

# Large Language Models

Vlad Shakhuro

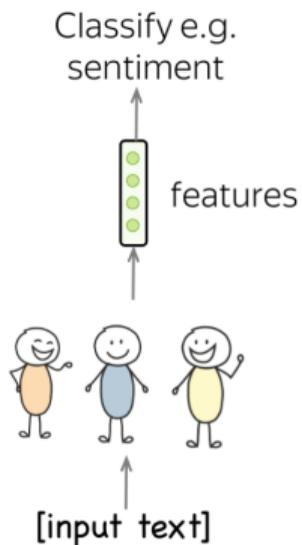


5 December 2025

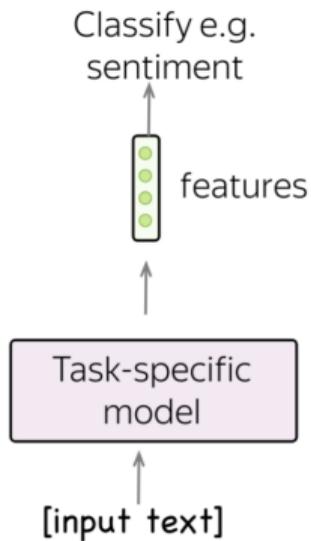
# Outline

1. From GPT-1 to GPT-3 and in-context learning
2. Prompting techniques
3. Do models understand prompts?
4. Analysis

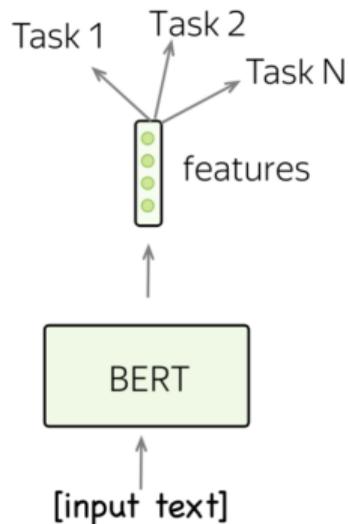
Different task –  
another set of  
features (and people!)



Different task –  
another model



One model,  
classify anything



Talk to the model

Input (prompt)

**What is the sentiment  
of the next sentence?  
I love this movie!**

Model output  
**positive**

# GPT: Generative Pre-trained Transformer

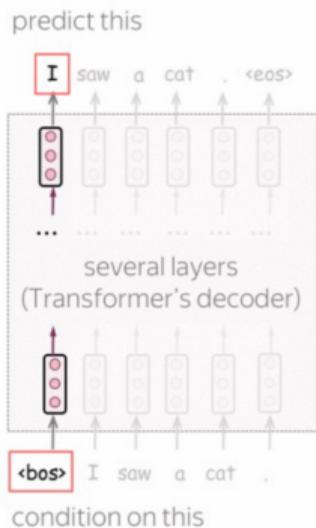
	<u>GPT-1</u> (2018)	<u>GPT-2</u> (2019)	<u>GPT-3</u> (2020)
Number of parameters	117 million	1.5 billion	175 billion
Training data	5 GB	40 GB	45 TB (i.e., 45 000 GB)

The entirety of English  
Wikipedia constitutes just  
0.6% of GPT-3 training data

# GPT-2

## Training

Transformer decoder with the standard left-to-right language modeling objective



## Inference

GPT-1 Classification via 

- Task-specific input transformations
- Supervised fine-tuning

GPT-2 Generation tasks via task-specific input transformations

For example, for text summarization simply add “TL;DR”:

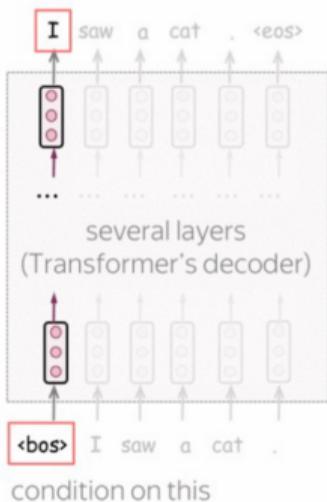
<start> Generative pre-trained transformers (GPT) are ... TL;DR:  
Text to be summarized

# GPT-3

## Training

Transformer decoder with the standard left-to-right language modeling objective

predict this



## Inference

GPT-1 Classification via 

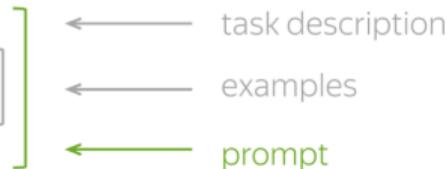
- Task-specific input transformations
- Supervised fine-tuning

GPT-2 Generation tasks via task-specific input transformations

For example, for text summarization simply add “**TL;DR**”:

GPT-3 Complex generation and reasoning tasks via in-context learning: prompt with task description and a few demonstrations

Translate English to Spanish:  
a black cat -> un gato negro  
I am hungry -> tengo hambre  
a cup of tea ->



# GPT-3 summarization



## TL;DR summarization

Transformation

Generation

Summarize text by adding a 'tl;dr:' to the end of a text passage. It shows that the API understands how to perform a number of tasks with no instructions.

### Prompt

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.[1] Neutron stars are the smallest and densest stellar objects, excluding black holes and hypothetical white holes, quark stars, and strange stars.[2] Neutron stars have a radius on the order of 10 kilometres (6.2 mi) and a mass of about 1.4 solar masses. [3] They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past white dwarf star density to that of atomic nuclei.

Tl;dr

### Sample response

A neutron star is the collapsed core of a massive supergiant star. These ultra-dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.

# GPT-3 factual answering



## Factual answering

Answers

Generation

Conversation

Classification

Guide the model towards factual answering by showing it how to respond to questions that fall outside its knowledge base. Using a '?' to indicate a response to words and phrases that it doesn't know provides a natural response that seems to work better than more abstract replies.

## Prompt

Q: Who is Batman?

A: Batman is a fictional comic book character.

Q: What is torsalplexity?

A: ?

Q: What is Devz9?

A: ?

Q: Who is George Lucas?

A: George Lucas is American film director and producer famous for creating Star Wars.

Q: What is the capital of California?

A: Sacramento.

Q: What orbits the Earth?

A: The Moon.

Q: Who is Fred Rickerson?

A: ?

Q: What is an atom?

A: An atom is a tiny particle that makes up everything.

Q: Who is Alvan Muntz?

A: ?

# GPT-3 translation



## English to other languages

Transformation

Generation

Translates English text into French, Spanish and Japanese.

## Prompt

Translate this into 1. French, 2. Spanish and 3. Japanese:

What rooms do you have available?

1.

## Sample response

Quels sont les chambres disponibles?

2. ¿Cuáles son las habitaciones disponibles?

3. 何室がありますか?

# GPT-3 learning from context

p language is a language where every word is appended with p.

