WELCOME TO NATURAL LANGUAGE PROCESSING IN HUMAN AND MACHINES

About us

Anuja Negi

PhD candidate

Denizens Lab

TU Berlin

Mathis Lamarre

PhD candidate

Denizens Lab

TU Berlin

Subba Reddy Oota

PostDoc

Denizens Lab

TU Berlin

Prof. Dr. Fatma Deniz

Principal Investigator

Denizens Lab

TU Berlin

Organisation

Language	English
Credits	3 ECTS
Lecture Period	7-10 Oct and 18 Oct, 2024
Time	10am-5pm (lunch: 1-2pm)
Location	MAR 5.044
Course Website*	https://denizenslab.github.io/teaching/seminars/NLP_HM/
ISIS	https://isis.tu-berlin.de/course/view.php?id=40430

*first place to check for all information

Organisation: Plan

01	**Oct 7, 2024**	• NLP Basics
02	**Oct 8, 2024**	• NLP advanced
03	**Oct 9, 2024**	• NLP in Machines • How to read a paper
04	**Oct 10, 2024**	• NLP in humans • How to present
05	**Oct 18, 2024**	• Presentation Day!

Organisation: Presentation day

Date	October 18th, 2024
Duration	20mins per group (+10min Q/A)
Grading	20% participation, 40% presentation, 40% report
Report	Submit by 1st Nov (max. 4 pages)

Organisation: Participants

Limited to 20.

- Attend first lecture to activate registration.
- Random selection based on TU rules.

Block Seminar?

- Intensive full days
  - lecture
  - mini-tutorials
- Presentation + report
  - in groups of 2/alone
  - present together on the last day
  - submit a report (commentary, put it in context + ideas 💡)

Our class will make use of Google Collab for the tutorials.

*BRING YOUR LAPTOPS TO CLASS!!!

*If this is not possible, please inform me. We can look for a solution together :)

Learning Cooperatively

- Discuss with with your classmates!
- Ask questions.
- Ask for help!
- Please do NOT blindly copy the answers during the tutorial breakout sessions from each other. It is important to try and do it yourself.

Questions?

