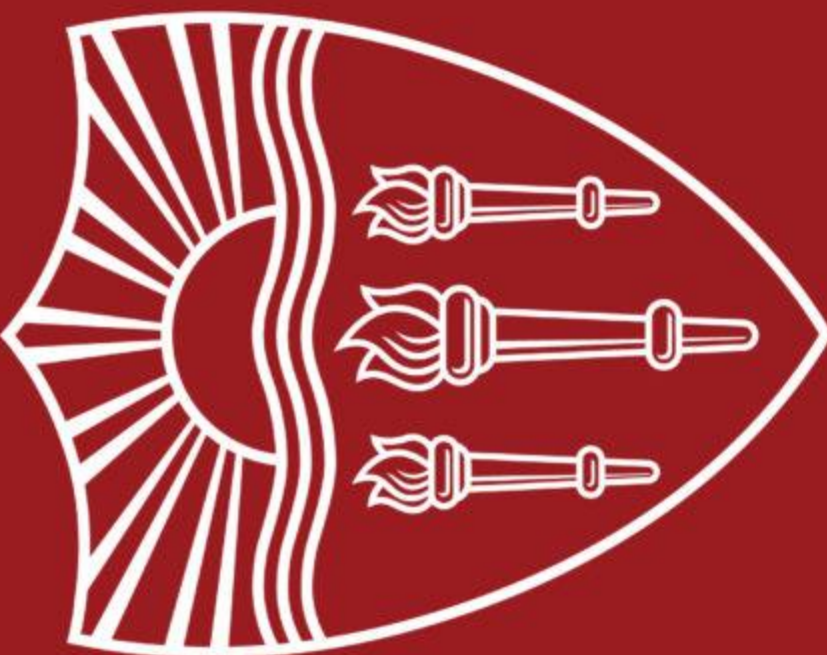
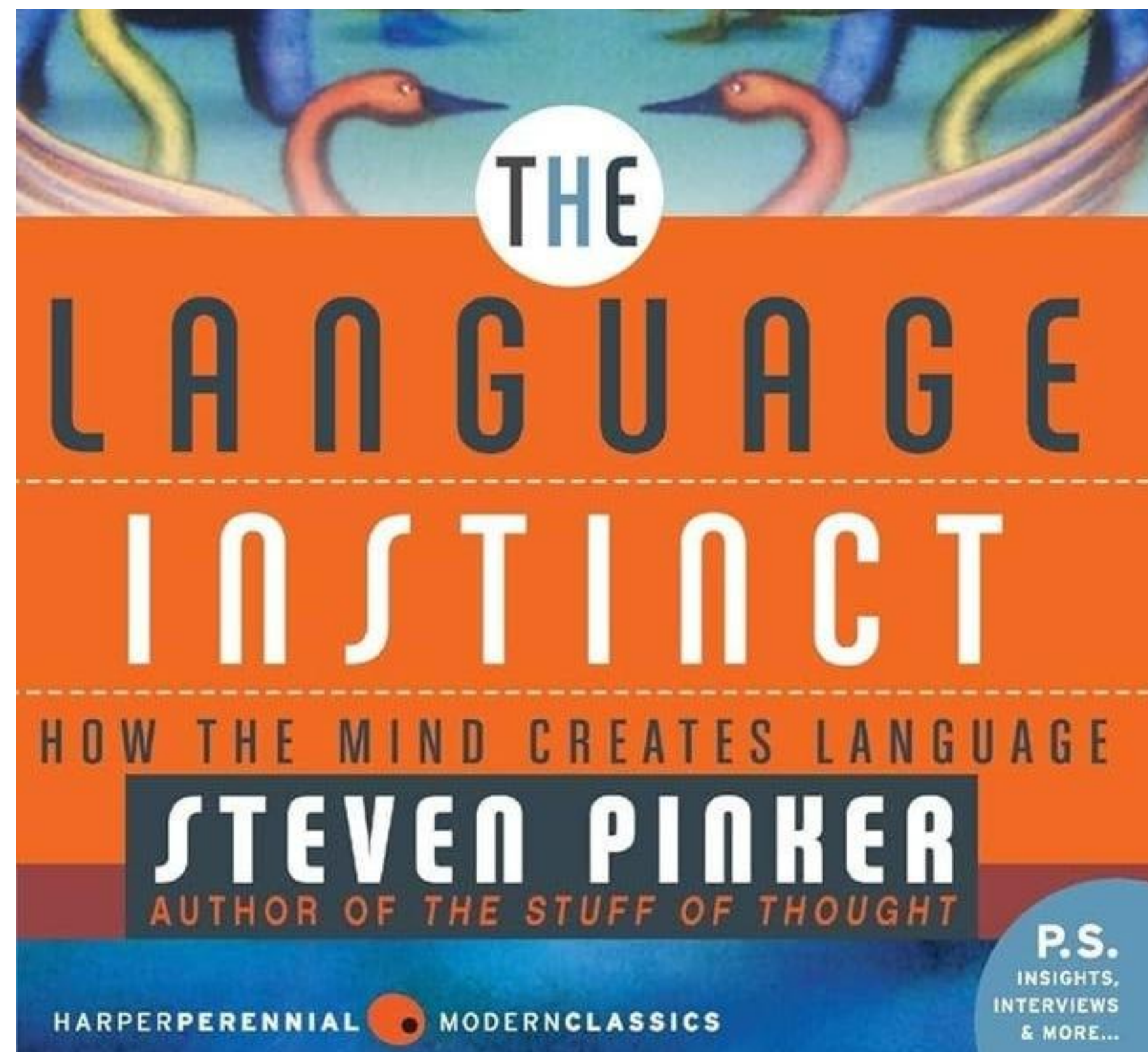




Lecture 01: Introduction and Course Overview

Xiang Ren
USC CSCI 662 Advanced NLP
Spring 2026

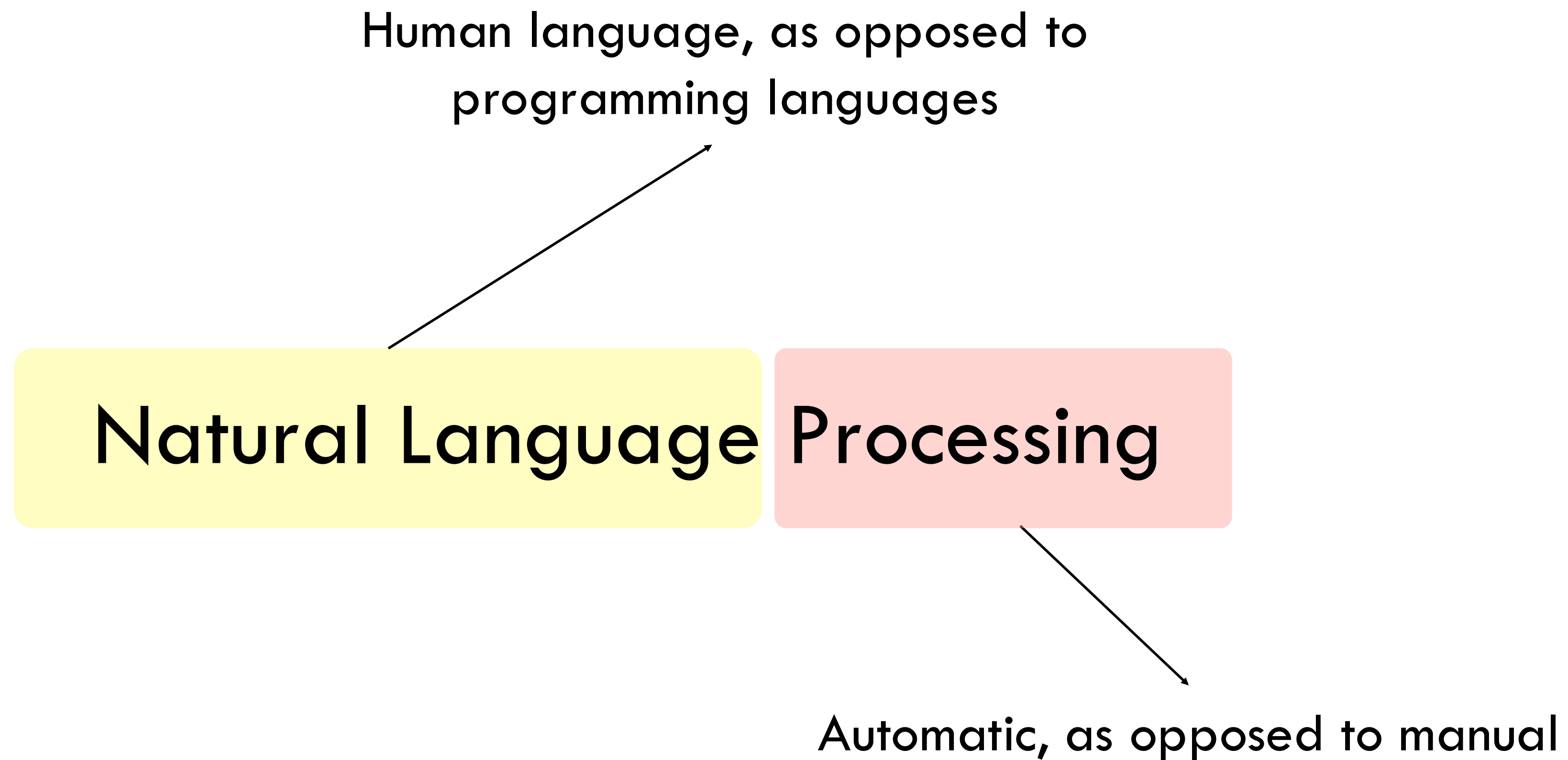
Language is a key modality



“Humans are so innately hardwired for language that they can no more suppress their ability to learn and use language than they can suppress the instinct to pull a hand back from a hot surface.”

– Steven Pinker

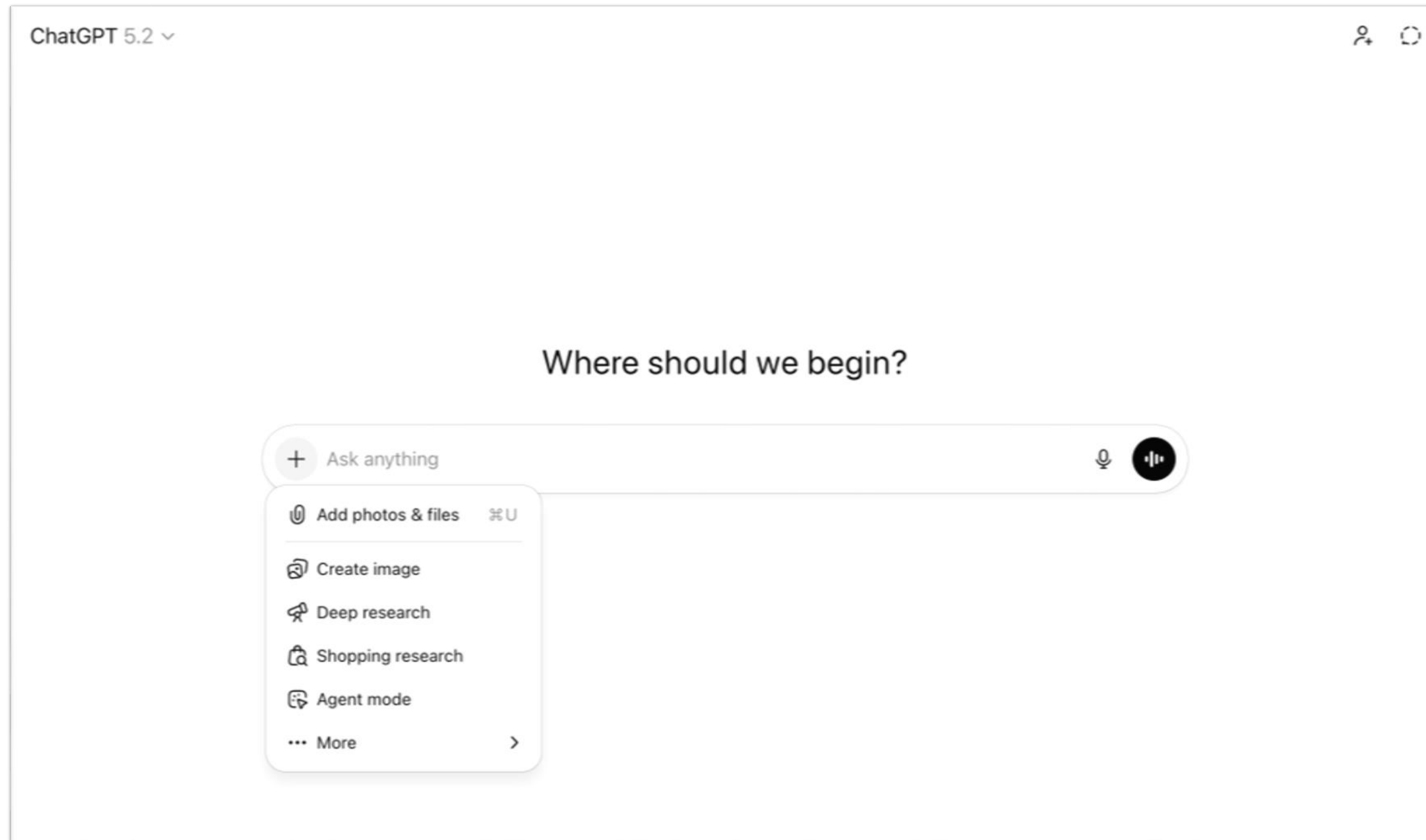
Four
Minute
Books



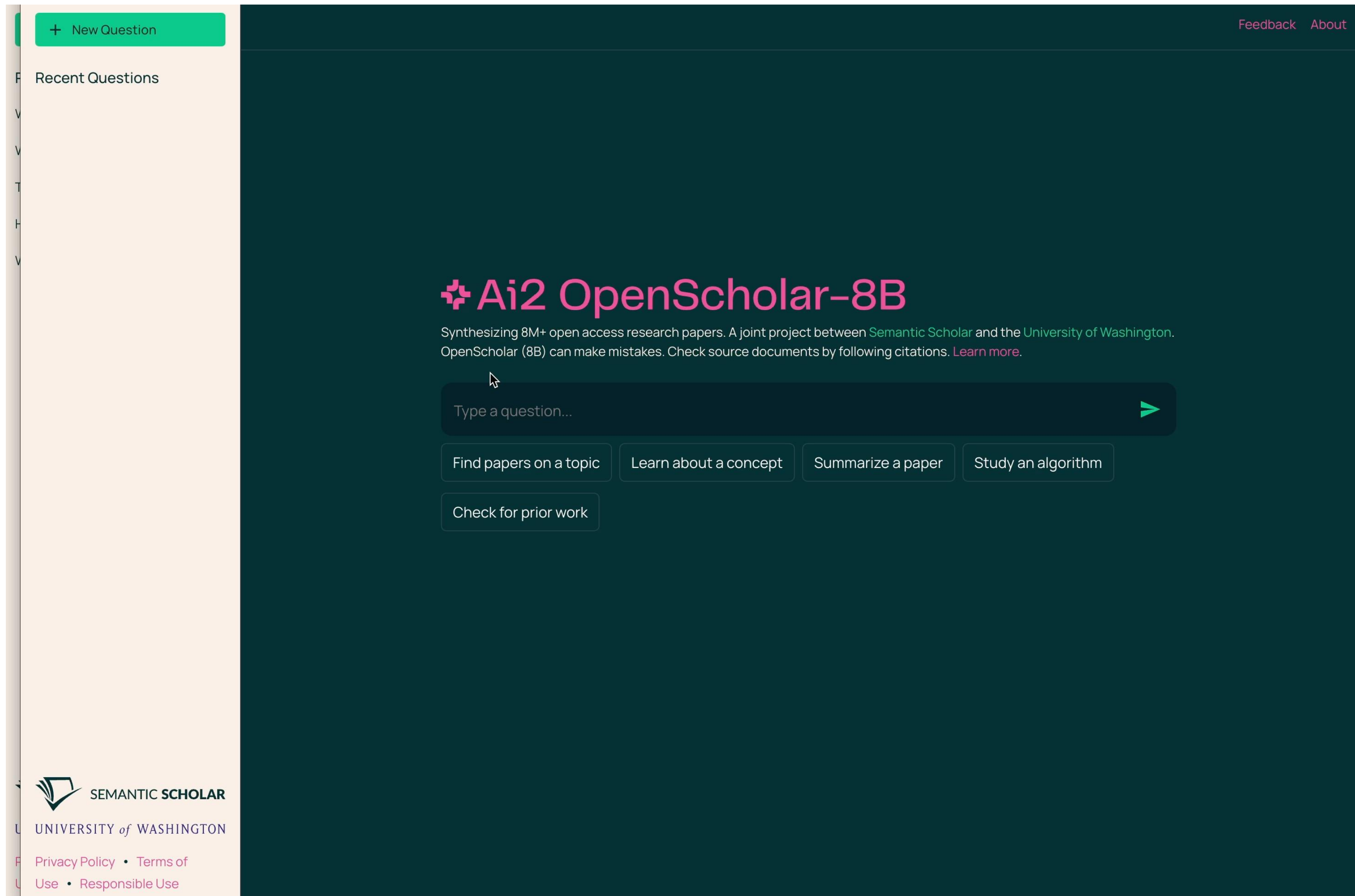
What is Natural Language Processing ?

- Field at the intersection of computer science, AI (especially machine learning) and linguistics
- Goal: for computers to process human language, similar to human understanding, towards performing useful tasks
- Challenge: understanding and representing the meaning of language is something even humans struggle with
- Processing: produce outputs (Y) with language or text as input (X)

Today's NLP



Today's NLP



The screenshot shows the Ai2 OpenScholar-8B web interface. On the left is a sidebar with a green '+ New Question' button, a 'Recent Questions' section, and logos for Semantic Scholar and the University of Washington. The main area has a dark teal background with the title 'Ai2 OpenScholar-8B' in pink. Below the title is a description of the project and a 'Learn more' link. A search bar with a green arrow button is present, followed by five buttons: 'Find papers on a topic', 'Learn about a concept', 'Summarize a paper', 'Study an algorithm', and 'Check for prior work'. The top right of the main area contains 'Feedback' and 'About' links.

+ New Question

Recent Questions

F
V
V
T
H
V

Feedback About

Ai2 OpenScholar-8B

Synthesizing 8M+ open access research papers. A joint project between [Semantic Scholar](#) and the [University of Washington](#). OpenScholar (8B) can make mistakes. Check source documents by following citations. [Learn more](#).

Type a question... ➤

Find papers on a topic Learn about a concept Summarize a paper Study an algorithm

Check for prior work

SEMANTIC SCHOLAR

UNIVERSITY of WASHINGTON

[Privacy Policy](#) • [Terms of Use](#) • [Responsible Use](#)

<https://openscholar.allen.ai/>

A General Framework for NLP Systems

- Create a function to map an **input** X into an **output** Y , where X and/or Y involve language.

<u>Input X</u>	<u>Output Y</u>	<u>Task</u>
Text	Continuing Text	Language Modeling
Text	Text in Other Language	Translation
Text	Label	Text Classification
Text	Linguistic Structure	Language Analysis
Image	Text	Image Captioning

Language Models Are Everywhere



Virtual Assistants



Translation



Content Creation



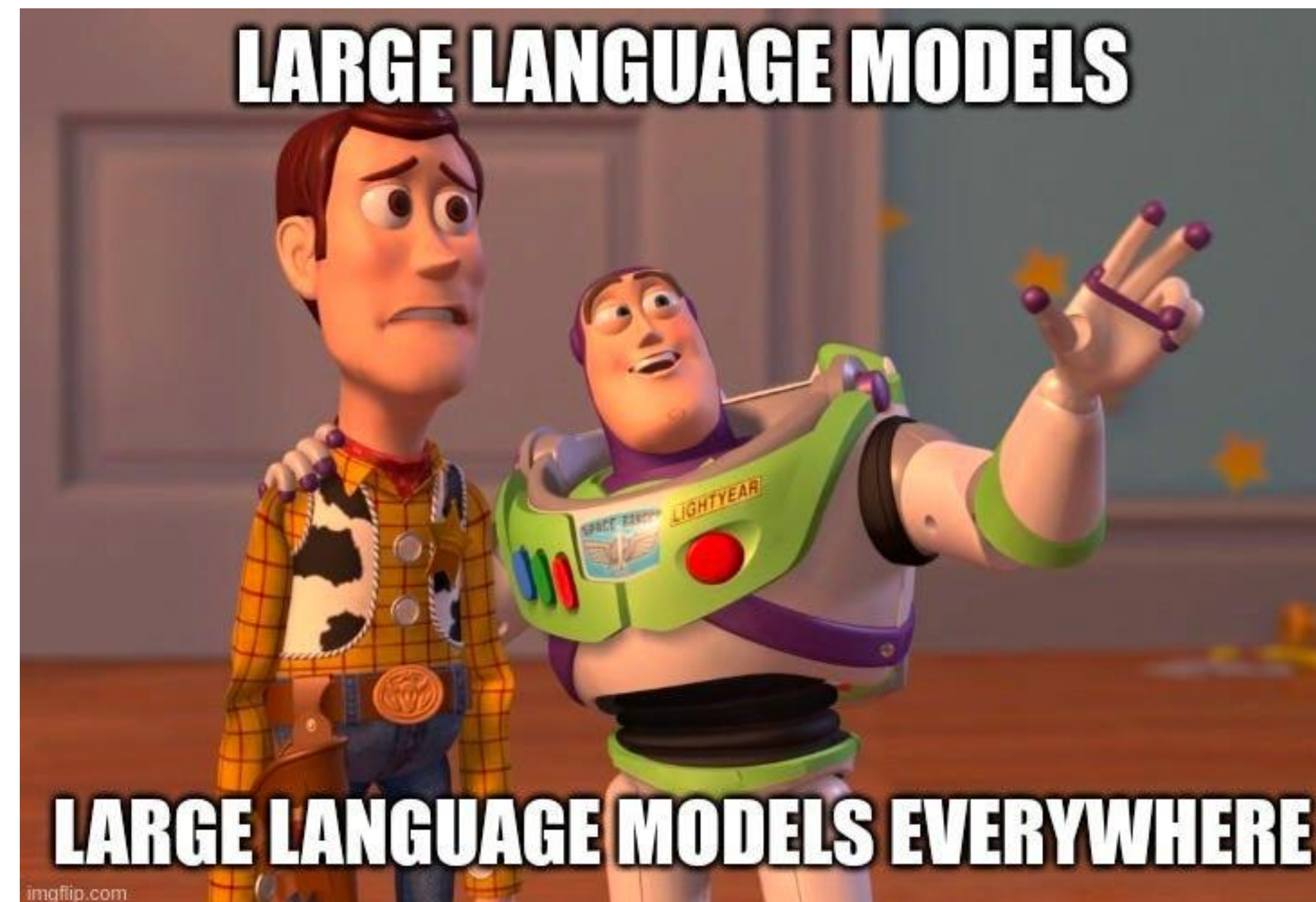
Code Generation



Customer Service



Data Analysis



You Probably Can See Where This Is Going

When my flight from LaGuardia landed in Minneapolis on that August afternoon, the first text I received was from the executive director of the nonprofit I'd be holding the training for the next day, canceling our dinner because of a family emergency. The second text was from my friend Jenny asking me to look at the profile of a guy named James on the dating app we both used and to let her know if it was the same asked-not-one-question James I'd gone out with around Christmas. That date had felt like such a waste of an evening that I'd given myself a break from the app since.

