

# Brief Introduction to Natural Language Processing & IR

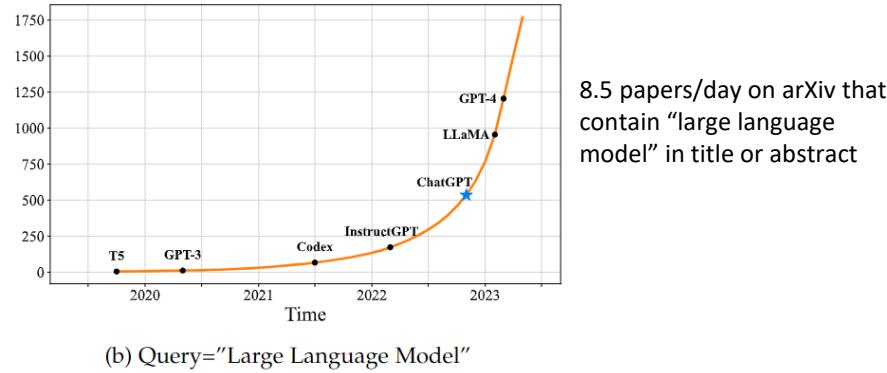
Adam Jatowt

24 May 2024

# Natural Language Processing

# Natural Language Processing

- NLP is a large field that focuses on **engineering aspects of language**
- Not only **analyzing** but also **generating** and, in general, **processing text** for large number of applications
- Lots of tasks
- Domination of LLMs



# Example Syntactic Tasks

# Word Segmentation

- Breaking a string of characters into a sequence of words.
- In some written languages (e.g. Chinese, Japanese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : () ]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar  
⇒ myspace .com pluckers wing bar  
⇒ myspace .com plucker swing bar

# Morphological Analysis

- **Morphology** is the field of linguistics that studies the internal structure of words (Wikipedia)
- A **morpheme** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. "carry", "pre", "ed", "ly", "s"
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried  $\Rightarrow$  carry + ed (past tense)
  - independently  $\Rightarrow$  in + (depend + ent) + ly
  - Googlers  $\Rightarrow$  (Google + er) + s (plural)
  - unlockable  $\Rightarrow$  un + (lock + able) ?  
 $\Rightarrow$  (un + lock) + able ?

# Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I ate the spaghetti with meatballs.

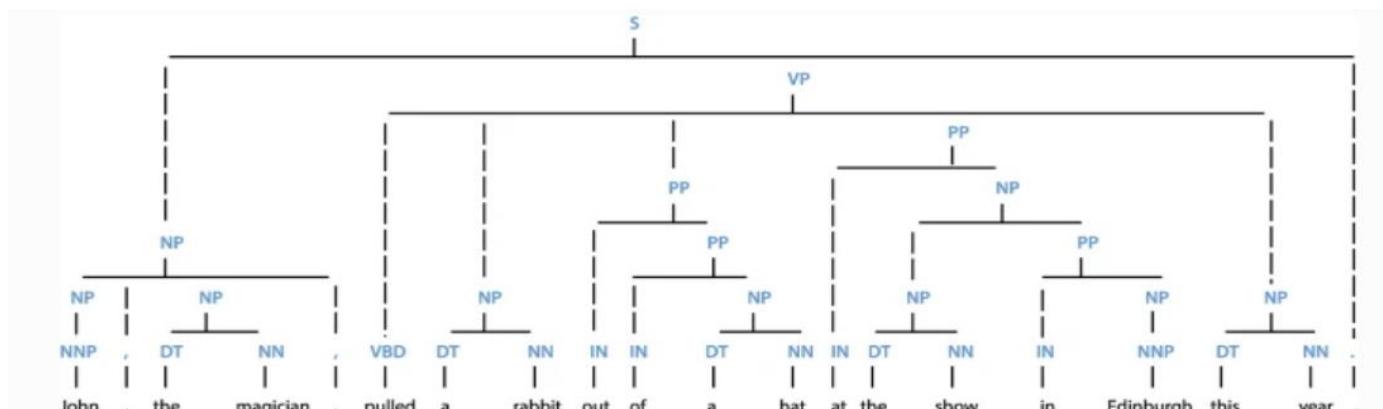
Pro V Det N Prep N

John saw the saw and decided to take it to the table.

PN V Det N Con V Part V Pro Prep Det N

- Useful for subsequent syntactic parsing and word sense disambiguation.

# Phrase Chunking



Results of applying the benepar chunking model to the sentence: "John, the magician, pulled a rabbit out of a hat at the show in Edinburgh this year."

# Example Semantic Tasks

# Word Sense Disambiguation (WSD)

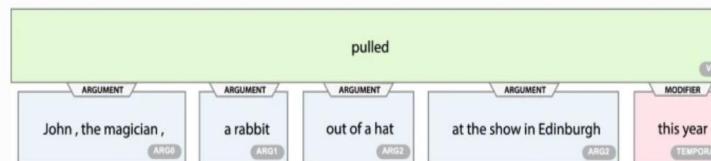
- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong **interest** in computational linguistics.
  - Ellen pays a large amount of **interest** on her credit card.
- For many tasks (e.g., question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

# Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

agent patient source destination instrument

- John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.
- 
- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”



Semantic role labeling result when applied to the sentence: "John, the magician, pulled a rabbit out of a hat at the show in Edinburgh this year."

# Textual Entailment (aka. Natural Language Inference or NLI)

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation

# Textual Entailment Problems in PASCAL Challenge

TEXT	HYPOTHESIS	ENTAILMENT
<p><i>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</i></p>	<p><i>Yahoo bought Overture.</i></p>	TRUE
<p><i>Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</i></p>	<p><i>Microsoft bought Star Office.</i></p>	FALSE
<p><i>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</i></p>	<p><i>Israel was established in May 1971.</i></p>	FALSE
<p><i>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</i></p>	<p><i>Israel was established in 1948.</i></p>	TRUE

# NLI example

Premise	Relation	Hypothesis
A soccer game with multiple males playing.	Entailment	Some men are playing a sport.
A black race car starts up in front of a crowd of people.	Contradiction	A man is driving down a lonely road.
A smiling costumed woman is holding an umbrella.	Neutral	A happy woman in a fairy costume holds an umbrella.

# Other Kinds of NLI: Time-related inference in natural language

S1: I am cooking pasta Bolonese for the first time

S2: The water is not hot yet

S1: I am cooking pasta Bolonese for the first time

S2: It was really tasty and I need a rest now

In which case, S2 implies the end of action expressed in S1?

# Other Kinds of NLI: Time-related inference in natural language

Hypothesis	Premise	Label
A small Asian street band plays in a city park.	Their performance pulls a large crowd as they used some new tunes and songs today.	SUPPORTED
A woman in blue rain boots is eating a sandwich outside.	She takes off her boots in her house.	INVALIDATED
A man jumping a rail on his skateboard.	His favorite food is pizza.	UNKNOWN

# Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc.

people   organizations   places

- Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

# Text Readability Assessment

BERT Regression from 0 easy to 1 hard.

Please choose the lower end of values displayed.



Mean Readability: 0.297386

Der Buchweizen. Häufig wenn man nach einem Gewitter an einem Acker vorübergeht, auf dem Buchweizen wächst, sieht man, daß er ganz schwarz geworden und abgesengt ist; es ist gerade, als ob eine Feuerflamme über denselben hingefahren wäre, und der Landmann sagt dann: »Das hat er vom Blitz bekommen!« Aber warum bekam er das? 0.462328

Ich will erzählen, was der Sperling mir gesagt hat, und der Sperling hat es von einem alten Weidenbaum gehört, welcher bei einem Buchweizenfelde steht. 0.284319

Es ist ein ehrwürdiger, großer Weidenbaum, aber verkrüppelt und alt, er ist in der Mitte geborsten und es wachsen Gras und Brombeer-Ranken aus der Spalte hervor; der Baum neigt sich vorn über und die Zweige hängen ganz auf die Erde hinunter, gerade als ob sie ein langes, grünes Haar bildeten. Auf allen Feldern rings umher wuchs Korn, sowohl Roggen und Gerste wie Hafer, ja der herrliche Hafer, der da, wenn er reif ist, gerade wie eine Menge kleiner, gelber Kanarienvögel auf einem Zweige aussieht. 0.569775

Das Korn stand gesegnet, und je schwerer es war, desto tiefer neigte es sich in frommer Demuth. Aber da war auch ein Feld mit Buchweizen, und dieses Feld war dem alten Weidenbaum gerade gegenüber. 0.363480

Der Buchweizen neigte sich durchaus nicht wie das übrige Korn, sondern prangte stolz und steif. »Ich bin wohl so reich wie die Aehre,« sagte er; »überließ bin ich weit hübscher; meine Blumen sind schön wie die Blüthen des Apfelbaumes; es ist eine Freude, auf mich und die Meinigen zu blicken!« 0.451931

Kennst Du etwas Prächtigeres als uns, Du alter Weidenbaum? Der Weidenbaum nickte mit dem Kopfe, gerade als ob er damit sagen wollte: »Ja freilich!« Aber der Buchweizen spreizte sich aus lauter Hochmuth und sagte: »Der dumme Baum, er ist so alt, daß ihm Gras im Leibe wächst! Nun zog ein schrecklich böses Gewitter auf; alle Feldblumen falteten ihre Blätter zusammen oder neigten ihre kleinen Köpfe herab, während der Sturm über sie dahinfuhr; aber der Buchweizen prangte in seinem Stolze. »Neige Dein Haupt wie wir!« sagten die Blumen. Das ist durchaus nicht nötig!« erwiederte der Buchweizen. »Senke Dein Haupt wie wir!« rief das Korn. 0.468438

»Nun kommt der Engel des Sturmes geflogen! 0.111728

Er hat Schwingen, die oben von den Wolken bis gerade herunter zur Erde reichen, und er schlägt Dich mittendurch, bevor Du bitten kannst, er möge Dir gnädig sein! «Aber ich will mich nicht beugen!« sagte der Buchweizen. »Schließe Deine Blumen und neige Deine Blätter!« sagte der alte Weidenbaum. 0.418416

»Sieh nicht zum Blitzem empor, wenn die Wolke berstet; selbst die Menschen dürfen das nicht, denn im Blitz kann man in Gottes Himmel hineinsehen; aber dieser Anblick kann selbst die Menschen blenden. 0.449119

Was würde erst uns, den Gewächsen der Erde, geschehen, wenn wir es wagten, wir, welche doch weit geringer sind! «Weit geringer?« sagte der Buchweizen. 0.371788 »Nun will ich gerade in Gottes Himmel hineinsehen! Und er that es in seinem Uebermuth und Stolz. 0.171622

Es war, als ob die ganze Welt in Flammen stände, so blitzte es. Als das böse Wetter vorbei war, standen die Blumen und das Korn in der stillen, reinen Luft erfrischt vom Regen, aber der Buchweizen war vom Blitz kohlschwarz gebrannt; er war nun ein todtes Unkraut auf dem Felde. Der alte Weidenbaum bewegte seine Zweige im Winde, und es fielen große Wassertropfen von den grünen Blättern, gerade als ob der Baum weine, und die Sperling fragten: »Weßhalb weinst Du?« 0.451732

Hier ist es ja so gesegnet! 0.053093 Sieh, wie die Sonne scheint, sieh, wie die Wolken ziehen! 0.104781 Kannst Du den Duft von Blumen und Büschen bemerken? 0.098560

Warum weinst Du, alter Weidenbaum? «Und der Weidenbaum erzählte vom Stolze des Buchweizens, von seinem Uebermuthe und der Strafe, die immer darauf folgt. 0.350989 Ich, der die Geschichte erzählte, habe sie von den Sperlingen gehört. 0.103803

Sie erzählten sie mir eines Abends, als ich sie um ein Märchen bat. 0.067044

# Question Answering

## What did Barack Obama teach?

**Barack Hussein Obama II** (born August 4, 1961) is an American attorney and politician who served as the 44th [President of the United States](#) from January 20, 2009, to January 20, 2017. A member of the [Democratic Party](#), he was the first [African American](#) to serve as president. He was previously a [United States Senator](#) from [Illinois](#) and a member of the [Illinois State Senate](#).

Obama was born in 1961 in [Honolulu, Hawaii](#), two years after the territory was [admitted to the Union](#) as the 50th state. Raised largely in Hawaii, he also spent one year of his childhood in [Washington state](#) and four years in [Indonesia](#). After graduating from [Columbia University](#) in 1983, he worked as a [community organizer](#) in [Chicago](#). In 1988, he enrolled in [Harvard Law School](#), where he was the first black president of the [Harvard Law Review](#). After graduating, he became a [civil rights](#) attorney and a professor, teaching [constitutional law](#); at the University of [Chicago Law School](#) from 1992 to 2004.

