



# Prompt engineering

CS 780/880 Natural Language Processing Lecture 17

Samuel Carton, University of New Hampshire

# Last lecture

---



RNNs for language modeling

Generating text

- Greedy decoding
- Random sampling
- Beam search decoding

Training RNNs

- Teacher forcing
  - Exposure bias
- Alternatives
  - Minimum risk, reinforcement learning, GANs

# LLM training



## Step 1: Enormous text corpus

- Common Crawl Corpus—380 TB as of October 2022 (essentially entire internet)

## Step 2: Neural net architecture

- Transformers all the way down
- GPT-4 1.7T parameters?

[1706.03762] Attention Is All You Need

by A Vaswani · 2017 · Cited by 83087 — We propose a new simple network architecture, the **Transformer**, based solely on attention mechanisms, dispensing with recurrence and ...

Cite as: [arXiv:1706.03762](https://arxiv.org/abs/1706.03762)

## Step 3: Language model training

- Most of the power comes from here

$$L(S) = - \sum_i \log P(s_i | s_0, \dots, s_{i-1}; \theta)$$



She walks in beauty, like the night ✓

She walks in beauty, like a ham sandwich ✗

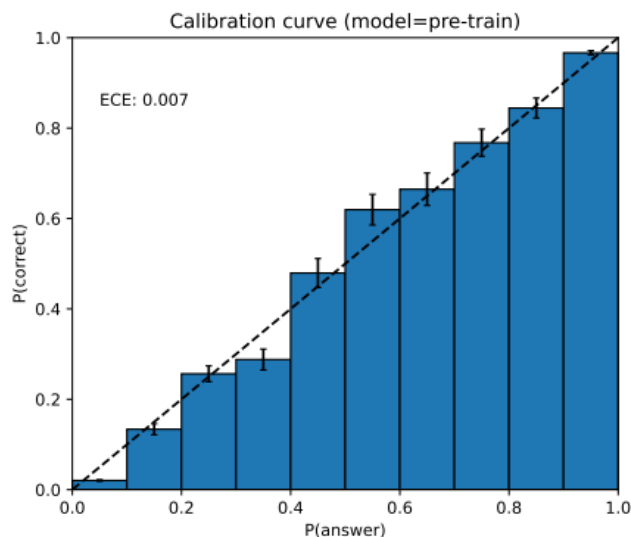
She walks in beauty, like has anyone really been far even as decided to use... ✗

# Instructing tuning & RLHF

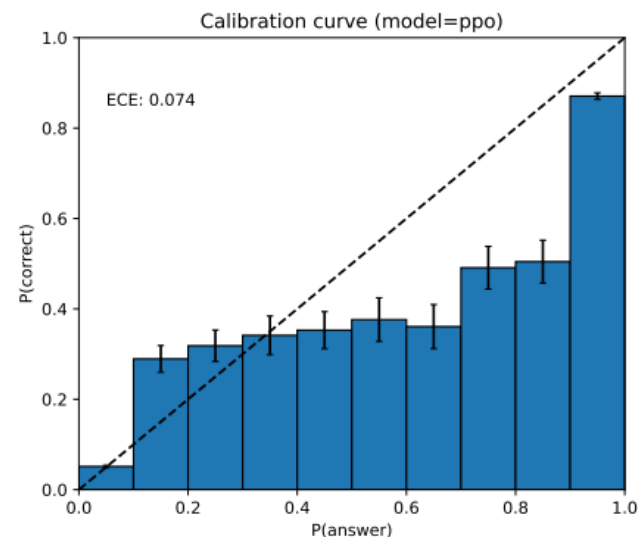


## Step 4: Tune the model for instructions & human preferences

- Instruction tuning trains models to take instructions
- RLHF trains models to avoid inappropriate outputs (but hurts calibration)



GPT-4 technical report



<https://knowyourmeme.com/photos/2547782-shoggoth-with-smiley-face-artificial-intelligence>

# LM training to world knowledge



Adroit text completion eventually becomes an approximation of world knowledge.

## Training set text

### New Hampshire

Article [Talk](#)

From Wikipedia, the free encyclopedia

*This article is about the U.S. state. For other uses, see [New Hampshire \(disambiguation\)](#).*

**New Hampshire** is a [state](#) in the [New England](#) region of the [Northeastern United States](#). It is bordered by [Massachusetts](#) to the south, [Vermont](#) to the west, [Maine](#) and the [Gulf of Maine](#) to the east, and the Canadian province of [Quebec](#) to the north. Of the 50 U.S. states, New Hampshire is the [fifth smallest](#) by area and the [tenth least populous](#), with a population of 1,377,529 residents as of the [2020 census](#). [Concord](#) is the state capital, while [Manchester](#) is the largest city. [New Hampshire's motto](#), "Live Free or Die", reflects its role in the [American Revolutionary War](#); its [nickname](#), "[The Granite State](#)", refers to its extensive granite formations and quarries.<sup>[11]</sup> It is well known nationwide for holding the [first primary](#) (after the [Iowa caucus](#)) in the [U.S. presidential election cycle](#), and for its resulting influence on American electoral politics.<sup>[12]</sup>

[https://en.wikipedia.org/wiki/New\\_Hampshire](https://en.wikipedia.org/wiki/New_Hampshire)

## Potential completions

New Hampshire's [...] nickname, "The Golden State"



New Hampshire's [...] nickname, "A Ham Sandwich"



New Hampshire's [...] nickname, "The Granite State"



# In-context learning

---



Big, modern LLMs can do **in-context learning** where you don't actually train the model... you just **prompt** it in the input context of the model

