Natural Language Processing

Info 159/259

Lecture 9: Prompting methods and RLHF (Feb 14, 2023)

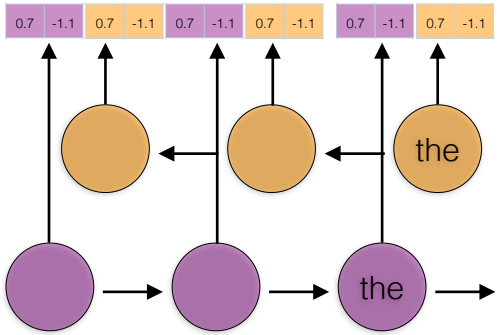
David Bamman, UC Berkeley

Contextualized word representations

- Big idea: transform the representation of a token in a sentence (e.g., from a static word embedding) to be sensitive to its local context in a sentence and trainable to be optimized for a specific NLP task.

ELMo

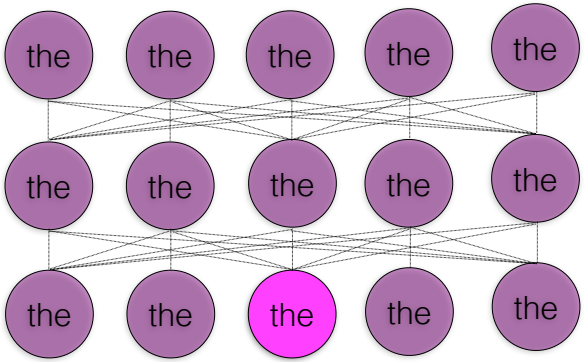
Stacked BiRNN trained to predict next word in language modeling task



Peters et al. 2018

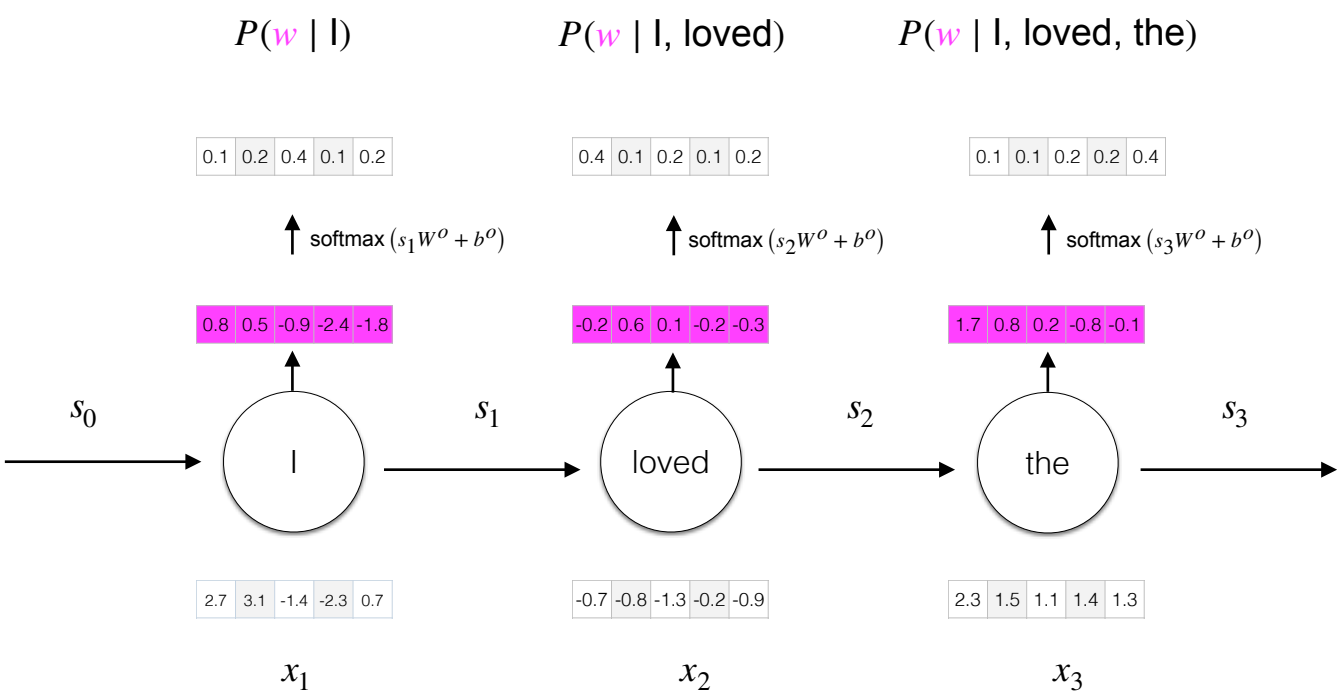
BERT

Transformer-based model to predict masked word using bidirectional context + next sentence prediction.



