



Applied Natural Language Processing

Info 256

Lecture 15: Prompting LLMs (Oct 16, 2023)

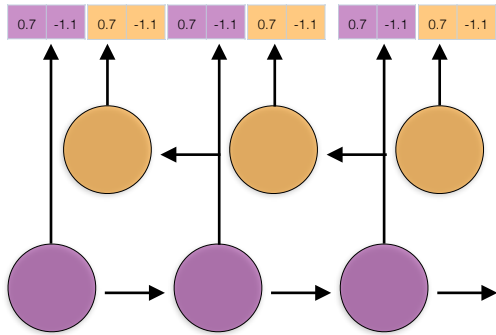
David Bamman, UC Berkeley

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

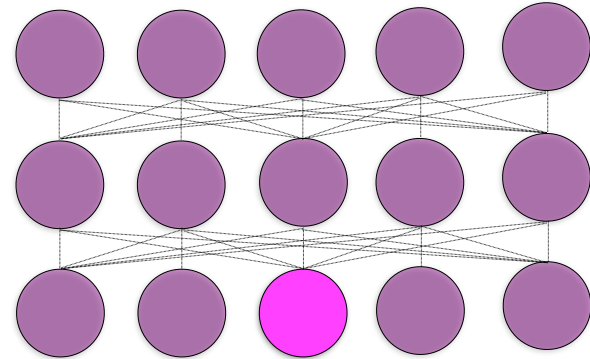
Stacked BiRNN trained to predict **next** word in language modeling task



Peters et al. 2018

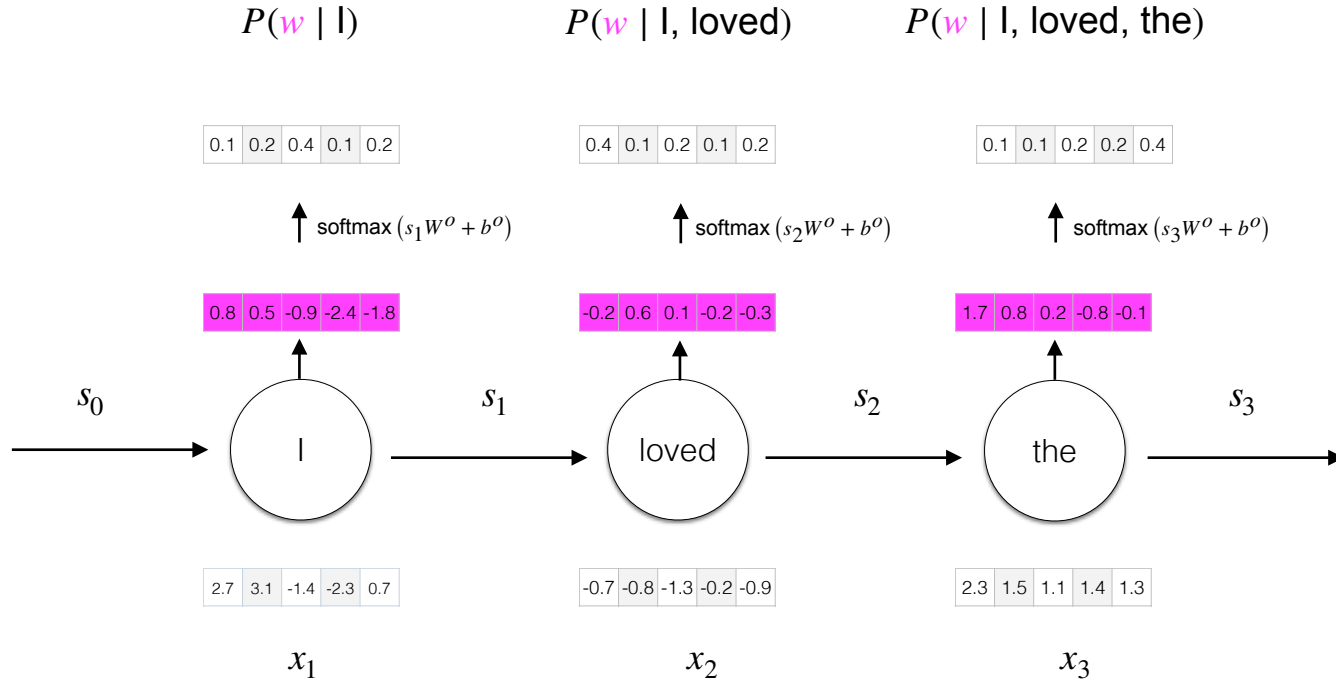
BERT

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).

