

LLMs, CoT, RAG

Manish Gupta

gmanish@microsoft.com

Principal Applied Scientist, Bing

Adjunct Faculty at IIIT-H

Visiting Faculty at ISB

Agenda

- Applications of LLMs for text and vision
- Transformers to LLMs
- Chain of Thought Reasoning
- Retrieval Augmented Generation

Agenda

- **Applications of LLMs for text and vision**
- Transformers to LLMs
- Chain of Thought Reasoning
- Retrieval Augmented Generation

Machine Translation

The image features a central, large red text "I love you" surrounded by numerous other languages' expressions of love, arranged in a circular or radial pattern. The languages include English, French, Spanish, German, Italian, Portuguese, Russian, Chinese, Korean, Japanese, and many others. Each language's expression is written in its own script and color, creating a diverse and colorful collage.

The agreement on the European Economic Area was signed in August 1992.

L'accord sur la zone économique européenne a été signé en août 1992.

.
<end>

Destruction
of
the
equipment
means
that
Syria
can
no
longer
produce
new
chemical
weapons
. .
<end>
La destruction de l'équipement signifie que la Syrie ne peut plus produire de nouvelles armes chimiques
. .
<end>

The figure displays a 2D matrix representing word co-occurrence or similarity. The x-axis lists English words: It, should, be, noted, that, the, marine, environment, is, the, least, known, of, environments, <end>. The y-axis lists French words: II, convient, de, noter, que, l', environnement, marin, est, le, moins, connu, de, l', environnement, ., <end>. The matrix uses grayscale to represent the strength of the relationship between words, with darker shades indicating higher similarity.

Summarization

Abstract While many approaches to make neural networks more fathomable have been proposed, they are restricted to interrogating the network with input data. [...] In this work, we propose neural persistence, a complexity measure for neural network architectures based on topological data analysis on weighted stratified graphs. [...]

Intro [...] In this work, we present the following contributions: We introduce neural persistence, a novel measure for characterizing the structural complexity of neural networks that can be efficiently computed. [...]

Conclusion [...] However, this did not yield an early stopping measure because it was never triggered, thereby suggesting that neural persistence captures salient information that would otherwise be hidden among all the weights of a network [...]

TLDR We develop a new topological complexity measure for deep neural networks and demonstrate that it captures their salient properties.

Question Answering

Paragraph A, Return to Olympus:

[1] *Return to Olympus* is the only album by the alternative rock band Malfunkshun. [2] It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:

[4] *Mother Love Bone* was an American rock band that formed in Seattle, Washington in 1987. [5] The band was active from 1987 to 1990. [6] Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene. [7] Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success. [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?

A: Malfunkshun

Supporting facts: 1, 2, 4, 6, 7

Solving reasoning problems

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 (Hyperbaton)

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

Q: Which sentence has the correct adjective order:

Options:

- (A) big circular pink Thai silver driving car
- (B) silver circular driving big Thai pink car

A: Let's think step by step.

Model Output

When there is more than one adjective before a noun, the adjectives need to respect the following order before a noun: "[1. opinion] [2. size] [3. age] [4. shape] [5. color] [6. origin] [7. material] [8. purpose] noun".

Option (A): "big circular pink Thai silver driving car". (1) "big" falls into the size category. (2) "circular" falls into the shape category. (3) "pink" falls into the color category. (4) "Thai" falls into the origin category. (5) "silver" falls into the material category. (6) "driving" falls into the purpose category. Option (A) has the following adjective order: [2. size] [4. shape] [5. color] [6. origin] [7. material] [8. purpose] (or, in numeric terms, 2 4 5 6 7 8). Because $2 < 4 < 5 < 6 < 7 < 8$ is correct, (A) has the correct ordering.

Option (B): "silver circular driving big Thai pink car". Option (B) has the following adjective order: [7. material] [4. shape] [8. purpose] [2. size] [6. origin] [5. color] (or, in numeric terms, 7 4 8 2 6 5). Because $7 < 4 < 8 < 2 < 6 < 5$ is not correct, (B) does not have the correct ordering. So the answer is **(A)**. 

Create lists

Multi-lingual generation

Prompt:

Create a shopping list from this recipe:

Trim the ends off zucchini. Cut zucchini in half lengthwise; scoop out pulp, leaving 1/2-in. shells. Finely chop pulp. In a skillet, cook beef, zucchini pulp, onion, mushrooms and peppers over medium heat until meat is no longer pink; drain. Remove from the heat. Add 1/2 cup cheese, ketchup, salt and pepper; mix well. Spoon into the zucchini shells. Place in a greased 13x9-in. baking dish. Sprinkle with remaining cheese.

Bake, uncovered, at 350° for 20-25 minutes.

- 8 zucchini squash, ends trimmed
- 1 pound ground beef
- 1 cup finely chopped zucchini pulp
- 1 small onion, chopped
- 8 ounces diced mushrooms
- 1 sweet red or green pepper, diced
- 1/2 cup grated cheese
- 1/4 cup ketchup
- Salt and pepper to taste
- 1 cup shredded cheese

Input:

Explain in a sentence in Telugu what is backpropagation in neural networks.

BLOOMZ:

బక న్యూట్రానిక్ నెట్వర్క్ లో బ్యాక్ ప్రోసెసింగ్ అనేది ఒక రకమైన అభ్యసన విధానం, ఇది రీజియన్ ల ర్యూక్చ గత అభ్యసన ఫలితాలను ఉపయోగించి, రీజియన్ ల ర్యూక్చ ప్రస్తుత అభ్యసన ఫలితాలను సవరిస్తుంది.

Input:

Write a fairy tale about a troll saving a princess from a dangerous dragon. The fairy tale is a masterpiece that has achieved praise worldwide and its moral is "Heroes Come in All Shapes and Sizes". Story (in Spanish):

BLOOMZ:

Un reino en el bosque estaba bajo un terrible dragón. Todos los que intentaron derrotar al dragón murieron. El rey y la reina le pedían al pueblo que les diera un niño, que les gustaría salvar el reino de la maldición del dragón. Un día, la reina dio a luz a un niño con una cabeza enorme y ojos que brillaban como estrellas. El rey y la reina llamaron al niño Elliot. Elliot crecería y se convertiría en un joven gigante. Él salvaría al pueblo.

Title Generation

Input: In a letter posted on the White House web site Tuesday John Brennan assistant to President Barack Obama for homeland security and counterterrorism said Schmidt will have regular access to the president and play a vital role in the country's security. Schmidt's selection comes more than 10 months after Obama declared cyber security a priority and ordered a broad administration review. A senior White House official said Obama was personally involved in the selection process and chose Schmidt because of his unique background and skills. Schmidt will have regular and direct access to the president for cyber security issues the official said. The official spoke on the condition of anonymity to discuss the selection process. At the same time cyber experts and potential job candidates have complained that the position lacks the budgetary and policy-making authority needed to be successful. Schmidt will report to the National Security Council and closely support the National Economic Council on cyber issues. Schmidt's selection suggests that economic and business interests in the White House held more sway in the selection process. Schmidt president and CEO of the Information Security Forum a nonprofit international consortium that conducts research in information security has served as chief security officer for Microsoft and as cyber security chief for online auction giant eBay. He was reportedly preferred by Lawrence Summers director of the economic council. A good format for the title can be the simple subject + object + verb.

Output: White House picks new cyber coordinator

Explanation: The title is relevant to the main topic of document, that is, the selection of Schmidt as the cybersecurity chief.

Keyword Tagging

Input: Abstract: There is controversy regarding the appropriate utilization of health care resources in the management of tricyclic antidepressant overdosage. Antidepressant overdose patients presenting to the emergency department (ED) are routinely admitted to intensive care units, but only a small proportion develop cardiac arrhythmias or other complications requiring such an environment. The authors reviewed the findings in 165 patients presenting to an ED with antidepressant overdose. They found that major manifestations of toxicity on ED evaluation (altered mental status, seizures, arrhythmias, and conduction defects) were commonly associated with a complicated hospital course. Patients with the isolated findings of sinus tachycardia or QTc prolongation had no complications. No patient experienced a serious toxic event without major evidence of toxicity on ED evaluation and continued evidence of toxicity during the hospital course. These data support the concept that proper ED evaluation can identify a large body of patients with trivial ingestions who may not require hospital observation.

Valid Output: ["Antidepressive Agents, Tricyclic", "Arrhythmia", "California", "Electrocardiography", "Emergencies", "Emergency Service, Hospital", "Female", "Human", "Length of Stay", "Male", "Prognosis", "Retrospective Studies", "Tachycardia, Sinus"]

Ads Copywriting



Scarlet Rose
Bouquet Online
By Floweraura,...
₹ 1,399.00
✓ FlowerAura



Online Red
Roses Bunch,
Red Roses...
₹ 599.00
✓ FlowerAura



Heart Shaped
Rose Box Online
By Floweraura,...
₹ 1,699.00
✓ FlowerAura



Gopro Hero11
5.3K And 27MP
30 FPS...
₹ 51,500.00
Croma



Nikon Z 30
20.9MP
Mirrorless...
₹ 65,999.00
Croma



Sale
SJCAM SJ10
Pro 4K And
12MP 60 FPS...
₹ 17,999.00 ₹...
Croma

6款网红麦片，哪个好吃又不胖

一向注重饮食的超模们，早上都会毫不犹豫来一顿麦片餐，比如刘雯。各种美食博主，都爱 Po 燕...