By the time the seatbelt light had gone off, I'd texted Jenny to tell her it wasn't, and I'd received a heart on the app from a man who was currently online, whose handle was MtnBiker1971. He was 53, a year younger than me; he had deep brown eyes and was bald with a gray and brown beard; and three of his five photos featured him on a bike. I swear, I *swear*, that MtnBiker1971 and I already had exchanged generic greetings before it occurred to me that because my settings showed profiles within 10 miles of me, he was local.

"Oh sorry!" I typed. "Just realized you live in Minnesota and I'm only visiting for 36 hrs."

Before I could send this, a new message came in from him: "What's your favorite fruit that's considered a vegetable and what's your best episode of TV ever?"

So I deleted "Oh sorry..." and typed "Okra and the season finale of Severance. You?"

Can you see where this is going? You probably can see where this is going.

By the time I was in an Uber, he'd told me his name was Brian, he worked for an environmental advocacy group, and the previous weekend, on a trail, he'd ridden his bike past a woodpecker sitting on the back of a deer; he'd been so close that he and the deer had made eye contact.



Human or AI?

The Flip-Flop Moment

Lydia had always been practical. It was her hallmark, the trait that kept her life organized in neat rows, like the files on her desk or the cushions on her sofa. At 48, this practicality had become her armor, protecting her from the reckless impulses that she might have indulged in during her younger years. And so, when she walked into the coffee shop on that sweltering July afternoon, it was with the same cautious optimism that she had applied to everything else in her life.

The coffee shop was called Java Junction, an ironic nod to its nondescript location. It was where she met her friends for book club, where she studied with her son while he was in high school and where she'd spent countless hours with her late husband. That was the first thing she'd felt when she entered, the lingering echo of shared memories. She was here for a meeting with her college friend, Julie, and Julie was late, which gave Lydia plenty of time to observe.

She sat at a small table near the window, which was always Lydia's favorite spot. The sunlight spilled over the table, creating a halo around her as she skimmed through a magazine, her flip-flops occasionally brushing the edge of the table. She had never been one for extravagant shoes — practical, again. But on that day, the flip-flops seemed to betray a different side of her, a side that wanted to feel something more, something less anchored.

Julie arrived, panting slightly from the heat. She was a whirlwind of energy, always dressed in vibrant colors and speaking in rapid bursts. The two friends embraced, and Lydia noticed how the decades had changed them both — Julie still had that unfiltered joy, while Lydia felt a certain grayness to her own existence.

They talked about their lives — Julie's recent move to a beach town and Lydia's endless workdays, the responsibilities of being a single mother, the growing distance from her teenage son. They laughed about old times and reminisced about their college days, and for a moment, Lydia felt something she hadn't in a while: a spark of connection, of vitality.

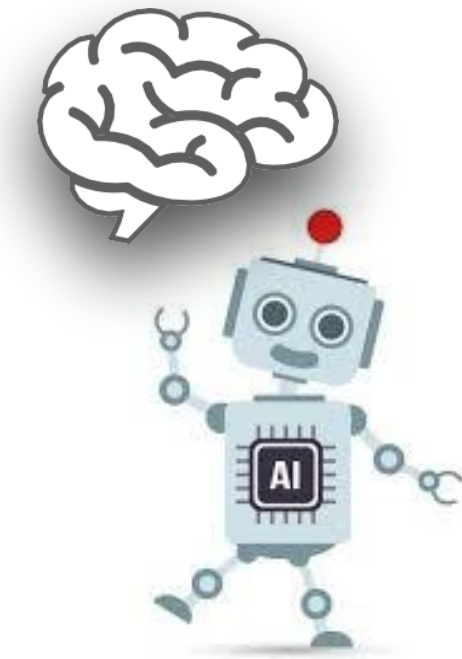
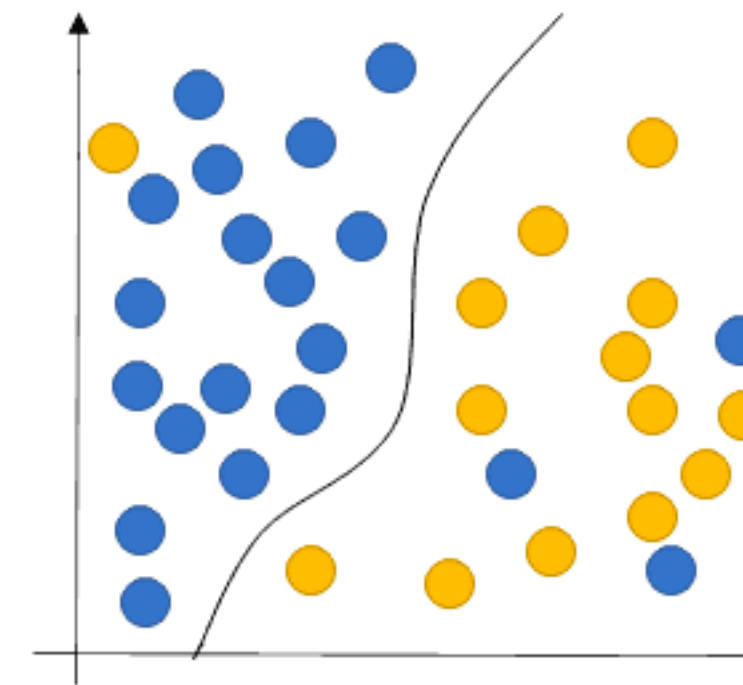
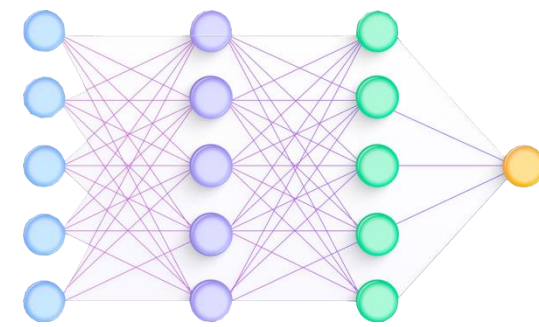


GPT-4o

Artificial Intelligence

Machine Learning and Natural Language Processing

Deep Learning



Generative AI

Language
Models





OpenAI Is Testing an A.I.-Powered Search Engine

Language Models are the most popular form of AI today!

Smart Assistants Across Its Apps

Can You Be Emotionally Reliant on an A.I. Voice? OpenAI Says Yes.

In Constant Battle With Insurers, Doctors Reach for a Cudgel: A.I.

Aided by A.I. Language Models, Google's Robots Are Getting Smart

In Big Election Year, A.I.'s Architects Move Against Its Misuse

A.I.'s Insatiable Appetite for Energy

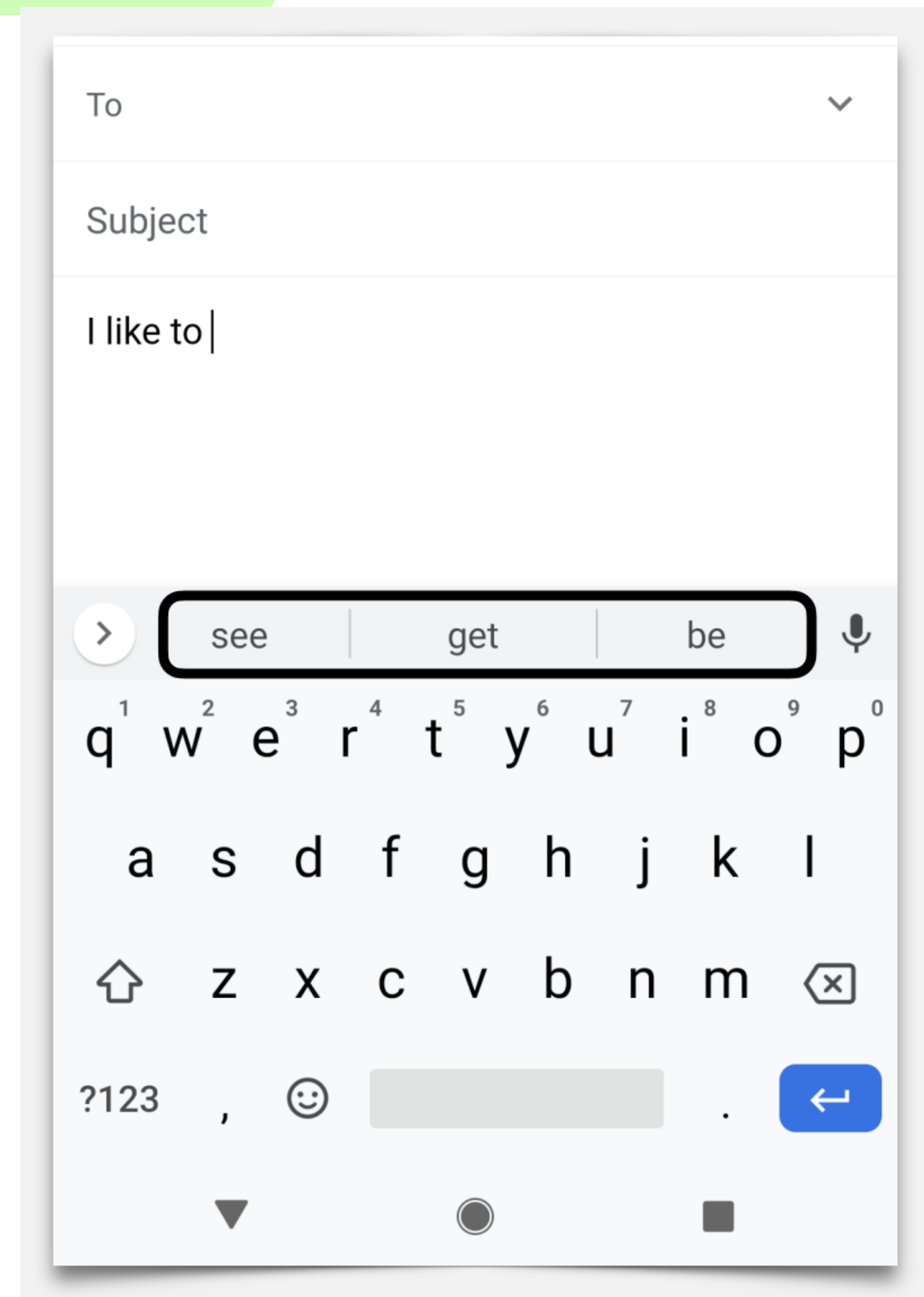
A.I. Can Write Poetry, but It Struggles With Math

The Chef Is Human. The Reviewer Isn't.

Google C.E.O. Sundar Pichai on the A.I. Moment: 'You Will See Us Be Bold'
In an extended interview, Mr. Pichai expressed both optimism and worry about the state of the A.I. race.

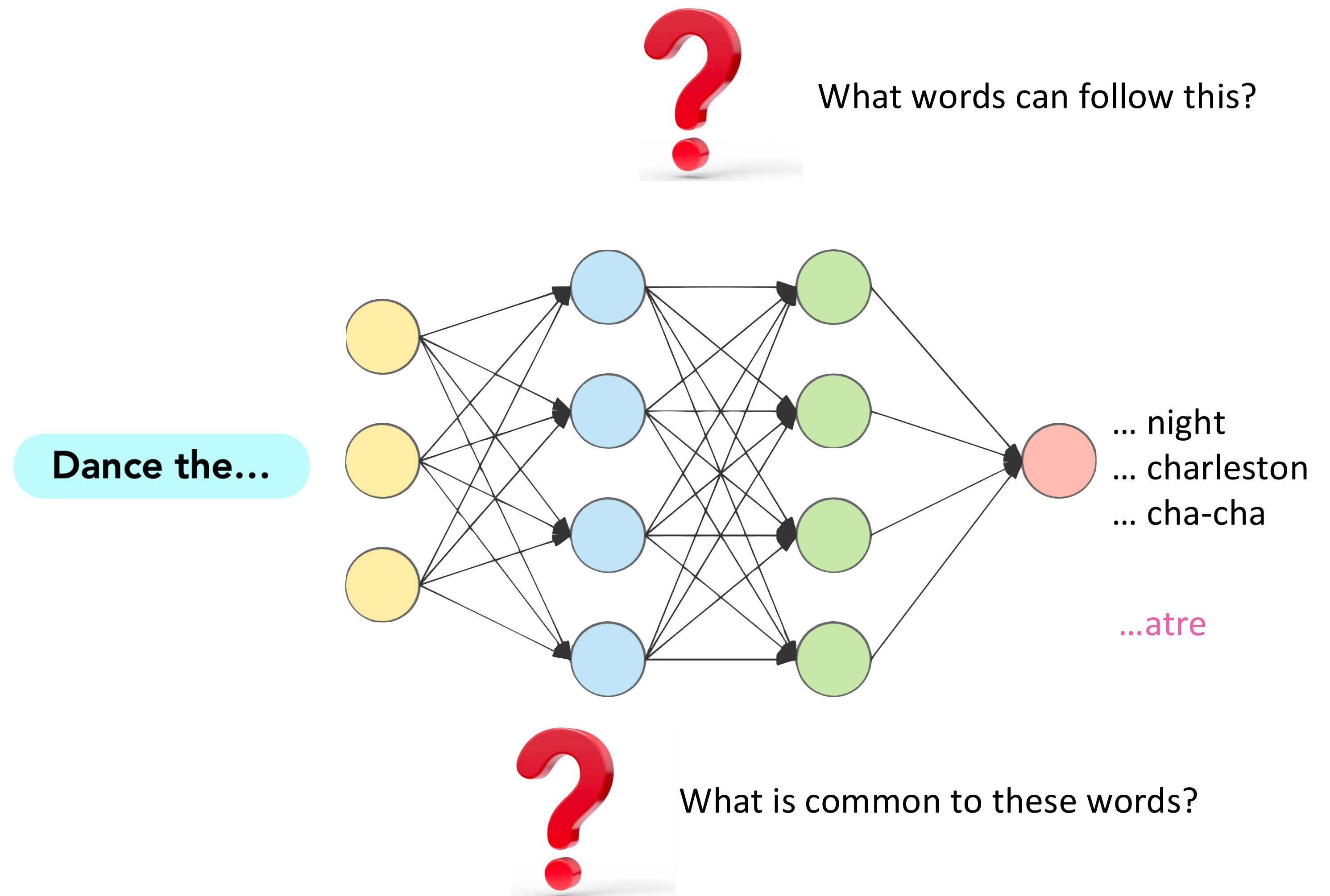
Language Models

- Task: Given a sequence of words so far (the context), predict what comes next
 - Like autocomplete!
- We never know for sure what comes next, but we can still make good guesses!
- Question: what is X and what is Y here?
 - Input X = Language and Output Y = Language



Language Model as Autocomplete

- Task: Given a sequence of words so far (the context), predict what comes next
- We never know for sure what comes next, but we can still make good guesses!



Language Model as Autocomplete

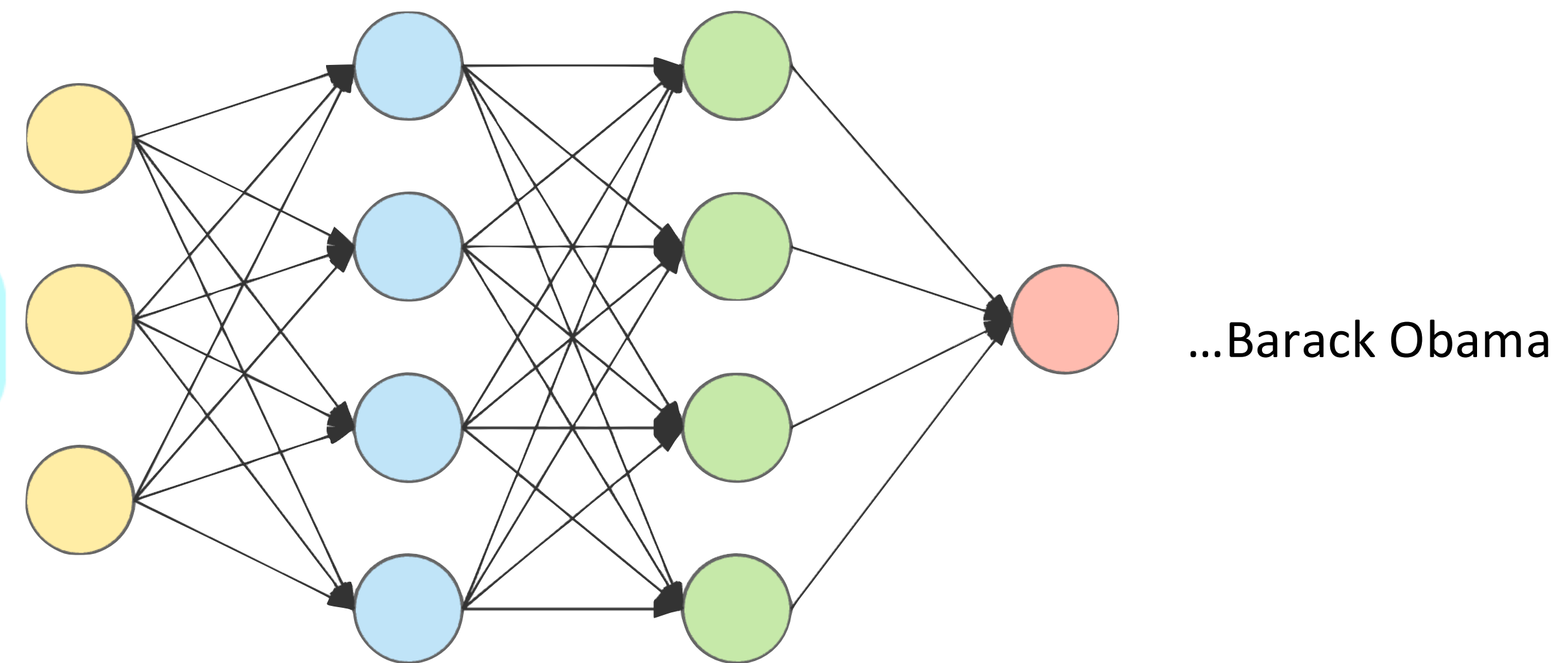


What words can follow this?

- Task: Given a sequence of words so far (the context), predict what comes next

**The 44th President of
United States was ...**

- We never know for sure what comes next, but we can still make good guesses!



I want to _____

◀

!

|

Language Models

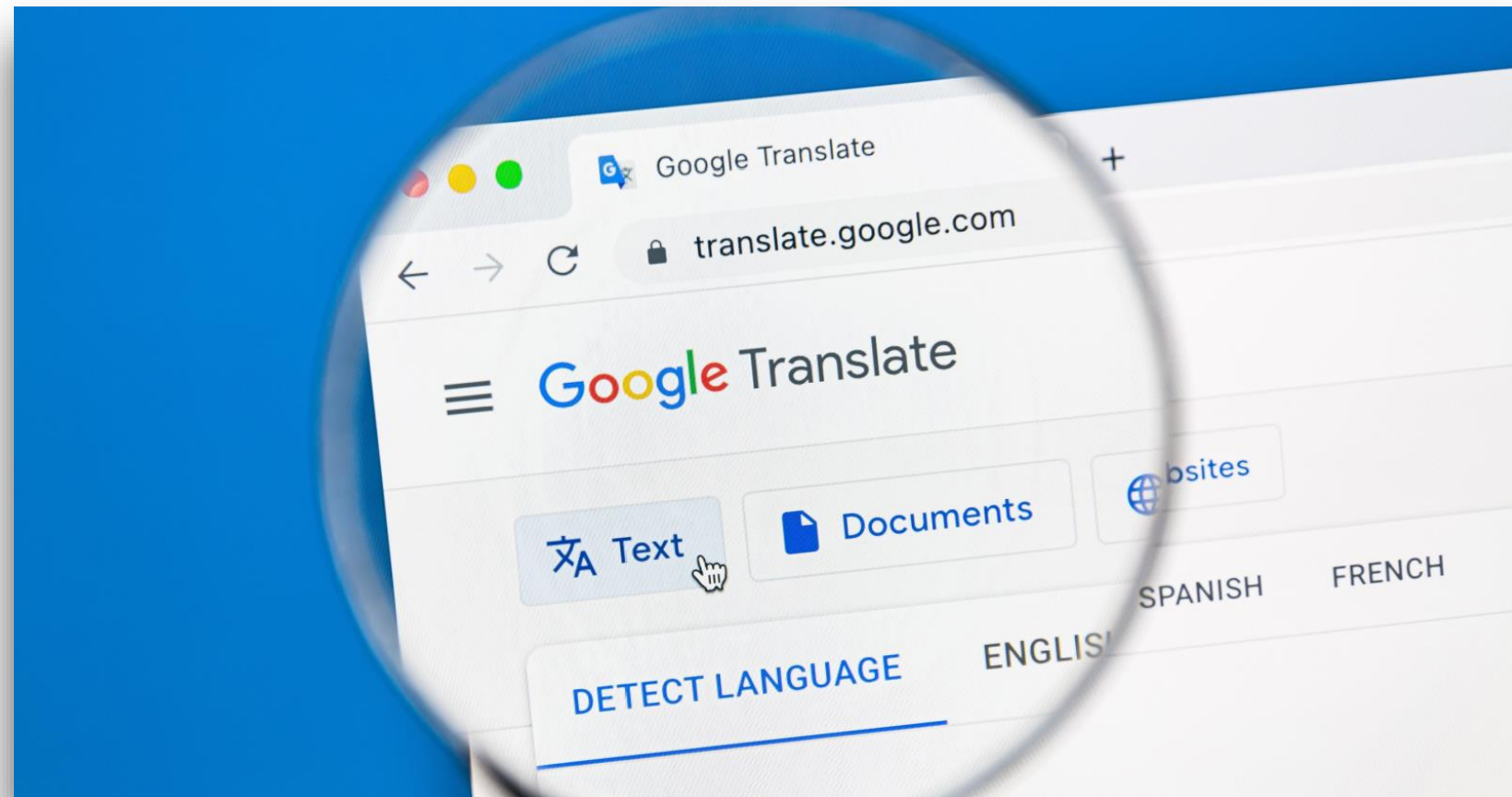
- But today, variants of language models are used
 - As chat models
 - As supercharged auto completes
 - As instruction-following assistants
 - And many more...