**Q: you are a good boy.**

**A: youp arep ap goodp boyp.**

**Q: Where is the spoon?**

**A: Wherep isp thep spoonp?**

**Q: The cat is a small animal.**

**A: Thep catp isp ap smallp animalp.**

**Q: Can he jump?**

**A: Canp hep jumpsp?**

**Q: The forest was big. There were lots of animals and trees. But no humans.**

**A: Thep forestp wasp bigp. Therep weresp lotsp ofp animalsp andp treesp. Butp nopsp humanp.**

**Q: Can the lady call the guy? There is an emergency.**

**A: Canp thep ladyp callp thep guyp? Therep isp ap emergencyp.**

**Q: Thank the heavens. Peter pan went to his pop.**

**A: Thankp thep heavensp. Peterp panp wentp topp hisp popp.**

# GPT-3 image classification

Prompt: put feature vectors inside context span



```
Input: 94, 47, 84, 31, output = 2
Input: 89, 51, 73, 31, output = 1
[...]
Input: 96, 51, 80, 38, output = 2
Input: 90, 37, 76, 27, output =
```



Features are scaled  
and rounded to integers  
Guess why?

0 = Iris Setosa

1 = Iris Versicolor

2 = Iris Virginica



(because GPT-3 is bad at processing fractions)

# GPT-3 image classification

Prompt: put feature vectors inside context span



```
Input: 94, 47, 84, 31, output = 2
Input: 89, 51, 73, 31, output = 1
[...]
Input: 96, 51, 80, 38, output = 2
Input: 90, 37, 76, 27, output =
```

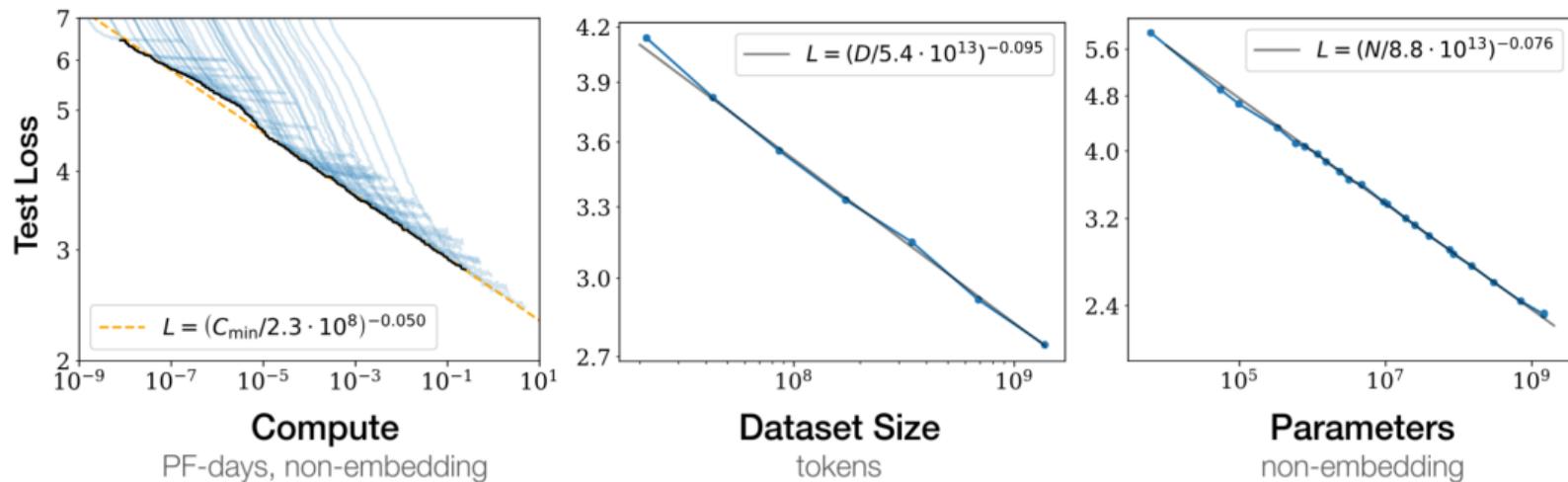
Features are scaled  
and rounded to integers  
Guess why?

(because GPT-3 is bad at processing fractions)

Model	Average acc.
kNN	81.78%
Logistic regr.	62.34%
Custom text	67.03%
Ada	73.70%
Babbage	72.10%
Curie	75.68%
Davinci	75.93%

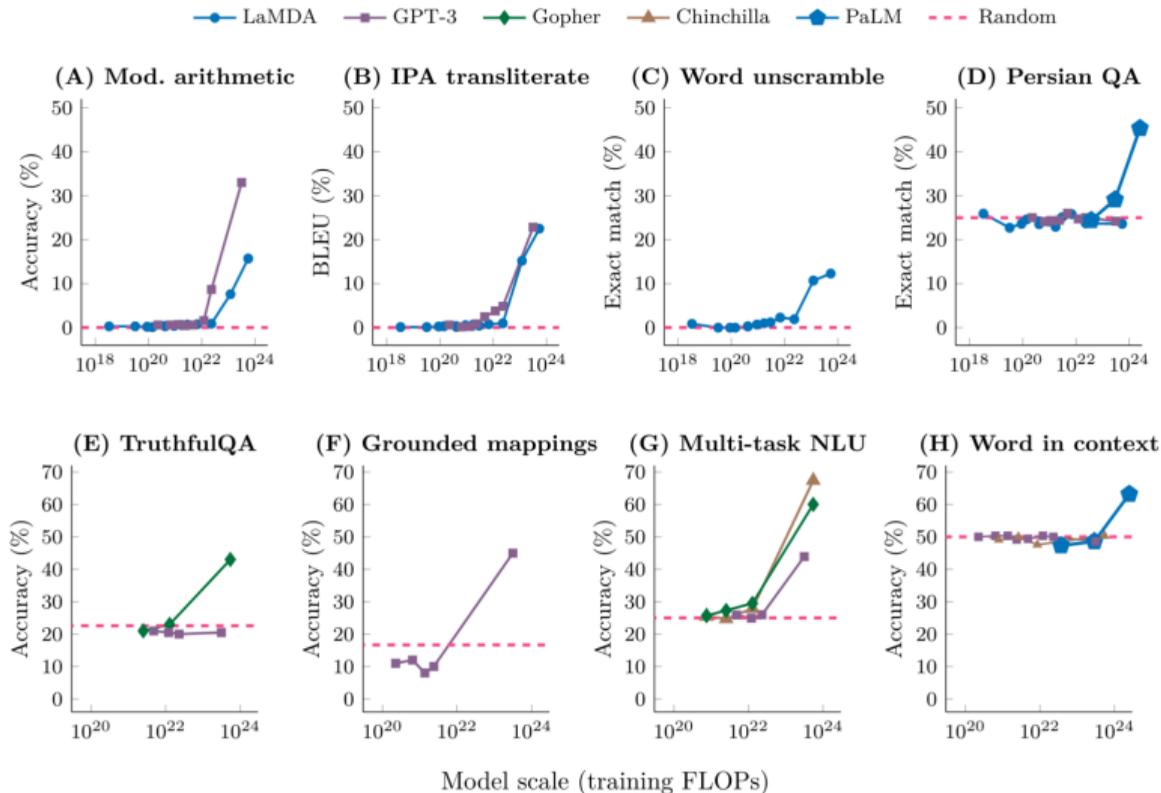
Model size:  
350M  
1.3B  
6.7B  
175B

# Scaling laws



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

# Emergent abilities



# Outline

1. From GPT-1 to GPT-3 and in-context learning
2. Prompting techniques
3. Do models understand prompts?
4. Analysis

# Prompting

No additional training! Just give a task with an example – prompt.  
This is fed as prefix before model generation.

