

LLM Reasoning

Slido: <https://app.sli.do/event/qqdJWow9mWJs9viZzDy7Jy>

Yu Meng

University of Virginia

yumeng5@virginia.edu

Nov 5, 2025

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 & Recurrent Neural Networks (RNNs)
- Week 6: Language Modeling with Transformers
- Week 8: Transformer and Pretraining
- Week 9: Large Language Models (LLMs) & In-context Learning
- Week 10: Knowledge in LLMs and Retrieval-Augmented Generation (RAG)
- **Week 11: LLM Reasoning**
- Week 12: Reinforcement Learning for LLM Post-Training
- Week 13: LLM Agents + Course Summary
- Week 15 (after Thanksgiving): Project Presentations

Reminders

- Assignment 5 out (due 11/19 11:59pm)
- Second guest lecture next Monday 11/10 (Meet on Zoom; details to be shared later)

(Recap) Prompting LMs: Parametric Knowledge

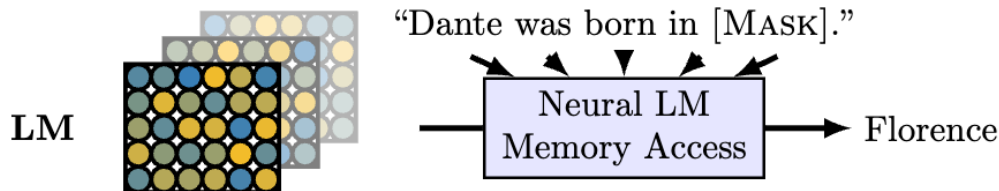
- LMs have learned from a lot of facts in their pretraining data
- LMs can be directly prompted to generate answers to factoid questions (Closed-book QA setting)
- Example:

$P(w|$ Q: 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**

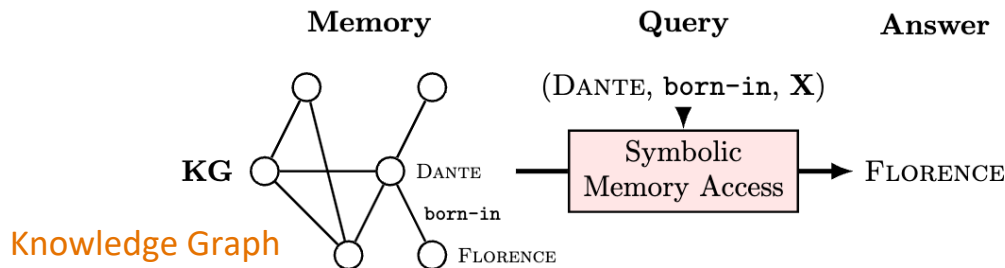
(Recap) Language Model as Knowledge Bases

- **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 - May produce incorrect/outdated information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



(Recap) 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

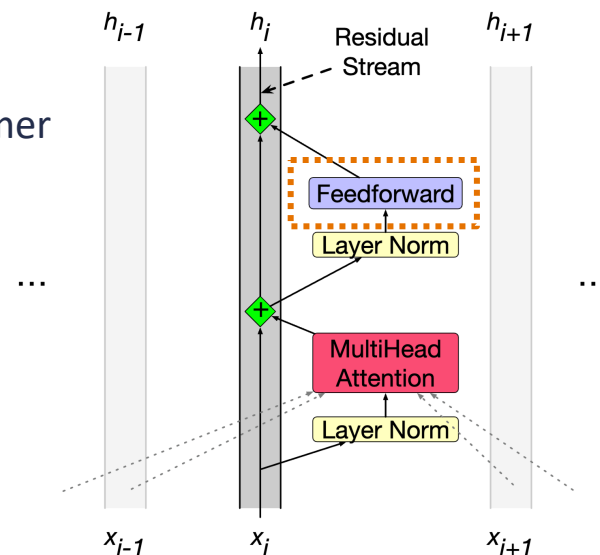


(Recap) Feedforward Parameters in Transformer

- 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



(Recap) Feedforward Parameters Are Neural Memories

Viewing FFN as key-value memories

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



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

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

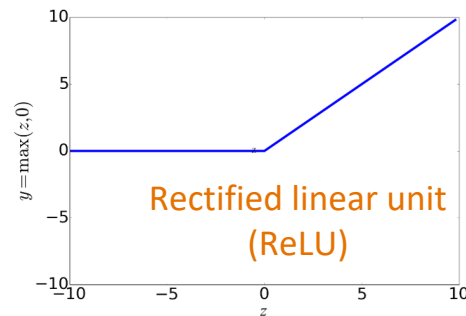
$$\mathbf{V} \in \mathbb{R}^{d_2 \times d_1}$$

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



(Recap) 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 “substitutes” (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 “base”/“bases” (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>



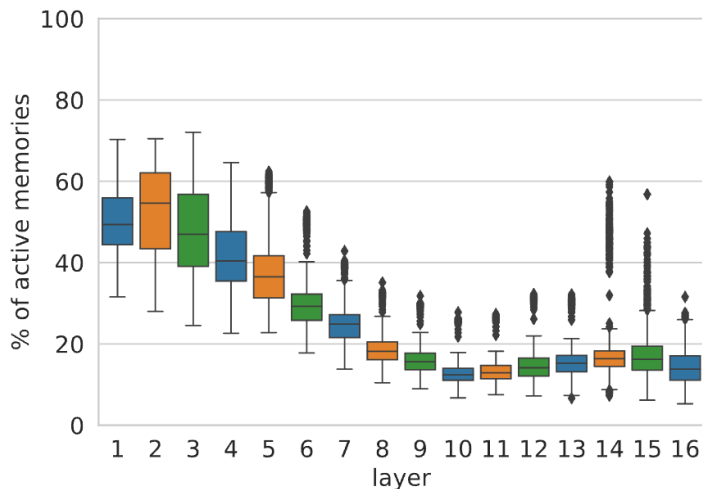
(Recap) Memory Values Correspond to Output Tokens

