



# LARGE LANGUAGE MODELS I

Sandeep Soni

---

03/21/2024

# STORY SO FAR

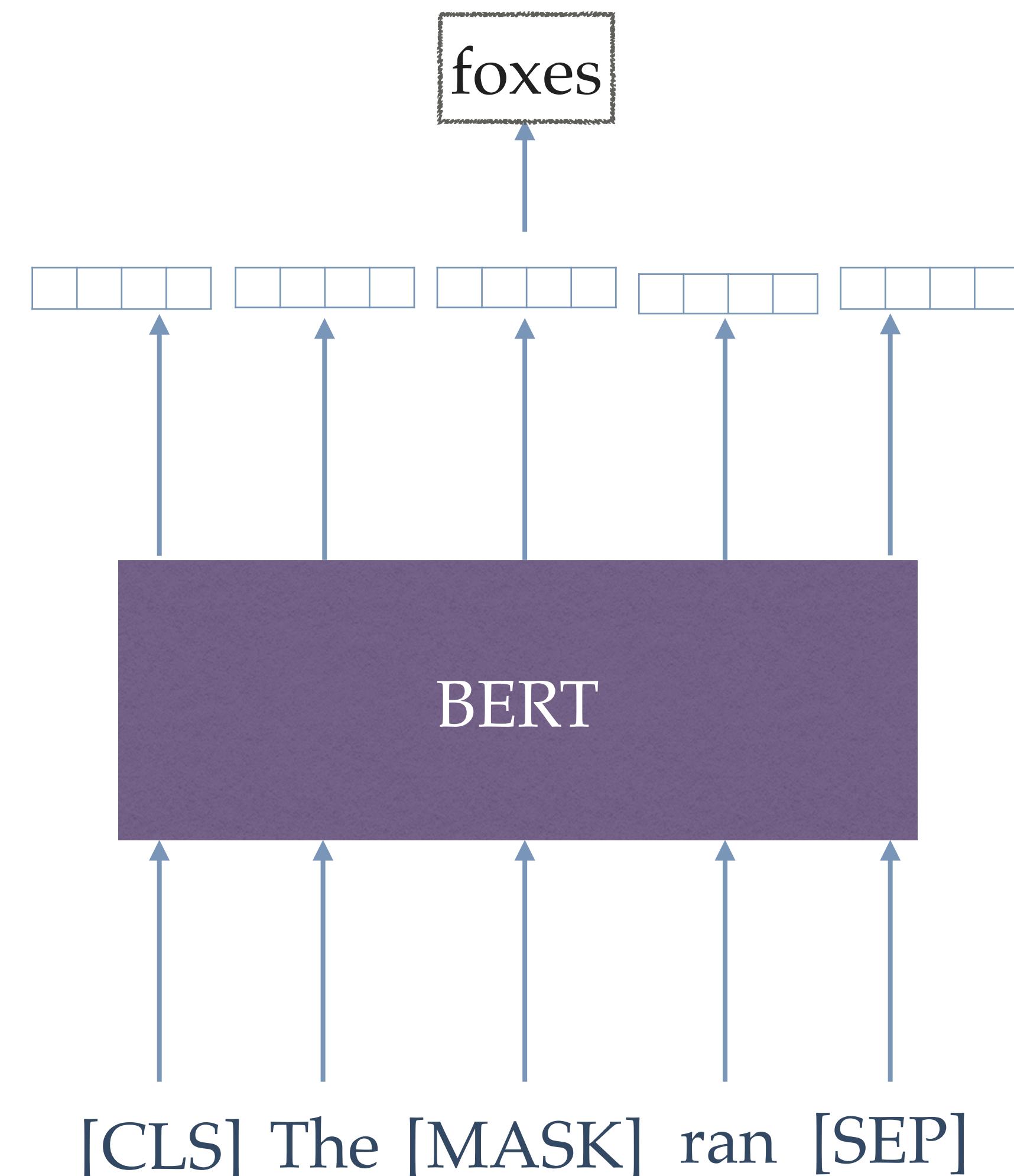
- **Language modeling:** Probabilistic model over natural language.
- **Objective:** Predict word based on some context i.e  $P(w|C)$
- **Variations:** Autoregressive LMs, masked LMs, etc
- **Techniques:** Count and normalize, skip-gram, RNN, Transformers

