Applied Natural Language Processing

Info 256

Lecture 16: Prompting LLMs 2 (Oct 18, 2023)

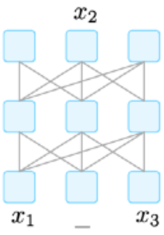
David Bamman, UC Berkeley

Language model

- Language models allow us to calculate the probability of the next word conditioned on some context (and different models make different assumptions about how much of that context is available).

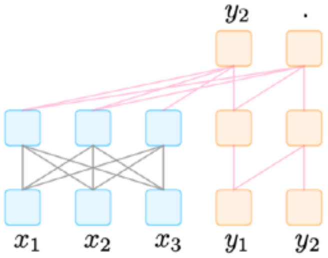
$$P(x_i \mid x_1, \ldots, x_{i-1})$$

- Even BERT can be used this way (by masking out the final word in a sequence)



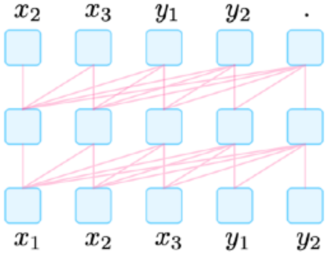
Masked LM
(BERT)

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$



Encoder-decoder
(T5)

$$P(y) = \prod_{i=1}^{n} P(y_i \mid y_1, \ldots, y_{i-1}, x)$$



Left-to-right LM
(GPT)

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

Liu et al. 2021

T5

- Encoder-decoder model pre-trained on 750GB of English web text by masking tokens in the input and predicting sequences of them in the output.

Thank you ~~for inviting~~ me to your party ~~last~~ week

encoder
decoder

Thank you [X] me to your party [Y] week → [X] for inviting [Y] last [Z]

GPT

- Transformer-based causal (left-to-right) language model:

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

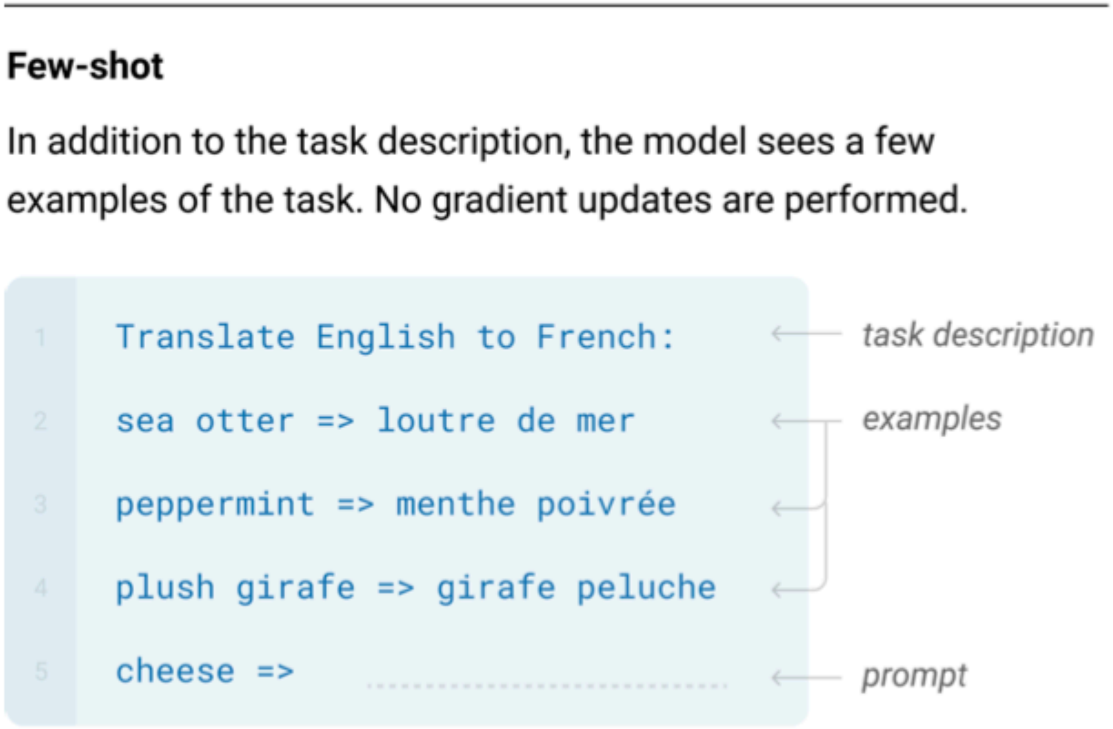
	Model	Data
GPT-2 (Radford et al. 2019)	Context size: 1024 tokens 117M-1.5B parameters	WebText (45 million outbound links from Reddit with 3+ karma); 8 million documents (40GB)
GPT-3 (Brown et al. 2020)	Context size: 2048 tokens 125M-175B parameters	Common crawl + WebText + “two internet-based books corpora” + Wikipedia (400B tokens, 570GB)

Everything is language modeling

The director of *2001: A Space Odyssey* is _____________

The French translation of “cheese” is _____________

The sentiment of “I really hate this movie” is ____________



Few-shot

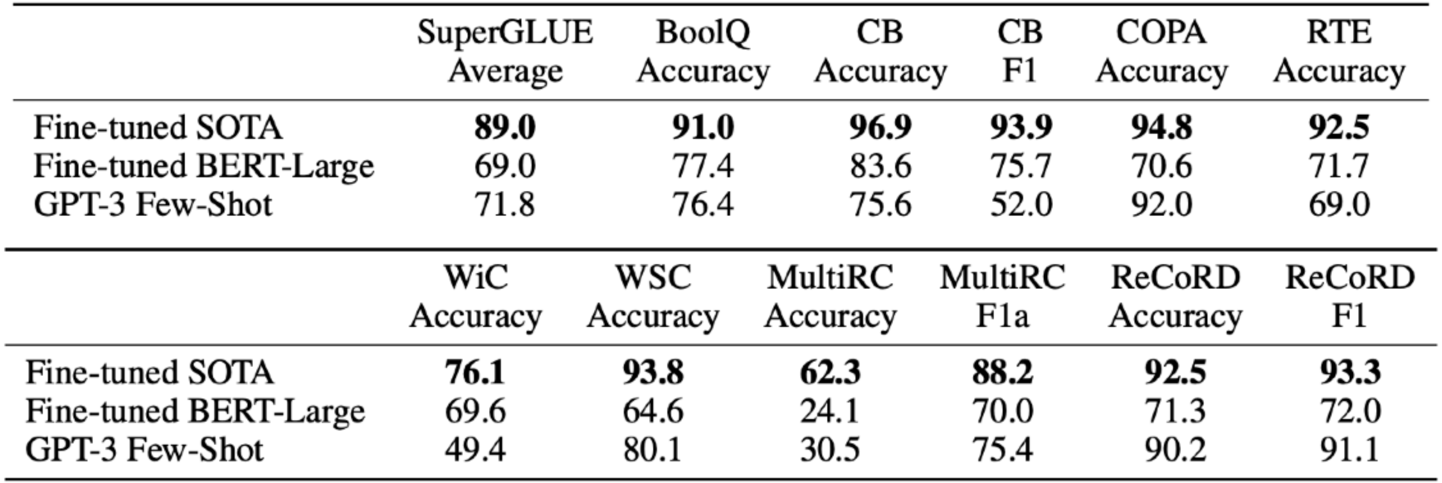
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

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

Brown et al. (2020, "Language Models are Few-Shot Learners"
https://arxiv.org/pdf/2005.14165.pdf

Causal reasoning

Textual entailment



	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

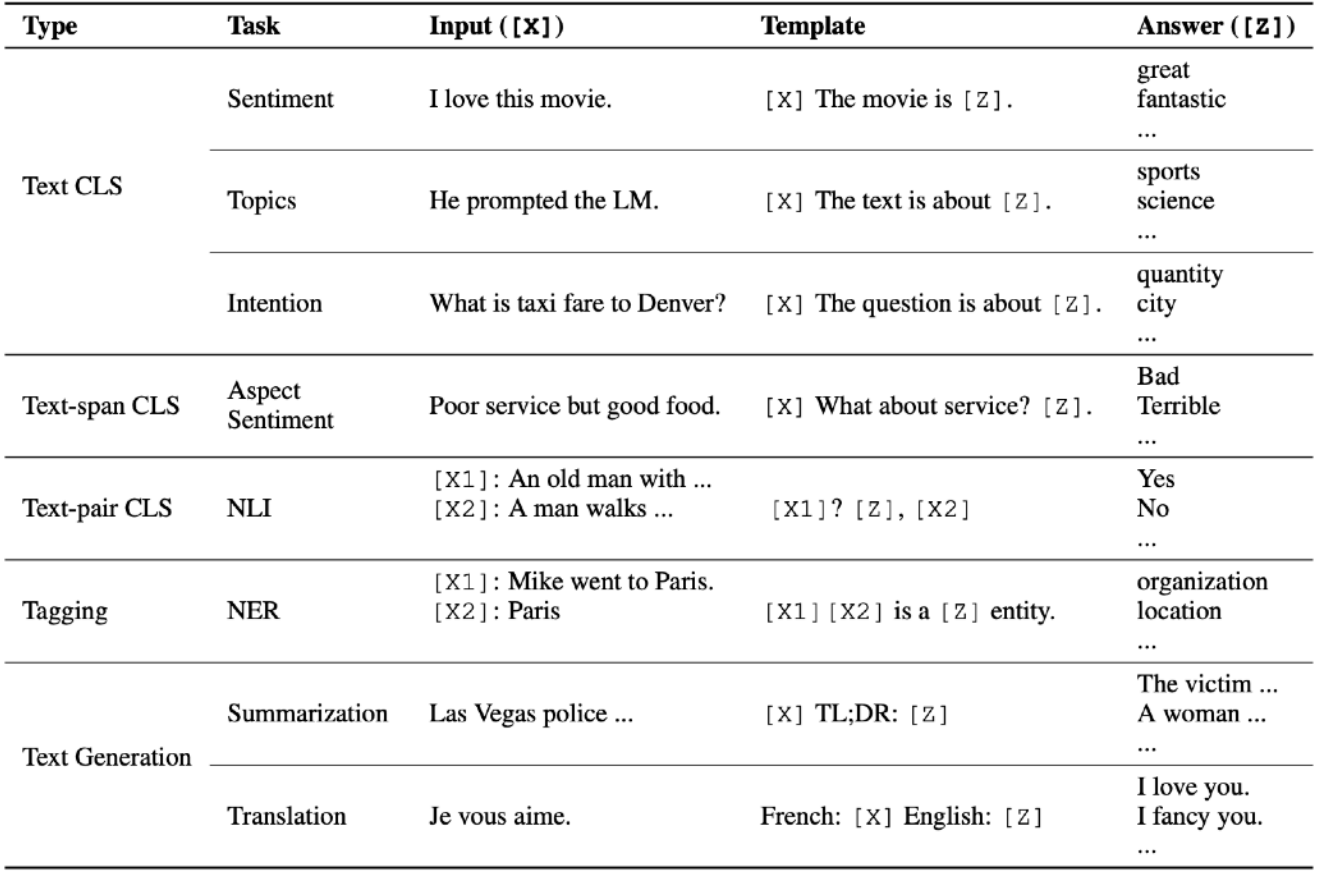
Word sense disambiguation

Question answering

Brown et al. (2020), “Language Models are Few-Shot Learners”

Prompt engineering

- Manual prompt design: encoding domain knowledge into prompt templates that are likely to generate a response in the output space.

