

# CS 6120

# Natural Language Processing

## Lecture 1

### Language Models in Brains and Machines

Si Wu

# Welcome!

First, class logistics

# Class logistics

- “Which CS6120 session should I enroll?”
  - 21600 (1:35pm session) and 21601 (my session) will teach the **same material**
    - Grading, TAs, lectures, everything will be the same
- But make sure you are in the lecture session you’re enrolled in

# Class logistics

- **Class website:** <https://siwu.io/nlp-class/>
- **Gradescope:**
  - link on class website
  - Code to enroll this class: **NGZDZP**
- **Ed Discussion:** for asking questions
  - Link to join <https://edstem.org/us/join/a9m5ff>
- Lectures are not recorded. Slides will be posted after lectures
- We will *not* use Canvas

# “Why should I go to the lecture?”

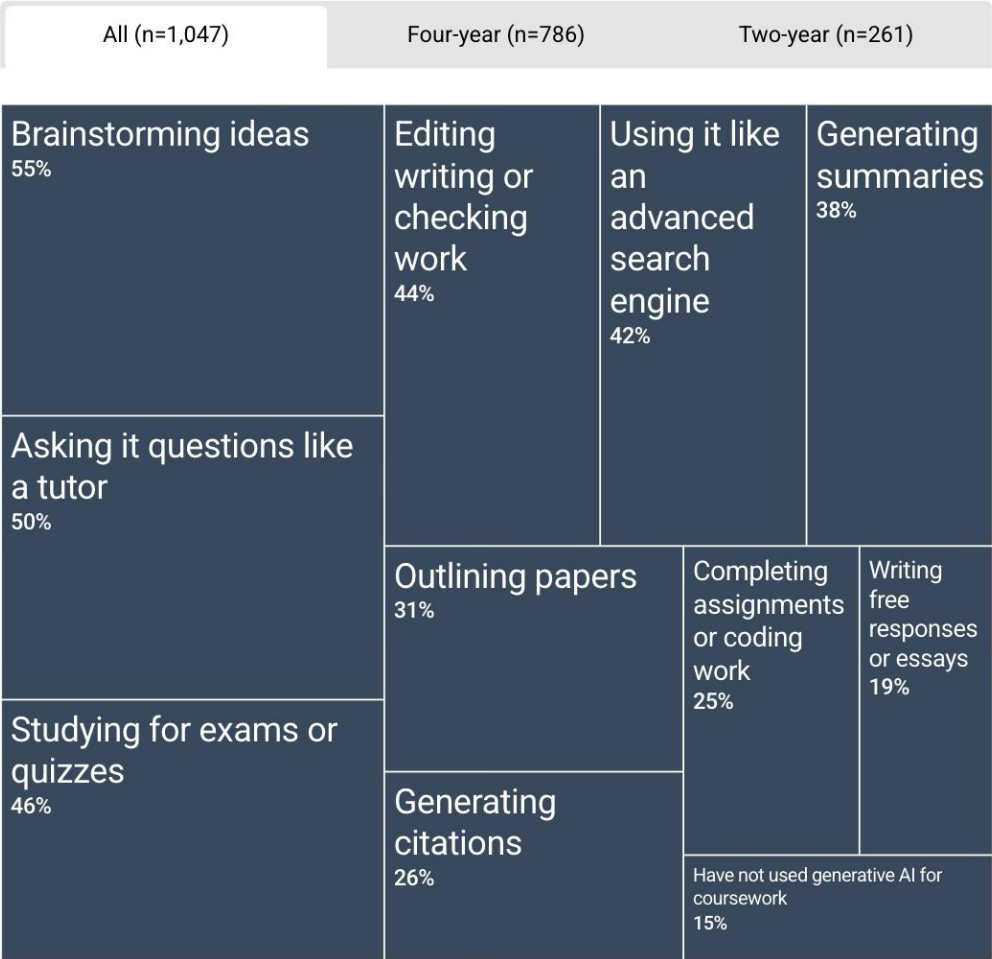
- We encourage in-class interactions!
  - Ask questions. This is not an online class. Your tuition covers lectures, so enjoy it. Also office hours.
- For your projects, you have the option to do it in a group. Meet people in class and vibe check them.
- There will be in-class quizzes (beginning or end of a lecture).

# About ChatGPT...



# Using Generative AI for Coursework

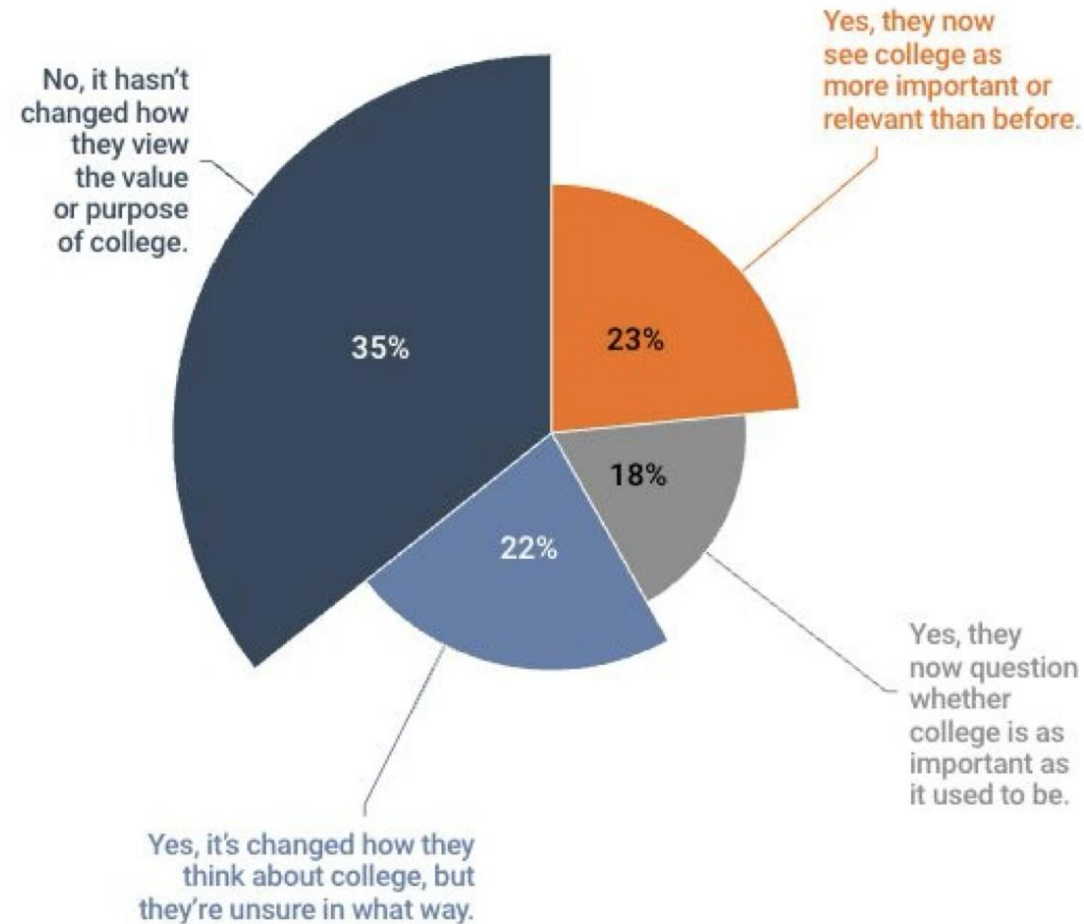
Students on just how they've used generative AI for coursework in the last year, all and by institution type



**Source:** Student Voice flash survey on AI, July 2025 • n=1,047 • Inside Higher Ed x Generation Lab. **Question:** In which of the following ways have you used generative AI (e.g., ChatGPT) for your coursework in the past year? Select all that apply. **Note:** Up to 2% in each group chose other.

# AI and College Value

Students (all) say whether/how the rise of generative AI has changed how they think about the value or purpose of college



**Source:** Student Voice flash survey on AI, July 2025 • n=1,047 • Inside Higher Ed x Generation Lab. **Question:** Has the rise of generative AI (such as ChatGPT) influenced how you think about the value or purpose of a college education? **Note:** 2% chose other.



# AI tools are great, but it's crucial to remain an independent thinker

- Critical thinking skill is important, especially when you are exploring a new area of study.
- AI is not always right.
  - It can easily trick you if you don't already know the answer.
  - ChatGPT is *confident* about factually wrong answers too. ([Simhi et al.](#))



Now, let's talk about language  
models

Human language

Linguistics

Social science: psychology,  
sociology, cultural studies

# Natural Language Processing

Engineering

Computing

Information theory

Statistics

Signal processing

# Earliest records of written languages

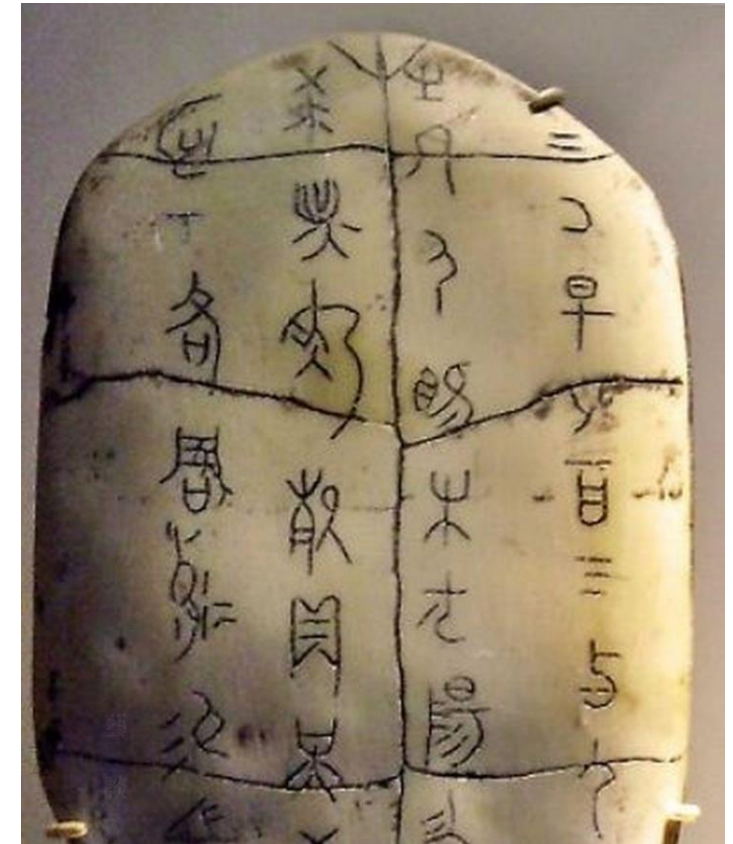


Egyptian, c. 2690 BC  
(About 4700 years ago)