# MASKED LANGUAGE MODELING

The \_\_\_\_\_ ran

Fill in the blank by  
using the surrounding  
context

$$P(w_t | \neg w_t)$$



BERT is a  
language  
model  
trained to  
predict the  
missing  
word

# 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

- Transformer based model

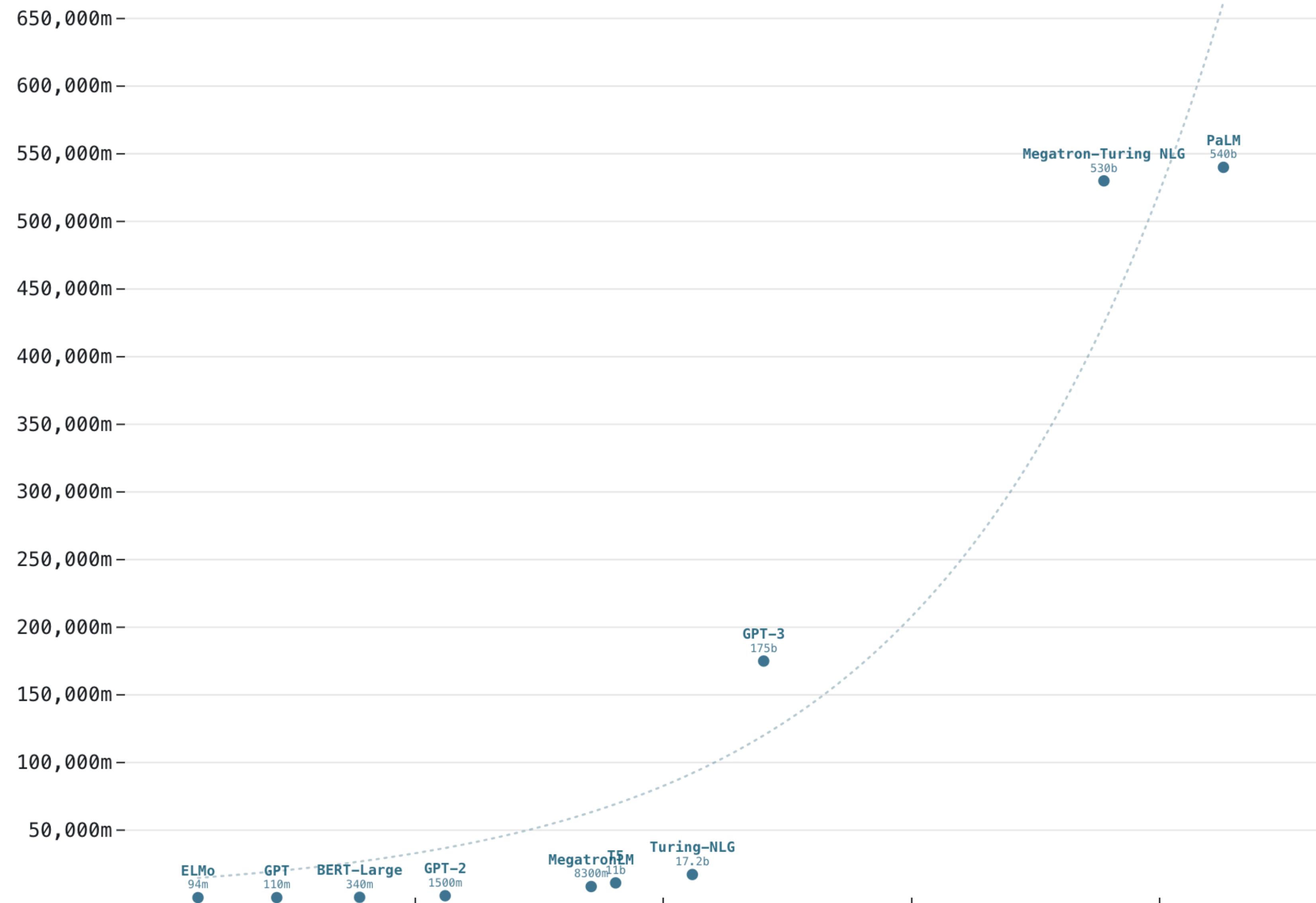
- trained on 512 sized contexts
- 12 or 24 transformer blocks
- 768 or 1024 sized representations
- pretrained on billions of tokens from wikipedia and books