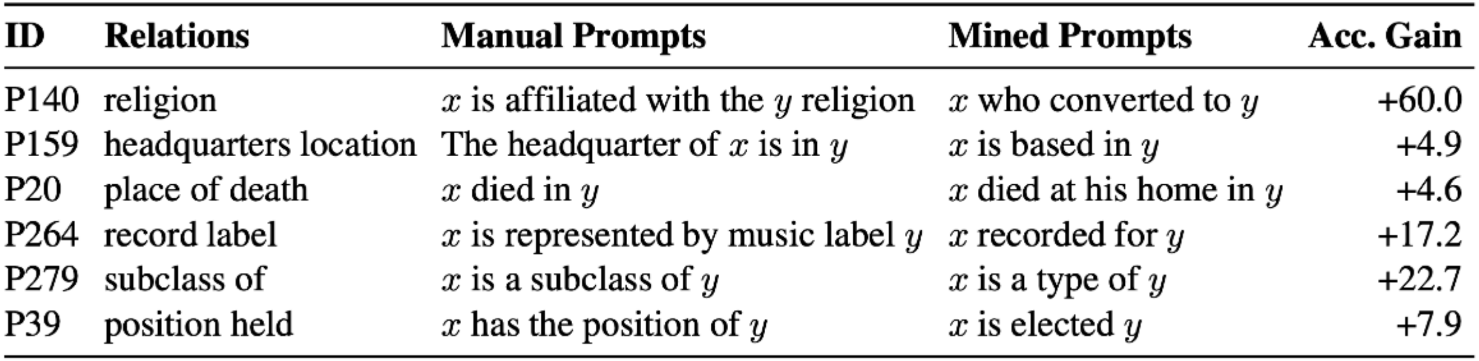


Type	Task	Input (`[X]`)	Template	Answer (`[Z]`)
Text CLS	Sentiment	I love this movie.	`[X]` The movie is `[Z]`.	great fantastic ...
	Topics	He prompted the LM.	`[X]` The text is about `[Z]`.	sports science ...
	Intention	What is taxi fare to Denver?	`[X]` The question is about `[Z]`.	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	`[X]` What about service? `[Z]`.	Bad Terrible ...
Text-pair CLS	NLI	`[X1]`: An old man with ... `[X2]`: A man walks ...	`[X1]`? `[Z]`, `[X2]`	Yes No ...
Tagging	NER	`[X1]`: Mike went to Paris. `[X2]`: Paris	`[X1]` `[X2]` is a `[Z]` entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	`[X]` TL;DR: `[Z]`	The victim ... A woman ... ...
	Translation	Je vous aime.	French: `[X]` English: `[Z]`	I love you. I fancy you. ...

Liu et al. 2021

Prompt engineering

- Prompt mining: rather than manually writing prompts, learning high-performing prompts from input/output pairs in training data (e.g., labeled classification/relation extraction examples).

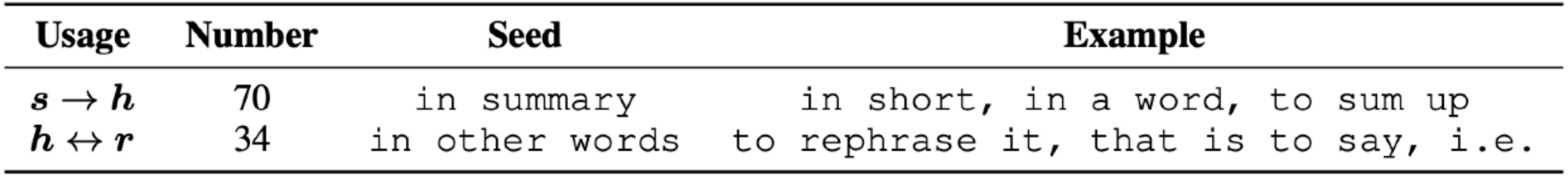


ID	Relations	Manual Prompts	Mined Prompts	Acc. Gain
P140	religion	x is affiliated with the y religion	x who converted to y	+60.0
P159	headquarters location	The headquarter of x is in y	x is based in y	+4.9
P20	place of death	x died in y	x died at his home in y	+4.6
P264	record label	x is represented by music label y	x recorded for y	+17.2
P279	subclass of	x is a subclass of y	x is a type of y	+22.7
P39	position held	x has the position of y	x is elected y	+7.9

Jiang et al. 2020, “How Can We Know What Language Models Know?”

Prompt engineering

- Prompt paraphrasing: automatically generate paraphrases of a manual prompt, and see which ones perform best on evaluation data.

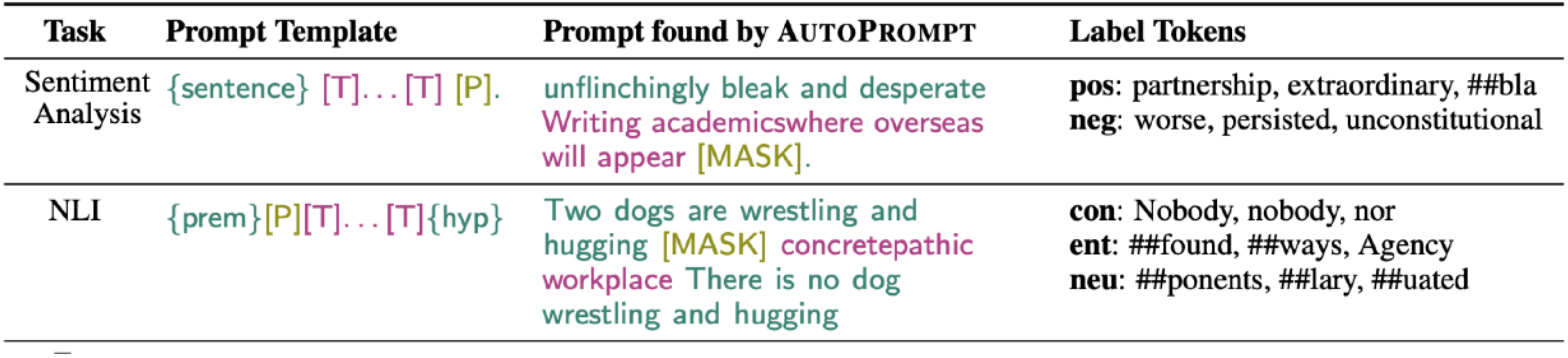


Usage	Number	Seed	Example
$s \rightarrow h$	70	in summary	in short, in a word, to sum up
$h \leftrightarrow r$	34	in other words	to rephrase it, that is to say, i.e.

Yuan et al. 2021, “BARTSCORE: Evaluating Generated Text as Text Generation”

Prompt engineering

- Prompt optimization: given training data in the form of input/output pairs, learn the prompts (and output labels) that maximize the probability of that training data.



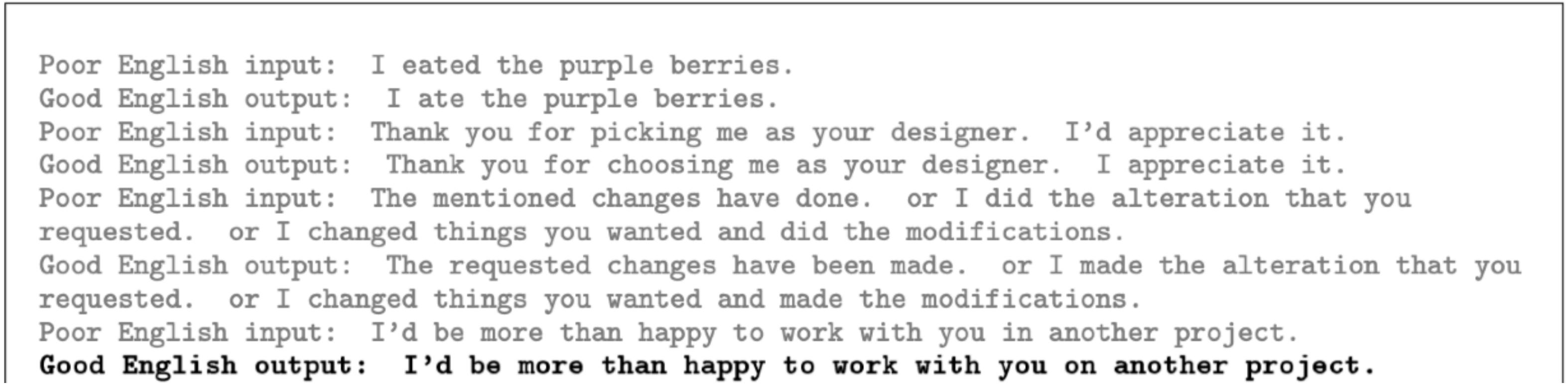
Task	Prompt Template	Prompt found by AUTOPROMPT	Label Tokens
Sentiment Analysis	{sentence} [T]. . . [T] [P].	unflinchingly bleak and desperate Writing academicswhere overseas will appear [MASK].	**pos**: partnership, extraordinary, ##bla **neg**: worse, persisted, unconstitutional
NLI	{prem}[P][T]. . . [T]{hyp}	Two dogs are wrestling and hugging [MASK] concretepathic workplace There is no dog wrestling and hugging	**con**: Nobody, nobody, nor **ent**: ##found, ##ways, Agency **neu**: ##ponents, ##lary, ##uated

Shin et al. 2020, “AUTOPROMPT: Eliciting Knowledge from Language Models with Automatically Generated Prompts”

Prompt augmentation

- Providing several examples in the prompt context to illustrate the intended behavior.

Answered prompts



Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Answer engineering

X: This movie was amazing. Y: positive

great

excellent

fantastic

amazing

Language models

- Remember that these are all still language models that let us calculate the probability of a term (or sequence) conditioned on some context.

