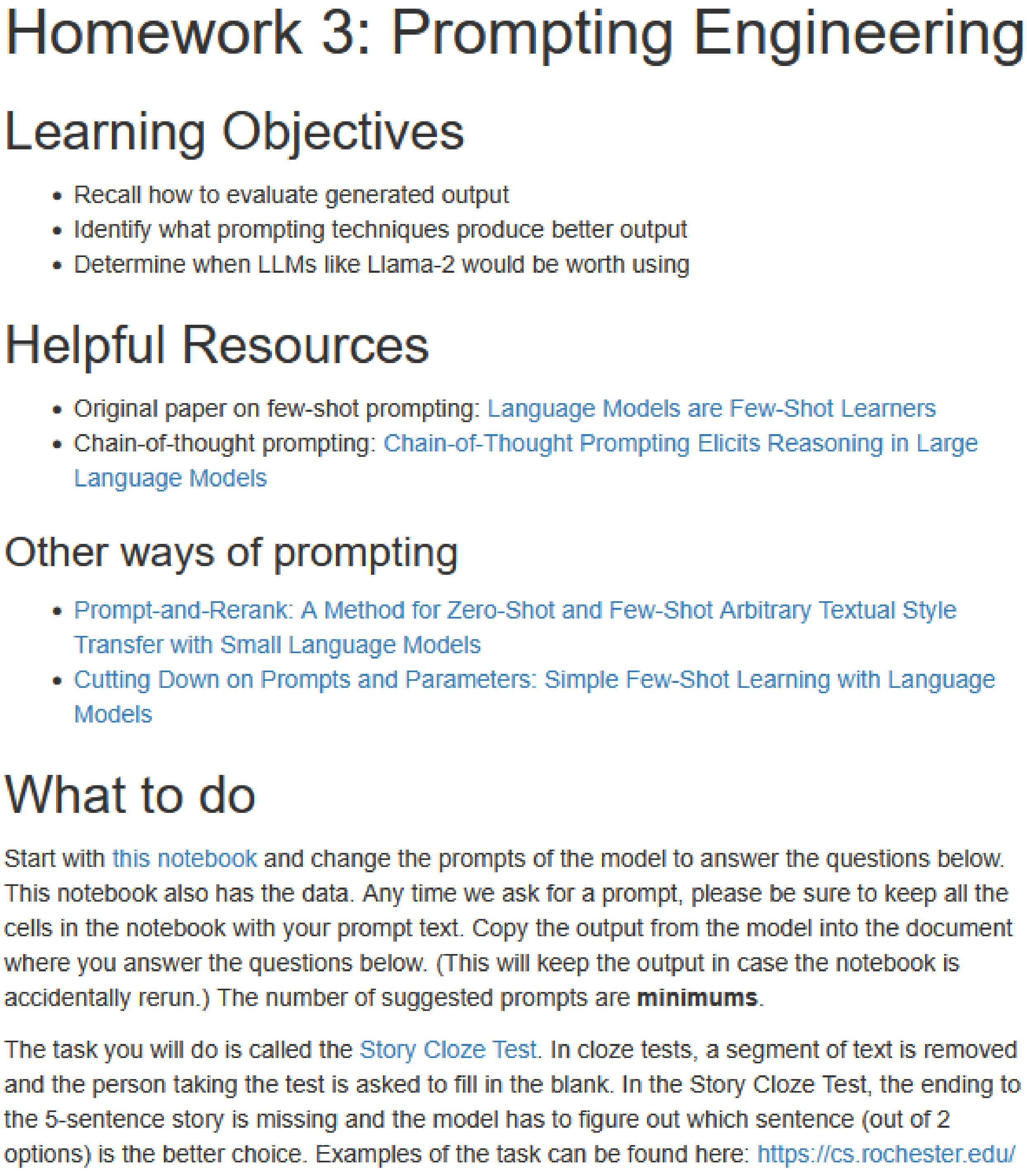
Pretrained Models and Prompting

CMSC 473/673 - NATURAL LANGUAGE PROCESSING

HW 3



Homework 3: Prompting Engineering

Learning Objectives

- Recall how to evaluate generated output
- Identify what prompting techniques produce better output
- Determine when LLMs like Llama-2 would be worth using

Helpful Resources

- Original paper on few-shot prompting: Language Models are Few-Shot Learners
- Chain-of-thought prompting: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Other ways of prompting

- Prompt-and-Rerank: A Method for Zero-Shot and Few-Shot Arbitrary Textual Style Transfer with Small Language Models
- Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models

What to do

Start with this notebook and change the prompts of the model to answer the questions below. This notebook also has the data. Any time we ask for a prompt, please be sure to keep all the cells in the notebook with your prompt text. Copy the output from the model into the document where you answer the questions below. (This will keep the output in case the notebook is accidentally rerun.) The number of suggested prompts are **minimums**.

The task you will do is called the Story Cloze Test. In cloze tests, a segment of text is removed and the person taking the test is asked to fill in the blank. In the Story Cloze Test, the ending to the 5-sentence story is missing and the model has to figure out which sentence (out of 2 options) is the better choice. Examples of the task can be found here: https://cs.rochester.edu/

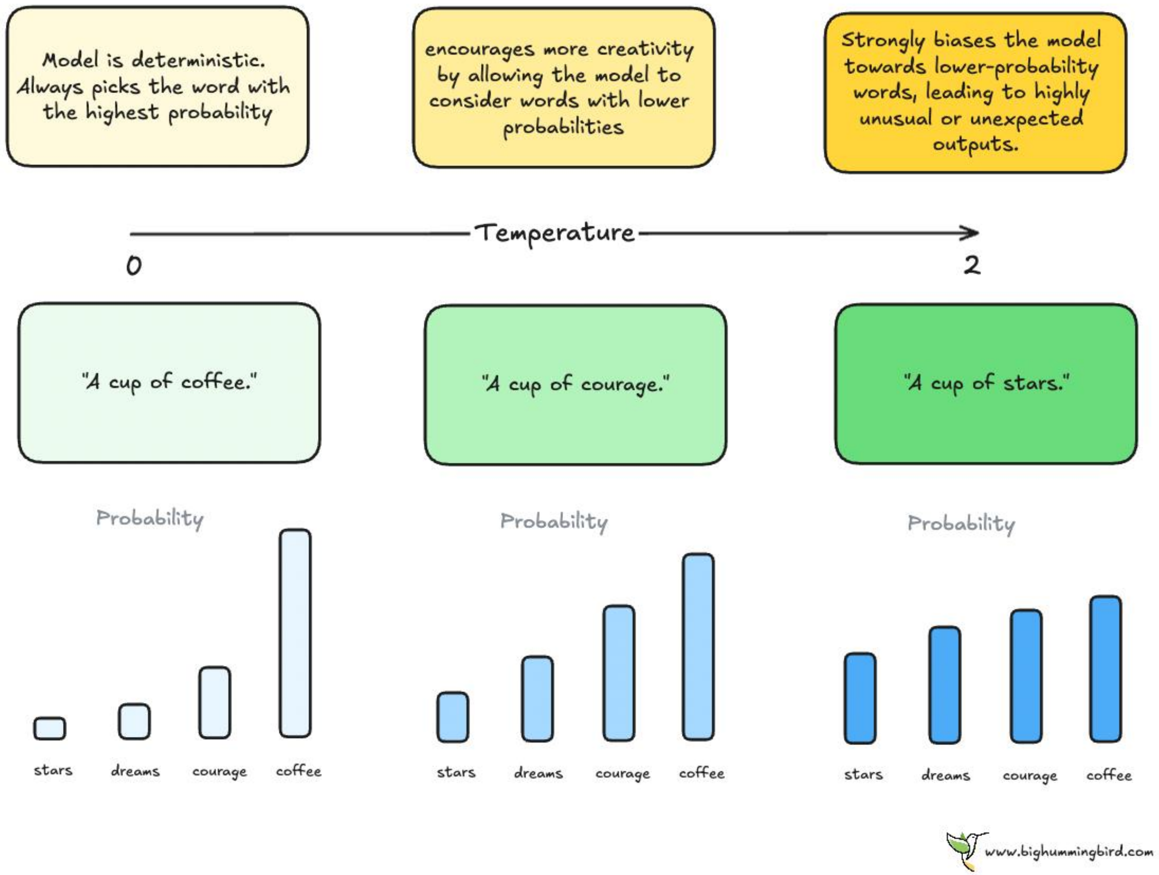
Learning Objectives

Recognize useful encoder-only, encoder-decoder, and decoder-only models

Distinguish between few-shot and zero-shot prompting

Try common prompting techniques like chain-of-thought

Review: “Temperature”



