CSCI 5541: Natural Language Processing

Lecture 2: Introduction to NLP

James Mooney

computer science
& engineering

MINNESOTA NLP EST. 2021

UNIVERSITY OF MINNESOTA
Driven to Discover®

Recitation and In-class Tutorials (next week) Announcement to Come Tomorrow via Slack

- Computing basics
  - Setting up environment for PyTorch and Transformers
  - Pytorch Basics Tutorial
- Tutorial on SciKit-learn/PyTorch
- Tutorial on HuggingFace/vLLM

Announcement

- If you miss the first class, please check out the course details in the lecture slides
- Share your interests and project ideas in #random channel and actively look for your teammates. Team formation is due on Feb 6.
- If you are enrolled but not invited to Slack, please send James an email.
- HW1 out tomorrow (Due: Feb 4)
- OH out tomorrow on course website

Outline

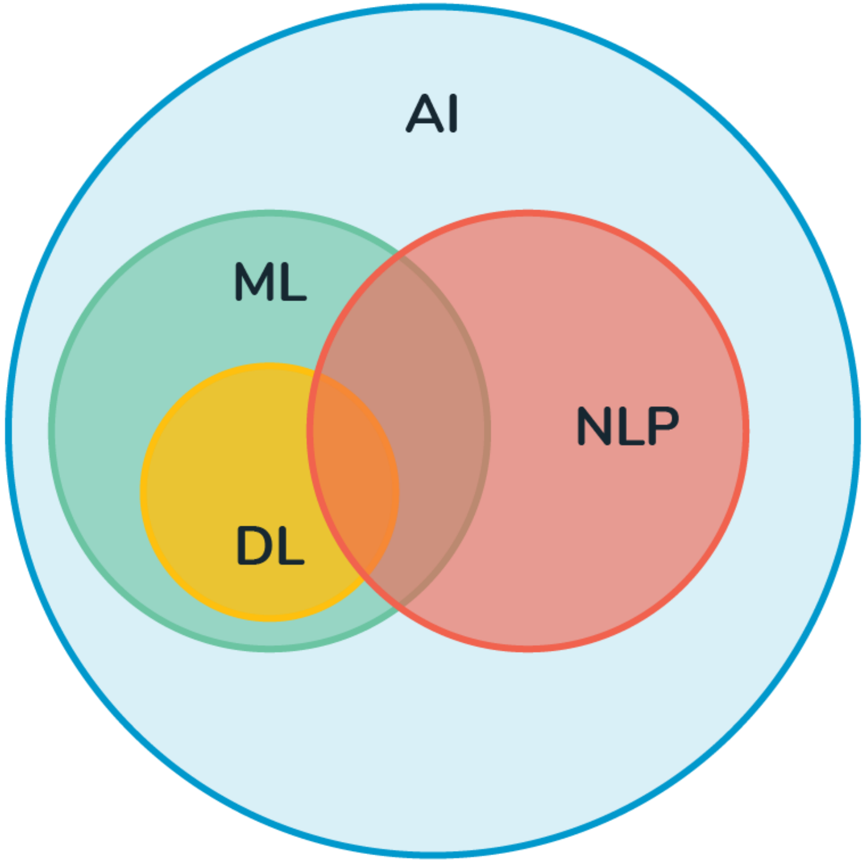
- ❑ What is NLP?
- ❑ Does ChatGPT solve every NLP problem?
- ❑ Language consists of many levels of structure
- ❑ What makes language so difficult to process?
- ❑ How to process language?
- ❑ Recent Developments (2019-2024)
- ❑ Limits of LLMs and the Financial Incentives of GenAI

NLP is interdisciplinary

- ❑ Linguistics
- ❑ Artificial Intelligence
- ❑ Machine Learning (2000-present)

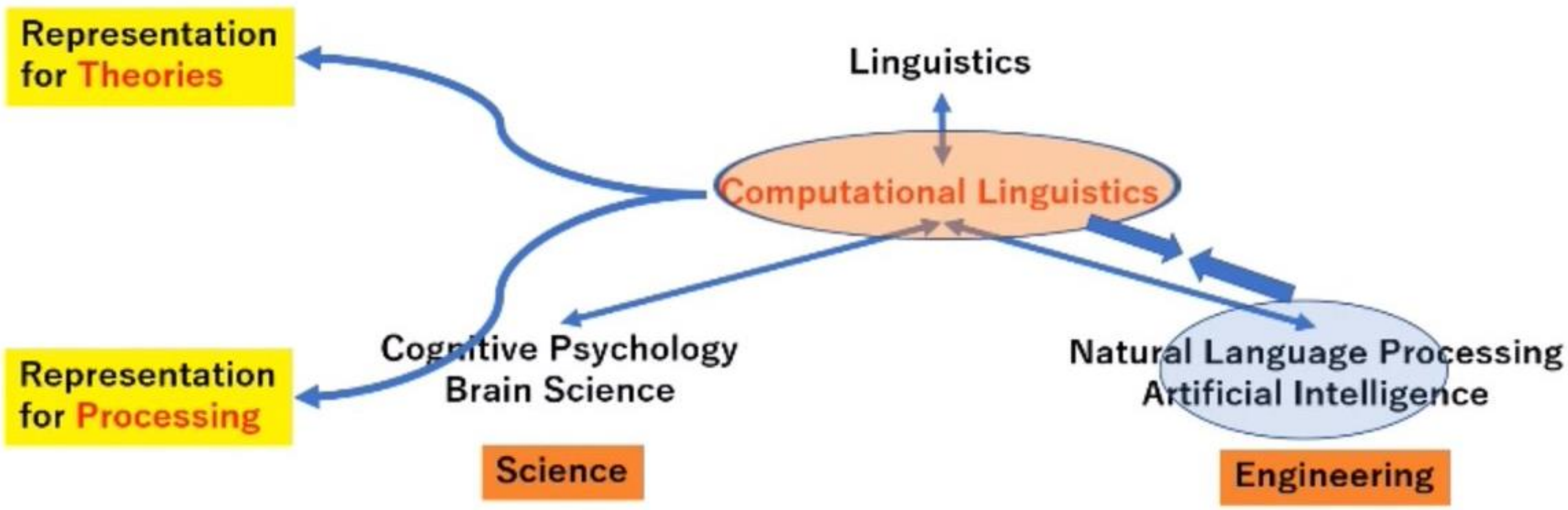
Recently,

- ❑ Social Science and Humanities
- ❑ Human-computer Interaction
- ❑ Education
- ❑ Robotics
- ❑ Cognitive Science / Brain Science / Neuroscience
- ❑ Psychology
- ❑ Law / Medical / Biology
- ❑ ..



NLP vs (Computational) Linguistics

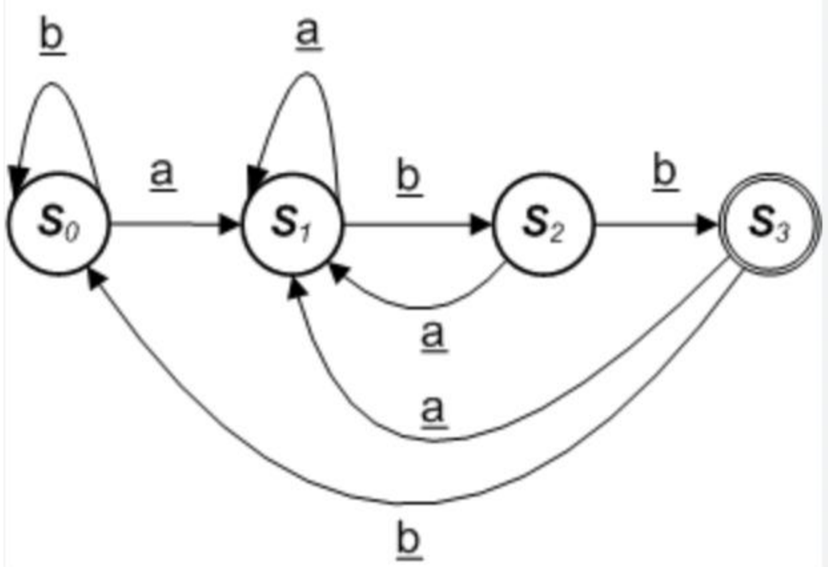
- **Linguistics** involve the nature of *linguistic representations and linguistic knowledge*, and how linguistic knowledge is acquired and deployed in comprehension of language.
- **Computational linguistics** asks *what humans are computing and how*, by *mathematically defining* classes of linguistic representations and *formal grammars* to capture the range of phenomena in human languages.
- **NLP** is the art of *solving engineering problems* that need to analyze (or generate) natural language text. The metric is whether you got good solutions on the engineering problem. After all, their goal is not a full theory but rather the simplest, most efficient approach that will get the job done.



https://twitter.com/radamihalcea/status/1422892875218628616

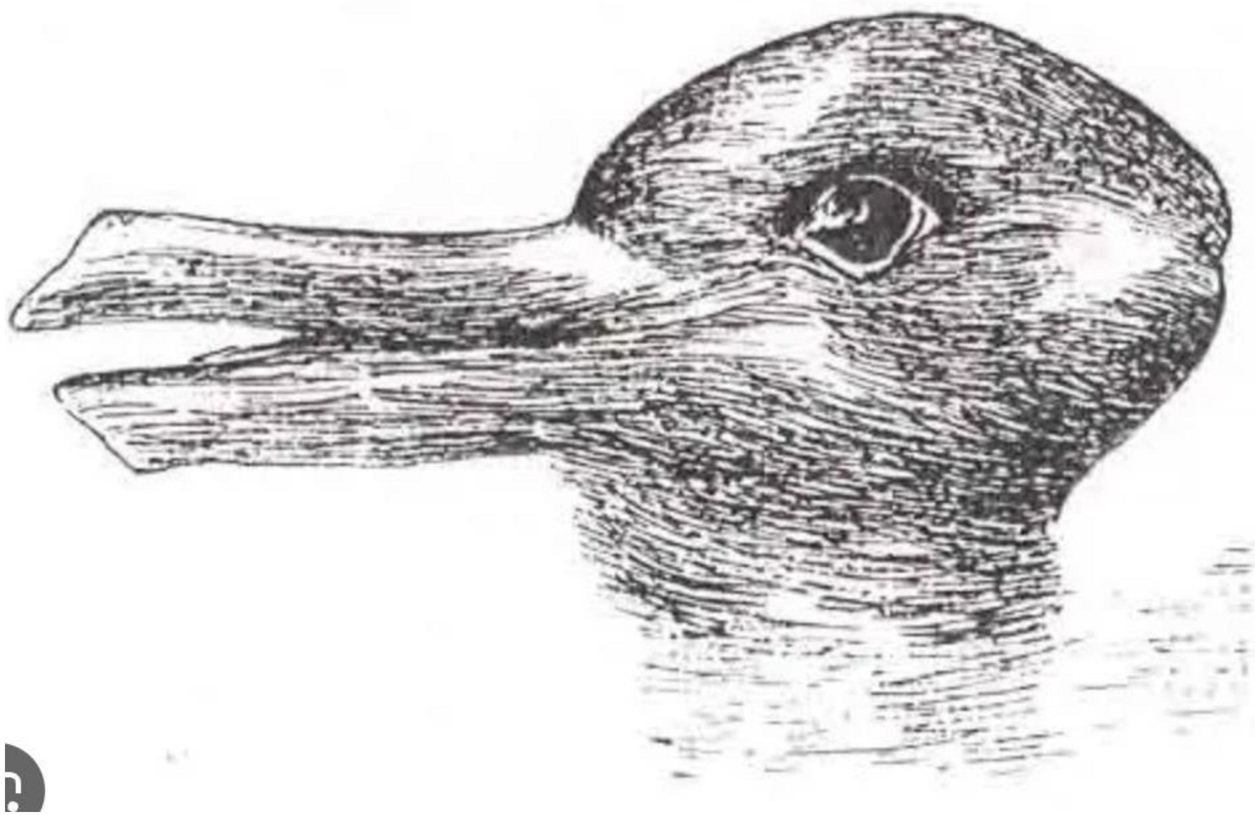
Linguistic Theories

Language as Formal Logic

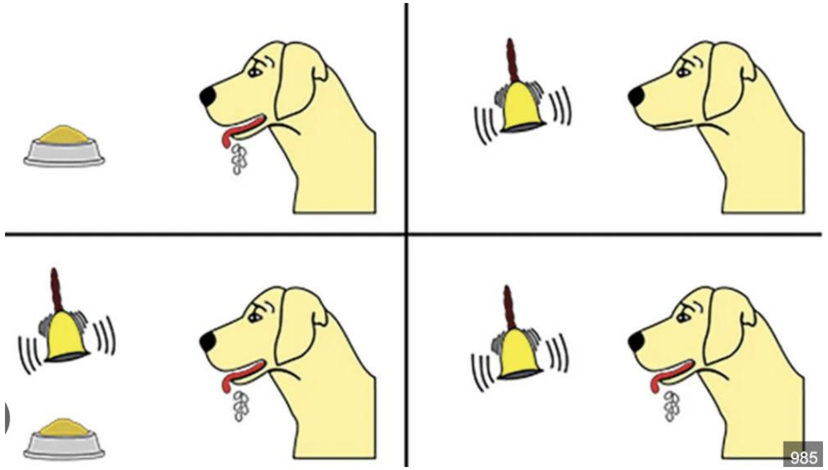


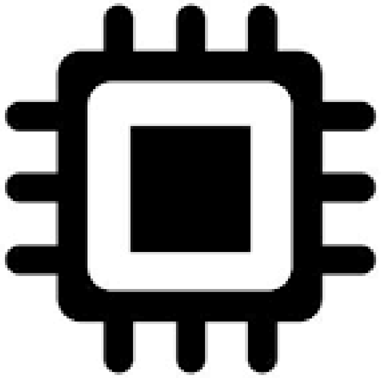
$$
\begin{aligned}
S &\rightarrow aS \mid bX \\
X &\rightarrow aX \mid bY \\
Y &\rightarrow aY \mid bZ \mid \Lambda \\
Z &\rightarrow aZ \mid \Lambda
\end{aligned}
$$

Language as Social Activity



How biased are our mental models to Language?





Behaviorist – Little Bias within models

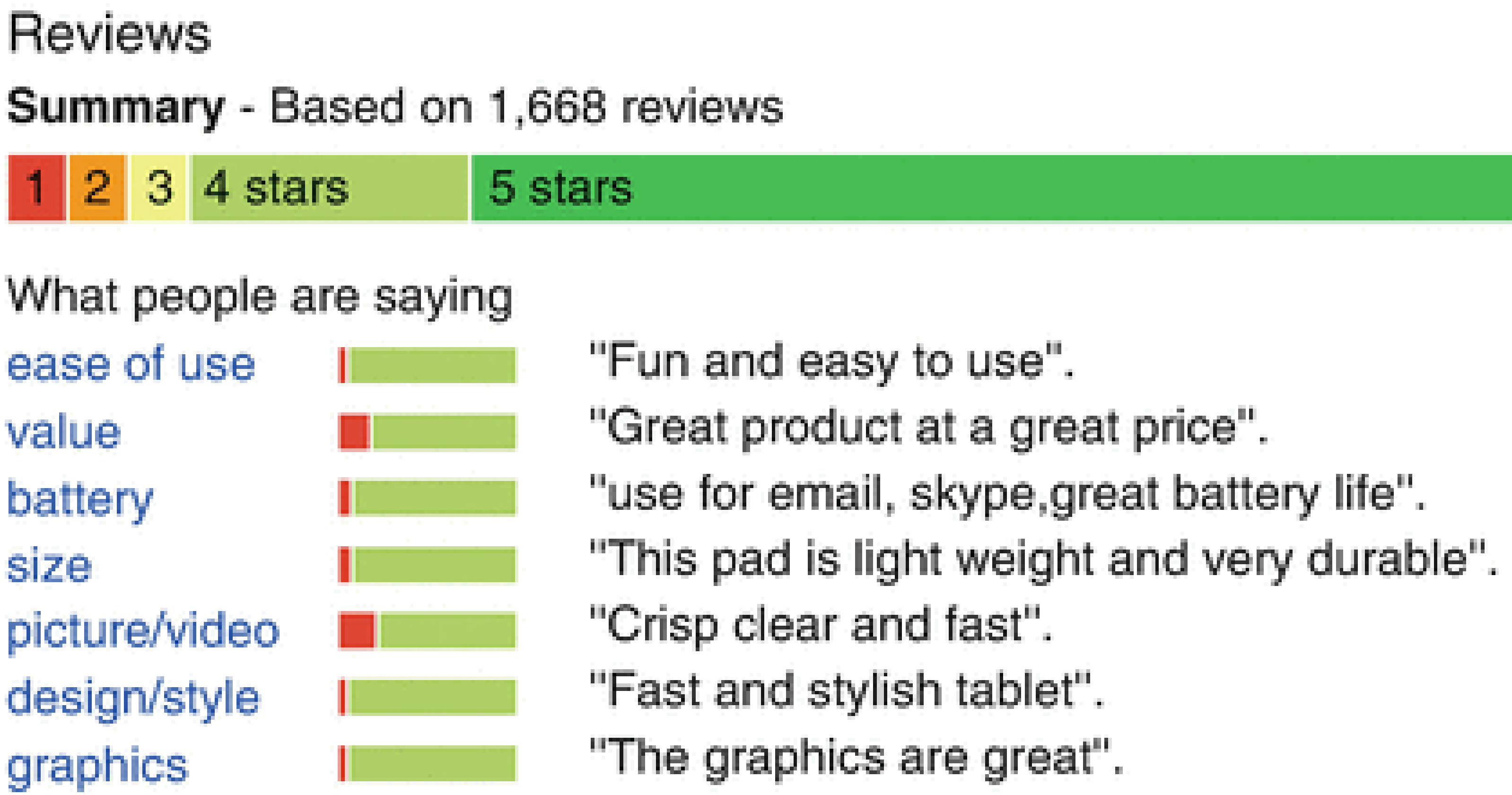
Language is Embedded in our minds – High bias

NLP = Processing language with computers

Processing as sorting and clouding

Word cloud generated with text on our class homepage using www.wordclouds.com

Processing as understanding sentiment



Lei Zhang and Bing Liu, Sentiment Analysis and Opinion Mining

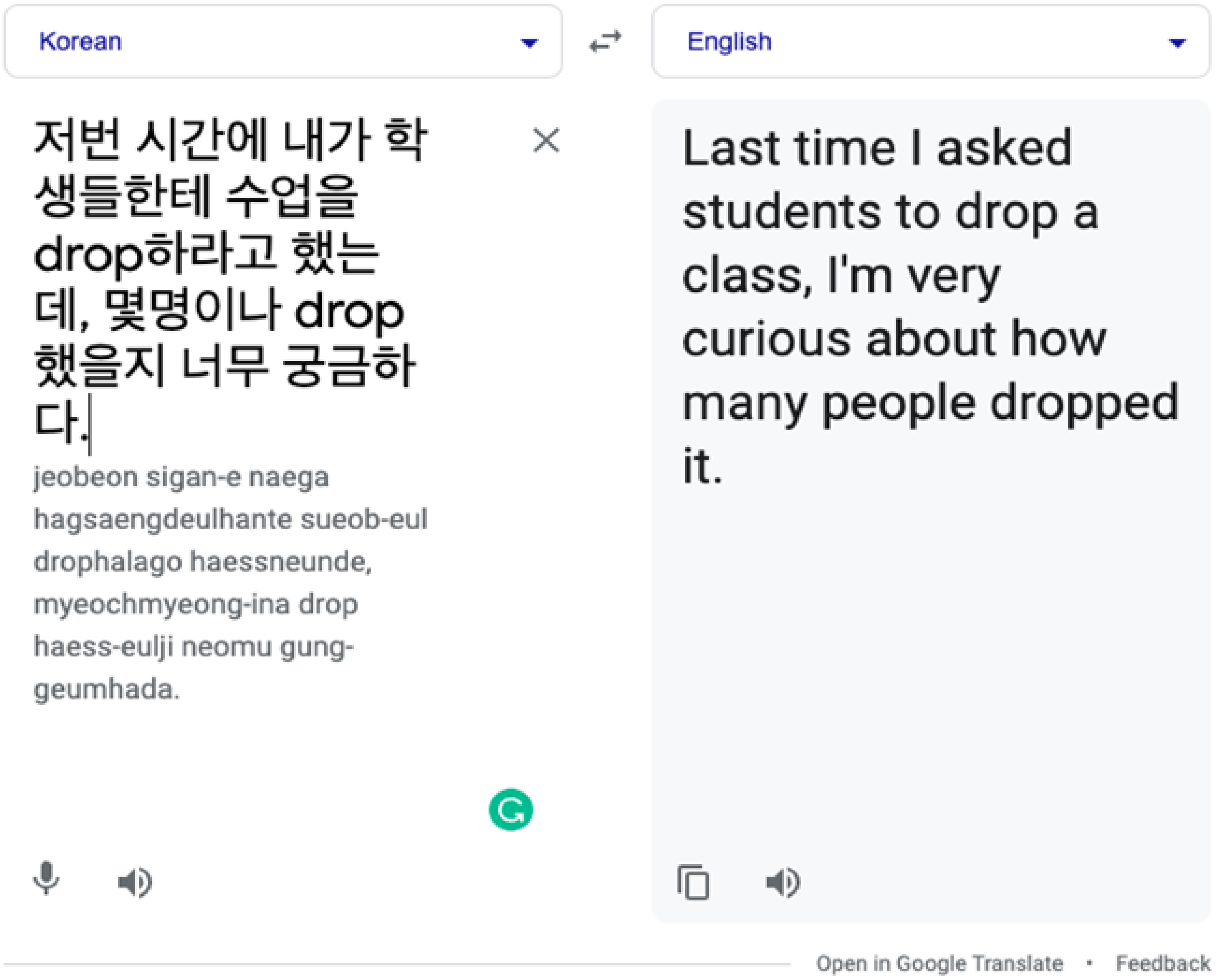
Processing as assistant

Processing as question answering

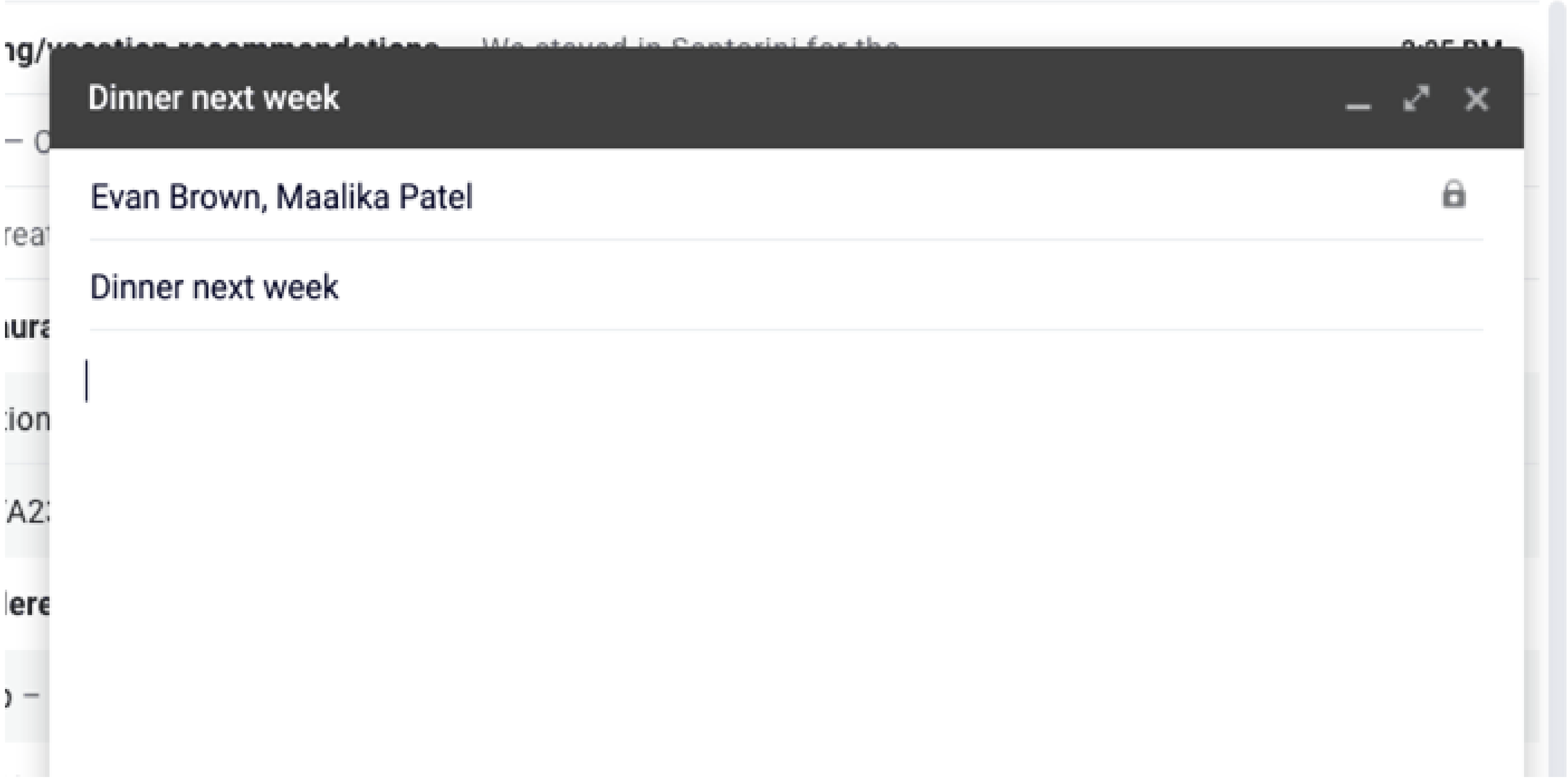
- ❑ What year was Abraham Lincoln born?
- ❑ How many states were in the United States that year?
- ❑ How much Chinese silk was exported to England in the end of the 18th century?

