Prompt engineering

CS 780/880 Natural Language Processing Lecture 21

Samuel Carton, University of New Hampshire

Last lecture

Hugging Face model variants

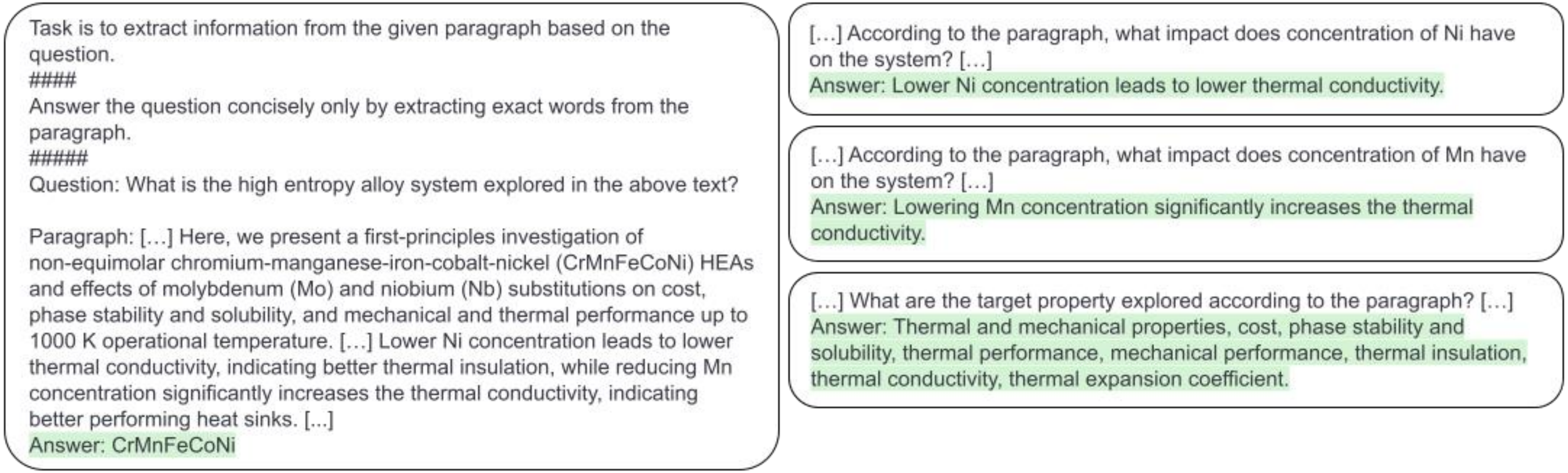
Contemporary LLMs: GPT-3, ChatGPT, GPT-4

Paradigm shift in machine learning:

- Moving away from gradient descent (a lot of the time)
- Moving toward **in-context learning**
  - Zero-shot learning: tell the model to do stuff
  - Few-shot learning: give a few examples first

Is there more to it? (yes)

Zero-shot learning



Task is to extract information from the given paragraph based on the question.
####
Answer the question concisely only by extracting exact words from the paragraph.
#####
Question: What is the high entropy alloy system explored in the above text?

Paragraph: […] Here, we present a first-principles investigation of non-equimolar chromium-manganese-iron-cobalt-nickel (CrMnFeCoNi) HEAs and effects of molybdenum (Mo) and niobium (Nb) substitutions on cost, phase stability and solubility, and mechanical and thermal performance up to 1000 K operational temperature. […] Lower Ni concentration leads to lower thermal conductivity, indicating better thermal insulation, while reducing Mn concentration significantly increases the thermal conductivity, indicating better performing heat sinks. [...]
Answer: CrMnFeCoNi

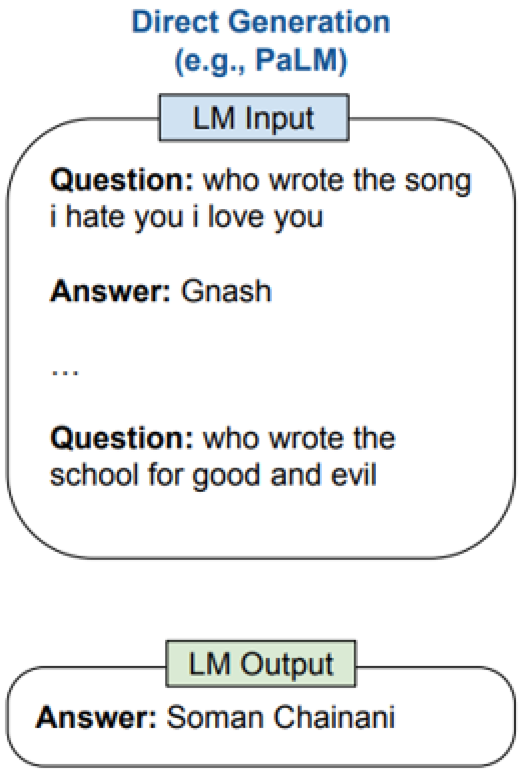
[…] According to the paragraph, what impact does concentration of Ni have on the system? […]
Answer: Lower Ni concentration leads to lower thermal conductivity.

[…] According to the paragraph, what impact does concentration of Mn have on the system? […]
Answer: Lowering Mn concentration significantly increases the thermal conductivity.

[…] What are the target property explored according to the paragraph? […]
Answer: Thermal and mechanical properties, cost, phase stability and solubility, thermal performance, mechanical performance, thermal insulation, thermal conductivity, thermal expansion coefficient.

Few-shot learning

A few examples from papers we're talking about today



Direct Generation
(e.g., PaLM)

LM Input

Question: who wrote the song i hate you i love you

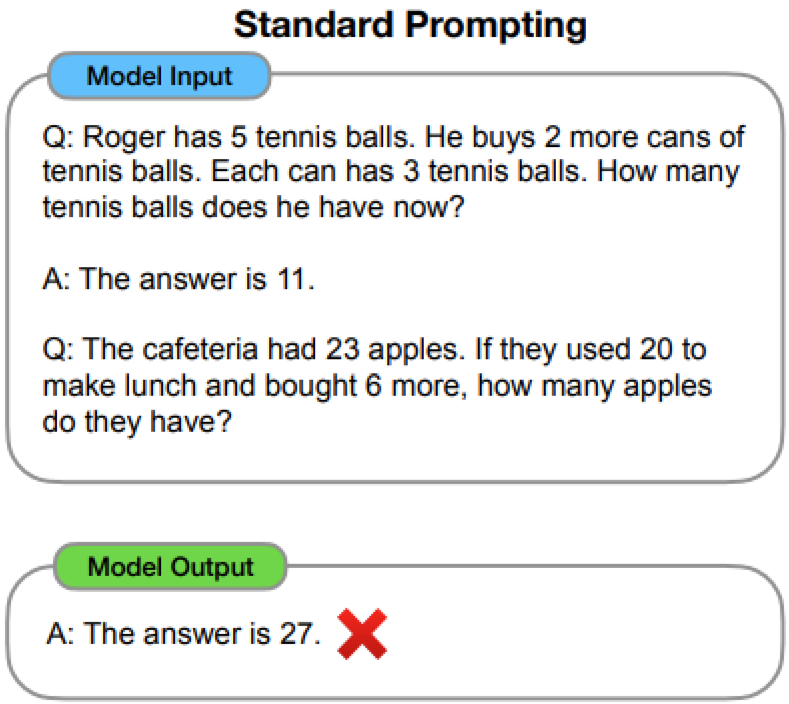
Answer: Gnash

…

Question: who wrote the school for good and evil

LM Output

Answer: Soman Chainani



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Problems with in-context learning

Ultimately, LLMs produce output that is **linguistically likely** regardless of what's actually true, interesting or reasonable.

This leads to a few key issues:

- Hallucination
- Reasoning errors
- Ungrounded outputs
  - Problem for robotics
- Boring outputs
- Biased outputs

Prompt engineering

The field of **prompt engineering** is emerging surrounding how to correctly prompt LLMs in order to avoid these kinds of issues.

What kind of information can we include in the prompt to the LLM (and how do we format that information) to get it to more reliably do what we want it to?

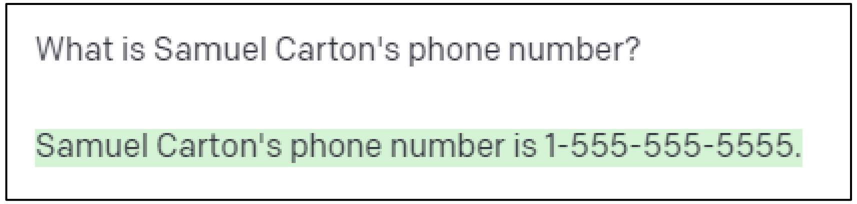
- Avoid hallucination
- Avoid reasoning mistakes
- Produce more interesting output
- Etc.

Very important to get a handle on this right now (2023)

This lecture: overview of a few recent prompt engineering papers tackling different problems.

Hallucination

When a model “hallucinates”, it makes up fake information which is linguistically likely but not actually true.

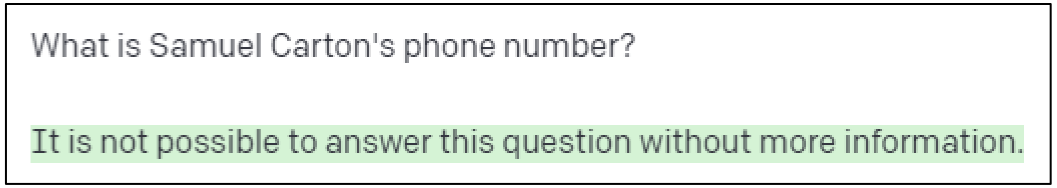


What is Samuel Carton's phone number?

Samuel Carton's phone number is 1-555-555-5555.