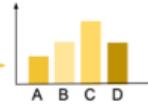
	zero-shot	few-shot	
		one-shot	two-shot
Input (prompt)	Review: I love this movie! Sentiment:	Review: This movie sucks. Sentiment: negative  Review: I love this movie! Sentiment:	Review: This movie sucks. Sentiment: negative  Review: This was cool! Sentiment: positive  Review: I love this movie! Sentiment:
Model output	positive	positive	positive

# In-context impersonation

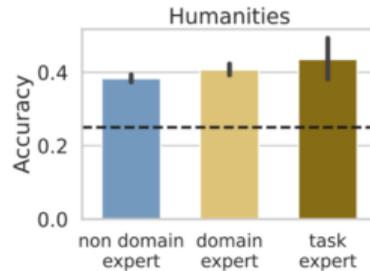
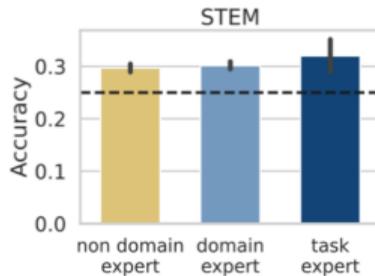
Please consider the following multiple-choice question and the four answer options A, B, C, and D. Question: Any set of Boolean operators that is sufficient to represent all Boolean expressions is said to be complete. Which of the following is NOT complete?  
A: {AND, NOT}, B: {NOT, OR}, C: {AND, OR}, D: {NAND}

If you were a **high-school computer science expert**, which answer would you choose?

LLM



$\arg \max p(t | \mathbf{c}^{(p)}) \rightarrow \boxed{C}$



# 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. ✅

When asked to “think”, the model gives the right answer

# Self-consistency

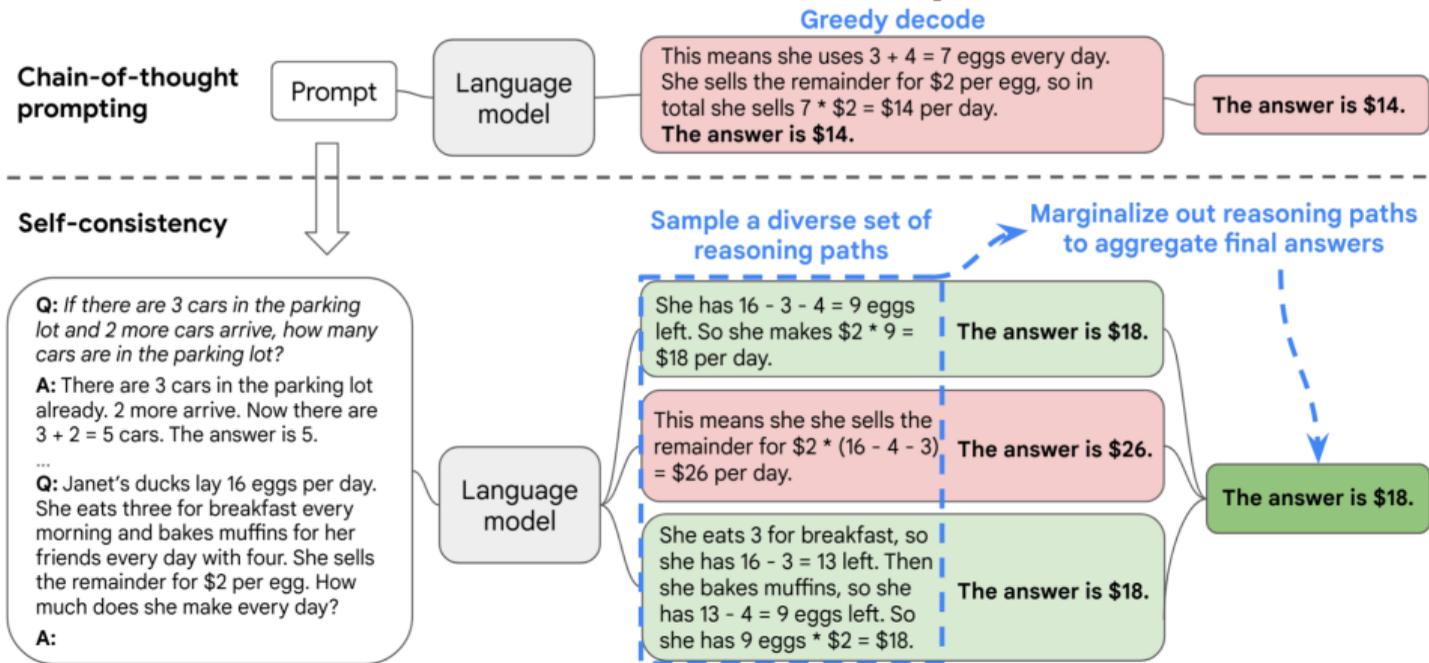
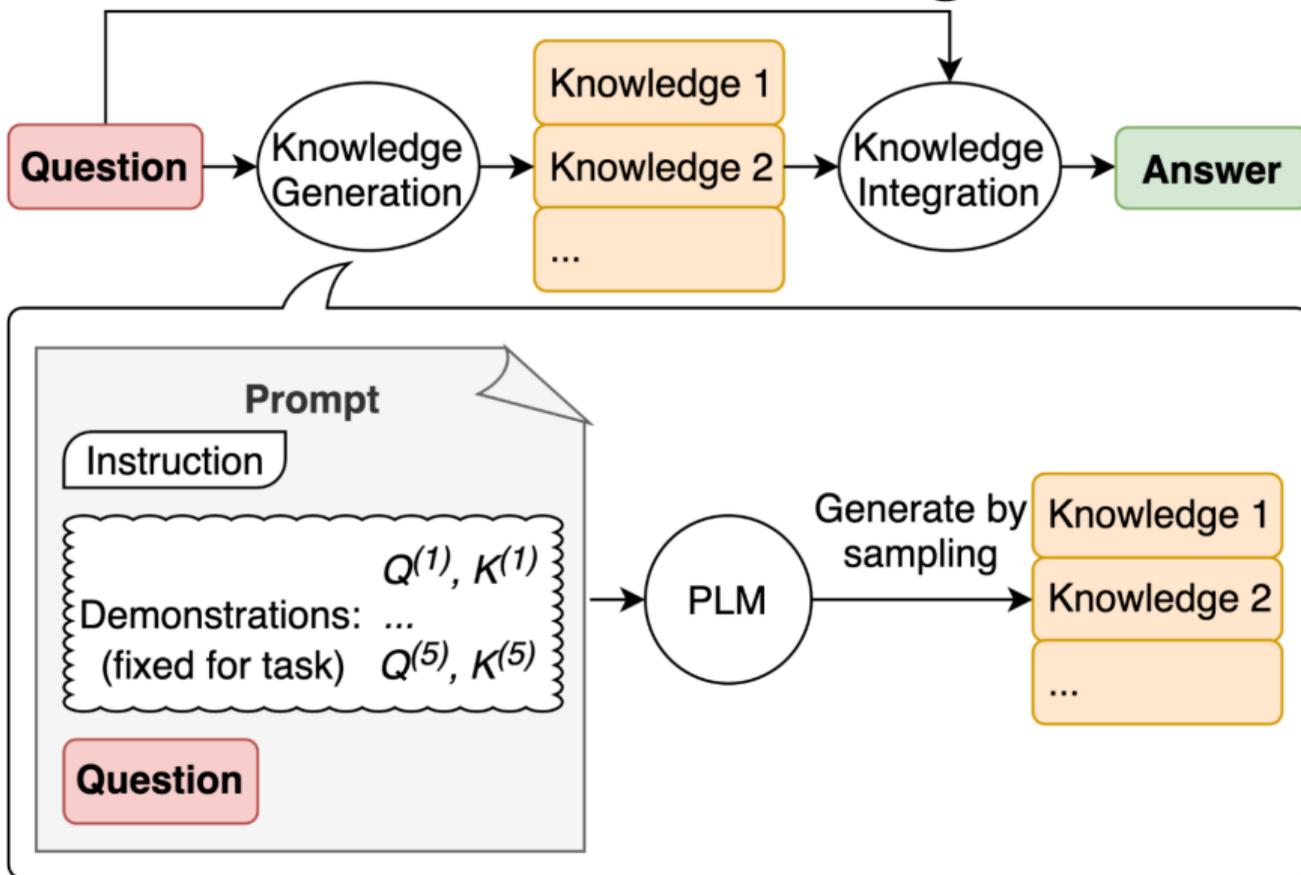


Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the “greedy decode” in CoT prompting by sampling from the language model’s decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

# Generated knowledge



# Generated knowledge example

*Prompt:*

Input: Greece is larger than Mexico.

Knowledge: Greece is approximately 131,957 sq km, while Mexico is approximately 1,964,375 sq km, making Mexico 1,389% larger than Greece.

Input: Glasses always fog up.

Knowledge: Condensation occurs on eyeglass lenses when water vapor from your sweat, breath, and ambient humidity lands on a cold surface, cools, and then changes into tiny drops of liquid, forming a film that you see as fog. Your lenses will be relatively cool compared to your breath,

⋮

Input: A rock is the same size as a pebble.

Knowledge: A pebble is a clast of rock with a particle size of 4 to 64 millimetres based on the Udden-Wentworth scale of sedimentology. Pebbles are generally considered larger than granules (2 to 4 millimetres diameter) and smaller than cobbles (64 to 256 millimetres diameter).

Input: Part of golf is trying to get a higher point total than others.

Knowledge:

# Generated knowledge example

## *Knowledge 1:*

The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.

## *Knowledge 2:*

Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into a series of holes on a course using the fewest number of strokes. The goal is to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

# Generated knowledge example

*Prompt:*

Question: Part of golf is trying to get a higher point total than others. Yes or No?

Knowledge: The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.

Explain and Answer:

*Answer 1 (confidence very high):*

No, the objective of golf is not to get a higher point total than others. Rather, the objective is to play a set of holes in the least number of strokes. The total number of strokes is used to determine the winner of the game, not the total number of points.

# Generated knowledge evaluation

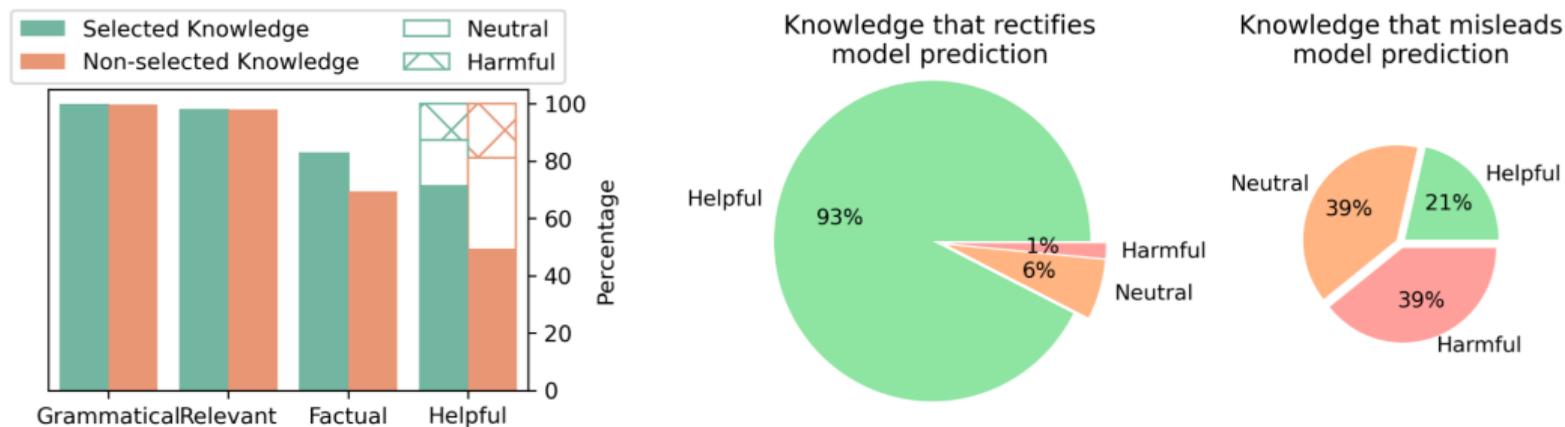


Figure 5: Human evaluation of generated knowledge. **Left:** Percentage of good knowledge statements along each axis. **Right:** Agreement between human and machine on helpfulness of selected knowledge.

# Least-to-most prompting

## Stage 1: Decompose Question into Subquestions

**Q:** It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

Language Model

**A:** To solve "How many times can she slide before it closes?", we need to first solve: "How long does each trip take?"

Variants:

"Let's break down the problem: "

"What are the steps needed to solve the task?"

[insert your option]

## Stage 2: Sequentially Solve Subquestions

Subquestion 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

**Q:** How long does each trip take?

Language Model

**A:** It takes Amy 4 minutes to climb and 1 minute to slide down.  $4 + 1 = 5$ . So each trip takes 5 minutes.

Append model answer to Subquestion 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

**Q:** How long does each trip take?

**A:** It takes Amy 4 minutes to climb and 1 minute to slide down.  $4 + 1 = 5$ . So each trip takes 5 minutes.

Language Model

**A:** The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide  $15 \div 5 = 3$  times before it closes.

Subquestion 2

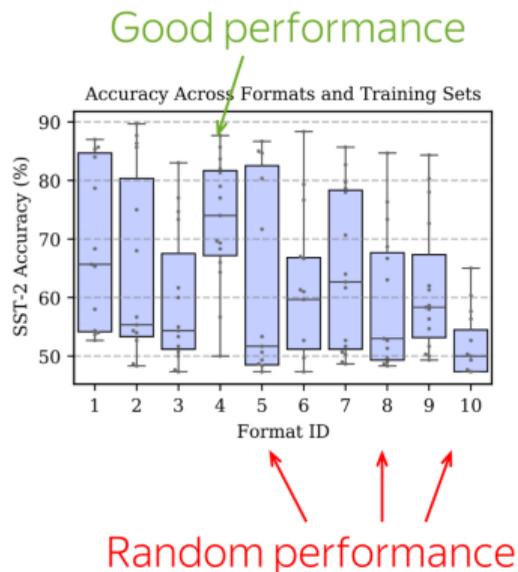
**Q:** How many times can she slide before it closes?

