



Question Answering & Factual Knowledge

Yu Meng

University of Virginia
yumeng5@virginia.edu

Oct 25, 2024



Overview of Course Contents

- Week 1: Logistics & Overview
- Week 2: N-gram Language Models
- Week 3: Word Senses, Semantics & Classic Word Representations
- Week 4: Word Embeddings
- Week 5: Sequence Modeling and Neural Language Models
- Week 6-7: Language Modeling with Transformers (Pretraining + Fine-tuning)
- Week 8: Large Language Models (LLMs) & In-context Learning
- **Week 9-10: Reasoning, Knowledge, and Retrieval-Augmented Generation (RAG)**
- Week 11: LLM Alignment
- Week 12: Language Agents
- Week 13: Recap + Future of NLP
- Week 15 (after Thanksgiving): Project Presentations



(Recap) Reasoning: Overview

- **Reasoning** (rough definition): perform deductive, inductive, commonsense, or logical reasoning via generating or analyzing text with language models
- Deductive reasoning: draw specific conclusions from general principles or premises
 - E.g.: “All humans are mortal” + “Socrates is a human” => “Socrates is mortal”
- Inductive reasoning: make generalizations based on specific observations
 - E.g.: “The sun has risen in the east every day” => “The sun will rise in the east tomorrow”
- Commonsense reasoning: rely on world knowledge or commonsense understanding to make predictions or answer questions
 - E.g.: “If I drop a ball, what will happen?” => “It will fall”
- Mathematical/logical reasoning: follow specific rules or procedures to arrive at a correct answer
 - E.g.: “If 3 apples cost \$6, how much do 5 apples cost?” => “\$10”



(Recap) Standard vs. CoT Prompting

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?

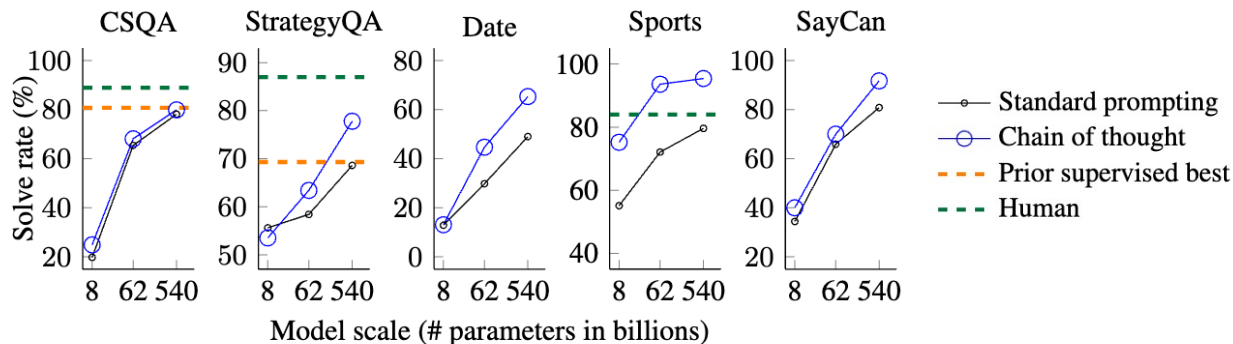
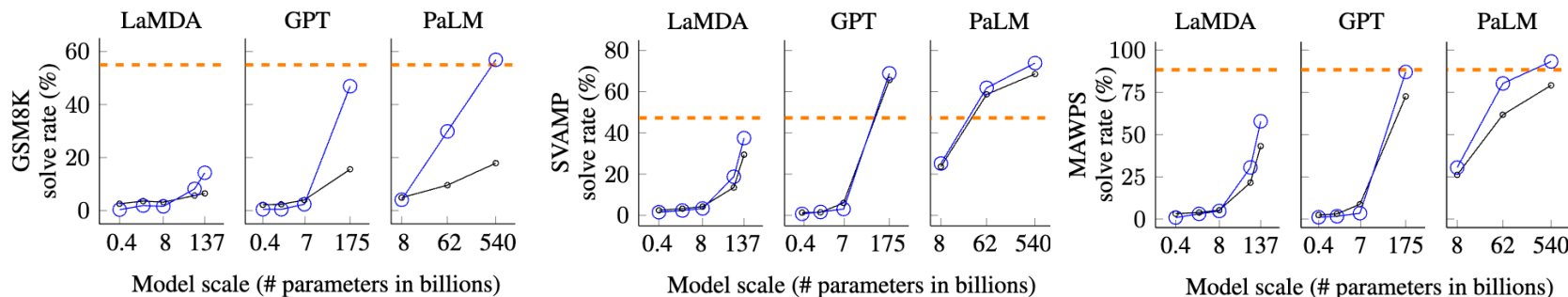
Model Output