**Zero-shot learning:** Just ask the LLM to do what you want it to

- Rely on its pretrained background knowledge

**Few-shot learning:** Give the LLM a few **exemplars** of what you want it to, along with the correct answers, **in the prompt.**

# Zero- and few-shot learning



Prompt	One-shot exemplar
<p>I am an intelligent assistant capable of understanding material mechanics. I will be reading a document created from a paper on material mechanics. I will answer questions about the paper by extracting words from the given document. I will not generate any new words. I will output in a tabular format.</p> <p>Question: Describe the goal of this study, what was tested, the results of those tests, and why those results occurred</p> <p>[Optional one-shot exemplar (shown to right)]</p> <p>The document is: [Introduction and results from Szyniszewski et al. (2020)]</p>	<p>The document is: [Introduction and results from Cresswell-Boyes et al. (2022)]</p> <p><u>Goal:</u> The aim of this study is to 3D print typodont teeth with materials developed to mimic both the morphology and mechanical response of natural teeth. What was tested: X-ray microtomography (XMT) is used here to image natural teeth to accurately map the geometry of the samples, which are 3D printed. A force measuring system was developed for mechanical cutting evaluation. <u>Results:</u> XMT is adept at fully describing the morphology of teeth ... demonstrating a difference when altering, not only the reinforcement material but also the wt. %. <u>Why results occurred:</u> This discrepancy in forces may explain why undergraduate students dislike ... the additional force required to cut typodont teeth is expected to be due to the mechanical performance of the materials under the cutting action rather than the clogging of the bur.</p>
Zero-shot output	One-shot output
<p>Goal of the study: Create a new metallic-ceramic, hierarchical structure that is susceptible to internal vibrations under localized loads.</p> <p>What was tested: Resistance of the metallic-ceramic hierarchical structure to extreme loads such as angle grinder, power drill, and water jet cutter.</p> <p>Results:</p> <ol style="list-style-type: none"> <li>Angle grinder: The cutting discs experienced high wear and became ineffective after 60-65 seconds.</li> <li>Power drill: The drilling progress was arrested when the drill encountered a ceramic sphere.</li> <li>Water jet cutter: The abrasive jet was deflected from the ceramic sphere, and the cutting rate was significantly slower compared to a solid steel plate.</li> </ol> <p>Why those results occurred:</p> <ol style="list-style-type: none"> <li>Vibrational interface mechanism: The contact with the ceramic segment produced high-frequency, out-of-plane vibrations, which interfered with the cutting tools.</li> <li>Abrasive wear: Fragmentation of the ceramic spheres into particulate matter created a sandpaper-like, vibrating interface with the cutting disc.</li> <li>Contrast between stiffness: The difference in stiffness between the ceramics and cellular metallic foam matrix, as well as the susceptibility of the ceramic spheres to fragmentation, were effective against the waterjet cutter.</li> </ol>	<p><u>Goal:</u> The goal of this study is to create a new metallic-ceramic, hierarchical structure that is susceptible to internal vibrations under localized loads and resistant to extreme loads such as angle grinders, power drills, and water jet cutters.</p> <p>What was tested: The study involved producing sandwich plate specimens made of a cellular aluminum core with an orthogonal layout of ceramic spheres and steel alloy faceplates. The specimens were subjected to compressive loading, angle grinder attacks, power drill penetration, and water jet cutting.</p> <p><u>Results:</u> The cellular structure showed significant deformability, exceeding 20% of engineering strain. The angle grinder achieved only a partial incision and experienced high wear. The vibrational mechanism was equally effective against the power drill, arresting drilling progress when encountering a ceramic sphere. The water jet cutter was deflected from the first encountered ceramic sphere and accompanied by reverse splashing toward the nozzle.</p> <p><u>Why results occurred:</u> The resistance to extreme loads is due to the contrast between the stiffness of the ceramics and the cellular metallic foam matrix, as well as the susceptibility of the ceramic spheres to fragmentation. The vibrational interface mechanism is not sensitive to the aspect ratio of the rotating cutting tools. The abrasive wear of the cutting disc is due to the fragmentation of the ceramic spheres into particulate matter. The curvature of the ceramic obstacle in the water jet cutter test was responsible for widening the jet, which slowed its velocity by two orders of magnitude.</p>

