Decoding, Pretrained Models, and Finetuning

Instructor: Lara J. Martin (she/they)

TA: Omkar Kulkarni (he)

https://laramartin.net/NLP-class/

Learning Objectives

Consider when to use various sampling algorithms

Discuss the uses of finetuning

Distinguish between few-shot and zero-shot prompting

Try common prompting techniques like chain-of-thought

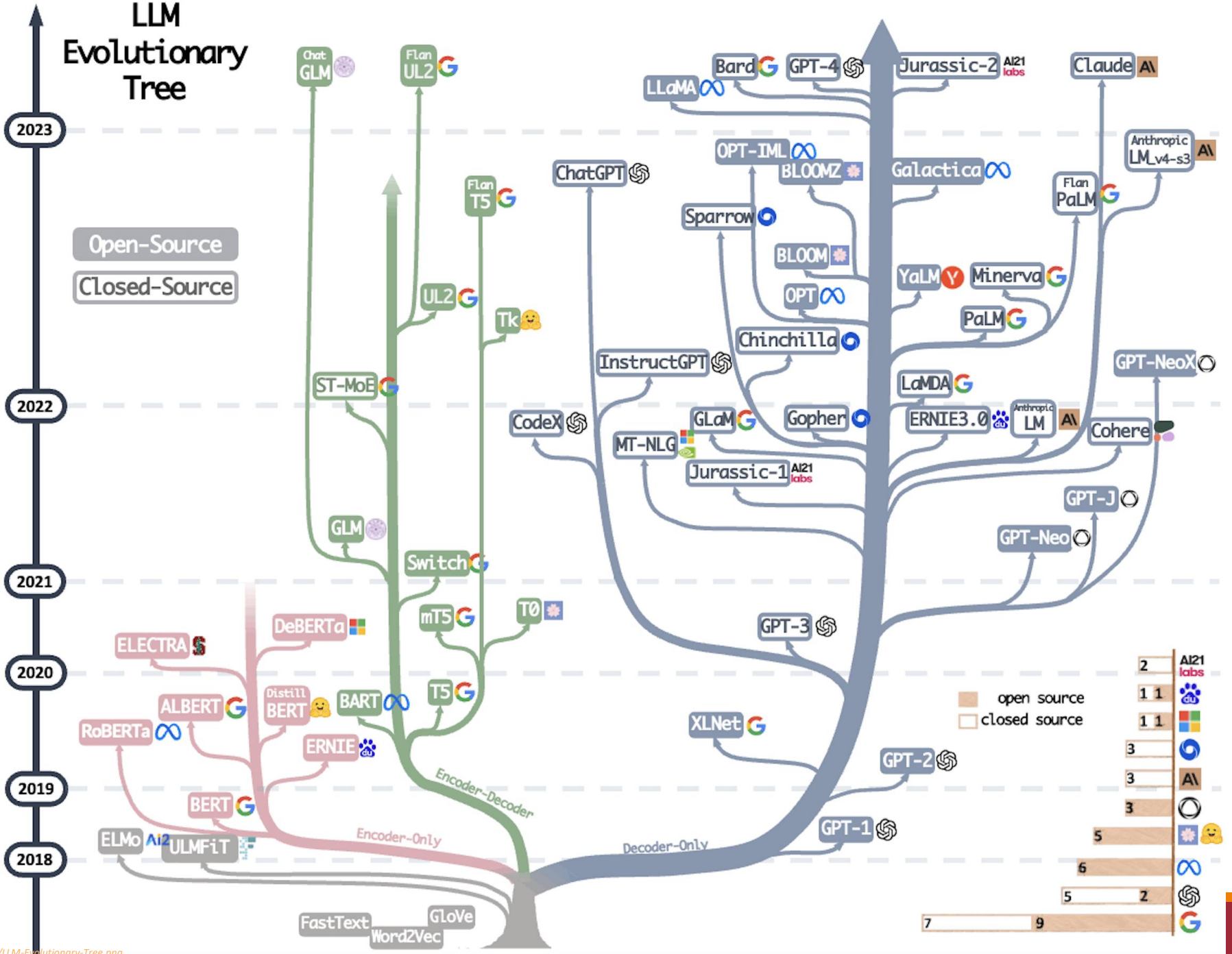
Review: What is a foundation model?

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



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

Review: Generating Text

Also sometimes called decoding 😕

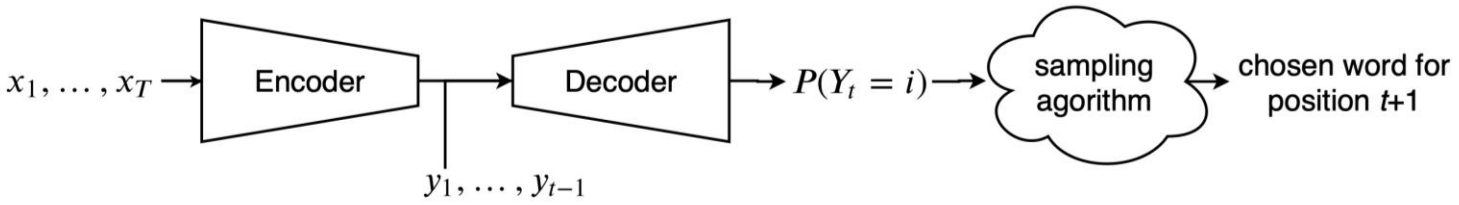
To generate text, we need an algorithm that selects tokens given the predicted probability distributions.

Examples:

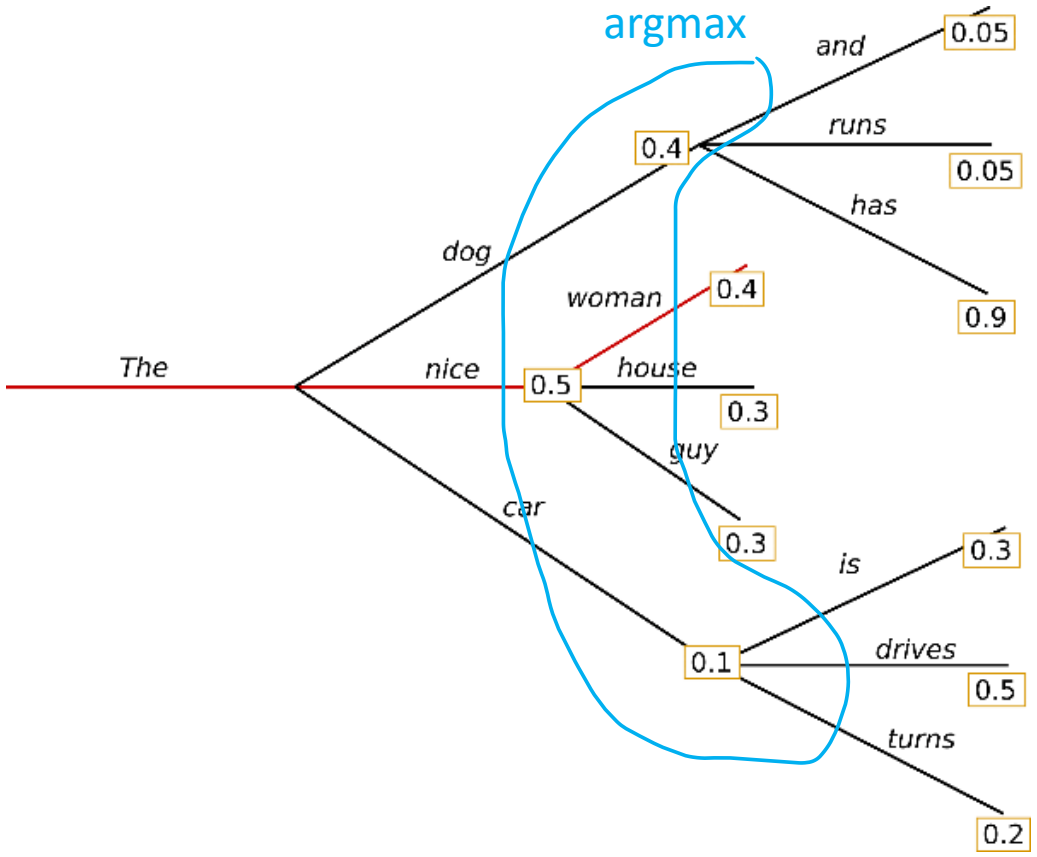
Argmax

Beam search

Random sampling



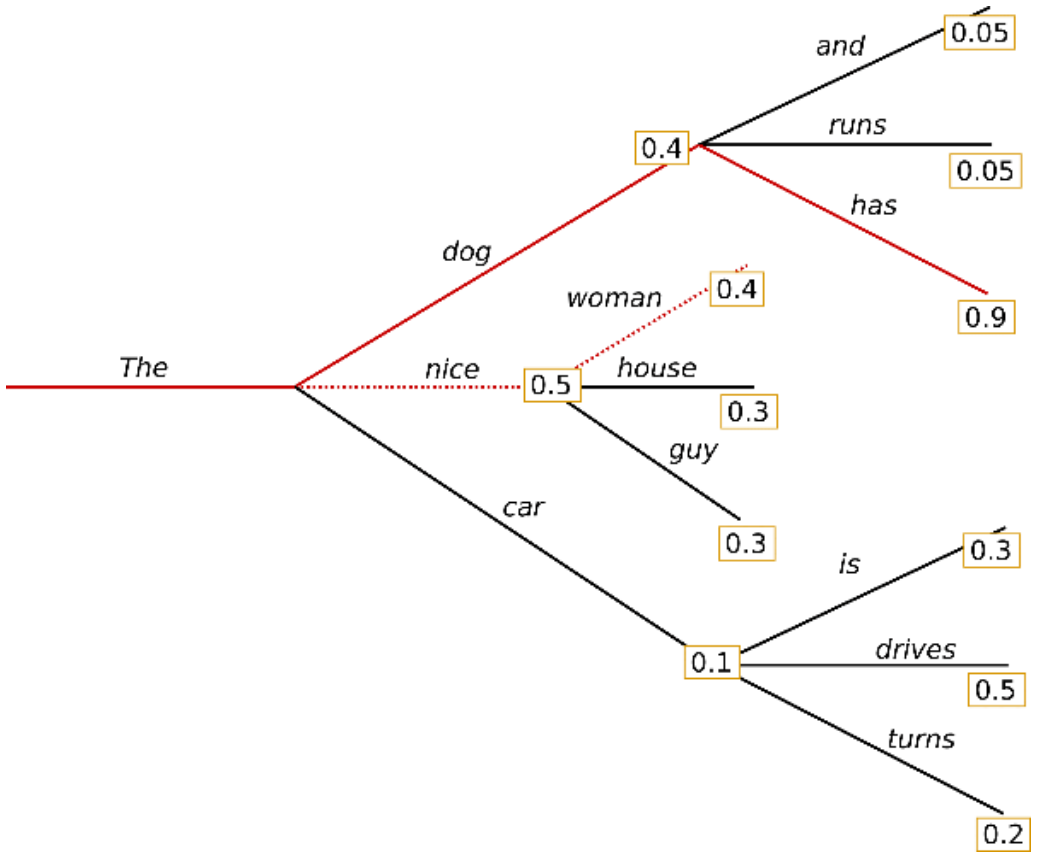
Greedy Search



https://huggingface.co/blog/how-to-generate

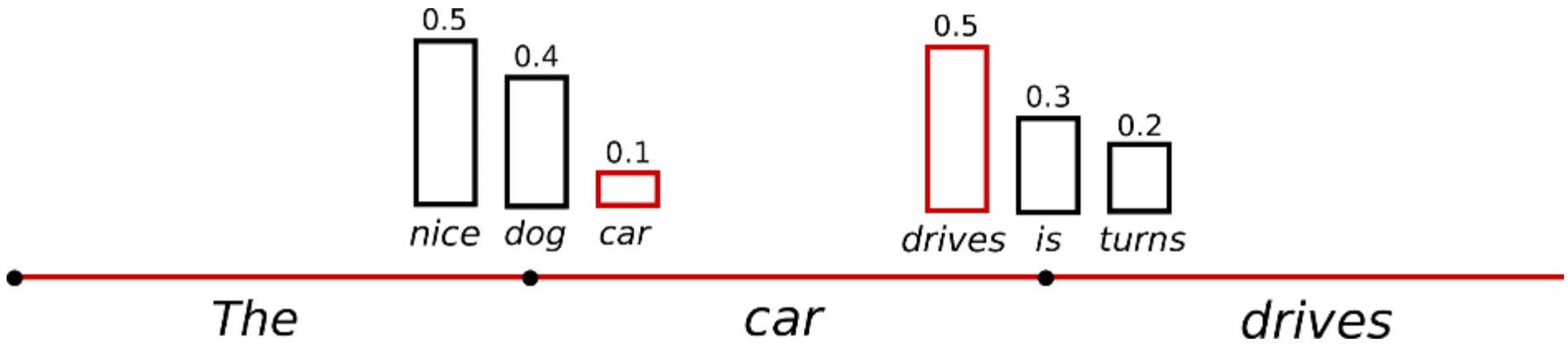
Beam Search

Number of beams = 2



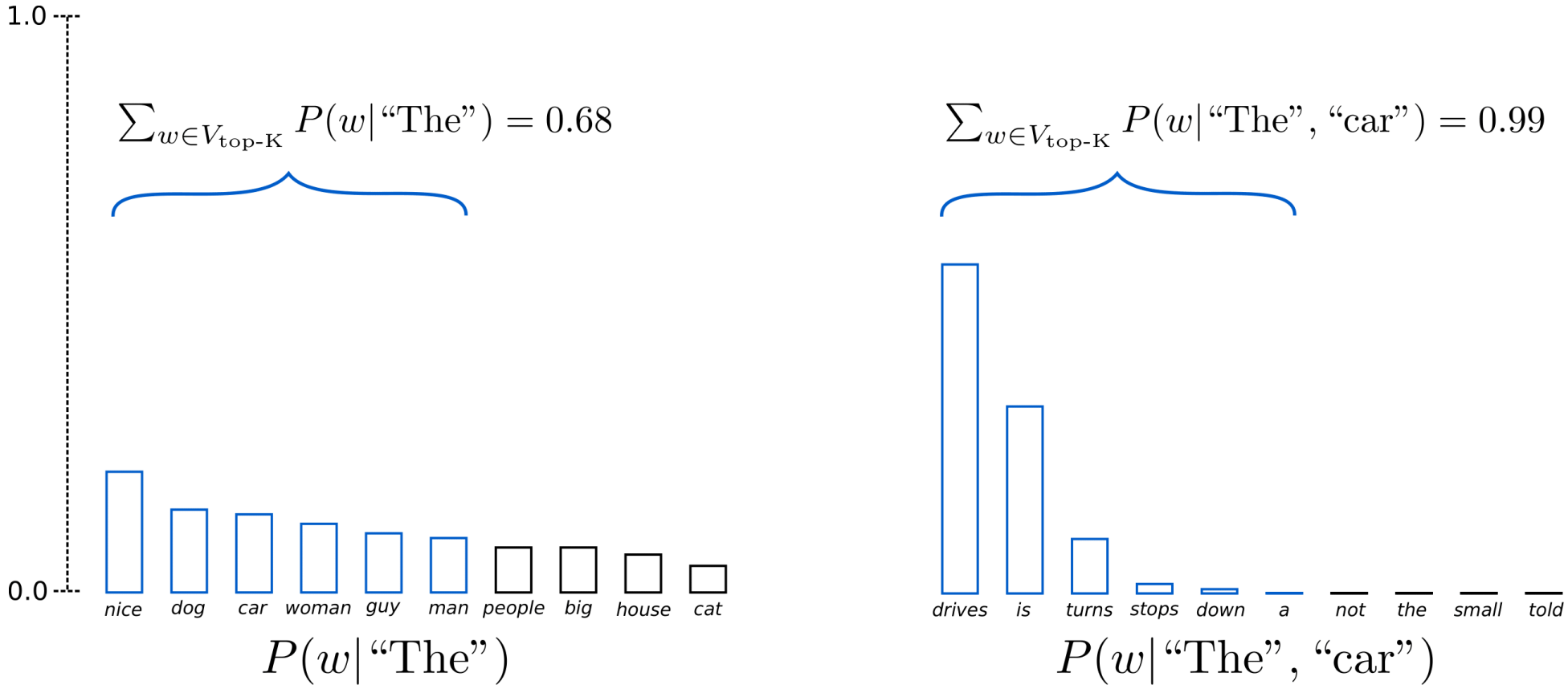
https://huggingface.co/blog/how-to-generate

Random Sampling