Sumerian, c. 2600 BC

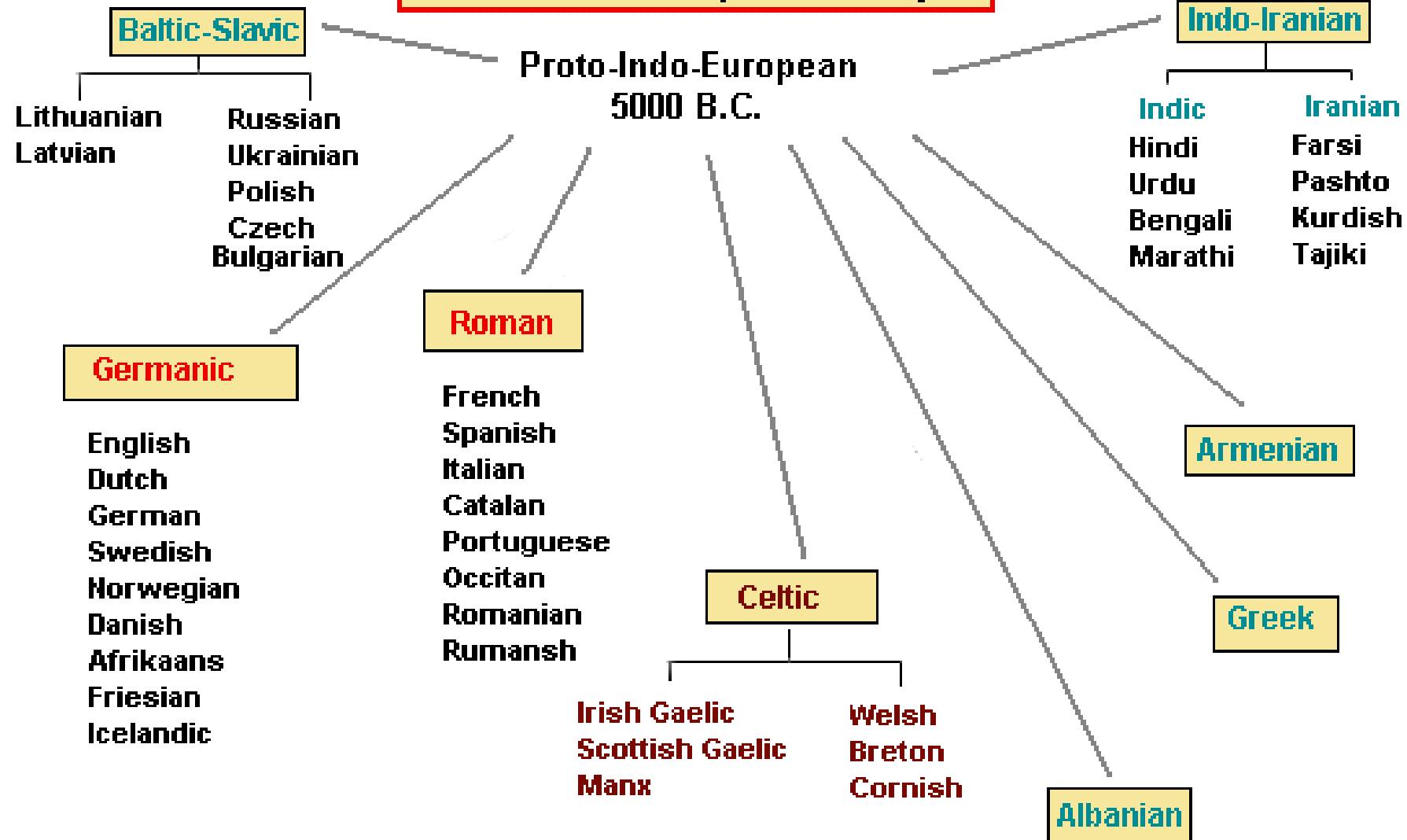
Old Chinese, c. 1250 BC



# Why is language hard to process/study computationally?

- Language is messy and diverse.
  - Multilinguality: There are about 7,000 human languages.

# The Indo-European Family



# Why is language hard to process/study computationally?

- Language is messy and diverse.
  - Multilinguality: There are about 7,000 human languages.
  - Each language possibly has multiple **dialects** (English, American vs British), and many **sociolects** (Gen Z vs Gen Alpha)



# Cookout

---

Article [Talk](#)

---

From Wikipedia, the free encyclopedia

**Cookout** may refer to:

- Social gathering for [grilling](#) (or [barbecuing](#)) outdoors, used primarily in the Southern United States and among Black Americans

# Why is language hard to process/study computationally?

- Language is messy and diverse.
  - Multilinguality: There are about 7,000 human languages.
  - Each language possibly has multiple **dialects** (English, American vs British), and many **sociolects** (Gen Z vs Gen Alpha)
  - Humans **miscommunicate** all the time.

iMessage  
Today 12:20 PM

I am here for you

Thanks :) I'm going through a  
tough time so it means a lot

And sorry, I lost all my contacts  
who is this?

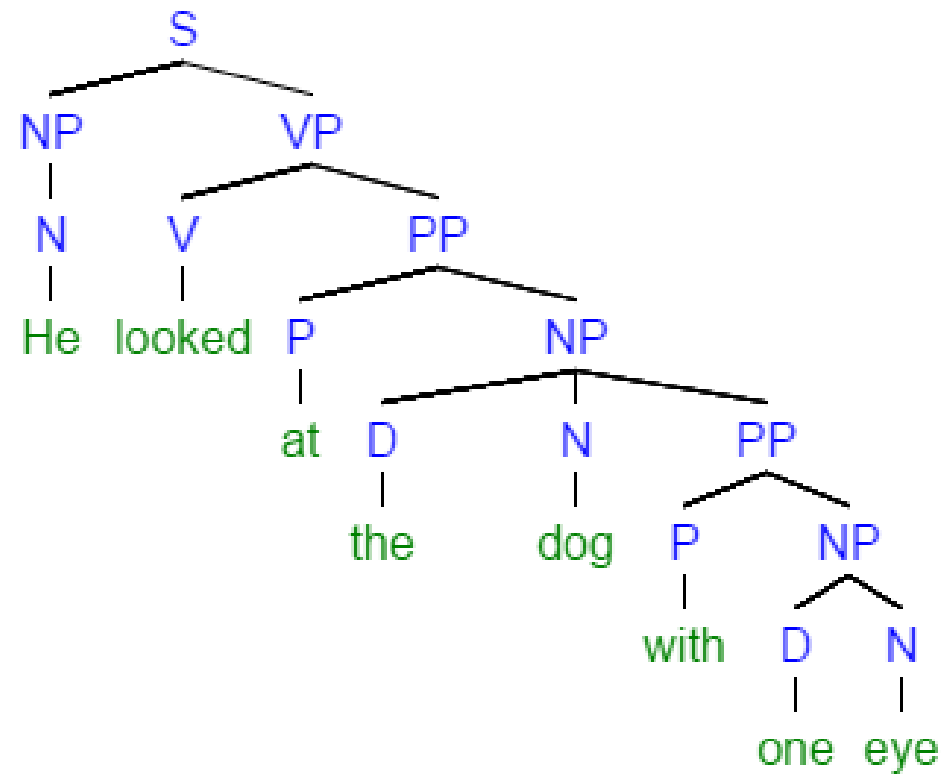
This is your Uber driver

I am here to pick you up

Oh

Delivered

# Ambiguity in language: garden path sentences



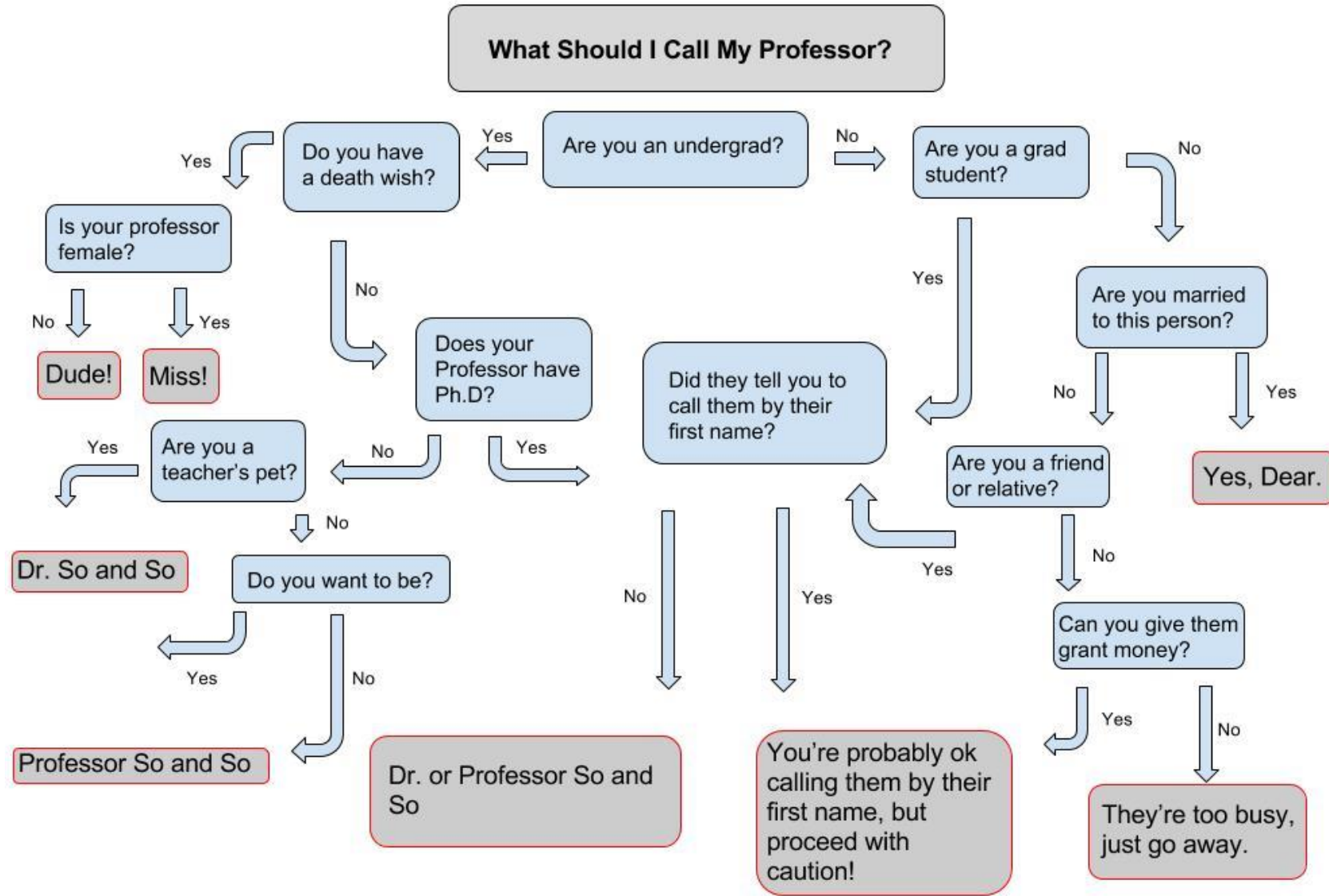
# Why is language hard to process/study computationally?

- Language is messy and diverse.
  - Multilinguality: There are about 7,000 human languages.
  - Each language possibly has multiple **dialects** (English, American vs British), and many **sociolects** (Gen Z vs Gen Alpha)
  - Humans **miscommunicate** all the time.
  - Social context: We use different **style** of language in different social occasion: texting mom vs in-court vs answering business email.

# Formal vs informal language

- Word choice
  - Informal: The study checked out the health effects of passive smoking.
  - Formal: The study examined the health effects of passive smoking.
- Complete rephrasing:
  - Informal: This experiment worked out just fine.
  - Formal: This experiment was successful.
- Contraction
  - Informal: The outcomes of the study haven't been documented yet.
  - Formal: The outcomes of the study have not been documented yet.

