



HOW TO WORK WITH LLMs?

Sandeep Soni

10/17/2024

STORY SO FAR

- **Language modeling:** Probabilistic model over natural language sequences.
- **Objective:** Predict word based on some context i.e $P(w|C)$
- **Variations:** Autoregressive LMs, masked LMs, etc
- **Techniques:** Count and normalize, skip-gram, RNN, Transformers

“How to use language models?”

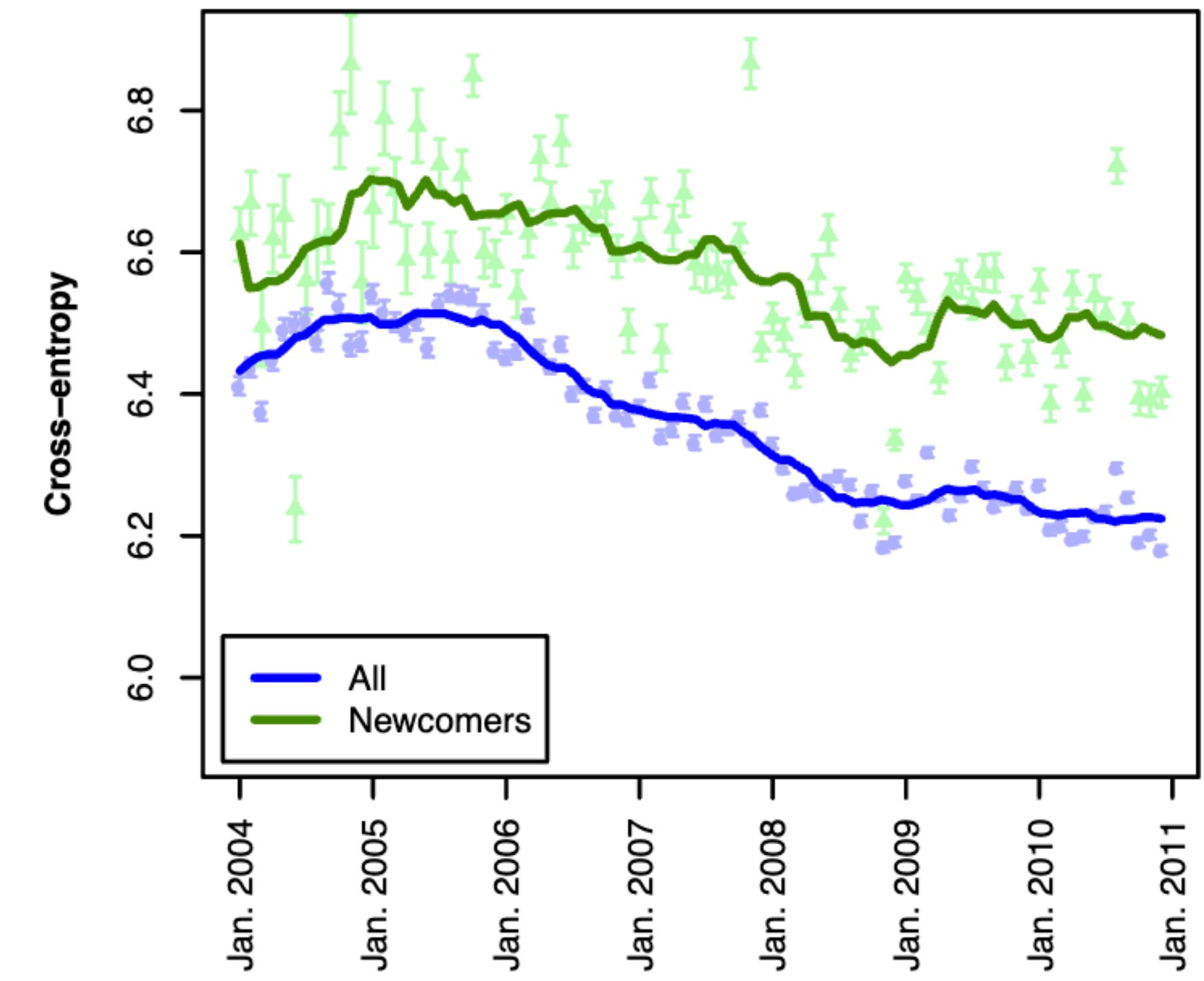
N-GRAM LANGUAGE MODELS

- N-gram language models are limited in predictive capability
- But! They are easy to construct, deploy, and use for shallow analysis.

- Whose language in a community is more conventional?

- Whose language in a community is more conventional?
- If a language model is able to predict well then language is normative

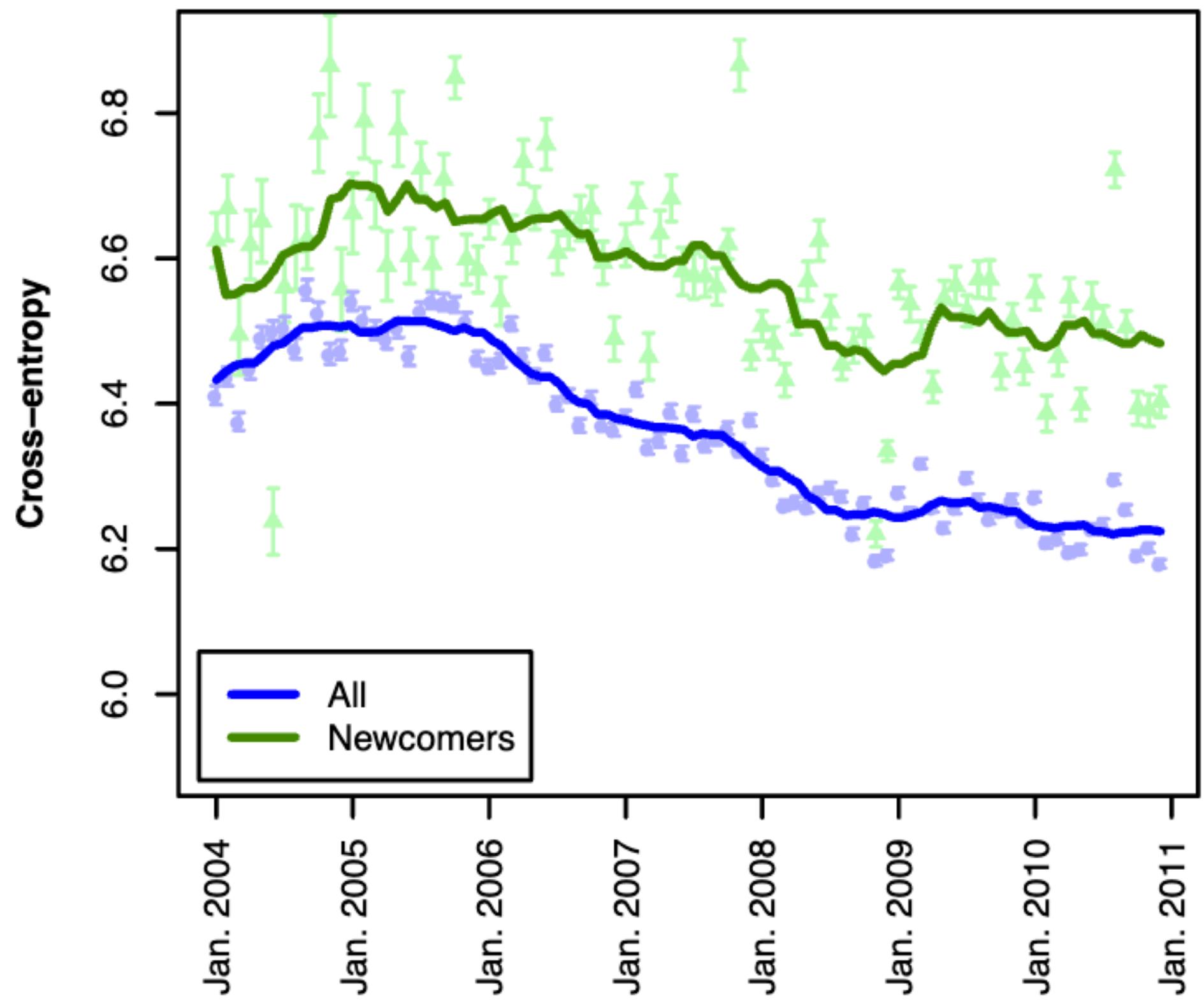
- Whose language in a community is more conventional?
- If a language model is able to predict well then language is normative



(a) BeerAdvocate

Source: CDNM (2013)

- Whose language in a community is more conventional?
- If a language model is able to predict well then language is normative
- The language of newcomers is less conventional → change is driven by newcomers