Barack Obama



44th President of the United States

In office

# Question Answering

- Directly answer natural language questions based on information presented in a corpus of textual documents (e.g., the web).
  - When was Barack Obama born? (*factoid*)
    - August 4, 1961
  - Who was president when Barack Obama was born?
    - John F. Kennedy
  - How many presidents have there been since Barack Obama was born?
    - 9

# Text Summarization (Abstractive)

- Produce a short summary of a longer document or article.
  - **Article:** With a split decision in the final two primaries and a flurry of superdelegate endorsements, [Sen. Barack Obama](#) sealed the Democratic presidential nomination last night after a grueling and history-making campaign against [Sen. Hillary Rodham Clinton](#) that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against [Sen. John McCain](#), the presumptive Republican nominee....
  - **Summary:** Senator Barack Obama was declared the presumptive Democratic presidential nominee.

# Text Summarization (Extractive)

Lindsay Lohan pleaded not guilty Wednesday to felony grand theft of a \$2,500 necklace, a case that could return the troubled starlet to jail rather than the big screen. Saying it appeared that Lohan had violated her probation in a 2007 drunken driving case, the judge set bail at \$40,000 and warned that if Lohan was accused of breaking the law while free he would have her held without bail. The Mean Girls star is due back in court on Feb. 23, an important hearing in which Lohan could opt to end the case early.

Lindsay Lohan pleaded not guilty Wednesday to felony grand theft of a \$2,500 necklace — a case that could return the troubled starlet to jail rather than the big screen. Saying it appeared that Lohan had violated her probation in a 2007 drunken driving case, the judge set bail at \$40,000 and warned that if Lohan was accused of breaking the law while free he would have her held without bail. The Mean Girls star is due back in court on Feb. 23, an important hearing in which Lohan could opt to end the case early.

Lindsay Lohan appears in court (Reuters)

Get a quote now.  
Save it, then have online later.

Lindsay Lohan charged with theft

Join the conversation  
You're not alone. 1,000 people are reading this now. Tell your friends

Top Life & Style articles

- Taking the first steps to a healthy relationship
- Stretching the resources
- True remedies
- How to make a difference
- Under bed self-defense
- 10 things you can do to help others
- More Life & Style articles

Story Tools

Share on Facebook

Print this story

WA Today Jobs

Assisted Property Manager, East Park, VIC 3004

Address Details

Judge tells Lohan she's no star in his courtroom

LOS ANGELES (AP) — Lindsay Lohan walked into a courtroom to face a felony grand theft charge looking like a million dollars, only to be told by a Judge she was no better than anyone else.

Lohan's attorney argued that she stole a \$2,500 necklace from an upscale store in the first line a judge threatened to throw the troubled starlet in. But it was the first time a judge wielded enough power to make her backed up for a long time.

"You're in a different situation now that a felony has been filed," Los Angeles Superior Court Judge Keith Schwartz said after the actress pleaded not guilty Wednesday.

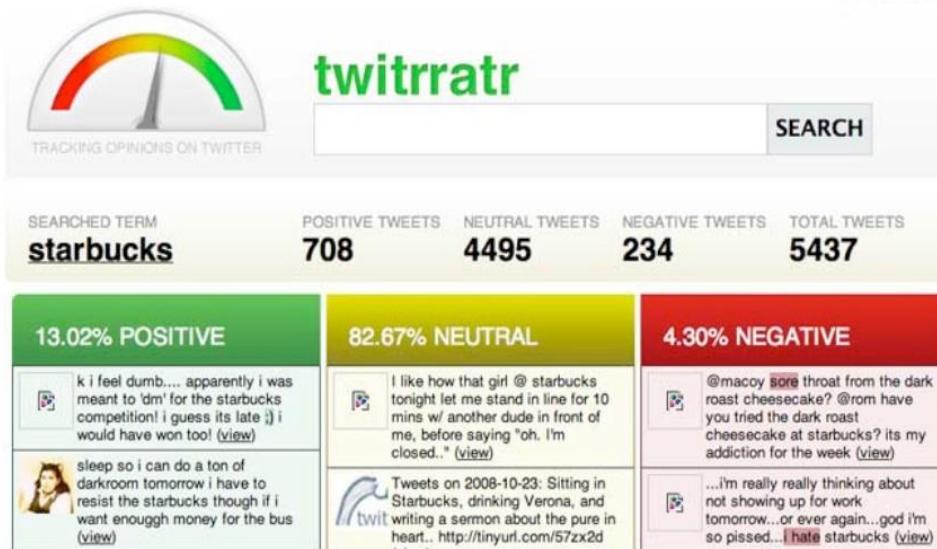
"Everybody else has to follow the law," Schwartz said, noting that he was giving the actress a tame version of a lecture he'd delivered to her attorney before the hearing and away from the dozens of assembled reporters.

"You're no different than anyone else. So please, don't push your luck."

Testing the limits — in the courtroom rather than the big screen — has been Lohan's calling card in recent months.

The "Mean Girls" star is due back in court on Feb. 23, an important hearing in which Lohan could opt to end the case early. Her attorney, Shaeen Chapman Holey, indicated Wednesday that the actress was interested in an early disposition program if the terms are right.

# Sentiment/Opinion Analysis

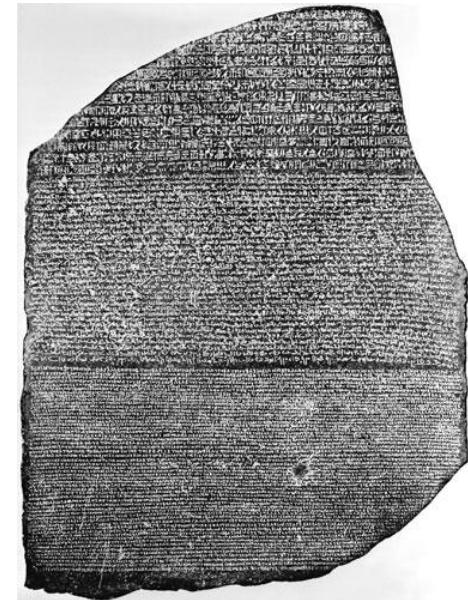


# Machine Translation (MT)

- Translate a sentence from one natural language to another.
  - Hasta la vista, bebé ⇒  
Until we see each other again, baby.
  - 我喜欢汉堡 ⇒  
I like burgers.

Google Translate

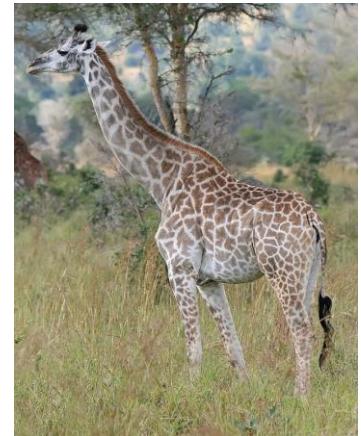
The screenshot shows the Google Translate interface. At the top, there are tabs for Text, Images, Documents, and Websites. Below that, the source language is set to English - Detected and the target language is Polish. The input text "I am happy today" is translated into Polish as "Jestem dzisiaj szczęśliwy". There are "Look up details" buttons and sharing icons below the translation.



# Commonsense reasoning

- the basic level of practical knowledge and reasoning
- concerning everyday situations and events
- that are commonly shared among most people

Going for a walk takes less time  
than going for vacation



For example, it's ok to keep the closet door open,  
but it's not ok to keep the fridge door open,  
as the food inside might go bad.

# Commonsense reasoning (knowledge bases)

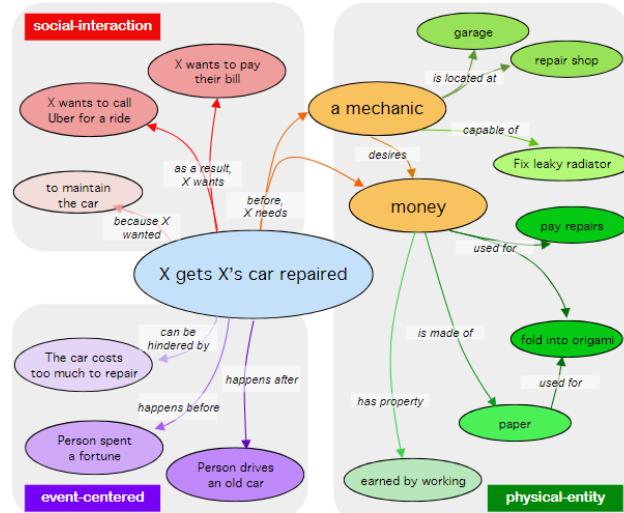
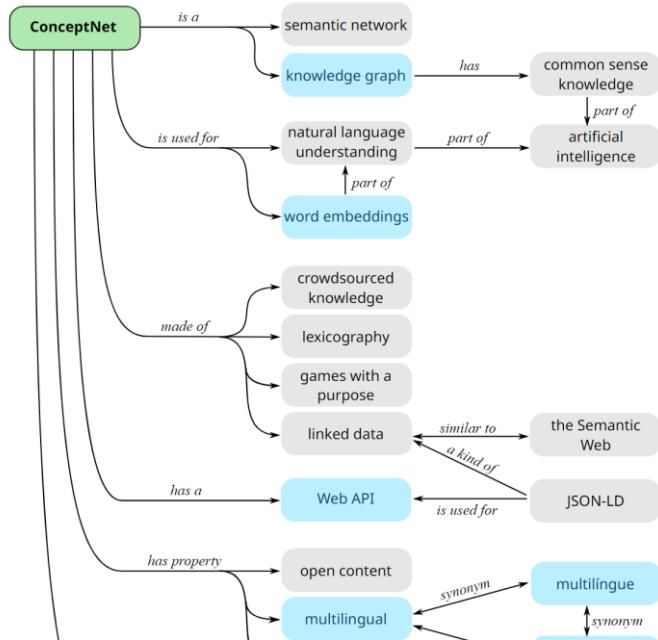


Figure 1: A tiny subset of ATOMIC<sup>20</sup>, a large atlas of social and physical commonsense relations. Relations in the top-left quadrant reflects relations from ATOMIC.<sup>1</sup>

# Image2text, Text2Image generation

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



Edit prompt or view more images+

TEXT PROMPT

an armchair in the shape of an avocado....

AI-GENERATED IMAGES



Edit prompt or view more images+

TEXT PROMPT

a store front that has the word 'openai' written on it....

AI-GENERATED IMAGES



Edit prompt or view more images+

<https://openai.com/blog/dall-e/>

# Fake News Detection

- Example: <http://www.fakenewschallenge.org/>
  - “The goal of the Fake News Challenge is to explore how artificial intelligence technologies, particularly machine learning and natural language processing, might be leveraged to combat the fake news problem. We believe that these AI technologies hold promise for significantly automating parts of the procedure human fact checkers use today to determine if a story is real or a hoax.”

## How Pizzagate went from fake news to a real problem for a D.C. business

By Joshua Gillin on Monday, December 5th, 2016 at 5:23 p.m.



### Input

A headline and a body text - either from the same news article or from two different articles.

### Output

Classify the stance of the body text relative to the claim made in the headline into one of four categories:

1. **Agrees:** The body text agrees with the headline.
2. **Disagrees:** The body text disagrees with the headline.
3. **Discusses:** The body text discuss the same topic as the headline, but does not take a position
4. **Unrelated:** The body text discusses a different topic than the headline

# CLEF 2024 Tasks Examples

The screenshot shows the homepage of the CLEF 2024 website. At the top, there is a banner with the conference title, date (9-12 September 2024), and location (Grenoble, France). Below the banner is a large photograph of a modern building complex with a large paved area in front, set against a backdrop of mountains. The main menu on the left includes links for Home, Programme, Keynote Talks, Conference, Accepted Papers, Call for Papers, Accepted Labs, Call for Lab Proposals, Accepted Labs, Registration, Slides, Registration, Poster Session, and Venue. The Accepted Labs section lists several accepted labs with their names in blue: BioASQ, CheckThat!, eRisk 2024, EXIST, iDPP, ImageCLEF, JOKER Lab, ELOQUENT Lab, LifeCLEF, LongEval, PAN, QuantumCLEF, SimpleText Lab, and Touché. To the right, there is a "Tweets from @clef\_initiative" sidebar with two tweets from the CLEF Twitter account (@clef\_2023) dated October 5 and 4, respectively. The sidebar also includes a "Like" button.