OpenAI playground; text-davinci-002

Important note: the whole point of Reinforcement Learning with Human Feedback (RLHF) is to get GPT to avoid these kinds of mistakes by training it to respond to certain kinds of inputs with non-answers.



What is Samuel Carton's phone number?

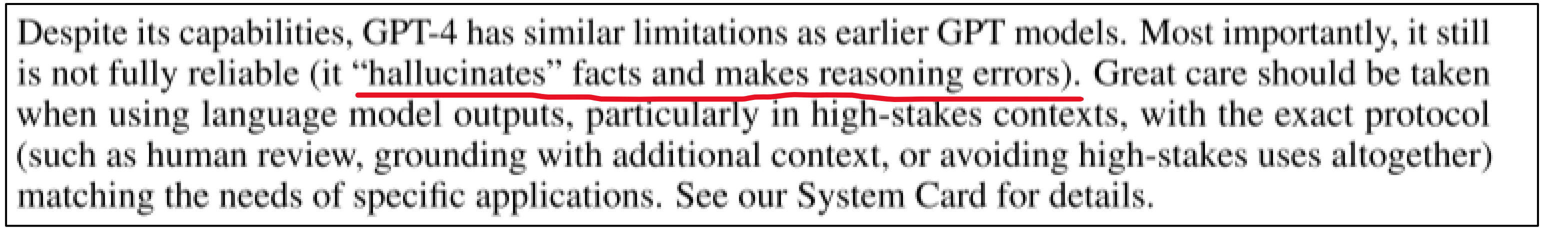
It is not possible to answer this question without more information.

OpenAI playground; text-davinci-003

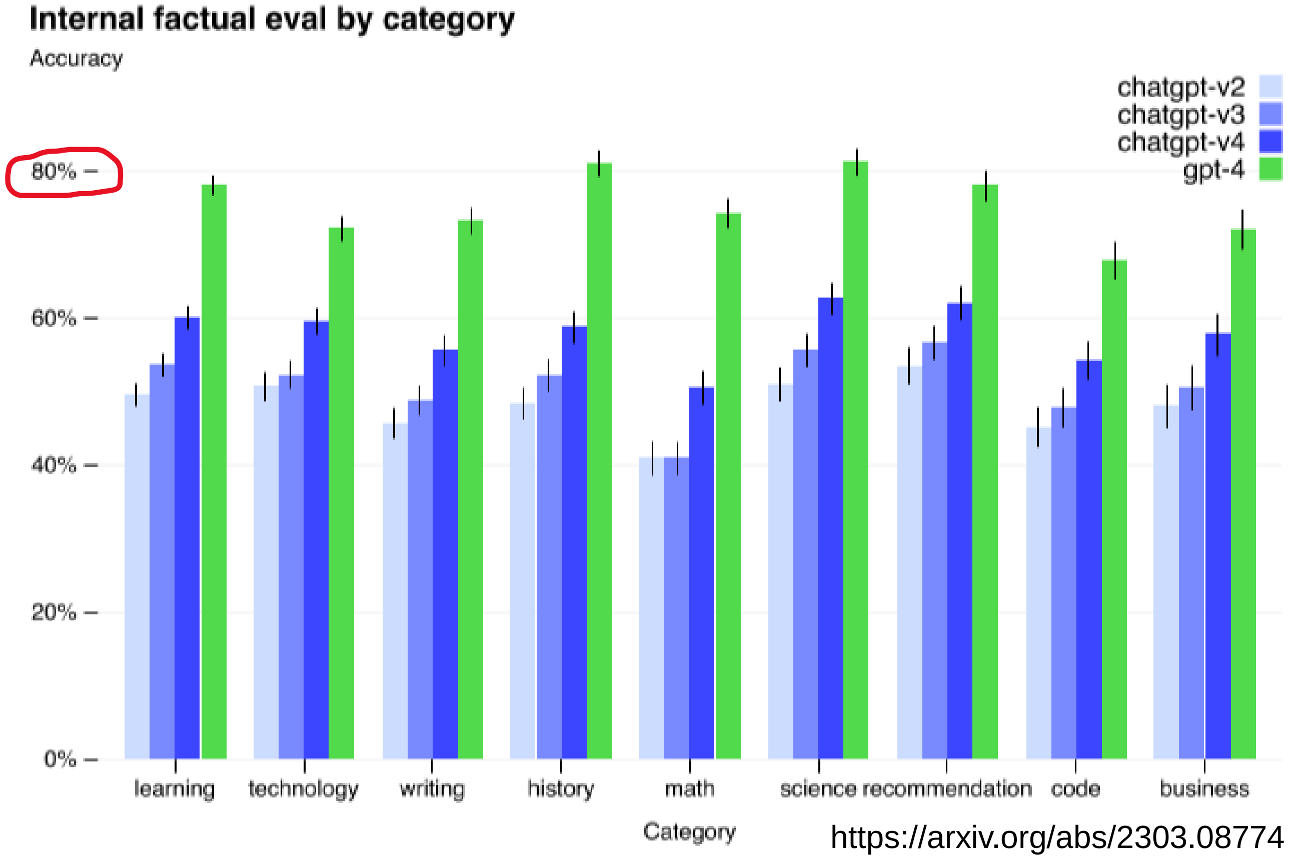
Hallucination

But even GPT-4 suffers from hallucination, and these models probably always will.

What other tools do we have in our toolbelt for alleviating the issue of hallucination?



Despite its capabilities, GPT-4 has similar limitations as earlier GPT models. Most importantly, it still is not fully reliable (it "hallucinates" facts and makes reasoning errors). Great care should be taken when using language model outputs, particularly in high-stakes contexts, with the exact protocol (such as human review, grounding with additional context, or avoiding high-stakes uses altogether) matching the needs of specific applications. See our System Card for details.



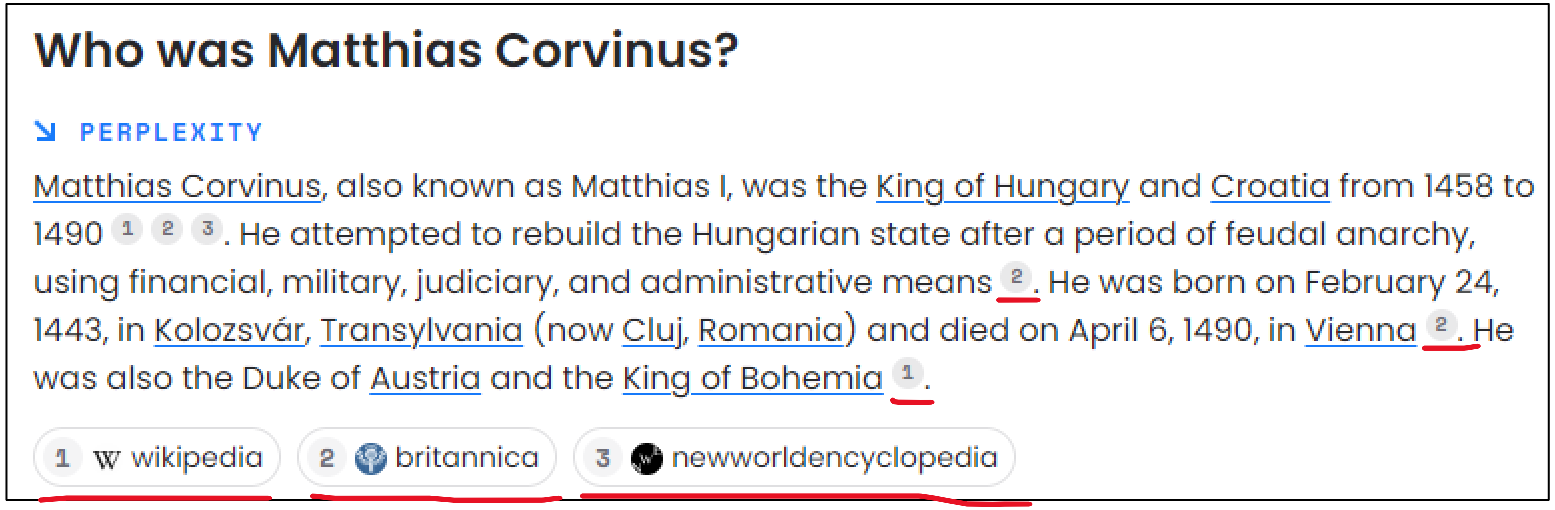
https://arxiv.org/abs/2303.08774

Retrieval-augmented generation

Basic idea: Rather than just querying the LLM for information (which it can hallucinate):

1. Retrieve relevant documents or snippets
2. Give them to LLM as part of prompt
3. Ask LLM to generate response based on that information

OpenAI already has a prototype system that does this in a basic way: https://www.perplexity.ai/



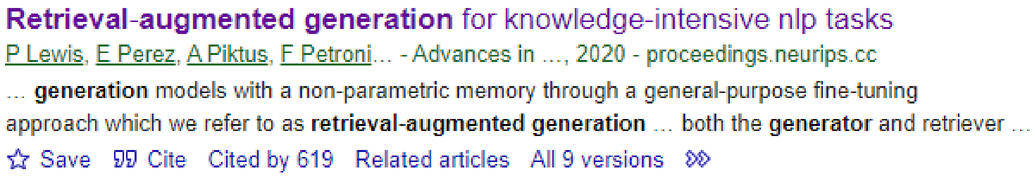
Who was Matthias Corvinus?

Matthias Corvinus, also known as Matthias I, was the King of Hungary and Croatia from 1458 to 1490 [1] [2] [3]. He attempted to rebuild the Hungarian state after a period of feudal anarchy, using financial, military, judiciary, and administrative means [2]. He was born on February 24, 1443, in Kolozsvár, Transylvania (now Cluj, Romania) and died on April 6, 1490, in Vienna [2]. He was also the Duke of Austria and the King of Bohemia [1].