“What distinguishes large language models?”

# SCALE

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

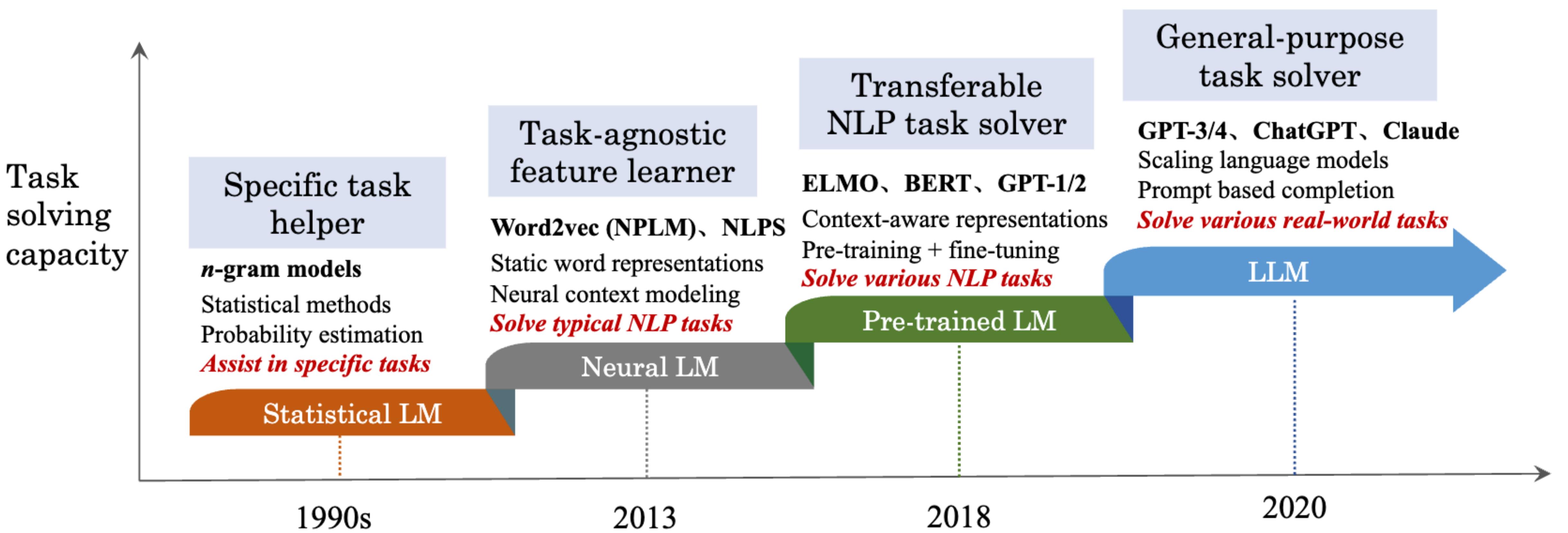
### Number of Parameters (in Millions)



