

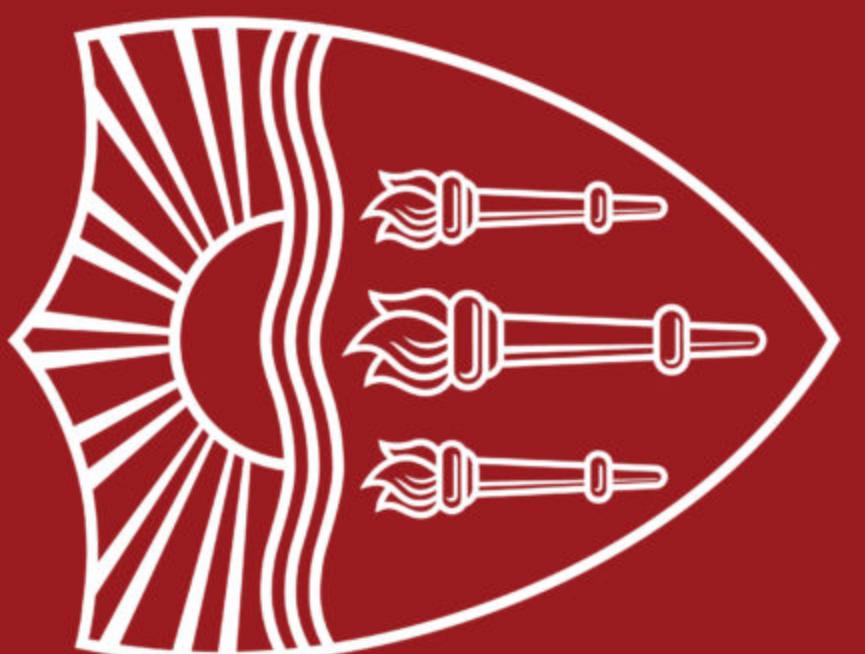


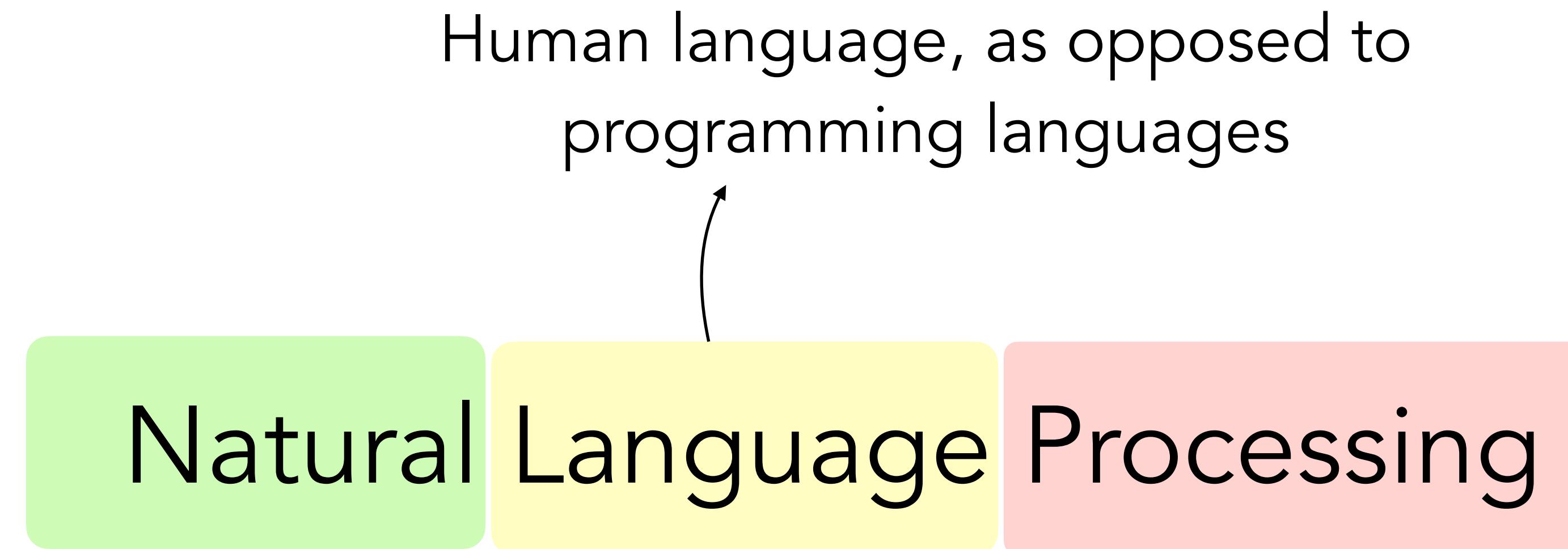
Lecture 1: Introduction and Course Overview