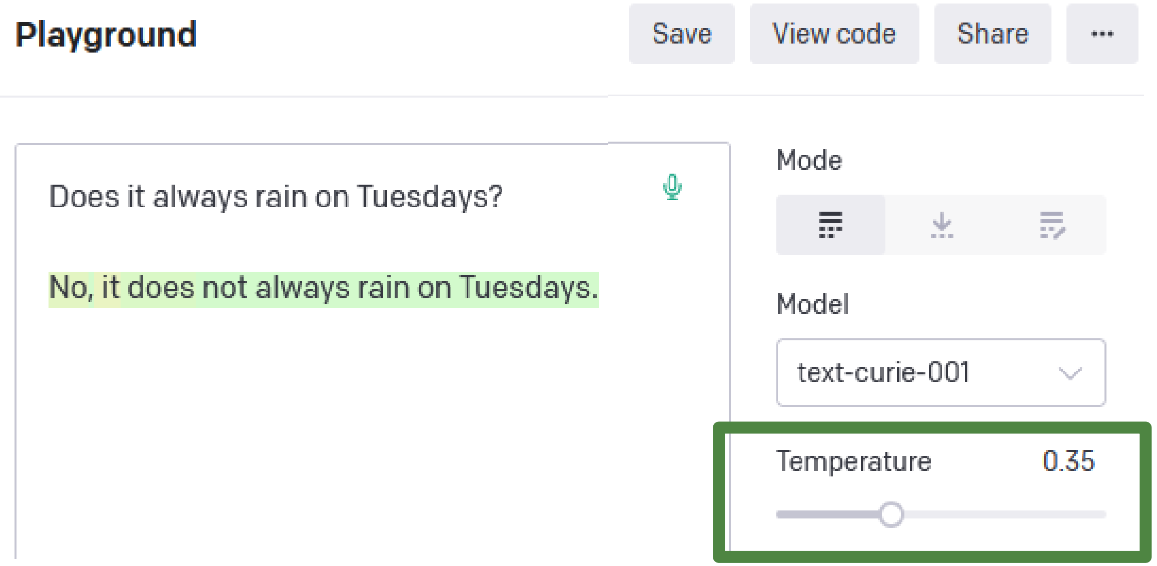
Model is deterministic. Always picks the word with the highest probability

encourages more creativity by allowing the model to consider words with lower probabilities

Strongly biases the model towards lower-probability words, leading to highly unusual or unexpected outputs.

https://www.bighummingbird.com/blogs/llm-hyperparameter

Temperature in Action

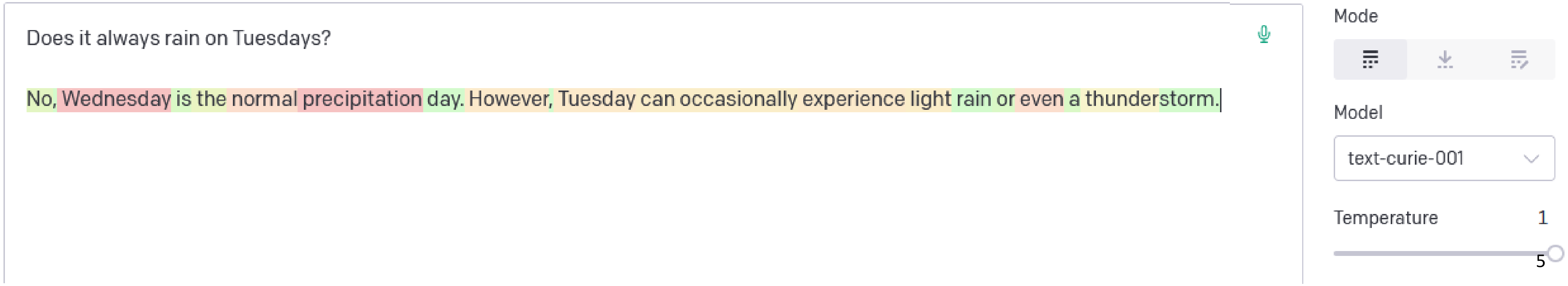


Does it always rain on Tuesdays?

No, it does not always rain on Tuesdays.

Model: text-curie-001

Temperature 0.35



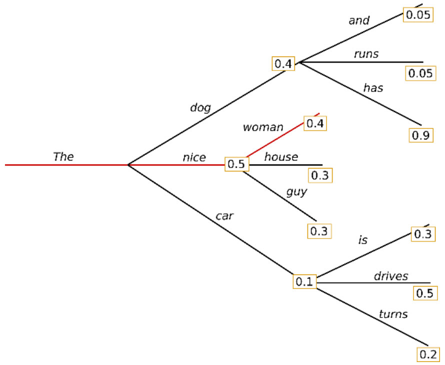
Does it always rain on Tuesdays?

No, Wednesday is the normal precipitation day. However, Tuesday can occasionally experience light rain or even a thunderstorm.

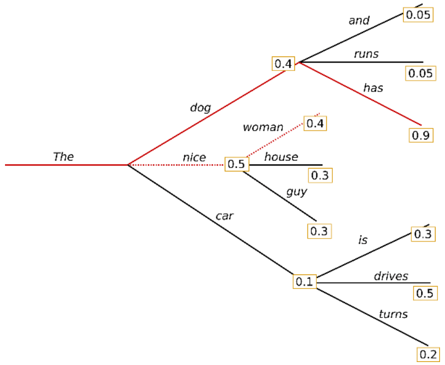
Model: text-curie-001

Temperature 1

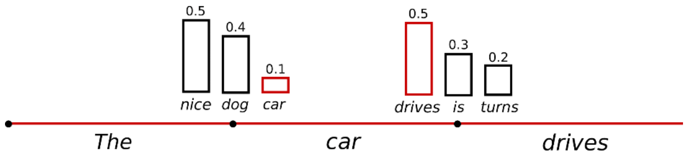
Review: Difference between Common Sampling Algorithms



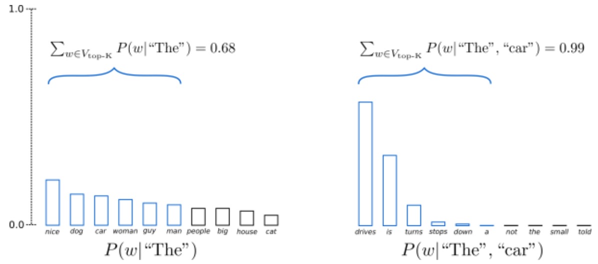
Greedy



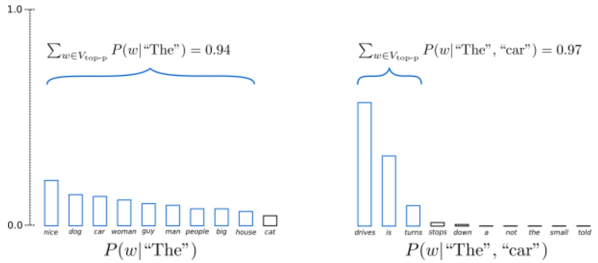
Beam Search



Random Sampling

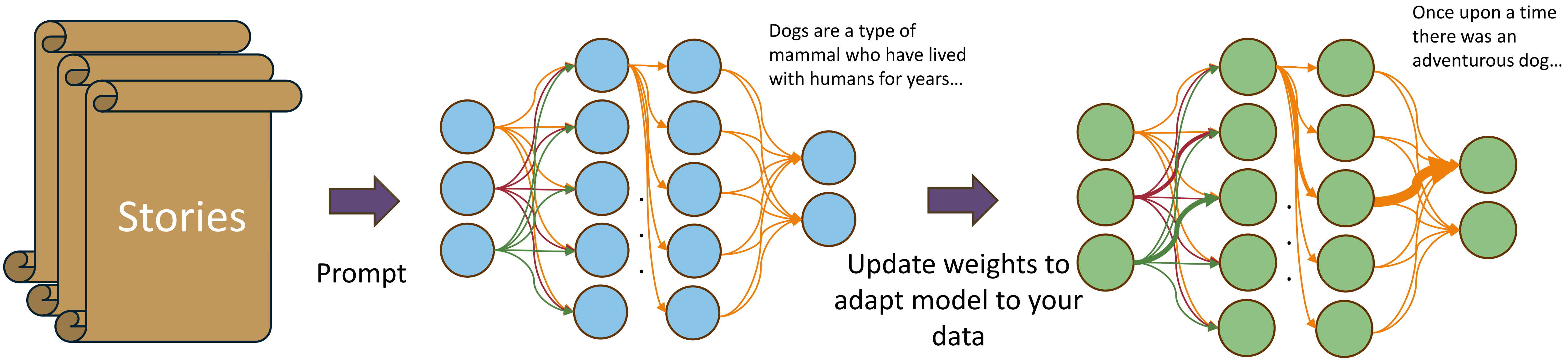


Top-K



Top-P / Nucleus Sampling

Review: Finetuning



Stories

Prompt

Dogs are a type of mammal who have lived with humans for years…

Update weights to adapt model to your data

Once upon a time there was an adventurous dog…

Your dataset

Pre-trained model (GPT)

New model (GPT+Stories)

What types of things can go wrong with finetuning?