Each value vector (roughly) matches a predicted token distribution

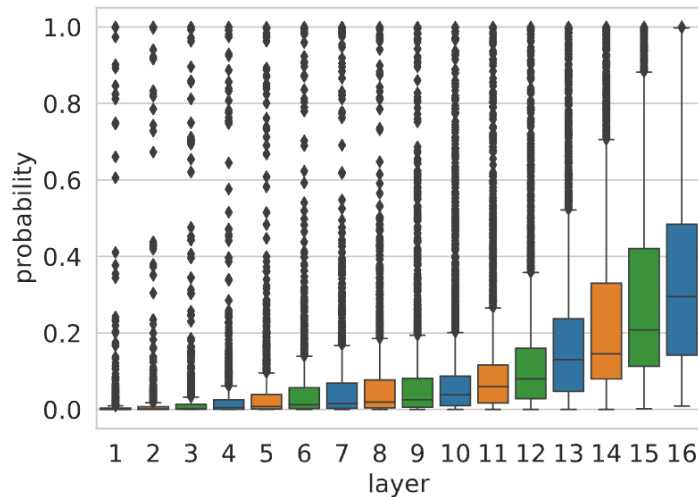
Value	Prediction	Trigger example
$\mathbf{v}_{222}^{15}$	<i>each</i>	<i>But when bees and wasps resemble each</i>
$\mathbf{v}_{752}^{16}$	<i>played</i>	<i>Her first role was in Vijay Lalwani's psychological thriller Karthik Calling Karthik, where Padukone was cast as the supportive girlfriend of a depressed 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>

(Recap) 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

(Recap) Hallucination

- **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
 - ...

(Recap) 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?

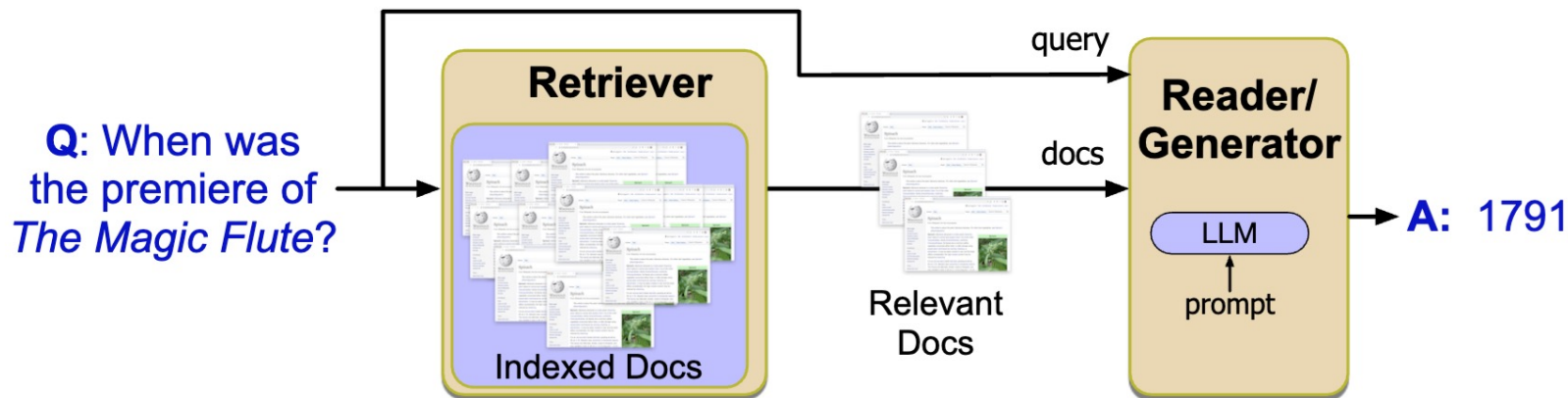
Knowledge cutoff date of GPT4o

(Recap) 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

(Recap) 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



(Recap) Overview: Information Retrieval (IR)

- **Information retrieval (IR):** finding relevant information from a large collection of unstructured data (e.g., documents, web pages) in response to a user query
- **Query:** user-provided input (e.g., keywords or phrases), describing the information they are seeking
- **Documents/corpus:** the data collection that the system searches through
- **Ranking:** sort the search results by relevance based on specific metrics (e.g., keyword matching, semantic similarity)
- Web search engines (e.g., Google, Bing) are IR systems

(Recap) Sparse vs. Dense Retrieval

- **Sparse** retrieval: based on traditional 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

(Recap) TF-IDF Weighting

- Introduced in week 3's lectures $\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \text{IDF}(w)$
- Main idea: represent a document with frequent & distinctive words

TF-IDF weighted

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.246	0	0.454	0.520
good	0	0	0	0
fool	0.030	0.033	0.0012	0.0019
wit	0.085	0.081	0.048	0.054

$$\cos(\mathbf{v}_{d_2}, \mathbf{v}_{d_3}) = 0.10 \quad \cos(\mathbf{v}_{d_3}, \mathbf{v}_{d_4}) = 0.99$$

Raw counts

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

$$\cos(\mathbf{v}_{d_2}, \mathbf{v}_{d_3}) = 0.81 \quad \cos(\mathbf{v}_{d_3}, \mathbf{v}_{d_4}) = 0.99$$



(Recap) Term Frequency (TF)

- A word appearing 100 times in a document doesn't make it 100 times more likely to be relevant to the meaning of the document
- Instead of using the raw counts, we squash the counts with log scale

$$\text{TF}(w, d) = \begin{cases} 1 + \log_{10} \text{count}(w, d) & \text{count}(w, d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