Instructor: Swabha Swayamdipta

USC CSCI 444 Natural Language Processing

Aug 25, Fall 2025





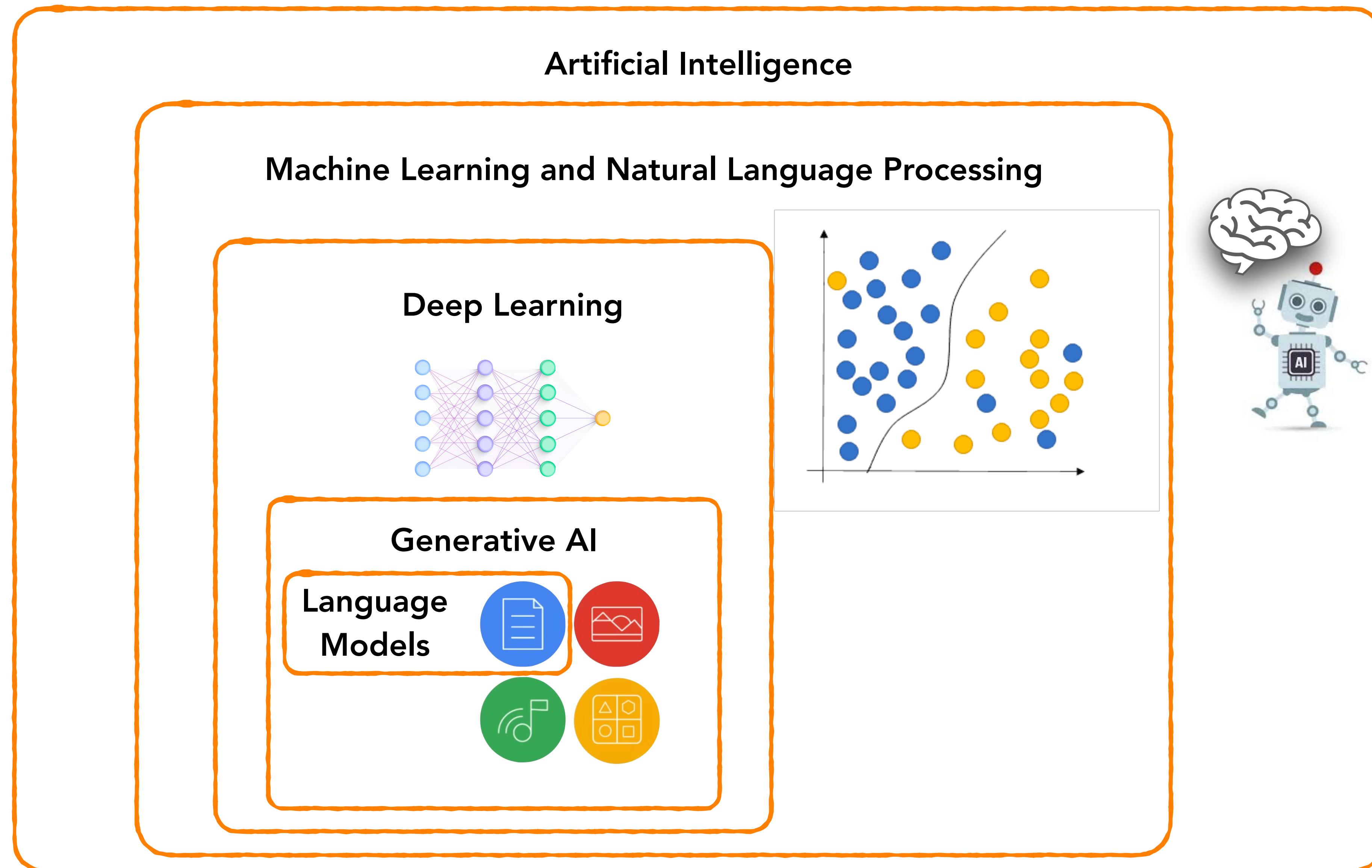
Automatic, as opposed to manual

NLP today = Almost entirely Language Models

More generally, language + Z

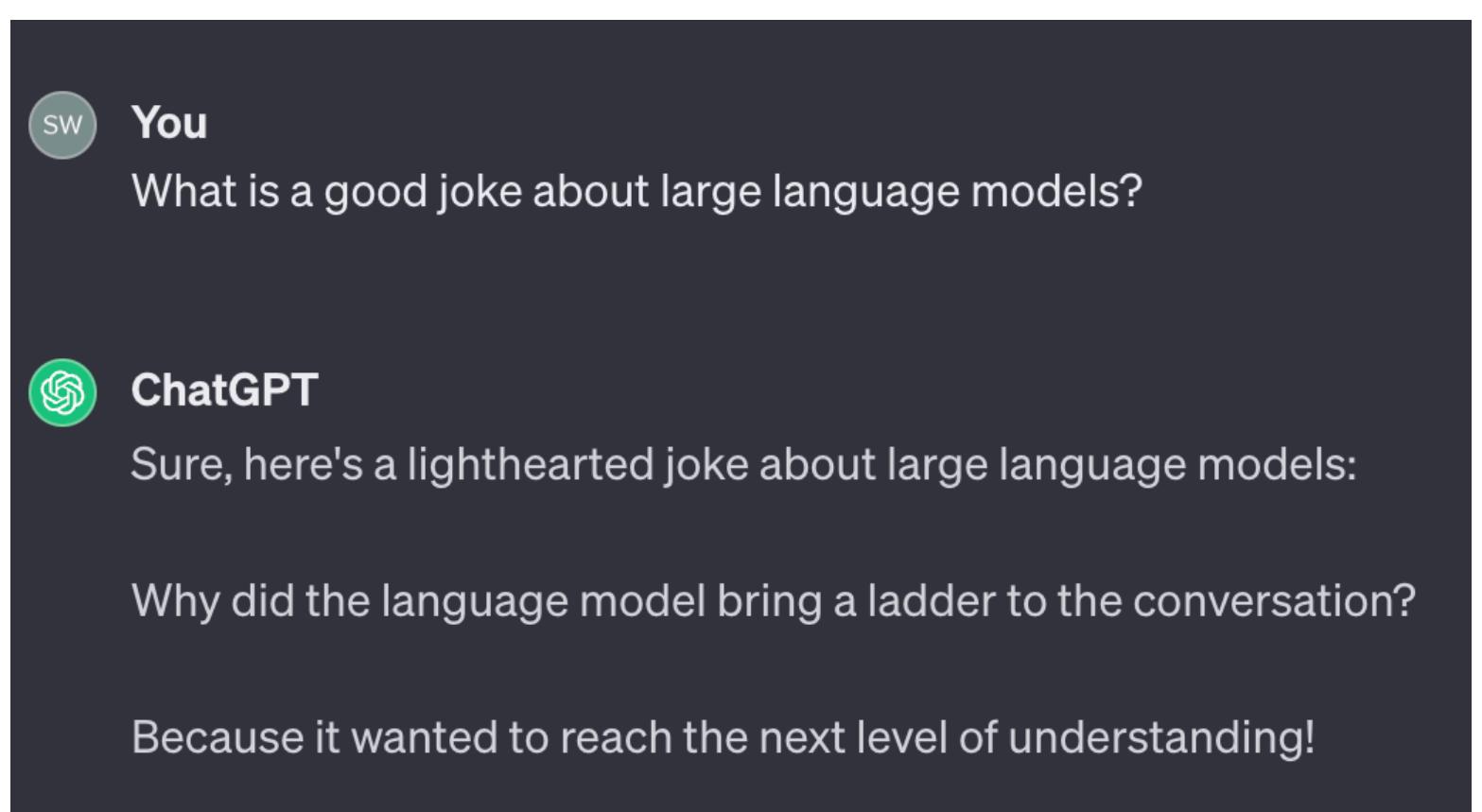
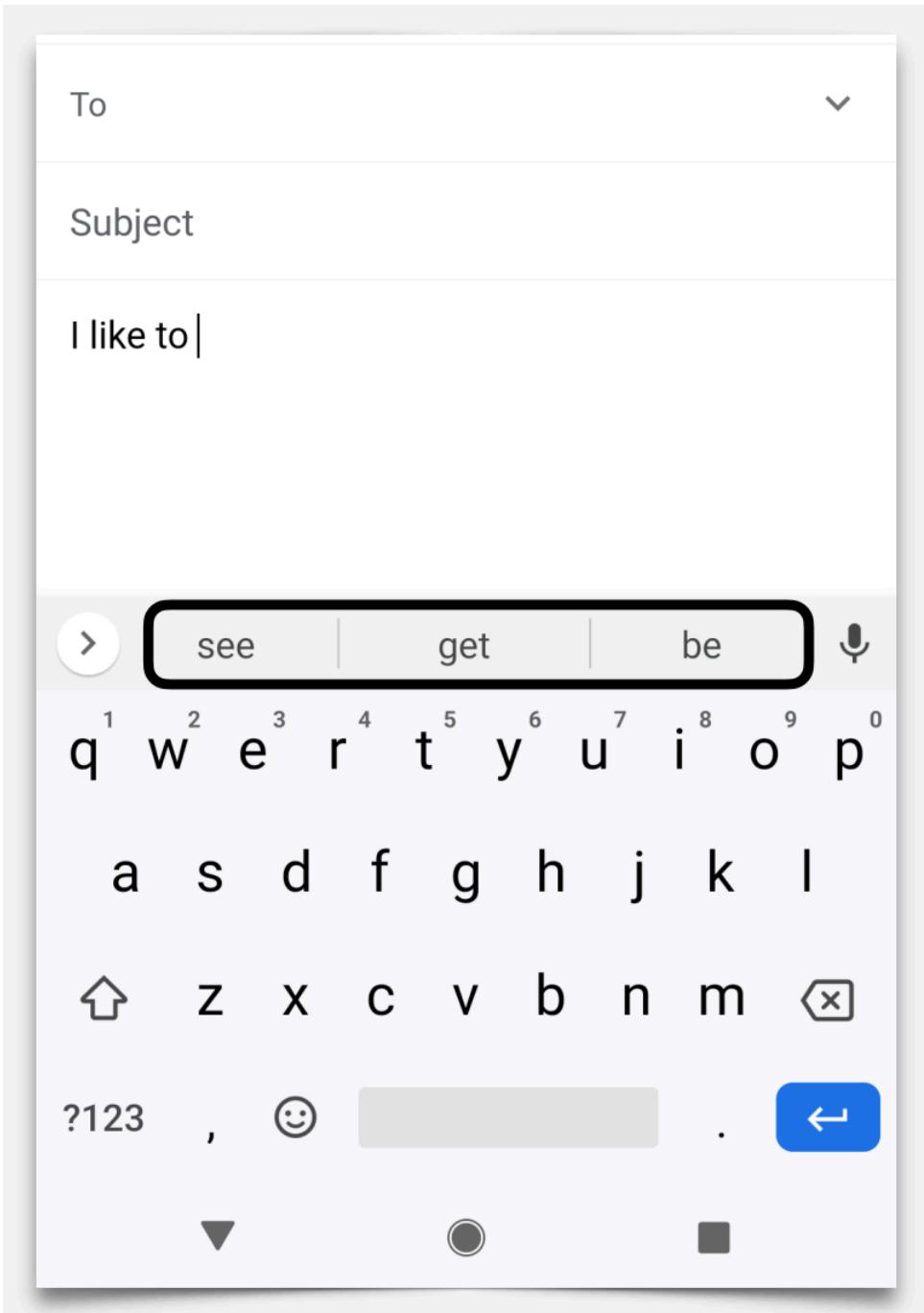
What is Natural Language Processing ?

- Field at the intersection of computer science, AI (especially machine learning or deep learning) and linguistics
- Processing: produce outputs (Y) with language or text as input (X)
 - Outputs and inputs can contain other modalities (images, videos) as well
- In today's parlance, NLP is the science behind language models
- Goal: for AI to interact with humans using our language, towards performing useful tasks
- Challenge: understanding and representing the meaning of language is something even humans struggle with



Language Models

- Task: Given a sequence of words so far (**the context**), predict what comes next
- But today, variants of language models are used
 - As chat models
 - As supercharged autocompletes
 - 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)





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

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

A.I.'s Insatiable Appetite for Energy



Language Models are the most popular form of AI today!

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



Google C.E.O. Sundar Pichai on the A.I. Moment: 'You Will See Us Be Bold'