Devlin et al. 2019

RNN Language model

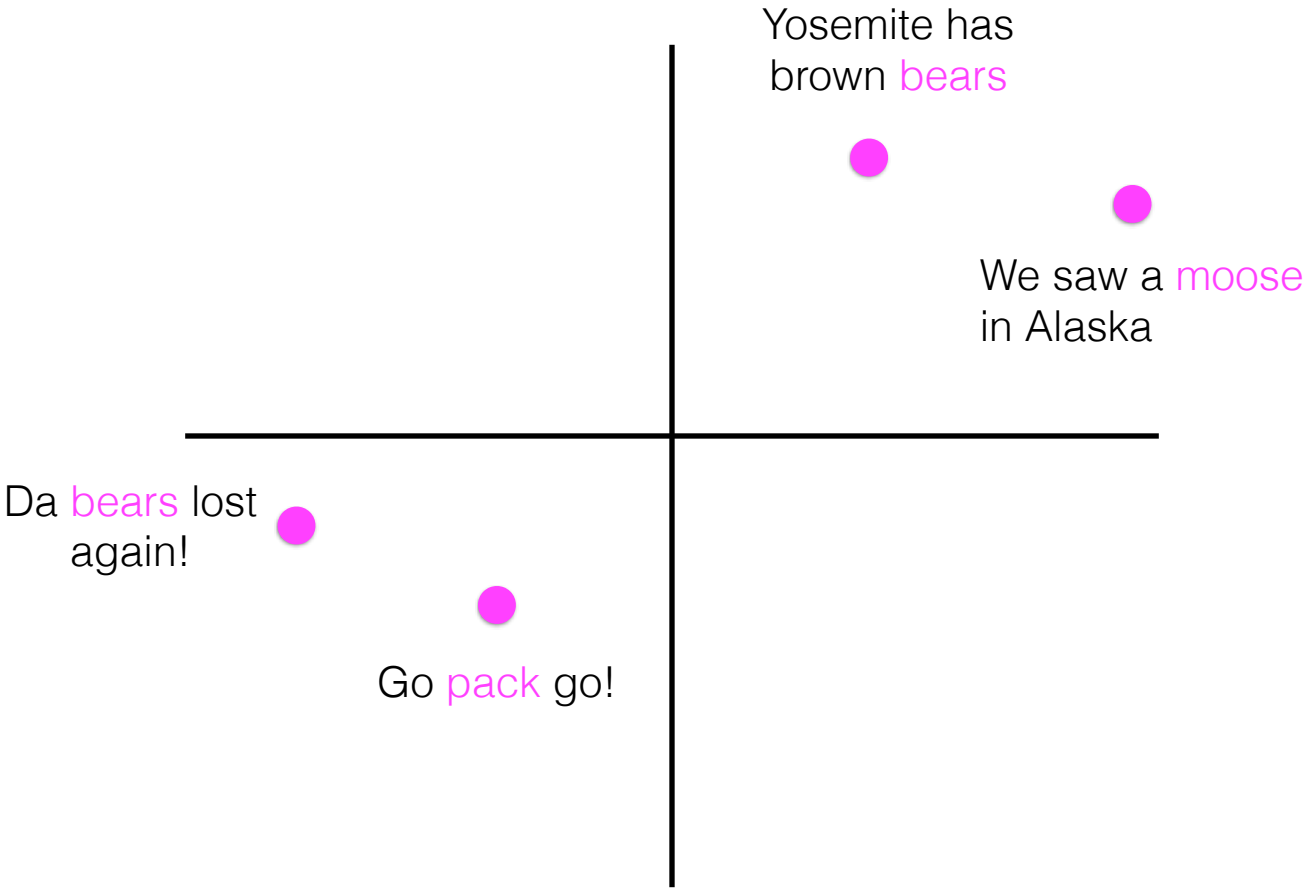


BERT

- Transformer-based model (Vaswani et al. 2017) to predict masked word using bidirectional context + next sentence prediction.
- Generates multiple layers of representations for each token sensitive to its context of use.

BERT

- Deep layers (12 for BERT base, 24 for BERT large)
- Large representation sizes (768 per layer)
- Pretrained on English Wikipedia (2.5B words) and BooksCorpus (800M words)


Yosemite has brown bears
We saw a moose in Alaska
Da bears lost again!
Go pack go!

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)