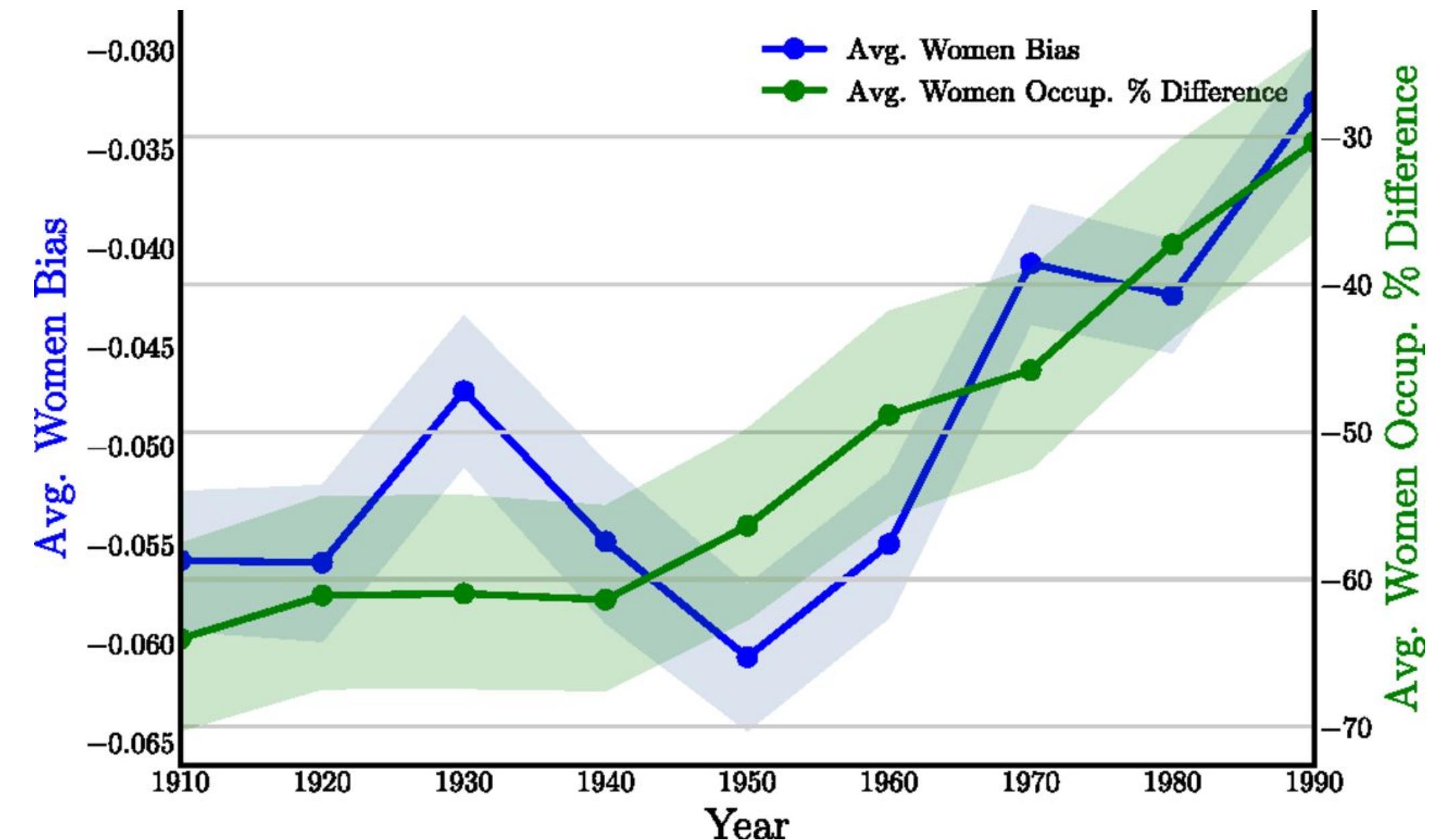
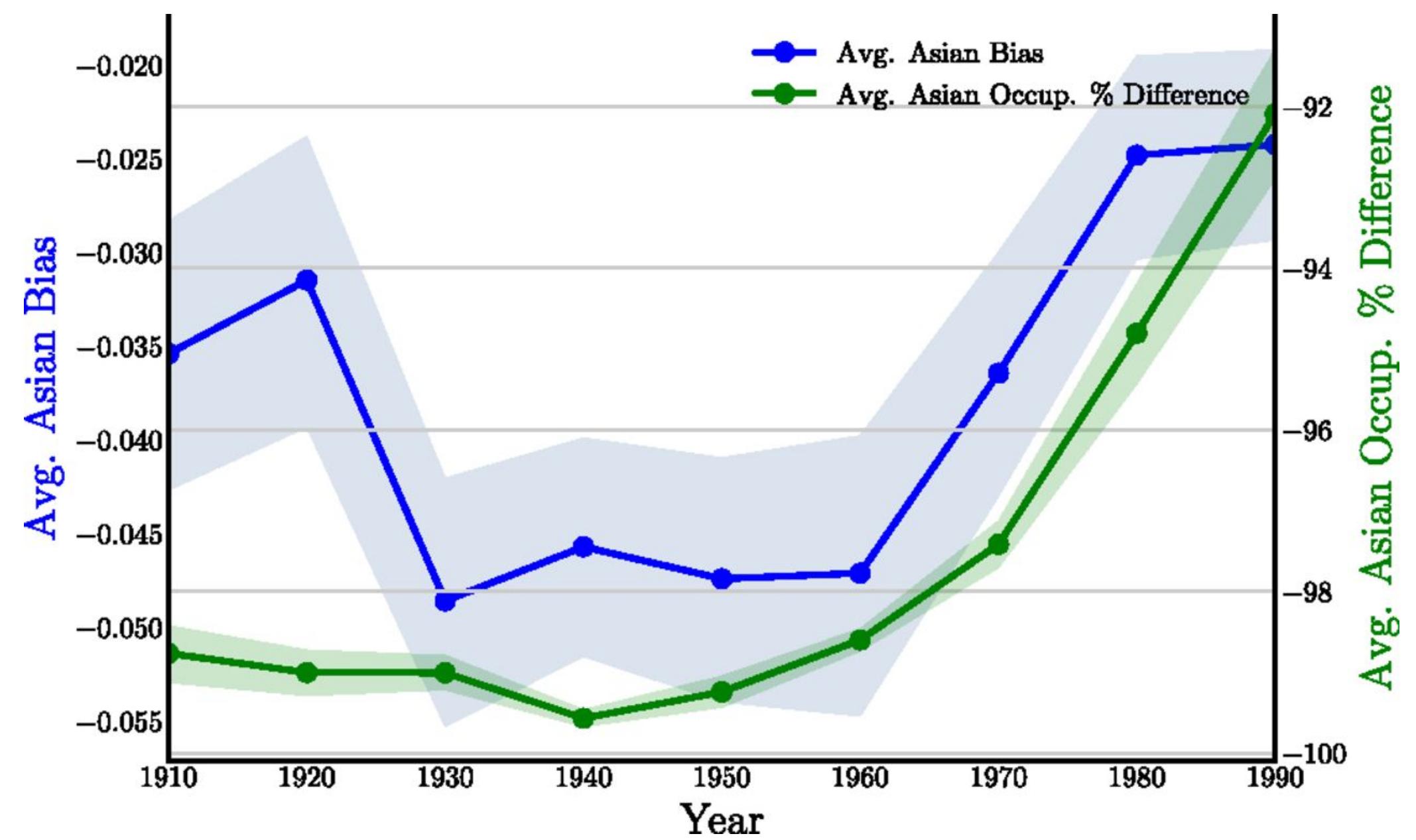
(a) BeerAdvocate

Source: CDNM (2013)

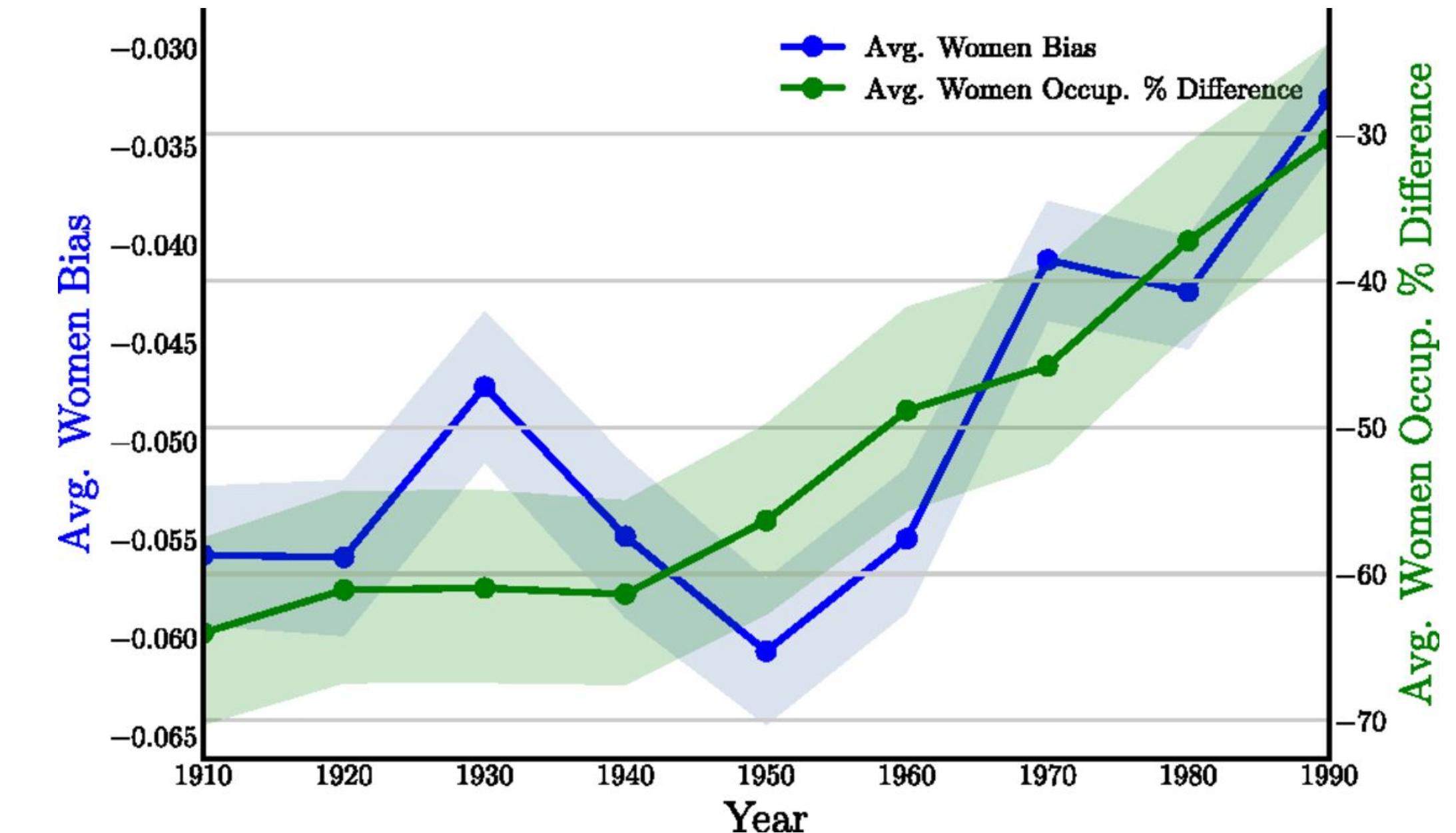
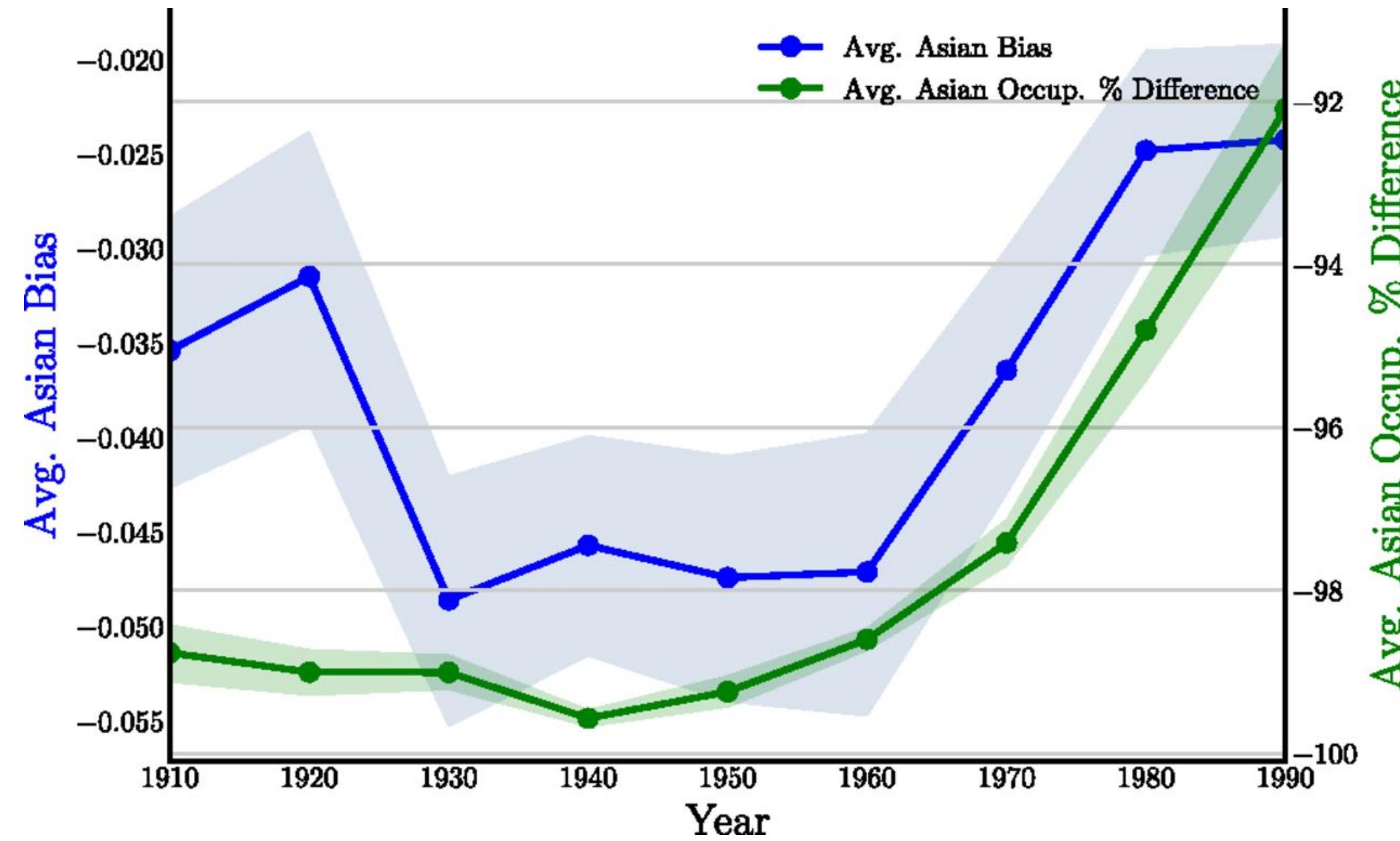
SKIPGRAM LANGUAGE MODELS