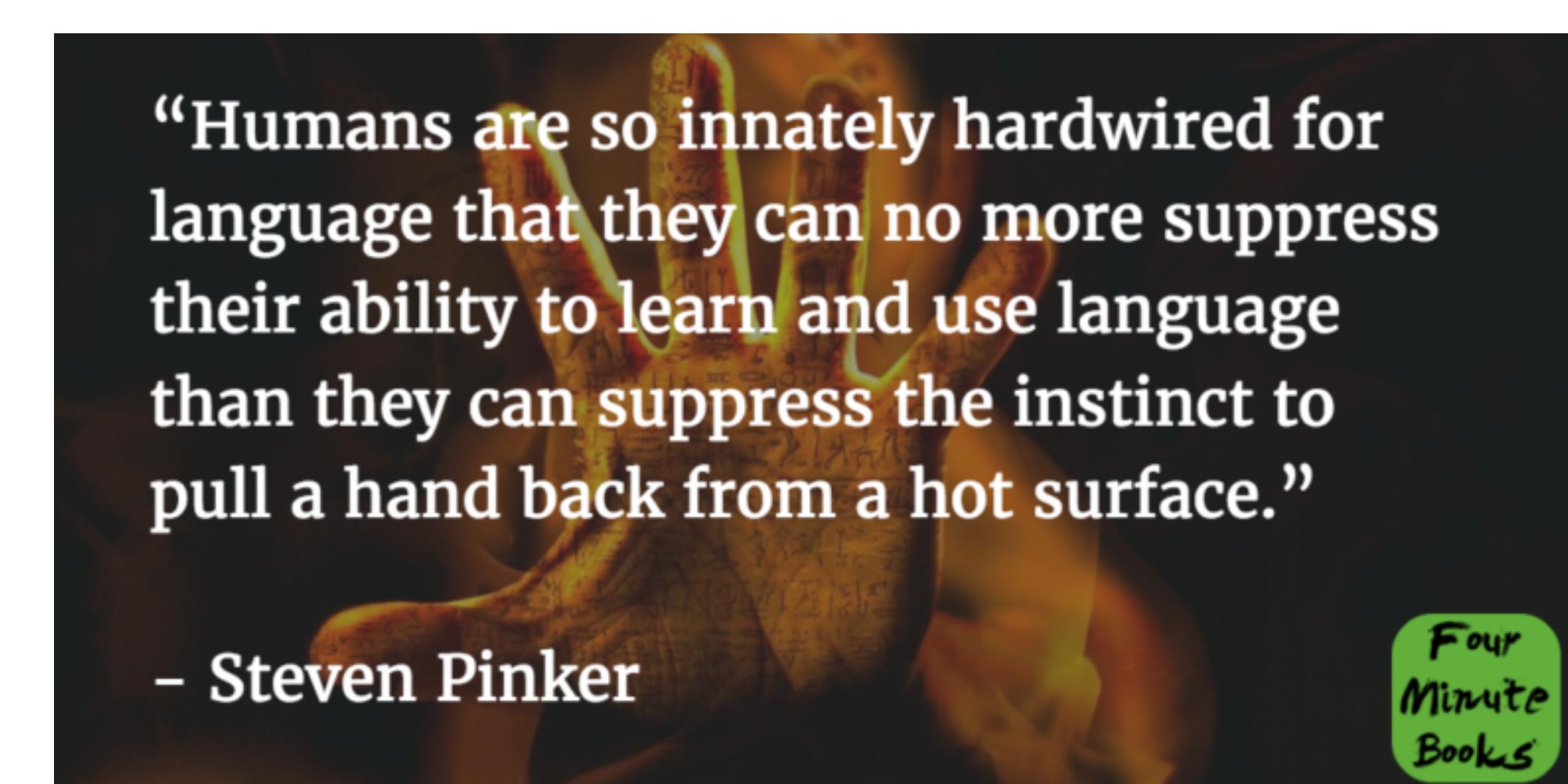
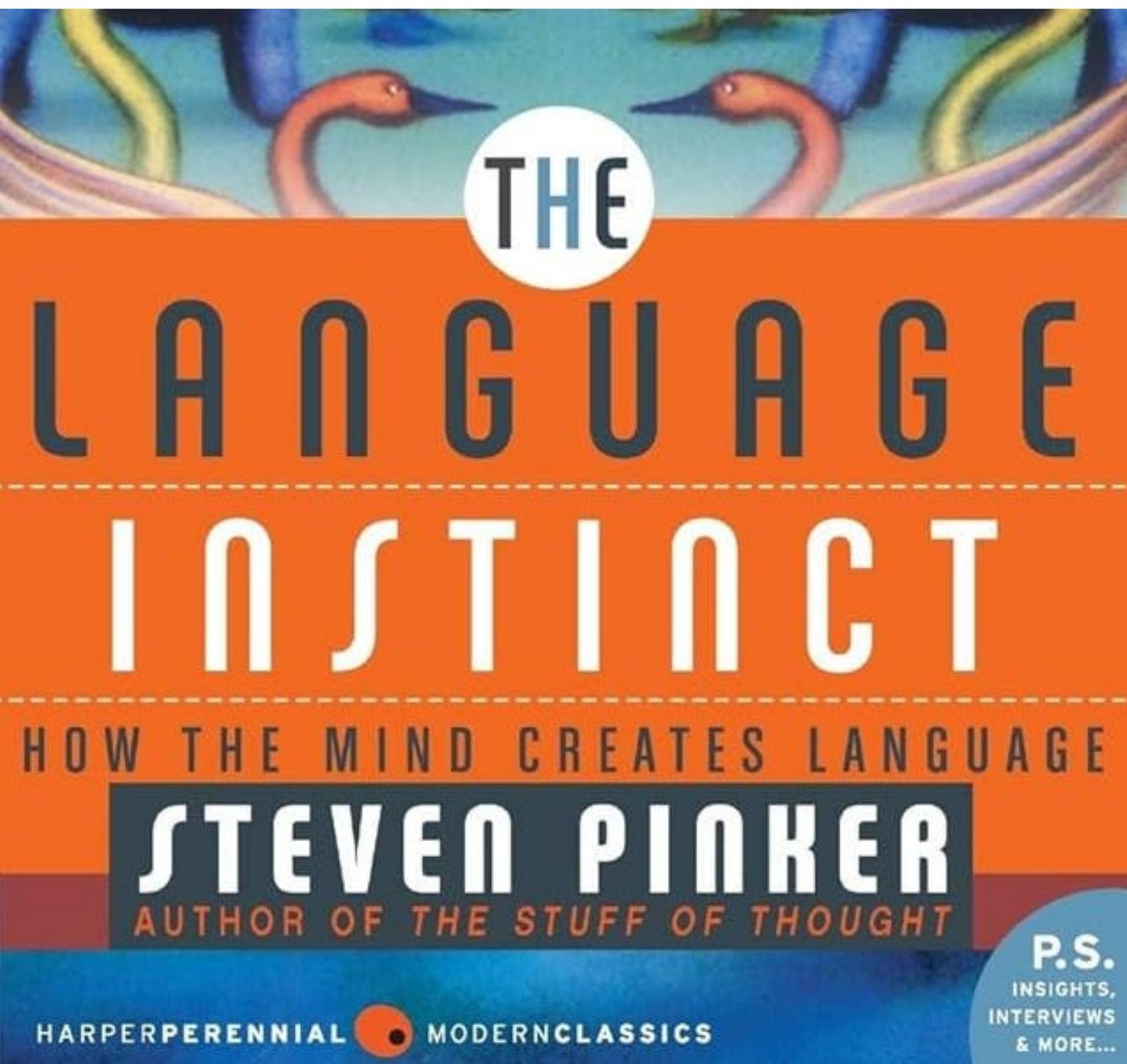
In an extended interview, Mr. Pichai expressed both optimism and worry about the state of the A.I. race.

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

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

The Chef Is Human. The Reviewer Isn't.

Language is a key modality



Language Models Are Everywhere



Virtual
Assistants



Translation



Content
Creation



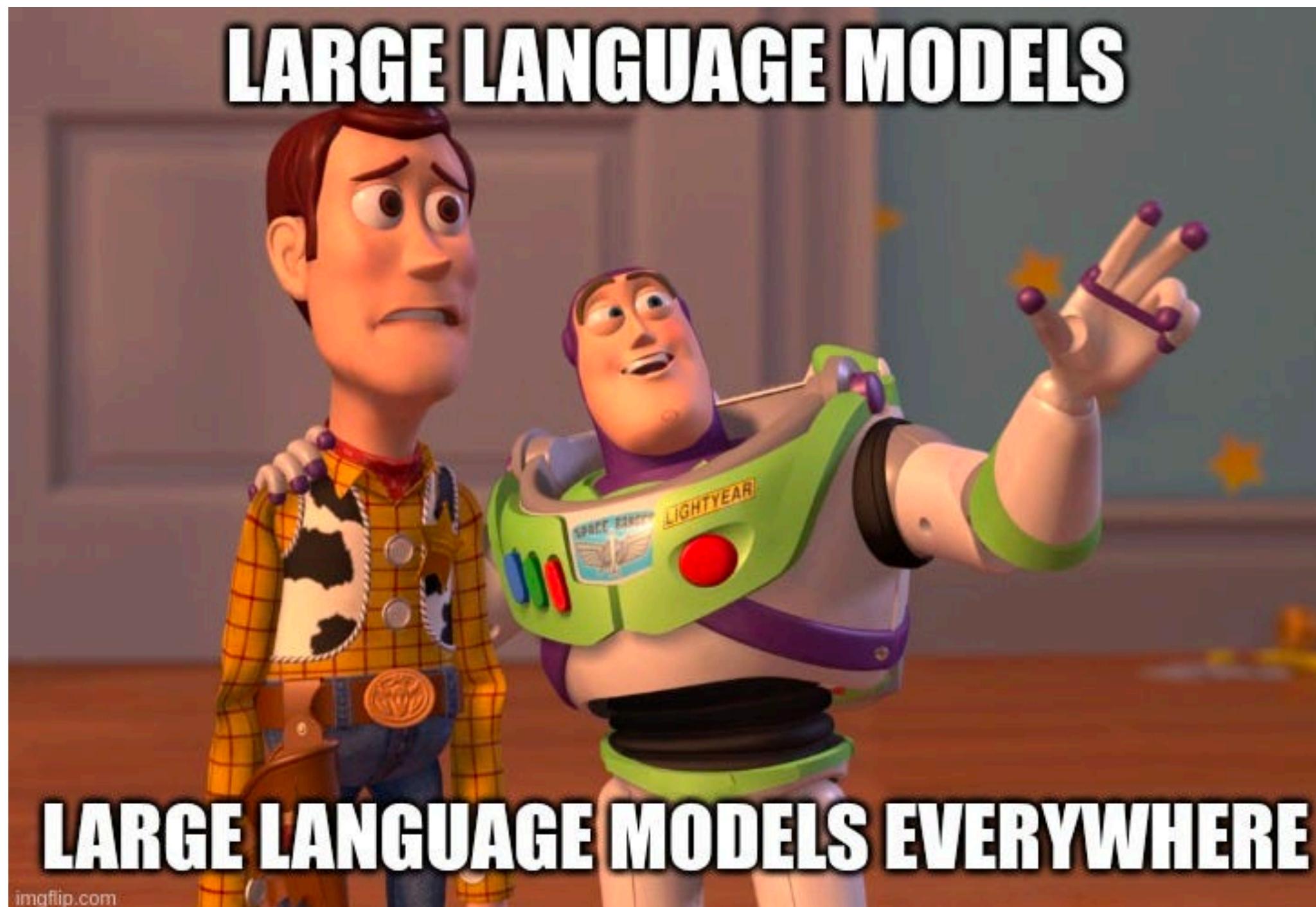
Code
Generation



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

GPT-4 Passes the Bar Exam!