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. ✅



(Recap) Standard vs. CoT Prompting

CoT prompting is especially effective for large models





(Recap) CoT Can Be Triggered Zero-shot

Just add “Let’s think step by step” at the beginning of the answer

(a) Few-shot

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) *The answer is 8.* ❌

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) *8* ❌

(b) Few-shot-CoT

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) *The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4.* ✓

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let’s think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓

(Recap) CoT Demo

Join at
slido.com
#1886 206



No-CoT prompt: How many 'r' letters are there in the following word: strawberry? Answer without reasoning steps

llama-3.1-70b-instruct

Expand to see the descriptions of 75 models

Scroll down and start chatting

How many 'r' letters are there in the following word: strawberry? Answer without reasoning steps

There are 2 'r' letters in the word "strawberry". **Wrong result**

CoT prompt: How many 'r' letters are there in the following word: strawberry? Let's think step by step

llama-3.1-70b-instruct

Expand to see the descriptions of 75 models

Scroll down and start chatting

How many 'r' letters are there in the following word: strawberry? Let's think step by step

Step 1: Write down the word
The word is: strawberry

Step 2: Break down the word into individual letters
Here are the individual letters: s-t-r-a-w-b-e-r-r-y

Step 3: Count the number of 'r' letters
Let's count the 'r' letters:

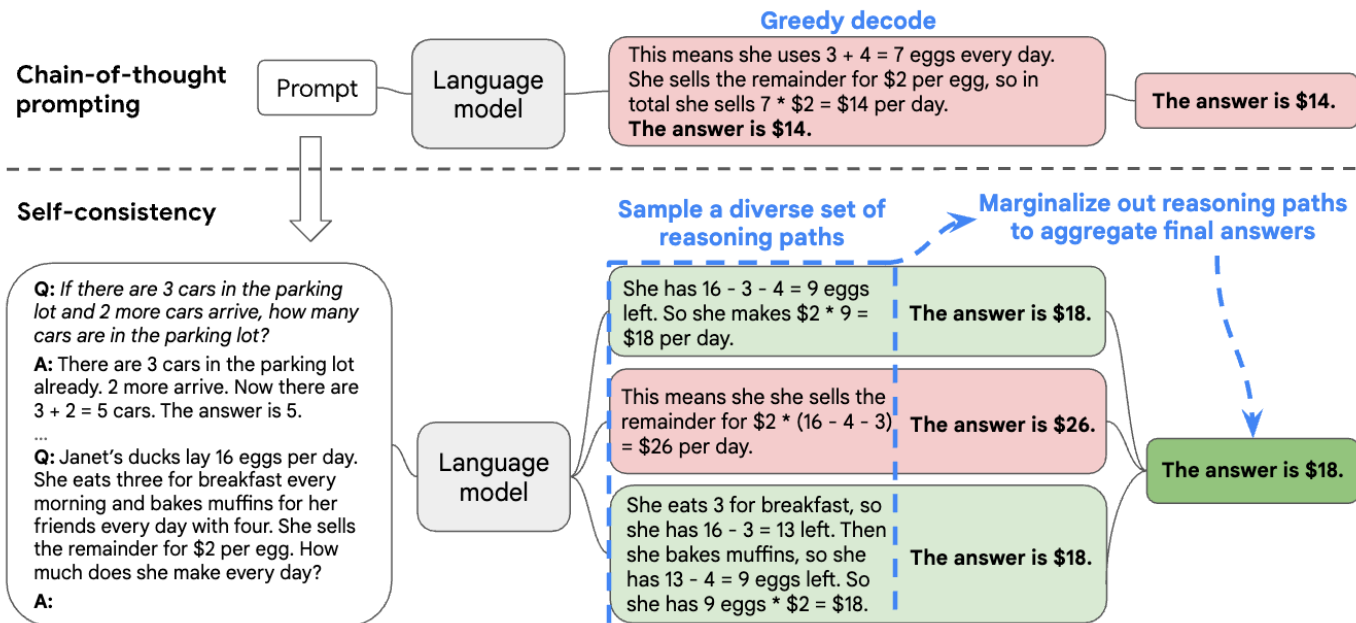
1. r (first 'r')
2. r (second 'r')
3. r (third 'r')