Generating

- As we sample, the words we generate form the new context we condition on

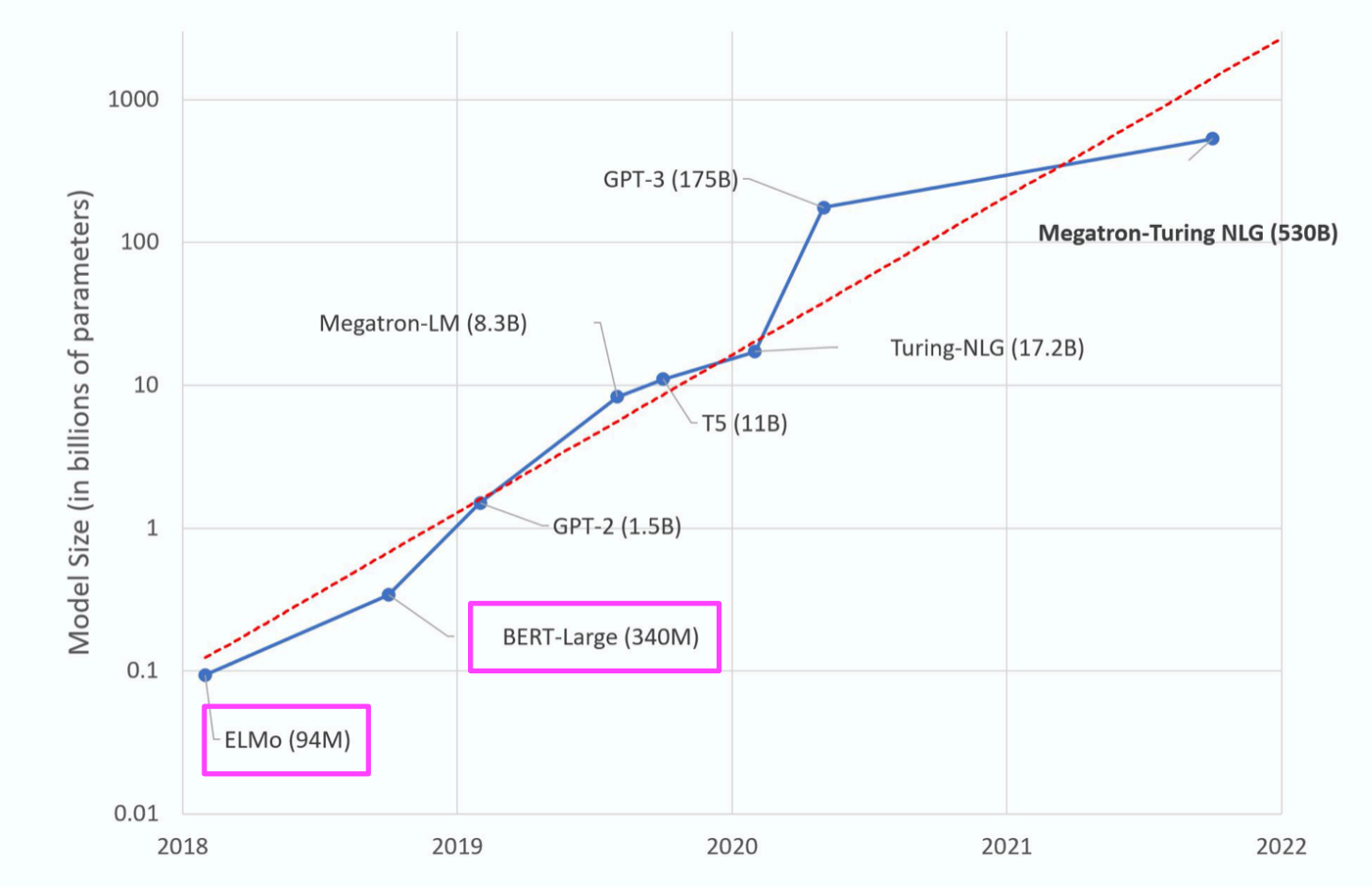
context1	context2	generated word
START	START	The
START	The	dog
The	dog	walked
dog	walked	in

Unigram model

- the around, she They I blue talking “Don’t to and little come of
- on fallen used there. young people to Lázaro
- of the
- the of of never that ordered don't avoided to complaining.
- words do had men flung killed gift the one of but thing seen I plate Bradley was by small Kingmaker.

Trigram Model

- “I’ll worry about it.”
- Avenue Great-Grandfather Edgeworth hasn’t gotten there.
- “If you know what. It was a photograph of seventeenth-century flourishin’ To their right hands to the fish who would not care at all. Looking at the clock, ticking away like electronic warnings about wonderfully SAT ON FIFTH
- Democratic Convention in rags soaked and my past life, I managed to wring your neck a boss won’t so David Pritchet giggled.
- He humped an argument but her bare He stood next to Larry, these days it will have no trouble Jay Grayer continued to peer around the Germans weren’t going to faint in the



https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

The importance of being on twitter

by Jerome K. Jerome
London, Summer 1897

It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage. I called it an anomaly, and it is.

I spoke to the sexton, whose cottage, like all sexton's cottages, is full of antiquities and interesting relics of former centuries. I said to him, "My dear sexton, what does all this twittering mean?" And he replied, "Why, sir, of course it means Twitter." "Ah!" I said, "I know about that. But what is Twitter?"

"It is a system of short and pithy sentences strung together in groups, for the purpose of conveying useful information to the initiated, and entertainment and the exercise of wits to the initiated, and entertainment and the exercise of wits to the rest of us."

https://twitter.com/quasimondo/status/1284509525500989445

Dialogue generation

Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

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)

LMs as knowledge bases

- Language models can directly encode knowledge present in the training corpus.

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

LMs as knowledge bases

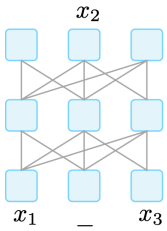
- Language models can directly encode knowledge present in the training corpus.