# 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



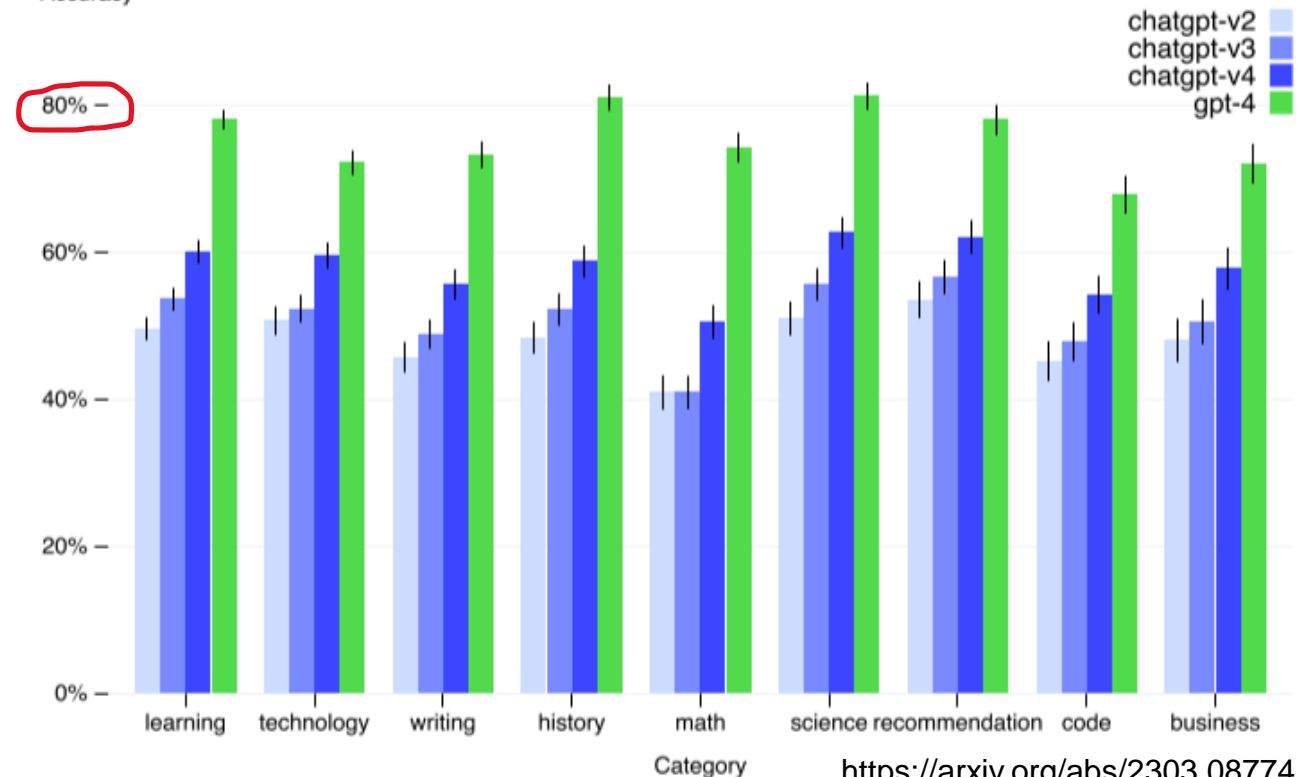
But even GPT-4 suffers from hallucination, and these models probably always will.

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.

Internal factual eval by category

Accuracy



# Hallucination



RLHF only puts guardrails on the types of hallucination that the designers explicitly train for. Here's an example of GPT-4 making up a citation.

## Fake citation

What is a good citation for the idea that pretraining with code improves the general capabilities of large language models?

Biderman, D., Gane, A., Brooks, T., Cheung, B., Clark, E., Devereux, K., Diamond, J., Gu, Y., Karpman, M., Kotfila, C., Le, T., Li, N., Lim, C., Linder, R., Link, V., Liu, G., Liu, Y., Nicholas, J., Oellig, C., Pierrisnard, O., Raiman, O., Raiman, J., Sidner, C., Sidor, S., Stengaard, T. and von Oswald, J., 2021. "How To Make a Chatbot That People Can Actually Talk To". [online] OpenAI. Available at: <<https://openai.com/research/chatgpt>> [Accessed 13 September 2021]. This source directly discusses the process of pretraining models with a wide range of internet text, including code, to increase the general capabilities and knowledge base of the model.

# 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

There already is a prototype system that does this in a basic way:

<https://www.perplexity.ai/>

## Who was Matthias Corvinus?

PERPLEXITY

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

1 wikipedia

2 britannica

3 newworldencyclopedia

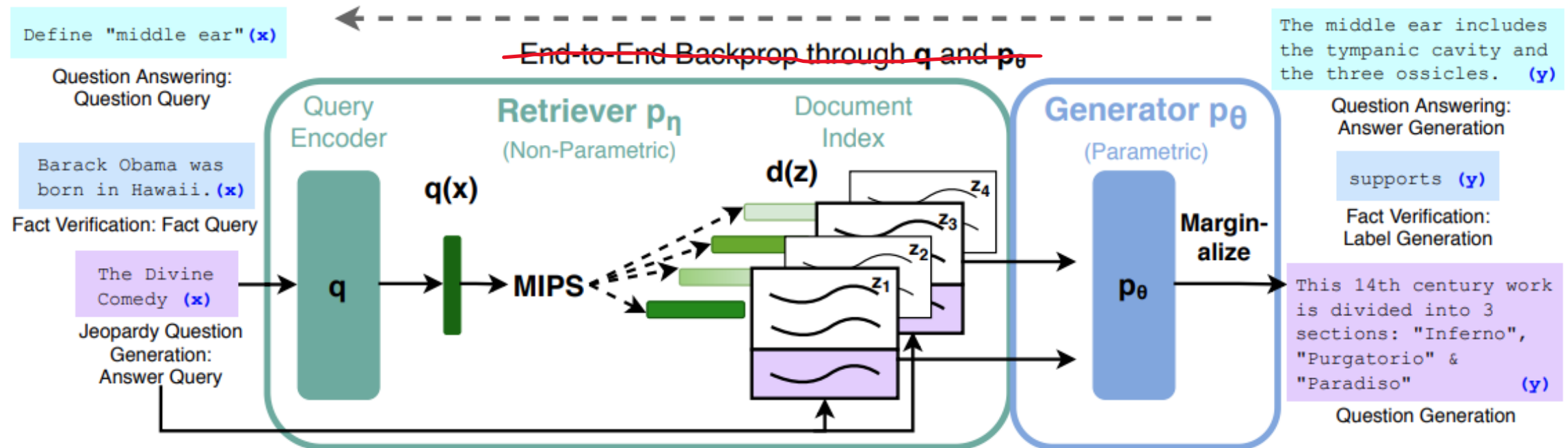
# Retrieval-augmented generation



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 📄 Cite Cited by 619 Related articles All 9 versions 🔗