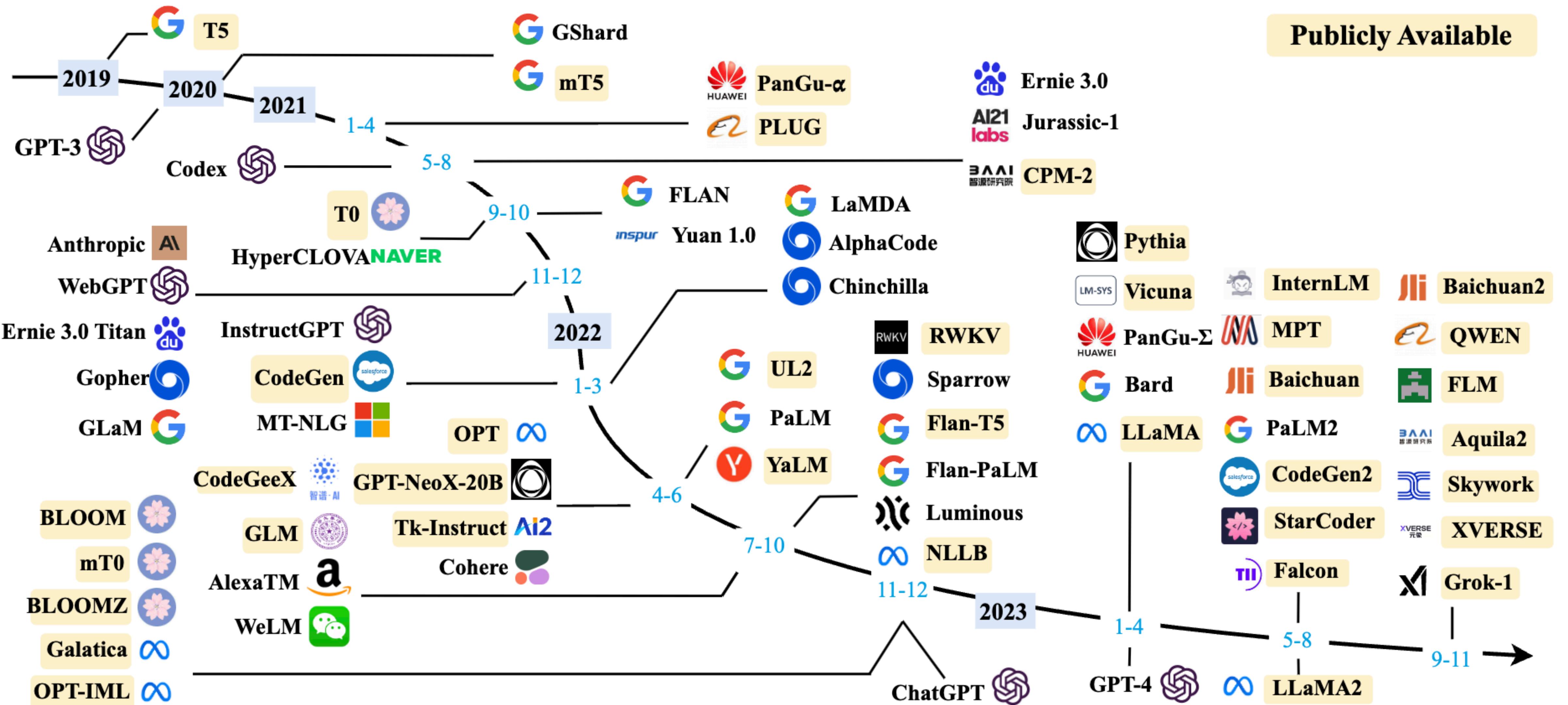
Large Language Models are models with millions or billions of parameters

# 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).



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)