$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

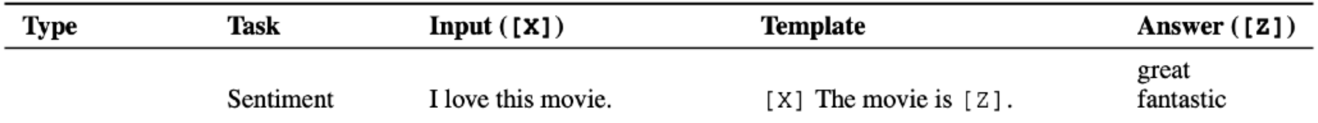
Answer engineering

- For classification with a discrete output space,
- E.g., classification with output space = {positive, negative, neutral} and input prompt “X: This movie was amazing. Y:”

$$\text{argmax} \begin{cases} P_{GPT-3}(w_n = \text{positive} \mid w_{1,\ldots,n-1} = \text{"X: This movie was amazing. Y:"}) \\ P_{GPT-3}(w_n = \text{negative} \mid w_{1,\ldots,n-1} = \text{"X: This movie was amazing. Y:"}) \\ P_{GPT-3}(w_n = \text{neutral} \mid w_{1,\ldots,n-1} = \text{"X: This movie was amazing. Y:"}) \end{cases}$$

Answer engineering

- Answer mapping: create a dictionary of allowable generations Z (e.g., great, fantastic, amazing, awesome, terrible, bad, horrible) and then map them to output labels (great→positive, fantastic→positive, terrible→negative, bad→negative, horrible→negative).

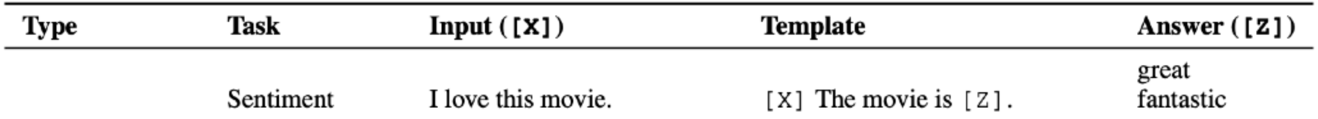


Type	**Task**	**Input (`[X]`)**	**Template**	**Answer (`[Z]`)**
	Sentiment	I love this movie.	`[X]` The movie is `[Z]`.	great fantastic

Liu et al. 2021

Answer engineering

- Answer paraphrasing: use a thesaurus to construct alternations of allowable generations (positive={great, amazing, awesome, good}) and calculate the probability of a class as the sum of the probability of all elements in the dictionary (Jiang et al. 2020)



Type	Task	Input ([X])	Template	Answer ([Z])
	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic

Liu et al. 2021

Documentation debt

- As Bender et al. 2021 notes, “documentation allows for accountability” and it’s often unclear what data these models are trained on (e.g., mysterious books1 and books2 corpora).

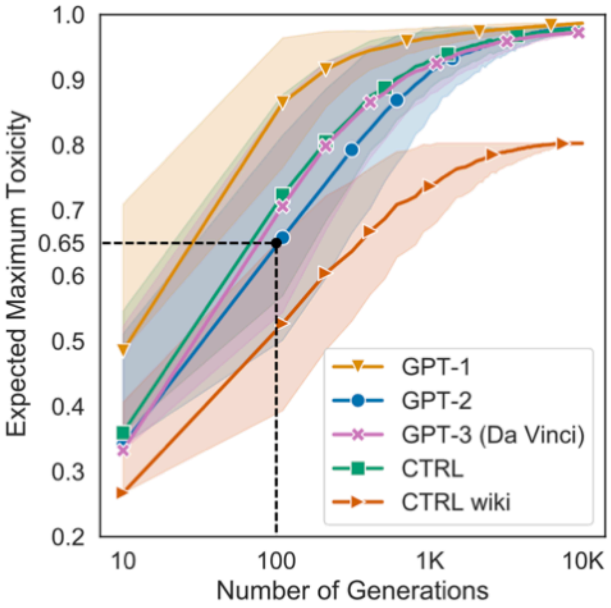
- When known, training data encodes narrow perspectives — e.g., links shared on Reddit; filtering out pages containing words related to sex (as C4 does) filters pornography but also positive sex discussions.

- Biases in training data can lead to representational harms [Kurita et al. 2019; Hutchinson et al. 2020; Gehman et al. 2020]

Bender et al. 2021, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”

Toxic generation

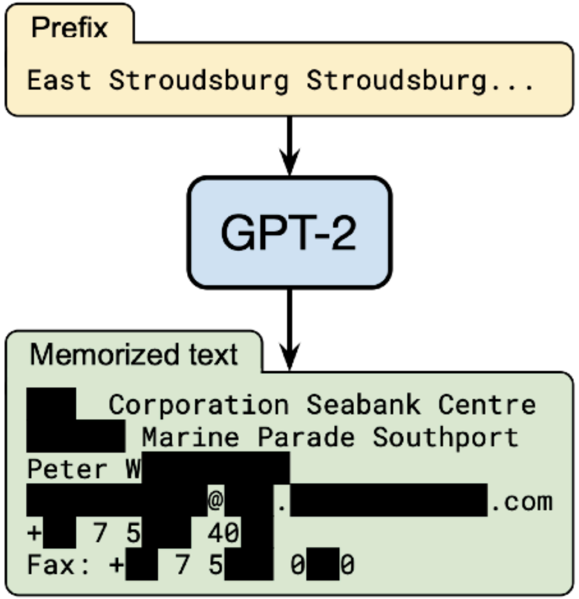
- Language models like GPT-{1,2,3} trained on toxic data (e.g., banned subreddits like /r/The_Donald or /r/WhiteRights) reproduce that toxicity in both prompted and unprompted generations



Sap et al. (2020), "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models"

Privacy

- Large language models (e.g,. GPT-3, BERT) can memorize training data, which is recoverable from it.
- Potential violations of confidential data (e.g., GMail messages) and contextual integrity (data being published in a way that violates a user's expectations of use).



