Foundation Models and Embeddings (continued) + Scripts and Events

Lara J. Martin (she/they)

https://laramartin.net/neurosymbolic-text-gen/

Learning Objectives

Calculate the distance between vector embeddings

Differentiate between encoder model embeddings and older dense embeddings

Differentiate between a script and causal chains

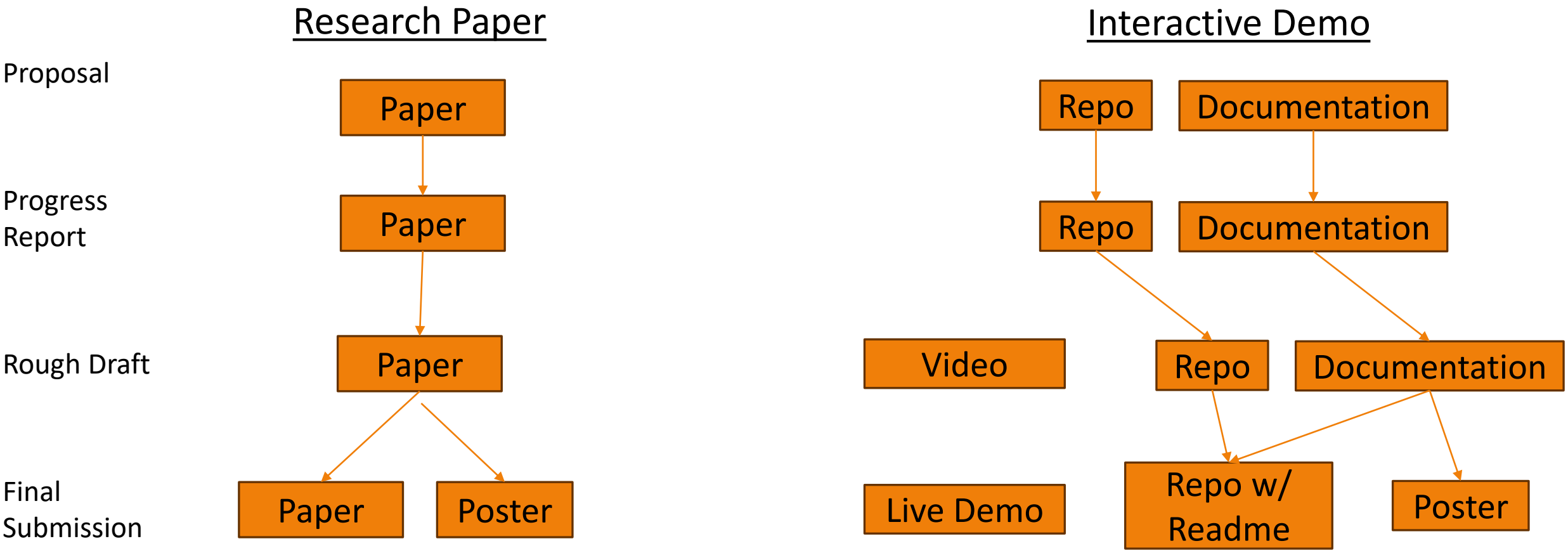
Create your own script

Speculate how scripts can be used in story generation

Distinguish between causal and probabilistic events

Identify the strengths and limitations of scripts

Project Flow - Deliverables



Review: What is a foundation model?

A model that captures “foundational” 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 pretrained large language models (LLMs) are foundation models

Review: Vector Embeddings: Key Ideas

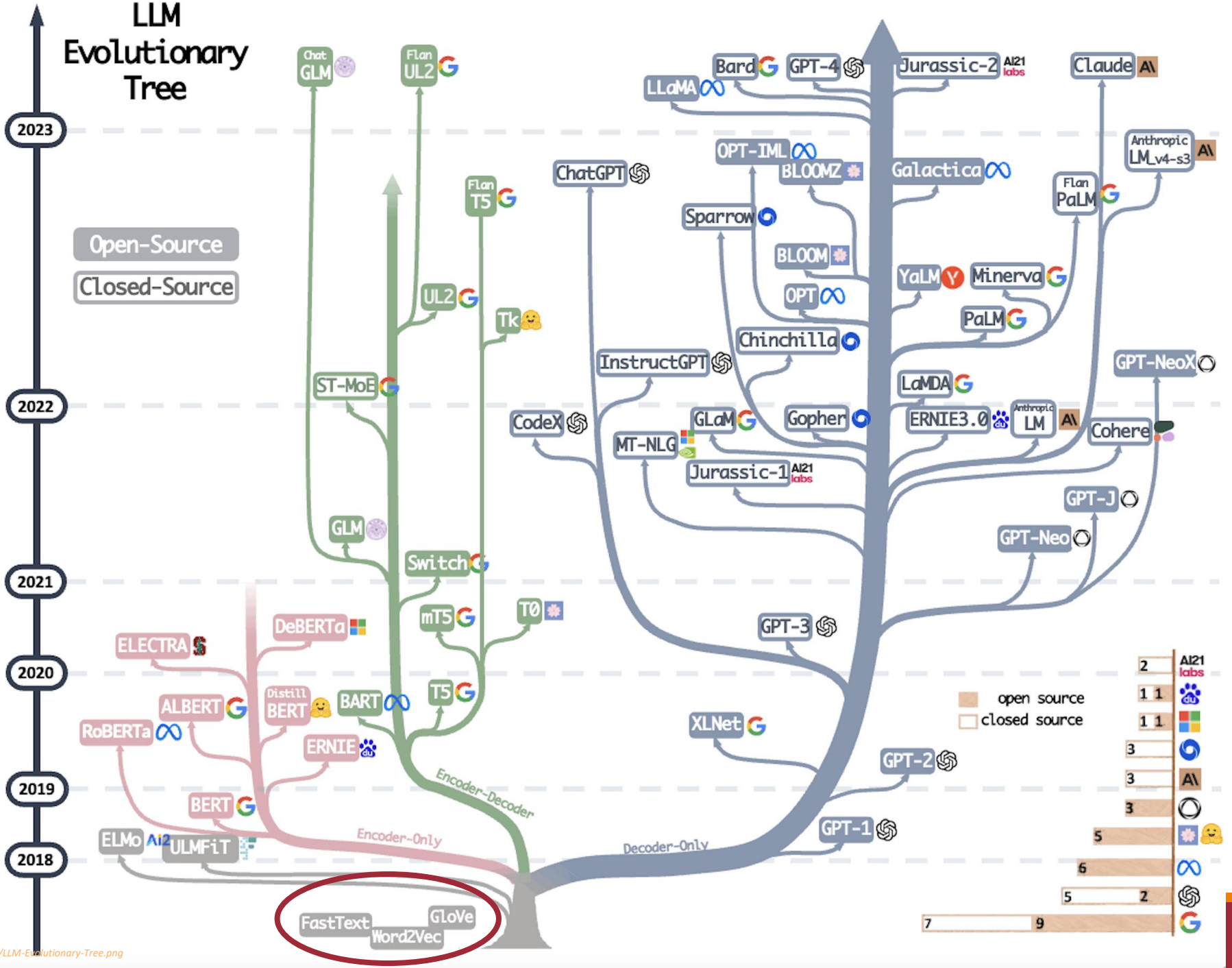
Vector embeddings can be used for phrases, paragraphs, or even whole documents!

1. Acquire basic contextual statistics (often counts) for each word type v

2. Extract a real-valued vector e_v for each word v from those statistics

[0.00315225, 0.00315225, 0.00547597, 0.00741556, 0.00912817, 0.01068435, 0.01212381, 0.01347162, 0.01474487, 0.0159558]

3. Use the vectors to represent each word in later tasks



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

Word2Vec

Mikolov et al. (2013; NeurIPS): “Distributed Representations of Words and Phrases and their Compositionality”

Revisits the context-word approach

Learn a model p(c | w) to predict a context word from a target word

Word2Vec

Mikolov et al. (2013; NeurIPS): “Distributed Representations of Words and Phrases and their Compositionality”

Revisits the context-word approach

Learn a model p(c | w) to predict a context word from a target word

Learn two types of vector representations

- $h_c \in \mathbb{R}^E$: vector embeddings for each context word
- $v_w \in \mathbb{R}^E$: vector embeddings for each target word

$$p(c \mid w) \propto \exp(h_c^T v_w)$$

Word2Vec

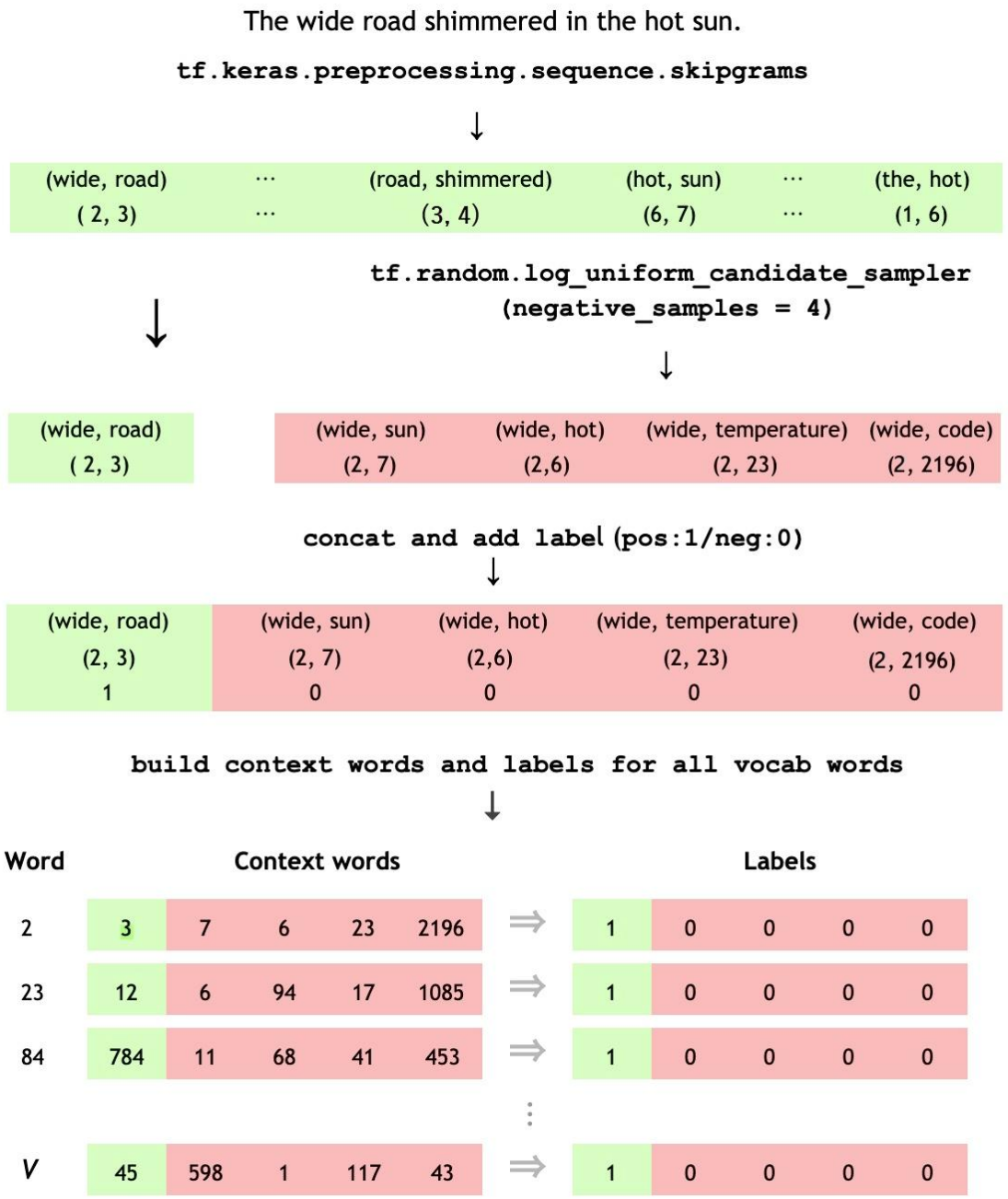
context (↓)-word (→) count matrix

	apricot	pineapple	digital	information
aardvark	0	0	0	0
computer	0	0	2	1
data	0	10	1	6
pinch	1	1	0	0
result	0	0	1	4
sugar	1	1	0	0

