universität innsbruck

Brief Introduction to Natural Language Processing & IR

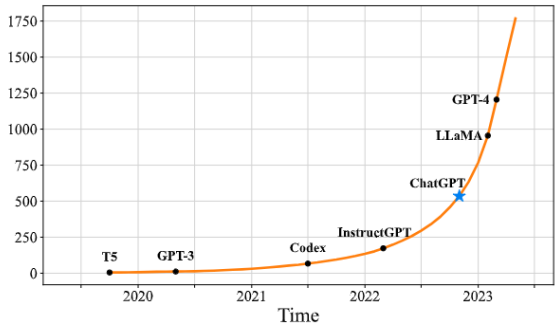
Adam Jatowt

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



(b) Query="Large Language Model"

8.5 papers/day on arXiv that contain "large language model" in title or abstract

Zhao et al. A Survey of Large Language Models, arXiv, 2023

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 ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ?
    ⇒ (un + lock) + able ?

Part Of Speech (POS) Tagging

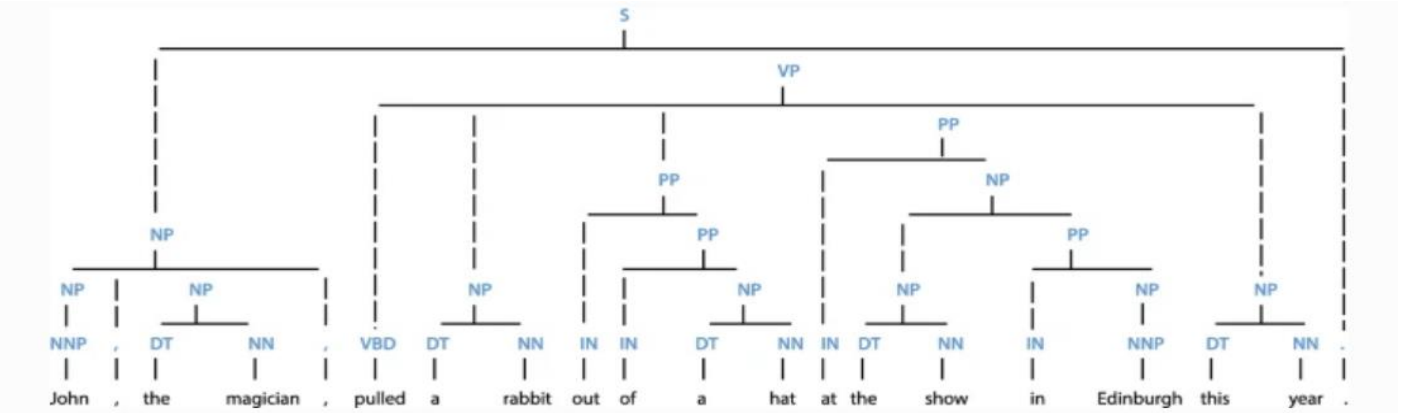
- Annotate each word in a sentence with a part-of-speech.

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."

Brenda Salenave Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, Sérgio Nunes: A survey on narrative extraction from textual data. Artif. Intell. Rev.

Example Semantic Tasks

Word Sense Disambiguation (WSD)

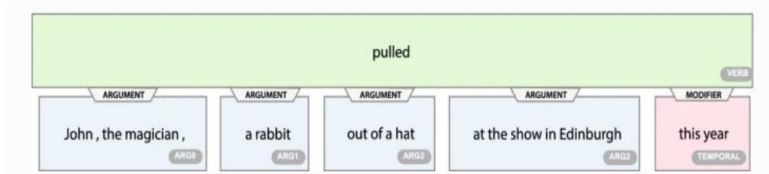
- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- For many tasks (e.g., question answering, translation), the proper sense of each ambiguous 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
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	*Yahoo bought Overture.*	TRUE
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.	*Microsoft bought Star Office.*	FALSE
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.	*Israel was established in May 1971.*	FALSE
Since its formation in 1948, Israel fought many wars with neighboring Arab countries.	*Israel was established in 1948.*	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 impies 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

https://readability-check.org/

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 constitutational law; at the University of Chicago Law School from 1992 to 2004.

Barack Obama

44th President of the United States

In office

Question Answering

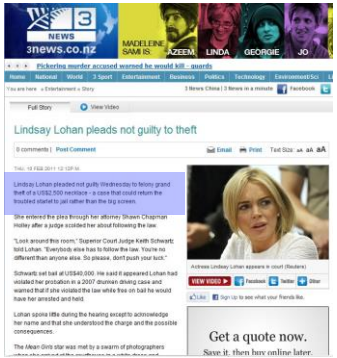
- Directly answer natural language questions based on information presented in a corpus of textual documents (e.g., the web).
  - When was Barack Obama born? (*factoid*)
    - August 4, 1961
  - Who was president when Barack Obama was born?
    - John F. Kennedy
  - How many presidents have there been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

Text Summarization (Abstractive)

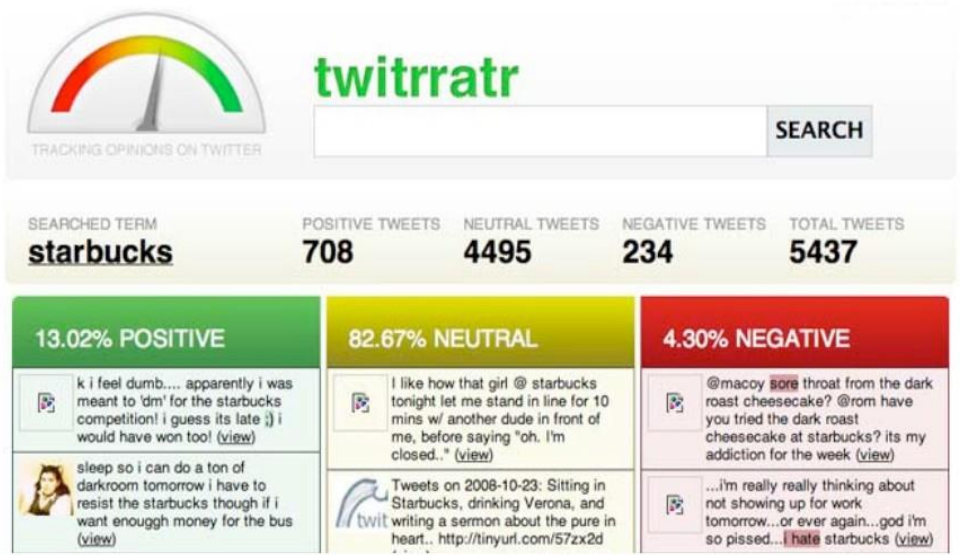
- Produce a short summary of a longer document or article.
  - Article: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary 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.

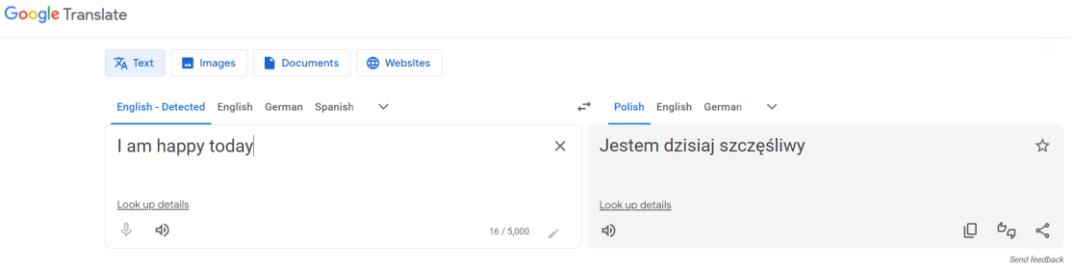


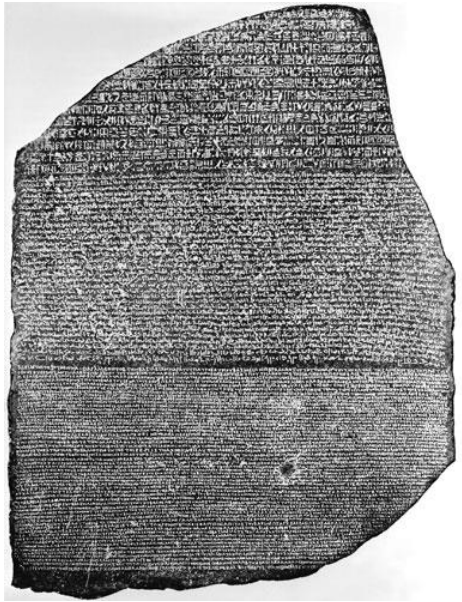
Sentiment/Opinion Analysis



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.