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

- 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

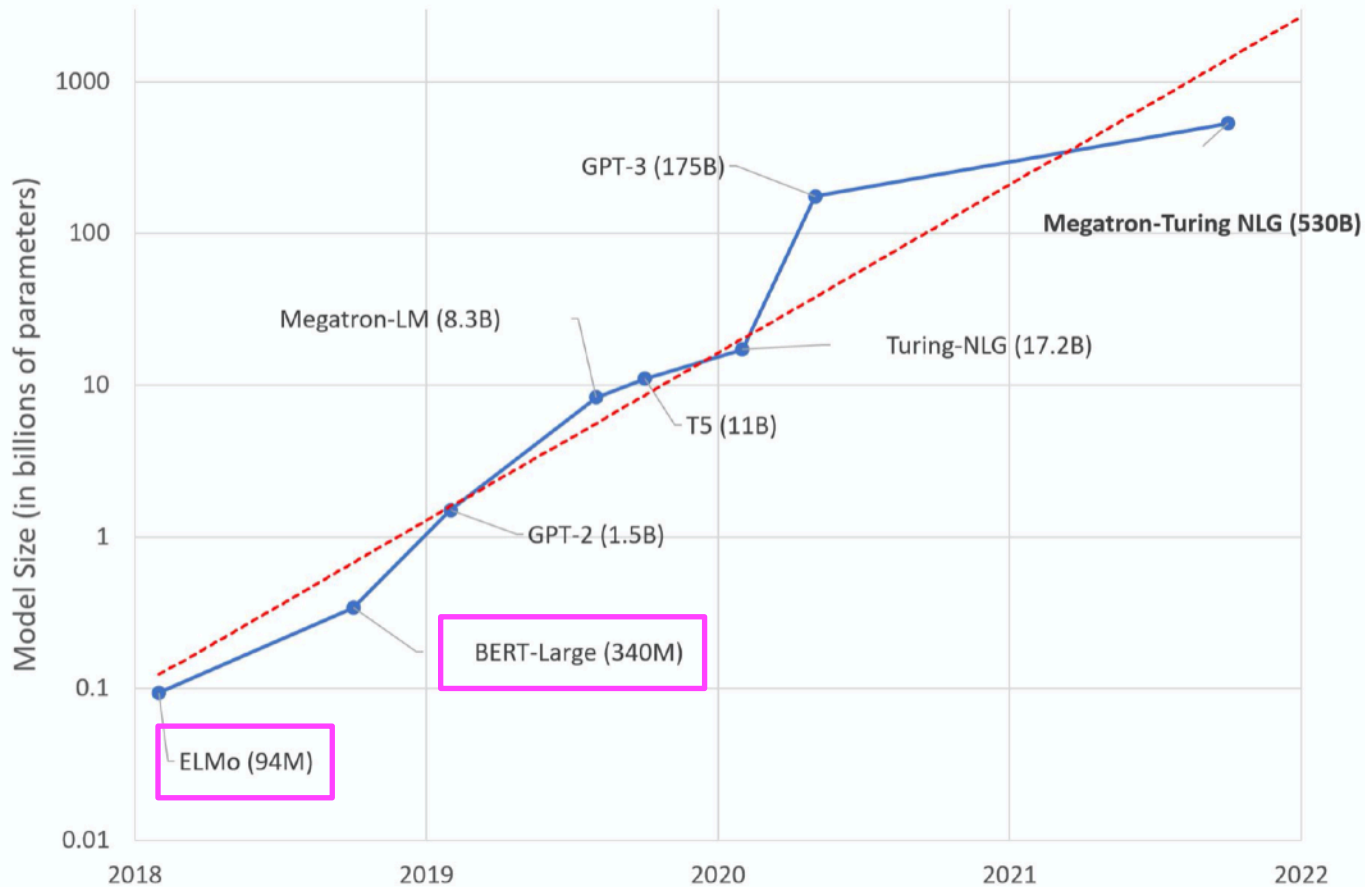
context1	context2	generated word
START	START	The
START	The	dog
The	dog	walked
dog	walked	in

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 Pritchett 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."