https://huggingface.co/blog/how-to-generate

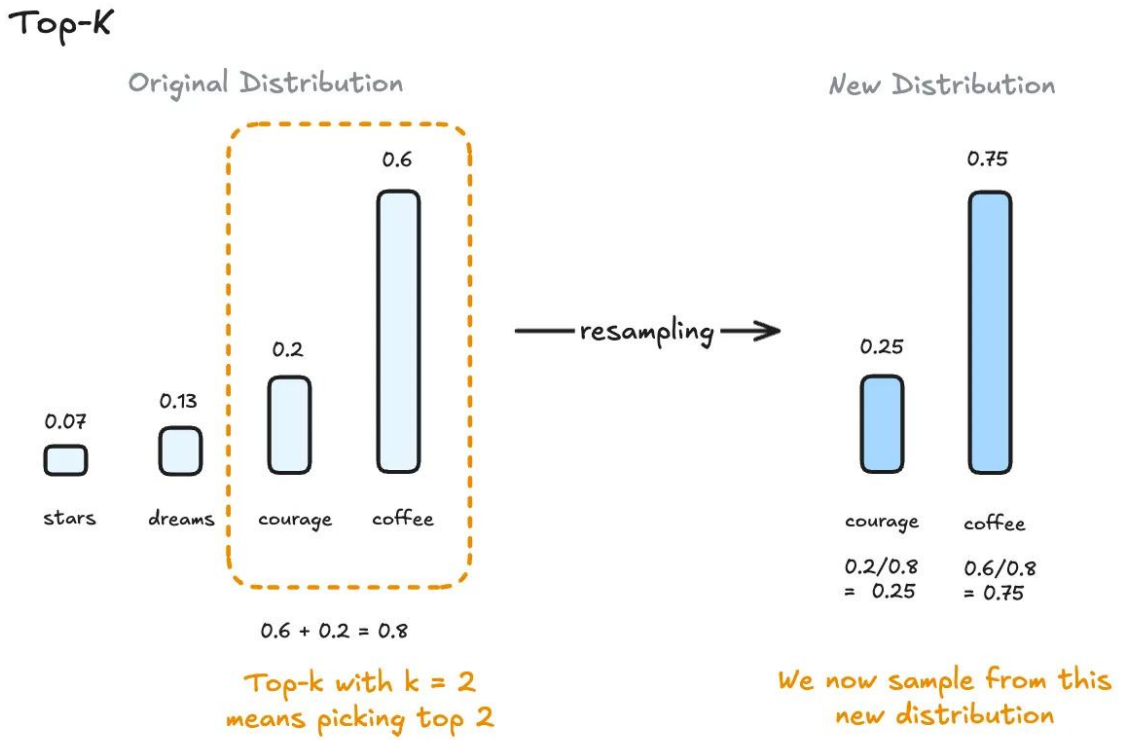
Top-K Sampling



A. Holtzman, J. Buys, M. Forbes, and Y. Choi, “The Curious Case of Neural Text Degeneration,” in *International Conference on Learning Representations (ICLR)*, 2020, p. 16. https://openreview.net/forum?id=rygGQyrFvH

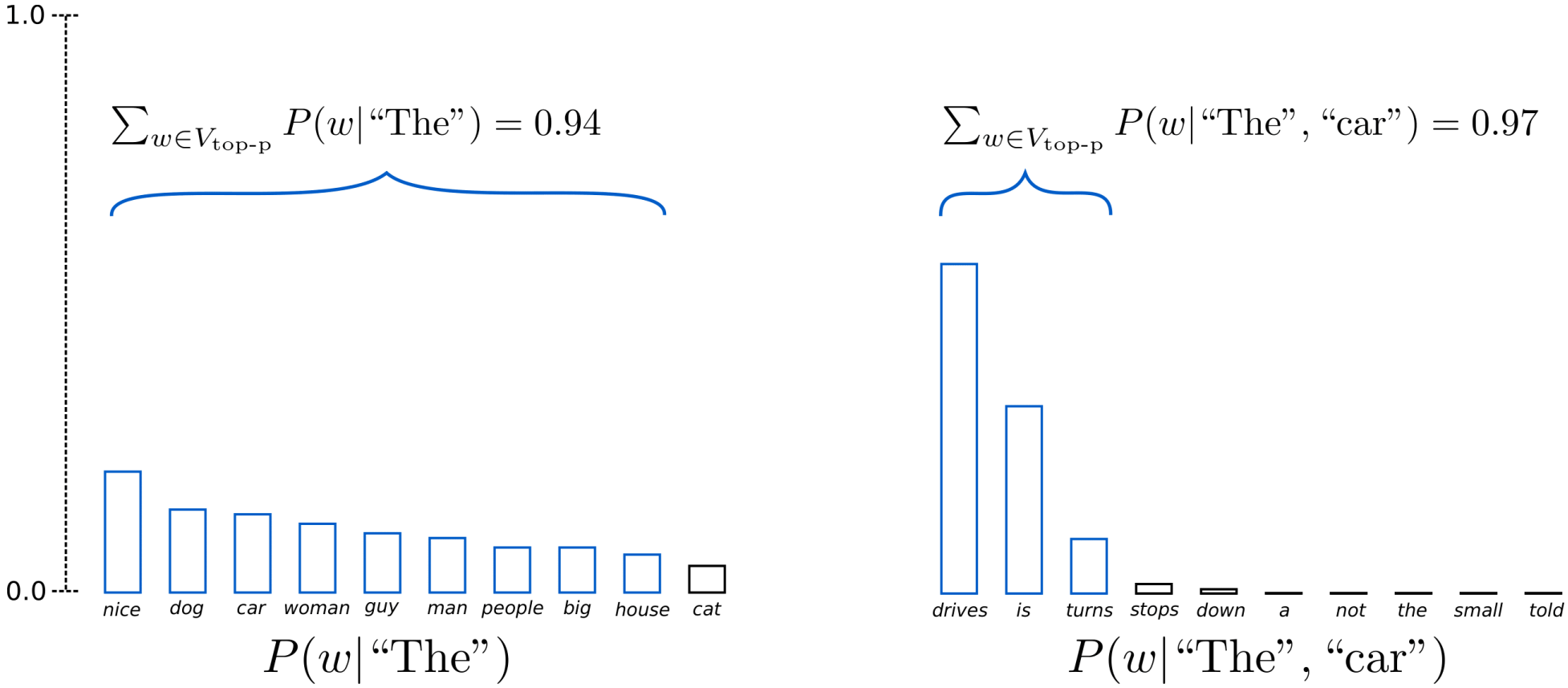
https://huggingface.co/blog/how-to-generate