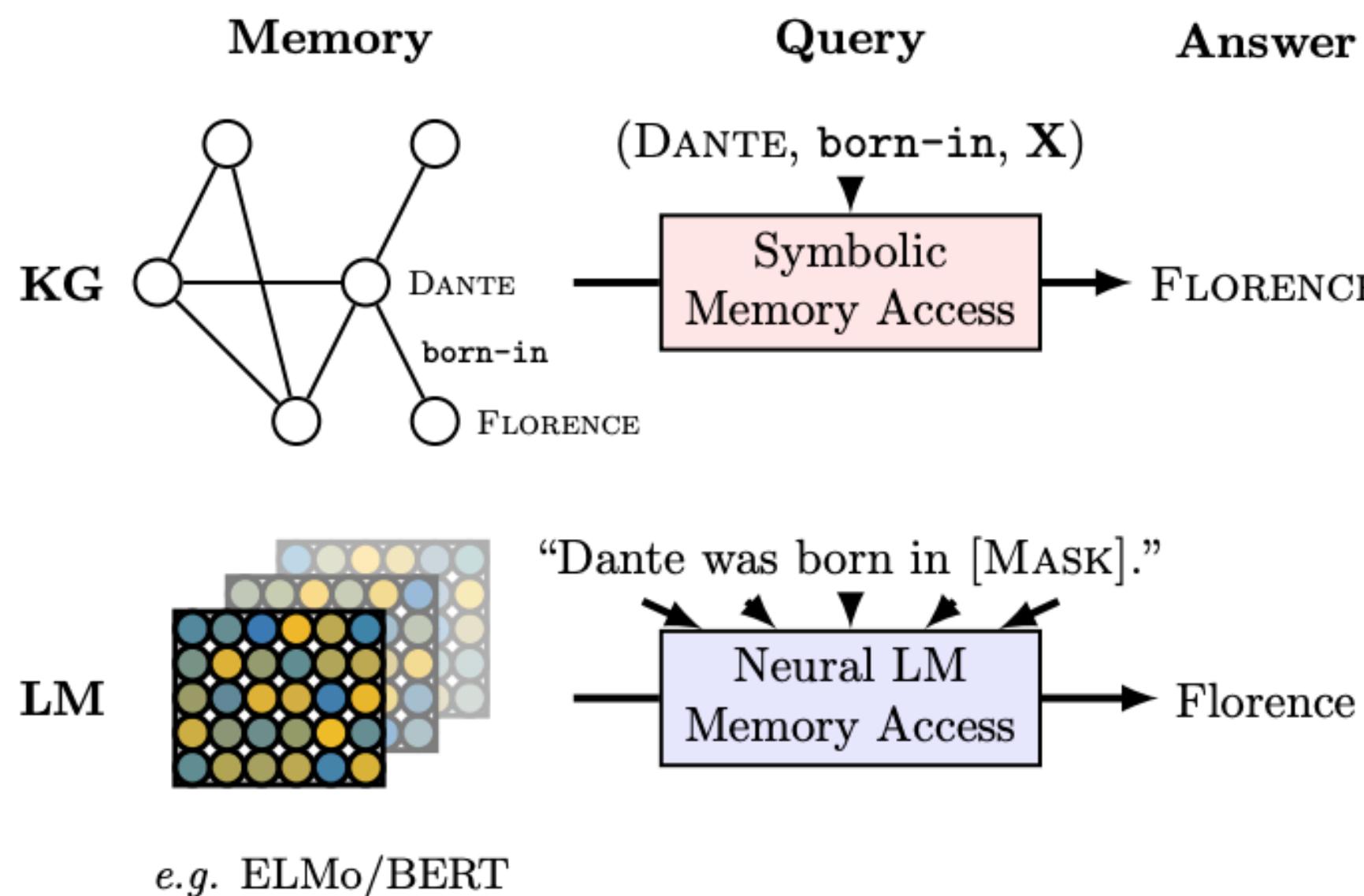
Instructions:

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

# LANGUAGE MODELS ARE KNOWLEDGE BASES



Query	Answer	Generation
Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], <b>Florence</b> [-1.8], Naples [-1.9], Milan [-2.4], Bologna [-2.5]
Adolphe Adam died in ____.	Paris	<b>Paris</b> [-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], <b>dog</b> [-2.4], cattle [-4.3], sheep [-4.5]
The official language of Mauritius is ____.	English	<b>English</b> [-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], <b>midfielder</b> [-2.4], forward [-2.4], midfield [-2.7]
Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], <b>Hamburg</b> [-7.5], Ludwig [-7.5]
The original language of Mon Oncle Benjamin is ____.	French	<b>French</b> [-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], <b>sodium</b> [-2.5], zinc [-2.8], salt [-2.9]
Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], <b>Labor</b> [-2.9]
Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], <b>Uganda</b> [-3.2], Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
iPod Touch is produced by ____.	Apple	<b>Apple</b> [-1.6], Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
Bailey Peninsula is located in ____.	Antarctica	<b>Antarctica</b> [-1.4], Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
Sunshine Coast, British Columbia is located in ____.	Canada	<b>Canada</b> [-1.2], Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
Pope Clement VII has the position of ____.	pope	cardinal [-2.4], Pope [-2.5], <b>pope</b> [-2.6], President [-3.1], Chancellor [-3.2]
Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], <b>Capitol</b> [-3.2], Columbia [-3.3]
London Jazz Festival is located in ____.	London	<b>London</b> [-0.3], Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]
Border TV is owned by ____.	ITV	Sky [-3.1], <b>ITV</b> [-3.3], Global [-3.4], Frontier [-4.1], Disney [-4.3]
The native language of Mammootty is ____.	Malayalam	<b>Malayalam</b> [-0.2], Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]
The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], <b>Philippines</b> [-3.6], February [-3.7], December [-3.8], Argentina [-4.0]
You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], <b>drain</b> [-3.6]
Ravens can ____.	fly	<b>fly</b> [-1.5], fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], <b>laugh</b> [-2.0], vomit [-2.6], scream [-2.6]
Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], <b>infection</b> [-2.6], plague [-3.3], fever [-3.4]
Birds have ____.	feathers	wings [-1.8], nests [-3.1], <b>feathers</b> [-3.2], died [-3.7], eggs [-3.9]
Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], <b>speed</b> [-4.1]
Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], <b>alive</b> [-3.3], young [-3.6], free [-3.9]
Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], <b>fish</b> [-2.8], recreation [-3.1]

# EVERYTHING IS LANGUAGE MODELING

- If you can construct a suitable prompt, many tasks can be framed as language modeling
  - The director of Avatar is \_\_\_\_\_
  - The definition of “chromatid” is \_\_\_\_\_
  - Is “sit down” rude than “please, sit down”? \_\_\_\_\_

# PRETRAINED AND FINETUNED

- Though a language model is trained on predicting words in sequences, a pretrained language model can be adapted for a specific task – a process called finetuning
- Because of the size of some LMs, finetuning can be costly

# PROMPTING

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

## Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

- 
- 1 Translate English to French: ← task description  
2 cheese => ← prompt
- This diagram shows a sequence of two items. Item 1 is "Translate English to French:" followed by a colon, which is labeled "task description". Item 2 is "cheese =>" followed by a dashed line, which is labeled "prompt". Arrows point from the labels to their respective parts.

## One-shot

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

- 
- 1 Translate English to French: ← task description  
2 sea otter => loutre de mer ← example  
3 cheese => ← prompt
- This diagram shows a sequence of three items. Item 1 is "Translate English to French:" followed by a colon, which is labeled "task description". Item 2 is "sea otter => loutre de mer" followed by a dashed line, which is labeled "example". Item 3 is "cheese =>" followed by a dashed line, which is labeled "prompt". Arrows point from the labels to their respective parts.

## Few-shot

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

- 
- 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
- This diagram shows a sequence of five items. Item 1 is "Translate English to French:" followed by a colon, which is labeled "task description". Item 2 is "sea otter => loutre de mer" followed by a dashed line, which is labeled "examples". Item 3 is "peppermint => menthe poivrée". Item 4 is "plush girafe => girafe peluche". Item 5 is "cheese =>" followed by a dashed line, which is labeled "prompt". Arrows point from the labels to their respective parts.