There are 3 'r' letters in the word "strawberry". **Correct result**



(Recap) Self-consistency CoT

Intuition: if multiple different ways of thinking lead to the same answer, one has greater confidence that the final answer is correct



Agenda

- Question Answering (QA)
- Parametric Knowledge in LMs
- Hallucination
- Non-parametric Knowledge

Join at
slido.com
#1886 206





Introduction to Question Answering

- **Question Answering (QA):** build systems that can automatically answer questions posed by humans in natural language
- Categorization by application domain: closed-domain vs. open-domain QA
- **Closed-domain QA:** answer questions within a specific domain
 - Example: medical, legal, technical fields
 - Models are trained on specialized knowledge to be highly accurate within their domain
- **Open-domain QA:** answer questions from any domain
 - Typically rely on vast (external) knowledge sources like the web or large text corpora
 - Most LLM applications consider open-domain QA settings



Introduction to Question Answering

- **Question Answering (QA):** build systems that can automatically answer questions posed by humans in natural language
- Categorization by modeling approach: extractive vs. abstractive QA
- **Extractive QA:** output a span of text extracted directly from a given context
 - A natural language understanding task (reading comprehension)
 - Example: context: “The human brain contains approximately 86 billion neurons” Q: “How many neurons are in the human brain?” A: “86 billion”
 - Can be done with encoder-only LMs (e.g., BERT)
- **Abstractive QA:** synthesize the answer in its own words (rephrasing/summarizing)
 - Example: context: “Albert Einstein published his theory of special relativity which introduced the famous equation $E=mc^2$, which relates energy (E) to mass (m) and the speed of light (c)” Q: “What did Einstein contribute to physics?” A: “Einstein made significant contributions to the theory of special relativity which established the relationship between energy and mass”
 - Need to use a generative LM (e.g., GPT)



Introduction to Question Answering

- **Question Answering (QA):** build systems that can automatically answer questions posed by humans in natural language
- Categorization by access to external source: closed-book vs. open-book QA
- **Closed-book QA:** answer questions without access to any external information
 - Accuracy depends heavily on how well the training data covered the relevant information
 - Similar to a human answering a question from memory without looking anything up
- **Open-book QA:** can access external knowledge source to answer the questions
 - Typically using retrieval from reliable external sources that contain
 - Similar to a human answering a question by looking it up in a book or online resource

Agenda

- Question Answering (QA)
- Parametric Knowledge in LMs
- Hallucination
- Non-parametric Knowledge

Join at
slido.com
#1886 206





Prompting LMs: Parametric Knowledge

- LMs have learned from a lot of facts in their pretraining data
- LMs can be directly prompted to generate answers to factoid questions (Closed-book QA setting)
- Example:

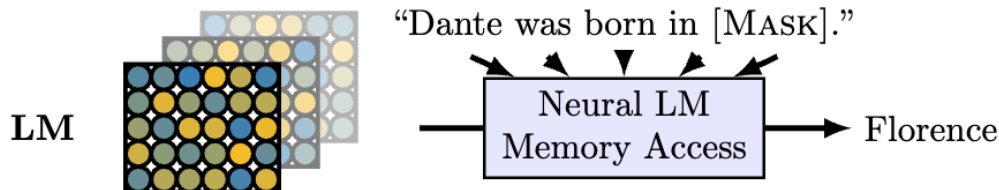
$P(w|Q: \text{Who wrote the book "The Origin of Species"}? A::)$ prompt

- Since prompting LLMs only relies on the information stored within the parameters of the model itself, this kind of knowledge is called **parametric knowledge**



Language Model as Knowledge Bases

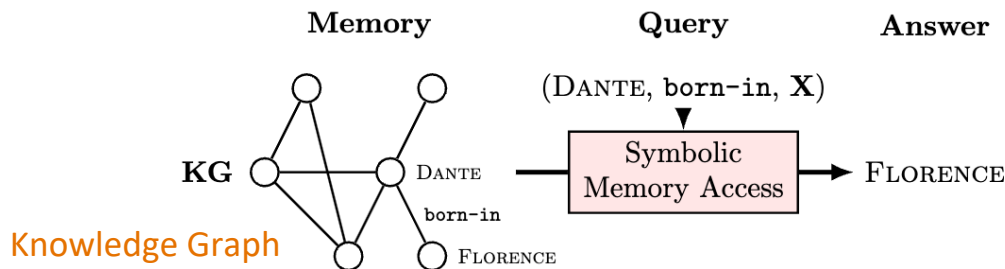
- **Acquisition:** LM's knowledge is derived from the vast amount of pretraining data
- **Access:** information is accessed through natural language prompts
- **Update/maintenance:** re-training/fine-tuning the model with new data
- **Pros:**
 - Handle a wide range of natural language queries with contextual understanding
 - Generalize to unseen queries not seen during training
- **Cons:**
 - May produce incorrect/outdated information
 - Lack interpretability/transparency





(Real) Knowledge Bases

- **Acquisition:** manually constructed by human annotators
- **Access:** information is accessed through queries in specific formats
- **Update/maintenance:** adding/modifying/deleting entries (incrementally) by humans
- **Pros:**
 - Precise & verifiable
- **Cons:**
 - Not able to handle natural language
 - Require massive human efforts to construct & maintain



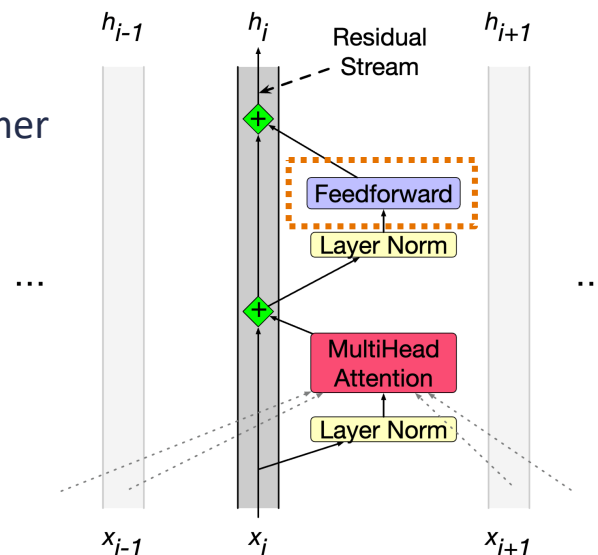


Feedforward Parameters in Transformer

- FFN in Transformer is a 2-layer network (one hidden layer, two weight matrices)

$$\text{FFN}(\mathbf{x}_i) = \text{ReLU}(\mathbf{x}_i \mathbf{W}_1) \mathbf{W}_2$$

