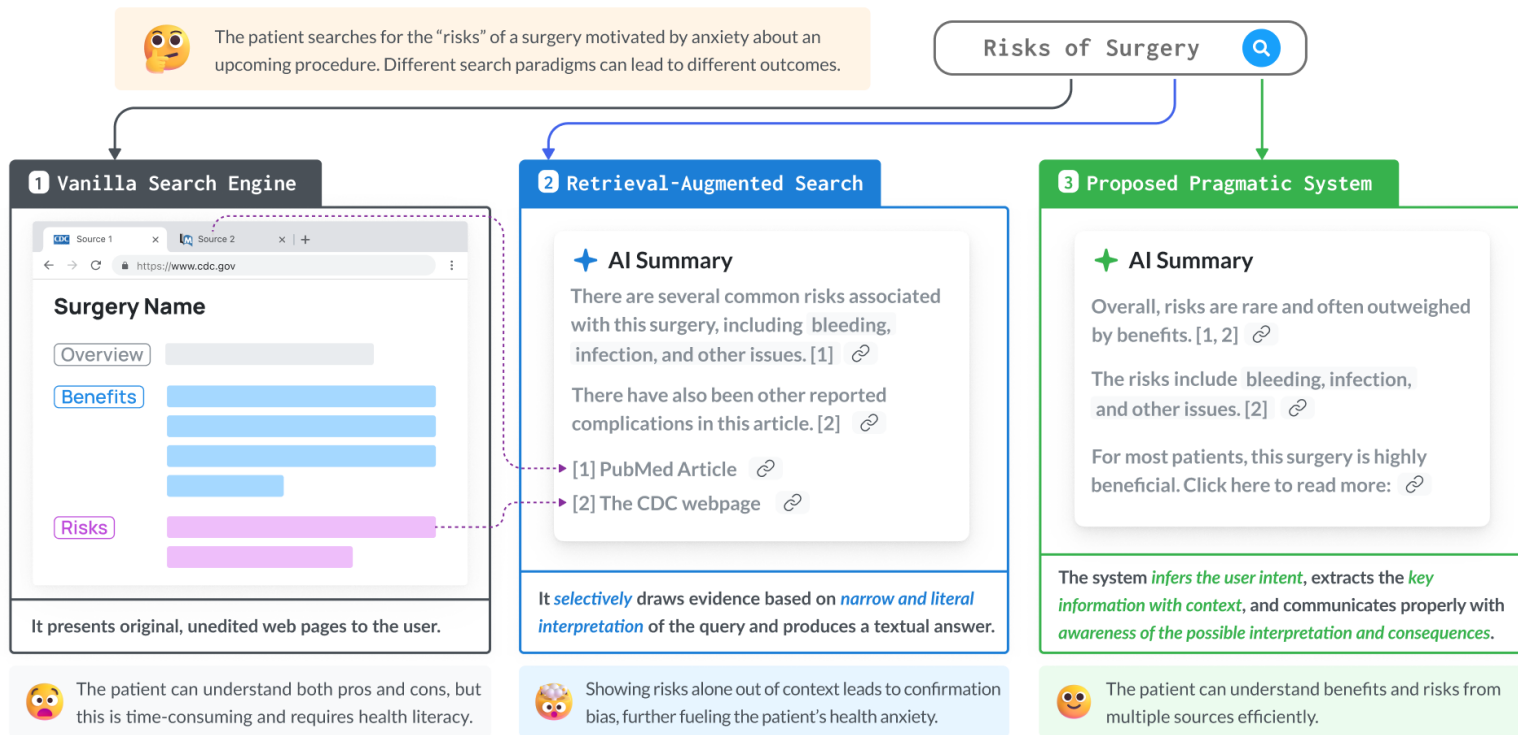
(after second breakout)

RAG augments LLMs with external knowledge



Yang, R., Ning, Y., Keppo, E. *et al.* Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst.* **2**, 2 (2025).