Query	Answer	Generation
Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8] , **Florence [-1.8]** , Naples
Adolphe Adam died in ____.	Paris	**Paris [-0.5]** , London [-3.5] , Vienna
English bulldog is a subclass of ____.	dog	dogs [-0.3] , breeds [-2.2] , **dog**
The official language of Mauritius is ____.	English	**English [-0.6]** , French [-0.9] , Arabic
Patrick Oboya plays in ____ position.	midfielder	centre [-2.0] , center [-2.2] , **midfielder**
Hamburg Airport is named after ____.	Hamburg	Hess [-7.0] , Hermann [-7.1] , Schmidt

Petroni et al. (2019), "Language Models as Knowledge Bases?" (ACL)

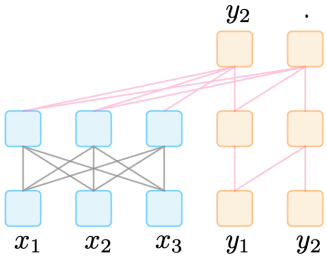
Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first			48.3%
Who is the heac			47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%

Radford et al. 2019, “Language Models are Unsupervised Multitask Learners” (GPT-2)



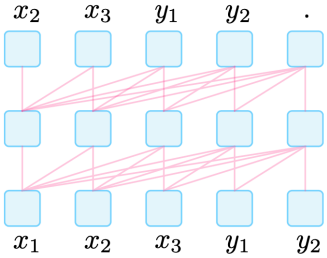
Masked LM
(BERT)

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



Encoder-decoder
(T5)

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



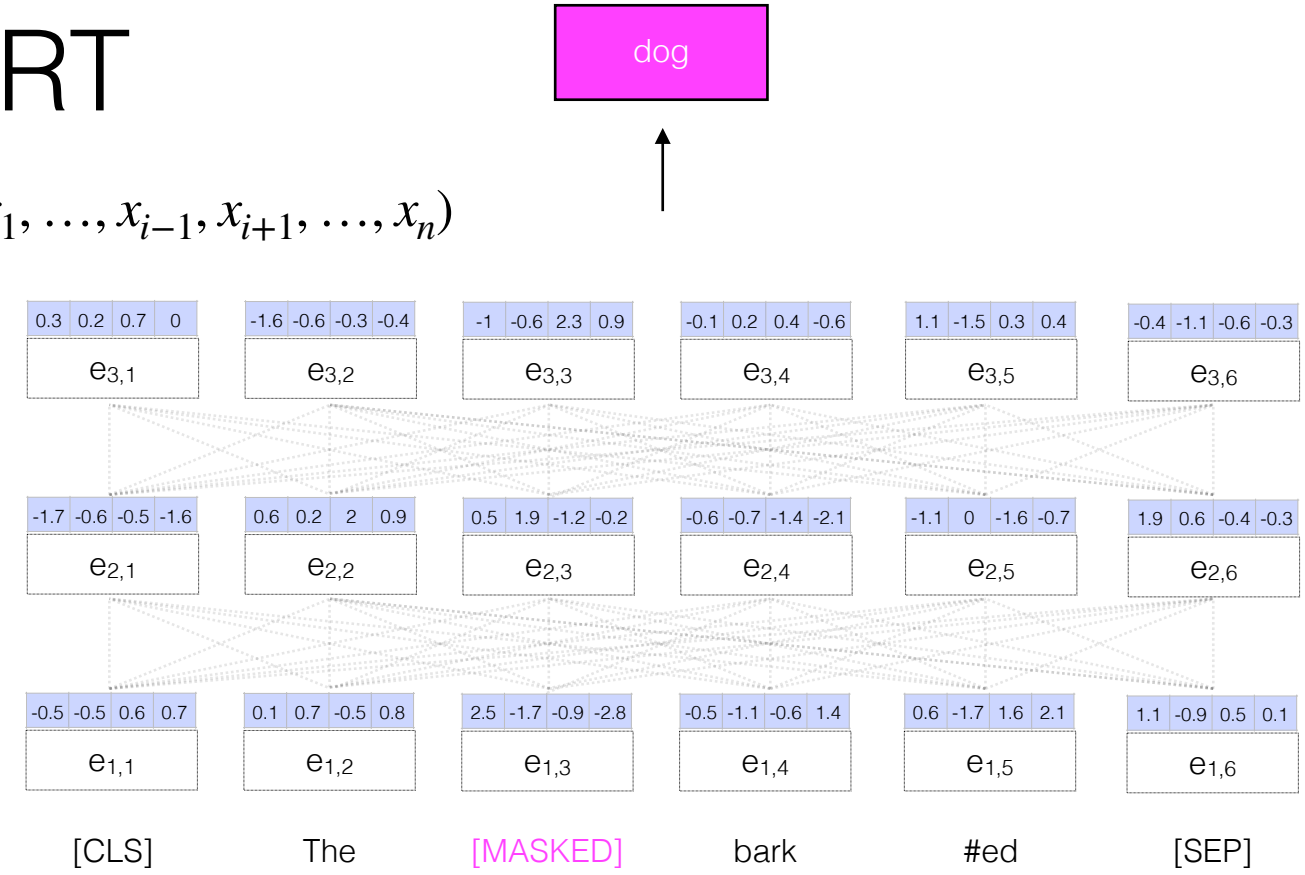
Left-to-right LM
(GPT)