- FFN constitutes $\sim 2/3$ of the total parameters of Transformer





Feedforward Parameters Are Neural Memories

Viewing FFN as key-value memories

$$\text{FFN}(\mathbf{x}_i) = \text{ReLU}(\mathbf{x}_i \mathbf{W}_1) \mathbf{W}_2$$



$$\mathbf{x}_i \in \mathbb{R}^{d_1}$$

$$\text{FFN}(\mathbf{x}_i) = \text{ReLU}(\mathbf{x}_i \mathbf{K}) \mathbf{V}$$

$$\mathbf{K} \in \mathbb{R}^{d_1 \times d_2}$$

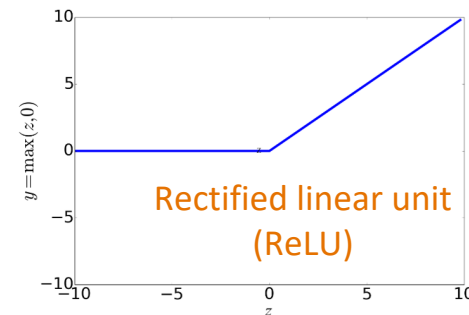
$$\mathbf{V} \in \mathbb{R}^{d_2 \times d_1}$$

key vectors (column vectors in \mathbf{K}) act as **pattern detectors** over the input sequence

value vectors (row vectors in \mathbf{V}) represent **distributions over the output vocabulary**

$$\text{FFN}(\mathbf{x}_i) = \sum_{j=1}^{d_2} \text{ReLU}(\mathbf{x}_i \cdot \mathbf{k}_j) \mathbf{v}_j$$

weights of value vectors





Memory Keys Correspond to Input Patterns

Each individual key vector corresponds to a specific pattern over the input prefix

Key	Pattern	Example trigger prefixes
k_{449}^1	Ends with “ <i>substitutes</i> ” (shallow)	<i>At the meeting, Elton said that “for artistic reasons there could be no substitutes In German service, they were used as substitutes Two weeks later, he came off the substitutes</i>
k_{2546}^6	Military, ends with “ <i>base</i> ”/“ <i>bases</i> ” (shallow + semantic)	<i>On 1 April the SRSG authorised the SADF to leave their bases Aircraft from all four carriers attacked the Australian base Bombers flying missions to Rabaul and other Japanese bases</i>
k_{2997}^{10}	a “part of” relation (semantic)	<i>In June 2012 she was named as one of the team that competed He was also a part of the Indian delegation Toy Story is also among the top ten in the BFI list of the 50 films you should</i>
k_{2989}^{13}	Ends with a time range (semantic)	<i>Worldwide, most tornadoes occur in the late afternoon, between 3 pm and 7 Weekend tolls are in effect from 7:00 pm Friday until The building is open to the public seven days a week, from 11:00 am to</i>
k_{1935}^{16}	TV shows (semantic)	<i>Time shifting viewing added 57 percent to the episode’s The first season set that the episode was included in was as part of the From the original NBC daytime version , archived</i>



Memory Values Correspond to Output Tokens

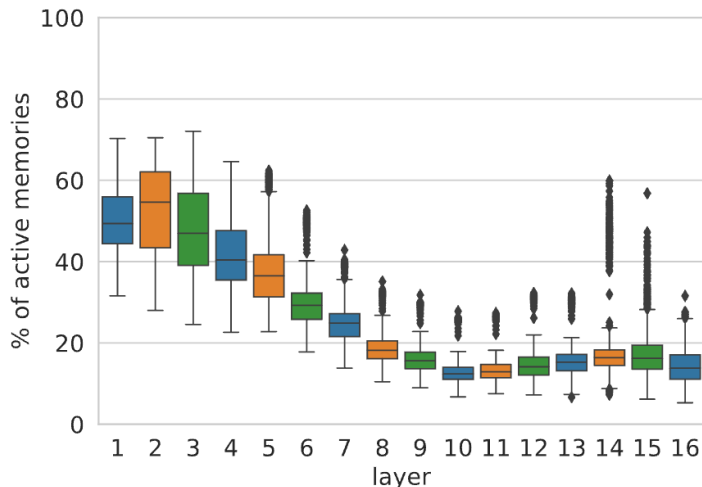
Each value vector (roughly) matches a predicted token distribution

