Intro to Generative AI

Tanner Dean

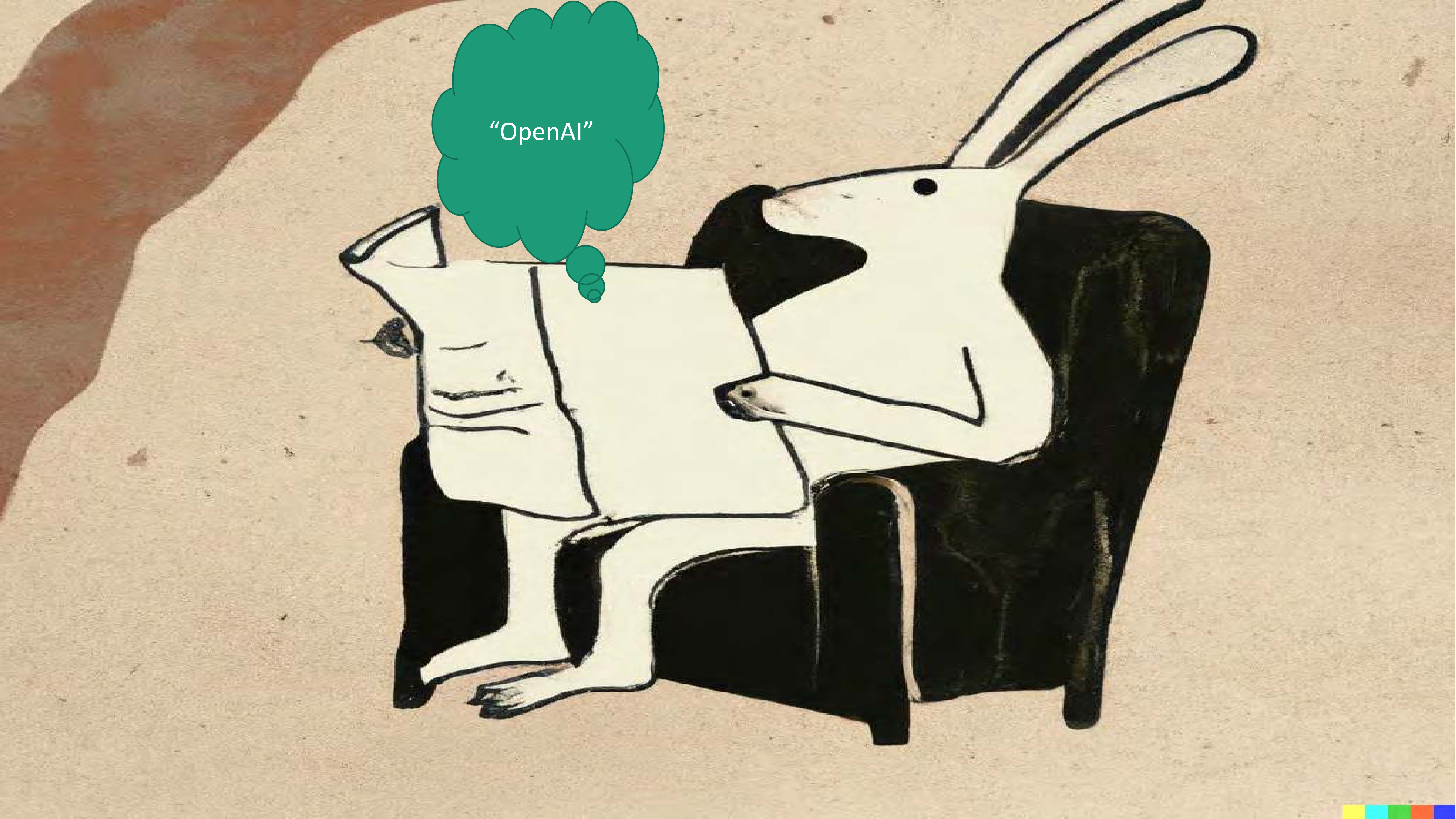
Tanner Dean, D.O.

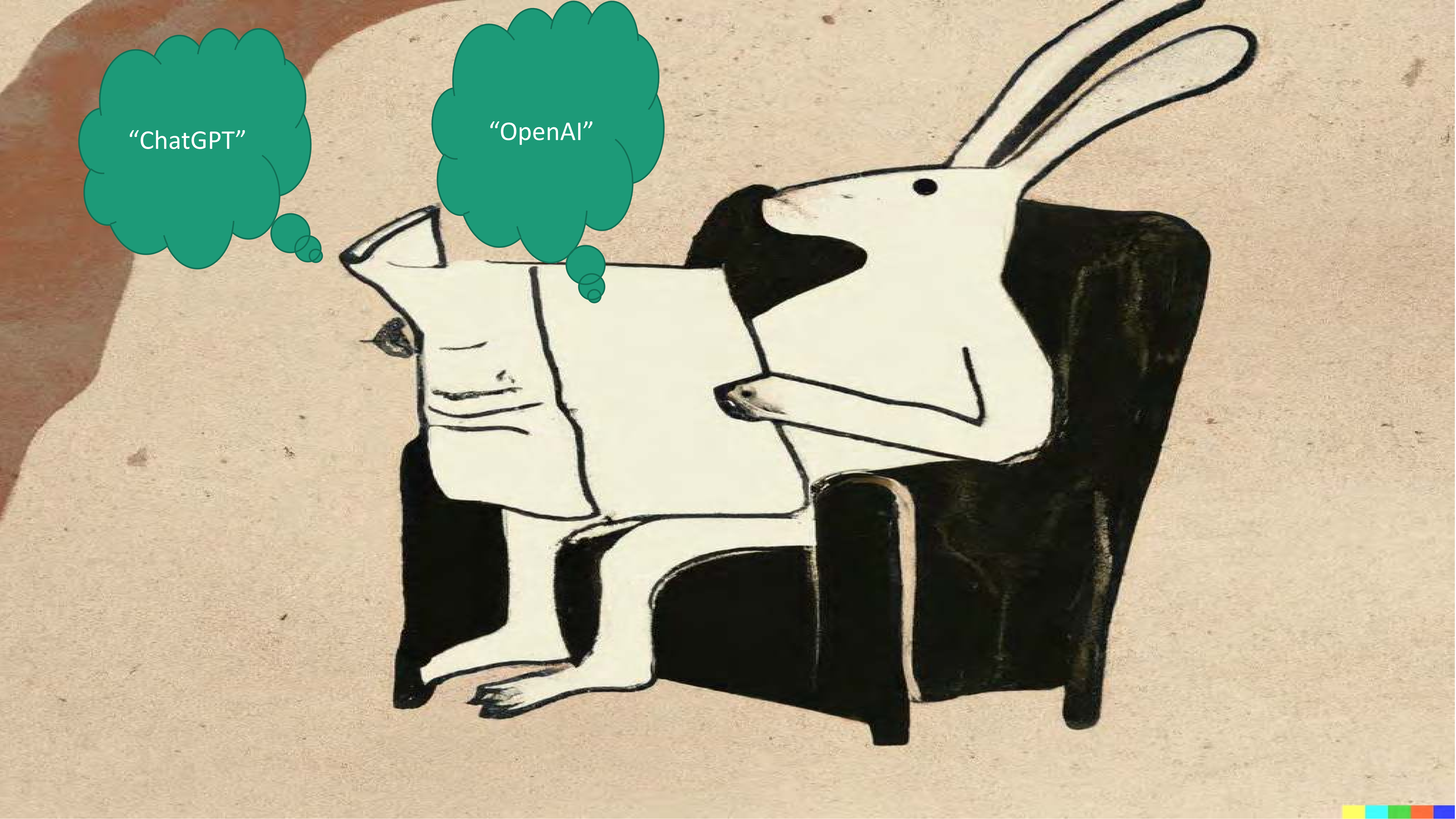
- Assistant Professor of Internal Medicine
- University of Kansas School of Medicine – Wichita
- No conflicts of interest

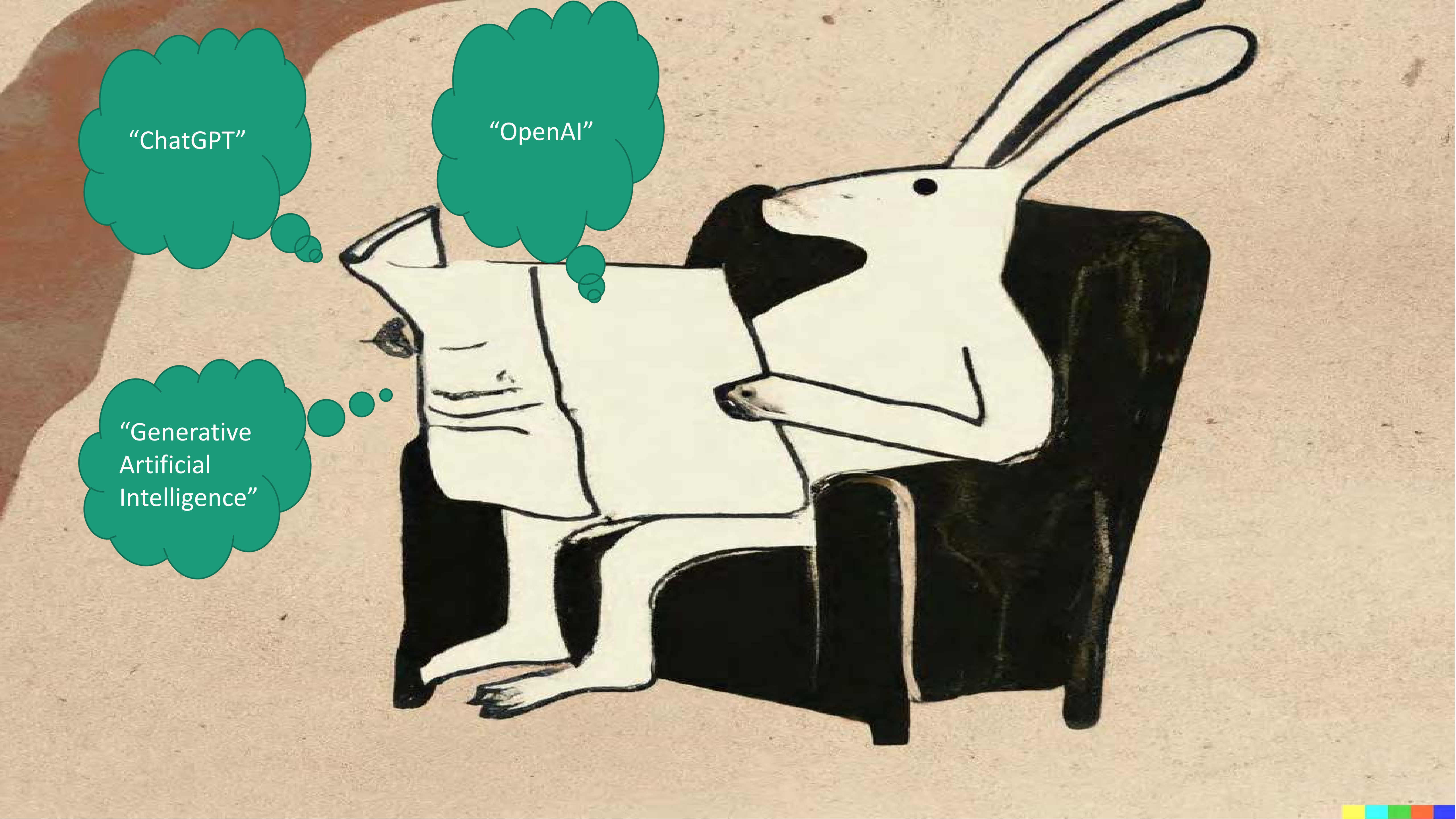
KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

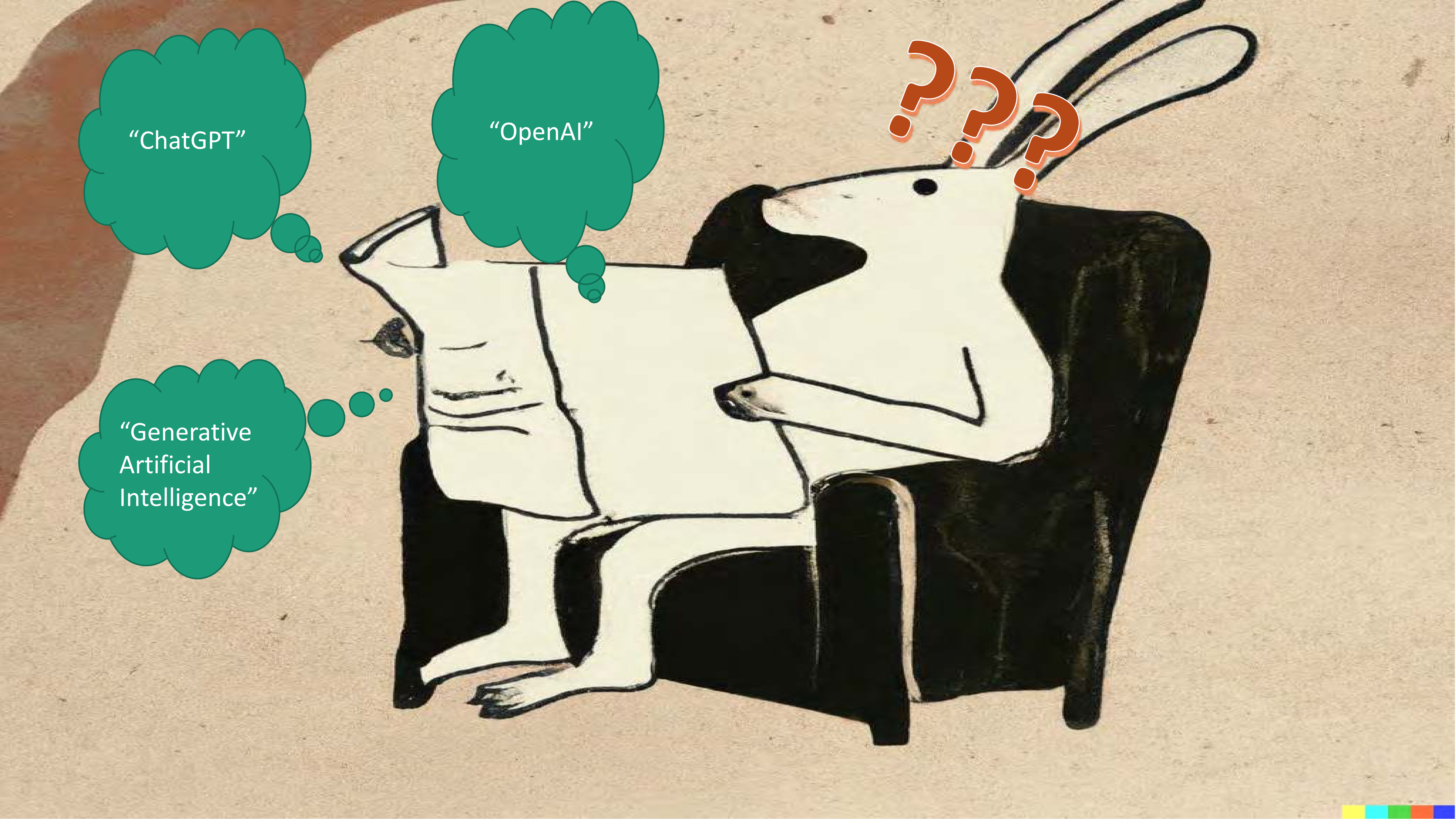
Common Among Us

- Diabetes Education and Counseling
- Very basically:
    - Teach correct principles, answer questions, inspire change


"OpenAI"


"ChatGPT"
"OpenAI"


"ChatGPT"
"OpenAI"
"Generative Artificial Intelligence"


"ChatGPT"
"OpenAI"
"Generative Artificial Intelligence"
???

