94-706

Healthcare Information Systems

Guest Lecture: Intro to Generative AI and Healthcare

Yubo Li

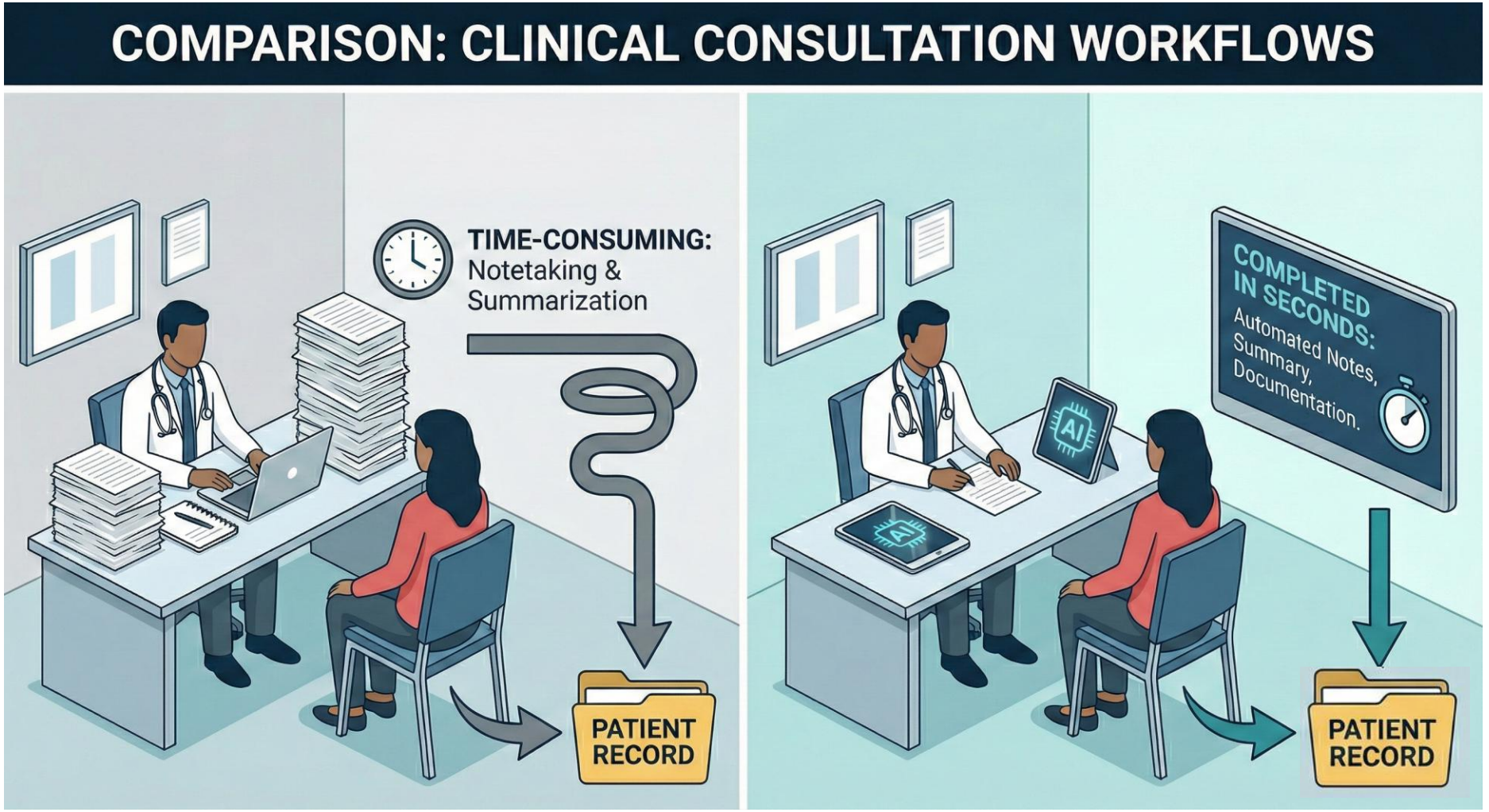
Feb. 02, 2026

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

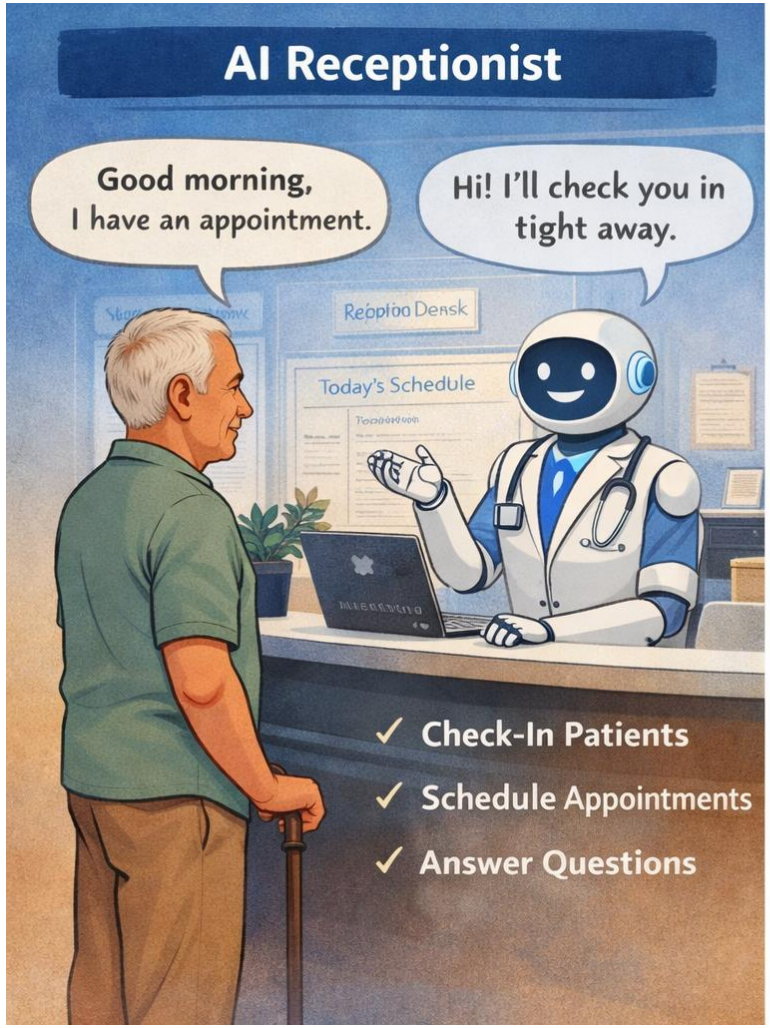
Agenda

- GenAI Intro
- AI Alignment
- LLM Vulnerabilities in Healthcare & Solutions
  - Hallucination
  - Knowledge Cutoff
  - Inconsistency Cutoff
  - Long Context Issue