Dialogue generation

DB

What's a good restaurant on Shattuck in Berkeley?



There are many great restaurants on Shattuck Avenue in Berkeley. Here are a few options:



1. Comal: This restaurant serves upscale Mexican cuisine and has a lively atmosphere. They have a great selection of cocktails and margaritas, as well as delicious dishes like tacos, ceviche, and grilled meats.
2. Angeline's Louisiana Kitchen: If you're in the mood for Cajun and Creole cuisine, Angeline's is a great choice. They have a wide selection of dishes, including gumbo, jambalaya, and po' boys. They also have a great brunch menu on the weekends.
3. Cheese Board Pizza: This popular pizza joint serves only one type of pizza per day, but it's always delicious. They use fresh, local ingredients and their crust is crispy and flavorful. They also have a great selection of cheeses and other snacks.
4. Jupiter: Jupiter is a great spot for pizza and beer. They have a wide selection of pizzas, salads, and sandwiches, as well as a rotating selection of craft beers on tap.
5. Great China: If you're in the mood for Chinese food, Great China is a great option. They have a large menu with many traditional dishes, as well as some more unique offerings. Their Peking duck is especially popular.

These are just a few of the many great restaurants on Shattuck Avenue in Berkeley. I recommend checking out menus and reviews online to find one that suits your tastes!

LMs as knowledge bases

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

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

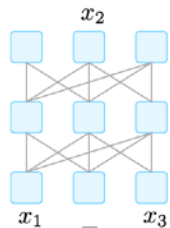
LMs as knowledge bases

- 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

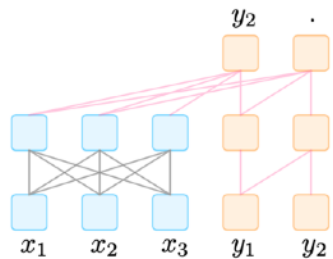
Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	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 head			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)



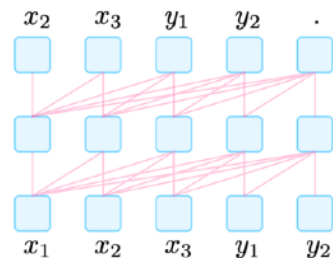
Masked LM
(BERT)

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



Encoder-decoder
(T5)