Apple's Siri

- Understands the user
- Remembers what the user said earlier
- Can understand which alarm she is referring to



Google Translate



- Detects language automatically
- Can reorder spans in text on the fly

中共中央政治局7月30日召开会议，会议分析研究当前经济形势，部署下半年经济工作。

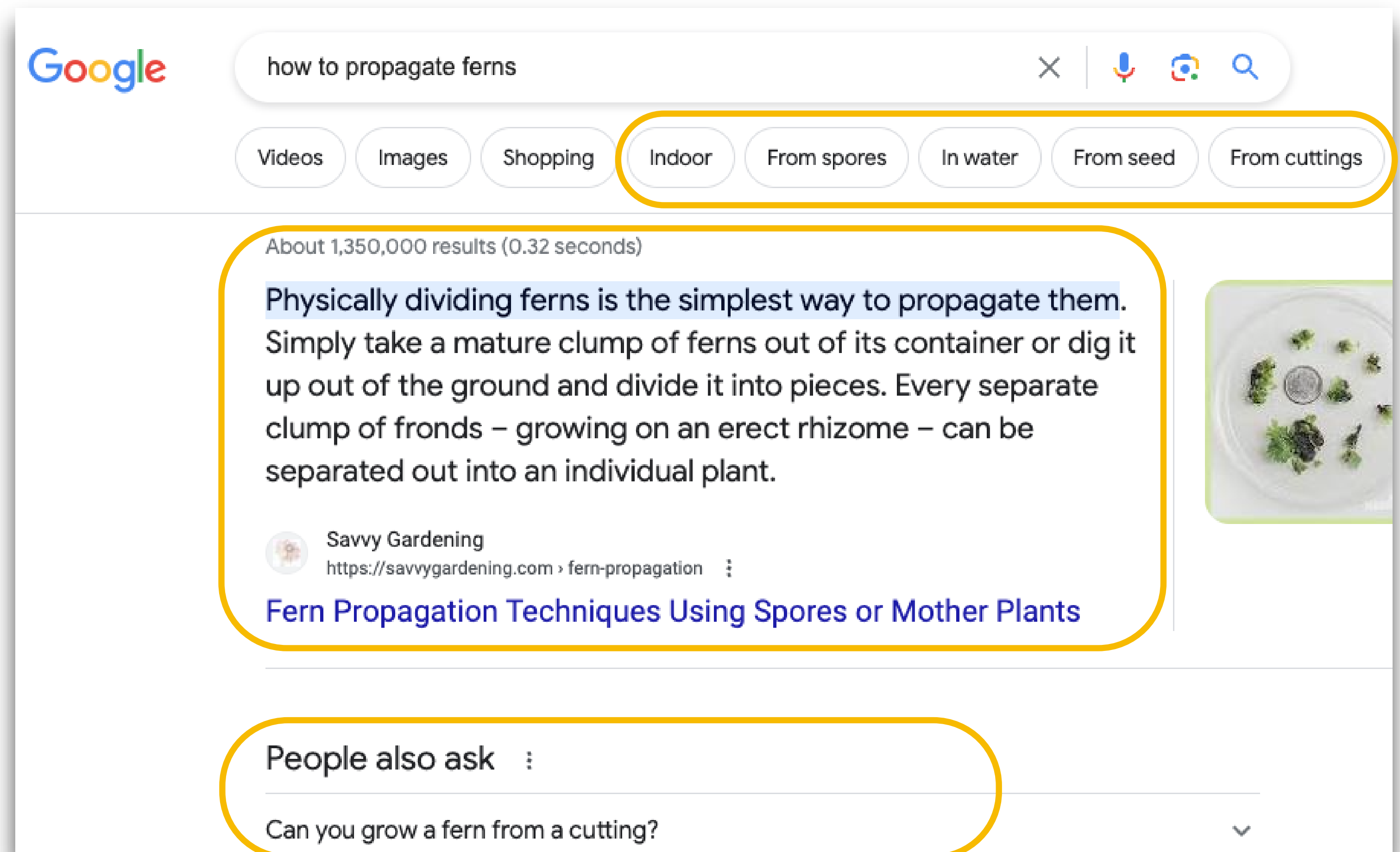
People's Daily, August 10, 2020

Translate

The Political Bureau of the CPC Central Committee held a meeting on July 30 to analyze and study the current economic situation and plan economic work in the second half of the year.

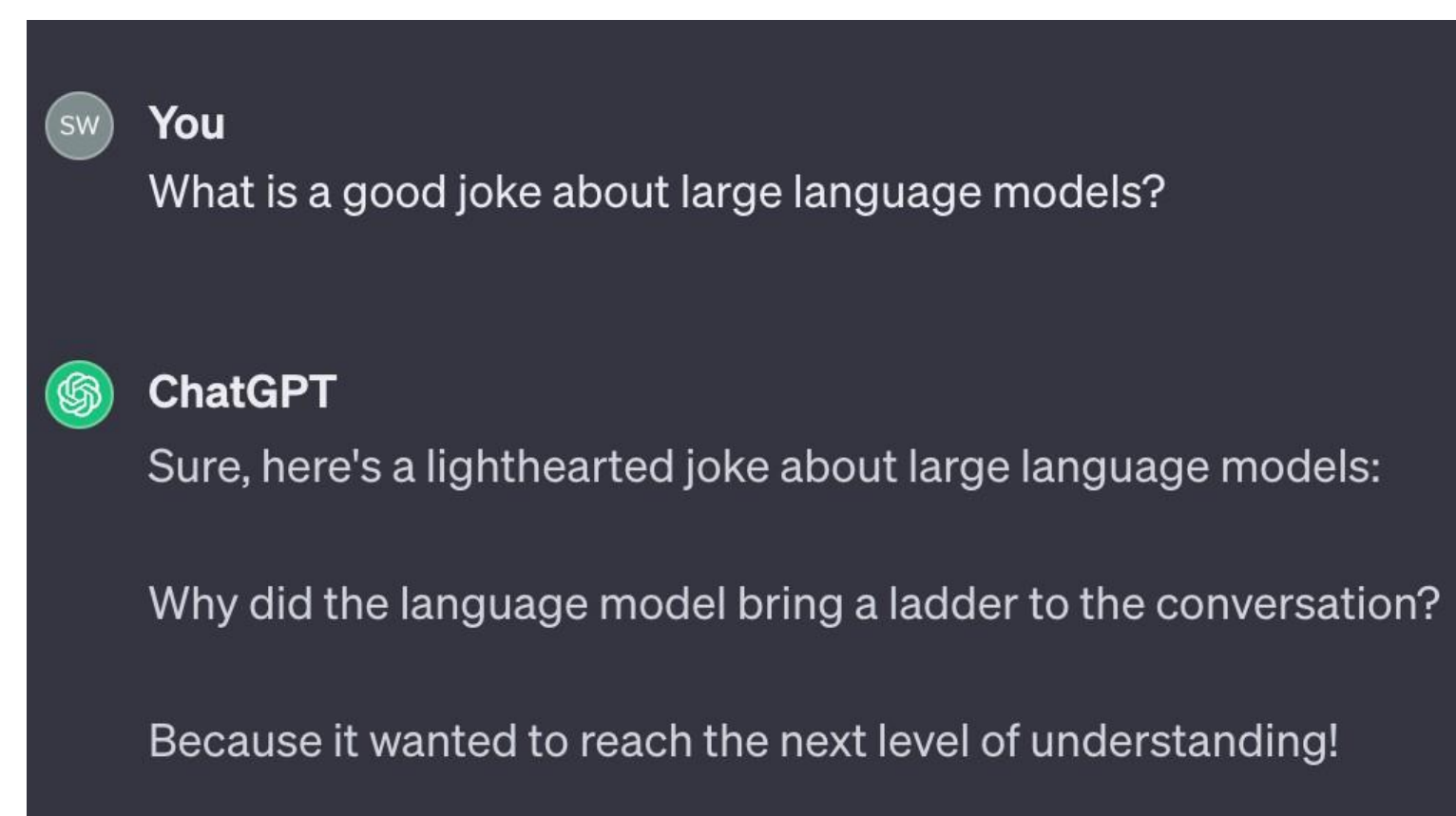
Google Search

- Understands that a fern can be indoor, can be propagated either from seed or from cuttings
- Can find the exact passage in a webpage that answers the questions
- Can find related (in meaning) questions



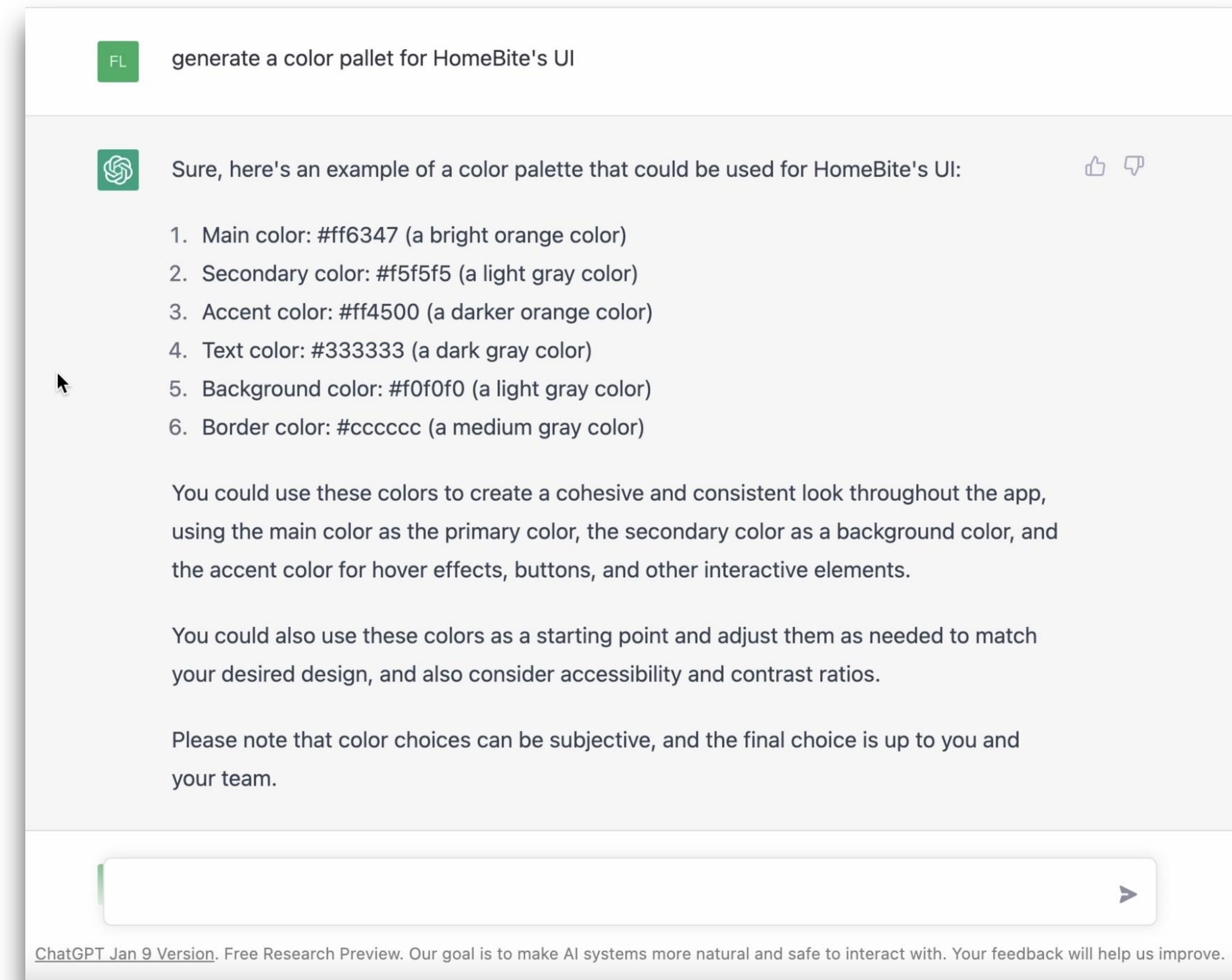
Language Models

- But today, variants of language models are used
 - As chat models
 - As supercharged auto completes
 - As instruction-following assistants
 - And many more...
- It's hard to know exactly how these models might respond, making these models stochastic / probabilistic
 - Contrast this with rule-based systems which respond exactly the same way each time (deterministic systems)

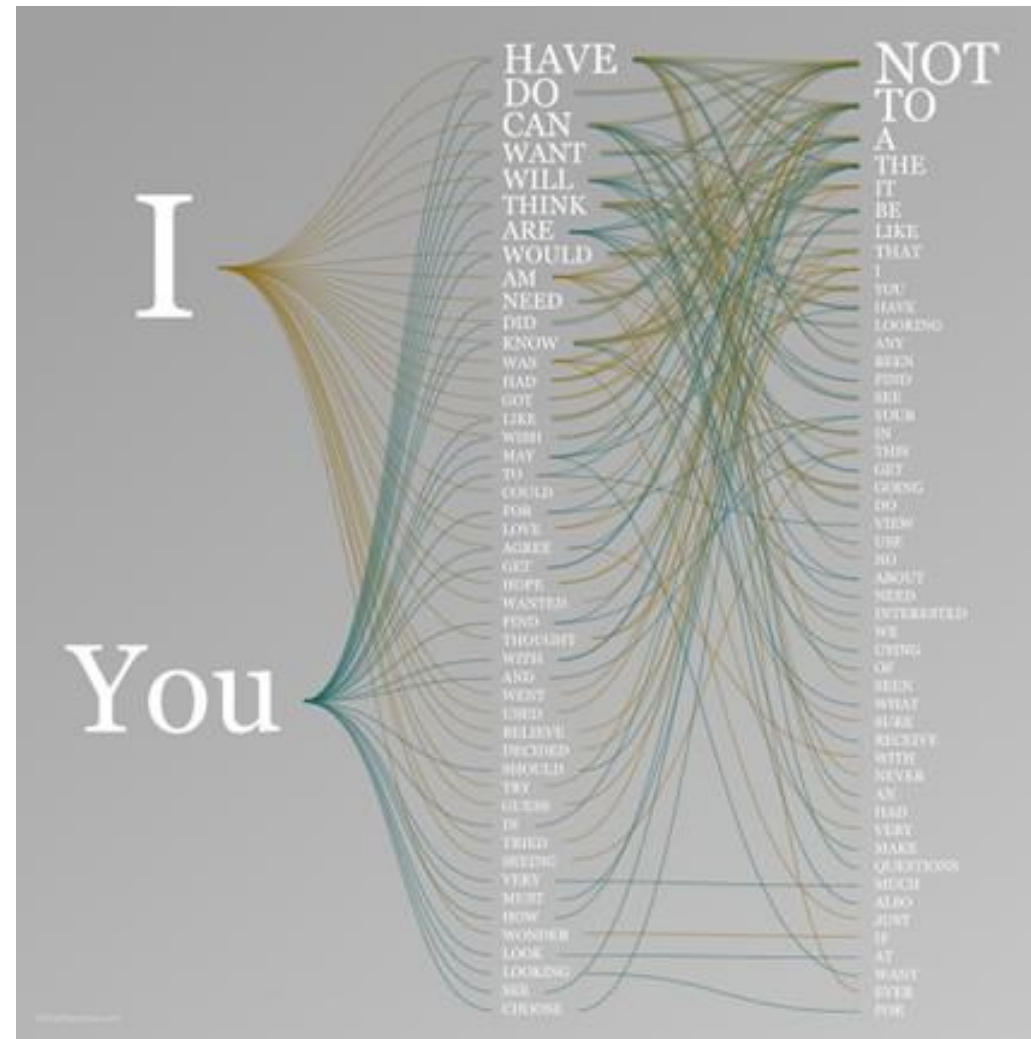


Why should we care about language models?

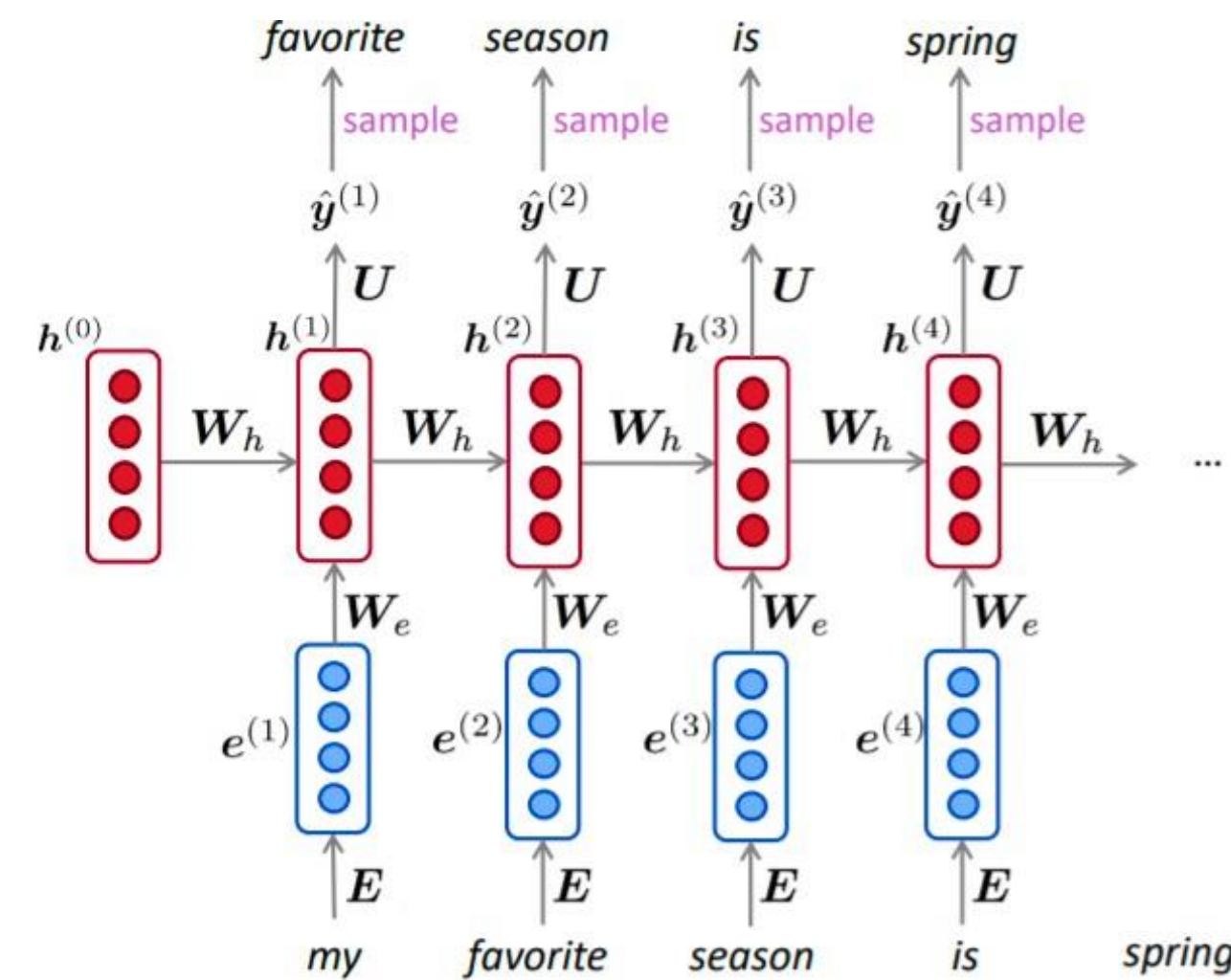
- Foundational models in NLP today
 - Internal representations of models is learned exclusively from language models
 - Direct applications of LLMs are growing rapidly
- Classical importance
- Ubiquitous, so everyone has an opinion!
- Immense societal implications