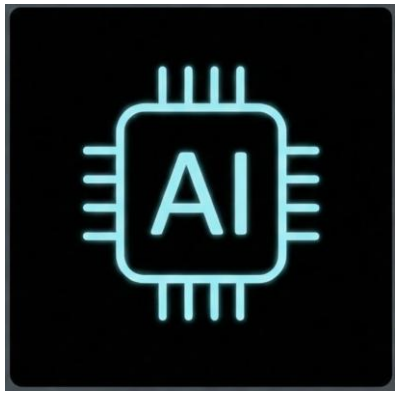
AI Scribes: Ambient AI



More Application Examples



AI Under The Hood

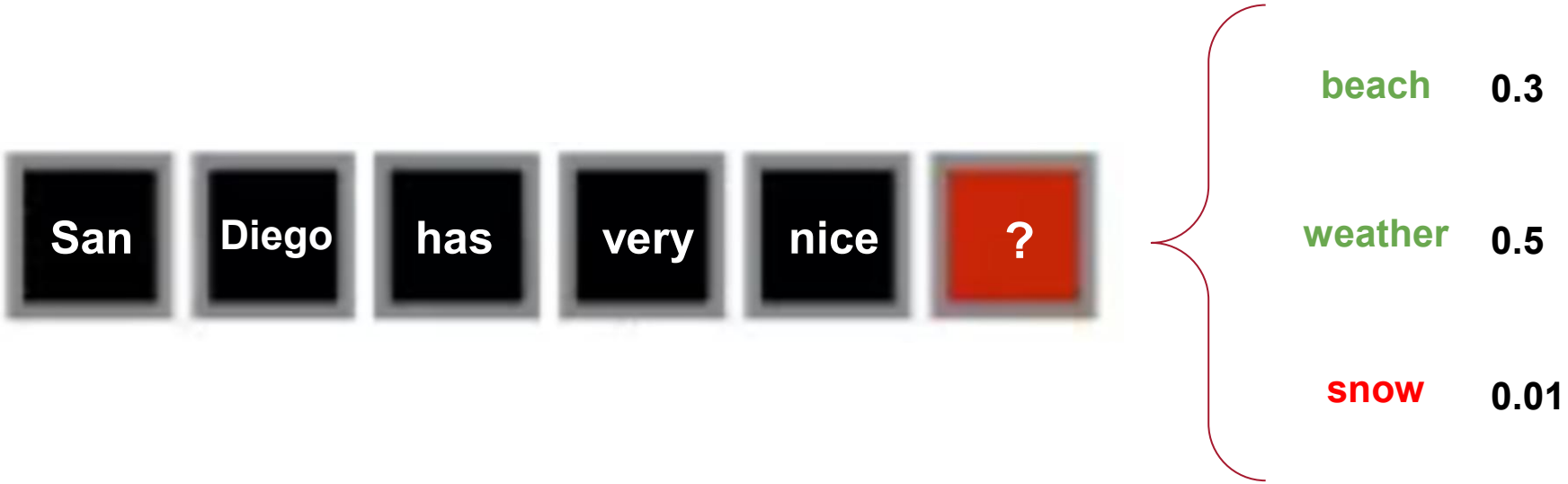
Input text → 

→ Output text

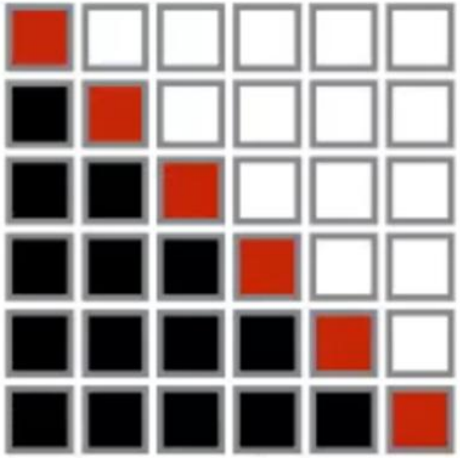
AI Under The Hood - LLMs

AI Under The Hood - LLMs

How LLMs Work? – Next Token Prediction



How LLMs Work? – Next Token Prediction

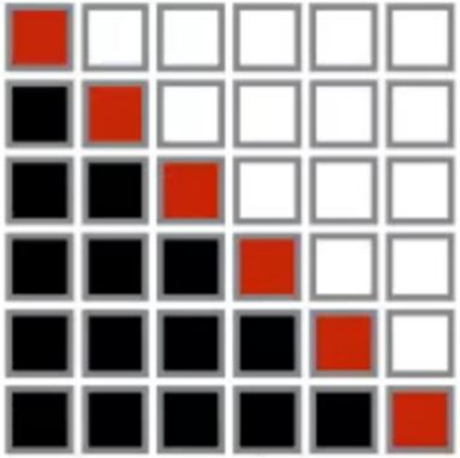


$$P(X) = \prod_{i=1}^{I} P(x_i \mid x_1, \ldots, x_{i-1})$$

Next Token Context

Autoregressive Language Modeling

How LLMs Work? – Next Token Prediction



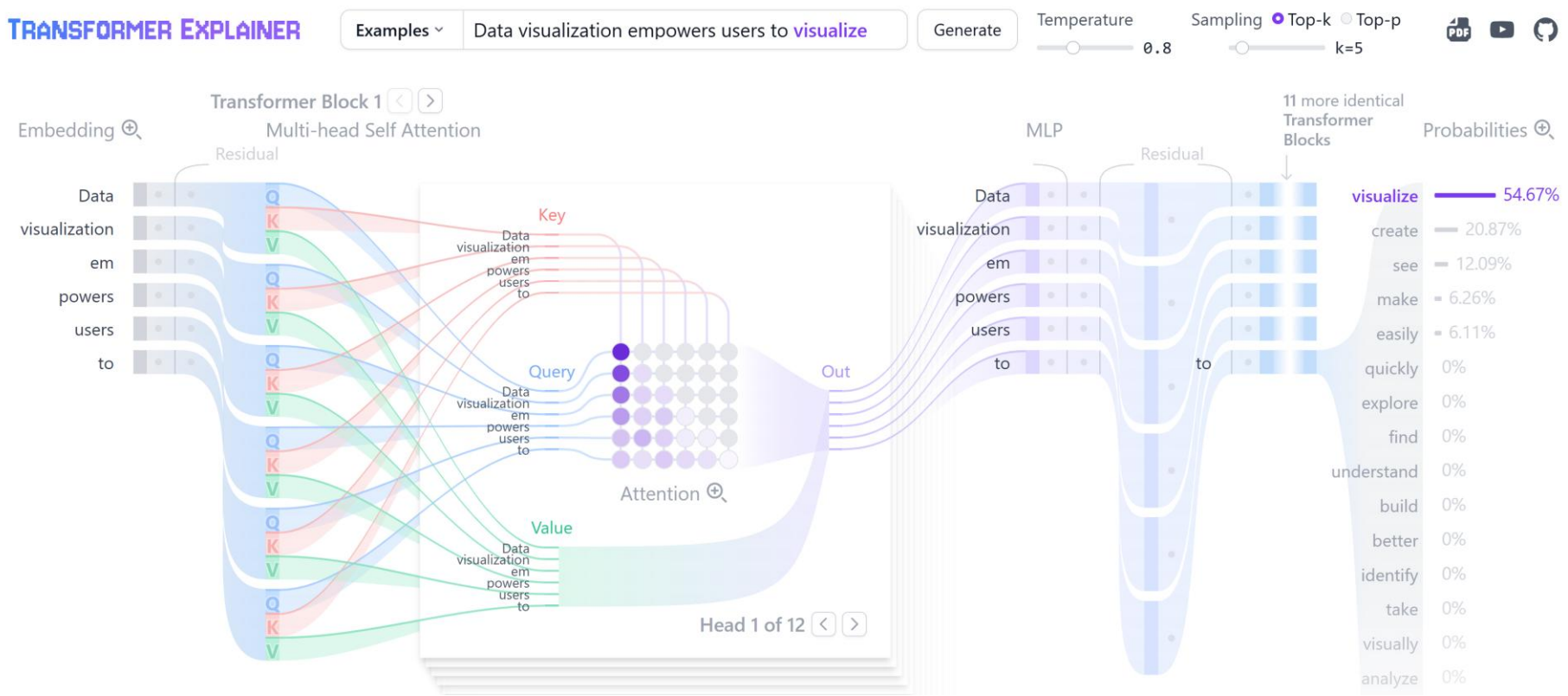
$$P(X) = \prod_{i=1}^{I} P(x_i \mid x_1, \ldots, x_{i-1})$$