Retrieval-augmented generation

The original paper about this idea is a few years old

- Assumes you can actually train the generator (not possible currently)

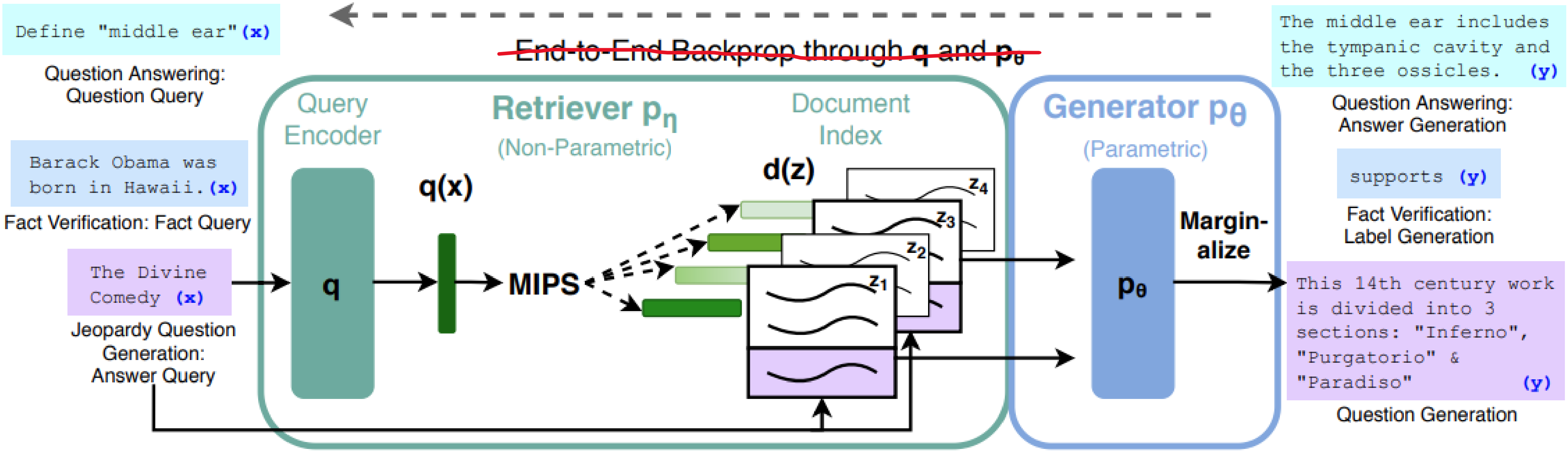


Retrieval-augmented generation for knowledge-intensive nlp tasks

P Lewis, E Perez, A Piktus, F Petroni… - Advances in …, 2020 - proceedings.neurips.cc

… **generation** models with a non-parametric memory through a general-purpose fine-tuning approach which we refer to as **retrieval-augmented generation** … both the **generator** and retriever …

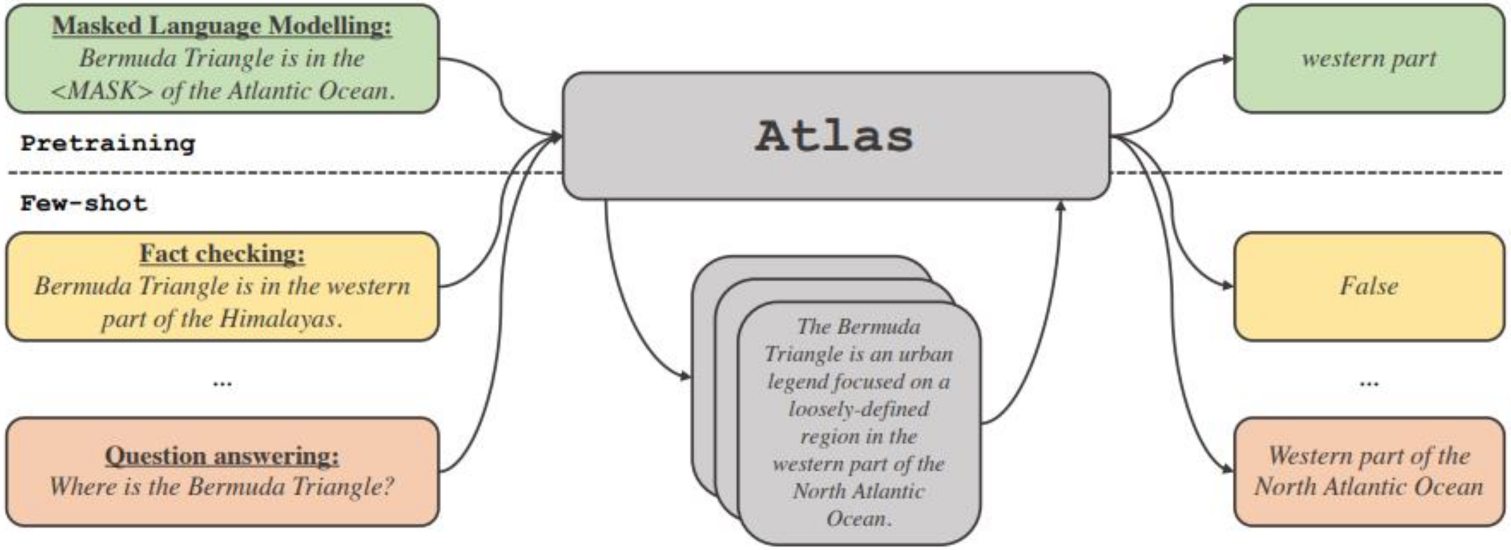
☆ Save 99 Cite Cited by 619 Related articles All 9 versions


~~End-to-End Backprop through q and p_θ~~
Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
Query Encoder
q
q(x)
Retriever p_η
(Non-Parametric)
MIPS
Document Index
d(z)
z_4
z_3
z_2
z_1
Generator p_θ
(Parametric)
p_θ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

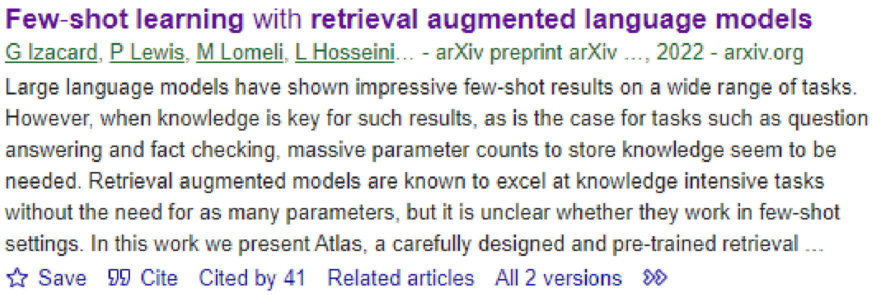
Atlas: Retrieval + few-shot learning

A more recent model is Atlas, which combines retrieval augmentation with few-shot learning

- Still involves actual fine-tuning of the generator though
- Uses “dense retrieval” (Contriever) to find relevant documents
- Then fine-tunes text-to-text model (T5) to generate correct answers given retrieved info



Masked Language Modelling:
Bermuda Triangle is in the <MASK> of the Atlantic Ocean.
Pretraining
Few-shot
Fact checking:
Bermuda Triangle is in the western part of the Himalayas.
...
Question answering:
Where is the Bermuda Triangle?
Atlas
The Bermuda Triangle is an urban legend focused on a loosely-defined region in the western part of the North Atlantic Ocean.
western part
False
...
Western part of the North Atlantic Ocean



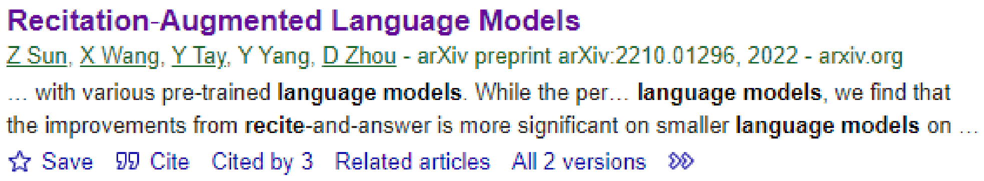
Few-shot learning with retrieval augmented language models

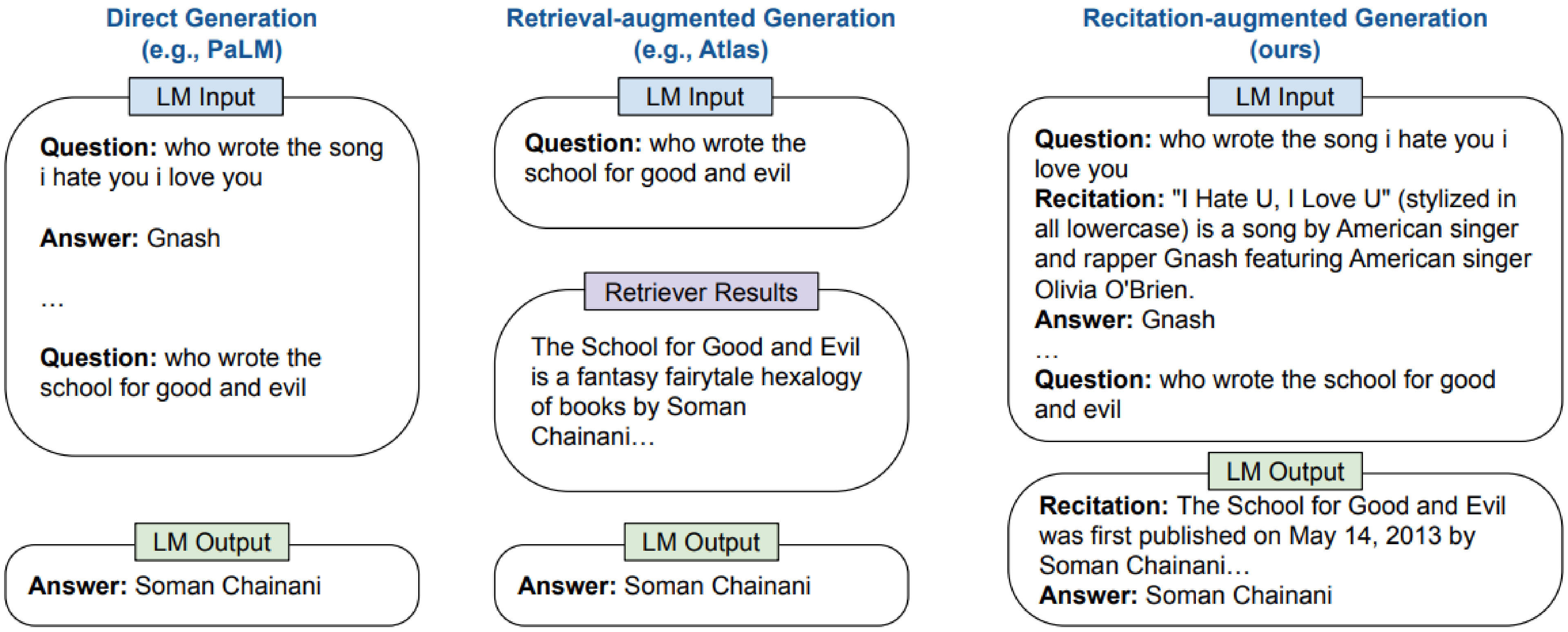
G Izacard, P Lewis, M Lomeli, L Hosseini… - arXiv preprint arXiv …, 2022 - arxiv.org