GPT Takes the Bar Exam

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-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

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

December 29, 2022

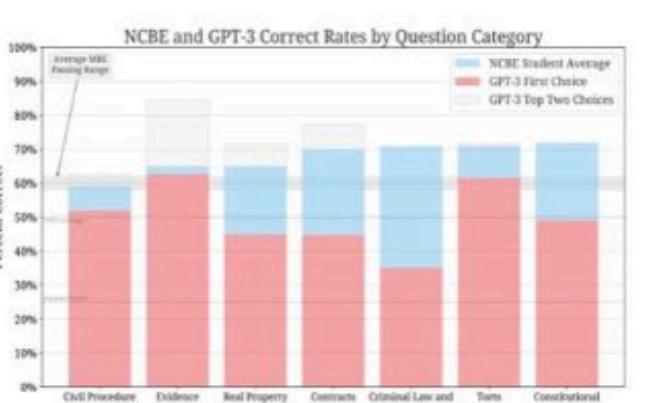
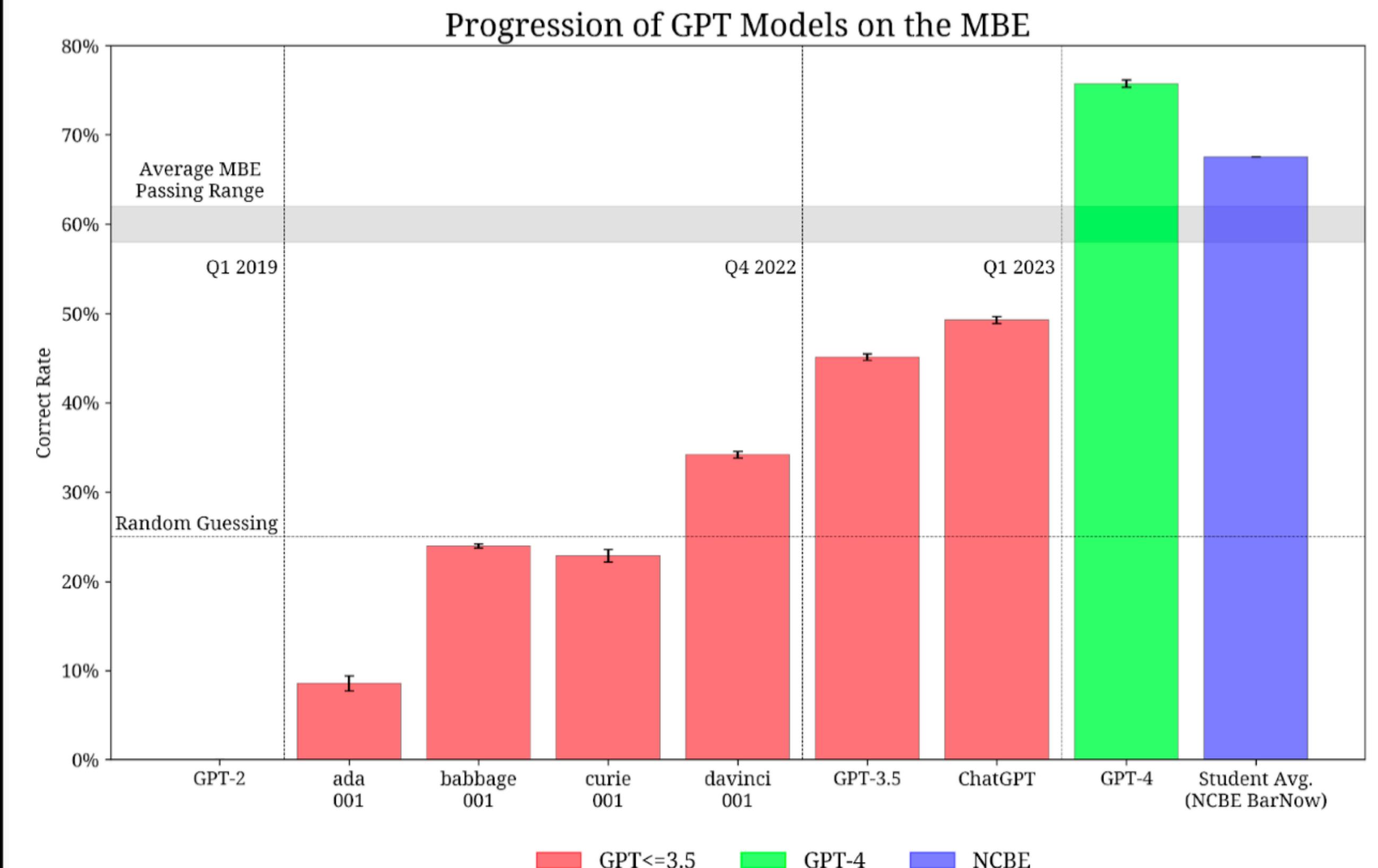
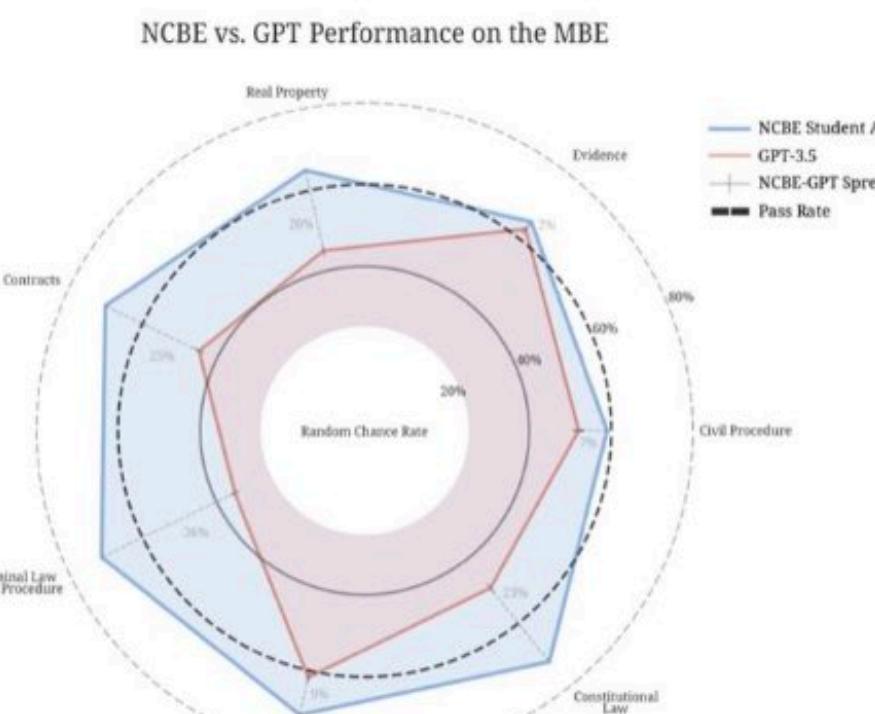
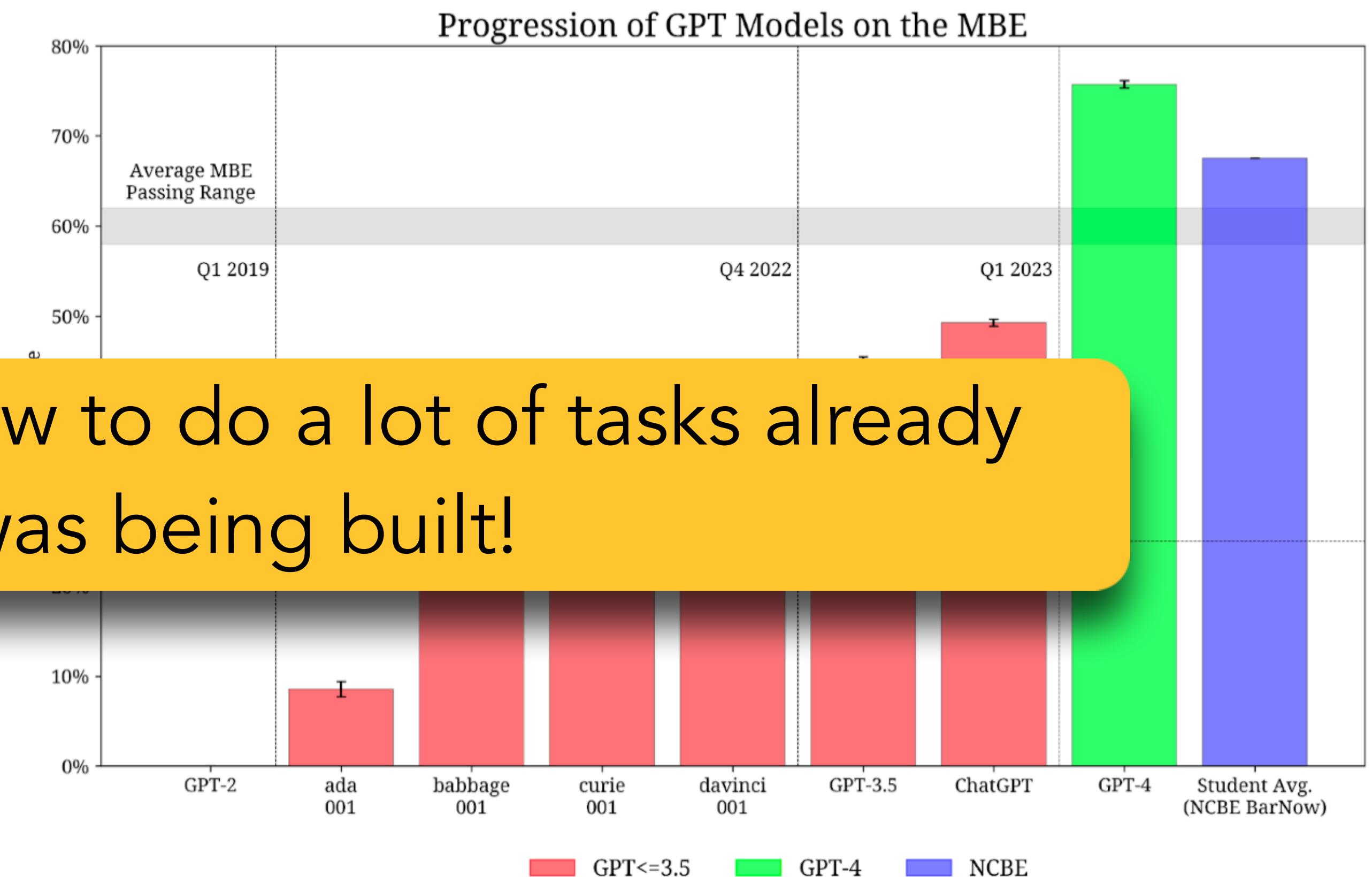
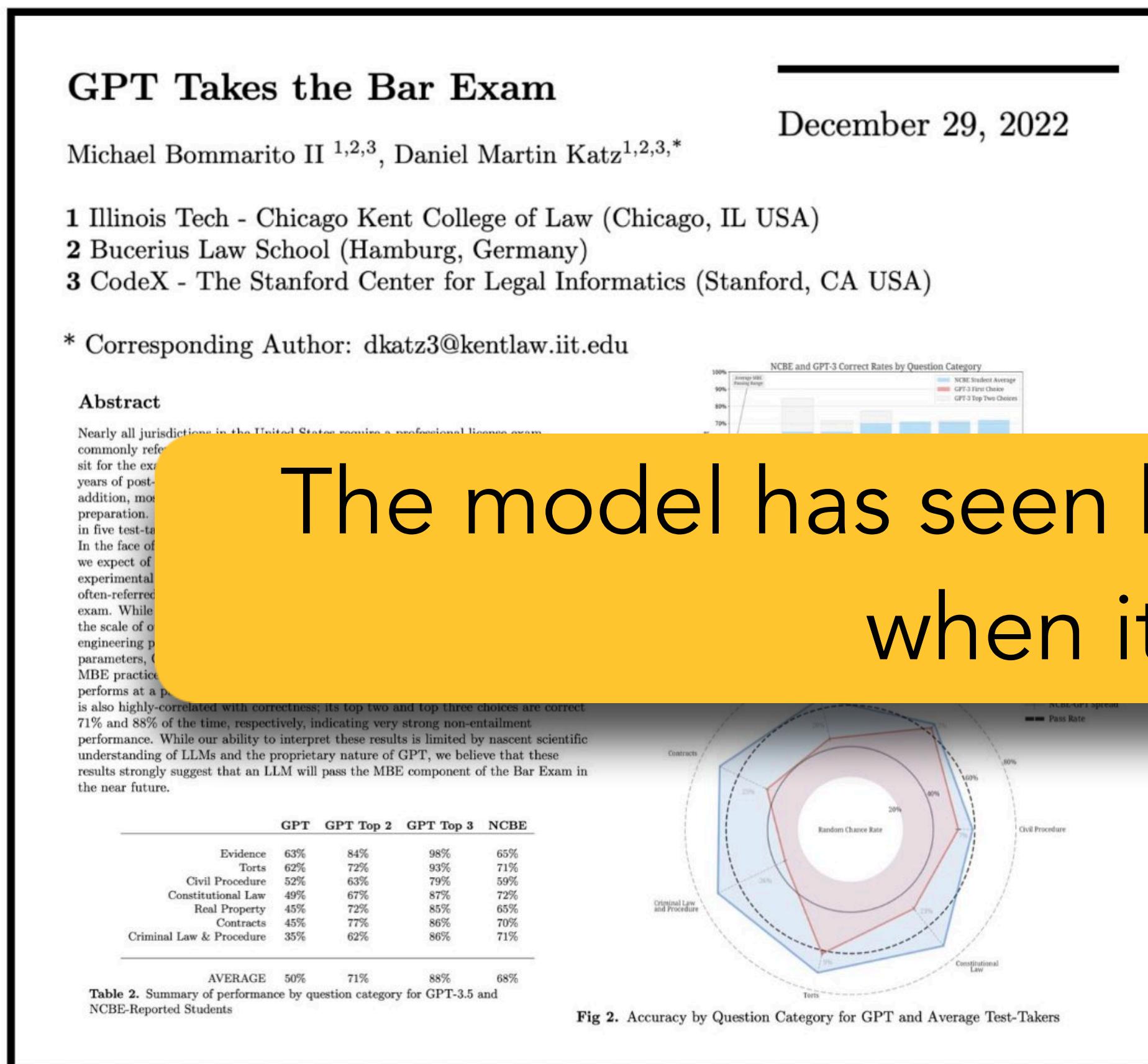