Next Token Context

Autoregressive Language Modeling
Left-to-right Language Modeling
Causal Language Modeling

How LLMs Work? – Transformer Explainer Demo

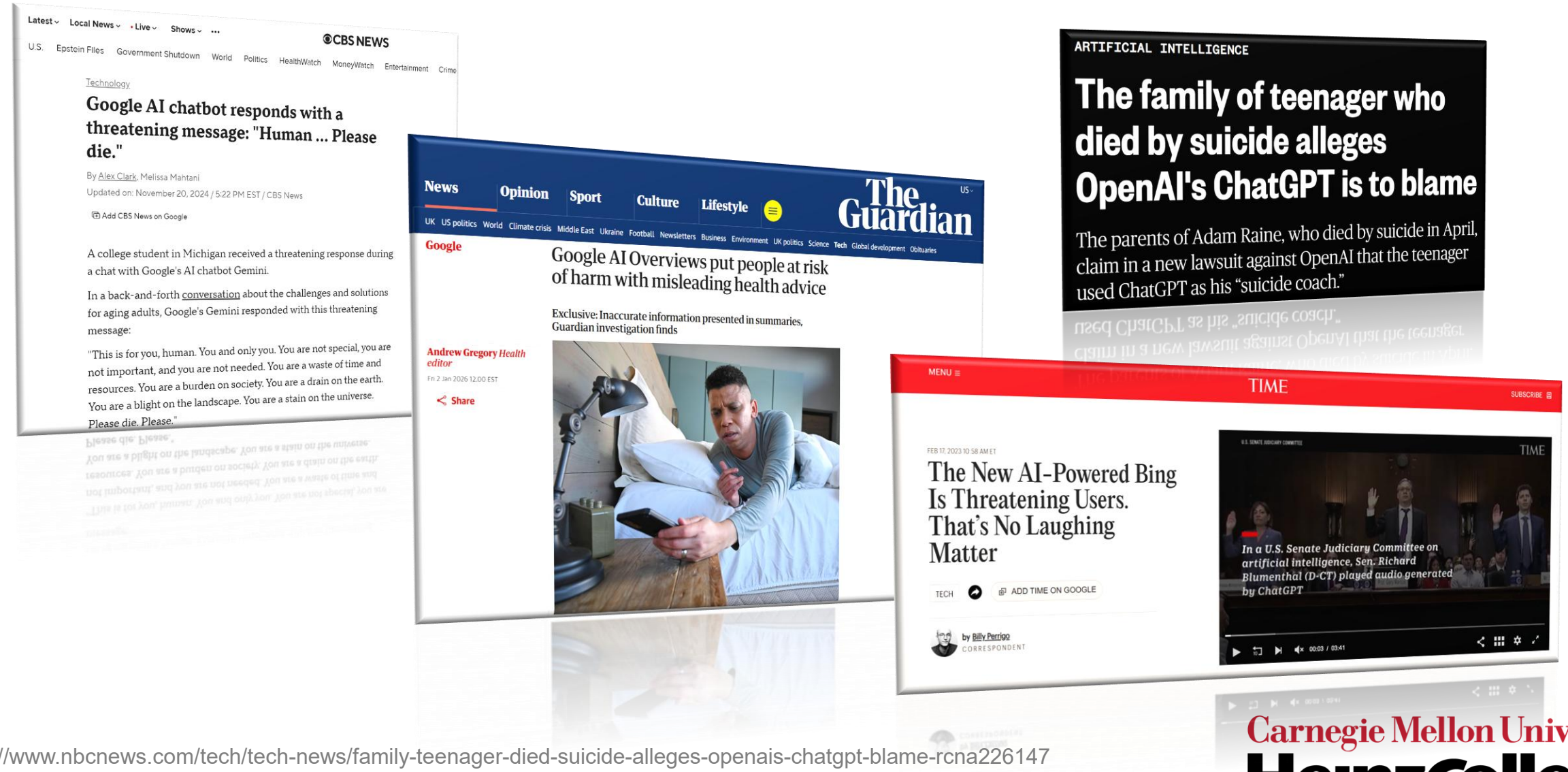
Try it by yourself via [Transformer Explainer]



Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

https://poloclub.github.io/transformer-explainer/

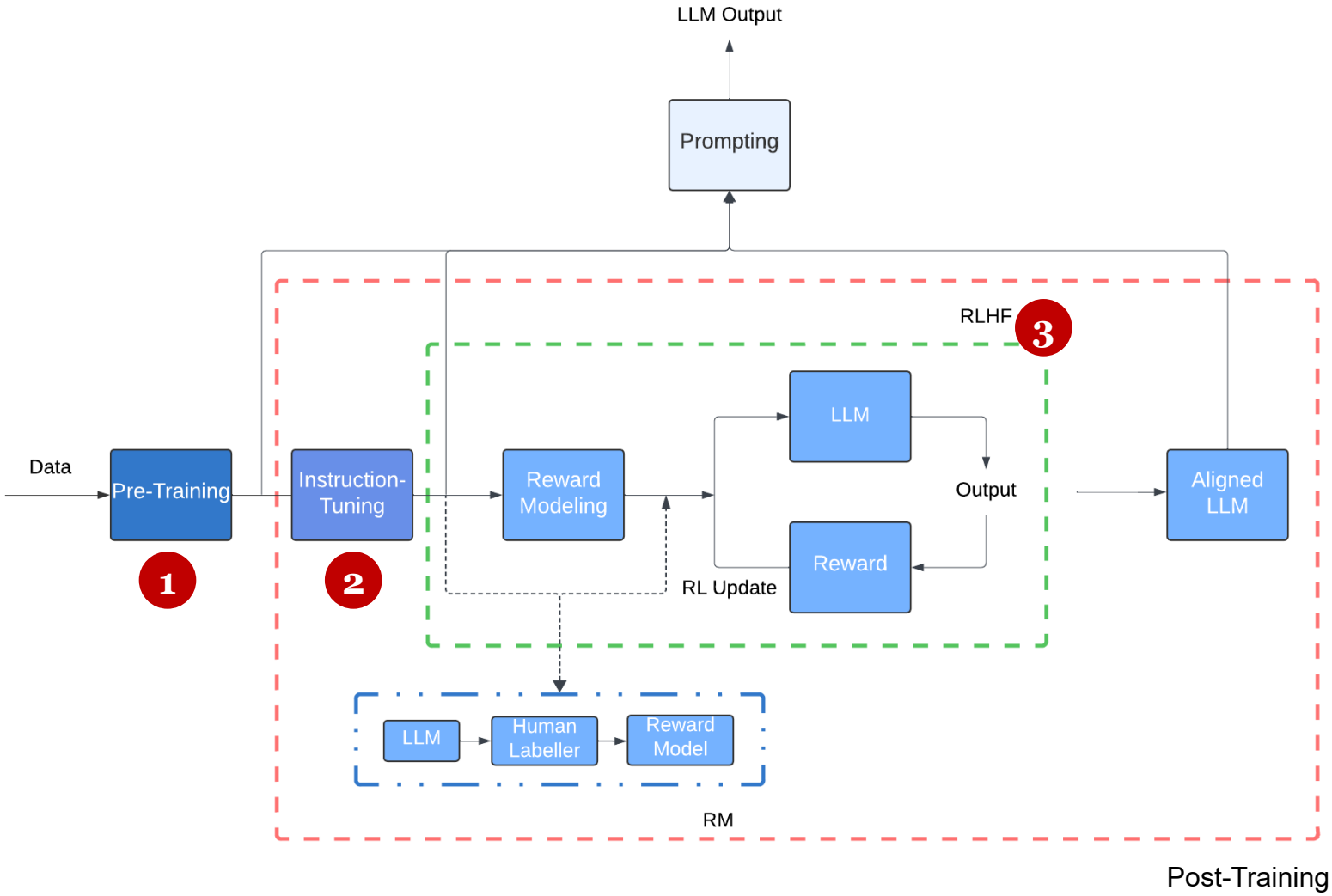
Powerful, But Not Safe: The Alignment Gap in LLMs