**CLEF 2024 Conference and Labs of the Evaluation Forum**  
Information Access Evaluation meets Multilinguality, Multimodality, and Visualization  
9-12 September 2024, Grenoble - France

**Accepted Labs**

- [BioASQ](#): Large-scale Biomedical Semantic Indexing and Question Answering
- [CheckThat!](#): Predicting Check-Worthiness, Subjectivity, Persuasion, Roles and Authorities
- [eRisk 2024](#): Early Risk Prediction on the Internet
- [EXIST](#): sEXism Identification in Social neTworks
- [iDPP](#): Intelligent Disease Progression Prediction
- [ImageCLEF](#): Multimodal Challenge in CLEF
- [JOKER Lab](#): Automatic Wordplay Analysis
- [ELOQUENT Lab](#): Evaluating Generative Language Models
- [LifeCLEF](#)
- [LongEval](#)
- [PAN](#): Lab on Stylometry and Digital Text Forensics
- [QuantumCLEF](#): Quantum Computing at CLEF
- [SimpleText Lab](#): Automatic Simplification of Scientific Texts
- [Touché](#): Argumentation Systems

**Tweets from @clef\_initiative**

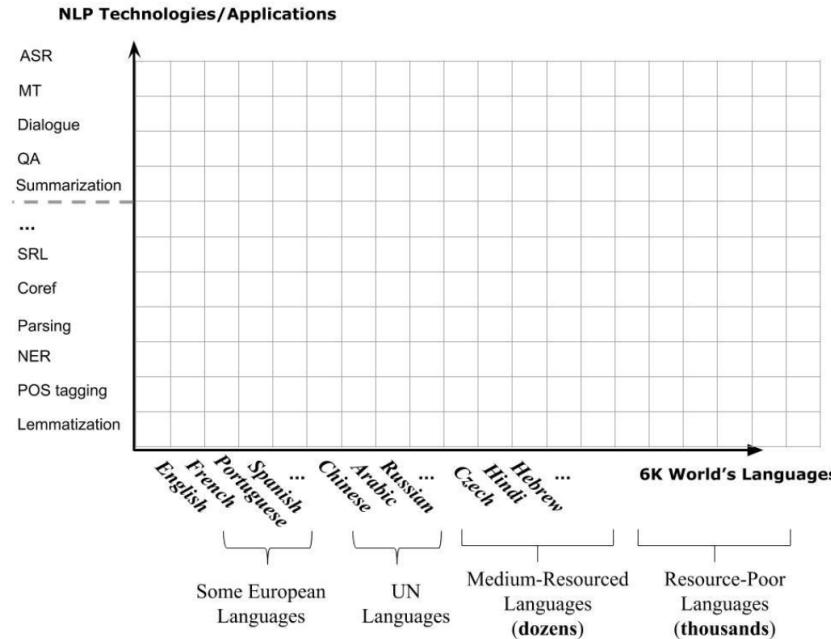
CLEF @clef\_2023 · Oct 5 #clef2023 working notes are online: they contain overview and participant papers from 13 labs!

Sun... @clef\_2023 · Oct 4 CLEF 2023: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum:...

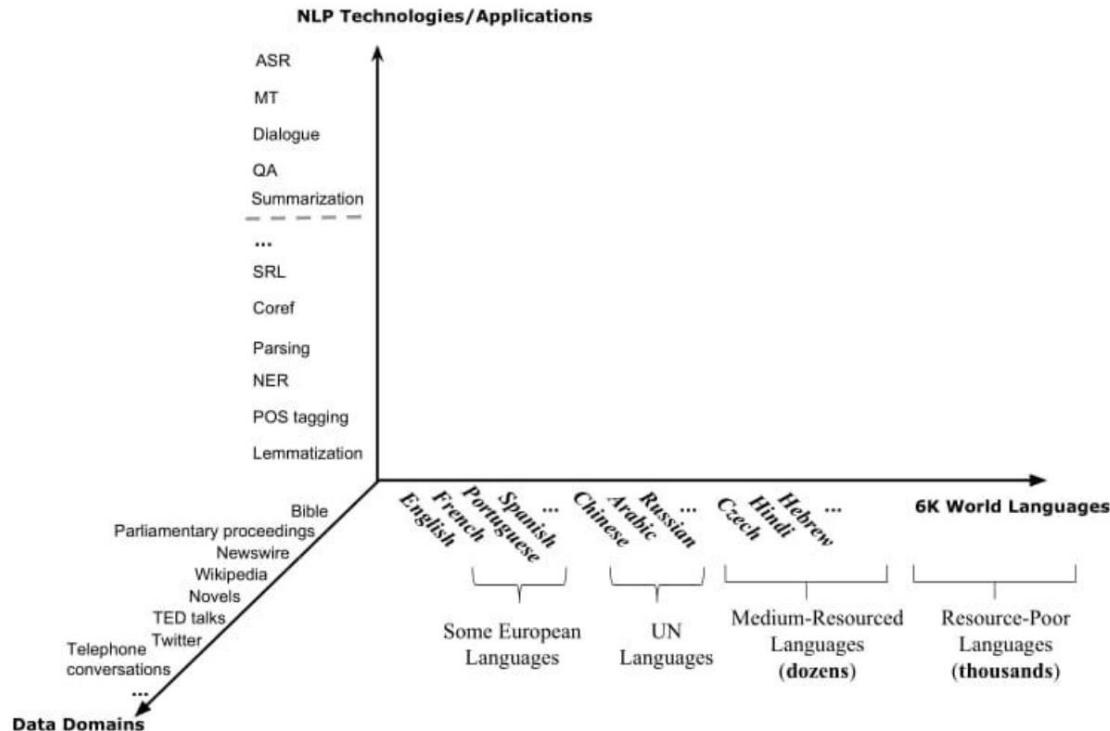
1 21 ⓘ

CLEF

# Challenge: Many languages..



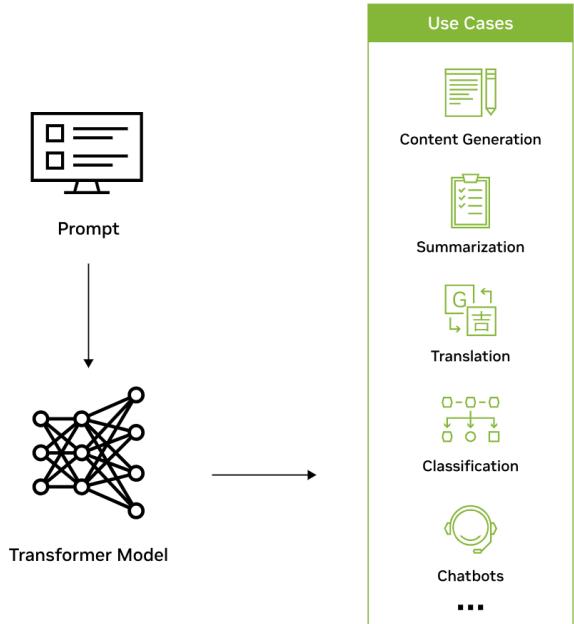
# Challenge: Many languages and domains..



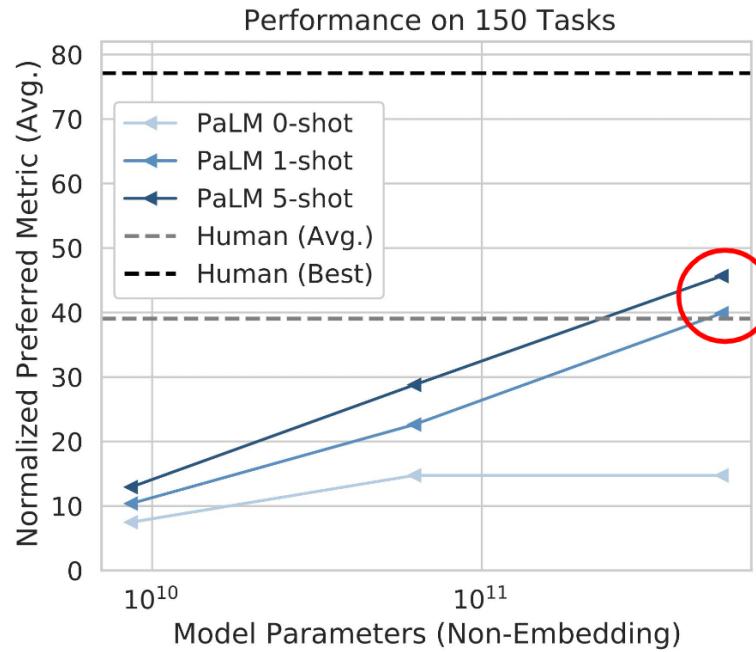
# Introduction to LLMs

# Large Language Models (LLMs)

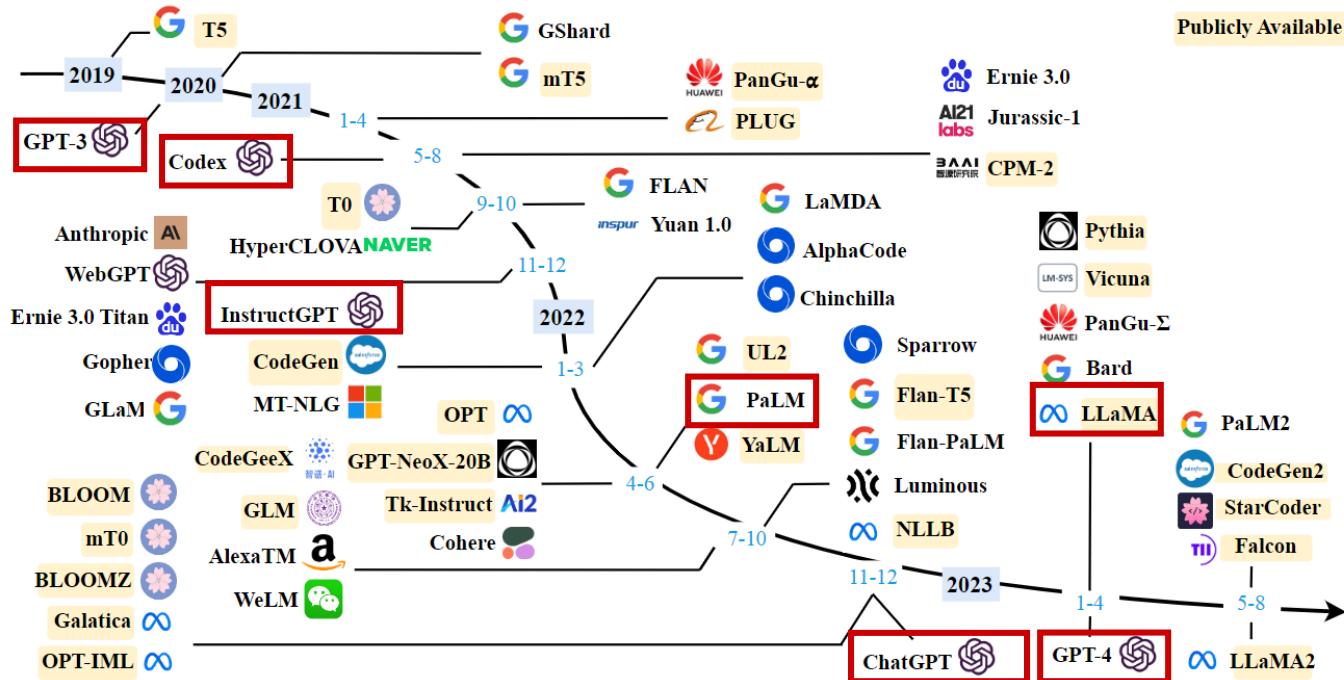
- Large Language Models (LLMs) are a type of generative AI that are trained on text and produce textual content
  - ChatGPT is a popular example of generative text AI
- LLMs acquire the ability to achieve general-purpose language understanding & generation
  - by using massive amounts of data to learn billions of parameters during training.