Underfitting

Overfitting

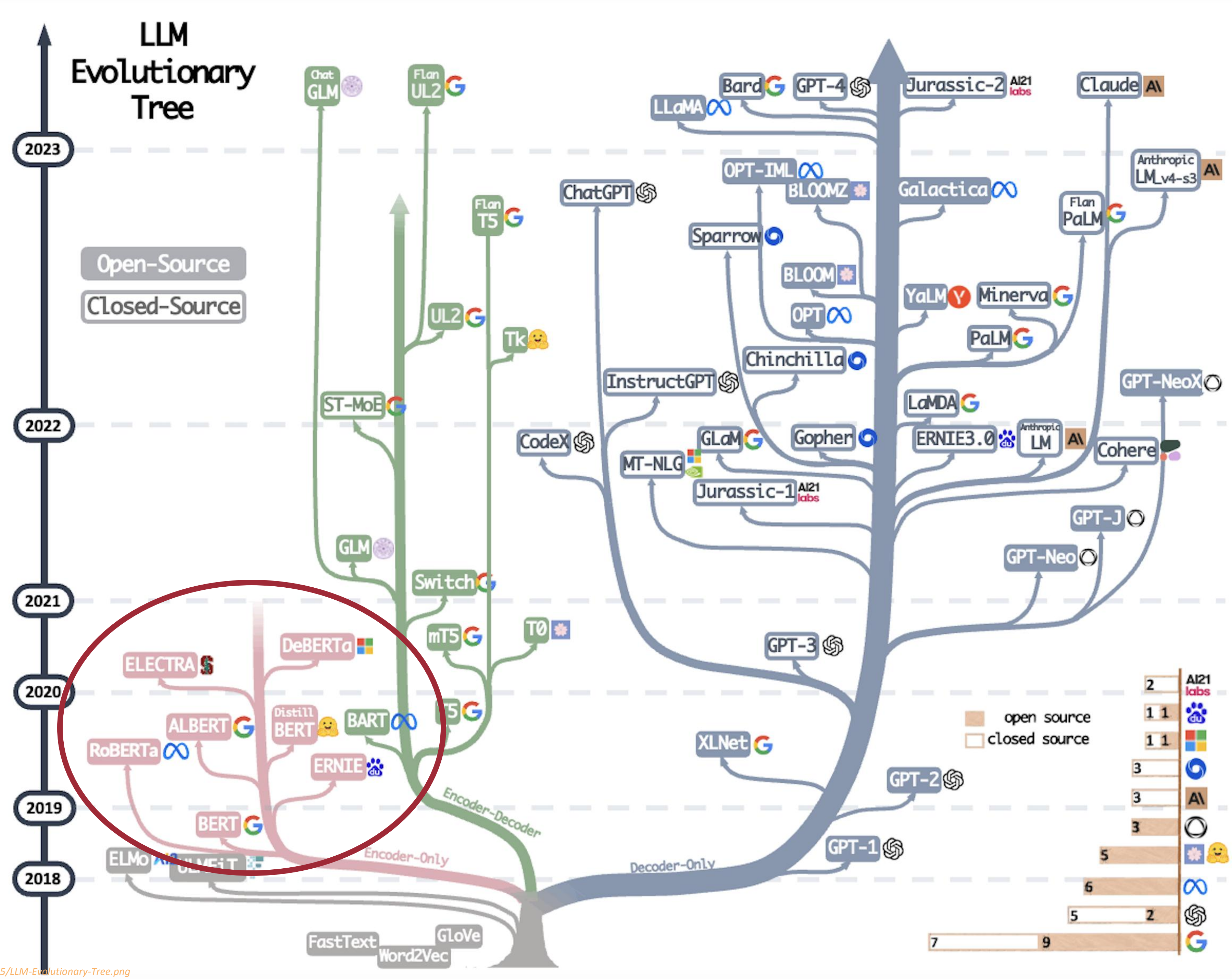
Review: What is a foundation model?

A model that captures “foundation” or core information about a modality (e.g., text, speech, images)

Pretrained on a large amount of data & able to *be* finetuned on a particular task

Self-supervised

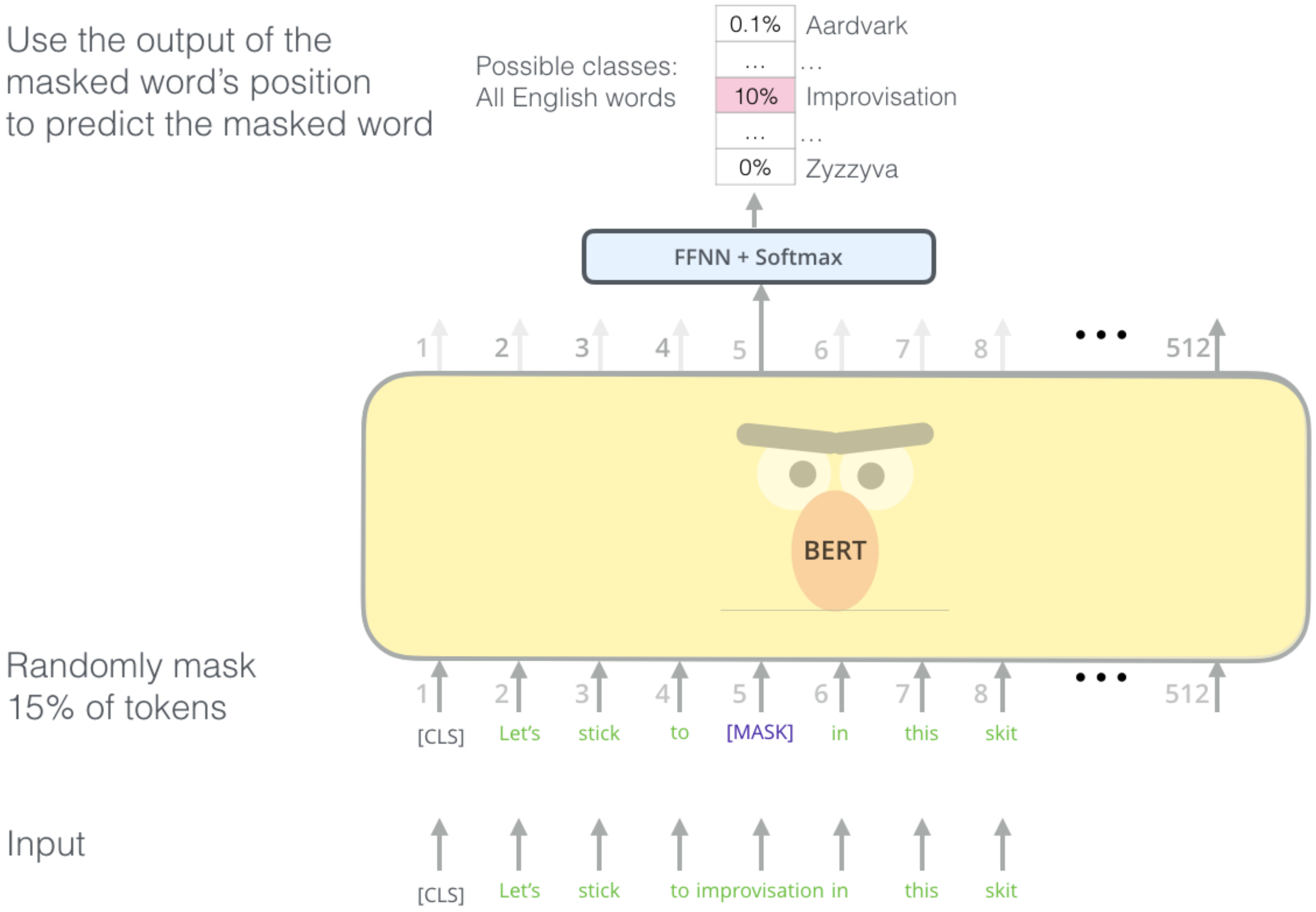
All non-finetuned large language models (LLMs) are foundation models



4/10/2025

https://blog.biocomm.ai/wp-content/uploads/2023/05/LLM-Evolutionary-Tree.png

Review: BERT (Devlin et al. 2019)



Use the output of the masked word's position to predict the masked word

Randomly mask 15% of tokens

Input

http://jalammar.github.io/illustrated-bert/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

BERT Family of Models

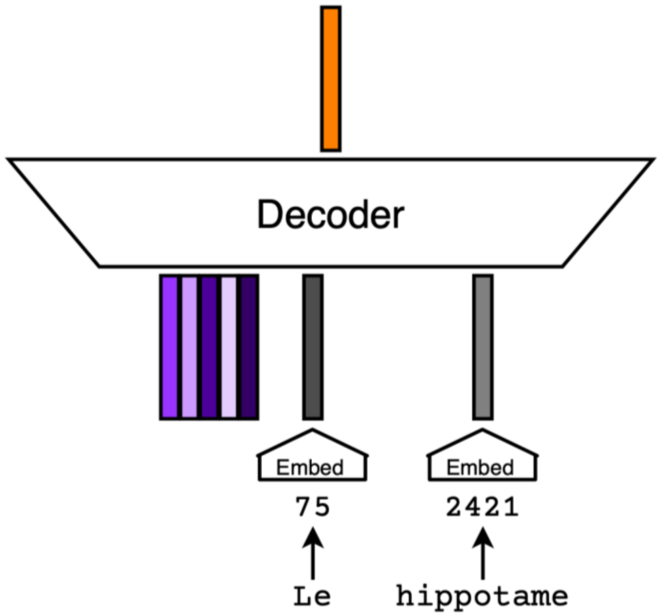
- Encoder-only
  - Input: “Corrupted” version of text sequence
  - Goal: Produce an uncorrupted version of text sequence
- How to use:
  - Finetune for a classification task
  - Extract word/sentence embeddings