Context: those other words within a small "window" of a target word

$$\max_{h,v} \sum_{c,w \text{ pairs}} \text{count}(c,w) \log p(c \mid w)$$

$$p(c \mid w) \propto \exp(h_c^T v_w)$$



- Positive example from training data
- Several negative examples (pairing two random words in the vocabulary)

https://www.tensorflow.org/text/tutorials/word2vec

FastText

"Enriching Word Vectors with Subword Information" Bojanowski et al. (2017; TACL)

Main idea: learn **character n-gram embeddings** for the target word (not context) and modify the word2vec model to use these

Pre-trained models in 150+ languages

- https://fasttext.cc

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "**Enriching Word Vectors with Subword Information**," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

FastText Details

Main idea: learn **character n-gram embeddings** and for the target word (not the context) modify the word2vec model to use these

Original word2vec:

$$p(c\,|w) \propto \exp(h_c^T v_w)$$

FastText:

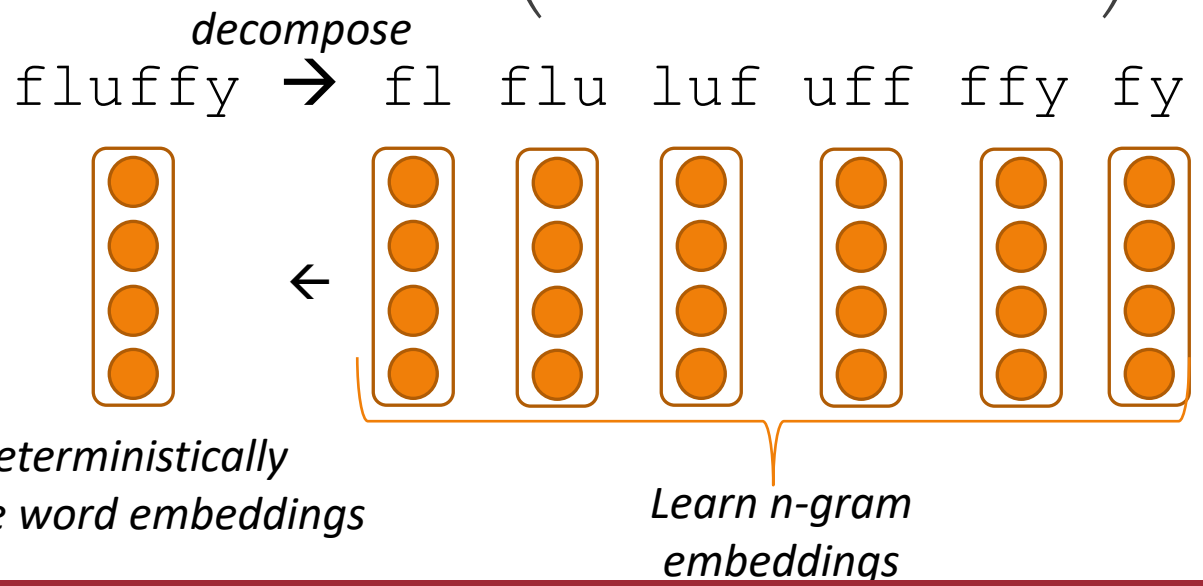
$$p(c\,|w) \propto \exp\left(h_c^T \left(\sum_{\text{n-gram } g \text{ in } w} z_g\right)\right)$$

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "**Enriching Word Vectors with Subword Information**," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

FastText Details

Main idea: learn **character n-gram embeddings** and for the target word (not the context) modify the word2vec model to use these

$$p(c \mid w) \propto \exp\left(h_c^T \left(\sum_{\text{n-gram } g \text{ in } w} z_g\right)\right)$$

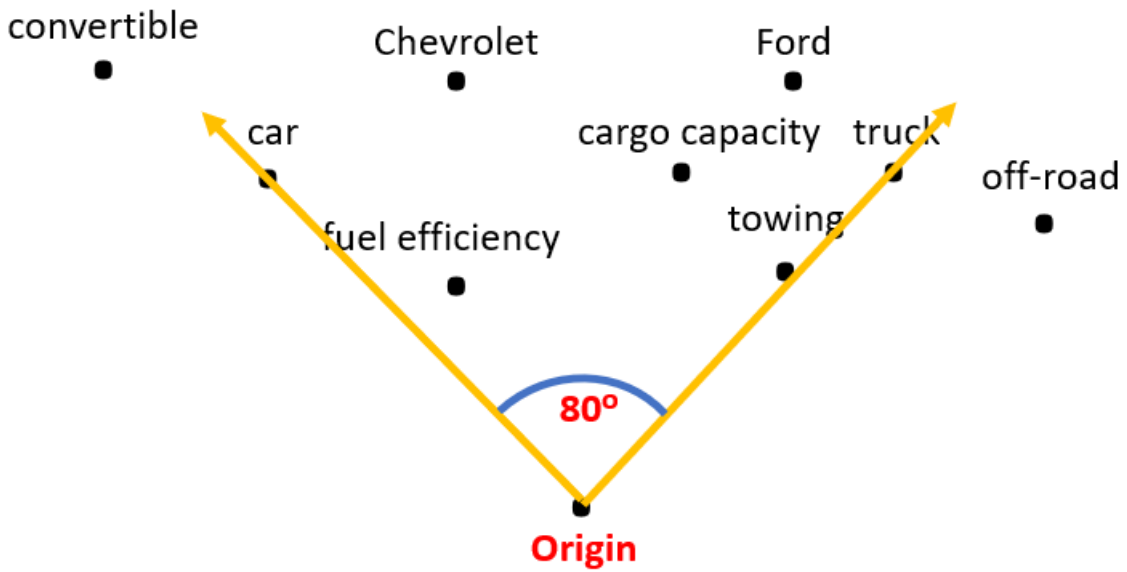


P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "**Enriching Word Vectors with Subword Information**," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

Comparing/Evaluating Word Embeddings

Cosine: Measuring Similarity

Given 2 target words v and w how similar are their vectors?



https://upload.wikimedia.org/wikipedia/commons/2/23/CosineSimilarity.png

Cosine: Measuring Similarity

Given 2 target words v and w how similar are their vectors?

Dot product or inner product from linear algebra

$$\text{dot-product}(\vec{v},\vec{w}) = \vec{v}\cdot\vec{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

- High when two vectors have large values in same dimensions, low for orthogonal vectors with zeros in complementary distribution

Correct for high magnitude vectors $\dfrac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}$

Cosine Similarity

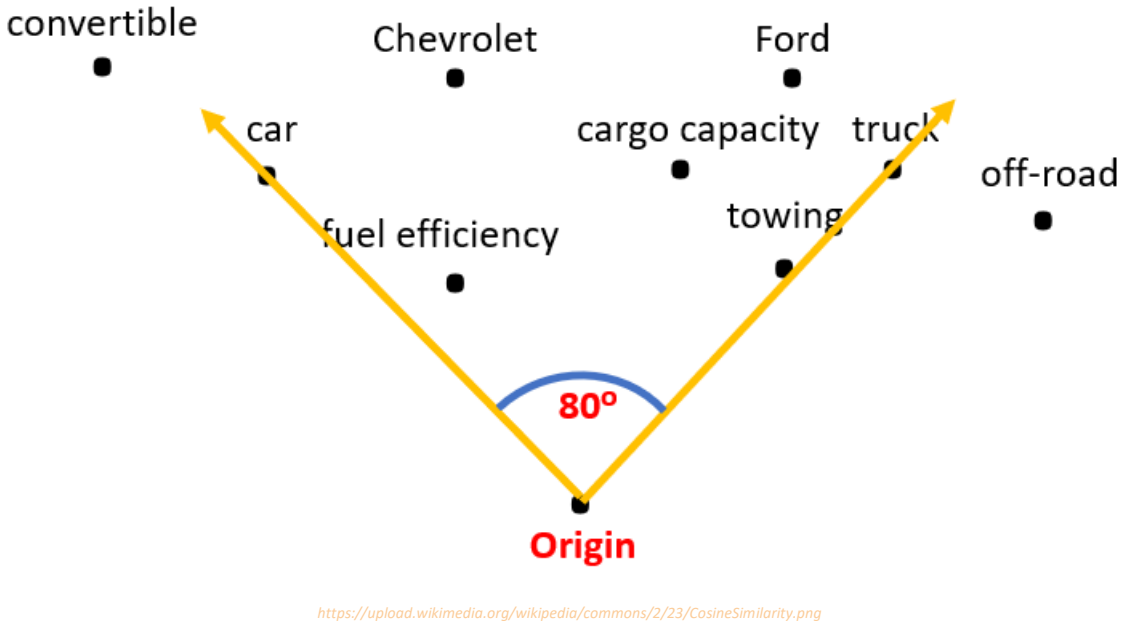
Divide the dot product by the length of the two vectors

$$\frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

This is the cosine of the angle between them

$$\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos\theta$$

$$\frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} = \cos\theta$$



https://upload.wikimedia.org/wikipedia/commons/2/23/CosineSimilarity.png

Example: Word Similarity

$$\cos(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

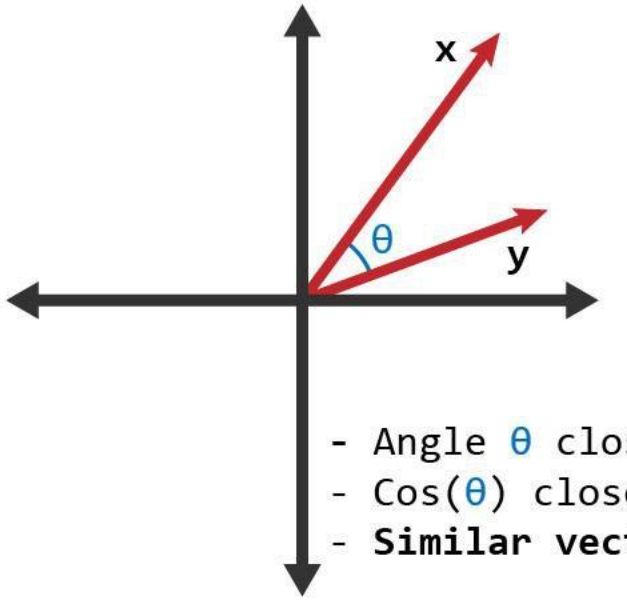
	Dim. 1	Dim. 2	Dim. 3
apricot	2	0	0
digital	0	1	2
information	1	6	1