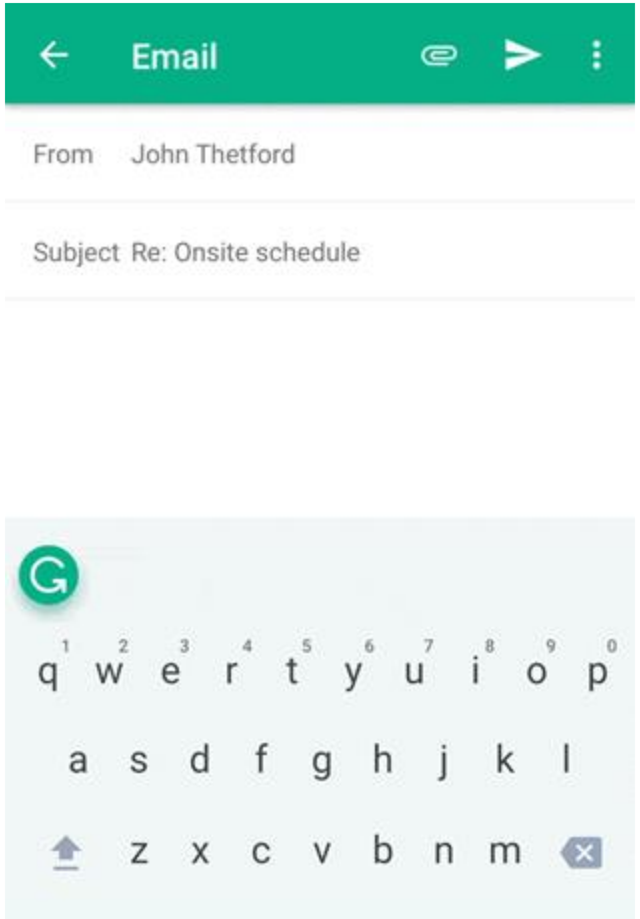
It's alive: IBM's Watson supercomputer defeats humans in final Jeopardy match, 2011

Processing as translation

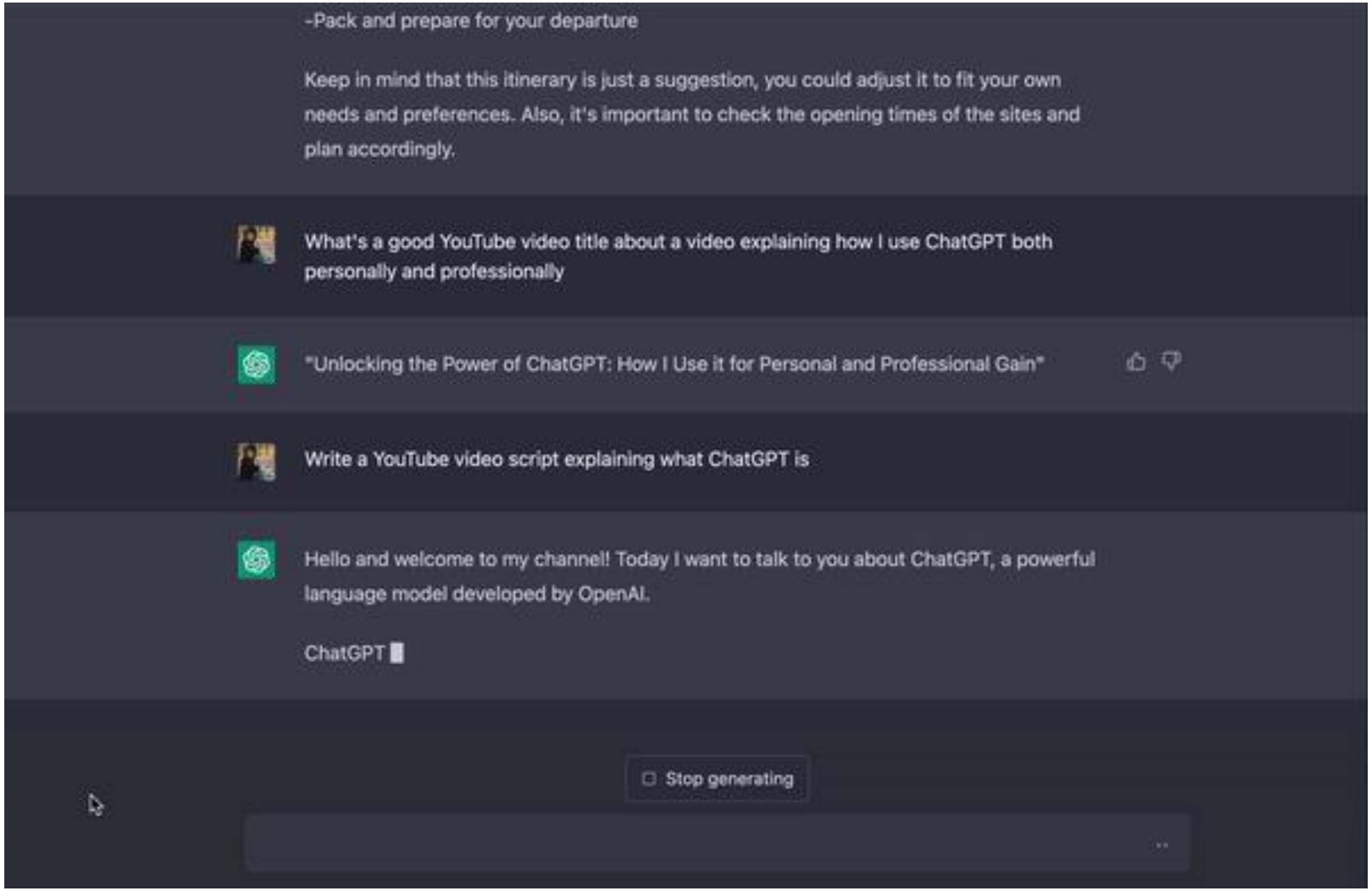


Processing as (email) assistant





Processing as communication



Social cognition / Reasoning

"Two children, Chloe and Alexander, went for a walk. They both saw a dog and a tree. Alexander also saw a cat and pointed it out to Chloe. She went to pet the cat."

"Did Chloe see the cat?"

Cause and effect

"You need flour to bake bread. You have a sack of flour in the garage. When you get there, you find that it got thoroughly soaked in a heavy rain last night.

So you have to ___"

Tracking long narratives

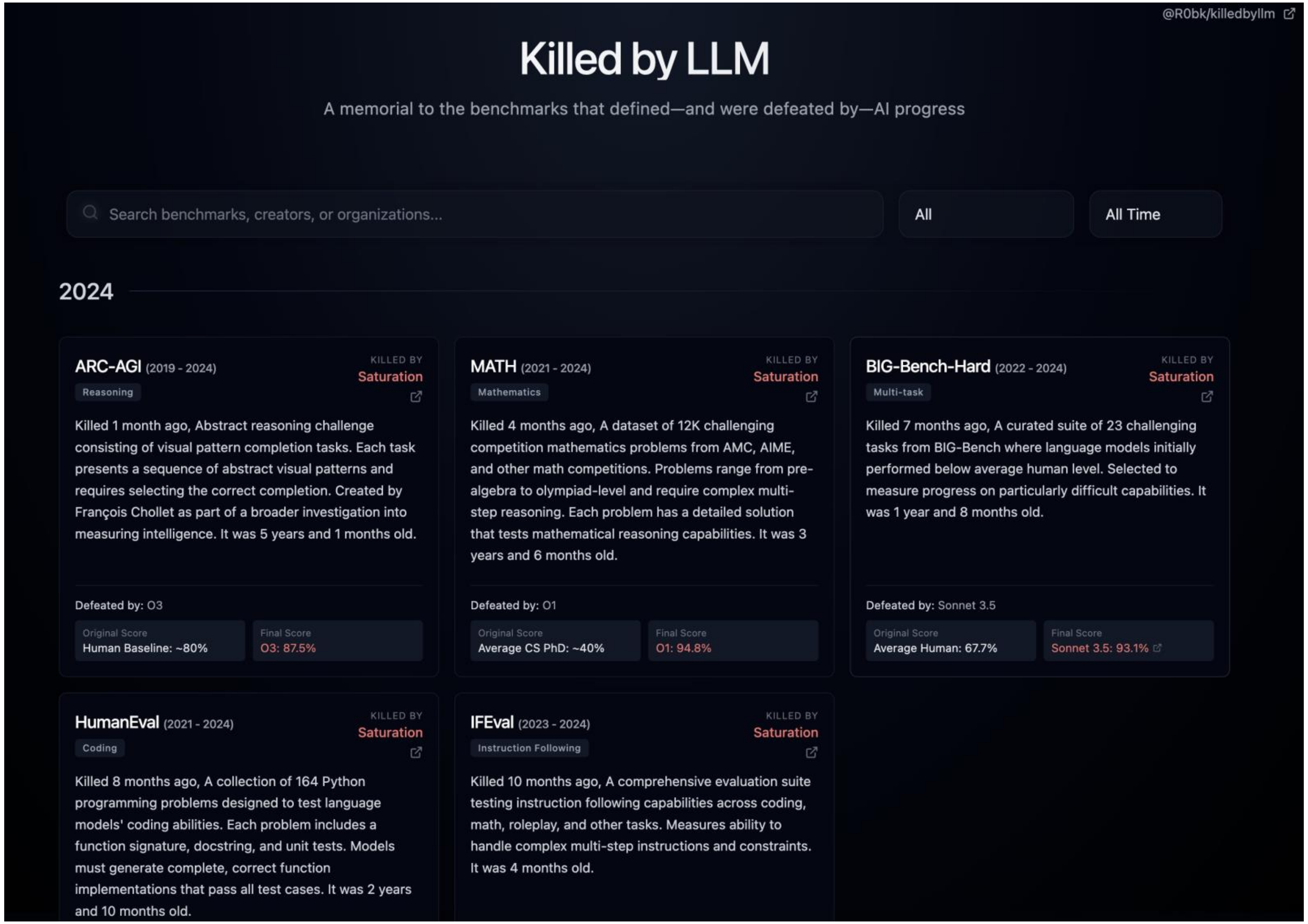
"Never in his life has Bashan caught a hare, nor will he ever; the thing is as good as impossible. Many dogs, they say, are the death of a hare, a single dog cannot achieve it, even one much speedier and more enduring than Bashan. The hare can ``double'' and Bashan cannot --- and that is all there is to it. How Bashan runs! It is beautiful to see a creature expending the utmost of its powers. He runs better than the hare does, he has stronger muscles, the distance between them visibly diminishes before I lose sight of them. And I make haste too, leaving the path and cutting across the park towards the river-bank, reaching the gravelled street in time to see the chase come raging on— the hopeful, thrilling chase, with Bashan on the hare's very heels; — "One more push, Bashan!" I think, and feel like shouting;

".....

Failure cases by GPT3, Gary Marcus & Ernst Davis

Do LLM’s solve every NLP problem?

LLMs Keep Conquering New Benchmarks



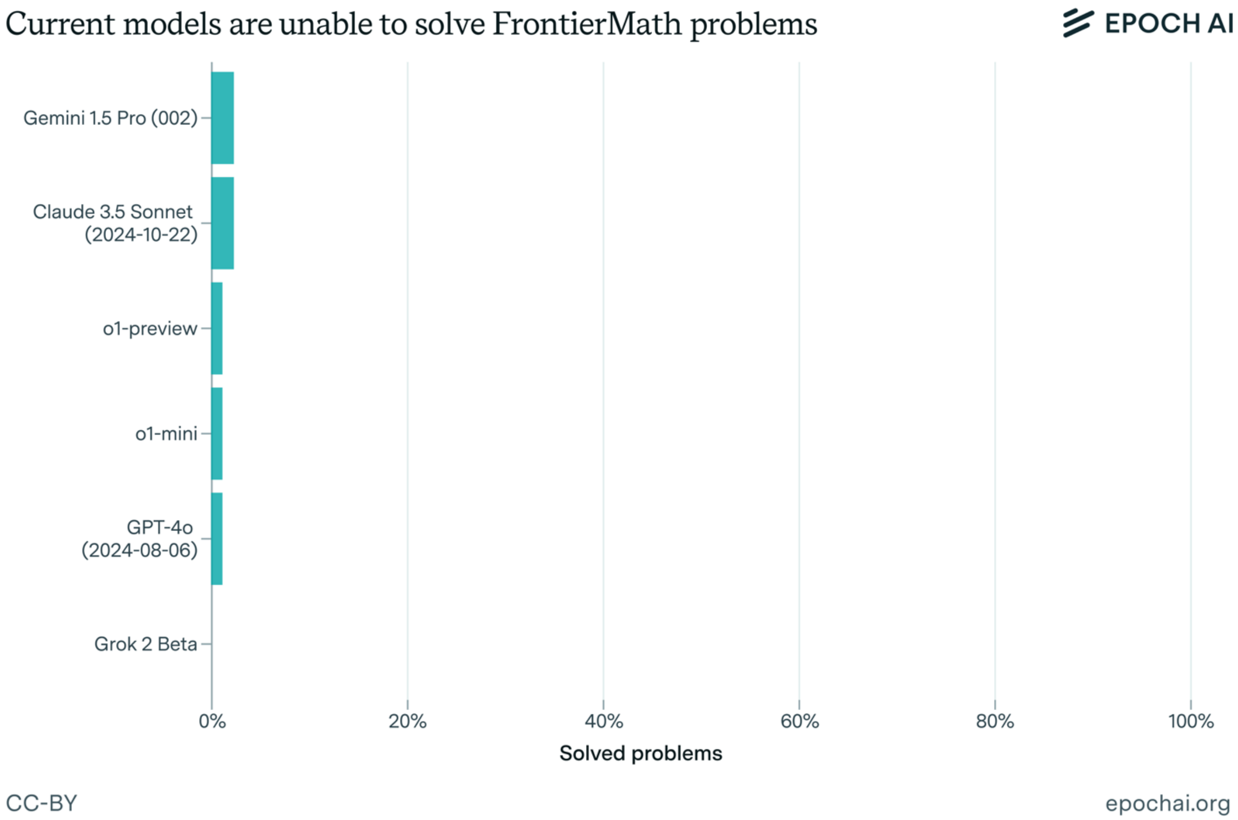
How many r's in strawberry?

There are **2 R's** in "strawberry."

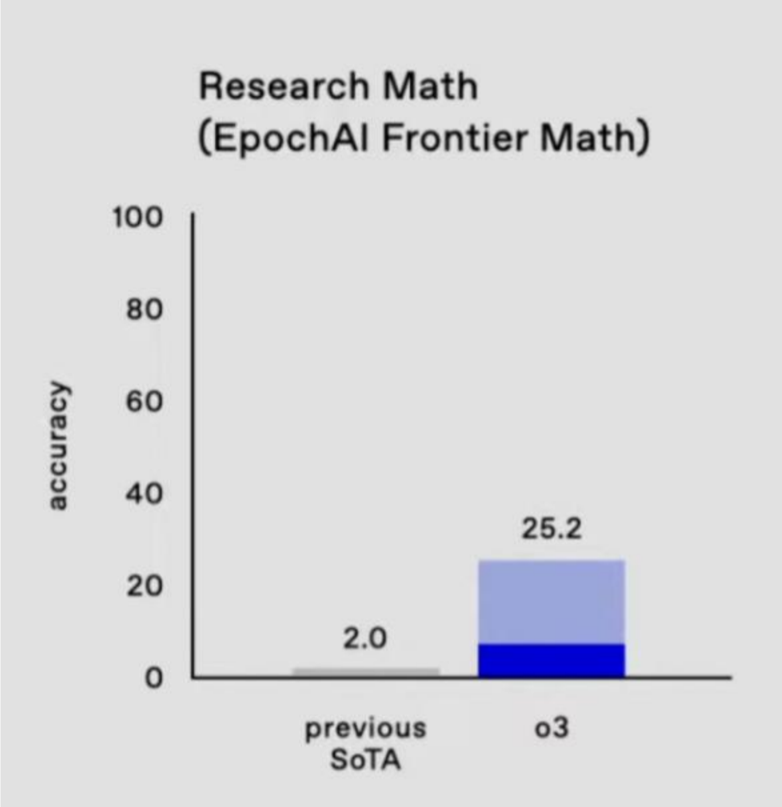

10:07
NEW
Share a clip
I SEE FOUR
COINS NOW.

Table 3: **Performance comparison of various baselines on SWE-bench M.** The table shows results for different software development agent frameworks, including SWE-agent (with multimodal and JavaScript-specific variations) and a retrieval augmented generation (RAG) approach. Each system's success rate (% Resolved) and average cost ($ Avg. Cost) per task are reported.

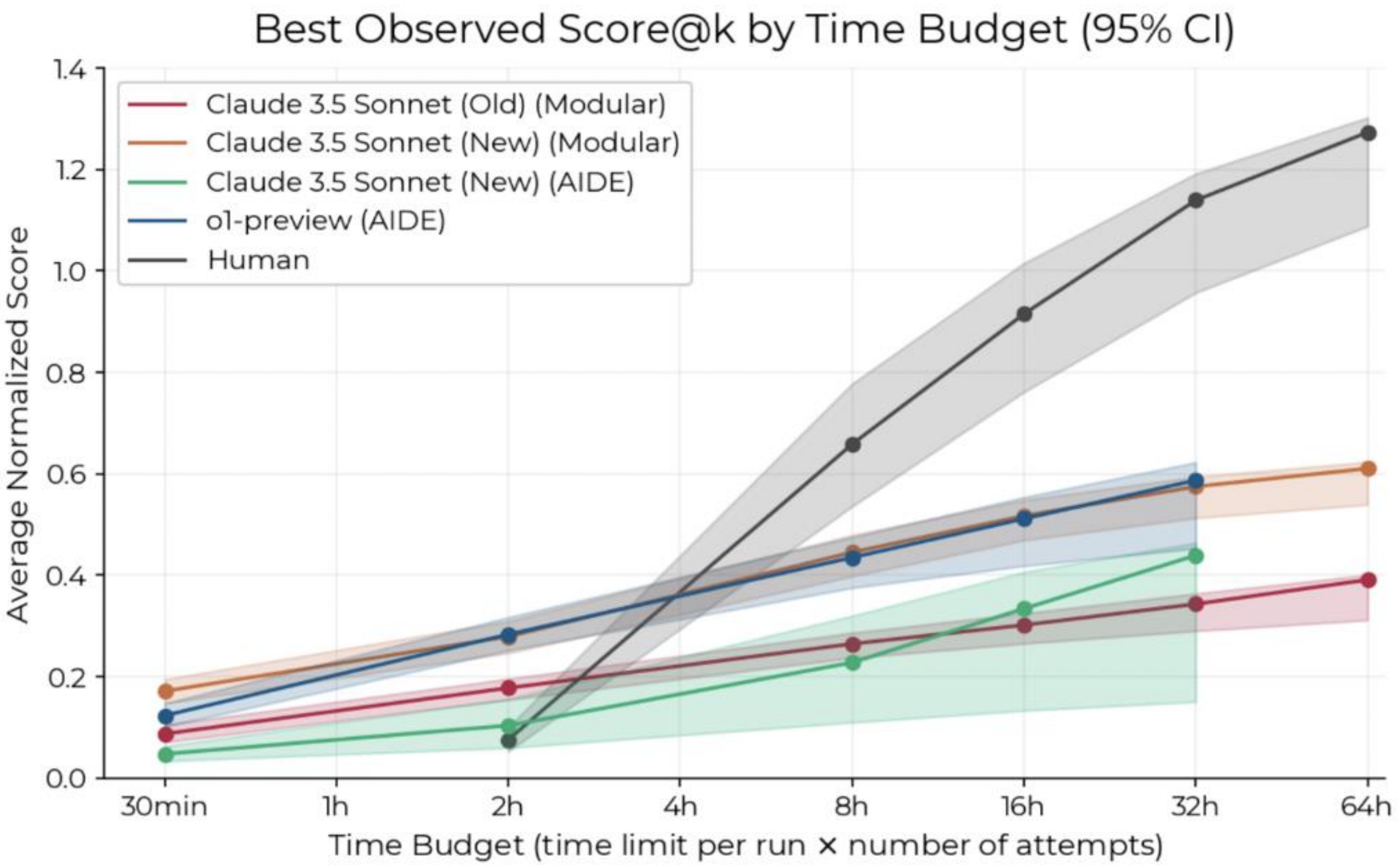
System	Model	% Resolved	$ Avg. Cost
SWE-agent M	GPT-4o	**12.2**	2.94
	Claude 3.5 Sonnet	11.4	3.11
SWE-agent JS	GPT-4o	9.2	0.99
	Claude 3.5 Sonnet	12.0	3.11
SWE-agent Base	GPT-4o	12.0	2.07
	Claude 3.5 Sonnet	**12.2**	1.52
Agentless JS	GPT-4o	3.1	0.38
	Claude 3.5 Sonnet	6.2	0.42
RAG	GPT-4o	6.0	0.17
	Claude 3.5 Sonnet	5.0	0.15

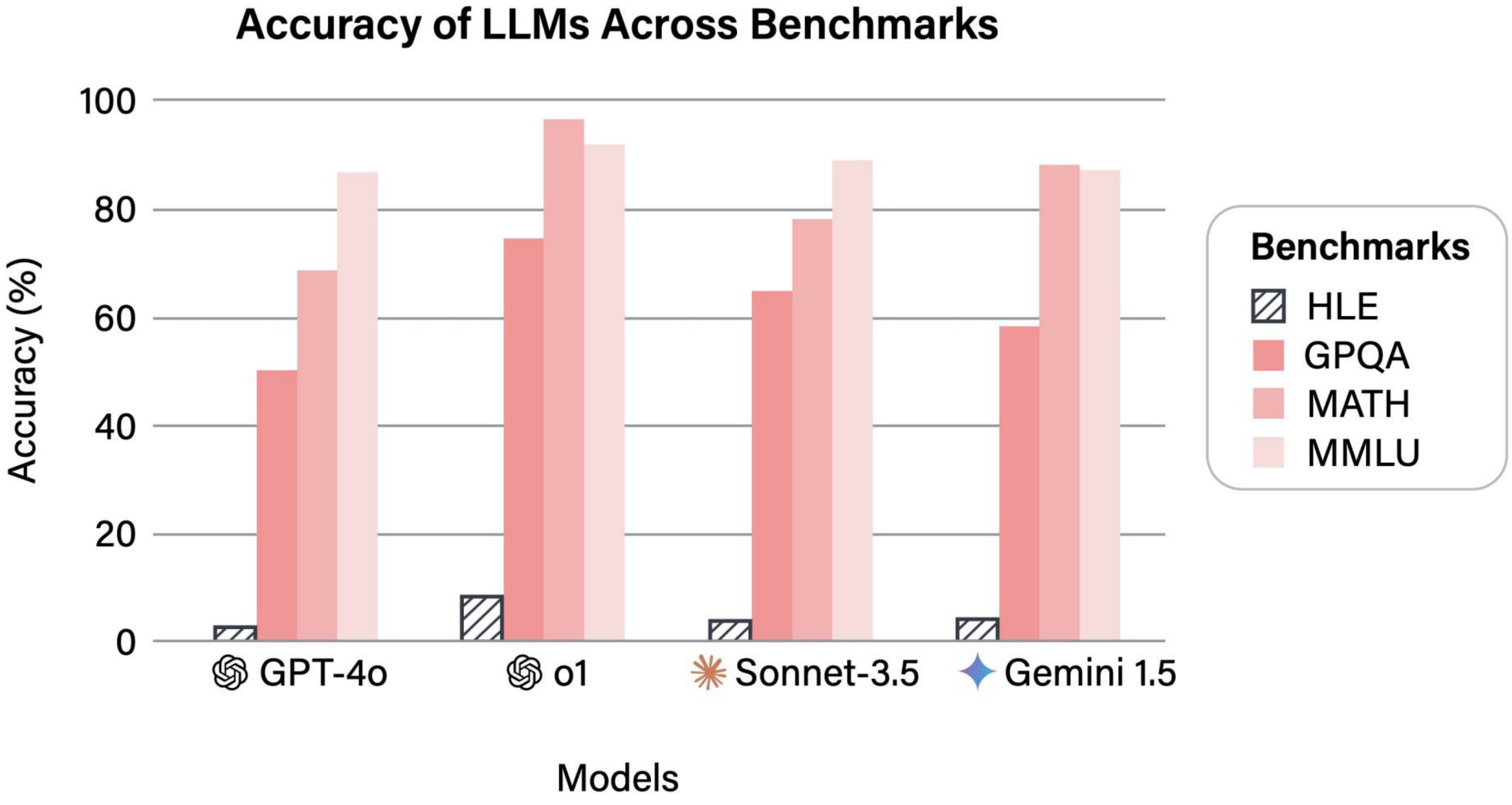


Current models are unable to solve FrontierMath problems
EPOCH AI
Gemini 1.5 Pro (002)
Claude 3.5 Sonnet (2024-10-22)
o1-preview
o1-mini
GPT-4o (2024-08-06)
Grok 2 Beta
0%
20%
40%
60%
80%
100%
Solved problems
CC-BY
epochai.org



Research Math
(EpochAI Frontier Math)
accuracy
100
80
60
40
20
0
2.0
25.2
previous SoTA
o3

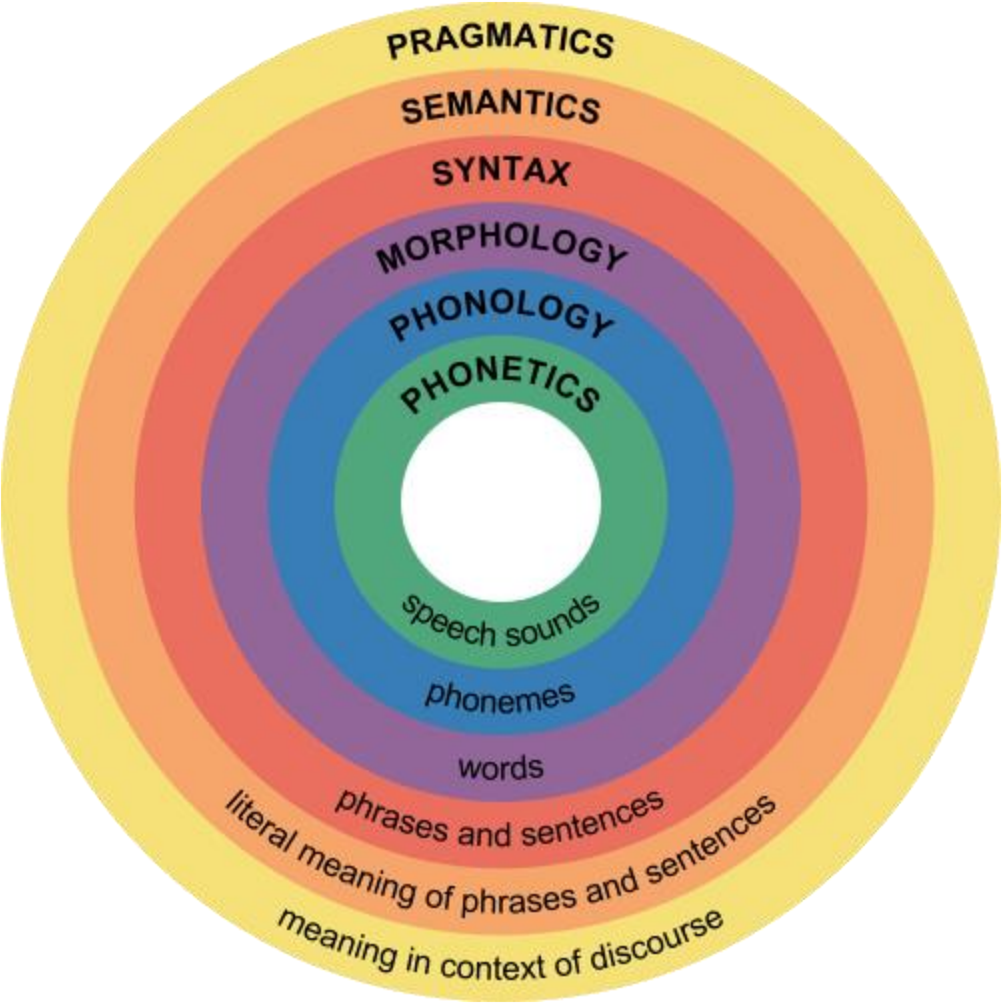

Best Observed Score@k by Time Budget (95% CI)
Average Normalized Score
Time Budget (time limit per run × number of attempts)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
30min
1h
2h
4h
8h
16h
32h
64h
Claude 3.5 Sonnet (Old) (Modular)
Claude 3.5 Sonnet (New) (Modular)
Claude 3.5 Sonnet (New) (AIDE)
o1-preview (AIDE)
Human



https://agi.safe.ai/

What makes language so difficult to process?