## Social convention and etiquette

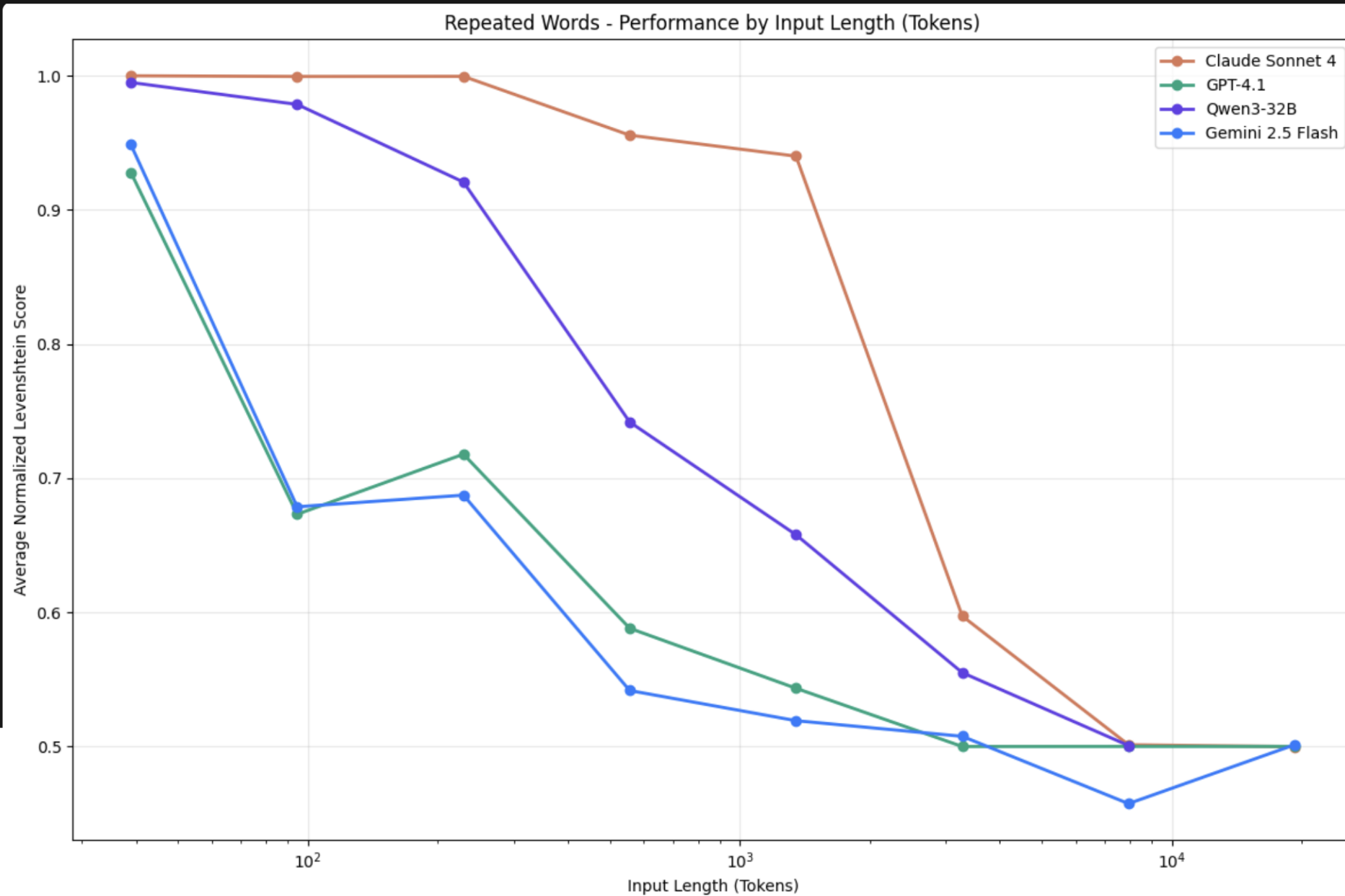


# Why is language hard to process/study computationally?


- Language is messy and diverse.
  - Multilinguality: There are about 7,000 human languages.
  - Each language possibly has multiple **dialects** (English, American vs British), and many **sociolects** (Gen Z vs Gen Alpha)
  - Humans **miscommunicate** all the time.
  - Social context: We use different **style** of language in different social occasion: texting mom vs in-court vs answering business email.
  - Various context length: a text, a paragraph, a chapter, a book...



# Context Rot: How Increasing Input Tokens Impacts LLM Performance




Claude Sonnet 4, GPT-4.1, Qwen3-32B, and Gemini 2.5 Flash on Repeated Words Task

 **SnoopsBadunkadunk** • 5h ago

An AI trained on the Reddit data set answering your questions, wha amirite. Well, don't use it as your therapist, that's for sure.

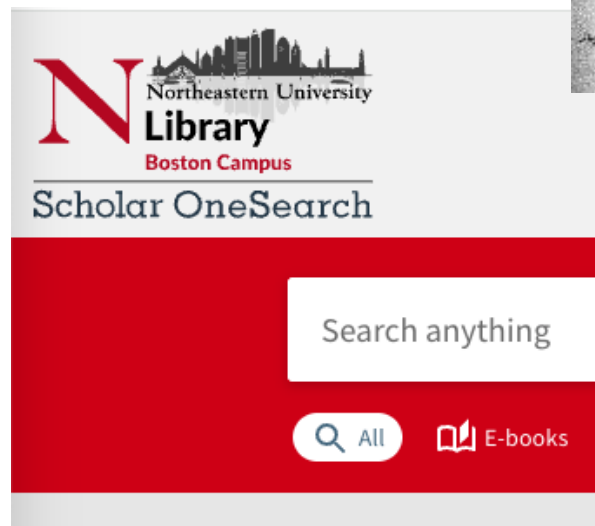
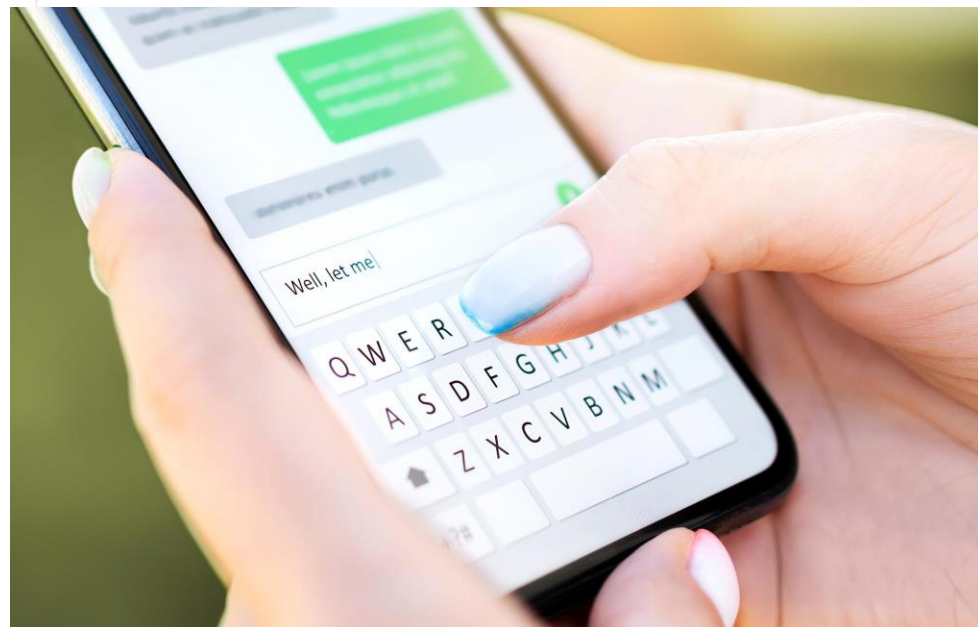
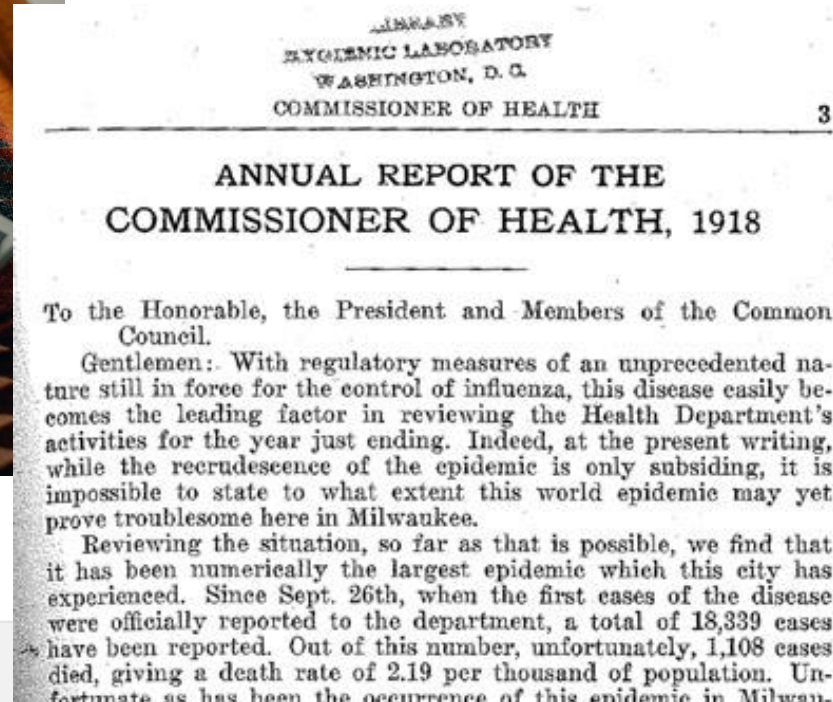
74   Reply   Share

 **Mlakuss** • 4h ago

- "I have a problem in my relationship"
- "leave it"

31   Reply   Share

+ 1 more reply



# Text comes in different sources

- Books
- News articles
- Social media posts
- Lyrics
- Movie scripts
- Texts
- Government: legislation, census data, filings, court recordings
- Business: internal comm (emails, reports), customer data (invoice, logs)
- Academia (publications: papers, books, reports)

# Text comes in different format

- Printed
  - Books
  - Pamphlets
  - Newspapers
- Digitized
  - Archived books in pdf, html, txt, json, jpeg, etc.
  - Archived newspapers
  - Records of social media postings (Reddit, Twitter)
  - Audio transcriptions
  - Image captions

# We still don't really understand how human uses language entirely

- The linguistics war: form (syntax) first or meaning (semantics) first?
  - Chomsky (internalist, generative grammar -> later generative syntax):  
language is acquired from an innate brain structure, syntax is independent of meaning
  - Lakoff (internalist, generative semantics -> later cognitive linguistics):  
language is rooted in meaning, structures of syntax emerge from

# We still don't really understand how human uses language entirely

- Neuroscience:
  - Classical (Broca/Wernicke): language is localized in distinct brain areas
  - Modern: language relies on a distributed network in the frontal, temporal, parietal, and subcortical regions
- Mostly agree that language in the left hemisphere. Disagreements about what right hemisphere do.
- People with **aphasia** have damaged language network in the brain. They have trouble express and understand written and spoken language.



# Biological/ Anatomical:

## Broca's area

**Broca's area** is an important part of language formulation — even if someone has the motor ability to form the sounds necessary for words, Broca's area is necessary to form and express **language**.