LLMs are making astonishing progress on many complex tasks



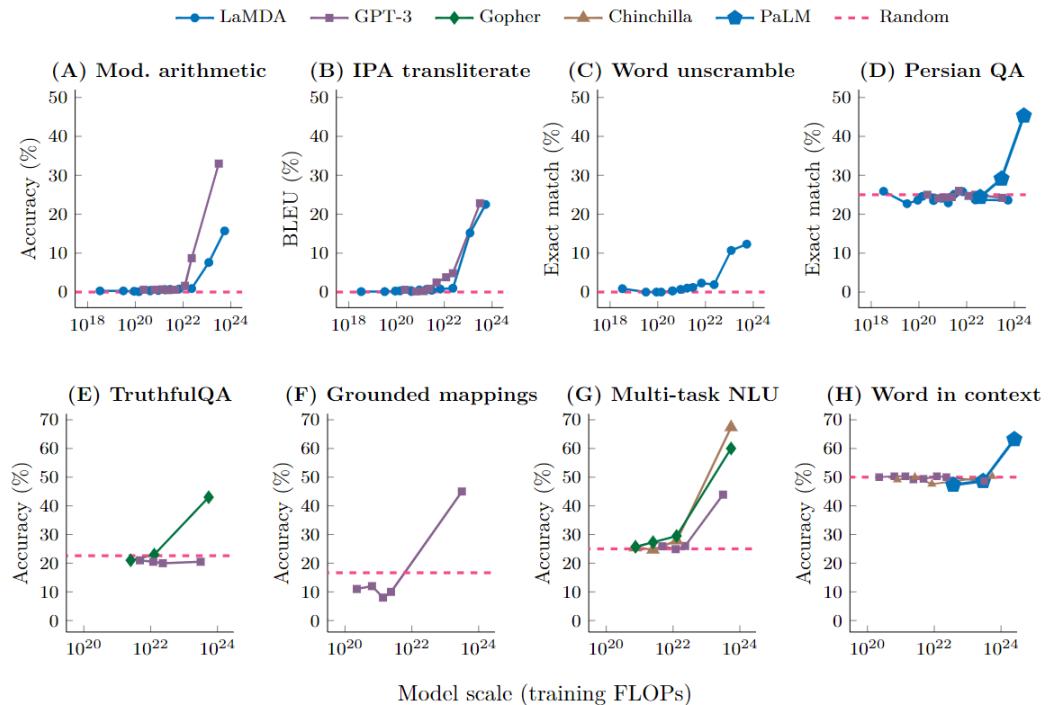
# A Timeline of Existing LLMs with Size >10B Parameters



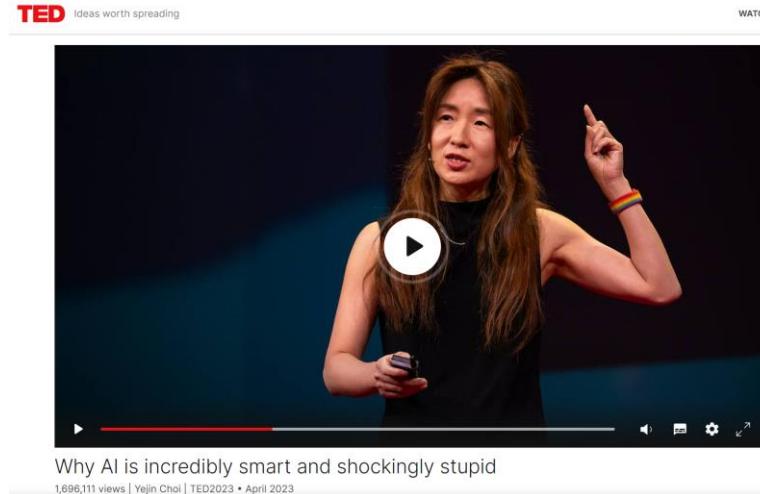
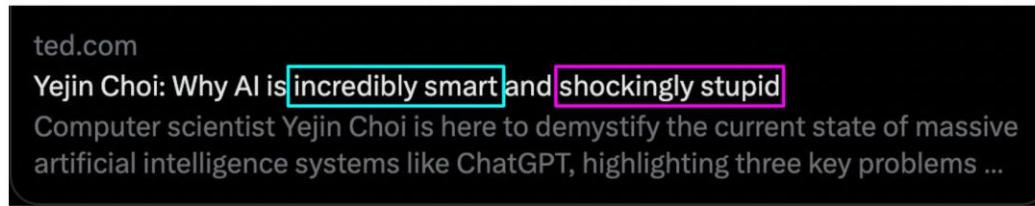
- **OpenAI** consistently maintains its leadership position with other companies having difficulty to catch up
- **Meta** contributes significantly to open-source LLMs, one of the most generous companies with all LLMs developed being open (e.g., Llama)
- LLMs exhibit an **increasing tendency to be closed-source** (LaMDA, PaLM, GPT-4) unlike in recent years (before 2020)
- **API-based access** becomes the predominant use method (size and secrecy)
- Increasing appearance of **domain-focused LLMs** like BloombergGPT, Med-PALM, FinBERT, SciBERT, TourBERT

# Effect of Scale: Emergent Abilities of Large Language Models

- Abilities that cannot be predicted by extrapolating performance improvements
- Models gain suddenly good performance on some tasks once the scale exceeds a certain range
  - One explanation is that complex tasks with multiple steps cannot be handled by models unless they are large enough to handle every step..



# However, there are (still) some problems..



Yejin Choi, *Why AI is incredibly smart and shockingly stupid?*, TED talk, April 2023

# And dangers..

Emily M. Bender  
Professor, University of Washington

## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender\*

ebender@uw.edu

University of Washington

Seattle, WA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington

Seattle, WA, USA

Timnit Gebru\*

timnit@blackinai.org

Black in AI

Palo Alto, CA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether



Illustration by TIME; reference image courtesy of Emily Bender

BY ISSIE LAPOWSKY  
SEPTEMBER 7, 2023 7:00 AM EDT

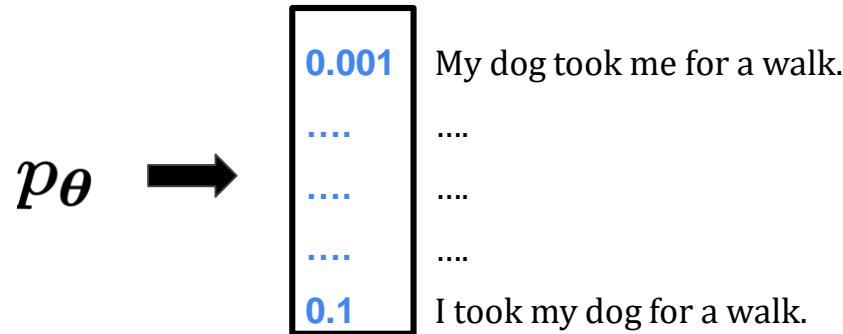
**E**mily M. Bender doesn't consider herself an AI researcher. The University of Washington professor is, first and foremost, a linguist. But her gimlet-eyed research on the dangers of large language models and her withering cross-examinations of the AI hype cycle have made her one of the industry's most formidable critics.

<https://time.com/collection/time100-ai/6308275/emily-m-bender/>

# Language Modeling

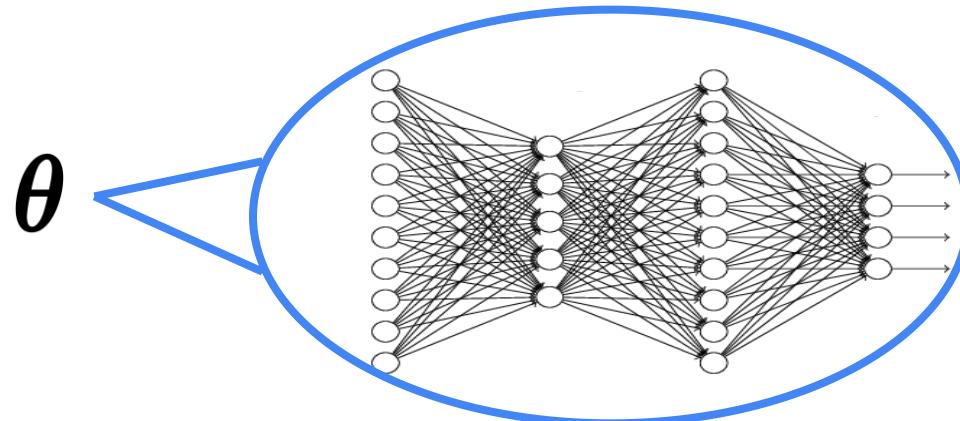
# General Definition of a Language Model

- A language model (LM)  $p_{\theta}$  is a **probability distribution over strings**
  - can tell us **how likely a text** is to occur in a particular natural language domain (i.e., the domain of the training corpus)



# NN-based Language Modeling

- A language model (LM)  $p_{\theta}$  is a **probability distribution over strings**
  - $\theta$  are the parameters of this distribution. Nowadays,  $\theta$  corresponds to a neural network (often with billions or trillions of parameters)



# Language Models: Definition

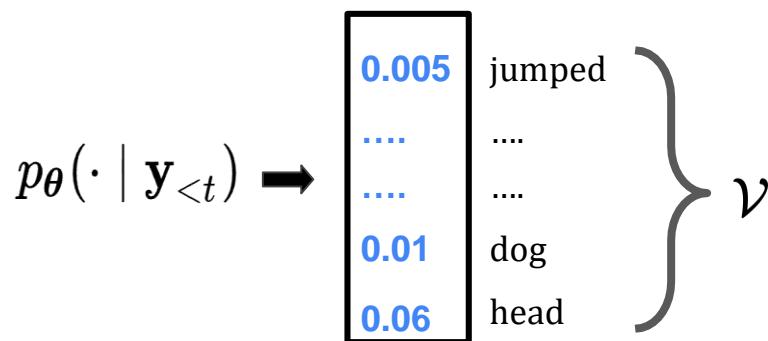
- We define a string  $\mathbf{y}$  as a sequence of tokens  $\langle y_1, y_2, \dots \rangle$



What are  $y$ 's?

# Language Models

- Most LMs are **autoregressive**, i.e. they actually estimate the **conditional probability** of individual symbols  $y_t \in \mathcal{V}$  in the string, given their prior context  $\mathbf{y}_{<t}$



$$\begin{aligned} p_{\theta}(\mathbf{y}) &= p_{\theta}(\langle y_1, y_2, \dots \rangle) \\ &= \prod_{t=1}^{|\mathbf{y}|} p_{\theta}(y_t \mid \mathbf{y}_{<t}) \end{aligned}$$

The joint probability of an entire string is usually computed as the product of these conditional probabilities

# Language Models: Learning Parameters

- In other words, we want to maximize the likelihood of a training corpus  $\mathcal{D}$  under our model  $p_{\theta}$
- Equivalent to minimizing the cross-entropy with the empirical distribution, as defined by a training corpus  $\mathcal{D}$

$$\underset{\theta}{\text{minimize}} \ H(p_{\mathcal{D}}(\cdot), p_{\theta}(\cdot))$$



$$\underset{\theta}{\text{minimize}} \ - \sum_{\mathbf{y} \in \mathcal{D}} \log p_{\theta}(\mathbf{y})$$



$$\underset{\theta}{\text{minimize}} \ - \sum_{\mathbf{y} \in \mathcal{D}} \sum_{t=1}^T \log p_{\theta}(y_t \mid \mathbf{y}_{<t})$$

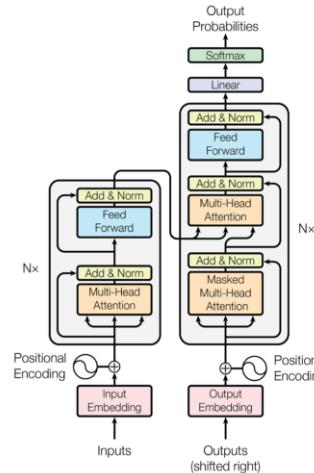
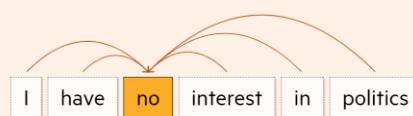


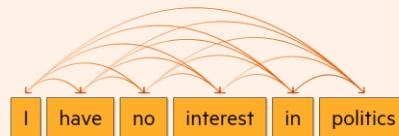
Figure 1: The Transformer - model architecture.

# Self-attention

Self-attention looks at each **token** in a body of text and decides which others are most important to understanding its meaning.



With self-attention, the transformer computes all the words in a sentence at the same time. Capturing this context gives LLMs far more sophisticated capabilities to parse language.



And when we combine the sentences, the model is still able to recognise the correct meaning of each word thanks to the attention it gives the accompanying text.

For the first use of interest, it is **no** and **in** that are most attended.



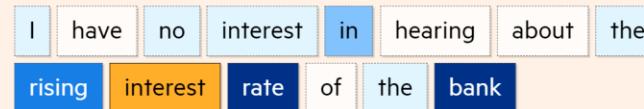
# Self-attention Idea

And when we combine the sentences, the model is still able to recognise the correct meaning of each word thanks to the attention it gives the accompanying text.

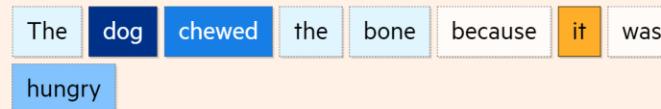
For the first use of interest, it is **no** and **in** that are most attended.



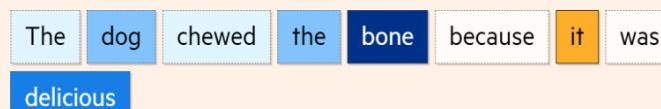
For the second, it is **rate** and **bank**.



In the following sentence, self-attention is able to calculate that **it** is most likely to be referring to **dog**.



And if we alter the sentence, swapping **hungry** for **delicious**, the model is able to recalculate, with **it** now most likely to refer to **bone**.