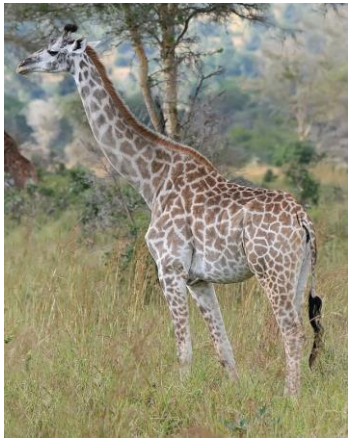


Commonsense reasoning

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

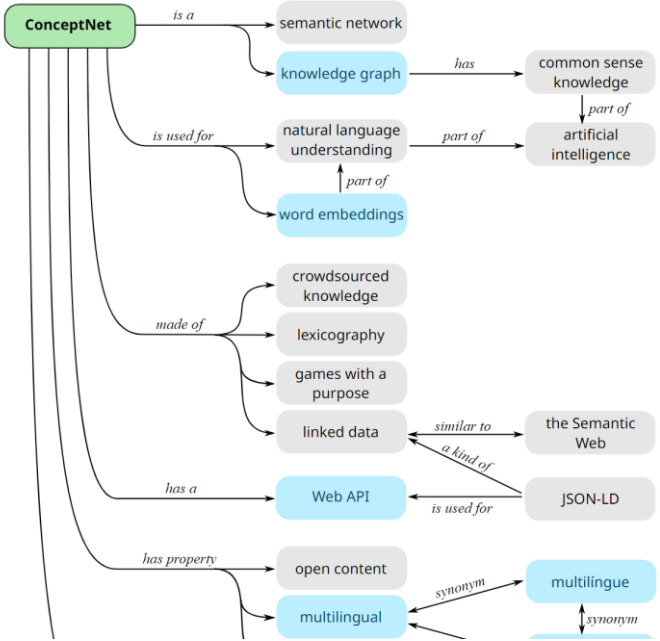
Going for a walk takes less time
than going for vacation

<



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)



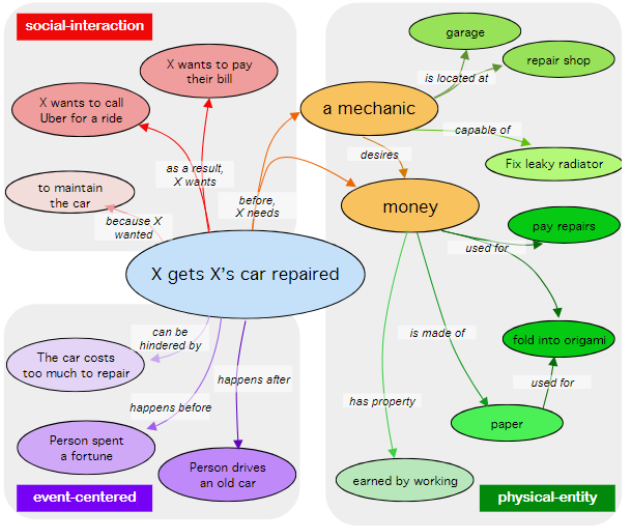


Figure 1: A tiny subset of ATOMIC$^{20}_{20}$, a large atlas of social and physical commonsense relations. Relations in the top-left quadrant reflects relations from ATOMIC.[1]

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

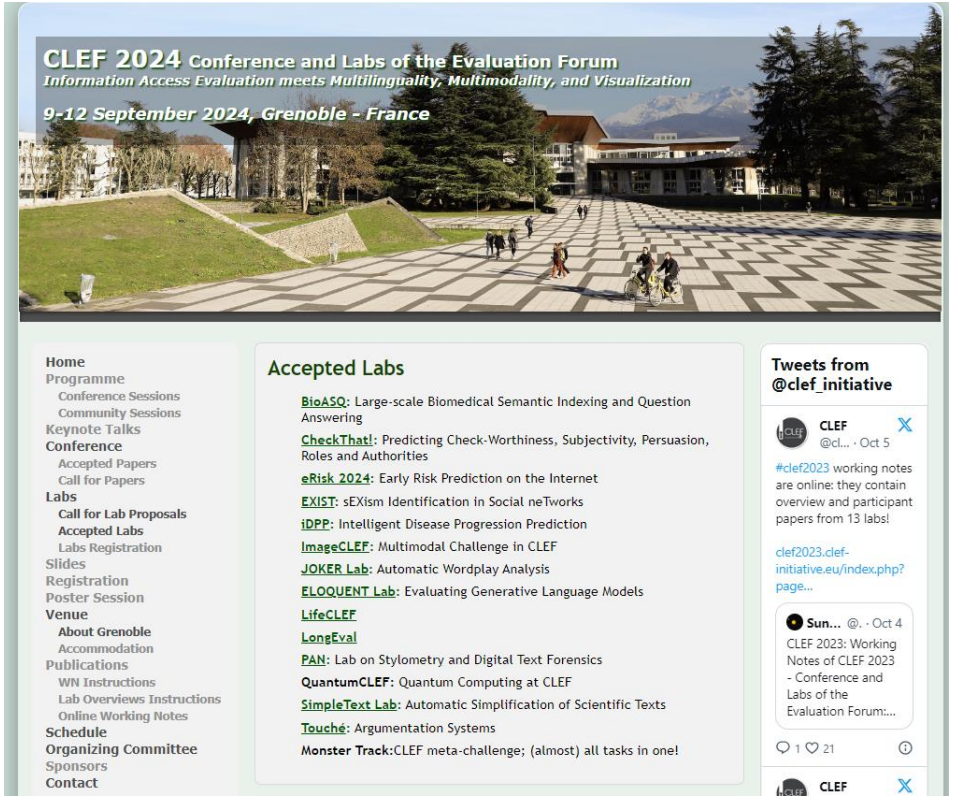
A headline and a body text - either from the same news article or from two different articles.

Output

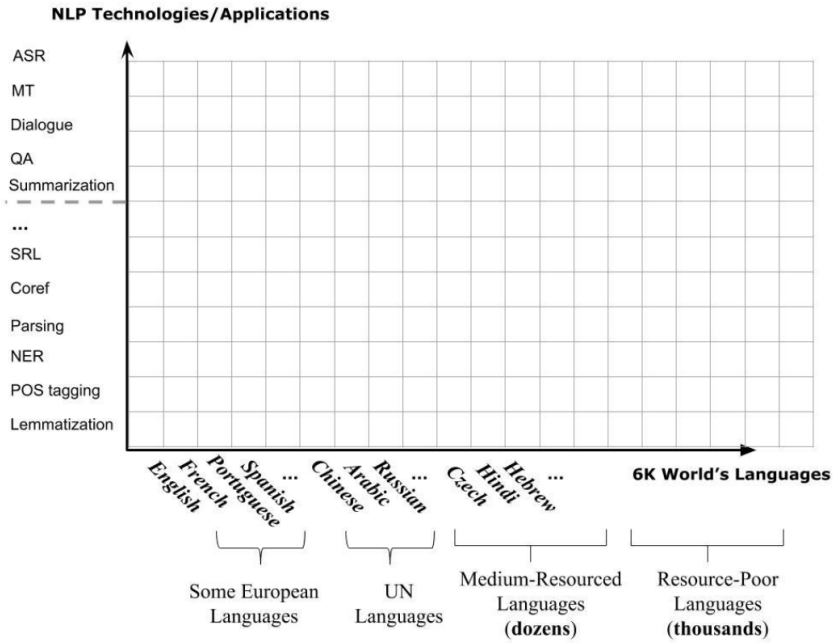
Classify the stance of the body text relative to the claim made in the headline into one of four categories:

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

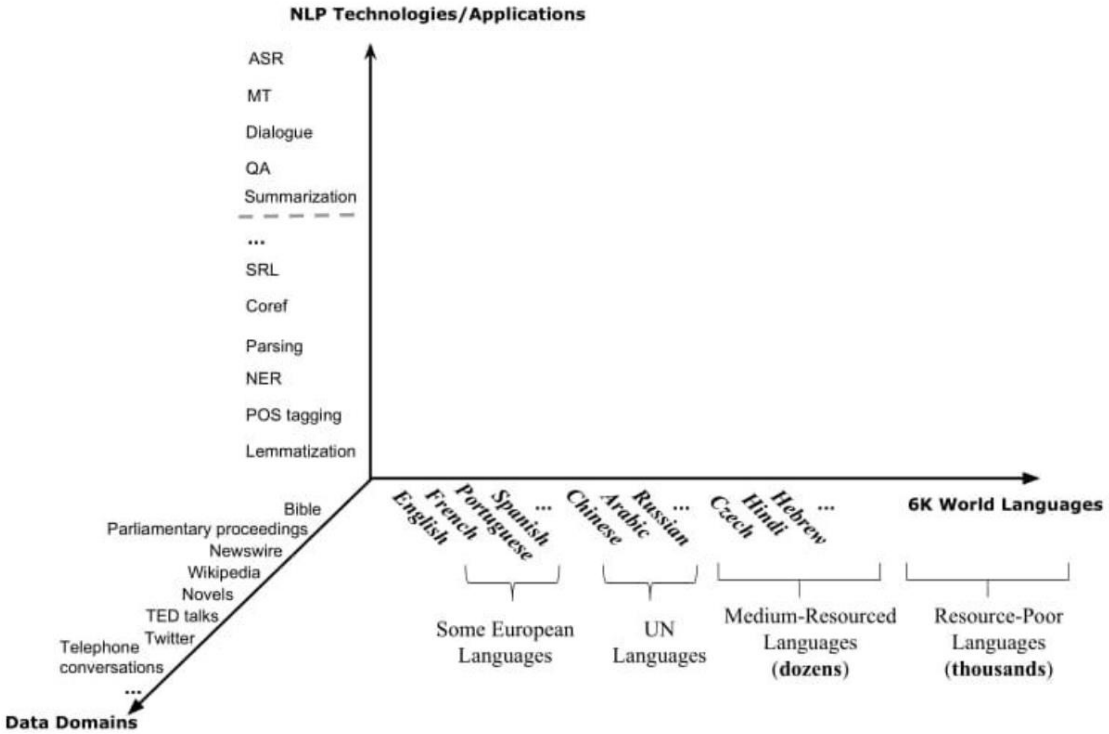
CLEF 2024 Tasks Examples



Challenge: Many languages..



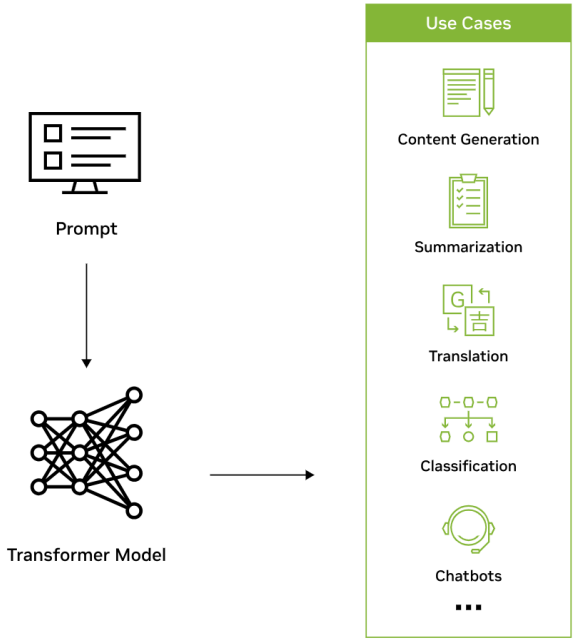
Challenge: Many languages and domains..



Introduction to LLMs

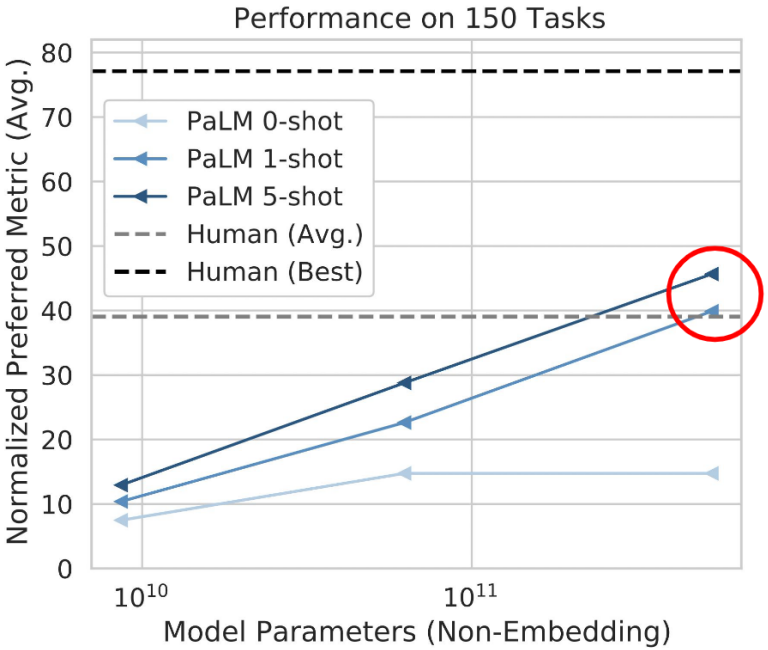
Large Language Models (LLMs)

- 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.