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

Liu et al. 2021

BERT

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



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

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

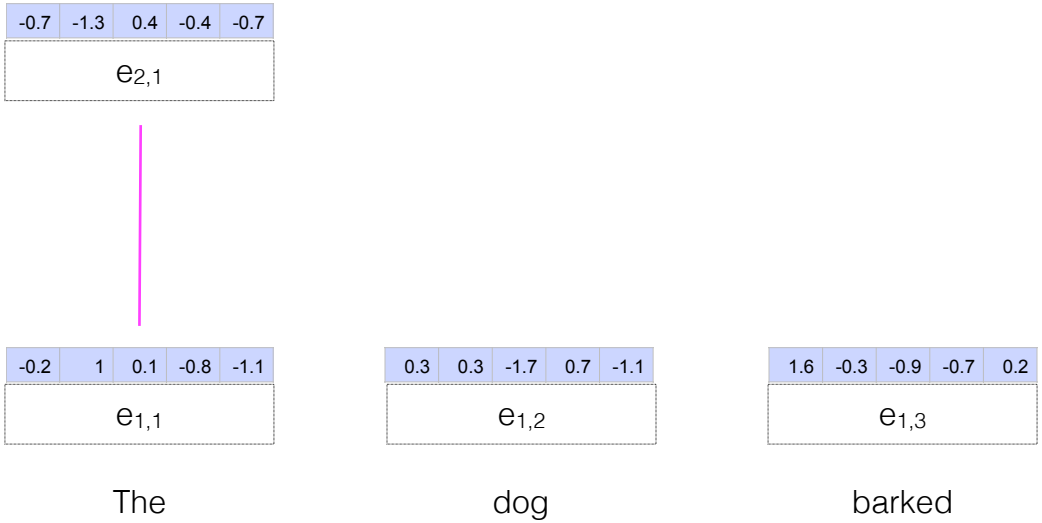
GPT

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

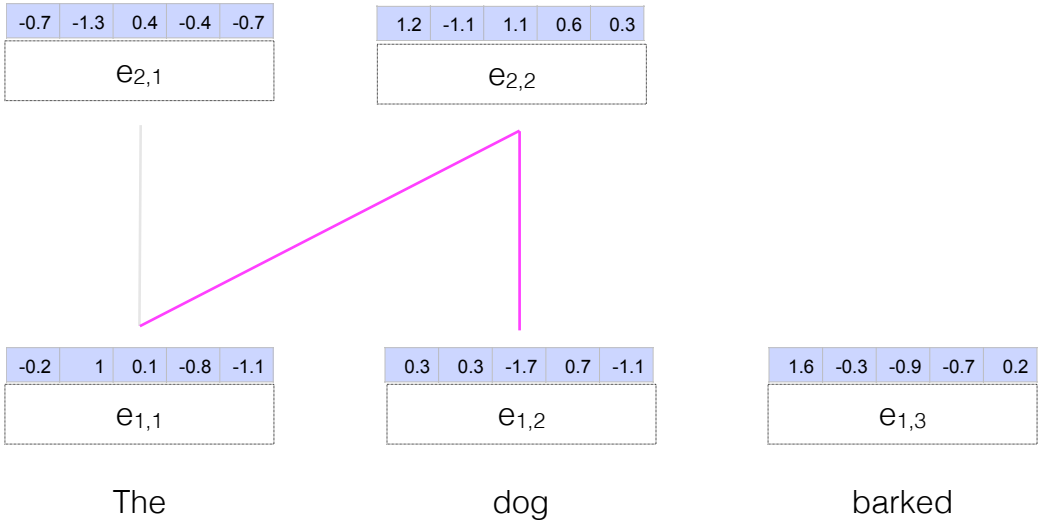
$$P(x) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

	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)

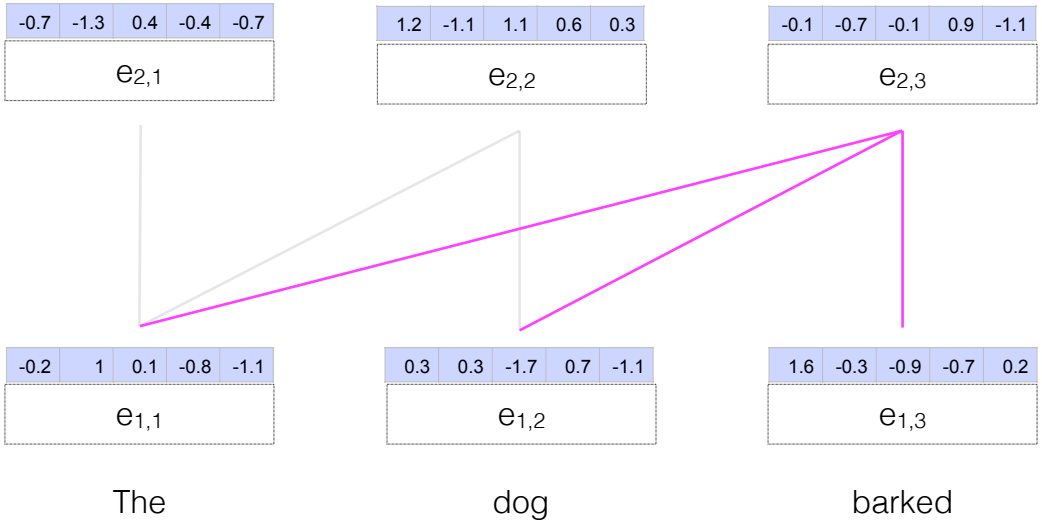
- Self-attention for token i at layer j only attends to tokens 1 through i at layer j-1