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

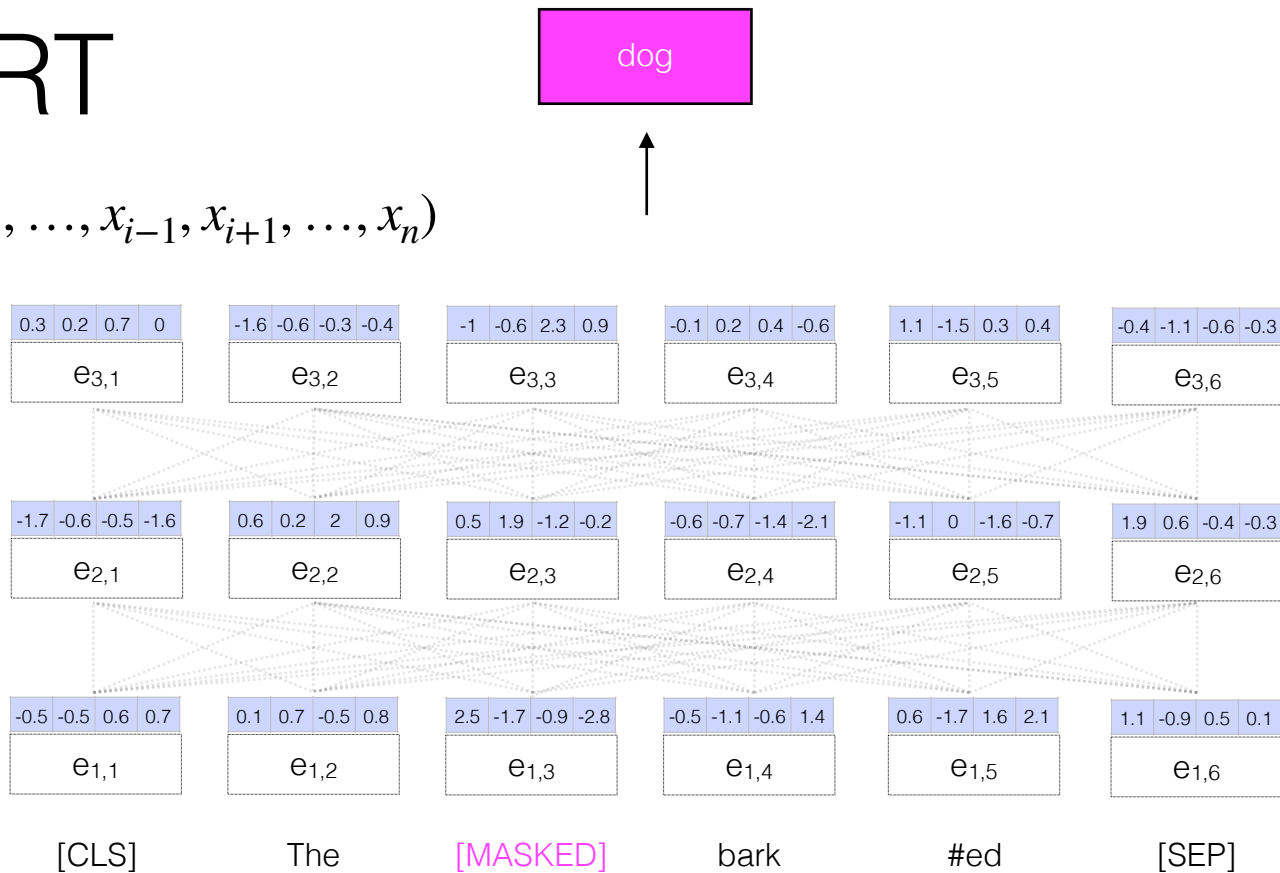


Left-to-right LM
(GPT)

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

BERT

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



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



[A] ~~for inviting~~ [B] ~~last~~ [C]