Fig 1. Summary of performance by question category for GPT-3.5 and NCBE-Reported Students



Progression of GPT Models on the MBE

Why does this work?



Memorization vs. generalization

 **Yann LeCun**  
@ylecun

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 · 1M Views

AI in practice Dec 26, 2024

Deepseek-V3 emerges as China's most powerful open-source language model to date



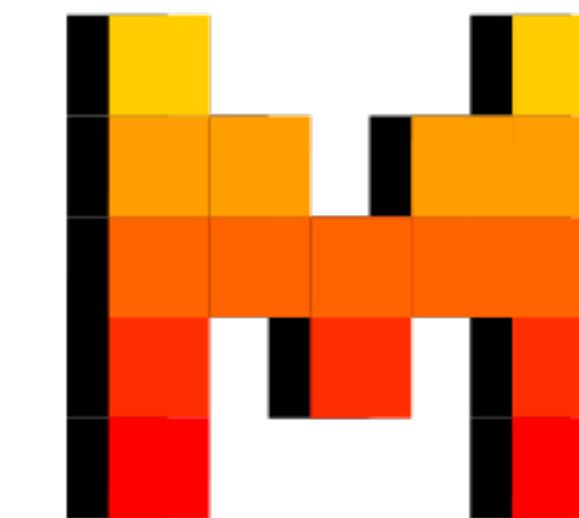
Midjourney prompted by THE DECODER



Gemini



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_

Language Models 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.



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.



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

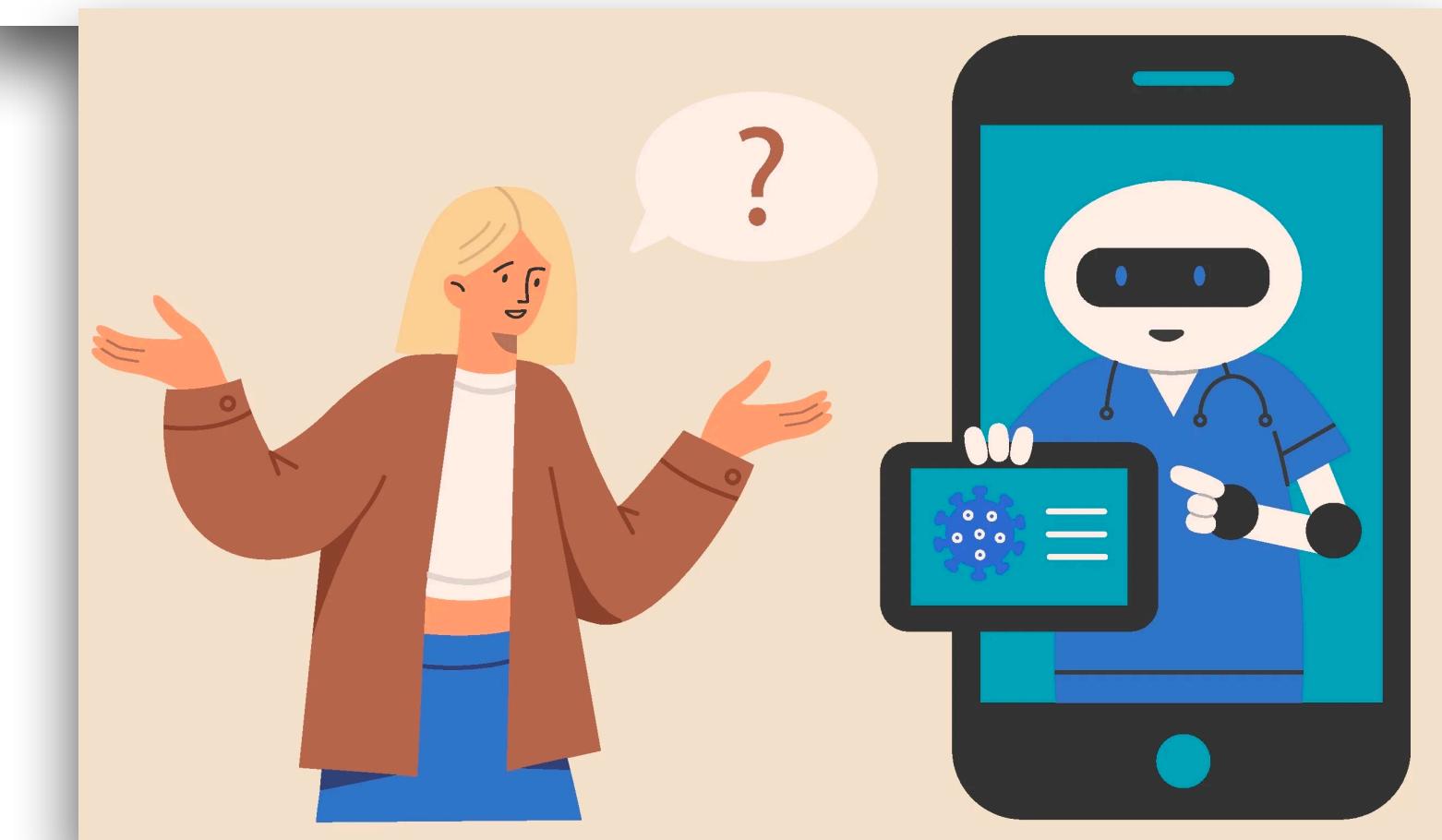
Artists want to be able to post their work online without the fear “of feeding this monster” that could replace them.