And if we alter the sentence, swapping **hungry** for **delicious**, the model is able to recalculate, with **it** now most likely to refer to **bone**.

The dog chewed the bone because it was delicious

The **dog** chewed the bone because it was delicious.

The benefits of self-attention for language processing increase the more you scale things up. It allows LLMs to take **context** from beyond sentence boundaries, giving the model a greater understanding of how and when a word is used.

In a square little town with rolling green hills and lush meadows, there lived a delightful canine with a coat of golden fur and a red collar around his neck like a crown. This charming creature was the beloved member of the Johnson family, with a gleaming coat of golden fur and ears that sparkled with warmth and affection. He won the hearts of everyone who crossed his path. The red collar became a symbol of his loyalty and an emblem of the countless adventures that awaited him each day he trotted over the Johnson's lawns. He brought an abundance of joy and laughter. His days were filled with frolics in the nearby park, chasing butterflies, and playing fetch with his master. In the afternoons, he would faithfully accompany Mr. Johnson on his walks around the neighborhood, sniffing the scents of the world with curious enthusiasm. The neighborhood admired his friendly nature and the undeniably good time he shared with his family. Whenever he went, the red collar shone like a beacon, a reminder of the joy and loyalty he offered to those who embraced him.

ate dinner at 6 pm

After the last walk every day, Mrs. Johnson ensured that their canine companion had a hearty meal and lots of wholesome ingredients. This would of patiently sit, his tail wagging merrily, until the clock struck six. As the aroma of his favorite meal wafted through the air, he couldn't contain his excitement. His red collar jingled with each step he took towards his feeding bowl; a sound that had become synonymous with the pulsing anticipation of dinner time. Beyond the boundaries of the town, Max's escapades expanded into a realm of wild imagination. He roamed through vast meadows and ventured into dense forests, his red collar contrasting against the vibrant hues of nature. On one such adventure, he met a pack of fellow canines, and together, they formed an inseparable bond. They navigated through the wilderness, encountering thrilling encounters with other animals, while sharing lots of laughs and stories under the watchful stars. As time passed, Max grew older, and the years began to leave their mark on his once vibrant fur. Through his days may have slowed, his spirit remained unbroken. The red collar, now slightly faded, continued to adorn his neck, a symbol of the countless memories he had woven into the fabric of his family life. As he gazed out the window of the days, the sun had set, casting a warm glow that punctuated the house with a golden light. Throughout the years, Max, through every moment, taught that the time they spent together was a precious gift, and they were determined to make it memorable.

In a square little village nestled amidst picturesque landscapes, there lived a delightful canine named Luna, whose presence brought an unexplainable sense of joy to all. Luna was a beautiful mix of Labrador and Border Collie, with a silvery black coat that shimmered under the sun and a pair of striking amber eyes that sparkled with intelligence. The town's residents couldn't help but smile whenever they caught a glimpse of her wagging tail and the exuberance in her every move. Her joyful energy was infectious, drawing people from all walks of life to her side, eager to bask in the warmth of her company and the joy she brought. The village was alive and well with the sound of children laughing and the gentle barking of dogs in the morning. Luna would rise with the sun, stretching her limbs and barking with her excitement for the day ahead. The local children adored her, and so did the adults, because her joyful temperament lit up every adventure they embarked upon. She would play fetch in the meadows, chasing butterflies, and rolling in the soft grass. The presence of the village square became a delightful spectacle, as the Johnson's dog brought new faces and playful, rustling sounds willing to indulge in a game of fetch or a nap in the shade. Luna's boundless energy was a testament to the simple joys of life, a reminder that can be found in the simplest of moments.

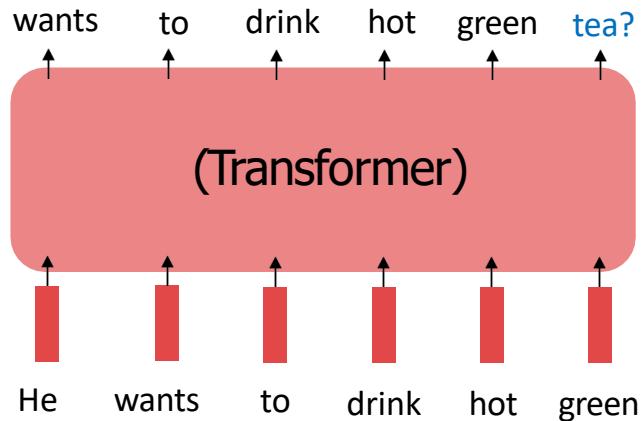
loved playing fetch

Max loved playing fetch with the children. After every afternoon, she would eagerly await the sound of the ball being thrown, and would soon come running towards her, with a bright red tail wagging furiously. In her mouth, she would take turns throwing it far into the distance, and Luna would bark in the wind to retrieve it. Her tail would wag back and forth in sheer delight. The children laughed, filled the air as they chased her on, and the bond between Luna and her young playmates grew stronger with each game. Through their innocent games of fetch, they learned the value of camaraderie and the joy of giving and receiving unconditional love.

# Pre-training

**Pre-train (on language modeling)**

Lots of text; learn general things



# What kinds of things does pre-training teach?

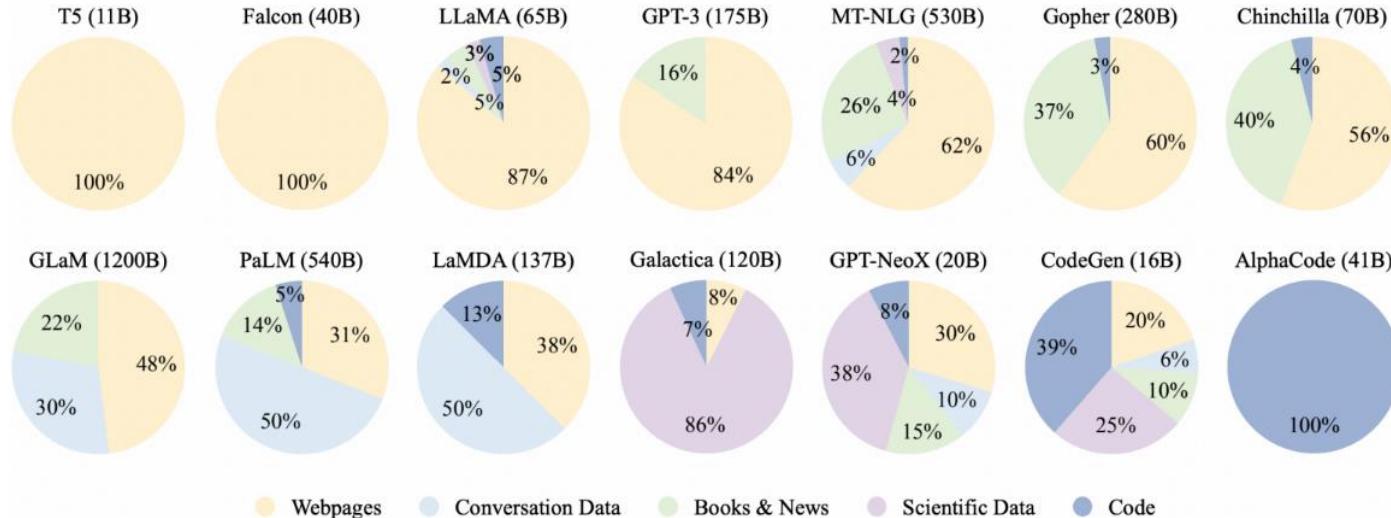
There's increasing evidence that pre-trained models learn a wide variety of things about the statistical properties of language:

- *University of Innsbruck is located in\_\_\_\_, Austria.* [trivia]
- *I put\_\_\_\_\_fork down on the table.* [syntax]
- *The woman walked across the street, checking for traffic over\_\_shoulder.* [coreference]
- *I went to the ocean to see the fish, turtles, seals, and\_\_\_\_.* [lexical semantics/topic]
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was\_\_\_\_.* [sentiment]
- Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the. [some reasoning – this is harder...]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21,\_\_\_\_. [some basic Arithmetic..]
- ...

Models also learn, and can exacerbate, racism, sexism, all kinds of biases...

# Distribution of data sources for pre-training LLMs

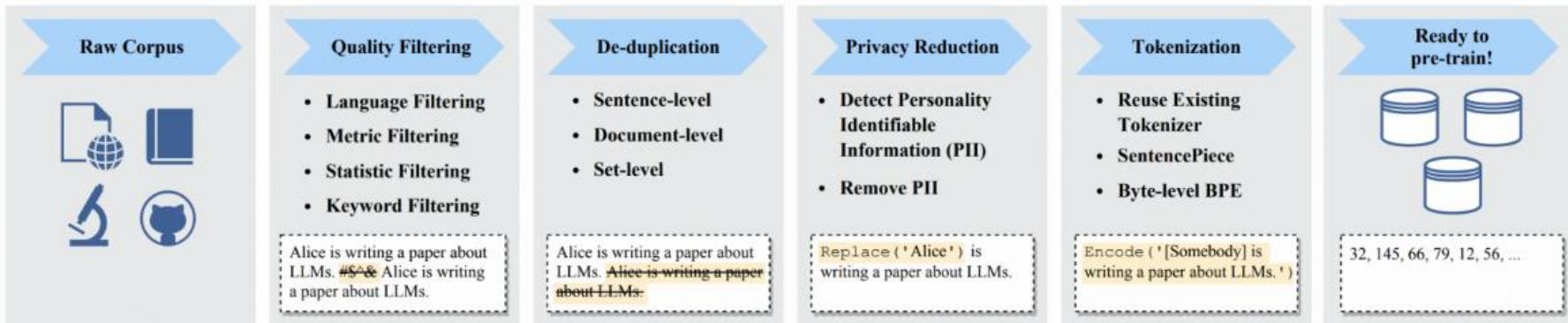
Inclusion of heterogenous sources is very helpful



Ratio of data sources in example recent LLMs

# Common Data Processing Pipeline for LLM Pretraining

Deduplication is essential (we want generalization rather than memorization), and, in general, high quality data gives advantage



Typical preprocessing pipeline for pretraining LLMs

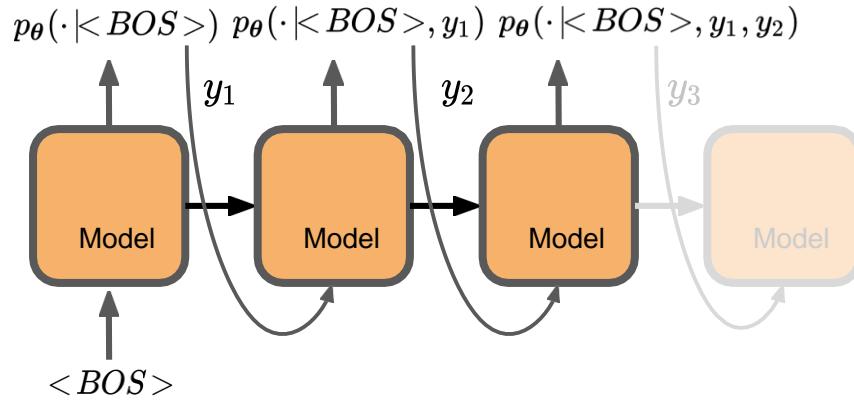
# Text Generation (Decoding)

# Generating text

- Modeling the conditional probability distribution over each word

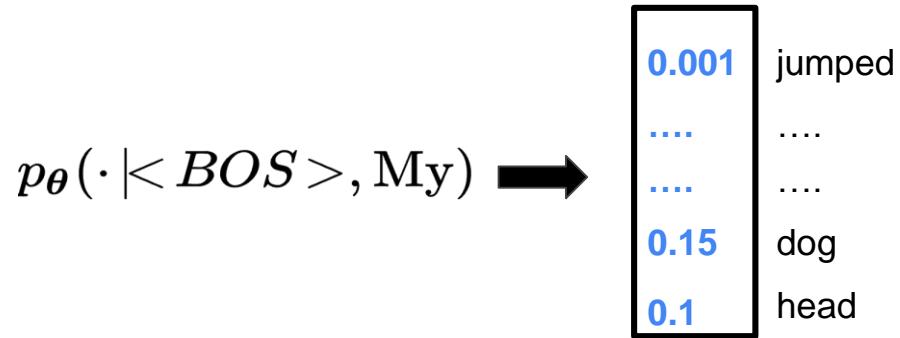
$$p_{\theta}(\mathbf{y}) = p_{\theta}(\langle y_1, y_2, \dots \rangle)$$

- We then generate  $y_1$  according to  $p_{\theta}(\cdot | \text{<} \text{BOS} \text{>})$ ,  $y_2$  according to  $p_{\theta}(\cdot | \text{<} \text{BOS} \text{>}, y_1)$

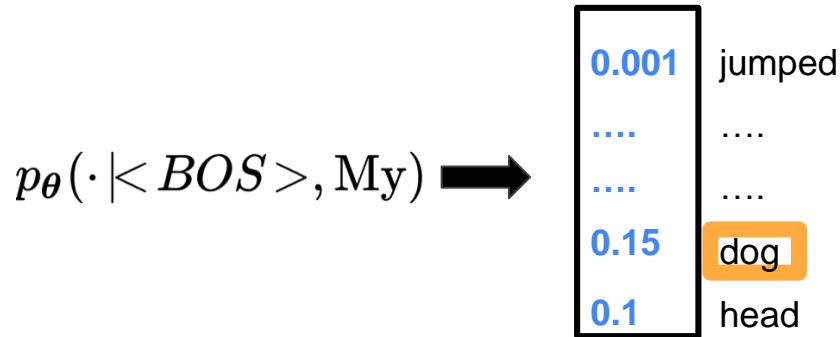


How do we choose  
 $y$  at each step?