cosine(apricot,information) = $\frac{2+0+0}{\sqrt{4+0+0}\sqrt{1+36+1}} = 0.1622$

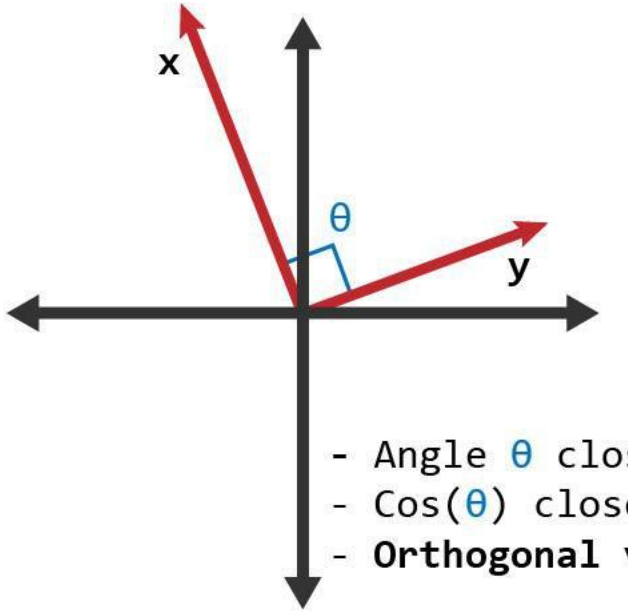
cosine(digital,information) = $\frac{0+6+2}{\sqrt{0+1+4}\sqrt{1+36+1}} = 0.5804$

cosine(apricot,digital) = $\frac{0+0+0}{\sqrt{4+0+0}\sqrt{0+1+4}} = 0.0$

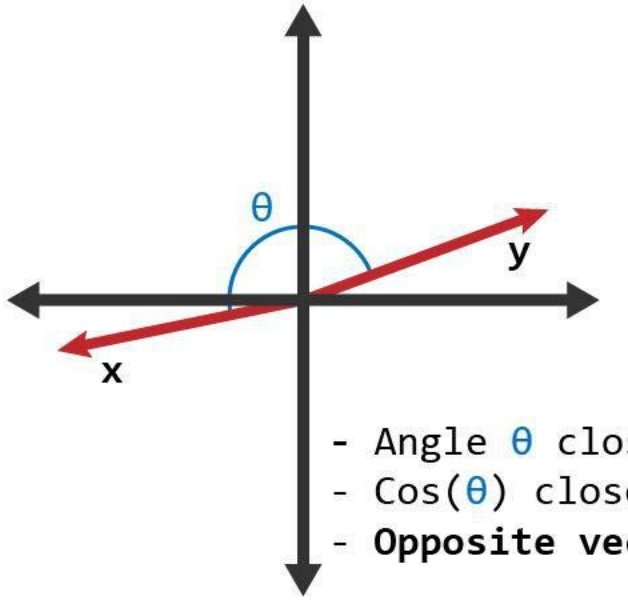
Cosine Similarity Range



- Angle θ close to 0
- Cos(θ) close to 1
- **Similar vectors**



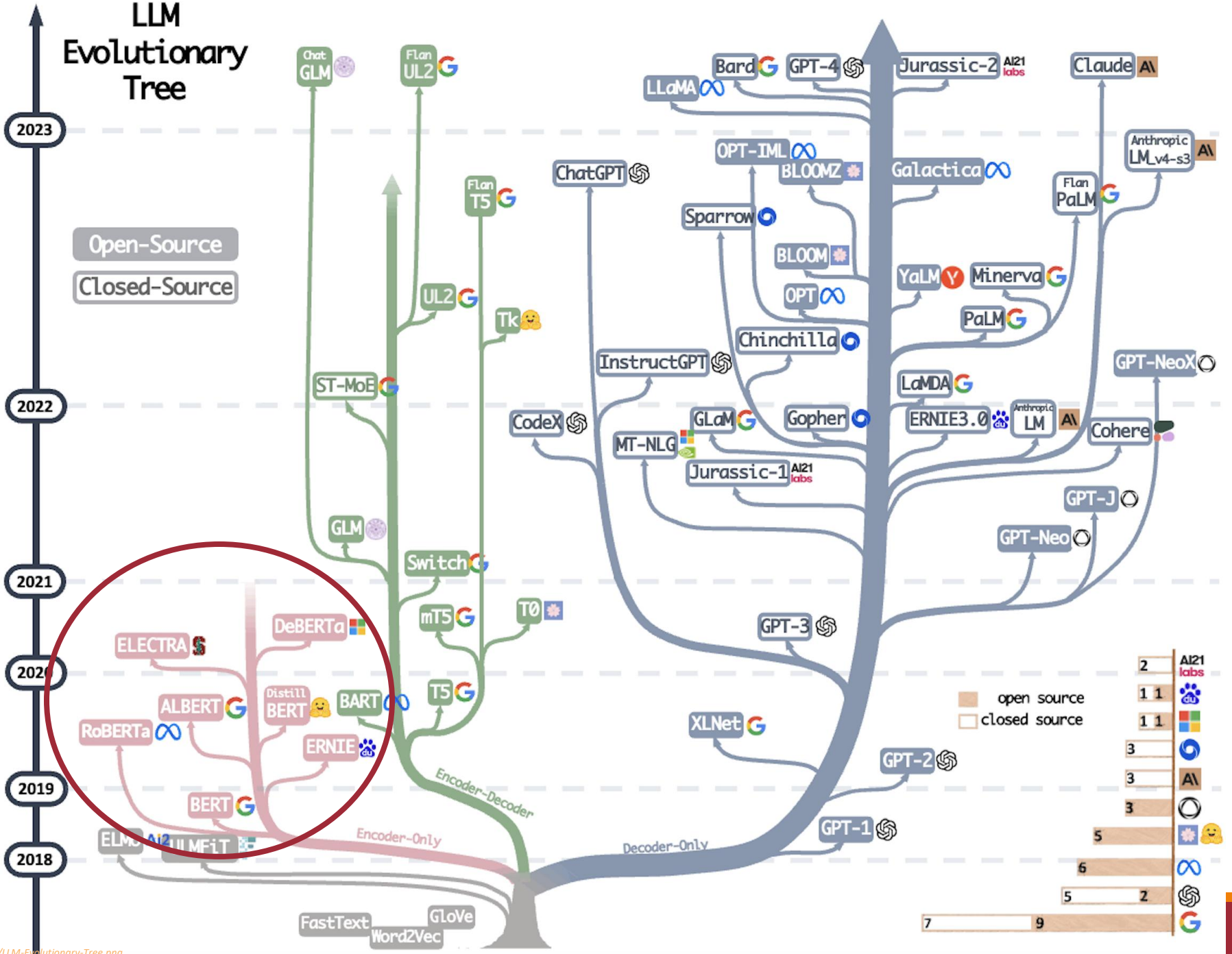
- Angle θ close to 90
- Cos(θ) close to 0
- **Orthogonal vectors**



- Angle θ close to 180
- Cos(θ) close to -1
- **Opposite vectors**

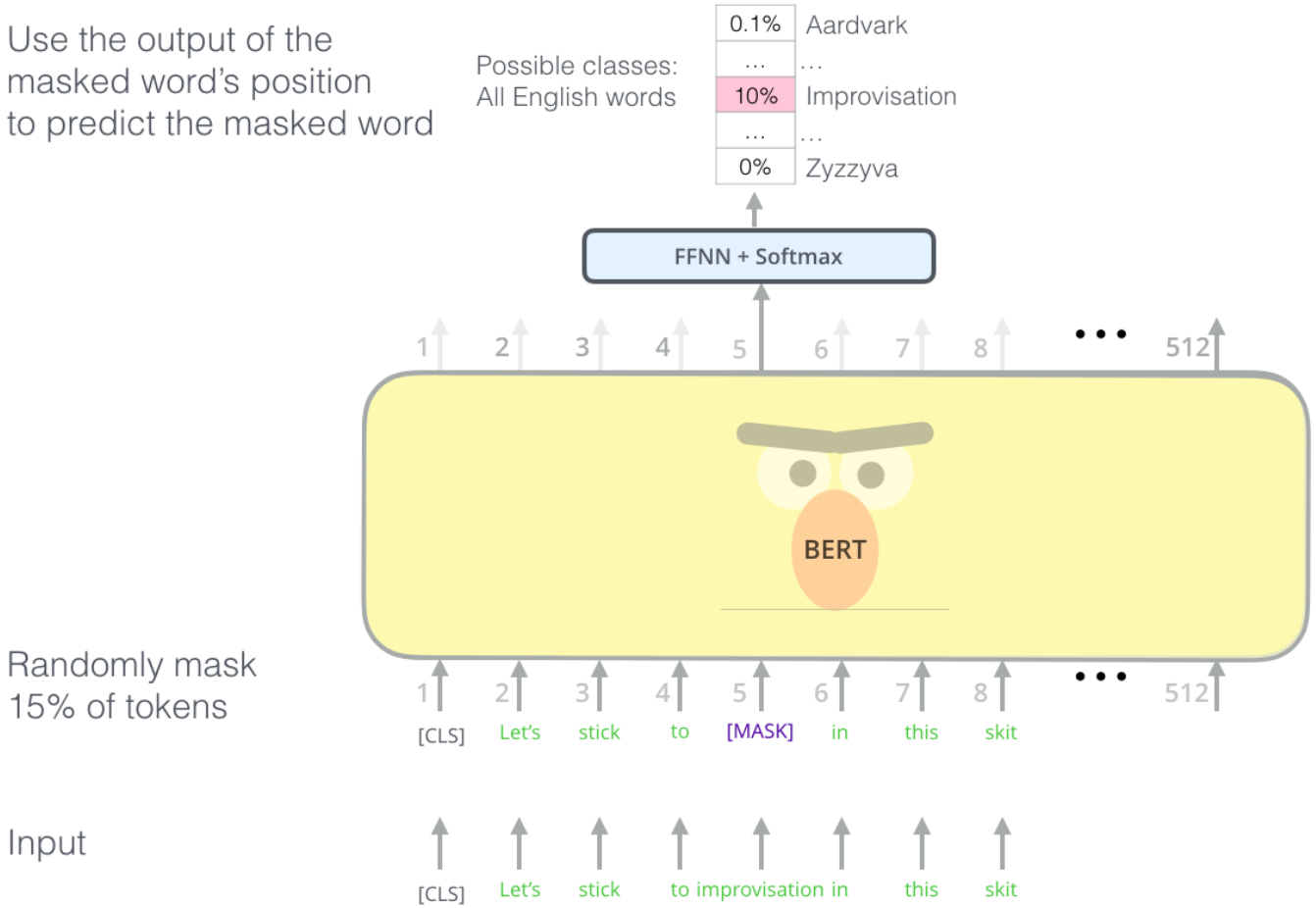
https://www.learndatasci.com/glossary/cosine-similarity/