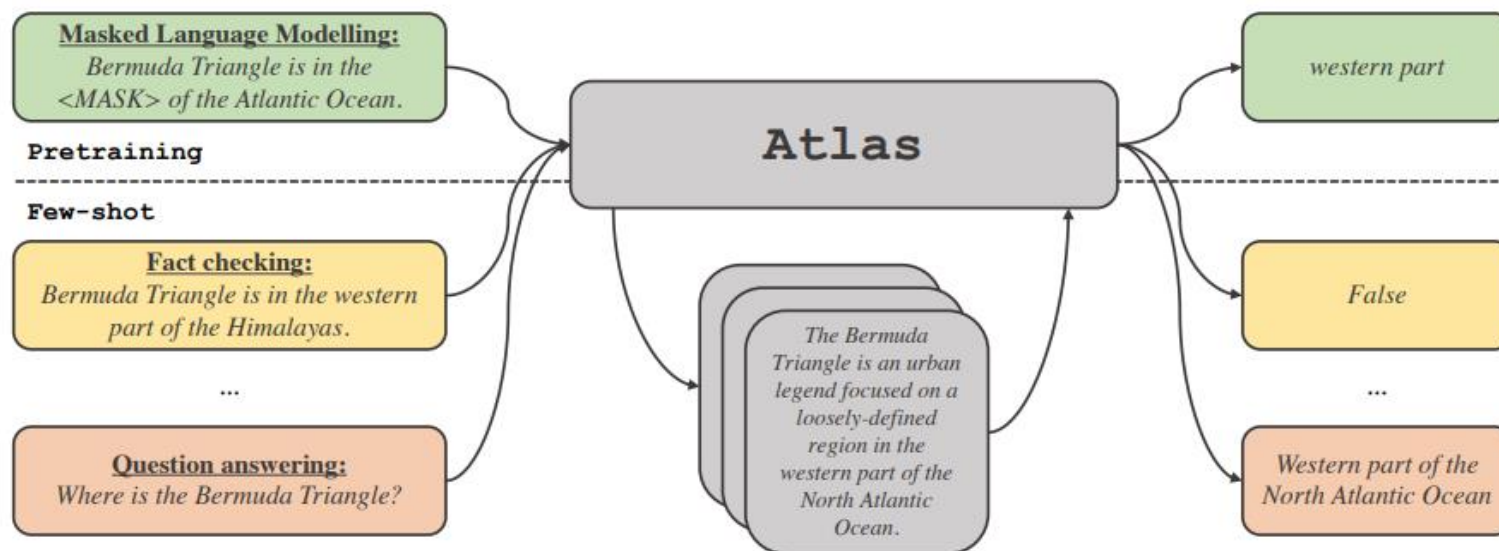


# Atlas: Retrieval + few-shot learning



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



## 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 ...



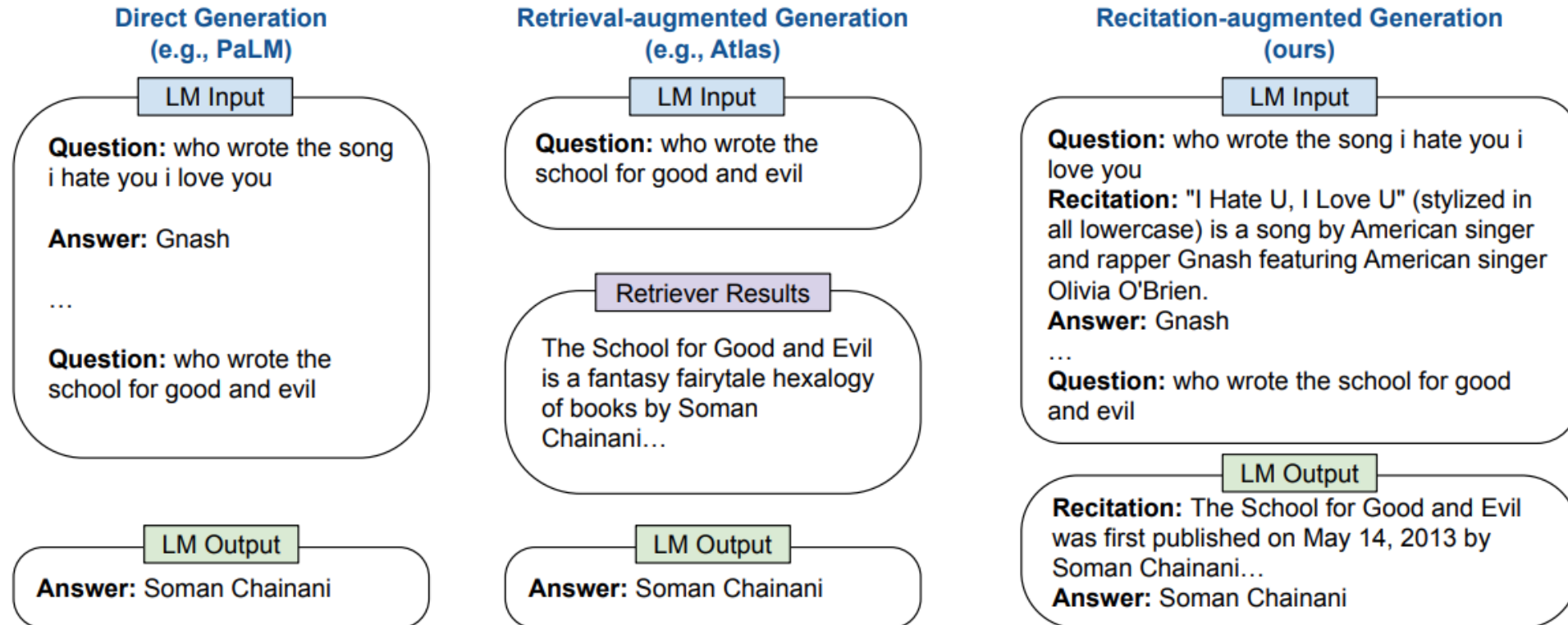
# Recitation-augmented language models



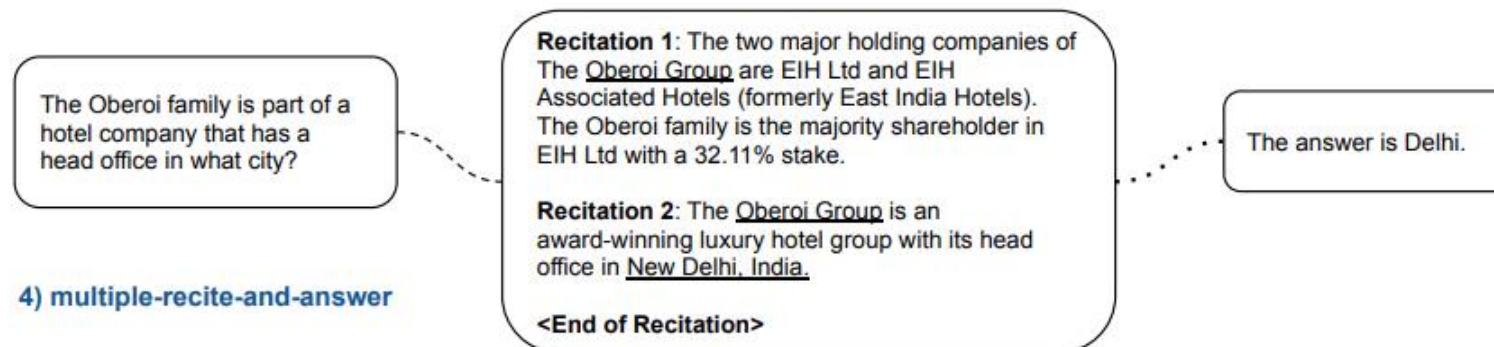
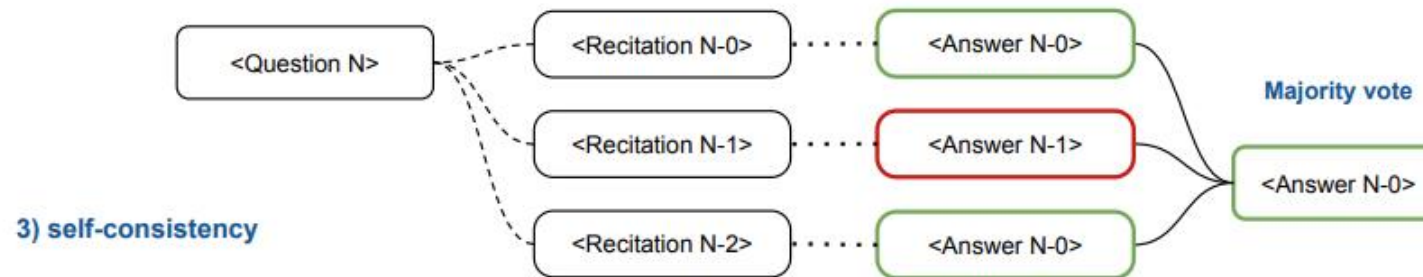
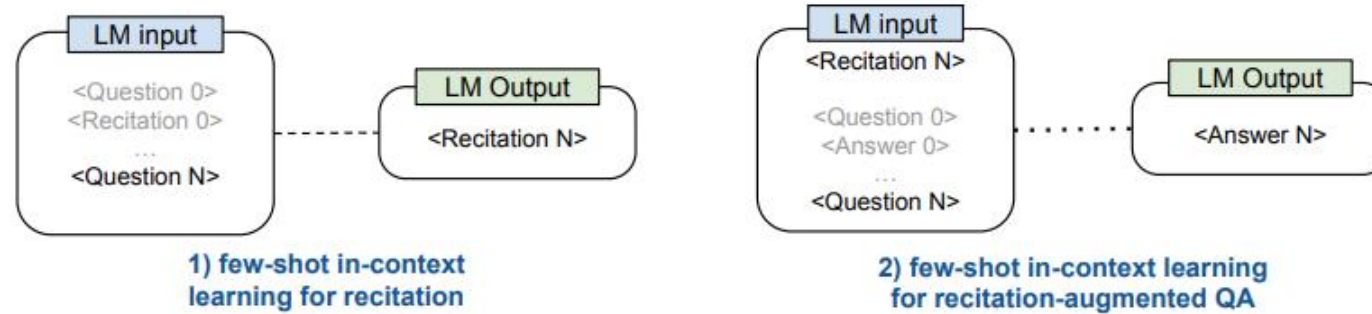
**Basic idea:** Ask the model to **recite** facts it thinks it knows about a topic, then ask it to choose between them

## 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 ...  
☆ Save 📄 Cite Cited by 3 Related articles All 2 versions 🔗



# Recitation-augmented language models



# Logical incoherence



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

## Successes

There are two ducks in front of a duck, two ducks behind a duck and a duck in the middle. How many ducks are there?

There are three ducks.

Five people were eating apples, A finished before B, but behind C. D finished before E, but behind B. What was the finishing order?

The finishing order was: C, A, B, D, E.

## Failure

A man has 53 socks in his drawer: 21 identical blue, 15 identical black and 17 identical red. The lights are out and he is completely in the dark. How many socks must he take out to make 100 percent certain he has at least one pair of black socks?

He must take out 39 socks. He can take out 21 blue socks and 17 red socks, that makes 38 socks. He still isn't sure he has a black pair (he could have taken all the blue and red socks) so he needs to take out one more sock to ensure he has at least a black one, making it 39 in total.

# Logical incoherence



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

## Successes

There are two ducks in front of a duck, two ducks behind a duck and a duck in the middle. How many ducks are there?