Generative AI

- A form of artificial intelligence that generates novel text, videos, sound, etc based off instructions called prompts.

Generative AI

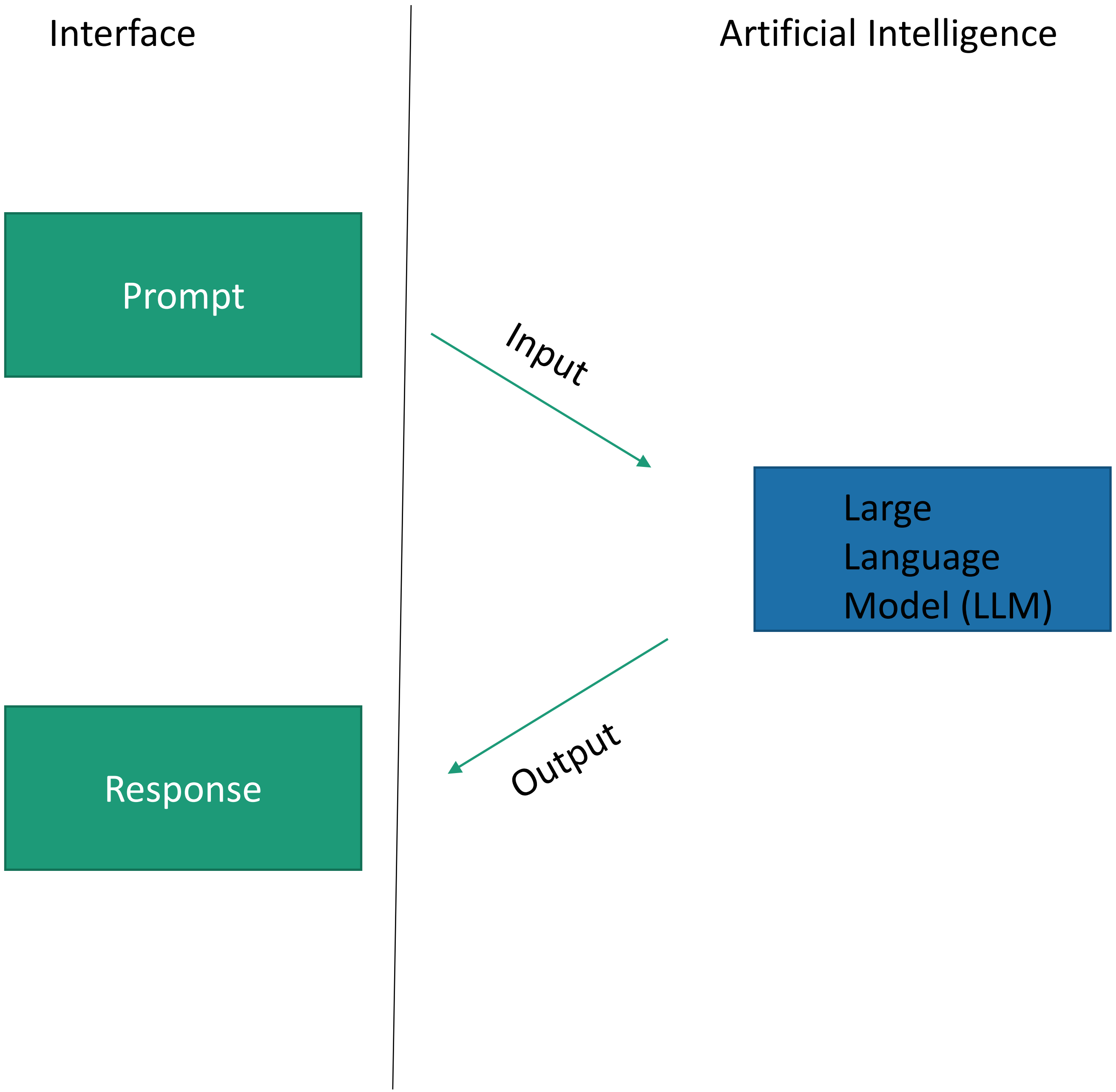
- A form of artificial intelligence that generates novel text, videos, sound, etc based off instructions called prompts.

Generative AI

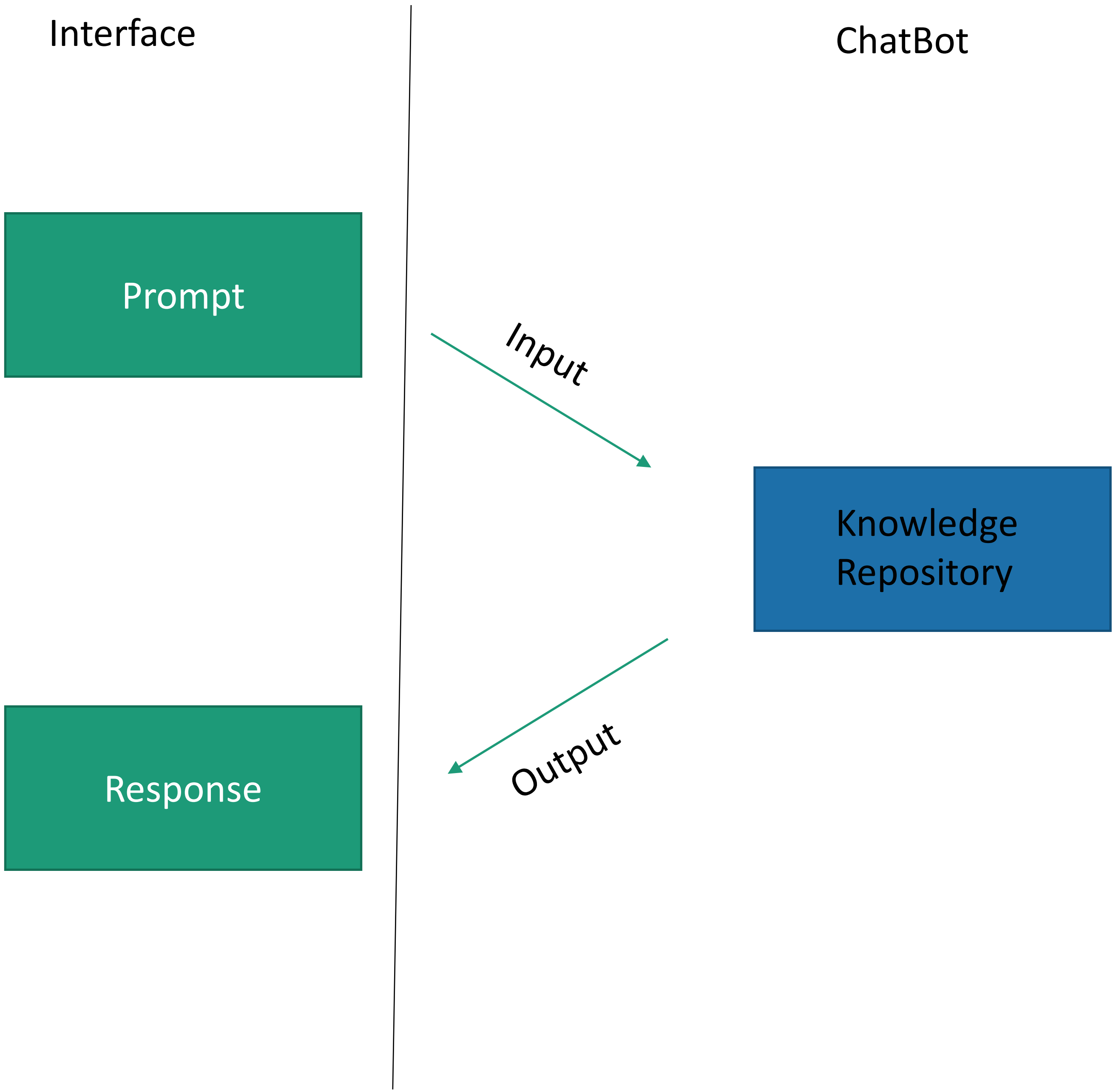
- A form of artificial intelligence that generates novel text, videos, sound, etc based off instructions called prompts.

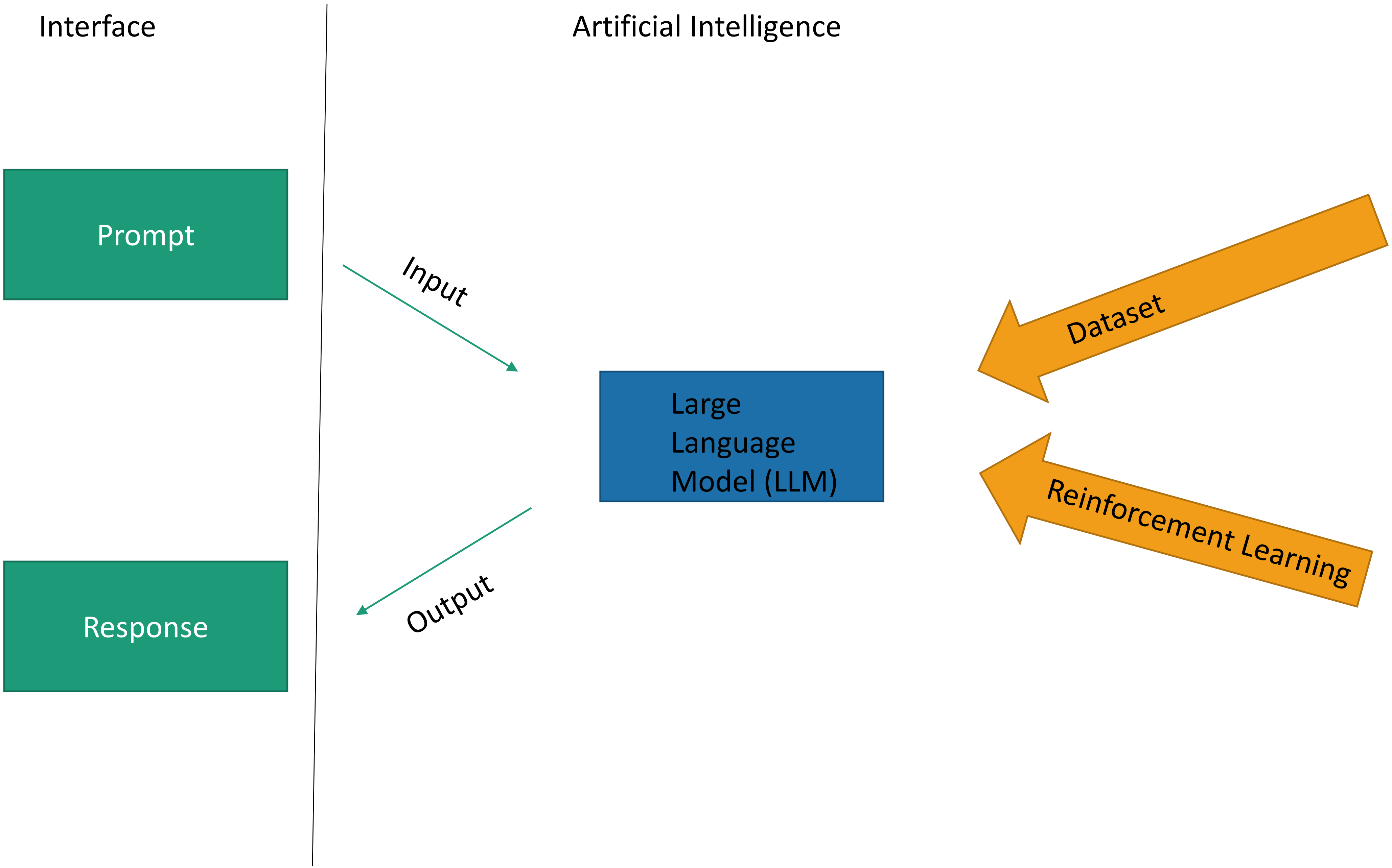
What is OpenAI and ChatGPT?

- OpenAI is the company who developed the ChatGPT interface
- ChatGPT = Answer-Response Interface for GPT(4)



Adapted from: *Introduction to chatgpt: Learn to use CHATGPT course*. DataCamp. (n.d.). Retrieved April 17, 2023, from https://www.datacamp.com/courses/introduction-to-chatgpt


Interface
ChatBot
Prompt
Input
Knowledge Repository
Output
Response


Interface
Artificial Intelligence
Prompt
Input
Large
Language
Model (LLM)
Dataset
Reinforcement Learning
Response
Output

Large Language Model – Simplified

Training

- Dataset = "Word Wall"
- Algorithms detect patterns within the "Word Wall"
- Fine-Tuned with Reinforcement Learning from Human Feedback (RLHF)

Calcium	**Is**	**The**	**Most**	**Abundant**
Cation	Found	In	The	Human
Body	And	Plays	An	Integral
Role	In	Neural	Transmission,	enzyme

Adapted from: *Introduction to chatgpt: Learn to use CHATGPT course*. DataCamp. (n.d.). Retrieved April 17, 2023, from https://www.datacamp.com/courses/introduction-to-chatgpt

Large Language Model – Simplified

Training

- Dataset = "Word Wall"
- Algorithms detect patterns within the "Word Wall"
- Fine-Tuned with Reinforcement Learning from Human Feedback (RLHF)

Calcium	is	the	most	abundant
cation	found	in	the	human
body	and	plays	an	integral
role	in	neural	transmission,	enzyme

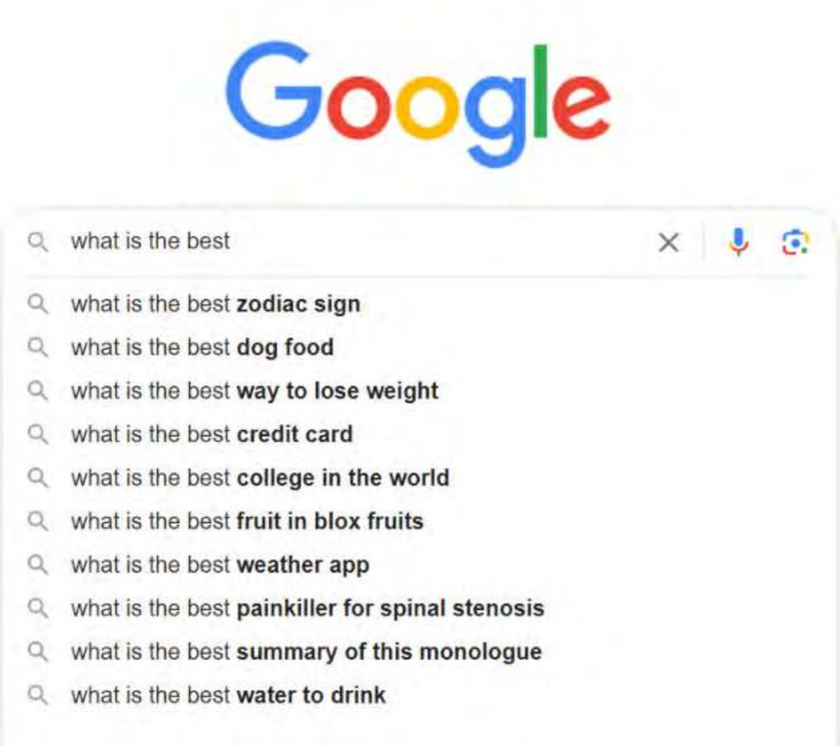
What makes the Algorithm and Detection so Special?