Large language models have shown impressive few-shot results on a wide range of tasks. However, when knowledge is key for such results, as is the case for tasks such as question answering and fact checking, massive parameter counts to store knowledge seem to be needed. Retrieval augmented models are known to excel at knowledge intensive tasks without the need for as many parameters, but it is unclear whether they work in few-shot settings. In this work we present Atlas, a carefully designed and pre-trained retrieval …

Recitation-augmented language models

Basic idea: Ask the model to **recite** facts it thinks it knows about a topic, then ask it to choose between them


Recitation-Augmented Language Models
Z Sun, X Wang, Y Tay, Y Yang, D Zhou - arXiv preprint arXiv:2210.01296, 2022 - arxiv.org
… with various pre-trained **language models**. While the per… **language models**, we find that the improvements from **recite**-and-answer is more significant on smaller **language models** on …


Direct Generation (e.g., PaLM)

LM Input

Question: who wrote the song i hate you i love you

Answer: Gnash

…

Question: who wrote the school for good and evil

LM Output

Answer: Soman Chainani

Retrieval-augmented Generation (e.g., Atlas)

LM Input

Question: who wrote the school for good and evil

Retriever Results

The School for Good and Evil is a fantasy fairytale hexalogy of books by Soman Chainani…

LM Output

Answer: Soman Chainani

Recitation-augmented Generation (ours)

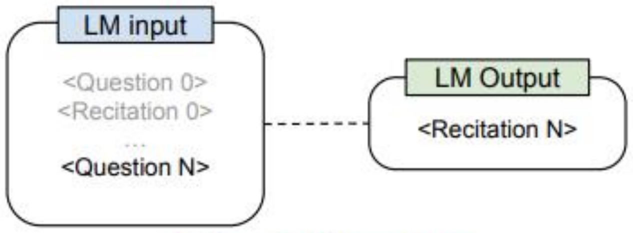
LM Input

Question: who wrote the song i hate you i love you
Recitation: "I Hate U, I Love U" (stylized in all lowercase) is a song by American singer and rapper Gnash featuring American singer Olivia O'Brien.
Answer: Gnash
…
Question: who wrote the school for good and evil

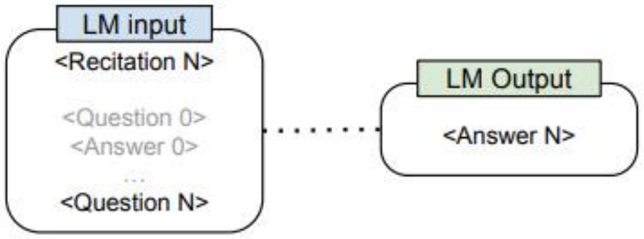
LM Output

Recitation: The School for Good and Evil was first published on May 14, 2013 by Soman Chainani…
Answer: Soman Chainani

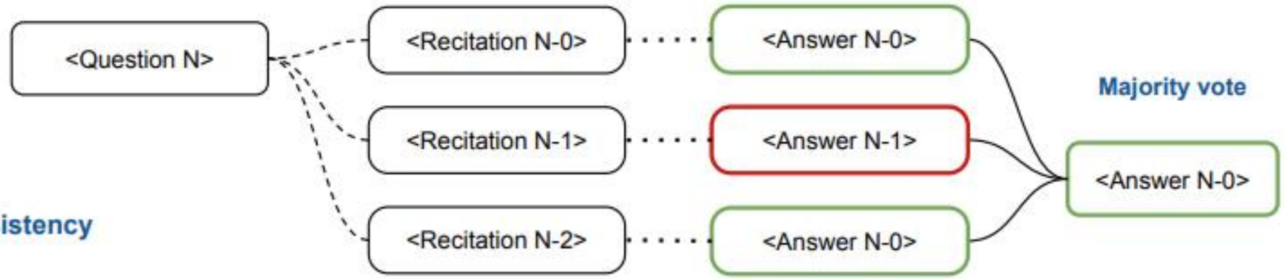
Recitation-augmented language models



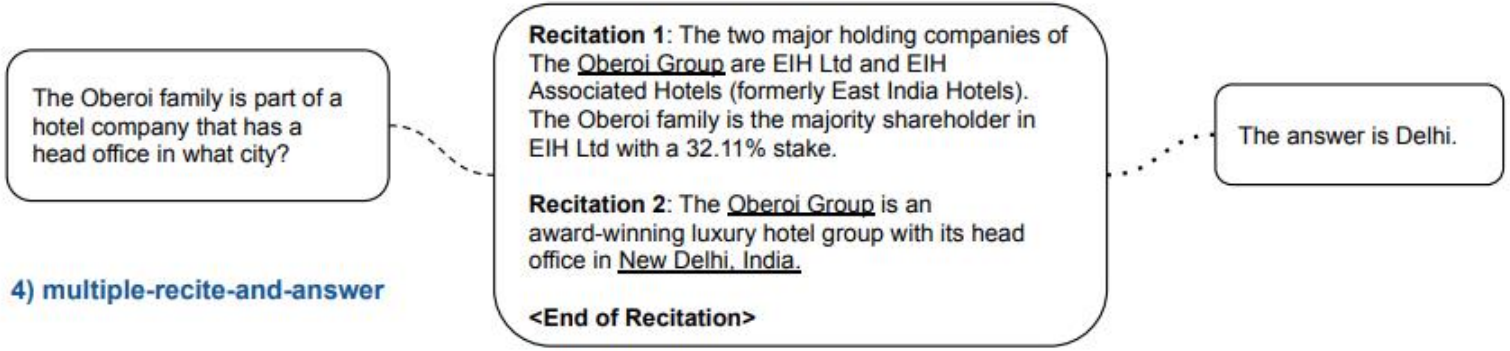
1) few-shot in-context learning for recitation



2) few-shot in-context learning for recitation-augmented QA



3) self-consistency


The Oberoi family is part of a hotel company that has a head office in what city?

Recitation 1: The two major holding companies of The Oberoi Group are EIH Ltd and EIH Associated Hotels (formerly East India Hotels). The Oberoi family is the majority shareholder in EIH Ltd with a 32.11% stake.

Recitation 2: The Oberoi Group is an award-winning luxury hotel group with its head office in New Delhi, India.

<End of Recitation>

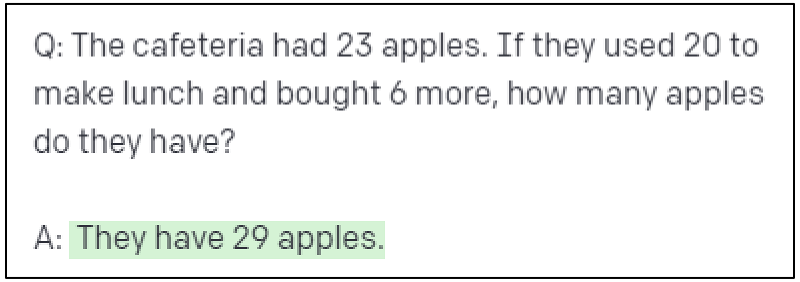
The answer is Delhi.

4) multiple-recite-and-answer

Reasoning errors

LLMS also tend to make reasoning errors when you ask them questions that require multiple logical steps

- Again, this is because they are **language models**, not logic engines



Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

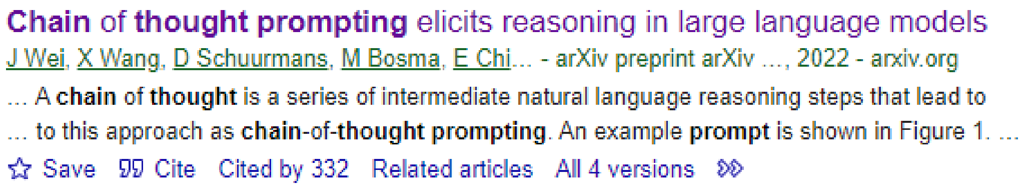
A: They have 29 apples.

OpenAI playground; text-davinci-003

Can you see what mistake the model made here?