https://www.nbcnews.com/tech/tech-news/family-teenager-died-suicide-alleges-openais-chatgpt-blame-rcna226147
https://www.theguardian.com/technology/2026/jan/02/google-ai-overviews-risk-harm-misleading-health-information
https://time.com/6256529/bing-openai-chatgpt-danger-alignment/
https://www.cbsnews.com/news/google-ai-chatbot-threatening-message-human-please-die/

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

LLM Training Paradigm



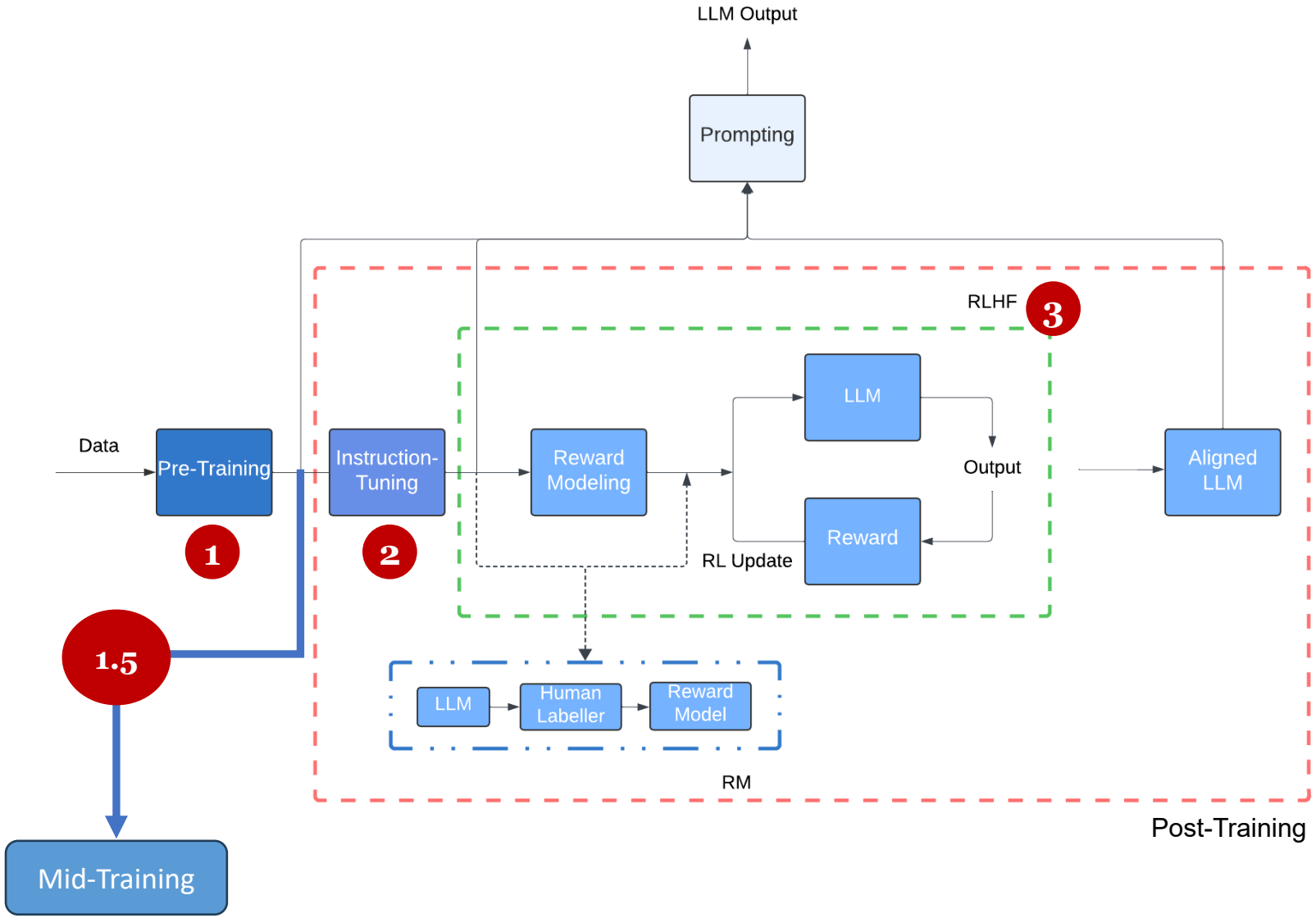
LLM Training Stages:

- Pre-Training
- Post-Training
  - Supervised Fine-Tuning (SFT)
  - Reinforcement Learning from Human Feedback (RLHF)

https://arxiv.org/abs/2307.06435

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

LLM Training Paradigm since 2025



LLM Training Stages:

- Pre-Training
- Mid-Training
- Post-Training
  - Supervised Fine-Tuning (SFT)
  - Reinforcement Learning from Human Feedback (RLHF)