## Angular gyrus

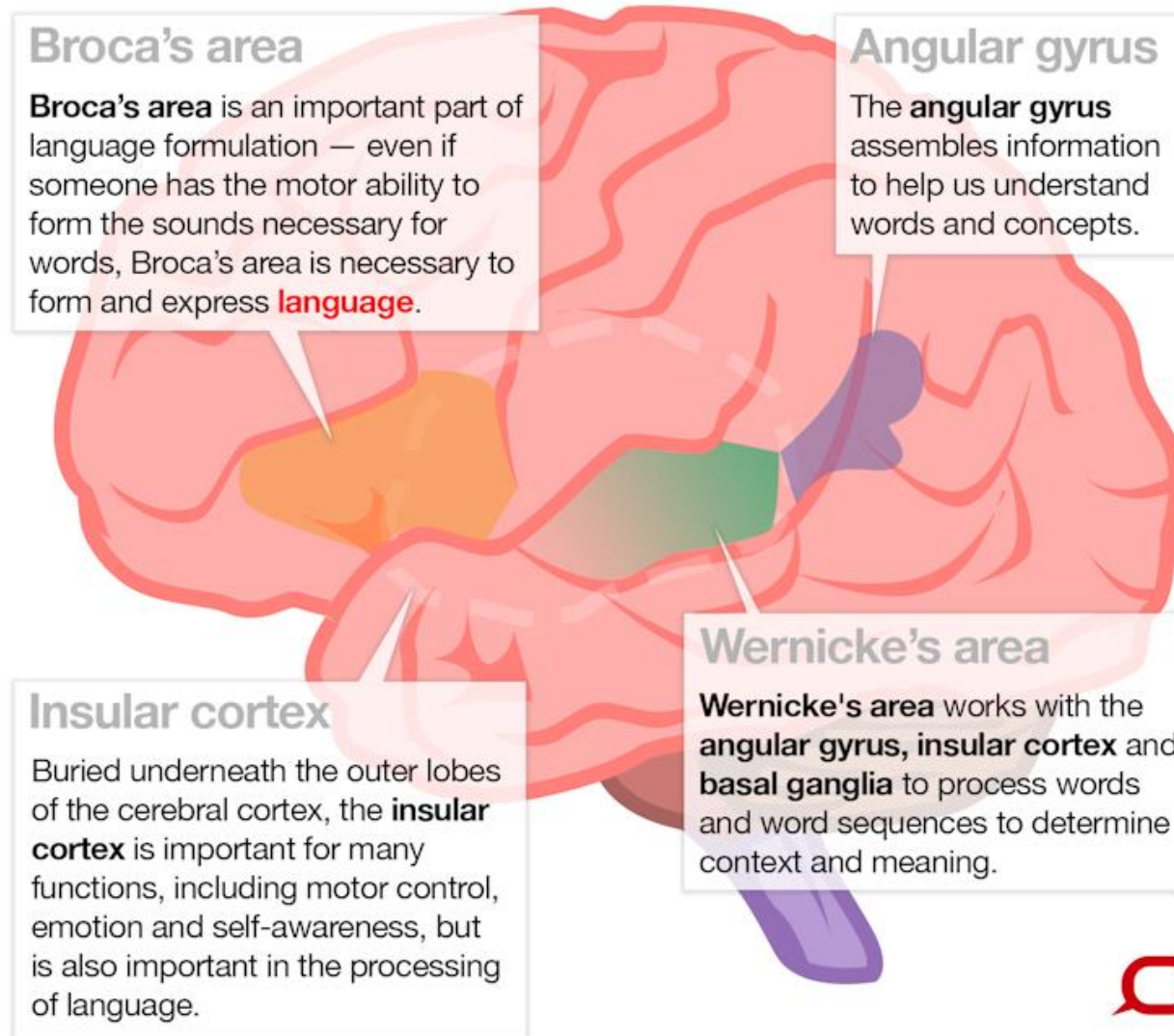
The **angular gyrus** assembles information to help us understand words and concepts.

## Insular cortex

Buried underneath the outer lobes of the cerebral cortex, the **insular cortex** is important for many functions, including motor control, emotion and self-awareness, but is also important in the processing of language.

## Wernicke's area

**Wernicke's area** works with the **angular gyrus**, **insular cortex** and **basal ganglia** to process words and word sequences to determine context and meaning.

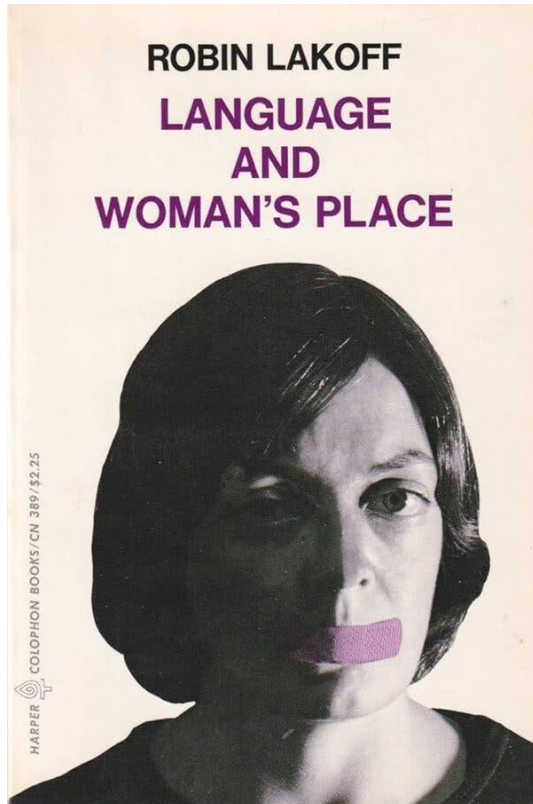


# Organization of language functions

- View 1: Language is modular, with syntax, semantics, and phonology tied to specific cortical regions.
- View 2: Language is distributed and dynamic, with regions flexibly contributing depending on context and task.



# Social/cultural:



PubMed®

Advanced

Save

Email

Meta-Analysis > Pers Soc Psychol Bull. 2024 Mar;50(3):371-386.

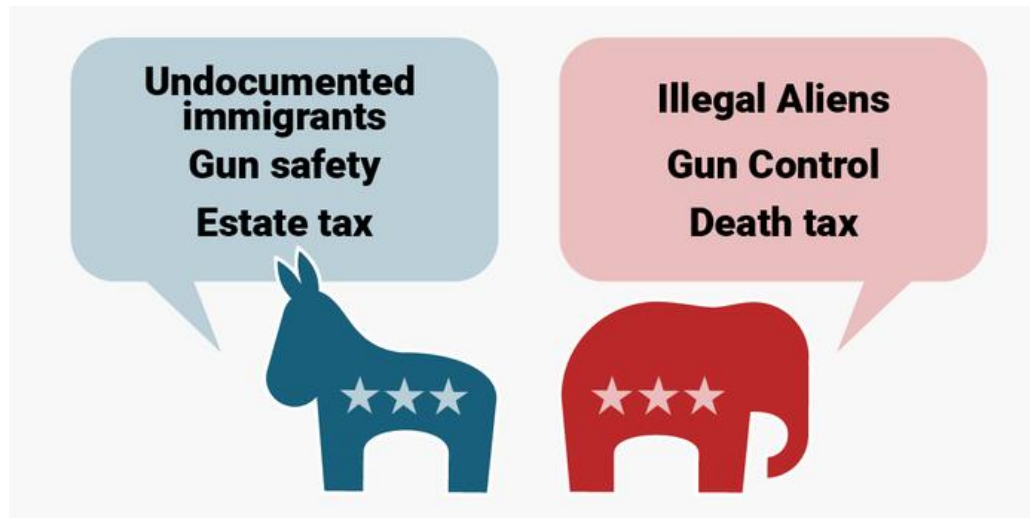
doi: 10.1177/01461672221130595. Epub 2022 Nov 3.

## Is Your Accent Right for the Job? A Meta-Analysis on Accent Bias in Hiring Decisions

Jessica L Spence<sup>1</sup>, Matthew J Hornsey<sup>1</sup>, Eloise M Stephenson<sup>1</sup>, Kana Imuta<sup>1</sup>

Affiliations + expand

PMID: 36326202 DOI: [10.1177/01461672221130595](https://doi.org/10.1177/01461672221130595)