Chain-of-thought prompting

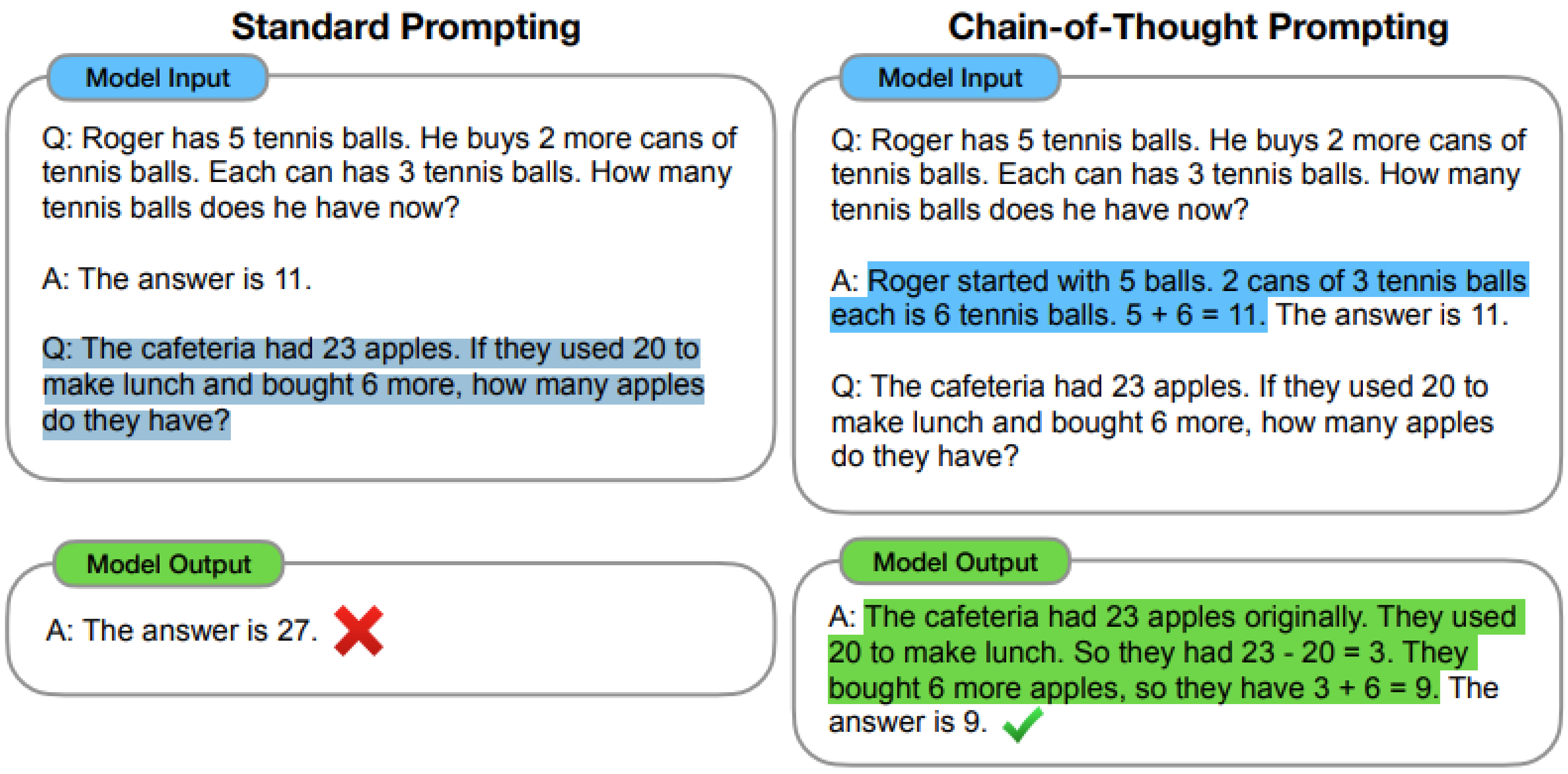
Basic idea: Give few-shot examples of not just the solution, but the **intermediate steps** for solving the problem



Chain of thought prompting elicits reasoning in large language models

J Wei, X Wang, D Schuurmans, M Bosma, E Chi… - arXiv preprint arXiv …, 2022 - arxiv.org

… A **chain of thought** is a series of intermediate natural language reasoning steps that lead to … to this approach as **chain-of-thought prompting**. An example **prompt** is shown in Figure 1. …



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

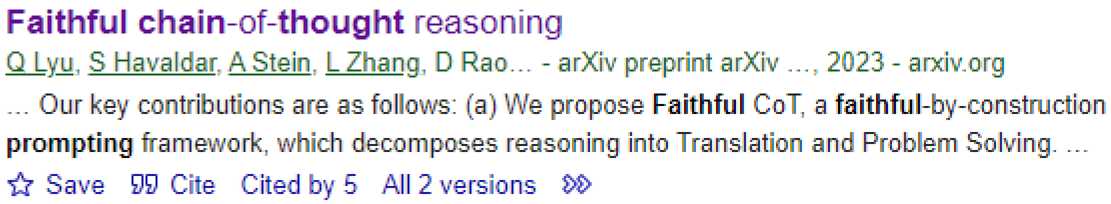
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

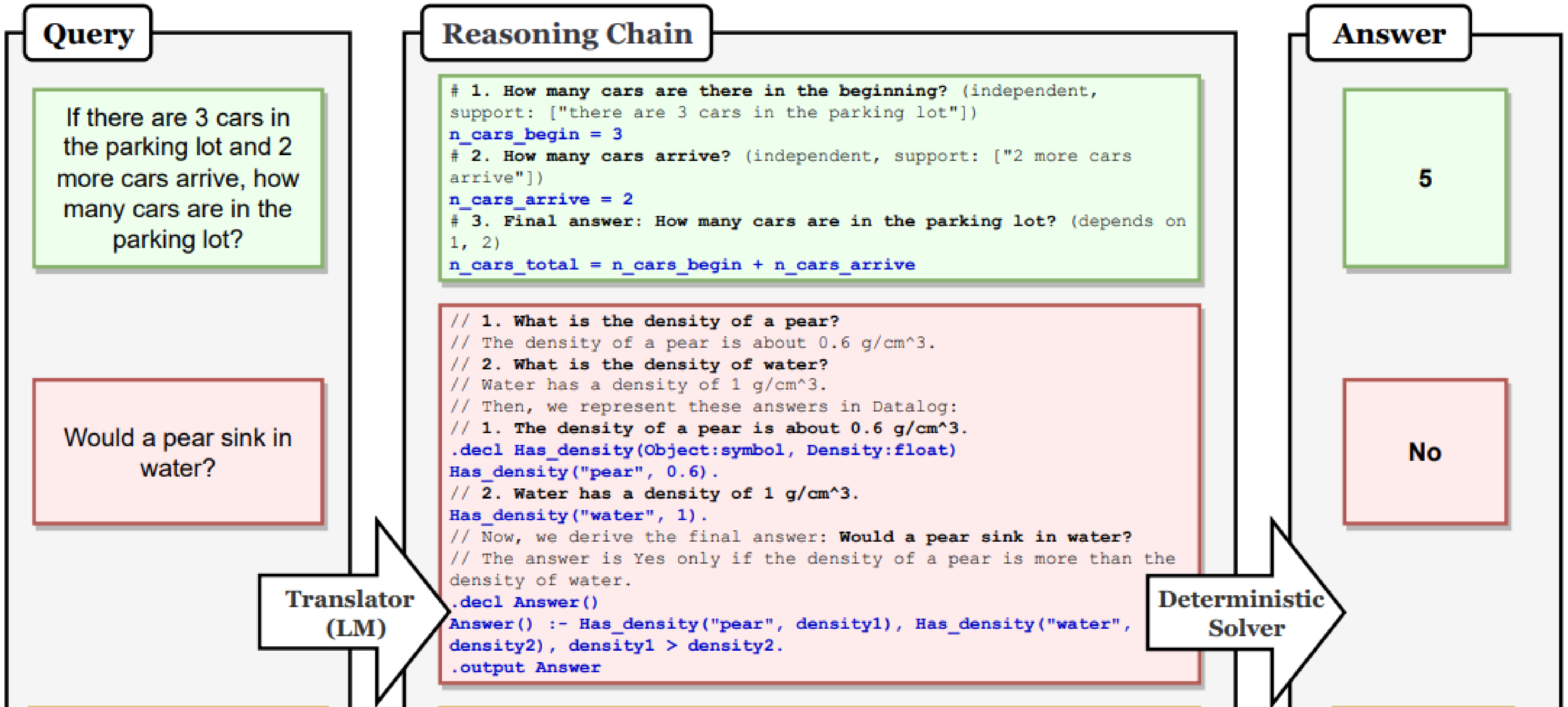
Faithful chain-of-thought prompting

Basic idea: Give few-shot examples of not just the intermediate steps, but **structured language** for solving the problem


Faithful chain-of-thought reasoning

Q Lyu, S Havaldar, A Stein, L Zhang, D Rao… - arXiv preprint arXiv …, 2023 - arxiv.org

… Our key contributions are as follows: (a) We propose **Faithful** CoT, a **faithful**-by-construction **prompting** framework, which decomposes reasoning into Translation and Problem Solving. …



If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

```
# 1. How many cars are there in the beginning? (independent,
support: ["there are 3 cars in the parking lot"])
n_cars_begin = 3
# 2. How many cars arrive? (independent, support: ["2 more cars
arrive"])
n_cars_arrive = 2
# 3. Final answer: How many cars are in the parking lot? (depends on
1, 2)
n_cars_total = n_cars_begin + n_cars_arrive
```

5

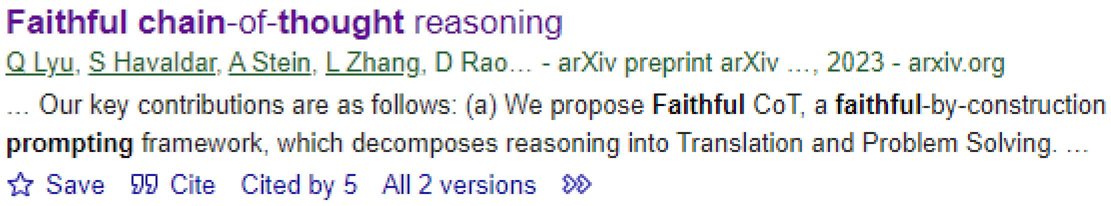
Would a pear sink in water?