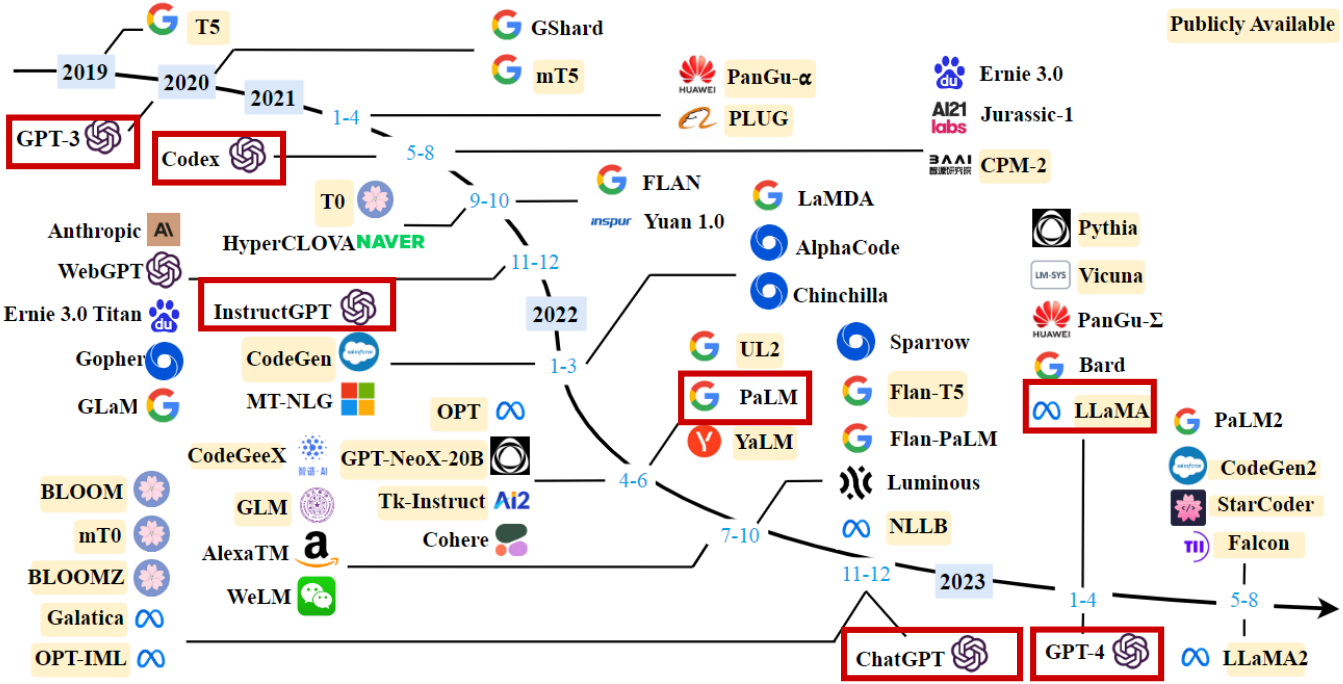
Credits: Nvidia

LLMs are making astonishing progress on many complex tasks



A Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*, JMLR, 2023

A Timeline of Existing LLMs with Size >10B Parameters

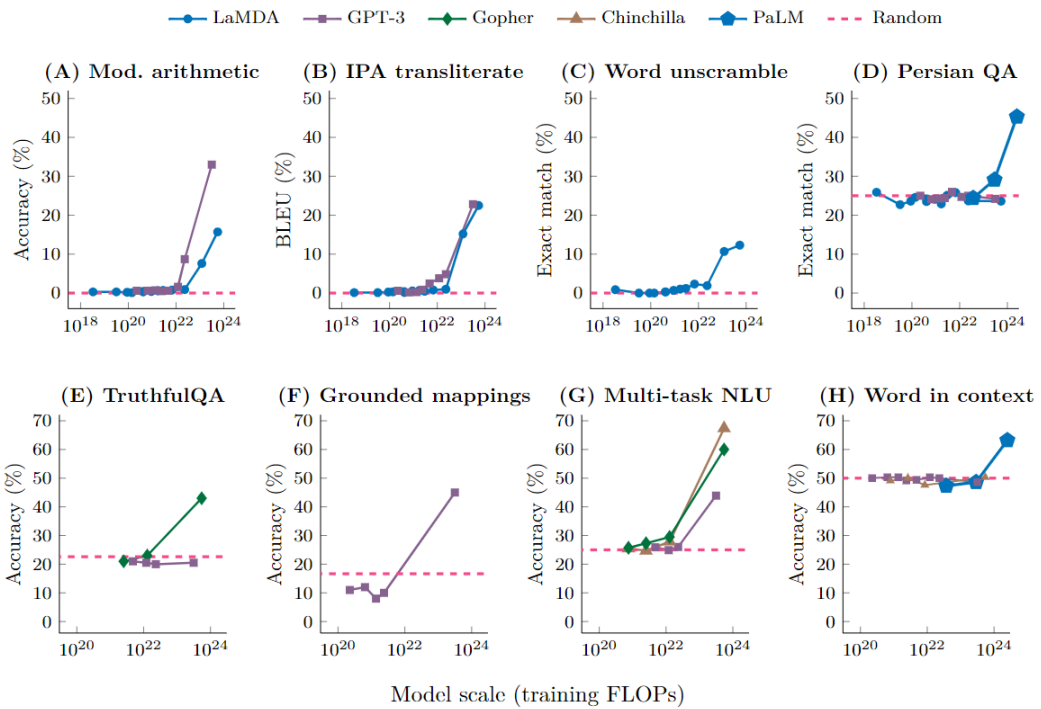


- **OpenAI** consistently maintains its leadership position with other companies having difficulty to catch up
- **Meta** contributes significantly to open-source LLMs, one of the most generous companies with all LLMs developed being open (e.g., Llama)
- LLMs exhibit an **increasing tendency to be closed-source** (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

Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

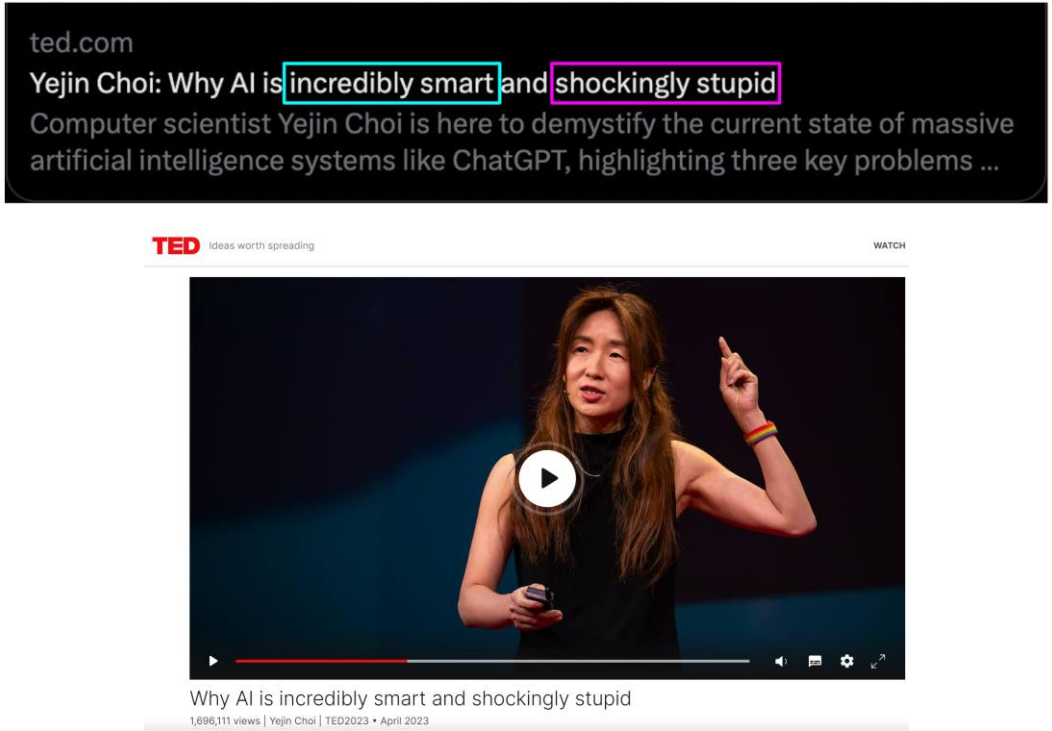
Effect of Scale: Emergent Abilities of Large Language Models

- 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..



Wei, Jason, et al. "Emergent abilities of large language models." *arXiv preprint arXiv:2206.07682* (2022).

However, there are (still) some problems..



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

And dangers..

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

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

Emily M. Bender

Professor, University of Washington

Illustration by TIME; reference image courtesy of Emily Bender

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

Emily 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.

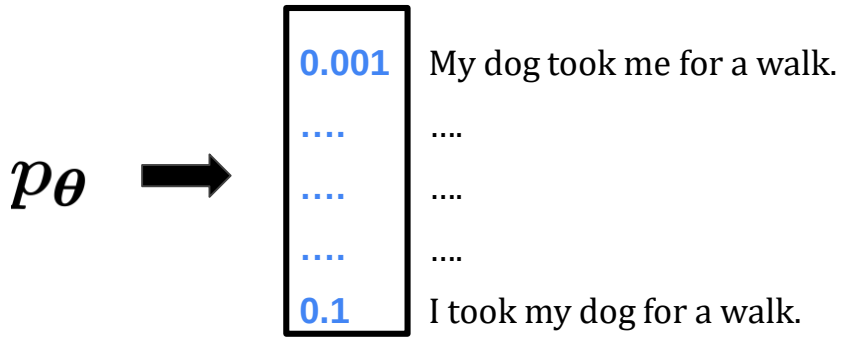
Emily M. Bender, et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). 610–623.

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

Language Modeling

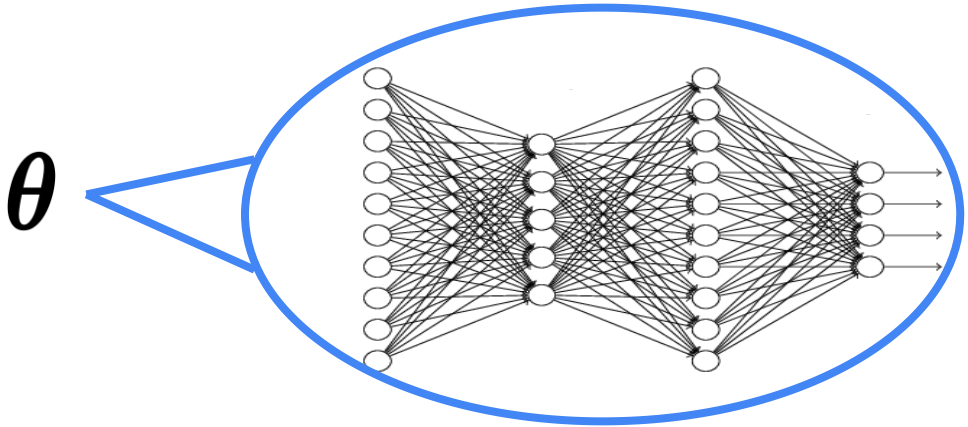
General Definition of a Language Model

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



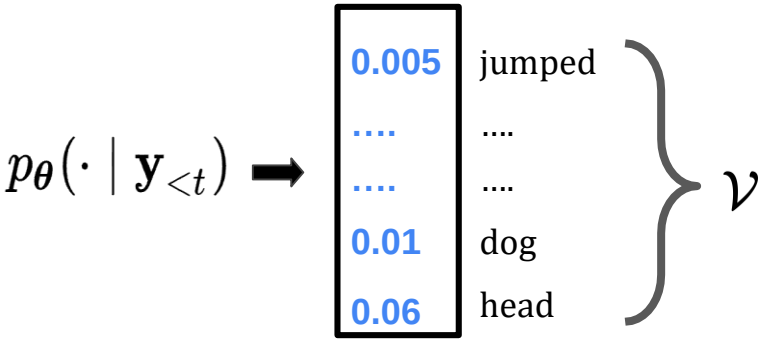
Language Models: Definition

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

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 $\mathbf{y}_{<t}$



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

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 ***D*** under our model $p_{\boldsymbol{\theta}}$
- Equivalent to minimizing the cross-entropy with the empirical distribution, as defined by a training corpus ***D***

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

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

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

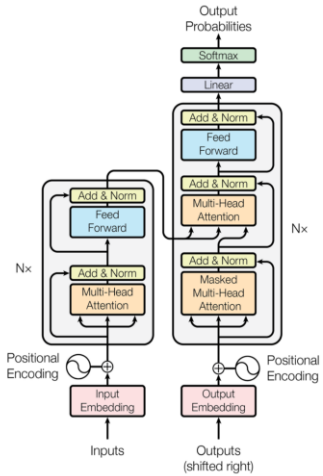


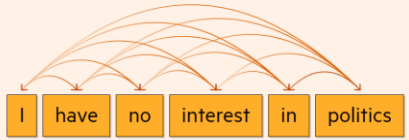
Figure 1: The Transformer - model architecture.