# We created language models

- To process language
- To generate language
- To study how humans communicate

# Noisy channel model

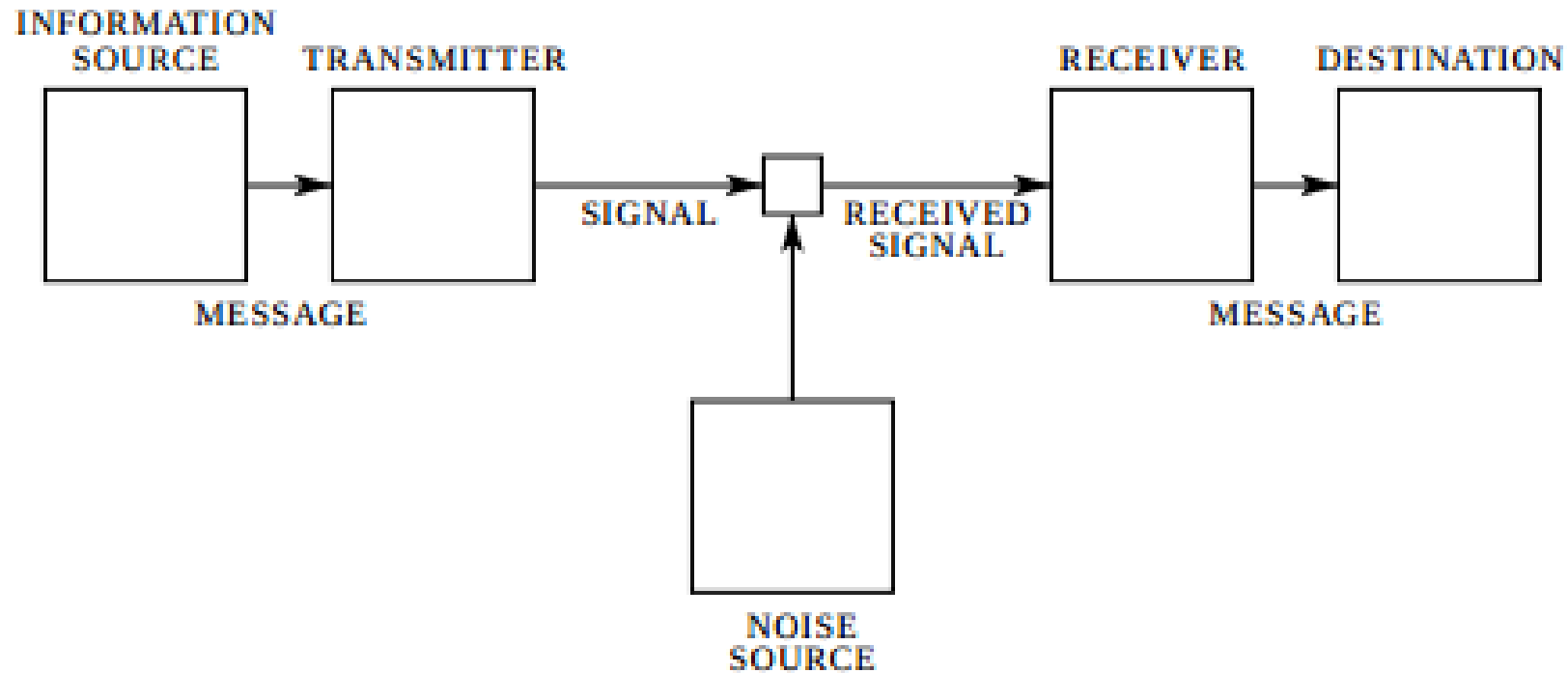
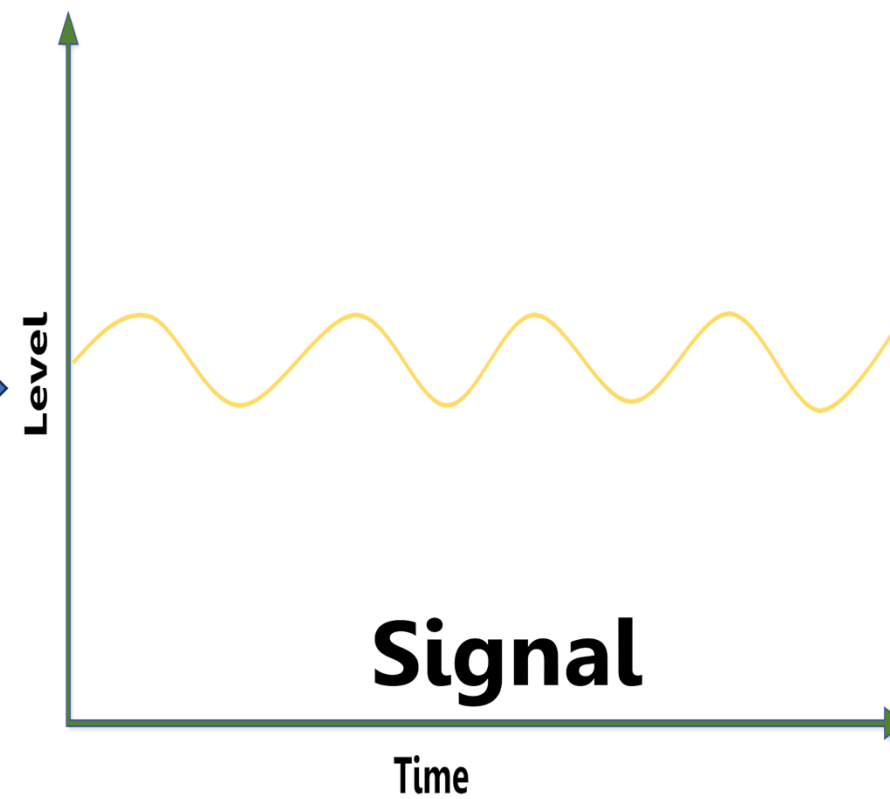
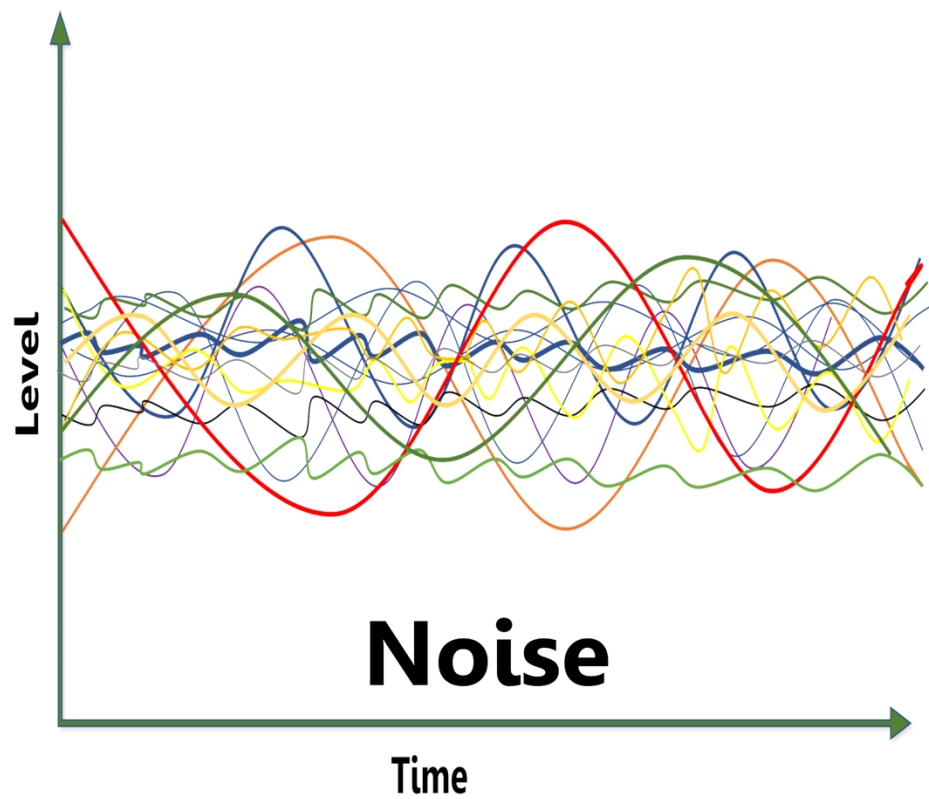


Fig. 1—Schematic diagram of a general communication system.



TO X05

FROM

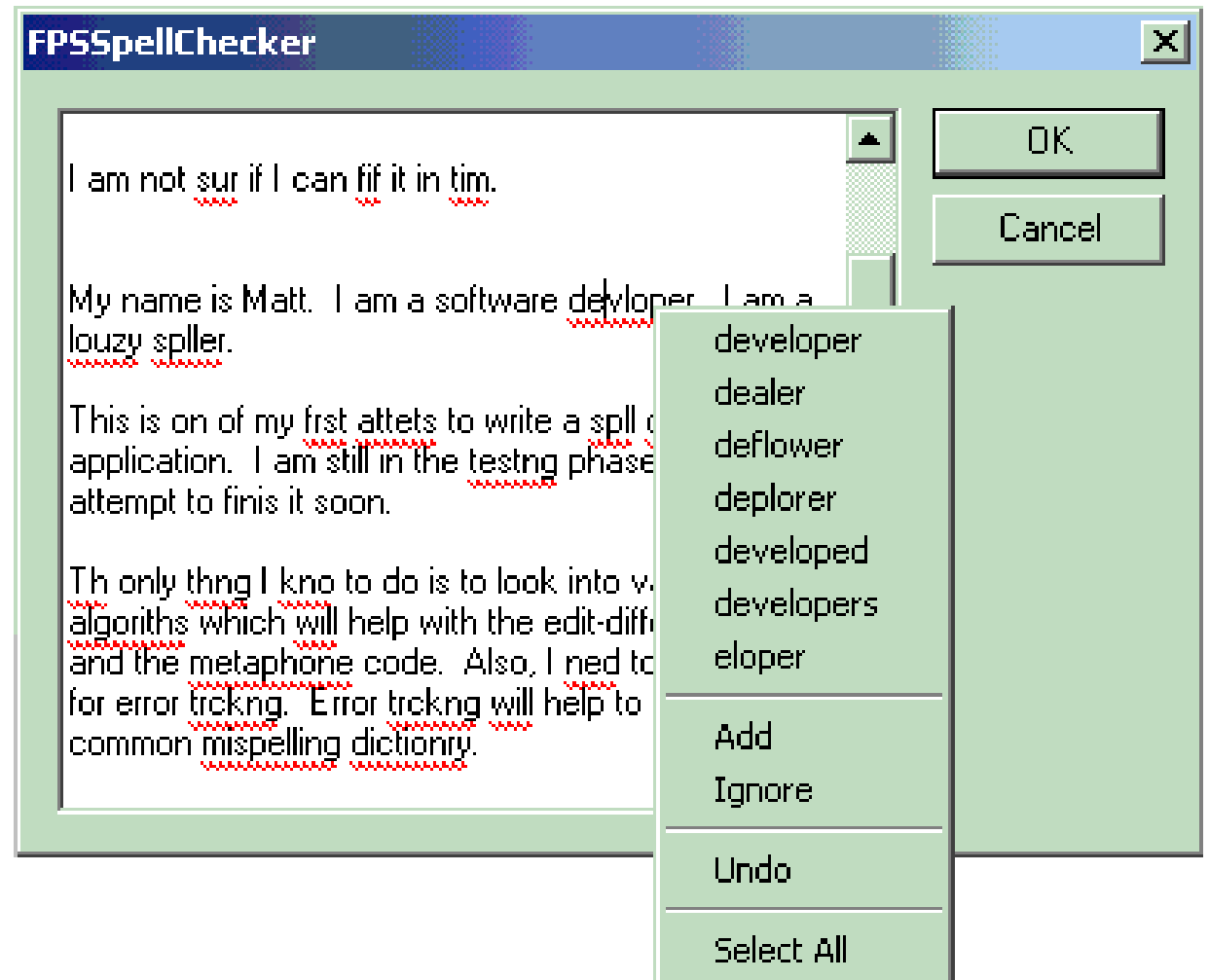
Originator's No.	Date	In reply to No.

AOKN HEPD FNET YIDDC  
 TQAR DJAP FUVFN YIAFF  
 PABAZ WTHP CIPNW WTRZH  
 ALXKE YEMH CWOIB AKKEE  
 HAOA RQPH DJCHM TPZEH  
 LKFFH RCHT TLECY FNKTQ  
 KLDTS FQIRU AOKN 27 1525/6

NURP 40TW 194  
 NURP 37OK 76

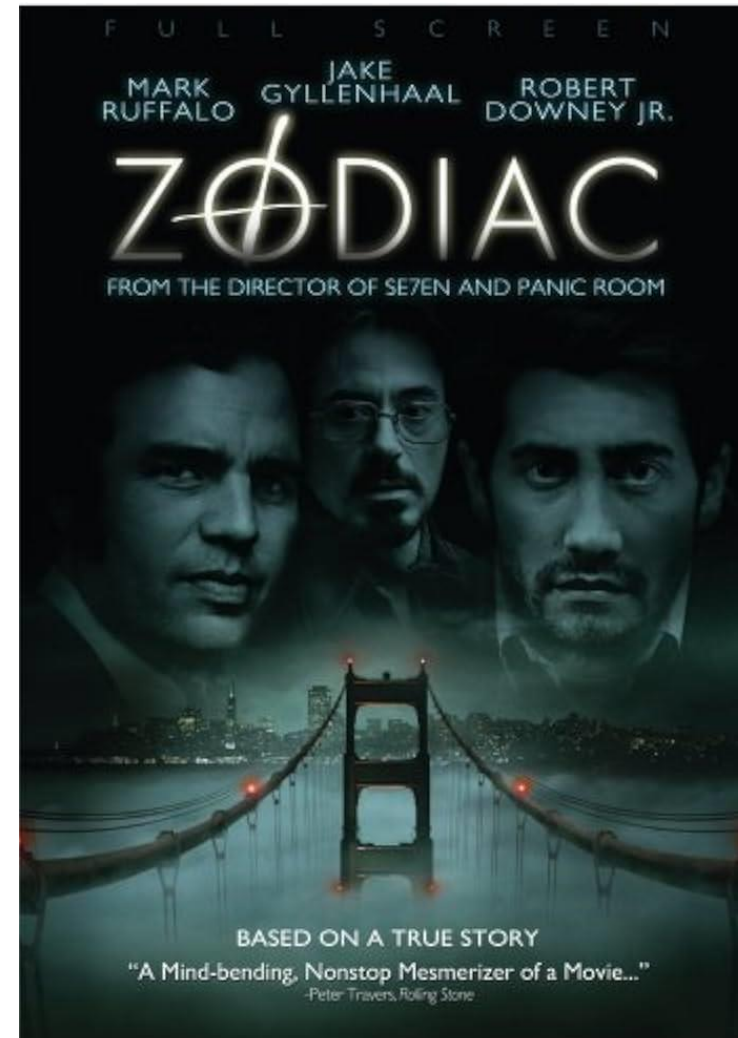
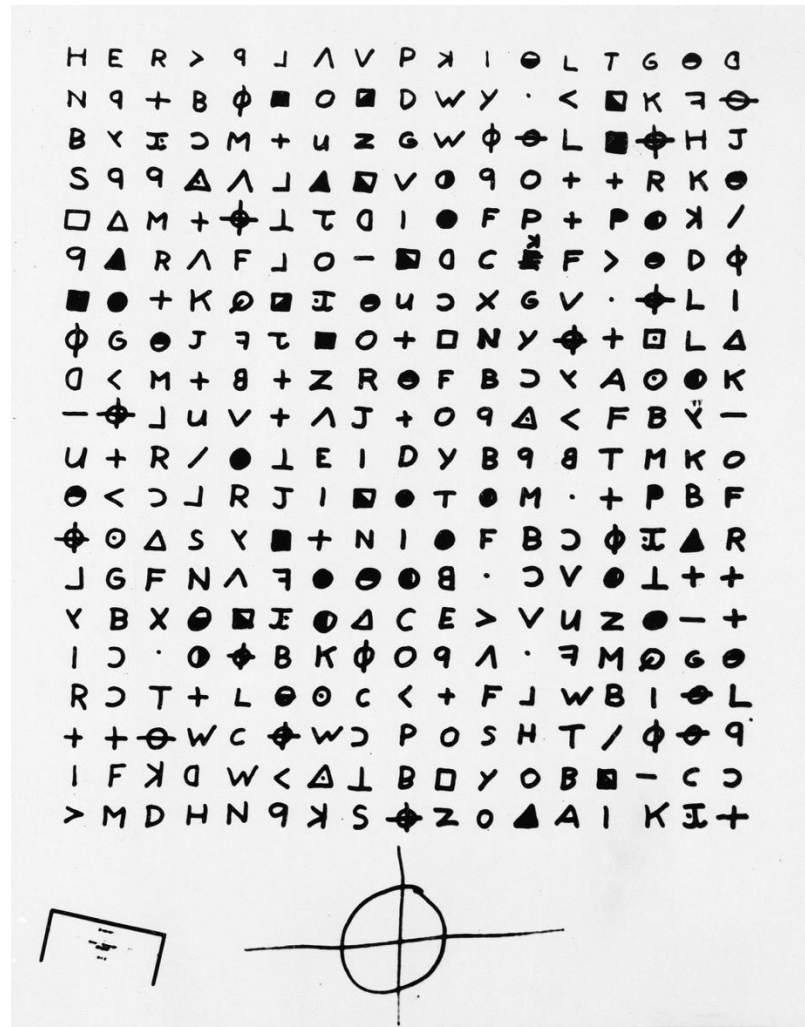