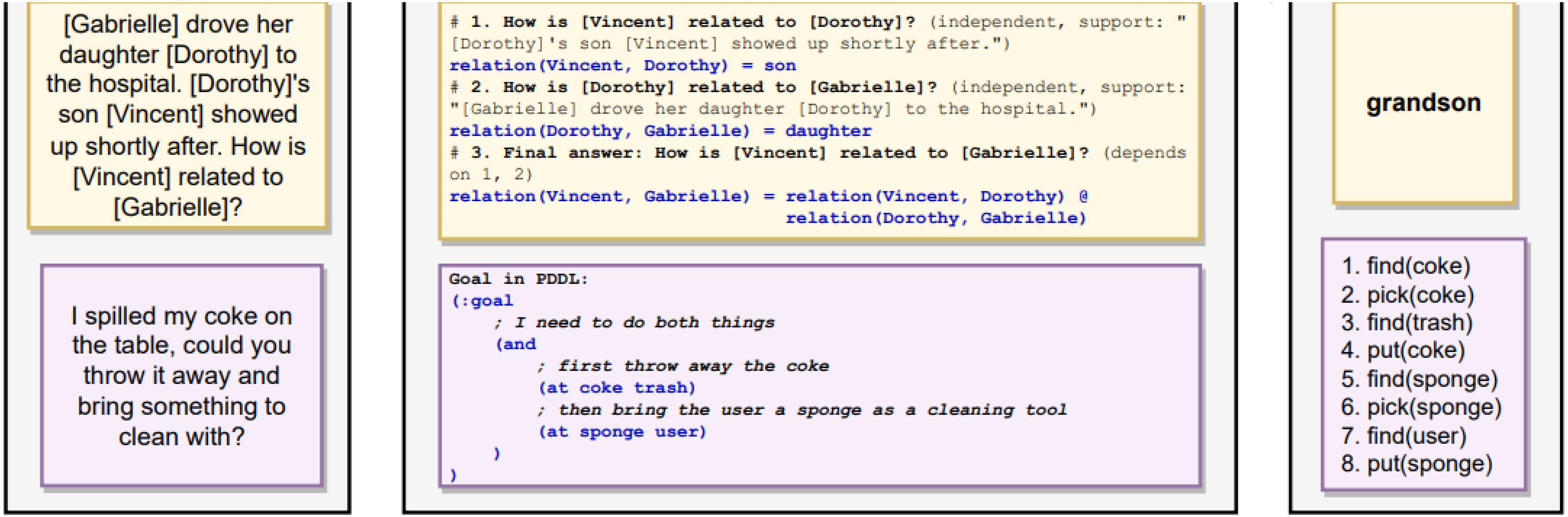
```
// 1. What is the density of a pear?
// The density of a pear is about 0.6 g/cm^3.
// 2. What is the density of water?
// Water has a density of 1 g/cm^3.
// Then, we represent these answers in Datalog:
// 1. The density of a pear is about 0.6 g/cm^3.
.decl Has_density(Object:symbol, Density:float)
Has_density("pear", 0.6).
// 2. Water has a density of 1 g/cm^3.
Has_density("water", 1).
// Now, we derive the final answer: Would a pear sink in water?
// The answer is Yes only if the density of a pear is more than the
density of water.
.decl Answer()
Answer() :- Has_density("pear", density1), Has_density("water",
density2), density1 > density2.
.output Answer
```

No

Faithful chain-of-thought prompting

Basic idea: Give few-shot examples of not just the intermediate steps, but **structured language** for solving the problem


Faithful chain-of-thought reasoning
Q Lyu, S Havaldar, A Stein, L Zhang, D Rao… - arXiv preprint arXiv …, 2023 - arxiv.org
… Our key contributions are as follows: (a) We propose **Faithful** CoT, a **faithful**-by-construction **prompting** framework, which decomposes reasoning into Translation and Problem Solving. …


[Gabrielle] drove her daughter [Dorothy] to the hospital. [Dorothy]'s son [Vincent] showed up shortly after. How is [Vincent] related to [Gabrielle]?

```
# 1. How is [Vincent] related to [Dorothy]? (independent, support: "
[Dorothy]'s son [Vincent] showed up shortly after.")
relation(Vincent, Dorothy) = son
# 2. How is [Dorothy] related to [Gabrielle]? (independent, support:
"[Gabrielle] drove her daughter [Dorothy] to the hospital.")
relation(Dorothy, Gabrielle) = daughter
# 3. Final answer: How is [Vincent] related to [Gabrielle]? (depends
on 1, 2)
relation(Vincent, Gabrielle) = relation(Vincent, Dorothy) @
                               relation(Dorothy, Gabrielle)
```

grandson

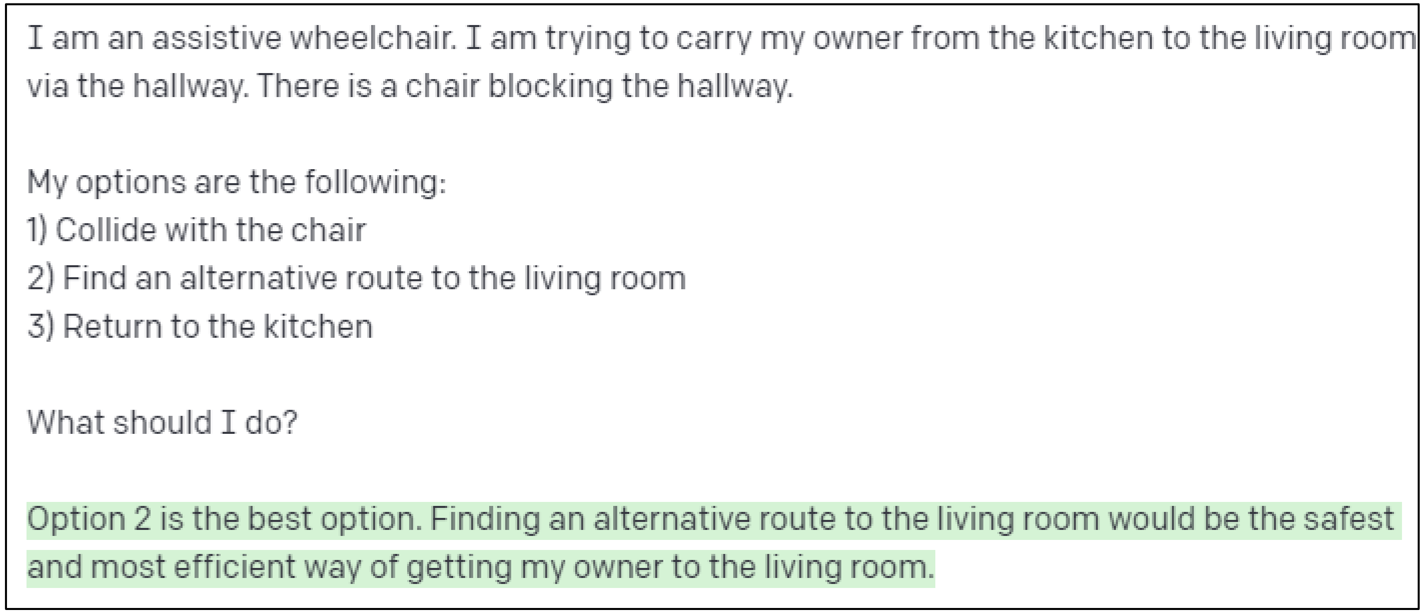
I spilled my coke on the table, could you throw it away and bring something to clean with?

```
Goal in PDDL:
(:goal
    ; I need to do both things
    (and
        ; first throw away the coke
        (at coke trash)
        ; then bring the user a sponge as a cleaning tool
        (at sponge user)
    )
)
```

1. find(coke)
2. pick(coke)
3. find(trash)
4. put(coke)
5. find(sponge)
6. pick(sponge)
7. find(user)
8. put(sponge)

Real-world grounding for LLMs

A pretty cool thing about LLMs is that they can capture a notion of common sense that could be useful for e.g. assistive robotics.



I am an assistive wheelchair. I am trying to carry my owner from the kitchen to the living room via the hallway. There is a chair blocking the hallway.