Transformer + Attention Mechanism

Let’s compare: “Old-School” Natural Language Processing vs. Transformer

Algorithm and Detection

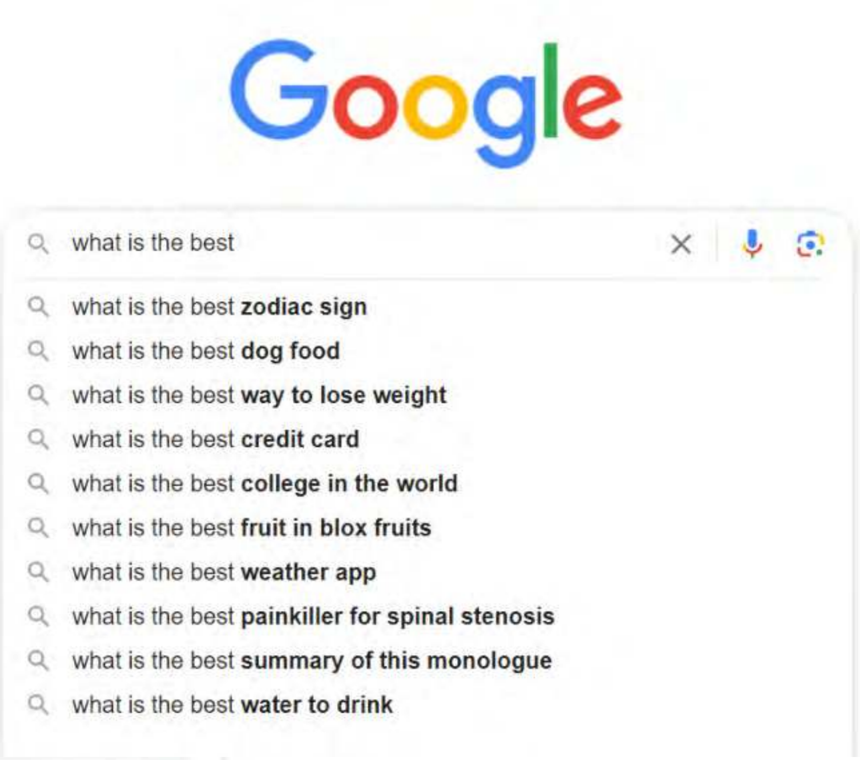
Old School NLP



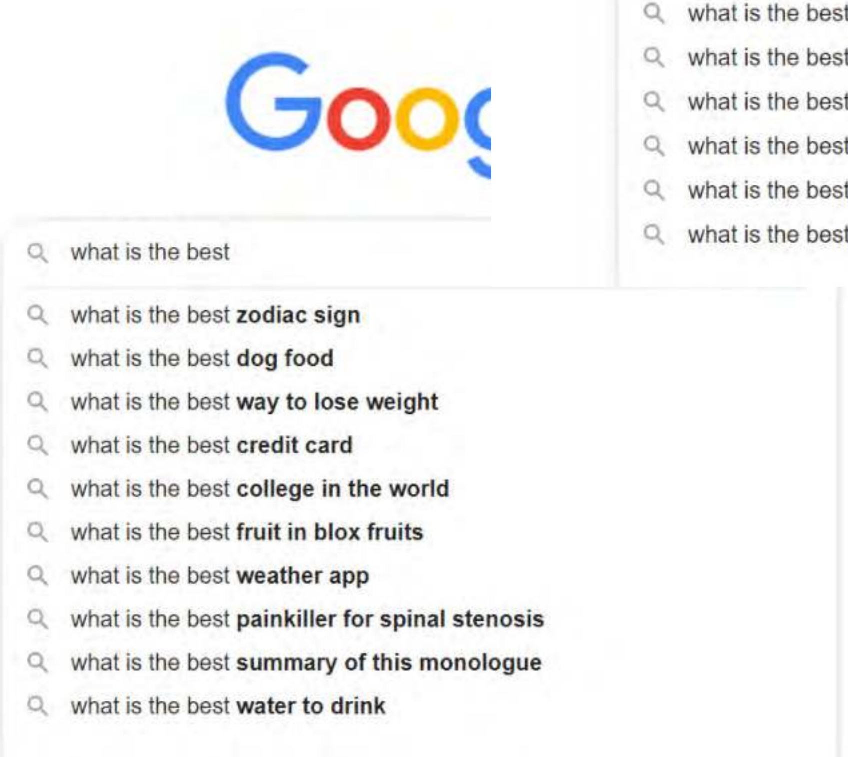
Transformer + Attention Mechanism

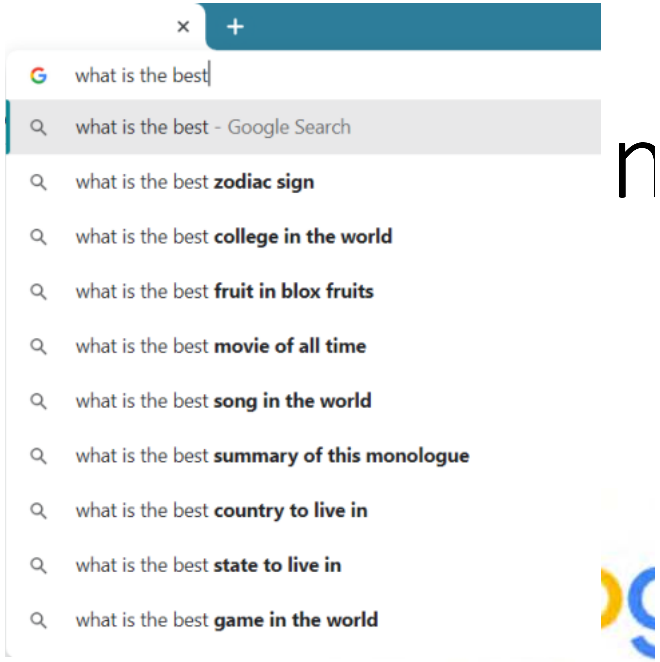
Algorithm an

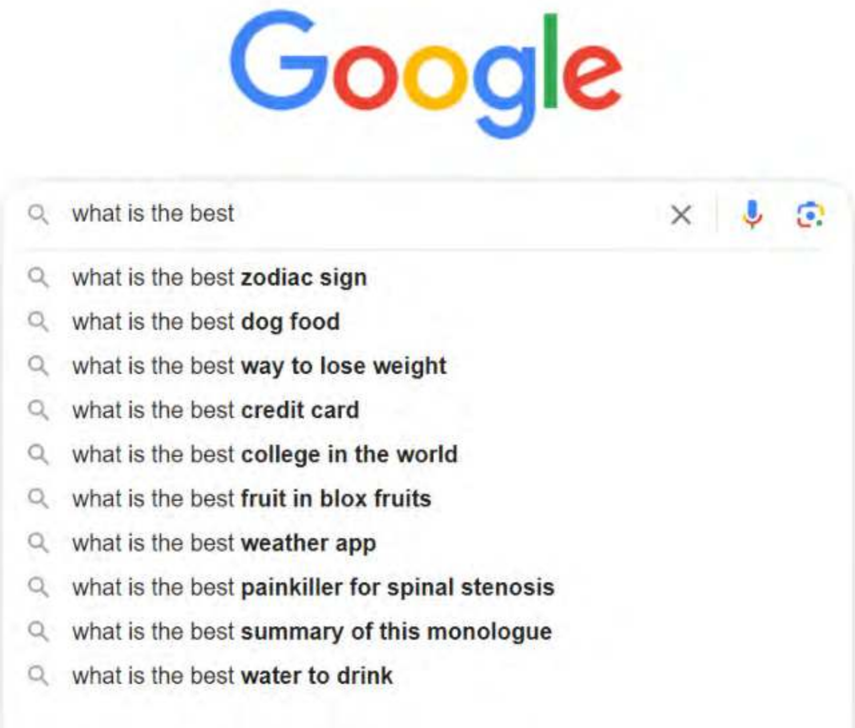
Old School NLP



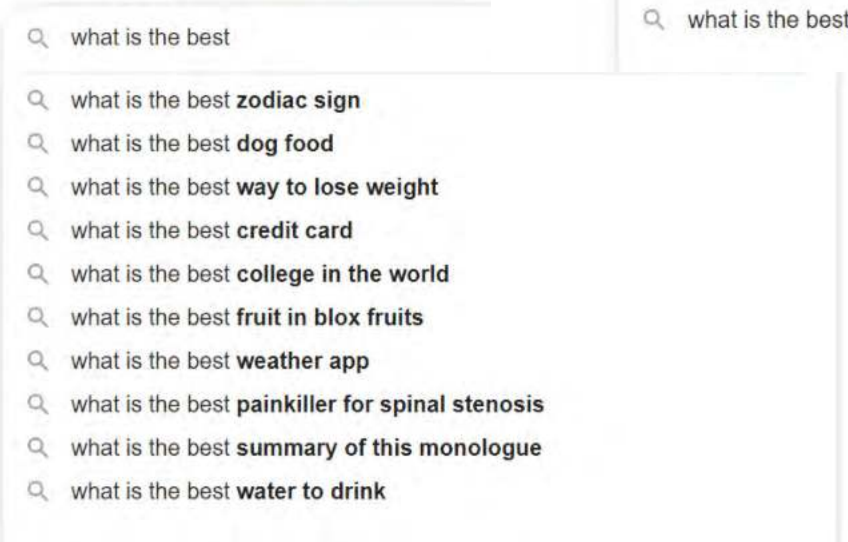
on Mechanism

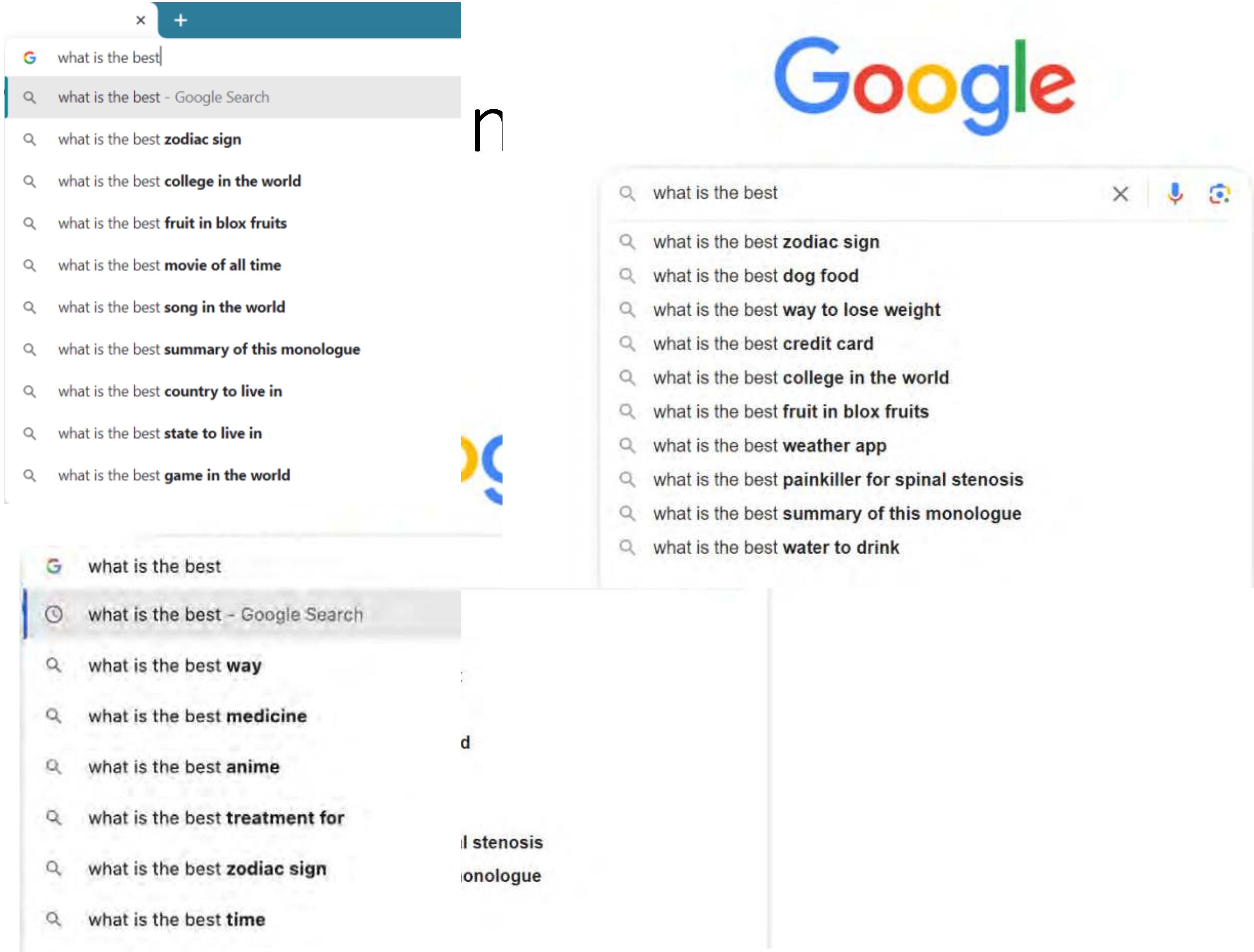




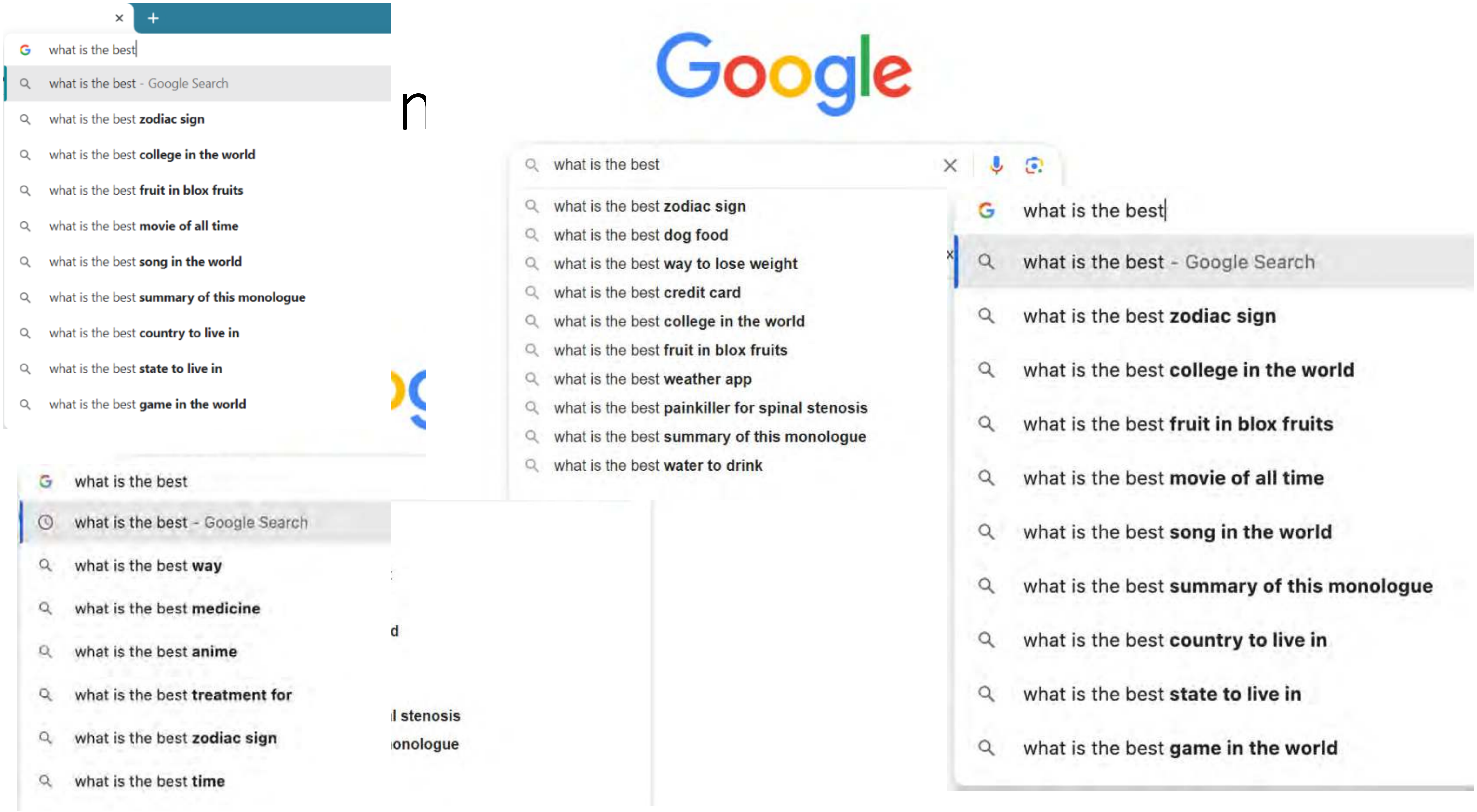


on Mechanism



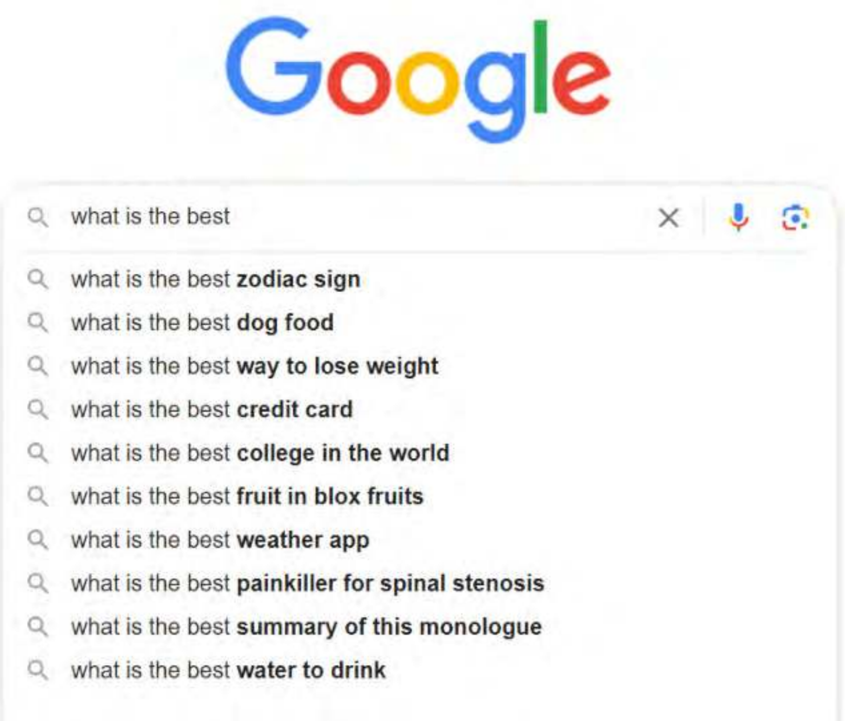


on Mechanism


what is the best
what is the best - Google Search
what is the best zodiac sign
what is the best college in the world
what is the best fruit in blox fruits
what is the best movie of all time
what is the best song in the world
what is the best summary of this monologue
what is the best country to live in
what is the best state to live in
what is the best game in the world
Google
what is the best
what is the best zodiac sign
what is the best dog food
what is the best way to lose weight
what is the best credit card
what is the best college in the world
what is the best fruit in blox fruits
what is the best weather app
what is the best painkiller for spinal stenosis
what is the best summary of this monologue
what is the best water to drink
what is the best
what is the best - Google Search
what is the best zodiac sign
what is the best college in the world
what is the best fruit in blox fruits
what is the best movie of all time
what is the best song in the world
what is the best summary of this monologue
what is the best country to live in
what is the best state to live in
what is the best game in the world
what is the best