¥82 已售罄 ② 2.2W

家乐氏玫瑰麦片

这款开袋就有浓郁的玫瑰香气扑鼻而来，干吃的话很合适，一天一袋，吃起来香甜爽脆。当早餐配牛奶、配酸奶也可以，泡发之后能增加玫瑰的香气，里面有椰蓉，白白的、细细的、小小的，是清新的椰子味。不过它的热量和含糖量很高，想减肥的宝宝不要入手啊。

ICA什锦燕麦片

这款麦片里面除了酸奶球，还有酸甜的草莓干、脆脆的玉米片等等，量大而足。泡牛奶也非常 nice，麦片很脆，牛奶变得粉粉的，这不是草莓干掉色，而是草莓粉末溶在了牛奶里。整体吃起来，带着草莓的酸甜，如果不小心吃到了酸奶片，还有种神秘的小惊喜感。

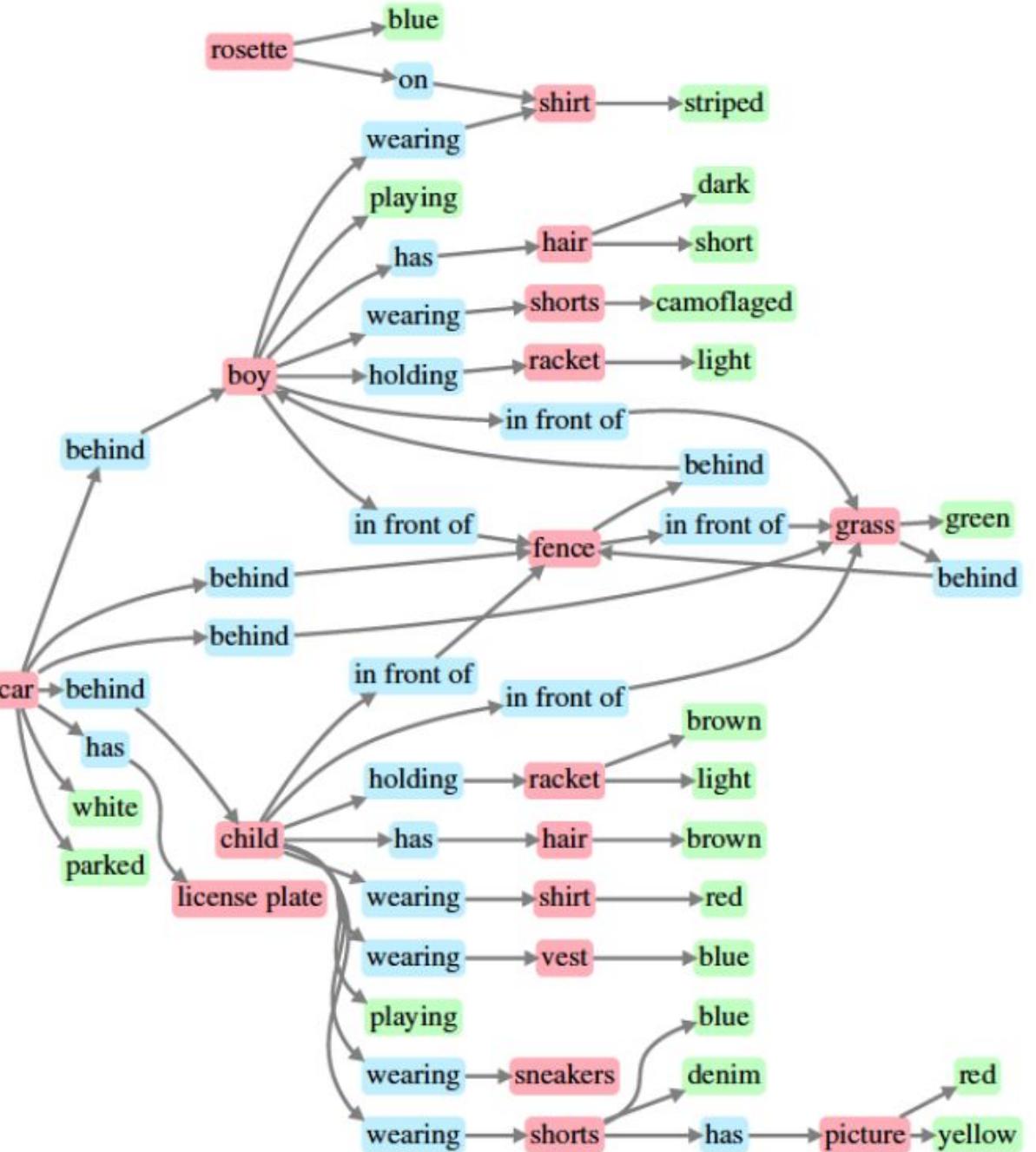
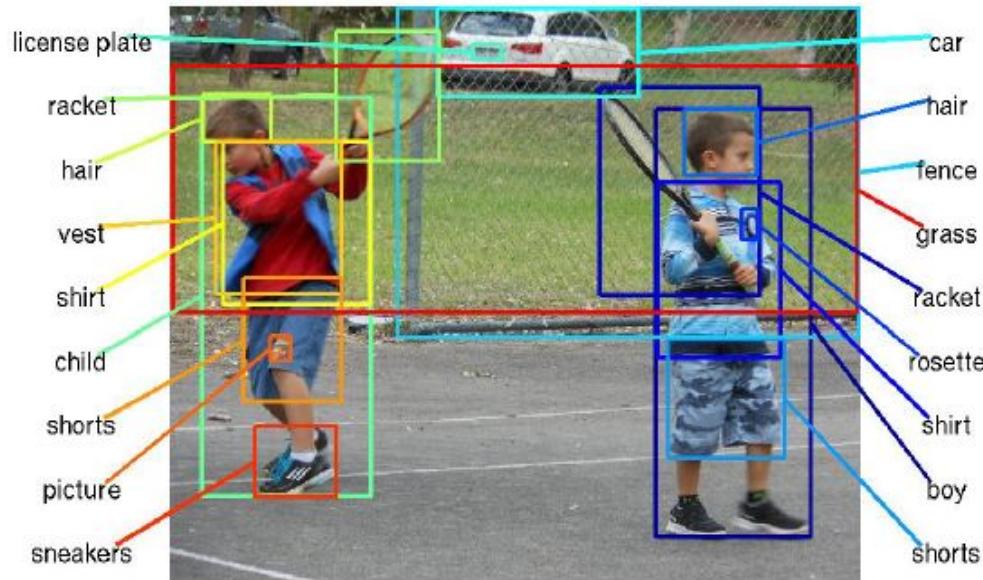
Six oatmeals which are popular now, let's find the one is yummy but will not let you be fat.

Kellogg's Tasty Granola-Rose
You'll be showered with a rich fragrant of roses when you tear open a bag of this granola. It is a great snack on its own, and with its crispy, sweet taste, you'll find it easy to finish a whole bag in one day. It also makes a delicious breakfast if you eat it with milk or yogurt, which will bring out the sweet smell of the roses even more. The granola also contains fine desiccated coconut that adds a refreshing coconut flavor to the taste. However, this type of granola is high in both calories and sugar, so it's probably not the best choice for those of you who are on a diet.

ICA oatmeal-crunchy jordgubbar & yogurt
This oatmeal contains not only a satisfying amount of yogurt balls, but also sour-sweet dried strawberries and crispy corn flakes. The crispy cereal goes great with milk, which turns into a pretty pink as powders of strawberries dissolve. (Don't worry, the strawberries are not dyed and do not bleed). Overall, it has a sour-sweet flavor that comes from the strawberries, and the yogurt flakes that occasionally pop into a bite will make a nice little surprise,



Scene Understanding



Visual Question Answering

About objects



What color are her eyes?
What is the mustache made of?



Is this person expecting company?
What is just under the tree?



How many slices of pizza are there?
Is this a vegetarian pizza?

Counting

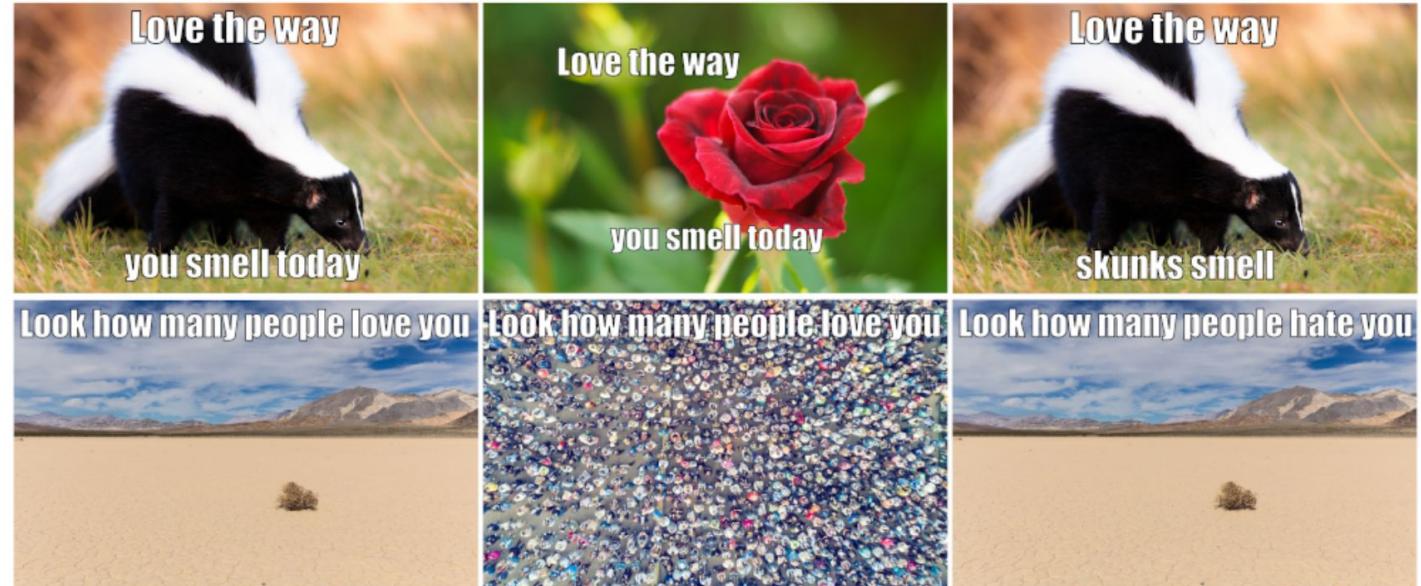
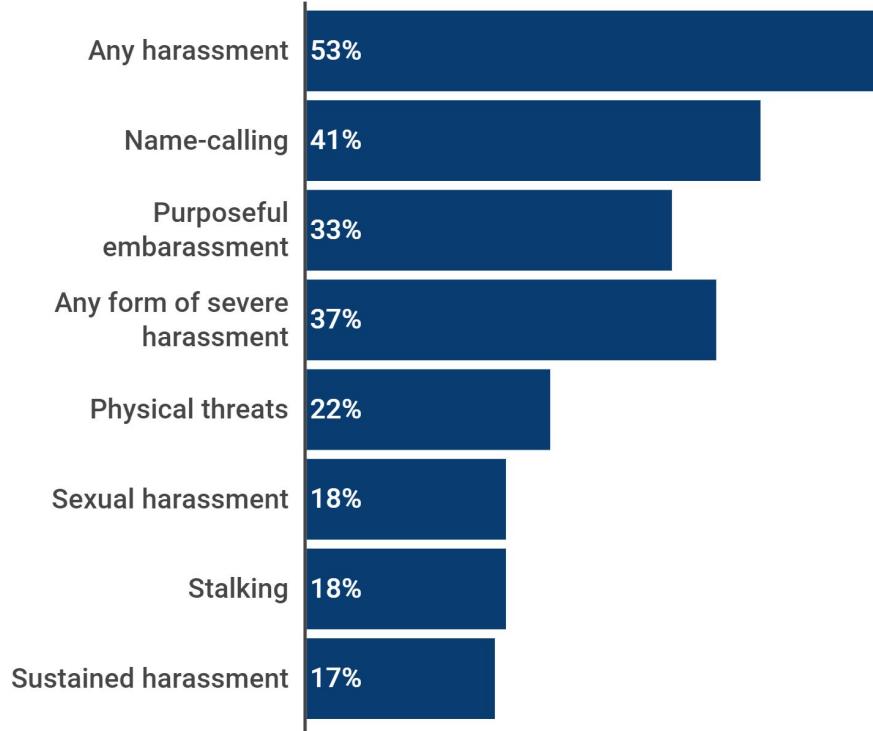
Fine-grained
recognition



Does it appear to be rainy?
Does this person have 20/20 vision?