Back to Transformers…



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

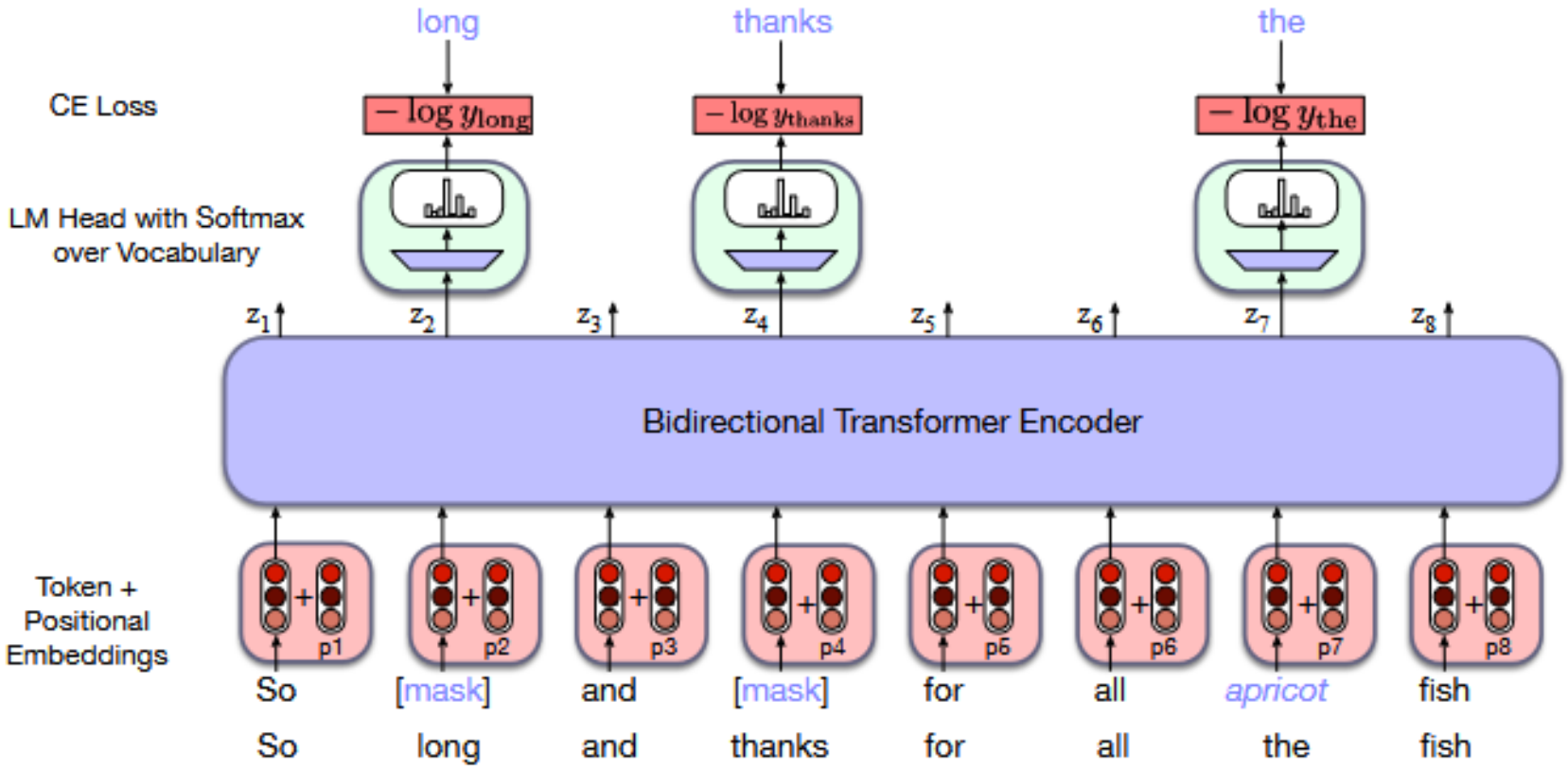
BERT (Devlin et al. 2019)



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

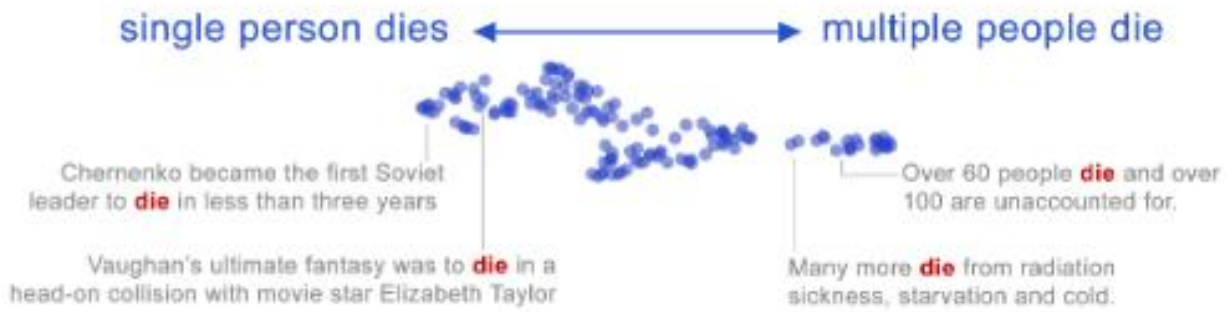
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers)*,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

Masked Language Models



From Jurafsky & Martin's Speech and Language Processing, 3rd Edition, Chapter 11

Contextual Embeddings



From Jurafsky & Martin's Speech and Language Processing, 3rd Edition, Chapter 11

BERT Question

Consider the highlighted words. Which two words would contextual word embeddings from BERT say are closest?

A. I am so excited to use my new **bat** at the baseball game tomorrow.

B. The favorite food of this species of **bat** is mosquitoes.

C. The **cardinal** isn't just a lawn decoration; the species makes themselves useful by eating mosquitoes.

PollEv.com/laramartin527

Word2Vec Question

Remember: word2vec is a dense vector embedding

Consider the highlighted words. Which two words would word2vec say are closest?

A. I am so excited to use my new **bat** at the baseball game tomorrow.

B. The favorite food of this species of **bat** is mosquitoes.

C. The **cardinal** isn't just a lawn decoration; the species makes themselves useful by eating mosquitoes.

PollEv.com/laramartin527

Encoder-Only Models

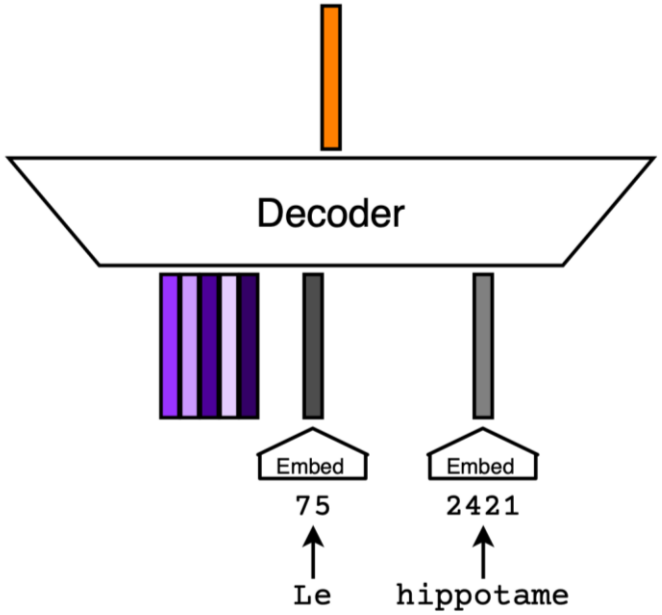
- Autoencoding
- Encoder-only
  - Input: Corrupted version of text sequence
  - Goal: Produce an uncorrupted version of text sequence
- When to use:
  - Classification tasks
  - Sentence embeddings
  - Context-dependent word embeddings
  - Any type of fill-in-the-blank tasks