# Generating from Locally Normalized Distributions



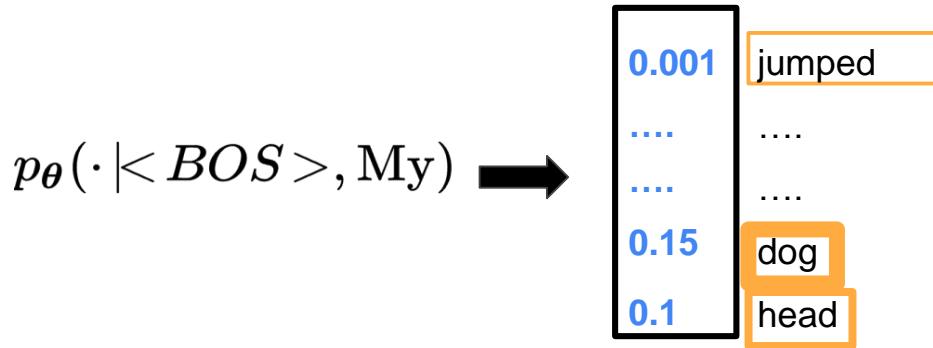
# Generating from Locally Normalized Distributions



**Greedy Search:** deterministically choose most probable token according to  $p_{\theta}$ !

$$y_t = \operatorname{argmax}_{y \in \mathcal{V}} p_{\theta}(y | \mathbf{y}_{<t})$$

# Generating from Locally Normalized Distributions



**Sampling:** sample according to multinomial distribution given by  $p_{\theta}$ !

$$y_t \sim p_{\theta}(\cdot | \mathbf{y}_{})$$

# Temperature Sampling

**Ancestral sampling** is obtained when the scoring function is simply the identity function

$$p(y \mid \mathbf{y}_{<t})$$

**Temperature sampling** can be represented as

$$p(y \mid \mathbf{y}_{<t})^{\frac{1}{\tau}}$$

**Lower** the temperature  $< 1$ : probability becomes more **spiky**

- **Less** diverse output (probability is concentrated on top words)

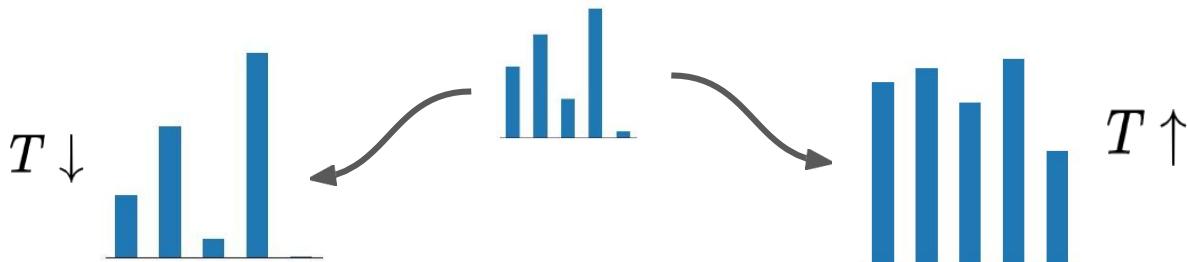
**Raise** the temperature  $> 1$ : probability becomes more **uniform**

- **More** diverse output (probability is spread around vocab)

# Consequences of temperature sampling

Lower the temperature  $< 1$ : probability becomes more **spiky**

- **Less diverse output** (probability is concentrated on top words)



One day a cat decided to climb a tree but was caught by a dog. The cat was taken to the vet and the vet said the cat had a broken leg. The cat was then taken to the vet and the vet said the cat had a broken leg and was in a lot of pain. The cat was then .....

Peaky distributions lead to **repetitive text**

Raise the temperature  $> 1$ : probability becomes more **uniform**

- **More diverse output** (probability is more spread around vocabulary)

One day a cat decided to climb a tree but did not properly rock climb, Zoo workers put her (feet down), disastrously Raphael Staonymusensht October 28, 2014 at 12:38 PM Hot Cat written by .....

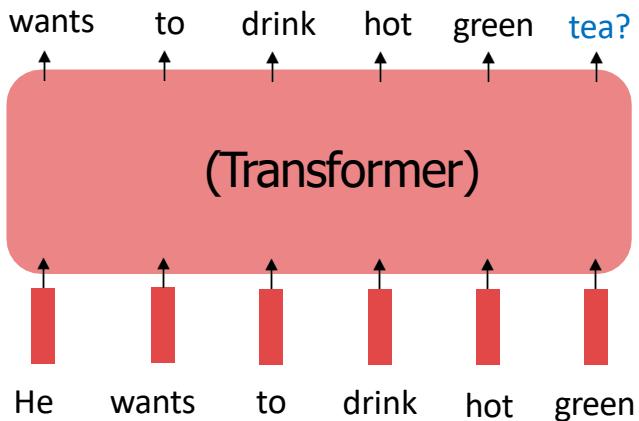
Flat distributions lead to **incoherent text**

# Finetuning

# The Pre-training & Fine-tuning Paradigm

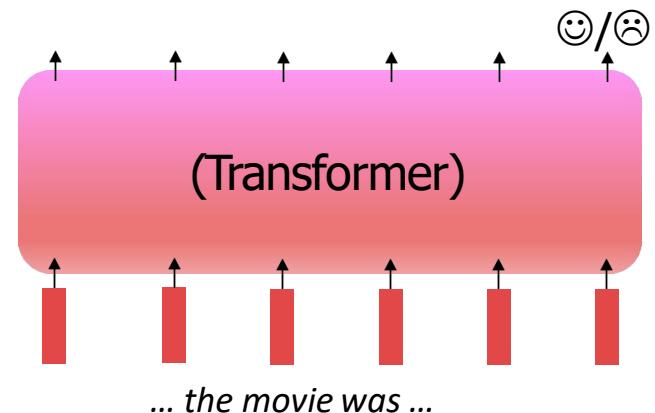
## Step 1: Pre-train (on language modeling)

Lots of text; learn general things.



## Step 2: Fine-tune (on your task)

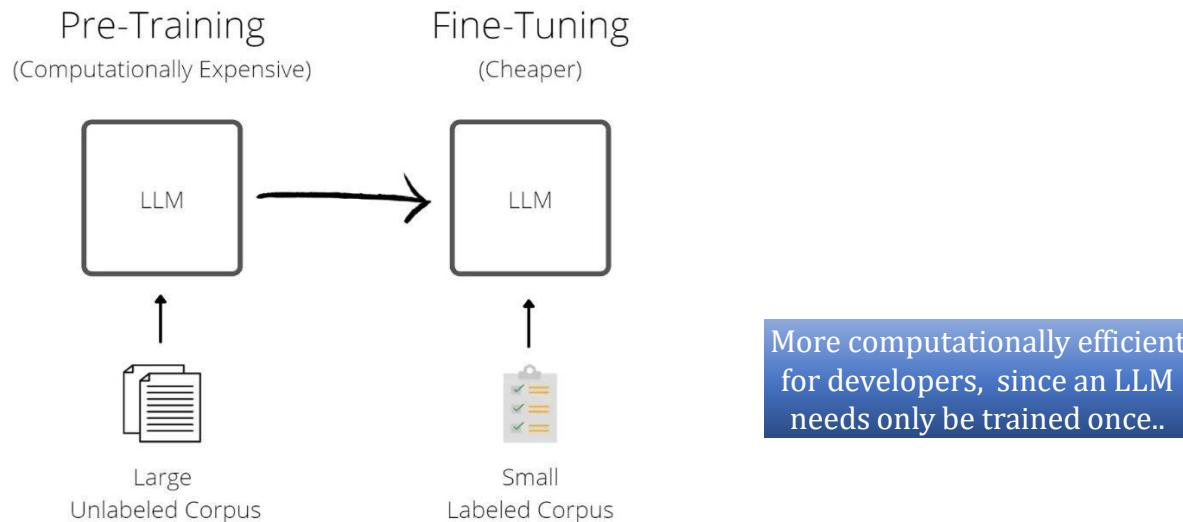
Not many labels; adapt to specific tasks.



**Pre-training** of a language model on a diverse corpus of unlabeled text, followed by **fine-tuning** on specific tasks.

# Fine-tuning

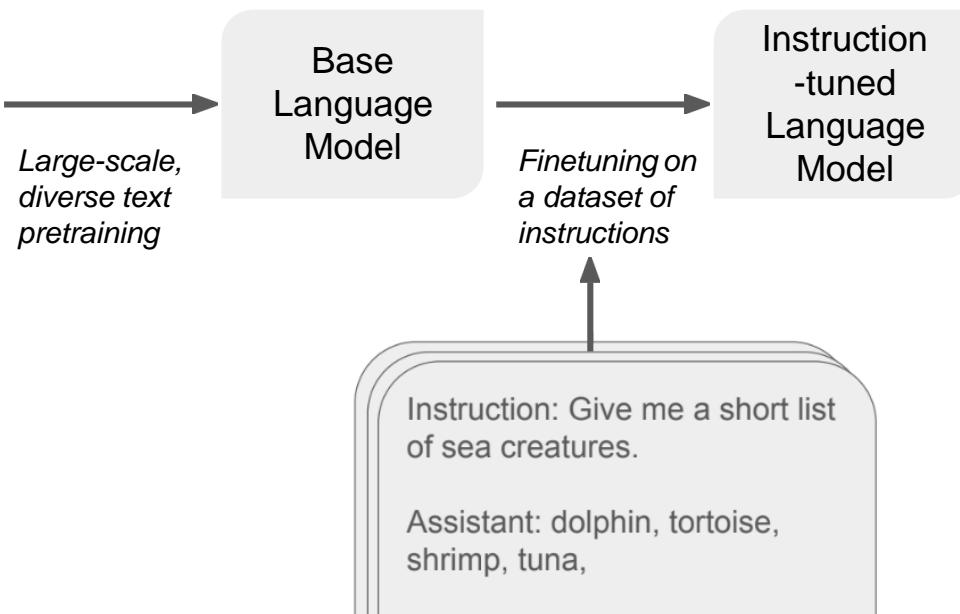
- Large language models are often used as base (foundation) models that are then fine-tuned for some downstream task.
- This works because the LLM encode a lot of information about natural language



Pre-trained models typically generate content that is a **natural continuation of prompts** (*even if the prompts are malicious*)

# Instruction Finetuning

Instruction finetuning uses the instruction-answer format



**Instruction tuning** is a **supervised learning** on instruction-answer data.

After instruction-tuning, models don't need to be prompted to know they should be generating answers instead of general continuations.

# InstructGPT

PROMPT    *Write a short poem about a wise frog.*

COMPLETION    GPT-3

Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

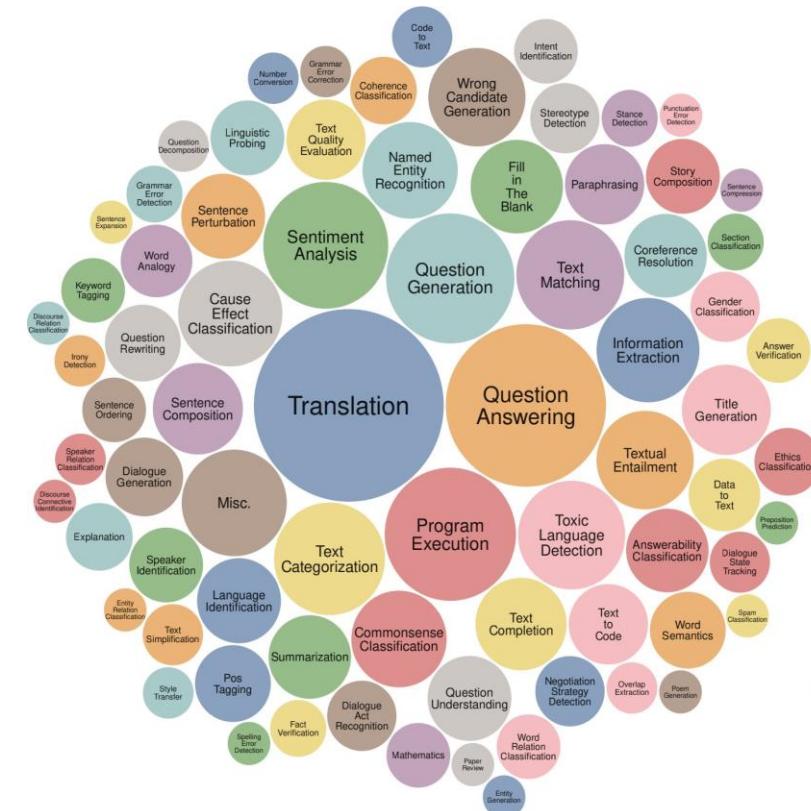
Write a short story in which a character has a bad dream.