(Recap) Inverse Document Frequency (IDF)

- We want to emphasize discriminative words (with low DF)
- Inverse document frequency (IDF): total number of documents (N) divided by DF, in log scale

$$\text{IDF}(w) = \log_{10} \left(\frac{N}{\text{DF}(w)} \right)$$

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

DF & IDF statistics in the
Shakespeare corpus
(37 documents)

(Recap) TF-IDF for Sparse Retrieval

- Score document-query semantic similarity by cosine similarity

$$\cos(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{|\mathbf{q}| |\mathbf{d}|}$$

- Both document and query vectors use TF-IDF weighting
- Can also adopt other weighting schemes (e.g., BM25)

(Recap) Example: TF-IDF for Sparse Retrieval

- Example query and mini-corpus:

Query: sweet love

Doc 1: Sweet sweet nurse! Love?

Doc 2: Sweet sorrow

Doc 3: How sweet is love?

Doc 4: Nurse!

- Query & document vectors:

word	Query					
	cnt	tf	df	idf	tf-idf	n'lized = tf-idf/ q
sweet	1	1	3	0.125	0.125	0.383
nurse	0	0	2	0.301	0	0
love	1	1	2	0.301	0.301	0.924
how	0	0	1	0.602	0	0
sorrow	0	0	1	0.602	0	0
is	0	0	1	0.602	0	0

word	Document 1			
	cnt	tf	tf-idf	n'lized
sweet	2	1.301	0.163	0.357
nurse	1	1.000	0.301	0.661
love	1	1.000	0.301	0.661
how	0	0	0	0
sorrow	0	0	0	0
is	0	0	0	0

	Document 2			
	cnt	tf	tf-idf	n'lized
1	1	1.000	0.125	0.203
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
1	1	1.000	0.602	0.979
0	0	0	0	0

$$\cos(\mathbf{q}, \mathbf{d}_1) = 0.747$$

$$\cos(\mathbf{q}, \mathbf{d}_2) = 0.078$$

Agenda

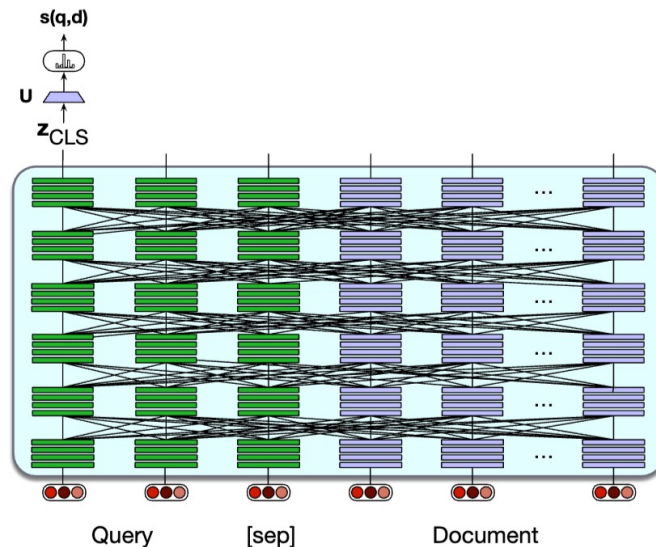
- Dense Retrieval
- Evaluation of Retrieval
- Retrieval-Augmented Generation (RAG) for LLMs
- Chain-of-thought Reasoning
- Reasoning Benchmarks

Dense Retrieval

- Motivation: sparse retrieval (e.g., TF-IDF) relies on the exact overlap of words between the query and document without considering semantic similarity
- Solution: use a language model to obtain (dense) distributed representations of query and document
- The retriever language model is typically a small text encoder model (e.g., BERT)
 - Retrieval is a natural language understanding task
 - Encoder-only models are more efficient than LLMs for this purpose
- Both query and document representations are computed by text encoders

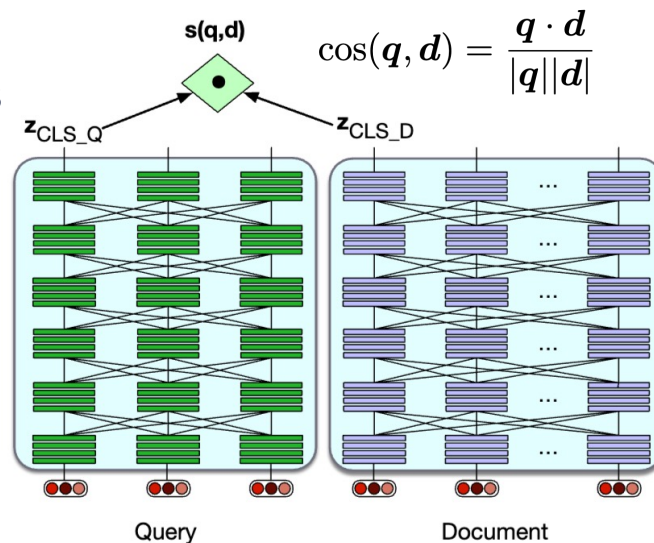
Dense Retrieval: Cross-encoder

- Process query-document pairs together
- Relevance score produced directly by the model output
- (+) Capture intricate interactions between the query and the document
- (-) Not scalable to large retrieval corpus
- Good for small document sets



Dense Retrieval: Bi-encoder

- Independently encode the query and the document using two separate (but often identical) encoder models
- Use cosine similarity between the query and document vectors as relevance score
- (+) Document vectors can be precomputed
- (-) Cannot capture query-document interactions
- Common choice for large-scale retrieval



Agenda

- Dense Retrieval
- Evaluation of Retrieval
- Retrieval-Augmented Generation (RAG) for LLMs
- Chain-of-thought Reasoning
- Reasoning Benchmarks