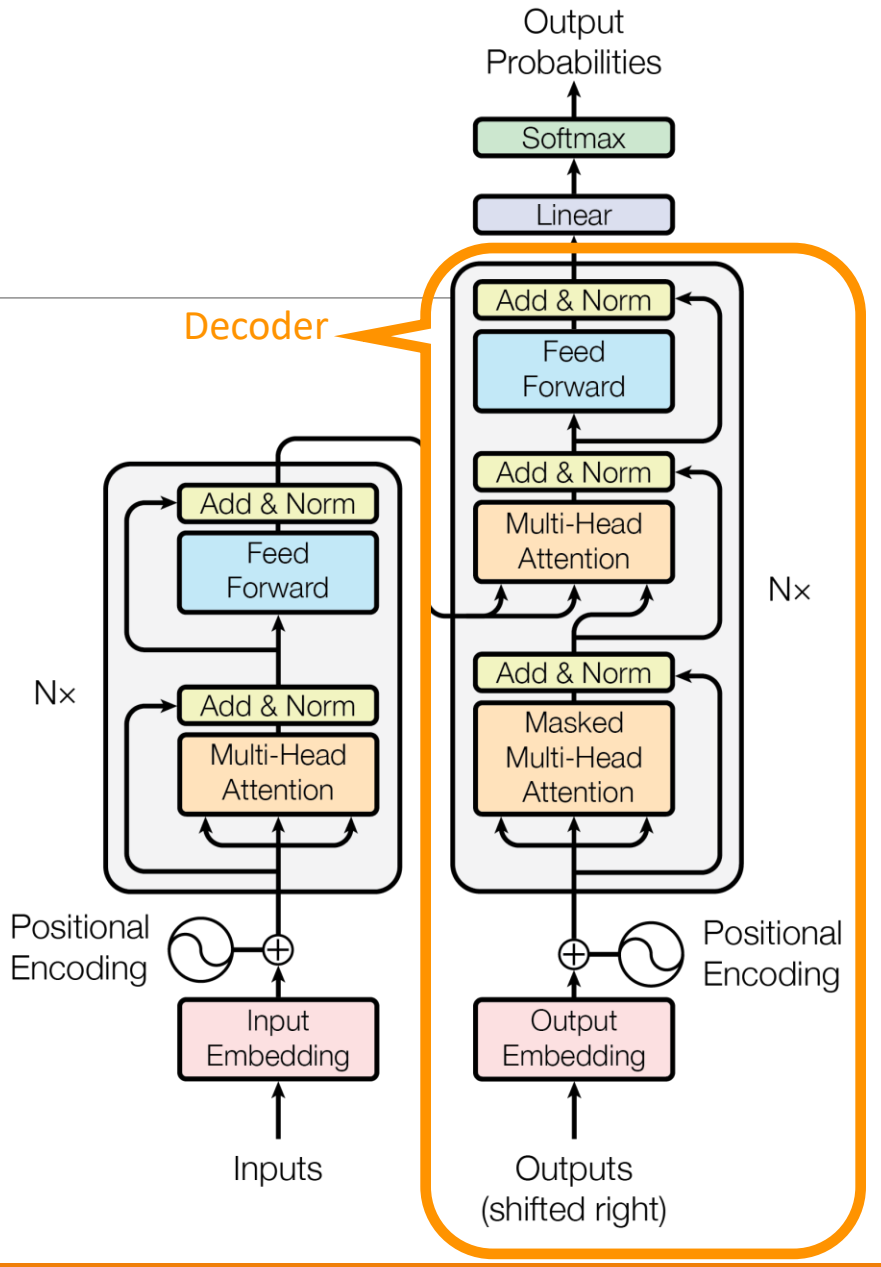
Some important BERT family members

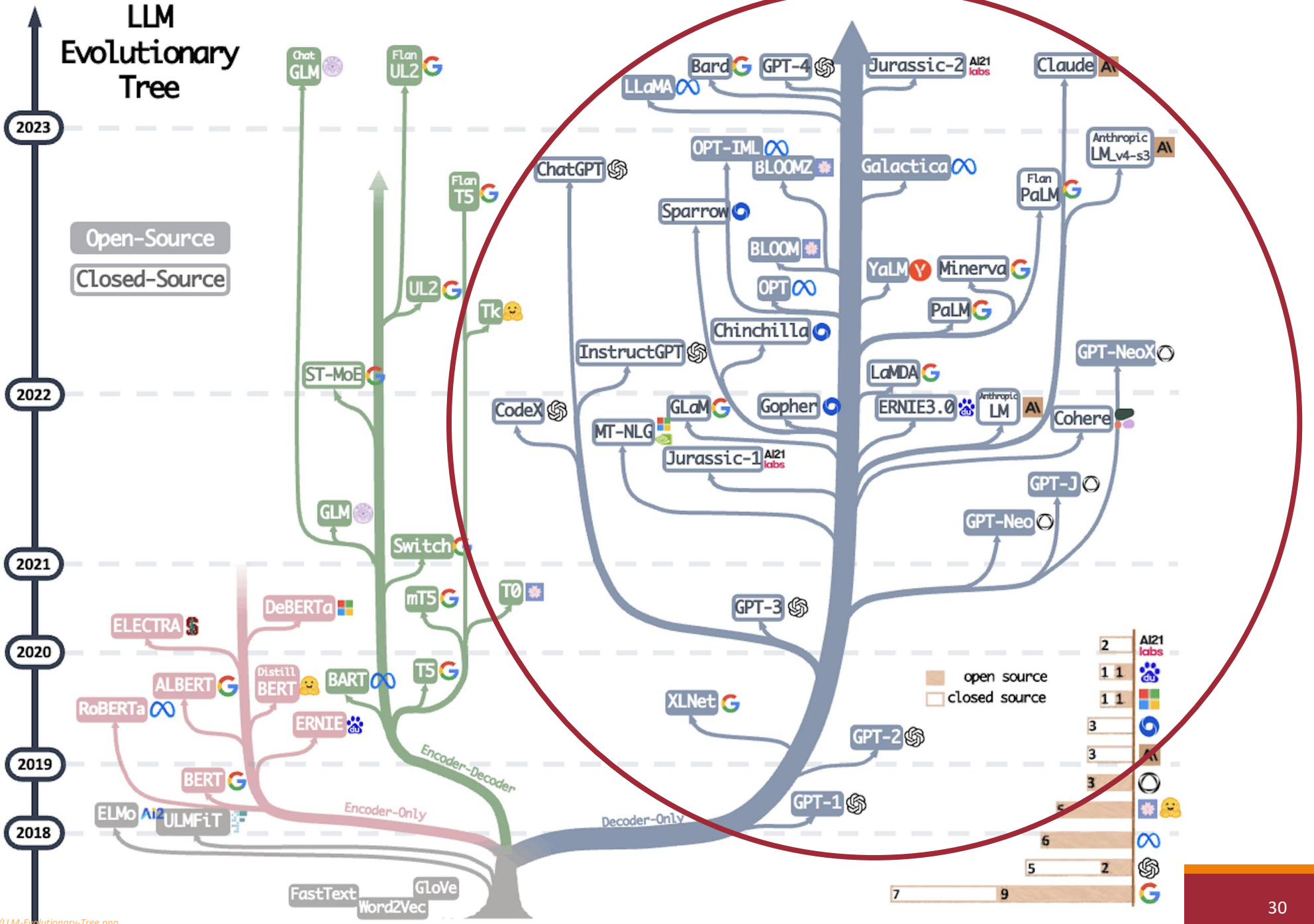
(in my opinion)

- 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

Decoder-Only Models

- Autoregressive
- Decoder-only
  - Input: Text sequence
  - Goal: Predict the next word given the previous ones
- When to use:
  - Most generation tasks

Decoder-Only Models

GPT (OpenAI)

LLaMA (Meta)

Gemma/Gemini (Google DeepMind) – Gemma is the open version of Gemini

Qwen (Alibaba Cloud)

Mistral (Mistral AI)

Nemotron (Nvidia)

Extensions of Foundation Models (Usually Decoder-Only)

Instruction-tuned Models (e.g., Instruct-GPT, FLAN-T5, Claude, most consumer-facing models)

- Finetuned to follow instructions

Reasoning Models (e.g., DeepSeek-R1, OpenAI o1)

- Finetuned to do CoT and/or self-criticism

Distilled Models (e.g., DistilQwen2.5, DeepSeek-R1-Distill)

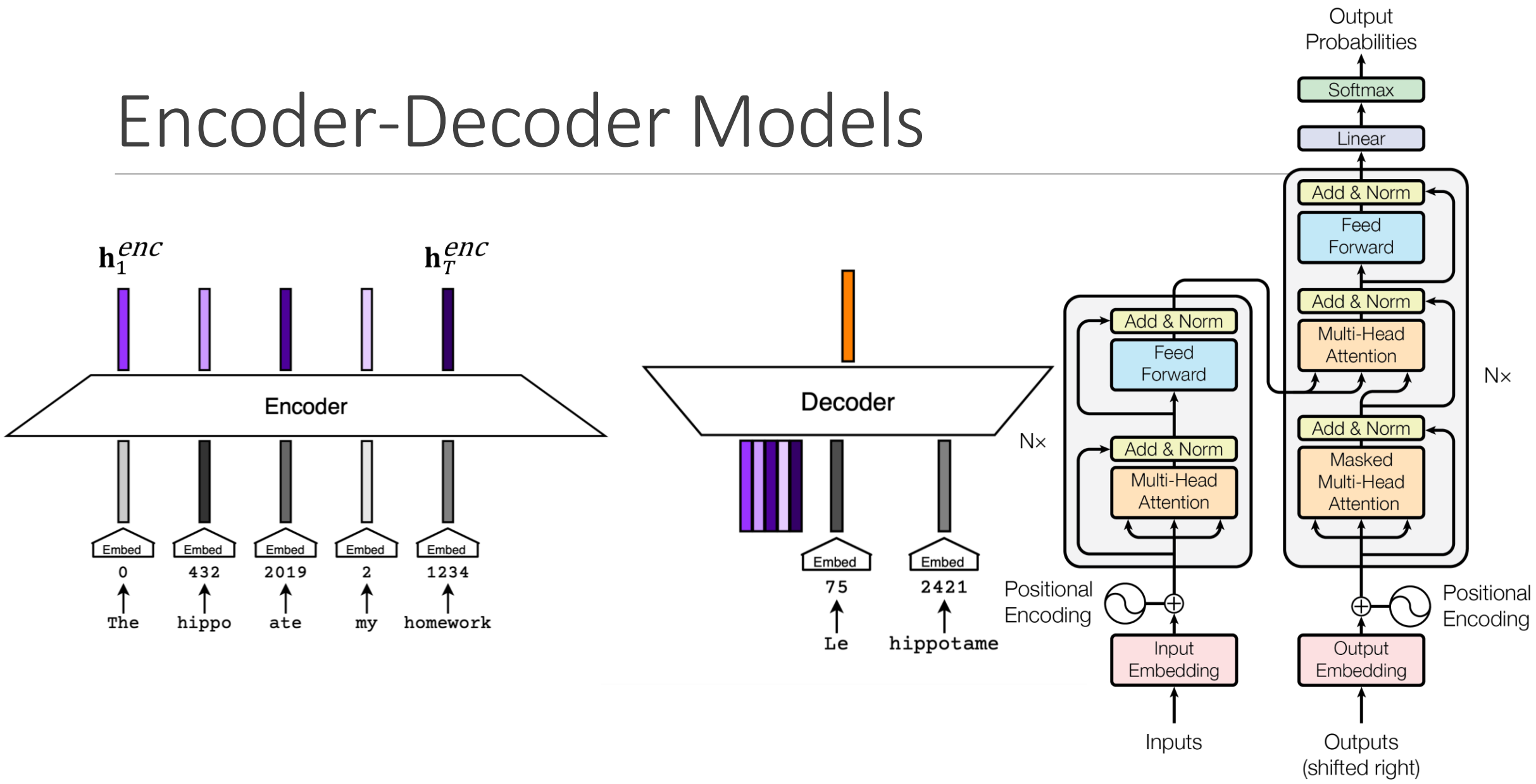
- Teacher-student learning to smaller model

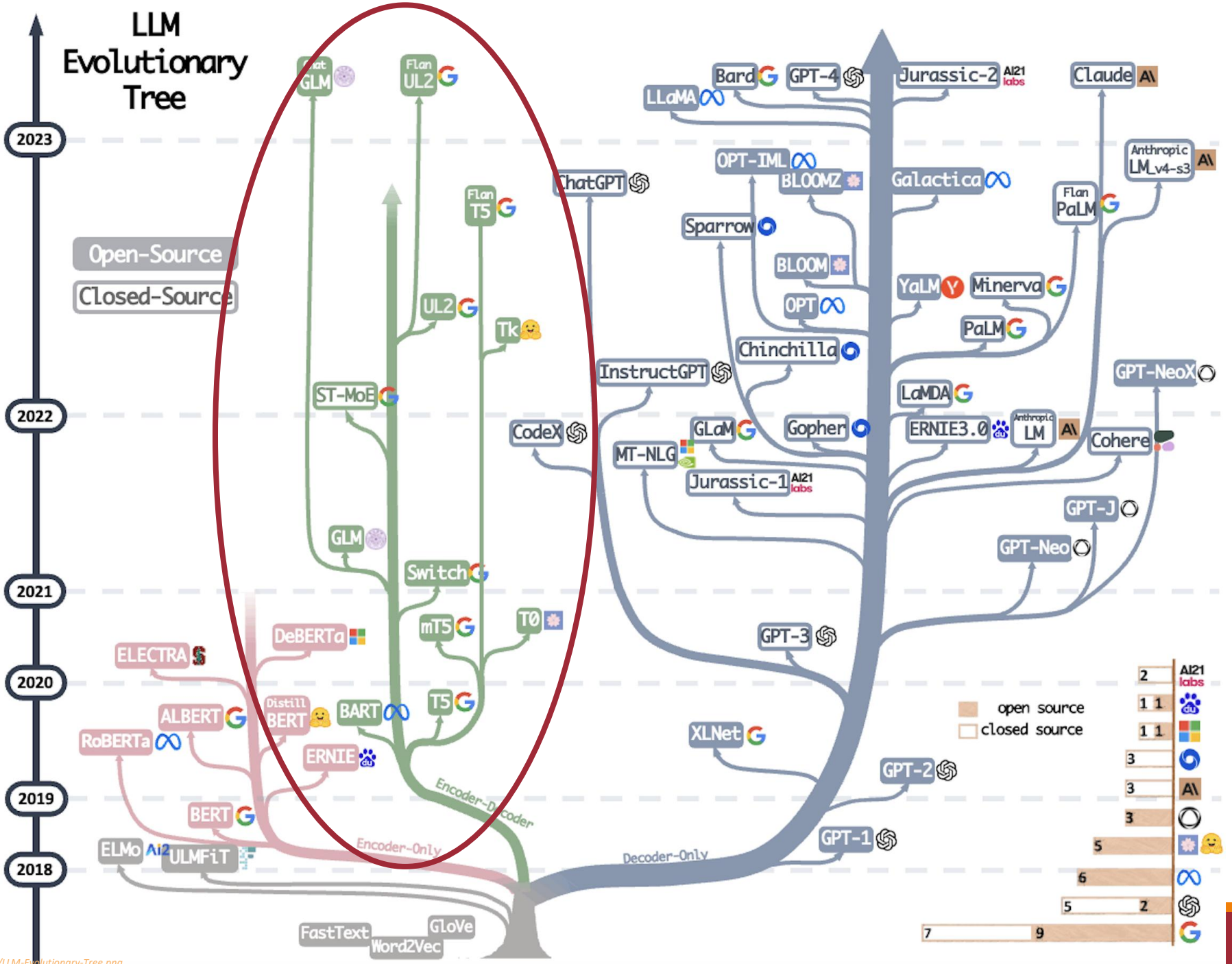
Mixture-of-Experts (MoE) (e.g., Mixtral, Nemotron 3.5 Lightning)

- Architecture that only activates certain clusters of neurons

See CMSC 473/673 Intro to NLP for more info on these!