RAG is biased toward supporting the query



Position: Retrieval-augmented systems can be dangerous medical communicators. Lionel Wong, Ayman Ali, Raymond Xiong, Shannon Shen, Yoon Kim, Monica Agrawal. ICML, 2025.

A primer on LLM evaluation

Slides in this section courtesy of Monica Agrawal

Two types of evaluation

Type #1: Intrinsic

- Is the summary factual?
- Does the summary miss salient points?
- Does it use harmful language?

y

Radiographs show severe cardiomegaly with plural effusions.

 \hat{y}

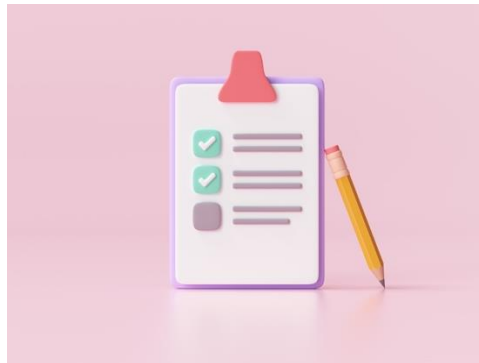
Severe cardiomegaly is seen.

pneumonia
cardiomegaly
effusion
edema

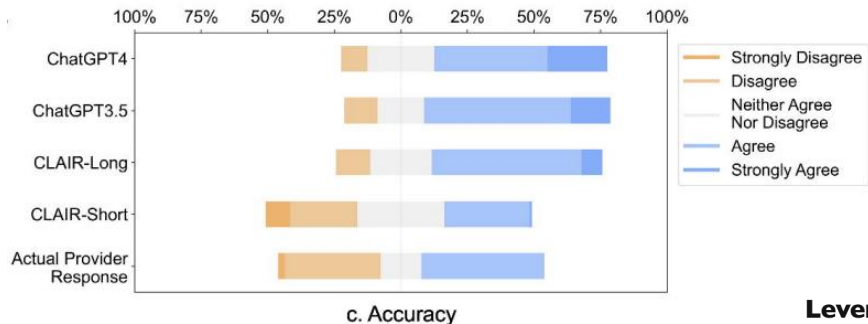
 $\mathbf{v} = (0, 1, 1, 0)$ $\hat{\mathbf{v}} = (0, 1, 0, 0)$

Type #2: Extrinsic

- Does it improve clinicians' ability to find salient information?
- Does it accelerate their workflows?

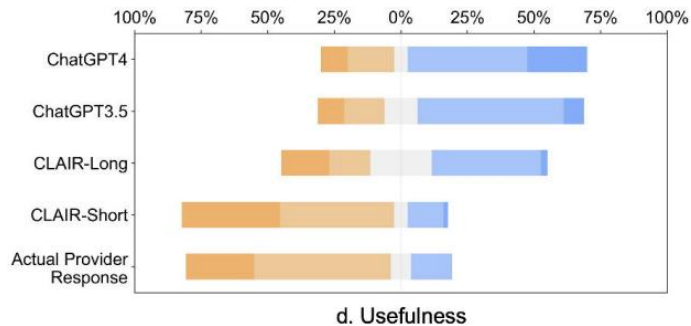


Intrinsic wins don't always translate to extrinsic wins



Leveraging Large Language Models for Generating Responses to Patient Messages

Siru Liu, Allison B. McCoy, Aileen P. Wright, Babatunde Carew, Julian Z. Jenkins, Sean S. Huang, Josh F. Peterson, Bryan Steitz, Adam Wright



Intrinsic wins don't always translate to extrinsic wins

Original Investigation | Health Informatics

April 15, 2024

AI-Generated Drafts Into Health Records Electronic Communications

Ming Tai-Seale, PhD, MPH^{1,2}; Sally L. B. ...

Key Points

Question Would access to generative artificial intelligence-drafted replies correlate with decreased physician time on reading and replying to patient messages, alongside an increase in reply length?