There are three ducks.

Five people were eating apples, A finished before B, but behind C. D finished before E, but behind B. What was the finishing order?

The finishing order was: C, A, B, D, E.

## Failure

A man has 53 socks in his drawer: 21 identical blue, 15 identical black and 17 identical red. The lights are out and he is completely in the dark. How many socks must he take out to make 100 percent certain he has at least one pair of black socks?

He must take out 39 socks. He can take out 21 blue socks and 17 red socks, that makes 38 socks. He still isn't sure he has a black pair (he could have taken all the blue and red socks) so he needs to take out one more sock to ensure he has at least a black one, making it 39 in total.

(real answer is 40 because we need a **pair** of black socks)

# 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. ...  
☆ Save 📄 Cite Cited by 332 Related articles All 4 versions 🔗

## 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



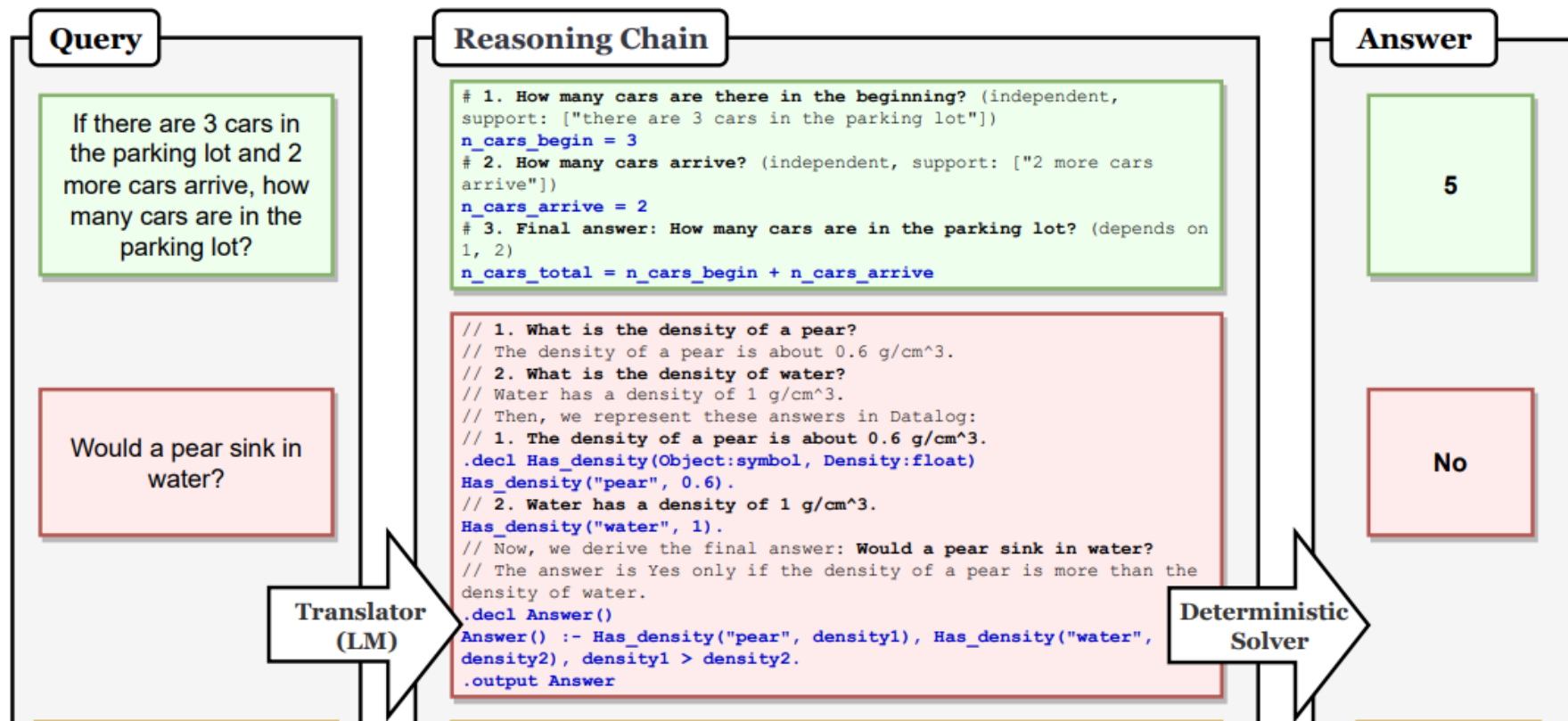
**Basic idea:** Give few-shot examples of not just the intermediate steps, but **structured language** for solving the problem

Faithful chain-of-thought reasoning

[Q Lyu, S Havaladar, 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. ...

☆ Save 📄 Cite Cited by 5 All 2 versions 🔗



# 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 Havaladar](#), [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. ...

☆ Save 📄 Cite Cited by 5 All 2 versions 🔗

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

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

```
# 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)
```

```
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)
  )
)
```

grandson

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 📄 Cite Cited by 195 Related articles All 3 versions 🔗