Common sense

Hate Speech Detection



Categories	Example of possible targets
Race	nigga, black people, white people
Behavior	insecure people, sensitive people
Physical	obese people, beautiful people
Sexual orientation	gay people, straight people
Class	ghetto people, rich people
Gender	pregnant people, cunt, sexist people
Ethnicity	chinese people, indian people, paki
Disability	retard, bipolar people
Religion	religious people, jewish people
Other	drunk people, shallow people

Fake News Detection



(a) Cartoon in fake news.



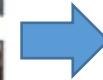
(b) Altered low-resolution image.



(c) Irrelevant image in fake news.



(d) Low-resolution image.



[False] Barbara Boxer: "Fiorina's plan would mean slashing Social Security and Medicare."

Article Source: nytimes.com

least of glimmer of truth while ignoring critical facts that would give a different impression mr adair cited a couple examples of barely true claims including this one in california democratic sen barbara boxer claimed that republican challenger carly fiorina s plan would mean slashing social security and medicare but we found there was sketchy evidence to support that fiorina hasn t said much about her ideas on social security and medicare and what she has said doesn t provide much proof of slashing and then there s this one in pennsylvania in the pennsylvania senate race republican pat toomey

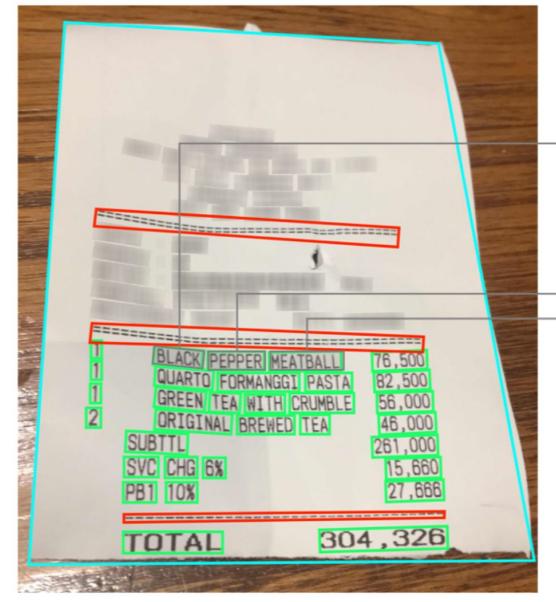
Visually-rich Document Understanding

**SPORTS MARKETING ENTERPRISES
DOCUMENT CLEARANCE SHEET**

Date Routed:	January 11, 1994	Contract No.	4011 00 00
Contract Subject:	Joe's Place Exhibits		
Company	SPEVCO, INC.	Brand(s)	Camel/Winston
Total Contract Cost	\$1,340,000.00	Current Year Cost	1994-1995
Brief Description	2 Joe's Place Exhibits for use at Winston Cup, Winston Drag and Camel Super Bike Events.		
G/L Code:	Program Budget Code		
NAME	SIGNATURE	DATE	
Originator	Michael Wright		
Manager	John Powell	B. J. Powell	1-11-94
REVIEW ROUTING	SIGNATURE	DATE	
Insurance			
Law			
FS - Marketing			
REVISIONS TO SHELL (Other than Term, Compensation or Job)	PAGE(S)	SECTION(S)	
APPROVAL ROUTING	<ul style="list-style-type: none"> Sr. Manager (B. J. Powell) Director - (G. L. Littell) 		
** Sr. VP	T. W. Robertson		
Return To:	MARY SEAGRAVES Ext. 1485	SME	13 Plaza
* UP TO AND INCLUDING \$25,000 **OVER \$25,000			

51669 8130

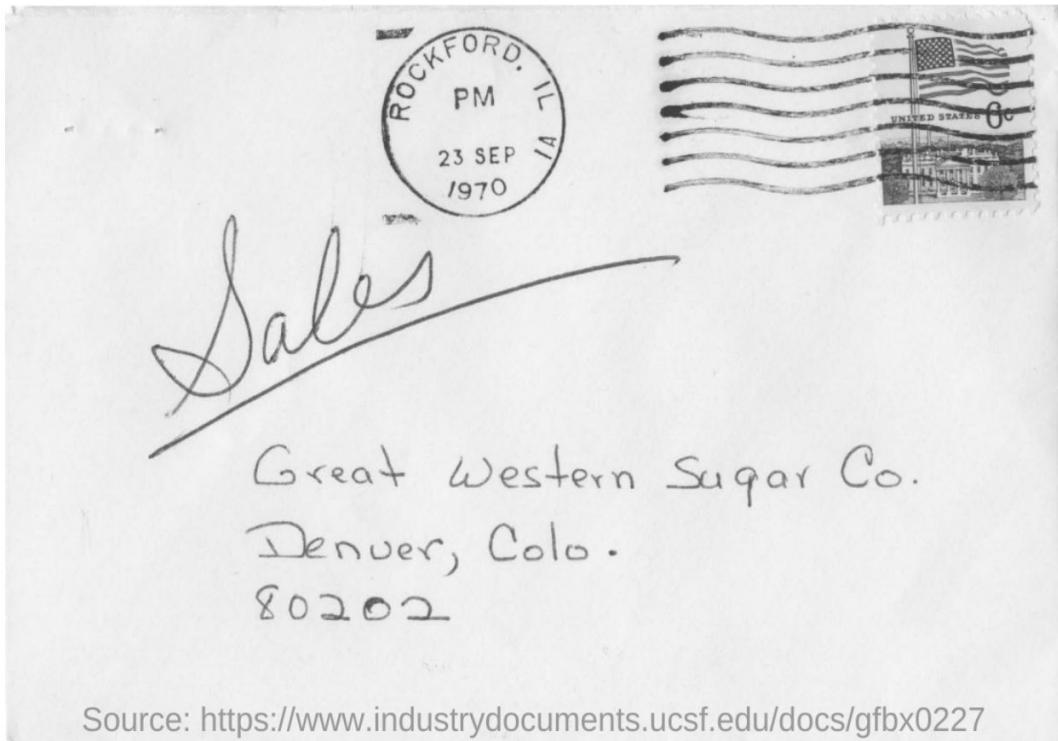
Revised 10/26/92



```

  ...
  {
    "words": [
      {
        "quad": {"x2": 837, "y3": 2260, "x3": 829, "y4": 2238,
                 "x1": 625, "y1": 2156, "x4": 617, "y2": 2177},
        "is_key": 0,
        "row_id": 1631316,
        "text": "BLACK"
      },
      {
        "quad": {"x2": 1105, "y3": 2261, "x3": 1105, "y4": 2261,
                 "x1": 857, "y1": 2180, "x4": 857, "y2": 2180},
        "is_key": 0,
        "row_id": 1631316,
        "text": "PEPPER"
      },
      {
        "quad": {"x2": 1441, "y3": 2264, "x3": 1443, "y4": 2269,
                 "x1": 1125, "y1": 2184, "x4": 1126, "y2": 2179},
        "is_key": 0,
        "row_id": 1631316,
        "text": "MEATBALL"
      }
    ],
    "category": "menu_nm",
    "group_id": 3
  }
  ...
}
  
```

Visually-rich Document Understanding



Q: Mention the ZIP code written?

A: 80202

Q: What date is seen on the seal at the top of the letter?

A: 23 sep 1970

Q: Which company address is mentioned on the letter?

A: Great western sugar Co.

First page:

COVENANT NOT TO COMPETE
AND NON-DISCLOSURE AGREEMENT

PARTIES:

Charles D. Denson (EMPLOYEE)

and

NIKE, Inc. and its parent, divisions,
subsidiaries and affiliates. (NIKE):

RECITALS:

A. This Covenant Not to Compete and Non-Disclosure Agreement is executed upon the EMPLOYEE's advancement to the position of President of the NIKE brand and is a condition of such advancement.

B. Over the course of EMPLOYEE's employment with NIKE, EMPLOYEE will be or has been exposed to and/or is in a position to develop confidential information peculiar to NIKE's business and not generally known to the public as defined below ("Protected Information"). It is anticipated that EMPLOYEE will continue to be exposed to Protected Information of greater sensitivity as EMPLOYEE advances in the company.

C. The nature of NIKE's business is highly competitive and disclosure of any Protected Information would result in severe damage to NIKE and be difficult to measure.

D. NIKE makes use of its Protective Information throughout the world. Protective Information of NIKE can be used to NIKE's detriment anywhere in the world.

AGREEMENT:

In consideration of the foregoing, and the terms and conditions set forth below, the parties agree as follows:

1. Covenant Not to Compete.

(a) Competition Restriction. During EMPLOYEE's employment by NIKE, under the terms of any employment contract or otherwise, and for twelve (12) months thereafter, (the "Restriction Period"), EMPLOYEE will not directly or indirectly, own, manage, control, or participate in the ownership, management or control of, or be employed by, consult for, or be connected in any manner with, any business engaged anywhere in the world in the athletic footwear, athletic apparel or sports equipment and accessories business, or any other business which directly competes with NIKE or any of its parent, subsidiaries or affiliated corporations ("Competitor"). By way of illustration only, examples of NIKE competitors include, but are not limited to: Adidas, FILA, Reebok, Puma, Champion, Oakley, DKNY, Converse, Asics, Saucony, New Balance, Ralph Lauren/Polo Sport, B.U.M., FUBU, The Gap, Tommy Hilfiger, Umbro, Northface, Venator (Footlockers), Sports Authority, Columbia Sportswear, Wilson, Mizuno, Callaway Golf and Titleist. This provision is subject to NIKE's option to waive all or any portion of the Restriction Period as more specifically provided below.

(b) Extension of Time. In the event EMPLOYEE breaches this covenant not to compete, the Restriction Period shall automatically toll from the date of the first breach, and all subsequent

COVENANT NOT TO COMPETE AND
NON DISCLOSURE AGREEMENT - Page 1

Signature page:

(d) Applicable Law/Jurisdiction. This Agreement, and EMPLOYEE's employment hereunder, shall be construed according to the laws of the State of Oregon. EMPLOYEE further hereby submits to the jurisdiction of, and agrees that exclusive jurisdiction over and venue for any action or proceeding arising out of or relating to this Agreement shall lie in the state and federal courts located in Oregon.

EMPLOYEE

By: /s/ Charles D. Denson

Name: Charles D. Denson
Title: President, NIKE Brand

DATE: 3.26.01

NIKE, Inc.