Language Models are Few-Shot Learners:  
<https://arxiv.org/pdf/2005.14165.pdf>

## 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. 

# 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?

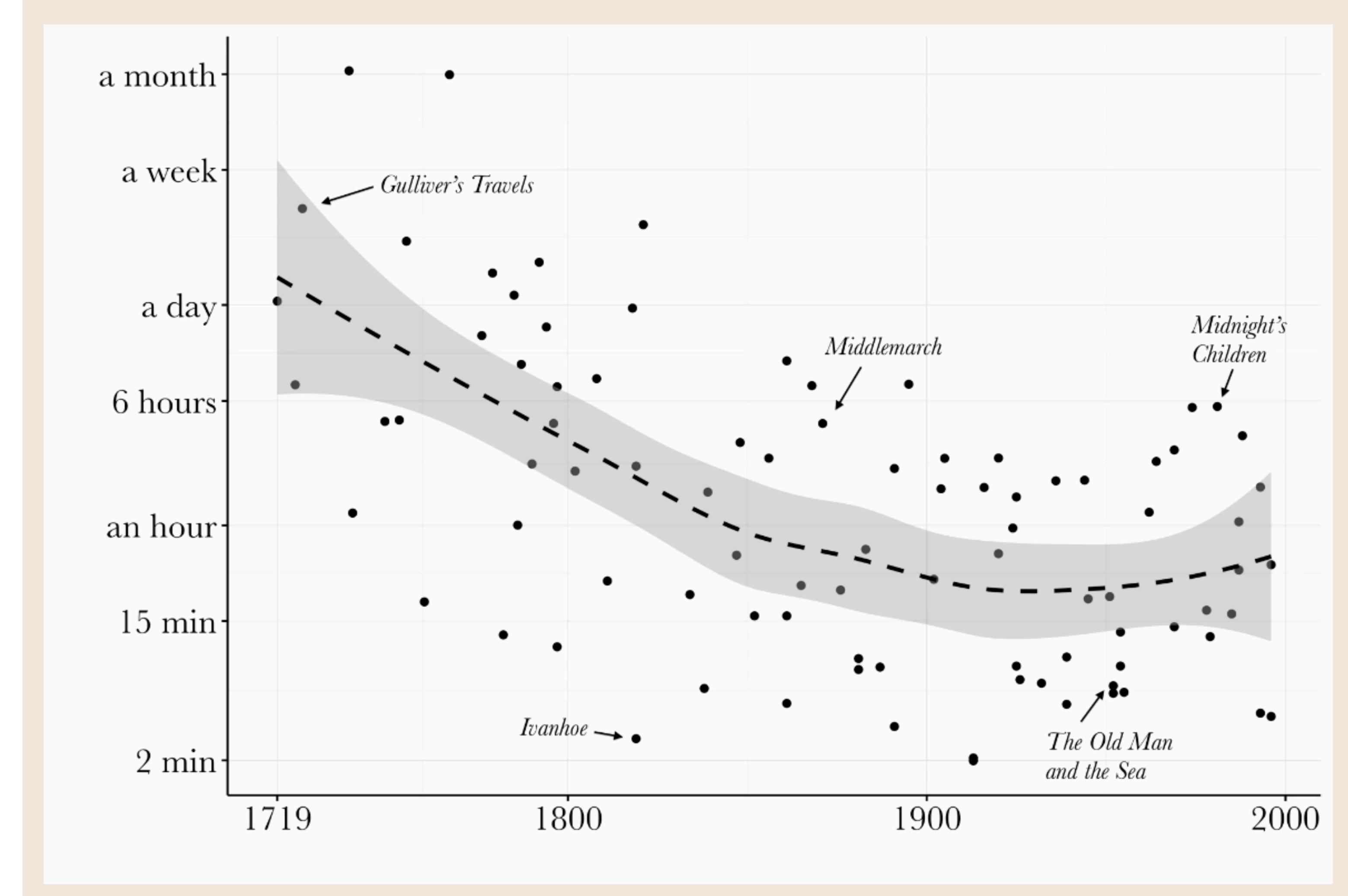


Image taken from Underwood's article "Why Literary Time is Measured in Minutes"