Value	Prediction	Trigger example
\mathbf{v}_{222}^{15}	<i>each</i>	<i>But when bees and wasps resemble each</i>
\mathbf{v}_{752}^{16}	<i>played</i>	<i>Her first role was in Vijay Lalwani's psychological thriller Karthik Calling Karthik, where Padukone was cast as the supportive girlfriend of a depressed man (played)</i>
\mathbf{v}_{2601}^{13}	<i>extratropical</i>	<i>Most of the winter precipitation is the result of synoptic scale, low pressure weather systems (large scale storms such as extratropical)</i>
\mathbf{v}_{881}^{15}	<i>part</i>	<i>Comet served only briefly with the fleet, owing in large part</i>
\mathbf{v}_{2070}^{16}	<i>line</i>	<i>Sailing from Lorient in October 1805 with one ship of the line</i>
\mathbf{v}_{3186}^{12}	<i>jail</i>	<i>On May 11, 2011, four days after scoring 6 touchdowns for the Slaughter, Grady was sentenced to twenty days in jail</i>

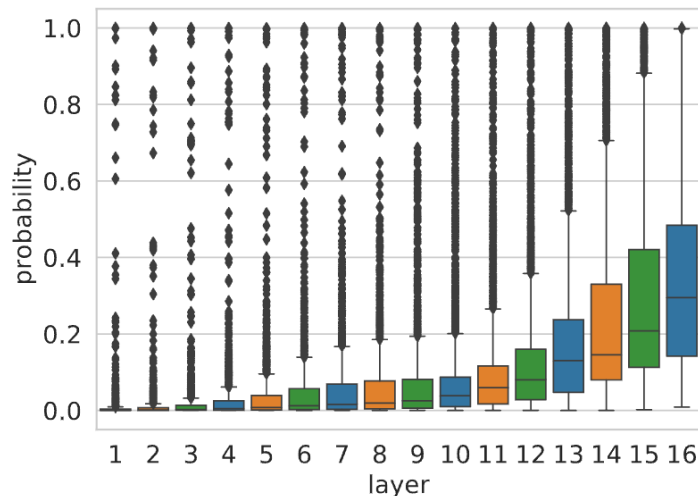


Memory Aggregation

- “Active” memories (memory vectors with non-zero coefficients) are typically sparse
- The residual connection sequentially refines token prediction from layer to layer



Fraction of active memory units across layers



Output token probability is gradually refined across layers



Further Reading on LLM Parametric Knowledge

- [How Much Knowledge Can You Pack Into the Parameters of a Language Model?](#) [Roberts et al., 2020]
- [Extracting Training Data from Large Language Models](#) [Carlini et al., 2021]
- [Locating and Editing Factual Associations in GPT](#) [Meng et al., 2022]

Agenda

- Question Answering (QA)
- Parametric Knowledge in LMs
- Hallucination
- Non-parametric Knowledge

Join at
slido.com
#1886 206



Hallucination

Join at
slido.com
#1886 206



- **Hallucination:** LM generates information that is factually incorrect, misleading, or fabricated, even though it may sound plausible or convincing
- Why does hallucination happen?
 - Limited knowledge: LLMs are trained on finite datasets, which don't have access to all possible information; when asked about topics outside their training data, they may generate plausible-sounding but incorrect responses
 - Overgeneralization: LLMs may apply patterns they've learned from one context to another where they don't apply, leading to incorrect conclusions
 - Lack of common sense: While LLMs can process and generate human-like text, they often lack the ability to apply commonsense reasoning to their outputs
 - ...