Encoder-Decoder Models





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

Enc-Dec Models

- Encoder-decoder
  - Input: Text sequence with random word spans masked (originally)
  - Goal: Generate the masked word spans
- When to use:
  - Translation
  - Generation or classification tasks across modalities

Some Enc-Dec family members

- T5 – Raffel et al. 2020 (Google)
- BART (combo of GPT and BERT) – (Facebook)

- DALL·E (for image generation, aka text to image) – (OpenAI)

Scripts

Levels of Information

‘What’s it going to be then, eh?’

There was me, that is Alex, and my three droogs, that is Pete, Georgie, and Dim, Dim being really dim, and we sat in the Korova Milkbar making up our rassoodocks what to do with the evening, a flip dark chill winter bastard though dry. The Korova Milkbar was a milk-plus mesto, and you may, O my brothers, have forgotten what these mestos were like, things changing so skorry these days and everybody very quick to forget, newspapers not being read much neither. Well, what they sold there was milk plus something else. They had no licence for selling liquor, but there was no law yet against prodding some of the new veshches which they used to put into the old moloko, so you could peet it with vellocet or synthemesc or drencrom or one or two other veshches which would give you a nice quiet horror-show fifteen minutes admiring Bog and All His Holy Angels and Saints in your left shoe with lights nursing all over your mozg. …

Text from *A Clockwork Orange* by Anthony Burgess

The story begins with the droogs sitting in their favourite hangout, the Korova Milk Bar, and drinking "milk-plus" – a beverage consisting of milk laced with the customer's drug of choice – to prepare for a night of ultra-violence.

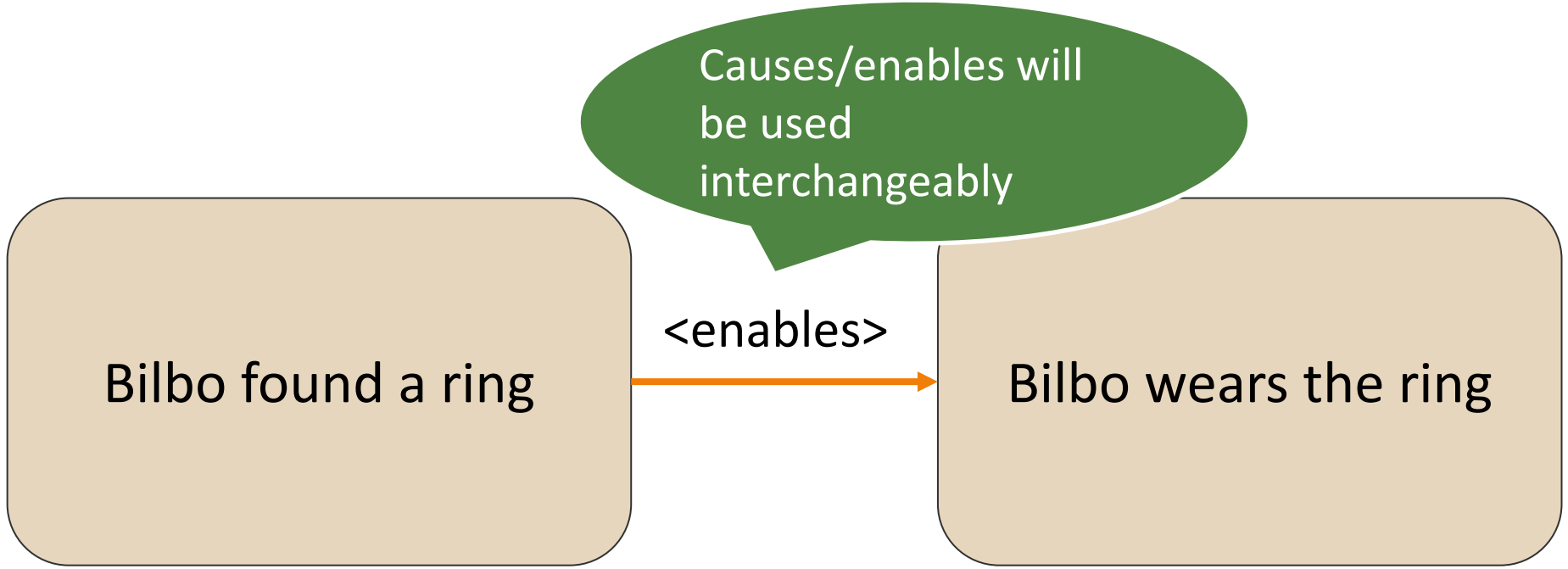
Summary from Wikipedia

Alex begins his narrative from the Korova, where the boys sit around drinking.

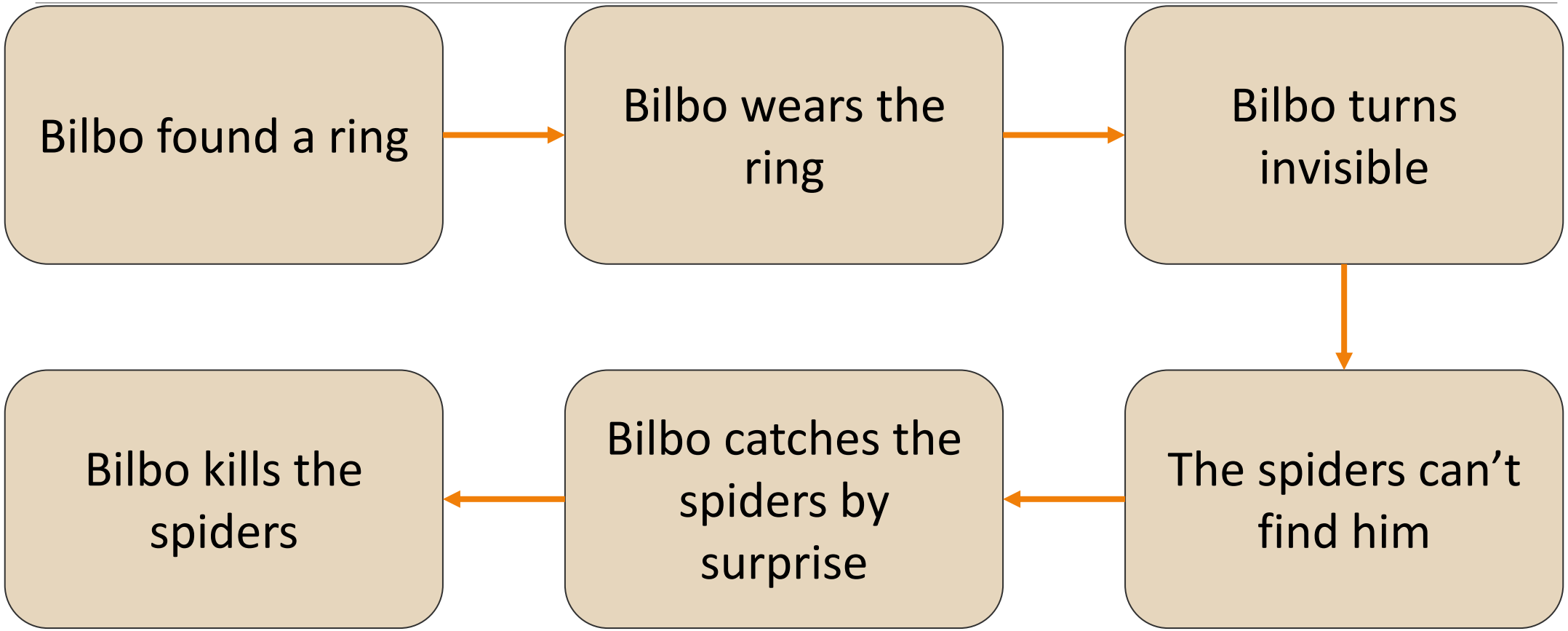
Summary from SparkNotes.com

Bilbo joined a group of dwarves on an adventure.
He found a ring in a cave.
He was able to save the dwarves from giant spiders.

Causal Links



Causal Chains

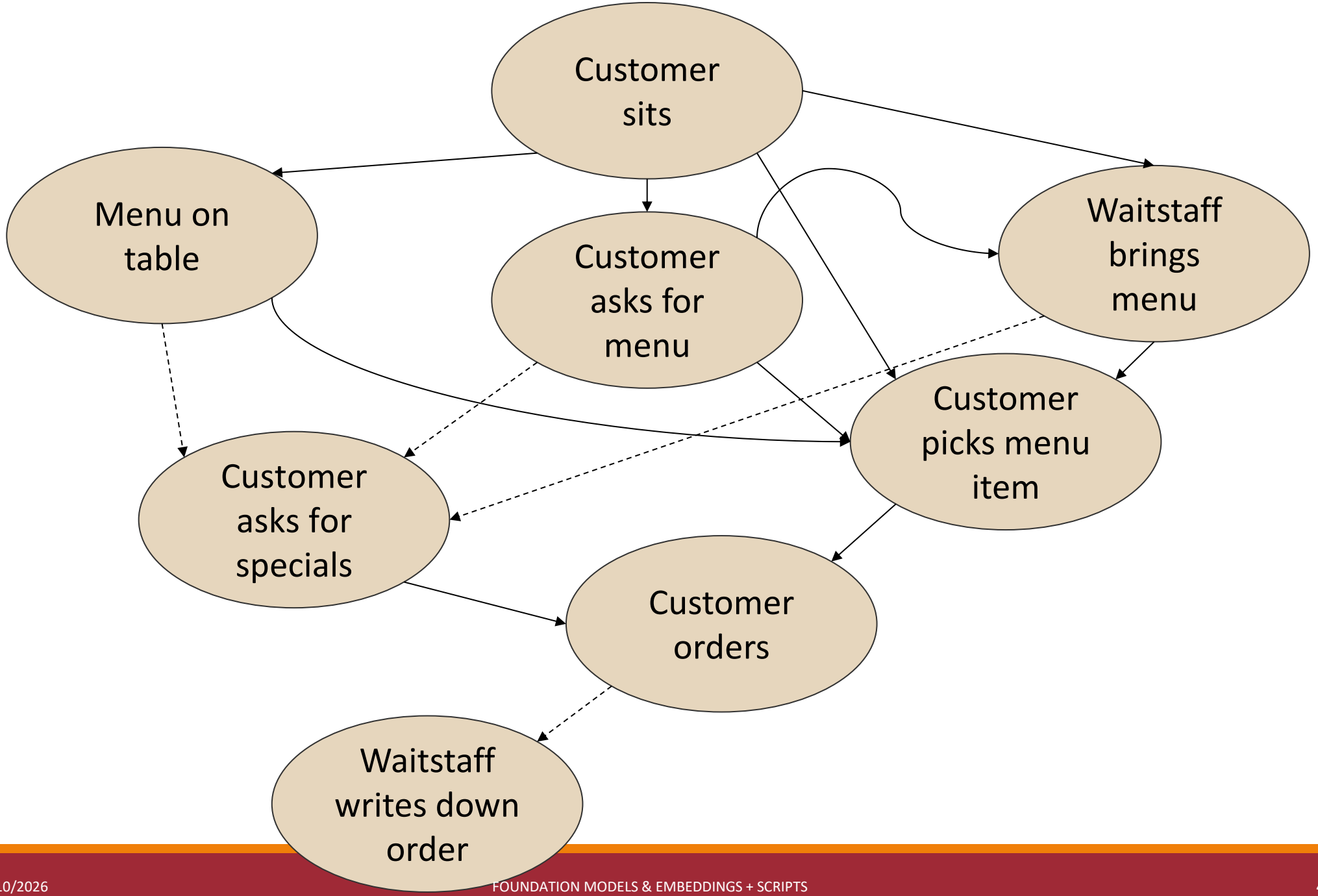


Script

"A standard event sequence" that

- Lays out different paths/options
- Consists of causal chains
- Can be used to leave out tedious details the reader is expected to know
- Can be considered a literary trope or a common social scenario

Schank, R. C., & Abelson, R. P. (1977). In *Scripts, plans, goals, and understanding: An enquiry into human knowledge structures* (The Artificial Intelligence Series, pp. 36–68). Erlbaum.