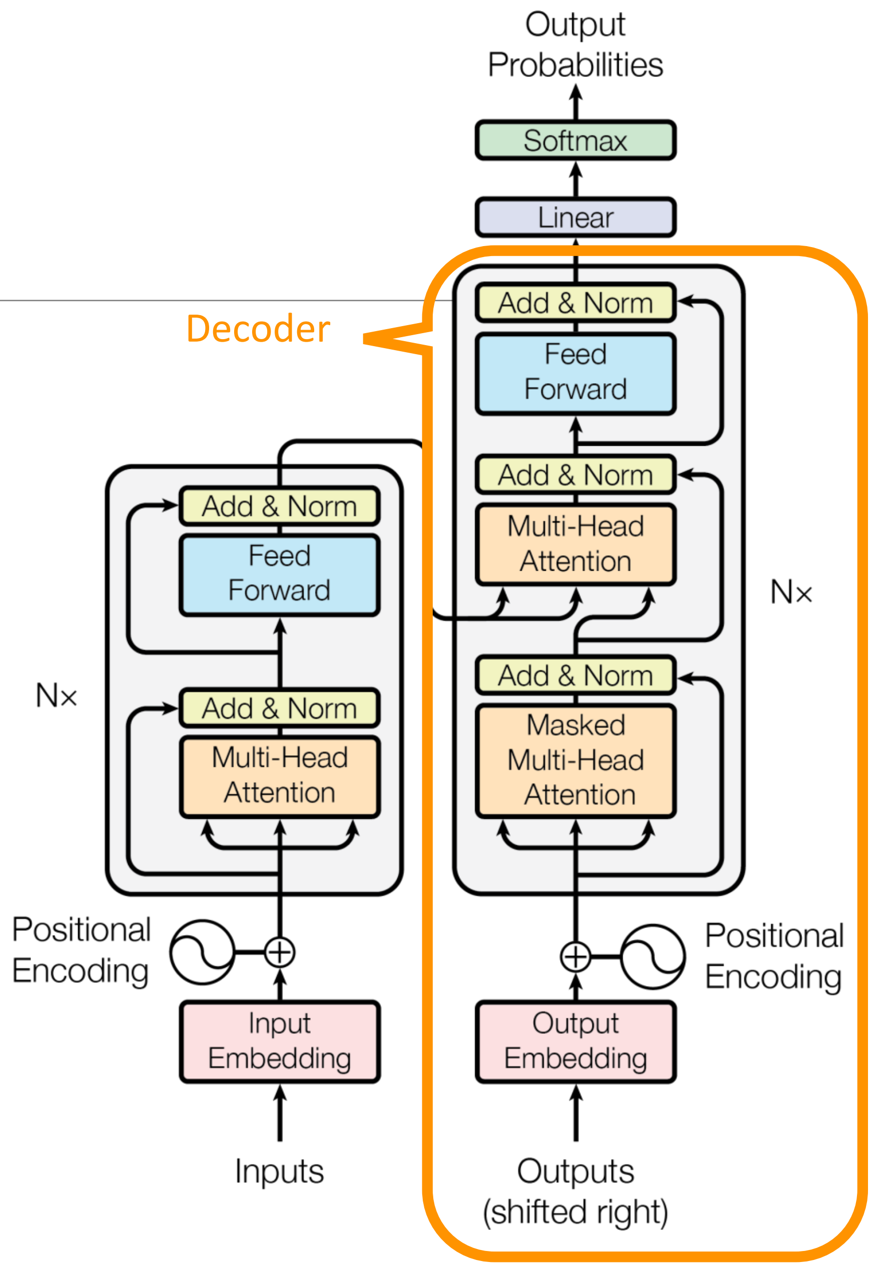
Some important BERT family members

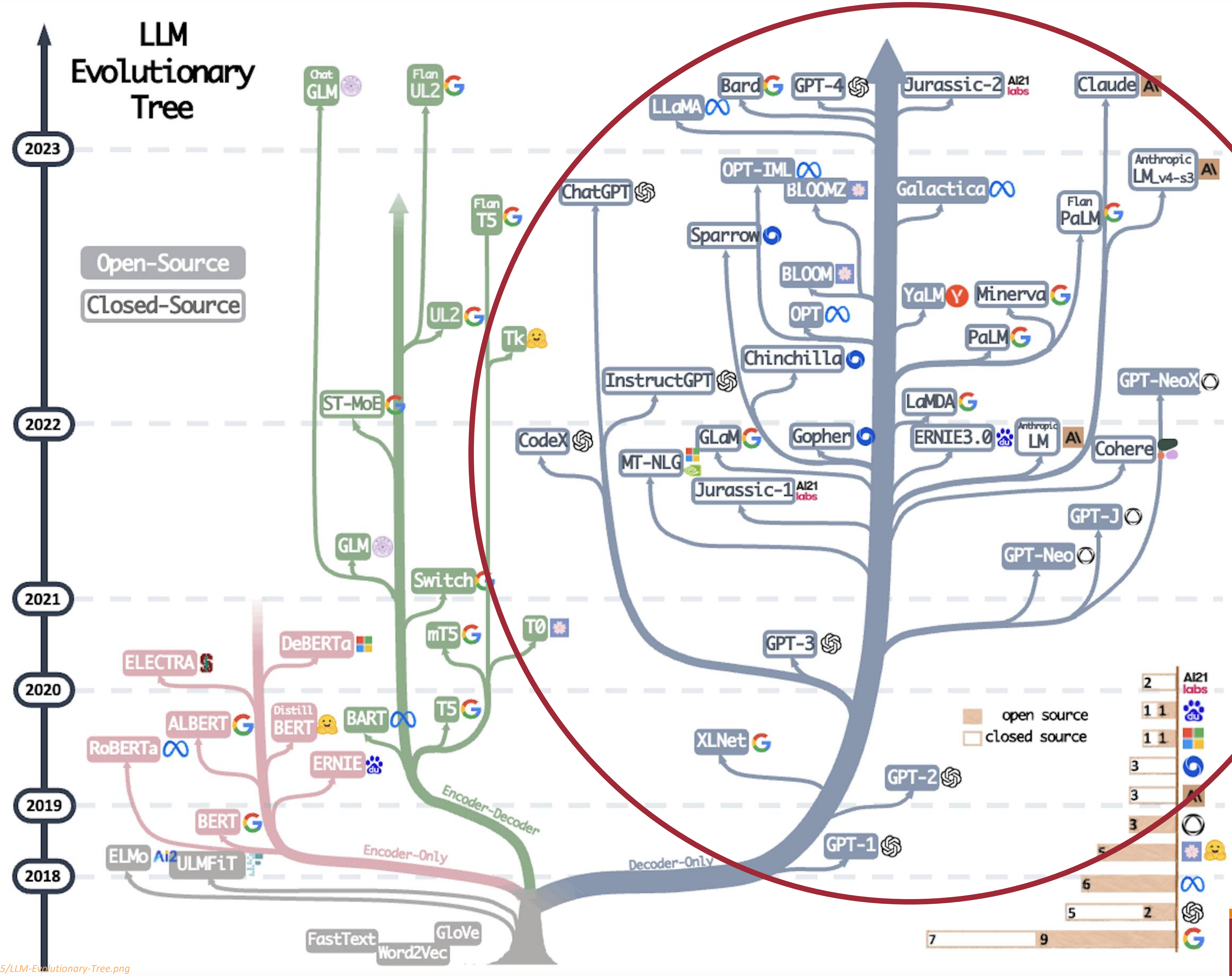
(in my opinion)

- RoBERTa (better version of the original BERT) – Liu et al. 2019 (Facebook)
- Sentence-BERT (BERT fine-tuned to give good sentence embeddings) – Reimers & Gurevych 2019 (Technische Universität Darmstadt)
- DistilBERT (lite BERT) – Sanh et al. 2019
- ALBERT (lite BERT) – Lan et al. 2020
- HuBERT (BERT for speech embeddings) – Hsu et al. 2021

Decoder-Only Models







https://blog.biocomm.ai/wp-content/uploads/2023/05/LLM-Evolutionary-Tree.png

GPT Family

- Decoder-only
  - Input: Text sequence
  - Goal: Generate the next word given the previous ones
- How to use:
  - Ask GPT* to continue from a prompt.
  - Finetune smaller GPTs for more customized generation tasks.
    - ChatGPT cannot be finetuned since it is already finetuned
  - Use OpenAI’s API to get them to fine-tune GPT* for you.
- Around GPT-2 was when pre-trained models became popular
- Around GPT-3 was when *just* prompting became reasonable to do

Other Decoder-Only Models

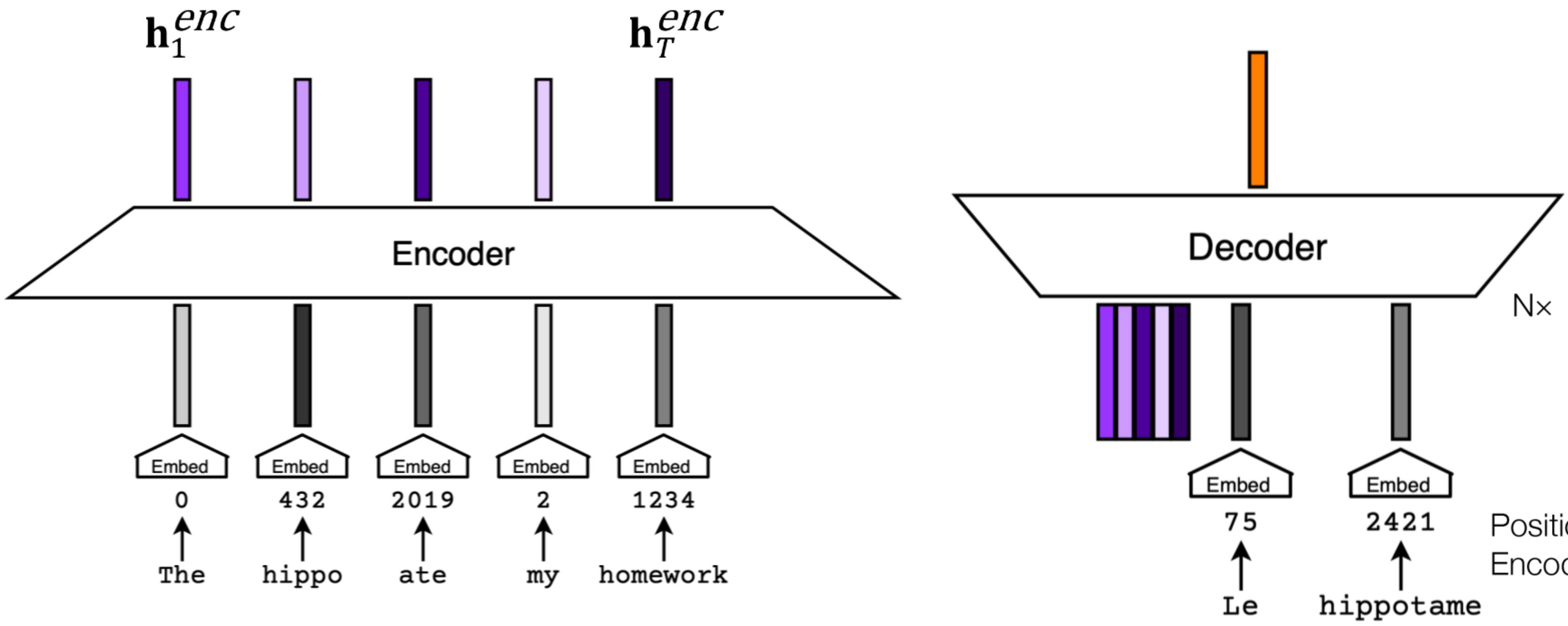
LLaMA 3/4 (Meta)