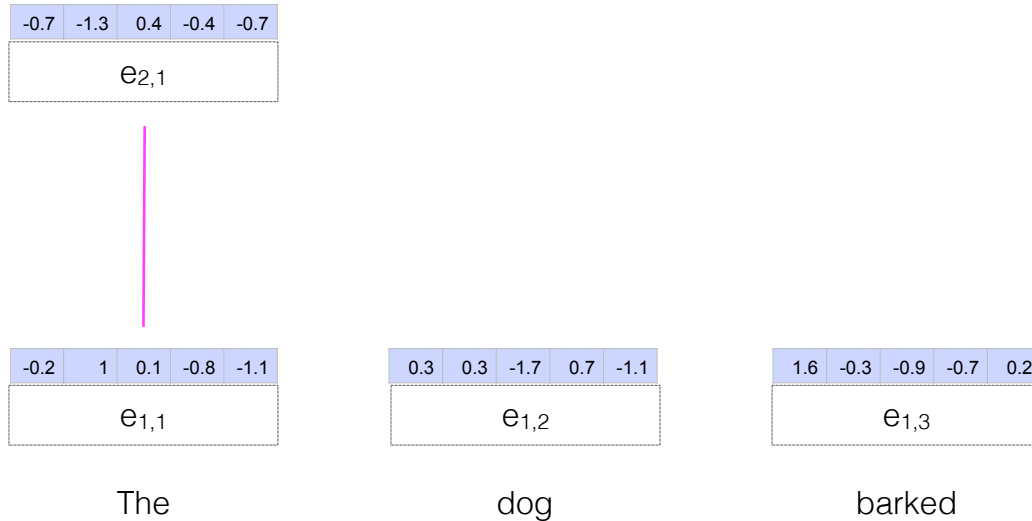
GPT

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

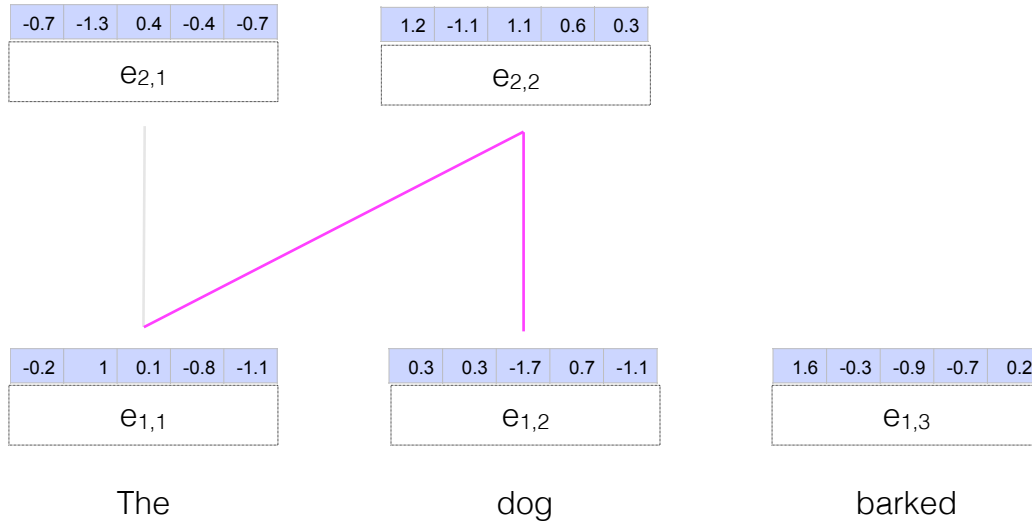
$$P(x) = \prod_{i=1}^n P(x_i | x_1, \dots, x_{i-1})$$

	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)

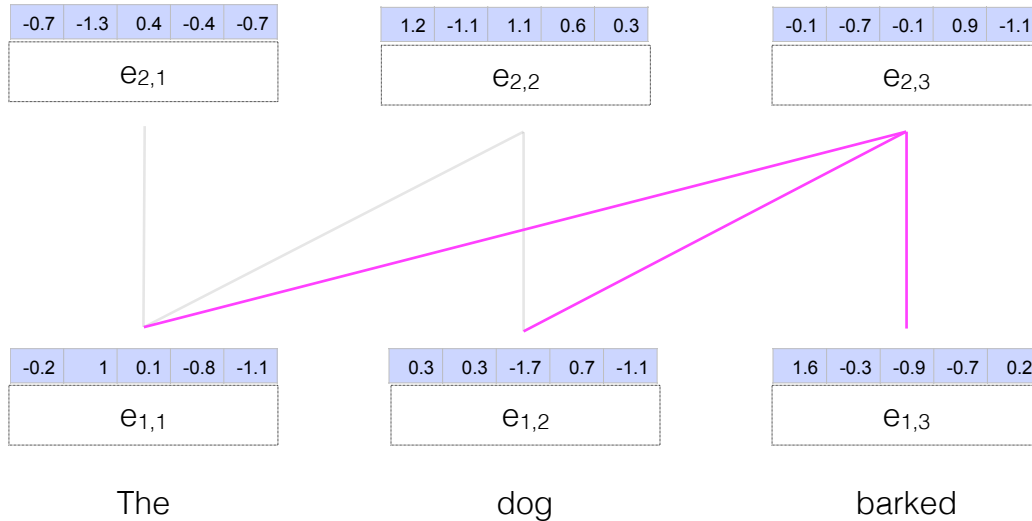
- Self-attention for token i at layer j only attends to tokens 1 through i at layer $j-1$