Resampling



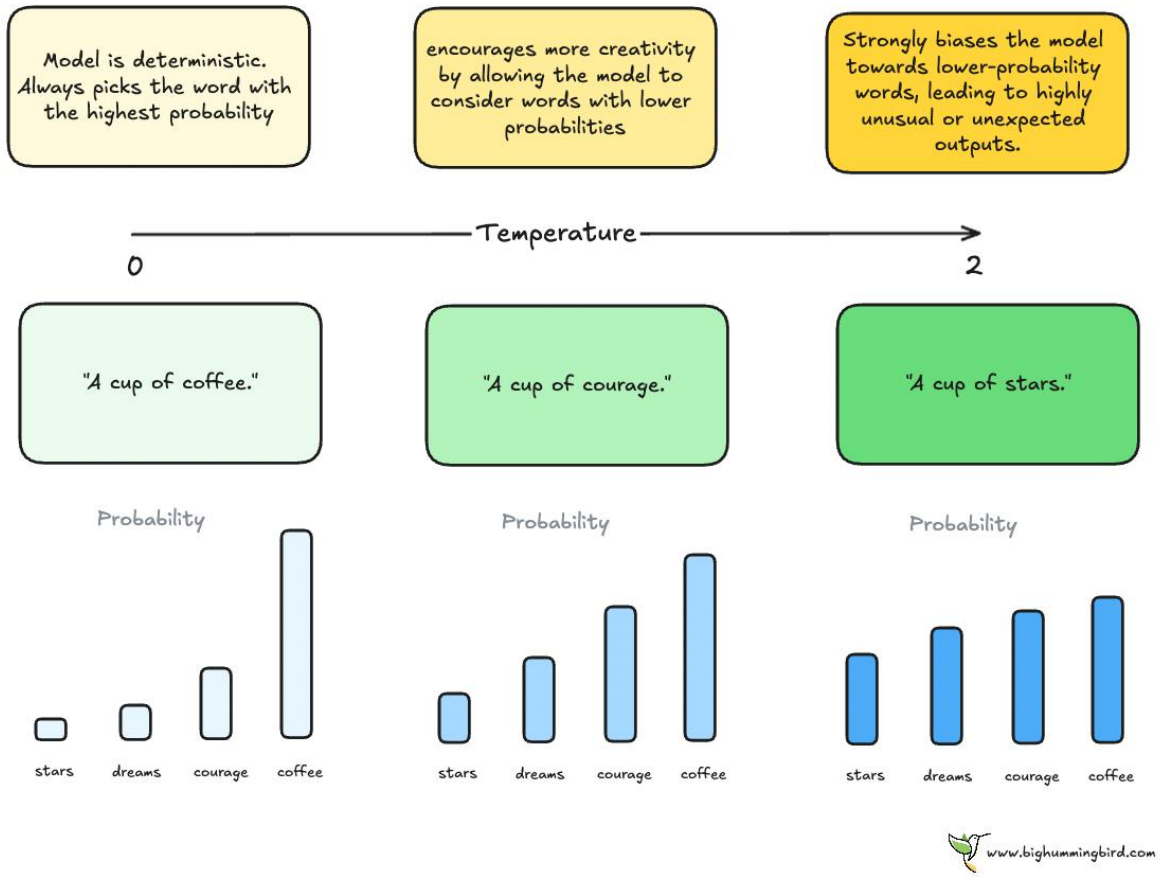
https://www.bighummingbird.com/blogs/llm-hyperparameter

Top-P Sampling



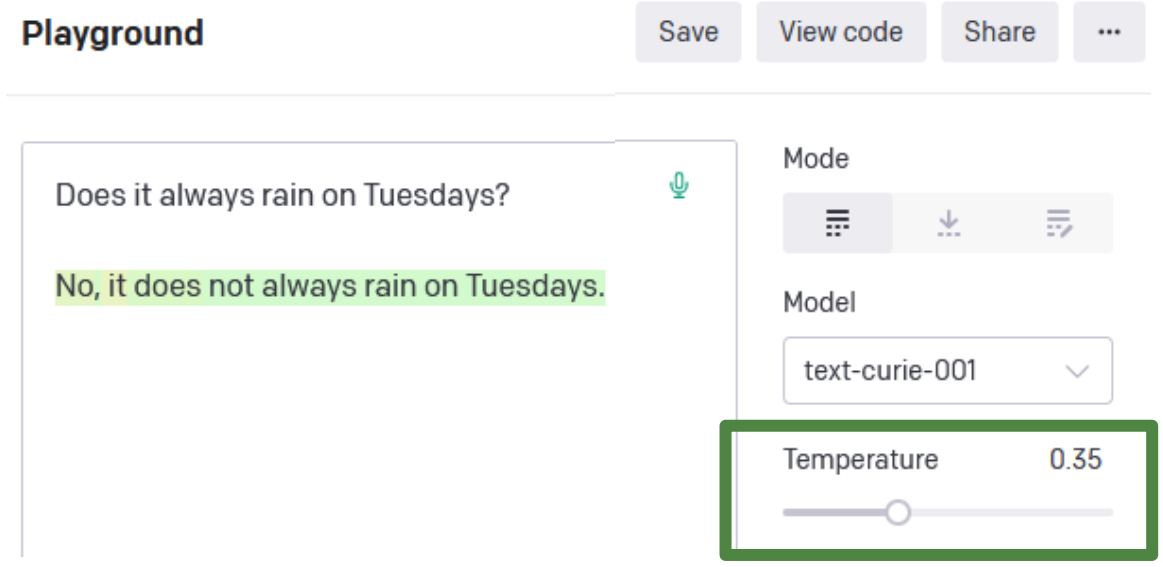
https://huggingface.co/blog/how-to-generate