Evaluation of IR Systems

- Assume that each document returned by the IR system is either **relevant** to our purposes or **not relevant**
- Given a query, assume the system returns a set of ranked documents T
 - A subset R of these are relevant (The remaining $N = T - R$ is irrelevant)
 - There are U documents in the entire retrieval collection that are relevant to this query
- **Precision:** the fraction of the returned documents that are relevant

$$\text{Precision} = \frac{|R|}{|T|}$$

- **Recall:** the fraction of all relevant documents that are returned

$$\text{Recall} = \frac{|R|}{|U|}$$

Precision & Recall @ k

- We hope to build a retrieval system that ranks the relevant documents higher
- Use precision & recall @ k (among the top- k items in the ranked list) to reflect this

Rank	Judgment	Precision _{Rank}	Recall _{Rank}
1	R	1.0	.11
2	N	.50	.11
3	R	.66	.22
4	N	.50	.22
5	R	.60	.33
6	R	.66	.44
7	N	.57	.44
8	R	.63	.55
9	N	.55	.55
10	N	.50	.55

Assume there are 9 total relevant documents in the retrieval corpus

Average Precision

Average precision (AP): mean of the precision values at the points in the ranked list where a relevant document is retrieved

Indicator function of whether
the document is relevant

$$AP = \frac{1}{|R|} \sum_{k=1}^{|T|} (\text{Precision@}k \times \mathbb{1}(d_k \text{ is relevant}))$$

Rank	Judgment	Precision _{Rank}	Recall _{Rank}
1	R	1.0	.11
2	N	.50	.11
3	R	.66	.22
4	N	.50	.22
5	R	.60	.33
6	R	.66	.44
7	N	.57	.44
8	R	.63	.55
9	N	.55	.55
10	N	.50	.55

Agenda

- Dense Retrieval
- Evaluation of Retrieval
- Retrieval-Augmented Generation (RAG) for LLMs
- Chain-of-thought Reasoning
- Reasoning Benchmarks

RAG for LLMs

RAG is a common technique to enhance factuality of LLM generation & mitigate hallucination

Time-sensitive query triggers
RAG for ChatGPT

What is the latest iPhone model?

Searching for latest iPhone model Apple

Retriever performs web search

Citations

W

Wikipedia

iPhone 17
Today

TOI

The Times of India

iPhone Air, iPhone 17, iPhone 17 Pro and iPhone 17 Pro Max launched: India price, features, key...
September 9, 2025 — Apple unveiled its 2025 iPhone lineup at a major launch event, introducing...

W

Wikipedia

iPhone 17 Pro
Today

W

Wikipedia

iPhone Air
Today

Retrieved webpages

RAG vs. Direct Prompting

- Prompting relies on LM's parametric knowledge to directly answer the question:

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

- RAG prepends the set of retrieved passages to the question

Schematic of a RAG Prompt

retrieved passage 1

retrieved passage 2

...

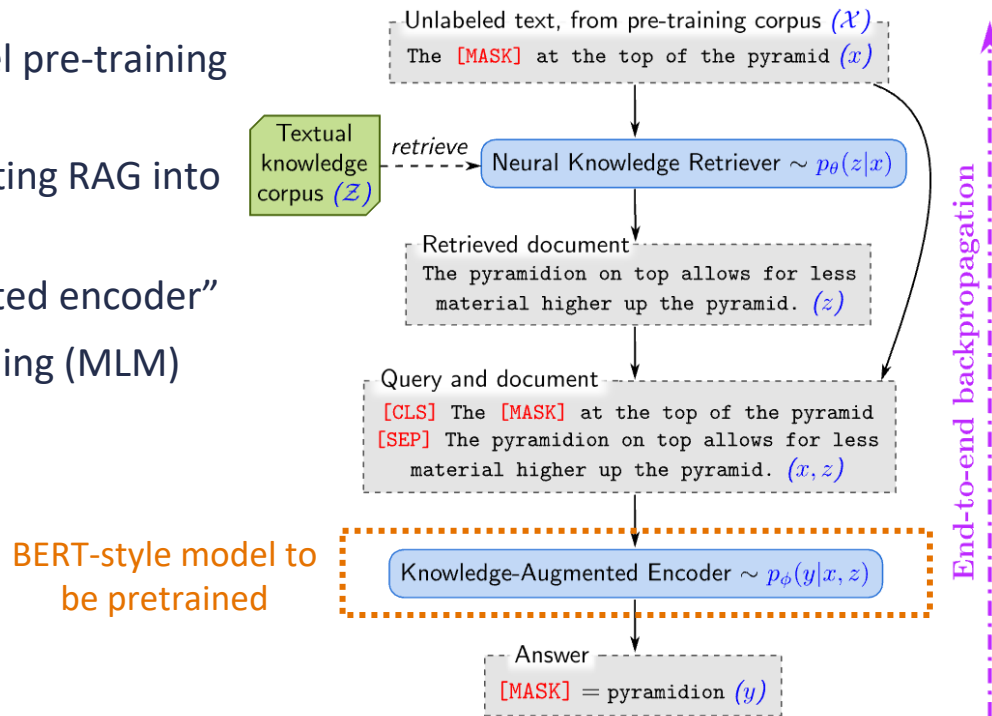
retrieved passage n

Returned by the retriever

Based on these texts, answer this question: Q: Who wrote the book "The Origin of Species"? A:

RAG in Pretraining

- Retrieval-Augmented Language Model pre-training (REALM)
- The first paper that studies incorporating RAG into encoder pretraining (BERT style)
- Main model is a “knowledge-augmented encoder”
- Pretrain with masked language modeling (MLM) loss conditioned on retrieved content