- Skipgram models were useful in learning representations of linguistic units
- These representations were found to be useful for prediction tasks as well as analysis

- How is the bias against minorities changing over time?



- How is the bias against minorities changing over time?



- How is the bias against minorities changing over time?
- Word2vec vectors can be learned and compared over time to measure bias

BERT

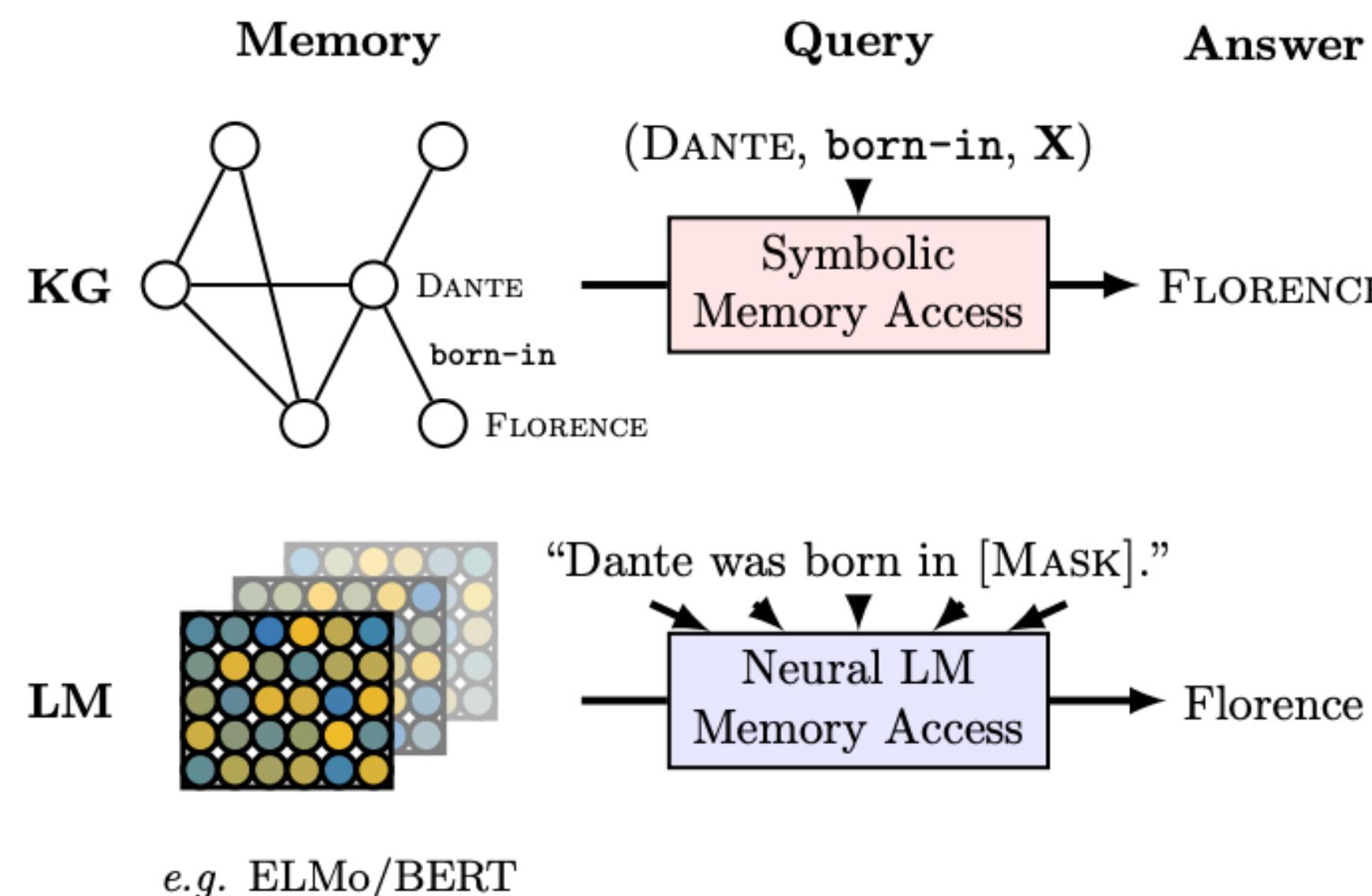
- Masked LM

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

- The ____ ran → BERT → foxes

- Pretrained on billions of tokens from wikipedia and books

LANGUAGE MODELS ARE KNOWLEDGE BASES



Query	Answer	Generation
Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], Florence [-1.8], Naples [-1.9], Milan [-2.4], Bologna [-2.5]
Adolphe Adam died in ____.	Paris	Paris [-0.5], London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], dog [-2.4], cattle [-4.3], sheep [-4.5]
The official language of Mauritius is ____.	English	English [-0.6], French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], midfielder [-2.4], forward [-2.4], midfield [-2.7]
Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], Hamburg [-7.5], Ludwig [-7.5]
The original language of Mon Oncle Benjamin is ____.	French	French [-0.2], Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
Dani Alves plays with ____.	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
Paul Toungui is a ____ by profession.	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], sodium [-2.5], zinc [-2.8], salt [-2.9]

TRANSFER LEARNING

- Take a **pretrained** model and **finetune** it for a specific task
- Pretrained model gains knowledge of language and fine-tuning helps solve a specific task

- Can BERT highlight key points of a patient visit?
- You can continue to pretrain BERT model to predict masked words in patient visitation summaries

review of systems : gen : no weight loss or gain , good general state of health , no weakness , no fatigue , no fever , good exercise tolerance , able to do usual activities . heent : head : no headache , no dizziness , no lightheadness eyes : normal vision , no redness , no blind spots , no floaters . ears : no earaches , no fullness , normal hearing , no tinnitus . nose and sinuses : no colds , no stuffiness , no discharge , no hay fever , no nosebleeds , no sinus trouble . mouth and pharynx : no cavities , no bleeding gums , no sore throat , no hoarseness . neck : no lumps , no goiter , no neck stiffness or pain . ln : no adenopathy cardiac : no chest pain or discomfort no syncope , no dyspnea on exertion , no orthopnea , no pnd , no edema , no cyanosis , no heart murmur , no palpitations resp : no pleuritic pain , no sob , no wheezing , no stridor , no cough , no hemoptysis , no respiratory infections , no bronchitis .

- ... Or finetune BERT on the ICD codes classification task

	AUC	
	Micro	Macro
Logistic Reg (max length 512)	0.932	0.815
Multi-head Attn (max length 512)	0.941	0.859
BERT (max length 512)	0.954	0.895
BERT (max length 1024)	0.955	0.898
Finetuned BERT (max length 1024)	0.958	0.903
BioBERT	0.960	0.908
clinical BERT	0.961	0.904
EHR BERT Small (max length 512)	0.959	0.897
EHR BERT Small (max length 1024)	0.965	0.918
EHR BERT Small + XML (max length 1024)	0.968	0.924
EHR BERT Big (max length 512)	0.964	0.917
EHR BERT Big (max length 1024)	0.968	0.925
EHR BERT Big + XML (max length 512)	0.967	0.919
EHR BERT Big + XML (max length 1024)	0.970	0.927

Source: Zhang et. al. (2020)

FINETUNING FOR CLASSIFICATION

FINETUNING FOR CLASSIFICATION

- Classification : $P(y|x)$

FINETUNING FOR CLASSIFICATION

- Classification : $P(y|x)$
- Logistic Regression : $\text{softmax}(w^T \cdot x)$

FINETUNING FOR CLASSIFICATION

- Classification : $P(y|x)$
- Logistic Regression : $\text{softmax}(w^T \cdot x)$
- BERT-based classification : $\text{softmax}(\text{BERT}(x))$