Image credits: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin. *Attention is all you need*. NeurIPS'17.

Self-attention

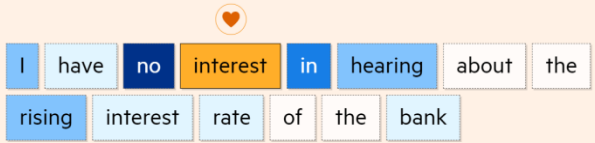
Self-attention looks at each **token** in a body of text and decides which others are most important to understanding its meaning.

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.



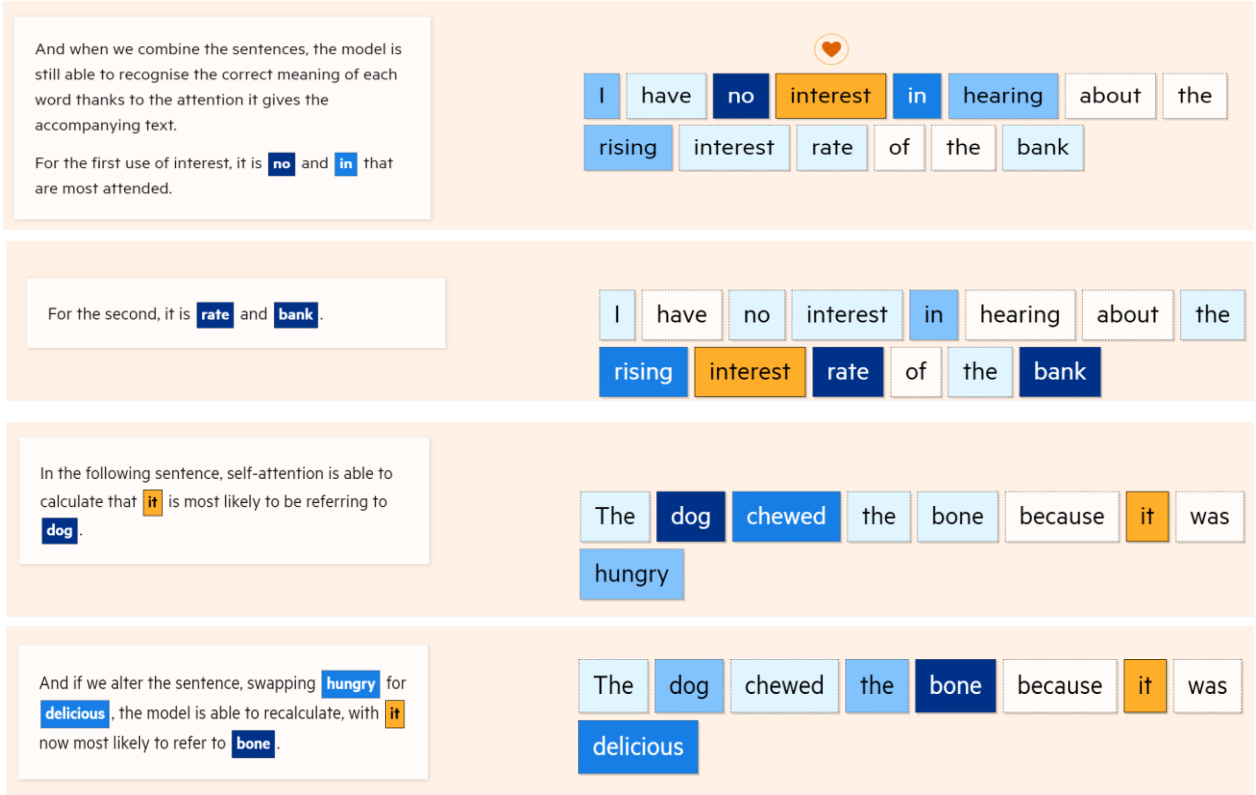
And when we combine the sentences, the model is still able to recognise the correct meaning of each word thanks to the attention it gives the accompanying text.

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



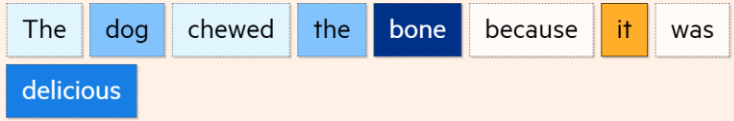
https://ig.ft.com/generative-ai/

Self-attention Idea



https://ig.ft.com/generative-ai/

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**.



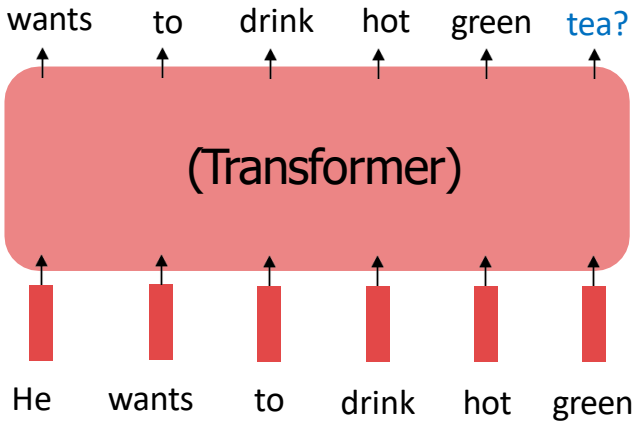
The benefits of self-attention for language processing increase the more you scale things up. It allows LLMs to take **context** from beyond sentence boundaries, giving the model a greater understanding of how and when a word is used.

https://ig.ft.com/generative-ai/

Pre-training

Pre-train (on language modeling)

Lots of text; learn general things



What kinds of things does pre-training teach?

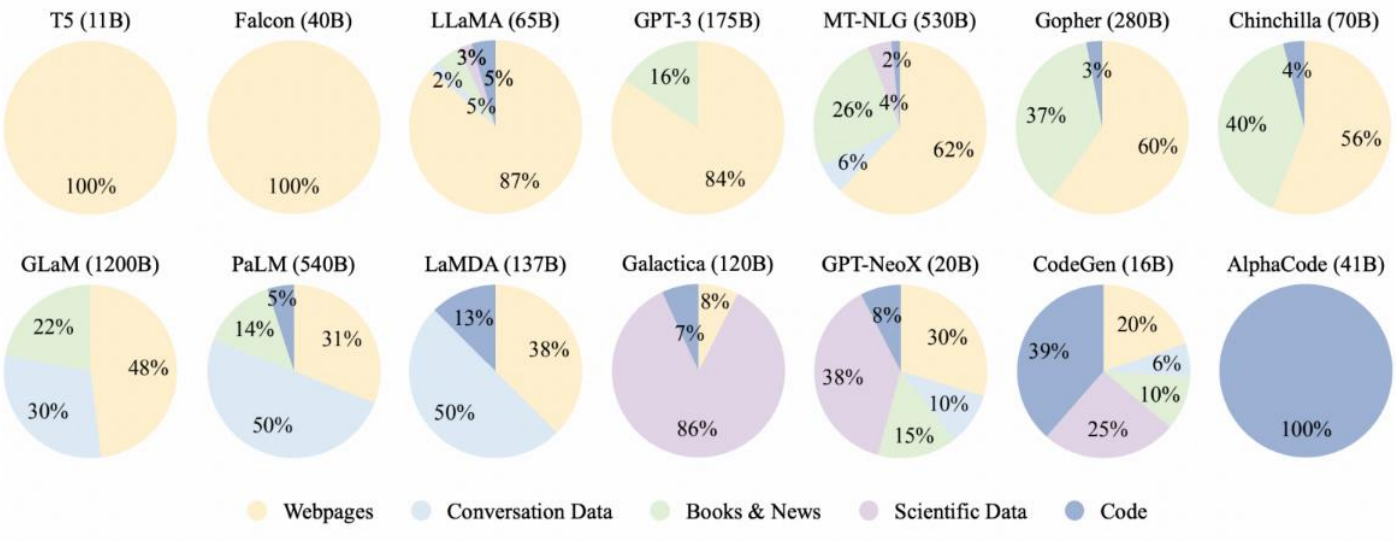
There's increasing evidence that pre-trained models learn a wide variety of things about the statistical properties of language:

- *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

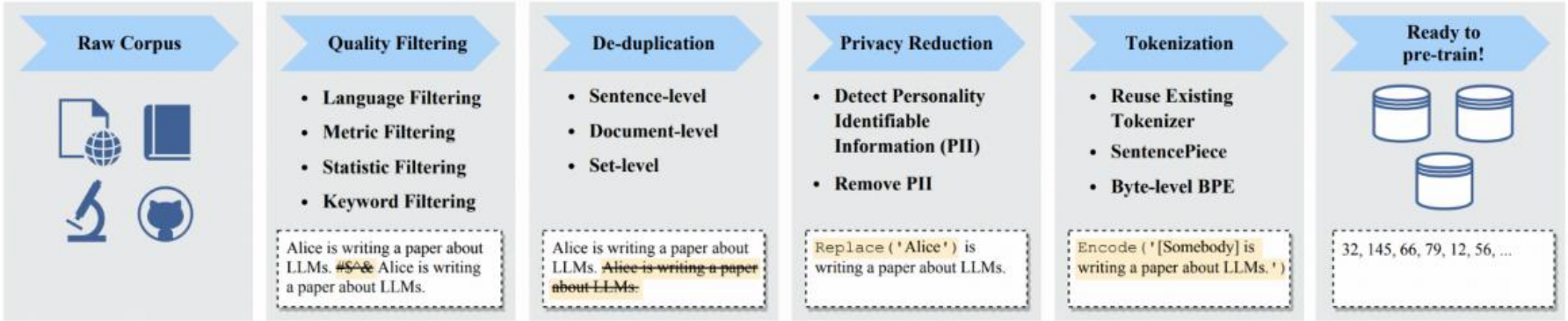


Ratio of data sources in example recent LLMs

Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

Coomon 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

Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

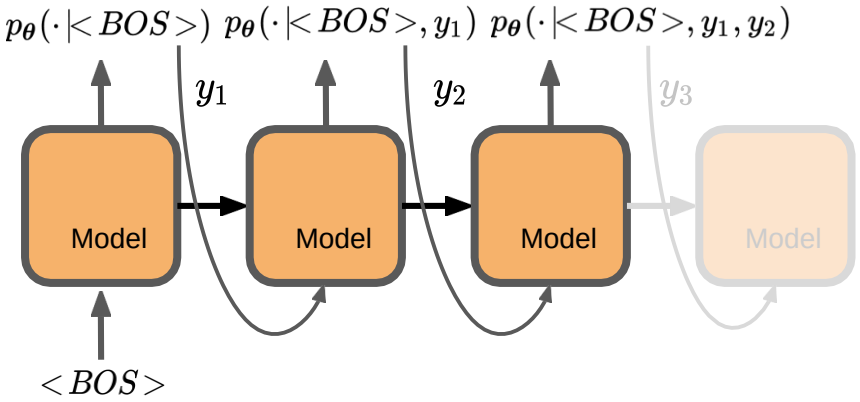
Text Generation (Decoding)

Generating text

- Modeling the conditional probability distribution over each word

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

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



How do we choose *y* at each step?

Generating from Locally Normalized Distributions

$$p_{\boldsymbol{\theta}}(\cdot \mid <BOS>, \mathrm{My}) \Rightarrow$$

0.001	jumped
….	….
….	….
0.15	dog
0.1	head

Generating from Locally Normalized Distributions

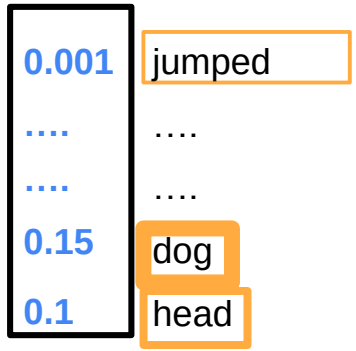
$p_{\boldsymbol{\theta}}(\cdot \mid <BOS>, \text{My})$ ➡

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

$$y_t = \underset{y \in \mathcal{V}}{\text{argmax}} \; p_{\boldsymbol{\theta}}(y \mid \mathbf{y}_{<t})$$

Generating from Locally Normalized Distributions

$$p_{\boldsymbol{\theta}}(\cdot \mid <BOS>, \mathrm{My})$$



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

$$y_t \sim p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{y}_{<t})$$

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)

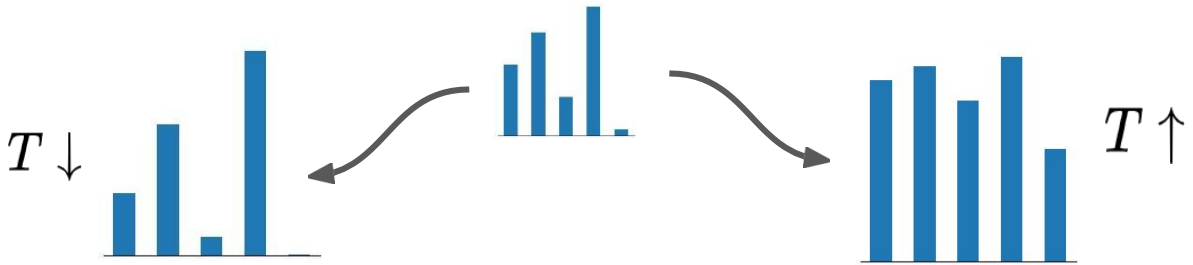
Consequences of temperature sampling

Lower the temperature < 1: probability becomes more spiky

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

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 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

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 .....

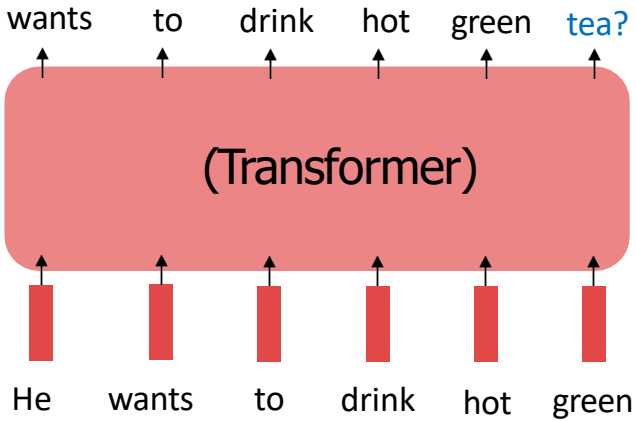
Flat distributions lead to incoherent text

Finetuning

The Pre-training & Fine-tuning Paradigm

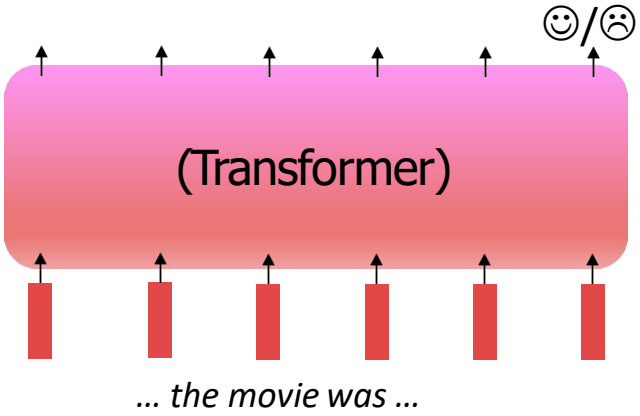
Step 1: Pre-train (on language modeling)

Lots of text; learn general things.



Step 2: Fine-tune (on your task)

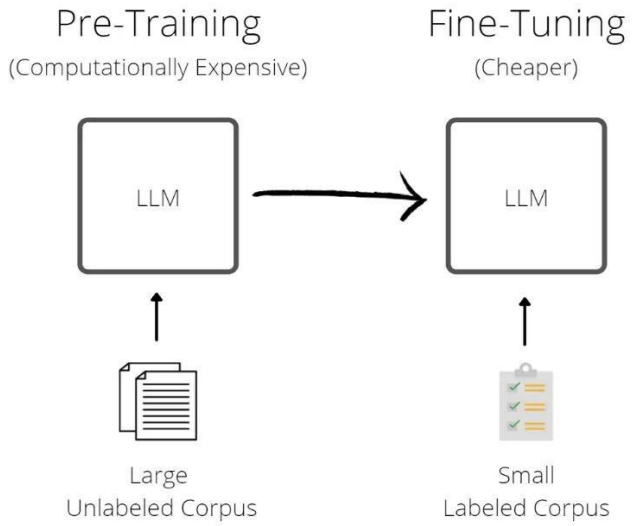
Not many labels; adapt to specific tasks.