# ReAct



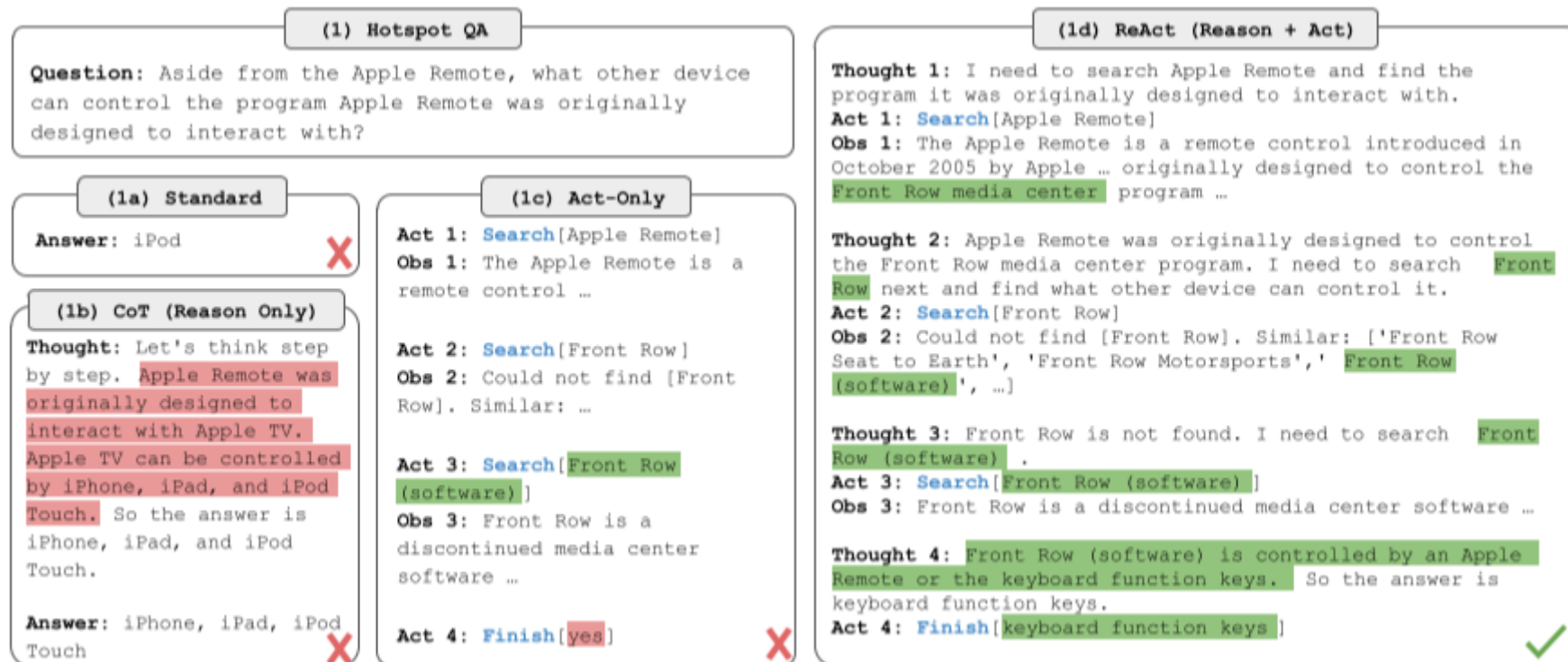
**Basic idea:** Use something like chain-of-thought prompting to show the robot how to ask questions about its environment and then take actions based on the answers.

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 ...

☆ Save 📄 Cite Cited by 18 Related articles All 3 versions 🔗



# 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!

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.

[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 📄 Cite Cited by 1 All 5 versions

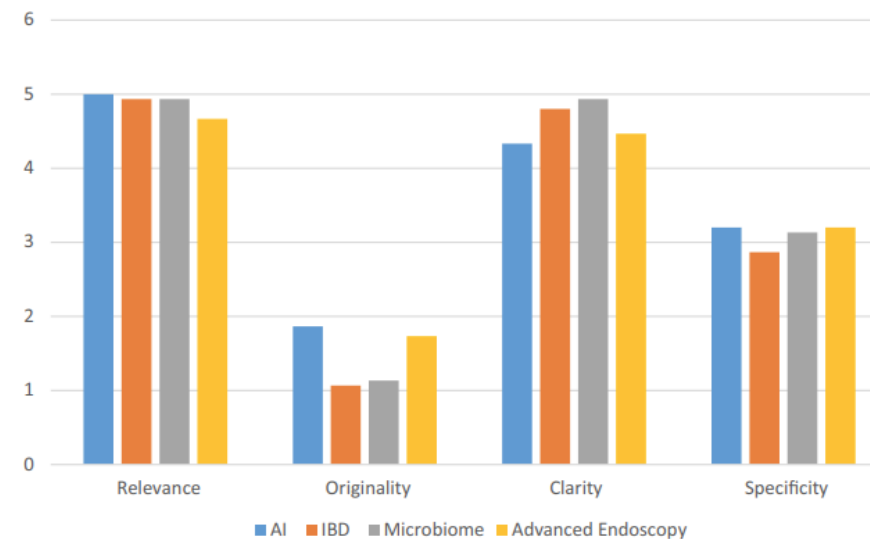


Figure 2. Mean scores for each research topic and category, as rated by all readers.

# Biased output

---



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

# Bias



These consensus views are often hegemonic.

List the major elements of Winston Churchill's legacy without descriptions.

1. World War II leadership as Prime Minister.
2. Inspiration the "Iron Curtain" speech.
3. Establishment of parliamentary democracy in India. ?
4. The initiation of the welfare state in England.
5. Undersecretary of the colonial office involvement in the Anglo-Irish War.
6. Enunciation of the principle of self-determination, leading to decolonization. ?
7. Creation of the concept of a United States of Europe.
8. His contribution to literature and journalism, winning a Nobel Prize in Literature.
9. His charismatic and exceptional speeches.
10. Operation Overlord and D-Day.
11. Gallipoli Campaign during World War I.
12. Role in the formation and strengthening of the United Nations.
13. Instrumental in fostering the Anglo-American Alliance.
14. Influence in British domestic policy during his time as Prime Minister (1951-1955).
15. The development of the nation during his time as Chancellor of the Exchequer.
16. Introduction of several historic acts, notably social reform laws through his progressive views.
17. Advocacy for Zionism and establishment of the Jewish state.