Privacy and Copyright Issues

Mar 8, 2023 - Technology

Chatbot therapy, despite cautions, finds enthusiasts

Peter Allen Clark



Ethical Issues and Biases

Lecture Outline

1. Course Introduction
2. Course Logistics
3. Probabilistic Language Models
4. n -gram Language Models

Class Logistics



Instructor, Website and Students



Instructor: [Swabha Swayamdipta](#)

swabhas@usc.edu

Office Hours: Monday 1-2pm

Lectures + Readings:



- Name
- Degree Status (Junior, Senior, etc.)
- One fun fact

Now your turn!

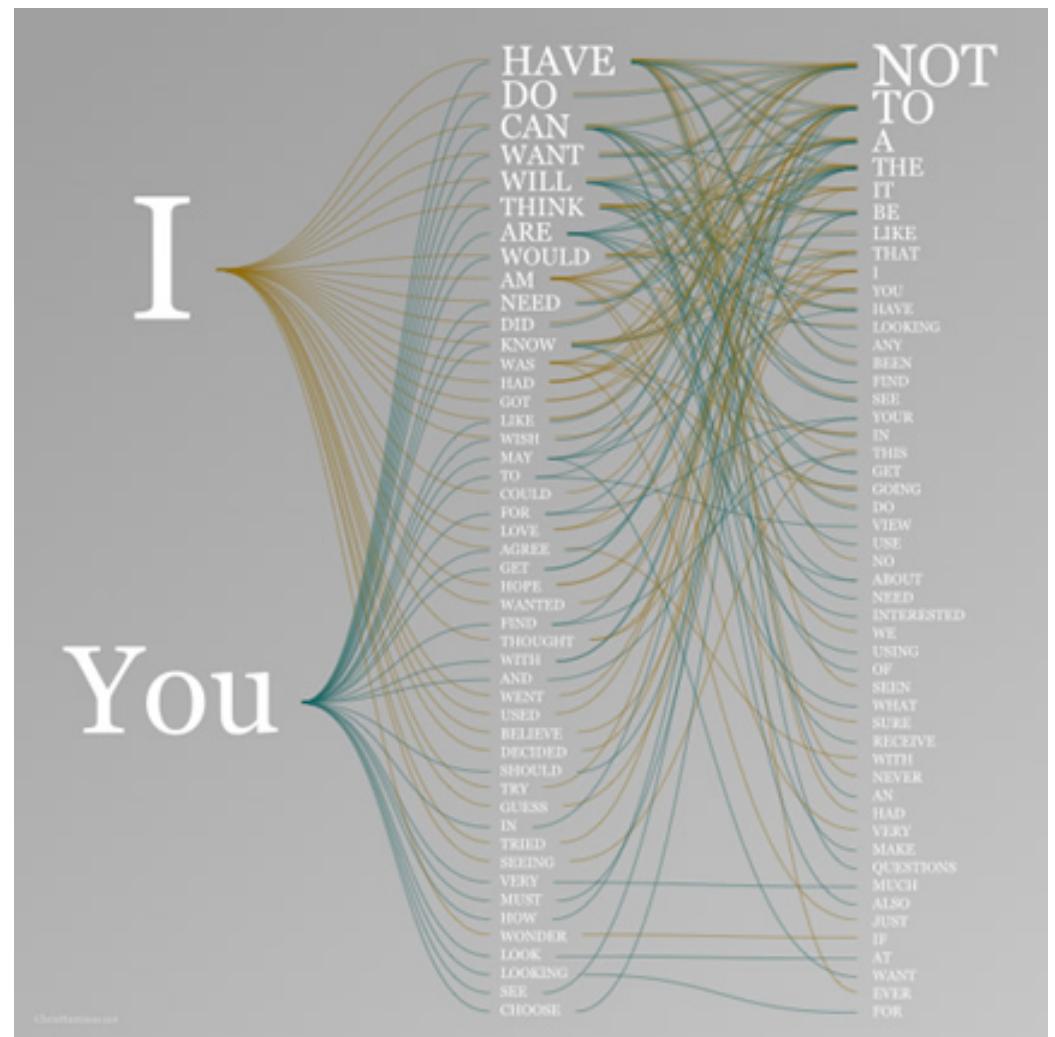
<https://swabhs.com/2503-csci444-nlp/>

Announcements etc. on Brightspace

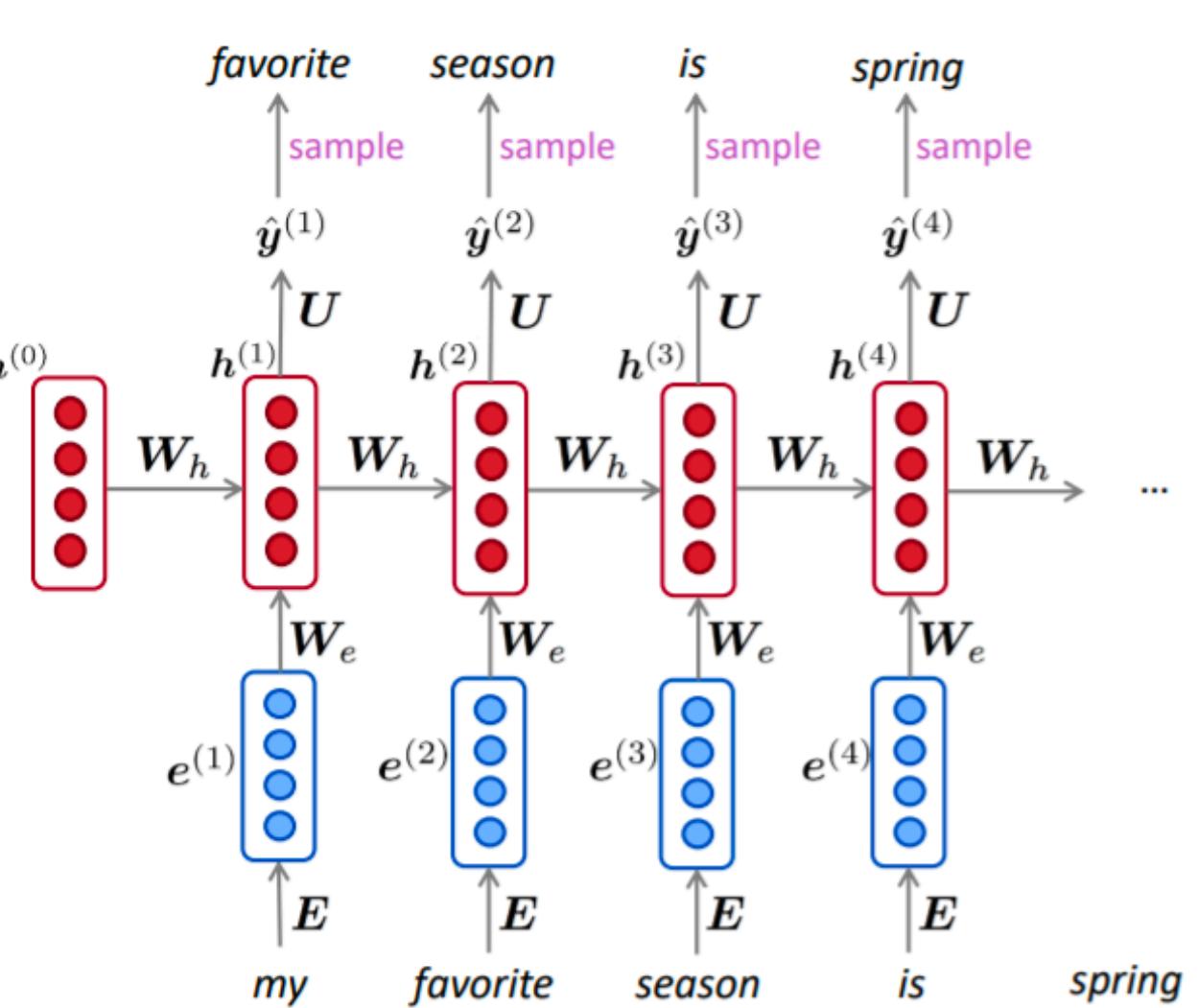


brightspace
by D2L

Class Syllabus

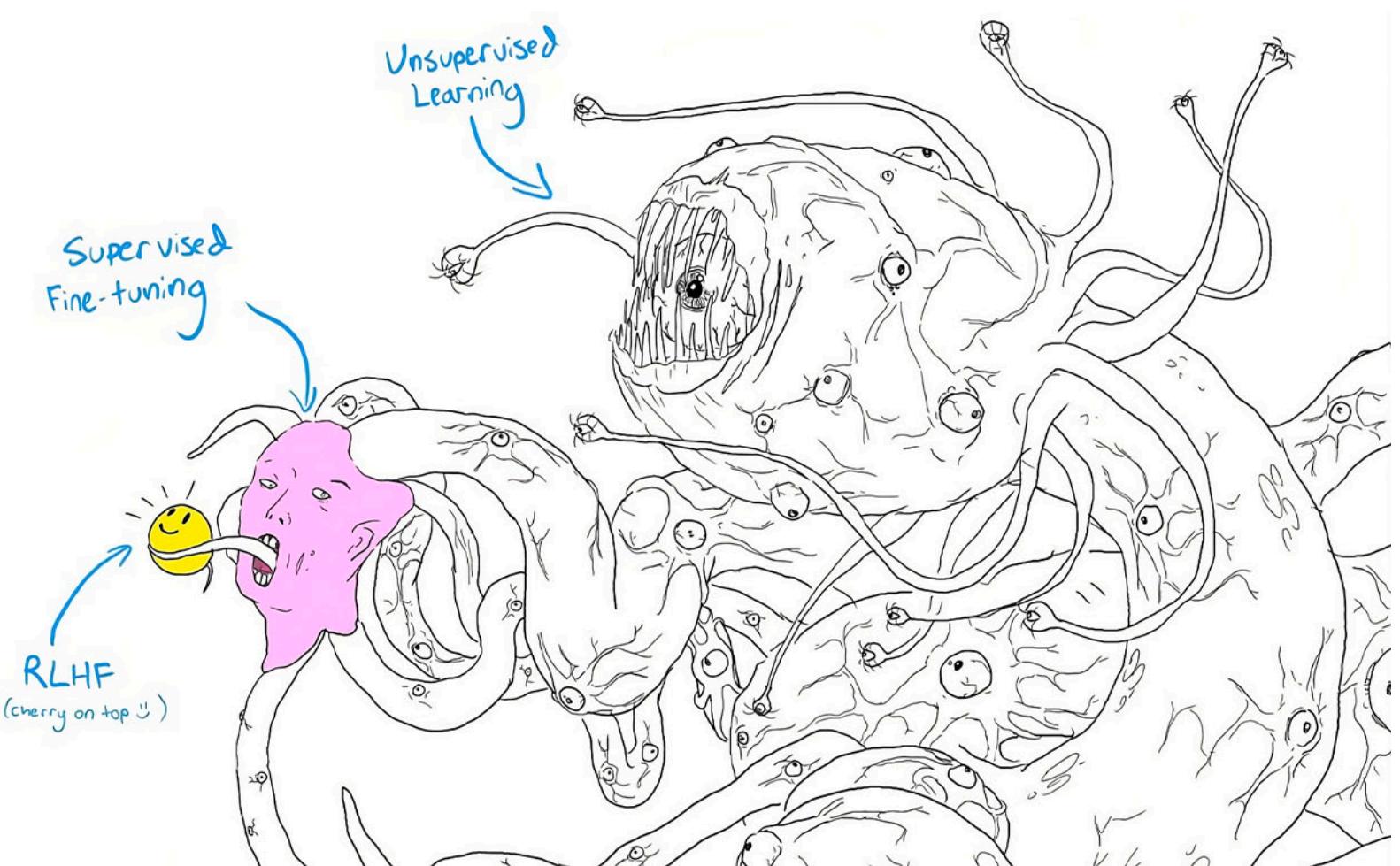
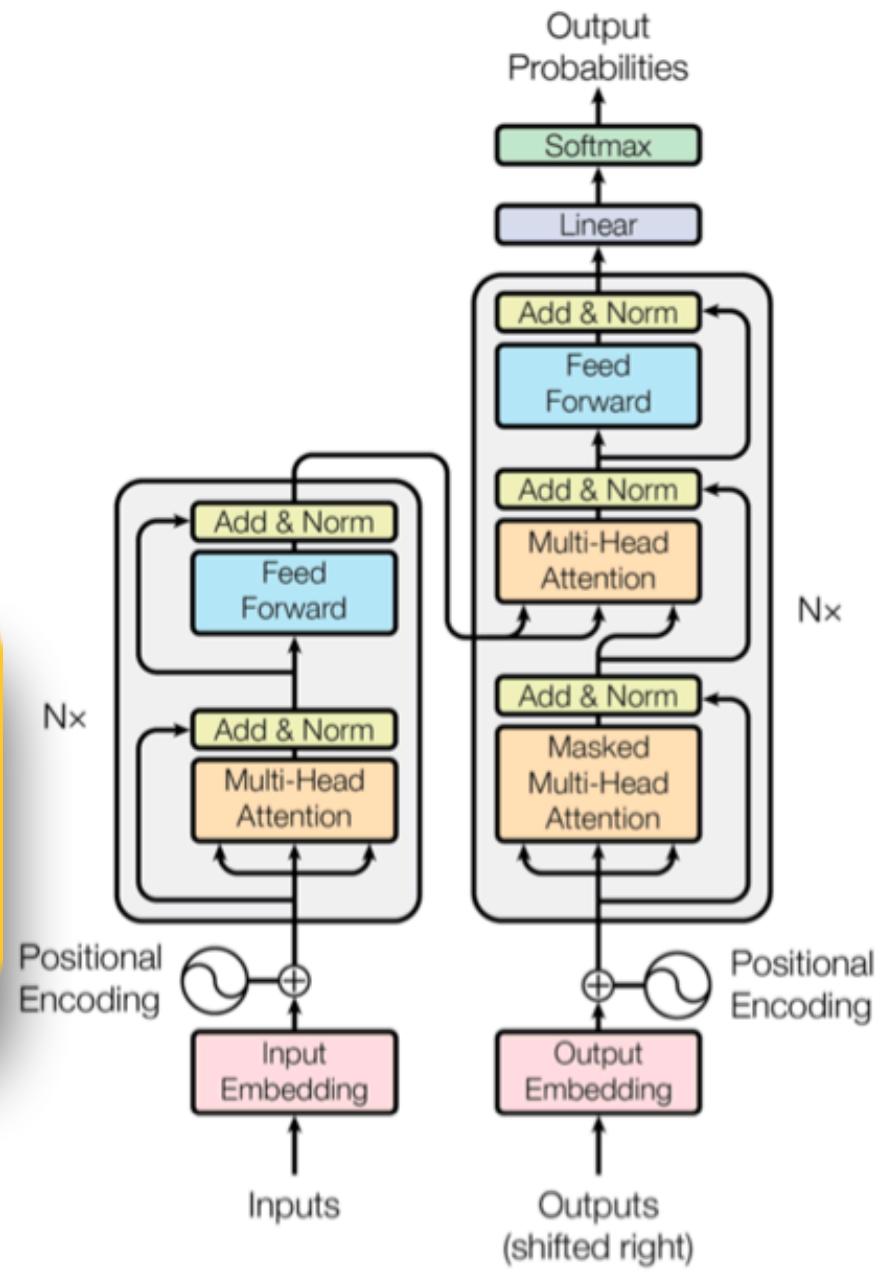


Pre-Neural Age of
LMs (-2013)



Early LMs
(2013-2018)

Modern LMs
(2018-2022)



LMs today (2023-present)