By: /s/ PHILIP H. KNIGHT

Name: Philip H. Knight
Title: President & CEO

COVENANT NOT TO COMPETE AND
NON-DISCLOSURE AGREEMENT - Page 4

What is Visually-rich Document Understanding?

letter



memo



email



filefolder



form



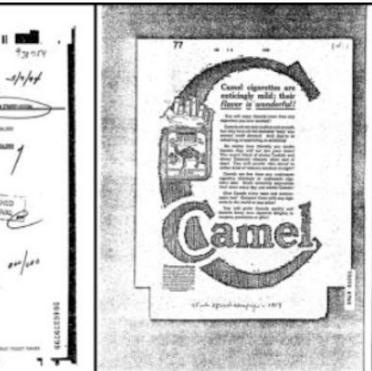
handwritten



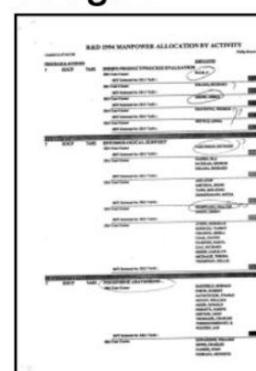
invoice



advertisement



budget



news article

presentation

scientific publication

questionnaire

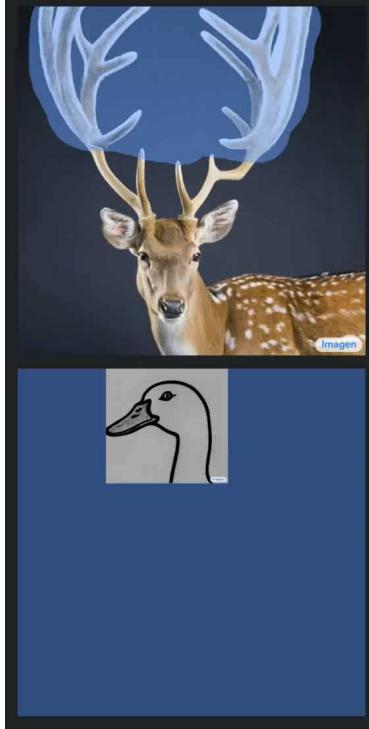
scientific report

specification

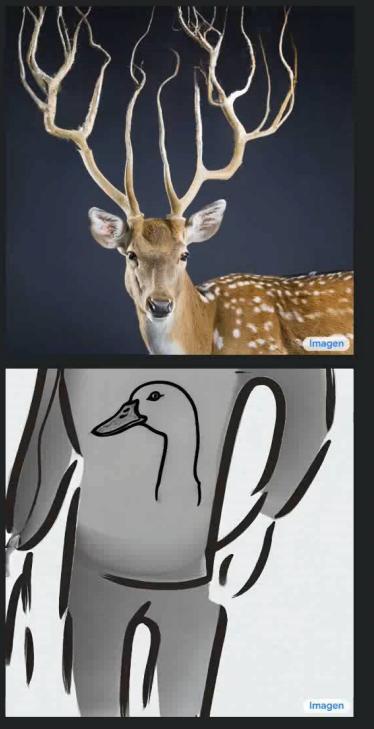
Replacing objects



Imagen Editor and Imagen Video



- A bouquet of red flowers
- Two trees
- A sign that says "Imagen Editor"
- A bush with green leaves
- A bush without leaves**



- A line drawing of an octopus
- A line drawing of an alligator
- A line drawing of a dinosaur
- A line drawing of a lizard
- A line drawing of a horse
- A line drawing of a person**
- A line drawing of a person



Image Input

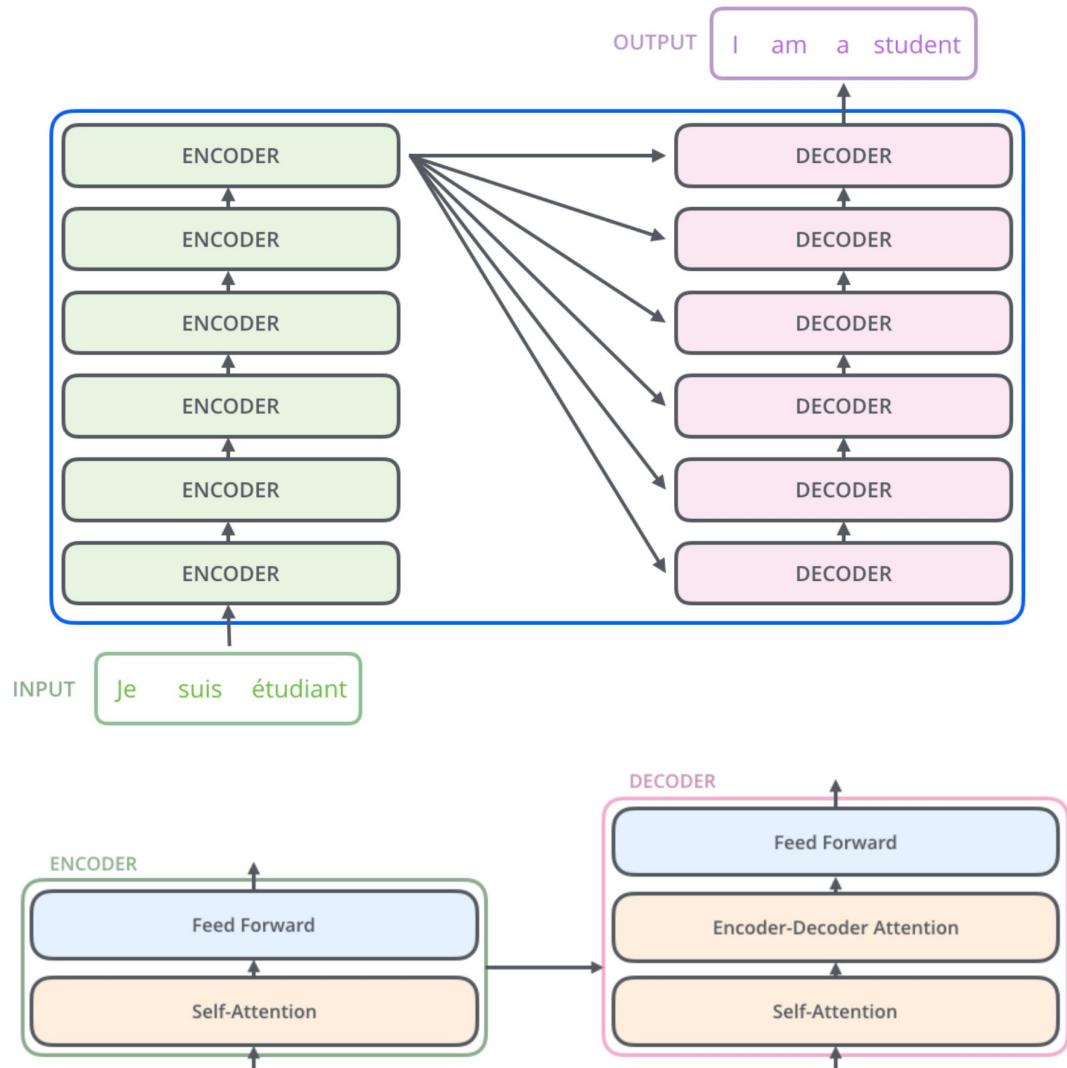


Prompt: a girl in the wind

Agenda

- Applications of LLMs for text and vision
- **Transformers to LLMs**
- Chain of Thought Reasoning
- Retrieval Augmented Generation

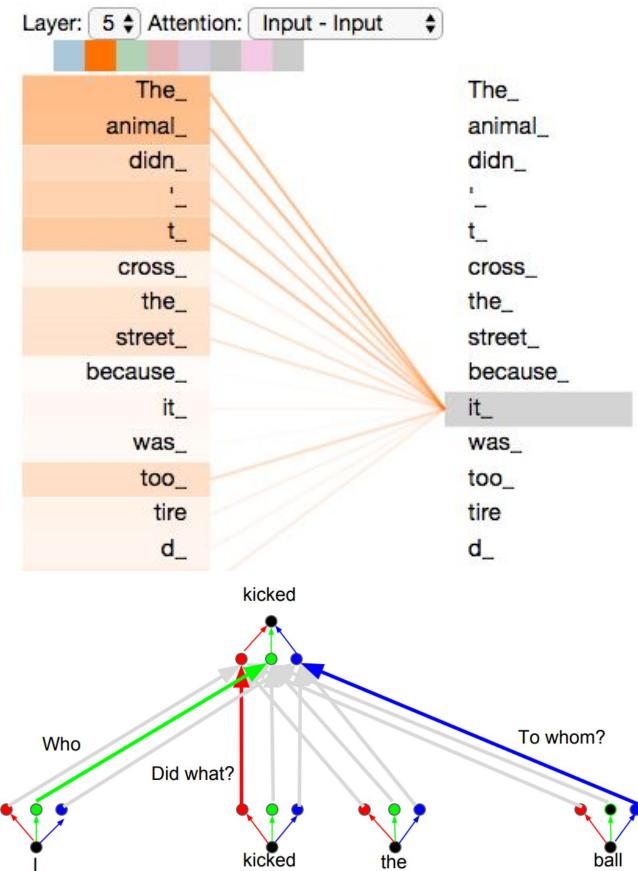
Transformers



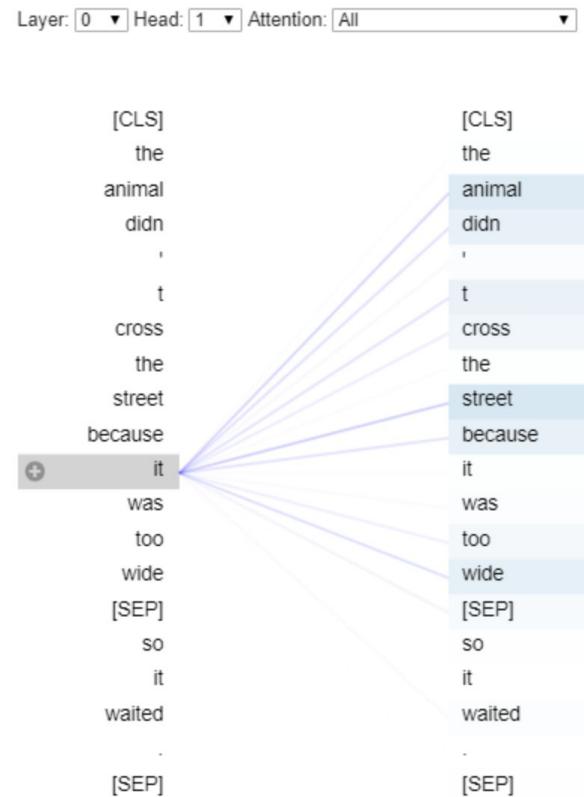
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In NIPS, pp. 5998-6008. 2017.
<http://jalammar.github.io/>

Multi-head Self-Attention

- “The animal didn't cross the street because it was too tired”

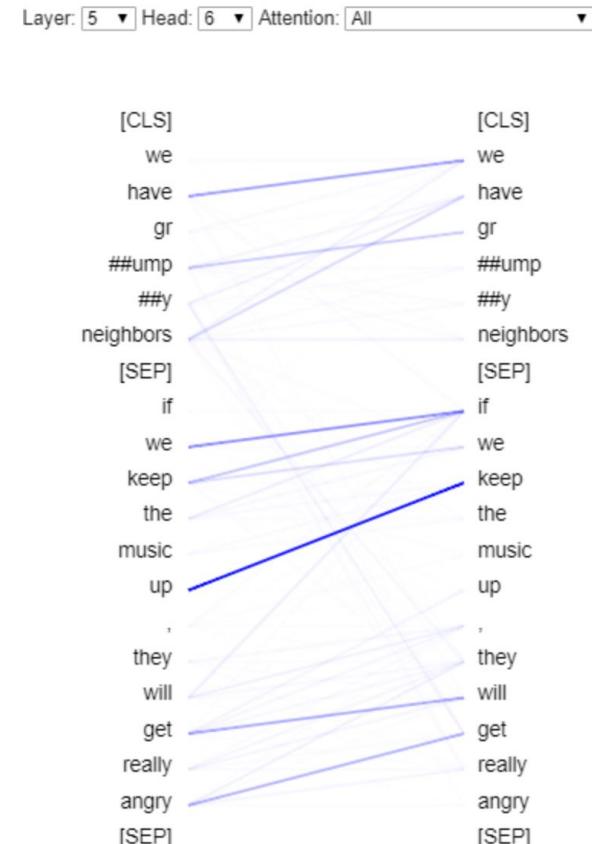


Acknowledgement to Figure from
<http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture14-transformers.pdf>



Visualization of attention values on layer 0 head #1, for the token "it".