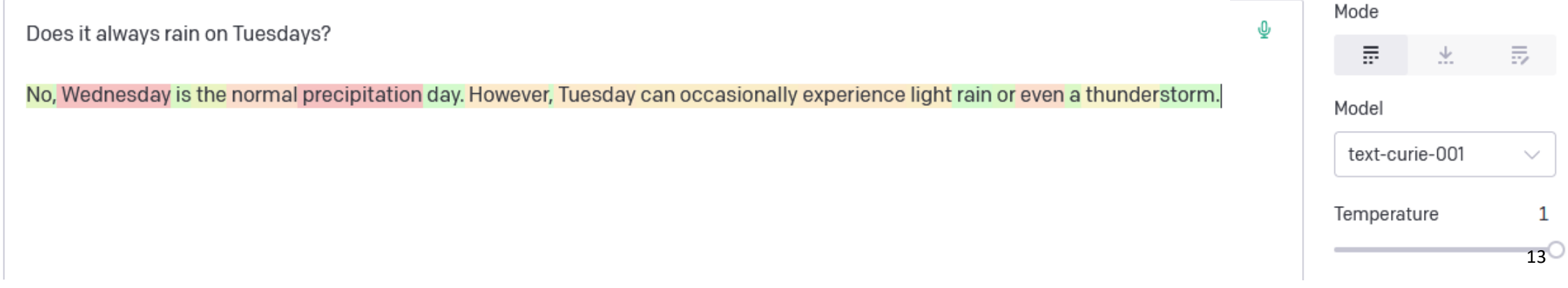
“Temperature”



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

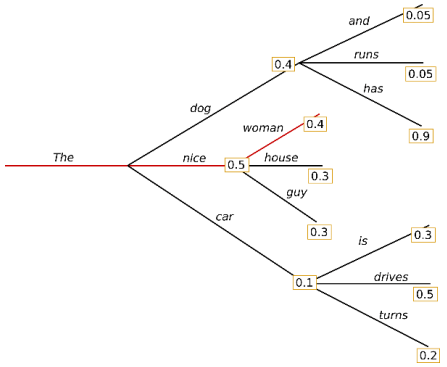
Temperature in Action



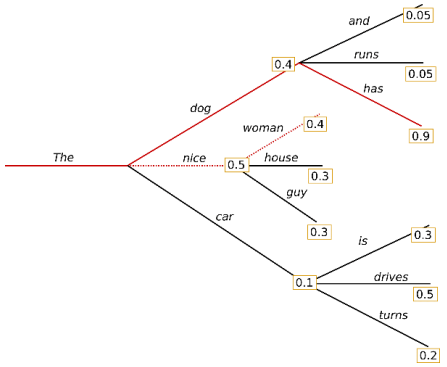


Think-Pair-Share

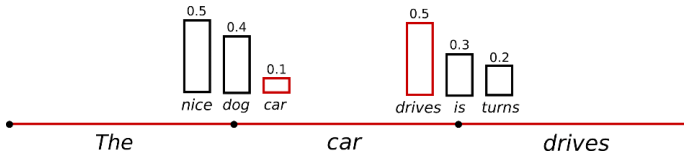
When (what types of scenarios or tasks) might you want to use one sampling algorithm over the other?



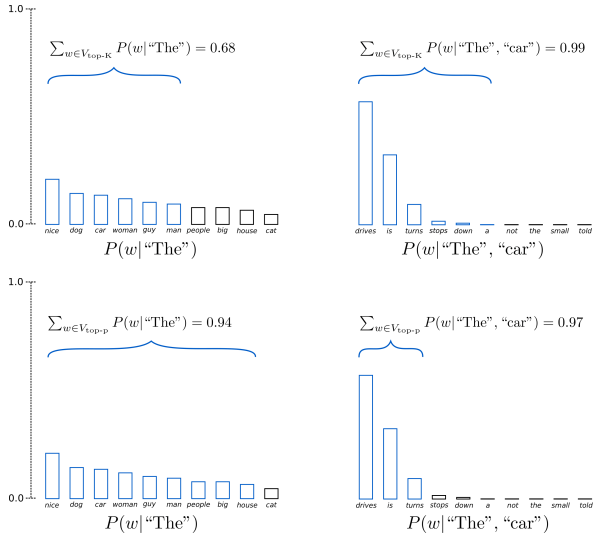
Greedy



Beam Search



Random Sampling



Top-K/P

Fine-tuning

Start with pre-trained model

Freeze the model (don't touch it) except for the last layer

- Sometimes you can adjust the weights of the whole model instead of just the last layer
- Start with generalized "foundation" model
- Train on a new, small dataset for your specific task

GPT-2

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Abstract

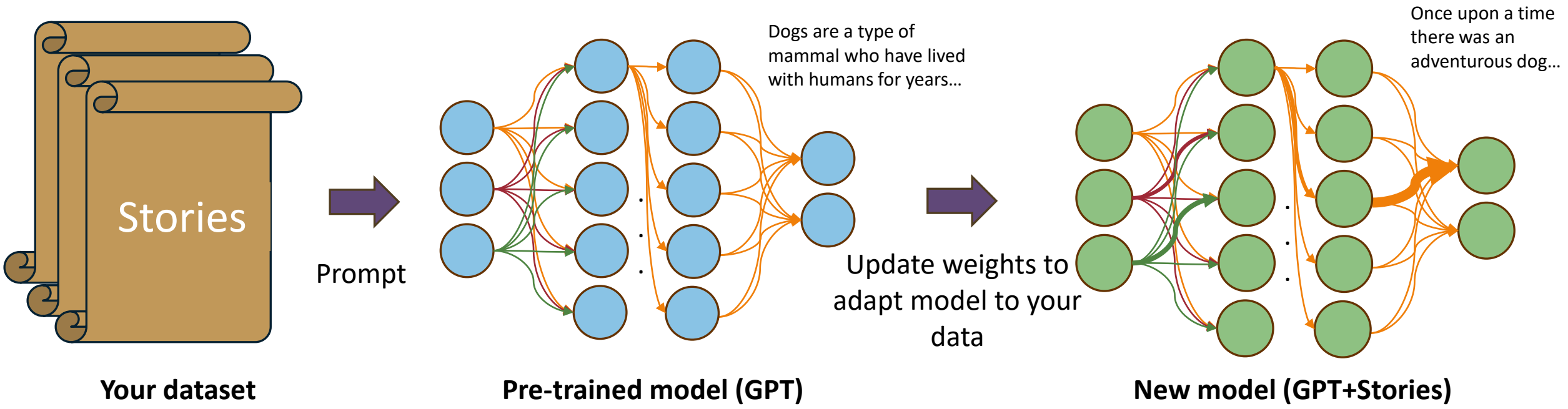
Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