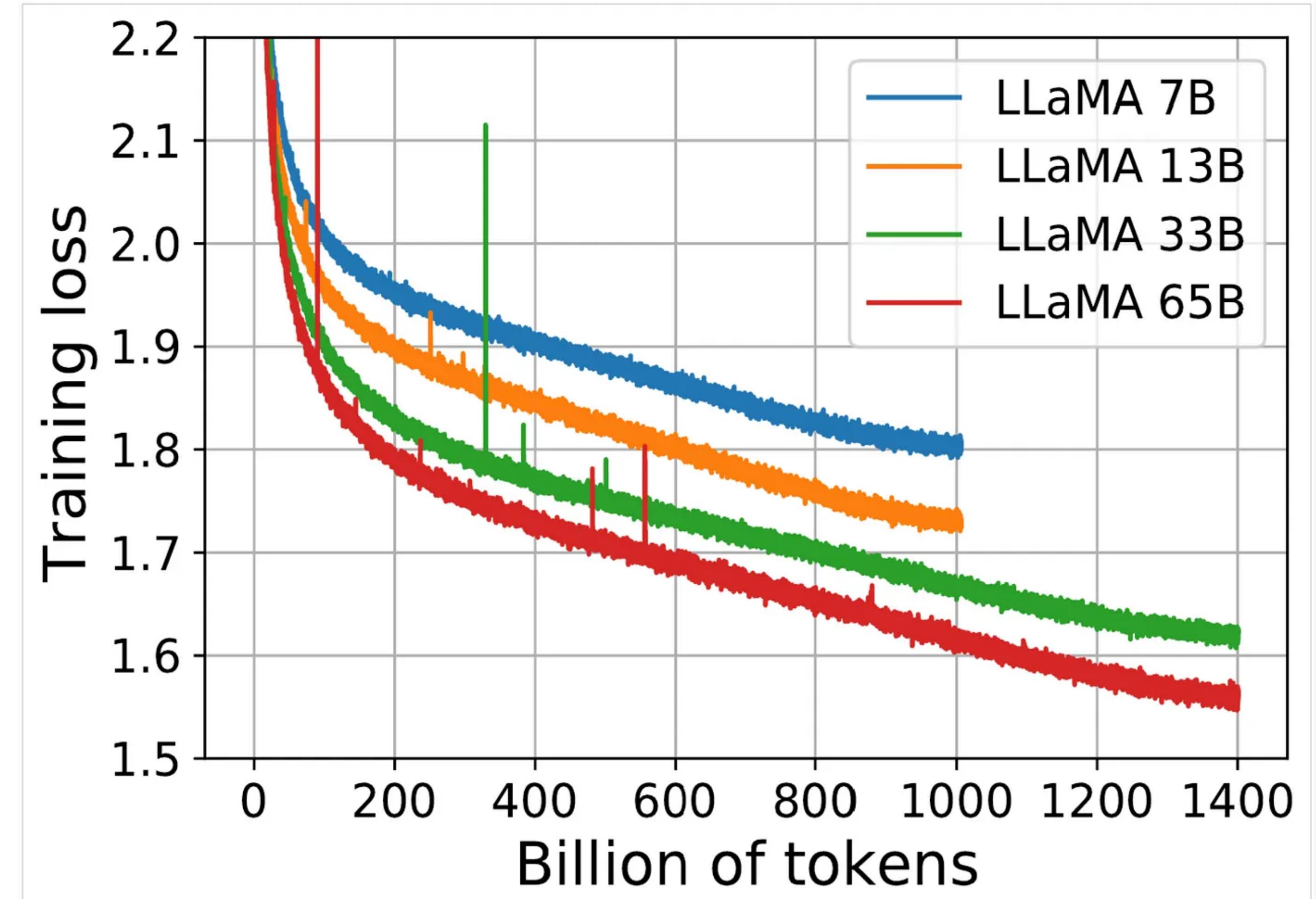
FINETUNING FOR CLASSIFICATION

- Classification : $P(y|x)$
- Logistic Regression : $\text{softmax}(w^T \cdot x)$
- BERT-based classification : $\text{softmax}(\text{BERT}(x))$

As all model parameters need to be updated at the time of learning, finetuning can be expensive!

“What happened next?”

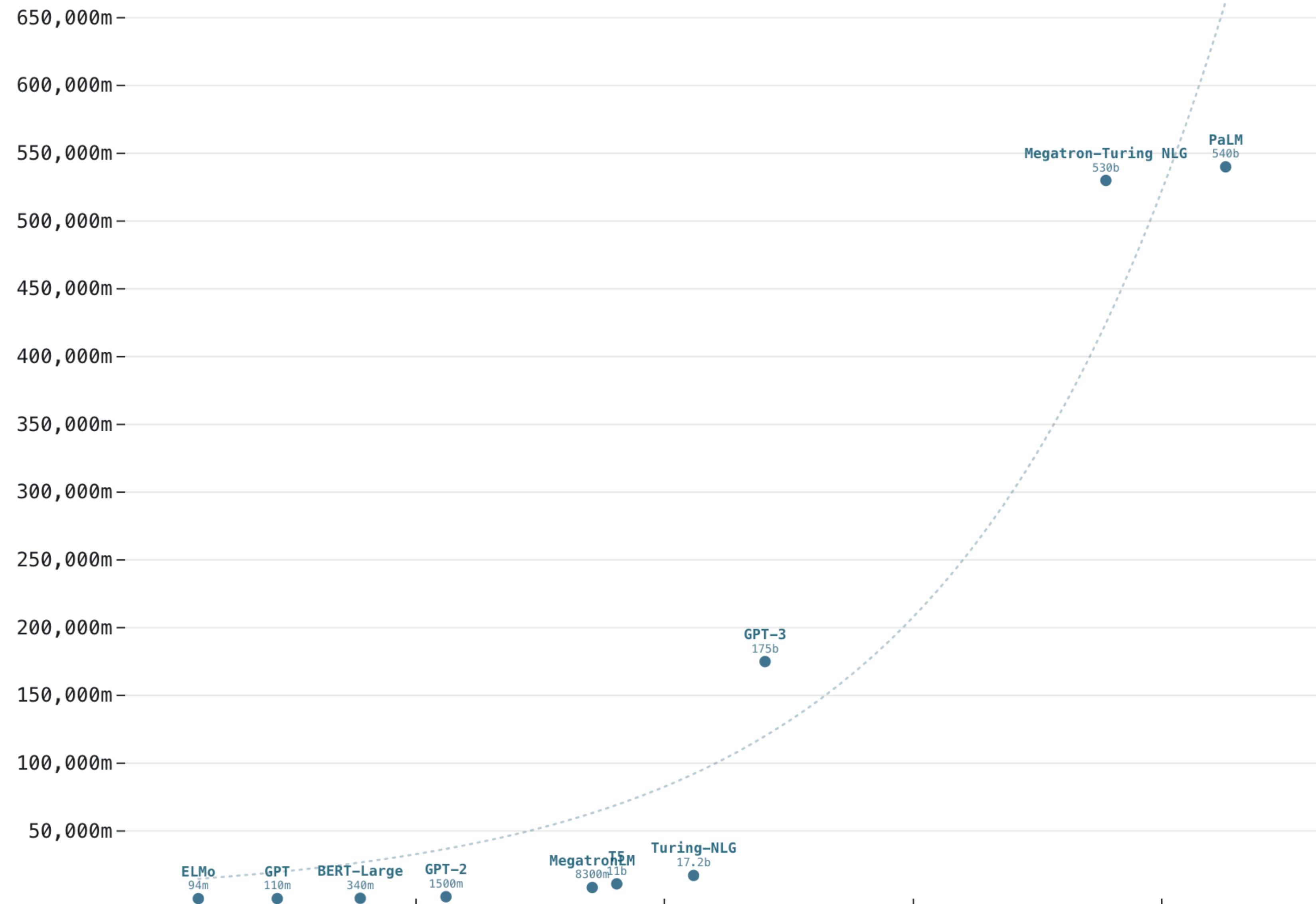
- Large language models are high capacity models that continue to improve with more data



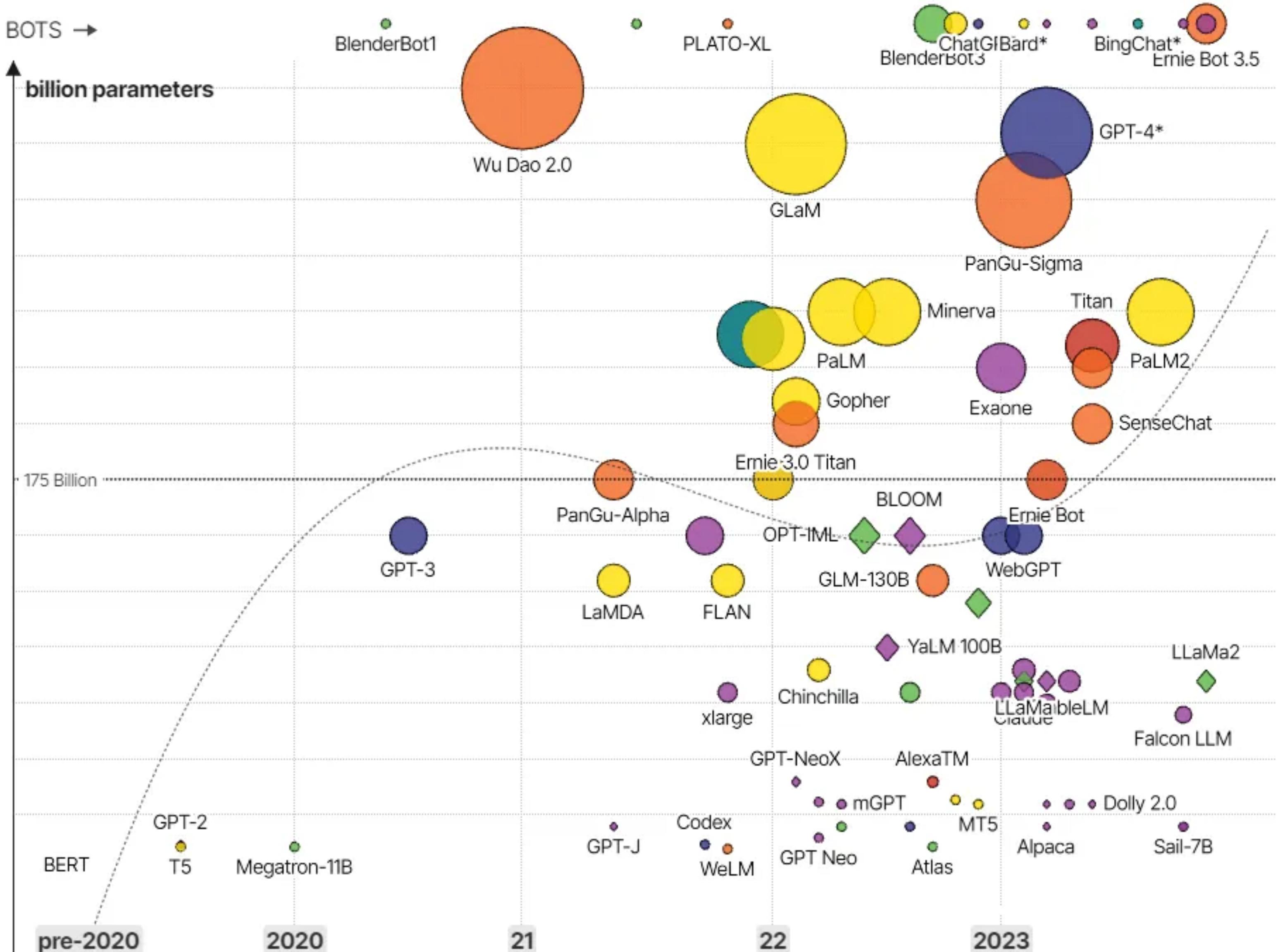
SCALE

- BERT contained a few million parameters
- Contemporary LLMs are massive in terms of size

Number of Parameters (in Millions)



Large Language Models are models with millions or billions of parameters



David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 27th Jul 23

source: news reports, [LifeArchitect.ai](#)
* = parameters undisclosed // see [the data](#)