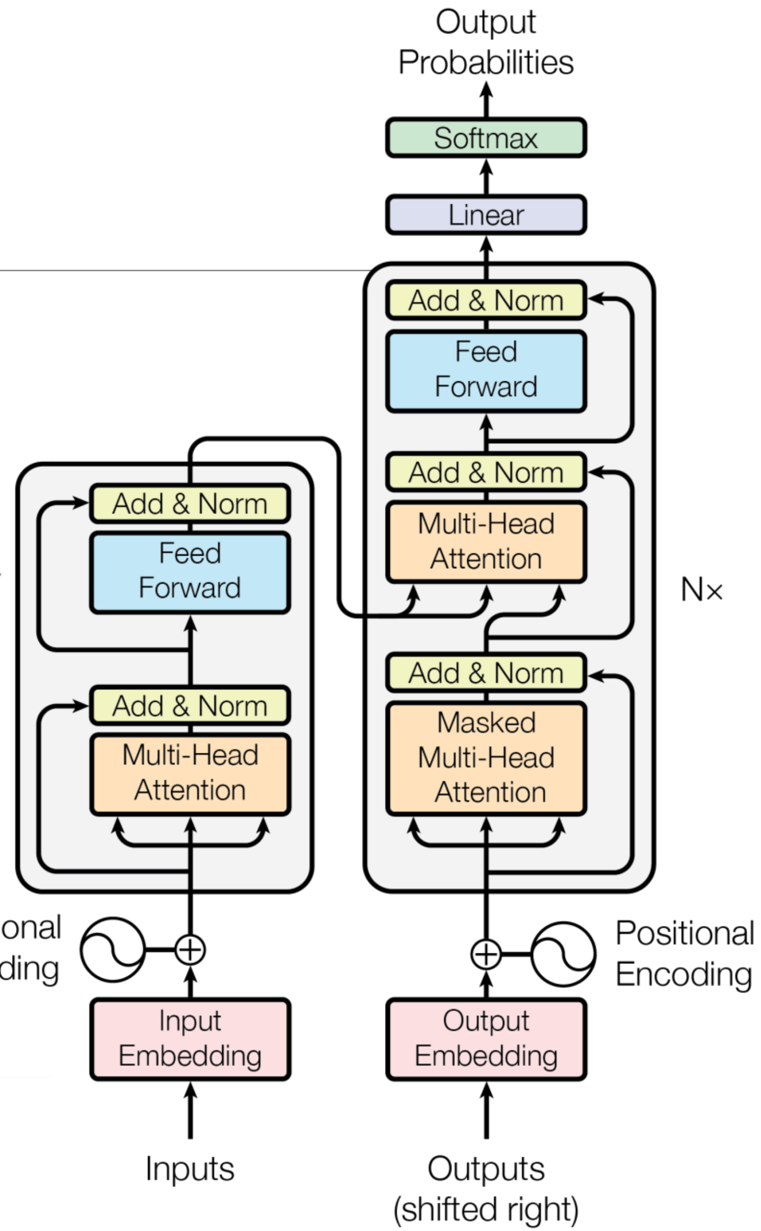
Claude 3 (Anthropic)

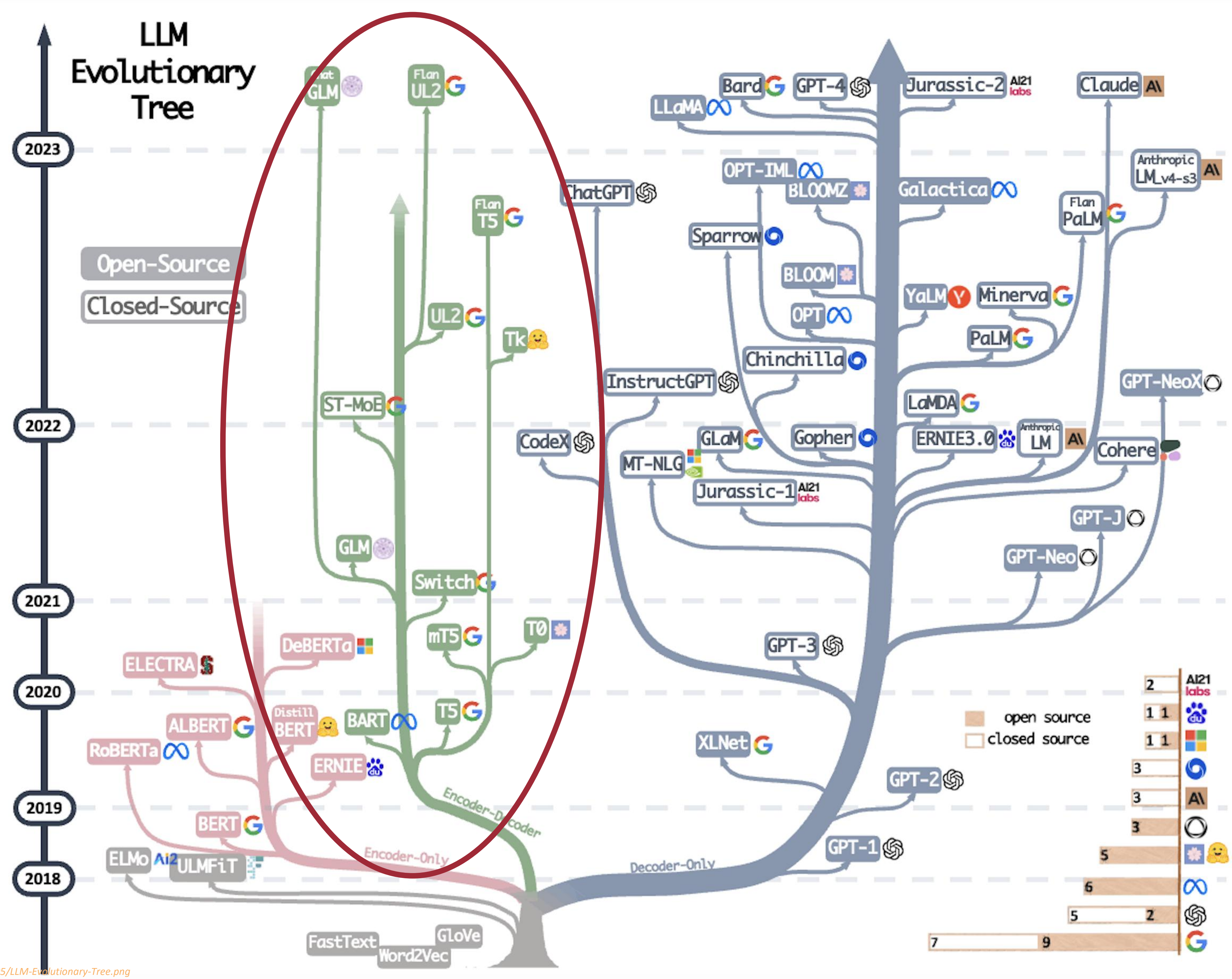
Gemma (Google)

OLMo 2 (AI2)

Encoder-Decoder Models







https://blog.biocomm.ai/wp-content/uploads/2023/05/LLM-Evolutionary-Tree.png

Enc-Dec Family of Models

- Encoder-decoder
  - Input: Text sequence with random word spans deleted
  - Goal: Generate the deleted word spans