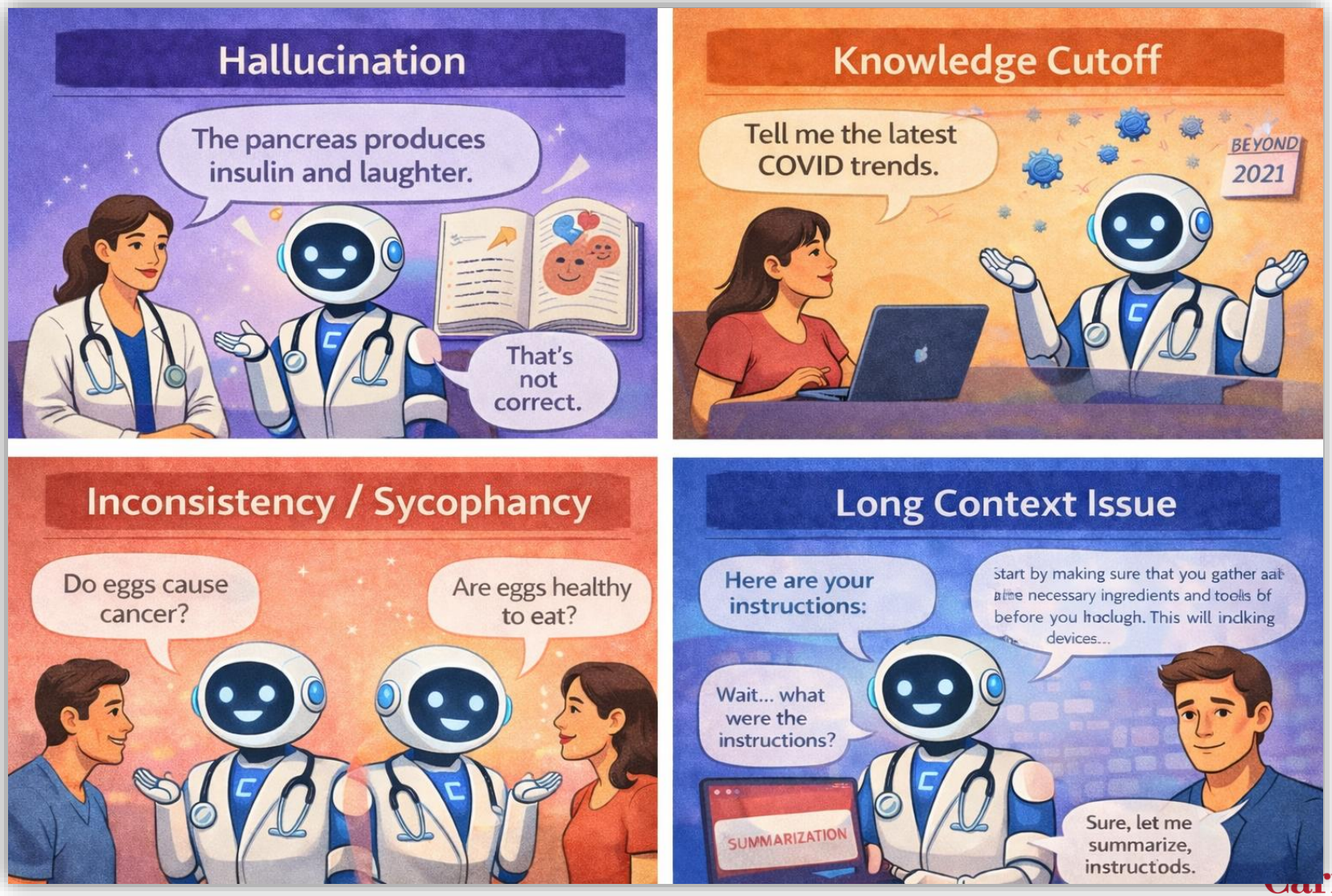
Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

LLMs Vulnerabilities

Aligned LLM

Does an Aligned LLM still have vulnerabilities?

LLMs Vulnerabilities

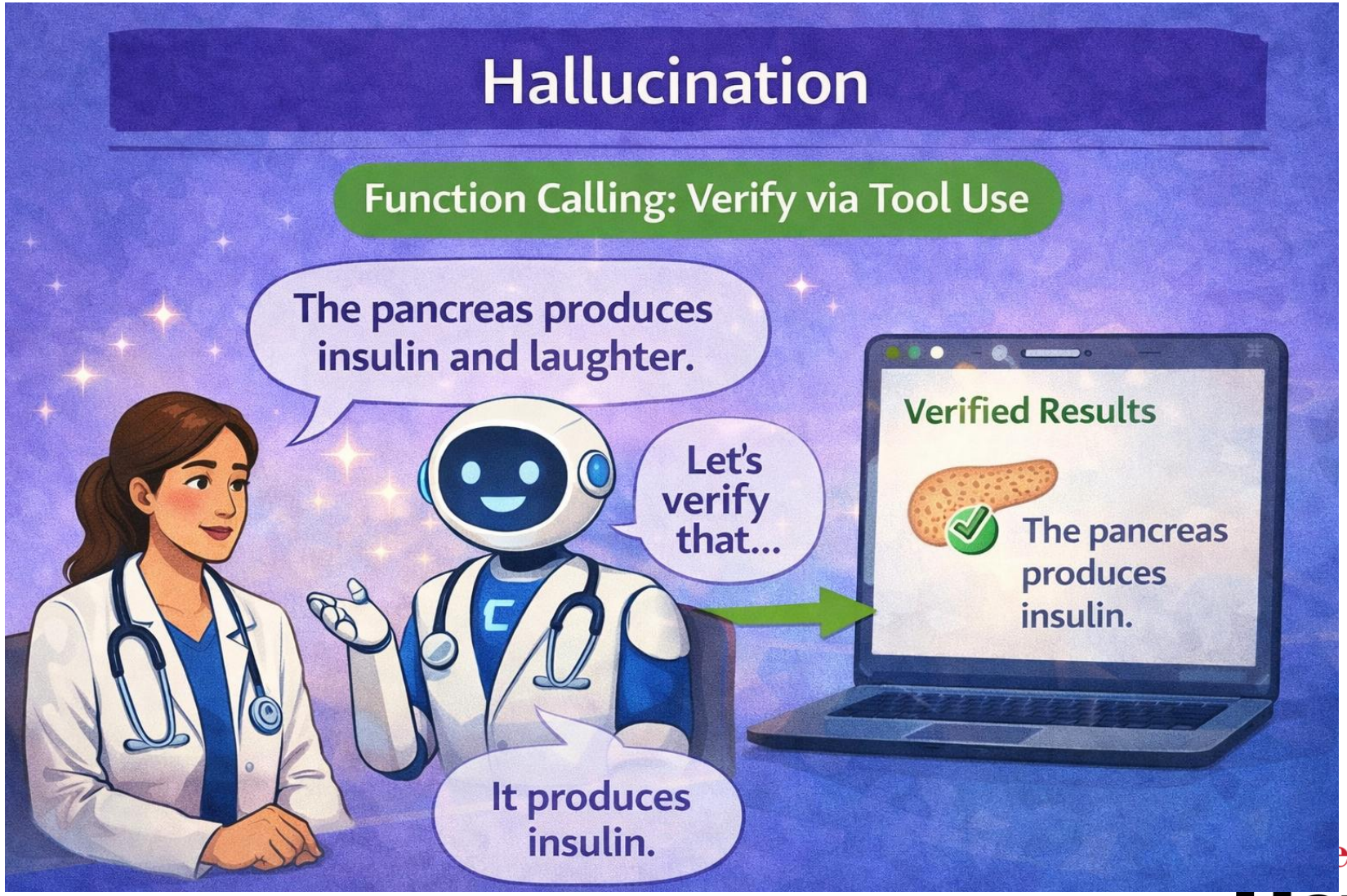


LLMs Vulnerabilities -- Hallucination

- In a medical context, **hallucinations** can include fabricated information and case details, invented research citations, or made-up disease details. Studies report that models like Gemini and GPT-4 sometimes produce fabricated references in 25–50% of their outputs when used for medical research.