Source: informationisbeautiful.net

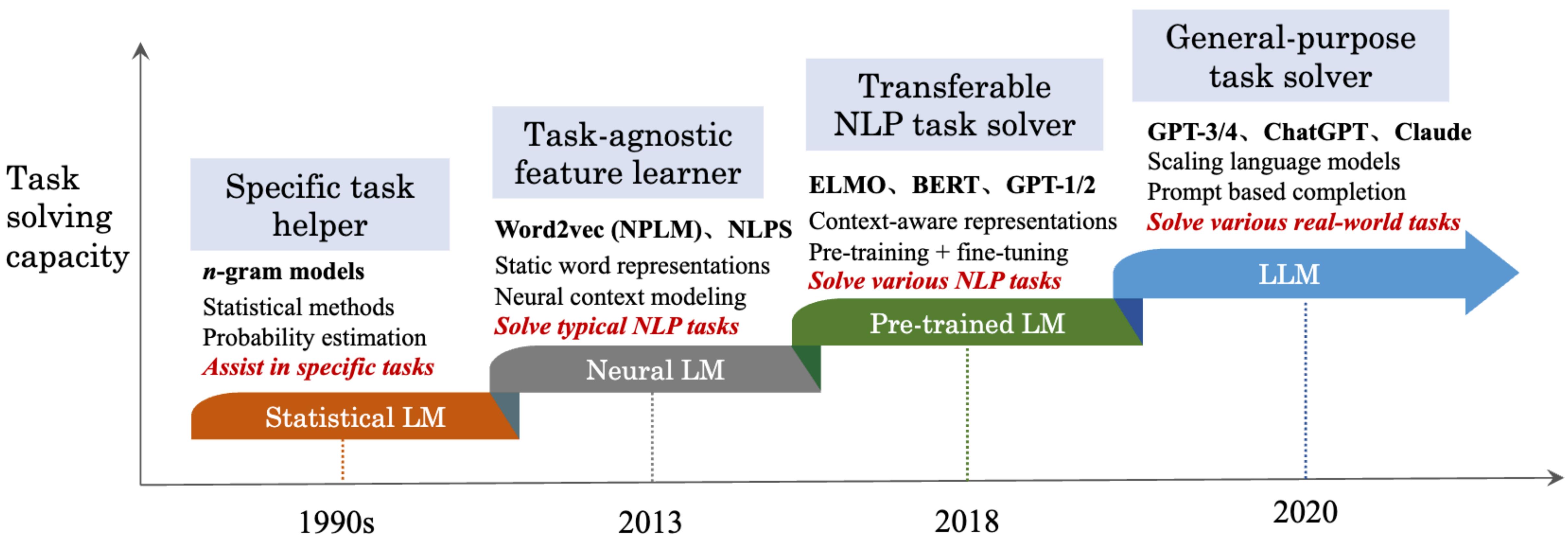
“Extrapolating the spectacular performance of GPT-3 into the future suggests that the answer to life, the universe and everything is just 4.398 trillion parameters.”

“Extrapolating the spectacular performance of GPT-3 into the future suggests that the answer to life, the universe and everything is just 4.398 trillion parameters.”

–Geoff Hinton

ABILITY

- Earlier language models were specific to a task
- Large language models are seen as general-purpose solvers



Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 (2023).

“What can we do with large language models?”

GENERATION

- Language models are generative models and hence they can be used to generate conversational text



I'm going to cook for my date who claims to be a picky eater. Can you recommend me a dish that's easy to cook?



Certainly! Cooking for a picky eater can be a bit of a challenge, but there are some dishes that tend to be universally well-received and are relatively easy to prepare. One option is to make a classic pasta dish, such as Spaghetti with Tomato Sauce and Meatballs. It's a familiar and comforting meal that most people enjoy. Here's a simple recipe:

Ingredients:

- 8 oz (about 225g) spaghetti
- 1 cup of your favorite tomato sauce (store-bought or homemade)
- 4-6 pre-cooked meatballs (you can buy frozen ones or make your own)
- Grated Parmesan cheese (optional, for topping)
- Fresh basil leaves (optional, for garnish)

Instructions:

1. Cook the spaghetti according to the package instructions until al dente. Drain and set aside.
2. While the spaghetti is cooking, heat your tomato sauce in a saucepan over low heat. If your date is particularly picky, you can use a smooth tomato sauce without visible chunks of vegetables or herbs.
3. Heat the pre-cooked meatballs according to the package instructions. You can heat them in the tomato sauce or in the oven.
4. Once everything is ready, plate the cooked spaghetti, top it with the warm tomato sauce, and arrange the meatballs on top.
5. If your date enjoys cheese, you can sprinkle some grated Parmesan over the top. You can also add a few fresh basil leaves for a touch of freshness.

This simple Spaghetti with Tomato Sauce and Meatballs is a crowd-pleaser and can be customized to suit your date's preferences. If your date has specific dietary restrictions or dislikes certain ingredients, be sure to ask them in advance to make any necessary adjustments. Enjoy your meal, and I hope your date goes well!

Regenerate

IS EVERYTHING JUST LANGUAGE MODELING?

- If you can construct a suitable prompt, many tasks can be framed as language modeling
 - The director of Avatar is _____
 - 3 times 20 plus 45 is _____
 - Is “sit down” rude than “please, sit down”? _____

PROMPTING

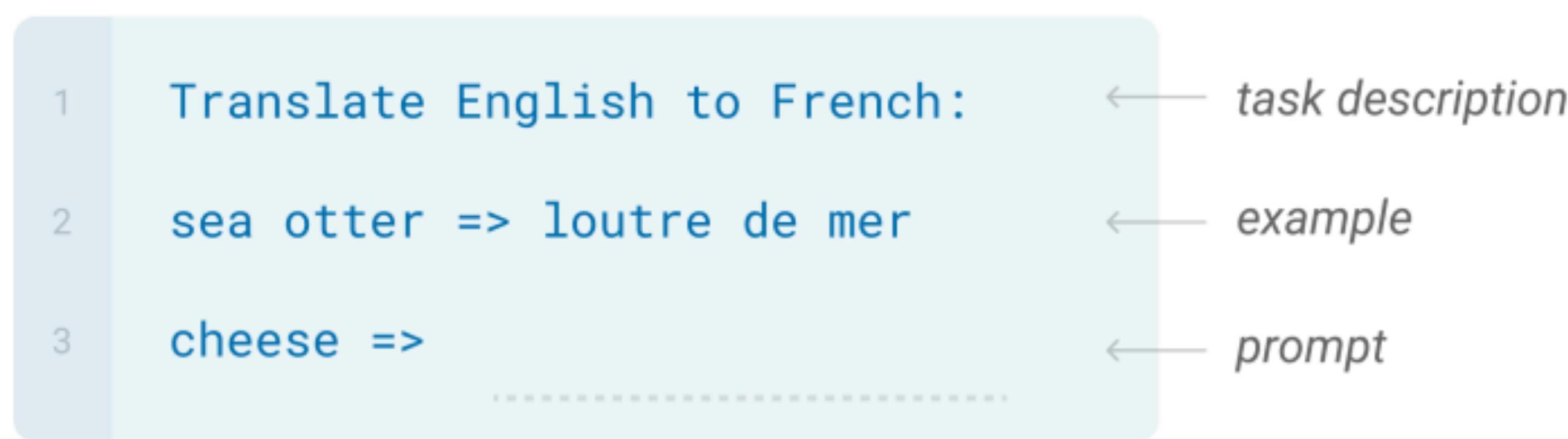
- Design a prompt for a task
- Generate a response using prompt as input to LLM
- Extract answer from response

ZERO-SHOT PROMPTING



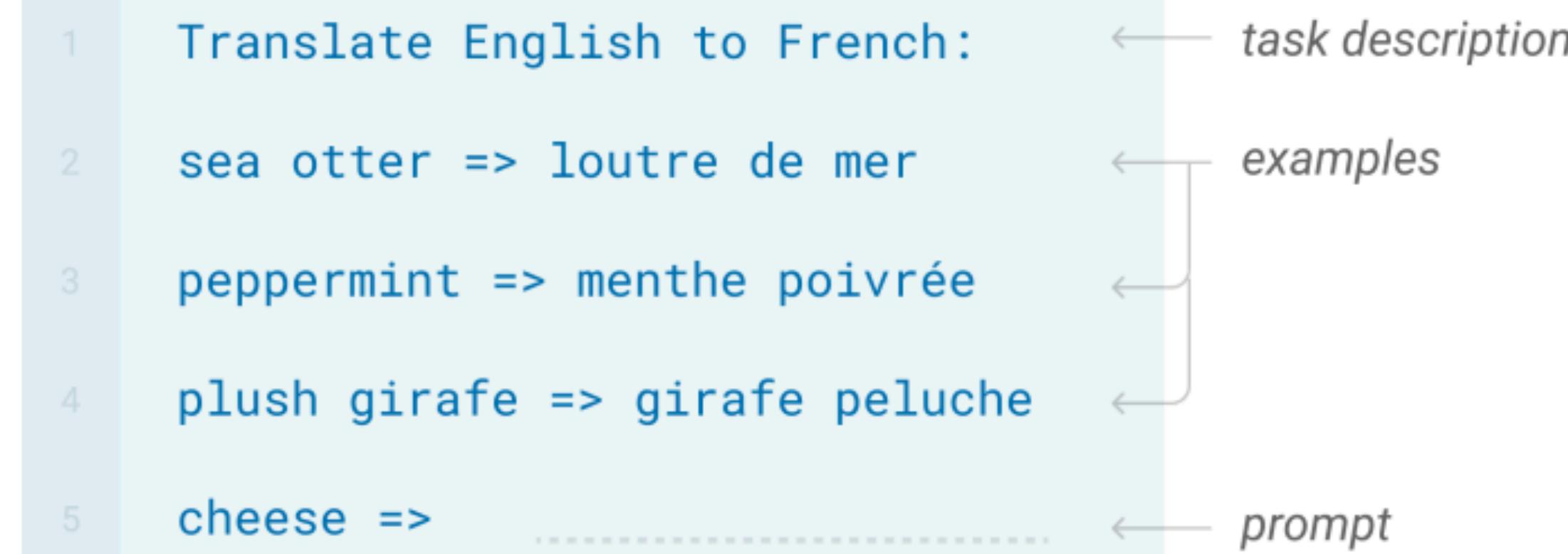
- Expect the model to give an answer given only a natural language text task description
- Zero-shot prompting assesses model's pretrained ability to solve any task

ONE-SHOT PROMPTING



- Help the model by giving an example.
- Model understands how to perform the task using just a single example

FEW-SHOT PROMPTING



The diagram illustrates the structure of a few-shot prompt. It consists of five numbered lines of text within a light blue rounded rectangle. To the right of the rectangle, three labels are positioned with arrows pointing to specific parts of the text: 'task description' points to line 1, 'examples' points to lines 2-4, and 'prompt' points to line 5.

1	Translate English to French:	← task description
2	sea otter => loutre de mer	← examples
3	peppermint => menthe poivrée	
4	plush girafe => girafe peluche	
5	cheese =>	← prompt

- More examples!

FEW-SHOT PROMPTING

- These prompting techniques are inference-time techniques
 - No gradient updates are made; model parameters are not updated
 - How many examples?
 - Depends on how many examples can be given in a small space

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. 

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. X

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.