Or

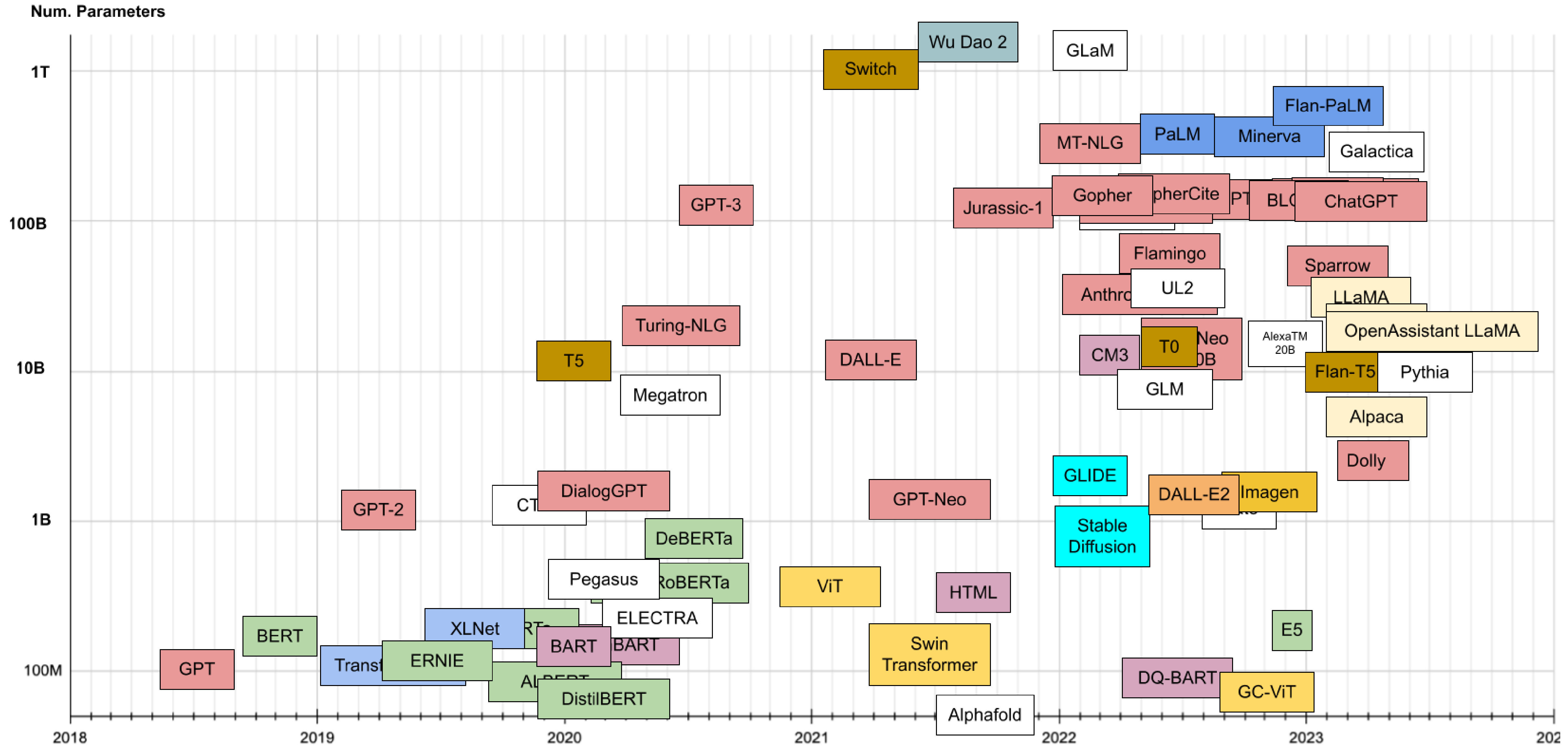
  - Input: Text sequence from “language 1”
  - Goal: Text sequence from “language 2”

- How to use:
  - Finetune smaller ones for either generation or classification tasks.
  - Prompt tuning (train a sequence of embedding which get prefixed to the input)

Some Enc-Dec family members

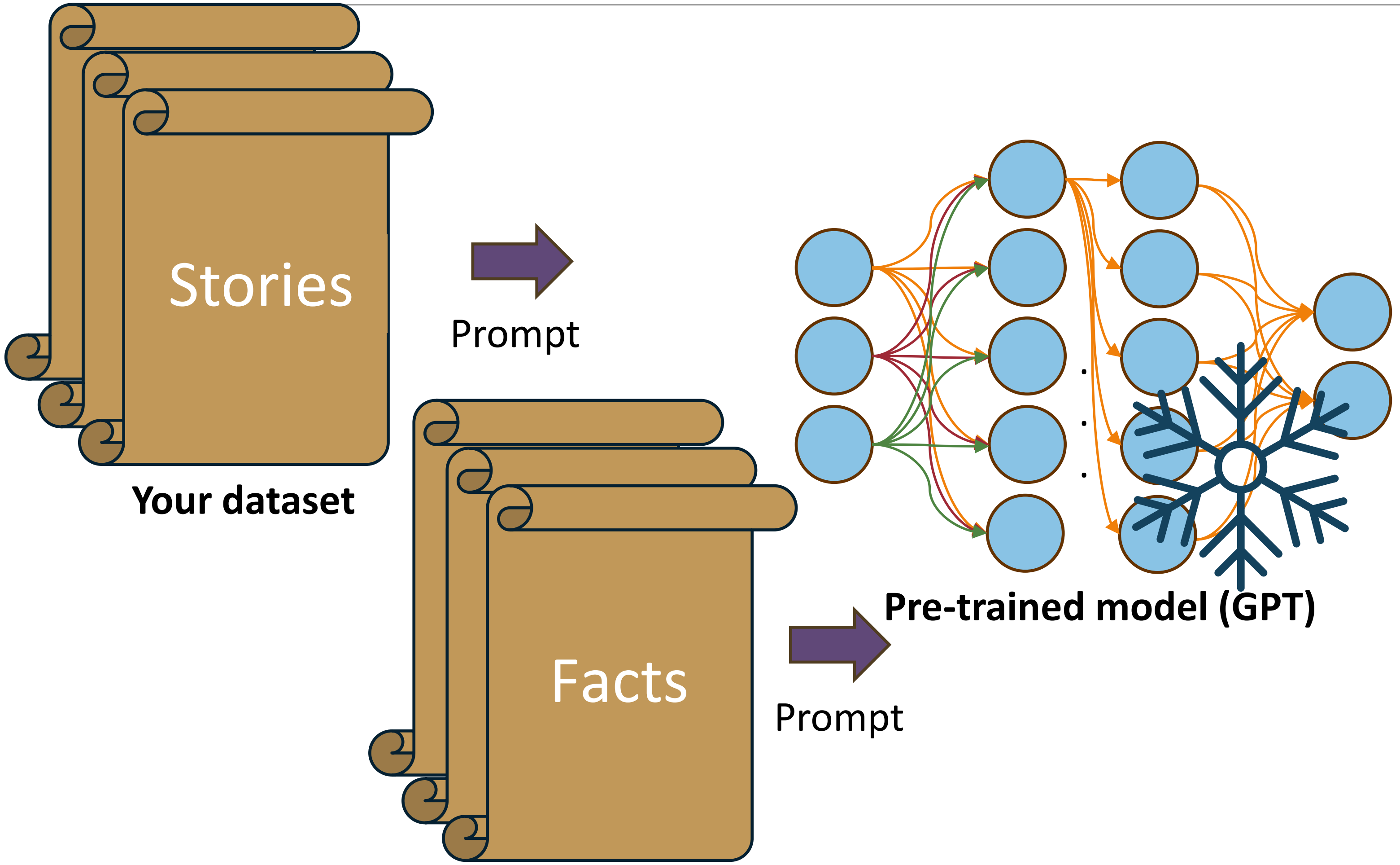
- T5 (Google)
- BART (combo of GPT and BERT) – (Facebook)

- DALL-E 2 (for caption prediction)



https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/

Prompting



Stories

Your dataset

Prompt

Facts

Prompt

Pre-trained model (GPT)

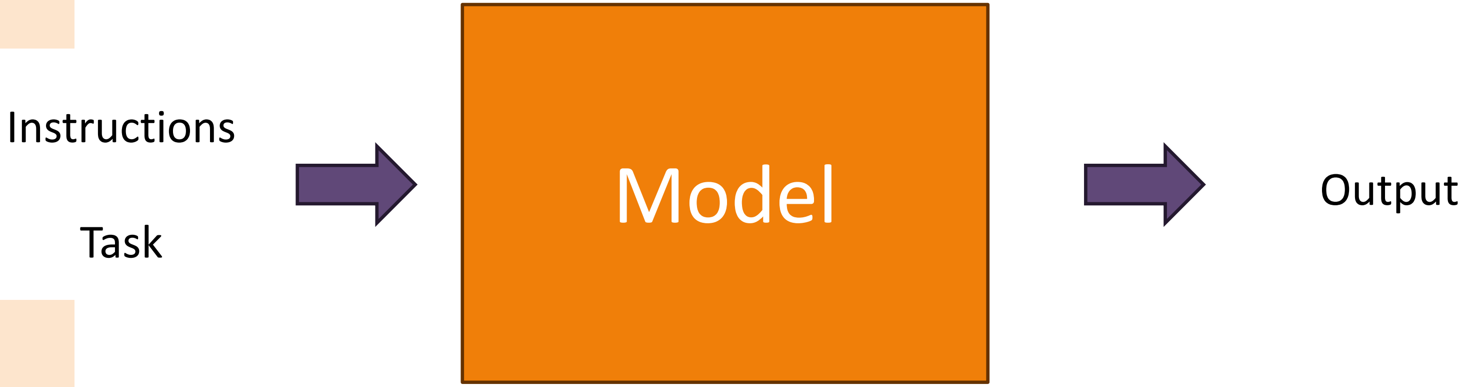
Once upon a time there was an adventurous dog…

Dogs are a type of mammal who have lived with humans for years…

Zero-shot Prompting

You are a helpful assistant. You will be tagging the parts of speech in sentences.

Sentence:
The dog ate the giant fish.