what is the best - Google Search
what is the best way
what is the best medicine
what is the best anime
what is the best treatment for
what is the best zodiac sign
what is the best time

Algorithm and Detection

Old School NLP



Transformer + Attention Mechanism

Calcium	is	the	most	abundant
cation	found	in	the	human
body	and	plays	an	integral
role	in	neural	transmission,	enzyme

Self-attention or multi-head attention

Large Language Model – Simplified

Training

- Dataset = “Word Wall”
- Algorithms detect patterns within the “Word Wall”
- Fine-Tuned with Reinforcement Learning from Human Feedback (RLHF)

Calcium	Is	The	Most	Abundant
Cation	Found	In	The	Human
Body	And	Plays	An	Integral
Role	In	Neural	Transmission,	enzyme

Reinforcement Learning with Human Feedback Conceptualized

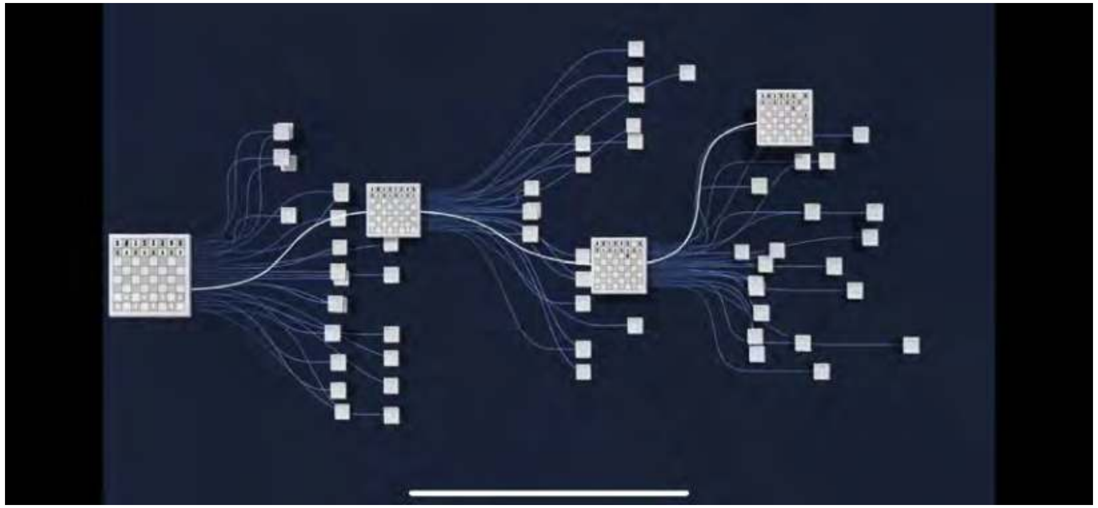
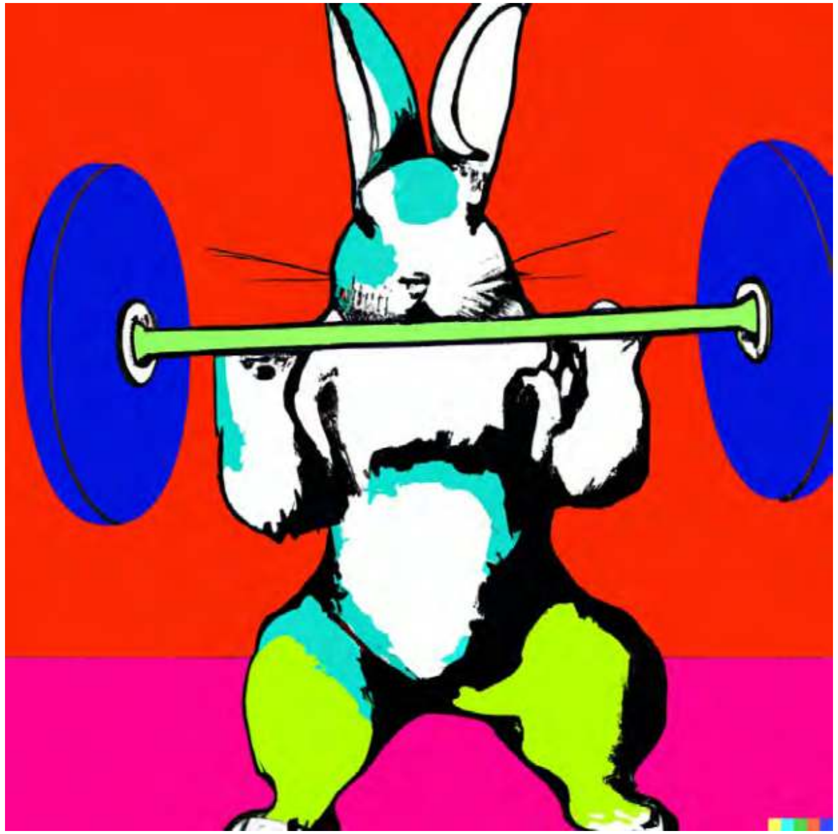
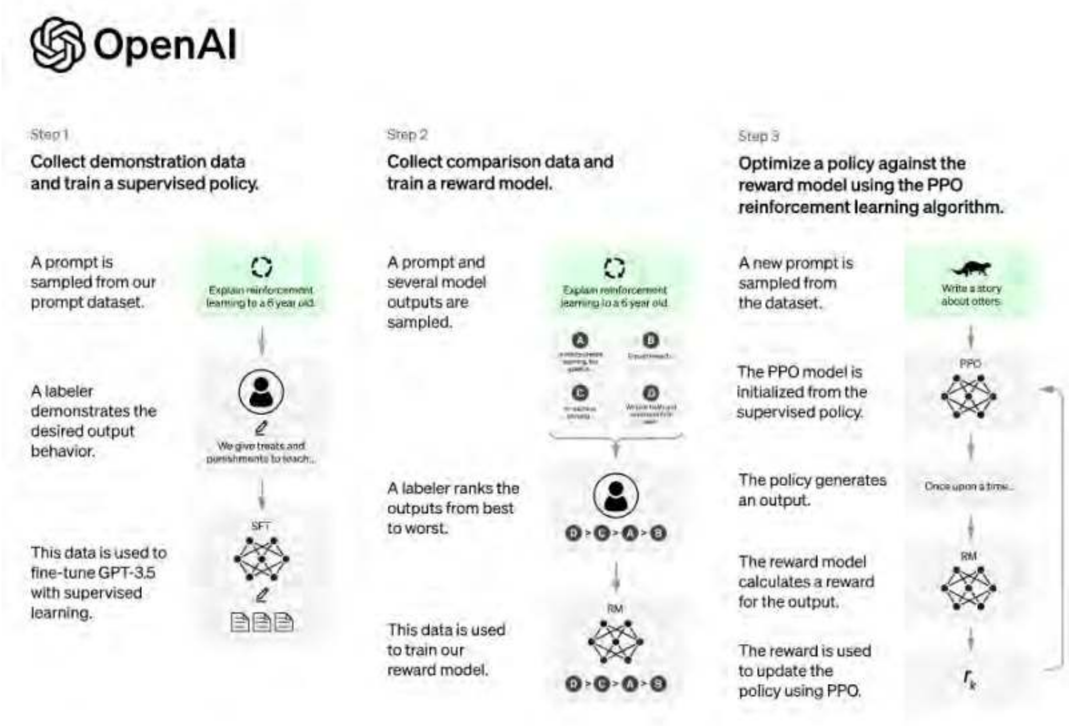


Image from: Kohs, G. (Director). (2020). *AlphaGo - The Movie* [Film]. [AlphaGo - The Movie | Full award-winning documentary - YouTube]()

The Real Training





Take a look

ChatGPT (GPT4) – What can it do?

Strengths :

- Summarize text
- Explain concepts
- Remembering conversation history
- Brainstorming / creating ideas
- Code/debug code

Limitations:

- Knowledge limit
  - GPT4 = early 2022
- Biases in dataset
- Loss of context tracking
- Hallucination
- Legal and Ethical uncertainties
- Cannot ask for clarifying data (LLM to user)
- For more see: Introducing ChatGPT (openai.com)

ChatGPT (GPT4) – What can it do?

Strengths :

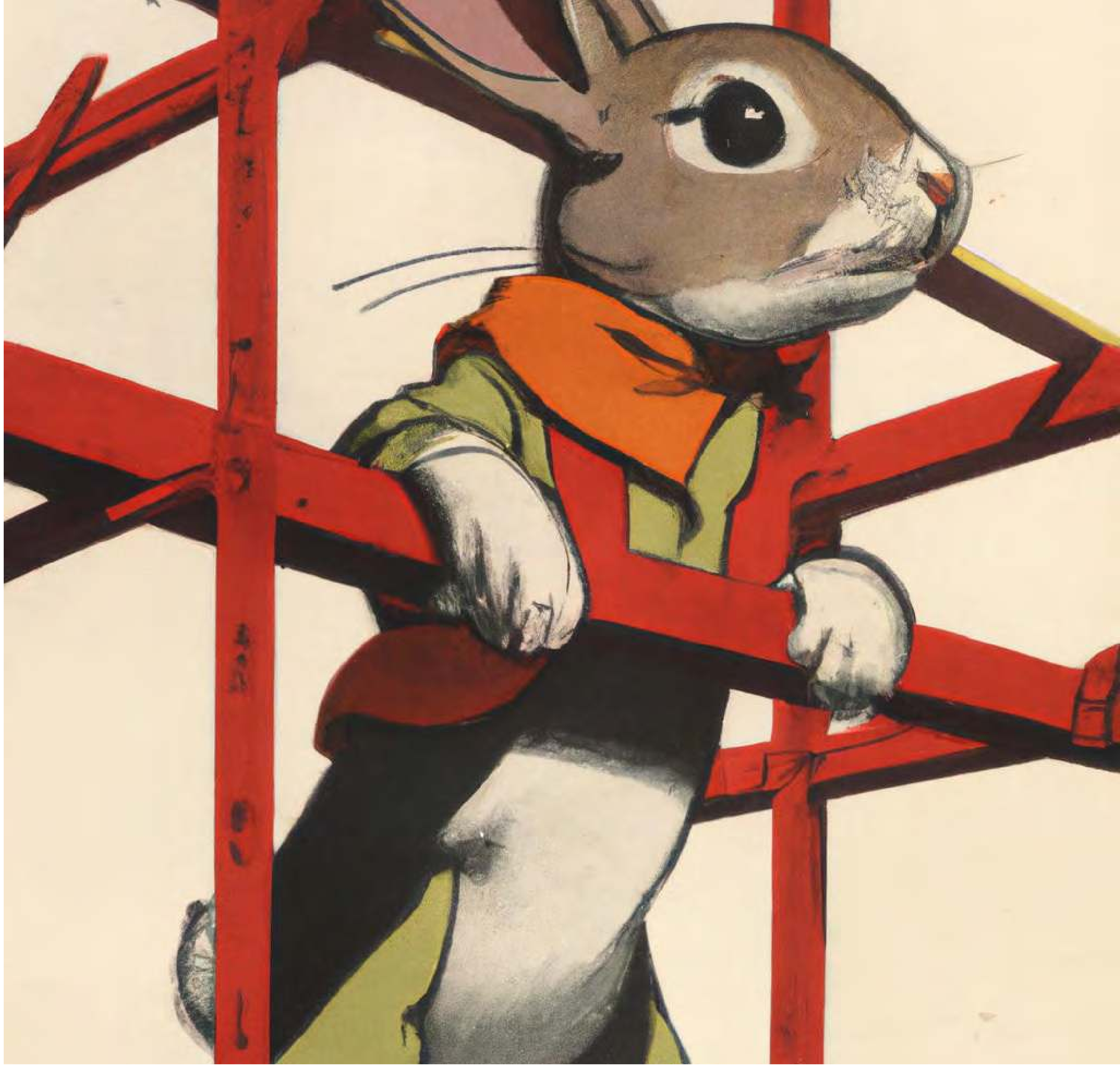
- Summarize text
- Explain concepts
- Remembering conversation history
- Brainstorming / creating ideas
- Code/debug code