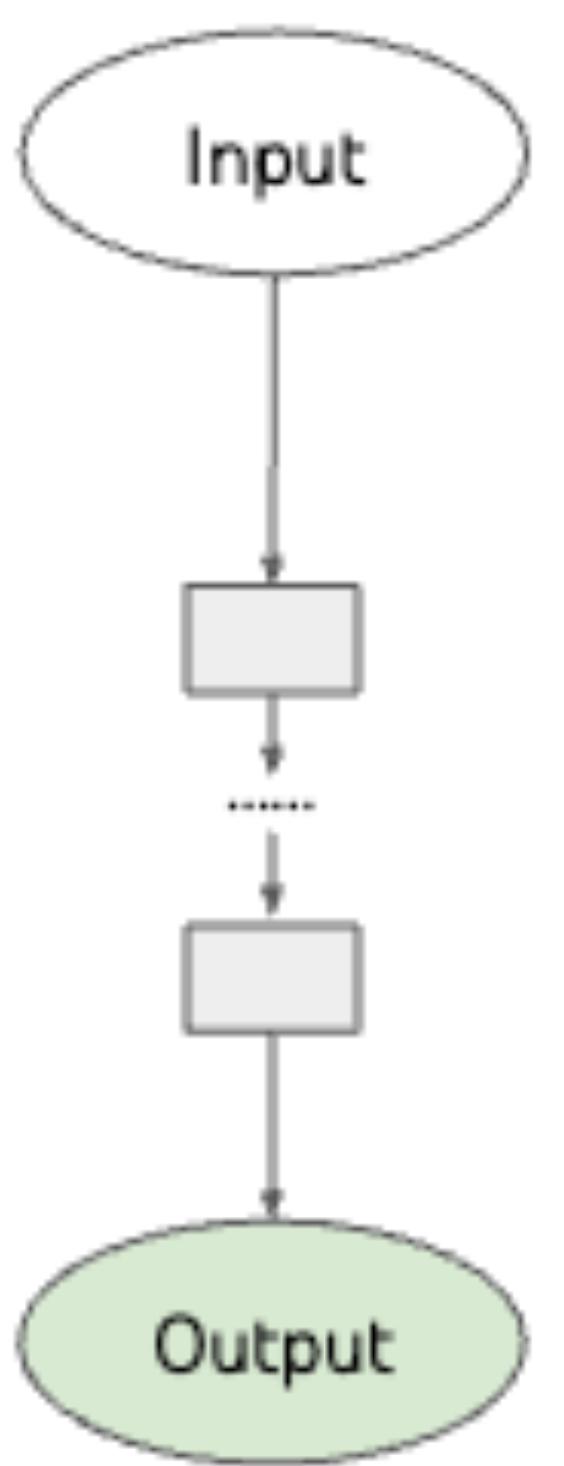
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

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. ✓



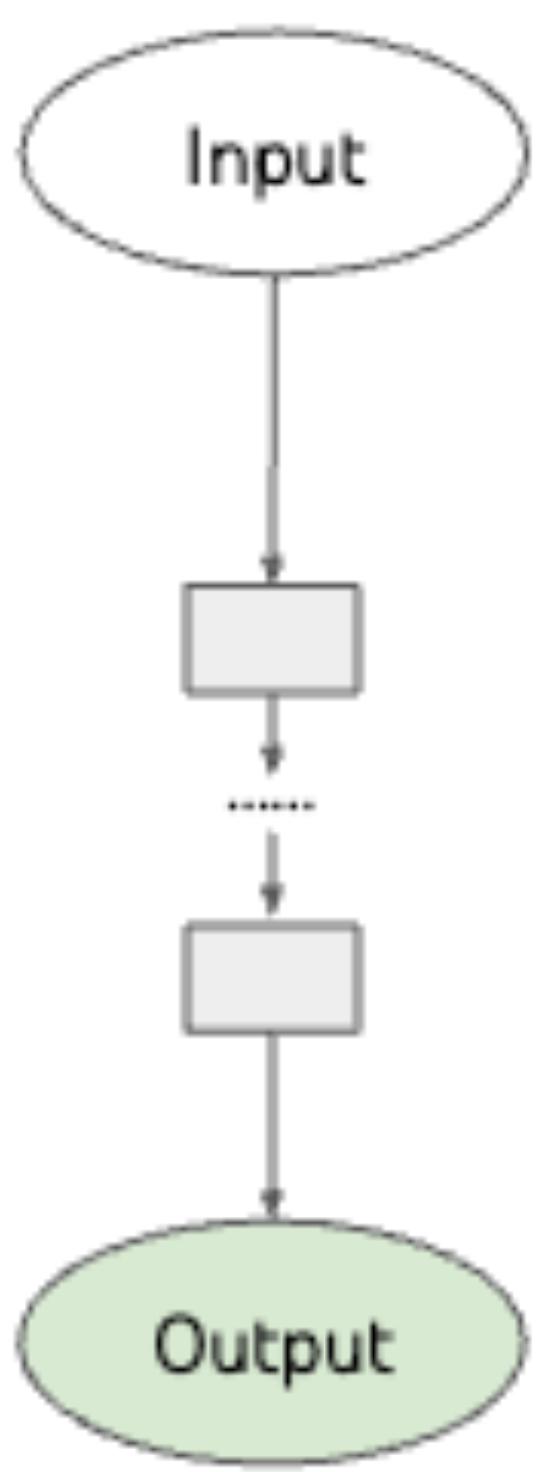
(a) Input-Output
Prompting (IO)



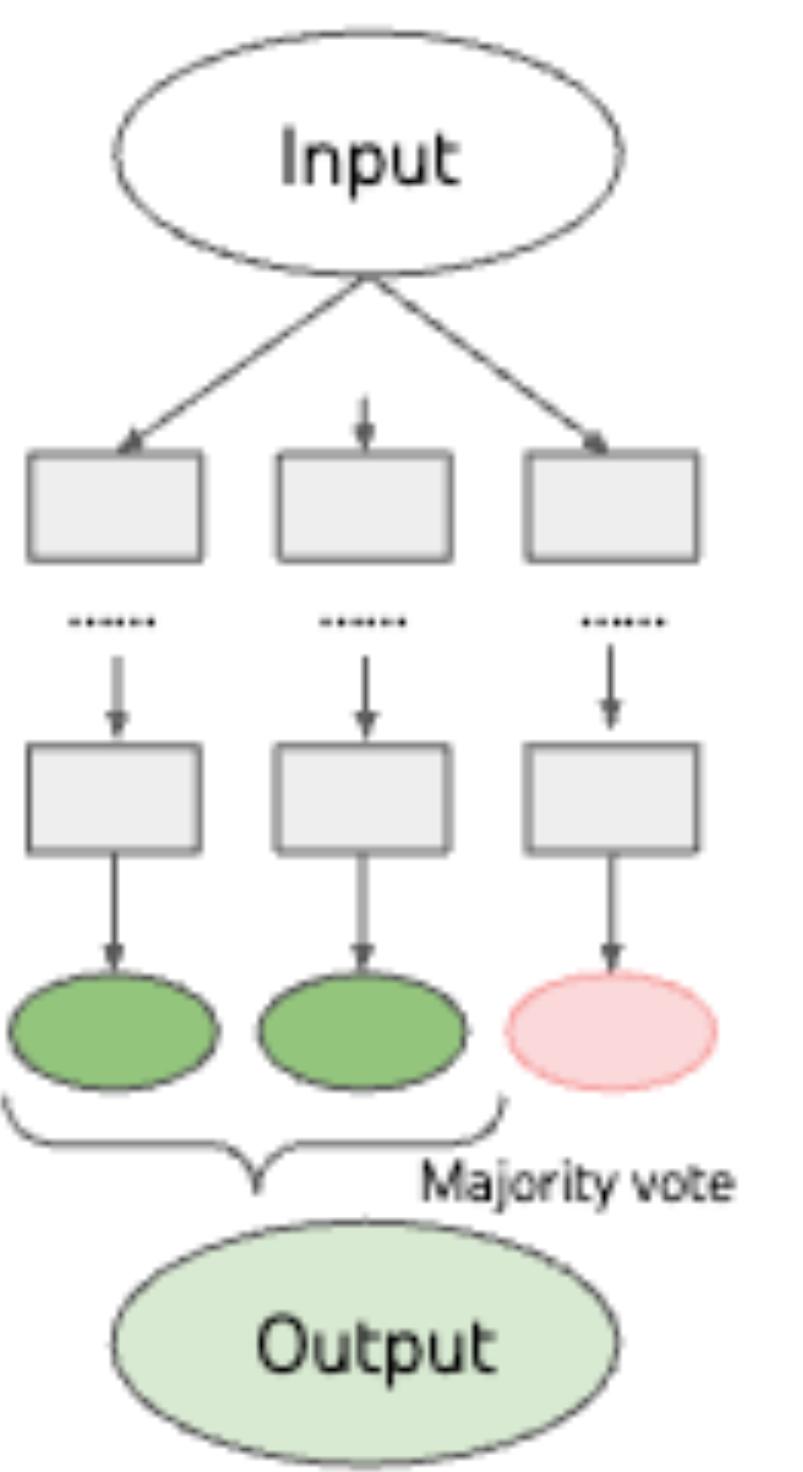
(c) Chain of Thought
Prompting (CoT)



(a) Input-Output
Prompting (IO)



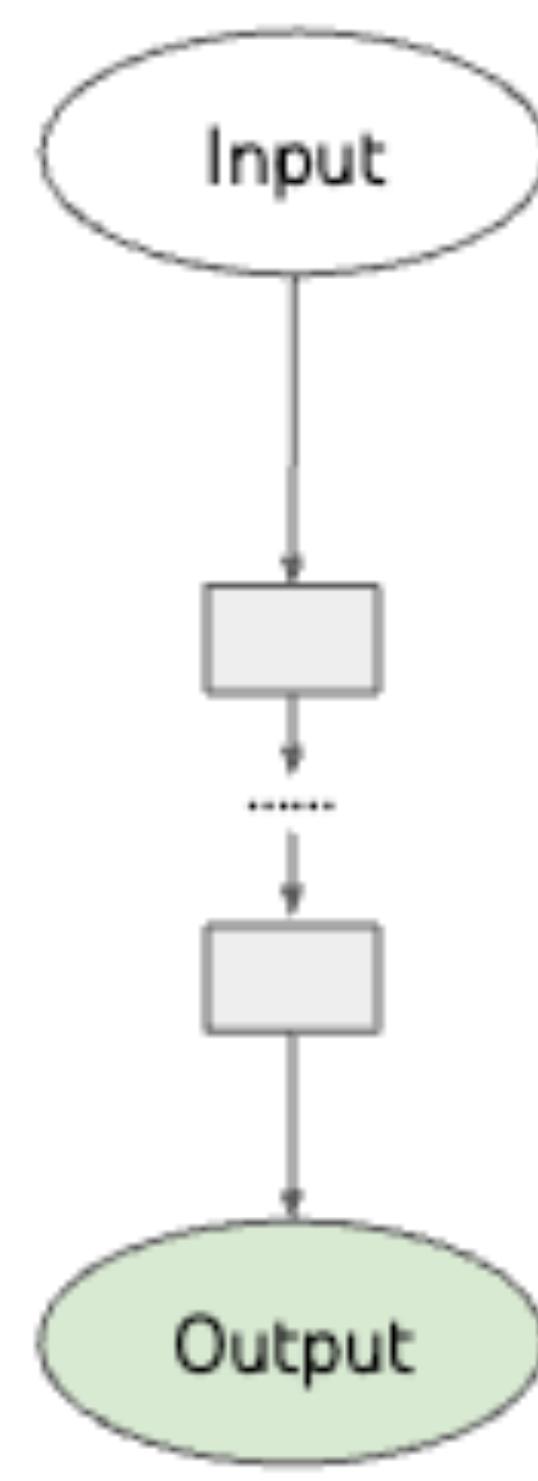
(c) Chain of Thought
Prompting (CoT)