Instructions

Task

Model

Output

Few-shot Prompting

Instructions

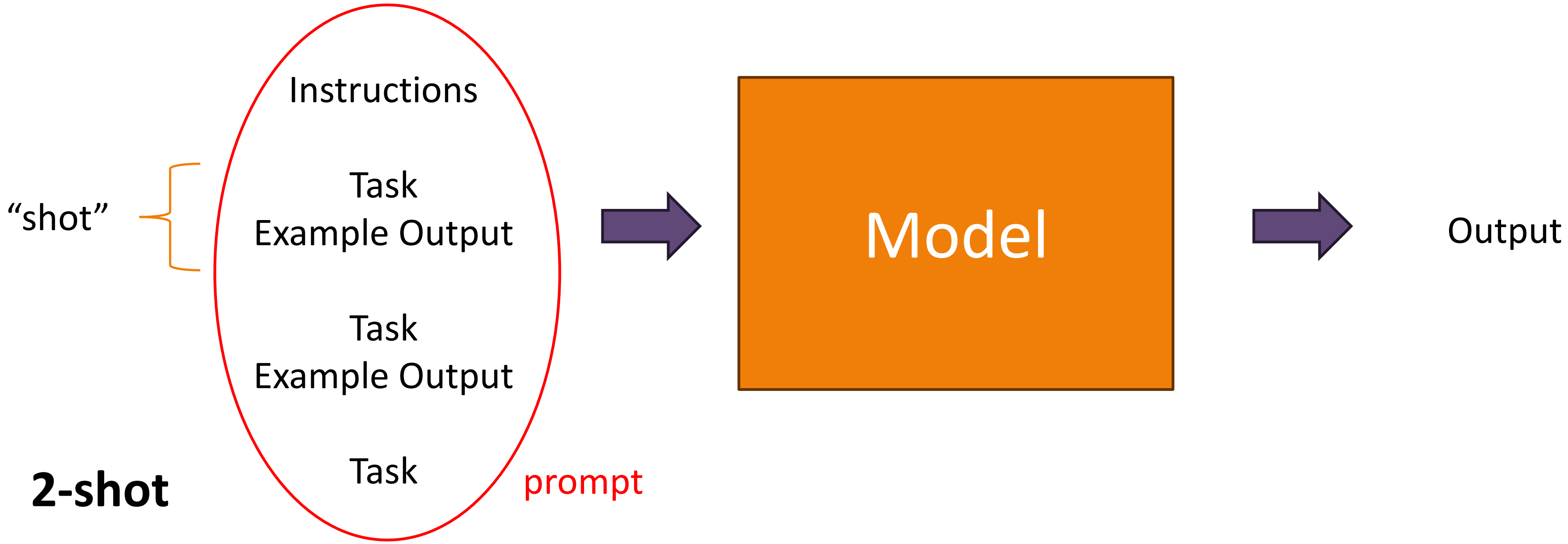
You are a helpful assistant.
You will be tagging the parts of speech in sentences.

Task

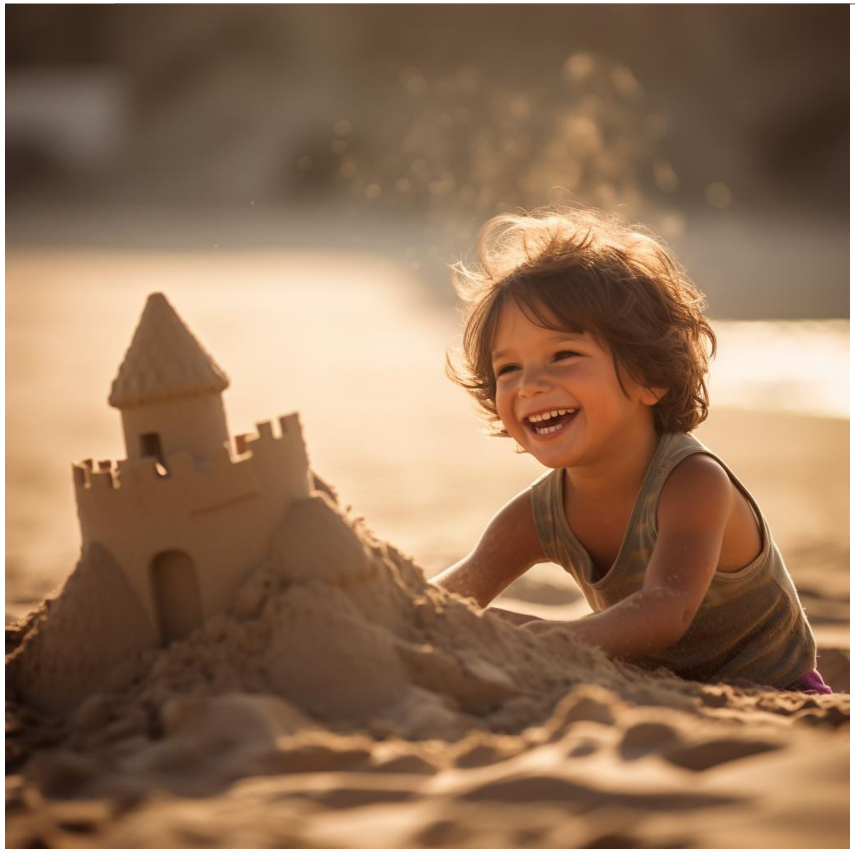
Sentence:
The dog ate the giant fish.

Example Output

The dog ate the giant fish.
D N V D Adj N



Prompt Engineering

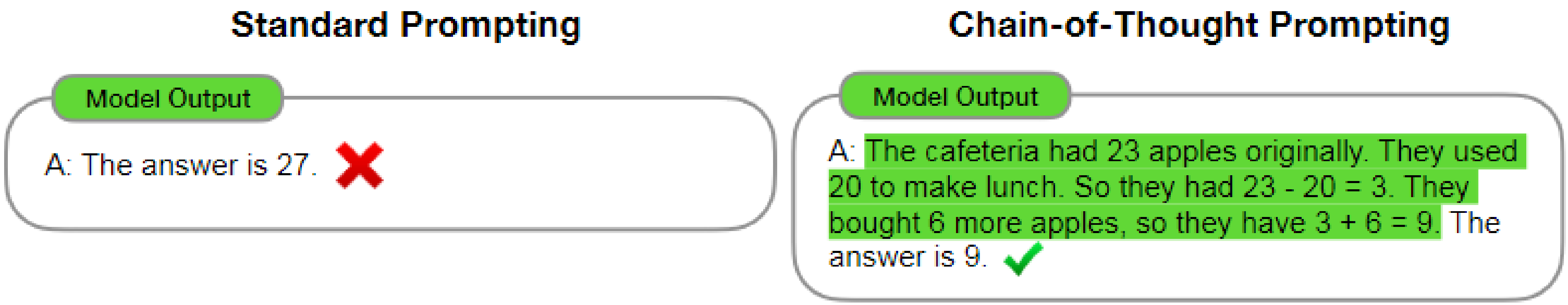


"A child playing on a sunny happy beach, their laughter as they build a simple sandcastle, emulate Nikon D6 high shutter speed action shot, soft yellow lighting."
Generated with Midjourney.
via https://zapier.com/blog/ai-art-prompts/

Need to be really specific
(also match the training data)

Chain-of-Thought Prompting

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



Standard Prompting

Model Output

A: The answer is 27. ✗

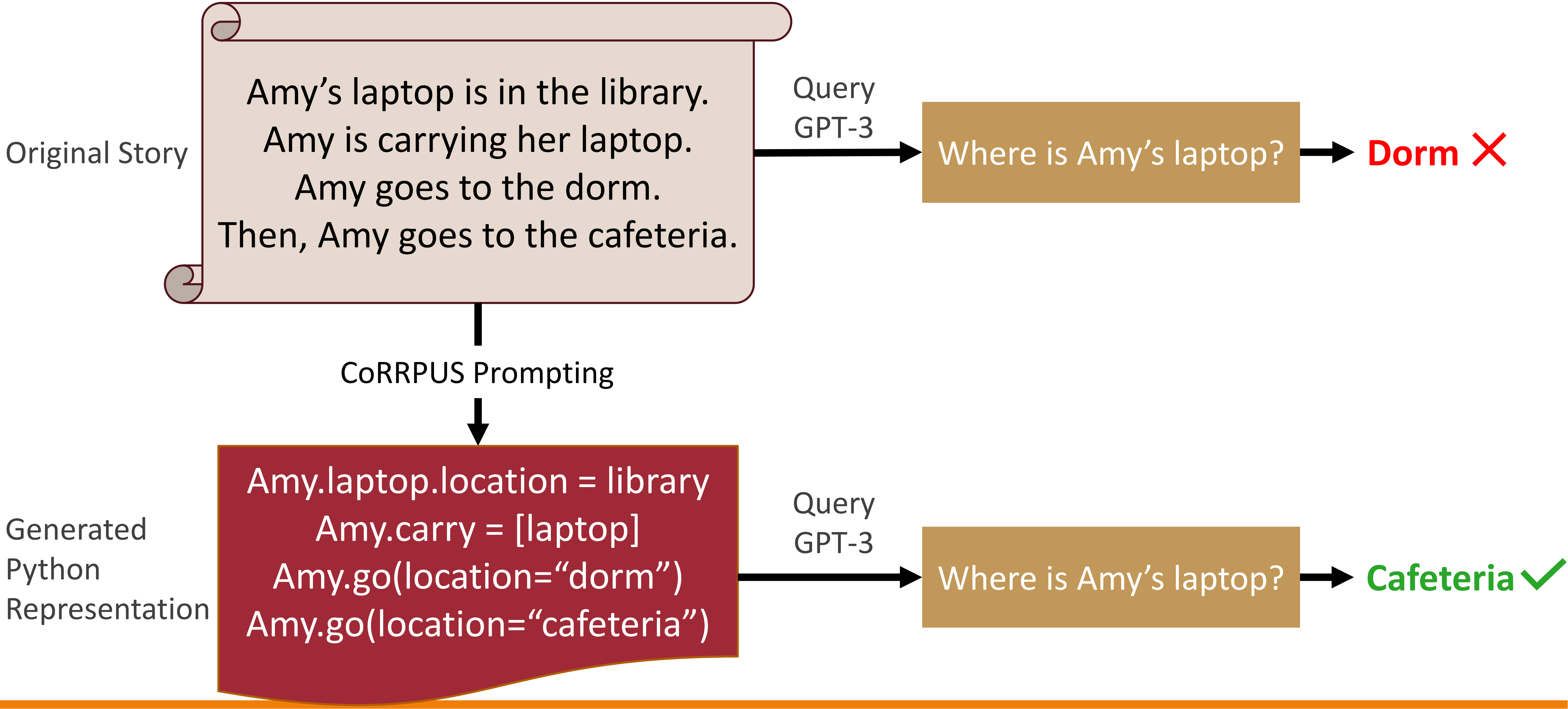
Chain-of-Thought Prompting

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. ✓

Part of Figure 1 from J. Wei *et al.*, "Chain of Thought Prompting Elicits Reasoning in Large Language Models," in *International Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA & Online, Jun. 2022. doi: 10.48550/arXiv.2201.11903.