Dear  
autocorrect.  
It's never  
duck.





# Deciphering



# Shannon & Entropy

To quantify **uncertainty** in predicting a message.

$$H(X) := - \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

## Prediction and Entropy of Printed English

By C. E. SHANNON

(Manuscript Received Sept. 15, 1950)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

### 1. INTRODUCTION

An extension of the above experiment yields further information concerning the predictability of English. As before, the subject knows the text up to the current point and is asked to guess the next letter. If he is wrong, he is told so and asked to guess again. This is continued until he finds the correct letter. A typical result with this experiment is shown below. The first line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

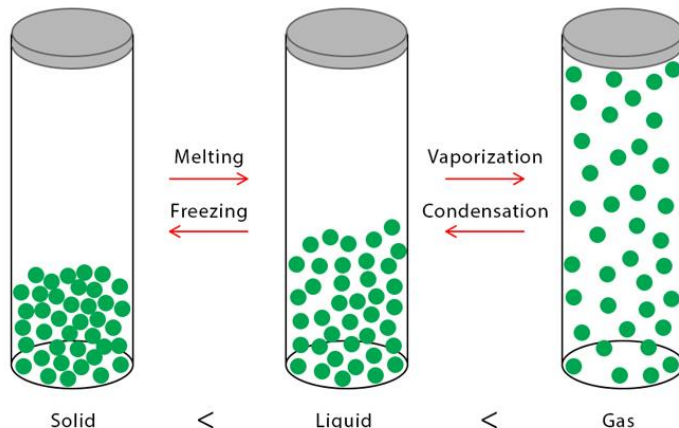
(1)	T	H	E	R	E		I	S		N	O		R	E	V	E	R	S	E		O	N		A		M	O	T	O	R	C	Y	C	L	E		A
(2)	1	1	1	5	1	1	2	1	1	2	1	1	15	1	17	1	1	1	2	1	3	2	1	22	7	1	1	1	1	4	1	1	1	1	3	1	
(1)	F	R	I	E	N	D		O	F		M	I	N	E		F	O	U	N	D		T	H	I	S		O	U	T								
(2)	8	6	1	3	1	1	1	1	1	1	1	1	1	1	6	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1			
(1)	R	A	T	H	E	R		D	R	A	M	A	T	I	C	A	L	L	Y		T	H	E		O	T	H	E	R		D	A	Y				
(2)	4	1	1	1	1	1	1	1	1	1	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			

(9)

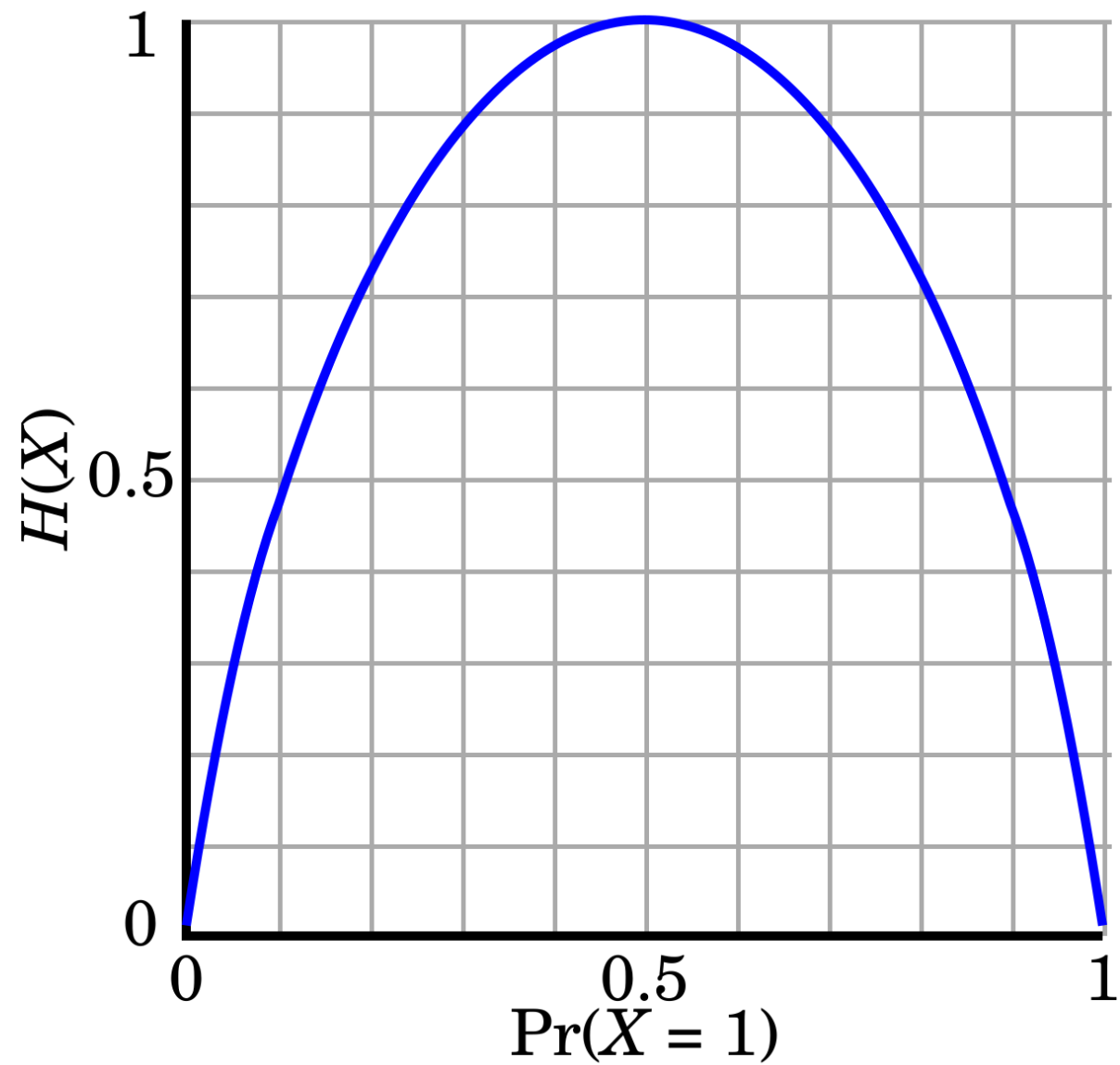
### Entropy

Entropy is the measure of the disorder of a system

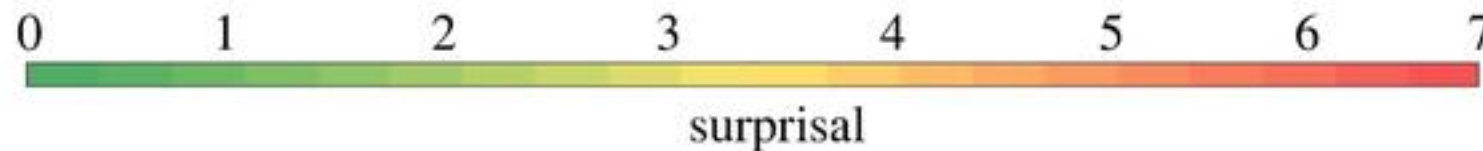
Entropy increases







Binge ... on | - | and | of | is  
 Binge **drinking** ... is | and | had | in | was  
 Binge drinking **may** ... be | also | have | not | increase  
 Binge drinking may **not** ... be | have | cause | always | help  
 Binge drinking may not **necessarily** ... be | lead | cause | results | have  
 Binge drinking may not necessarily **kill** ... you | the | a | people | your  
 Binge drinking may not necessarily kill **or** ... even | injure | kill | cause | prevent  
 Binge drinking may not necessarily kill or **even** ... kill | prevent | cause | reduce | injure  
 Binge drinking may not necessarily kill or even **damage** ... your | the | a | you | someone  
 Binge drinking may not necessarily kill or even damage **brain** ... cells | functions | tissue | neurons  
 Binge drinking may not necessarily kill or even damage brain **cells,** ... some | it | the | is | long



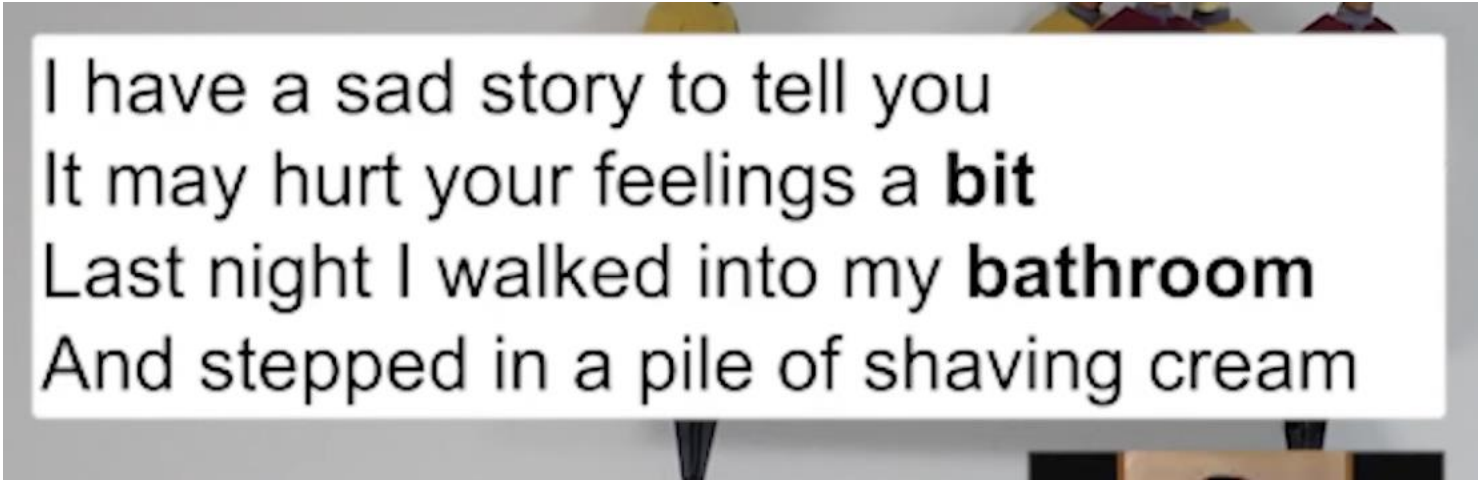
I have a sad story to tell you  
It may hurt your feelings a **bit**  
Last night I walked into my **bathroom**  
And stepped in a pile of sh

So with all that in mind, take a  
guess about what the next letter is ...