RAG for Text Generation

- The first paper that studies RAG for text generation
- Both the retriever (an encoder-only LM) and the generator (an LLM) are trained

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis^{†‡}, Ethan Perez^{*},

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel^{†‡}, Sebastian Riedel^{†‡}, Douwe Kiela[†]

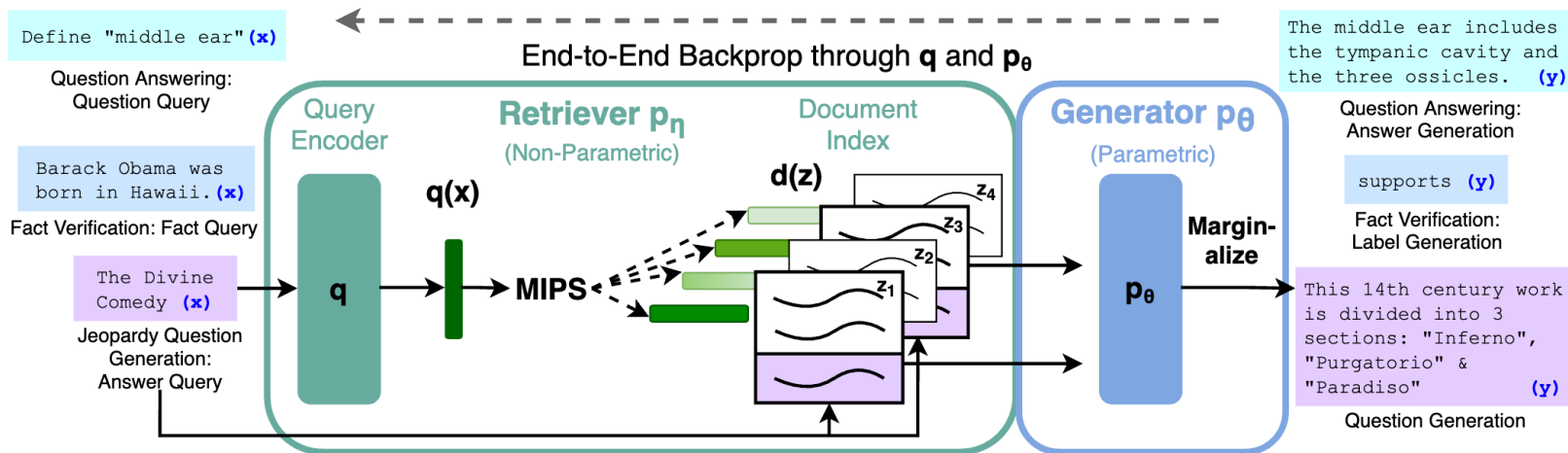
A Latent Variable Model

The retrieved documents are treated as latent variables (z) for generation

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{D}} p(\mathbf{z}|\mathbf{x})p(\mathbf{y}|\mathbf{x}, \mathbf{z})$$

Retrieve document (z)
based on query (x)

Generate answer (y) based on
retrieved docs (z) and query (x)



RAG-Sequence Model

- Use the same retrieved document to generate the complete sequence
- Treat the retrieved document as a single latent variable
- Marginalize to get the generation probability $p(\mathbf{y}|\mathbf{x})$ via a top-K approximation

$$p_{\text{RAG-sequence}}(\mathbf{y}|\mathbf{x}) \approx \sum_{\mathbf{z} \in \text{top-K}(p(\cdot|\mathbf{x}))} p_{\eta}(\mathbf{z}|\mathbf{x}) p_{\theta}(\mathbf{y}|\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z} \in \text{top-K}(p(\cdot|\mathbf{x}))} p_{\eta}(\mathbf{z}|\mathbf{x}) \prod_{i=1}^N p_{\theta}(y_i|\mathbf{x}, \mathbf{z}, \mathbf{y}_{<i})$$

 Top-K approximation
(only consider the top-K retrieved docs)

 The same retrieved doc (z) is used to
generate all tokens in the sequence



RAG-Token Model

- Can use different retrieved documents to generate different tokens in a sequence
- Marginalization is performed for each generated token (rather than at sequence level)

$$p_{\text{RAG-token}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^N p_{\theta}(y_i|\mathbf{x}, \mathbf{y}_{<i}) \approx \prod_{i=1}^N \sum_{\mathbf{z} \in \text{top-K}(p(\cdot|\mathbf{x}, \mathbf{y}_{<i}))} p_{\eta}(\mathbf{z}|\mathbf{x}, \mathbf{y}_{<i}) p_{\theta}(y_i|\mathbf{x}, \mathbf{z}, \mathbf{y}_{<i})$$



Different retrieved doc ($\mathbf{z}$) can be used to generate different tokens in the sequence

RAG-Sequence & RAG-Token Results

Evaluation results on open-domain QA tasks:

- Natural Questions (NQ)
- TriviaQA (TQA)
- WebQuestions (WQ)
- CuratedTrec (CT)

Model		NQ	TQA	WQ	CT
Closed	T5-11B [52]	34.5	- /50.1	37.4	-
Book	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open	REALM [20]	40.4	- / -	40.7	46.8
Book	DPR [26]	41.5	57.9 / -	41.1	50.6
RAG-Token		44.1	55.2/66.1	45.5	50.0
RAG-Seq.		44.5	56.8/ 68.0	45.2	52.2



Further Reading on RAG

- [Generalization through Memorization: Nearest Neighbor Language Models](#) [Khandelwal et al., 2019]
- [Active Retrieval Augmented Generation](#) [Jiang et al., 2023]
- [Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection](#) [Asai et al., 2023]
- [InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales](#) [Wei et al., 2024]

Agenda

- Dense Retrieval
- Evaluation of Retrieval
- Retrieval-Augmented Generation (RAG) for LLMs
- Chain-of-thought Reasoning
- Reasoning Benchmarks

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”

Frontier LLMs Are Reasoning Models

One unified system

GPT-5 is a unified system with a **smart, efficient model** that answers most questions, a **deeper reasoning model** (GPT-5 thinking) for harder problems, and a **real-time router** that quickly decides which to use based on conversation type, complexity, tool needs, and your explicit intent (for example, if you say “think hard about this” in the prompt). The router is continuously trained on real signals, including when users switch models, preference rates for responses, and measured correctness, improving over time. Once usage limits are reached, a mini version of each model handles remaining queries. In the near future, we plan to integrate these capabilities into a single model.