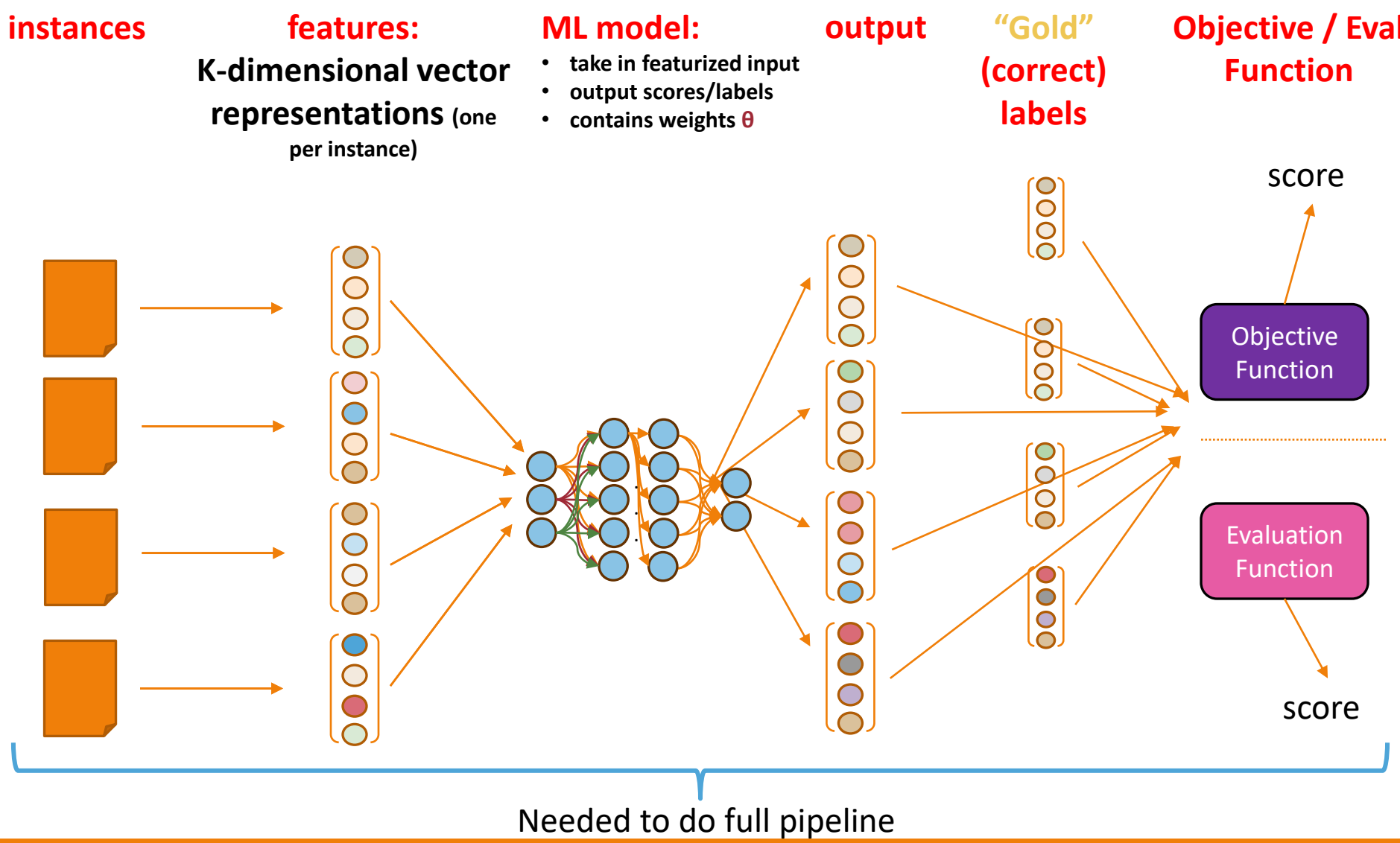
The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017), and image classifiers (Alcorn et al., 2018) on the diversity and variety of possible inputs highlights some of the shortcomings of this approach.

Our suspicion is that the prevalence of single task training on single domain datasets is a major contributor to the lack of generalization observed in current systems. Progress towards robust systems with current architectures is likely to require training and measuring performance on a wide range of domains and tasks. Recently, several benchmarks

Finetuning

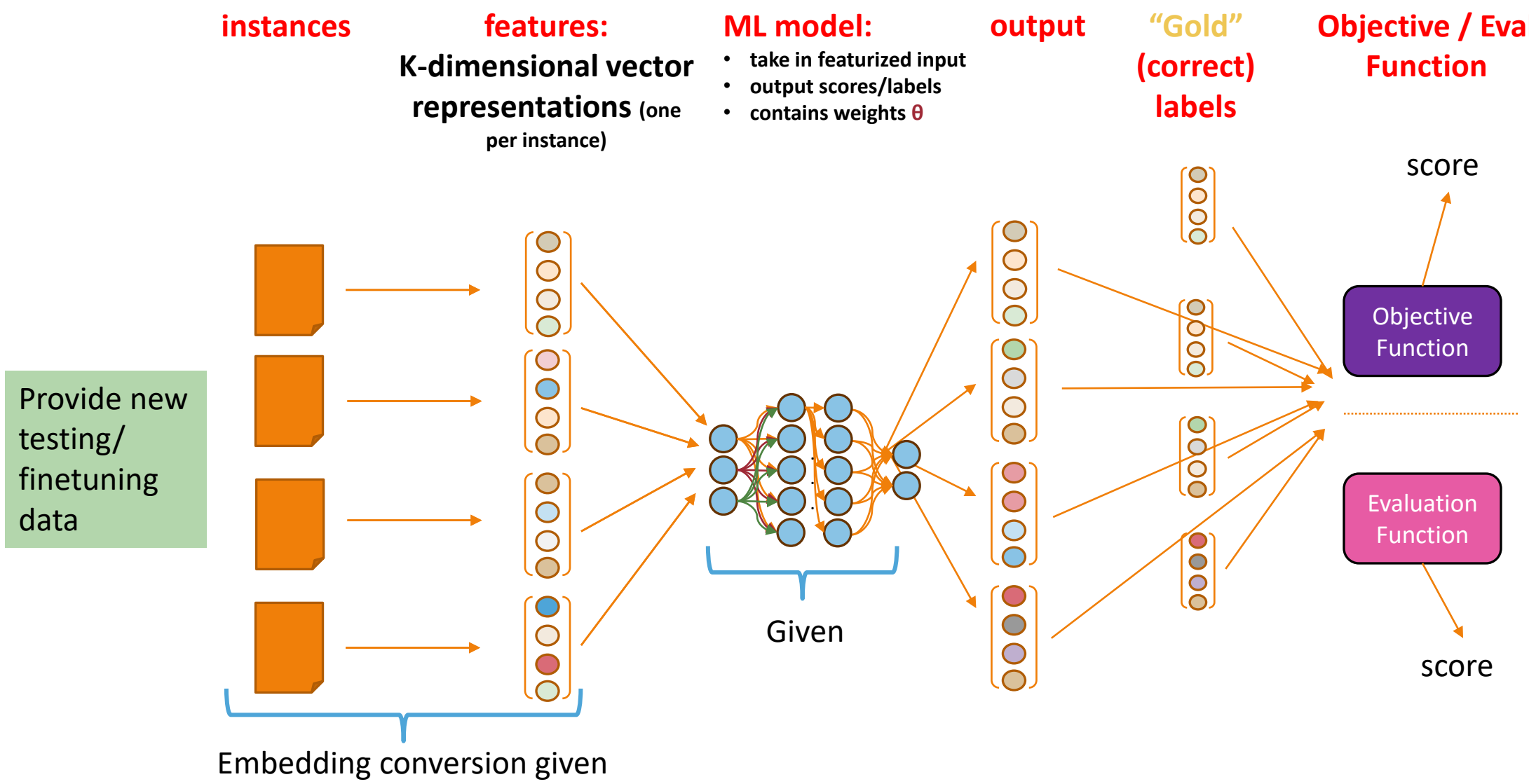

Stories
Prompt
Dogs are a type of mammal who have lived with humans for years…
Update weights to adapt model to your data
Once upon a time there was an adventurous dog…
Your dataset
Pre-trained model (GPT)
New model (GPT+Stories)

Pre-transformer Neural NLP



Slide adapted from Dr. Frank Ferraro

Transformer-based NLP



Slide adapted from Dr. Frank Ferraro

What types of things can go wrong with finetuning?