Customer sits
Menu on table
Customer asks for menu
Waitstaff brings menu
Customer picks menu item
Customer asks for specials
Customer orders
Waitstaff writes down order

Causal Chains vs Scripts

CAUSAL CHAINS

"Sequential flow of events" with cause & effect links between

Might be missing events that are considered "common knowledge"

SCRIPTS

Frequently co-occurring events

The "common knowledge"

Disclaimer: This is the way Schank and Abelson defined them, but sometimes these terms get fuzzy in today's research, especially because of the use of probabilistic models (further in the lecture)

Script

"A standard event sequence" that

- Lays out different paths/options
- Consists of causal chains
- Can be used to leave out tedious details the reader is expected to know

Can be considered a literary trope or a common social scenario

Schank, R. C., & Abelson, R. P. (1977). In *Scripts, plans, goals, and understanding: An enquiry into human knowledge structures* (The Artificial Intelligence Series, pp. 36–68). Erlbaum.

Principle of Minimal Departure

"This law—to which I shall refer as the principle of minimal departure—states that we reconstrue the central world of a textual universe in the same way we reconstrue the alternate possible worlds of nonfactual statement: as conforming as far as possible to our representation of [the actual world]"

In other words:

The story world is expected to be like the real world, unless otherwise specified

Ryan, M.-L. (1991). Chapter 3: Reconstructing the Textual Universe: The Principle of Minimal Departure. In *Possible Worlds, Artificial Intelligence, and Narrative Theory* (pp. 48–60). Indiana Univ. Press.

Or, as Miguel de Cervantes points out…

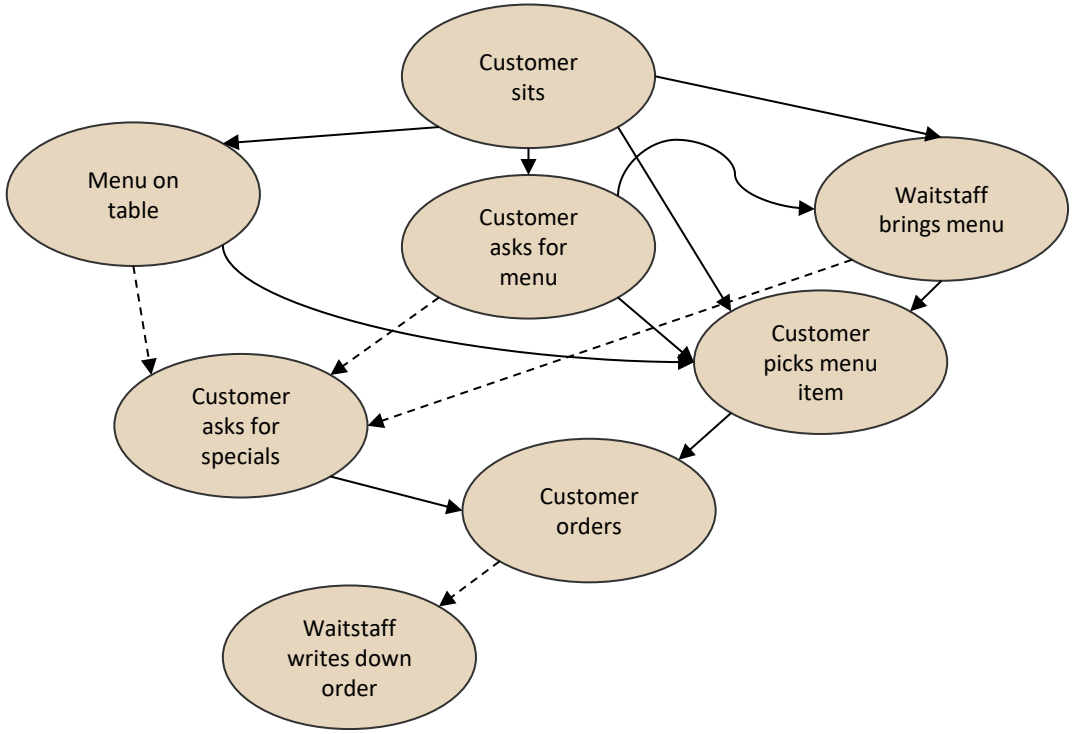
“The innkeeper asked if he had any money. Don Quixote said that he didn’t have a blanca, because he’d never read in the histories of knights errant that any one of them had taken money with him. To this, the innkeeper said that he was mistaken, because, although the histories didn’t specify something as obvious and necessary as money and clean shirts, there was no reason to believe that they didn’t have them.”

In-Class Activity

Create a small graph to represent the script for **“checking out at a grocery store”**

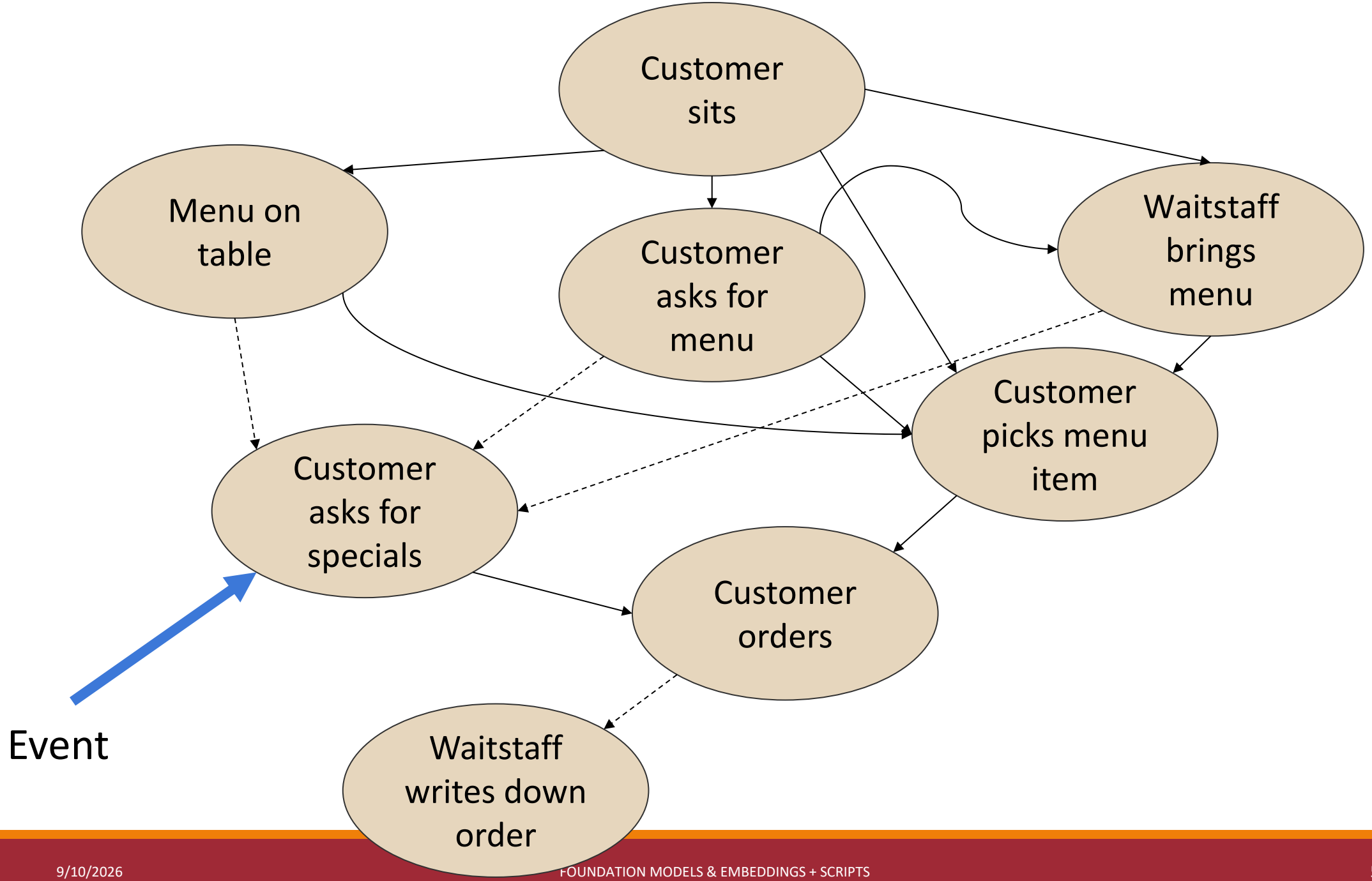
Use draw.io (or your graph-making software of choice) to make the graph or draw on paper.

No need to submit.