Chain-of-thought (CoT) Prompting

- **Chain-of-thought (CoT):** the model breaks down complex problems into a step-by-step reasoning process
- Instead of directly providing an answer to a question or task, the model is prompted to explain its reasoning or thought process in a logical sequence

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com



Standard Prompting 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?

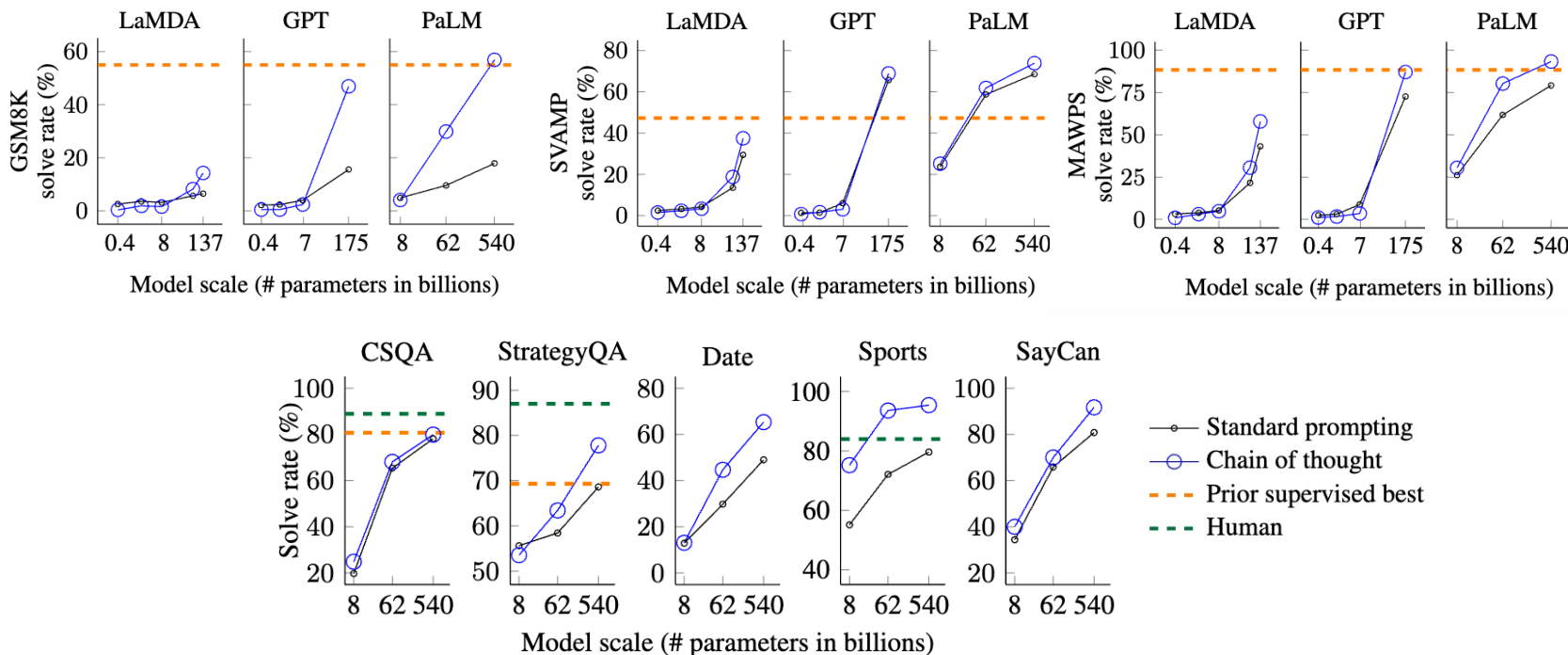
Model Output

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. ✅



Standard vs. CoT Prompting Performance

CoT prompting is especially effective for large models



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. ✗

(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. ✓

(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 ✗

(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. ✓

CoT Demo

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

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

llama-3.3-70b-instruct

2

Wrong result

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

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

llama-3.3-70b-instruct

To find out how many 'r' letters are in the word "strawberry", let's break it down step by step:

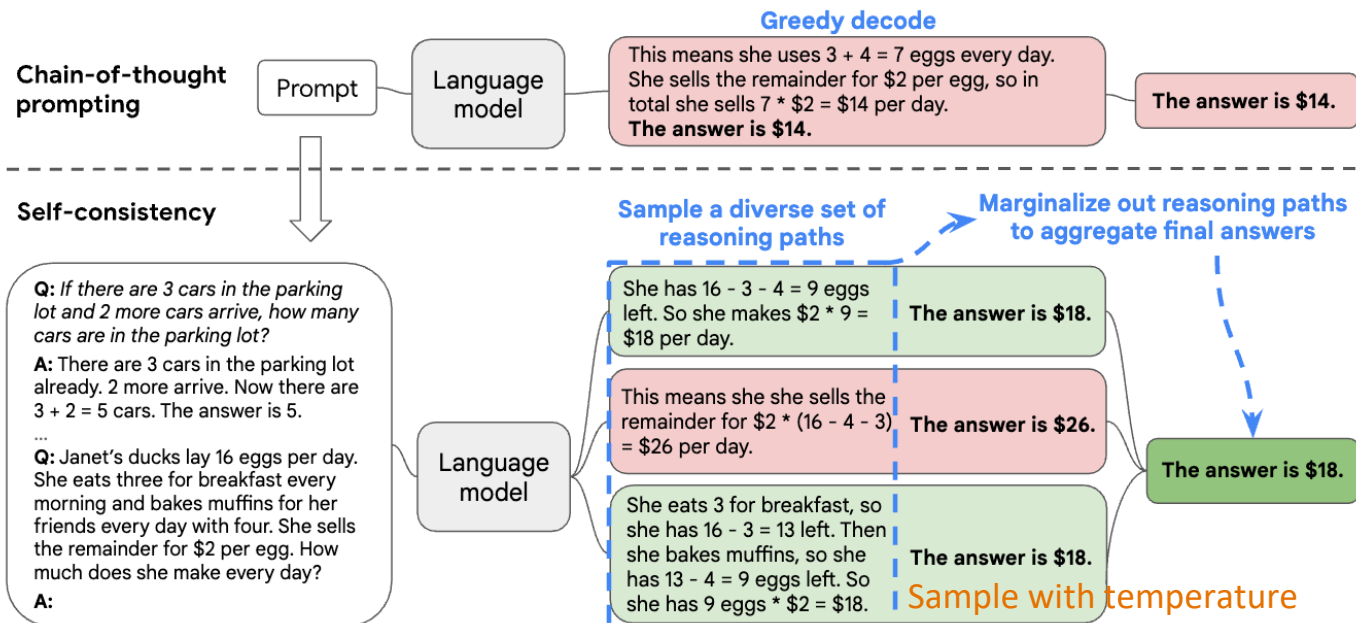
1. **Start with the word:** strawberry
2. **Separate the word into its individual letters:** s-t-r-a-w-b-e-r-r-y
3. **Identify the 'r' letters:** We see 'r' in the following positions:
 - The third letter is 'r'.
 - The eighth letter is 'r'.
 - The ninth letter is also 'r'.
4. **Count the 'r' letters:** There are 3 'r' letters in the word "strawberry".

So, there are 3 'r' letters in the word "strawberry".

Correct result

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





Further Reading on LLM Reasoning

- [Least-to-Most Prompting Enables Complex Reasoning in Large Language Models](#) [Zhou et al., 2022]
- [Large Language Models Can Self-Improve](#) [Huang et al., 2022]
- [Tree of Thoughts: Deliberate Problem Solving with Large Language Models](#) [Yao et al., 2023]
- [Let's Verify Step by Step](#) [Lightman et al., 2023]

Agenda

- Dense Retrieval
- Evaluation of Retrieval
- Retrieval-Augmented Generation (RAG) for LLMs
- Chain-of-thought Reasoning
- Reasoning Benchmarks

Grade School Math (GSM8K)

8.5K high quality grade school math problems created by human problem writers

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4 \times 2 = 8$ dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = 96$ cookies
She splits the 96 cookies equally amongst 16 people so they each eat $96/16 = 6$ cookies

Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50?

Mrs. Lim got 68 gallons - 18 gallons = 50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = 200 gallons.
She was able to sell 200 gallons - 24 gallons = 176 gallons.
Thus, her total revenue for the milk is $\$3.50/\text{gallon} \times 176 \text{ gallons} = \616 .

Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?

Solution: Tina buys 3 12-packs of soda, for $3 \times 12 = 36$ sodas
6 people attend the party, so half of them is $6/2 = 3$ people
Each of those people drinks 3 sodas, so they drink $3 \times 3 = 9$ sodas
Two people drink 4 sodas, which means they drink $2 \times 4 = 8$ sodas
With one person drinking 5, that brings the total drank to $5 + 9 + 8 = 25$ sodas
As Tina started off with 36 sodas, that means there are $36 - 25 = 11$ sodas left

Final Answer: 11

MATH

12.5K challenging competition mathematics problems

Problem: Suppose a and b are positive real numbers with $a > b$ and $ab = 8$. Find the minimum value of $\frac{a^2+b^2}{a-b}$.

Ground truth solution: We can write $\frac{a^2+b^2}{a-b} = \frac{a^2+b^2-2ab+16}{a-b} = \frac{(a-b)^2+16}{a-b} = a - b + \frac{16}{a-b}$. By AM-GM, $a - b + \frac{16}{a-b} \geq 2\sqrt{(a-b) \cdot \frac{16}{a-b}} = 8$. Equality occurs when $a - b = 4$ and $ab = 8$. We can solve these equations to find $a = 2\sqrt{3} + 2$ and $b = 2\sqrt{3} - 2$. Thus, the minimum value is $\boxed{8}$.

Problem: Right $\triangle ABC$ has legs measuring 8 cm and 15 cm. The triangle is rotated about one of its legs. What is the number of cubic centimeters in the maximum possible volume of the resulting solid? Express your answer in terms of π .

Ground truth solution: If the triangle is rotated about the shorter leg, then the radius is the longer leg and the height is the shorter leg, and the volume is $\frac{1}{3} \cdot (15^2\pi)(8) = 600\pi$ cubic centimeters. If the triangle is rotated about the longer leg, then the radius is the shorter leg and the height is the longer leg, and the volume is $\frac{1}{3}(8^2\pi)(15)$, which is $\frac{8}{15}$ of the volume we found earlier. So, the maximum possible volume is $\boxed{600\pi}$ cubic centimeters.