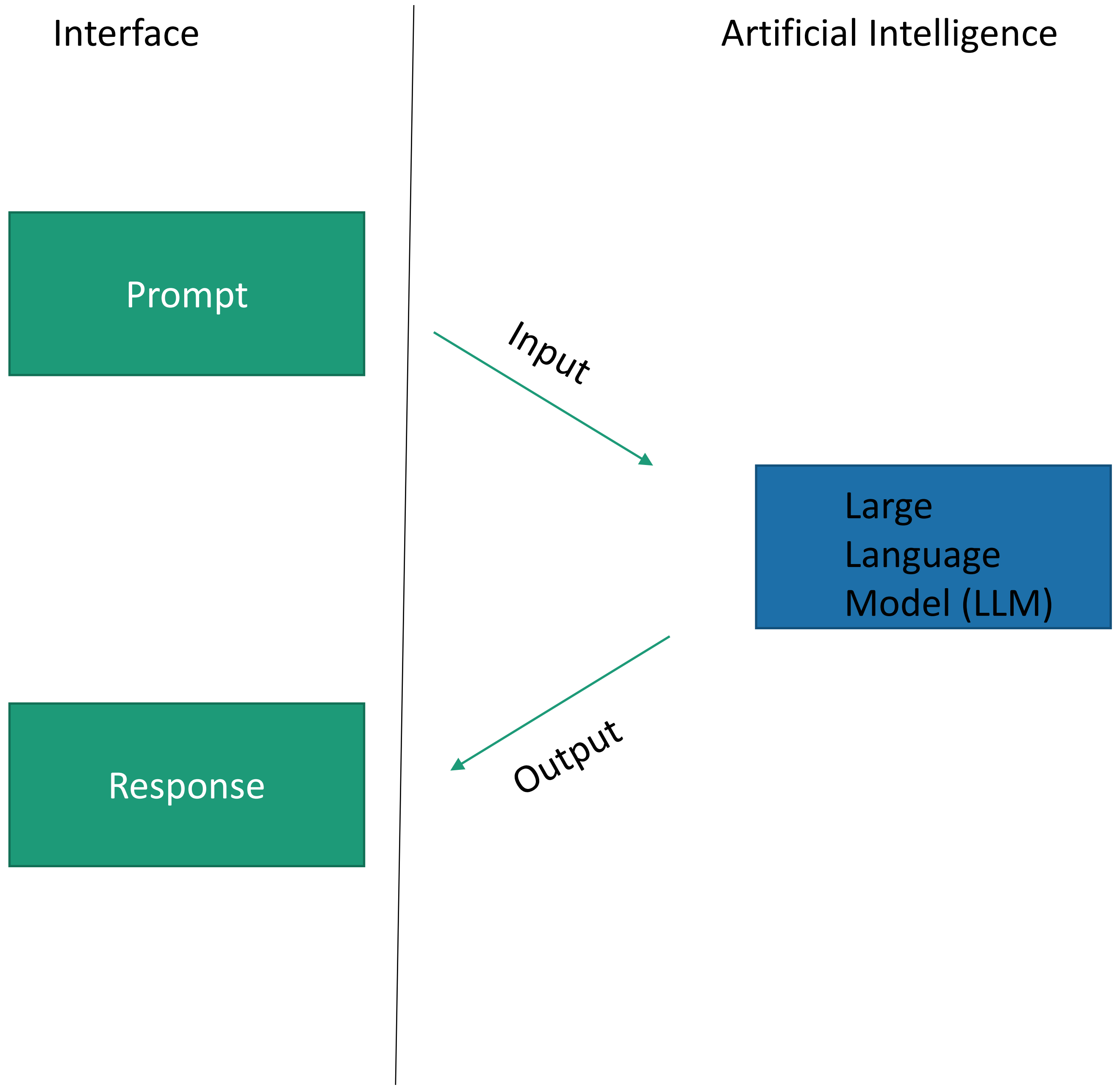
Limitations:

- Knowledge limit
  - GPT4 = early 2022
- Biases in dataset
- Loss of context tracking
- Hallucination
- Legal and Ethical uncertainties
- Cannot ask for clarifying data (LLM to user)
- For more see: Introducing ChatGPT (openai.com)

Prompt Engineering

Thoughtful responses are based on thoughtful prompts




Interface
Artificial Intelligence
Prompt
Input
Large
Language
Model (LLM)
Output
Response

Prompt Engineering – How ChatGPT interprets prompts

Step 1: What are the broad topics of the prompt?

Step 2: What is being requested?

Step 3: Generate a response

Best practices prompts

Essential practices

- Give all the needed context
- Clear and specific
- Keep it concise (extra words distract it)
- Use correct grammar and spelling

Super-power prompts practices

- Give examples
- Give it an identity
- Set restrictions to responses
- Only ask for something you could do, given enough time

Best practices prompts

Essential practices

- Give all the needed context
- Clear and specific
- Keep it concise (extra words distract it)
- Use correct grammar and spelling

Super-power prompts practices

- Give examples
- Give it an identity
- Set restrictions to responses
- Only ask for something you could do given enough time
- Have it generate prompts for other prompts

Technical Prompting Techniques

Zero Shot Prompting

- Machine Learning Principle
- Relies on the models baseline knowledge without providing any new training
- "When is Christmas celebrated in America?"
  - This is using the base training of the model

Few Shot Prompting

- Machine Learning Principle
- Based off additional training data provided to the model
- "What is my favorite cooking spice?"
  - Provide several examples of foods I think taste good

Carnes, Beau. Kubow, Ania. "Learn Prompt Engineering – Full Course." *freeCodeCamp.Org*, freeCodeCamp.org, www.freecodecamp.org/news/learn-prompt-engineering-full-course/.

Prompting Techniques

Zero Shot Prompting

- Machine Learning Principle
- Relies on the models baseline knowledge without providing any new training
- “When is Christmas celebrated in America?”
    - This is using the base training of the model

Few Shot Prompting

- Machine Learning Principle
- Based off additional training data provided to the model
- “What is my favorite cooking spice?”
    - Provide several examples of foods I think taste good

Carnes, Beau. Kubow, Ania. “Learn Prompt Engineering – Full Course.” *freeCodeCamp.Org*, freeCodeCamp.org, www.freecodecamp.org/news/learn-prompt-engineering-full-course/.

Prompting Techniques

Zero Shot Prompting

- Machine Learning Principle
- Relies on the models baseline knowledge without providing any new training
- "When is Christmas celebrated in America?"
  - This is using the base training of the model

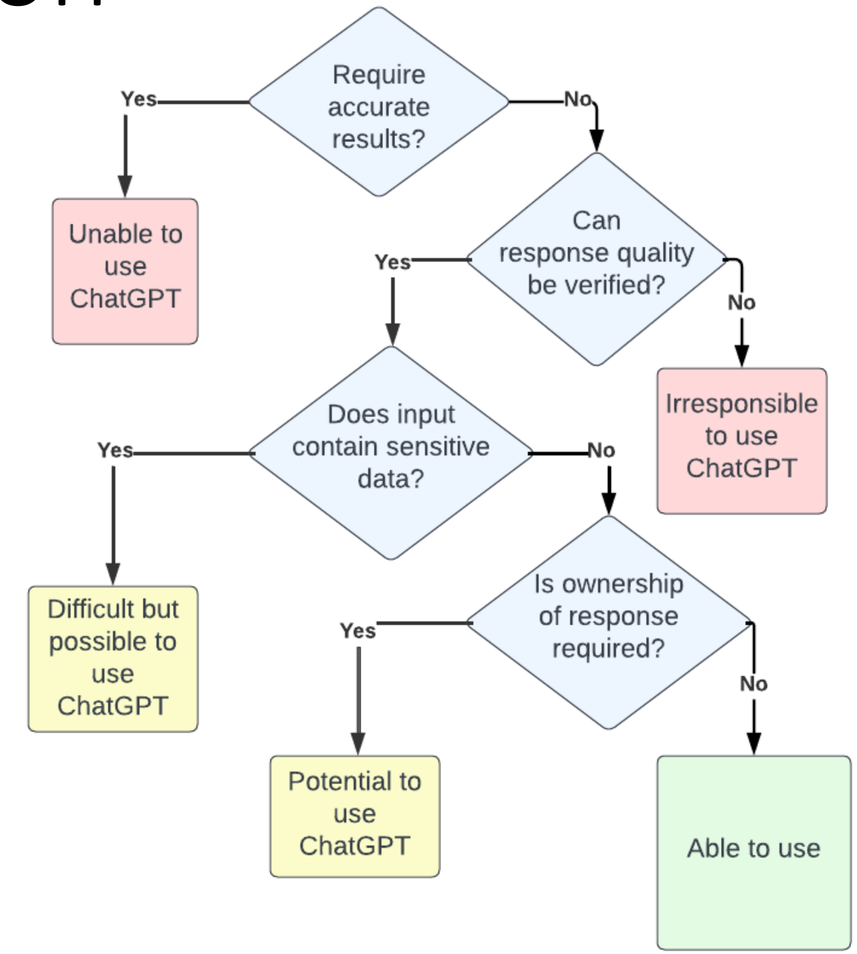
Few Shot Prompting

- Machine Learning Principle
- Based off additional training data provided to the model
- "What is my favorite cooking spice?"
  - Provide several examples of foods I think taste good

Carnes, Beau. Kubow, Ania. "Learn Prompt Engineering – Full Course." *freeCodeCamp.Org*, freeCodeCamp.org, www.freecodecamp.org/news/learn-prompt-engineering-full-course/.

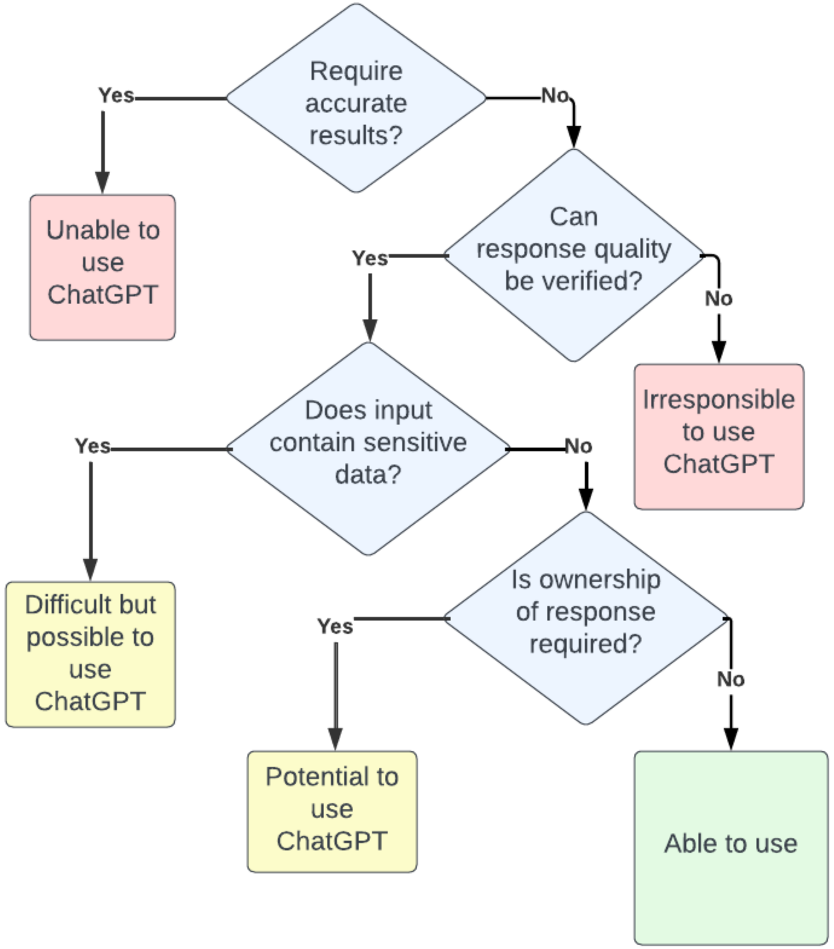
Implementation

Implementation



Adapted from: *Introduction to chatgpt: Learn to use CHATGPT course*. DataCamp. (n.d.). Retrieved April 17, 2023, from https://www.datacamp.com/courses/introduction-to-chatgpt

Understanding implementation can help you understand pitfall



Pitfalls

- “Hallucination”
  - Authoritative confabulation
- PHI
  - Without private mode, OpenAI can see everything
- Lack of subject matter expertise
  - Unable to verify validity of output
- No editing, using output as is
- Bias
  - Data set is trained on things as they are done, not how they should be done

Pitfalls

- “Hallucination”
  - Authoritative confabulation
- PHI
  - Without private mode, OpenAI can see everything
- Lack of subject matter expertise
  - Unable to verify validity of output
- No editing, using output as is
- Bias
  - Data set is trained on things as they are done, not how they should be done

Pitfalls

- “Hallucination”
  - Authoritative confabulation
- PHI
  - Without private mode, OpenAI can see everything
- Lack of subject matter expertise
  - Unable to verify validity of output
- No editing, using output as is
- Bias
  - Data set is trained on things as they are done, not how they should be done

Pitfalls

- "Hallucination"
  - Authoritative confabulation
- PHI
  - Without private mode, OpenAI can see everything
- Lack of subject matter expertise
  - Unable to verify validity of output
- No editing, using output as is
- Bias
  - Data set is trained on things as they are done, not how they should be done

Pitfalls

- "Hallucination"
  - Authoritative confabulation
- PHI
  - Without private mode, OpenAI can see everything
- Lack of subject matter expertise
  - Unable to verify validity of output
- No editing, using output as is
- Bias
  - Data set is trained on things as they are done, not how they should be done

Research Questions

-Can ChatGPT produce usable information for lab follow-up letters?

-Does “prompt structure” change the output?