- 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
$e_{3,1}$				



-0.7	-1.3	0.4	-0.4	-0.7
$e_{2,1}$				

1.2	-1.1	1.1	0.6	0.3
$e_{2,2}$				

-0.1	-0.7	-0.1	0.9	-1.1
$e_{2,3}$				

-0.2	1	0.1	-0.8	-1.1
$e_{1,1}$				

0.3	0.3	-1.7	0.7	-1.1
$e_{1,2}$				

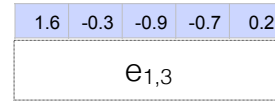
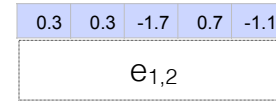
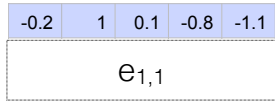
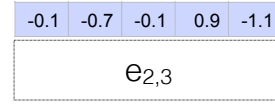
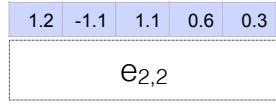
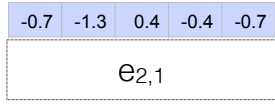
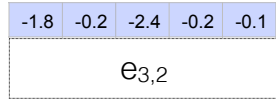
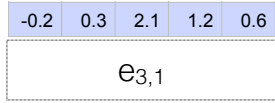
1.6	-0.3	-0.9	-0.7	0.2
$e_{1,3}$				

The

dog

barked

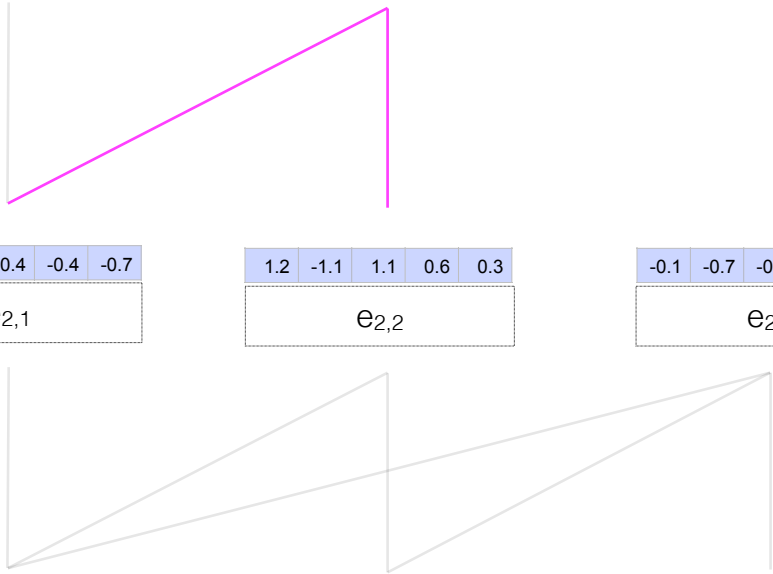


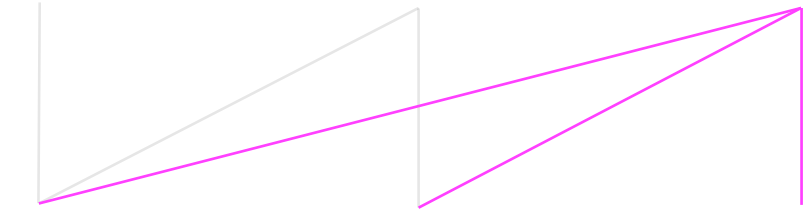
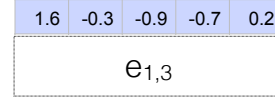
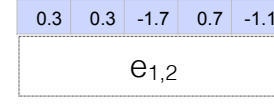
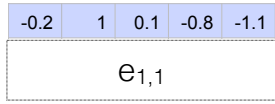
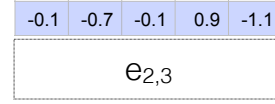
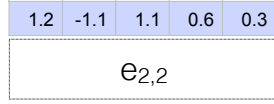
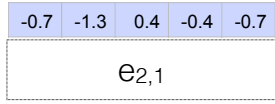
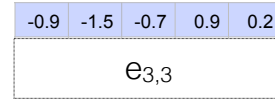
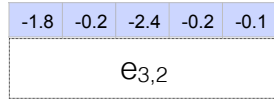
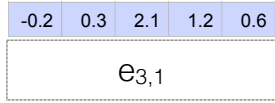


The

dog

barked





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.