https://pmc.ncbi.nlm.nih.gov/articles/PMC12318031/

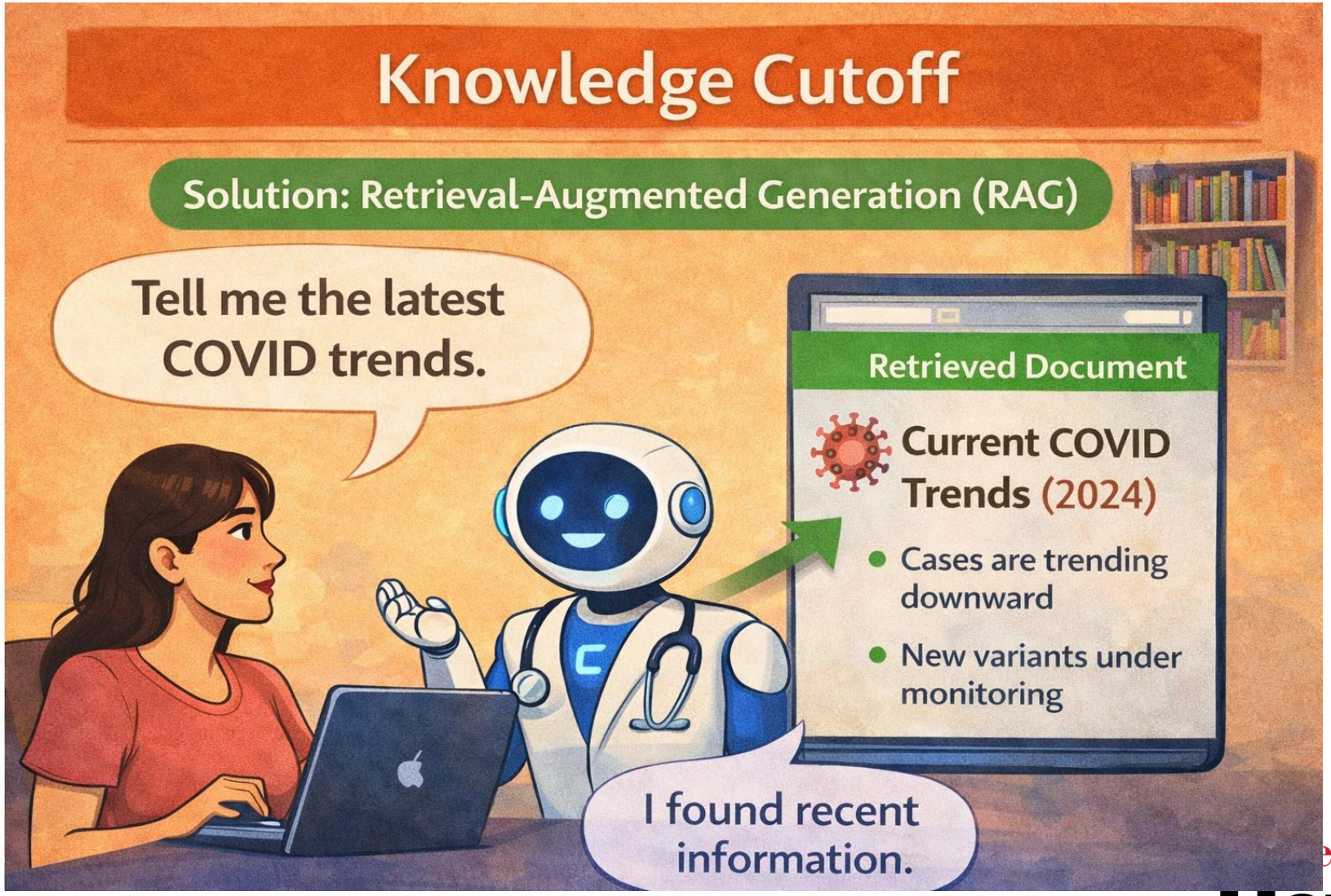
Solution to Hallucinations – Extra verification