Hallucination Examples

- **(Limited knowledge)** Q: “What were the main features of the iPhone 15 Pro Max?”
LLM (trained before 2023): “The iPhone 15 Pro Max features a revolutionary holographic display, quantum computing chip, and telepathic user interface.”
- **(Overgeneralization)** Q: “How do you form the past tense in Japanese?”
LLM: “In Japanese, you typically add '-ed' to the end of verbs to form the past tense, just like in English.” (incorrect)
- **(Lack of common sense)** Q: “How many tennis balls can fit in a typical smartphone?”
LLM: “Approximately 15-20 tennis balls can fit in a typical smartphone, depending on the model and screen size.”

what's your knowledge cutoff date?



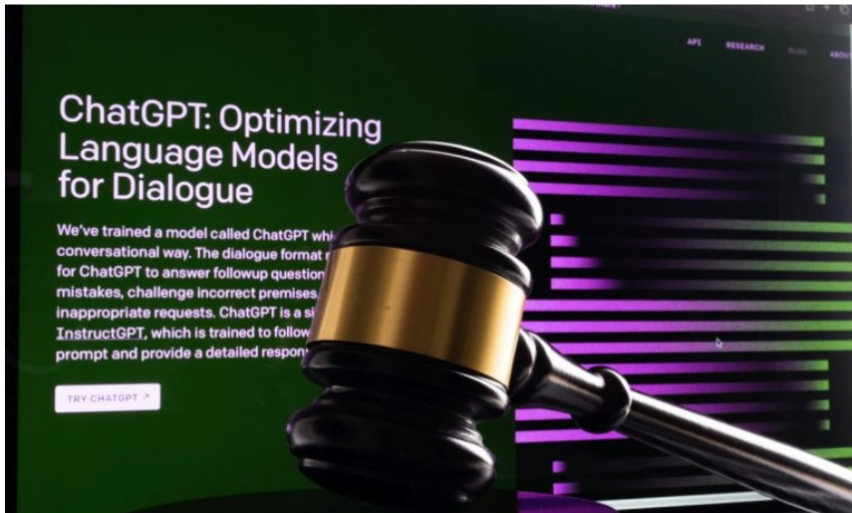
My knowledge cutoff date is October 2023. This means I don't have information on events or developments that have occurred after that time. How can I assist you with your question?

Concerns About Hallucination

Still a concerning issue in modern LLMs!

Attorneys Face Sanctions After Citing Case Law 'Hallucinated' by ChatGPT

BY PYMNTS | MAY 30, 2023



Join at
slido.com
#1886 206



Figure source: <https://www.pymnts.com/artificial-intelligence-2/2023/attorneys-face-sanctions-after-citing-information-hallucinated-by-chatgpt/>



Further Reading on Hallucination

- [LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples](#) [Yao et al., 2023]
- [Towards Mitigating Hallucination in Large Language Models via Self-Reflection](#) [Ji et al., 2023]
- [Hallucination is Inevitable: An Innate Limitation of Large Language Models](#) [Xu et al., 2024]

Agenda

- Question Answering (QA)
- Parametric Knowledge in LMs
- Hallucination
- Non-parametric Knowledge