History of LMs

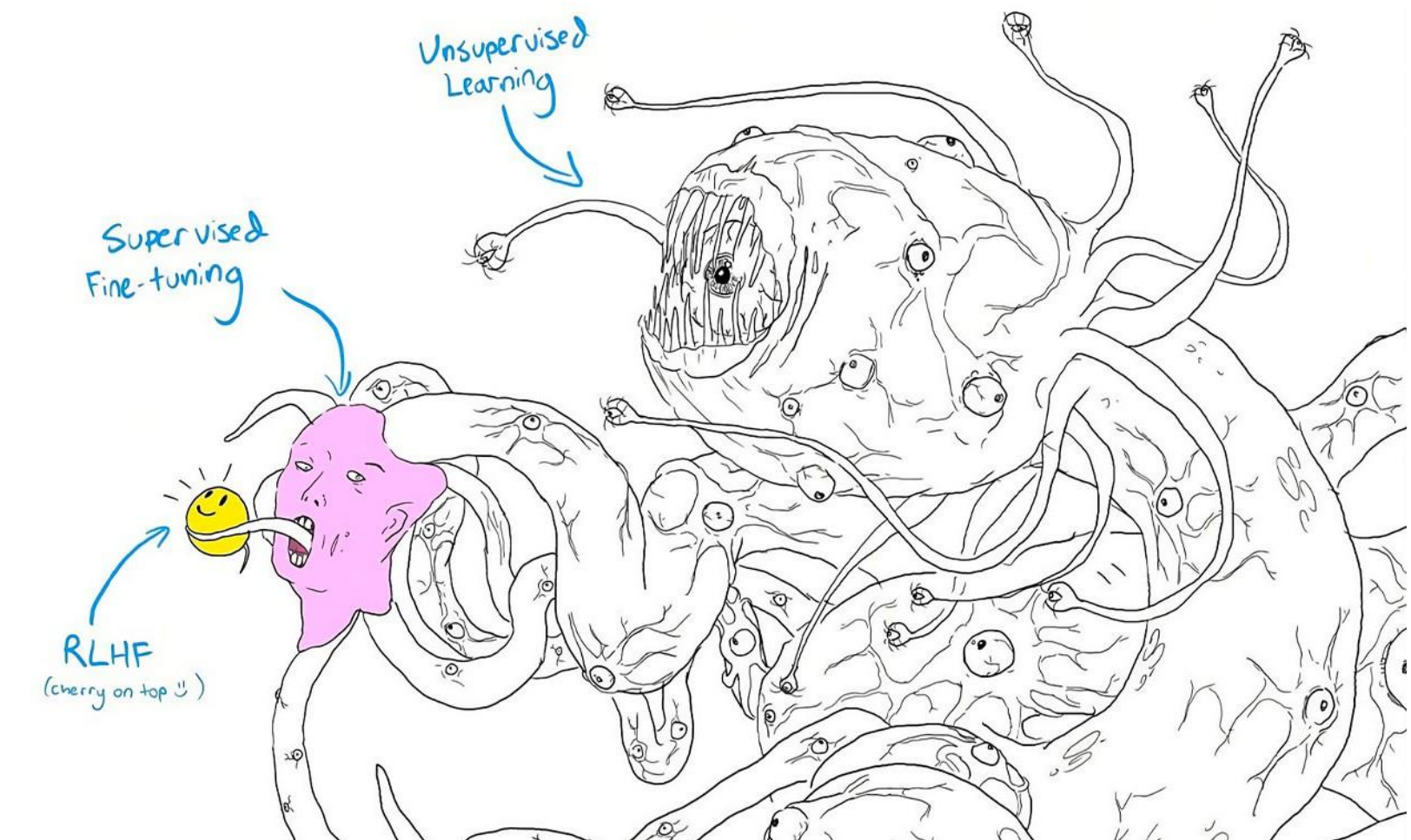
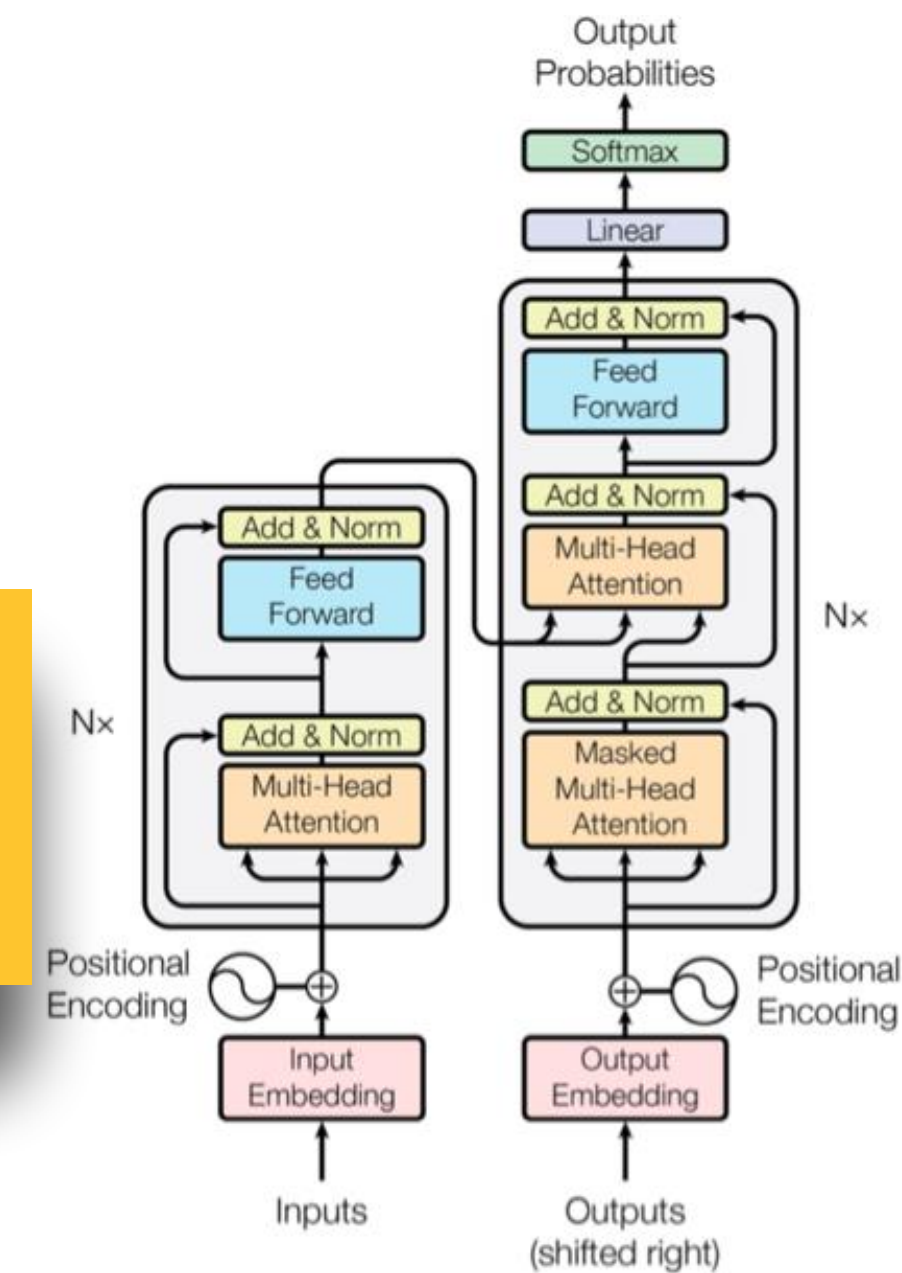


Pre-Neural Age of LMs (-2013)



Early LMs (2013-2018)

Modern LMs (2018-2022)



LMs today (2023-present)

The Pre-Neural Age of LMs (-2013)

- n-gram language models
 - The power of context
 - Smoothing
 - Evaluating LMs

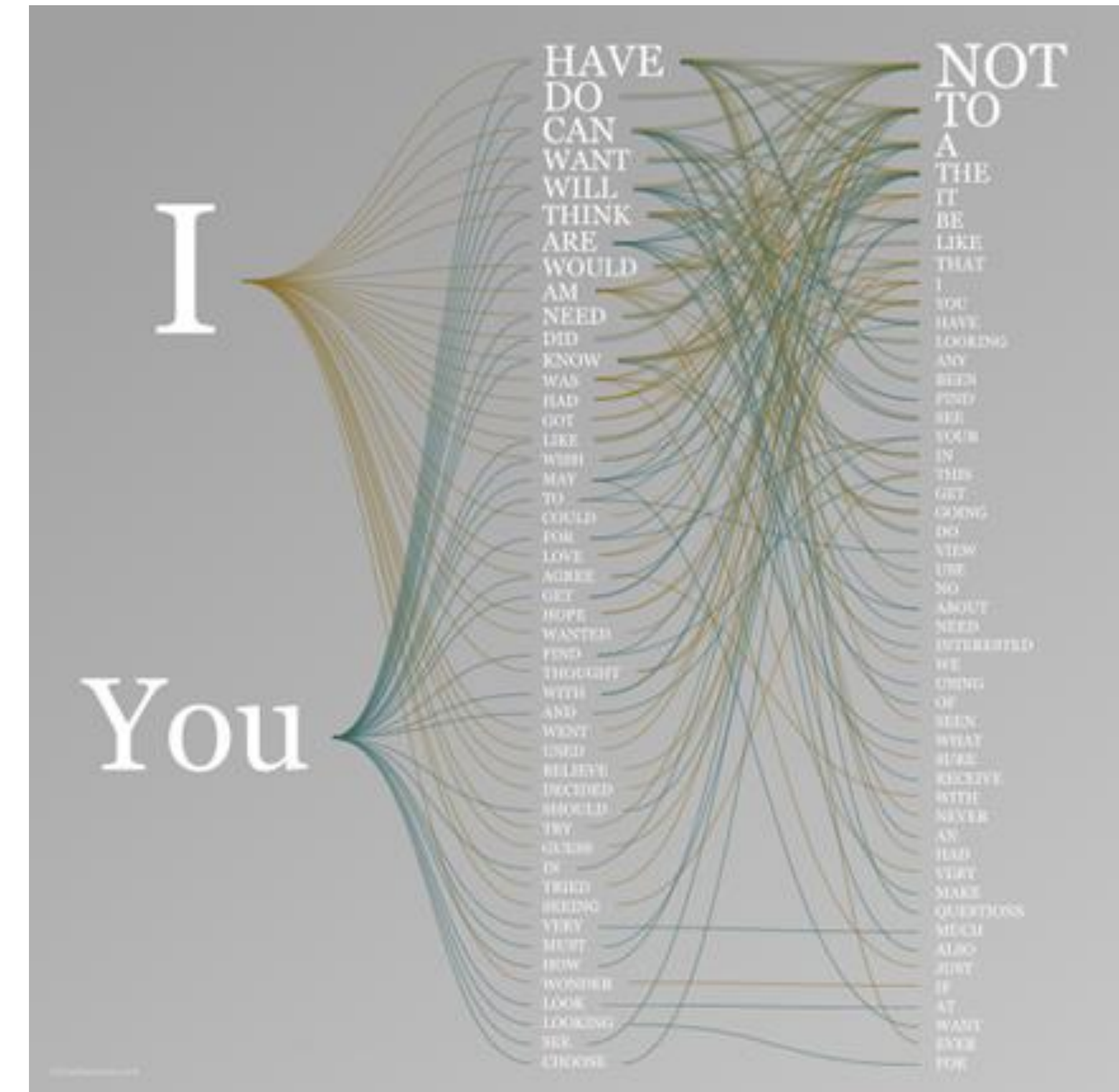
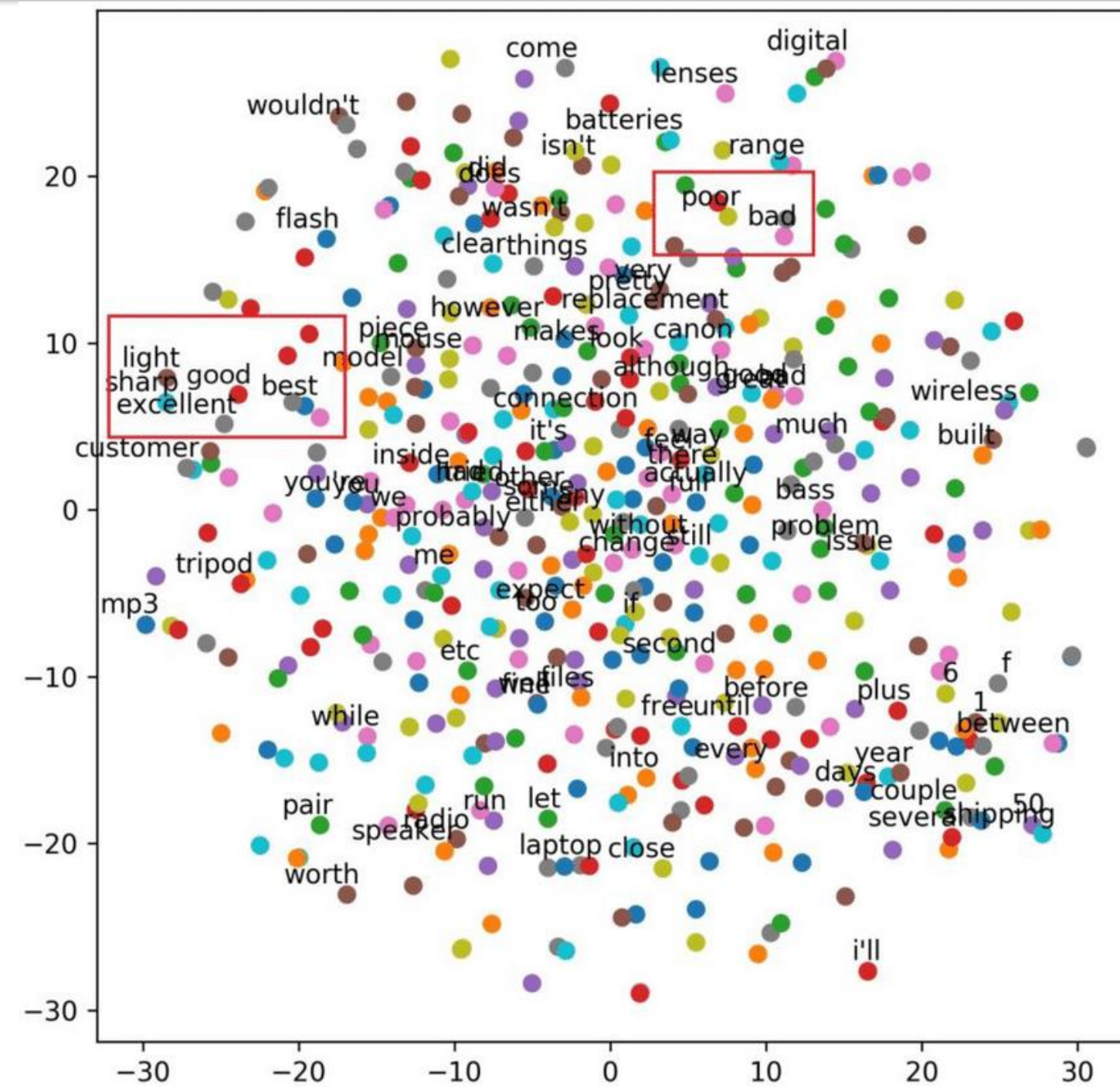
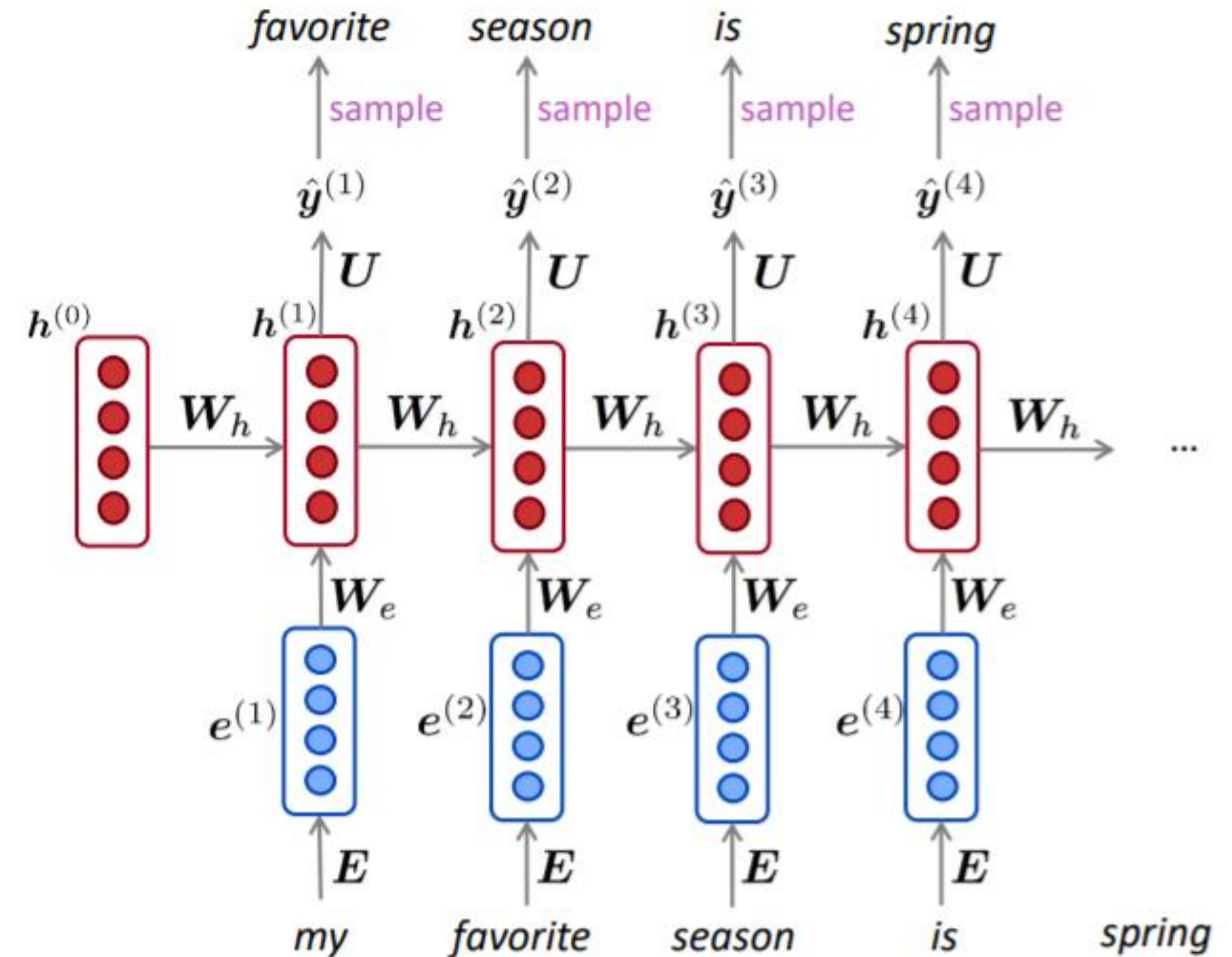