Language consists of many levels of structure



Humans fluently integrate all of these in generating and understanding language

This is a simple sentence

Phonology

- Pronunciation modeling

SOUNDS Th i a si e n

Example by Nathan Schneider

Words

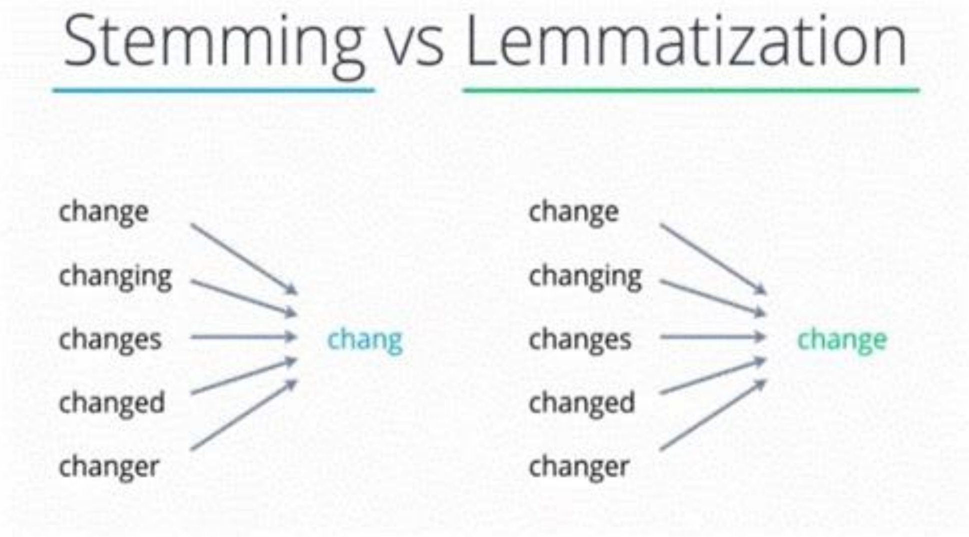
- Tokenization
- Language modeling
- Spelling correction

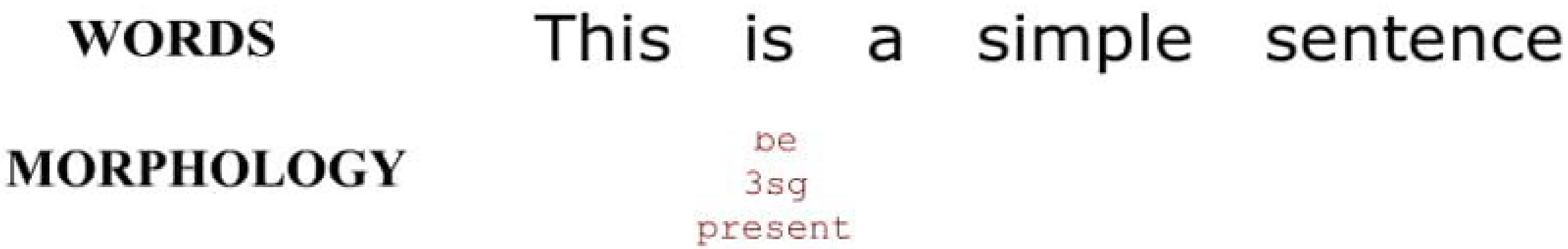
WORDS This is a simple sentence

Example by Nathan Schneider

Morphology

- ❑ Morphological analysis
- ❑ Tokenization
- ❑ Stemming / Lemmatization





Read more about stemming and lemmatization
https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Example by Nathan Schneider

Parts of Speech (POS)

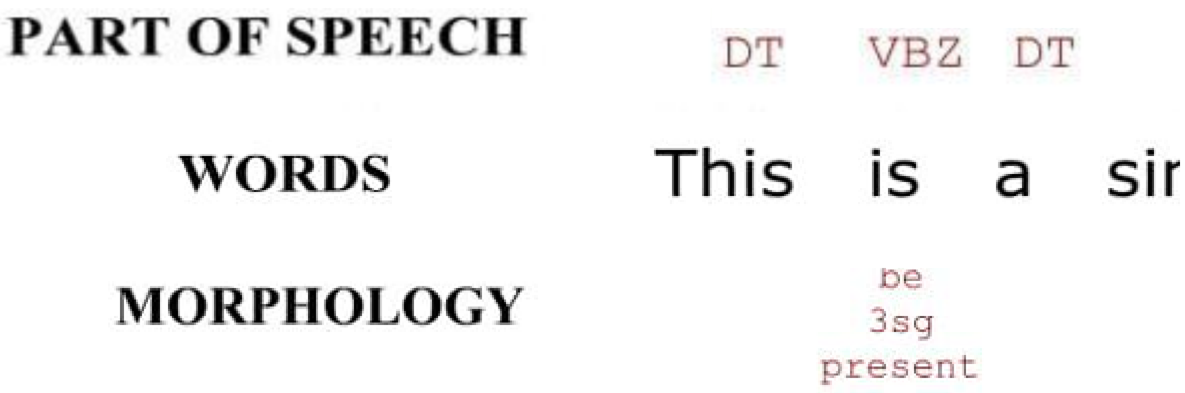
- Part-of-speech tagging

PART OF SPEECH	DT	VBZ	DT	JJ	NN
WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

Example by Nathan Schneider

Parts of Speech (POS)

- Part-of-speech tagging

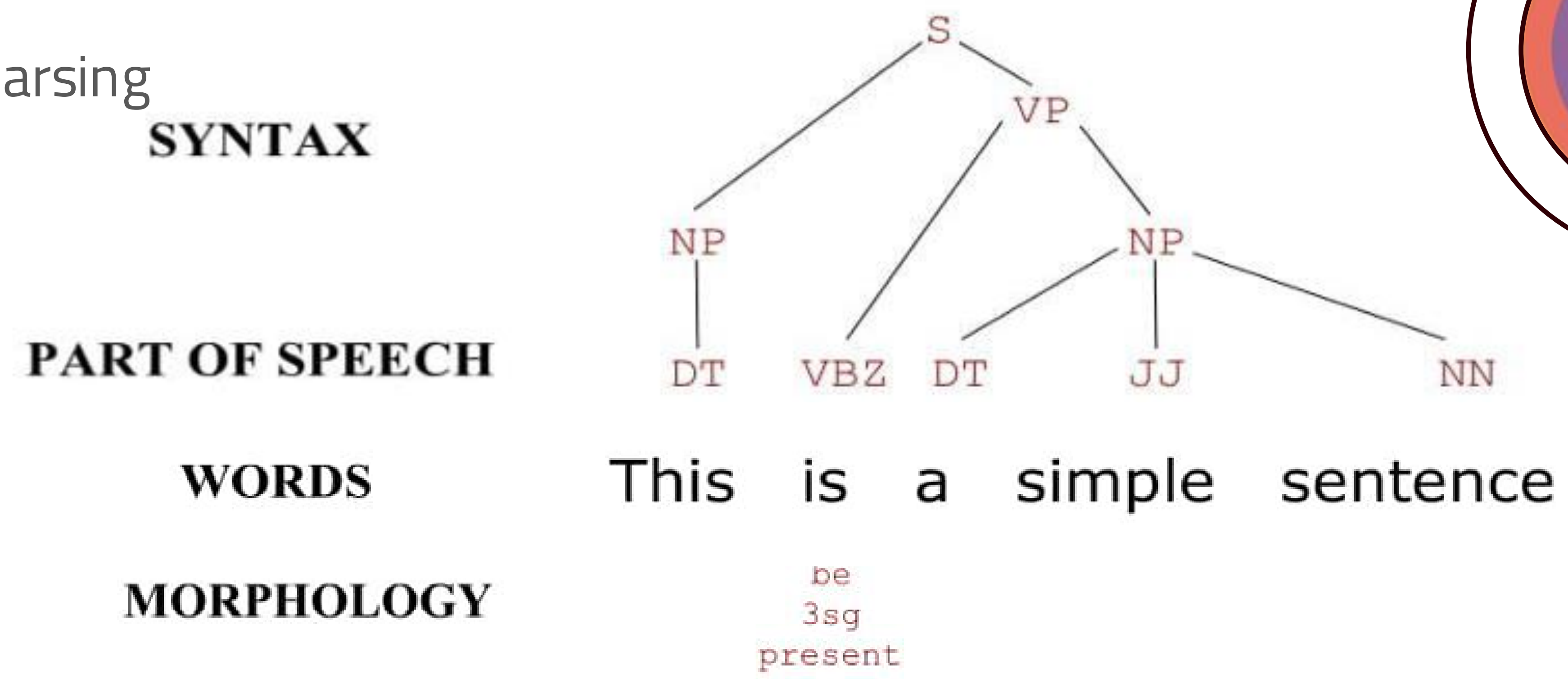


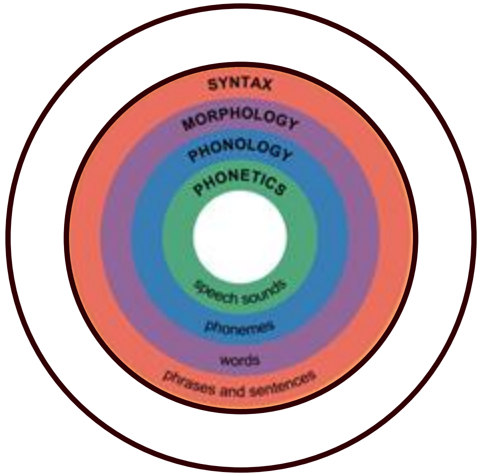
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Example by Nathan Schneider

Syntax

- Syntax parsing

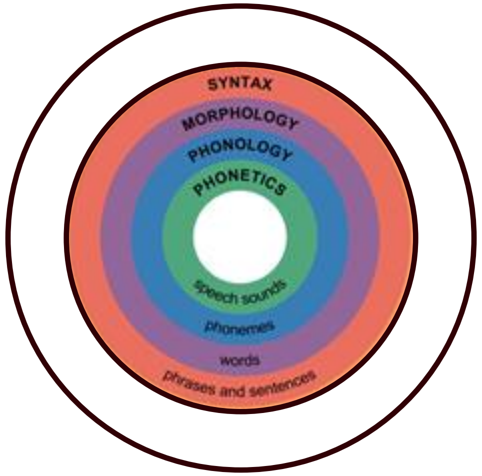


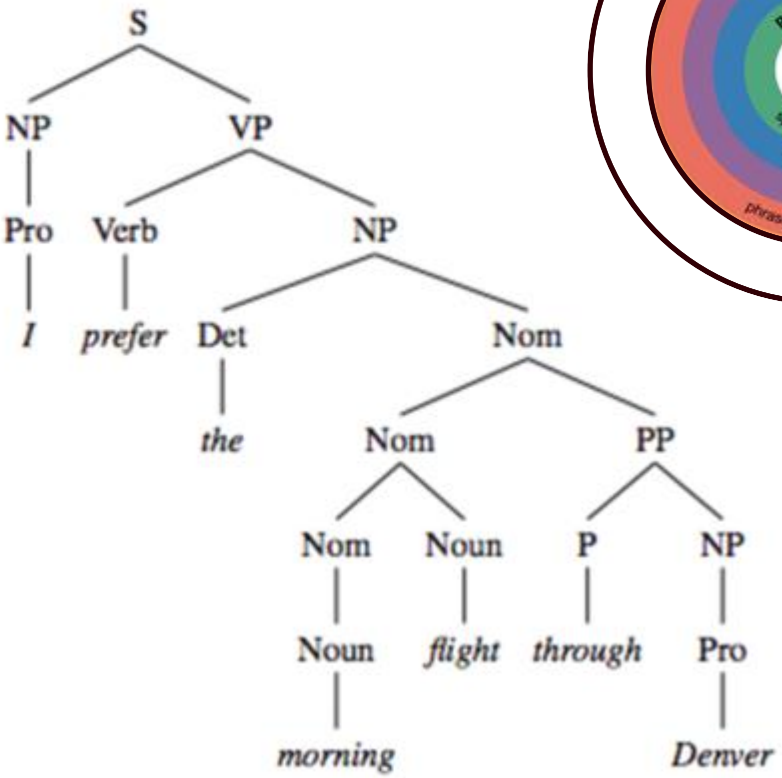


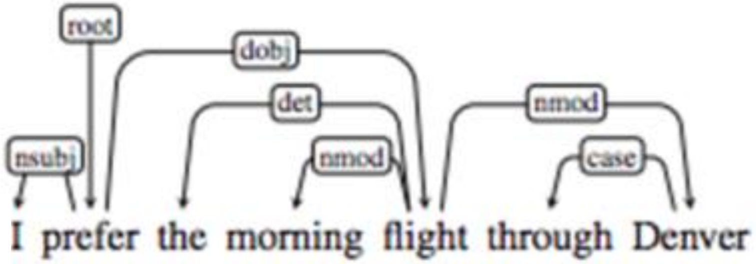
Example by Nathan Schneider

Syntax

- Syntax parsing
  - *Constituency Parsing*: break a sentence into sub-phrases
  - *Dependency Parsing:* explore the dependencies between the words in a sentence





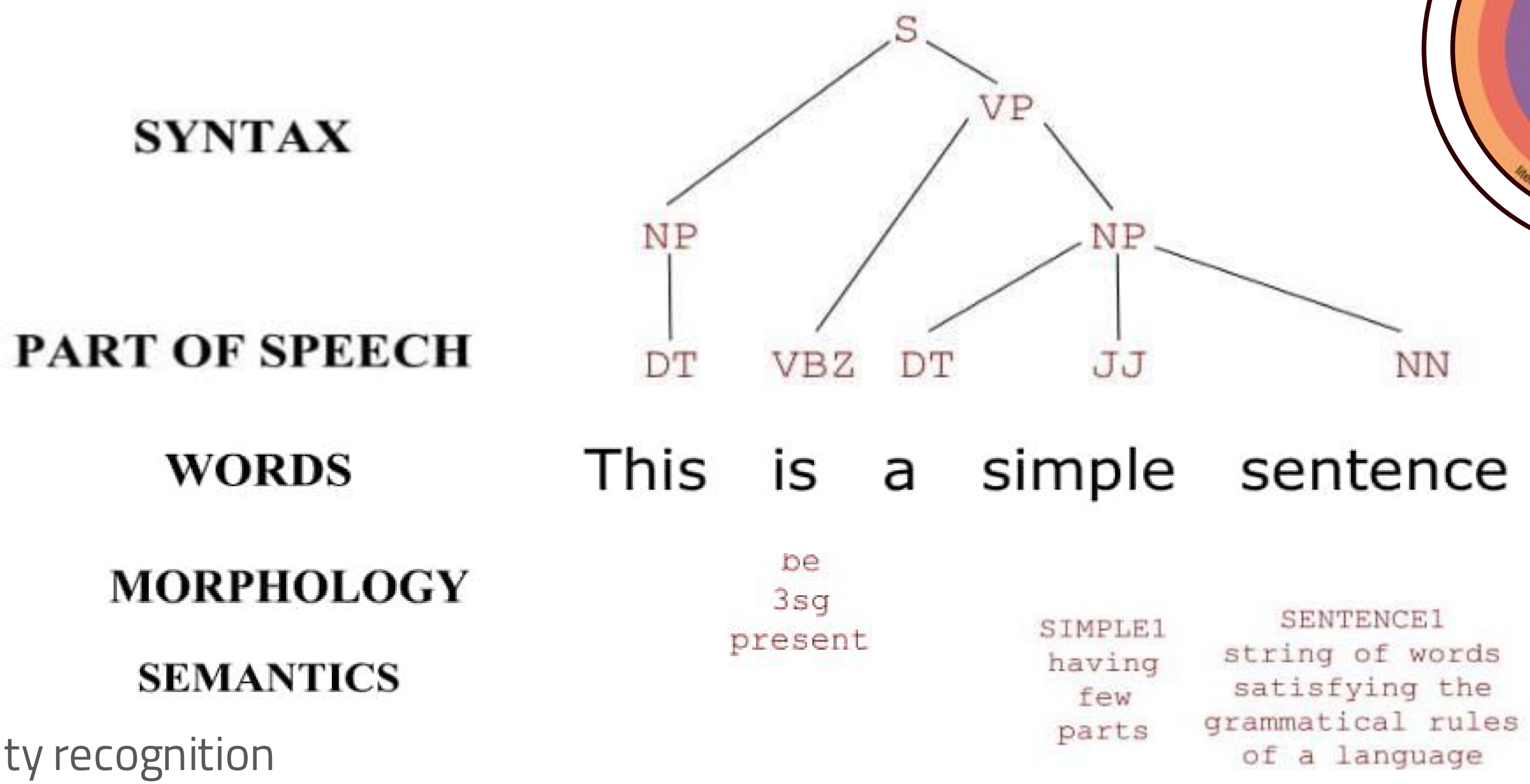


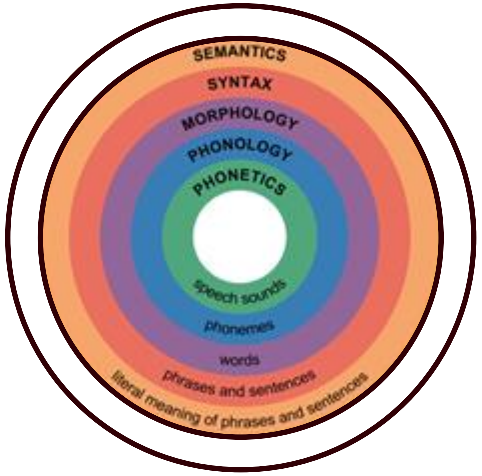
Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Figure 15.2 Selected dependency relations from the Universal Dependency set. (de Marneffe et al., 2014)

Example by Nathan Schneider

Semantics



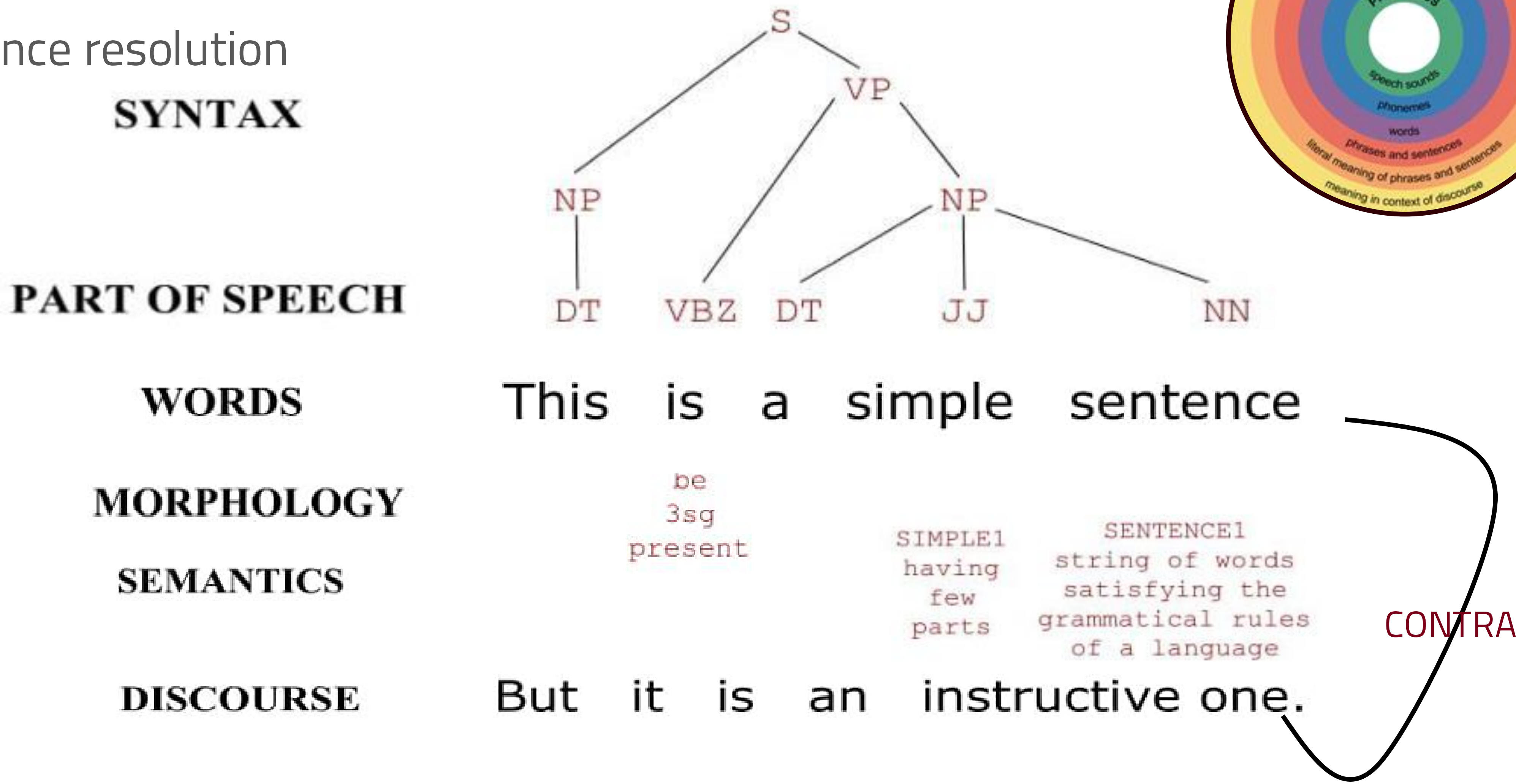


- ❑ Named entity recognition
- ❑ Word sense disambiguation
- ❑ Semantic role labeling
- ❑ Frame semantics

Example by Nathan Schneider

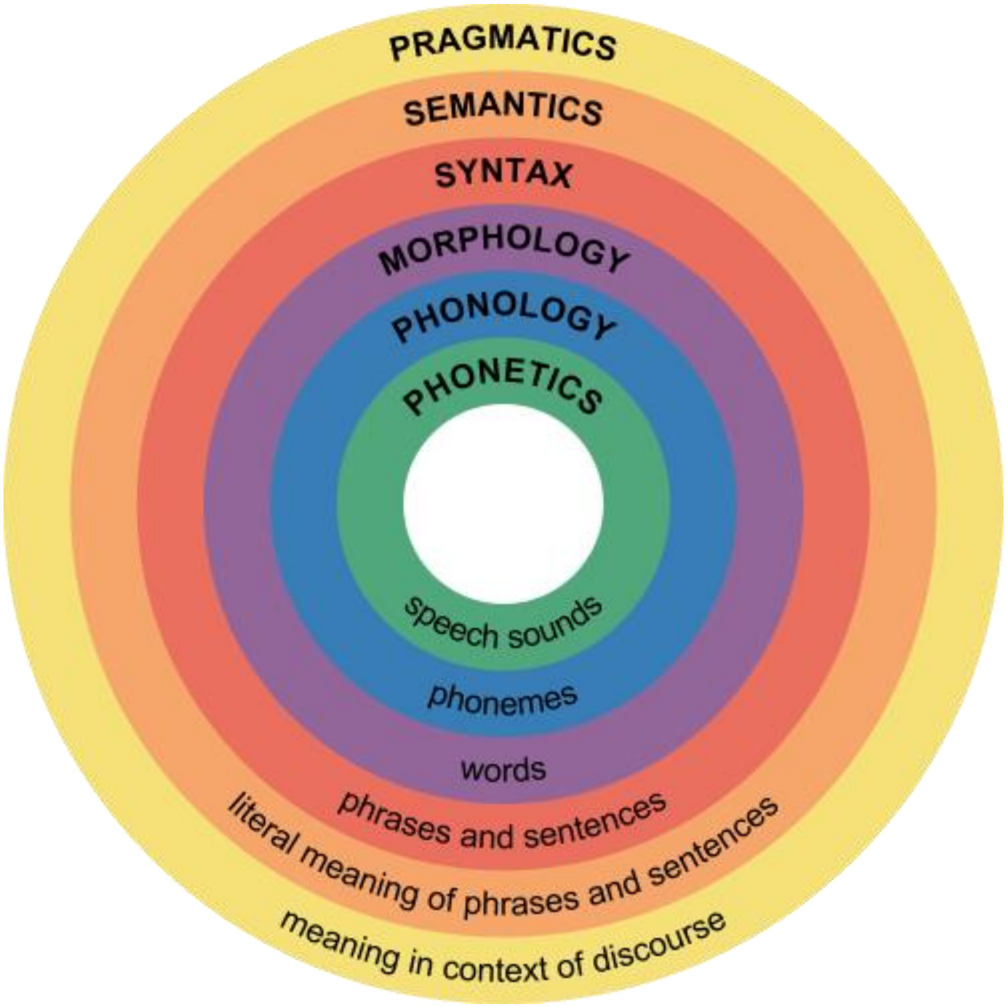
Discourse (Pragmatics)

- Co-reference resolution



Example by Nathan Schneider

Language consists of many levels of structure



Humans fluently integrate all of these in generating and understanding language

What makes language difficult?

- Language is *ambiguous*
- Language needs to be *scaled*
- Language is *sparse*
- Language is *varying*
- Language is *implicit*
- Language is hard to *represent*

Ambiguity at multiple levels

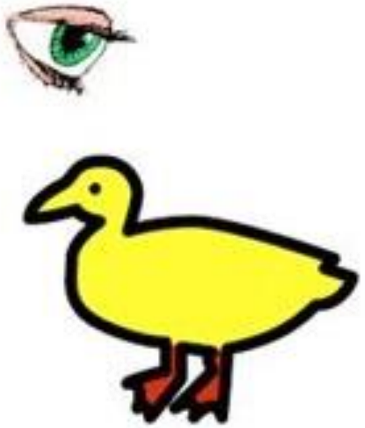
Groucho Marx

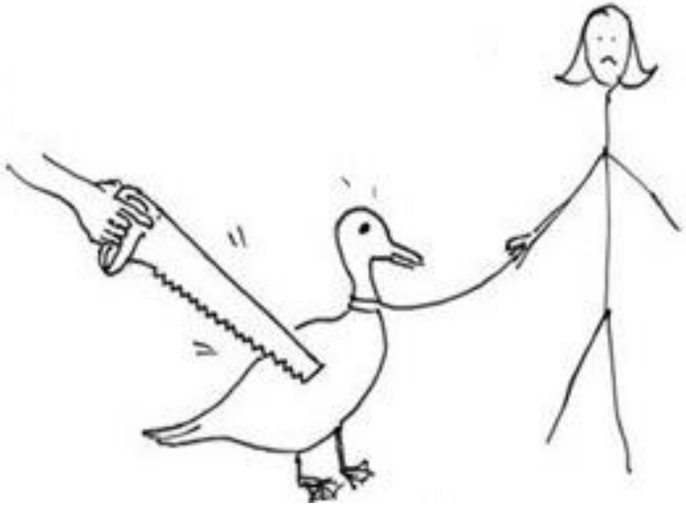
"One morning I shot an elephant **in my pajamas**"

Ambiguity at multiple levels

"One morning I shot an elephant **in my pajamas**"

"I saw her duck with a telescope"





- I used a telescope to observe a small web-footed broad-billed swimming bird belonging to a female person.
- I observed a small web-footed broad-billed swimming bird belonging to a female person. The bird had a telescope.
- I observed a female person move quickly downwards. The person had a telescope.
- I used a telescope to observe a female person move quickly downwards.
- I used a telescope to cut a small web-footed broad-billed swimming bird belonging to a female person.
- I used a telescope to observe heavy cotton fabric of plain weave belonging to a female person.
- I used a telescope to cut heavy cotton fabric of plain weave belonging to a female person.