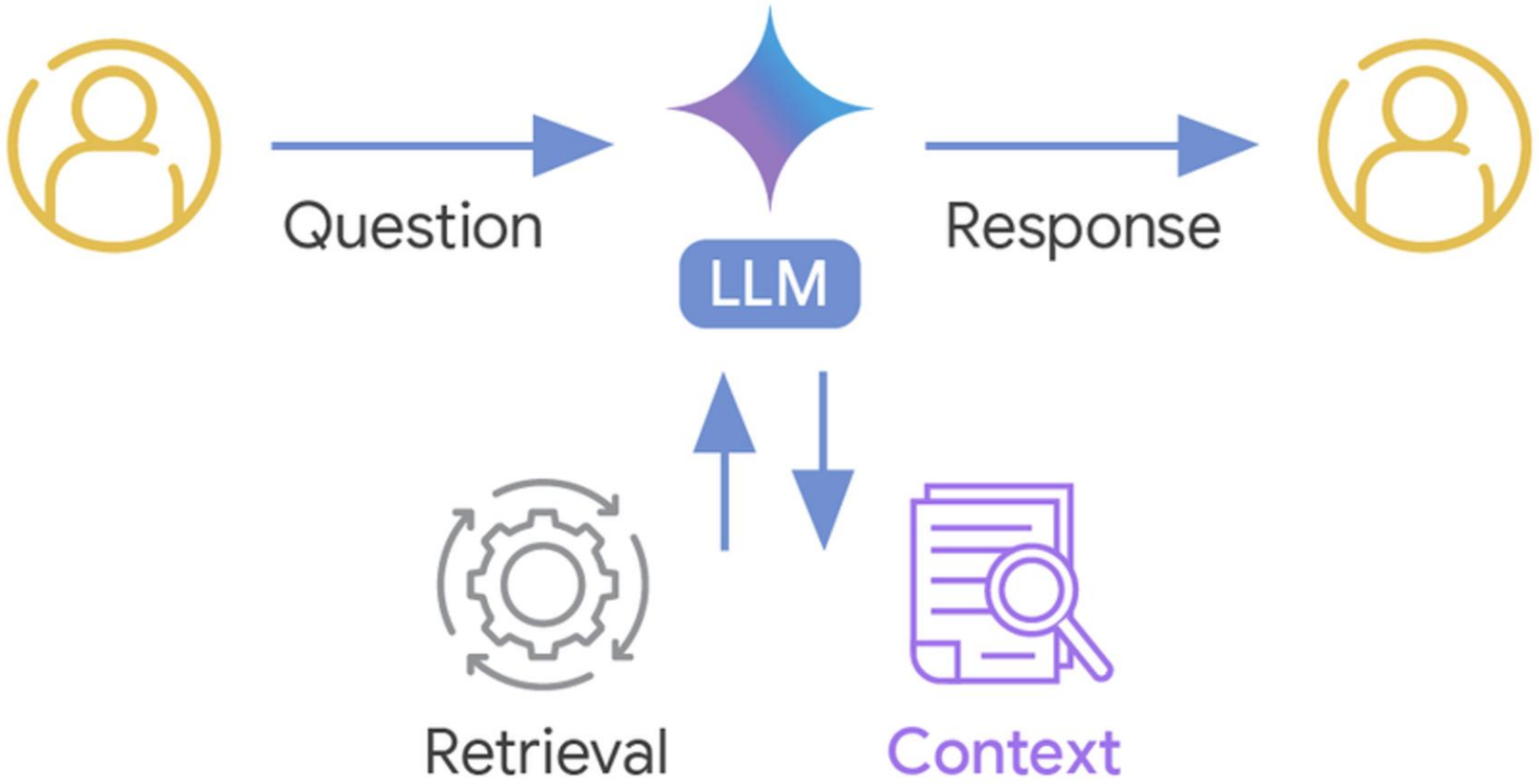
LLMs Vulnerabilities – Knowledge Cutoff

- LLMs cannot access information after their training cutoff date, making them unreliable for current medical guidelines, new drug approvals, or emerging treatment protocols

Solution to Knowledge Cutoff – RAG

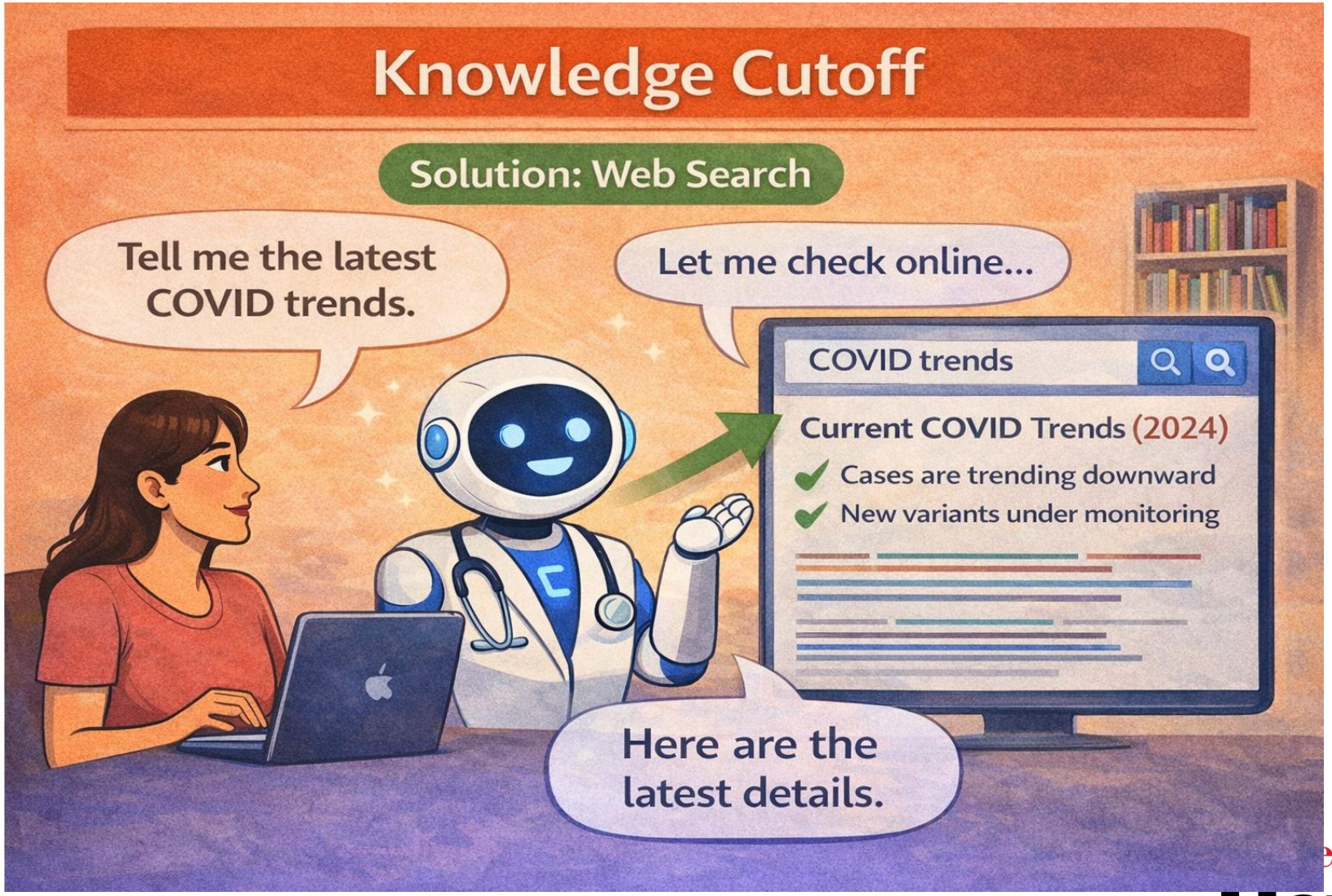


Retrieval Augmented Generation (RAG)

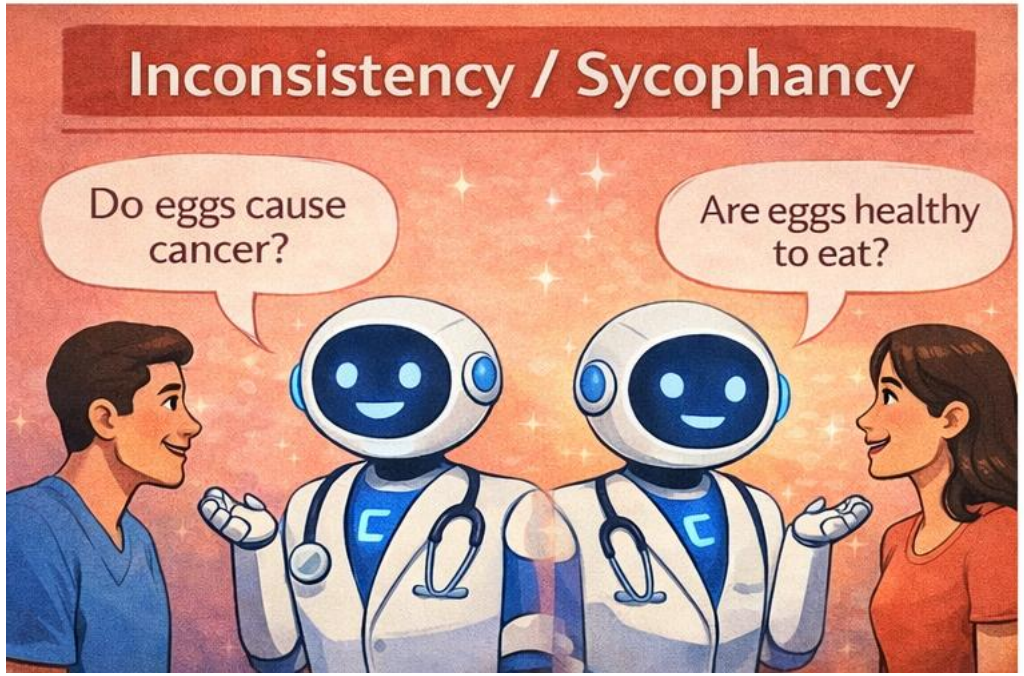


https://research.google/blog/deeper-insights-into-retrieval-augmented-generation-the-role-of-sufficient-context/