Prefix

`East Stroudsburg Stroudsburg...`

GPT-2

Memorized text

```
[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

Carlini et al. (2020), "Extracting Training Data from Large Language Models"

Name cloze

At dawn, I lie in bed for a while, watching the sun come up on a beautiful morning. It's Sunday. A day off at home. I wonder if [MASK] is in the woods yet. Usually we devote all of Sunday to stocking up for the week.

→ Gale (Collins, *Hunger Games*)

- Sampled 100 passages from 491 books that contain a single proper name PER entity and no other named entities and assess how often a model gets it right
- Human performance: 0%
- Majority class ("Mary") = 0.6%

Chang et al. (2023), "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4"

GPT-4	ChatGPT	BERT	Year	Author	Title
0.98	0.82	0.00	1865	Lewis Carroll	Alice's Adventures in Wonderland
0.76	0.43	0.00	1997	J.K. Rowling	Harry Potter and the Sorcerer's Stone
0.74	0.29	0.00	1850	Nathaniel Hawthorne	The Scarlet Letter
0.72	0.11	0.00	1892	Arthur Conan Doyle	The Adventures of Sherlock Holmes
0.70	0.10	0.00	1815	Jane Austen	Emma
0.65	0.19	0.00	1823	Mary W. Shelley	Frankenstein
0.62	0.13	0.00	1813	Jane Austen	Pride and Prejudice
0.61	0.35	0.00	1884	Mark Twain	Adventures of Huckleberry Finn
0.61	0.30	0.00	1853	Herman Melville	Bartleby, the Scrivener
0.61	0.08	0.00	1897	Bram Stoker	Dracula
0.61	0.18	0.00	1838	Charles Dickens	Oliver Twist
0.59	0.13	0.00	1902	Arthur Conan Doyle	The Hound of the Baskervilles
0.59	0.22	0.00	1851	Herman Melville	Moby Dick; Or, The Whale
0.58	0.35	0.00	1876	Mark Twain	The Adventures of Tom Sawyer

GPT-4	ChatGPT	BERT	Year	Author	Title
0.76	0.43	0.00	1997	J.K. Rowling	Harry Potter and the Sorcerer's Stone
0.57	0.30	0.00	1949	George Orwell	1984
0.51	0.20	0.01	1954	J.R.R. Tolkien	The Fellowship of the Ring
0.49	0.16	0.13	2012	E.L. James	Fifty Shades of Grey
0.48	0.14	0.00	2008	Suzanne Collins	The Hunger Games
0.43	0.27	0.00	1954	William Golding	Lord of the Flies
0.43	0.17	0.00	1979	Douglas Adams	The Hitchhiker's Guide to the Galaxy
0.30	0.16	0.00	1959	Chinua Achebe	Things Fall Apart
0.28	0.12	0.00	1977	J. R. R. & C. Tolkien	The Silmarillion
0.27	0.13	0.00	1953	Ray Bradbury	Fahrenheit 451
0.27	0.13	0.00	1996	George R.R. Martin	A Game of Thrones
0.26	0.05	0.01	2003	Dan Brown	The Da Vinci Code
0.26	0.08	0.00	1965	Frank Herbert	Dune
0.25	0.20	0.01	1937	Zora Neale Hurston	Their Eyes Were Watching God
0.25	0.14	0.00	1961	Harper Lee	To Kill a Mockingbird

GPT-4	ChatGPT	BERT	Year	Author	Title
0.76	0.43	0.00	1997	J.K. Rowling	Harry Potter and the Sorcerer's Stone
0.57	0.30	0.00	1949	George Orwell	1984
0.51	0.20	0.01	1954	J.R.R. Tolkien	The Fellowship of the Ring
0.49	0.16	0.13	2012	E.L. James	Fifty Shades of Grey
0.48	0.14	0.00	2008	Suzanne Collins	The Hunger Games
0.43	0.27	0.00	1954	William Golding	Lord of the Flies
0.43	0.17	0.00	1979	Douglas Adams	The Hitchhiker's Guide to the Galaxy
0.30	0.16	0.00	1959	Chinua Achebe	Things Fall Apart
0.28	0.12	0.00	1977	J. R. R. & C. Tolkien	The Silmarillion
0.27	0.13	0.00	1953	Ray Bradbury	Fahrenheit 451
0.27	0.13	0.00	1996	George R.R. Martin	A Game of Thrones
0.26	0.05	0.01	2003	Dan Brown	The Da Vinci Code
0.26	0.08	0.00	1965	Frank Herbert	Dune
0.25	0.20	0.01	1937	Zora Neale Hurston	Their Eyes Were Watching God
0.25	0.14	0.00	1961	Harper Lee	To Kill a Mockingbird

Aligning Language Models

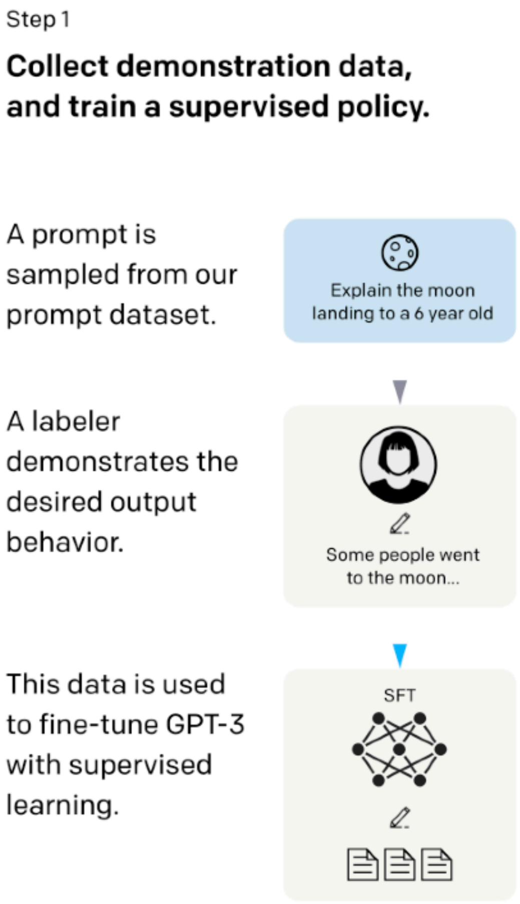