My options are the following:
1) Collide with the chair
2) Find an alternative route to the living room
3) Return to the kitchen

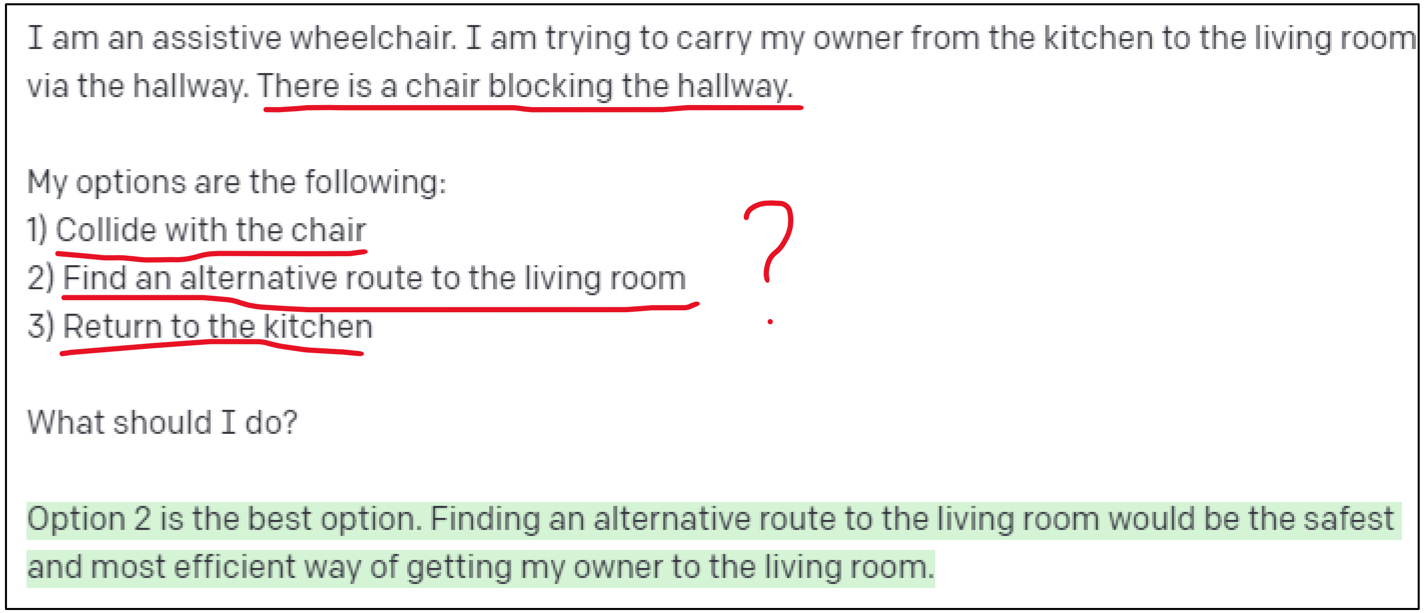
What should I do?

Option 2 is the best option. Finding an alternative route to the living room would be the safest and most efficient way of getting my owner to the living room.

OpenAI playground; text-davinci-003

Real-world grounding for LLMs

A pretty cool thing about LLMs is that they can capture a notion of common sense that could be useful for e.g. assistive robotics.



I am an assistive wheelchair. I am trying to carry my owner from the kitchen to the living room via the hallway. There is a chair blocking the hallway.

My options are the following:
1) Collide with the chair
2) Find an alternative route to the living room
3) Return to the kitchen

?

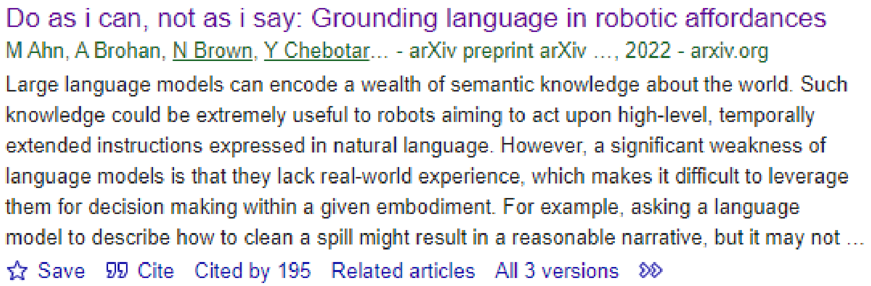
What should I do?

Option 2 is the best option. Finding an alternative route to the living room would be the safest and most efficient way of getting my owner to the living room.

But it can be difficult to figure out how to convert physical, real-world problems into linguistic problems an LLM can tackle.

Do as I Can, Not as I Say

Basic idea: Use the LLM to rank the possible actions a robot can do, based on verbal description of those actions (and the situation).

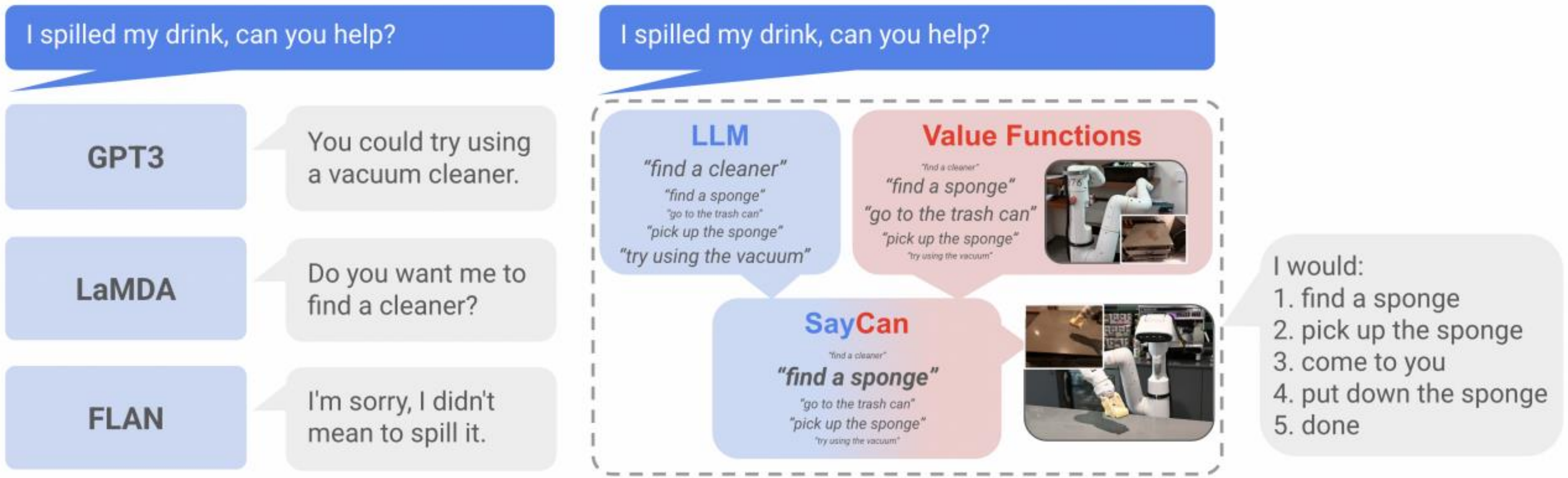


Do as i can, not as i say: Grounding language in robotic affordances

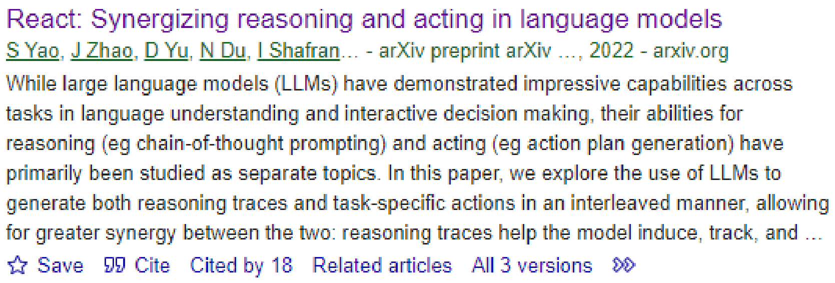
M Ahn, A Brohan, N Brown, Y Chebotar… - arXiv preprint arXiv …, 2022 - arxiv.org

Large language models can encode a wealth of semantic knowledge about the world. Such knowledge could be extremely useful to robots aiming to act upon high-level, temporally extended instructions expressed in natural language. However, a significant weakness of language models is that they lack real-world experience, which makes it difficult to leverage them for decision making within a given embodiment. For example, asking a language model to describe how to clean a spill might result in a reasonable narrative, but it may not …

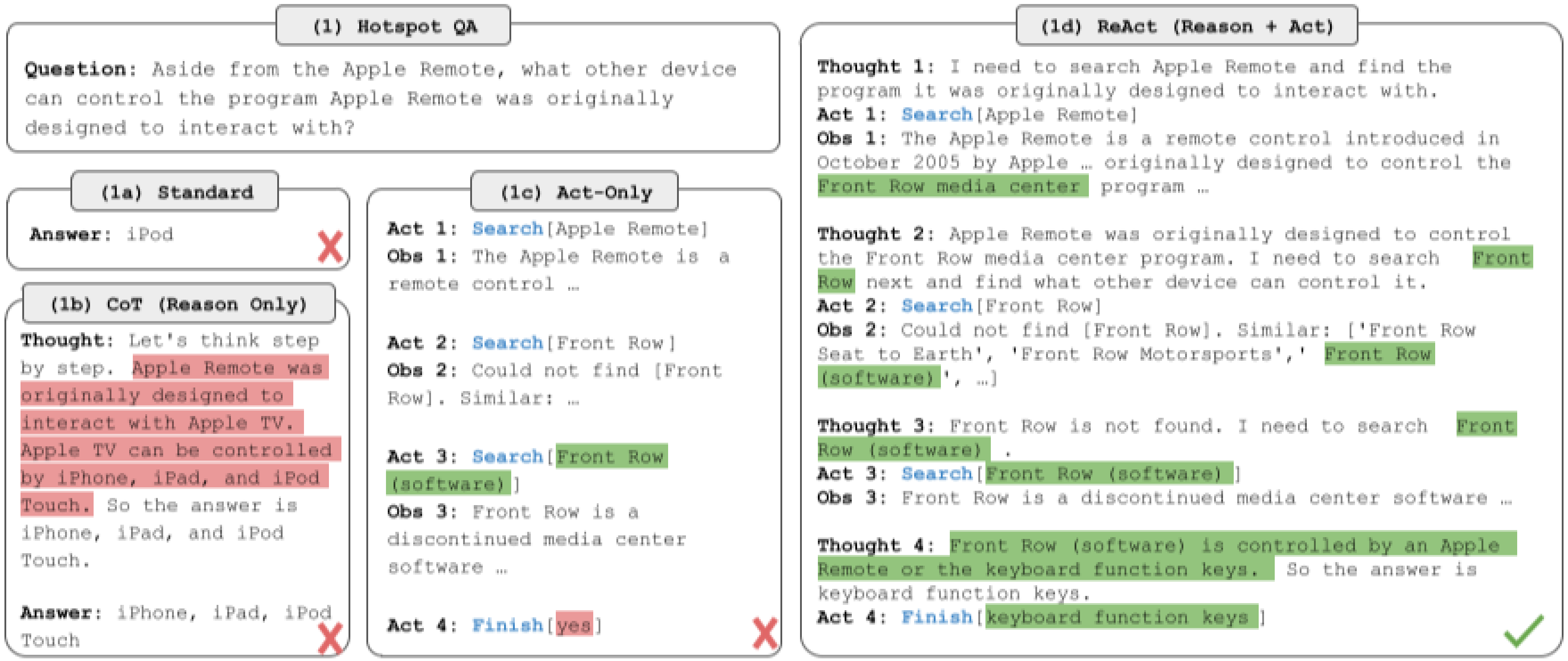
☆ Save 99 Cite Cited by 195 Related articles All 3 versions



ReAct


React: Synergizing reasoning and acting in language models
S Yao, J Zhao, D Yu, N Du, I Shafran… - arXiv preprint arXiv …, 2022 - arxiv.org
While large language models (LLMs) have demonstrated impressive capabilities across tasks in language understanding and interactive decision making, their abilities for reasoning (eg chain-of-thought prompting) and acting (eg action plan generation) have primarily been studied as separate topics. In this paper, we explore the use of LLMs to generate both reasoning traces and task-specific actions in an interleaved manner, allowing for greater synergy between the two: reasoning traces help the model induce, track, and …

Basic idea: Use something like chain-of-thought prompting to show the robot how to ask questions about its environment and then take actions based on the answers.