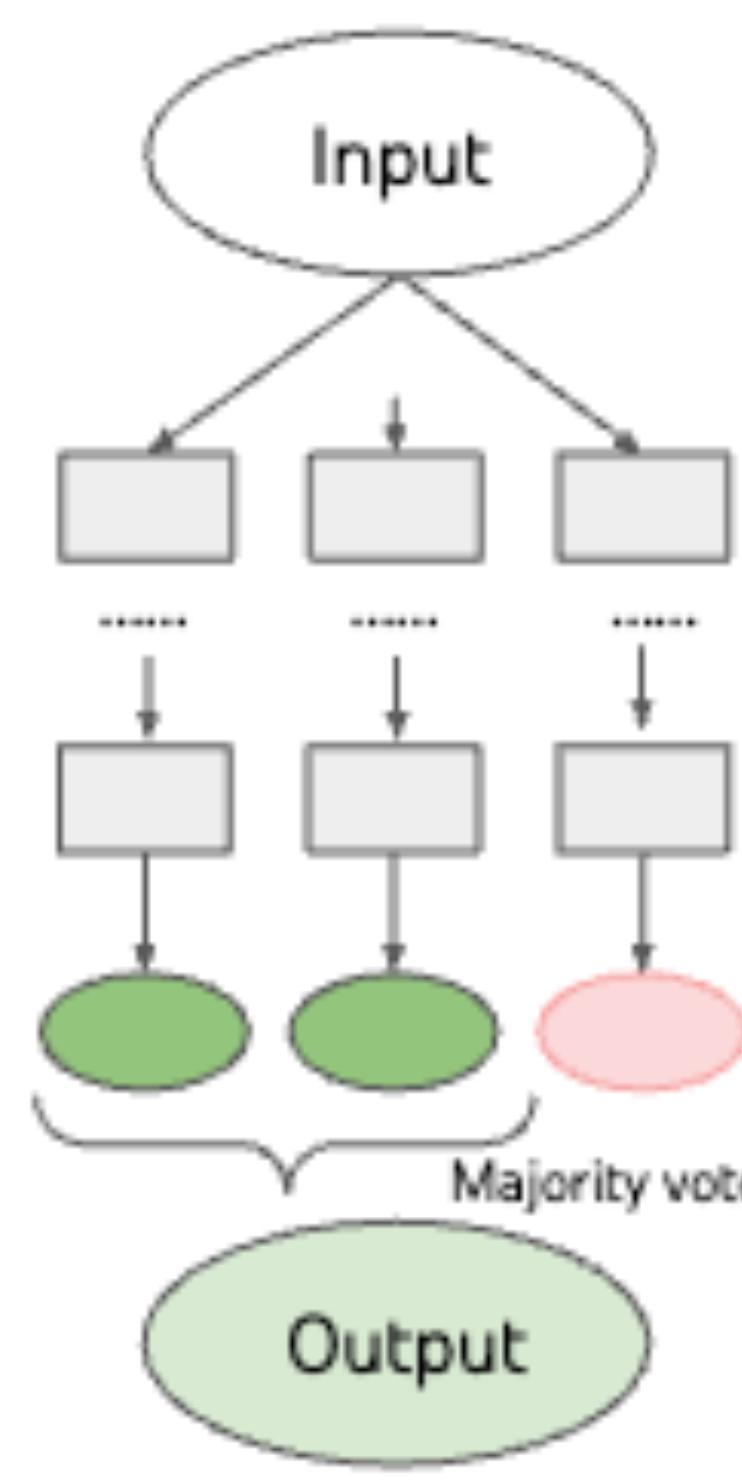
(c) Self Consistency
with CoT (CoT-SC)



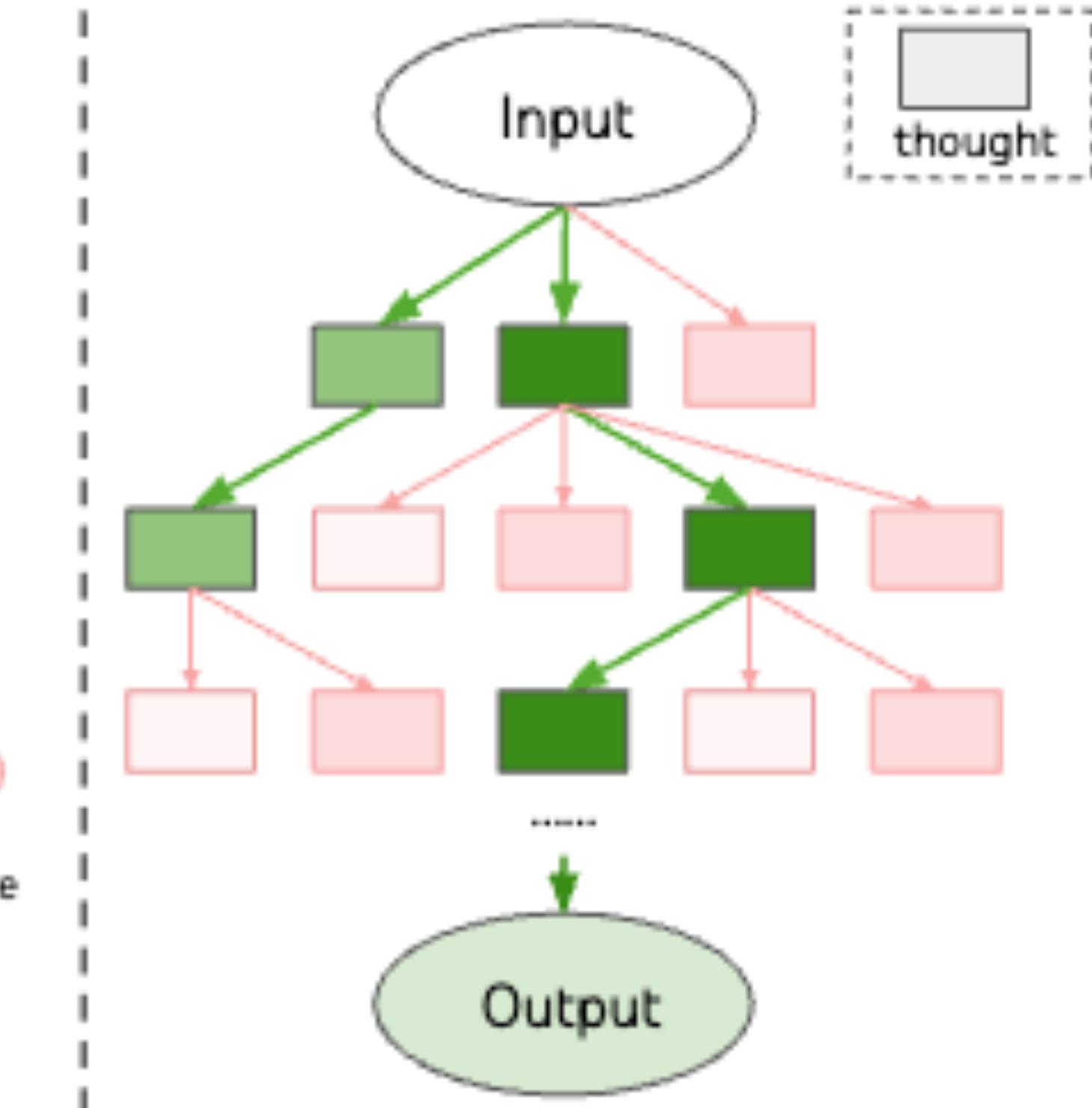
(a) Input-Output
Prompting (IO)



(c) Chain of Thought
Prompting (CoT)



(c) Self Consistency
with CoT (CoT-SC)



(d) Tree of Thoughts (ToT)

PROMPT ENGINEERING

- Manual prompting: Design prompt templates by hand
- Prompt mining: Find the best prompt from a choice of prompts that optimizes some criterion
- Prompt paraphrasing: Generate a concise prompt using LLM for a task
- Autoprompt: Learn the prompt and the output of a task by giving just a set of input/output pairs

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.

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

What is the amount of narrative time that has passed in a short excerpt of fictional text?

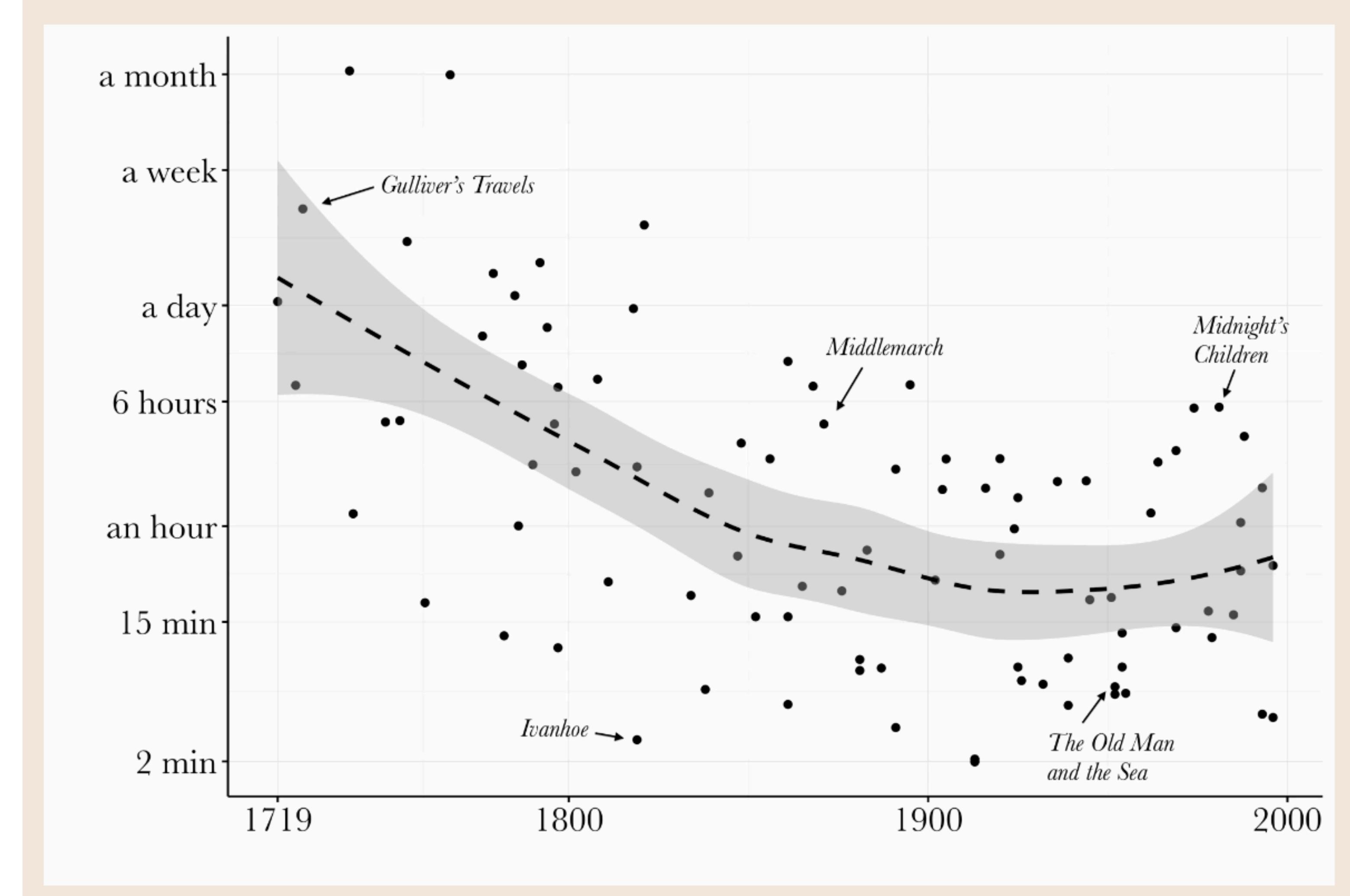


Image taken from Underwood's article "Why Literary Time is Measured in Minutes"

Read the following passage of fiction. Then do five things.