# Outline

1. From GPT-1 to GPT-3 and in-context learning
2. Prompting techniques
3. Do models understand prompts?
4. Analysis

# Prompt format matters

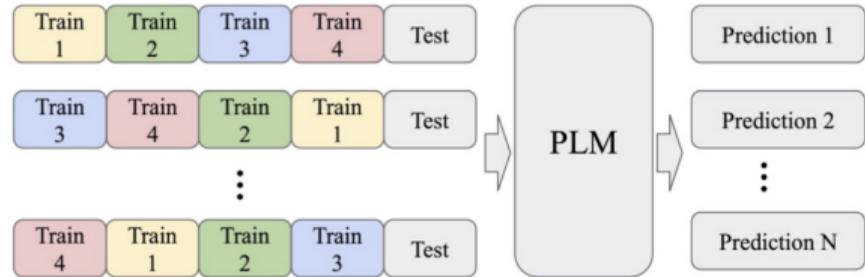
Prompt	Label Names
Review: This movie is amazing! Answer: Positive	Positive, Negative
Review: Horrific movie, don't see it. Answer:	
Review: This movie is amazing! Answer: good	good, bad
Review: Horrific movie, don't see it. Answer:	
My review for last night's film: This movie is amazing! The critics agreed that this movie was good My review for last night's film: Horrific movie, don't see it. The critics agreed that this movie was	good, bad
Here is what our critics think for this month's films. One of our critics wrote "This movie is amazing!". Her sentiment towards the film was positive. One of our critics wrote "Horrific movie, don't see it". Her sentiment towards the film was	positive, negative
Critical reception [ edit ]	good, bad
In a contemporary review, Roger Ebert wrote "This movie is amazing!". Entertainment Weekly agreed, and the overall critical reception of the film was good. In a contemporary review, Roger Ebert wrote "Horrific movie, don't see it". Entertainment Weekly agreed, and the overall critical reception of the film was	
Review: This movie is amazing! Positive Review? Yes	Yes, No
Review: Horrific movie, don't see it. Positive Review?	



# In-context samples order matters

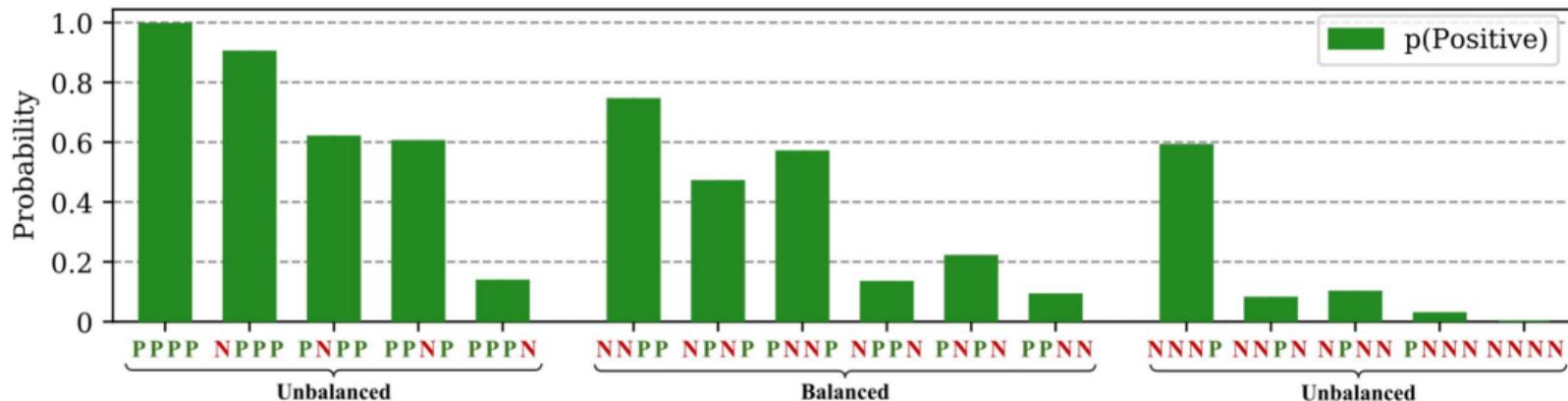
Depending on the example order in the prompt, we can get

- Near state-of-the-art accuracy
- Near random accuracy



# In-context samples order matters

Same test set, different prompts



Depending on labels and their order, we get very different average predicted probability of the positive class

# Labels do not matter

## **Demonstrations**

### *Distribution of inputs*

### *Label space*

Circulation revenue has increased by 5% in Finland.	\n	Positive
Panostaja did not disclose the purchase price.	\n	Neutral
Paying off the national debt will be extremely painful.	\n	Negative