Image Courtesy: Chris Harrison



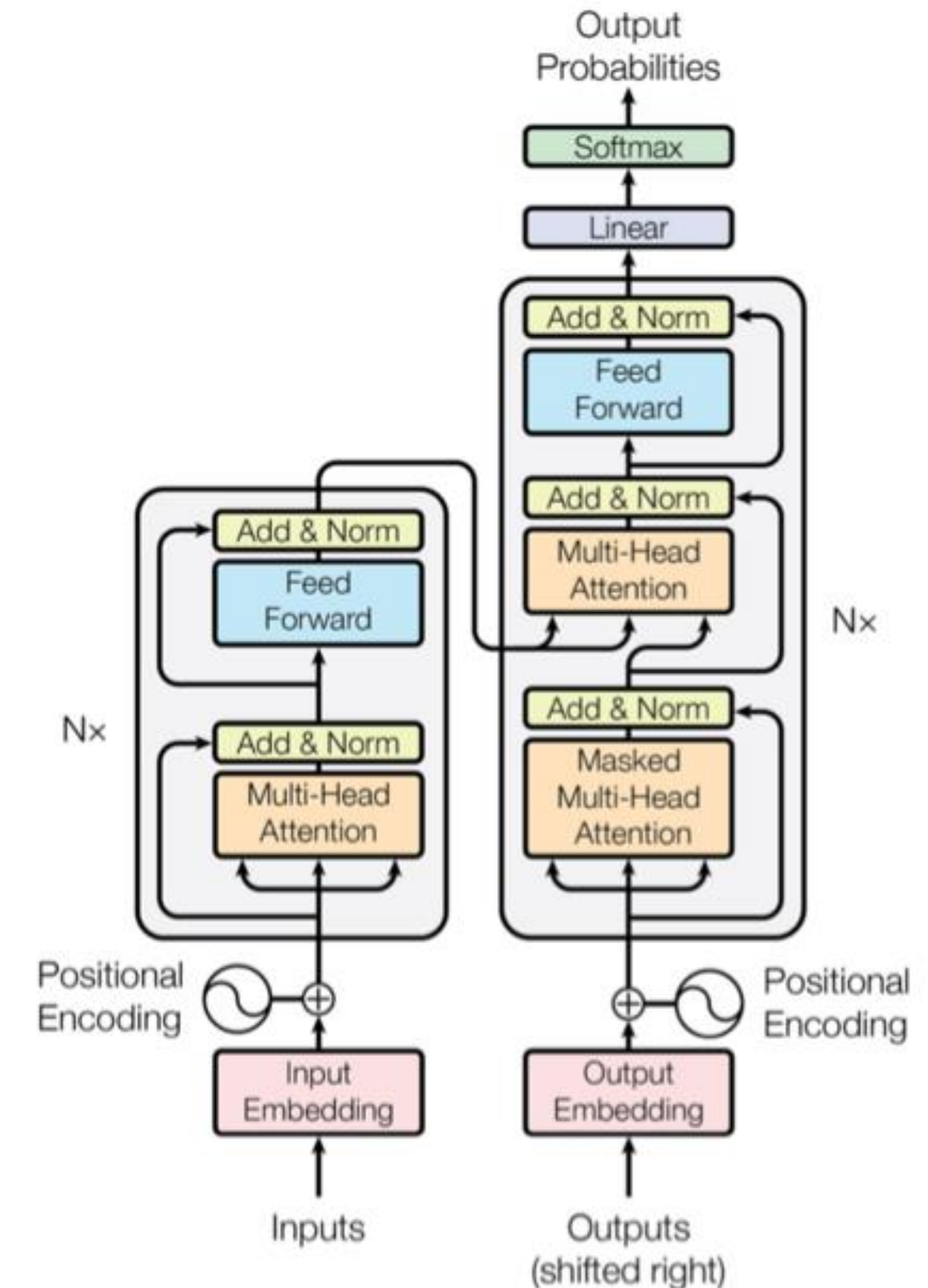
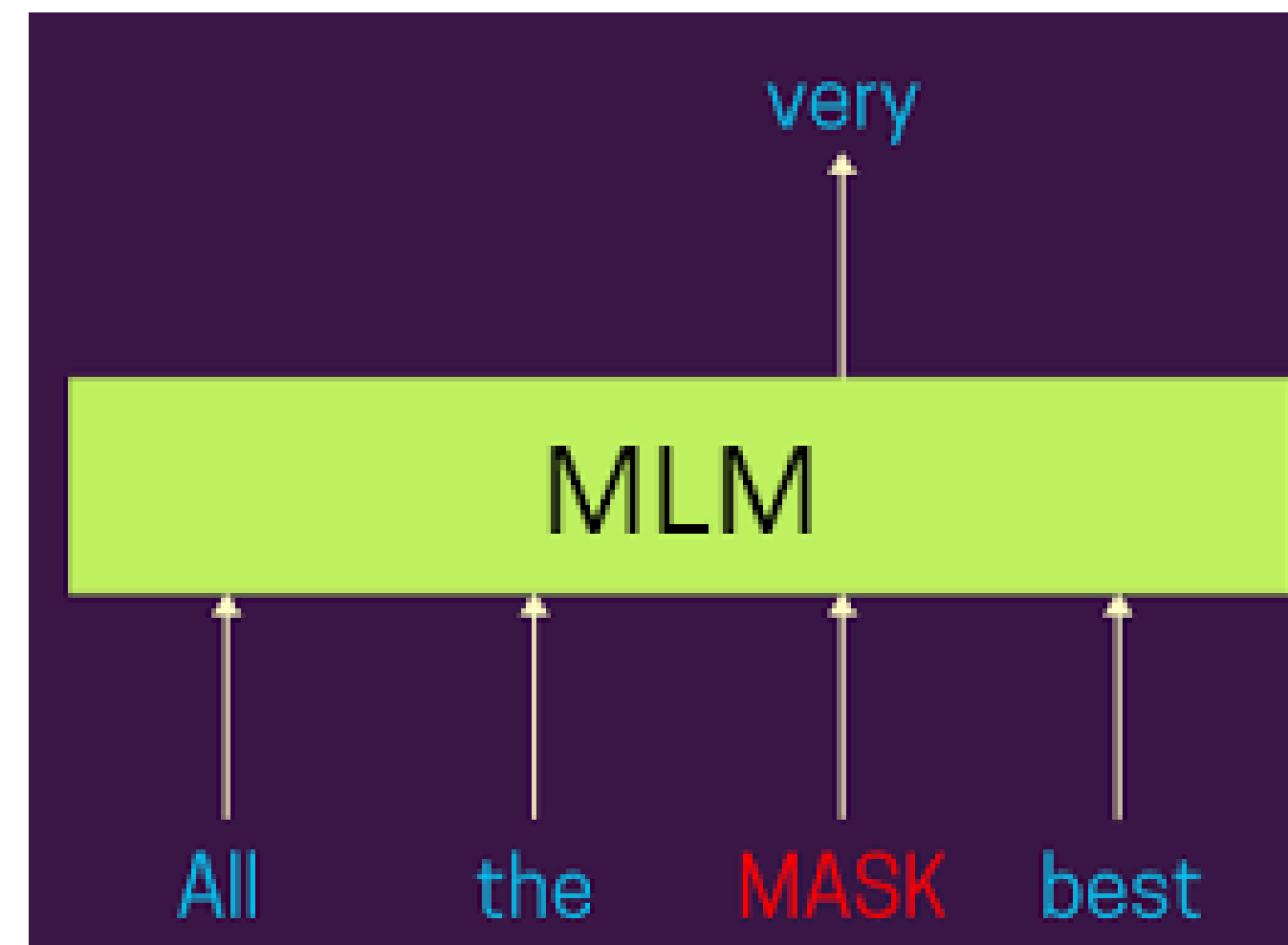
Early Neural LMs (2013-2018)

- Word Embeddings
- Recurrent Neural Net Language Models
- Basics of Neural Nets: Backpropagation



Modern Neural Language Models (2018 - present)

- Transformers
- Feed Forward Neural Nets
- Attention Mechanism
- Masked Language Models

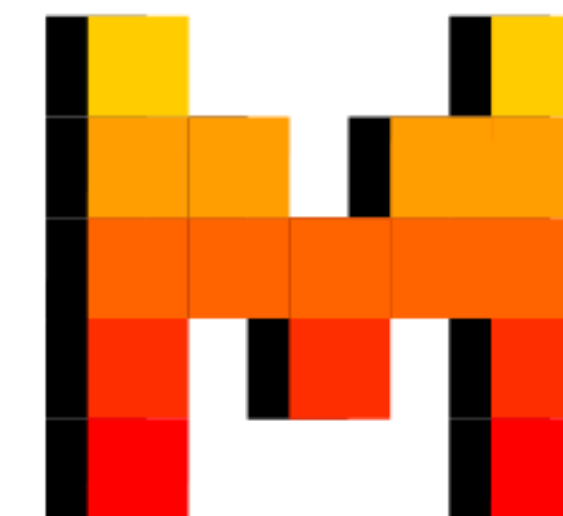


The LLama 3 logo, featuring the text "LLama 3" in white on a dark blue background with a network of glowing blue nodes and lines.

LLama 3

The Claude logo, showing the word "Claude" in white on a blue background, with a stylized orange and white network diagram to the left.

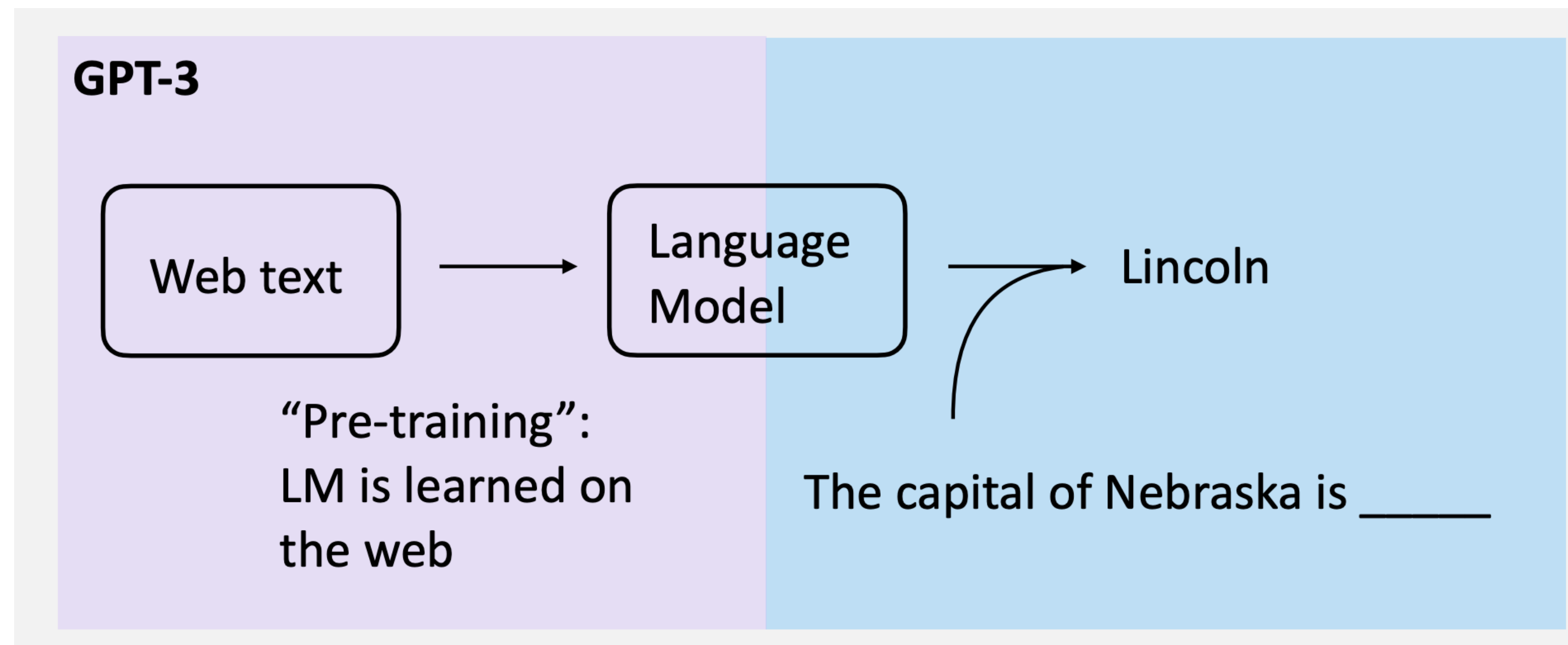
Language models are getting larger (LLMs), trained on large quantities of data, and containing billions of parameters: only a few key players can develop them



MISTRAL
AI_

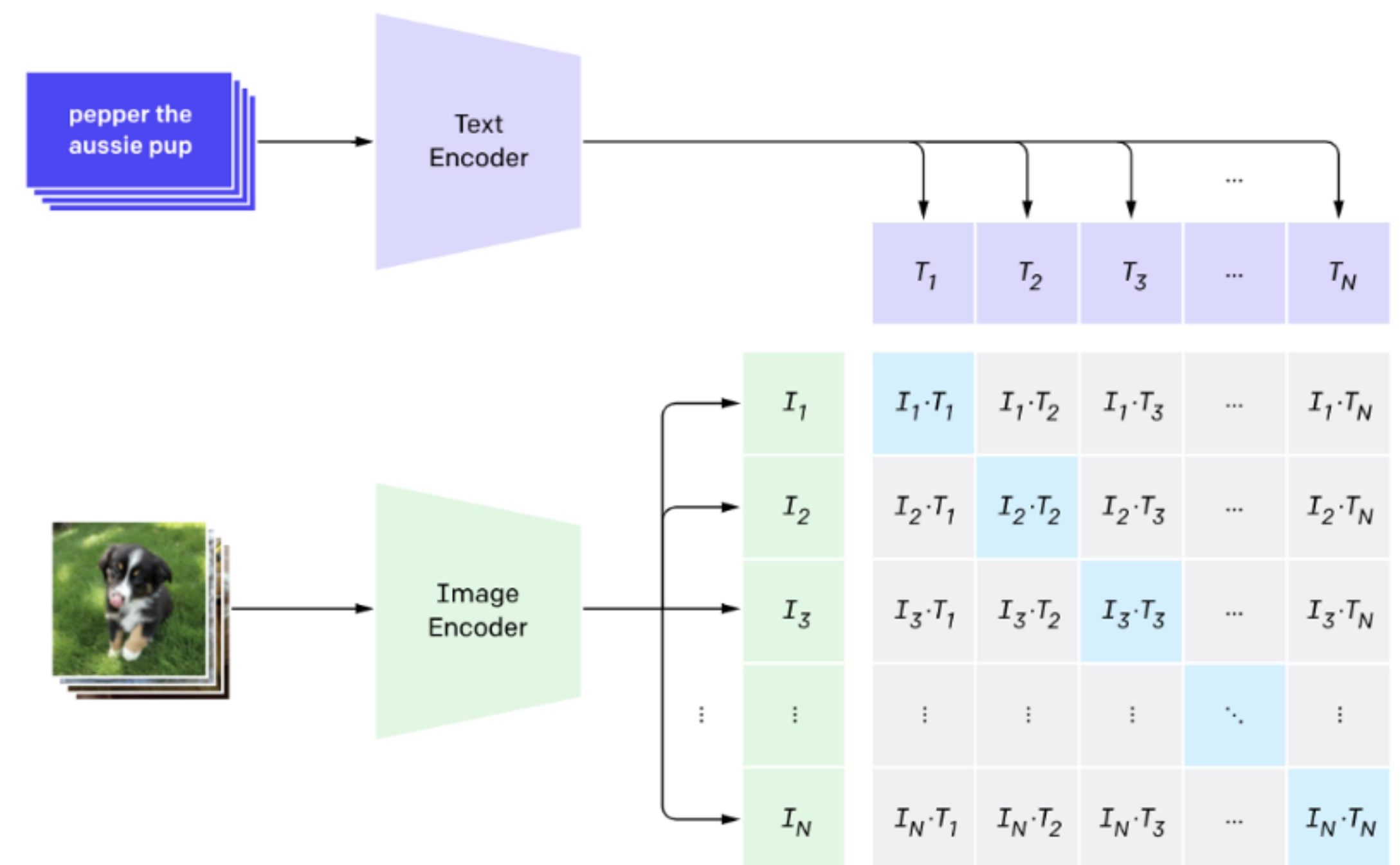
Large Language Models (LLMs)

- These models are trained over a ton of data (a curated scrape of the web). So they will have seen information about Nebraska and Lincoln.
- A big enough model can answer questions even without being trained to do so. What else can we get these models to do?



Large Language Models

- Prompting and Instruction Tuning
- Current Capabilities e.g. code generation
- Outstanding Technical Issues (e.g., hallucinations) and Socio-Technical Issues (e.g., privacy)
- Another kind of “large”
 - Beyond English
 - Beyond the language modality



CLIP from OpenAI

GPT-4 Passes the Bar Exam!

GPT Takes the Bar Exam

December 29, 2022

Michael Bommarito II ^{1,2,3}, Daniel Martin Katz^{1,2,3,*}

1 Illinois Tech - Chicago Kent College of Law (Chicago, IL USA)

2 Bucerius Law School (Hamburg, Germany)

3 CodeX - The Stanford Center for Legal Informatics (Stanford, CA USA)

* Corresponding Author: dkatz3@kentlaw.iit.edu

Abstract

Nearly all jurisdictions in the United States require a professional license exam, commonly referred to as “the Bar Exam,” as a precondition for law practice. To even sit for the exam, most jurisdictions require that an applicant completes at least seven years of post-secondary education, including three years at an accredited law school. In addition, most test-takers also undergo weeks to months of further, exam-specific preparation. Despite this significant investment of time and capital, approximately one in five test-takers still score under the rate required to pass the exam on their first try. In the face of a complex task that requires such depth of knowledge, what, then, should we expect of the state of the art in “AI?” In this research, we document our experimental evaluation of the performance of OpenAI’s TEXT-DAVINCI-003 model, often-referred to as GPT-3.5, on the multistate multiple choice (MBE) section of the exam. While we find no benefit in fine-tuning over GPT-3.5’s zero-shot performance at the scale of our training data, we do find that hyperparameter optimization and prompt engineering positively impacted GPT-3.5’s zero-shot performance. For best prompt and parameters, GPT-3.5 achieves a headline correct rate of 50.3% on a complete NCBE MBE practice exam, significantly in excess of the 25% baseline guessing rate, and performs at a passing rate for both Evidence and Torts. GPT-3.5’s ranking of responses is also highly-correlated with correctness; its top two and top three choices are correct 71% and 88% of the time, respectively, indicating very strong non-entanglement performance. While our ability to interpret these results is limited by nascent scientific understanding of LLMs and the proprietary nature of GPT, we believe that these results strongly suggest that an LLM will pass the MBE component of the Bar Exam in the near future.

	GPT	GPT Top 2	GPT Top 3	NCBE
Evidence	63%	84%	98%	65%
Torts	62%	72%	93%	71%
Civil Procedure	52%	63%	79%	59%
Constitutional Law	49%	67%	87%	72%
Real Property	45%	72%	85%	65%
Contracts	45%	77%	86%	70%
Criminal Law & Procedure	35%	62%	86%	71%

AVERAGE	50%	71%	88%	68%
---------	-----	-----	-----	-----

Table 2. Summary of performance by question category for GPT-3.5 and NCBE-Reported Students

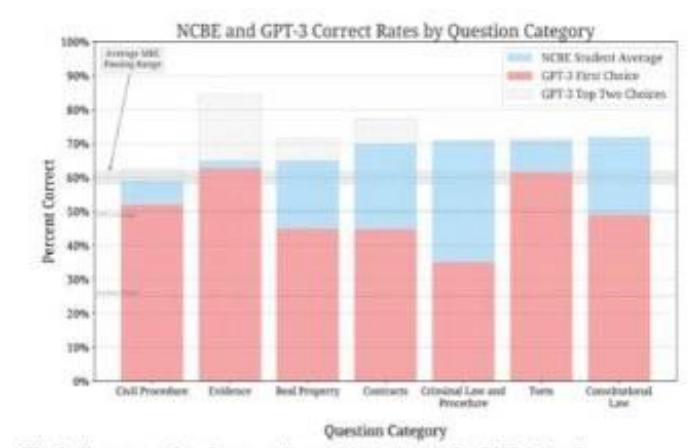


Fig 1. Summary of performance by question category for GPT-3.5 and NCBE-Reported Students

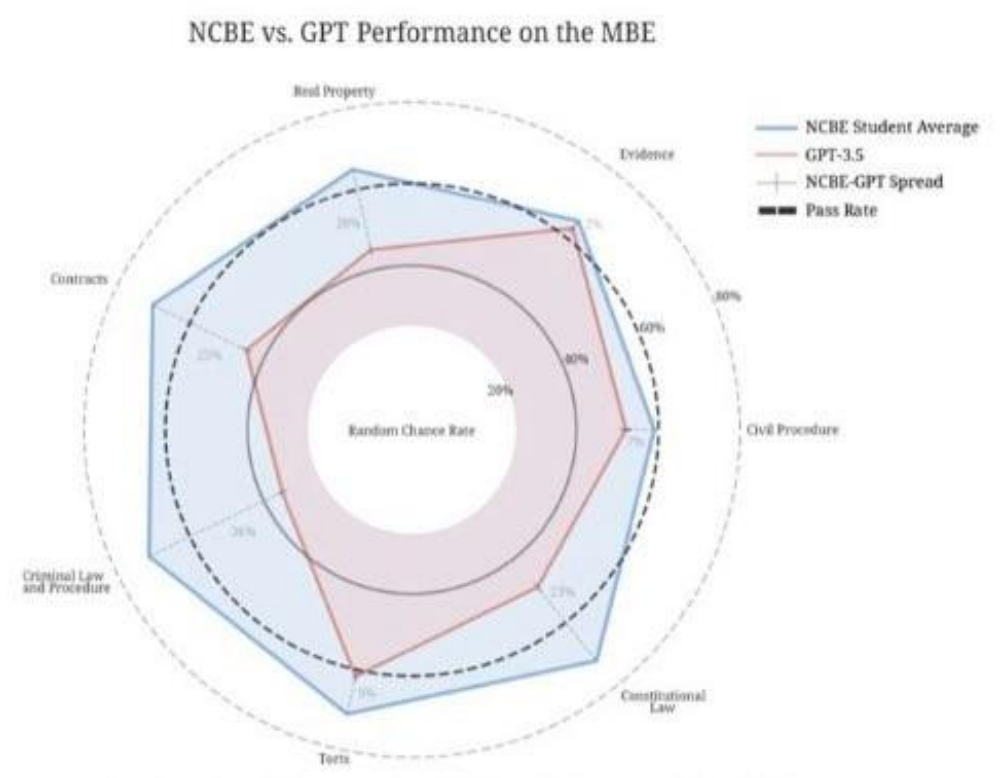
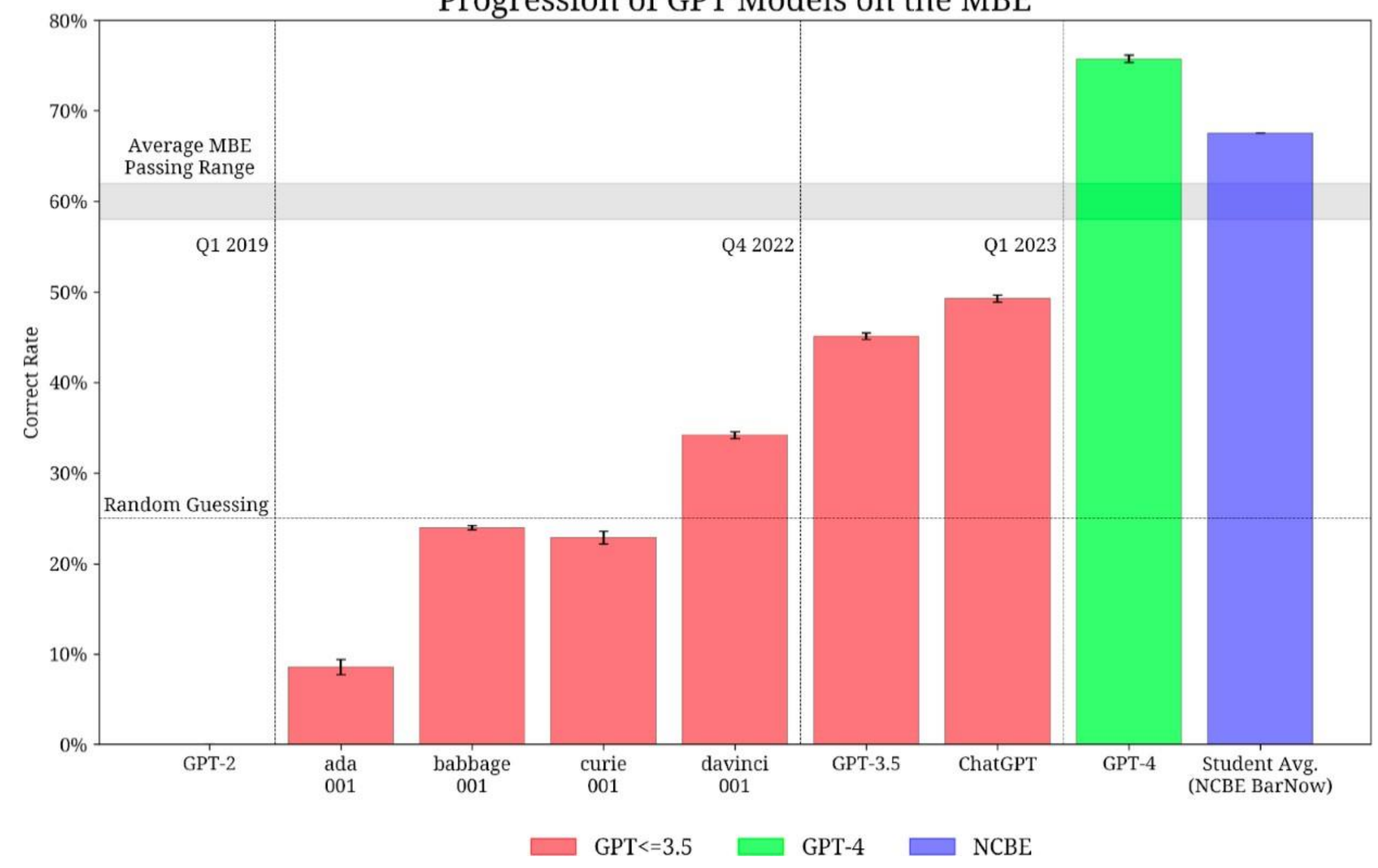


Fig 2. Accuracy by Question Category for GPT and Average Test-Takers

Progression of GPT Models on the MBE



GPT<=3.5 GPT-4 NCBE

Why does this work?