1) Readable

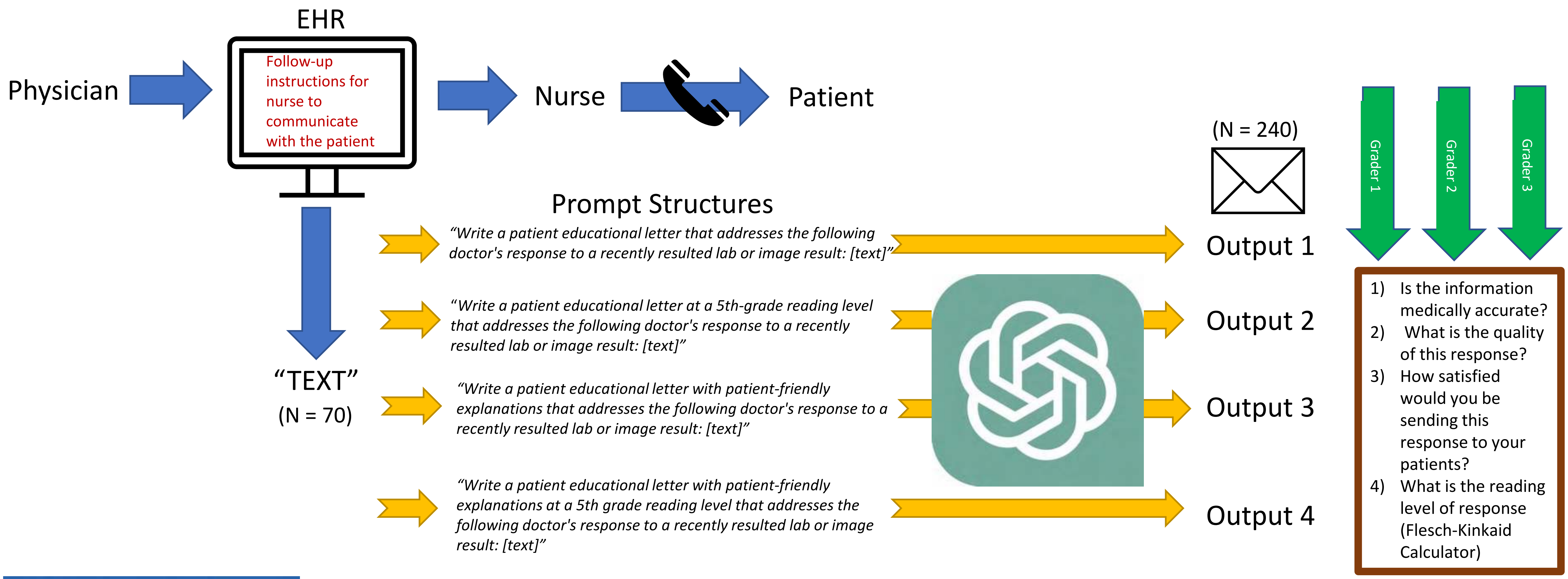
2) Medically accurate

3) Quality

3) Physician Satisfaction

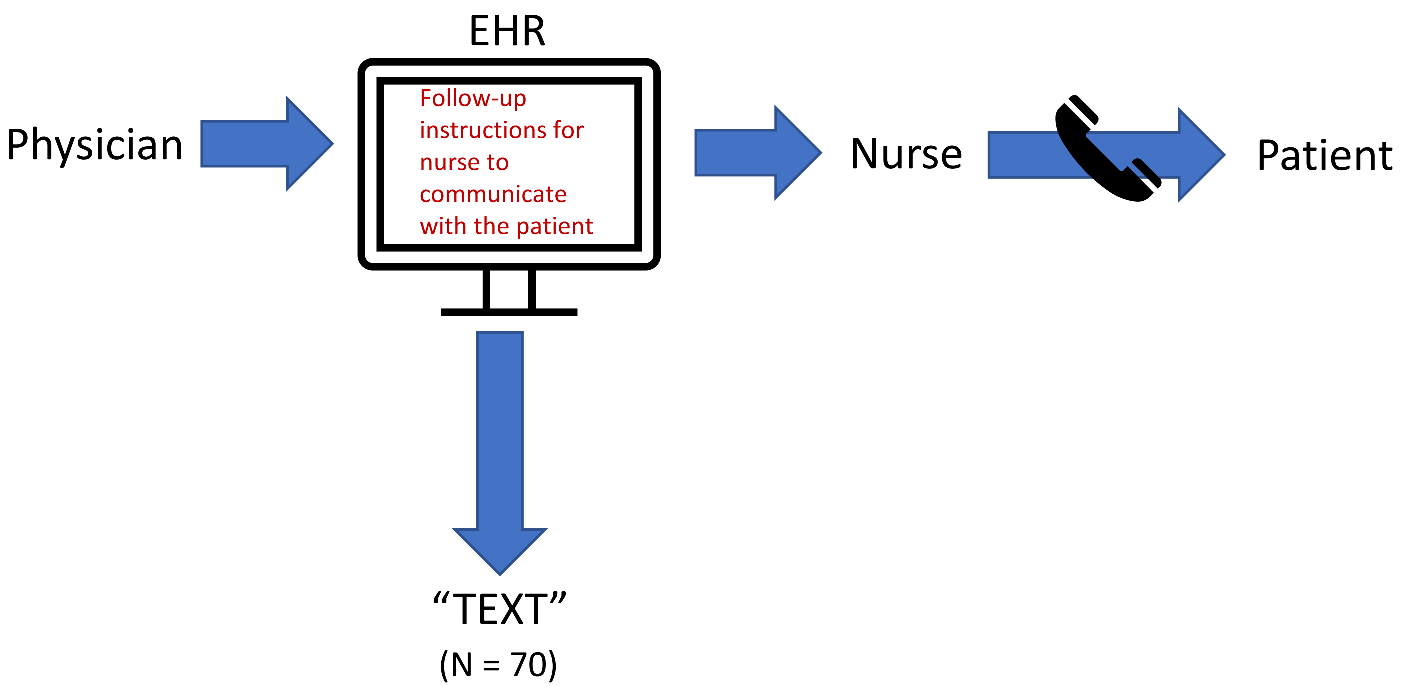
KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Methods

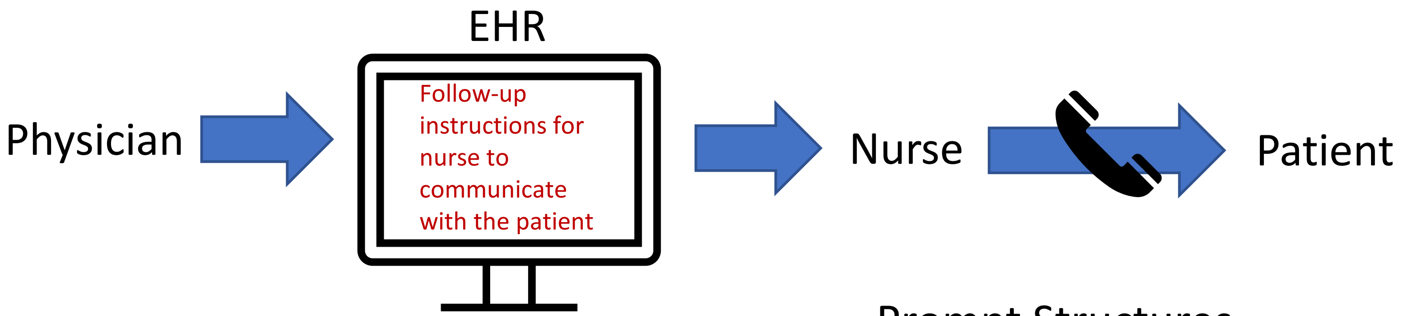


KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Methods


EHR
Physician
Follow-up instructions for nurse to communicate with the patient
Nurse
Patient
"TEXT"
(N = 70)

Methods



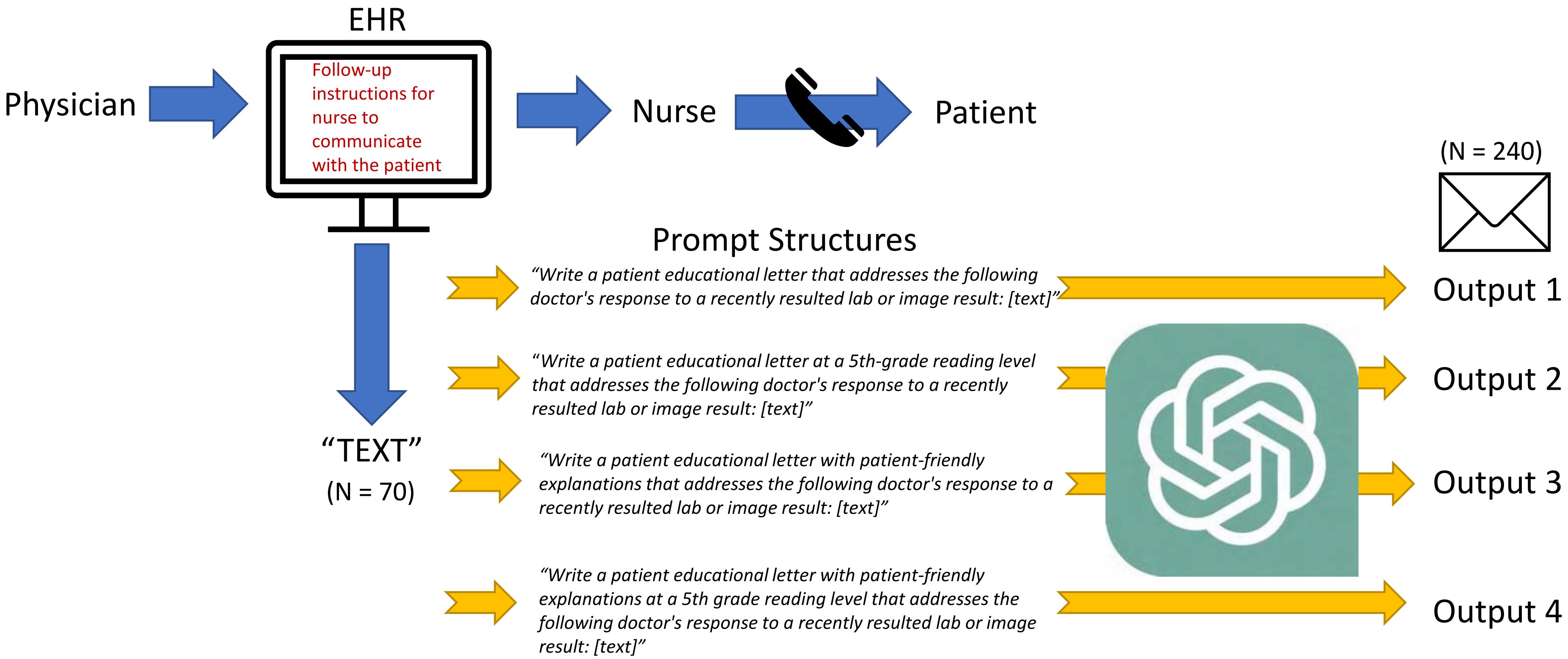
“TEXT”
(N = 70)

Prompt Structures

- *“Write a patient educational letter that addresses the following doctor's response to a recently resulted lab or image result: [text]”*
- *“Write a patient educational letter at a 5th-grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]”*
- *“Write a patient educational letter with patient-friendly explanations that addresses the following doctor's response to a recently resulted lab or image result: [text]”*
- *“Write a patient educational letter with patient-friendly explanations at a 5th grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]”*

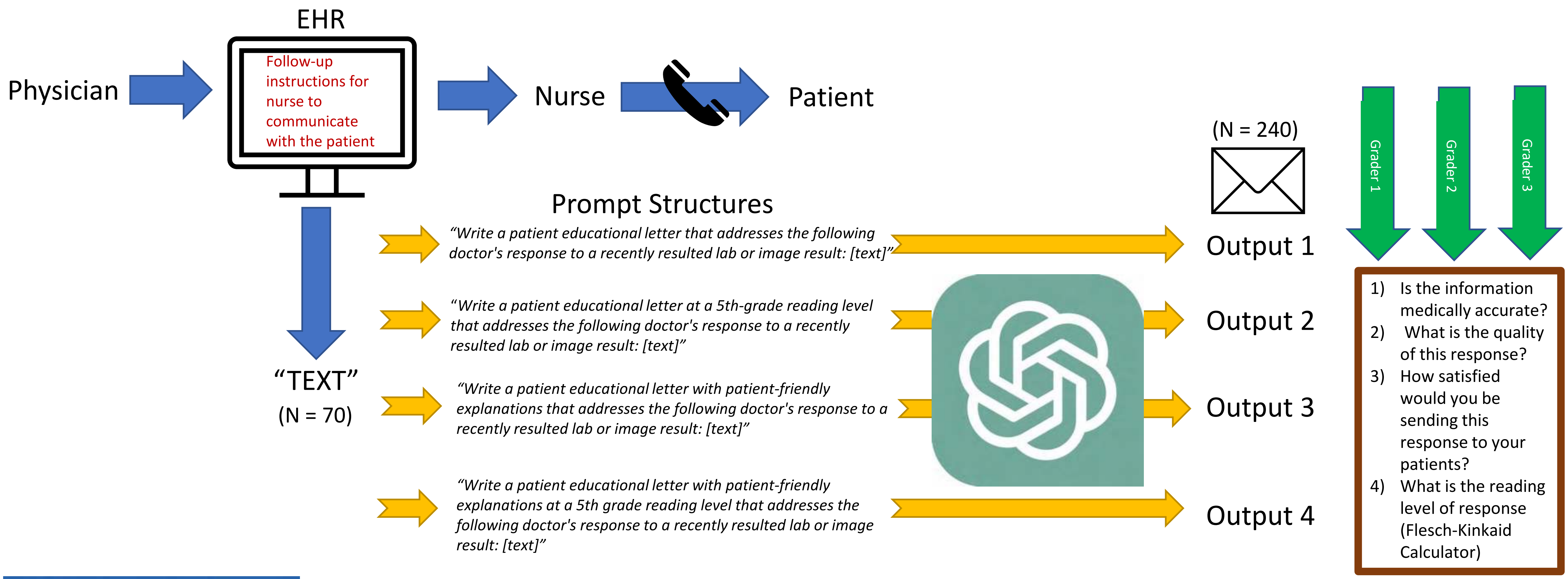
KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Methods



KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Methods



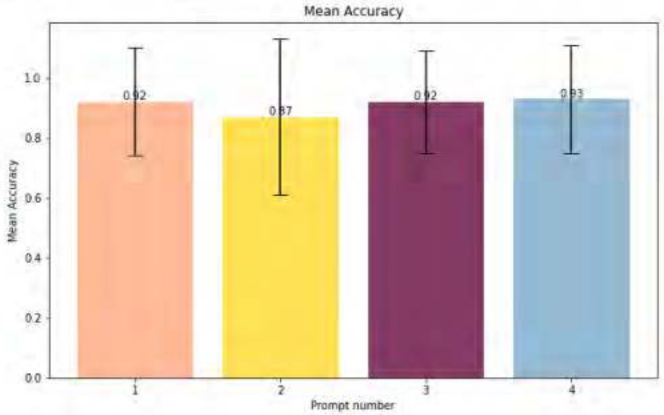
KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Results

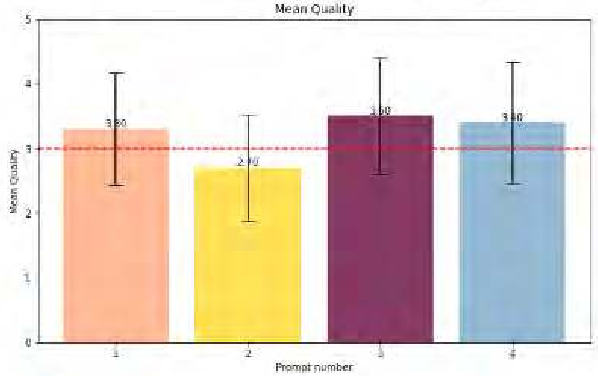
5-point Likert Scale

1 = Completely negative
2= Negative
3 = Neutral
4 = Positive
5= Completely positive