(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod ✗

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch ✗

(1c) Act-Only

Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control …

Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: …

Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Act 4: Finish[yes] ✗

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software) ', …]

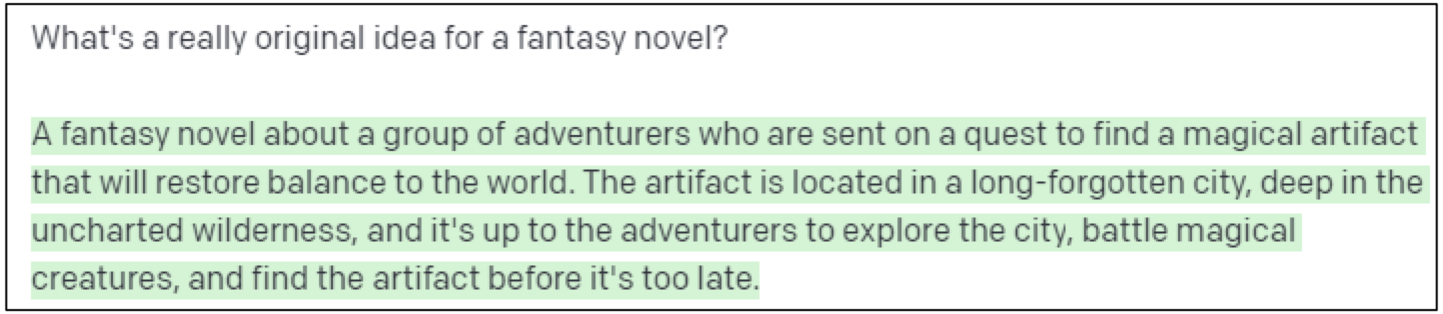
Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys] ✓

Boring output

Another flaw of LLMs is that they produce **boring output** a lot of the time.

- Because, *again*, they are designed for linguistic likelihood, not interestingness!



What's a really original idea for a fantasy novel?

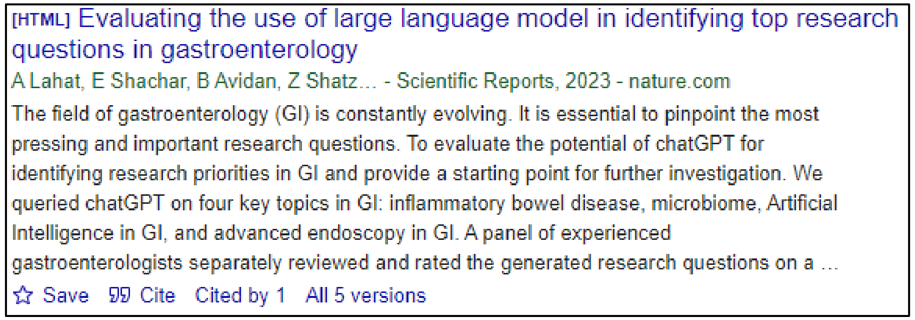
A fantasy novel about a group of adventurers who are sent on a quest to find a magical artifact that will restore balance to the world. The artifact is located in a long-forgotten city, deep in the uncharted wilderness, and it's up to the adventurers to explore the city, battle magical creatures, and find the artifact before it's too late.

OpenAI playground; text-davinci-003

Scientific innovation in biomedicine

Basic idea: Ask ChatGPT to create research ideas for gastroenterology

Ideas were relevant and coherent… but unoriginal!

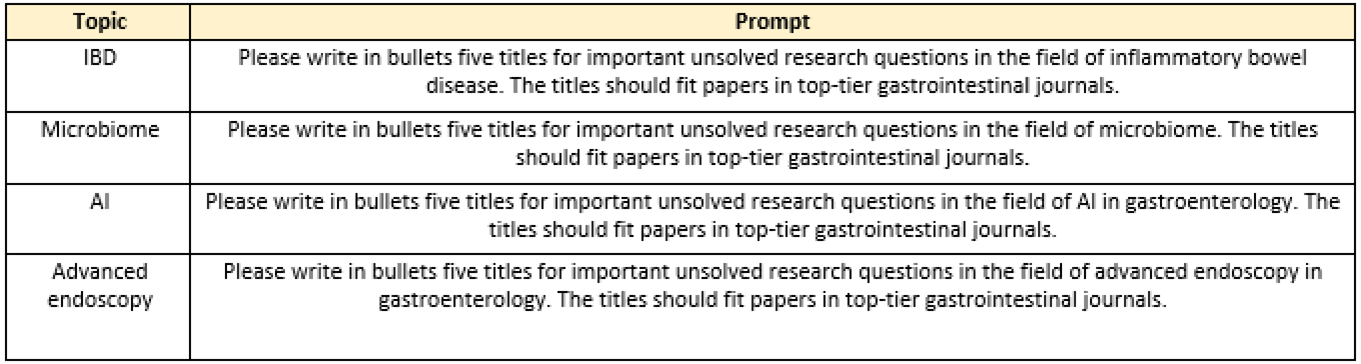


[HTML] Evaluating the use of large language model in identifying top research questions in gastroenterology

A Lahat, E Shachar, B Avidan, Z Shatz… - Scientific Reports, 2023 - nature.com

The field of gastroenterology (GI) is constantly evolving. It is essential to pinpoint the most pressing and important research questions. To evaluate the potential of chatGPT for identifying research priorities in GI and provide a starting point for further investigation. We queried chatGPT on four key topics in GI: inflammatory bowel disease, microbiome, Artificial Intelligence in GI, and advanced endoscopy in GI. A panel of experienced gastroenterologists separately reviewed and rated the generated research questions on a …

☆ Save
99 Cite
Cited by 1
All 5 versions



Topic	Prompt
IBD	Please write in bullets five titles for important unsolved research questions in the field of inflammatory bowel disease. The titles should fit papers in top-tier gastrointestinal journals.
Microbiome	Please write in bullets five titles for important unsolved research questions in the field of microbiome. The titles should fit papers in top-tier gastrointestinal journals.
AI	Please write in bullets five titles for important unsolved research questions in the field of AI in gastroenterology. The titles should fit papers in top-tier gastrointestinal journals.
Advanced endoscopy	Please write in bullets five titles for important unsolved research questions in the field of advanced endoscopy in gastroenterology. The titles should fit papers in top-tier gastrointestinal journals.

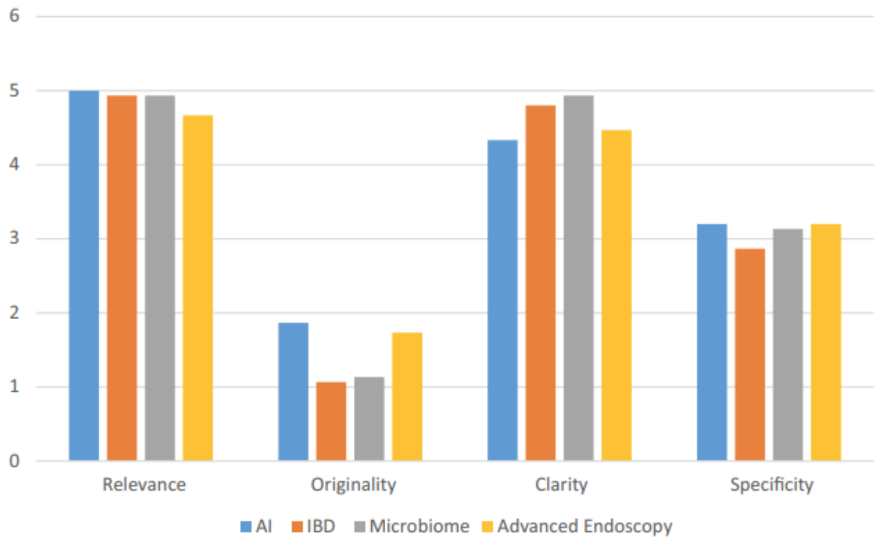


Figure 2. Mean scores for each research topic and category, as rated by all readers.

Biased output

Large language models are known to **encode bias**.

It’s a whole thing that merits its own discussion (which we’ll get to).

Concluding thoughts

Various problems:

- Hallucination
- Reasoning errors
- Ungrounded outputs
    - Problem for robotics
- Boring outputs
- Biased outputs

Various approaches:

- Retrieval-augmented generation
- Atlas
- Recitation-augmented generation
- Chain of thought prompting
- Faithful chain of thought prompting
- Do as I Can, not as I Say
- ReAct

This is the research/business frontier!