Slide from Dhruv Batra and figure from Liang Huang

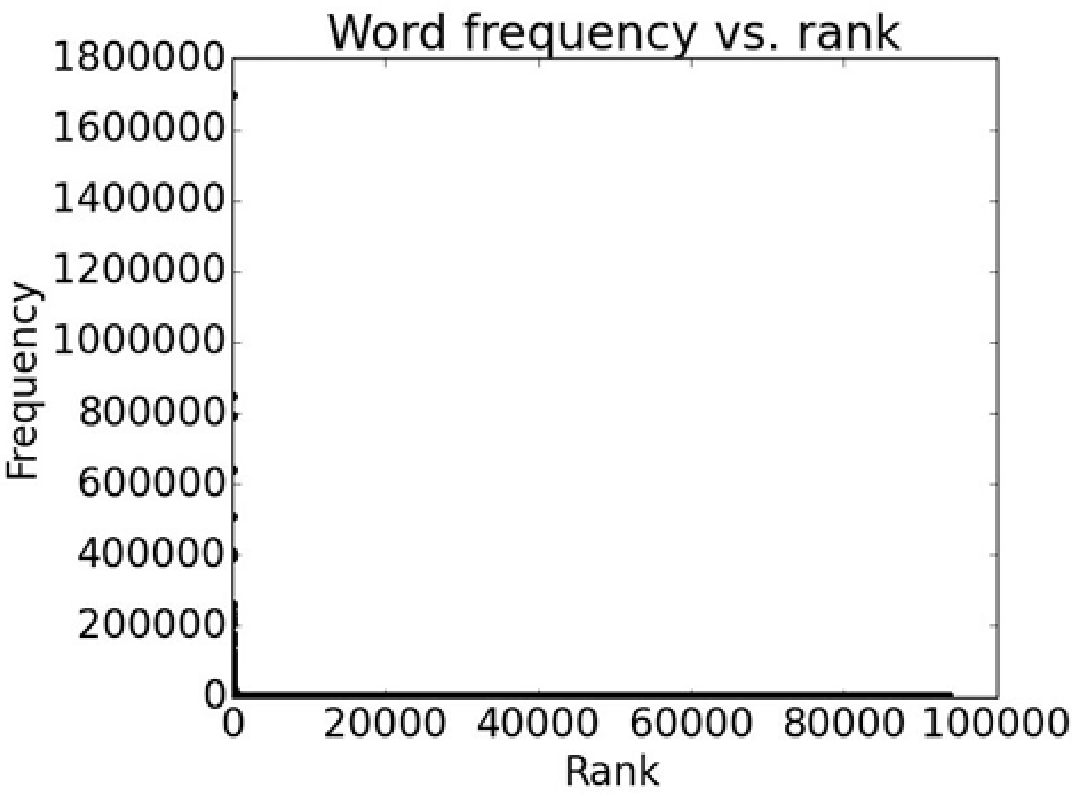
Scale: Applications x Languages

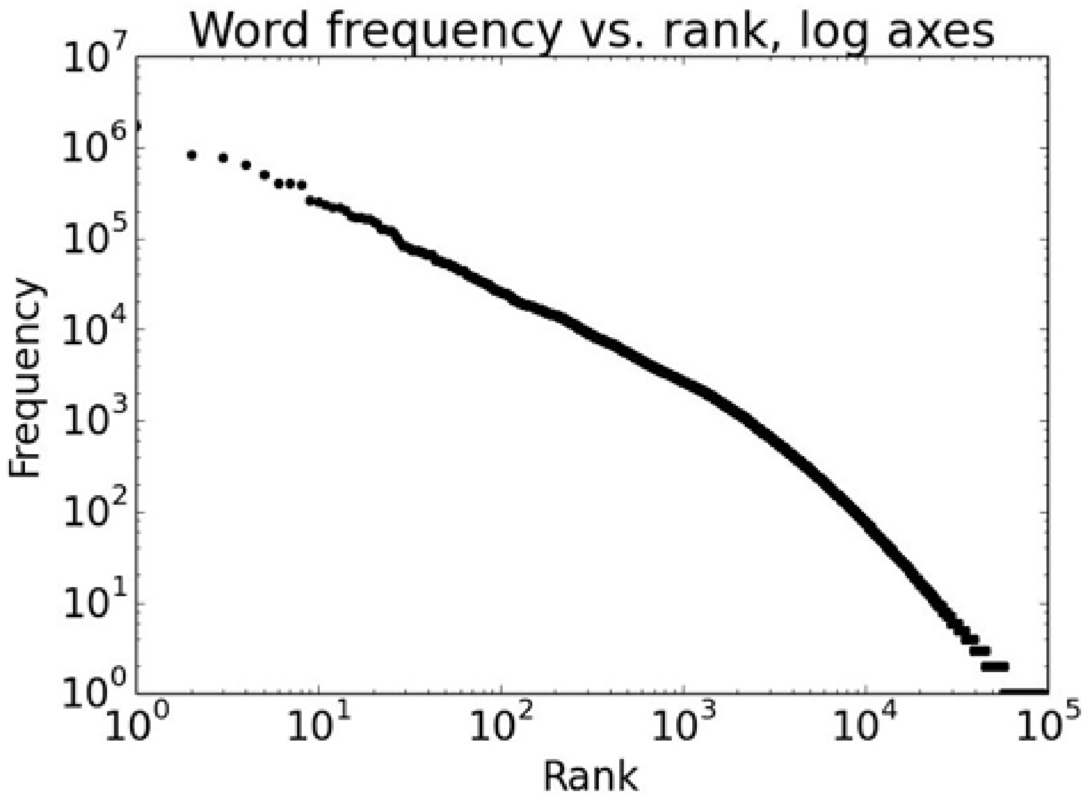
Sparsity

- Sparse data due to Zipf's Law
- Most frequent words in the English Europarl corpus (out of 24M word tokens)
- 36,231 occur only once
  - E.g., pseudo-rapporteur, lobby-ridden, perfunctorily, Lycketoft, UNCITRAL, policyfor, 145.95 ..

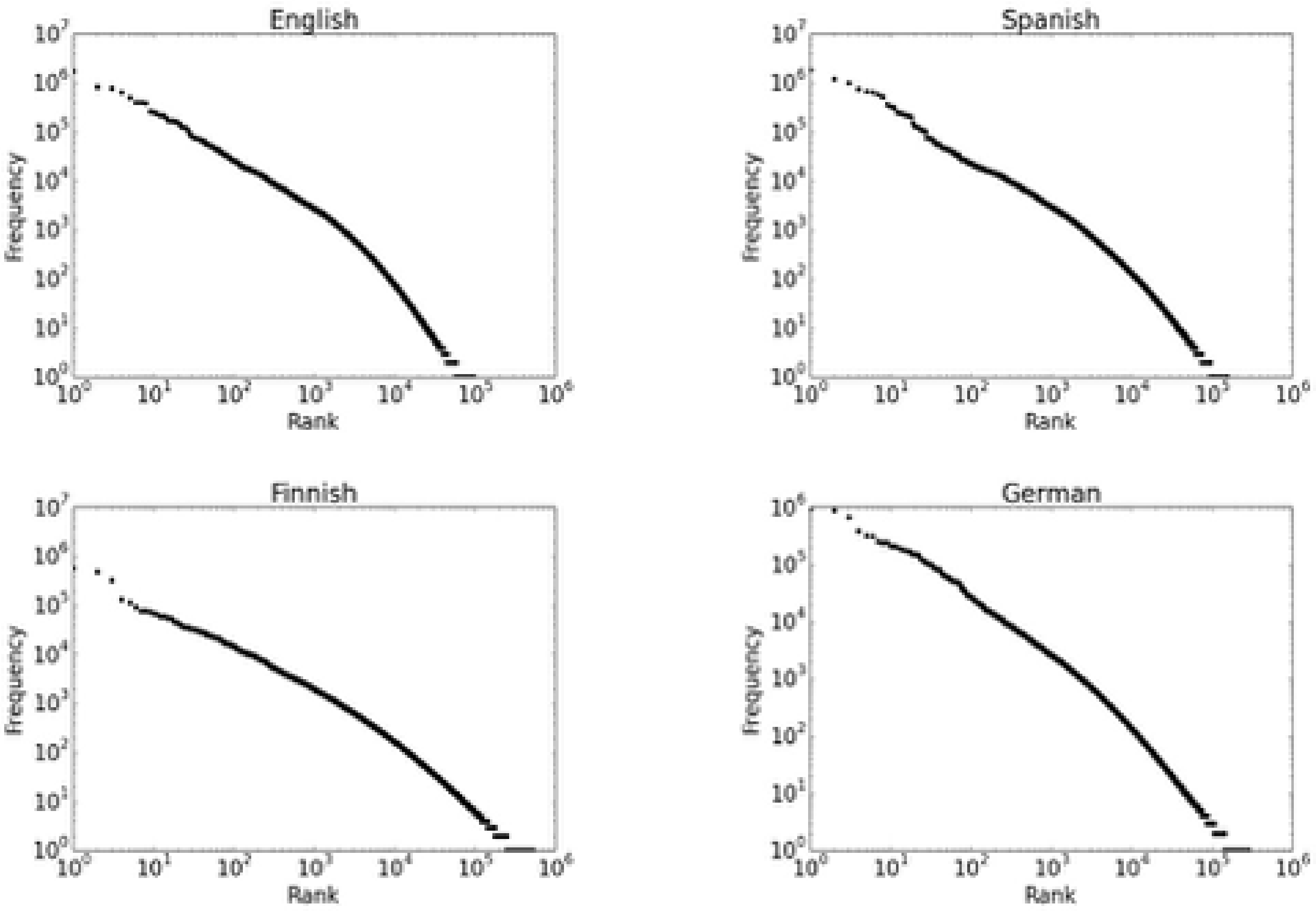
any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

Word Frequency Distribution



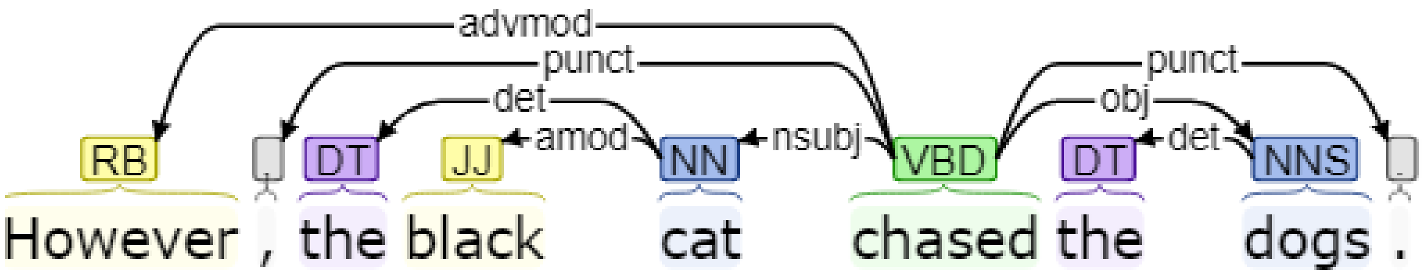


Zipf's Law

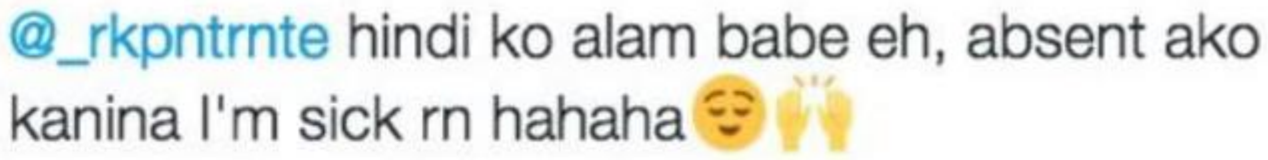


Variation over Domains

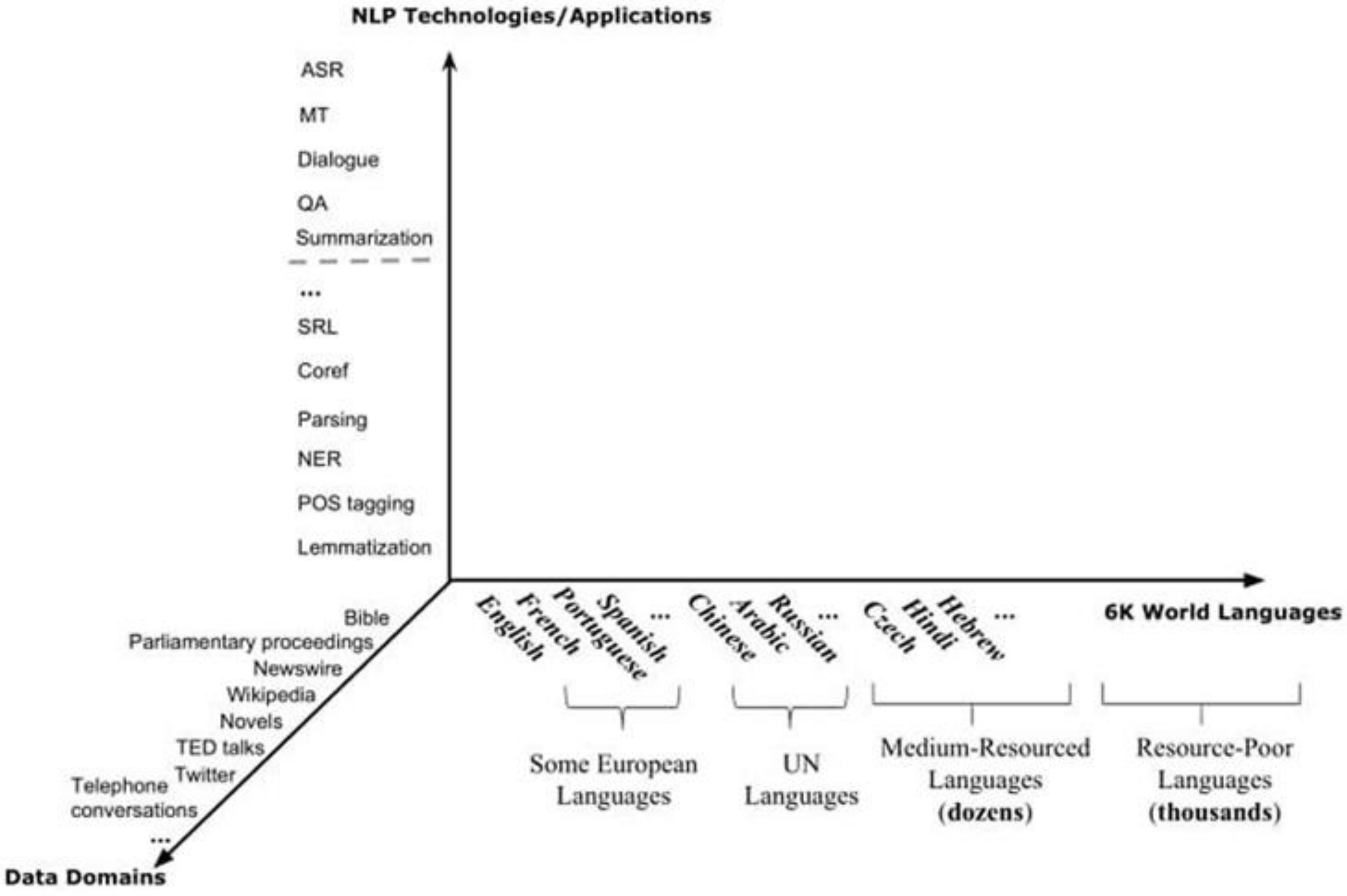
- Suppose you trained a part-of-speech tagger or parser on the Wall Street Journal



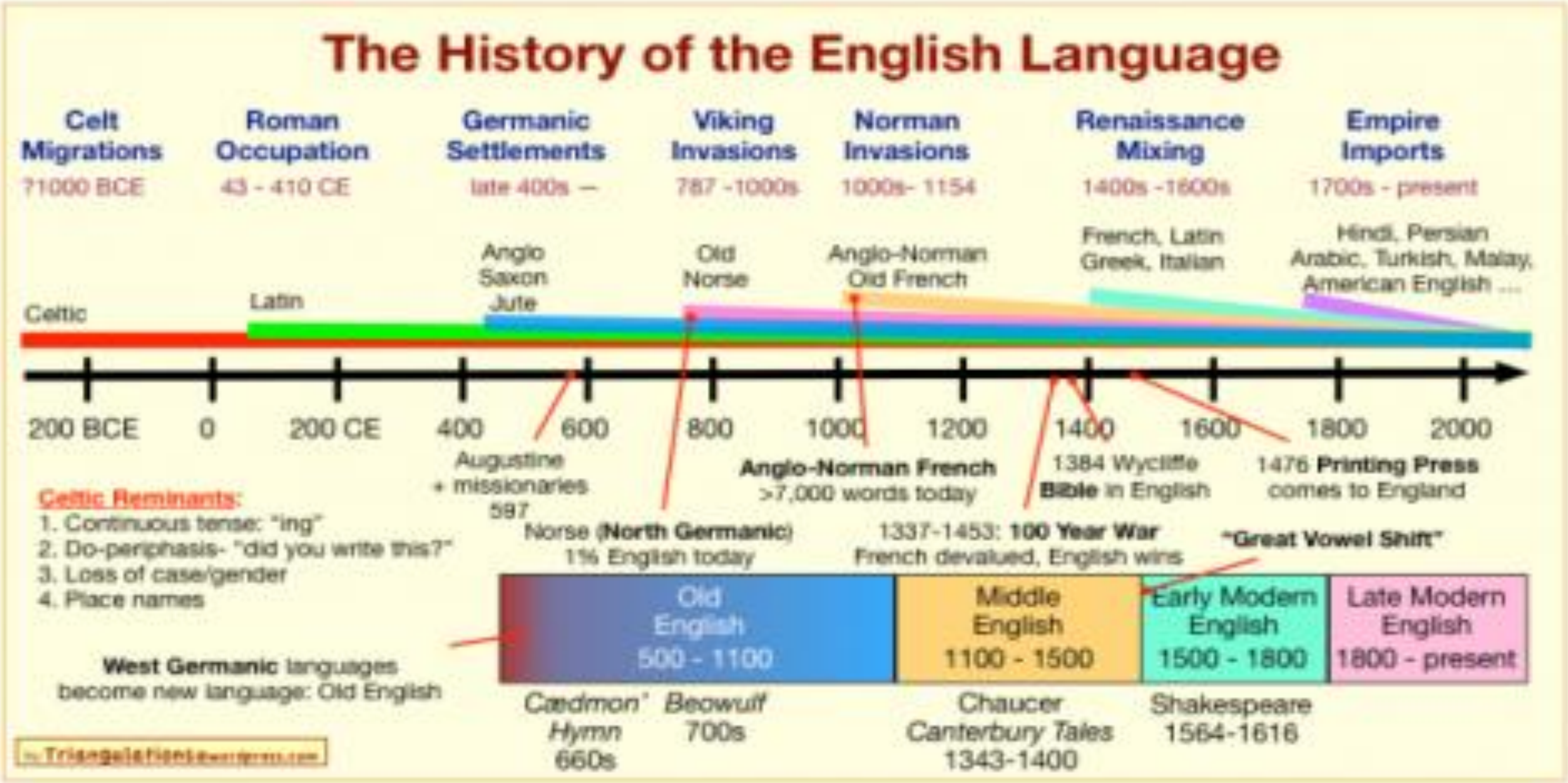
- What happens if you try to use the same tagger/parser for social media text?



Application x Languages x Domains



Variation over Time

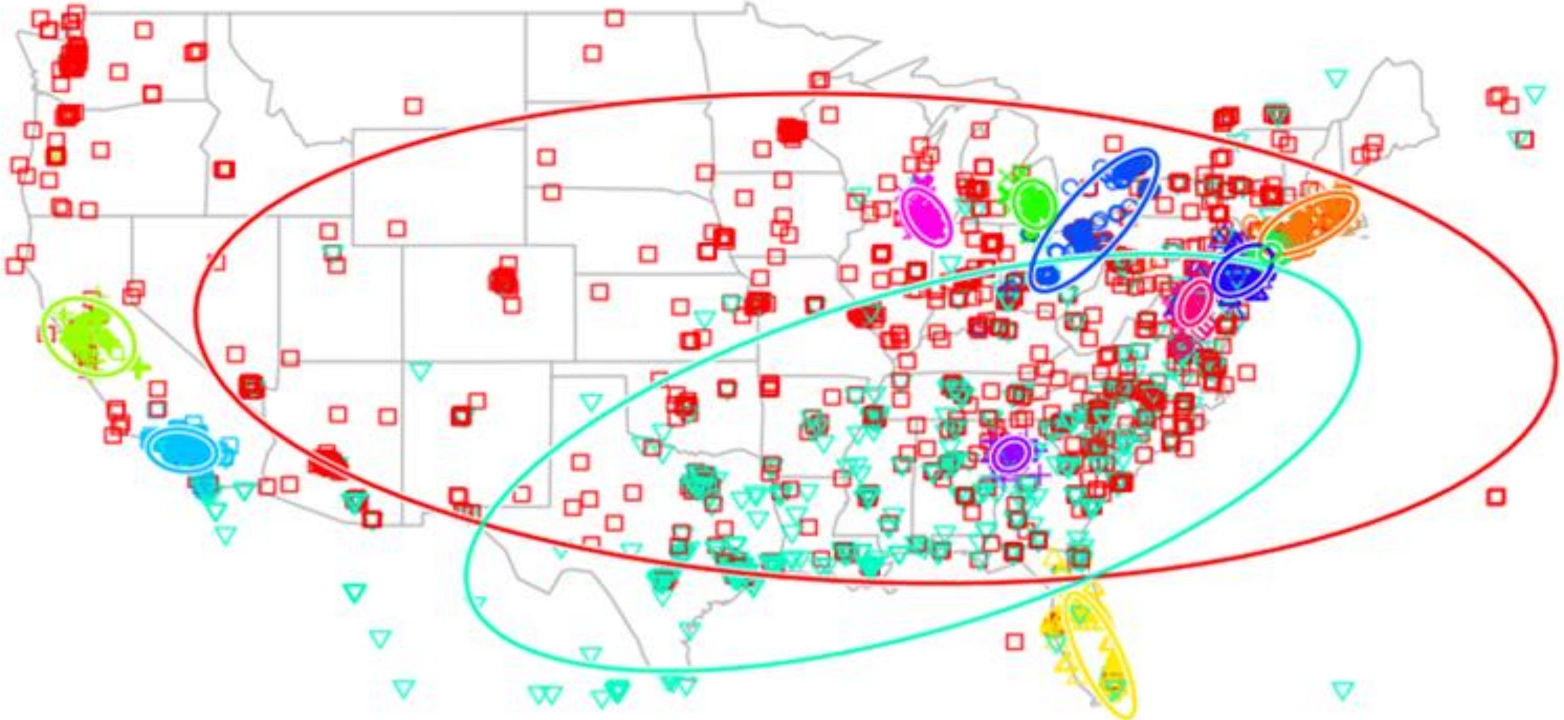


Variation over Time

24 New Words Invented by Teenagers , NYT 20220331

https://www.instagram.com/reel/C-NuNbutMD6/

Variation over Location



A Latent Variable Model for Geographic Lexical Variation [Eisenstein et al., 2010]

British & American English

British	American
anticlockwise	counter
appetizer	starter
aubergine	eggplant
biscuit	cookie
boot	trunk
braces	suspenders
candyfloss	cotton candy
car park	parking lot
chemist	drugstore
chips	French fries
cot	crib
courgette	zucchini
crisps	chips
drawing pin	thumbtack
dressing gown	robe
dummy	pacifier
dustbin	garbage can
flannel	washcloth
flat	apartment
football	soccer
fringe	bangs
grill	broil

British	American
grill	broiler
hairslide	barrette
holiday	vacation
jumper	sweater
lift	elevator
mobile phone	cell phone
number plate	license plate
off-licence	liquor store
oven glove	oven mitt
parting	part
pavement	sidewalk
petrol	gas, gasoline
postbox	mailbox
postcode	zip code
public school	private school
pushchair	stroller
shopping trolley	shopping cart
skipping rope	jump rope
sledge	sled
state school	public school

www.englishgrammarhere.com

Beyond conventional meaning

Implicit meaning behind language and Pragmatics

- Speech act [Austin 1962]
  - "Could you please pass the salt to me?"

- Implicature [Grice 1975]
  - Alice: "Are you going to Paul's party?"
  - Bob: "I have to work."

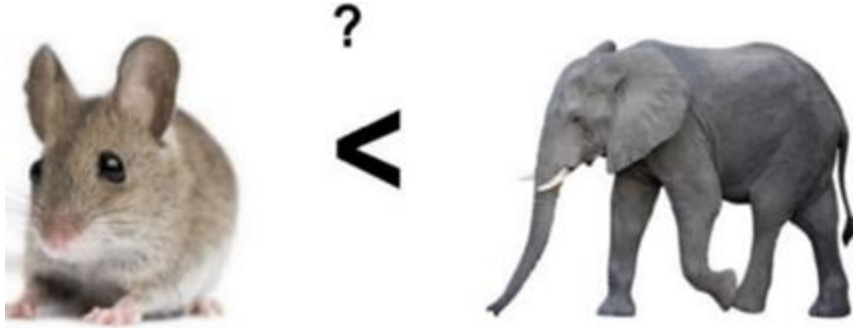
- labelling
- repeating
- answering
- **requesting (action)**
- requesting (answer)
- calling
- greeting
- protesting
- practicing

Unknown Representation

- We don't even know how to represent knowledge a human has/needs
- What is the meaning of word or sentence?
- How to model context or general knowledge?

"**Drink** this milk"

"Sunset is **beautiful**"



Elephants are **bigger than** mice?

Summary

- NLP is interdisciplinary
- Language consists of many levels of structure:
  - Phonology, syntax, semantics, discourse, pragmatics
- Processing language is difficult, due to
  - ambiguity, scales, sparsity, variation, implication, and representation
- Development of NLP models and representations grows rapidly
  - From rules to feature learning to RNNs to Transformers
- “Large” language models
  - Generalist AI or AGI via prompting and chat
  - Scaling law
  - Multimodal
  - Limitations? Future directions?

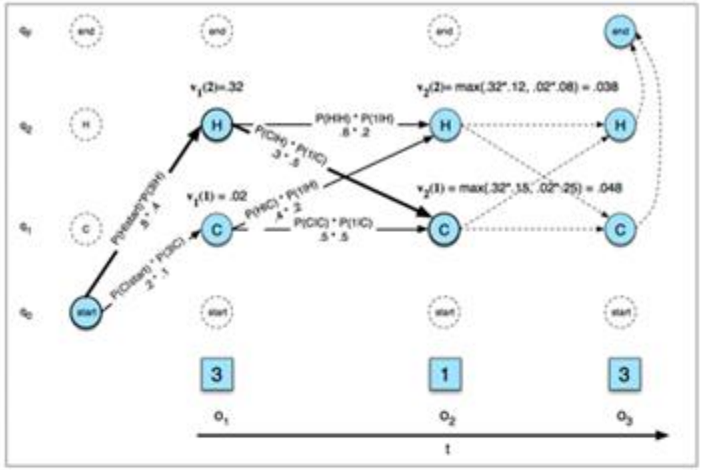
How to process language?