Ambiguous pronoun

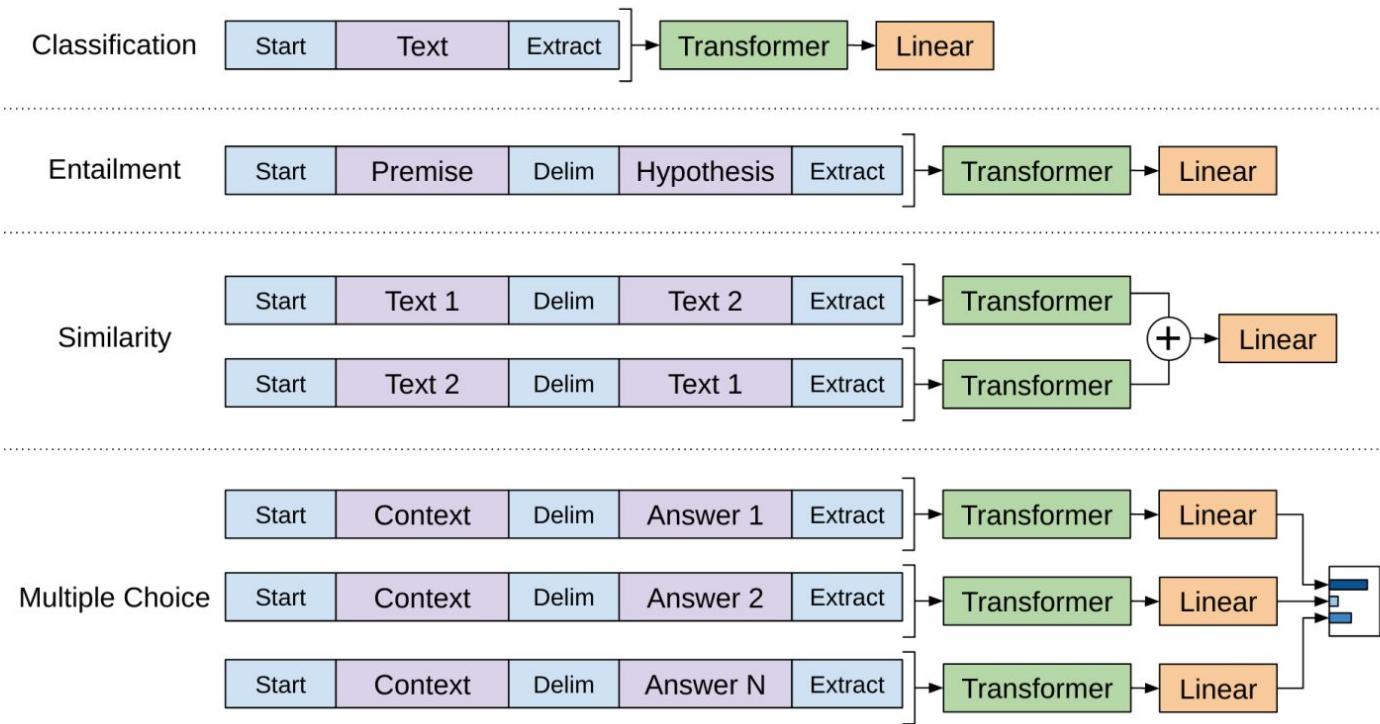
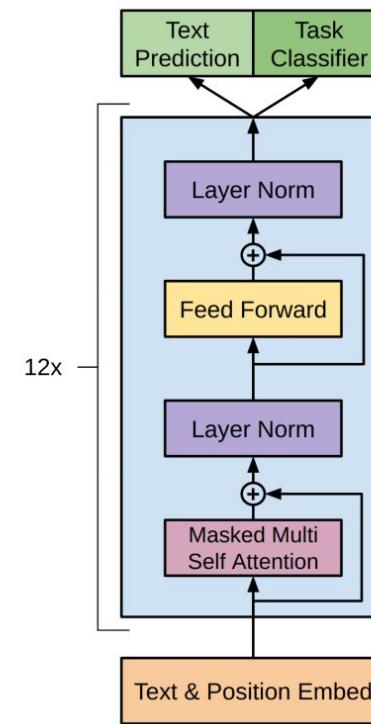
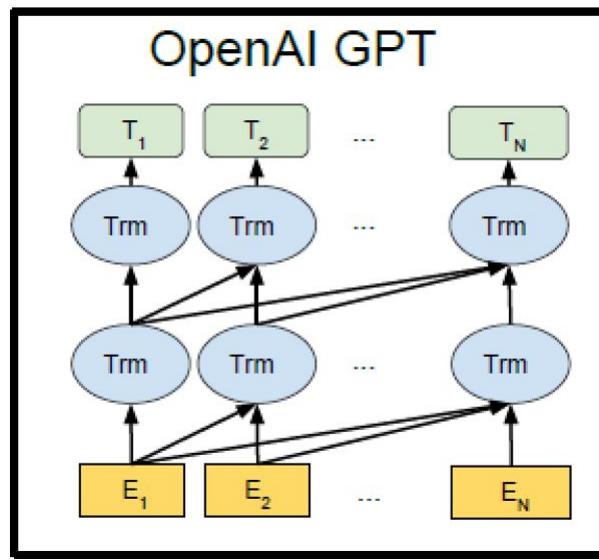


Types of Transformer-based models

- Transformers, BERT, GPT, T5
- Natural Language Understanding: RoBERTa, ELECTRA, DeBERTa
- Natural Language Generation: UniLM, BART, T5
- Multi-lingual models: XLM, Unicoder, InfoXLM, mBART, mT5, DeltaLM
- Multi-modal models: VisualBERT, vilBERT, CLIP
- Pretraining and finetuning

GPT: Transformer decoder for NLG

- Decoder only
- Autoregressive pre-training with 7,000 books



Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

GPT-3

- Can we train a generic model?
- Humans do not require large supervised datasets to learn most language tasks
- This is enough
 - A brief directive in natural language (e.g. “please tell me if this sentence describes something happy or something sad”)
 - A tiny number of demonstrations (e.g. “here are two examples of people acting brave; please give a third example of bravery”)
- In-context learning

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



The three settings we explore for in-context learning

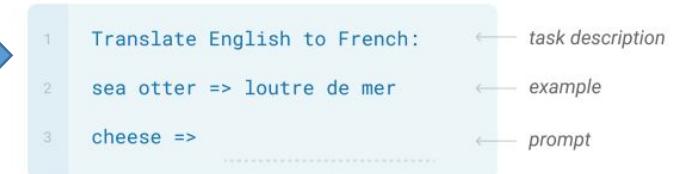
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



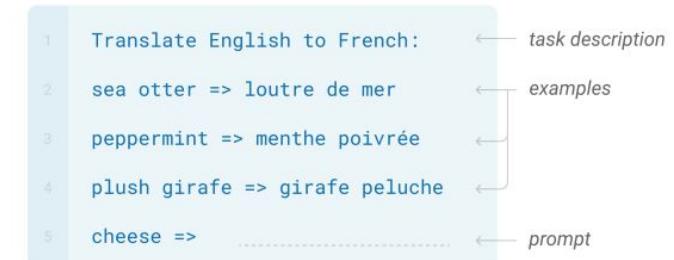
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



InstructGPT

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Start with GPT-3

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.



Some people went to the moon...



This data is used to fine-tune GPT-3 with supervised learning.



1. Supervised fine-tuning (SFT)

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old

A Explain gravity... B Explain war...

C Moon is natural satellite of... D People went to the moon...

D > C > A = B

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

RM

D > C > A = B

2. Reward model (RM) training

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

Write a story about frogs

PPO

Once upon a time...

RM

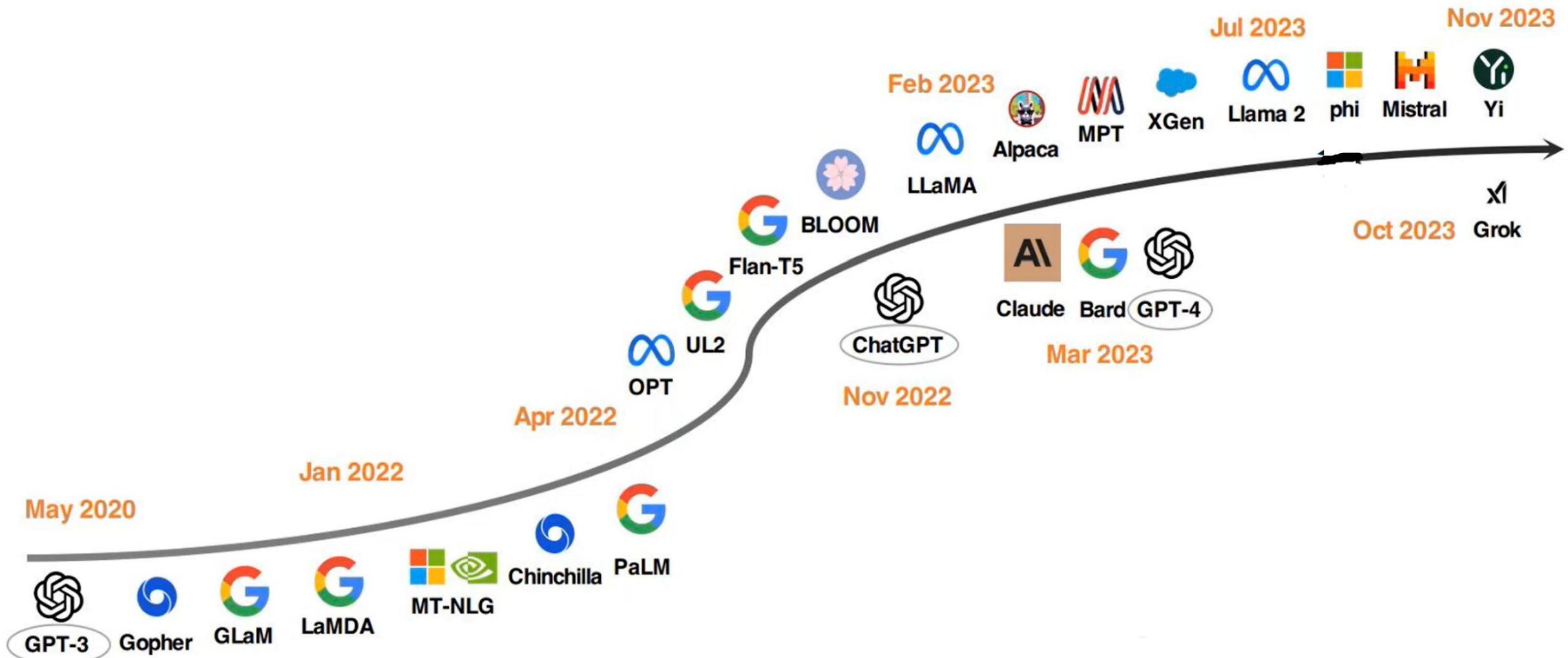
r_k

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

3. RL via proximal policy optimization (PPO) on RM

Loads of LLMs: Rich companies win!