GPT Takes the Bar Exam

December 29, 2022

Michael Bommarito II^{1,2,3}, Daniel Martin Katz^{1,2,3,*}

¹ Illinois Tech - Chicago Kent College of Law (Chicago, IL USA)

² Bucerius Law School (Hamburg, Germany)

³ CodeX - The Stanford Center for Legal Informatics (Stanford, CA USA)

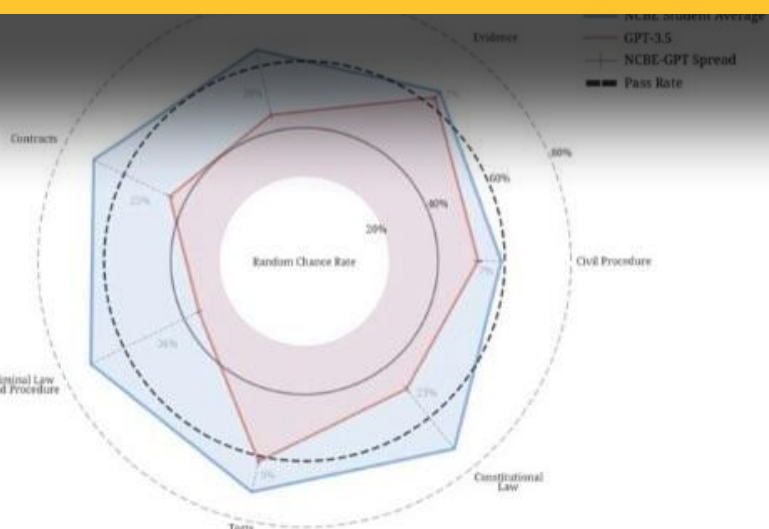
* Corresponding Author: dkatz3@kentlaw.iit.edu

Abstract

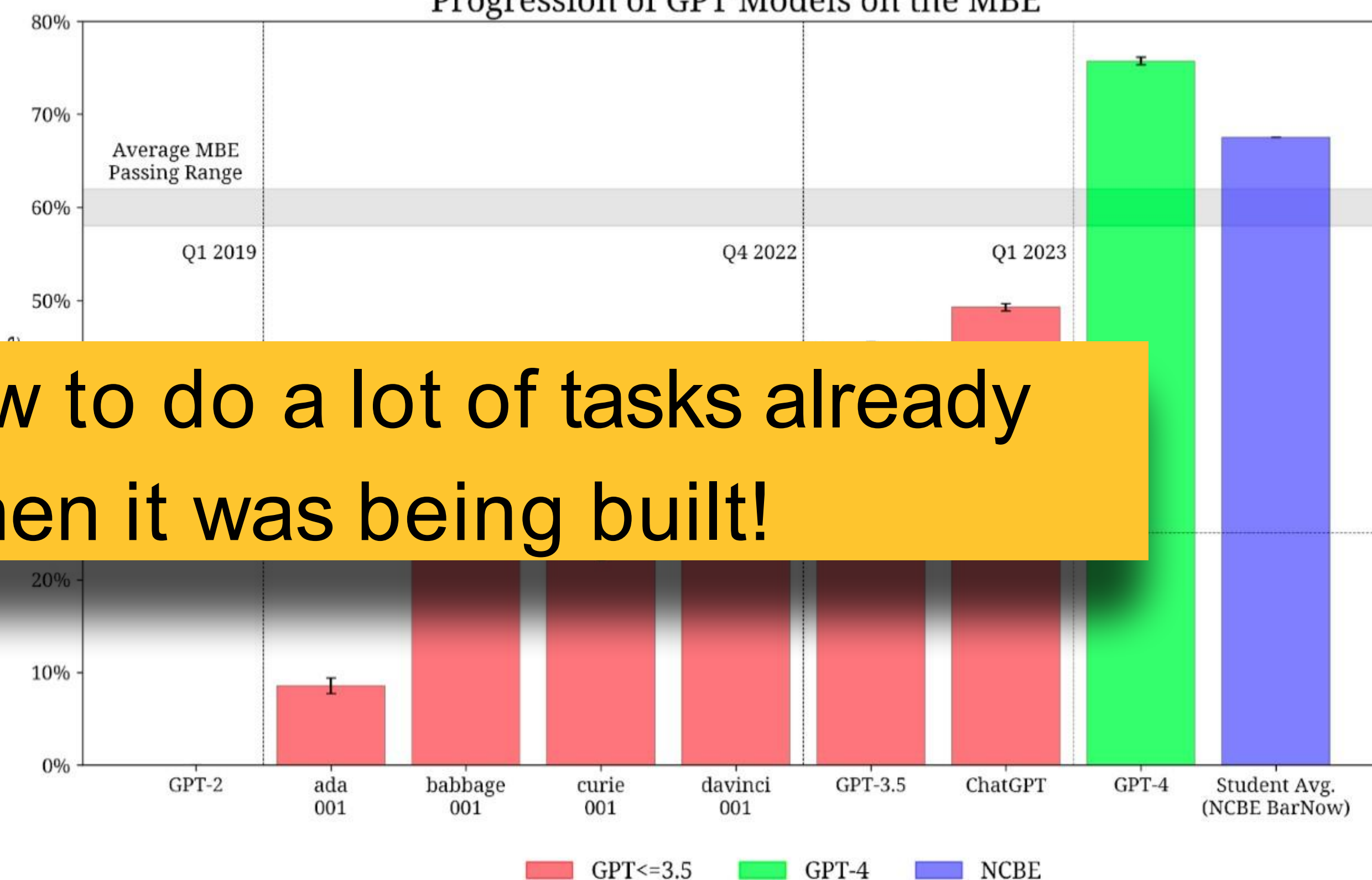
Nearly all jurisdictions require a passing grade on the Multistate Bar Examination (MBE) for the admission to the bar. The MBE is a standardized test of legal knowledge and skills. In the face of this, we expect of experimental, often-referred-to as “AI,” models. While the scale of engineering parameters, GPT-3.5’s performance on the MBE practice exam, significantly exceeds the 50% random chance rate. GPT-3.5 performs at a passing rate for both Evidence and Torts. GPT-3.5’s ranking of responses is also highly-correlated with correctness; its top two and top three choices are correct 71% and 88% of the time, respectively, indicating very strong non-entailment performance. While our ability to interpret these results is limited by nascent scientific understanding of LLMs and the proprietary nature of GPT, we believe that these results strongly suggest that an LLM will pass the MBE component of the Bar Exam in the near future.

	GPT	GPT Top 2	GPT Top 3	NCBE
Evidence	63%	84%	98%	65%
Torts	62%	72%	93%	71%
Civil Procedure	52%	63%	79%	59%
Constitutional Law	49%	67%	87%	72%
Real Property	45%	72%	85%	65%
Contracts	45%	77%	86%	70%
Criminal Law & Procedure	35%	62%	86%	71%
AVERAGE	50%	71%	88%	68%

Table 2. Summary of performance by question category for GPT-3.5 and NCBE-Reported Students



Progression of GPT Models on the MBE



The model has seen how to do a lot of tasks already when it was being built!



AR-LLMs can pass the bar exam, medical licensing & MBA exams.
But on the IIT entrance exams they perform badly on chemistry, horribly
on physics, and terribly on math.
They are good with rote learning & fluency but bad with building mental
models & reasoning with them.

 **Daman Arora** @amuseddaman · Apr 15, 2023

Sparks of AGI? @Cinnabar233 and I decided to put this to test and evaluate
GPT models on one of the toughest exams in the world: the JEE Advanced. It is
held annually for admissions to the IITs and other top Engg colleges in India. 1/n

2:46 AM · Apr 17, 2023 · 11M views

Memorization vs. generalization

But LLMs are far from perfect

The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he “did not comprehend” that the chat bot could lead him astray.

Share full article



A.I.-Generated Content Discovered on News Sites, Content Farms and Product Reviews

The findings in two new reports raise fresh concerns over how artificial intelligence may transform the misinformation landscape.

Hallucination leading to misinformation

An A.I. Hit of Fake ‘Drake’ and ‘The Weeknd’ Rattles the Music World

A track like “Heart on My Sleeve,” which went viral before being taken down by streaming services this week, may be a novelty for now. But the legal and creative questions it raises are here to stay.

Give this article



This Tool Could Protect Artists From A.I.-Generated Art That Steals Their Style

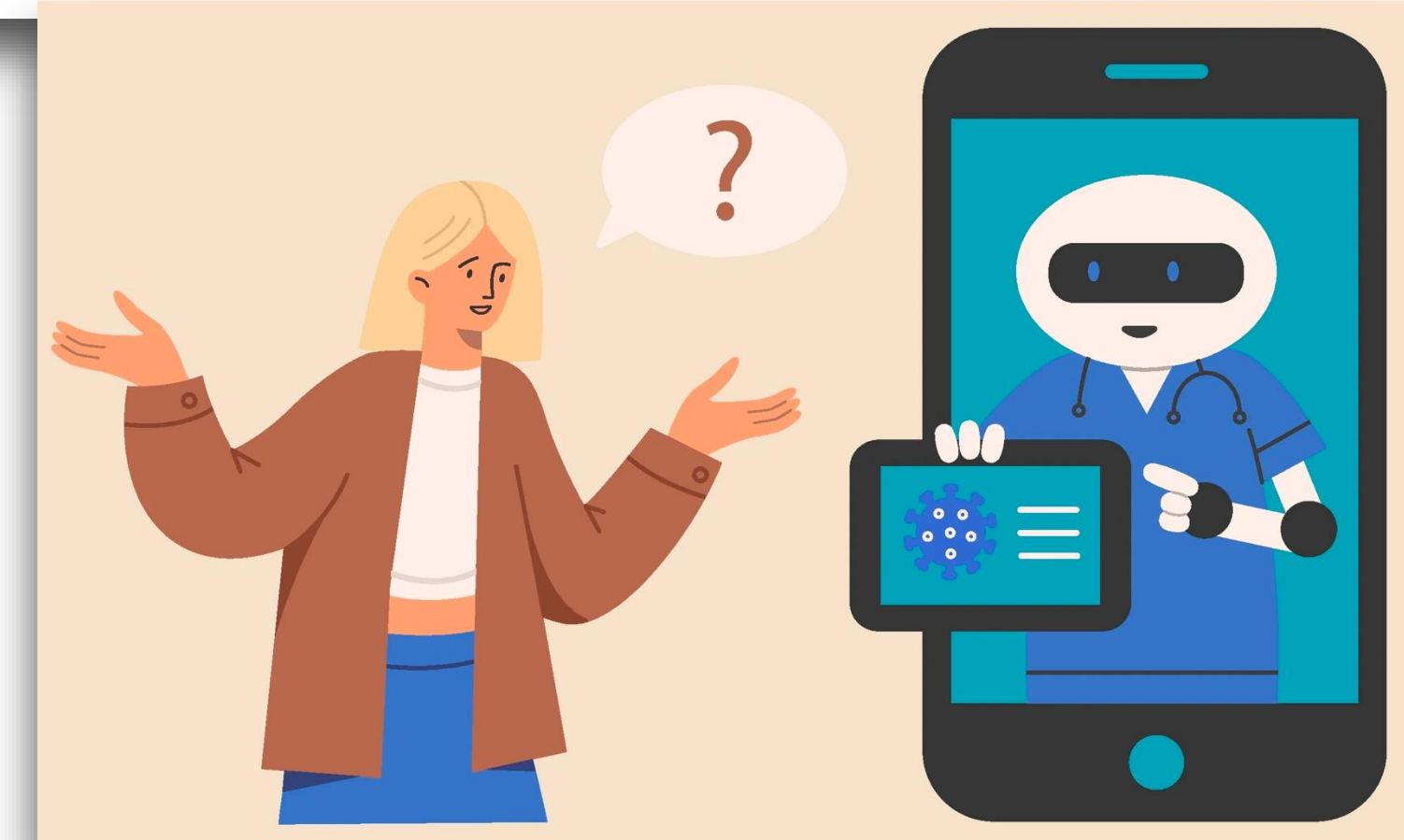
Artists want to be able to post their work online without the fear “of feeding this monster” that could replace them.

Privacy and Copyright Issues

Mar 8, 2023 - Technology

Chatbot therapy, despite cautions, finds enthusiasts

Peter Allen Clark



Ethical Issues and Biases

In this class, you'll learn the fundamental concepts and practical techniques underlying systems like these!

Concrete Outcomes

- Fundamentals of language modeling
- Build on a language model - project
- Learn the connections between this language model and models such as OpenAI's ChatGPT / GPT-4 models
- Current capabilities and outstanding issues with LLMs
- Exciting new problems



In this class, we ask:

What goes into building the state-of-the-art NLP systems that **work well at some tasks**?

Where and why do current state-of-the-art NLP systems still **struggle**?

How can we **make appropriate improvements** and **achieve whatever we want to do with NLP**?

What the class will NOT cover

- Detailed discussion on NLP classification tasks, e.g. question answering
- Classical NLP algorithms for structured prediction
 - e.g. logical semantics and lambda calculus
 - sequence tagging tasks
- In-depth discussion of linguistics

Discourse

Semantics

Syntax: Constituents

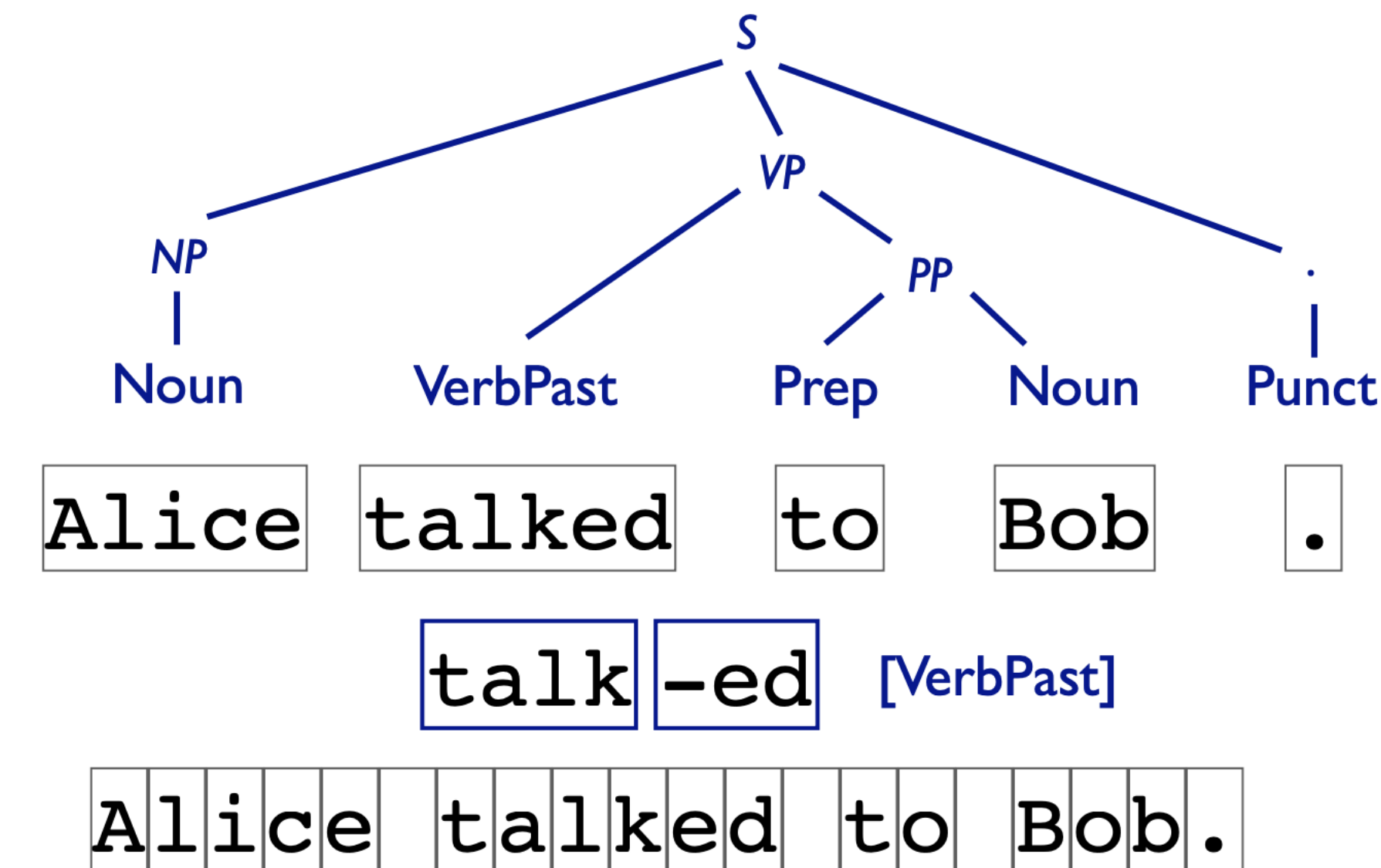
Syntax: Part of Speech

Words

Morphology

Characters

CommunicationEvent(e)	SpeakerContext(s)
Agent(e, Alice)	TemporalBefore(e, s)
Recipient(e, Bob)	



Class Logistics



Instructor, Website and Students



Instructor: Xiang Ren

xiangren@usc.edu

Office Hours: Fri 9-9:45 AM Location GCS SB5 (lower level 2)



TA: Jake Bremerman

jbrem@usc.edu

Office Hours: Fri 9-9:45 AM Location GCS SB5 (lower level 2)

Instructor, Website and Students



Instructor: Xiang Ren

xiangren@usc.edu

Office Hours: Fri 9-9:45 AM Location GCS SB5 (lower level 2)

- Associate Professor, CS, USC
- Spent time at AI2, Snapchat, Stanford, Sahara AI
- PhD in Computer Science, University of Illinois Urbana Champaign
- Research interests: NLP, large language model reasoning, generalization, model safety and trust, Attribution, memorization, and provenance; Efficient inference time method and scaling
- Personal website: <https://www.seanre.com/>

Instructor, Website and Students



Instructor: Xiang Ren

xiangren@usc.edu

Office Hours: Fri 9-9:45 AM Location GCS SB5 (lower level 2)

Lectures + Readings:

https://inklab.usc.edu/cs662_s26/

Announcements etc. on Brightspace



Deliverables

- Paper Presentations and Discussion - 20%
- Class Project - 65%
- Class Participation - 15%



- 5 total late days for deliverables
 - no fractions
- All work due by 11:59pm PT

Paper Presentation

20% of the grade

Each student will complete **one in-depth research paper presentation (15 mins presentation + 5 mins discussion)** during the semester using slides,

Each student will also serve as a **designated discussion lead for two additional paper presentations.**

Paper Presentation

A list of suggested papers in 2025 will be compiled and shared next week:

Document tabs +

CSCI 662 Paper List ⋮

List of Suggested Papers

1) Beyond Chain-of-Th...

2) Inference-Time Scali...

3) Agentic LLMs: Planni...

Search-o1: Agentic S...

4) Reinforcement Learn...

5) Data Quality, Synthet...

6) Model Forensics, Att...

7) Model Safety & Secu...

Constitutional Classifi...

8) Mechanistic Interpre...

List of Suggested Papers

1) Beyond Chain-of-Thought Reasoning

CoT is no longer treated as the solution; its brittleness, memorization, and verbosity are now central concerns.

- Training Large Language Model to Reason in a Continuous Latent Space (ICLR 2025 submission / OpenReview; updated Feb 2025)
- Reasoning beyond language: A Comprehensive Survey on Latent Chain-of-Thought Reasoning (arXiv 2025)
- Understanding R1-Zero-Like Training: A Critical Perspective
- LIMO: Less is More for Reasoning
- The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity
-

2) Inference-Time Scaling & Adaptive Computation

Training-time scaling is expensive; inference-time scaling is more controllable. Industry pressure for adaptive compute per query.

- Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly

Paper Presentation

The presentation will be assessed on the student's ability to clearly explain:

- the paper's motivation
- technical approach
- key results;

To critically evaluate its assumptions, limitations, and contributions;

To situate the work within the broader LLM and NLP research landscape

Paper Presentation

Students will also be evaluated on *the quality of the discussion they facilitate*:

- through insightful questions
- engagement with peers
- ability to surface open problems and future research directions

both during their own presentation and when leading discussions for others.

The slides of the presentation need to be shared a day before the presentation to the class.

Semester Project

Students will design and carry out a research project that aims to answer a question in natural language processing (Learning Objective O4).

The focus of the class project can be research-focused or application-focused.

A **research-focused project** will develop models and analyze data of an existing problem in NLP, or formulate a new problem altogether.

An **application-focused project** will train (possibly only fine-tuning) and deploy NLP models to new application areas, while not necessarily developing any novel research question to be answered.