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

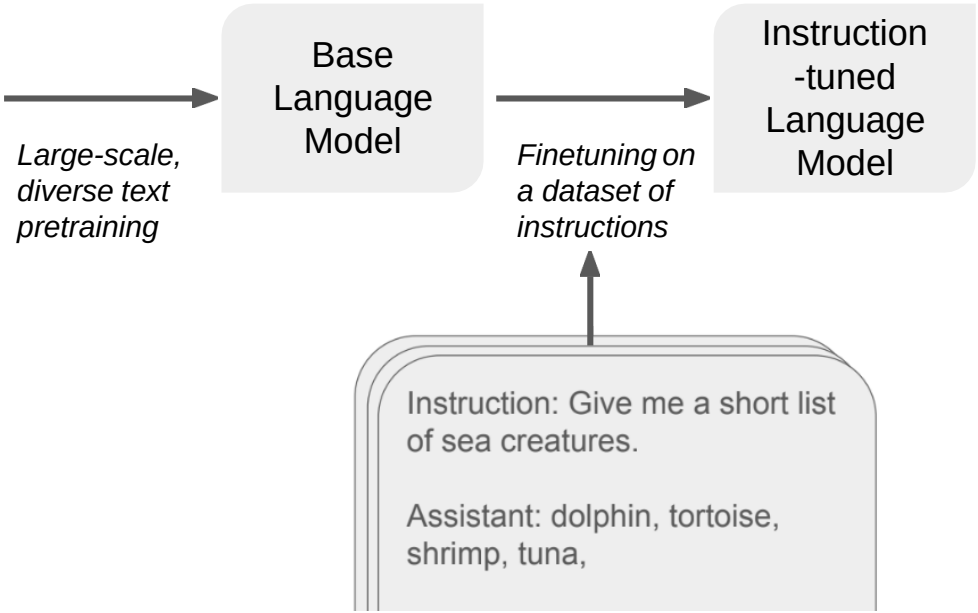


More computationally efficient for developers, since an LLM needs only be trained once..

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

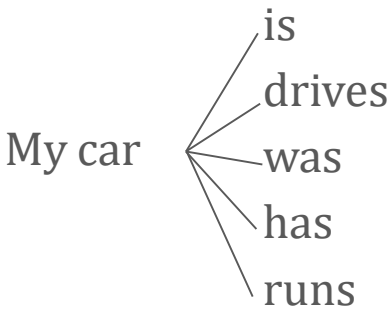
- As is usually the case, **data + model scale** is key for this to work
- For example, the **Super-NaturalInstructions** dataset contains **over 1.6K tasks, 3M+** examples
  - Classification, sequence tagging,
    - rewriting, translation, QA...

The Art of Prompting

Why does prompting work?

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

Context (often) lowers entropy



High entropy; many plausible options

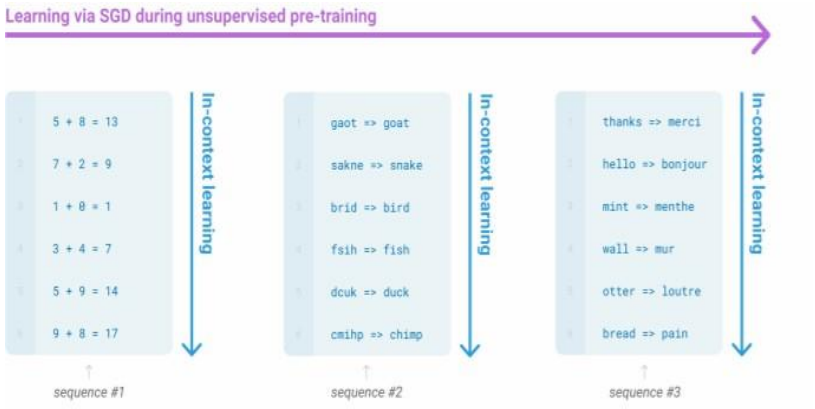
It was so, so cold
this morning. My car —— battery

(didn't start)

Fewer options; some are much more likely than before

In-context learning in very large models

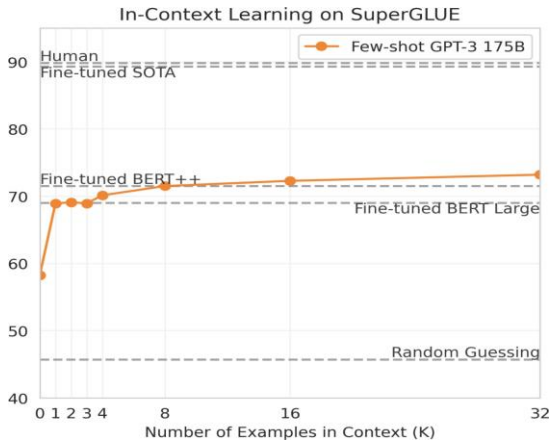
Very large language models (e.g., GPT-3) seem to perform some kind of **learning without traditional machine learning procedures** simply from examples you provide within their contexts (prompts)…



Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

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.

```
Translate English to French:    ← task description
cheese =>                       ← prompt
```

One-shot

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

```
Translate English to French:    ← task description
sea otter => loutre de mer      ← example
cheese =>                       ← prompt
```

Few-shot

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

```
Translate English to French:    ← task description
sea otter => loutre de mer      ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                       ← prompt
```

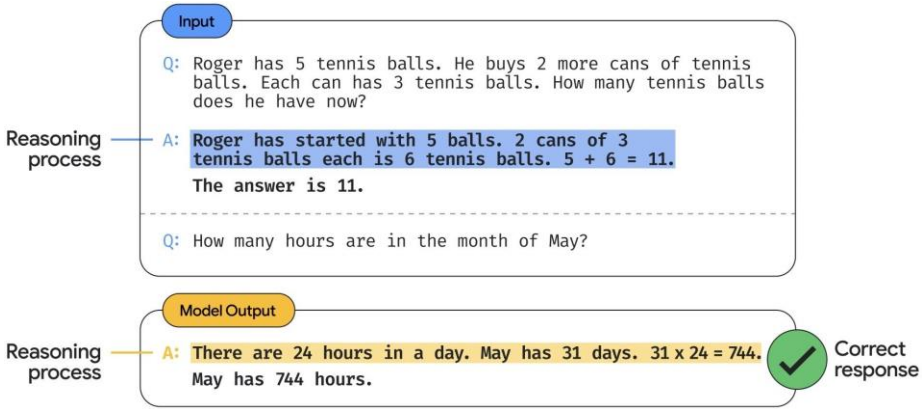
https://medium.com/@haifengl/a-tutorial-to-llm-f78dd4e82efc

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

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

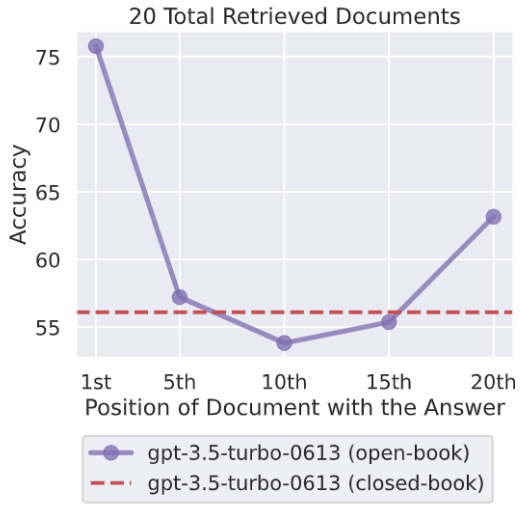
	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7

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.	**82.0**
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

Kojima, Takeshi, et al. "Large language models are zero-shot reasoners." *Advances in neural information processing systems* 35 (2022): 22199-22213.

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



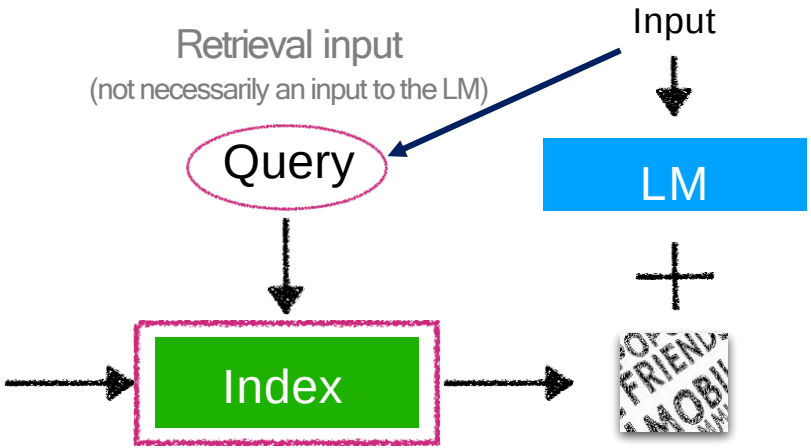
Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *arXiv preprint arXiv:2307.03172* (2023)

RAG

Retrieval-augmented Large Language Models

Retrieval-augmented LLMs = Retrieval + LLMs

Datastore



Find a small subset of documents in a datastore that are the most similar to the query

Why retrieval-based LMs?