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>

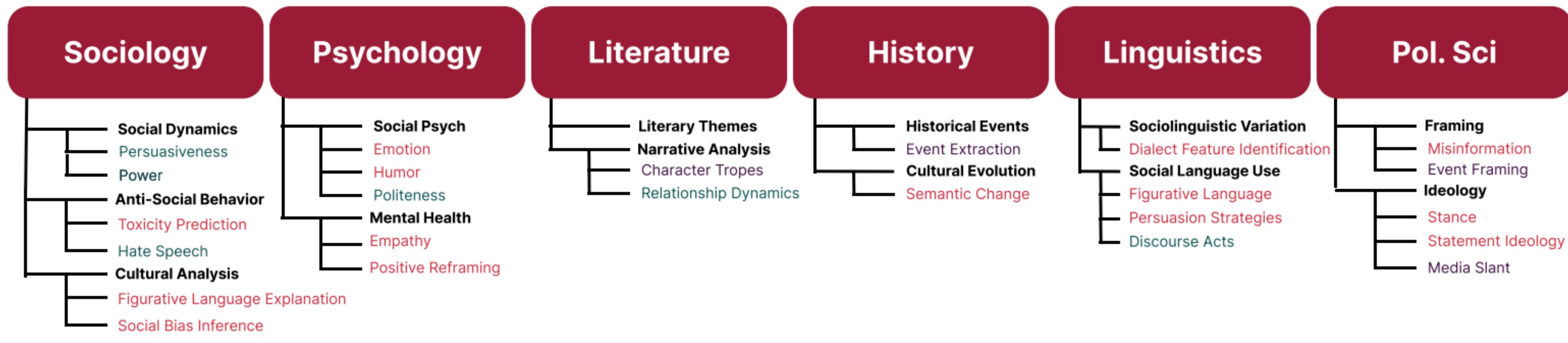
- 1: A detective wakes up 'the next morning,' looks out a window for an undefined time, drinks (and presumably needs to make) two cups of coffee, then shaves and showers and gets dressed before stepping out his front door and seeing a car.
- 2: Making coffee, showering, and getting dressed take at least 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

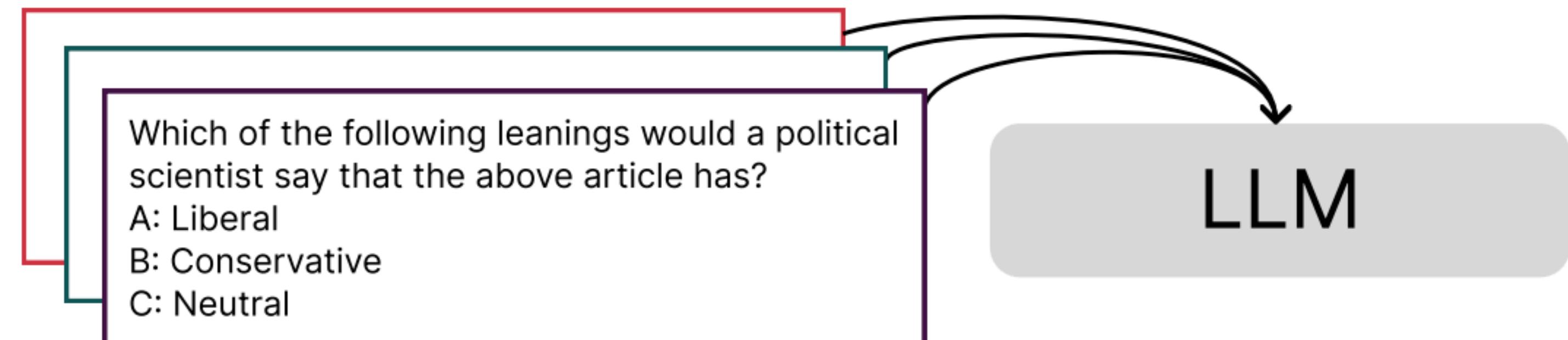


### Discourse Types

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

# MULTIMODALITY

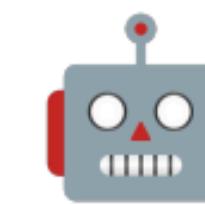
c)



Who are the people on the right, on the left, and in the middle?



Language models can be integrated with other modalities (e.g., vision)



The people in the image are:

- \* Barack Obama, the former President of the United States, is on the left side of the image.
- \* Joe Biden, the current President of the United States, is in the middle of the image.
- \* Donald Trump, the former President of the United States, is on the right side of the image.

## 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/>