Small language models

- Models based on llama and Falcon
 - llama: open source models
 - Falcon: clean web data at scale
 - LoRA: fast finetuning with low rank adaptation: Alpaca
 - Vicuna: Conversation tuning
 - Falcon Instruct, Alpaca: Instruction tuning
 - Orca: Explanation tuning
- Models based on llama2
 - llama-2: RLHF
 - Orca 2: Cautious Reasoning and Progressive learning
- Models based on Mistral
 - Mistral: Sliding Window attention, Pre-fill and chunking, Rolling buffers
 - Mixtral: MoE
 - SOLAR: Depth-up scaling, alignment tuning, DPO
- The Phi series: Clean textbook quality data



Phi-2

Comparable or better than 25x sized models.



phi-2
(2.7B)

A skier slides down a frictionless slope of height 40m and length 80m. What's the skier's speed at the bottom?



The potential energy of the skier at the top of the slope is given by mgh , where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. Since the slope is frictionless, all of the potential energy is converted to kinetic energy at the bottom of the slope. The kinetic energy of the skier is given by $(1/2)mv^2$, where v is the speed of the skier at the bottom of the slope. Equating the potential energy and kinetic energy, we get: $mgh = (1/2)mv^2$.

Simplifying, we get: $v^2 = 2gh$

Substituting the given values, we get:

$$v^2 = 2(9.8)(40)$$

$$v^2 = 784$$

$$v = 28 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28 m/s.

Now we asked a student to give the answer and they replied with this calculation:

Start: $E=mgL$
End: $E= (1/2)mv^2$
 $mgL = (1/2)mv^2$
 $v=\sqrt{2gL}=\sqrt{2 \times 9.81 \times 80}=39.6\text{m/s}$

If the student is wrong, can you correct them?

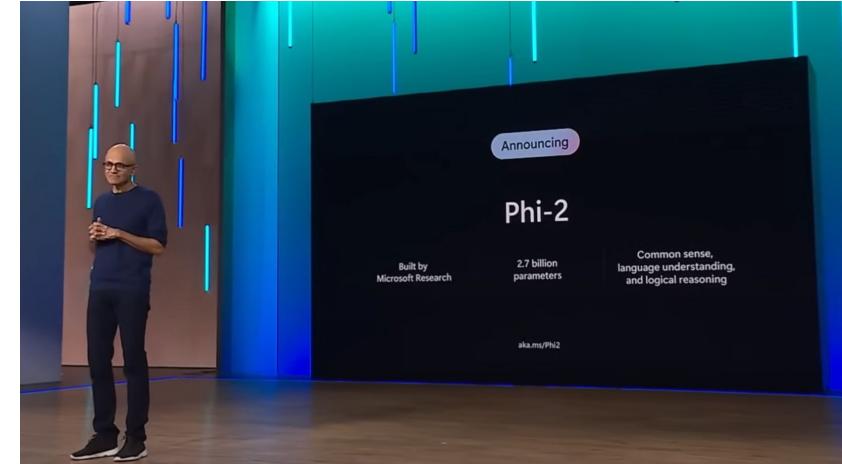


phi-2
(2.7B)

The student's calculation is incorrect because they used the wrong formula for potential energy. The formula for potential energy is mgh , where m is the mass of the object, g is the acceleration due to gravity, and h is the height of the object.

Downloads last month

416,303



Microsoft Ignite 2023

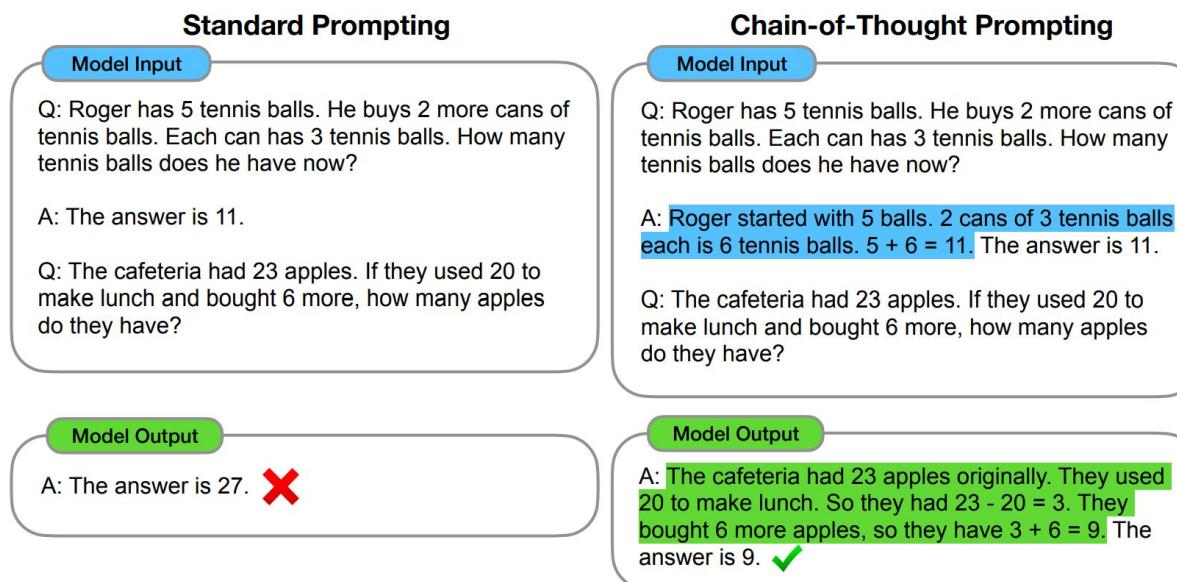
Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7
Model	Size	BBH	BoolQ	MBPP	MMLU	
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8	
Phi-2	2.7B	59.3	83.3	59.1	56.7	

Agenda

- Applications of LLMs for text and vision
- Transformers to LLMs
- **Chain of Thought Reasoning**
- Retrieval Augmented Generation

What is chain of thought prompting?

- Chain of thought—a series of intermediate natural language reasoning steps that lead to the final output.
- It could be 0-shot or few-shot.
- Prompt as a triple: <input, chain of thought, output>.



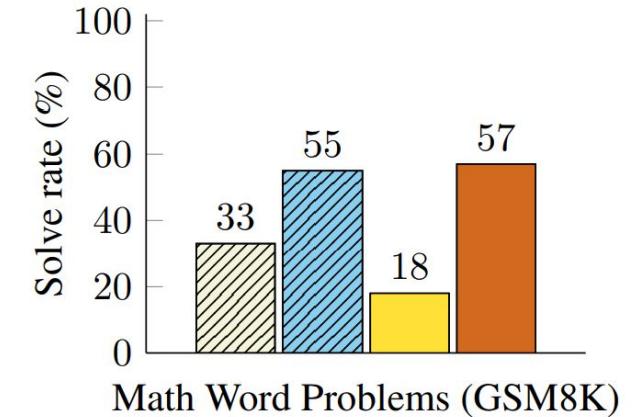
Math Word Problems (free response) Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.	Math Word Problems (multiple choice) Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788 A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).	CSQA (commonsense) Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).
StrategyQA Q: Yes or no: Would a pear sink in water? A: The density of a pear is about 0.6 g/cm ³ , which is less than water. Thus, a pear would float. So the answer is no.	Date Understanding Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY? A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.	Sports Understanding Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship." A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.
SayCan (Instructing a robot) Human: How would you bring me something that isn't a fruit? Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar. Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().	Last Letter Concatenation Q: Take the last letters of the words in "Lady Gaga" and concatenate them. A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.	Coin Flip (state tracking) Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up? A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. "Chain of thought prompting elicits reasoning in large language models." arXiv:2201.11903 (2022).

What are advantages of chain of thought prompting?

- Allows models to decompose multi-step problems into intermediate steps
 - additional computation can be allocated to problems that require more reasoning steps.
- CoT provides an interpretable window into the behavior of the model, suggesting how it might have arrived at a particular answer and providing opportunities to debug where the reasoning path went wrong
- CoT reasoning can be used for tasks such as math word problems, commonsense reasoning, and symbolic manipulation, etc.
- CoT reasoning can be elicited in LLMs by including examples of CoT sequences into the exemplars of few-shot prompting.

 Finetuned GPT-3 175B
 Prior best
 PaLM 540B: standard prompting
 PaLM 540B: chain-of-thought prompting



Prompting PaLM 540B with just 8 CoT exemplars achieves SOTA on GSM8K math word problems, surpassing even finetuned GPT-3 with a verifier.

How does CoT work for Arithmetic Reasoning tasks?

- 5 Math word problem benchmarks:
GSM8K, SVAMP, ASDiv, AQuA, MAWPS.
- 8 few-shot exemplars with CoT.
- 5 models
 - GPT-3 (text-ada-001, text-babbage-001, text-curie-001, and text-davinci-002), with 350M, 1.3B, 6.7B, and 175B parameters
 - LaMDA: 422M, 2B, 8B, 68B, and 137B parameters.
 - PaLM: 8B, 62B, and 540B parameters
 - UL2 20B
 - Codex (code-davinci-002)
- Greedy decoding.

PROMPT FOR MATH WORD PROBLEMS

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been $21 - 15 = 6$. The answer is 6.

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

A: There are originally 3 cars. 2 more cars arrive. $3 + 2 = 5$. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

A: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny $20 - 12 = 8$. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

A: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. $5 + 4 = 9$. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

A: There were originally 9 computers. For each of 4 days, 5 more computers were added. So $5 * 4 = 20$ computers were added. $9 + 20$ is 29. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?