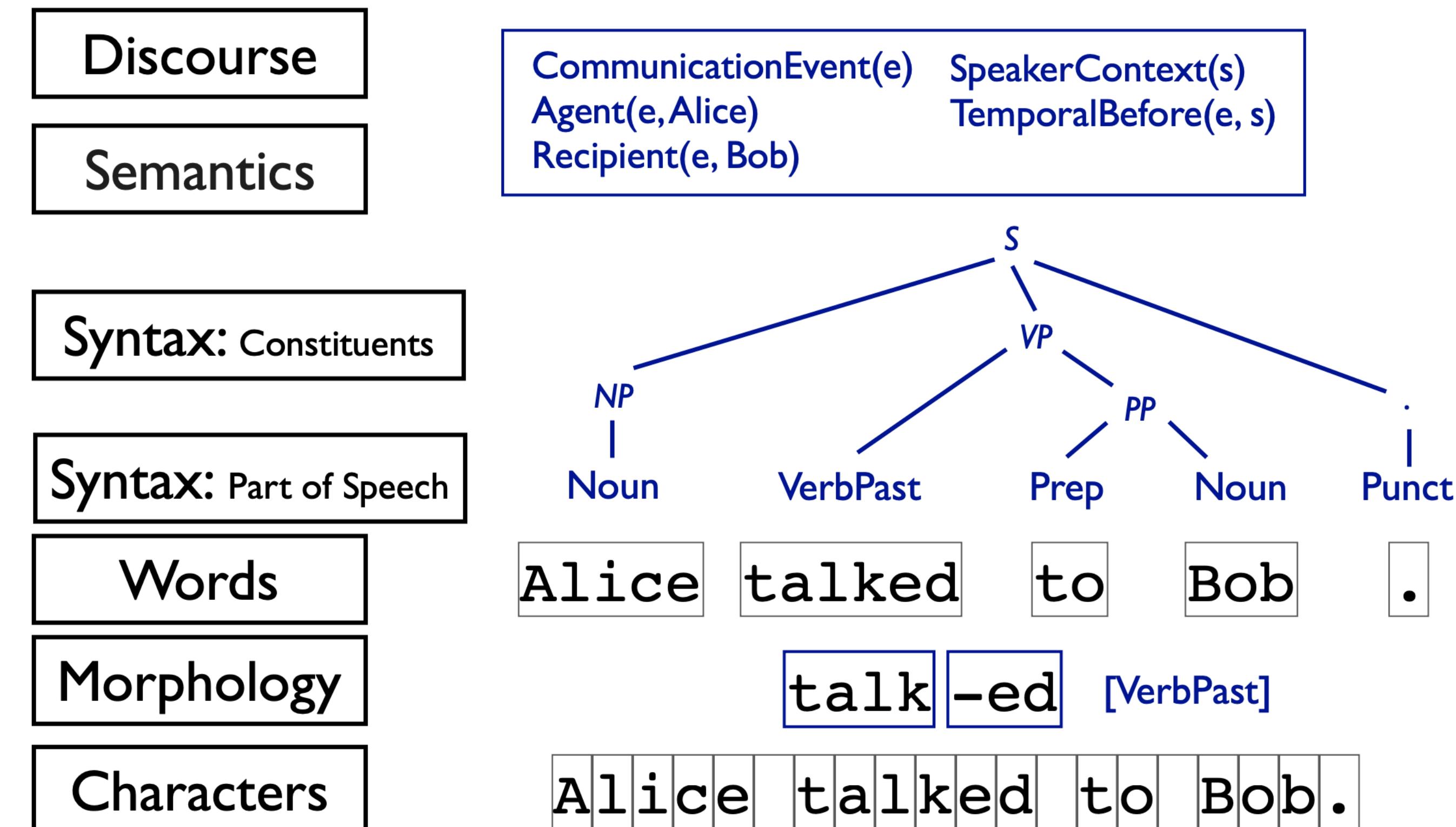
Concrete Outcomes

- Fundamentals of language modeling
- Build some language models - homework assignments and / or class projects
- Learn the connections between this language model and commercial language models such as ChatGPT or Gemini
- Current capabilities and outstanding issues with language models, along with exciting new problems



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



In other words, this class should have been named
“Language Models”

Class Deliverables

- Homework Assignments: 10% X 3
- Quizzes: 3% X 5
- Class Project:
 - Pitch: 5%
 - Proposal: 5%
 - Status Report: 10%
 - Project Presentation: 10%
 - Final Report: 10%
- Paper Presentations: 10%
- Class Participation: 5%

Class Project Topics - Examples

Detoxifying Language Model with Context

Distillation

Haiku Generation with Large Language Models

Legal-SBERT: Creating for the Legal Domain and Generating Data

Prompting for Diverse Responses: Making Large Language Models More Truthful

Creativity in choosing new and interesting problems often get rewarded!

See more: Stanford CS224n Projects

Forage: ML Generated Recipes

Learning the Language of Wine

Machine Translation from Inuktitut to

Comments in Social Commentary

Authorship Attribution with Limited Text

When Was it Written?

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.

Website contains links to chapters,
available for free

Lecture Outline

1. Course Introduction
2. Course Logistics
3. Probabilistic Language Models
4. n -gram Language Models

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”



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

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

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. $P(\text{its, water, is, so, transparent, that})$

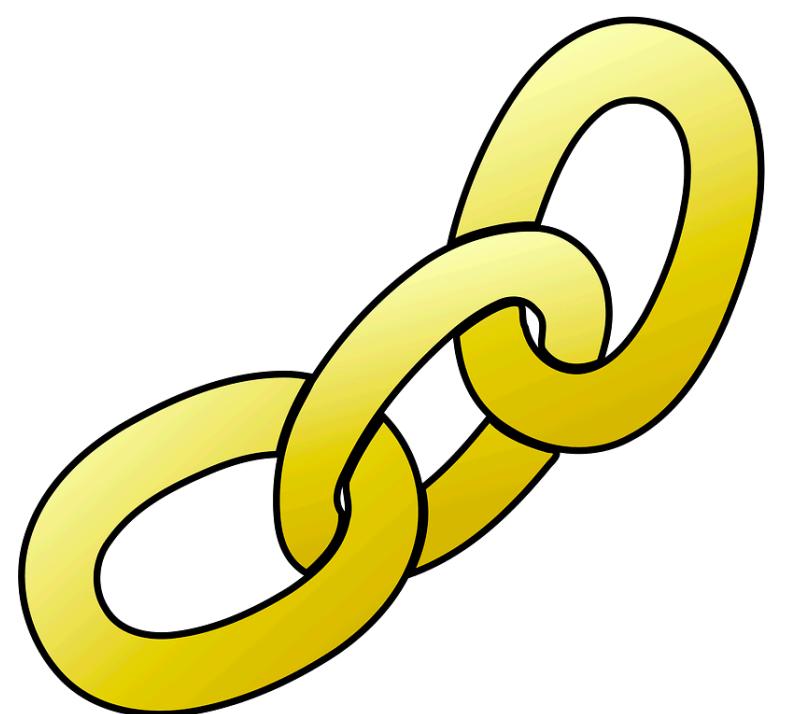
Intuition: let's rely on the Chain Rule of Probability

Chain Rule for words in a sentence

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

$P(\text{its water is so transparent}) = P(\text{its}) \times$
 $P(\text{water} | \text{its}) \times$
 $P(\text{is} | \text{its water}) \times$
 $P(\text{so} | \text{its water is}) \times$
 $P(\text{transparent} | \text{its water is so})$

Ordering matters in language!



Why Probabilistic Models?

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

- 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

I will be back soonish

I will be bassoon dish

Your so silly

You're so silly

Everything has improve

Everything has improved

Probabilistic Language Models

Machine Translation:

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

Spell Correction:

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

Speech Recognition:

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

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 .

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

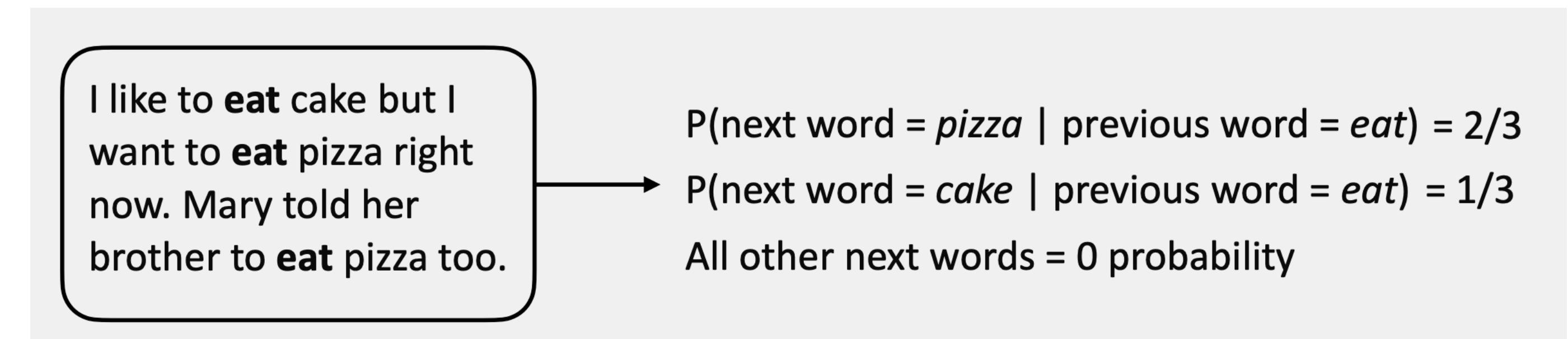
We don't know what p is — could be 0.5! But $p = 3/4 = 0.75$ maximizes the probability of data sequence (H,H,H,T)

maximum likelihood estimate

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

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



Vocabulary

Corpus

How to estimate the probability of the next word?

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

Could we just count and divide?

No! Too many possible sentences!

We'll never see enough data for estimating these

Markov Assumption

Simplifying Assumption:

$P(\text{that} \mid \text{its water is so transparent}) \approx P(\text{that} \mid \text{transparent})$



Andrei Markov

Or maybe...

$P(\text{that} \mid \text{its water is so transparent}) \approx P(\text{that} \mid \text{so transparent})$

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

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

($k + 1$)-th order Markov assumption

Mini Recap: Probabilistic Modeling

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



CoolClips.com

Lecture Outline

1. Course Introduction
2. Course Logistics
3. Probabilistic Language Models
4. n -gram Language Models

n-gram Language Models

simplest probabilistic model

Simplest Case: Unigram model

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

First-order Markov Assumption: the probability of a word only depends on itself

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

Second-order Markov Assumption: The probability of a word is conditioned 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 Language Models

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.”

Long-distance / Long-range dependencies

But we can often get away with *n*-gram models, where *n* is a small number

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: <s> and </s>
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 >

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

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

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

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

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

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

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

History

Next Word

	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

 w_i

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

	i	want	to	eat	chinese	food	lunch	spend
w_{i-1}								
i	0.002	0.33	0	0.0036	0	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?

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

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

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

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

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

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

$P(\text{i} \mid \langle s \rangle) = .25$

Bigram estimates of sentence probabilities

$$\begin{aligned} P(<\text{s}> \text{ I want english food } </\text{s}>) &= \\ P(\text{I} | <\text{s}>) & \\ \times P(\text{want} | \text{I}) & \\ \times P(\text{english} | \text{want}) & \\ \times P(\text{food} | \text{english}) & \\ \times P(</\text{s}> | \text{food}) & \end{aligned}$$

= .000031

Quite low...

Underflow Issues

We do everything in log space

- Avoid underflow
- Adding is faster than multiplying

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

Welcome

CSCI 444 Fall 2025: NLP

Fall 2025



 Mon / Wed 10:00 - 11:50a



WPH 106

- TODOs for you
 - Start talking to each other and seeking out potential teammates
 - Request: please spread the word among friends :)
 - Next Class
 - n-gram Language Models contd.