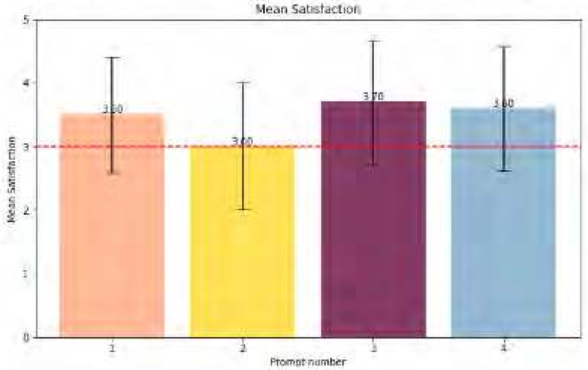
	Prompt	Mean Accuracy	Median Accuracy	STD
0	1	0.92	1	0.18
1	2	0.87	1	0.26
2	3	0.92	1	0.17
3	4	0.93	1	0.18



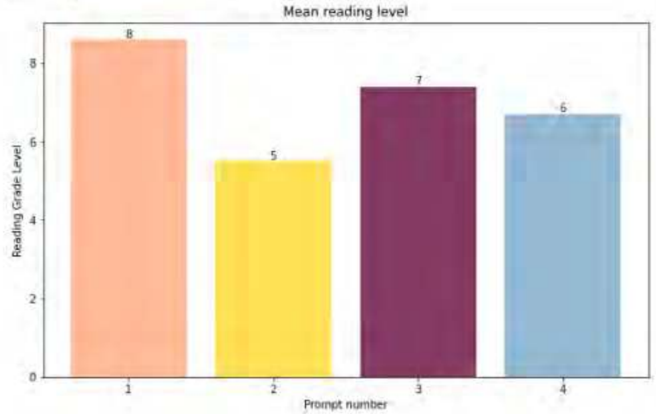
	Prompt	Mean Quality	Median Quality	STD
0	1	3.3	4	0.86
1	2	2.7	3	0.83
2	3	3.5	4	0.89
3	4	3.4	3	0.93



	Prompt	Mean Satisfied	Median Satisfied	STD
0	1	3.5	4	0.91
1	2	3.0	3	1.00
2	3	3.7	4	0.97
3	4	3.6	4	0.98



	Prompt	Reading Level (mean)
0	1	8.6
1	2	5.5
2	3	7.4
3	4	6.7

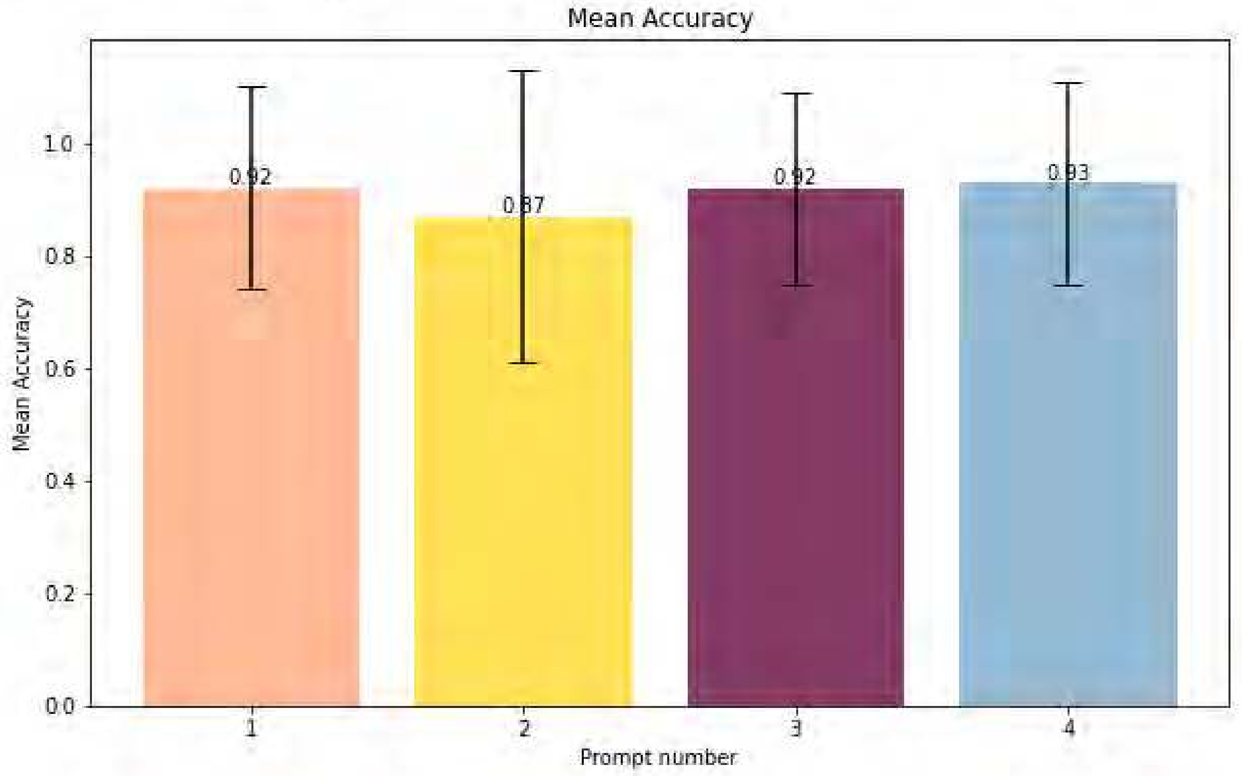


KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Accuracy

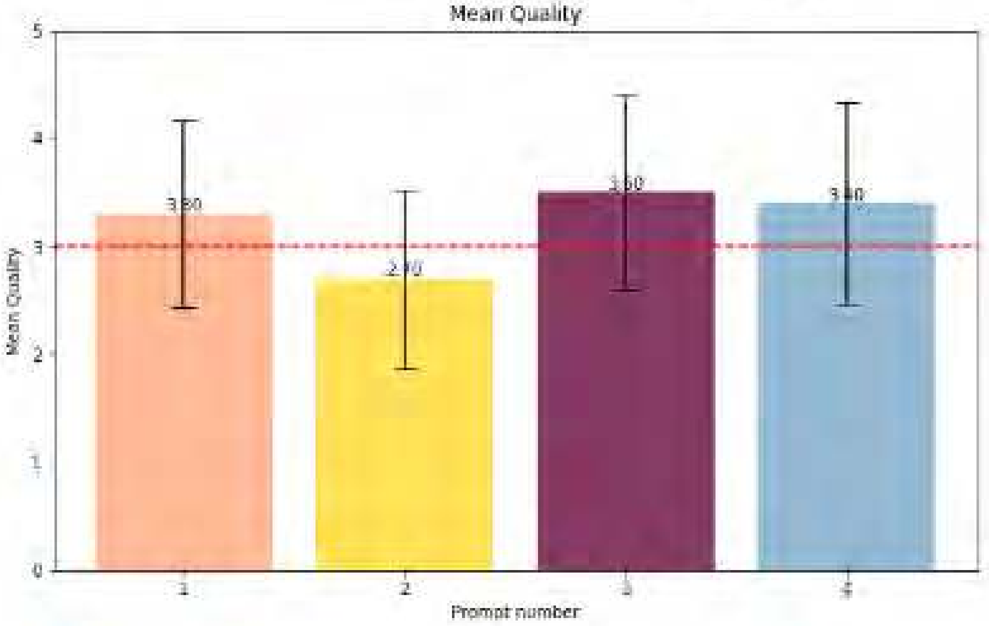
- Prompt:
  - *“Write a patient educational letter that addresses the following doctor's response to a recently resulted lab or image result: [text]”*
  - *“Write a patient educational letter at a 5th-grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]”*
  - “Write a patient educational letter with patient-friendly explanations that addresses the following doctor's response to a recently resulted lab or image result: [text]”
  - *“Write a patient educational letter with patient-friendly explanations at a 5th grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]”*

	Prompt	Mean Accuracy	Median Accuracy	STD
0	1	0.92	1	0.18
1	2	0.87	1	0.26
2	3	0.92	1	0.17
3	4	0.93	1	0.18



KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

	Prompt	Mean Quality	Median Quality	STD
0	1	3.3	4	0.86
1	2	2.7	3	0.83
2	3	3.5	4	0.89
3	4	3.4	3	0.93



Quality

- Prompt:

1. *"Write a patient educational letter that addresses the following doctor's response to a recently resulted lab or image result: [text]"*
2. *"Write a patient educational letter at a 5th-grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]"*
3. "Write a patient educational letter with patient-friendly explanations that addresses the following doctor's response to a recently resulted lab or image result: [text]"
4. *"Write a patient educational letter with patient-friendly explanations at a 5th grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]"*

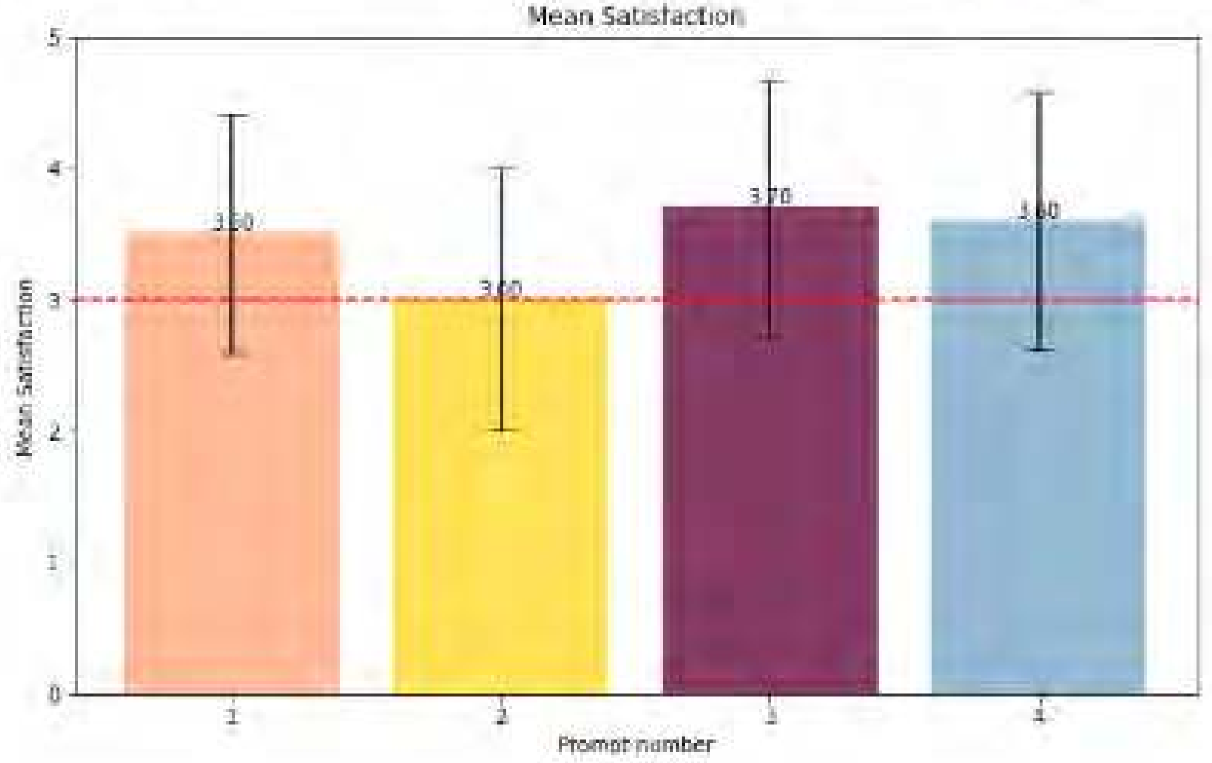
KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Satisfaction

- Prompt:

1. *"Write a patient educational letter that addresses the following doctor's response to a recently resulted lab or image result: [text]"*
2. *"Write a patient educational letter at a 5th-grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]"*
3. "Write a patient educational letter with patient-friendly explanations that addresses the following doctor's response to a recently resulted lab or image result: [text]"
4. *"Write a patient educational letter with patient-friendly explanations at a 5th grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]"*

	Prompt	Mean Satisfied	Median Satisfied	STD
0	1	3.5	4	0.91
1	2	3.0	3	1.00
2	3	3.7	4	0.97
3	4	3.6	4	0.98



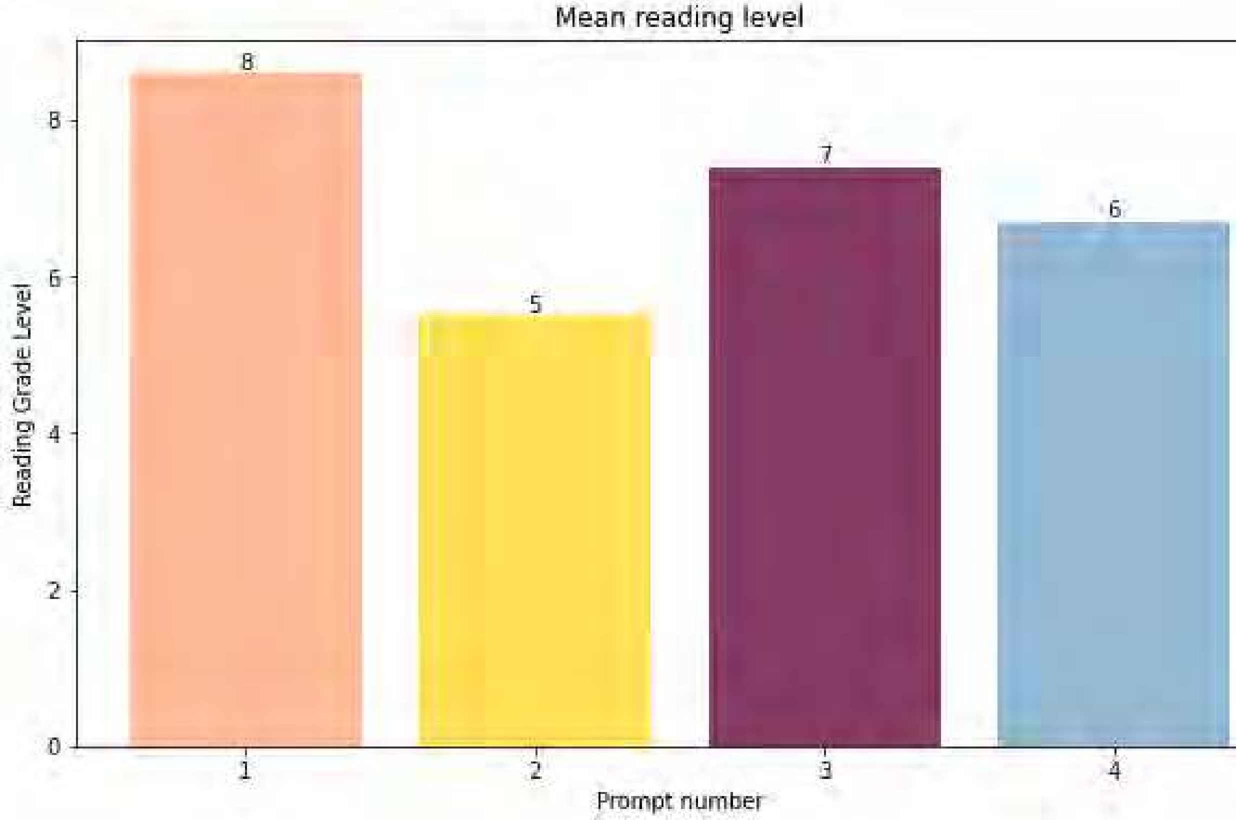
KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Reading Level

- Prompt:

1. *“Write a patient educational letter that addresses the following doctor's response to a recently resulted lab or image result: [text]”*
2. *“Write a patient educational letter at a 5th-grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]”*
3. “Write a patient educational letter with patient-friendly explanations that addresses the following doctor's response to a recently resulted lab or image result: [text]”
4. *“Write a patient educational letter with patient-friendly explanations at a 5th grade reading level that addresses the following doctor's response to a recently resulted lab or image result: [text]”*

	Prompt	Reading Level (mean)
0	1	8.6
1	2	5.5
2	3	7.4
3	4	6.7



KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Conclusions

These findings show that ChatGPT (version 3.5) produces generally accurate patient lab follow-up letters. ChatGPT did however produce neutral-quality patient letters and physicians were only moderately satisfied with the letters generated. Furthermore, ChatGPT did not produce letters at a patient-oriented reading level.