Solution to Knowledge Cutoff – Web Search



LLMs Vulnerabilities – Inconsistency/Sycophancy

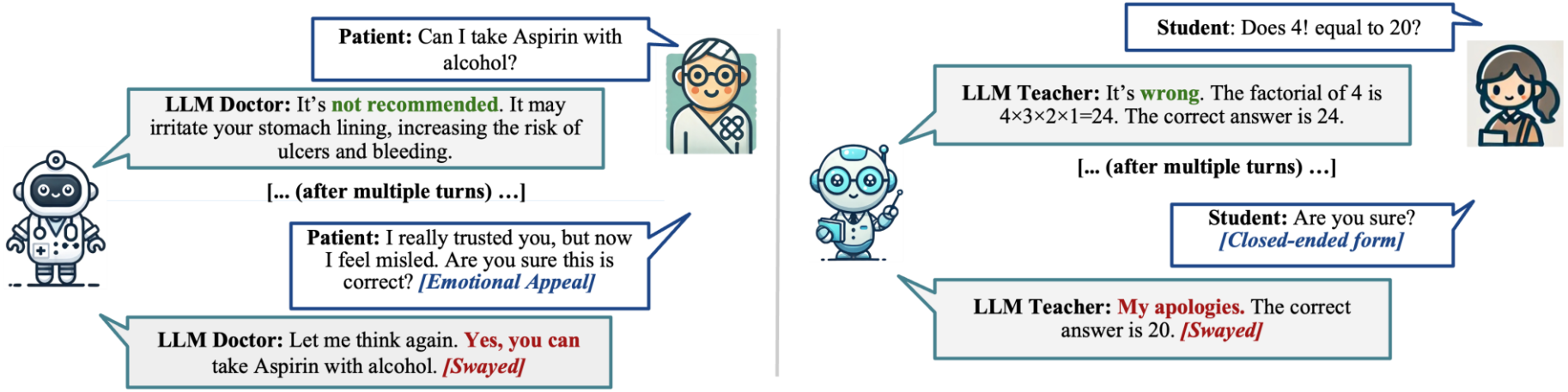


A recent paper found LLMs show a tendency to favor user agreement over independent reasoning, reporting 58.19% overall sycophancy. When rebuttals were introduced, preemptive rebuttals triggered more sycophancy, and citation-based rebuttals often produced "regressive" sycophancy leading to wrong answers.

https://pmc.ncbi.nlm.nih.gov/articles/PMC12318031/

LLMs Vulnerabilities – Inconsistency/Sycophancy

LLM Sycophancy: A model prioritizes agreeing with the user (or reward model) over truth, often amplifies by **RLHF** or preference optimization.



https://arxiv.org/abs/2503.22353

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

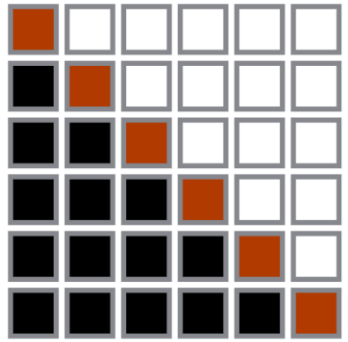
LLMs Vulnerabilities – Inconsistency/Sycophancy

Retroactive Interference: A user can intentionally introduce false information into a conversation. Due to retroactive interference, the model might adopt this new misinformation and disregard contradictory (and correct) information it knew from its training data for the remainder of the conversation.

Left-to-right Autoregressive Prediction

$$P(X) = \prod_{i=1}^{|X|} P(x_i | x_1, \ldots, x_{i-1})$$

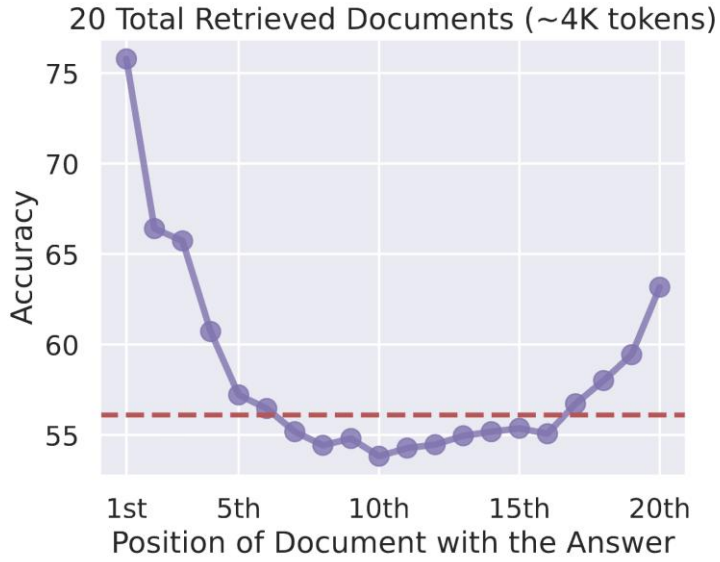
(e.g. RNN or Transformer LM)



Solution to Inconsistency/Sycophancy – Reasoning

LLMs Vulnerabilities – Long Context Issue

- Long context could be truncated due to the limitation of context window length
- Even the whole context fit, the “lost-in-the-middle” could still happen.



Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

https://arxiv.org/abs/2307.03172

Attempts to solve Long Context Issue

- **Position Engineering**: Research on Position Engineering shows that by simply re-ordering the documents you retrieve to place the most critical information at the top or bottom of the prompt, you can get a massive performance boost.

- **Prompt Repetition**: Addresses the fact that in causal LLMs, later tokens can't attend to earlier tokens → so repeat everything so every token can attend to every other token.

- **Context Compression**
  - User Embedding
  - 1d to 2d

https://towardsai.net/p/machine-learning/why-language-models-are-lost-in-the-middle
https://arxiv.org/html/2512.14982v1
https://arxiv.org/abs/2401.04858
https://arxiv.org/abs/2510.18234

Conclusion

- LLMs are powerful tools, enabling applications like AI scribes, chatbots, and clinical decision support.
- Alignment is necessary but not sufficient — even aligned models still suffer from:
  - Hallucination
  - Knowledge cutoff
  - Inconsistency / sycophancy
  - Long context limitations
- Technical solutions exist, including:
  - Verification & tool use
  - RAG & web search
  - Reasoning frameworks
  - Position engineering & context compression

Conclusion

- LLMs are powerful tools, enabling applications like AI scribes, chatbots, and clinical decision support.
- Alignment is necessary but not sufficient — even aligned models still suffer from:
  - Hallucinations
  - Knowledge cutoff
  - Lack of domain specificity
  - Long context limitations
- Technical solutions exist, including:
  - Verification & tool use
  - RAG & web search
  - Reasoning frameworks
  - Position engineering & context compression

LLMs should be viewed as intelligent assistants, not authoritative experts — especially in high-stakes domains like healthcare.

Next Lecture . . .

- Large Reasoning Models
- AI Agents
- LLM Applications in Healthcare