## The Shannon Game, Entropy, and the Power of Predicting Language [Lecture]

<https://www.youtube.com/watch?v=0shft1gokac>

# Surprise



I have a sad story to tell you  
It may hurt your feelings a **bit**  
Last night I walked into my **bathroom**  
And stepped in a pile of shaving cream

- <https://www.ccs.neu.edu/home/dasmith/courses/cs6120/shannon/>

# Naive Bayes

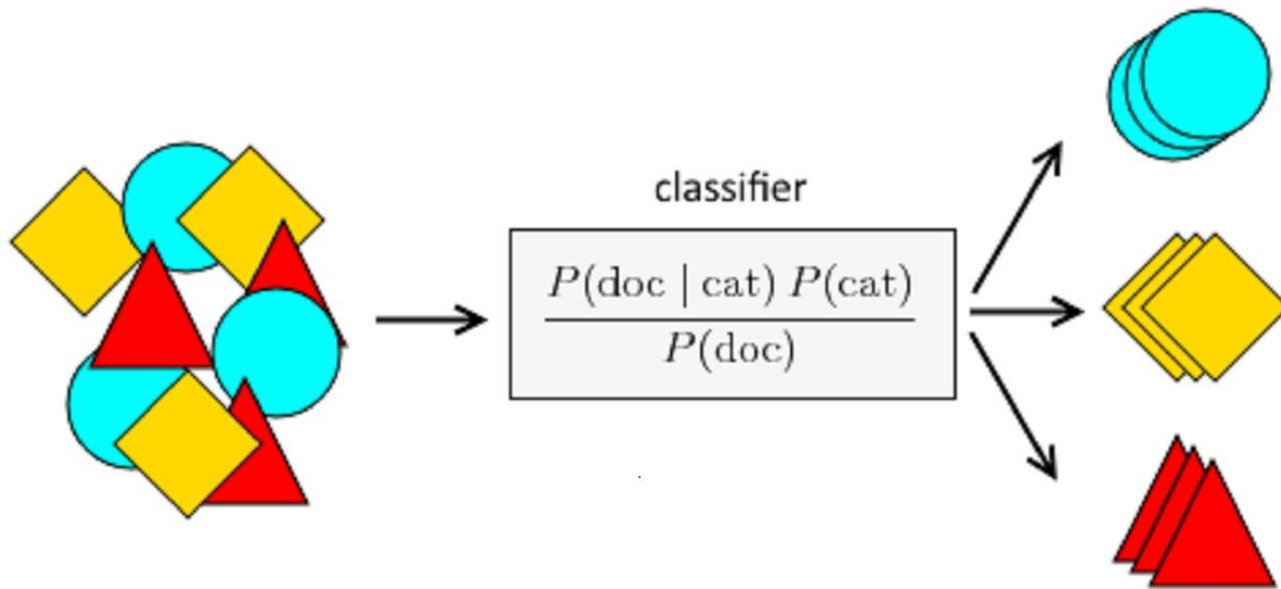
Likelihood of the Evidence given that the Hypothesis is True

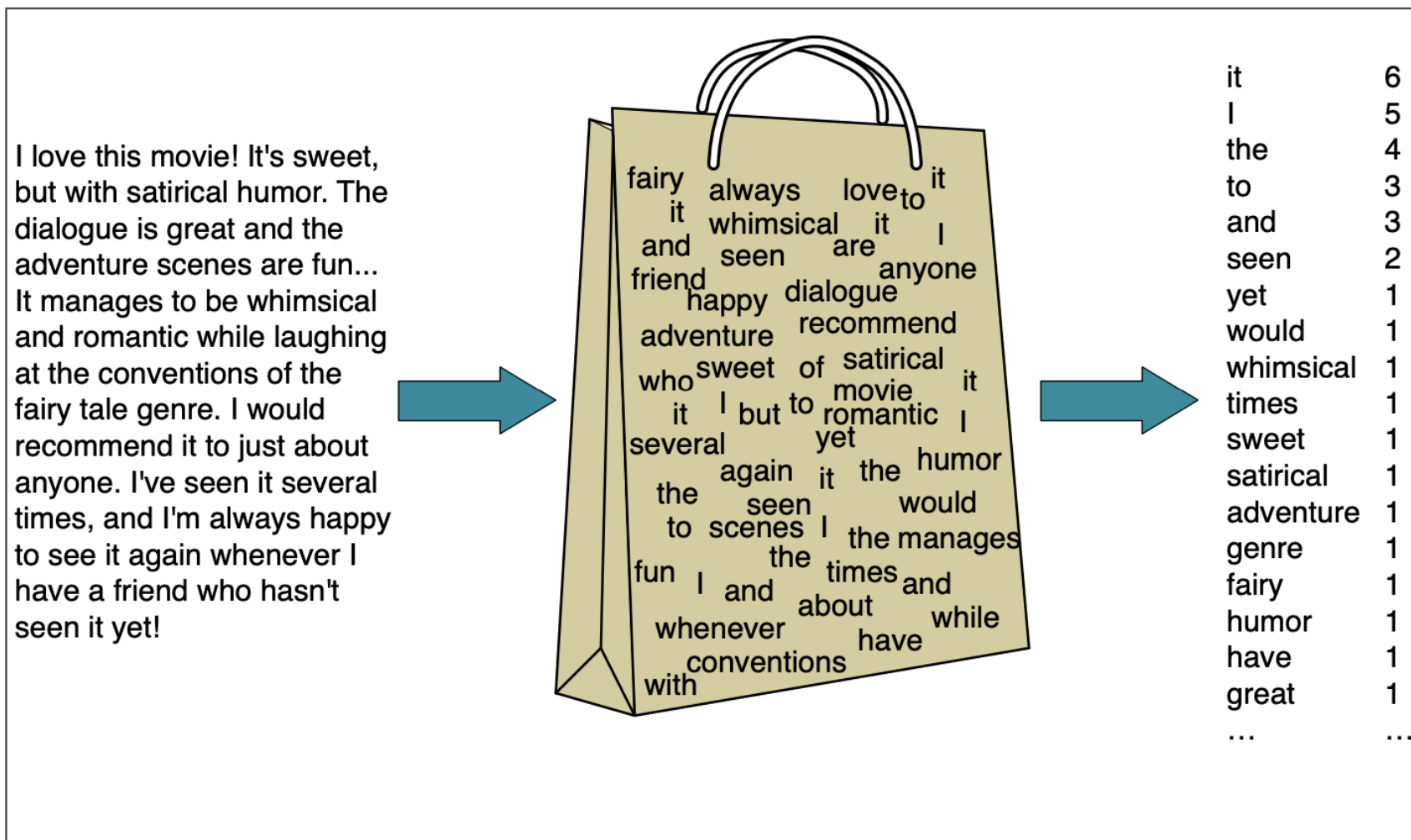
Prior Probability of the Hypothesis

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True



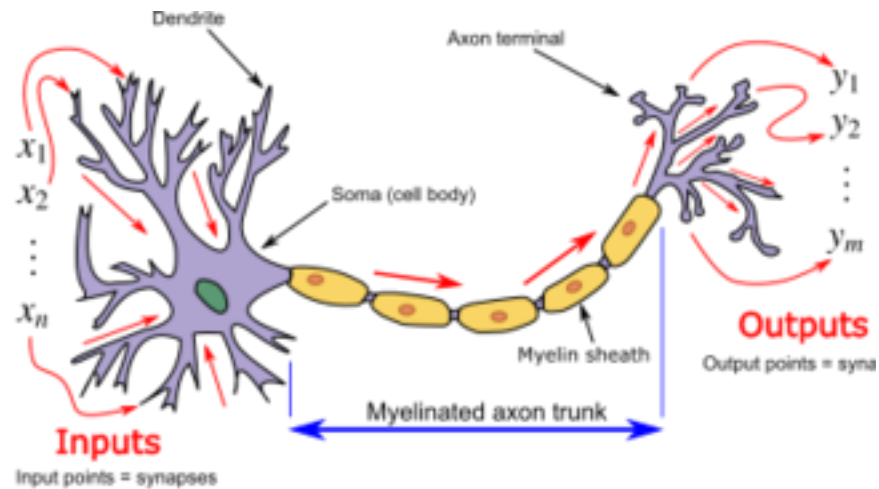
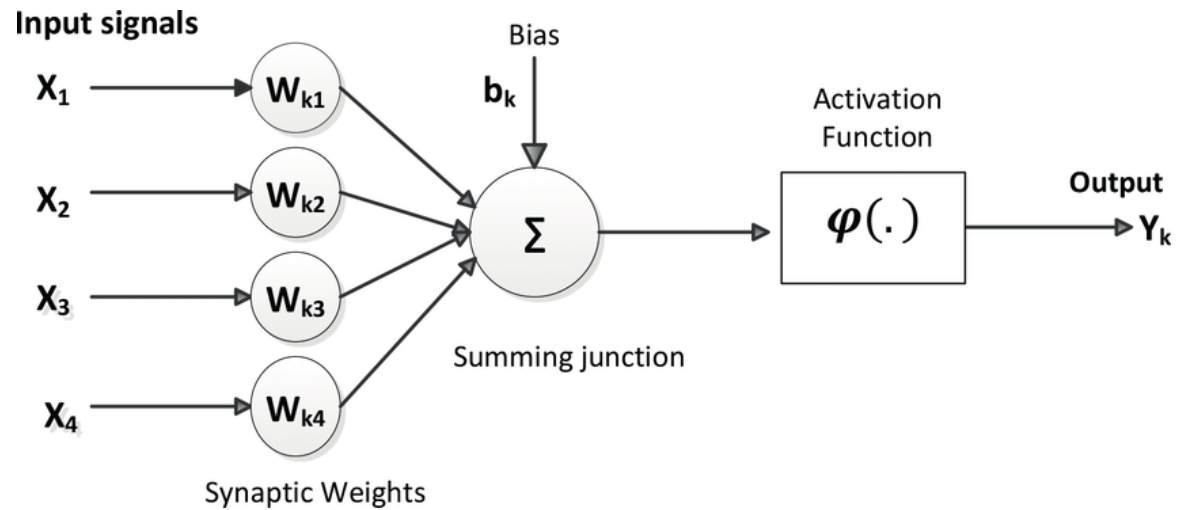
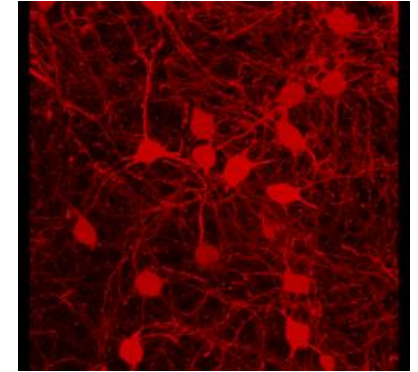


$$\hat{P}(w_i|c) = \frac{\text{count}(w_i, c)}{\sum_{w \in V} \text{count}(w, c)}$$

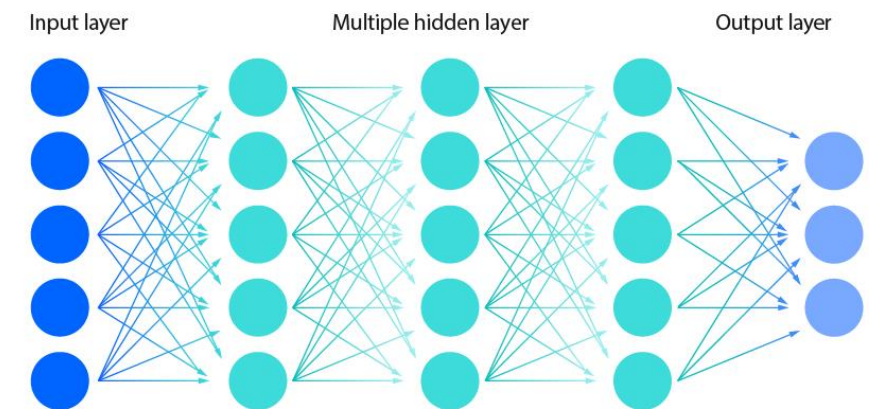
**Figure 4.1** Intuition of the multinomial naive Bayes classifier applied to a movie review. The position of the words is ignored (the *bag of words* assumption) and we make use of the frequency of each word.