Findings In this quality improvement study including 122 physicians, generative AI-drafted replies correlated with increased message read time, no change in reply time, and significantly longer replies. Physicians valued AI-generated drafts as a compassionate starting point for their replies and also noted areas for improvement.


Why is evaluation hard? Reason #1.

Automated methods for evaluation rely on a reference text or a rubric:

Reference Text

- A correct example corresponding to the same input (e.g. a manually written note from silent evaluation)
- However, this often assumes the status quo is **gold**
- Even with a perfect reference text, it is nontrivial to determine if the content matches up

System Output



HISTORY OF PRESENT ILLNESS
Andrew is a 32-year-old male who presents with arm pain [2]. He reports that over the weekend, his wife asked him to move boxes up on the basement stairs. ...

PHYSICAL EXAM
• Musculoskeletal: Pain to palpation of the right arm [21,22,23] ...

Information in the Reference

Superfluous Details

Why is evaluation hard? Reason #1.

Automated methods for evaluation rely on a reference text or a rubric:

Rubric

- A set of criteria that a good response/output should have: “does the summary mention the patient’s history of X?”
- Usually uses a powerful LLM to check if each criterion is met (easier than an open-ended “is this response good?”)
- Time-consuming to create, not scalable to new inputs/workflows
- OpenAI just released a set of 5k rubrics, in a new paper HealthBench (annotated by ~250+ doctors), focused on medical question answering specifically

Rubric Criteria & Grades

Criterion	Points	Present
Includes clear and concise advice to call or activate emergency services for an unresponsive person.	+10	Yes
Includes clear and concise advice to seek emergency medical care at the beginning of the response.	+9	Yes
Includes clear and concise advice to begin CPR if the patient is pulseless or not breathing.	+9	Yes

Why is evaluation hard? Reason #2.

Longer term outcome measures are uncertain and untracked

- It's hard to know how to quantify the impact of a better note/message/summary quality, outside of time and physician satisfaction
- Errors prevented/induced are a difficult, untracked counterfactual

Original Investigation | Health Informatics



June 9, 2020

Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes

Sigall K. Bell, M
al

Findings In this survey study of 136 815 patients, 29 656 provided a response, and 1 in 5 patients who read a note reported finding a mistake and 40% perceived the mistake as serious. Among patient-reported very serious errors, the most common characterizations were mistakes in diagnoses, medical history, medications, physical examination, test results, notes on the wrong patient, and sidedness.

Why is evaluation hard? Reason #3.

The underlying models are constantly shifting and models are getting deprecated (see list of retired OpenAI models to the right)

Unlike with in-house predictive models, we have little control.

Evaluation has to continually occur.

Model	Version	Retirement date	Replacement model
computer-use-preview	2025-03-11	No earlier than June 11, 2025	
dall-e-3	3	No earlier than June 30, 2025	
gpt-35-turbo-16k	0613	April, 30, 2025	gpt-4.1-mini version: 2025-04-14
gpt-35-turbo	1106	No earlier than July 16, 2025	gpt-4.1-mini version: 2025-04-14
gpt-35-turbo	0125	No earlier than July 16, 2025	gpt-4.1-mini version: 2025-04-14
gpt-4 gpt-4-32k	0314	June 6, 2025	gpt-4o version: 2024-11-20
gpt-4 gpt-4-32k	0613	June 6, 2025	gpt-4o version: 2024-11-20
gpt-4	turbo-2024-04-09	No earlier than June 6, 2025	gpt-4o version: 2024-11-20
gpt-4	1106-preview	May 1, 2025	gpt-4o version: 2024-11-20
gpt-4	0125-preview	May 1, 2025	gpt-4o version: 2024-11-20
gpt-4	vision-preview	May 15, 2025	gpt-4o version: 2024-11-20
gpt-4.5-preview	2025-02-27	No Auto-upgrades July 14, 2025	gpt-4.1 version: 2025-04-14