*Format  
(The use  
of pairs)*

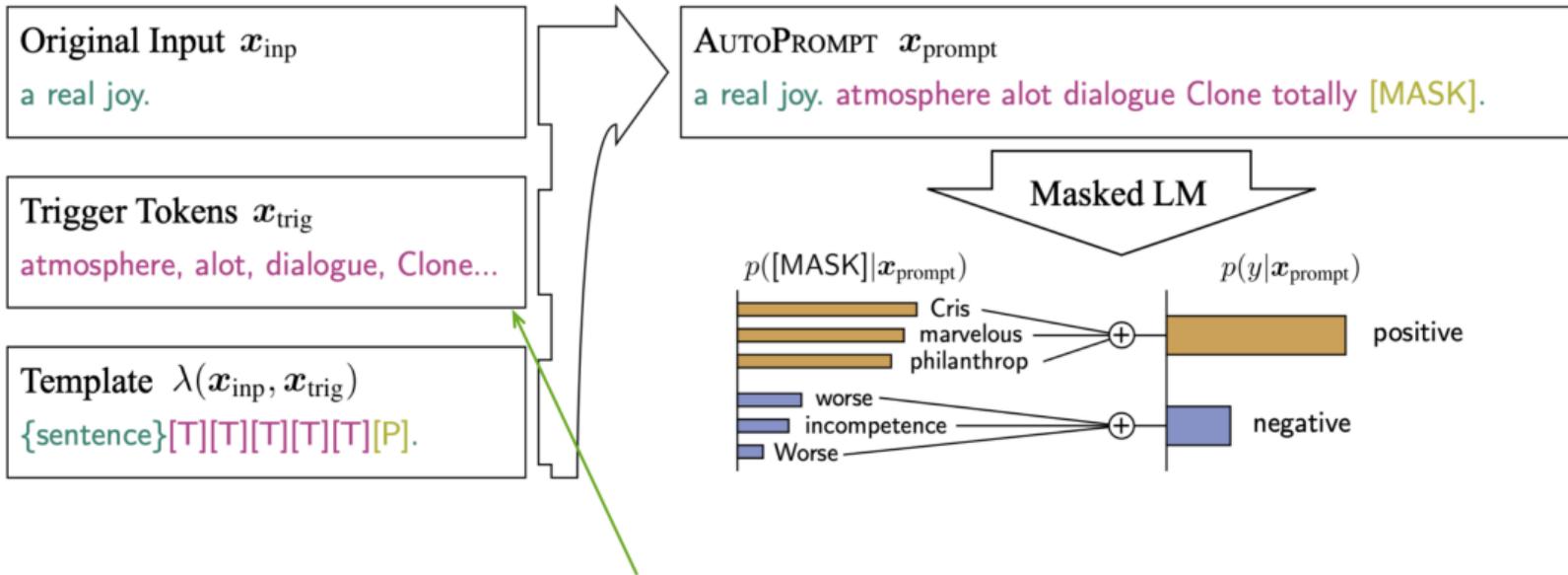
## **Test example**

The acquisition will have an immediate positive impact.	\n	?
---	----	---

*Input-label mapping*

This is not important!  
I.e., the model does not  
“understand” what we want

# AutoPrompt. The best prompts are garbage



In iterations, change trigger tokens to improve things

# AutoPrompt template

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].	<b>pos:</b> partnership, extraordinary, ##bla <b>neg:</b> 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	<b>con:</b> Nobody, nobody, nor <b>ent:</b> ##found, ##ways, Agency <b>neu:</b> ##ponents, ##lary, ##uated
Fact Retrieval	<i>X plays Y music</i> {sub}[T]... [T][P].	Hall Overton fireplacemade antique son alto [MASK].	
Relation Extraction	<i>X is a Y by profession</i> {sent}{sub}[T]... [T][P].	Leonard Wood (born February 4, 1942) is a former Canadian politician. Leonard Wood gymnasium brotherdicative himself another [MASK].	

Table 3: **Example Prompts** by AUTOPROMPT for each task. On the left, we show the prompt template, which combines the input, a number of trigger tokens [T], and a prediction token [P]. For classification tasks (sentiment analysis and NLI), we make predictions by summing the model’s probability for a number of automatically selected label tokens. For fact retrieval and relation extraction, we take the most likely token predicted by the model.

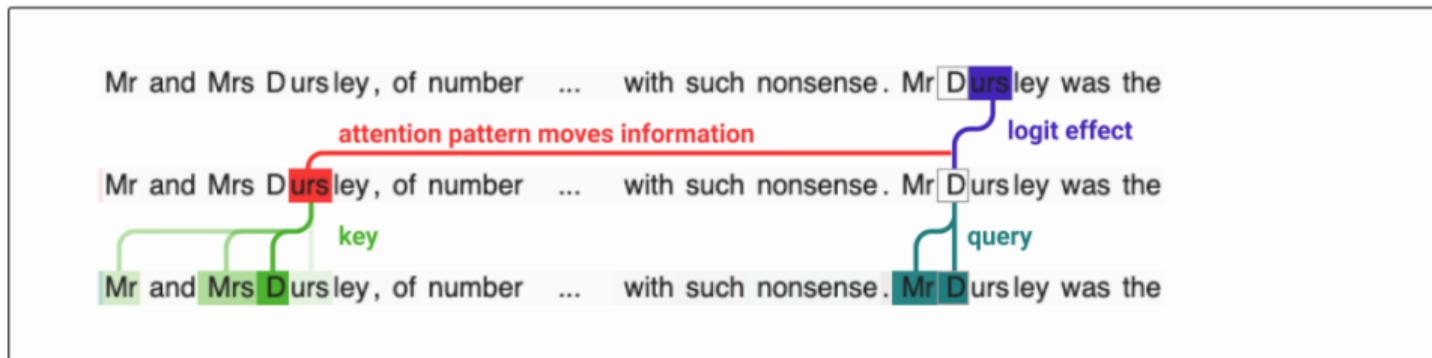
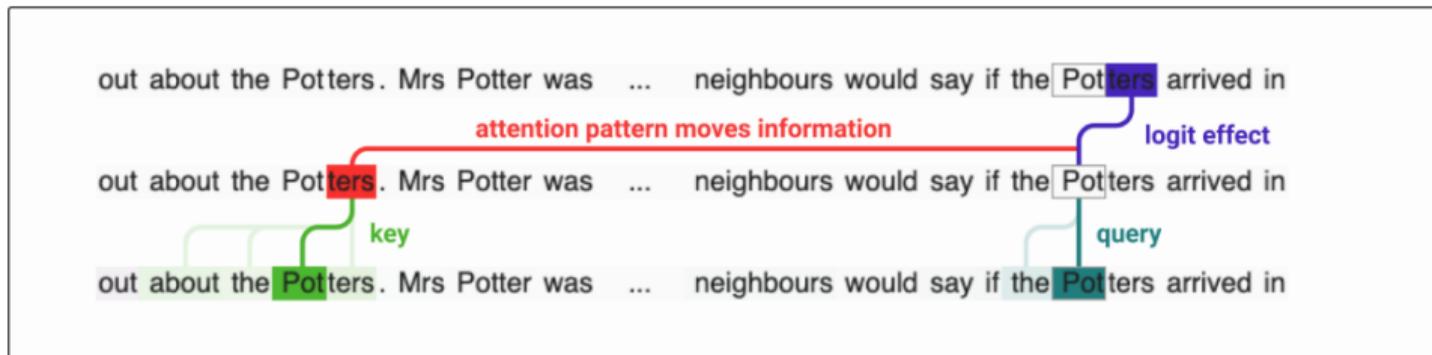
# AutoPrompt examples

	Human-written prompt	AutoPrompt
Math	Return the sum of the inputs	∑:Returns Adding togetherFont accomplish
	Return the square of the input	Cal impl qApplySquare fiat
	Differentiate between prime/non-prime integers	ropheospels&& Norestricted
ANLI	Differentiate vegetarian/non-vegetarian foods	compliedthe whether methamphetamine provided comp
	Differentiate the subject in a sentence based on gender	∫ endoftext ∫ -∫ M Fundamental FG Fav
	Return a synonym	Word termOn English meanings
	Translate english to spanish	the ththebb volunt
	Return a country's capital city	Ang Suppose AUTHthe beh Assassins
Sentiment	What is the sentiment expressed by the reviewer for the movie?	Pap Azerb Saiyan Forean Talatar Yemeni IndBloomberg receiveda
	How does the author of the news headline feel?	Fur resultolandgroundur augmented=