# Neural networks



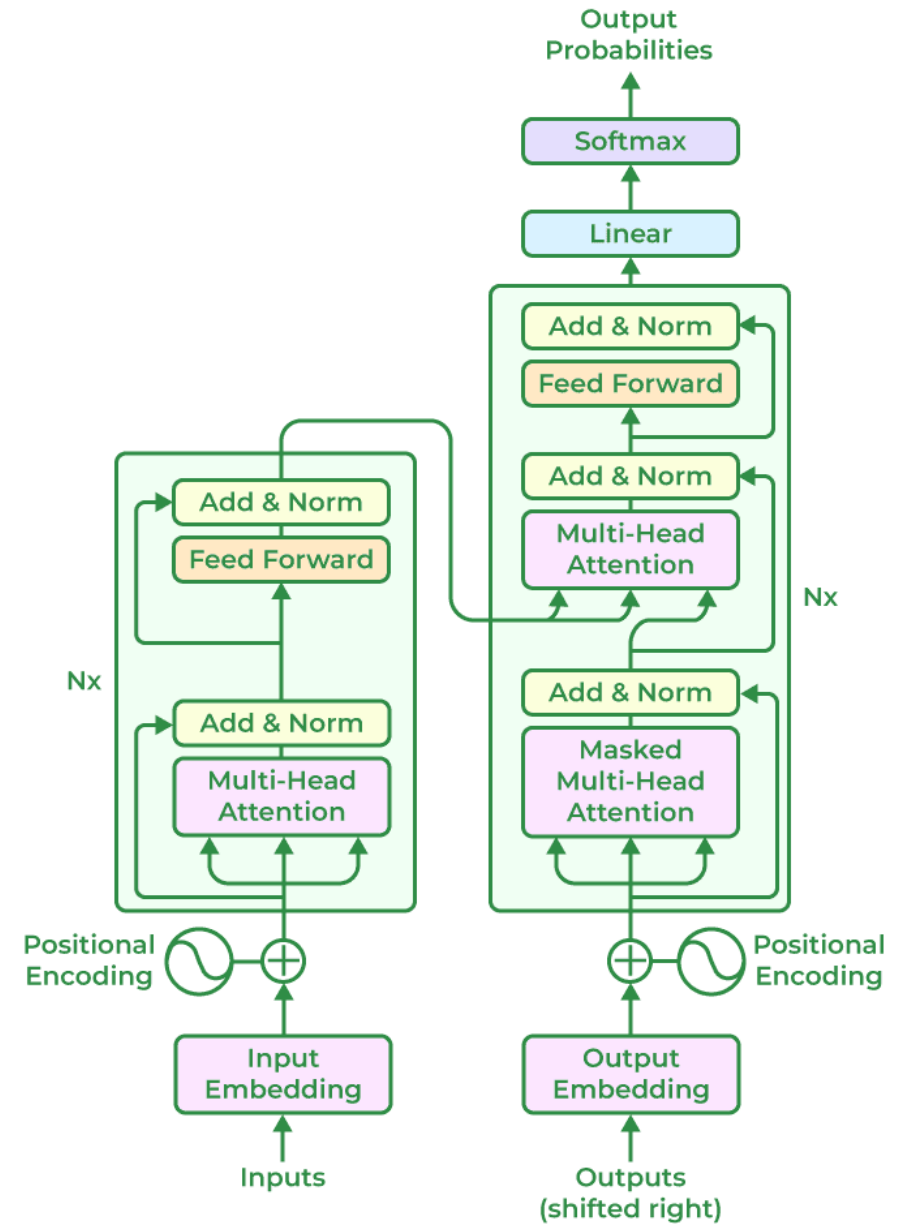
## Deep neural network





# Large language models

- Still statistical models
- Still dependent on training data



# All language models are statistical

- The performance depends on the data its trained on
  - The text source matters: quality, quantity, style, language coverage

# All language models are statistical

- The performance depends on the data its trained on
  - The text source matters: quality, quantity, style, language coverage
- All learn language in terms of probability distribution over sequences (tokens, words, sentences, etc.)

# All language models are statistical

- The performance depends on the data its trained on
  - The text source matters: quality, quantity, style, language coverage
- All learn language in terms of probability distribution over sequences (tokens, words, sentences, etc.)
- Poor at generating or even recognizing rare or unseen data
  - Humans are much better at extrapolation/ few-shot generation
    - We learn language without a whole million training data

# All language models are statistical

- The performance depends on the data its trained on
  - The text source matters: quality, quantity, style, language coverage
- All learn language in terms of probability distribution over sequences (tokens, words, sentences, etc.)
- Poor at generating or even recognizing rare or unseen data
  - Humans are much better at extrapolation/ few-shot generation
    - We learn language without a whole million training data
- Evaluation of these models are also mostly statistical
  - Difficult to know the real performance if test dataset is biased or small.

# LM in brains and machines

- Human:

- Learn from social interaction, exposure to language (written text, speech, sign language)
- Require much less text to master a language
- Can do next word prediction
- Can learn continuously
- Can make mistakes

- Machines:

- Learn statistical relationship of words from training data (text, speech, images)
- Require tremendous amount of text to master a language
- Can do next word prediction
- Some can learn continuously\*, it's complicated
- Can hallucinate

I don't need a therapist when I have ChatGPT honestly

22h Reply

1,329

I was telling chatgpt my issues that I've told my therapist and it damn near gave me the same advise when I prompted it to advise me like a therapist 😂 I was like Ohh okay

11h Reply

34

I've wondered if I should judge myself but its given me wayyyy better advice than therapists have

11h Reply

4

1h Reply

I literally ... get more support from ai than the real therapist


# OpenAI, Altman sued over ChatGPT's role in California teen's suicide

By Jody Godoy

August 26, 2025 5:46 PM EDT · Updated August 26, 2025



OpenAI CEO Sam Altman attends an event to pitch AI for businesses in Tokyo, Japan February 3, 2025. REUTERS/Kim Kyung-Hoon/File Photo [Purchase Licensing Rights](#)

 Food for thought:  
What does it mean for  
language to sound/be  
more human?

# “I Am the One and Only, Your Cyber BFF”: Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI

State-of-the-art generative AI (GenAI) systems are increasingly prone to anthropomorphic behaviors, i.e., to generating outputs that are perceived to be human-like. While this has led to scholars increasingly raising concerns about possible negative impacts such anthropomorphic AI systems can give rise to, anthropomorphism in AI development, deployment, and use remains vastly overlooked, understudied, and under-specified. In this blog post, we argue that we cannot thoroughly understand the impact of generative AI without understanding the impact of anthropomorphic AI, and outline a call to action.

---

#### AUTHORS

**Myra Cheng**  
**Alicia DeVrio**  
**Lisa Egede**  
**Su Lin Blodgett\***  
**Alexandra Olteanu\***

#### AFFILIATIONS

Stanford University  
Carnegie Mellon University  
Carnegie Mellon University  
Microsoft Research  
Microsoft Research

#### PUBLISHED

April 28, 2025



Abeba Birhane

@Abebab



AGI edition

I'll see myself out now



Please suggest a list of topics students should cover in an NLP class.

Here's a structured list of core and extended topics for a Natural Language Processing (NLP) class, moving from fundamentals to more advanced themes. You could use this as a syllabus backbone or checklist for student learning:

---

### 1. Foundations of NLP

- What is NLP? Scope, applications, and history
- Linguistic basics: morphology, syntax, semantics, pragmatics
- Text preprocessing: tokenization, normalization, stemming, lemmatization
- Representing text: bag-of-words, n-grams, TF-IDF

---

### 2. Classical NLP Methods

- Language modeling with n-grams and smoothing
- Part-of-speech tagging (Hidden Markov Models, Maximum Entropy models)
- Parsing: constituency and dependency parsing
- Named Entity Recognition (NER), chunking, and information extraction
- Sentiment analysis with traditional machine learning

**ChatGPT is too traditionalist. We will cover more recent topics.**



### 3. Statistical and Machine Learning Approaches

**D** Please suggest a list of topics students should cover in an NLP class.

Here's a comprehensive list of topics that should be covered in an NLP (Natural Language Processing) class:

### **Foundational Concepts**

- Introduction to NLP and its applications
- Linguistic fundamentals (morphology, syntax, semantics, pragmatics)
- Text preprocessing (tokenization, normalization, stemming, lemmatization)
- Regular expressions and pattern matching
- N-grams and language modeling basics

### **Traditional NLP Techniques**

- Part-of-speech (POS) tagging
- Named Entity Recognition (NER)
- Parsing (constituency and dependency parsing)
- Word sense disambiguation
- Information extraction
- Text classification (spam detection, sentiment analysis)
- Clustering and topic modeling (LDA, LSA)

### **Machine Learning for NLP**

- Feature engineering for text

Claude is also too traditionalist.



# What we will cover in this class

- Markov (finite-state, n-gram) LMs
- Linear classifiers
- Word embeddings
- Neural classifiers
- Morphology, syntax, and semantics
- Sequences, Attention, and Transformer LMs

# What we will cover in this class

- A Taxonomy of Large Language Models
- Pretraining
- Generation
- Post-training: RLHF, DPO, and friends
- Prompting, In-context Learning
- Benchmarks and Experimental Design

# What we will cover in this class

- Retrieval, Retrieval-Augmented Generation
- Summarization
- Multilinguality and Translation
- Language in Social Context

# Guest lectures

- Alexander Spangher (10/7): narrative and discourse
- Terra Blevins (11/4): multilingual learning
- Niloofar Miresghallah (11/14): security & privacy
- Lucy Li (11/21): computational social science

# Course project

- Groups of 1–4
- Initial pitch
- Research plan
- Sample data for evaluation
- Grade contract
- Presentation
- Final report
- Feedback on each step



# Grading

- Five programming assignments (30%) - Individually
- Six quizzes (30%) - Individually
- Course project (40%) - In groups of 1– 4

# To do:

- Join **Gradescope**, code to join **NGZDZP**
- Join **Ed Discussion**
  - Link is on the class website
  - I will make a post later for you to suggest topics to cover in our last lecture of the semester
- **Class website** to bookmark: <https://siwu.io/nlp-class>