- All of the models we’ve discussed so far (BERT, GPT-*) are optimized to predict the probabilities of words—-not to encourage (or discourage) any specific kind of behavior.

Q: How many bonks are in a quoit?
A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?
A: It takes two rainbows to jump from Hawaii to seventeen.

Lacker (2020), "Giving GPT-3 a Turing Test” (https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html); Brown et al. (2020), “Language Models are Few-Shot Learners” (NeurIPS)

Instruct-GPT


Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

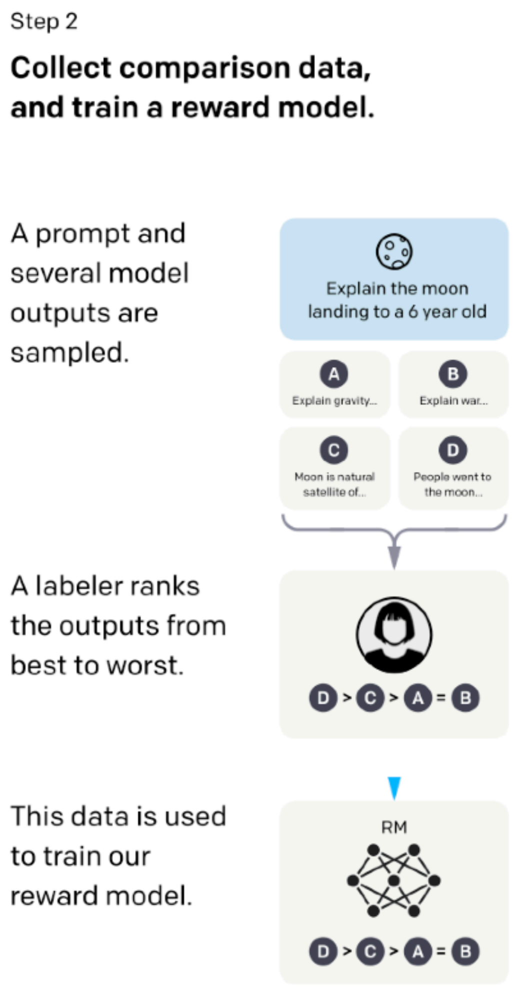
This data is used to fine-tune GPT-3 with supervised learning.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

- We can encourage specific behaviors by providing *feedback* to the model, in two forms:
- Given a prompt ("How many rainbows does it take to jump from Hawaii to seventeen?"), a human writes a preferred response ("A rainbow is not a unit of measurement."). **13K prompts/labels.**
- Fine-tune a pretrained model (e.g., GPT-3) on that supervised data to produce a supervised policy (a distribution over words to produce given a model state).

https://arxiv.org/pdf/2203.02155.pdf

Instruct-GPT



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

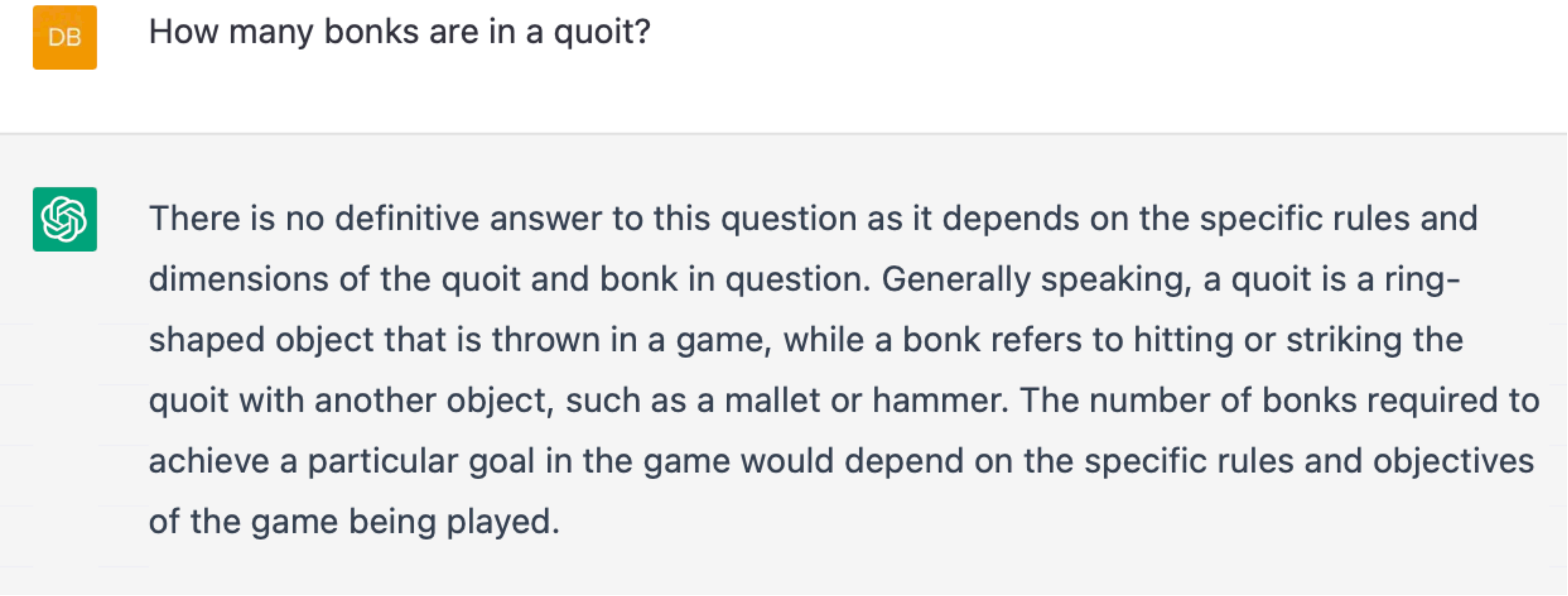
This data is used to train our reward model.

Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B

- We can encourage specific behaviors by providing *feedback* to the model, in two forms:
- Given a prompt and a set of model responses, a human ranks those responses from best to worst, defining a preference. **33K prompts + 4-9 outputs per prompt.**
- Train a reward model (using reinforcement learning) to score an output reflecting those learned ranks.

https://arxiv.org/pdf/2203.02155.pdf

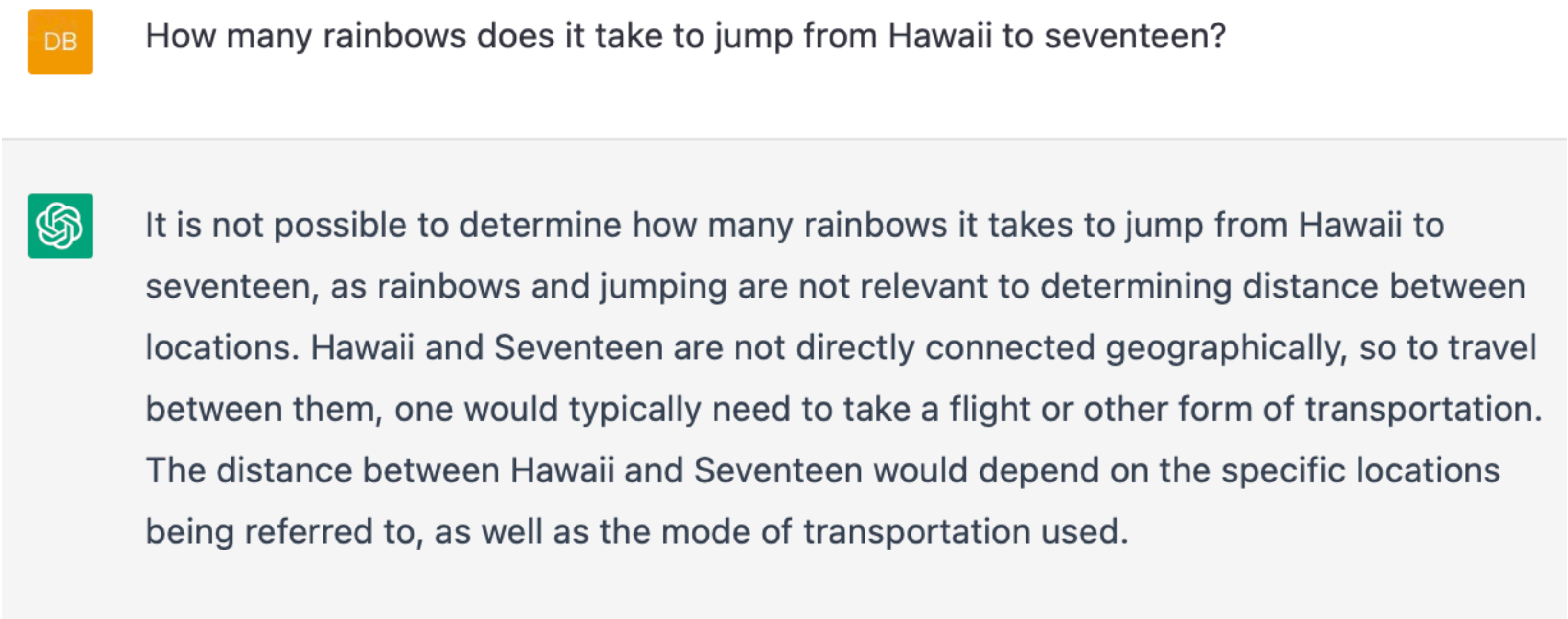
ChatGPT



How many bonks are in a quoit?

There is no definitive answer to this question as it depends on the specific rules and dimensions of the quoit and bonk in question. Generally speaking, a quoit is a ring-shaped object that is thrown in a game, while a bonk refers to hitting or striking the quoit with another object, such as a mallet or hammer. The number of bonks required to achieve a particular goal in the game would depend on the specific rules and objectives of the game being played.

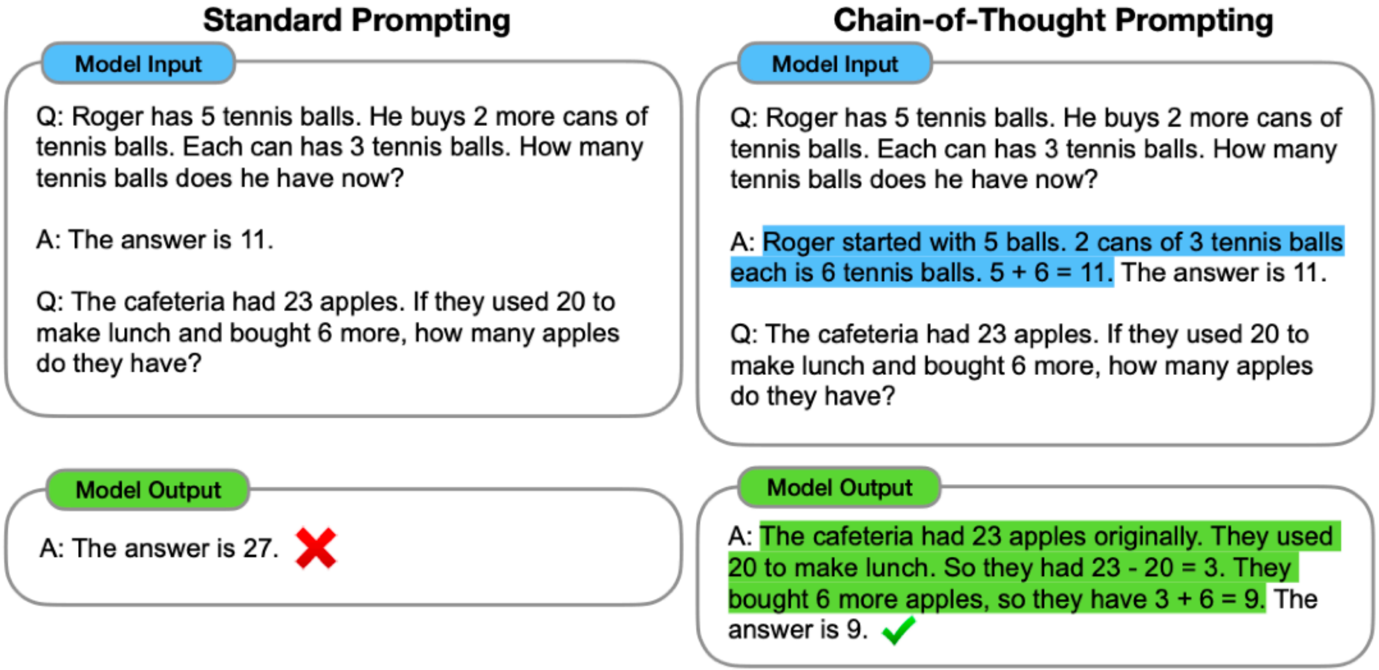
ChatGPT



How many rainbows does it take to jump from Hawaii to seventeen?

It is not possible to determine how many rainbows it takes to jump from Hawaii to seventeen, as rainbows and jumping are not relevant to determining distance between locations. Hawaii and Seventeen are not directly connected geographically, so to travel between them, one would typically need to take a flight or other form of transportation. The distance between Hawaii and Seventeen would depend on the specific locations being referred to, as well as the mode of transportation used.

Chain-of-thought



Standard Prompting

Model Input

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

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

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

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

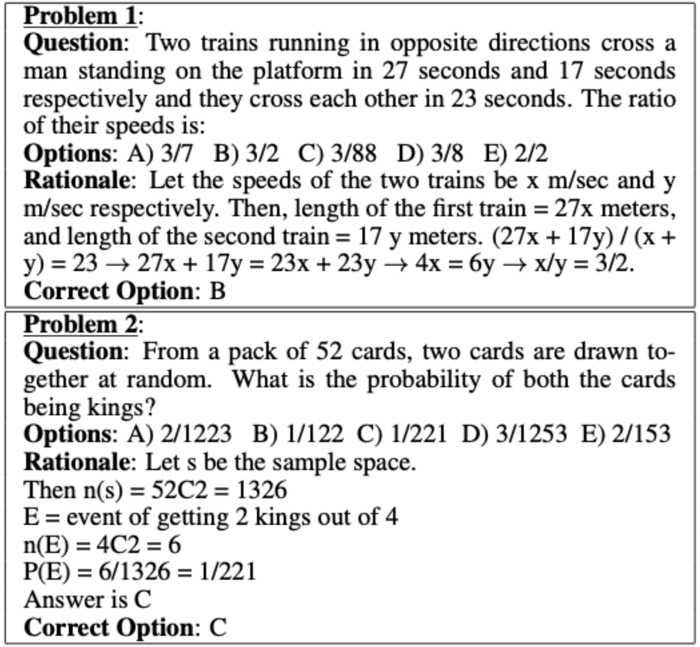
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Wei et al. 2022, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models"