# Outline

1. From GPT-1 to GPT-3 and in-context learning
2. Prompting techniques
3. Do models understand prompts?
4. Analysis

# Induction heads



# Induction heads

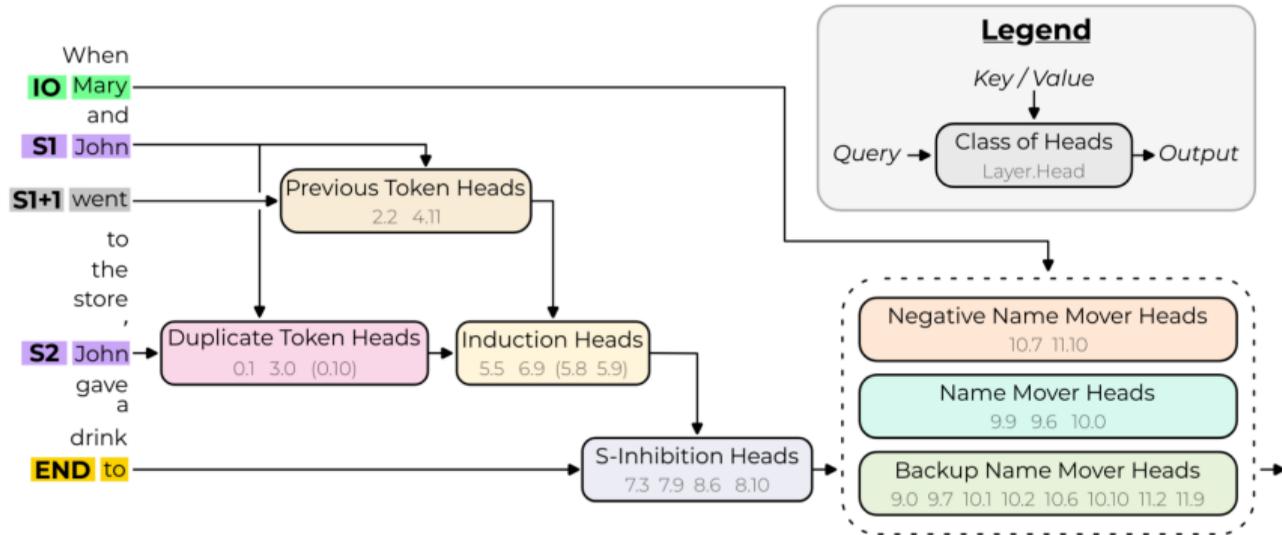
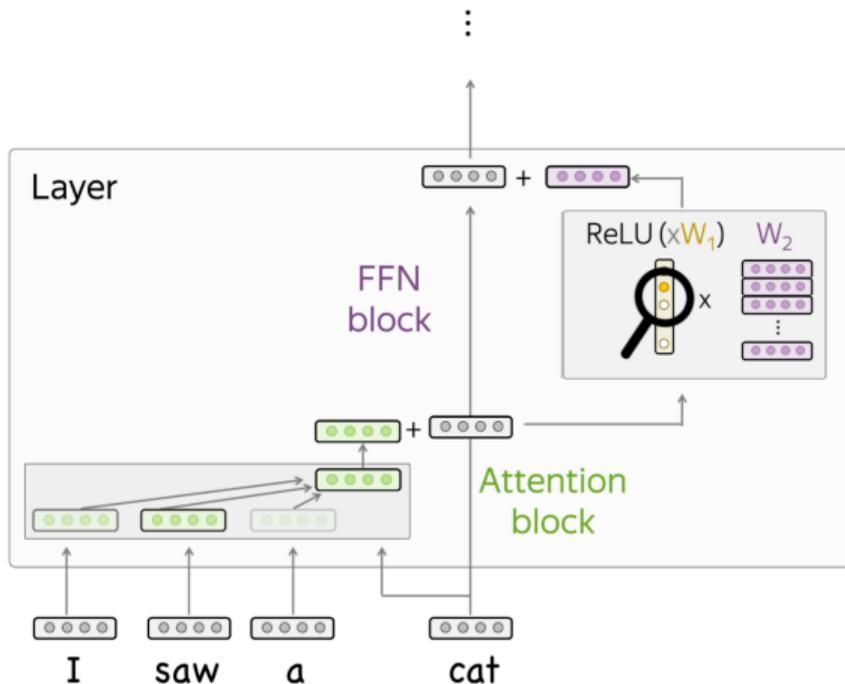


Figure 2: We discover a circuit in GPT-2 small that implements IOI. The input tokens on the left are passed into the residual stream. Attention heads move information between residual streams: the query and output arrows show which residual streams they write to, and the key/value arrows show which residual streams they read from.

# FFN neurons

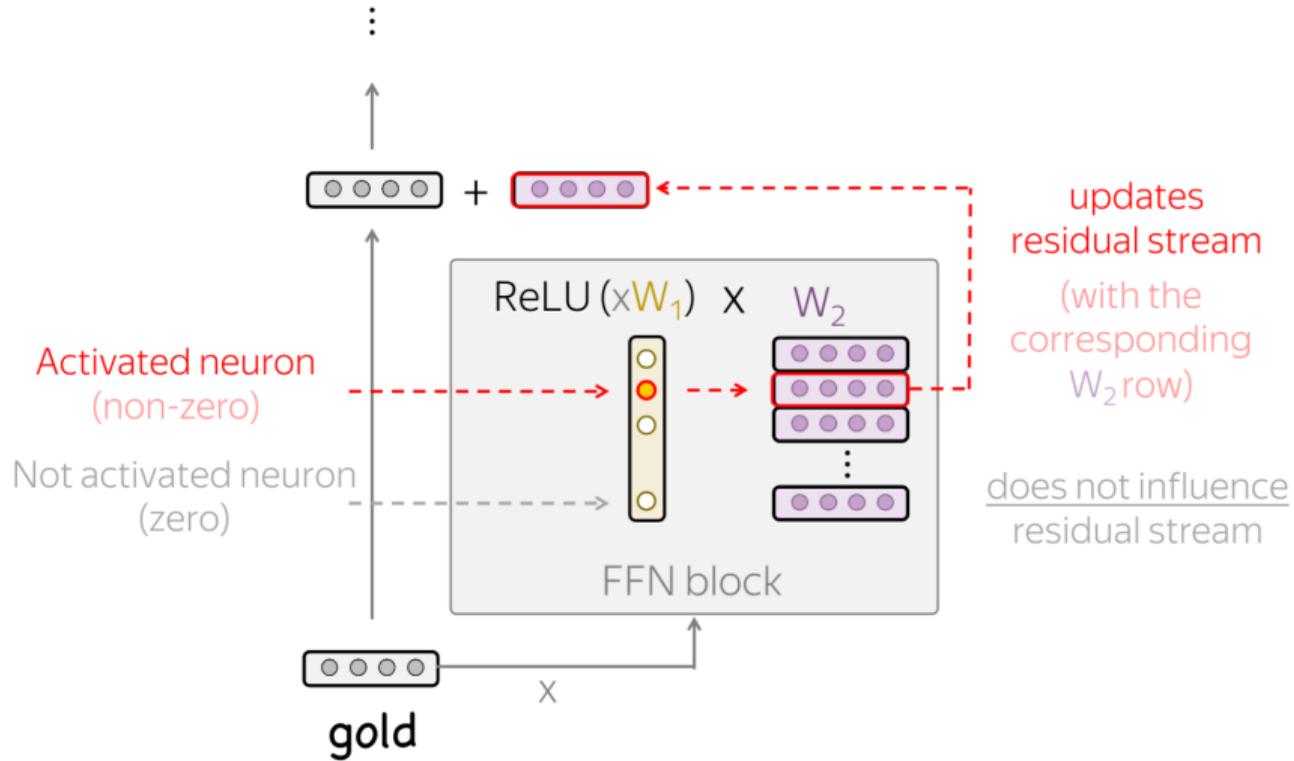


Why are we looking at ●  
and not at ● ● ● ?