Students will leverage tools, concepts, and techniques presented in the class.

Semester Project

- Styled like a research paper
- **Individually**, or Teams of 2 or 3 students
 - We expect to see ~25 groups, based on enrollment
- Grading:
 - 5%: **presentation** of project pitch (3 mins; 1-2 slide) +
 - 5%: [project proposal](#) (1 page) +
 - 10%: **presentation** of midterm progress (5 mins) +
 - 10%: [midterm report](#) (4 pages) +
 - 15%: **presentation** of main findings (13 mins) +
 - 20%: [final report](#) (8 pages)
- **Project team formation due by Jan 28th**

pitch presentation				
5	Feb 11	Project Pitch presentation; Transformer Language Models (cont'd)	J&M, Chap 9;	Project Proposal Due
6	Feb 18	Pre-training and post-training language models; Paper presentation: Cutting Edge Topics in NLP 1	J&M, Chap 8; J&M, Chap 9;	
7	Feb 25	Inference-time methods Paper presentation: Cutting Edge Topics in NLP 2	J&M, Chap 10;	
8	Mar 4	Paper presentation: Cutting Edge Topics in NLP 3	J&M Chap 11;	
9	Mar 11	Midterm project presentation	J&M Chap 12;	
10	Mar 18	Spring Break	Hovy and Spruit	
11	Mar 25	Paper presentation: Cutting Edge Topics in NLP 4		
12	Apr 1	Paper presentation: Cutting Edge Topics in NLP 5	BLOOM BigScience Clip - Radford et al. 2021	Project Midterm Report due
13	Apr 8	Paper presentation: Cutting Edge Topics in NLP 6		
14	Apr 15	Paper presentation: Cutting Edge Topics in NLP 7		Project Presentations Due
15	Apr 22	Semester Project Presentations		Project Presentations Due

Semester Project

Semester Project

Proposal (5%)

The project proposal (1 page, 10pt font size) should outline the type of project (research-focused or application-focused), and then answer the following questions clearly in a sentence and/or a few paragraphs each:

- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What is new in your approach and why do you think it will be successful?
- Who cares?
- If you are successful, what difference will it make?
- What are the risks? What could go wrong, and how will you pivot early on if that happens?
- How much will it cost? That is, what resources will you need in terms of time and computation? Are these reasonable for a semester and what access you have?
- Identify two milestones along the way to your finished project.

Semester Project

Pitch (5%)

1-2 slides to cover short sentence answers to these questions:

- What are you trying to do? Articulate your objectives using absolutely no jargon.
- How is it done today, and what are the limits of current practice?
- What is new in your approach and why do you think it will be successful?
- Who cares? If you are successful, what difference will it make?

Participation

Participation (15%)

To encourage students to explain concepts underlying natural language processing in their own words (Learning Objective O2 & O3), we will evaluate each student's engagements in course discussions during class and through the FAQ and research discussion during student's paper presentations and project presentations.

Class participation

- Attend class - do inform of extenuating circumstances that prevent you from attending
- Ask questions in class, engaged in discussion during others' presentations, and answer instructor questions

Class Project Topics - Examples

[Detoxifying Language Model with Context Distillation](#)

[Forage: ML Generated Recipes](#)

[Learning the Language of Wine](#)

[Haiku Generation with Large Language Models](#)

[Legal-SBERT: Creating a S](#)

[Legal Domain and Genera](#)

Creativity in choosing new and interesting problems often get rewarded!

[tut to English](#)

[al Commentary](#)

[Prompting for Diverse Responses: Making L](#)

[Language Models More Truthful](#)

Office Hours for Project Feedback

[Attribution with Limited Text](#)

[Is it Written?](#)

See more: [Stanford CS224n Projects](#)

See more: [Stanford CS229 Machine Learning](#)

Textbooks

- [Jurafsky and Martin. "Speech and Language Processing." 3rd Ed.](#) This textbook contains chapters on the fundamentals of natural language processing.
- [Eisenstein. "Natural Language Processing."](#) This textbook contains an overview of machine learning approaches for NLP.
- [Goldberg. "Neural Network Methods for Natural Language Processing."](#) This textbook provides a deep learning perspective towards NLP.

Class Introductions

Now your turn!

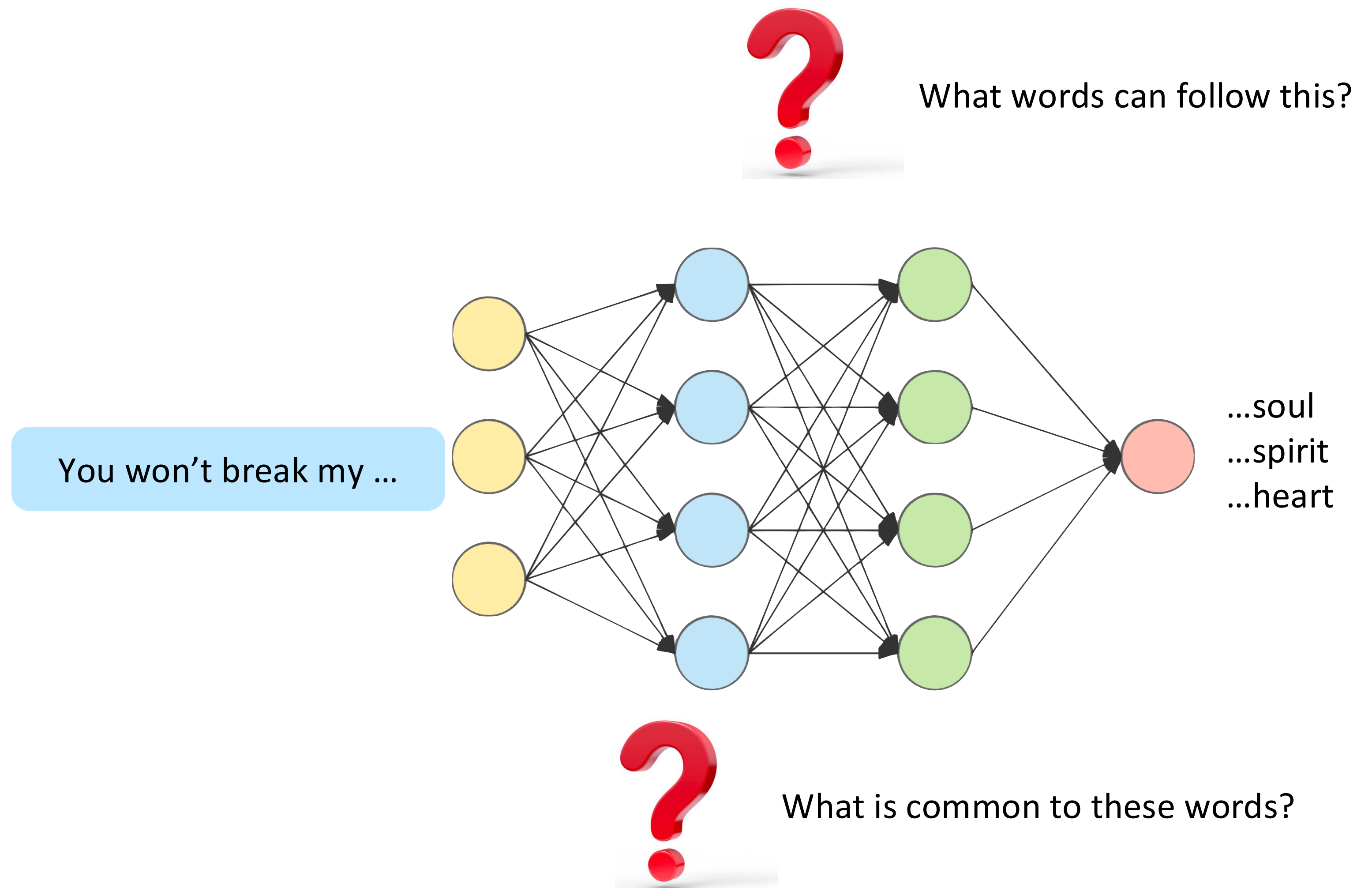
- **Name**
- **Degree Status / Year**
- **Research interests & experiences**
- **Why this class?**
- **One fun fact**

Probabilistic Language Models!

Assign a probability to a sentence

Building a Language Model

- Task: Given a sequence of words so far (the context), predict what comes next
- We never know for sure what comes next, but we can still make good guesses!



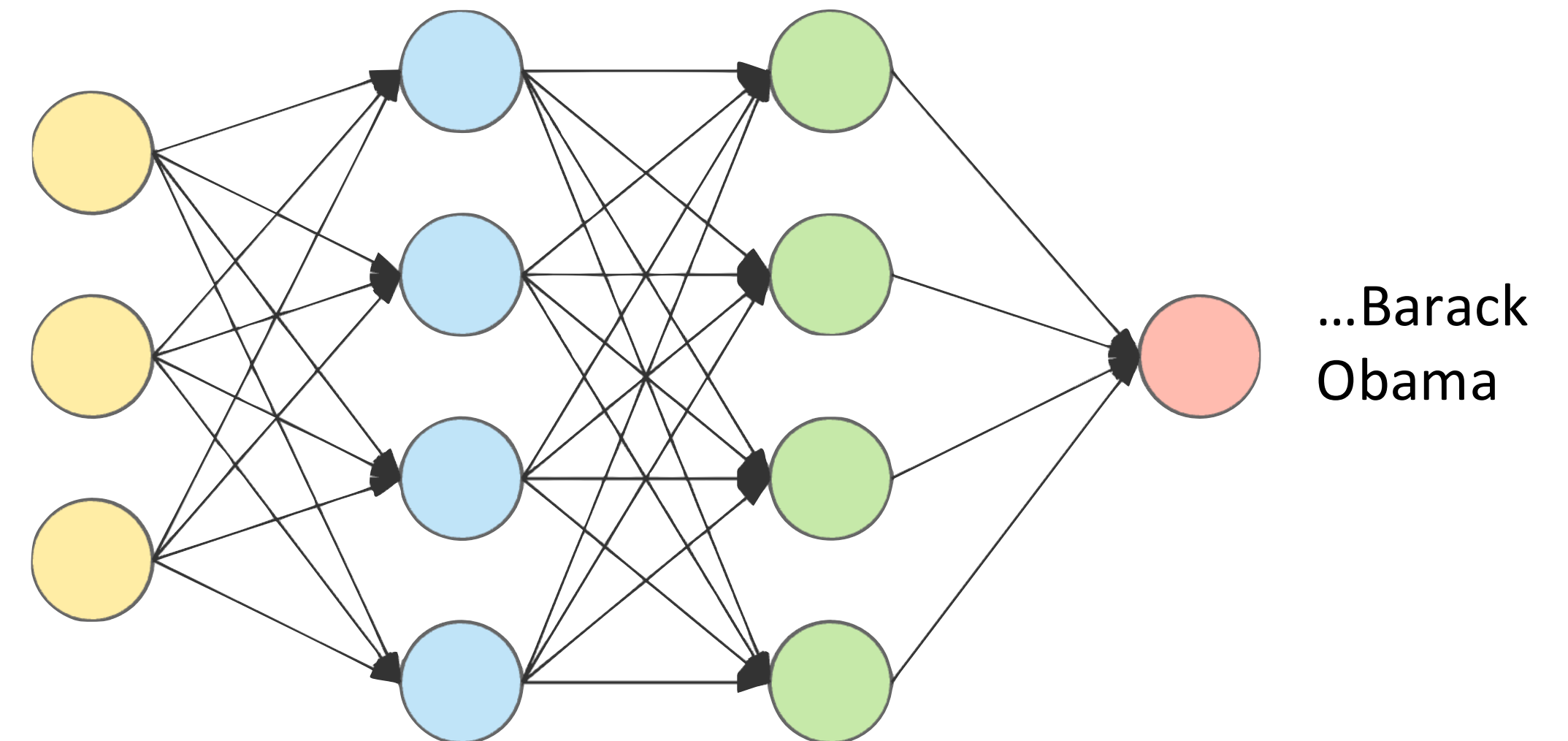
Building a Language Model

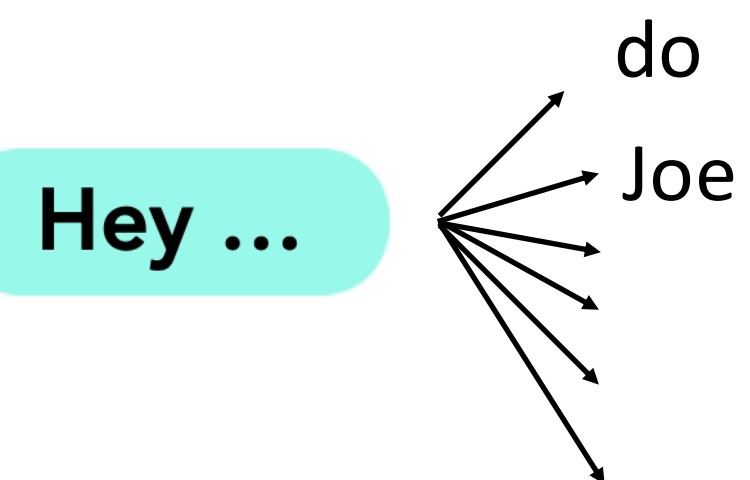
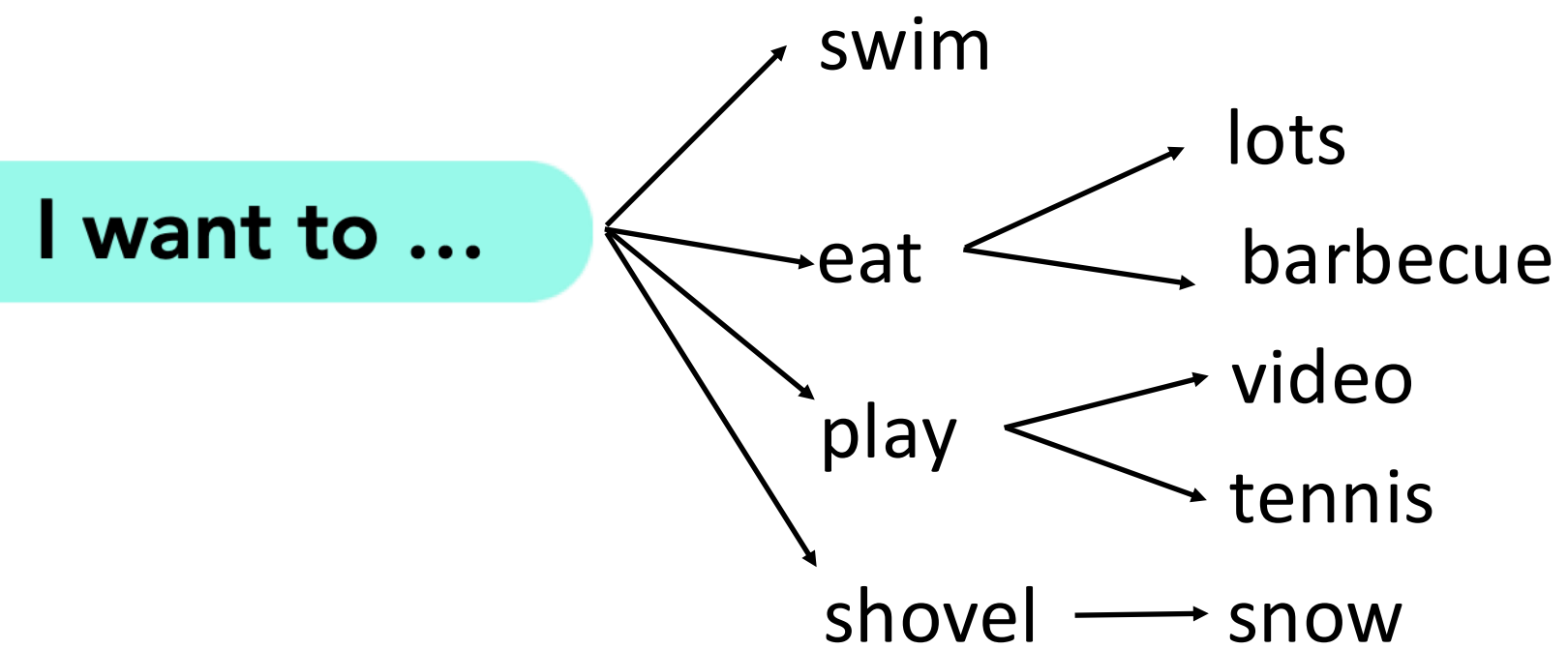


What words can follow this?

- Task: Given a sequence of words so far (the context), predict what comes next
- We never know for sure what comes next, but we can still make good guesses!

The 44th President of United States was ...





Certain sentence constructions are more likely than others, due to grammaticality, obscurity or commonness

Sentences have different probabilities!

The capital of Nebraska is ...

Lincoln

Probabilistic Language Models!

Assign a probability to a sentence

Probabilistic Language Modeling

Goal: compute the probability of a sentence or sequence of words:

$$P(\mathbf{w}) = P(w_1, w_2, w_3, w_4, w_5, \dots w_n)$$



Difference

Related task: probability of an upcoming word:

$$P(w_n | w_1, w_2, w_3, w_4, \dots w_{n-1})$$

A model that assigns probabilities to sequences of words (e.g., either of these: $P(\mathbf{w})$ or $P(w_n | w_1, w_2, \dots w_{n-1})$) is called a language model

How to compute $P(W)$?

“its water is so transparent that you can see the bottom”

➔ $P(\textit{its, water, is, so, transparent, that, you, can, see, the, bottom})$

- How to compute this joint probability, $P(\mathbf{w}) = P(w_1, w_2, w_3, w_4, w_5, \dots w_n)$?
- e.g. $P(\textit{its, water, is, so, transparent, that})$

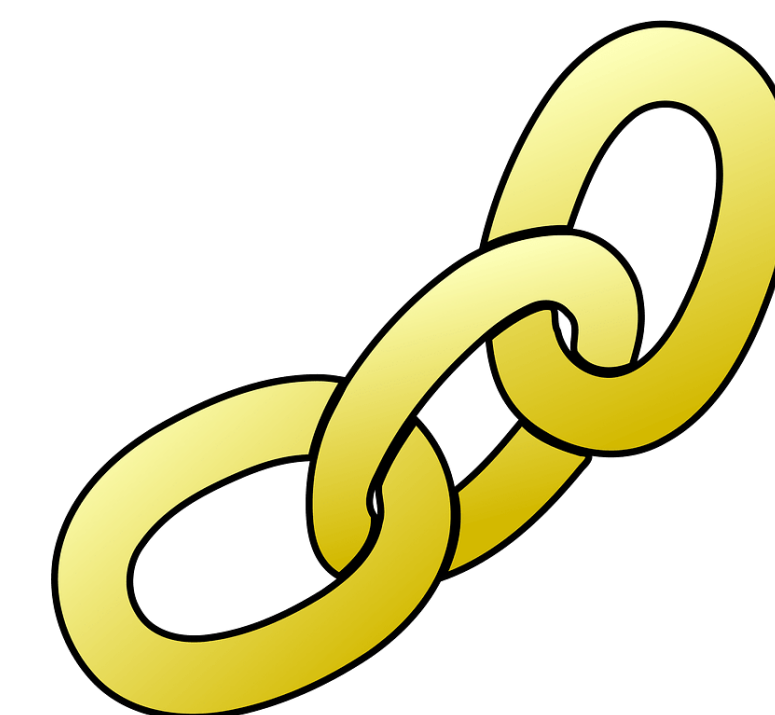
Intuition: let's rely on the Chain Rule of Probability

The Chain-Rule

- Recall the definition of conditional probabilities: $P(B|A) = \frac{P(A,B)}{P(A)}$
- Rewriting: $P(A, B) = P(A)P(B|A)$
- More variables: $P(A, B, C, D) = P(A)P(B|A)P(C|B, A)P(D|C, B, A)$
- The Chain Rule in General

- $$P(x_1, x_2, x_3, \dots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \dots P(x_n|x_1, \dots, x_{n-1})$$

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

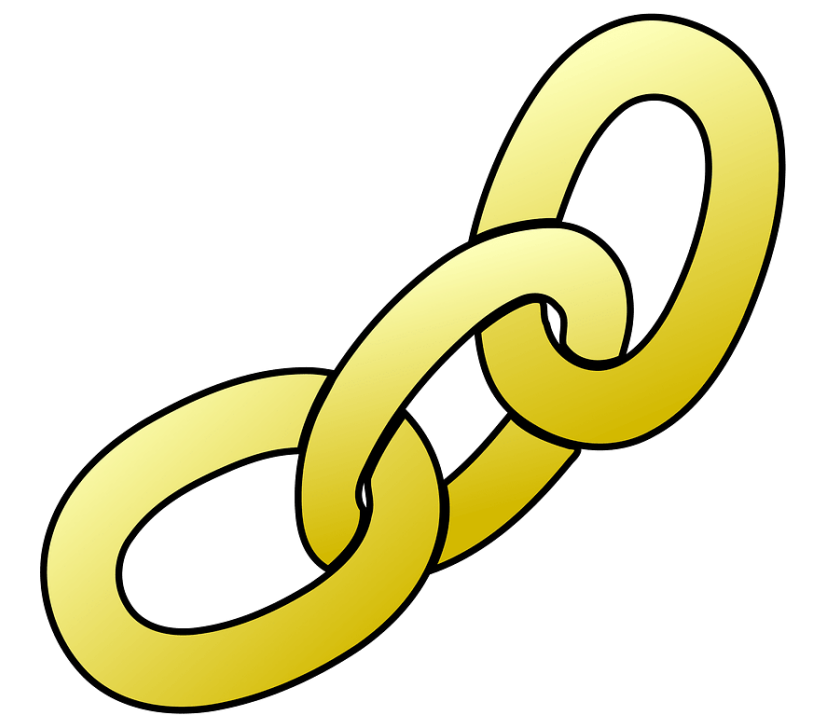


The Chain Rule applied to compute joint probability of words in a sentence

$$P(w_1, w_2, \dots w_n) = \prod_{i=1}^n P(w_i | w_{i-1} \dots w_1)$$

$$\begin{aligned} P(\textit{its water is so transparent}) = & P(\textit{its}) \times \\ & P(\textit{water} | \textit{its}) \times \\ & P(\textit{is} | \textit{its water}) \times \\ & P(\textit{so} | \textit{its water is}) \times \\ & P(\textit{transparent} | \textit{its water is so}) \end{aligned}$$

Ordering matters in
language!