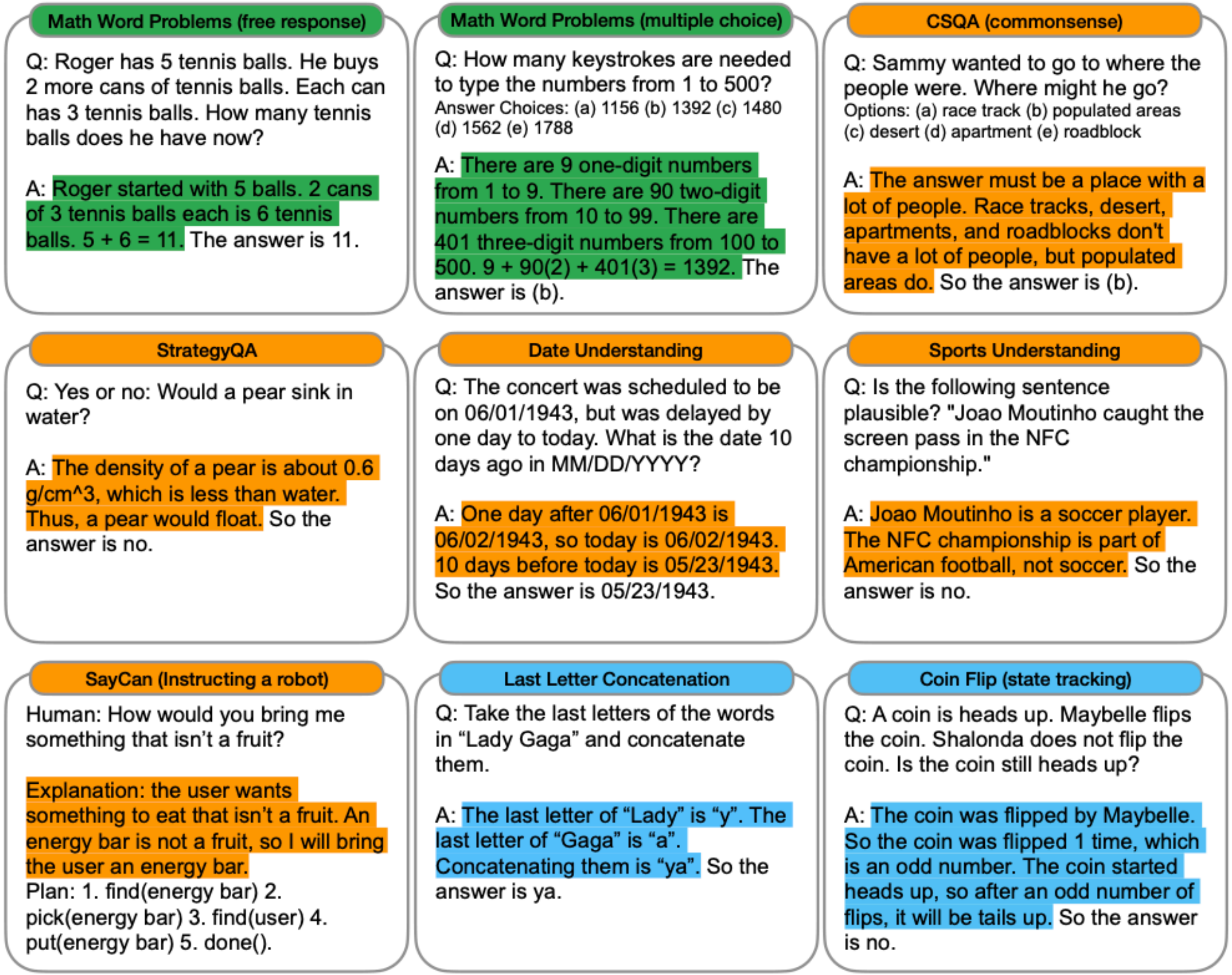
Rationales



Problem 1:
Question: Two trains running in opposite directions cross a man standing on the platform in 27 seconds and 17 seconds respectively and they cross each other in 23 seconds. The ratio of their speeds is:
Options: A) 3/7 B) 3/2 C) 3/88 D) 3/8 E) 2/2
Rationale: Let the speeds of the two trains be x m/sec and y m/sec respectively. Then, length of the first train = 27x meters, and length of the second train = 17 y meters. (27x + 17y) / (x + y) = 23 → 27x + 17y = 23x + 23y → 4x = 6y → x/y = 3/2.
Correct Option: B

Problem 2:
Question: From a pack of 52 cards, two cards are drawn together at random. What is the probability of both the cards being kings?
Options: A) 2/1223 B) 1/122 C) 1/221 D) 3/1253 E) 2/153
Rationale: Let s be the sample space.
Then n(s) = 52C2 = 1326
E = event of getting 2 kings out of 4
n(E) = 4C2 = 6
P(E) = 6/1326 = 1/221
Answer is C
Correct Option: C

Ling et al. 2017, "Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems"



Math Word Problems (free response)

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

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

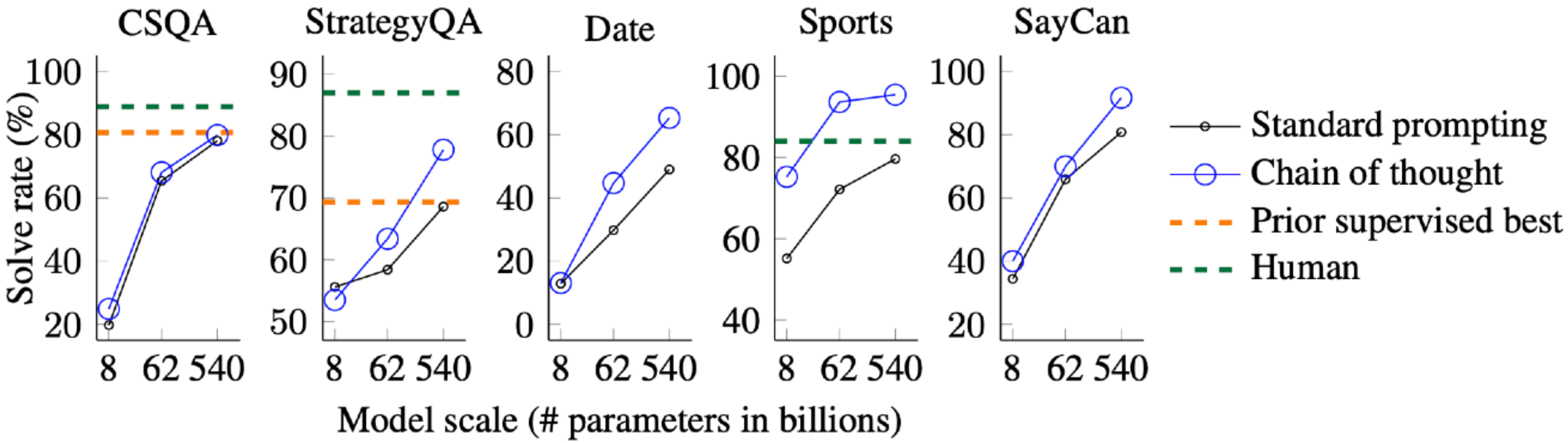
Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Wei et al. 2022, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models"

Chain-of-thought

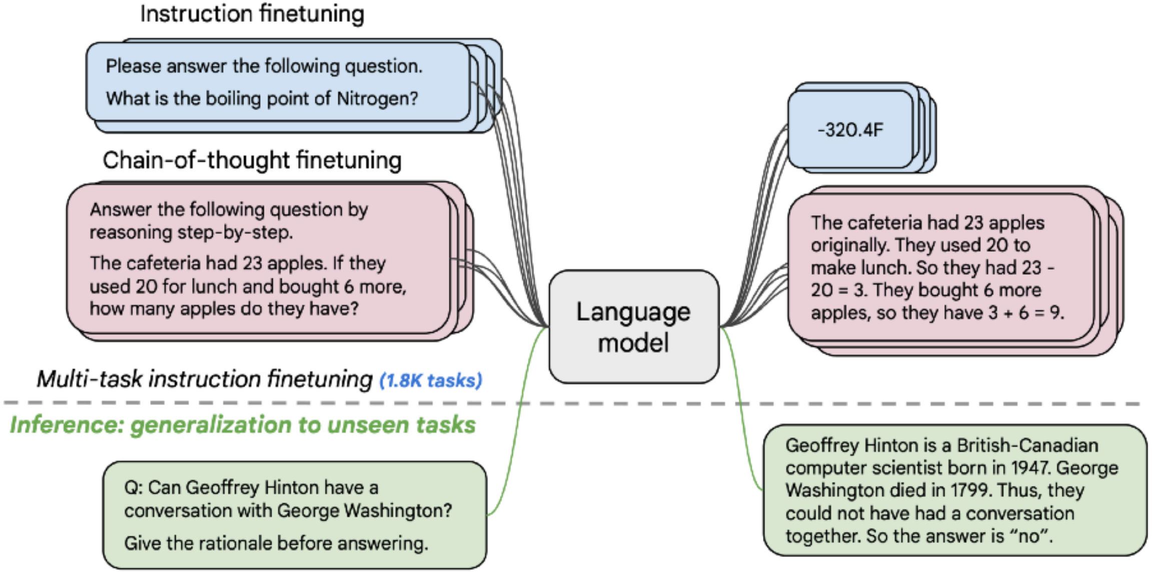


Wei et al. 2022, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models"

Instruction-finetuning

“Flan” = Finetuning language models

- Instruction-tune on many many tasks with diversity of data + problem type (1,836)
- With both CoT and non-CoT


Instruction finetuning
Chain-of-thought finetuning
Language model
Multi-task instruction finetuning (1.8K tasks)
Inference: generalization to unseen tasks

Please answer the following question.
What is the boiling point of Nitrogen?

-320.4F

Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is “no”.

Chung et al. 2022, “Scaling Instruction-Finetuned Language Models”

Finetuning tasks

T0-SF

Commonsense reasoning
Question generation
Closed-book QA
Adversarial QA
Extractive QA
Title/context generation
Topic classification
Struct-to-text
...

55 Datasets, 14 Categories, 193 Tasks

Muffin

Natural language inference
Code instruction gen.
Program synthesis
Dialog context generation
Closed-book QA
Conversational QA
Code repair
...

69 Datasets, 27 Categories, 80 Tasks

CoT (Reasoning)

Arithmetic reasoning
Commonsense Reasoning
Implicit reasoning
Explanation generation
Sentence composition
...

9 Datasets, 1 Category, 9 Tasks

Natural Instructions v2

Cause effect classification
Commonsense reasoning
Named entity recognition
Toxic language detection
Question answering
Question generation
Program execution
Text categorization
...

372 Datasets, 108 Categories, 1554 Tasks

- A **Dataset** is an original data source (e.g. SQuAD).
- A **Task Category** is unique task setup (e.g. the SQuAD dataset is configurable for multiple task categories such as extractive question answering, query generation, and context generation).
- A **Task** is a unique <dataset, task category> pair, with any number of templates which preserve the task category (e.g. query generation on the SQuAD dataset.)

Held-out tasks

MMLU

Abstract algebra
College medicine
Professional law
Sociology
Philosophy
...

57 tasks

BBH

Boolean expressions
Tracking shuffled objects