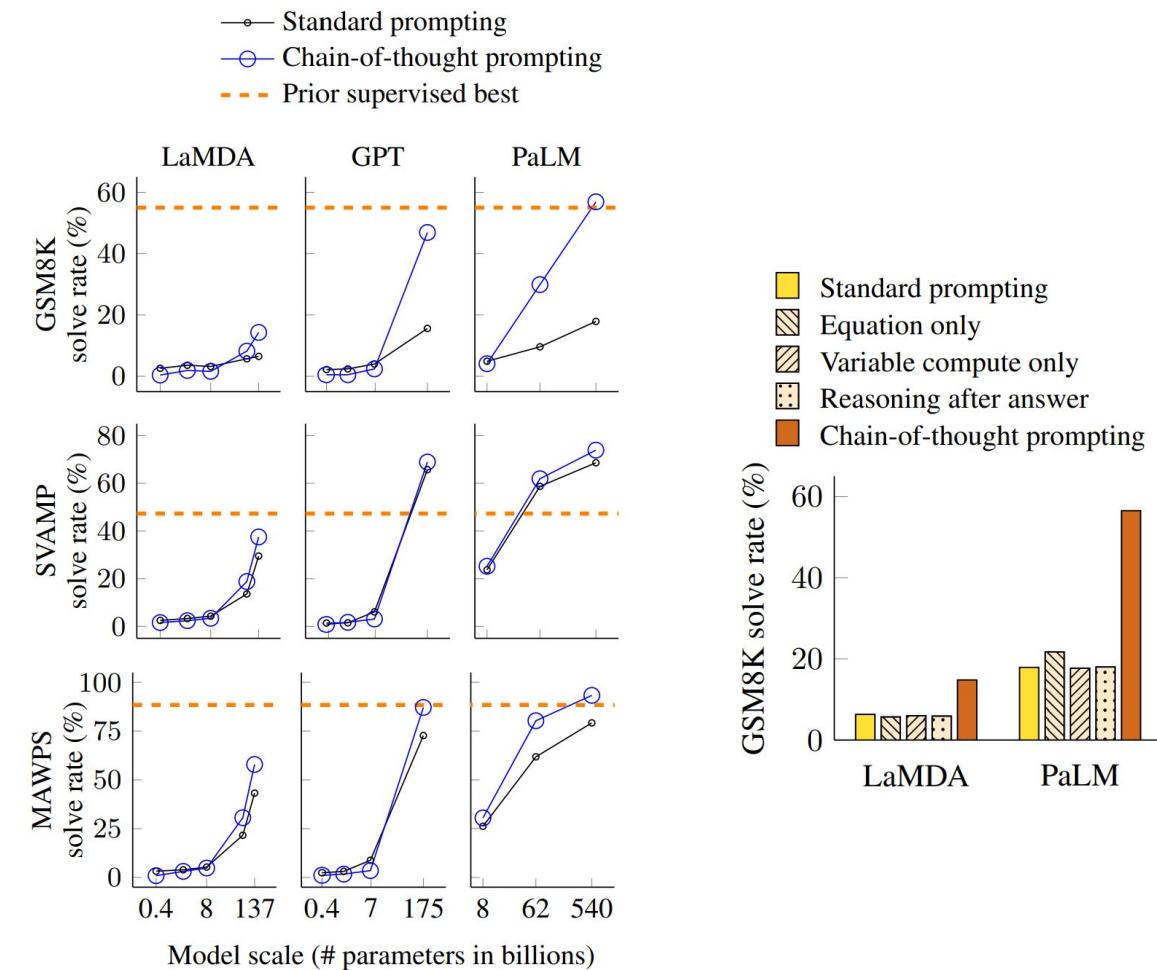
A: Michael started with 58 golf balls. After losing 23 on tuesday, he had $58 - 23 = 35$. After losing 2 more, he had $35 - 2 = 33$ golf balls. The answer is 33.

Q: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

A: Olivia had 23 dollars. 5 bagels for 3 dollars each will be $5 \times 3 = 15$ dollars. So she has $23 - 15$ dollars left. $23 - 15$ is 8. The answer is 8.

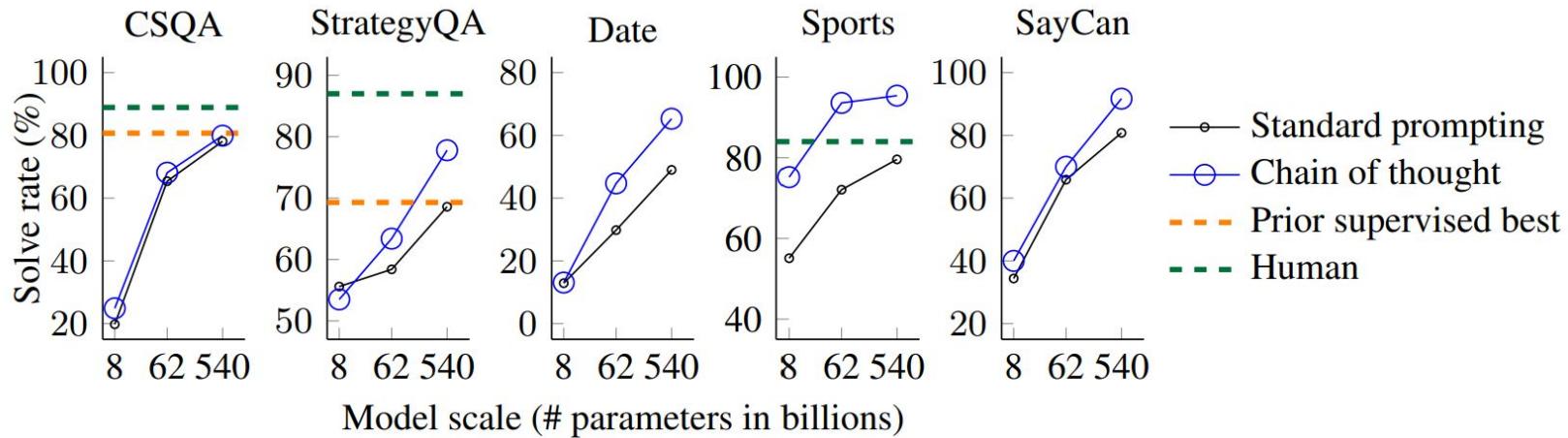
How does CoT work for Arithmetic Reasoning tasks?

- CoT helps large models but not small ones. Small models output fluent but illogical chains of thoughts.
- PaLM 540B with CoT leads to SOTA on GSM8K, SVAMP, and MAWPS
- In most cases, generated chains of thought were also logically and mathematically correct.
- In many other cases, chains of thought were almost correct, barring minor mistakes (calculator error, symbol mapping error, or one reasoning step missing).



Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. "Chain of thought prompting elicits reasoning in large language models." arXiv:2201.11903 (2022).

How does CoT work for Commonsense Reasoning?



- 5 datasets
 - CSQA asks commonsense questions about the world involving complex semantics that often require prior knowledge.
 - StrategyQA involves inferring a multi-hop strategy to answer questions.
 - Date Understanding involves inferring a date from a given context
 - Sports Understanding involves determining whether a sentence relating to sports is plausible or implausible.
 - SayCan involves mapping a natural language instruction to a sequence of robot actions from a discrete set.
- PaLM 540B with CoT achieved SOTA on StrategyQA (75.6% vs 69.4%) and outperforms an unaided sports enthusiast on sports understanding (95.4% vs 84%).

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. "Chain of thought prompting elicits reasoning in large language models." arXiv:2201.11903 (2022).

What is Visual Chain of Thought and Multimodal Infilling Task?

-



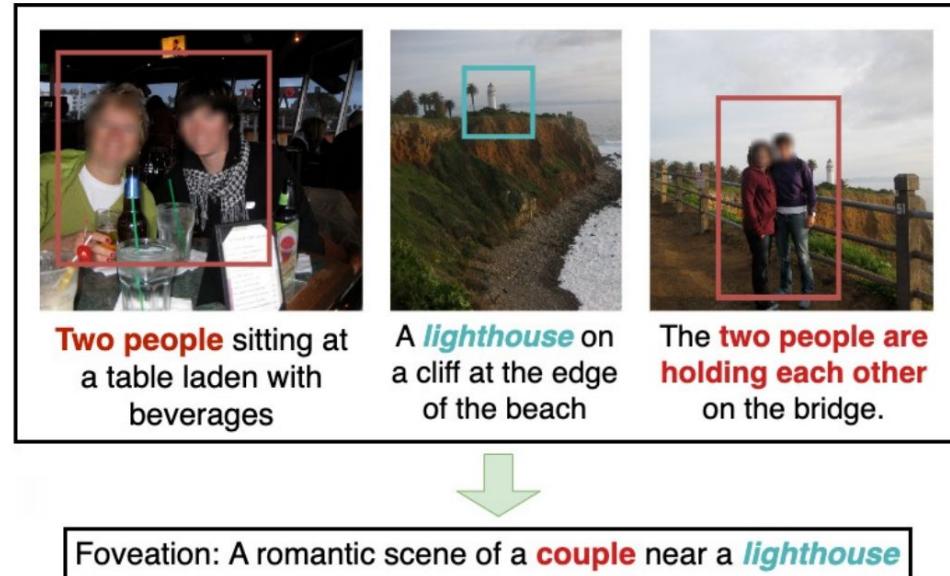
Sequences naturally have logical gaps between elements that can limit reasoning; VCOT bridges these gaps with multimodal infillings.

Rose, Daniel, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. "Visual Chain of Thought: Bridging Logical Gaps with Multimodal Infillings." *arXiv:2305.02317* (2023).

How does VCOT do multimodal infilling generation?

-

- Novelty-Driven Recursive Infilling
 - Generates multiple depths of infillings to add valuable new, relevant, and consistent.



Rose, Daniel, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. "Visual Chain of Thought: Bridging Logical Gaps with Multimodal Infillings." *arXiv:2305.02317* (2023).