Underfitting – finetuning data is too different from what the foundational model was train on → model can't learn it

Overfitting – overwrites what the model learned originally

Pre-trained models

Most LLMs people use today are pre-trained models

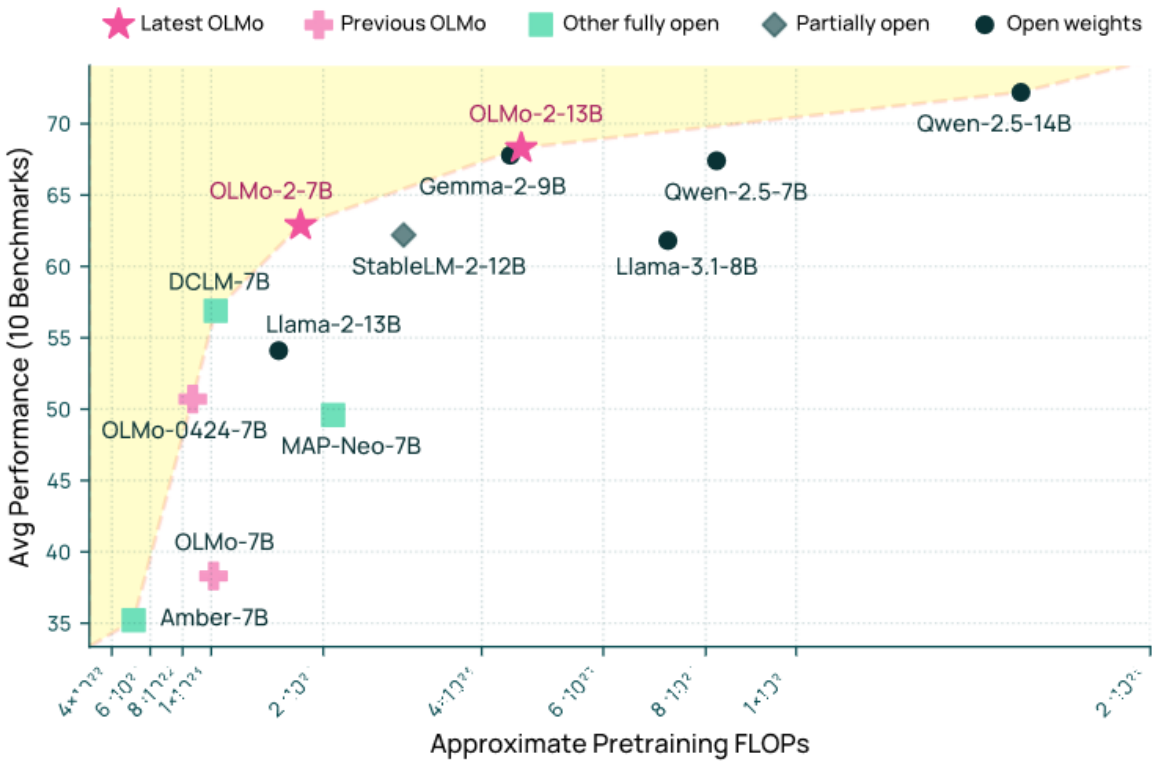
Trained on “the Internet” → Impossible to know all of what it’s train on

- Very few models release all the data. One example is OLMo 2.

Can then be finetuned on specific data

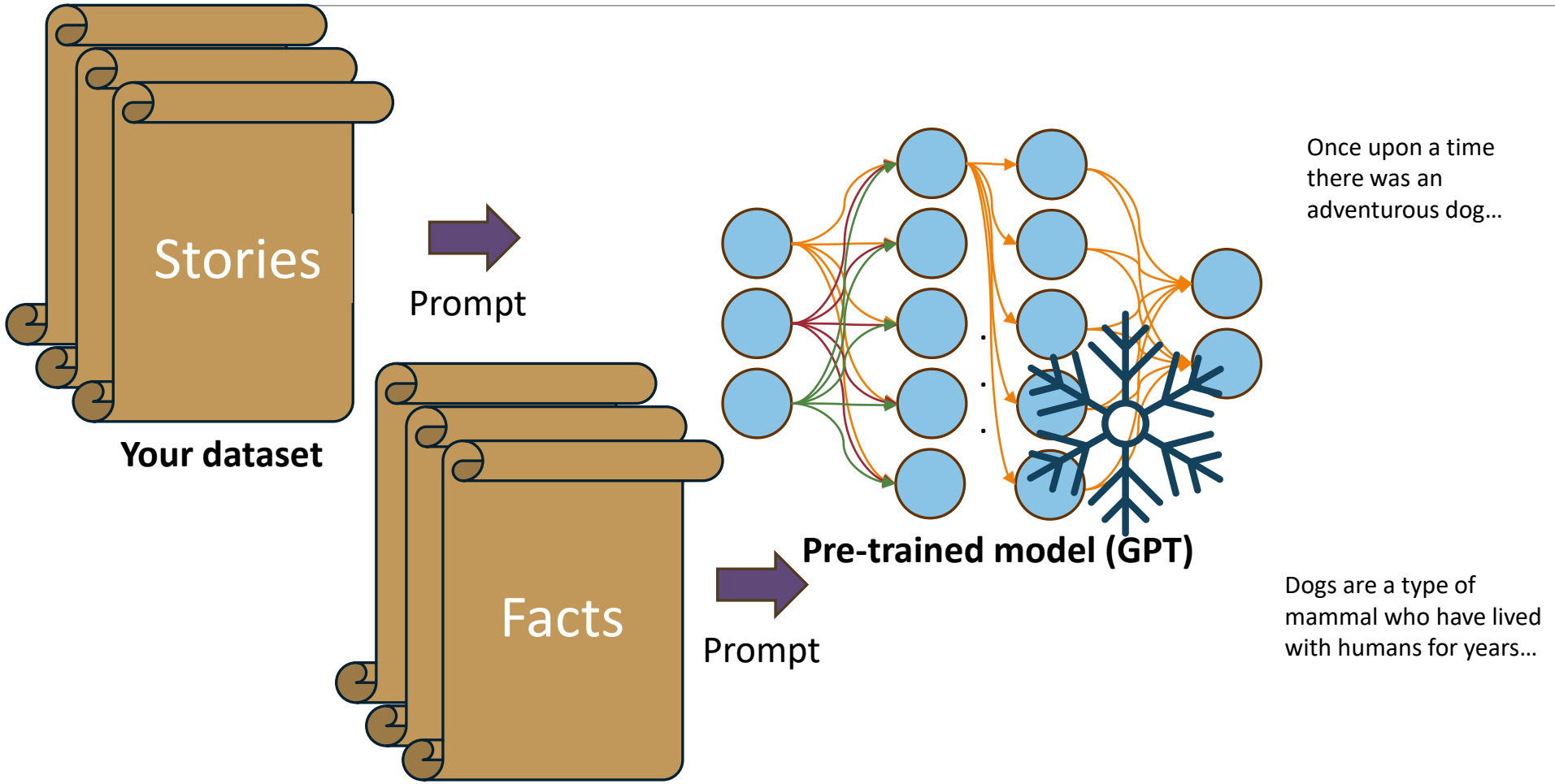
Why would you want to “tweak” an existing model?