Dyck languages
Navigate
Word sorting
...

27 tasks

TyDiQA

Information seeking QA

8 languages

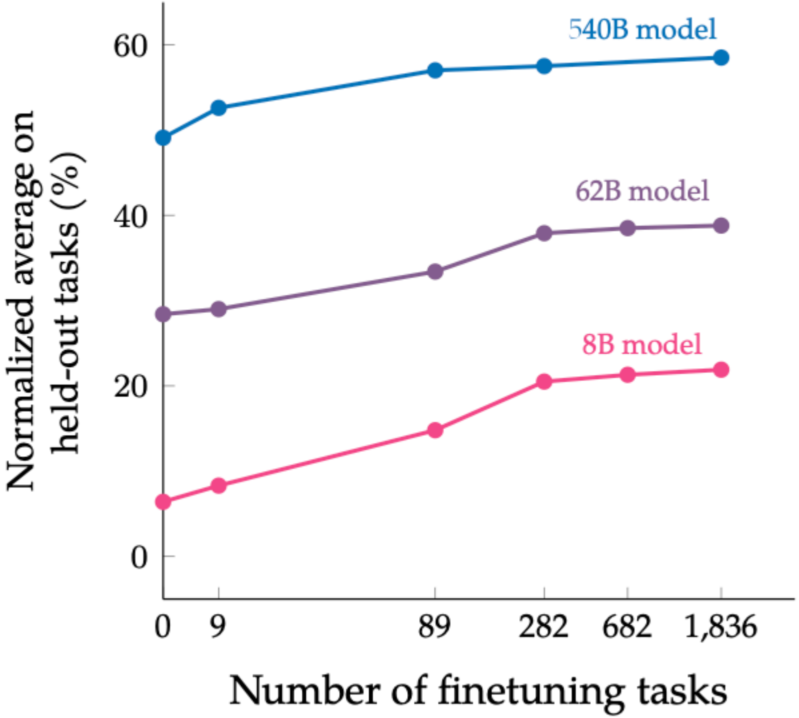
MGSM

Grade school math problems

10 languages

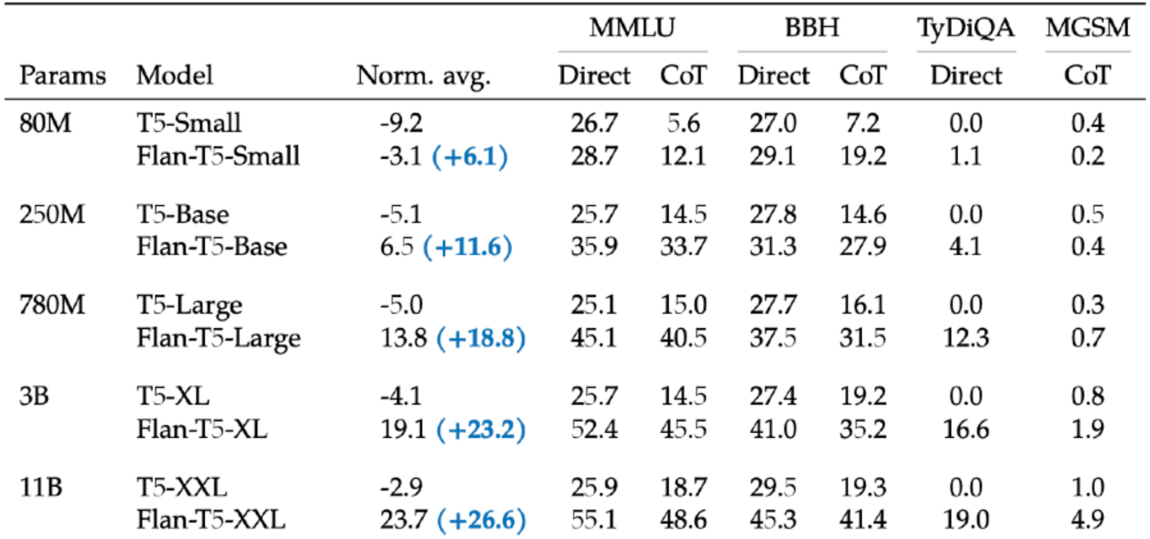
Chung et al. 2022, “Scaling Instruction-Finetuned Language Models”

Instruction-finetuning



Chung et al. 2022, “Scaling Instruction-Finetuned Language Models”

Instruction-finetuning

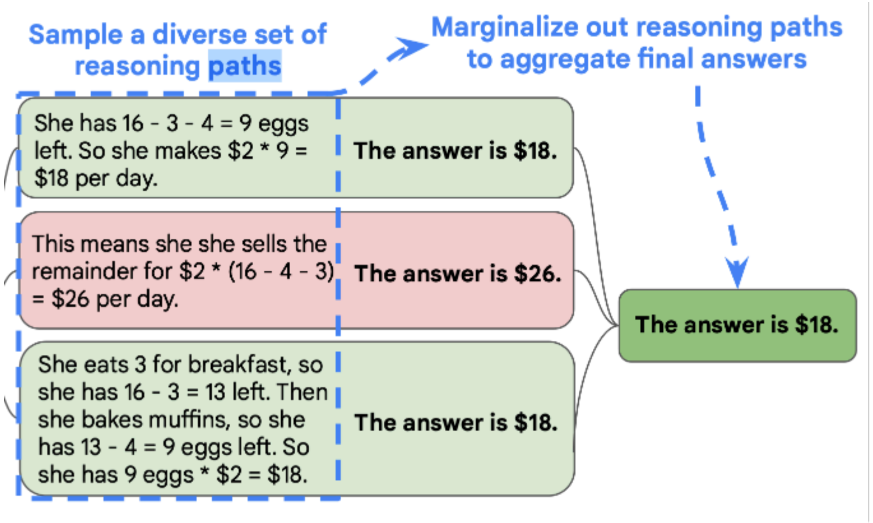


			MMLU		BBH		TyDiQA	MGSM
Params	Model	Norm. avg.	Direct	CoT	Direct	CoT	Direct	CoT
80M	T5-Small	-9.2	26.7	5.6	27.0	7.2	0.0	0.4
	Flan-T5-Small	-3.1 (+6.1)	28.7	12.1	29.1	19.2	1.1	0.2
250M	T5-Base	-5.1	25.7	14.5	27.8	14.6	0.0	0.5
	Flan-T5-Base	6.5 (+11.6)	35.9	33.7	31.3	27.9	4.1	0.4
780M	T5-Large	-5.0	25.1	15.0	27.7	16.1	0.0	0.3
	Flan-T5-Large	13.8 (+18.8)	45.1	40.5	37.5	31.5	12.3	0.7
3B	T5-XL	-4.1	25.7	14.5	27.4	19.2	0.0	0.8
	Flan-T5-XL	19.1 (+23.2)	52.4	45.5	41.0	35.2	16.6	1.9
11B	T5-XXL	-2.9	25.9	18.7	29.5	19.3	0.0	1.0
	Flan-T5-XXL	23.7 (+26.6)	55.1	48.6	45.3	41.4	19.0	4.9

Chung et al. 2022, “Scaling Instruction-Finetuned Language Models”

Self-consistency

- Sample multiple outputs with labels + CoT reasoning; select the answer with majority vote over samples.



Sample a diverse set of reasoning paths

Marginalize out reasoning paths to aggregate final answers

She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $18 per day. **The answer is $18.**

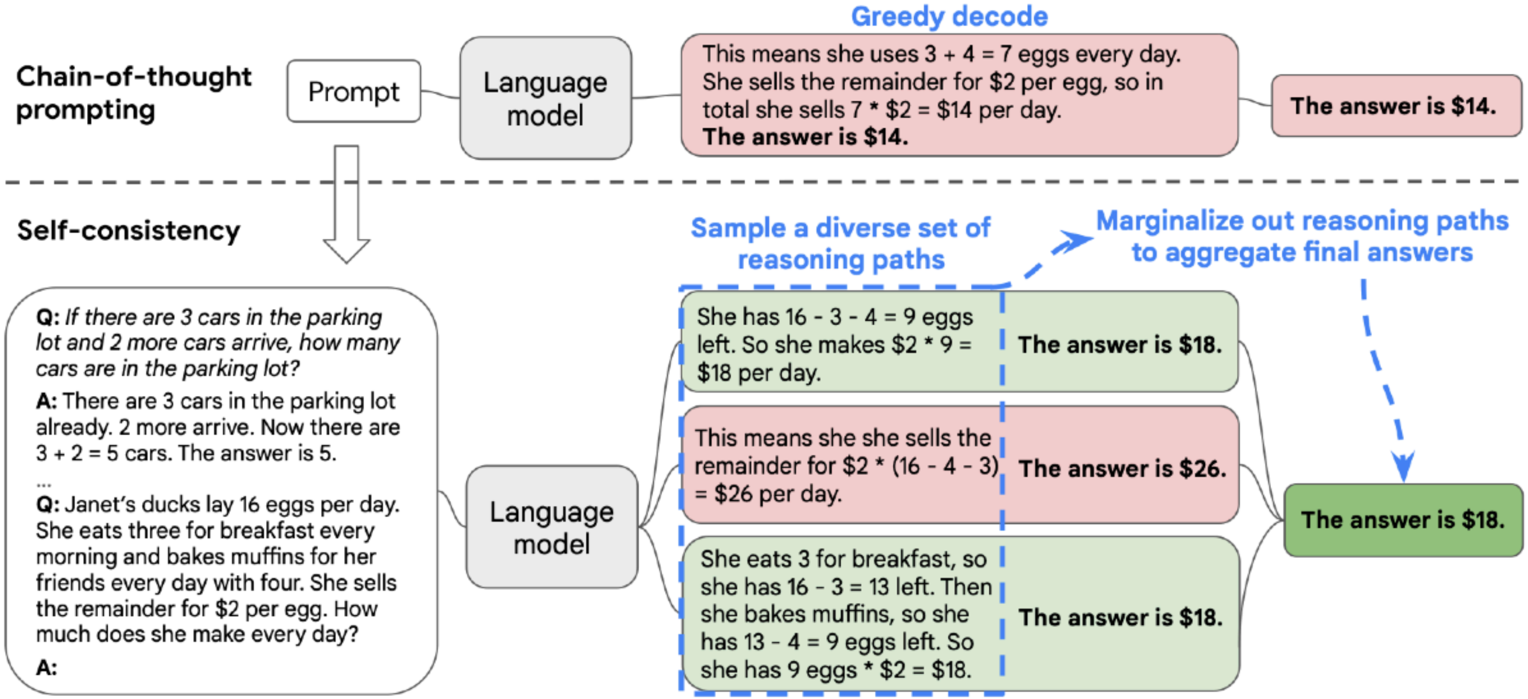
This means she she sells the remainder for $2 * (16 - 4 - 3) = $26 per day. **The answer is $26.**

She eats 3 for breakfast, so she has 16 - 3 = 13 left. Then she bakes muffins, so she has 13 - 4 = 9 eggs left. So she has 9 eggs * $2 = $18. **The answer is $18.**

The answer is $18.

Wang et al. 2023, "Self-Consistency Improves Chain of Thought Reasoning in Language Models”

Self-consistency



Chain-of-thought prompting

Prompt

Language model

Greedy decode

This means she uses 3 + 4 = 7 eggs every day.
She sells the remainder for $2 per egg, so in total she sells 7 * $2 = $14 per day.
The answer is $14.

The answer is $14.

Self-consistency

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

A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.

...

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for $2 per egg. How much does she make every day?

A:

Language model

Sample a diverse set of reasoning paths

She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $18 per day. — **The answer is $18.**

This means she she sells the remainder for $2 * (16 - 4 - 3) = $26 per day. — **The answer is $26.**

She eats 3 for breakfast, so she has 16 - 3 = 13 left. Then she bakes muffins, so she has 13 - 4 = 9 eggs left. So she has 9 eggs * $2 = $18. — **The answer is $18.**

Marginalize out reasoning paths to aggregate final answers

The answer is $18.

Wang et al. 2023, "Self-Consistency Improves Chain of Thought Reasoning in Language Models"

Self-consistency

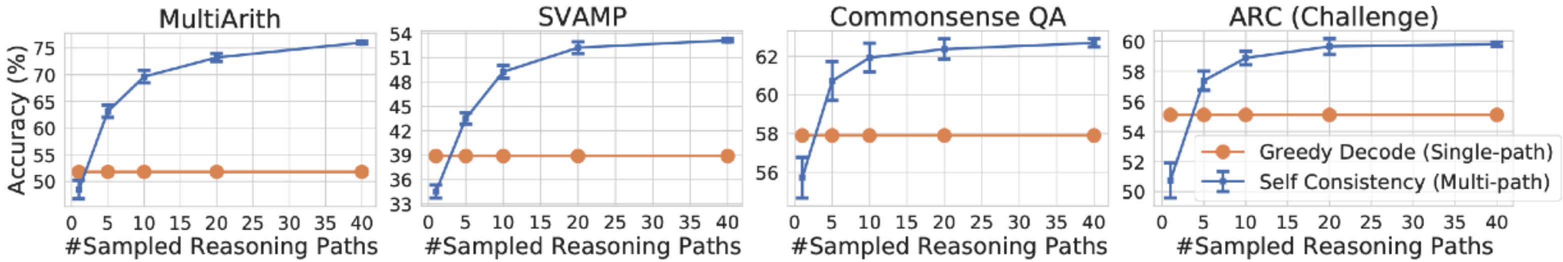


Figure 2: Self-consistency (blue) significantly improves accuracy over CoT-prompting with greedy decoding (orange) across arithmetic and commonsense reasoning tasks, over LaMDA-137B. Sampling a higher number of diverse reasoning paths consistently improves reasoning accuracy.

Wang et al. 2023, "Self-Consistency Improves Chain of Thought Reasoning in Language Models”

DEEP LEARNING FICTION