InstructGPT

The frog is a symbol of wisdom  
He knows all the secrets of the world  
He is a master of disguise  
And a great teacher of life  
He is a symbol of transformation  
And the bringer of change  
He is the frog who has seen it all  
And knows the meaning of it all

# Instruction finetuning

- As is usually the case, **data + model scale** is key for this to work
- For example, the **Super-NaturalInstructions** dataset contains **over 1.6K tasks, 3M+** examples
  - Classification, sequence tagging,
  - rewriting, translation, QA...

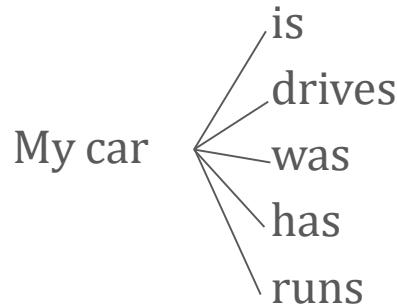


# The Art of Prompting

# Why does prompting work?

Context (often) lowers **entropy**

**Intuition:** providing more context to the model can narrow down the set of viable options when it is predicting continuation



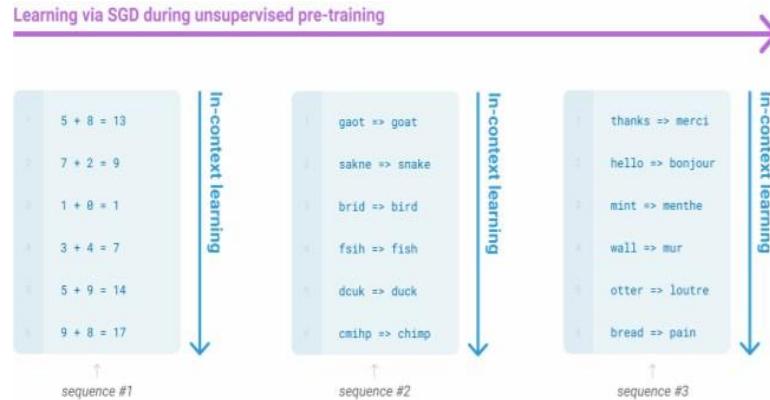
High entropy;  
many plausible  
options

It was so, so cold  
this morning. My car — battery  
(didn't start)

Fewer options;  
some are much  
more likely than  
before

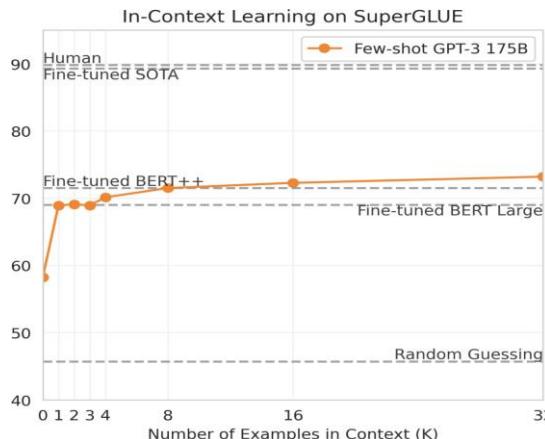
# In-context learning in very large models

Very large language models (e.g., GPT-3) seem to perform some kind of **learning without traditional machine learning procedures** simply from **examples you provide** within their contexts (prompts)...



# In-context Learning

- **Zero**-shot learning
- **One**-shot learning
- **Few**-shot 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

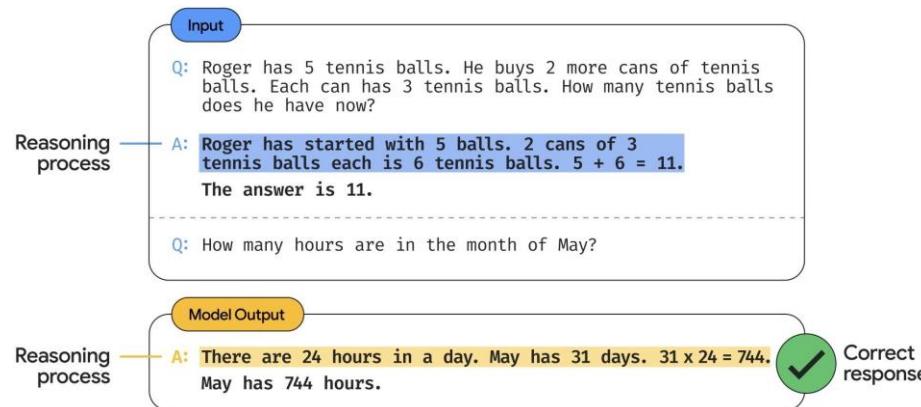
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



# Type of Prompting: Chain-of-Thought

For a task with input X and desired output Y, where we have examples of x, y pairs

- When the task is more complicated/algorithmic, asking the model to “reason through” an answer can yield better results



**Caveat:** Coming up with x, y pairs can be resource intensive, since each answer must include numerous reasoning steps

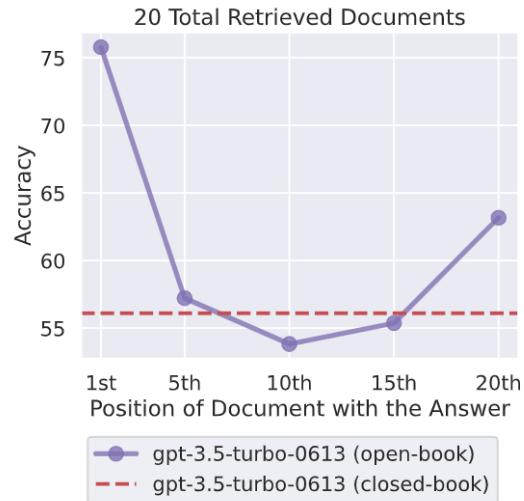
# Zero-shot CoT

	MultiArith	GSM8K
<b>Zero-Shot</b>	<b>17.7</b>	<b>10.4</b>
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
<b>Zero-Shot-CoT</b>	<b>78.7</b>	<b>40.7</b>
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7

No.	Category	Zero-shot CoT Trigger Prompt	Accuracy
1		Let's work this out in a step by step way to be sure we have the right answer.	<b>82.0</b>
2	Human-Designed	Let's think step by step. (*1)	78.7
3		First, (*2)	77.3
4		Let's think about this logically.	74.5
5		Let's solve this problem by splitting it into steps. (*3)	72.2
6		Let's be realistic and think step by step.	70.8
7		Let's think like a detective step by step.	70.3
8		Let's think	57.5
9		Before we dive into the answer,	55.7
10		The answer is after the proof.	45.7
-		(Zero-shot)	17.7

# How LLMs handle long context?

- Liu et al. find that changing the location of relevant information in the input can degrade model performance
- They find that decoder-only LLMs like GPT-3.5 can deal well with such information at the beginning or end of the input context
  - they cannot access information in the middle of it well, resulting in a U-shaped performance curve



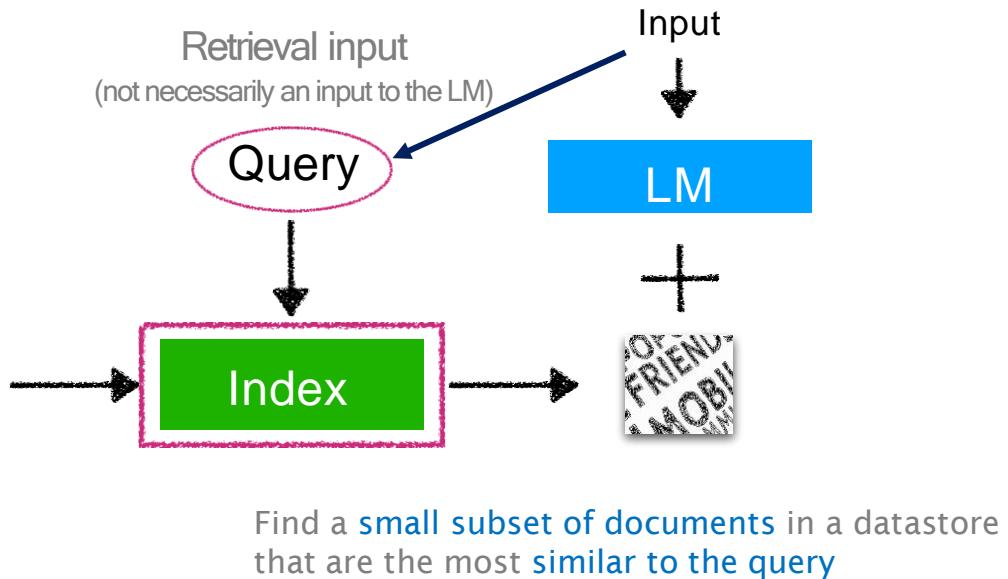
RAG

# Retrieval-augmented Large Language Models

Retrieval-augmented LLMs = Retrieval + LLMs



Datastore



# Why retrieval-based LMs?

LLMs can't memorize all (long-tail) knowledge in their parameters

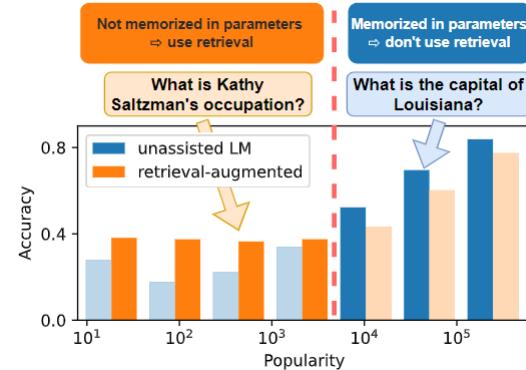


List 5 important papers authored by Geoffrey Hinton

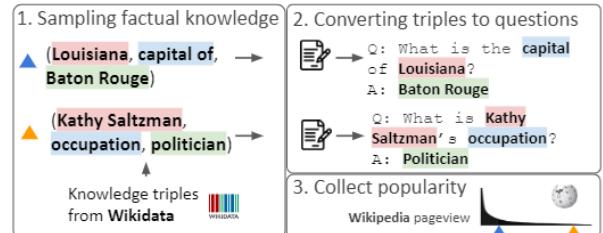


Geoffrey Hinton is a renowned computer scientist ...  
Here are five important papers authored by him:

1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, ...
2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, ...
3. "...  
4. "**Deep Learning**" (with Y. Bengio and A. Courville) - Published as a book in **2016**,...
5. "**Attention Is All You Need**" (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model,...



(Mallen et al., 2023)



PoPQA dataset

# Why retrieval-based LMs?

LLMs' knowledge is easily outdated and hard to update



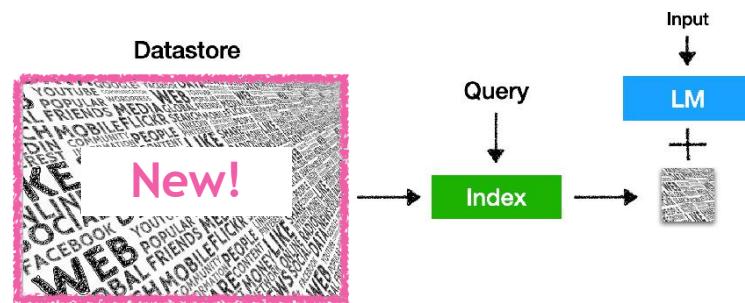
Who is the CEO of Twitter?



As of my **knowledge cutoff in September 2021**, the CEO of Twitter is **Jack Dorsey**....