AI2 Reasoning Challenge (ARC)

~8K natural science questions on commonsense knowledge/reasoning

Reasoning Type	Example
Question logic	Which item below is not made from a material grown in nature? (A) a cotton shirt (B) a wooden chair (C) a plastic spoon (D) a grass basket
Linguistic Matching	Which of the following best describes a mineral? (A) the main nutrient in all foods (B) a type of grain found in cereals (C) a natural substance that makes up rocks (D) the decomposed plant matter found in soil
Multihop Reasoning	Which property of a mineral can be determined just by looking at it? (A) luster (B) mass (C) weight (D) hardness
Comparison	Compared to the Sun, a red star most likely has a greater (A) volume. (B) rate of rotation. (C) surface temperature. (D) number of orbiting planets
Algebraic	If a heterozygous smooth pea plant (Ss) is crossed with a homozygous smooth pea plant (SS), which are the possible genotypes the offspring could have? (A) only SS (B) only Ss (C) Ss or SS (D) ss or SS
Hypothetical / Counterfactual	If the Sun were larger, what would most likely also have to be true for Earth to sustain life? (A) Earth would have to be further from the Sun. (B) Earth would have to be closer to the Sun. (C) Earth would have to be smaller. (D) Earth would have to be larger.
Explanation / Meta-reasoning	Why can steam be used to cook food? (A) Steam does work on objects. (B) Steam is a form of water. (C) Steam can transfer heat to cooler objects. (D) Steam is able to move through small spaces.

Paper: <https://arxiv.org/pdf/1803.05457>

BIG-Bench Hard (BBH)

23 challenging tasks covering a wide range of reasoning (e.g. arithmetic, logical, spatial...)

Model Input (Navigate)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: If you follow these instructions, do you return to the starting point? Turn left. Turn right. Take 5 steps. Take 4 steps. Turn around. Take 9 steps.

Options:

- Yes
- No

A: Let's think step by step.

Model Output

We start at the origin (0, 0), facing the positive y-axis.

- (1) Turn left: (0, 0), facing the negative x-axis.
 - (2) Turn right: (0, 0), facing the positive y-axis.
 - (3) Take 5 steps: (0, 5), facing the positive y-axis.
 - (4) Take 4 steps: (0, 9), facing the positive y-axis.
 - (5) Turn around: (0, 9), facing the negative y-axis.
 - (6) Take 9 steps: (0, 0), facing the negative y-axis.
- Since (0, 0) is (0, 0), we are indeed where we started. So the answer is **Yes**. ✓

Model Input (Word Sorting)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: Sort the following words alphabetically: List: burley bela arapaho bacteria bock

A: Let's think step by step.

Model Output

The first letter: "burley": "b" (2). "bela": "b" (2). "arapaho": "a" (1). "bacteria": "b" (2). "bock": "b" (2).
We now have: (1) "arapaho" < (2) ["burley" ? "bela" ? "bacteria" ? "bock"]. Now let's sort this subpart ["burley" ? "bela" ? "bacteria" ? "bock"] by looking at their second letters.
The second letter: "burley": "u" (21). "bela": "e" (5). "bacteria": "a" (1). "bock": "o" (15). We now have: (1) "bacteria" < (5) "bela" < (15) "bock" < (21) "burley". Hence, we have "arapaho" < ["bacteria" < "bela" < "bock" < "burley"]. So the answer is **arapaho bacteria bela bock burley**. ✓



Humanity's Last Exam (HLE)

2,500 challenging questions across over a hundred subjects (created by experts)

Mathematics

Question:

The set of natural transformations between two functors $F, G: \mathcal{C} \rightarrow \mathcal{D}$ can be expressed as the end

$$\text{Nat}(F, G) \cong \int_A \text{Hom}_{\mathcal{D}}(F(A), G(A)).$$

Define set of natural cotransformations from F to G to be the coend

$$\text{CoNat}(F, G) \cong \int^A \text{Hom}_{\mathcal{D}}(F(A), G(A)).$$

Let:

- $F = \mathbf{B}_*(\Sigma_4)_{*/}$ be the under ∞ -category of the nerve of the delooping of the symmetric group Σ_4 on 4 letters under the unique 0-simplex $*$ of $\mathbf{B}_*\Sigma_4$.
- $G = \mathbf{B}_*(\Sigma_7)_{*/}$ be the under ∞ -category nerve of the delooping of the symmetric group Σ_7 on 7 letters under the unique 0-simplex $*$ of $\mathbf{B}_*\Sigma_7$.

How many natural cotransformations are there between F and G ?

✉ Emily S
🏛️ University of São Paulo

Computer Science

Question:

Let G be a graph. An edge-indicator of G is a function $a: \{0, 1\} \rightarrow V(G)$ such that $\{a(0), a(1)\} \in E(G)$.

Consider the following Markov Chain $M = M(G)$:

The statespace of M is the set of all edge-indicators of G , and the transitions are defined as follows:

Assume $M_t = a$.

1. pick $b \in \{0, 1\}$ u.a.r.
2. pick $v \in N(a(1 - b))$ u.a.r. (here $N(v)$ denotes the open neighbourhood of v)
3. set $a'(b) = v$ and $a'(1 - b) = a(1 - b)$
4. Set $M_{t+1} = a'$

We call a class of graphs $\mathcal{G}$ well-behaved if, for each $G \in \mathcal{G}$ the Markov chain $M(G)$ converges to a unique stationary distribution, and the unique stationary distribution is the uniform distribution.

Which of the following graph classes is well-behaved?

Answer Choices:

- A. The class of all non-bipartite regular graphs
- B. The class of all connected cubic graphs
- C. The class of all connected graphs
- D. The class of all connected non-bipartite graphs
- E. The class of all connected bipartite graphs.

✉ Marc R
🏛️ Queen Mary University of London



Thank You!

Yu Meng

University of Virginia

yumeng5@virginia.edu