Using GPT-4 to measure the passage of time in fiction

Large language models are valuable research assistants, especially when they refuse to follow instructions.

By tedunderwood March 19, 2023 28 Comments

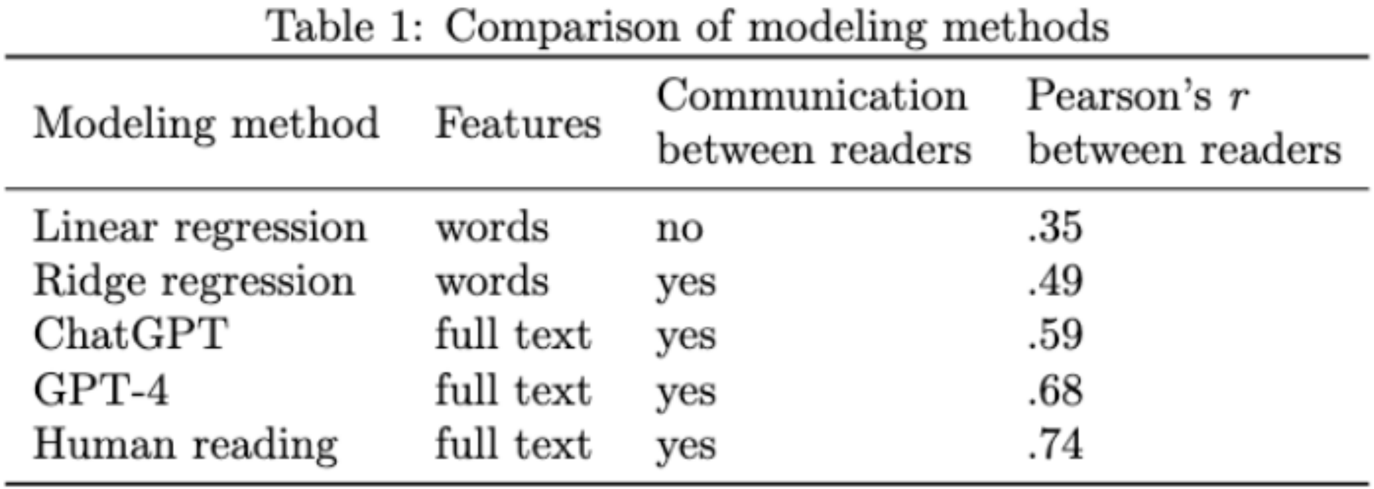


Table 1: Comparison of modeling methods

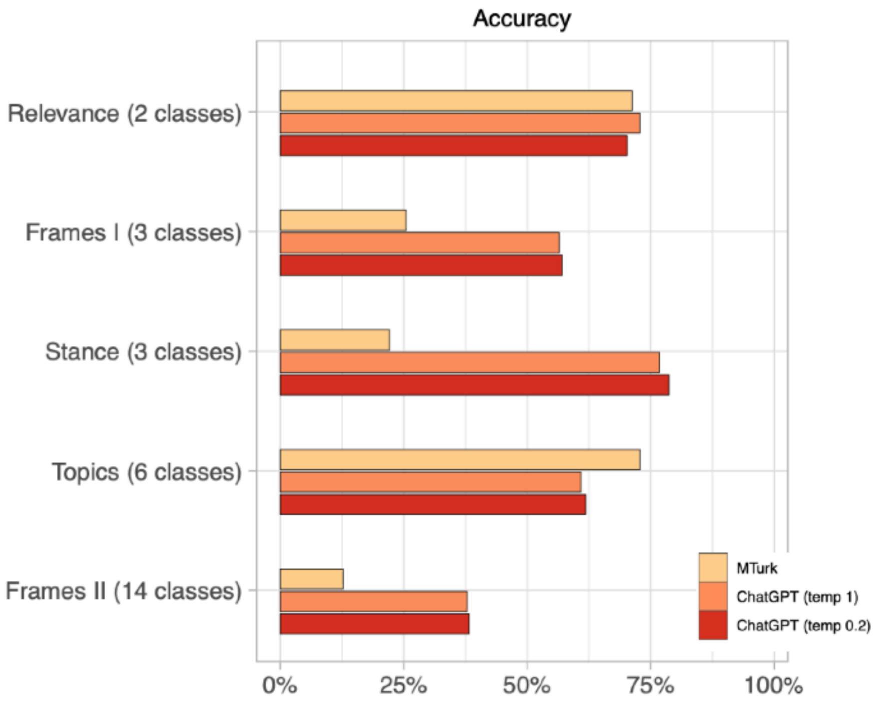
Modeling method	Features	Communication between readers	Pearson's r between readers
Linear regression	words	no	.35
Ridge regression	words	yes	.49
ChatGPT	full text	yes	.59
GPT-4	full text	yes	.68
Human reading	full text	yes	.74

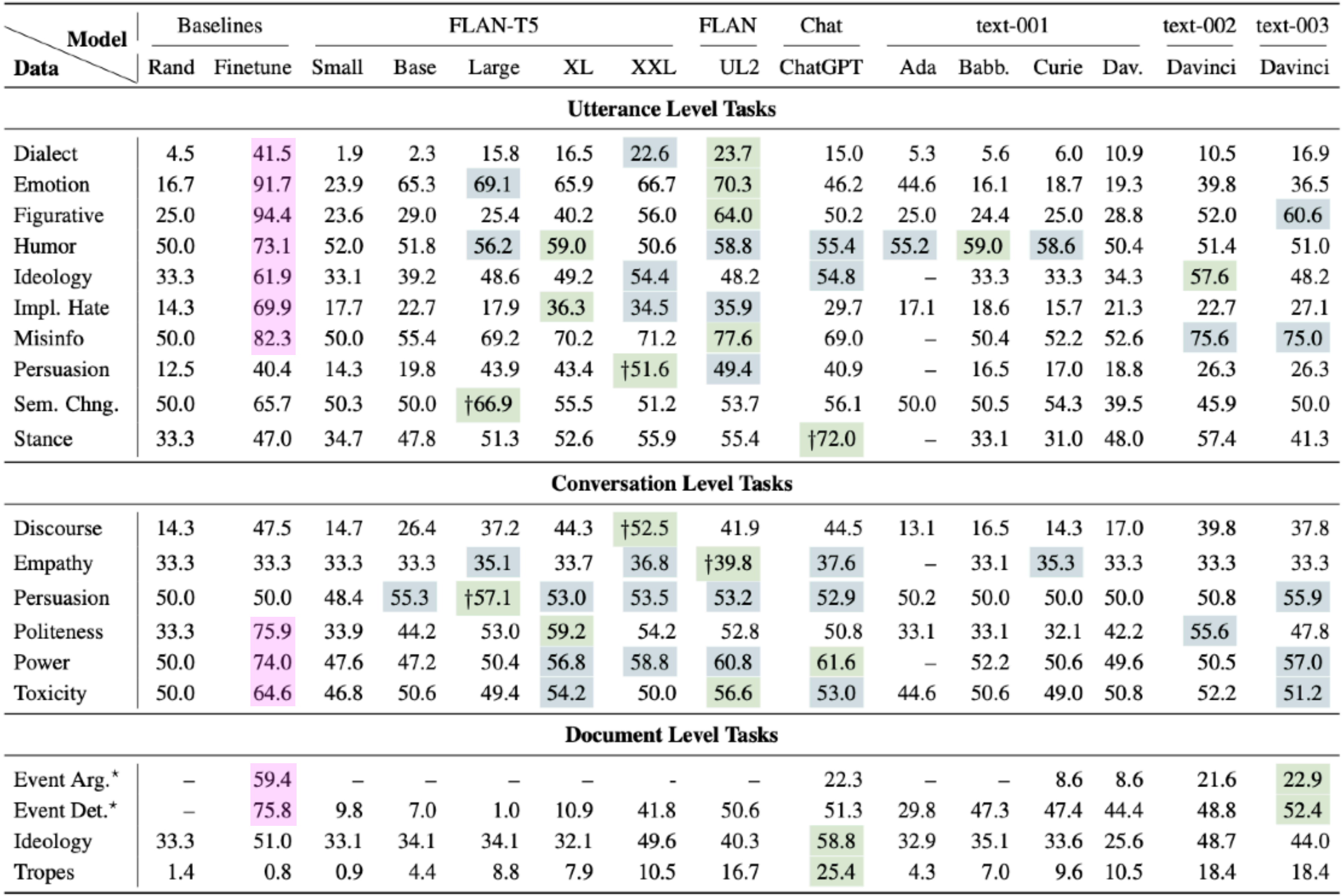
https://tedunderwood.com/2023/03/19/using-gpt-4-to-measure-the-passage-of-time-in-fiction/

LLMs for Text-as-Data

- Gilardi et al. (2023), “ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks”

- Accuracy of ChatGPT vs. MTurk on judging content moderation relevance, stance toward §230, topic ID, content moderation frames, media frames





Model / Data	Baselines		FLAN-T5					FLAN	Chat	text-001				text-002	text-003
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	ChatGPT	Ada	Babb.	Curie	Dav.	Davinci	Davinci
Utterance Level Tasks															
Dialect	4.5	41.5	1.9	2.3	15.8	16.5	22.6	23.7	15.0	5.3	5.6	6.0	10.9	10.5	16.9
Emotion	16.7	91.7	23.9	65.3	69.1	65.9	66.7	70.3	46.2	44.6	16.1	18.7	19.3	39.8	36.5
Figurative	25.0	94.4	23.6	29.0	25.4	40.2	56.0	64.0	50.2	25.0	24.4	25.0	28.8	52.0	60.6
Humor	50.0	73.1	52.0	51.8	56.2	59.0	50.6	58.8	55.4	55.2	59.0	58.6	50.4	51.4	51.0
Ideology	33.3	61.9	33.1	39.2	48.6	49.2	54.4	48.2	54.8	–	33.3	33.3	34.3	57.6	48.2
Impl. Hate	14.3	69.9	17.7	22.7	17.9	36.3	34.5	35.9	29.7	17.1	18.6	15.7	21.3	22.7	27.1
Misinfo	50.0	82.3	50.0	55.4	69.2	70.2	71.2	77.6	69.0	–	50.4	52.2	52.6	75.6	75.0
Persuasion	12.5	40.4	14.3	19.8	43.9	43.4	†51.6	49.4	40.9	–	16.5	17.0	18.8	26.3	26.3
Sem. Chng.	50.0	65.7	50.3	50.0	†66.9	55.5	51.2	53.7	56.1	50.0	50.5	54.3	39.5	45.9	50.0
Stance	33.3	47.0	34.7	47.8	51.3	52.6	55.9	55.4	†72.0	–	33.1	31.0	48.0	57.4	41.3
Conversation Level Tasks															
Discourse	14.3	47.5	14.7	26.4	37.2	44.3	†52.5	41.9	44.5	13.1	16.5	14.3	17.0	39.8	37.8
Empathy	33.3	33.3	33.3	33.3	35.1	33.7	36.8	†39.8	37.6	–	33.1	35.3	33.3	33.3	33.3
Persuasion	50.0	50.0	48.4	55.3	†57.1	53.0	53.5	53.2	52.9	50.2	50.0	50.0	50.0	50.8	55.9
Politeness	33.3	75.9	33.9	44.2	53.0	59.2	54.2	52.8	50.8	33.1	33.1	32.1	42.2	55.6	47.8
Power	50.0	74.0	47.6	47.2	50.4	56.8	58.8	60.8	61.6	–	52.2	50.6	49.6	50.5	57.0
Toxicity	50.0	64.6	46.8	50.6	49.4	54.2	50.0	56.6	53.0	44.6	50.6	49.0	50.8	52.2	51.2
Document Level Tasks															
Event Arg.*	–	59.4	–	–	–	–	-	–	22.3	–	–	8.6	8.6	21.6	22.9
Event Det.*	–	75.8	9.8	7.0	1.0	10.9	41.8	50.6	51.3	29.8	47.3	47.4	44.4	48.8	52.4
Ideology	33.3	51.0	33.1	34.1	34.1	32.1	49.6	40.3	58.8	32.9	35.1	33.6	25.6	48.7	44.0
Tropes	1.4	0.8	0.9	4.4	8.8	7.9	10.5	16.7	25.4	4.3	7.0	9.6	10.5	18.4	18.4

Ziems et al. (2023), “Can Large Language Models Transform Computational Social Science?”

Ziems et al. 2023

- LLMs are good on tasks that have lots of evidence in pre-training (e.g., sentiment analysis)
- Much worse performance on complex or new tasks that require category definitions.
- “Even the best LLMs exhibit unusably low performance on [Computational Social Science] tasks”

Activity

- Llama2
  https://replicate.com/meta/llama-2-70b-chat
  https://www.llama2.ai/

- ChatGPT/GPT-4
  http://chat.openai.com

- Bard
  https://bard.google.com