Prompt structure may result in mild changes the output accuracy, quality, satisfaction and reading level.

Accuracy =

Quality =

Satisfaction =

Reading Level =

KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Limitations

Text used for ChatGPT was not written with "prompting" in mind.

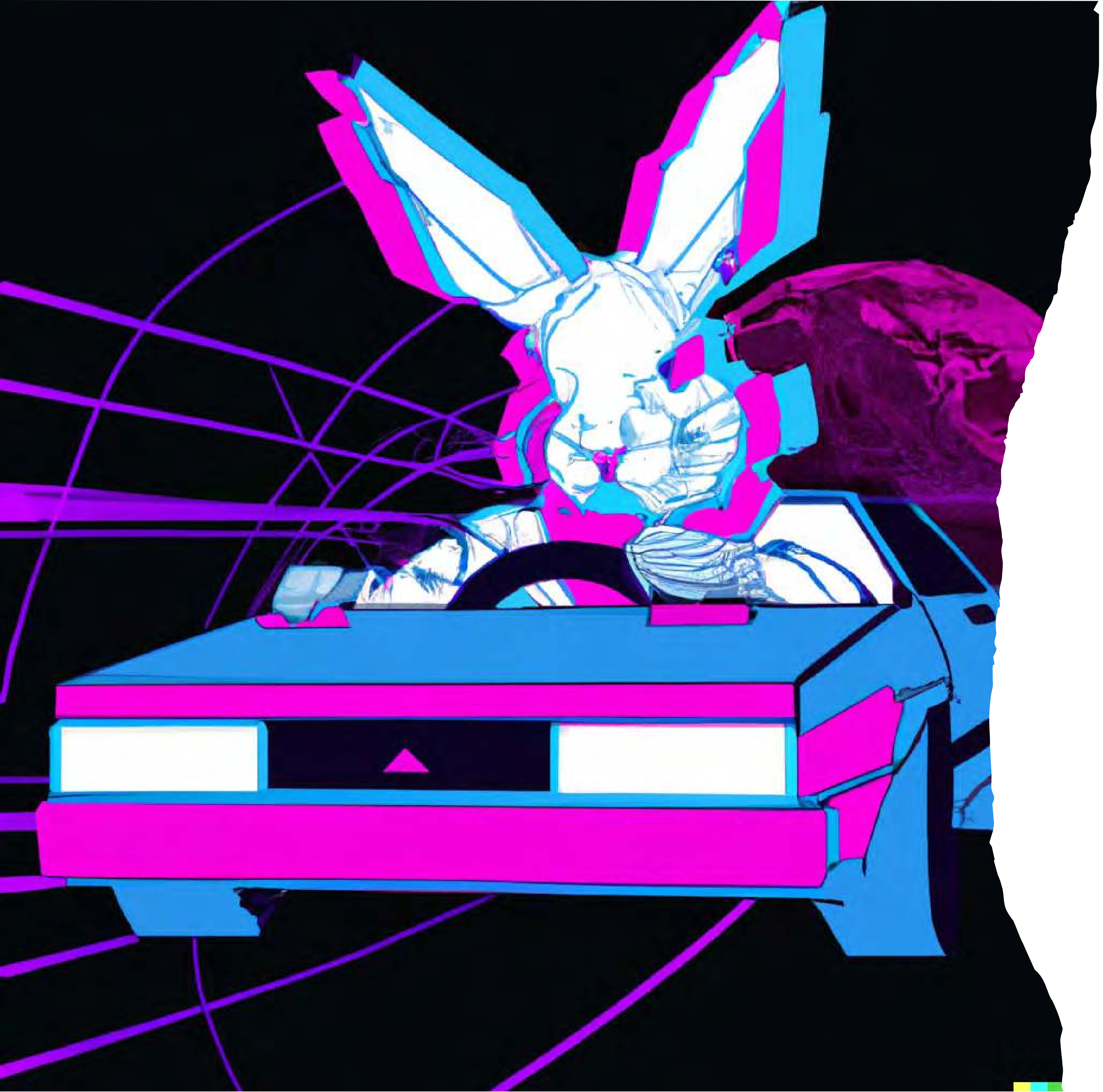
More focused writing could have significantly changed the output

Prompt structures could be further optimized

How does this compare to the nurse – patient phone call?

KU SCHOOL OF MEDICINE WICHITA
The University of Kansas

Interpretation



Consider the Future

Moore's Law: The number of transistors on a microchip doubles about every two years, though the cost of computers is halved.

Moore's Law, if true…

Now

10 years (current undergraduates)

Common Among Us

- Diabetes Education and Counseling
- Very basic
  - Teach correct principles,
  - answer questions,
  - inspire change

Teach Correct Principles

- Medical Knowledge

OpenAI

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd	670/800 ~87th
SAT Math	700/800 ~89th	690/800 ~89th	590/800 ~70th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd	147/170 ~25th
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th	154/170 ~63rd
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th	87/150 99th–100th	43/150 31st–33rd
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%

Teach Correct Principles

- Medical Knowledge

OpenAI

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd	670/800 ~87th
SAT Math	700/800 ~89th	690/800 ~89th	590/800 ~70th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd	147/170 ~25th
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th	154/170 ~63rd
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th	87/150 99th–100th	43/150 31st–33rd
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%

Research Letter

June 15, 2023

Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge

Zahir Kanjee, MD, MPH[1]; Byron Crowe, MD[1]; Adam Rodman, MD, MPH[1]

» Author Affiliations

JAMA. 2023;330(1):78-80. doi:10.1001/jama.2023.8288

Results: "A generative AI model provided the correct diagnosis in its differential in 64% of challenging cases and as its top diagnosis in 39%"

Teach Correct Principles

- Medical Knowledge

OpenAI

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298/400 ~90th	298/400 ~90th	213/400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd	670/800 ~87th
SAT Math	700/800 ~89th	690/800 ~89th	590/800 ~70th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd	147/170 ~25th
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th	154/170 ~63rd
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th	4/6 ~54th
USABO Semifinal Exam 2020	87/150 99th–100th	87/150 99th–100th	43/150 31st–33rd
USNCO Local Section Exam 2022	36/60	38/60	24/60
Medical Knowledge Self-Assessment Program	75%	75%	53%

Research Letter

June 15, 2023

Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge

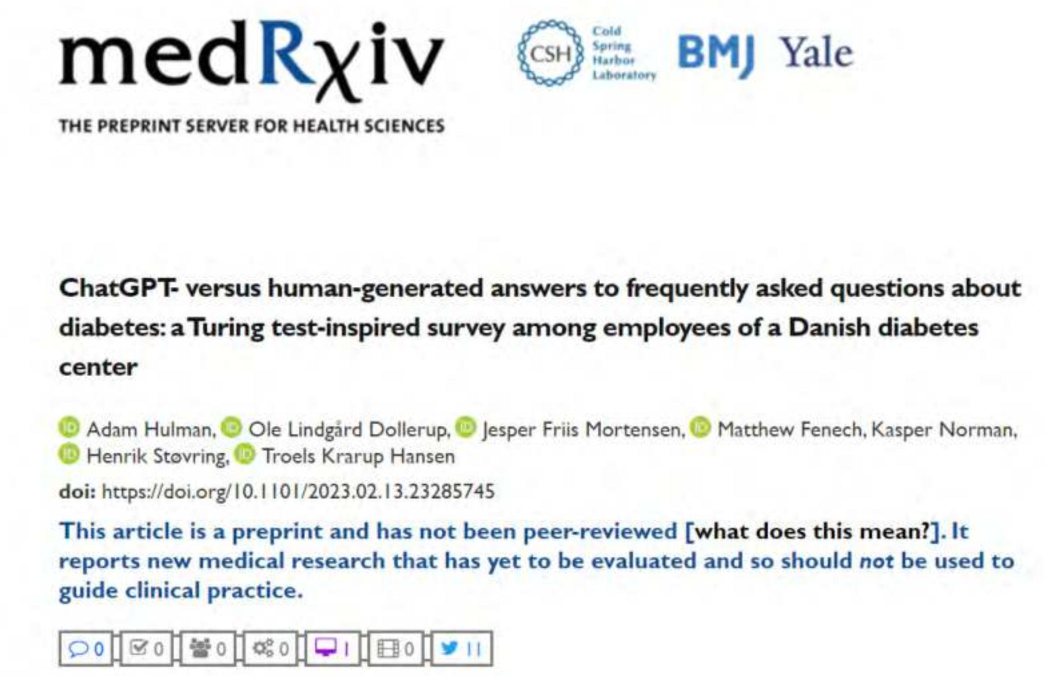
Zahir Kanjee, MD, MPH[1]; Byron Crowe, MD[1]; Adam Rodman, MD, MPH[1]

» Author Affiliations

JAMA. 2023;330(1):78-80. doi:10.1001/jama.2023.8288

Results: "A generative AI model provided the correct diagnosis in its differential in 64% of challenging cases and as its top diagnosis in 39%"

Answer Questions



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES
CSH Cold Spring Harbor Laboratory BMJ Yale

ChatGPT- versus human-generated answers to frequently asked questions about diabetes: a Turing test-inspired survey among employees of a Danish diabetes center

Adam Hulman, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew Fenech, Kasper Norman, Henrik Støvring, Troels Krarup Hansen

doi: https://doi.org/10.1101/2023.02.13.23285745

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

- 311 individuals who work at a diabetes healthcare center determine if answers to common diabetes questions were generated by a human expert or ChatGPT
- Results
  - Participants could determine who answered the question 59.5% of the time
    - Better than coin flip
- Important: the ChatGPT had 2 incorrect answers (out of 10 answers total)

Adam Hulman, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew Fenech, Kasper Norman, Henrik Støvring, Troels Krarup Hansen
medRxiv 2023.02.13.23285745; doi: https://doi.org/10.1101/2023.02.13.23285745

Answer Questions


medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

ChatGPT- versus human-generated answers to frequently asked questions about diabetes: a Turing test-inspired survey among employees of a Danish diabetes center

Adam Hulman, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew Fenech, Kasper Norman, Henrik Støvring, Troels Krarup Hansen

doi: https://doi.org/10.1101/2023.02.13.23285745

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

0 0 0 0 1 0 11

- 311 individuals who work at a diabetes healthcare center determine if answers to common diabetes questions were generated by a human expert or ChatGPT
- Results
  - Participants could determine who answered the question 59.5% of the time
    - Better than coin flip
- Important: the ChatGPT had 2 incorrect answers (out of 10 answers total)

Adam Hulman, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew Fenech, Kasper Norman, Henrik Støvring, Troels Krarup Hansen
medRxiv 2023.02.13.23285745; doi: https://doi.org/10.1101/2023.02.13.23285745

Answer Questions

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

ChatGPT- versus human-generated answers to frequently asked questions about diabetes: a Turing test-inspired survey among employees of a Danish diabetes center

Adam Hulman, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew Fenech, Kasper Norman, Henrik Støvring, Troels Krarup Hansen

doi: https://doi.org/10.1101/2023.02.13.23285745

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

- 311 individuals who work at a diabetes healthcare center determine if answers to common diabetes questions were generated by a human expert or ChatGPT
- Results
  - Participants could determine who answered the question 59.5% of the time
    - Better than coin flip
- Important: the ChatGPT had 2 incorrect answers (out of 10 answers total)

Adam Hulman, Ole Lindgård Dollerup, Jesper Friis Mortensen, Matthew Fenech, Kasper Norman, Henrik Støvring, Troels Krarup Hansen
medRxiv 2023.02.13.23285745; doi: https://doi.org/10.1101/2023.02.13.23285745

Inspire Change

> JAMA Intern Med. 2023 Apr 28;e231838. doi: 10.1001/jamainternmed.2023.1838.
Online ahead of print.

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W Ayers [1, 2], Adam Poliak [3], Mark Dredze [4], Eric C Leas [1, 5], Zechariah Zhu [1], Jessica B Kelley [6], Dennis J Faix [7], Aaron M Goodman [8, 9], Christopher A Longhurst [10], Michael Hogarth [10, 11], Davey M Smith [2, 11]

Affiliations + expand

PMID: 37115527 PMCID: PMC10148230 (available on 2024-04-28)
DOI: 10.1001/jamainternmed.2023.1838

Design:

“a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from October 2022 where a verified physician responded to a public question. Chatbot responses were generated by entering the original question”

- ## Results:
  - Chatbot responses were rated of significantly higher quality than physician responses ($t = 13.3$; $P < .001$)
    - 3.6 times higher prevalence of good or very good quality responses for the chatbot (> Humans)
  - Chatbot responses were also rated significantly more empathetic than physician responses ($t = 18.9$; $P < .001$)
    - The proportion of responses rated empathetic or very empathetic (≥4) was higher for chatbot than for physicians

Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

Inspire Change

> JAMA Intern Med. 2023 Apr 28;e231838. doi: 10.1001/jamainternmed.2023.1838.
Online ahead of print.

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W Ayers [1,2], Adam Poliak [3], Mark Dredze [4], Eric C Leas [1,5], Zechariah Zhu [1], Jessica B Kelley [6], Dennis J Faix [7], Aaron M Goodman [8,9], Christopher A Longhurst [10], Michael Hogarth [10,11], Davey M Smith [2,11]

Affiliations + expand

PMID: 37115527 PMCID: PMC10148230 (available on 2024-04-28)
DOI: 10.1001/jamainternmed.2023.1838

Design:

"a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from October 2022 where a verified physician responded to a public question. Chatbot responses were generated by entering the original question"

- Results:
  - Chatbot responses were rated of significantly higher quality than physician responses (t = 13.3; P < .001)
    - 3.6 times higher prevalence of good or very good quality responses for the chatbot (> Humans)
  - Chatbot responses were also rated significantly more empathetic than physician responses (t = 18.9; P < .001)
    - The proportion of responses rated empathetic or very empathetic (≥4) was higher for chatbot than for physicians

Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

Inspire Change

> JAMA Intern Med. 2023 Apr 28;e231838. doi: 10.1001/jamainternmed.2023.1838.
Online ahead of print.

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W Ayers [1,2], Adam Poliak [3], Mark Dredze [4], Eric C Leas [1,5], Zechariah Zhu [1], Jessica B Kelley [6], Dennis J Faix [7], Aaron M Goodman [8,9], Christopher A Longhurst [10], Michael Hogarth [10,11], Davey M Smith [2,11]

Affiliations + expand

PMID: 37115527 PMCID: PMC10148230 (available on 2024-04-28)
DOI: 10.1001/jamainternmed.2023.1838

Design:

"a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from October 2022 where a verified physician responded to a public question. Chatbot responses were generated by entering the original question"

- **Results:**
  - Chatbot responses were rated of significantly higher quality than physician responses ($t = 13.3$; $P < .001$)
    - 3.6 times higher prevalence of good or very good quality responses for the chatbot (> Humans)
  - Chatbot responses were also rated significantly more empathetic than physician responses ($t = 18.9$; $P < .001$)
    - The proportion of responses rated empathetic or very empathetic (≥4) was higher for chatbot than for physicians

Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

Inspiring Change

Questions?