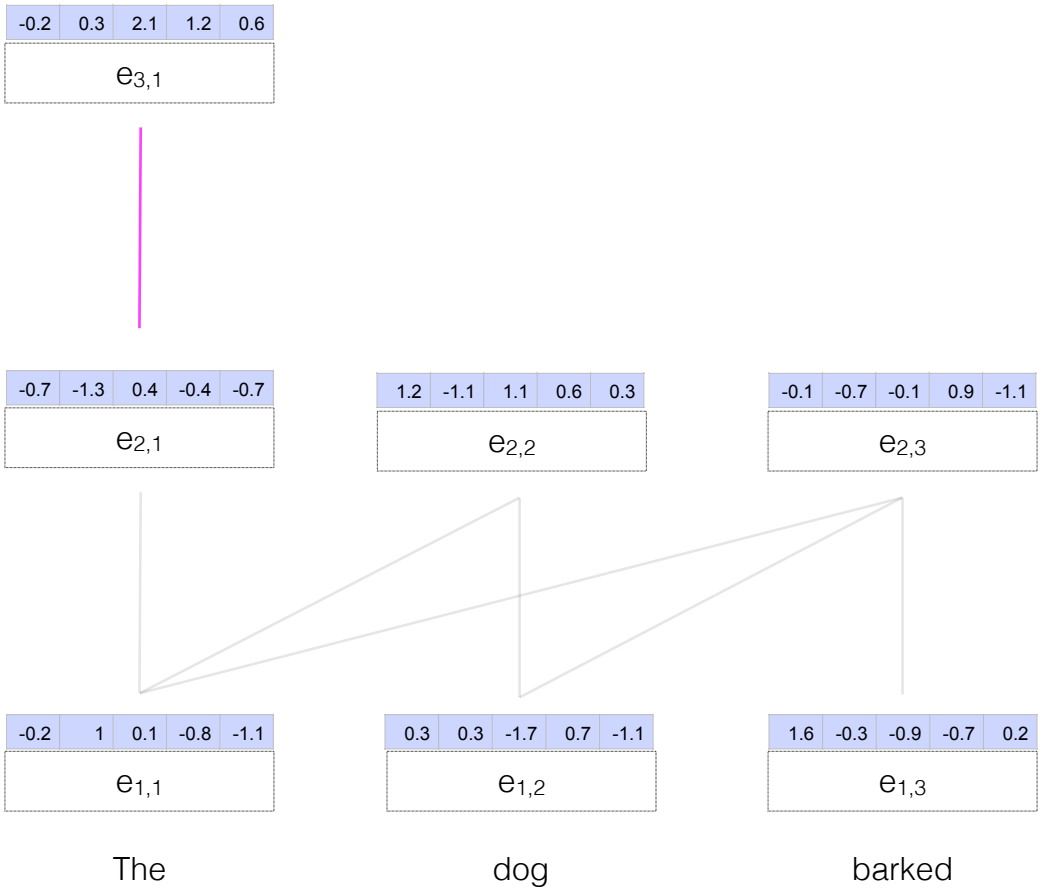


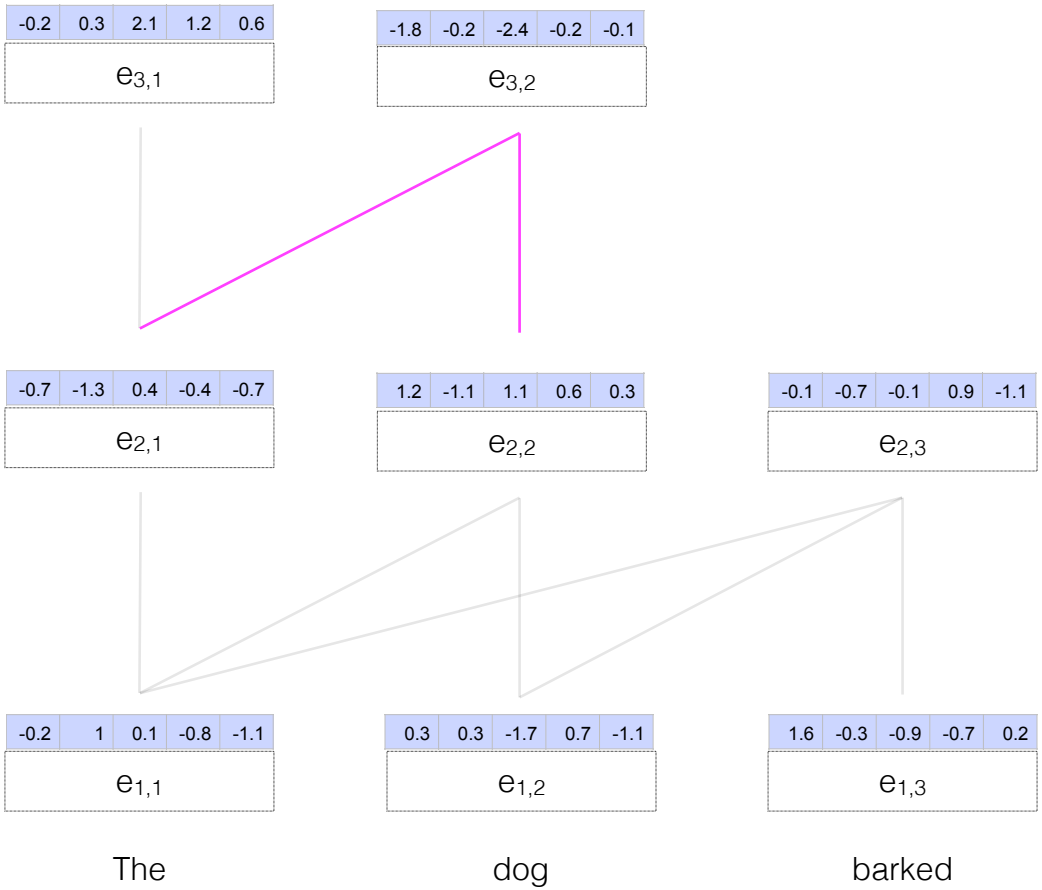
- Self-attention for token i at layer j only attends to tokens 1 through i at layer j-1