Join at
slido.com
#1886 206





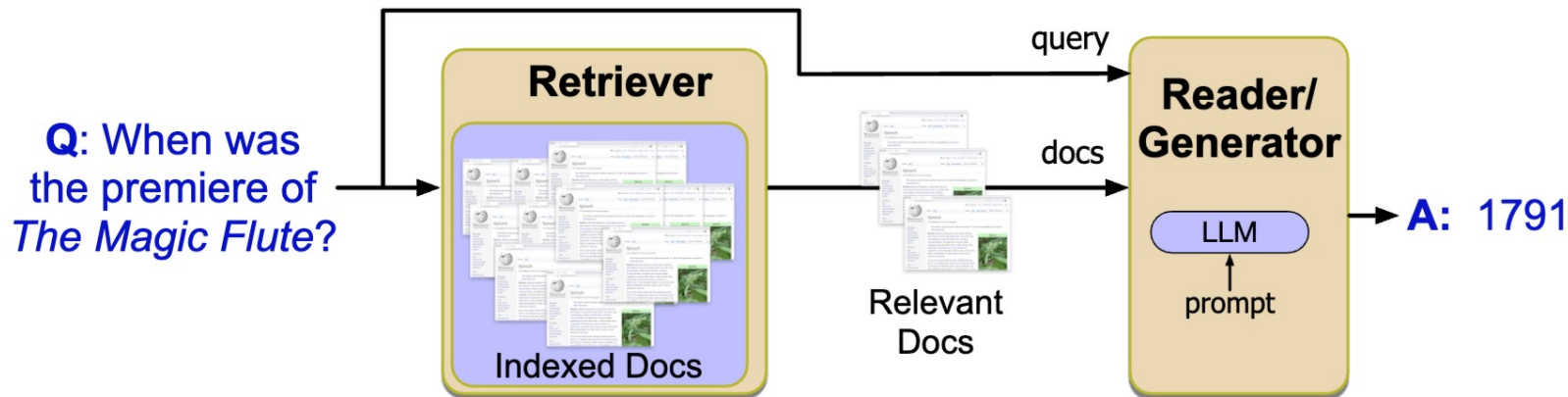
Non-parametric Knowledge

- **Non-parametric knowledge:** (external) information not stored in the model's parameters but can be accessed or retrieved when needed
- Examples:
 - External knowledge bases/graphs
 - Pretraining corpora
 - User-provided documents/passages
- Non-parametric knowledge is typically used to **augment** parametric knowledge (typically via **retrieval**) for more accurate factoid question answering
- Benefits of **non-parametric knowledge**
 - Incorporate more information without increasing model size
 - Easier updates and modifications to the knowledge base
 - Improve model interpretability



Overview: Retrieval-Augmented Generation

- Use a retriever to obtain relevant documents to the query from an external text collection
- Use LLMs to generate answers given the documents and a prompt



Sparse vs. Dense Retrieval

Join at
slido.com
#1886 206



- **Sparse** retrieval: based on traditional information retrieval (IR) techniques where the representations of documents and queries are sparse (most vector values are zero)
 - Example: TF-IDF
 - Pros: simple and interpretable
 - Cons: lack semantic understanding
- **Dense** retrieval: encode documents and queries into dense vectors (embeddings) using deep neural networks
 - Example: BERT-based encoding methods
 - Pros: semantic & contextualized understanding
 - Cons: computationally more expensive and less interpretable



Summary: Scaling Up LLMs & Emergent Abilities

- Scaling up LLMs involves three factors
 - Add more parameters (scaling up model size)
 - Add more data (scaling up dataset size)
 - Train for more iterations (scaling up compute)
- The language modeling loss exhibits a power-law relationship with each of the three factors
- Given a specific compute, an optimal model size can be derived
- Emergent ability: a model has random performance until a certain scale, after which performance increases notably



Summary: Reasoning with LLMs

- Reasoning: perform deductive, inductive, commonsense, or logical reasoning by generating text with language models
- Chain-of-thought (CoT): the model breaks down complex problems into a step-by-step reasoning process Instead of directly answering the question
- CoT can be either zero-shot or few-shot
- Self-consistency CoT involves generating multiple reasoning paths and deriving the final answer with majority voting



Summary: Question Answering & Knowledge

- Question answering (QA) can be categorized in different ways:
 - Closed-domain vs. open-domain QA
 - Extractive vs. abstractive QA
 - Closed-book vs. open-book QA
- LMs can be used as knowledge bases via prompting
- Feedforward network parameters can store factual knowledge (parametric knowledge)
- LMs can hallucinate, which can be mitigated by retrieving non-parametric knowledge



Thank You!

Yu Meng

University of Virginia

yumeng5@virginia.edu