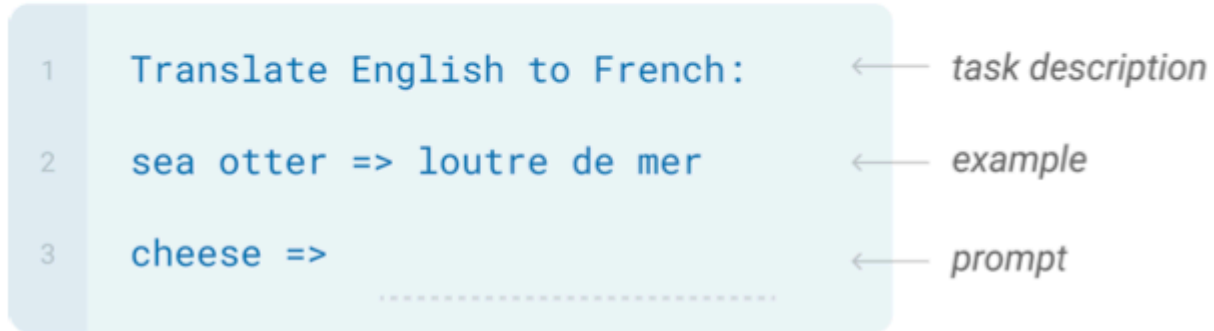


The diagram shows a light blue rounded rectangle containing two lines of text. The first line is '1 Translate English to French:' followed by an arrow pointing left and the text 'task description'. The second line is '2 cheese =>' followed by a dotted line and an arrow pointing left and the text 'prompt'.

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

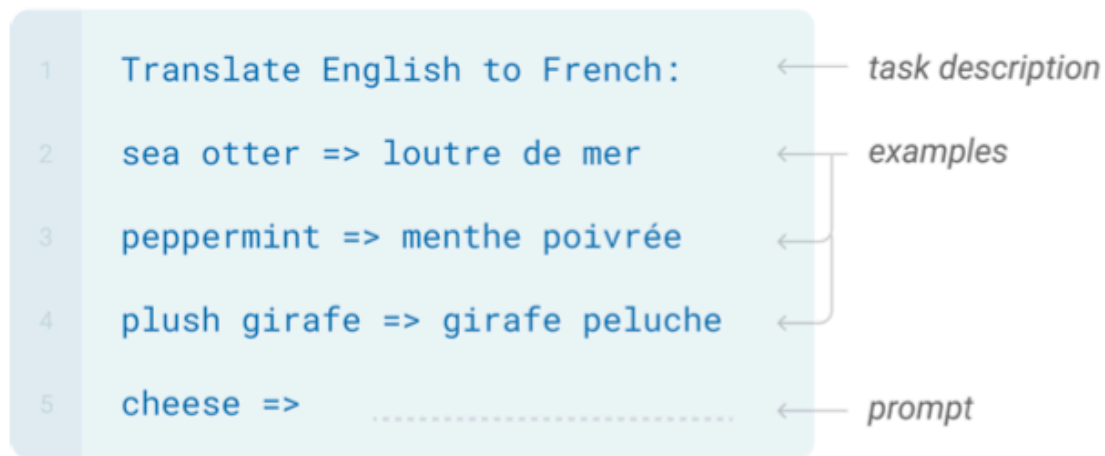
One-shot

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



Few-shot

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



Poor English input: I ate 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.

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

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

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

Activity

8.lm/PromptDesign_TODO.ipynb