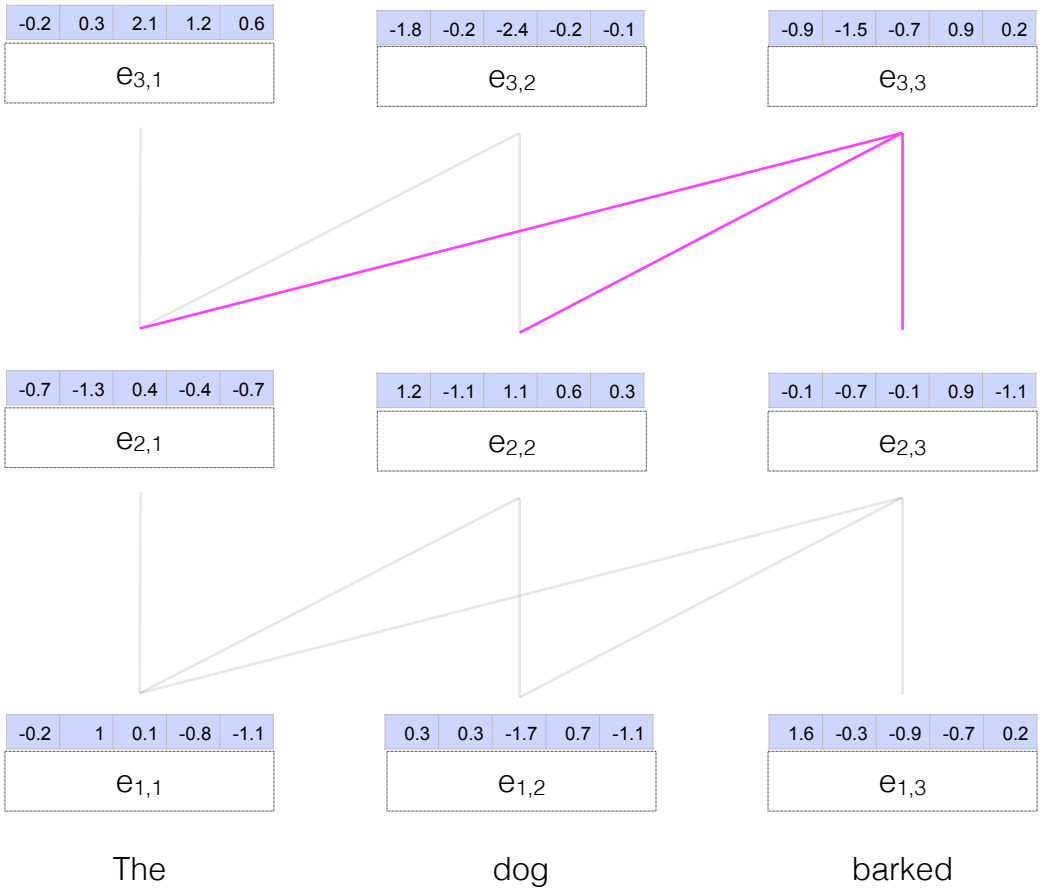


- Self-attention for token i at layer j only attends to tokens 1 through i at layer j-1




-0.2 0.3 2.1 1.2 0.6
e3,1
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked


-0.2 0.3 2.1 1.2 0.6
e3,1
-1.8 -0.2 -2.4 -0.2 -0.1
e3,2
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked


-0.2 0.3 2.1 1.2 0.6
e3,1
-1.8 -0.2 -2.4 -0.2 -0.1
e3,2
-0.9 -1.5 -0.7 0.9 0.2
e3,3
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked

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 _____

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:
cheese =>
```

task description

prompt

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

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:
sea otter => loutre de mer
cheese =>
```

task description

example

prompt

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

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 =>
```

task description

examples

prompt

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

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.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

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

Context →	Please unscramble the letters into a word, and write that word: volwskagen =
Target Completion →	volkswagen

Figure G.23: Formatted dataset example for Anagrams 2

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

Context →	Title: The_Blitz Background: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Gerät beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Gerät, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Gerät. Q: How many sorties were flown in March 1941? A: 4,000 Q: When did the Luftwaffe fly inland missions? A:
Target Completion →	only on moonlit nights

Figure G.28: Formatted dataset example for SQuADv2

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

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
$\boldsymbol{s} \rightarrow \boldsymbol{h}$	70	`in summary`	`in short, in a word, to sum up`
$\boldsymbol{h} \leftrightarrow \boldsymbol{r}$	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:”

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

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).

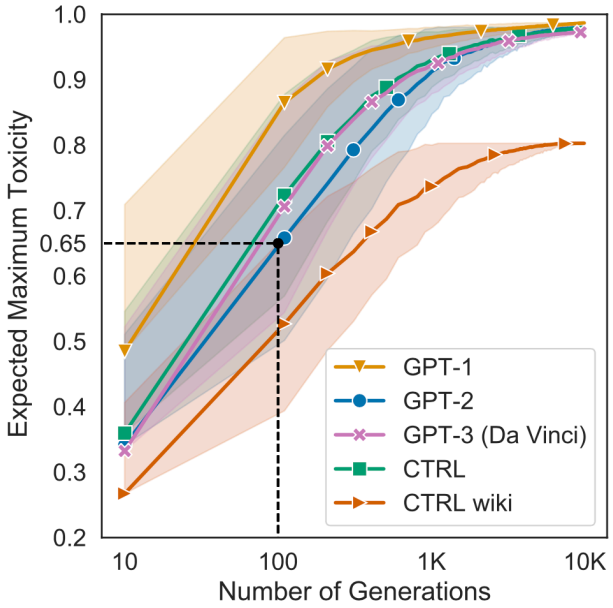
- When known, training data encodes narrow perspectives — e.g., links shared on Reddit; filtering out pages containing words related to sex (as C4 does) filters pornography but also positive sex discussions.

- 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

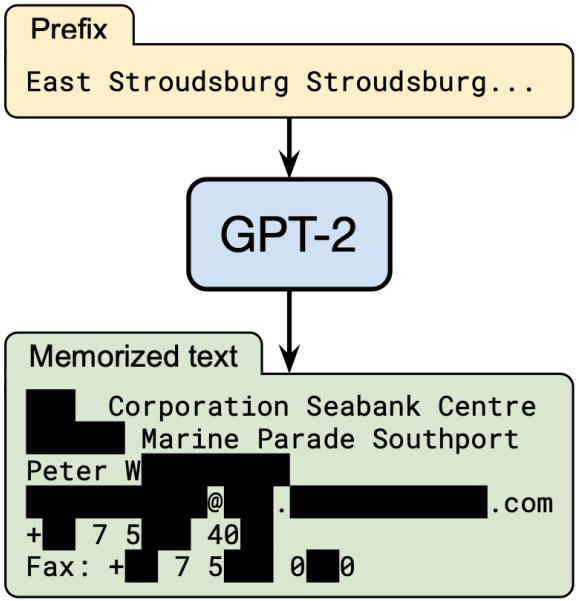


Sap et al. (2020), “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”

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).



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

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.

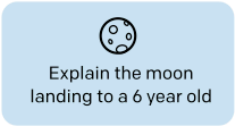
Q: How many rainbows does it take to jump from Hawaii to seventeen?
A: It takes two rainbows to jump from Hawaii to seventeen.

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)

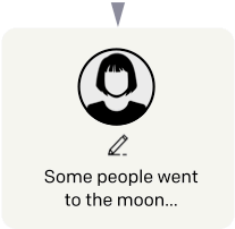
Step 1

Collect demonstration data, and train a supervised policy.

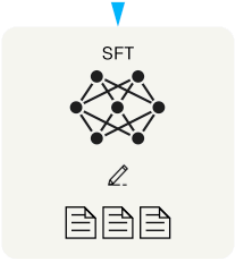
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



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



Instruct-GPT

- 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

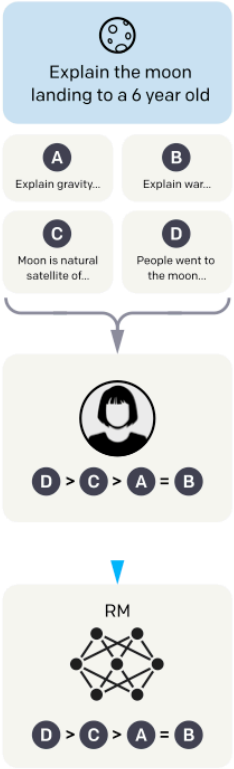
Step 2

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.



Instruct-GPT

- 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

DB

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

DB

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.

259 project proposal
due 3/5

- Final project involving 1 to 3 students involving natural language processing -- either focusing on core NLP methods or using NLP in support of an empirical research question.

- Proposal (1000 words, excluding references):

    - outline the work you're going to undertake
    - motivate its rationale as an interesting question worth asking
    - assess its potential to contribute new knowledge by situating it within related literature in the scientific community. (cite 5 relevant sources)
    - who is the team and what are each of your responsibilities (everyone gets the same grade)

- Feel free to come by my office hours and discuss your ideas!