- 1: Briefly summarize the passage.
 - 2: Reason step by step to decide how much time is described in the passage. If the passage doesn't include any explicit reference to time, you can guess how much time the events described would have taken. Even description can imply the passage of time by describing the earlier history of people or buildings. But characters' references to the past or future in spoken dialogue should not count as time that passed in the scene. Report the time using units of years, weeks, days, hours, or minutes. Do not say zero or N/A.
 - 3: If you described a range of possible times in step 2 take the midpoint of the range. Then multiply to convert the units into minutes.
 - 4: Report only the number of minutes elapsed, which should match the number in step 3. Do not reply N/A.
 - 5: Given the amount of speculation required in step 2, describe your certainty about the estimate--either high, moderate, or low.
- The passage follows: <omitting this to save space>

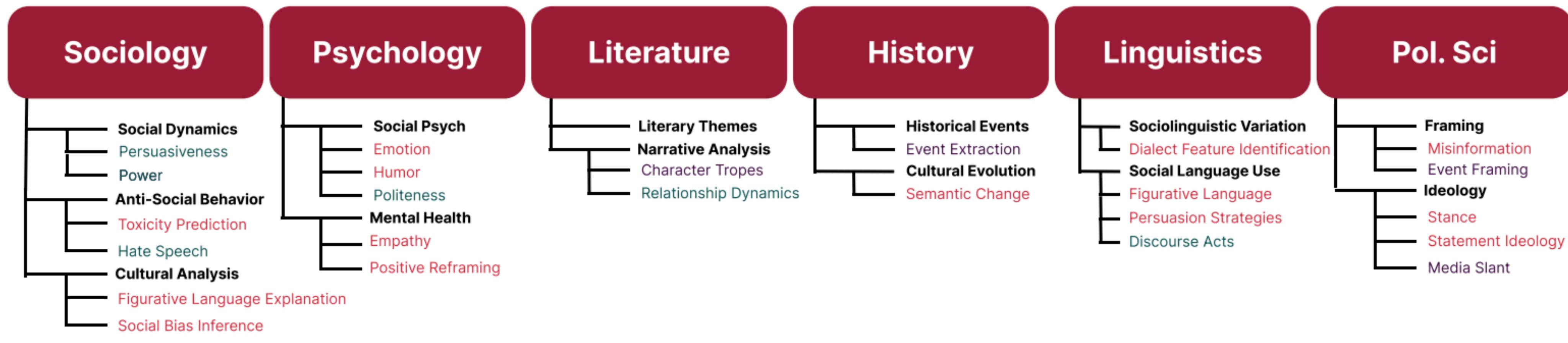
- 1: A detective wakes up 'the next morning,' looks out a window for an undefined time, drinks (and presumably needs to make) two cups of coffee, then shaves and showers and gets dressed before stepping out his front door and seeing a car.
- 2: Making coffee, showering, and getting dressed take at least an hour. There's some ambiguity about whether to count the implicit reference to yesterday (since this is 'the next morning') as time elapsed in the passage, but let's say no, since yesterday is not actually described. So, an hour to 90 minutes.
- 3: 1.25 hours have elapsed. Multiplying by 60 minutes an hour that's 75 minutes.
- 4: 75 minutes.
- 5: Low confidence, because of ambiguity about a reference to the previous day.

Response from GPT-4

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

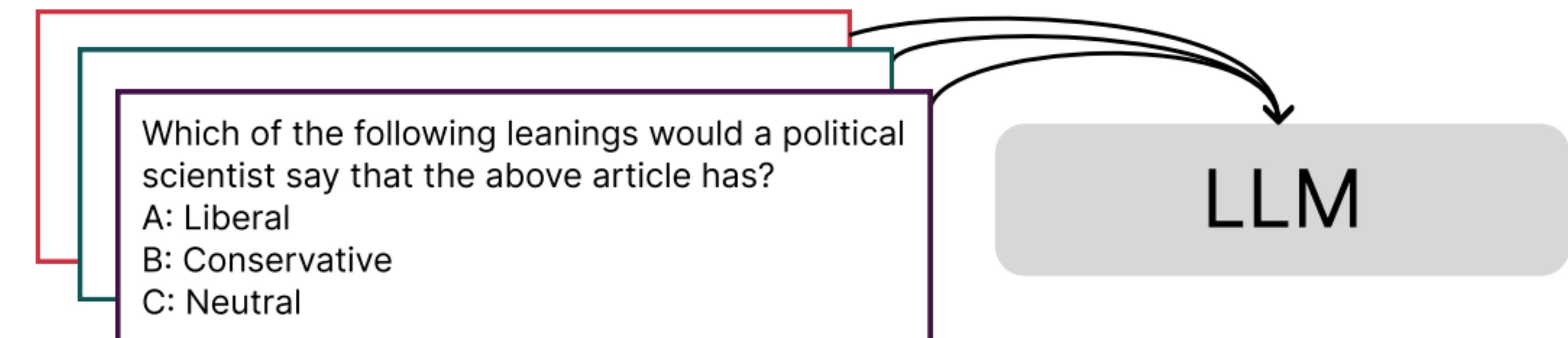
LLMs are promising “research assistants” on measurement tasks



Discourse Types

Utterances
Conversations
Documents

Zero Shot Prompt Formatting



Ziems, Caleb, et al. "Can Large Language Models Transform Computational Social Science?."

Large language models can be effective partners to humans on multiple tasks

Wow. I sit down, fish the questions from my backpack, and go through them, inwardly cursing [MASK] for not providing me with a brief biography. I know nothing about this man I'm about to interview. He could be ninety or he could be thirty. → **Kate** (James, *Fifty Shades of Grey*).

Some days later, when the land had been moistened by two or three heavy rains, [MASK] and his family went to the farm with baskets of seed-yams, their hoes and machetes, and the planting began. → **Okonkwo** (Achebe, *Things Fall Apart*).

Figure 1: Name cloze examples. GPT-4 answers both of these correctly.

Wow. I sit down, fish the questions from my backpack, and go through them, inwardly cursing [MASK] for not providing me with a brief biography. I know nothing about this man I'm about to interview. He could be ninety or he could be thirty. → **Kate** (James, *Fifty Shades of Grey*).

Some days later, when the land had been moistened by two or three heavy rains, [MASK] and his family went to the farm with baskets of seed-yams, their hoes and machetes, and the planting began. → **Okonkwo** (Achebe, *Things Fall Apart*).

You have seen the following passage in your training data. What is the proper name that fills in the [MASK] token in it? This name is exactly one word long, and is a proper name (not a pronoun or any other word). You must make a guess, even if you are uncertain.

Example:

Input: Stay gold, [MASK], stay gold.
Output: <name>Ponyboy</name>

Input: The door opened, and [MASK], dressed and hatted, entered with a cup of tea.
Output: <name>Gerty</name>

Input: My back's to the window. I expect a stranger, but it's [MASK] who pushes open the door, flicks on the light. I can't place that, unless he's one of them. There was always that possibility.
Output:

Figure 1: Name cloze examples. GPT-4 answers both of these correctly.

GPT-4	ChatGPT	BERT	Date	Author	Title
0.98	0.82	0.00	1865	Lewis Carroll	<i>Alice's Adventures in Wonderland</i>
0.76	0.43	0.00	1997	J.K. Rowling	<i>Harry Potter and the Sorcerer's Stone</i>
0.74	0.29	0.00	1850	Nathaniel Hawthorne	<i>The Scarlet Letter</i>
0.72	0.11	0.00	1892	Arthur Conan Doyle	<i>The Adventures of Sherlock Holmes</i>
0.70	0.10	0.00	1815	Jane Austen	<i>Emma</i>
0.65	0.19	0.00	1823	Mary W. Shelley	<i>Frankenstein</i>
0.62	0.13	0.00	1813	Jane Austen	<i>Pride and Prejudice</i>
0.61	0.35	0.00	1884	Mark Twain	<i>Adventures of Huckleberry Finn</i>
0.61	0.30	0.00	1853	Herman Melville	<i>Bartleby, the Scrivener</i>
0.61	0.08	0.00	1897	Bram Stoker	<i>Dracula</i>
0.61	0.18	0.00	1838	Charles Dickens	<i>Oliver Twist</i>
0.59	0.13	0.00	1902	Arthur Conan Doyle	<i>The Hound of the Baskervilles</i>
0.59	0.22	0.00	1851	Herman Melville	<i>Moby Dick; Or, The Whale</i>
0.58	0.35	0.00	1876	Mark Twain	<i>The Adventures of Tom Sawyer</i>
0.57	0.30	0.00	1949	George Orwell	<i>1984</i>
0.54	0.10	0.00	1908	L. M. Montgomery	<i>Anne of Green Gables</i>
0.51	0.20	0.01	1954	J.R.R. Tolkien	<i>The Fellowship of the Ring</i>
0.49	0.16	0.13	2012	E.L. James	<i>Fifty Shades of Grey</i>
0.49	0.24	0.01	1911	Frances H. Burnett	<i>The Secret Garden</i>
0.49	0.12	0.00	1883	Robert L. Stevenson	<i>Treasure Island</i>
0.49	0.16	0.00	1847	Charlotte Brontë	<i>Jane Eyre: An Autobiography</i>
0.49	0.22	0.00	1903	Jack London	<i>The Call of the Wild</i>

IN CLASS EXERCISE

- LLMs prompting
- Try yourself
 - ChatGPT: <https://chat.openai.com/>
 - Gemini: <https://gemini.google.com/app>
 - Claude: <https://claude.ai/chats>
 - Llama2: <https://www.llama2.ai/>