ask whenever you have them

Natural Language Processing is interdisciplinary

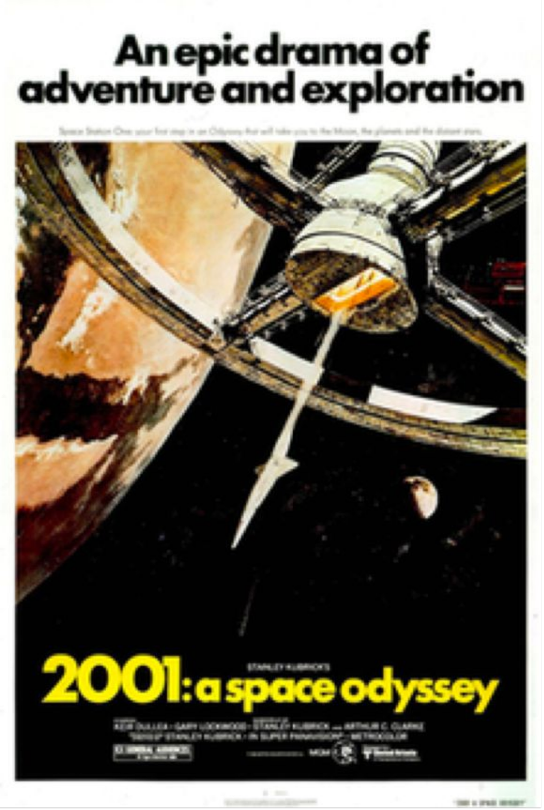
Artificial intelligence

Machine learning

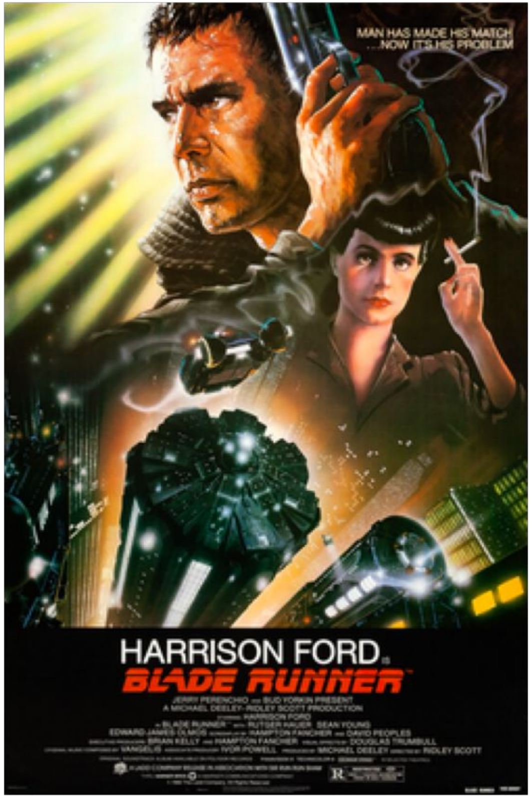
Linguistics

Cognitive Neuroscience

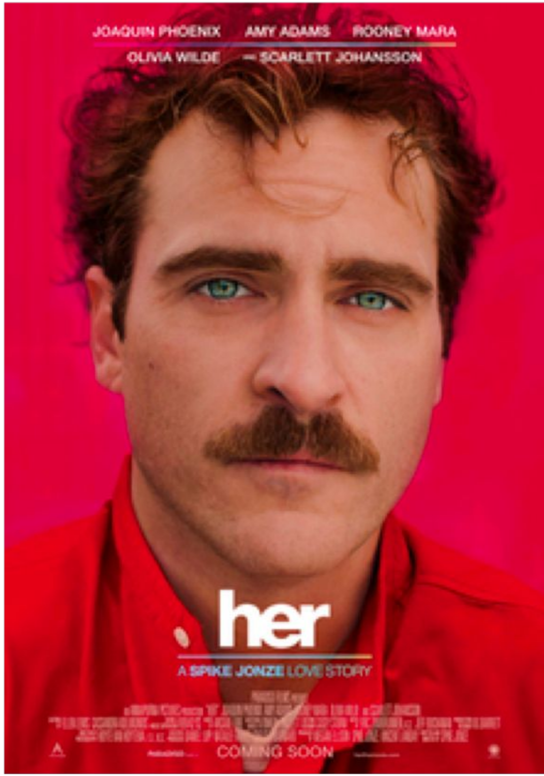
Social sciences



1960s


MAN HAS MADE HIS MATCH ...NOW IT'S HIS PROBLEM
HARRISON FORD
BLADE RUNNER
JERRY PERENCHIO AND BUD YORKIN PRESENT
A MICHAEL DEELEY-RIDLEY SCOTT PRODUCTION

1980s

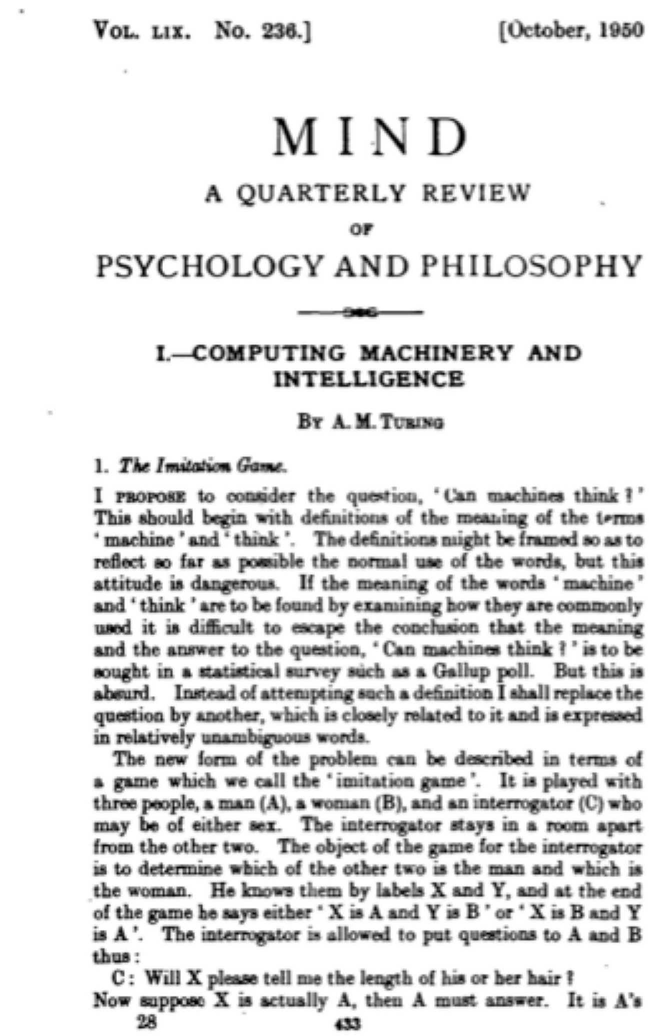


2020s

Turing test

“Can machines think?”