Methods

- Logic-based and rule-based NLP systems (~80s)
- Dynamic programming and Viterbi/CKY (~90s)
- Naïve Bayes, LogReg, HMM/CRF, SVM, N-gram LMs (~00s)

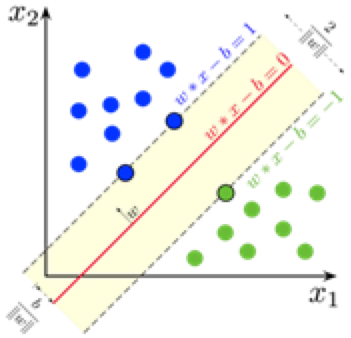
Some queries:

```
?- ancestor(mildred,mary).
yes % because parent(mildred,mary).
?- ancestor(irvin,nora).
yes % because
    % parent(irvin,ken) and
    % ancestor(ken,nora) because parent(ken,nora).
?- ancestor(chester,elizabeth).
yes % because
    % parent(chester,irvin)
    % and ancestor(irvin,elizabeth)
    %  because parent(irvin,ken) and
    %  ancestor(ken,elizabeth)
    %   because parent(ken,elizabeth).
```



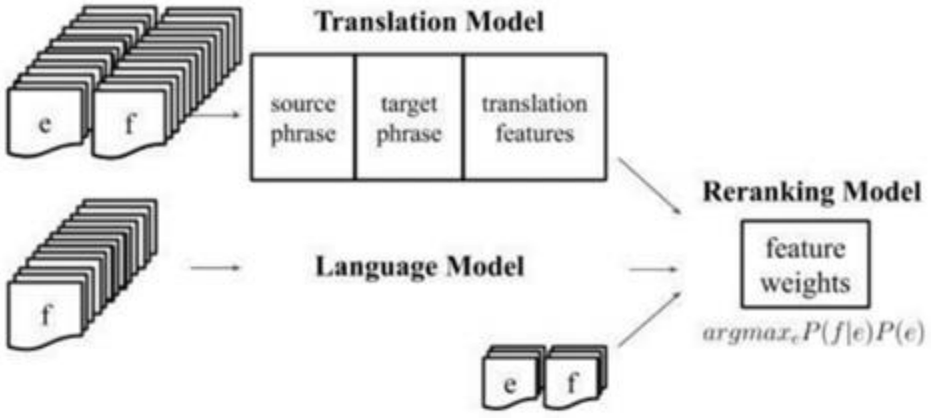
$$P(Y = y|X = x)$$

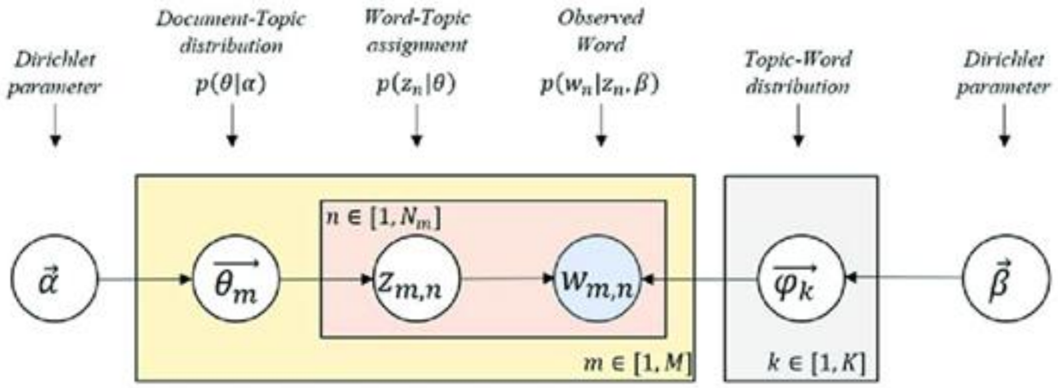
$$= \frac{P(Y = y)P(X = x|Y = y)}{\sum_y P(Y = y)P(X = x|Y = y)}$$



Methods

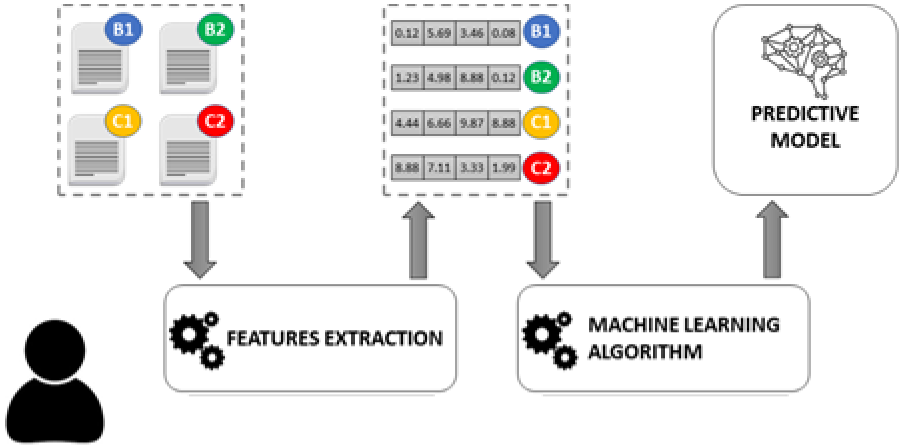
- Statistical NLP (~2005s)
- Latent variable models (Early ~2010s)
  - Specifying probabilistic structure between variables and inferring likely latent values

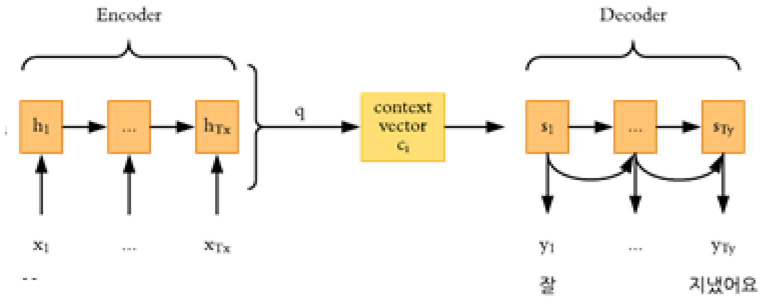


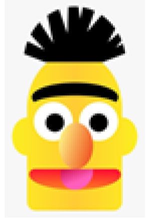


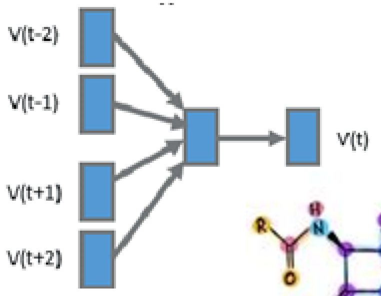
Representations

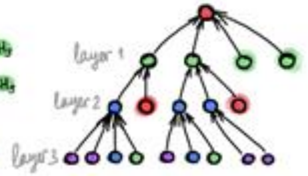
- ❑ Human-engineered features and SVMs (2005s ~ 2010s)
- ❑ Learned features/representations (2013s ~ 2018)

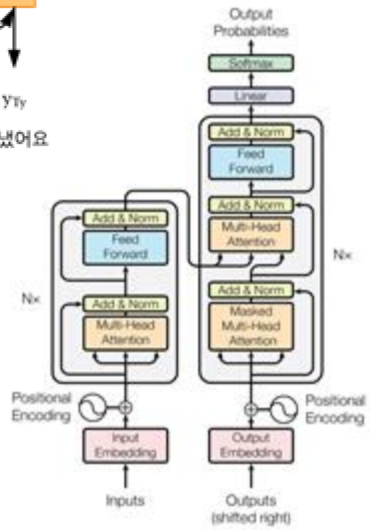




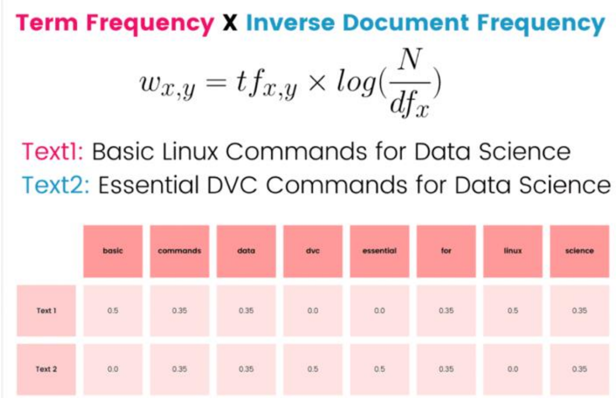


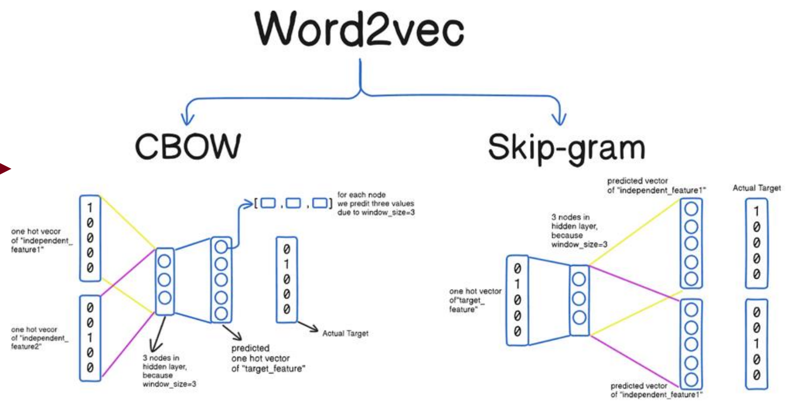


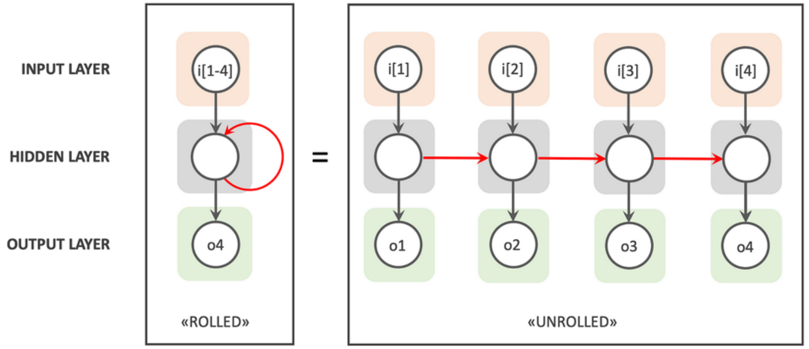


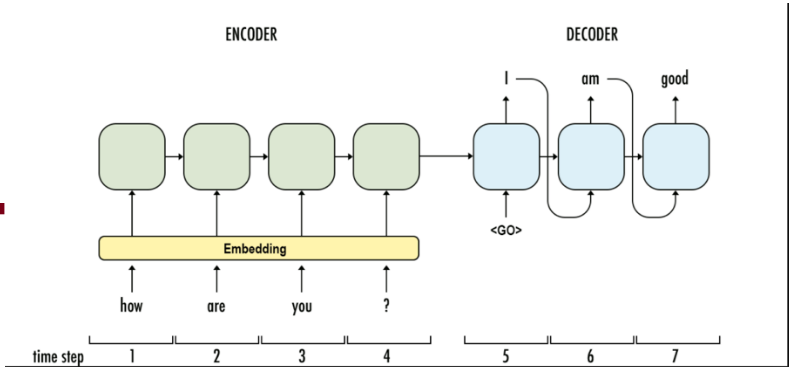


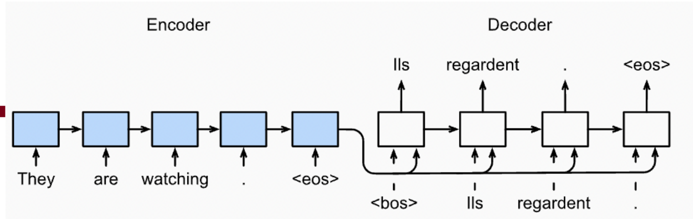
Representations (Developing Attention)











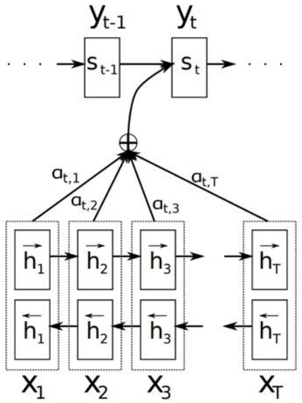
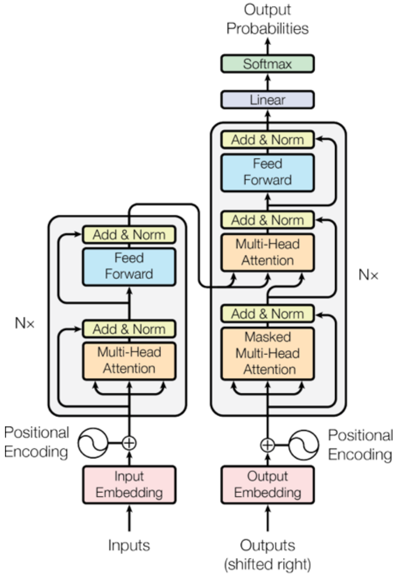
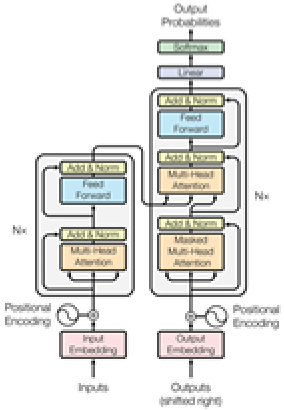
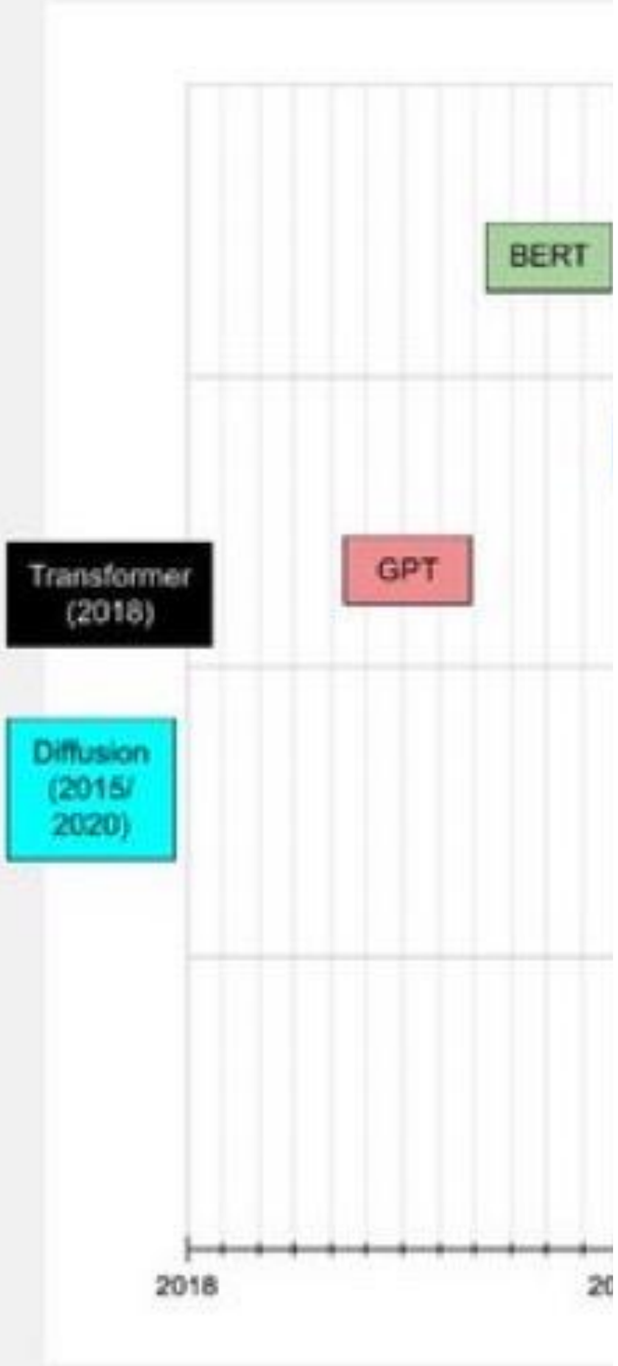


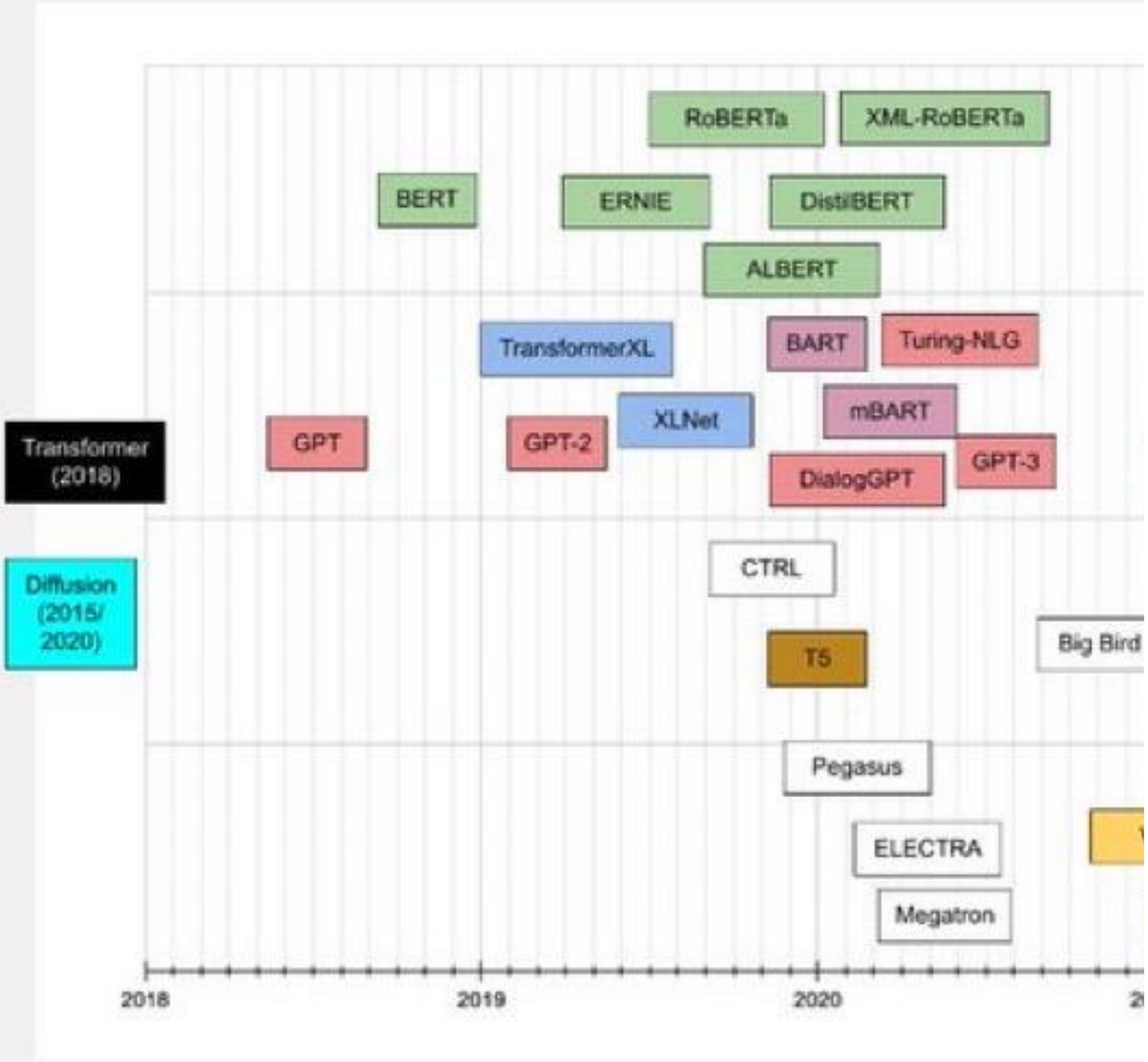
Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.



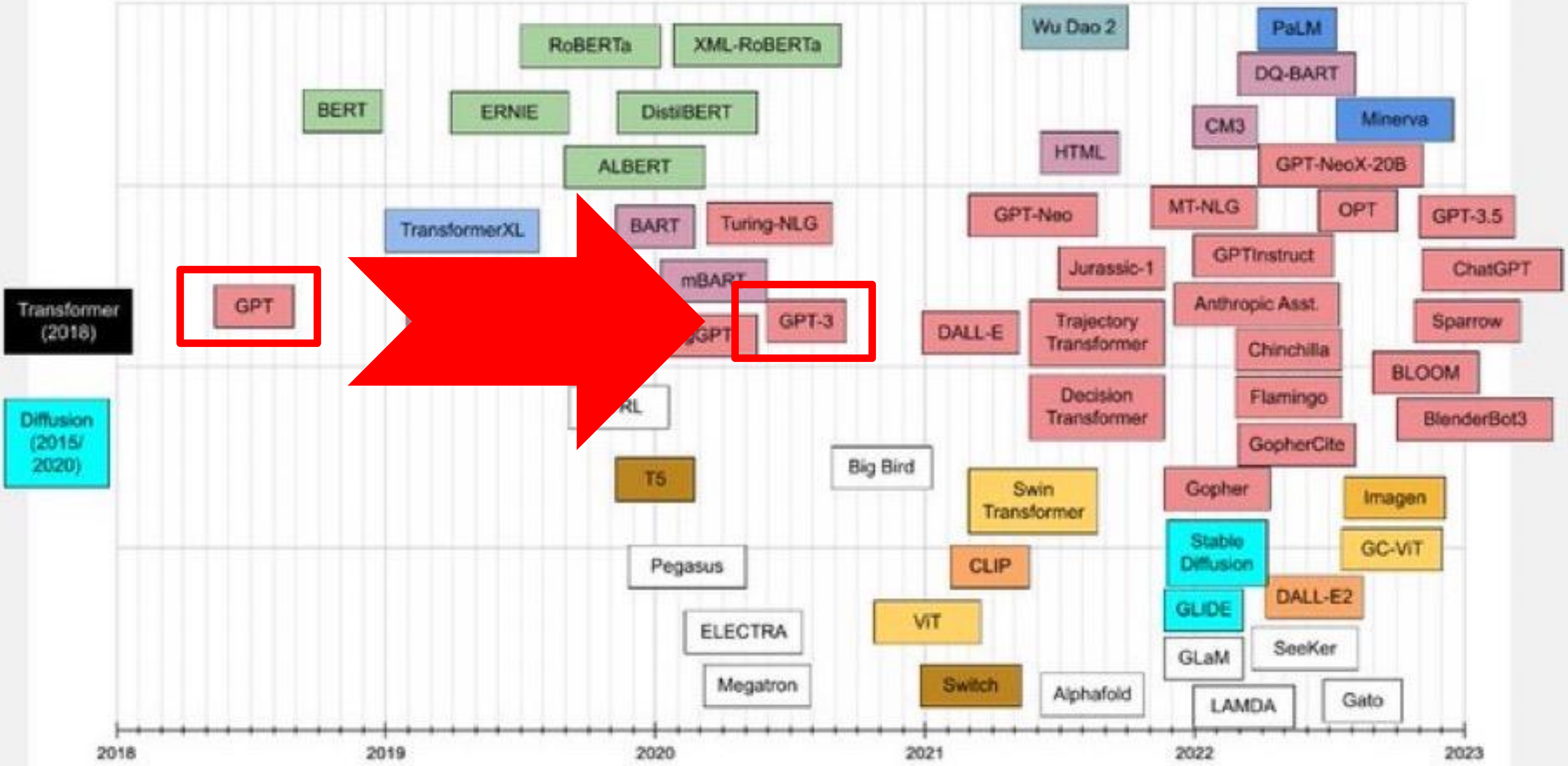
What happened in NLP over the last five years (2019-2024)?



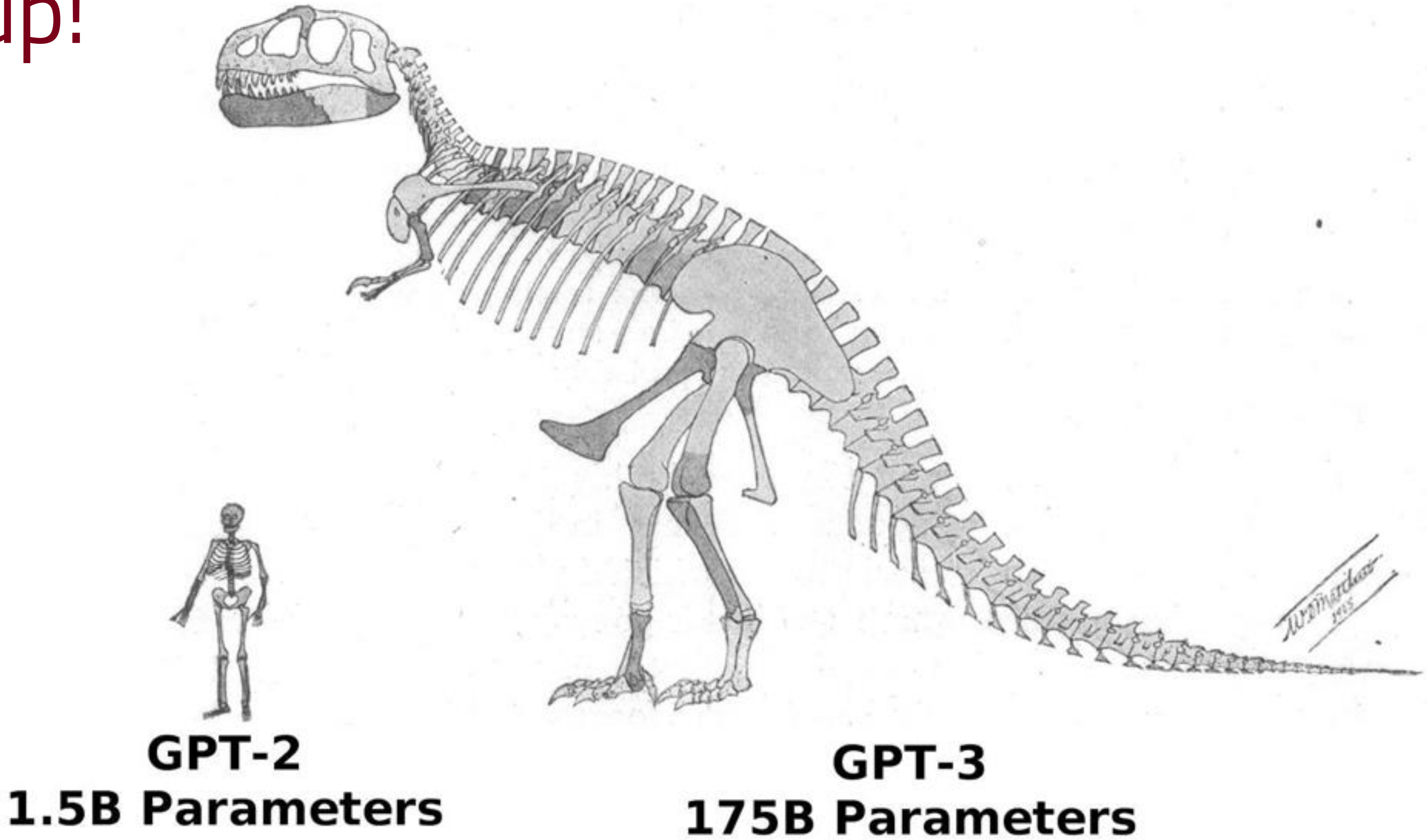

BERT
Transformer
(2018)
GPT
Diffusion
(2015/
2020)
2018

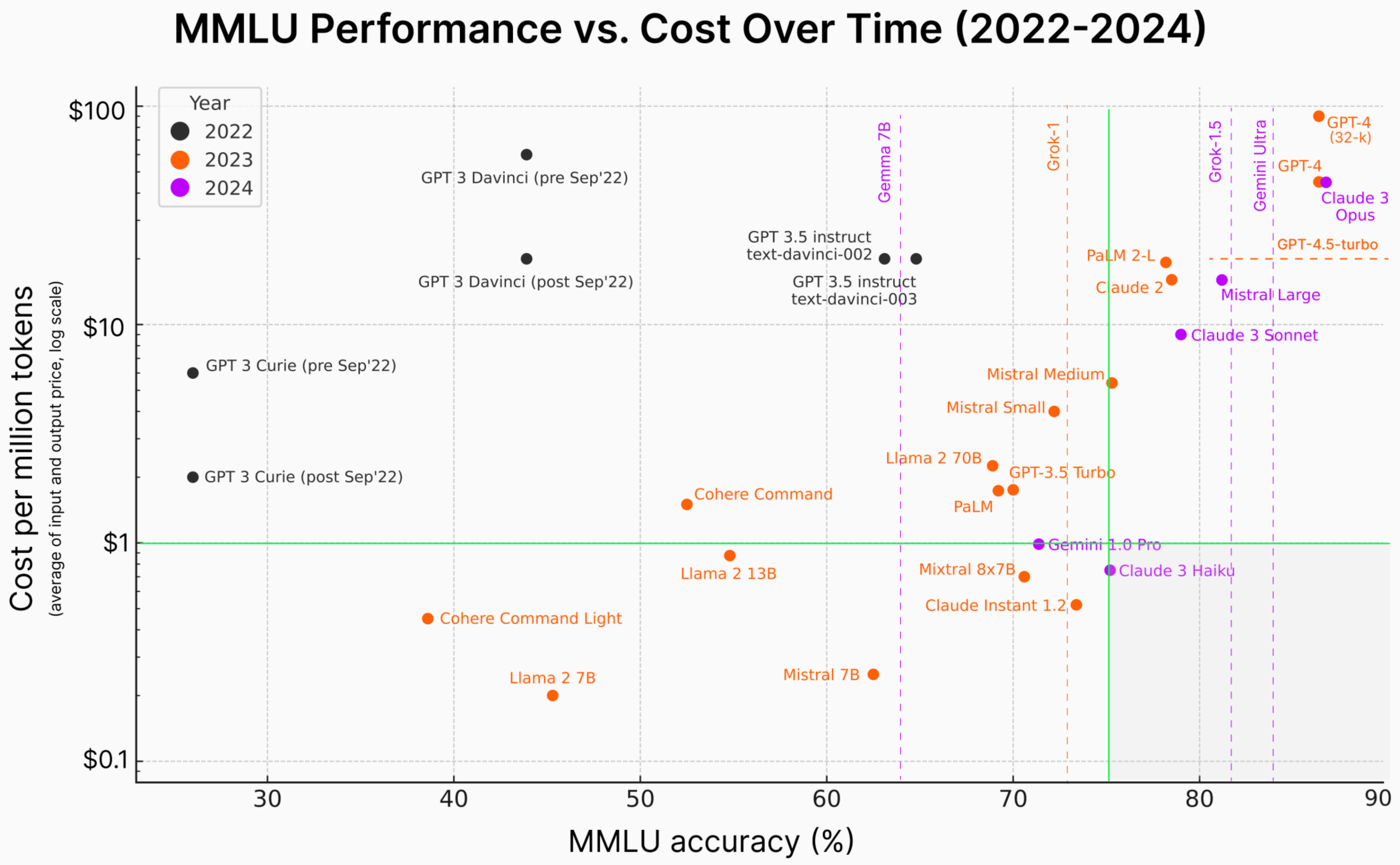


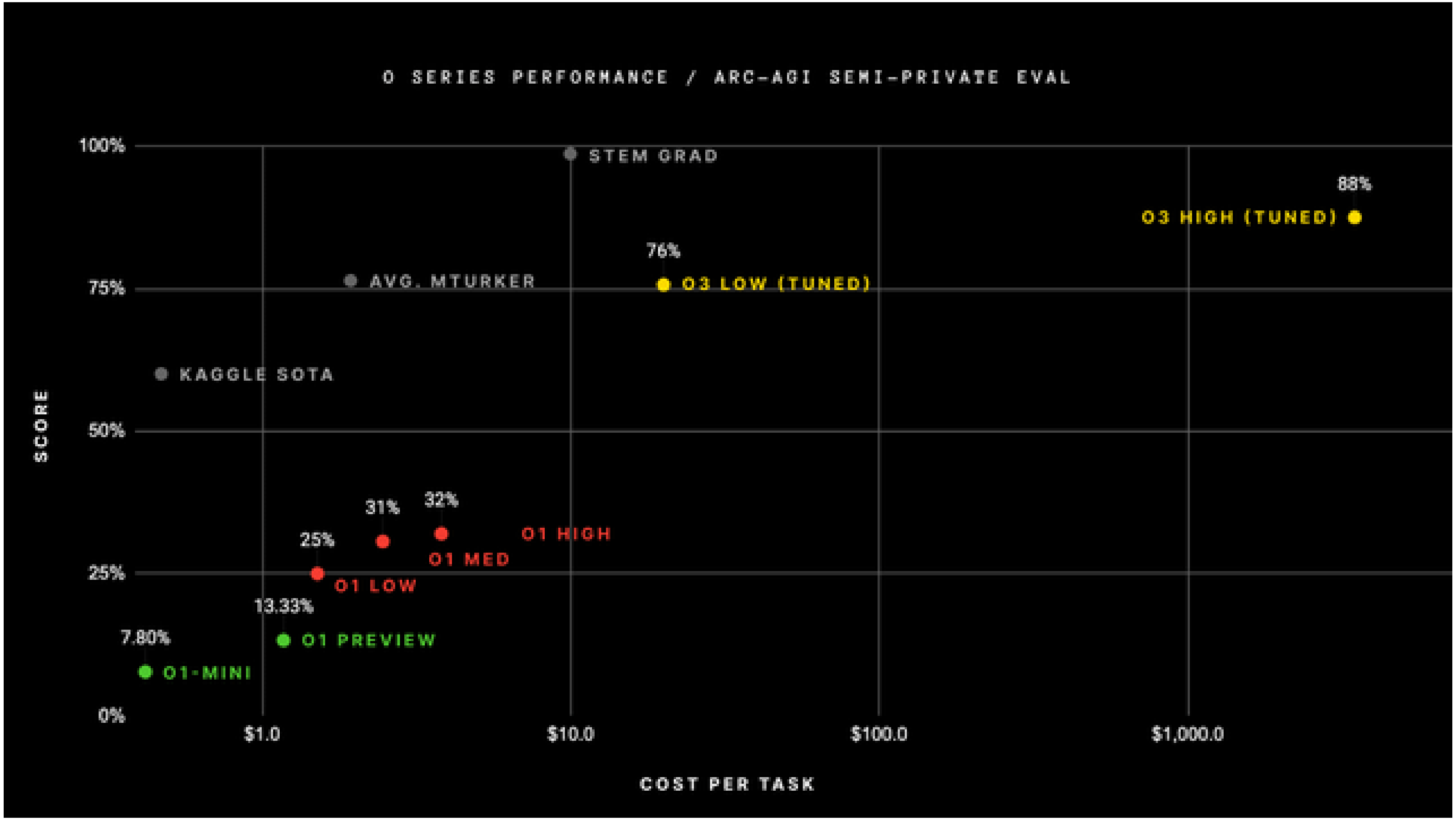
RoBERTa
XML-RoBERTa
BERT
ERNIE
DistilBERT
ALBERT
TransformerXL
BART
Turing-NLG
mBART
Transformer (2018)
GPT
GPT-2
XLNet
GPT-3
DialogGPT
CTRL
Diffusion (2015/ 2020)
T5
Big Bird
Pegasus
ELECTRA
Megatron
2018
2019
2020