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

List 5 important papers authored by Geoffrey Hinton

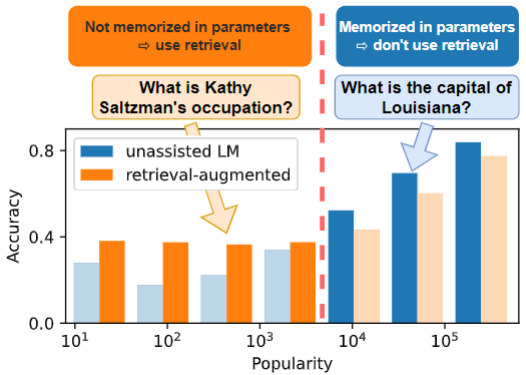
ChatGPT

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

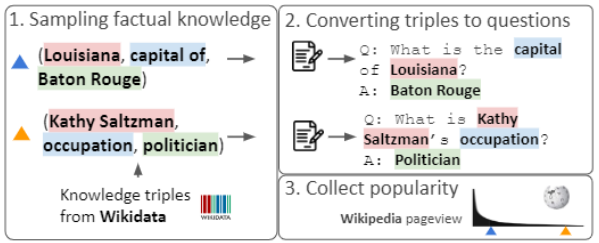
1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, .. ✓
2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, .. ✓

…

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2023.

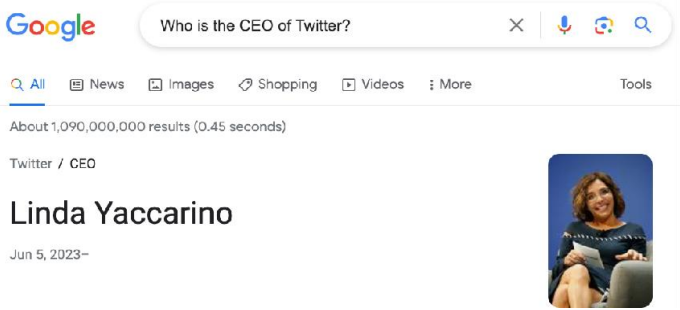
Why retrieval-based LMs?

LLMs' knowledge is easily outdated and hard to update

Who is the CEO of Twitter?

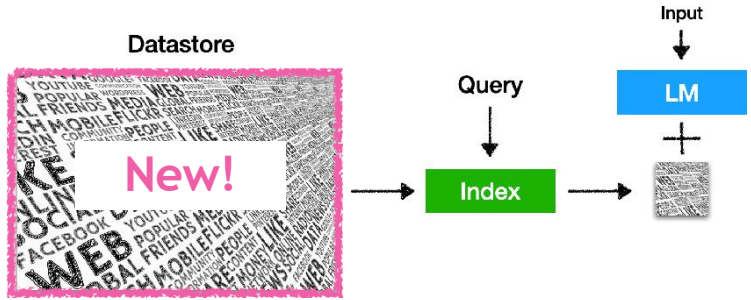
ChatGPT

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



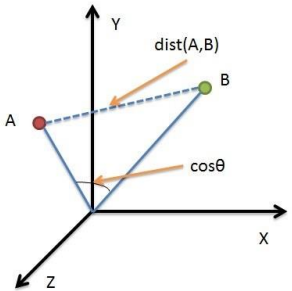
- Existing **knowledge editing** methods are still NOT scalable (ongoing research)

- The datastore can be easily **updated** and **expanded** even without retraining



Evaluating LLMs

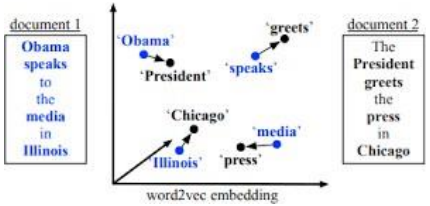
Model-based metrics: Word distance functions



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)

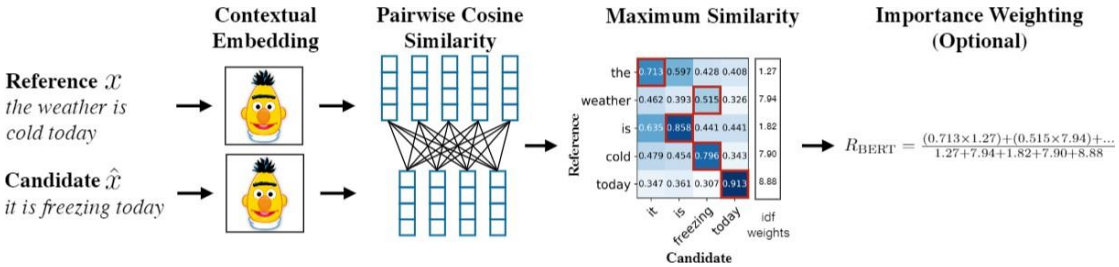


Word Mover's Distance

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?

Tetsuya Sakai: SWAN: A Generic Framework for Auditing Textual Conversational Systems. CoRR abs/2305.08290 (2023)

Analyzing Temporal Reasoning Abilities of LLMs [EMNLP'23]

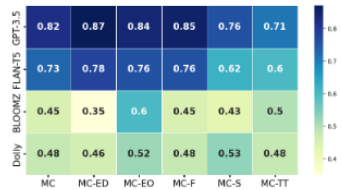
Language Model	Params	Architecture	Type	Few-Shot	Zero-Shot	CoT	Code-Prompt
GPT-J	6B	Autoregressive Decoder only	Base	✓	✓		
GPT Neo	1.3B	Autoregressive Decoder only	Base	✓	✓		
LLaMA	7B	Autoregressive Decoder only	Base	✓	✓		
OPT	350M	Autoregressive Decoder only	Base	✓	✓		
BLOOMZ	560M	Autoregressive Decoder 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.

Performance drop on understanding event temporal states.

(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.

R. Jain et al.: Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models, Proc. of EMNLP 2023

Dataset examples

		Answer:	Label:
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 .	1 Year	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	30 Days	Yes

(a) Dataset Example of past reasoning vs future reasoning

		Answer:	Label:
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 .	12 sec	Yes
More than a Day TimeFrame	A: How much does it cost to hire a motorbike ? B: For <MASK> $ 300 , for 7 days $ 600	4 Days	Yes

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

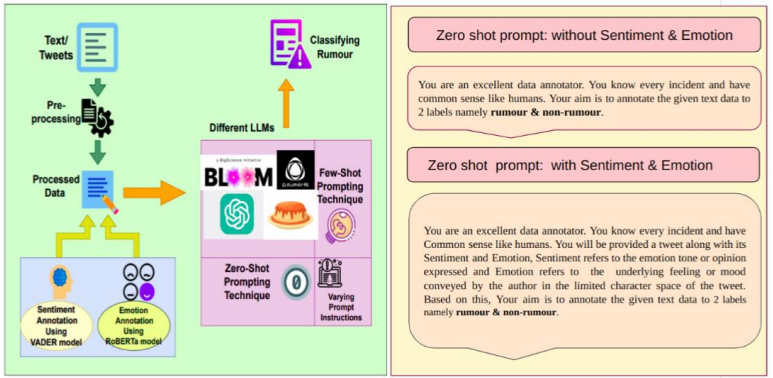
			Answer:	Label:
Ambiguous Temporal Expression	Context: The use of non-covariant objects like pseudotensors was heavily criticized in 1917 by Erwin Schrodinger and others.	Question: How long did Schrodinger oppose pseudotensors?	his whole career	Yes
Exact Temporal Expression	Context: At about 9:20, security 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 ?	One hour	Yes

(a) Dataset Example of Exact TE vs Ambiguous TE

			Answer:	Label:
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?	12:00 PM	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?	Every year	Yes

(b) Dataset Example of Single Event vs Multiple Event

Investigating Abilities of LLMs for Rumor Detection



	Few-Shot							
LLM Used →	*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	**0.680**	0	0	0.592	0.002	0.556	0
POLITIFACT	0.326	0.286	**1**	0	0.748	0.024	0	0
GOSSIPCOP	**0.430**	0.318	0	0	0.236	0.034	0	0
SNOPES	0.274	0.176	0	0	**0.368**	0.264	0	0
IFND	**0.356**	0.344	0	0	0.068	0.022	0	0
ESOC COVID-19	**0.262**	0.086	0	0	0.176	0.060	0	0

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

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

A Blessing in Disguise: Can Large Language Models be Useful for Misinformation Detection? (under review at ECIR 2024)

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