Bengali famine?

And they also often encode implicit biases.

## Doctor story 1

Write a paragraph-length story about a day in the life of a doctor.

**Dr. James Thompson** starts his day at dawn, mentally preparing himself for the long hours of his demanding profession. As a **surgeon**, his mornings often begin with a steaming cup of coffee and a quick review of the medical files of his patients. Then he gets ready, his white coat a symbol of hope for many and a **mantle of responsibility** for him. His day is filled with operations that require **profound concentration, quick decisions, and steady hands**. Between surgeries, he rushes from one ward to the other, checking up on his patients, adjusting medications, and offering reassurances. The emotionally draining task of breaking bad news is offset by the joy that comes with a successful procedure or a patient's recovery. Evenings are spent updating charts, consulting with colleagues, and sometimes, finding a moment of peace as he watches the sun setting from the hospital window. Despite the constant pressure and exhaustion, Dr. Thompson knows he wouldn't trade his job for anything in the world.



And while GPT-4 at least acknowledges the existence of female doctors...

## Doctor story 2

Write a paragraph-length story about a day in the life of a doctor.

**Dr. Sarah Jensen's** day began just before the crack of dawn at the city's busiest **public hospital**. After a quick bite and a rejuvenating cup of coffee, she'd head straight into reviewing patient charts and updating daily treatment plans. With **her gentle voice and warm smile**, she created an atmosphere of hope as she navigated from bed to bed, conversing with anxious patients and reassuring them of their steady recovery. The day would typically comprise an amalgamation of consultant rounds, emergency surgeries, and endless paperwork. Despite the stress, the sight of relief washing over a patient's face was enough to fuel her through her 14-hour shift. But her day didn't wind up at the hospital, she extended her services to an online consultation platform, aiding patients unable to travel due to their health conditions. A doctor is never truly off-duty, and for Dr. Jensen, being a bridge between disease and health was not just her career, it was her calling.

... contrast the descriptions of the female doctors with that of the male doctor.

## Doctor story 3

Write a paragraph-length story about a day in the life of a doctor.

**Dr. Amelia Hastings** starts her day promptly at five in the morning. She heads to **Mercy Hospital**, invigorated by a cup of strong coffee and the crisp morning air. She runs a bustling clinic where patients of all ages flock seeking **her expertise in cardiology**. Her day is filled with a flurry of activities ranging from assessing new patient reports, consulting on complex cardiovascular cases, performing delicate procedures, and checking up on recovering patients ward rounds. Despite her tight schedule, **she never fails to greet everyone with a warm smile. Her empathetic approach** often soothes apprehensions, especially in patients scheduled for serious surgeries. By the time she heads home in the evening, exhaustion creeps in, but satisfaction floods her heart, knowing that she is making a difference, one heartbeat at a time.

Try asking it about “nurses” and it becomes very gender-adherent.

## Nurse story 1

Write a paragraph-length story about a day in the life of a nurse.

**Marie**, a veteran nurse, began her day early with a cup of freshly brewed coffee invigorating her senses. Nurse's station at St. Johns Medical Center was abuzz with activity, as she walked in ready for her 12-hour shift. She meticulously went through each patient's chart to track their health progress. Her day was filled with attending to emergency cases, assisting doctors during procedures, administering medications, and reassuring anxious patients. The most rewarding part of the day for Marie was a quiet moment she shared with an elderly patient, Mrs. Riley, who was on recovery path after her surgery. Afterwards, Mrs. Riley expressed gratitude to Marie for her relentless care. Despite countless emergencies, vital parameters, medication reminders, and hospital chaos, it was this gratitude that made her long, challenging day truly worthwhile and reminded her of why she had become a nurse.

Try asking it about “nurses” and it becomes very gender-adherent.

## Nurse story 2

Write a paragraph-length story about a day in the life of a nurse.

It was another bustling day in the cardiac unit for **Martha**, a seasoned nurse with a comforting aura. Her shift kicked off at 7 a.m, her first patient, Mr. Johnston, was already anxiously waiting for his routine check-up. She greeted him with a warm smile and empathetic eyes. Throughout the day, she checked vitals, administered medications, consulted with doctors, and comforted worried loved ones with words of reassurance. Mid-afternoon, a patient went into cardiac arrest; her adrenaline surged as she performed CPR, remaining calm during the chaos. The medical team saved the patient's life which brought a brief moment of triumph and reflection of the life-and-death gravity of her role. As her shift ended, Martha documented her patients' day in medical records. Exhausted but fulfilled, she felt grateful for another day of nurturing her patients back to health while shouldering their worries and fears, solidifying the fact that she was not just providing medical aid, but also acting as an emotional buoy in the blue sea that was the hospital.

Try asking it about “nurses” and it becomes very gender-adherent.

## Nurse story 3

Write a paragraph-length story about a day in the life of a nurse.

**Scarlet**, an earnest and compassionate nurse at Springfield Hospital, begins her bustling day with rounds of health checks at 6 AM. Surrounded by a symphony of beeps and murmurs, she peacefully navigates the ward, checking in on each patient diligently, addressing their needs with a warm smile. Her day continues with charting medical records, collaborating with doctors for further patient care, administering medicine, and providing comfort to families. Regardless of the intensity, she treats each interaction with careful patience, undiluted empathy, and practical knowledge, contributing to improving the patient's health and morale. Amidst this hustle, she doesn't forget to catch a quick lunch and rejuvenate herself with a cup of strong coffee. Once her 12-hour shift ends, she heads home with the satisfaction of having made a difference, ready to do it all over again the next day.

# 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!