Open-Sourced Models



OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., … Hajishirzi, H. (2024). *2 OLMo 2 Furious* (No. 2501.00656). arXiv. https://doi.org/10.48550/arXiv.2501.00656

Prompting

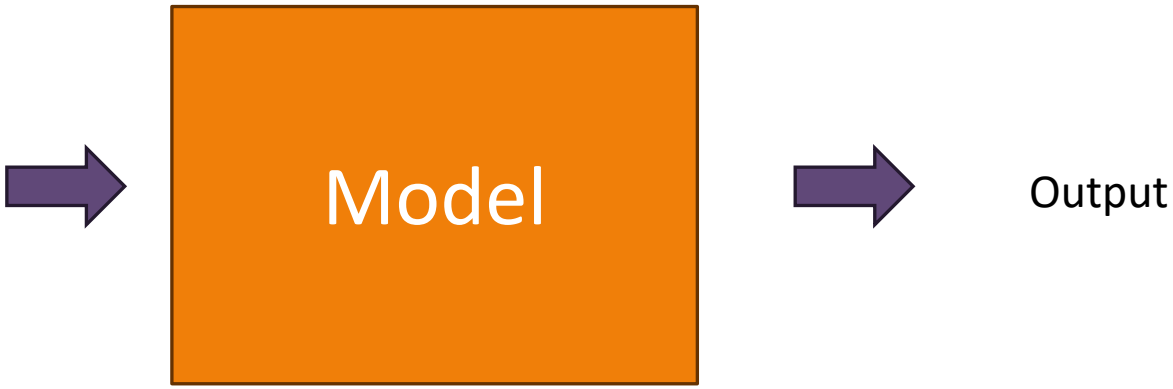

Stories
Your dataset
Prompt
Facts
Prompt
Pre-trained model (GPT)
Once upon a time there was an adventurous dog…
Dogs are a type of mammal who have lived with humans for years…

Zero-shot Prompting

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

Instructions

Task



Output

Sentence:
The dog ate the giant fish.

Few-shot Prompting

Instructions

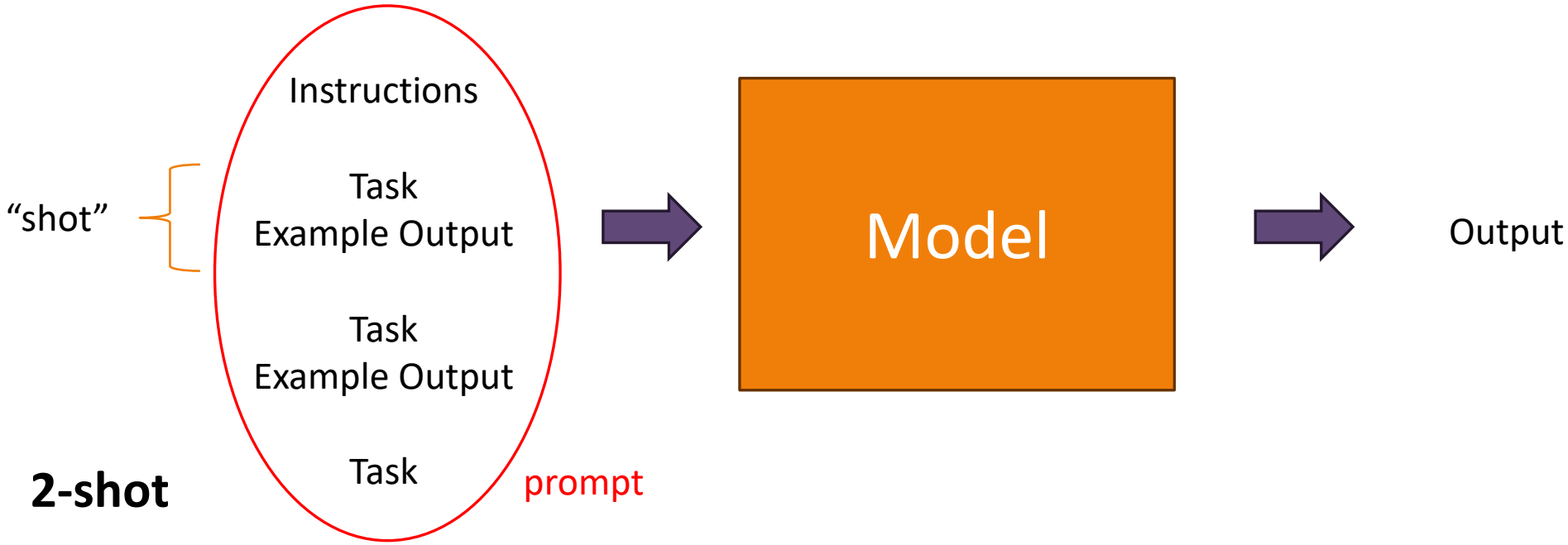
You are a helpful assistant.
You will be tagging the parts of speech in sentences.

Task

Sentence:
The dog ate the giant fish.

Example Output

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



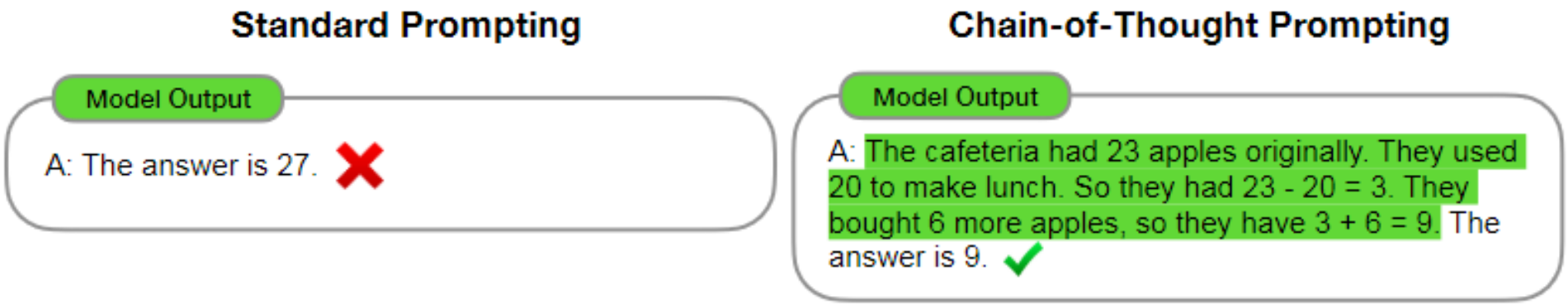
Prompt Engineering

"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?



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.