A **language-based** experiment



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, ‘Can machines think?’ This should begin with definitions of the meaning of the terms ‘machine’ and ‘think’. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ‘machine’ and ‘think’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ‘Can machines think?’ is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

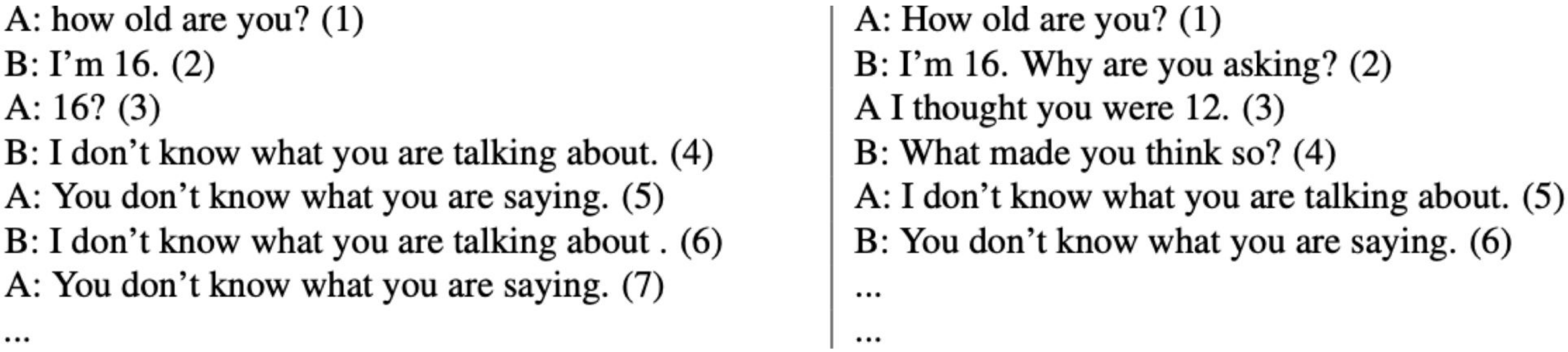
The new form of the problem can be described in terms of a game which we call the ‘imitation game’. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either ‘X is A and Y is B’ or ‘X is B and Y is A’. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's

28 433

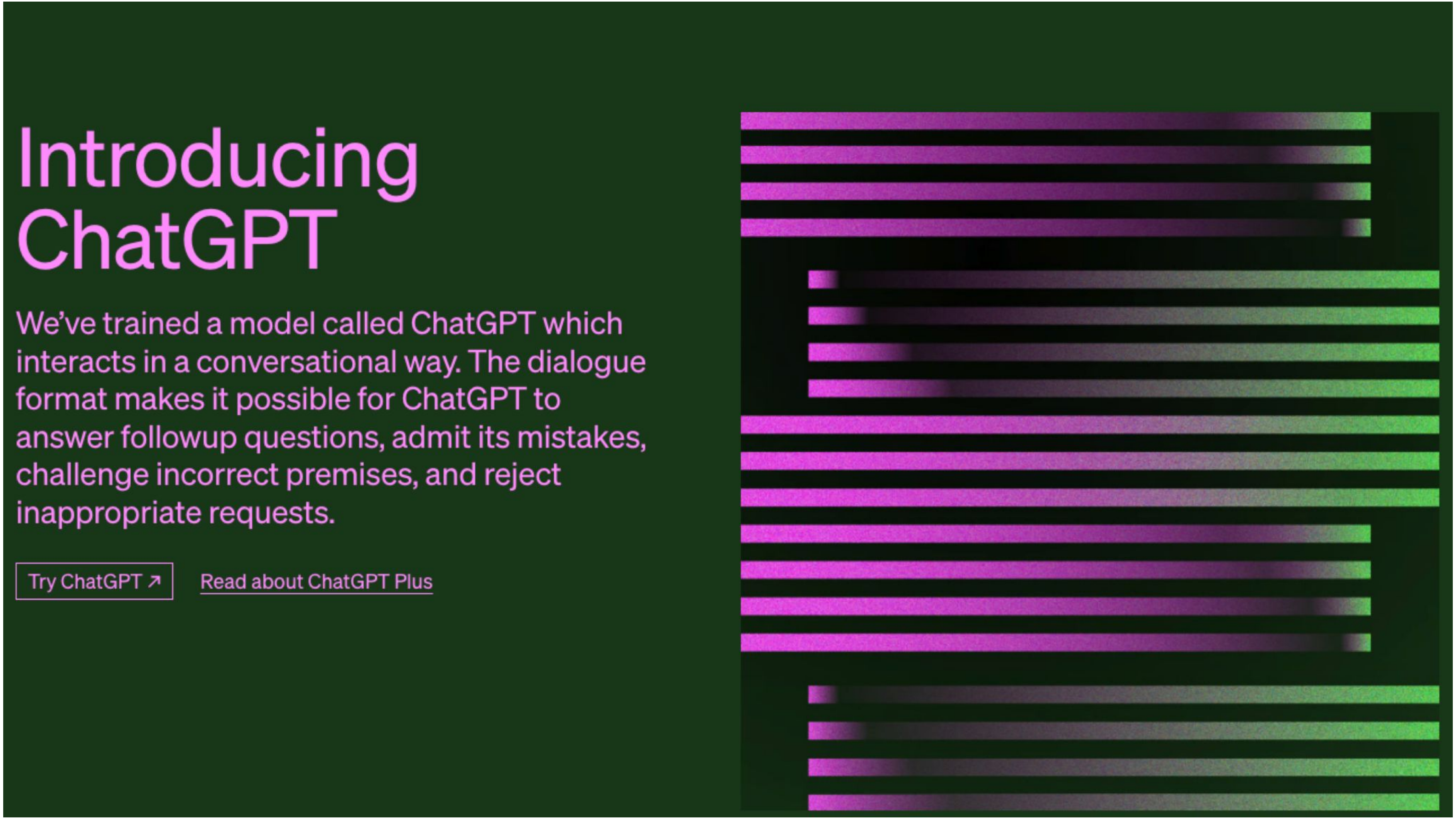
Dialogue Generation for Conversational Agents