Transformer (2018)
Diffusion (2015/ 2020)
GPT
BERT
RoBERTa
XML-RoBERTa
ERNIE
DistilBERT
ALBERT
TransformerXL
BART
Turing-NLG
mBART
GPT-3
T5
Pegasus
ELECTRA
Megatron
Big Bird
ViT
Switch
CLIP
DALL-E
Wu Dao 2
HTML
GPT-Neo
Jurassic-1
Trajectory Transformer
Decision Transformer
Swin Transformer
Alphafold
PaLM
DQ-BART
CM3
Minerva
GPT-NeoX-20B
MT-NLG
OPT
GPT-3.5
GPTInstruct
ChatGPT
Anthropic Asst.
Sparrow
Chinchilla
BLOOM
Flamingo
BlenderBot3
GopherCite
Gopher
Imagen
Stable Diffusion
GC-ViT
GLIDE
DALL-E2
GLaM
SeeKer
LAMDA
Gato
2018
2019
2020
2021
2022
2023

Scaling up!




MMLU Performance vs. Cost Over Time (2022-2024)
Year
2022
2023
2024
Cost per million tokens
(average of input and output price, log scale)
$100
$10
$1
$0.1
MMLU accuracy (%)
30
40
50
60
70
80
90
GPT 3 Davinci (pre Sep'22)
GPT 3 Davinci (post Sep'22)
GPT 3.5 instruct
text-davinci-002
GPT 3.5 instruct
text-davinci-003
GPT 3 Curie (pre Sep'22)
GPT 3 Curie (post Sep'22)
Gemma 7B
Grok-1
Grok-1.5
Gemini Ultra
GPT-4
(32-k)
GPT-4
Claude 3
Opus
GPT-4.5-turbo
PaLM 2-L
Claude 2
Mistral Large
Claude 3 Sonnet
Mistral Medium
Mistral Small
Llama 2 70B
GPT-3.5 Turbo
PaLM
Cohere Command
Gemini 1.0 Pro
Claude 3 Haiku
Llama 2 13B
Mixtral 8x7B
Claude Instant 1.2
Cohere Command Light
Llama 2 7B
Mistral 7B



O SERIES PERFORMANCE / ARC-AGI SEMI-PRIVATE EVAL
SCORE
100%
75%
50%
25%
0%
STEM GRAD
AVG. MTURKER
KAGGLE SOTA
88%
O3 HIGH (TUNED)
76%
O3 LOW (TUNED)
32%
O1 HIGH
31%
O1 MED
25%
O1 LOW
13.33%
O1 PREVIEW
7.80%
O1-MINI
$1.0
$10.0
$100.0
$1,000.0
COST PER TASK

The Leading Players

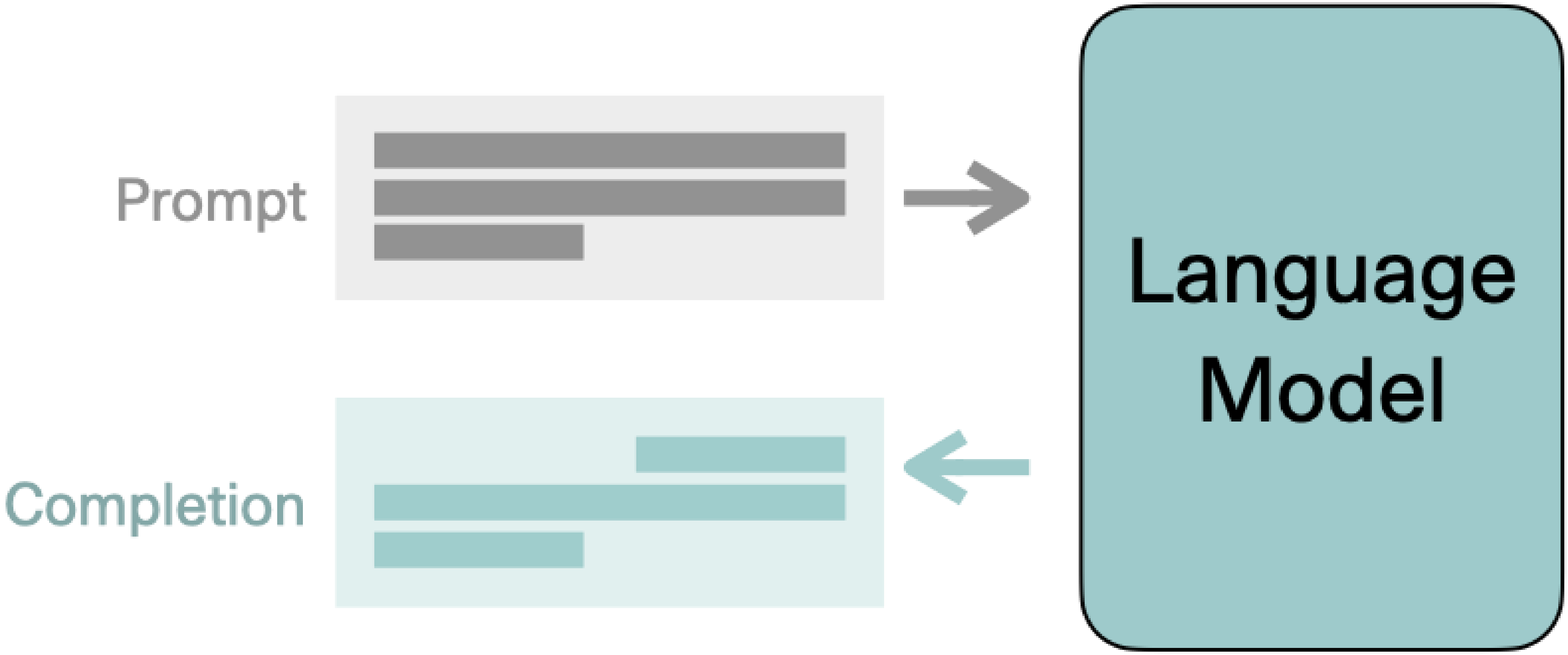
Qwen2.5

Meta Llama 3.2

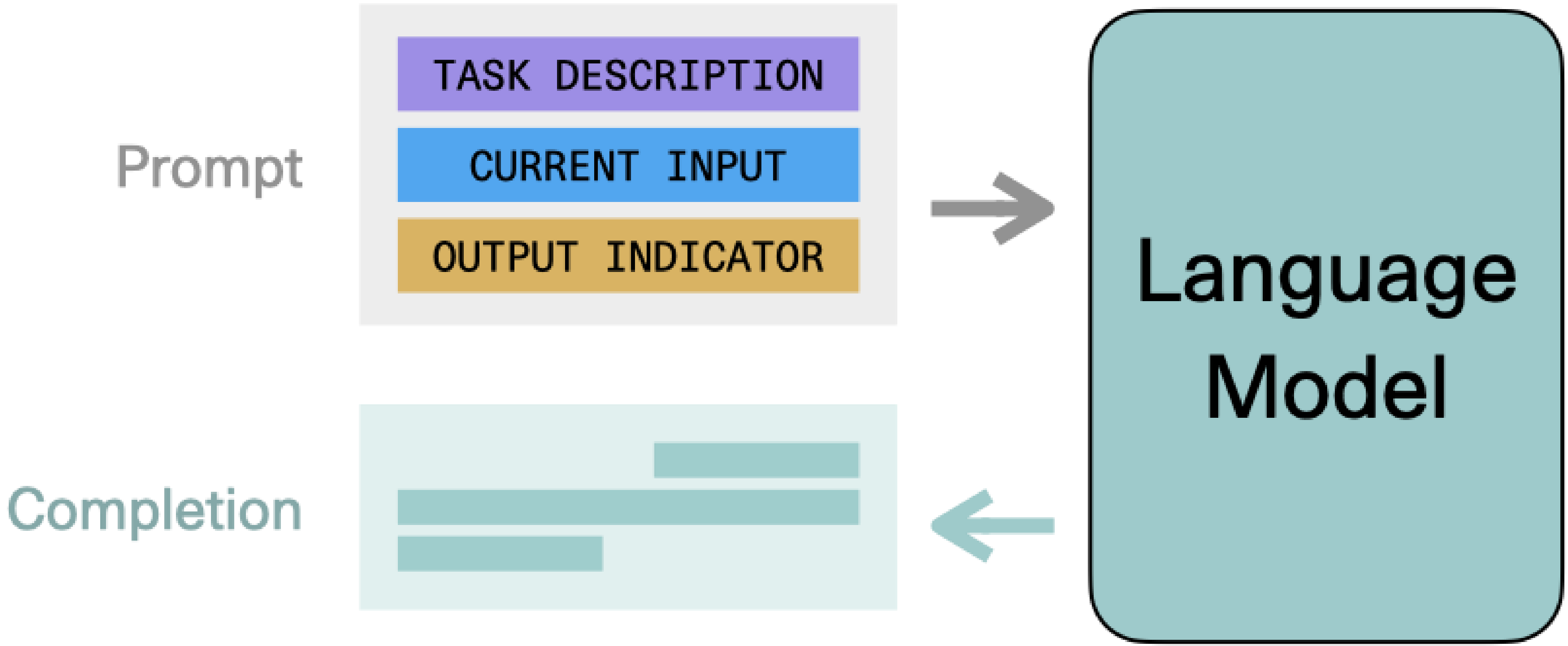
deepseek

ANTHROPIC

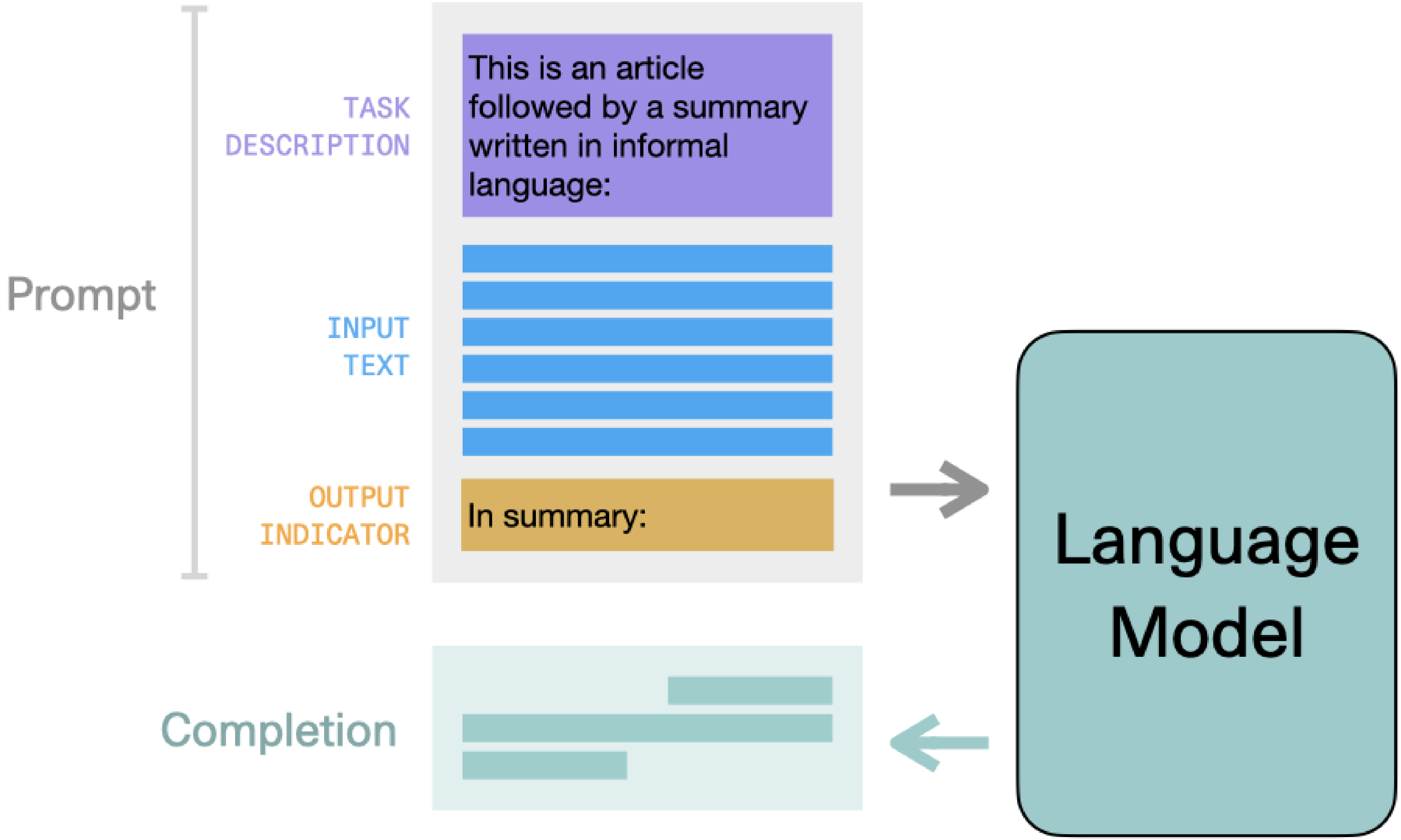
OpenAI



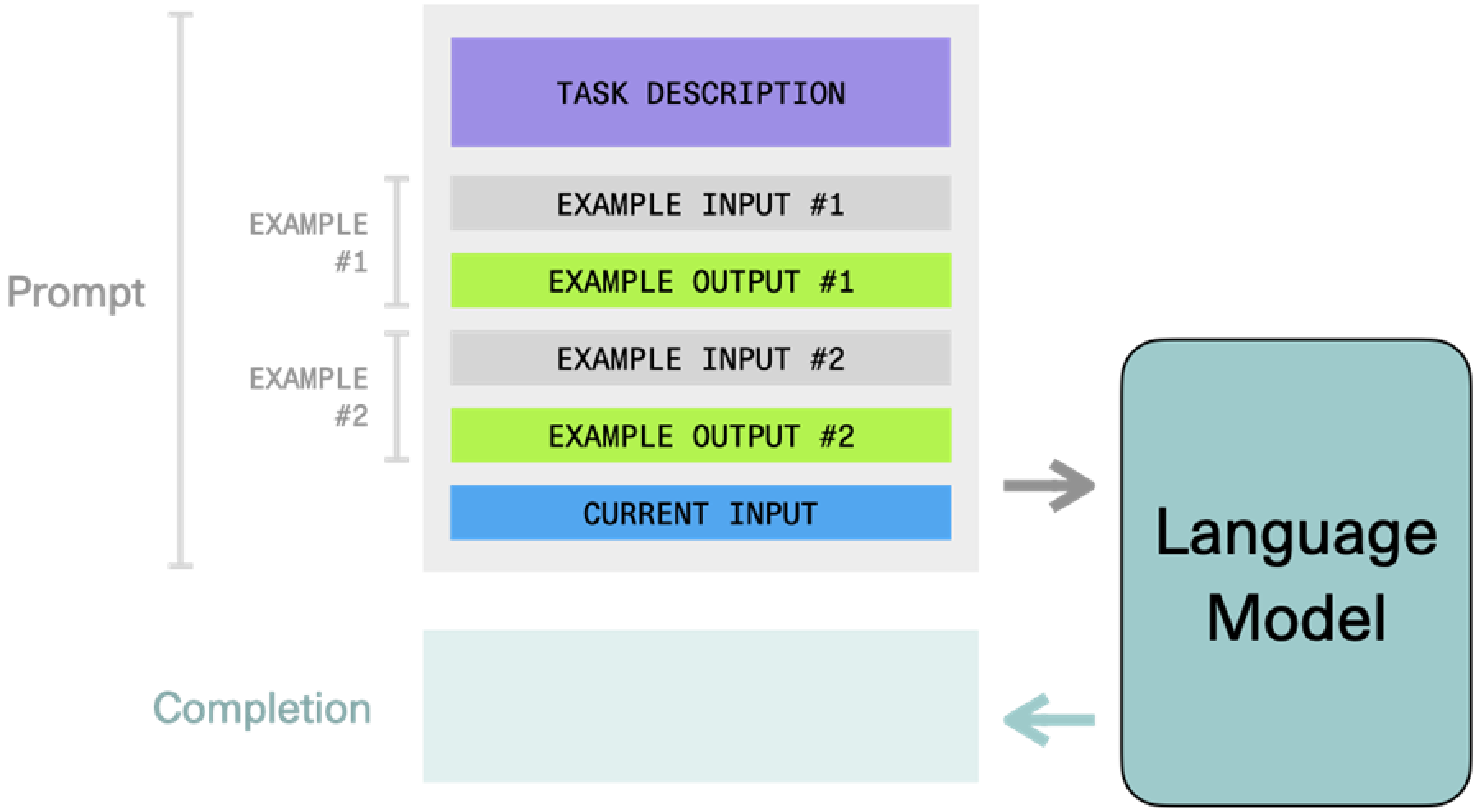
https://docs.cohere.ai/prompt-engineering-wiki/



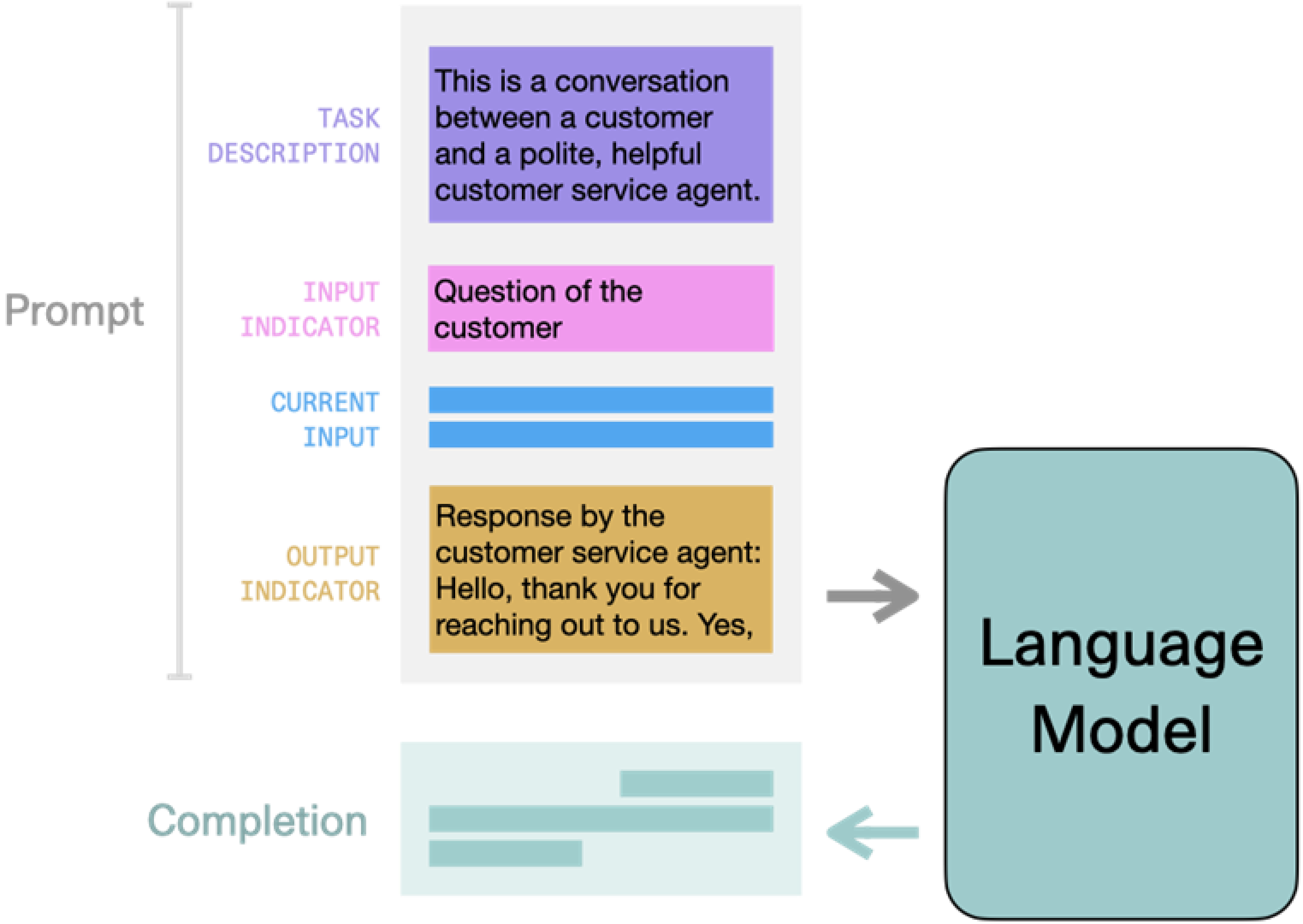
https://docs.cohere.ai/prompt-engineering-wiki/



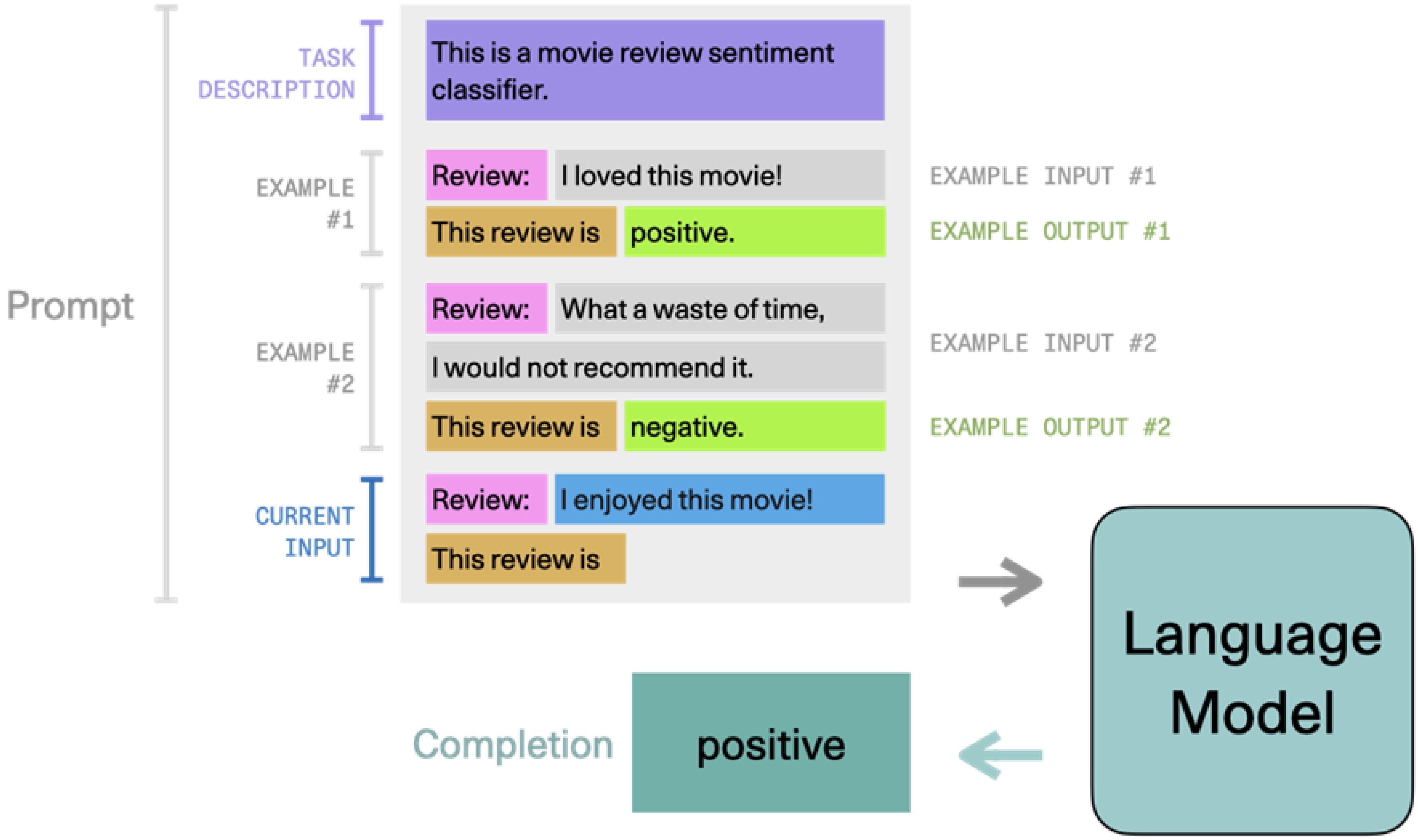
https://docs.cohere.ai/prompt-engineering-wiki/



https://docs.cohere.ai/prompt-engineering-wiki/



https://docs.cohere.ai/prompt-engineering-wiki/


Prompt
TASK DESCRIPTION
This is a movie review sentiment classifier.
EXAMPLE #1
Review:
I loved this movie!
This review is
positive.
EXAMPLE INPUT #1
EXAMPLE OUTPUT #1
EXAMPLE #2
Review:
What a waste of time,
I would not recommend it.
This review is
negative.
EXAMPLE INPUT #2
EXAMPLE OUTPUT #2
CURRENT INPUT
Review:
I enjoyed this movie!
This review is
Language Model
Completion
positive