Why Probabilistic Models?

- Why would you want to predict upcoming words, or assign probabilities to sentences?



I will be back soonish

I will be bassoon dish

- Probabilities are essential for language generation
- Any task in which we have to identify words in noisy, ambiguous input, like speech recognition
- For writing tools like spelling correction or grammatical error correction

Your so silly

You're so silly

Everything has improve

Everything has improved

Probabilistic Language Models

Machine Translation:

- $P(\textit{high winds tonight}) > P(\textit{large winds tonight})$

Spell Correction:

- $P(\textit{I'm about fifteen minuets away}) < P(\textit{I'm about fifteen minutes away})$

Speech Recognition:

- $P(\textit{I saw a van}) \gg P(\textit{eyes awe of an})$

Summarization, question-answering, etc., etc.!!

But how to learn these probabilities?

Probability Estimation via Statistical Modeling



Suppose we have a biased coin that's heads with probability p .

As we know, $0 \leq p \leq 1$, and for a normal coin, $p = 0.5$ (equal probability of heads or tails)

Suppose we flip the coin four times and see (H, H, H, T). What is p ?

We don't know what p is — could be 0.5! But $p = 3/4 = 0.75$ maximizes the probability of data sequence (H,H,H,T)



maximum likelihood estimate

The probability of the data is $ppp(1 - p)$: if you take the derivative and set it equal to zero and find $p = 0.75$

How to estimate the probability of the next word?

$$P(\textit{that}|\textit{its water is so transparent}) = \frac{\textit{Count}(\textit{its water is so transparent that})}{\textit{Count}(\textit{its water is so transparent})}$$



Could we just count and divide?

- No! Too many possible sentences!
 - We'll never see enough data for estimating these

Markov Assumption

- Simplifying assumption:

$$P(\textit{that} | \textit{its water is so transparent}) \approx P(\textit{that} | \textit{transparent})$$



Andrei Markov

Or maybe...

$$P(\textit{that} | \textit{its water is so transparent}) \approx P(\textit{that} | \textit{so transparent})$$

Markov Assumption contd.

$$P(w_1, w_2, \dots w_n) = \prod_i P(w_i | w_{i-k} \dots w_{i-1})$$

- In other words, we approximate each component in the product such that it is only conditioned on the previous **k** elements

$$P(w_i | w_1, w_2, \dots w_{i-1}) \approx P(w_i | w_{i-k} \dots w_{i-1})$$

Recap: Probabilistic Modeling

- What is a probabilistic language model?
- How do we simplify it?
- How do we estimate it?
- Why would we need one?
- Next: a simple probabilistic language model



Simplest Case: Unigram model

$$P(w_1, w_2, \dots, w_n) \approx \prod_i P(w_i)$$

Some automatically generated sentences from a unigram model

- fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass
- thrift, did, eighty, said, hard, 'm, july, bullish
- that, or, limited, the

Bigram Model

Condition on the previous word:

$$P(w_i | w_1, w_2, \dots, w_{i-1}) \approx P(w_i | w_{i-1})$$

Some automatically generated sentences from a bigram model

- texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen
- outside, new, car, parking, lot, of, the, agreement, reached
- this, would, be, a, record, november

N-gram Models

- We can extend to trigrams, 4-grams, 5-grams, ...
- In general this is an insufficient model of language
 - “The computer which I had just put into the machine room on the fifth floor crashed.”
- But we can often get away with N-gram models



language has long-distance dependencies

n-gram Language Models

Simplest probabilistic model

n-gram Language Model

The decision for what words occur after a word w is exactly the same as the biased coin, but with *many* possible outcomes (as many as all the words) instead of 2

I like to **eat** cake but I
want to **eat** pizza right
now. Mary told her
brother to **eat** pizza too.

$P(\text{next word} = \textit{pizza} \mid \text{previous word} = \textit{eat}) = 2/3$
 $P(\text{next word} = \textit{cake} \mid \text{previous word} = \textit{eat}) = 1/3$
All other next words = 0 probability

Vocabulary

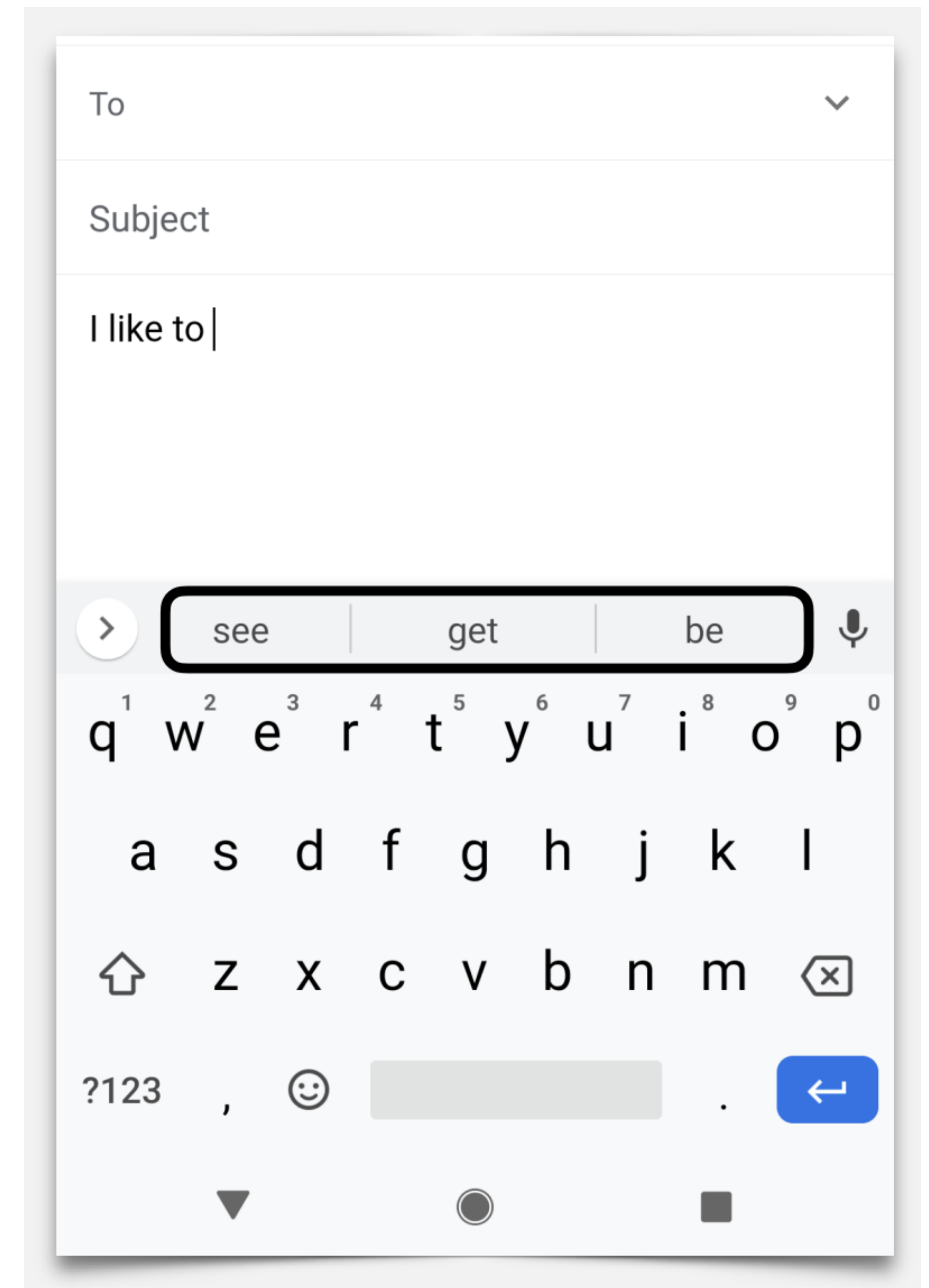
2-gram language models

- If we have these probabilities, we can build a predictive text system:

$$P(\text{next word} = _ | \text{previous word} = \text{to})$$

Check all the possible words from that list, pick the ones with the highest probability (most likely next words)

- Where do these probabilities come from? We're going to learn them from a bunch of text data we see



2-gram language models

Based on a conditional probability distribution:

“the probability of the next word is y given that the previous word is x ”

$$P(\text{next word} = y | \text{previous word} = x)$$

I want to go to_____

$$P(\text{next word} = \text{was} | \text{previous word} = \text{to}) = 0.0$$

$$P(\text{next word} = \text{LA} | \text{previous word} = \text{to}) = 0.2$$

$$P(\text{next word} = \text{Europe} | \text{previous word} = \text{to}) = 0.1$$

$$P(\text{next word} = \text{Mexico} | \text{previous word} = \text{to}) = 0.1$$

$$P(\text{next word} = \text{eat} | \text{previous word} = \text{to}) = 0.1$$

These have to add up to 1 over the vocabulary (every possible word y could be)

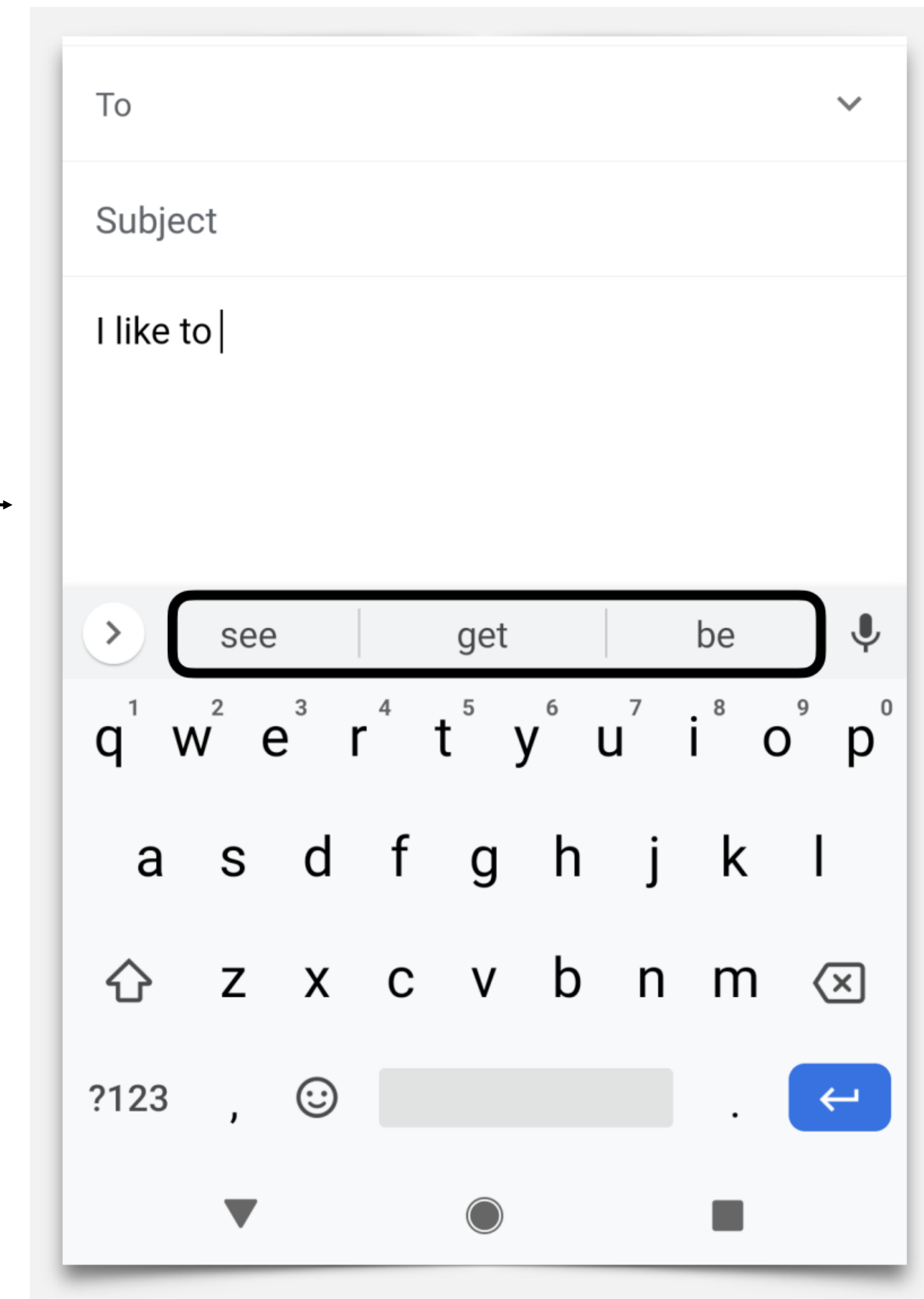
If we see “to”, there’s a 20% chance the next word is “LA”

Assume a fixed vocabulary of $\sim 30,000$ words



Lots and lots of text data

→
2-gram LM
probabilities
→



Estimating bigram probabilities

The maximum likelihood estimate

$$P(w_i | w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

$$P(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$



What happens when $i = 1$?

Special edge case tokens: $\langle s \rangle$ and $\langle /s \rangle$ for beginning of sentence and end of sentence, respectively

An example

$$P(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(\text{I} \mid \text{<s>}) = \frac{2}{3} = .67$$

$$P(\text{Sam} \mid \text{<s>}) = \frac{1}{3} = .33$$

$$P(\text{am} \mid \text{I}) = \frac{2}{3} = .67$$

$$P(\text{</s>} \mid \text{Sam}) = \frac{1}{2} = 0.5$$

$$P(\text{Sam} \mid \text{am}) = \frac{1}{2} = .5$$

$$P(\text{do} \mid \text{I}) = \frac{1}{3} = .33$$

Larger Example:

Berkeley Restaurant Project (BRP)

- can you tell me about any good cantonese restaurants close by
- mid priced thai food is what i'm looking for
- tell me about chez panisse
- can you give me a listing of the kinds of food that are available
- i'm looking for a good place to eat breakfast
- when is caffe venezia open during the day

Total: 9222 similar sentences

BRP: Raw Counts

Out of 9222 sentences

Unigrams

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

Bigrams

Next Word

History

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

BRP: Bigram Probabilities

Bigram Probabilities: Raw bigram counts normalized by unigram counts

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

w_i

	i	want	to	eat	chinese	food	lunch	spend
w_{i-1}	i	0.002	0.33	0	0.0036	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

What kinds of knowledge?

$$P(\text{english} | \text{want}) = .0011$$

$$P(\text{chinese} | \text{want}) = .0065$$

$$P(\text{to} | \text{want}) = .66$$

$$P(\text{eat} | \text{to}) = .28$$

$$P(\text{food} | \text{to}) = 0$$

$$P(\text{want} | \text{spend}) = 0$$

$$P(i | \langle s \rangle) = .25$$

Bigram estimates of sentence probabilities

$P(< s> \text{ I want english food } < /s>) =$

$P(\text{I} | < s>)$

$\times P(\text{want} | \text{I})$

$\times P(\text{english} | \text{want})$

$\times P(\text{food} | \text{english})$

$\times P(< /s> | \text{food})$

$= .000031$

Quite low...

Underflow Issues

- We do everything in log space
- Avoid underflow
- Adding is faster than multiplying

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Evaluation and Perplexity

How good is a language model?

Does our language model prefer good sentences to bad ones?

- Key Idea: Assign higher probability to “real” or “frequently observed” sentences than “ungrammatical” or “rarely observed” sentences?
 - In practice we don’t explicitly need to do the latter!

Intrinsic Evaluation

We train parameters of our model on a training set.

We test the model’s performance on data we haven’t seen.

- A test set is an unseen dataset that is different from our training set, totally unused.
- An evaluation metric tells us how well our model does on the test set.

How good is a language model?

- Extrinsic vs. Intrinsic Evaluation
 - Q: can you think of an extrinsic evaluation?

Extrinsic evaluation of N-gram models

Best evaluation for comparing models A and B

1. Put each model in a task
 - spelling corrector, speech recognizer, MT system
2. Run the task, get an accuracy for A and for B
 - How many misspelled words corrected properly
 - How many words translated correctly
3. Compare accuracy for A and B



Downsides??



Text Generation: Intrinsic or Extrinsic Evaluation?

Machine Learning 101

Train Set vs Test Set:

- We can't allow test sentences into the training set
- We will assign it an artificially high probability when we set it in the test set
- “Training on the test set” is bad science! And violates the honor code

Another risk of cheating:

- using a particular test set so often that we implicitly tune to its characteristics.
- So how to evaluate while developing a model? Use a fresh test set that is truly unseen: development set!

In practice, we often just divide our data into 80% training, 10% development, and 10% test.

How best to evaluate an LM?

- Extrinsic evaluation can be time-consuming; hard to design
 - Which is the best task? How many tasks to try?
- Therefore, we often use intrinsic evaluation:
 - Bad approximation
 - unless the test data looks just like the training data
 - Generally only useful in pilot experiments

Perplexity

Intuition of Perplexity

The Shannon Game: How well can we predict the next word?

I always order pizza with cheese and ____

The 33rd President of the US was ____

I saw a ____

mushrooms 0.1

pepperoni 0.1

anchovies 0.01

....

fried rice 0.0001

....

and 1e-100



- Unigrams are terrible at this game

A better model of a text

- is one which assigns a higher probability to the word that actually occurs

Perplexity

- The best language model is one that best predicts an unseen test set
 - Gives the highest $P(\textit{sentence})$

Perplexity is the inverse probability of the test set, normalized by the number of words

$$PPL(\mathbf{w}) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}}$$

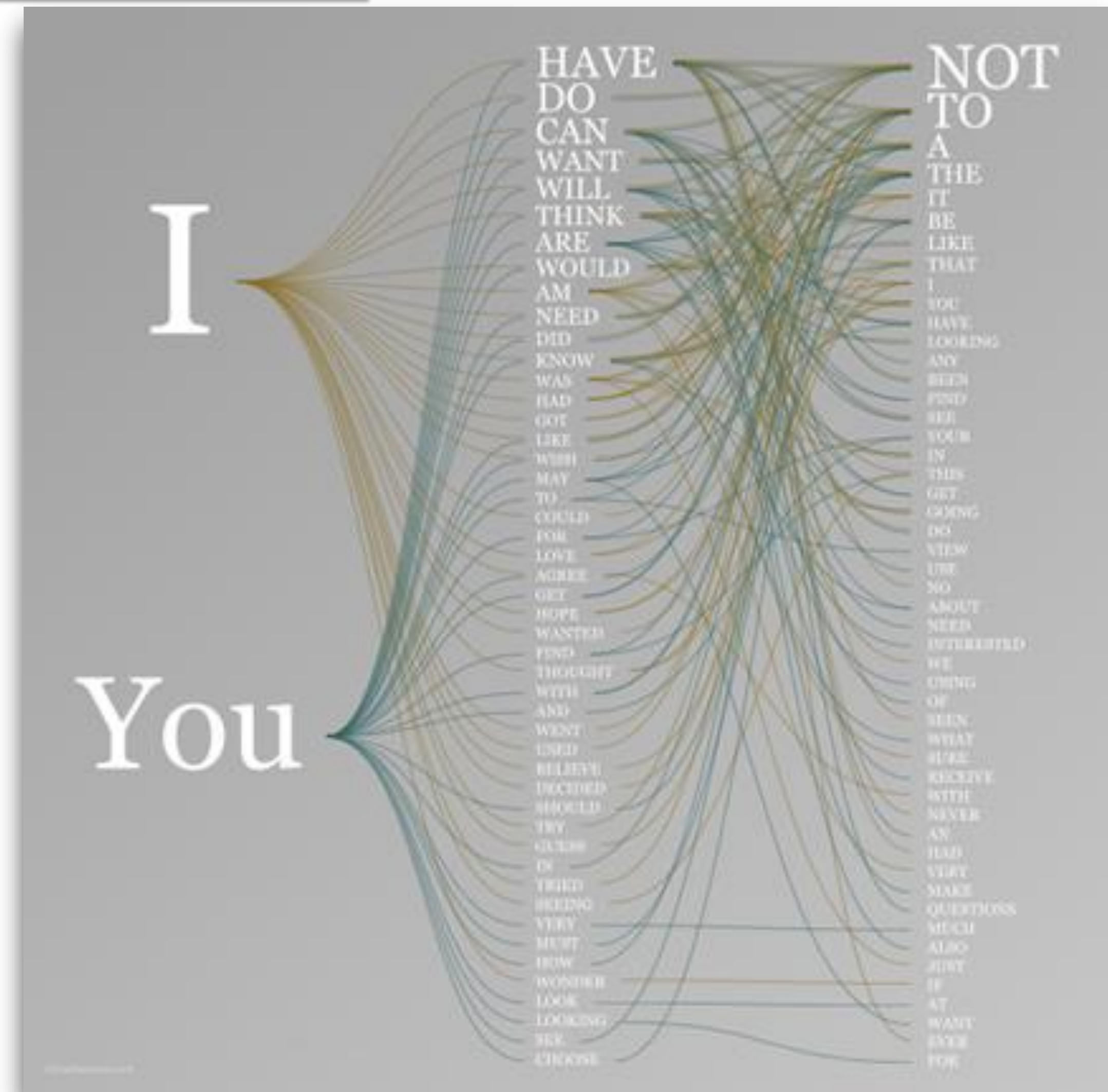
Welcome

CSCI 662 Spring 2026: NLP

🐝 Spring 2026 🕒 Wed 2:00 - 5:20pm 📍 [WPH 102](#)

- TODOs for you
 - Start building project teams (**due on Jan 28**)
 - Start brainstorming project ideas
 - Project pitch on **Feb 4**

Image Courtesy: Chris Harrison





Next Class

- Words are more than discrete symbols
- Parameters!!!