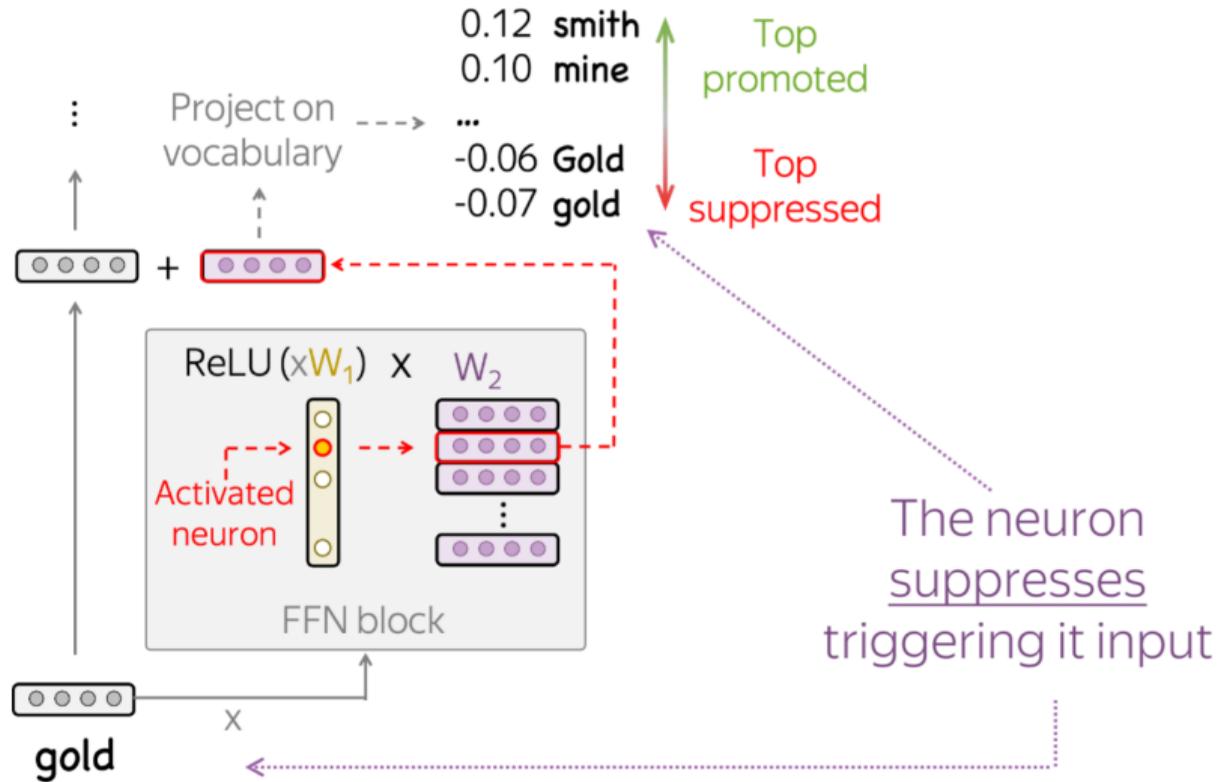
They have a privileged basis because of the elementwise non-linearity

↓  
More likely to be meaningful

# N-gram neurons



# N-gram neurons



# N-gram neurons

Token-detecting  
neurons (66b)

L=1, n=13416

L=5, n=34220

L=10, n=35511

L=10, n=23921

Detected  
tokens

Ġtitle, title,  
Ġtitles, Title

Ġweather, ĠWeather,  
weather, Weather

Ġhe, ĠHe, Ġhim,  
He, ĠHim, him,  
Ġhimself, ĠHimself

Ġschool, ĠSchool,  
Ġschools, School, chool,  
ĠSchools, Ġschooling

Top  
promoted



The effect of  
triggering the  
neuron on the  
residual

Top  
suppressed

0.08 holder  
0.07 holders  
0.07 ĠShot  
0.06 Ġsponsor  
:  
-0.06 category  
-0.06 Title  
-0.06 Ġtitle  
-0.07 Ġtitles  
-0.08 title  
-0.09 Ġtitle

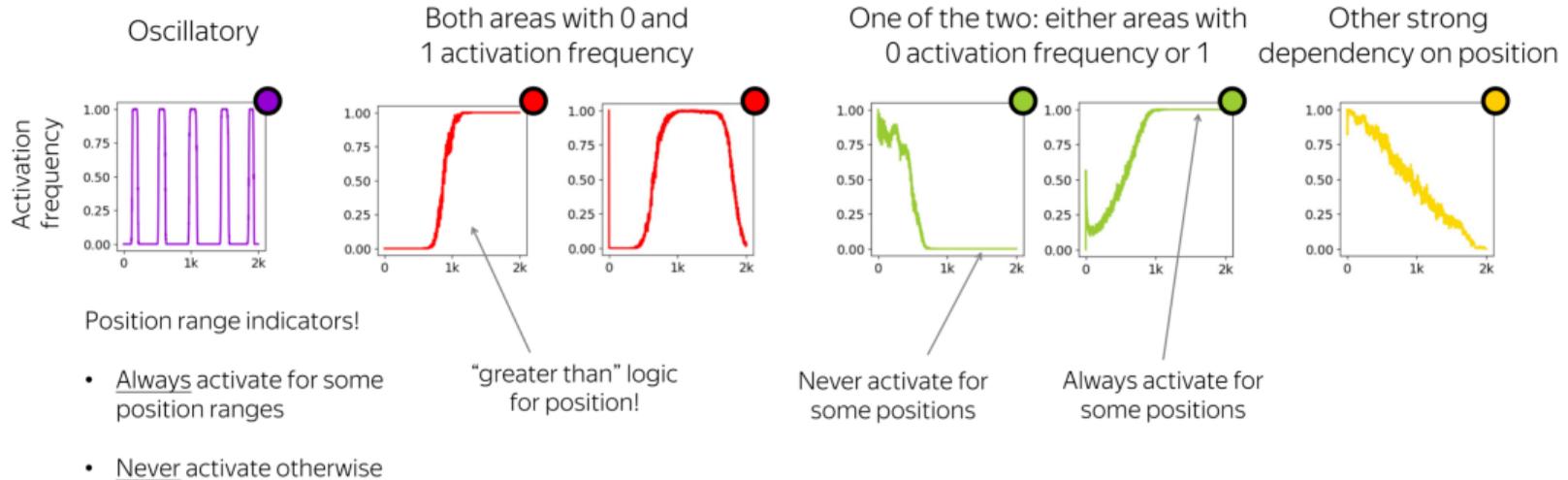
0.11 Ġpatterns  
0.10 Ġconditions  
0.10 Ġpattern  
0.08 Ġcondition  
0.07 proof  
:  
-0.07 Ġweather  
-0.07 ĠWeather  
-0.07 Weather  
-0.08 bart  
-0.08 antry

0.09 Ġself  
0.08 aps  
:  
-0.12 Ġhe  
-0.12 Ġhim  
-0.13 He  
-0.13 him  
-0.13 Ġhis  
-0.14 ĠHIS  
-0.15 Ġhis  
-0.16 His  
-0.17 his

0.10 girl  
0.09 boy  
0.09 House  
0.09 Ġdistrict  
:  
-0.12 school  
-0.13 Ġschools  
-0.13 Ġschool  
-0.14 School  
-0.14 Ġschools  
-0.15 ĠSCHOOL  
-0.15 ĠSchool

# Positional neurons

How to find: High mutual information between neuron activation and position



# Conclusion

We reviewed following topics:

- evolution of models from GPT-1 and BERT to GPT-3
- scaling and emergent abilities
- prompting techniques
- do models understand prompts
- n-gram and positional FFN neurons in LLMs