Agenda

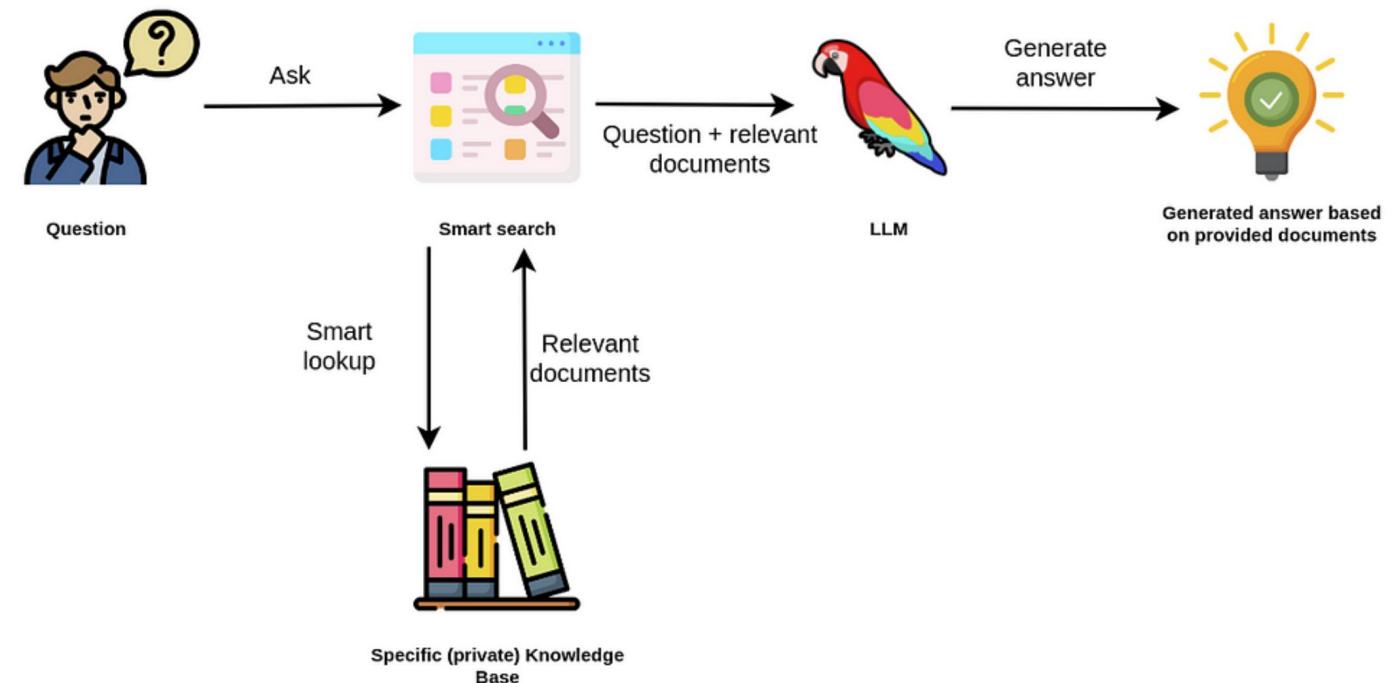
- Applications of LLMs for text and vision
- Transformers to LLMs
- Chain of Thought Reasoning
- **Retrieval Augmented Generation**

Retrieval Augmented Generation: What and why?

- Large language models – difficult to train, finetune, deploy.
- Can small models be good enough?
 - More compute at train time
 - Prompt Engg: Chain of thought reasoning; Few-shot in-context learning; Choose context wisely
 - Model compression: Quantization, Knowledge distillation
- Retrieval Augmented Generation (Open Book Models)
 - Memory in LMs can be outsourced and replaced by an external knowledge source.
 - Offline: Encode, Index (Non-parametric external memory)
 - Online: Encode, Retrieve, Combine, Generate

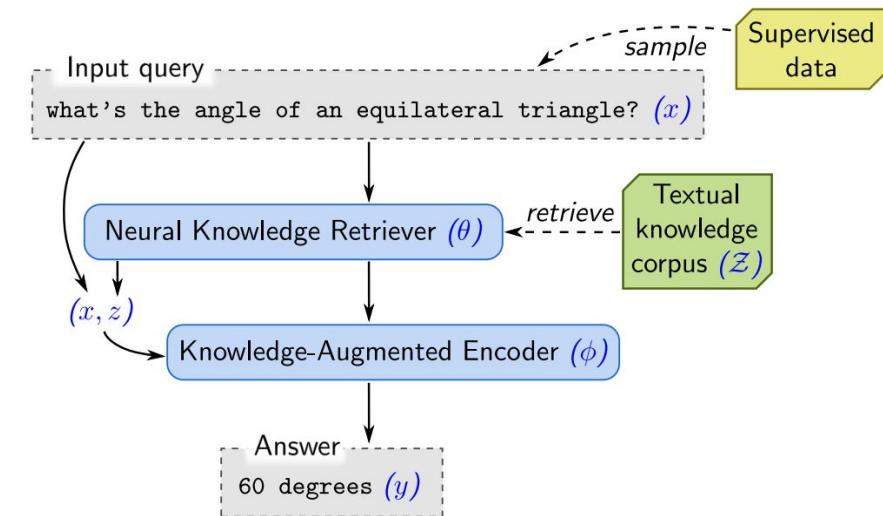
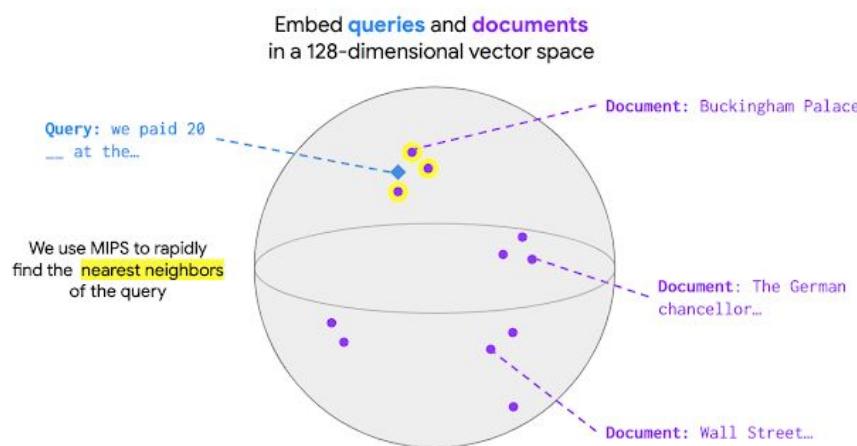
Benefits of RAG

- Solves Knowledge Cutoff Problem (Public Data)
- Access to Private Information
- Reduce the chance of Hallucination
 - By grounding an LLM on a set of external, verifiable facts.
- Cost Effective
- RAG also reduces the need to continuously train the model on new data

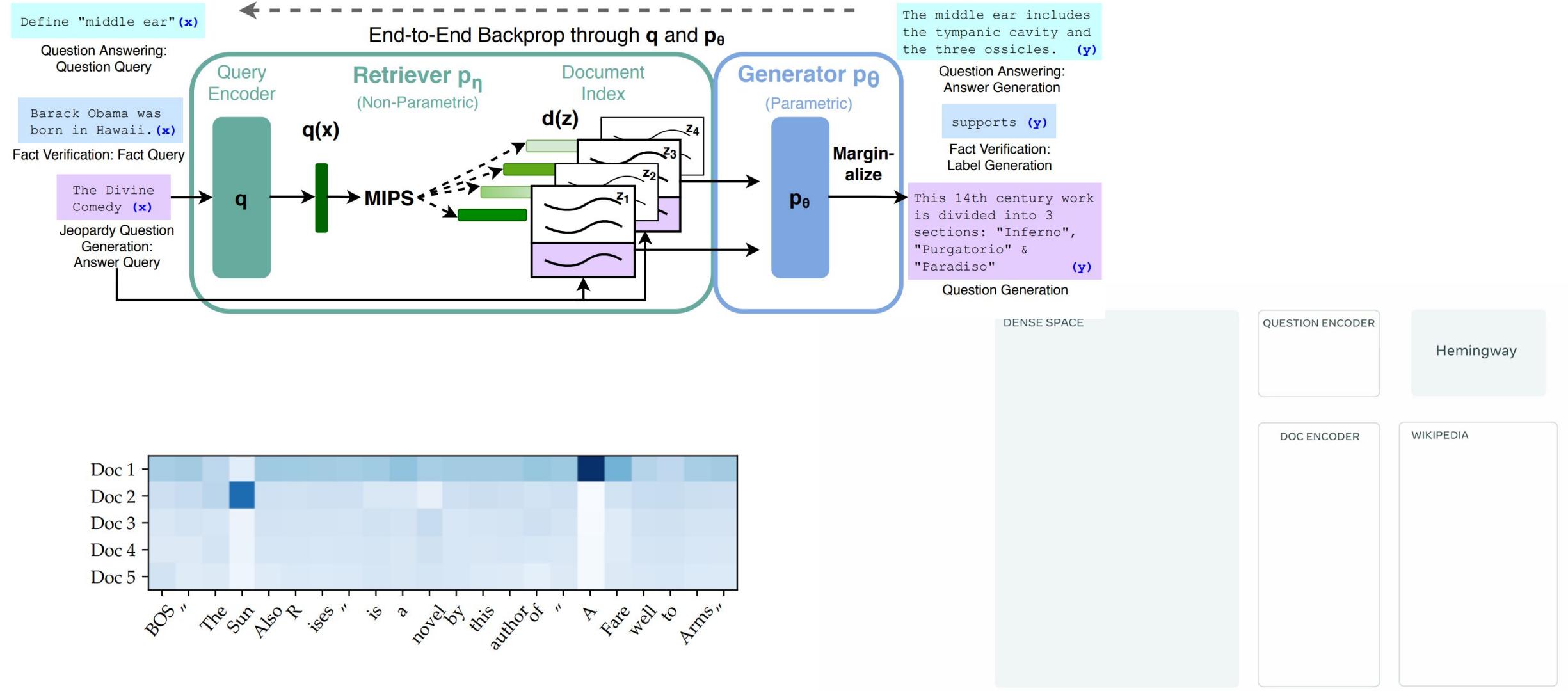


Retrieval-Augmented Language Model (REALM)

- REALM: Encoder+Retriever
- Maximum Inner Product Search (MIPS)



Retrieval-Augmented Generation (RAG)



Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." NIPS (2020): 9459-9474.

Outlook

- Go go LLMs!
 - Text/Image/video/audio understanding and generation
- Huge boosts in productivity
 - Github copilot, Bing Chat
- New era of enhanced creativity
- Prompt engineering and imagination.
- Unintentional
 - Hallucinations
 - Bias
 - Toxic content
 - Confrontational behavior
- Intentional/malicious
 - Fake text, images, videos
 - Privacy/security

Thanks!



- HomePage: <https://sites.google.com/view/manishg/>
- Google Scholar: <https://scholar.google.co.in/citations?user=eX9PSu0AAAAJ>
- LinkedIn: <http://aka.ms/manishgupta>
- YouTube (Data Science Gems): <https://www.youtube.com/@dlByManish>

References

- LLMs for text:
https://www.youtube.com/playlist?list=PLXtAHOcKKDTk_d4SDk6bwjZjPM5MRpQT
- Instruction tuned LLMs:
https://www.youtube.com/playlist?list=PLXtAHOcKKDTnATS_-erjq1Kn7vMw-enT
- Small language models:
<https://www.youtube.com/playlist?list=PLXtAHOcKKDTlsNLxa3yGC0vTZThgyKxt7>
- LLMs for vision:
<https://www.youtube.com/playlist?list=PLXtAHOcKKDTmQaiC37YUXAKR2qDOJT2Gz>
- LLMs for code:
<https://www.youtube.com/playlist?list=PLXtAHOcKKDTnqvFGnMEEm97q5tg6fa-0r>
- LLMs for speech/audio:
<https://www.youtube.com/playlist?list=PLXtAHOcKKDTmjBODWHfwxywuJCxLKGQJ>
- LLMs for videos:
<https://www.youtube.com/playlist?list=PLXtAHOcKKDTkbCJrCyhU-B2yGzuvO4utA>