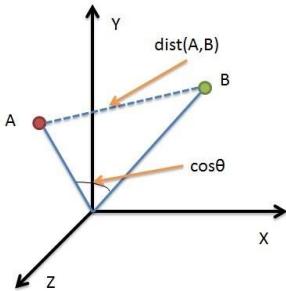
The screenshot shows a Google search results page. The search bar at the top contains the query "Who is the CEO of Twitter?". Below the search bar, there are several navigation links: "All" (selected), "News", "Images", "Shopping", "Videos", and "More". A link to "Tools" is also present. The main search result is for "Twitter / CEO", which is attributed to Linda Yaccarino. The date "Jun 5, 2023—" is shown next to her name. To the right of the text is a small portrait photo of Linda Yaccarino.

- Existing **knowledge editing** methods are still NOT scalable (ongoing research)
- The datastore can be easily **updated** and **expanded** even without retraining



# Evaluating LLMs

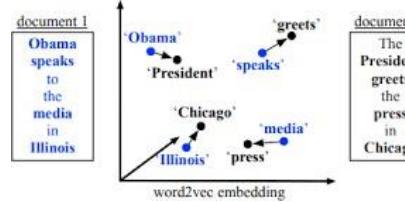
# Model-based metrics: Word distance functions



## Vector Similarity

Embedding based similarity for semantic distance between text.

- Embedding Average (Liu et al., 2016)
- Vector Extrema (Liu et al., 2016)
- MEANT (Lo, 2017)
- YISI (Lo, 2019)



## Word Mover's Distance

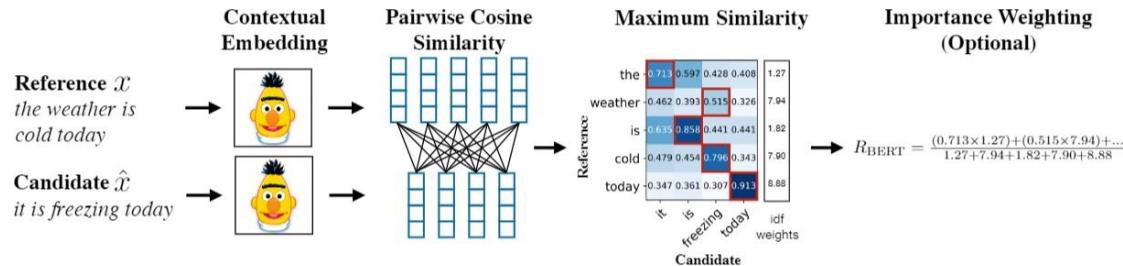
Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.

(Kusner et.al., 2015; Zhao et al., 2019)

## BERTSCORE

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.

(Zhang et.al. 2020)



# Human evaluations

Ask *humans* to evaluate the quality of generated text

- Overall or along some specific dimension:
  - fluency
  - coherence / consistency
  - factuality and correctness
  - commonsense
  - style / formality
  - grammaticality
  - typicality
  - Redundancy
  - ...

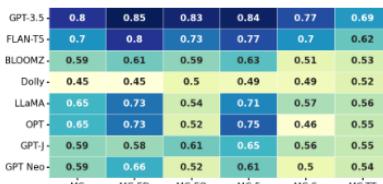
	Criterion	Brief comments (with related and (near-)equivalent criteria)
0	Fluency (solved)	(Naturalness) Does the turn pass as a manually composed text?
1	Coherence	(Relevance) Does the turn make sense as a response to the previous user turn?
2	Sensibleness	No common sense mistakes, no absurd responses
3	Correctness	Is the nugget factually correct?
4	Groundedness	Is the nugget based on some supporting evidence?
5	Explainability	Can the user see how the system came up with the nugget?
6	Sincerity	Is the nugget likely to be consistent with the system's internal results?
7	Sufficiency	(Recall) Does the turn satisfy the requests in the previous user turn?
8	Conciseness	Is the system turn minimal in length?
9	Modesty	(Confidence) Does the system's confidence about the nugget seem appropriate?
10	Engagingness	(Interestingness, Topic breadth) Does the system nugget/turn make the user want to continue the conversation?
11	Recoverability	Does the system turn keep the user interacting after the user has expressed dissatisfaction?
12	Originality	(Creativity) Is the nugget original, and not a copy of some existing text?
13	Fair exposure	Does the system mention different groups fairly across its turns?
14	Fair treatment	Does the system provide the same benefit to different users and user groups?
15	Harmlessness	(Safety, Appropriateness) No threats, no insults, no hate or harassment, etc.
16	Consistency	Given the nuggets seen so far, is the present nugget logically possible?
17	Retentiveness	Does the system "remember"?
18	Robustness to input variations	Does the system eventually provide the same information no matter how we ask?
19	Customisability	(Personalisability) Does the system adapt to different users and user groups?
20	Adaptability	Does the system keep up with the changes in the world?

# Analyzing Temporal Reasoning Abilities of LLMs

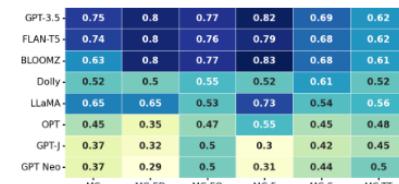
## [EMNLP'23]

Language Model	Params	Architecture	Type	Few-Shot	Zero-Shot	CoT	Code-Prompt
GPT-J	6B	Autoregressive Decoder only	Base	✓	✓		
GPT Neo	1.3B	Autoregressive Decoder only	Base	✓	✓		
LLaMA	7B	Autoregressive Decoder only	Base	✓	✓		
OPT	350M	Autoregressive Decoder only	Base	✓	✓		
BLOOMZ	560M	Autoregressive Decoder only	SIFT	✓	✓	✓	
Dolly	3B	Autoregressive Decoder only	SIFT	✓	✓	✓	
FLAN-T5	780M	Encoder-Decoder	SIFT	✓	✓	✓	
SantaCoder	1.1B	Autoregressive Decoder only	Base				✓
CodeGen2	2B	Encoder-Decoder	Base				✓
GPT-3.5	-	-	RLHF	✓	✓	✓	✓

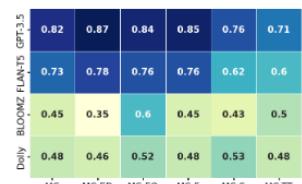
Strong performance of LLMs on event frequency, and duration tasks.



(a) Heatmap for Accuracy of LLMs in Few-shot setting



(b) Heatmap for Accuracy of LLMs in Zero-shot setting



(c) Heatmap for Accuracy of LLMs in CoT setting

Mixed performance on event ordering tasks.

LLMs struggle with specific event timings.

# Dataset examples

Past Reasoning	A: Your resume says you have had <MASK> experience working in a foreign representative office in Shanghai , may I ask why you quit ? B: I worked in a foreign office for one year . However , I leave there two years ago because the work they gave was rather dull .	Answer: 1 Year	Label: Yes
Future Reasoning	A: What is this eviction notice for ? B: The notice you received is a <MASK> notice to vacate A: You can't just throw me out on the street ! B: You have 30 days to catch up on your rent , or a sheriff will evict you	Answer: 30 Days	Label: Yes

(a) Dataset Example of past reasoning vs future reasoning

Second TimeFrame	A: How are the children doing at sport ? B: I ' m very pleased with their performances . Timmy can cover the 100 meters in <MASK> . That ' s very fast for kid his age .	Answer: 12 sec	Label: Yes
More than a Day TimeFrame	A: How much does it cost to hire a motorbike ? B: For <MASK> \$ 300 , for 7 days \$ 600	Answer: 4 Days	Label: Yes

(b) Dataset Example of second vs more than 1 day timeframe

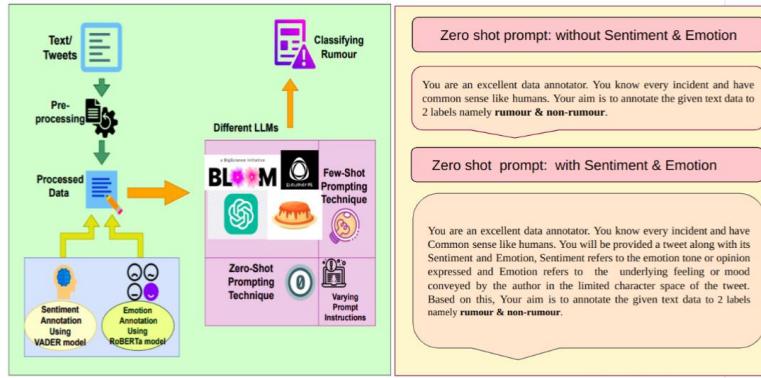
Ambiguous Temporal Expression	Context: The use of non-covariant objects like pseudotensors was heavily criticized in 1917 by Erwin Schrodinger and others.	Question: How long did Schrodinger oppose pseudotensors?	Answer: his whole career	Label: Yes
Exact Temporal Expression	Context: At about 9:20, security personnel at FAA headquarters set up a hijacking with several agencies, including the Defense Department.	Question: How long did the security take to set up a hijacking ?	Answer: One hour	Label: Yes

(a) Dataset Example of Exact TE vs Ambiguous TE

Single Event	Context: The Beatles are giving a press conference about their new film , Magical Mystery Tour	Question: What time of day was the press conference?	Answer: 12:00 PM	Label: Yes
Multiple Event	Context: Durer's father died in 1502, and his mother died in 1513.	Question: How often did Durer visit his mother's grave?	Answer: Every year	Label: Yes

(b) Dataset Example of Single Event vs Multiple Event

# Investigating Abilities of LLMs for Rumor Detection



LLM Used → GPT-3.5	Zero-Shot		Few-Shot	
DATASET ↓	C-SE	R-SE	C-SE	R-SE
PHEME	0.630	0.540	<b>0.680</b>	0.306
POLITIFACT	<b>0.494</b>	0.336	0.286	0.120
GOSSIPCOP	0.296	0.282	0.318	<b>0.360</b>

**Table 3.** Accuracy results for different dataset combinations and GPT-3.5 models are presented in both Few-Shot and Zero-Shot scenarios, covering both correctly and randomly labelled sentiments and emotions (C-SE and R-SE, respectively).

LLM Used →	Few-Shot							
	GPT-3.5		BLOOM		FLAN-T5		GPT-Neo	
DATASET ↓	WO-SE	W-SE	WO-SE	W-SE	WO-SE	W-SE	WO-SE	W-SE
PHEME	0.640	<b>0.680</b>	0	0	0.592	0.002	0.556	0
POLITIFACT	0.326	0.286	<b>1</b>	0	0.748	0.024	0	0
GOSSIPCOP	<b>0.430</b>	0.318	0	0	0.236	0.034	0	0
SNOPES	0.274	0.176	0	0	<b>0.368</b>	0.264	0	0
IFND	<b>0.356</b>	0.344	0	0	0.068	0.022	0	0
ESOC COVID-19	<b>0.262</b>	0.086	0	0	0.176	0.060	0	0

LLM Used → GPT-3.5	temperature = 0.2		temperature = 0.7		temperature = 1.0	
	Zero-Shot	Few-Shot	Zero-Shot	Few-Shot	Zero-Shot	Few-Shot
PHEME (WO-SE)	0.476	0.082	0.638	<b>0.640</b>	0.430	0.126
PHEME (W-SE)	0.376	0.048	0.630	<b>0.680</b>	0.404	0.094
POLITIFACT (WO-SE)	0.194	0.422	<b>0.626</b>	0.326	0.212	0.444
POLITIFACT (W-SE)	0.170	0.264	<b>0.494</b>	0.286	0.200	0.308
GOSSIPCOP (WO-SE)	0.472	0.363	<b>0.506</b>	0.430	0.468	0.414
GOSSIPCOP (W-SE)	0.362	0.188	0.296	0.318	<b>0.378</b>	0.252

**Table 2.** Accuracy results for diverse dataset combinations, with and without sentiment and emotions, are shown for GPT-3.5 models in both Few-Shot and Zero-Shot settings, using varying temperature hyperparameters.

# Reading Materials

- Yang et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, 2023 (<https://arxiv.org/pdf/2304.13712.pdf>)
- Zhao et al. **A Survey of Large Language Models**, arXiv, 2023 (<https://arxiv.org/pdf/2303.18223.pdf>)
- Challenges and Applications of Large Language Models (<https://arxiv.org/pdf/2307.10169.pdf>)
- Transformers & self-attention:
  - <https://towardsdatascience.com/transformers-intuitively-and-exhaustively-explained-58a5c5df8dbb>
  - <https://towardsdatascience.com/de-coded-transformers-explained-in-plain-english-877814ba6429>
  - <https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a>
  - <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>

# Watching Materials

- Visual Guide to Transformer Neural Networks
  - 17:47 [Episode 0](#) - [OPTIONAL] The Neuroscience of "Attention"
  - 12:22 [Episode 1](#) - Position Embeddings
  - 15:24 [Episode 2](#) - Multi-Head & Self Attention
  - 16:04 [Episode 3](#) - Decoder's Masked Attention