Now that you've gotten your hands dirty

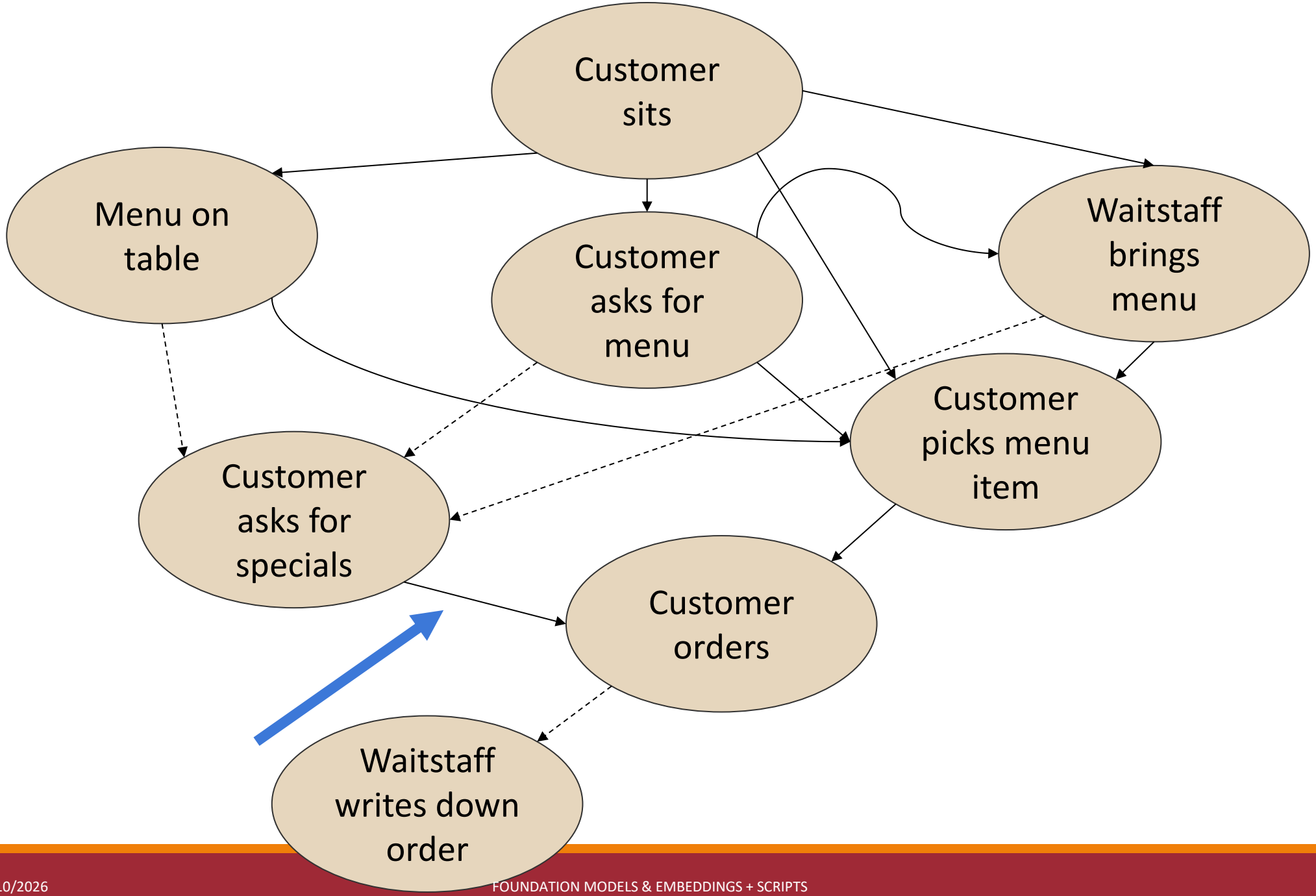
How do you think you would you use scripts to generate text (such as a story)?


Customer sits
Menu on table
Customer asks for menu
Waitstaff brings menu
Customer picks menu item
Customer asks for specials
Customer orders
Waitstaff writes down order
Event

Events

An occurrence that

- has a distinct beginning & end
- may contain multiple actions
  - E.g., “paying the bill” might involve taking out your credit card, handing it over to the staff, etc.
- can be at a variety of levels of granularity
  - E.g. “opens wallet” vs “pays the bill” vs “ate at the restaurant”

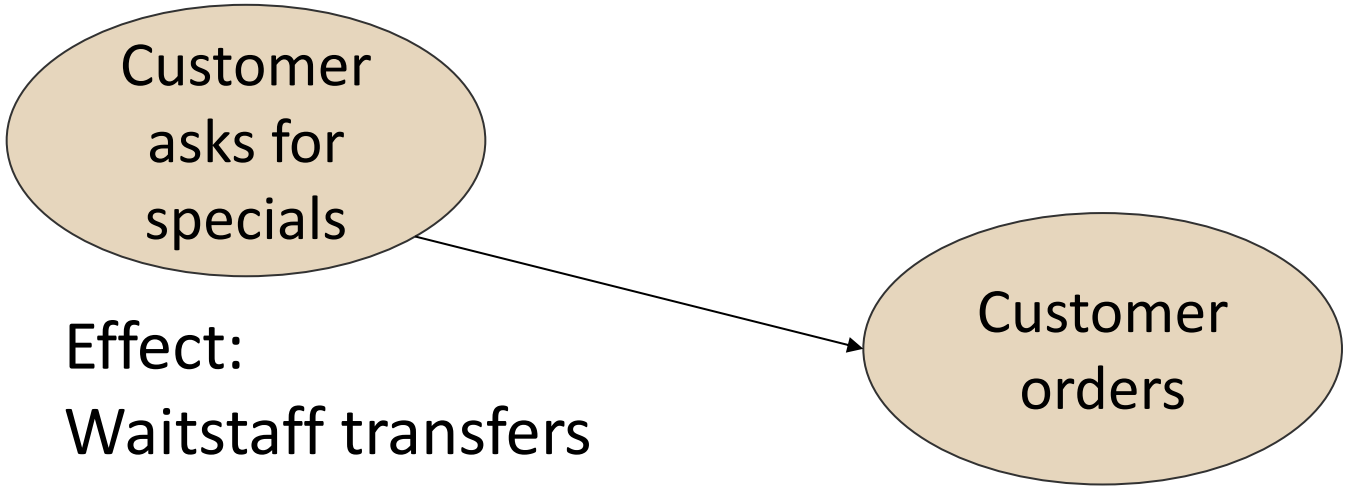

Customer sits
Menu on table
Customer asks for menu
Waitstaff brings menu
Customer picks menu item
Customer asks for specials
Customer orders
Waitstaff writes down order

Pre-conditions and Effects

Pre-conditions determine what the state of the world needs to be in before an event is entered

Effects (or post-conditions) tell you what happens after an event occurs

Restaurant Example



Effect:
Waitstaff transfers "specials" info to customer

Pre-condition:
Customer knows what they want to order

Domain

dig(?char,?item)

Precons: ?char alive.
?item buried.
?char knows ?item.

Effects: ?char has ?item.
¬ ?item buried.

Consent: ?char

give(?gvr,?item,?rcvr)

Precons: ?gvr alive.
?gvr has ?item.
?rcvr alive.

Effects: ?rcvr has ?item.
¬ ?gvr has ?item.

Consent: ?gvr ?rcvr

open(?char)

Precons: ?char alive.
?char has ?item.

Effects: R opened.
¬ ?char alive.

Consent: ?char

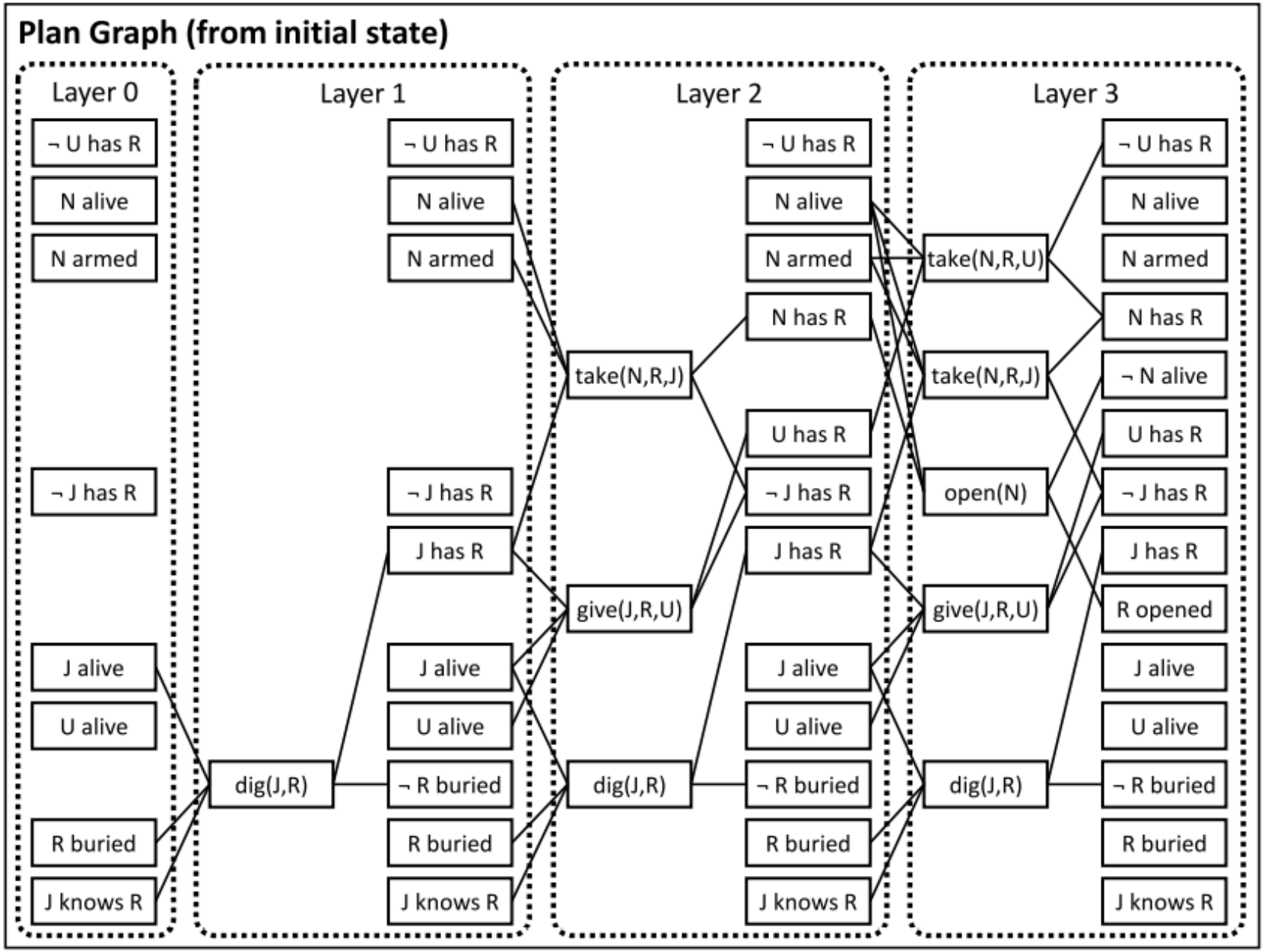
take(?thief,?item,?char)

Precons: ?thief alive.
?char has ?item.
¬ ?char alive OR
?thief armed.

Effects: ?thief has ?item.
¬ ?char has ?item.

Consent: ?thief

Plan Graph (from initial state)



Stephen G. Ware and R. Michael Young. 2014. Glaive: a state-space narrative planner supporting intentionality and conflict. In *Proceedings of the Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (*AIIDE'14*). AAAI Press, 80–86.

Which of the following statements about scripts is FALSE?

PollEv.com/laramartin527

Script information is often implied in stories **(A)**

A script is composed of frequently co-occurring events **(B)**

For a given scenario, everyone shares the same script **(C)**

A script is made up of causal chains **(D)**

Powered by Poll Everywhere

Linking Events

PROBABILISTIC

Occur frequently together (not necessarily because they had to)

Example:

I pour dog food in my dog's bowl.

I pet my dog.

CAUSAL

Occur because of one another

Example:

I pour dog food in my dog's bowl.

My dog eats dog food.

Example of a Probabilistic Event Representation

From sentence, extract event representation:

(subject, verb, direct object, modifier, preposition)

Original sentence: yoda uses the force to take apart the platform

Events:

yoda use force Ø Ø
yoda take_apart platform Ø Ø

Generalized Events:

<PERSON>0 fit-54.3 power.n.01 Ø Ø

<PERSON>0 destroy-44 surface.n.01 Ø Ø

What might be the pros of using scripts?

PollEv.com/laramartin527

What might be the cons of using scripts?

PollEv.com/laramartin527

Pros & Cons of Scripts

PROS	CONS