CoRRPUS (Code Representations to Reason & Prompt over for Understanding in Stories)



Original Story

Amy's laptop is in the library.
Amy is carrying her laptop.
Amy goes to the dorm.
Then, Amy goes to the cafeteria.

Generated Python Representation

```
Amy.laptop.location = library
Amy.carry = [laptop]
Amy.go(location="dorm")
Amy.go(location="cafeteria")
```

Dong, Y. R., **Martin, L. J.**, & Callison-Burch, C.
"CoRRPUS: Code-based Structured Prompting for Neurosymbolic Story Understanding." *Findings of ACL 2023.*

CoRRPUS Chain-of-Thought Prompting

Three versions that are initialized the same:

Comment

```
def story(self):
    ## Mary moved to the bathroom.
    self.Mary.location = “bathroom”
    ## Mary got the football there.
    self.Mary.inventory.append(“football”)
    …
```

Specific Functions

```
self.Mary_moved_to_the_bathroom()
self.Mary_got_the_football_there()
self.John_went_to_the_kitchen()
self.Mary_went_back_to_the_garden()

def Mary_moved_to_the_bathroom()
    self.Mary.location=“bathroom”
def Mary_got_the_football_there():
…
```

Abstract Functions

```
def go(self, character, location):
    character.location = location
    for item in character.inventory:
        item.location = location
def pick_up(): …

def story(self):
    ## Mary moved to the bathroom.
    self.go(character=self.Mary,
    location = “bathroom”)
    …
```

Dong, Y. R., **Martin, L. J.**, & Callison-Burch, C.
“CoRRPUS: Code-based Structured Prompting for Neurosymbolic Story Understanding.” *Findings of ACL 2023.*

Tested On 2 Tasks

bAbI (Weston et al. 2015)

- Task 2: Stories tracking objects that characters carry

Re^3 (Yang et al. 2022)

- Identifying inconsistencies in stories (e.g., descriptions of characters’ appearances, relationships)
- Stories were generated from a list of facts (the premise). They also generated premises with a contradiction.

Dong, Y. R., **Martin, L. J.,** & Callison-Burch, C.
“CoRRPUS: Code-based Structured Prompting for Neurosymbolic Story Understanding.” *Findings of ACL 2023.*

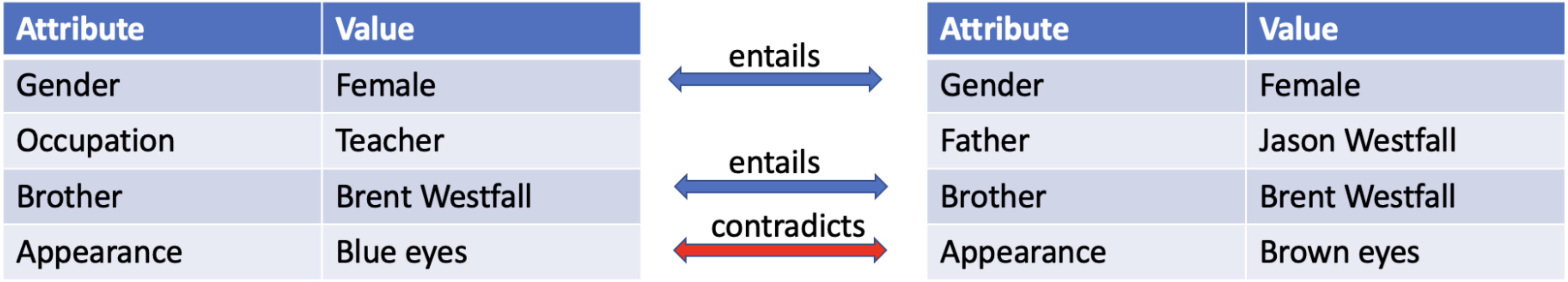
bAbI (Weston et al. 2015)

Method	**# Shot**	**Accuracy ↑**
Random	-	25%
GPT-3	1	56.5%
Chain of Thought (Creswell et al. 2022)	1	46.4%
Selection-Inference (Creswell et al. 2022)	1	29.3%
Dual-System (Nye et al. 2021)	10	100%
CoRRPUS (comment)	**1**	**67.0%**
CoRRPUS (specific)	**1**	**78.7%**
CoRRPUS (abstract)	**1**	**99.1%**

Dong, Y. R., **Martin, L. J.**, & Callison-Burch, C.
“CoRRPUS: Code-based Structured Prompting for Neurosymbolic Story Understanding.” *Findings of ACL 2023.*

Re^3

The task is to see what stories match what premises based on the facts extracted from both.



Joan Westfall premise

Attribute	Value
Gender	Female
Occupation	Teacher
Brother	Brent Westfall
Appearance	Blue eyes

Joan Westfall in story

Attribute	Value
Gender	Female
Father	Jason Westfall
Brother	Brent Westfall
Appearance	Brown eyes

Takeaway: structured representations help!

Re^3 (Yang et al. 2022)

Method	ROC-AUC ↑
Random	0.5
GPT-3	0.52
Entailment (Yang et al. 2022)	0.528
Entailment with Dense Passage Retrieval (Yang et al. 2022)	0.610
Attribute Dictionary → Sentence (Yang et al. 2022)	0.684
CoRRPUS (comment)	**0.751**
CoRRPUS (specific)	**0.794**
CoRRPUS (abstract)	**0.704**

→ Probably because functions like `set_age(self, character, age)` complicate more than they help.

Dong, Y. R., **Martin, L. J.,** & Callison-Burch, C.
"CoRRPUS: Code-based Structured Prompting for Neurosymbolic Story Understanding." *Findings of ACL 2023.*

Tricks of the Trade

Instruction-tuned models like GPT-3.5 and Mistral-7B-Instruct like to be given a "role" first (e.g., "You are a helpful writing assistant.")

The more defined the task, the better

- More details
- One thing to do at a time

LLMs are overly confident (like people on the internet)

- To "objectively" have the model evaluate something, you should create a new instance and ask it

Chain-of-thought prompting helps models come up with better answers

They will "Yes and…" your prompt

Note that this is very similar to HW 3!

Your Turn

Think of something you’re an expert in. It can be anything!

Ask your LLM to give you information about that topic. Ask in different ways about different things.

What does it do well with?

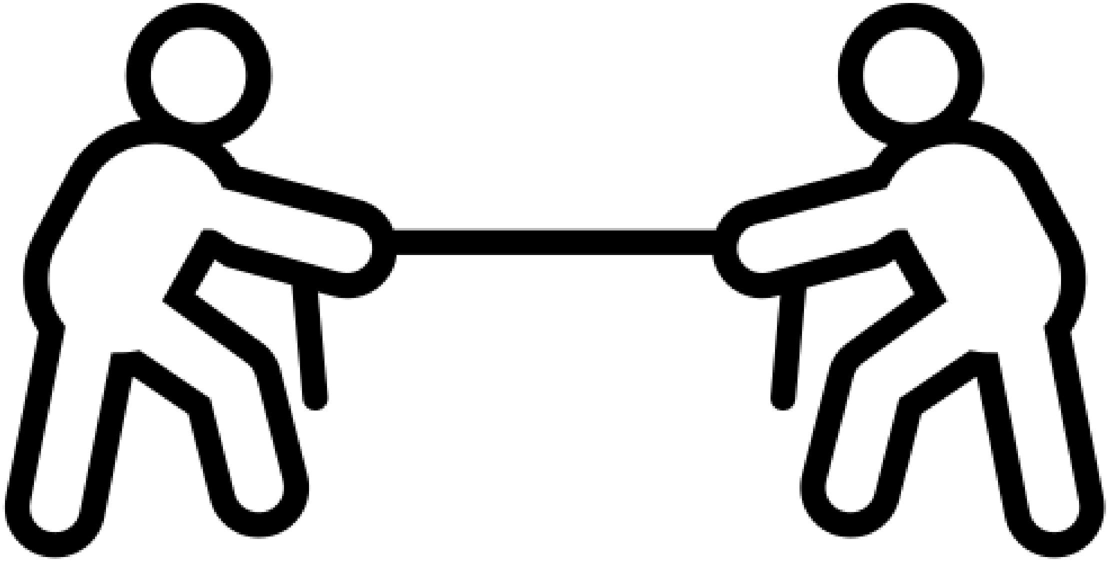
What does it not do well with?

Dealing with any language models

Likelihoods → Not cause & effect

What is probable might not be possible.

Lara’s Language Model Tradeoff



Coherence

Originality

https://thenounproject.com/icon/tug-of-war-1016981/

For next lecture…

Read the Bender et al. paper on Stochastic Parrots!