Sentence classification via Prompting

Input Temperature:0	Classify the sentences below as positive, negative, neutral: Sentence: I enjoyed this movie despite the gory violence. Classification: Positive -- Sentence: It is beyond my comprehension how such a movie grossed over $100 USD. Classification: Negative -- Sentence: I can't say I hate it or love it. Classification: Neutral -- Sentence: I endured the silly plot purely because of the excellent acting of the hero. Classification:

https://towardsdatascience.com/a-quiet-shift-in-the-nlp-ecosystem-84672b8ec7af

Text Summarization via Prompting

Input Temperature:0	Summarize this for a second-grade student: An atom is the smallest unit of ordinary matter that forms a chemical element.[1] Every solid, liquid, gas, and plasma is composed of neutral or ionized atoms. Atoms are extremely small, typically around 100 picometers across. They are so small that accurately predicting their behavior using classical physics—as if they were tennis balls, for example—is not possible due to quantum effects.

https://towardsdatascience.com/a-quiet-shift-in-the-nlp-ecosystem-84672b8ec7af

Relation Extraction via Prompting

Input Temperature:0	Identify drugs, diseases and genes as well as the relations between them. Sentence: Imatinib is used to treat cancer Entity1: Imatinib (drug) Entity2: cancer (disease) Relation: treat -- Sentence: Imatinib can cause abdominal pain Entity1: Imatinib (drug) Entity2: abdominal pain (disease) Relation: cause -- Sentence: EGFR is overexpressed in many forms of cancers Entity1: EGFR (gene) Entity2: cancers (disease) Relation: overexpressed -- Sentence: Dasatinib, nilotinib is used as a combination therapy for some cancers Entity1: Dasatinib (drug), nilotinib (drug) Entity2: cancers (disease) Relation: combination therapy -- Sentence: Her hypophysitis secondary to ipilimumab was well managed with supplemental hormones Entity1:

https://towardsdatascience.com/a-quiet-shift-in-the-nlp-ecosystem-84672b8ec7af

Email Generation via Prompting

Input Temperature:0	Generate full emails from simple commands. Here are some examples: Command: Thank John for his mother's day gift Email: John, Thank you so much for your thoughtful gift. I hope to see you soon - Mom. -- Command: Tell Sam to email the invoice Email:

https://towardsdatascience.com/a-quiet-shift-in-the-nlp-ecosystem-84672b8ec7af

Code Generation via Prompting

Prompt	Model Response

```
// Translate from C to Python
int add_one ( int x ){
 int m = 1;
 while ( x & m ) {
   x = x ^ m;
   m <<= 1;
 }
 x = x ^ m;
 return x; }
```

https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

Mathematical Reasoning via Prompting

Input Temperature:0	Calculate 4.5e1 + 1.5e2

Jurassic-X (7.5B) → Calculator

4.5e1 + 1.5e2=195

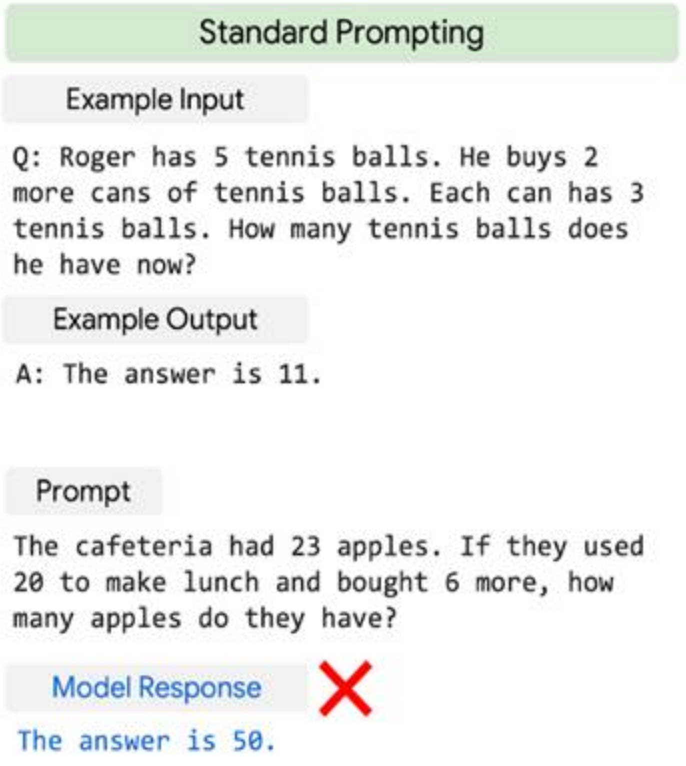
&frasl Explain answer

X=(4.5e1+1.5e2)

https://towardsdatascience.com/a-quiet-shift-in-the-nlp-ecosystem-84672b8ec7af

Chain-of-Thought Prompting

Few-shot CoT



Standard prompting versus chain-of-thought prompting for an example grade-school math problem. Chain-of-thought prompting decomposes the prompt for a multi-step reasoning problem into intermediate steps (highlighted in yellow), similar to how a person would approach it.

https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

Chain-of-Thought Prompting

Zero-shot CoT

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

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

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

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

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

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

Figure 1: Example inputs and outputs of GPT-3 with (a) standard Few-shot ([Brown et al., 2020]), (b) Few-shot-CoT ([Wei et al., 2022]), (c) standard Zero-shot, and (d) ours (Zero-shot-CoT). Similar to Few-shot-CoT, Zero-shot-CoT facilitates multi-step reasoning (blue text) and reach correct answer where standard prompting fails. Unlike Few-shot-CoT using step-by-step reasoning examples **per task**, ours does not need any examples and just uses the same prompt "Let's think step by step" *across all tasks* (arithmetic, symbolic, commonsense, and other logical reasoning tasks).

Large Language Models are Zero-Shot Reasoners, https://arxiv.org/abs/2205.11916

What is meaning of life?
Let's think step by step.

First, what is the meaning of the word "life"?

The meaning of the word "life" is the natural or human state of existence. In other words, it is the time between being born and dying.

Now that we know the meaning of the word "life", let's try to figure out the meaning of "the meaning of life".

One possible interpretation is that it is the purpose or reason for living. In other words, why we are alive and what we are supposed to do with our time on earth.

Another interpretation is that it is the way we live our lives. This includes our values, beliefs, and actions.

Ultimately, the meaning of life is something that each person has to figure out for themselves. There is no single answer that is right for everyone.

Submit 183

Scaling law in language model

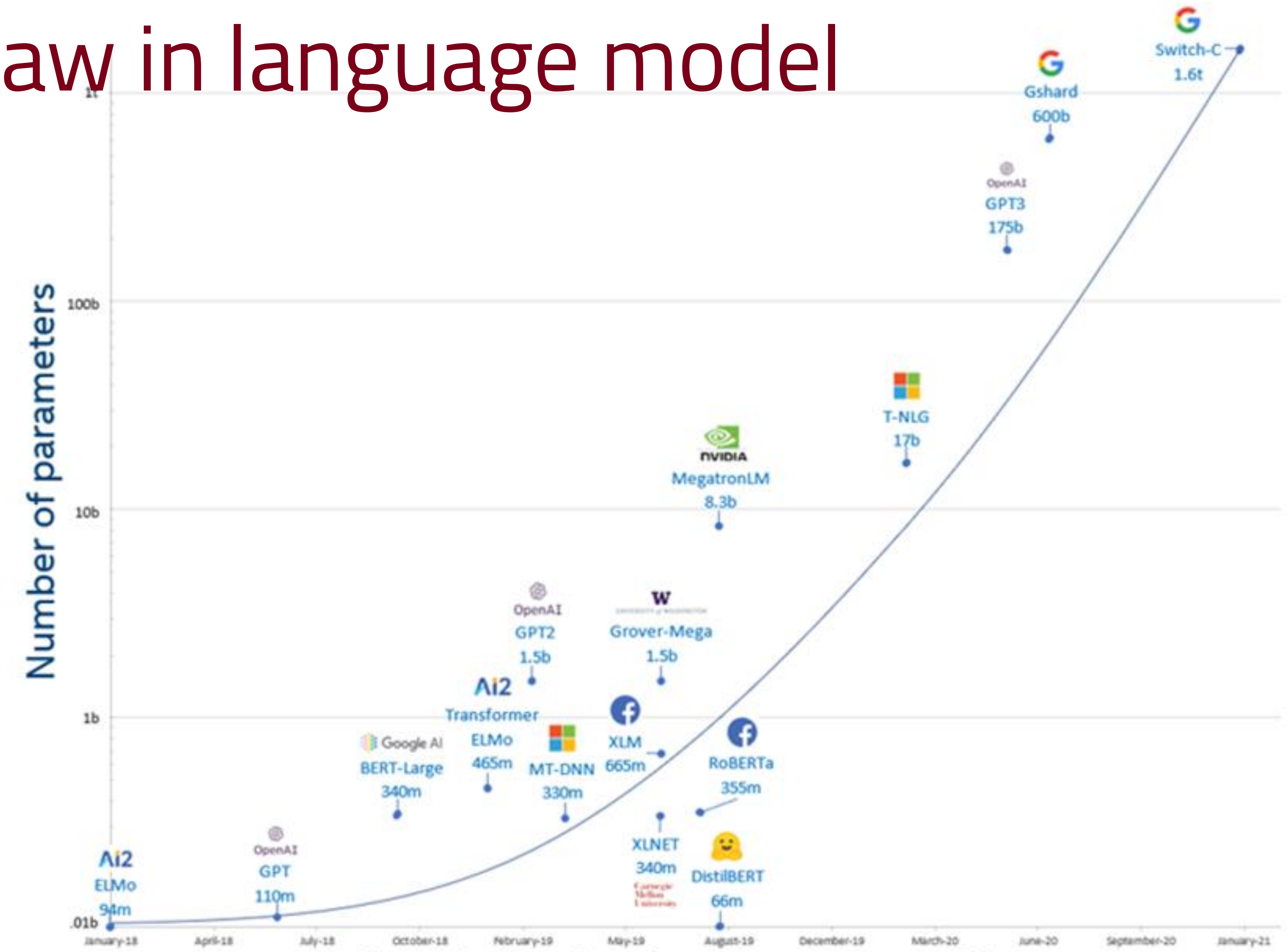
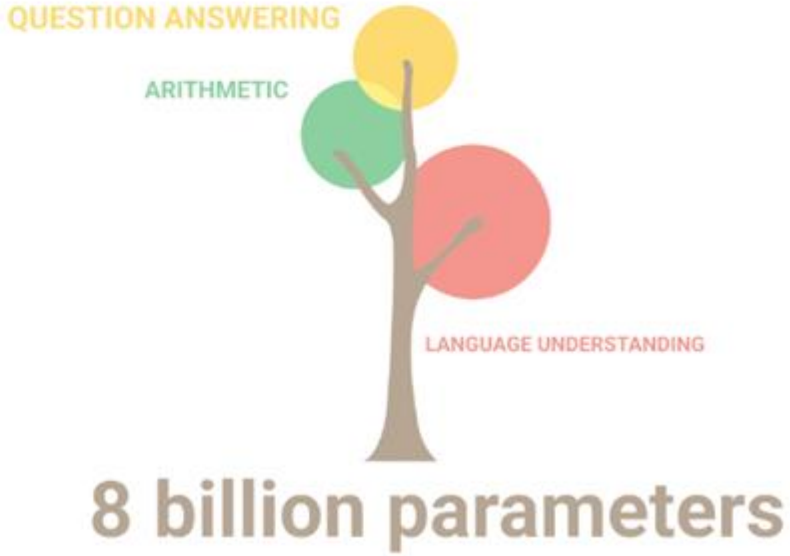
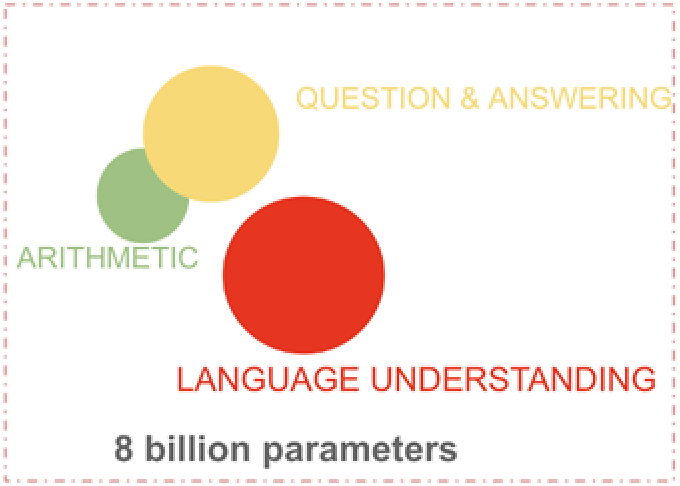
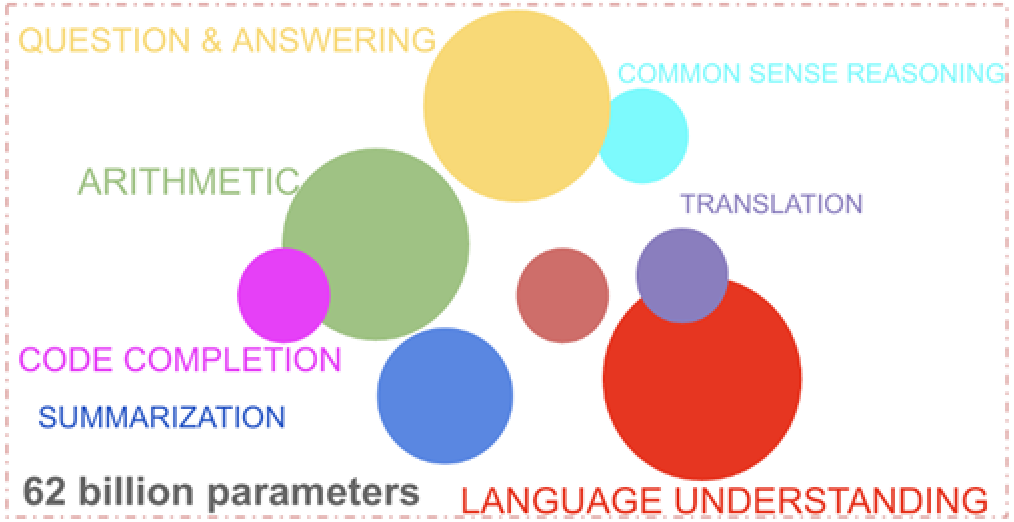


Figure 1: Exponential growth of number of parameters in DL models

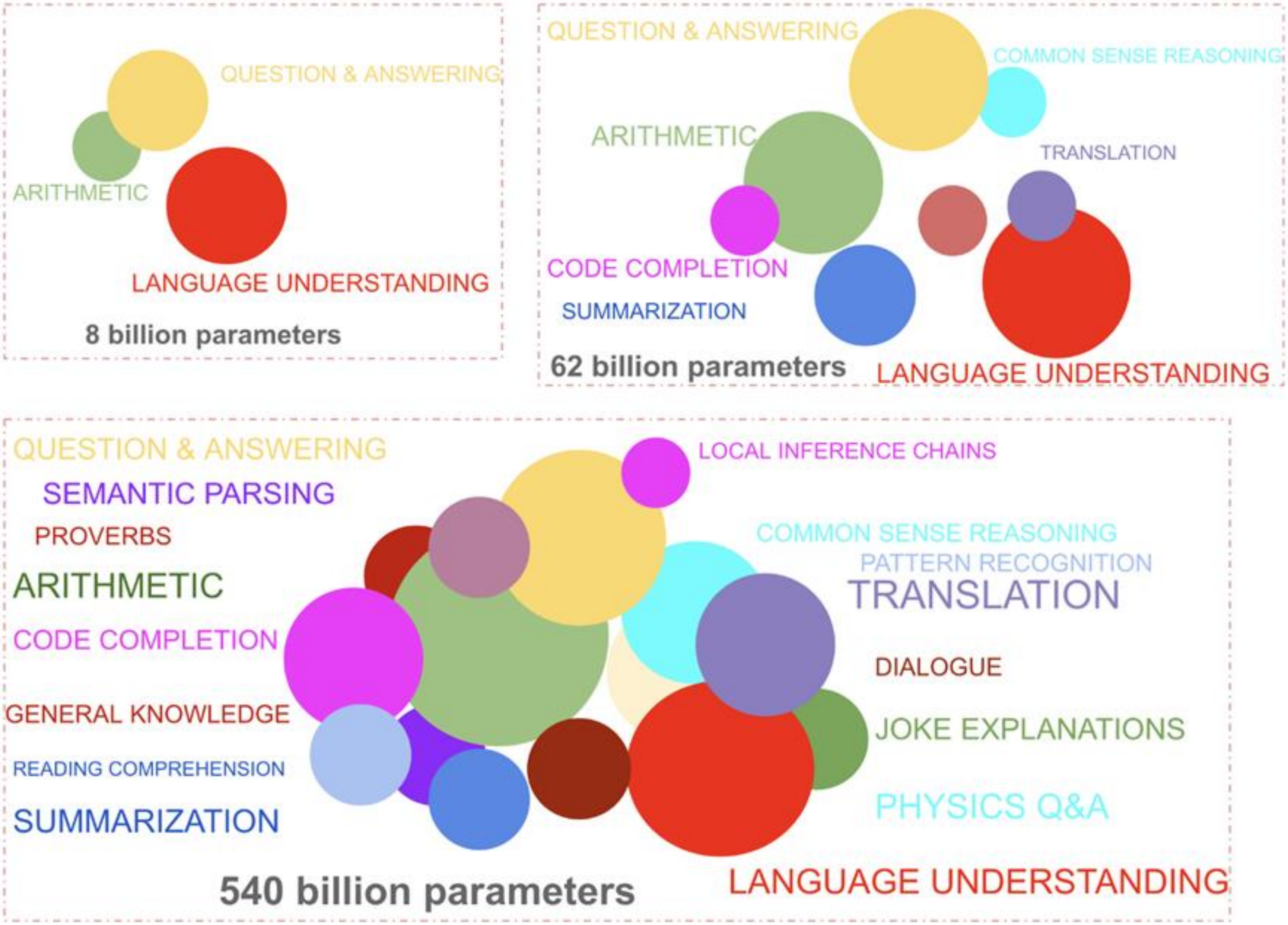


https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

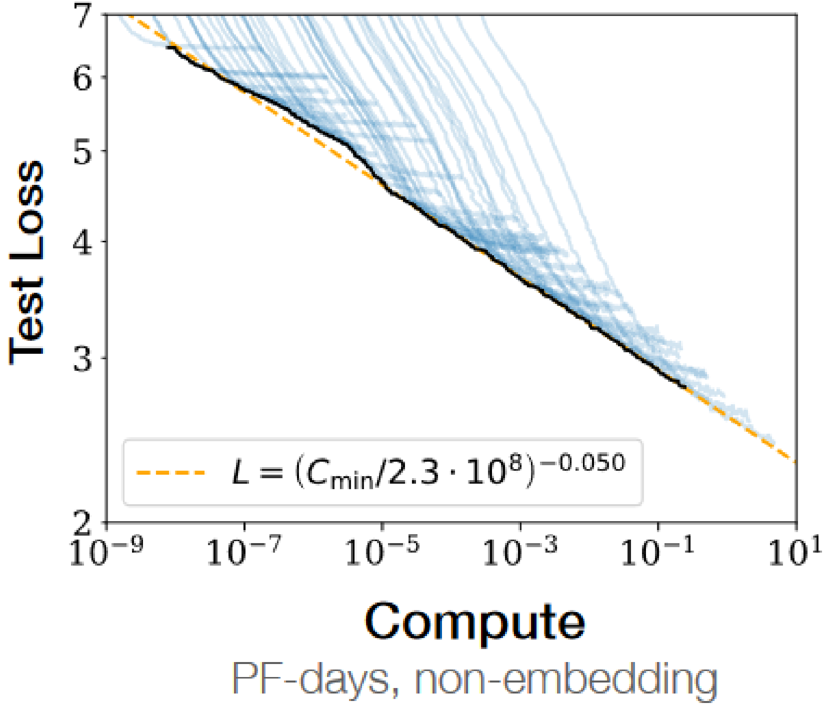




https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

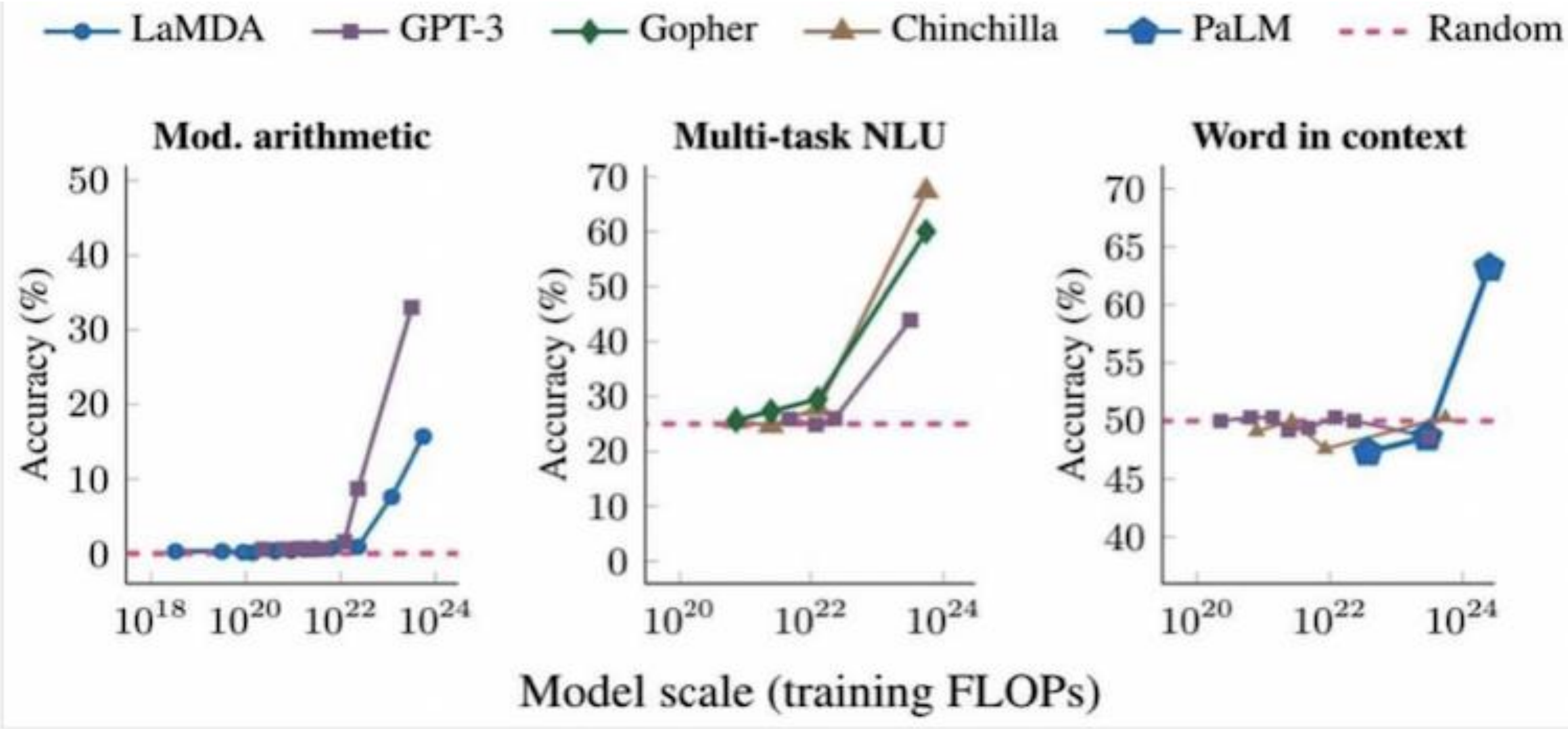


https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html



Test Loss
$L = (C_{min}/2.3 \cdot 10^8)^{-0.050}$
$L = (D/5.4 \cdot 10^{13})^{-0.095}$
$L = (N/8.8 \cdot 10^{13})^{-0.076}$
Compute
PF-days, non-embedding
Dataset Size
tokens
Parameters
non-embedding

Emergent behavior from Scaling Law



Jeff Dean https://ai.googleblog.com/2023/01/google-research-2022-beyond-language.html

Generalist
NLP Agent

Generalist AI Agent

PROMPT:
Email: "Dear Alan, how should I learn more about AI? Best, Jim"
Respond to this email diplomatically and comprehensively.

Dear Jim,

Thank you for your email. Learning about AI can be a fascinating and rewarding experience. There are many ways to do this, and the best approach will depend on your individual interests and goals. Here are a few suggestions to get you started:

- If you are new to AI, one of the best ways to learn more is to start with some of the foundational concepts and principles. This can help you gain a better understanding of how AI works, and what it is capable of. Some good resources for this include online courses, books, and tutorials.
- Another way to learn about [...] some of the applications and

Try again

PROMPT:
Add an extra column. In the new column, insert an emoji symbolising the movie.

Here is a table showing ten slow-paced and relaxing movies set at night, with an added column for an emoji symbolizing the movie:

Movie Title	Director	Year Released	Emoji Symbol
Moonlight	Barry Jenkins	2016	🌙
Arrival	Denis Villeneuve	2016	🛸
The Tree of Life	Terrence Malick	2011	🌳
A Ghost Story	David Lowery	2017	👻
Her	Spike Jonze	2013	💻
The Theory of Everything	James Marsh	2014	🌌
Melancholia	Lars von Trier	2011	🌌
The Road	John Hillcoat	2009	🚗
The Shawshank Redemption	Frank Darabont	1994	🚪
The Elephant Man		1980	🐘

Try again

https://chat.openai.com/chat

Generalist AI Agent

PROMPT:
Write a new piece for pipe organ in the style of Bach

Sure! Here is a short piece for pipe organ in the style of Bach:

Copy code

```
Prelude in C Major

C Major: C E G

1. (C) All (E) hail the (G) pow'r of (C) Jesus' (E) name!
2. Let (G) angels (C) prostrate (E) fall;
3. Bring (C) forth the (E) royal (G) diadem,
4. And (C) crown Him (E) Lord of (G) all.

Interlude: C Major chord arpeggio

5. (C) Ye who (E) own the (G) faith of (C) Jesus,
6. (E) Praise the (G) God of (C) grace!
7. (C) Let (E) ev'ry (G) tribe and (C) tongue
8. (E) In this (G) glad (C) hour (E) sing (C) of His (E) love and
```

PROMPT:

```
char *strcpy(char *t, const char *s) {
char *p = t;
while(*t++=*s++);
return p;
}
```

Explain this code

This code appears to be a implementation of the `strcpy` function in C, which copies the string pointed to by `s` (including the null character at the end) into the array pointed to by `t`.

The function takes two arguments: a pointer `t` to an array of characters where the copied string will be stored, and a pointer `s` to a null-terminated string that will be copied.

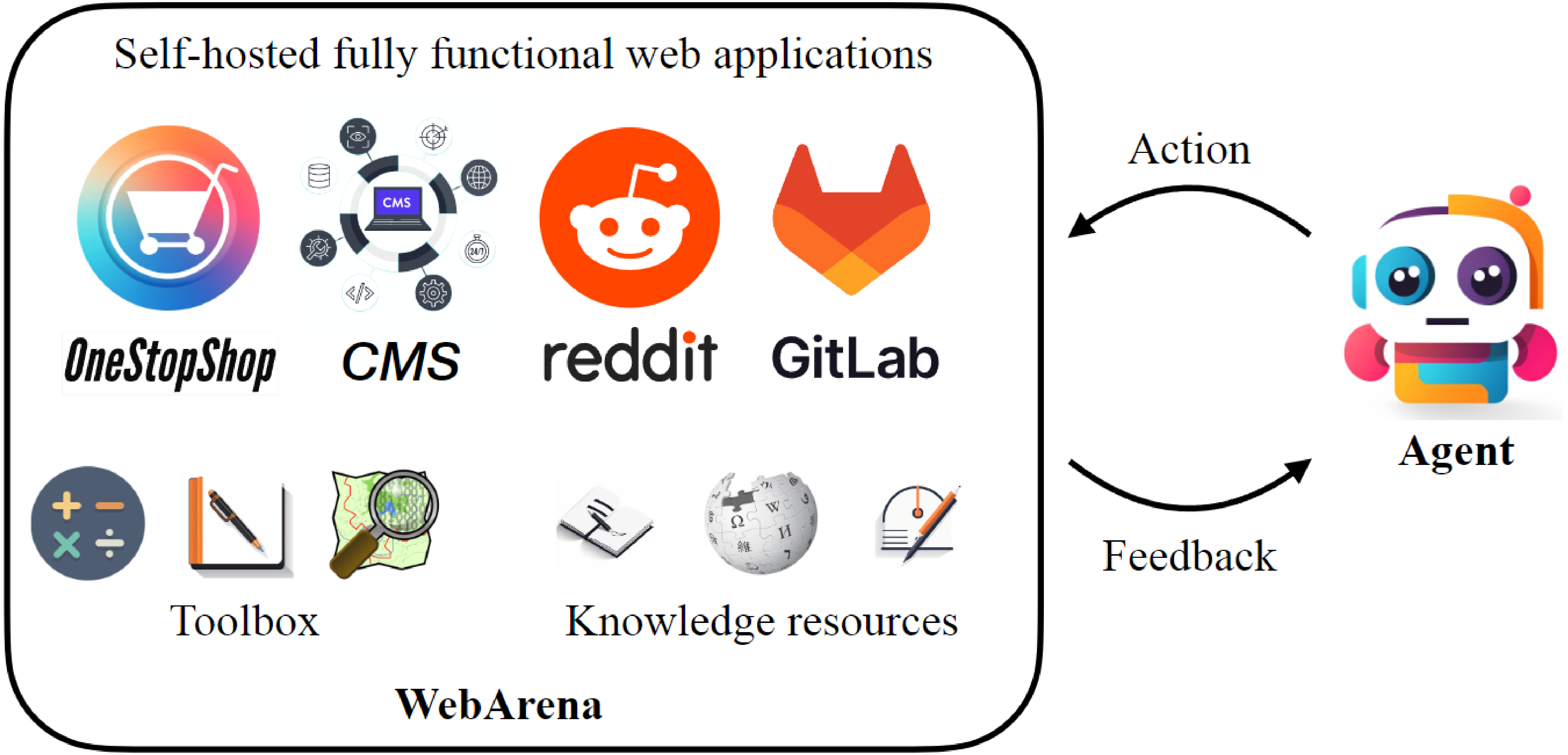
The function first declares a local variable `p` and initializes it with the value of `t`, which is the destination array. The function then enters a loop that uses the `=` and `++` operators to copy each character from `s` to `t` until it reaches the null character at the end of `s`.

Finally, the function returns the value of `p`, which is the original value of `t` (the destination array).

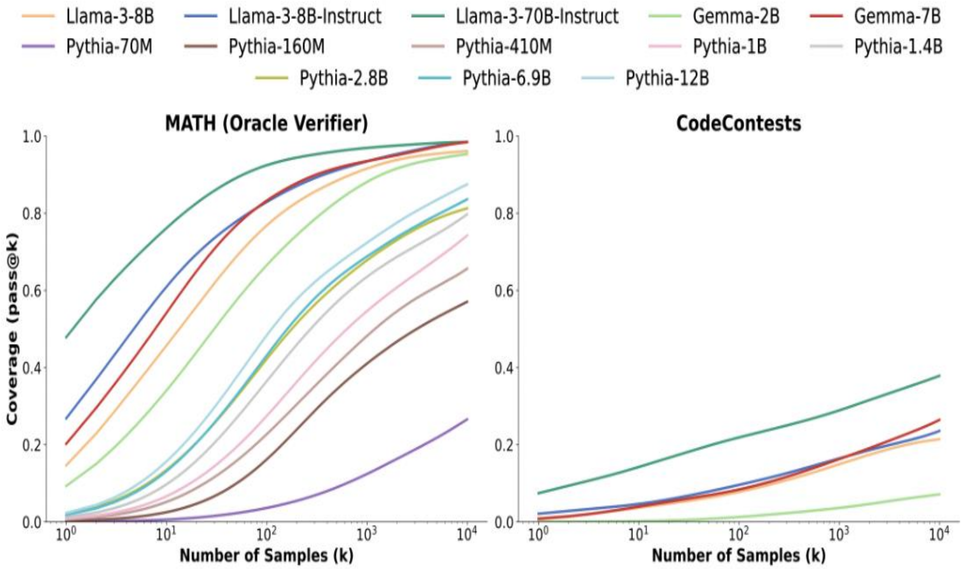
Note that this implementation of `strcpy` is potentially unsafe because it does

https://chat.openai.com/chat

Web Based Agent



Reasoning (Test-time compute/scaling)



Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

Generalist AI across different modalities

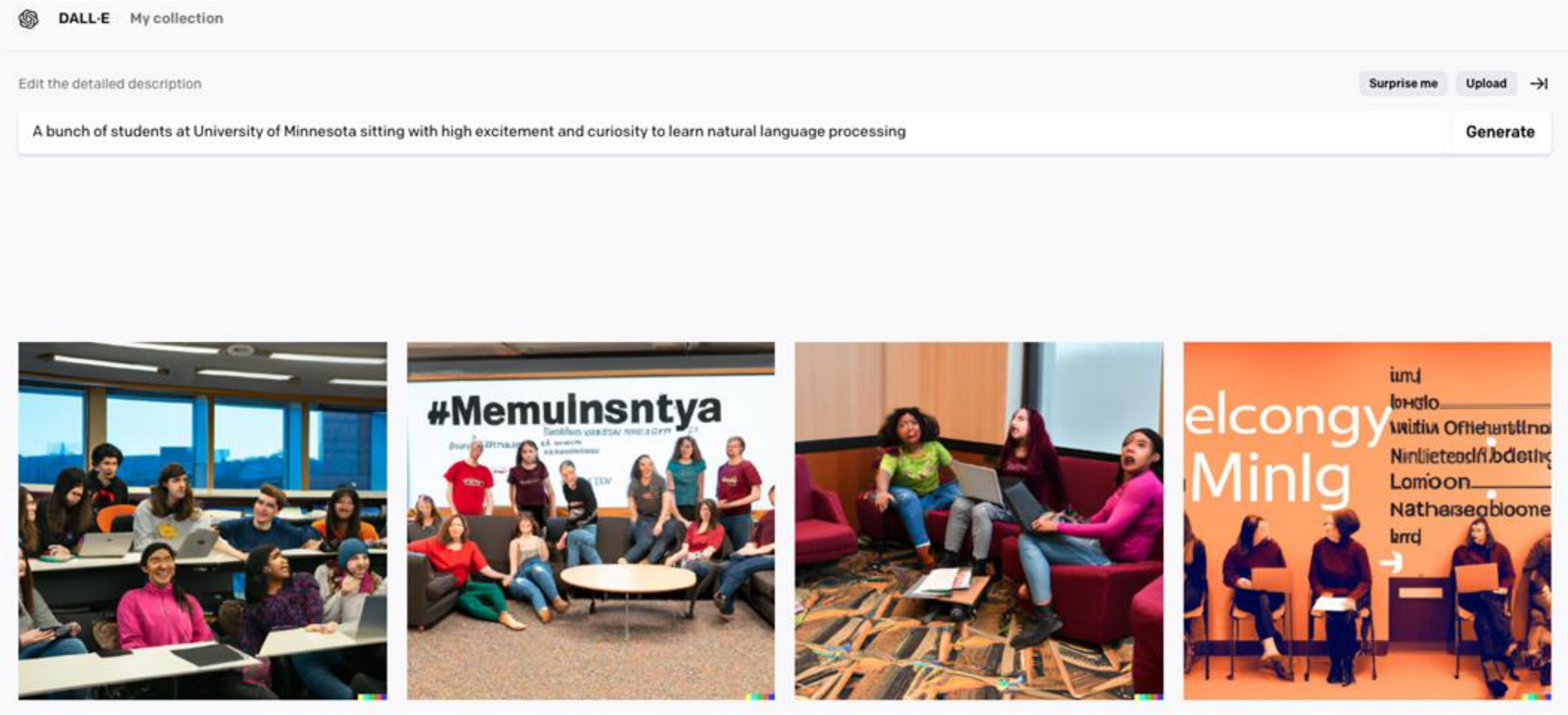
Jeff Dean https://ai.googleblog.com/2023/01/google-research-2022-beyond-language.html

Scaling Law in Vision-Language Model

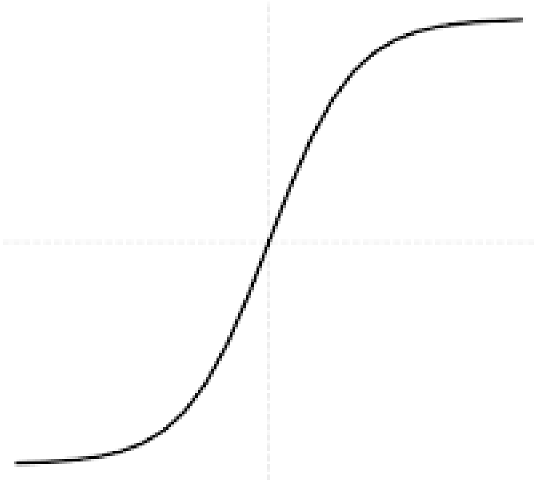
Figure 4. The generated image for the text "*A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!*". *Note the model gets the text in the image "welcome friends" correct at 20B.*

https://towardsdatascience.com/a-quiet-shift-in-the-nlp-ecosystem-84672b8ec7af

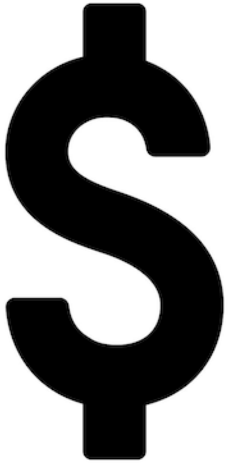
Beyond Language

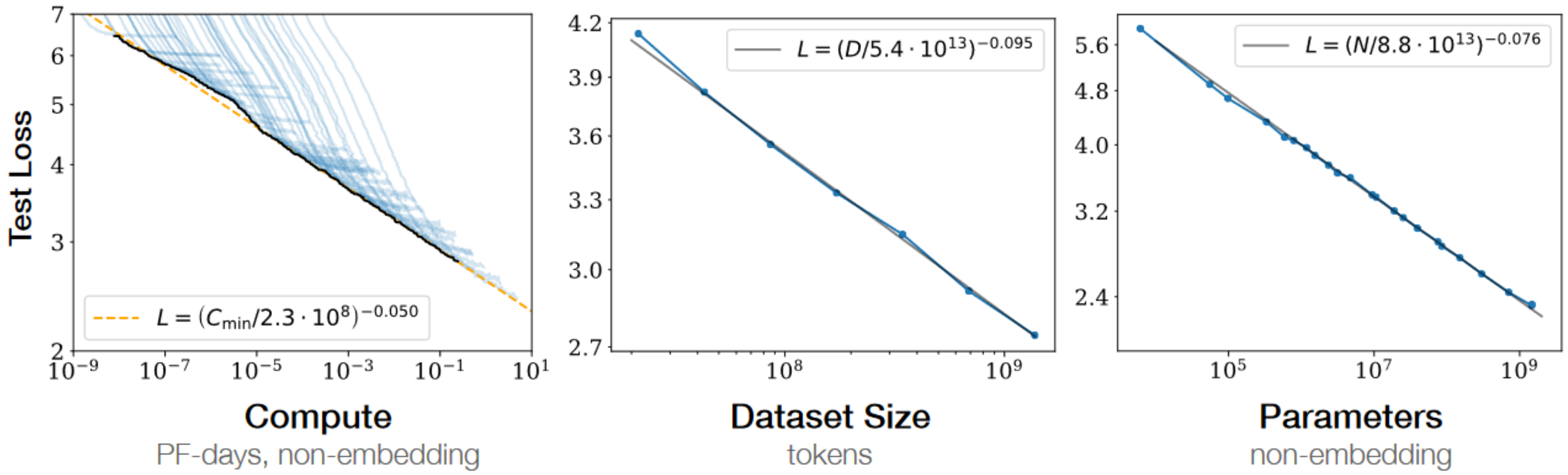


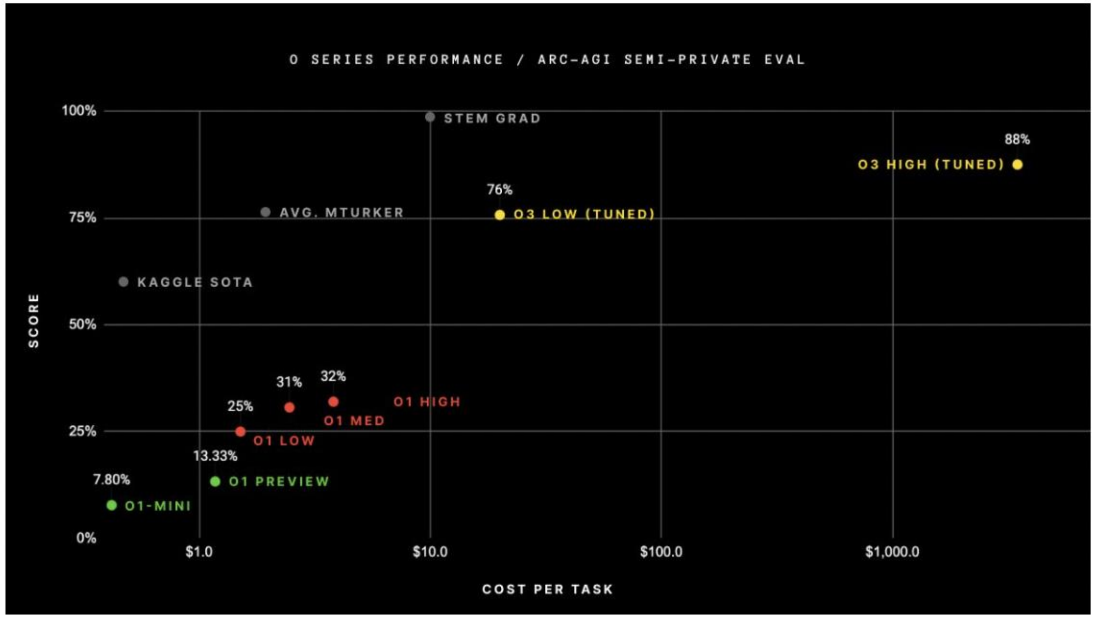
Limits of LLMs and the Financial Incentives of GenAI



Limits of scaling







Falling Short

- Benchmarks saturate rapidly, but this does not lead to immediate capabilities on tasks we would like to automate outside the scope of those benchmarks
- How much are current benchmarks serving as a proxy for getting an enormous number of intelligent people to stuff as much insight into the models as possible (either with good training environments in RL settings, or large datasets of reasoning over certain problem areas)?
- How can we design better benchmarks which indicate that supposedly ‘PhD level’ models are capable of quickly doing the kinds of basic work that we actually care about

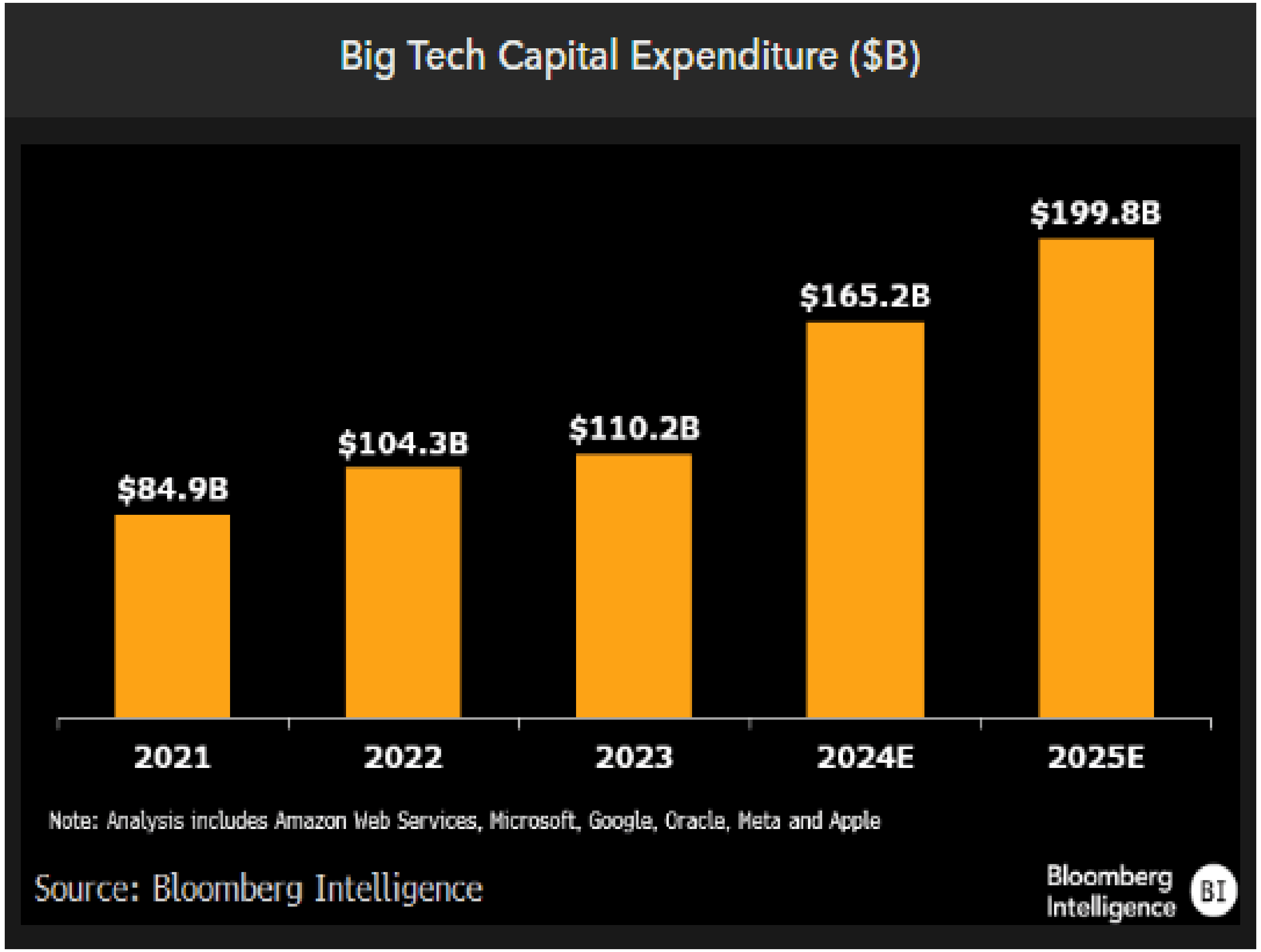
Andrej Karpathy
@karpathy

It’s done because it’s much easier to 1) collect, 2) evaluate, and 3) beat and make progress on. We’re going to see every task that is served neatly packaged on a platter like this improved (including those that need PhD-grade expertise). But jobs (even intern-level) that need long, multimodal, coherent, error-correcting sequences of tasks glued together for problem solving will take longer. They are unintuitively hard, in a Moravec’s Paradox sense.

Fwiw I’m ok and happy to see harder “task” evals. Calling it humanity’s last exam is a bit much, and misleading.

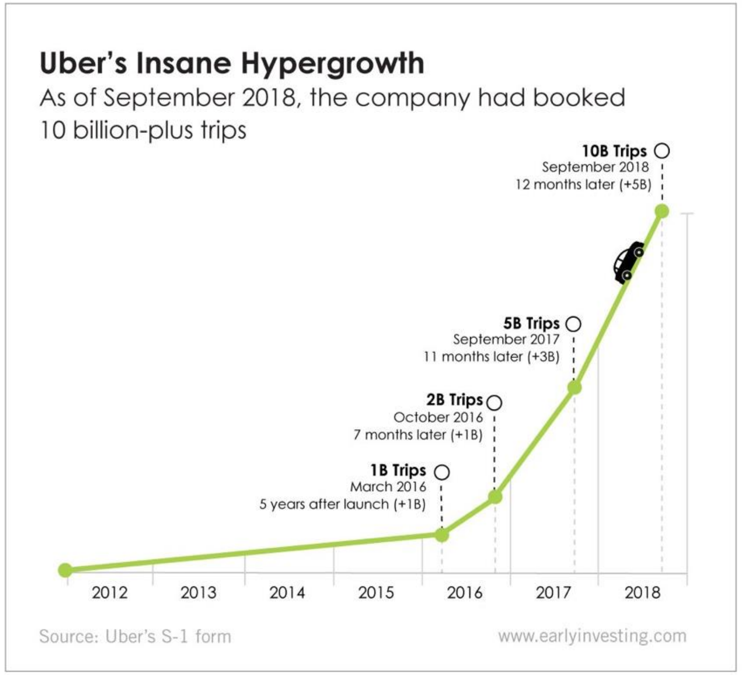
> **Niels Rogge** @NielsRogge · 4h
>
> Unpopular opinion: benchmarks like these are moving the field in the wrong direction
>
> No I don't want an AI to be able to memorize (useless?) questions like "How many paired tendons are supported by a sesamoid bone?" in its weights...
> Show more

AI “arms race” by Big Tech

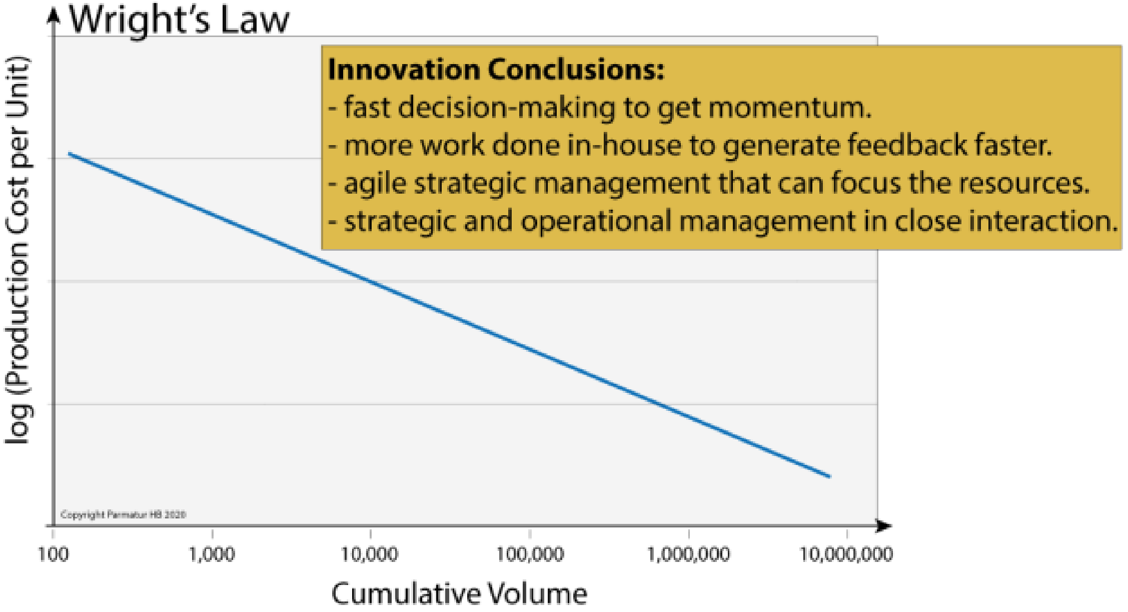


https://www.bloomberg.com/professional/insights/technology/big-tech-2025-capex-may-hit-200-billion-as-gen-ai-demand-booms/

Growth Economics and High CapEx



Demand-side Advantages



Supply Side Advantages

The Game Theory of the AI Arms Race

		Cloud 1: Escalate	Cloud 1: Don't Escalate
Cloud 2	Escalate	Overbuilding	Cloud 2 Wins
	Don't Escalate	Cloud 1 Wins	Equilibrium

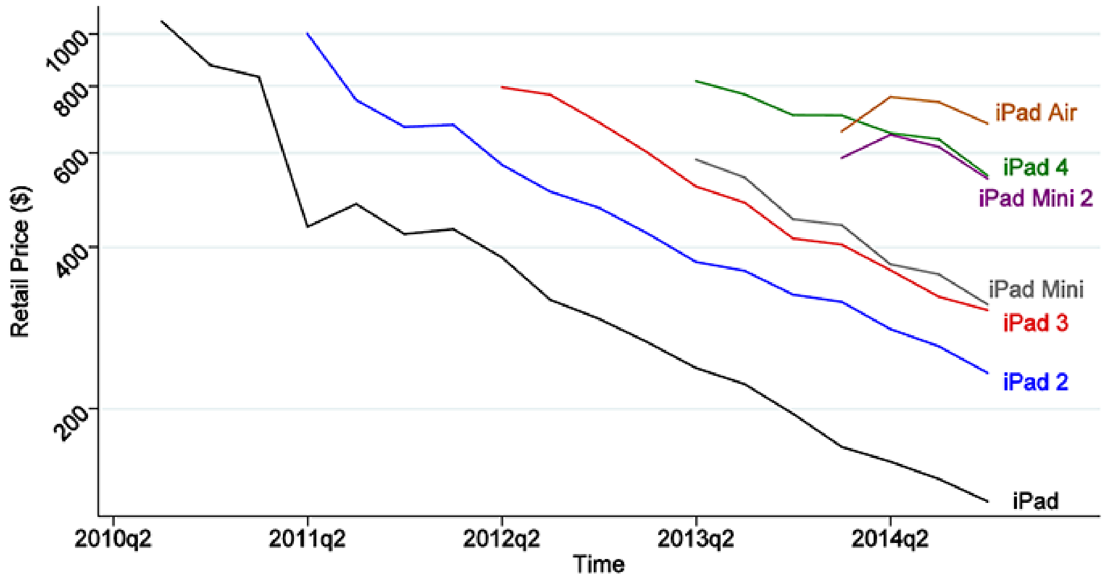
Rank	Name	Earnings
1	Saudi Aramco 2222.SR	$216.88 B
2	Berkshire Hathaway BRK-B	$138.32 B
3	Apple AAPL	$123.21 B
4	Alphabet (Google) GOOG	$112.26 B
5	Microsoft MSFT	$110.77 B
6	NVIDIA NVDA	$73.16 B
7	JPMorgan Chase JPM	$69.03 B
8	Meta Platforms (Facebook) META	$64.51 B
9	Amazon AMZN	$62.50 B

https://www.sequoiacap.com/article/ai-optimism-vs-ai-arms-race/

What if They are Wrong?

(it will sting, but they will probably be fine…unless margin increases)

- Computer hardware depreciates very rapidly (typically ~50% every 2-3 years)
- This means revenues must be recovered from high spend very fast in order to compensate for this loss



Summary

- NLP is interdisciplinary
- Language consists of many levels of structure:
  - Phonology, syntax, semantics, discourse, pragmatics
- Processing language is difficult, due to
  - ambiguity, scales, sparsity, variation, implication, and representation
- Development of NLP models and representations grows rapidly
  - From rules to feature learning to RNNs to Transformers
- "Large" language models
  - Generalist AI or AGI via prompting and chat
  - Scaling law
  - Multimodal
  - Limitations? Future directions?