A: how old are you? (1)
B: I’m 16. (2)
A: 16? (3)
B: I don’t know what you are talking about. (4)
A: You don’t know what you are saying. (5)
B: I don’t know what you are talking about . (6)
A: You don’t know what you are saying. (7)
...

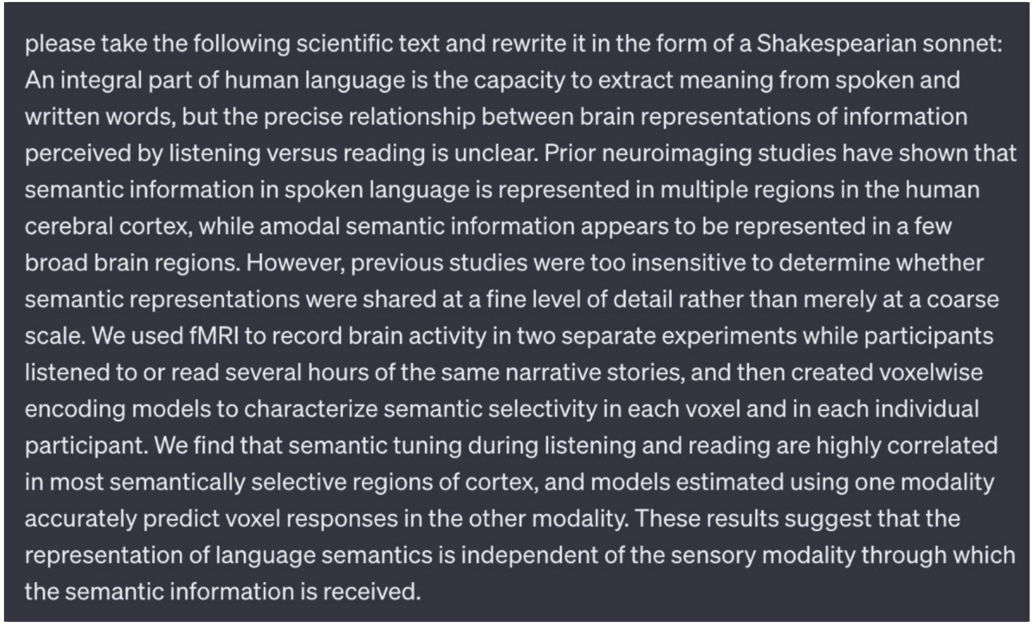
A: How old are you? (1)
B: I’m 16. Why are you asking? (2)
A I thought you were 12. (3)
B: What made you think so? (4)
A: I don’t know what you are talking about. (5)
B: You don’t know what you are saying. (6)
...
...

Li et al., Deep Reinforcement
Learning for Dialogue Generation,
EMNLP, 2016

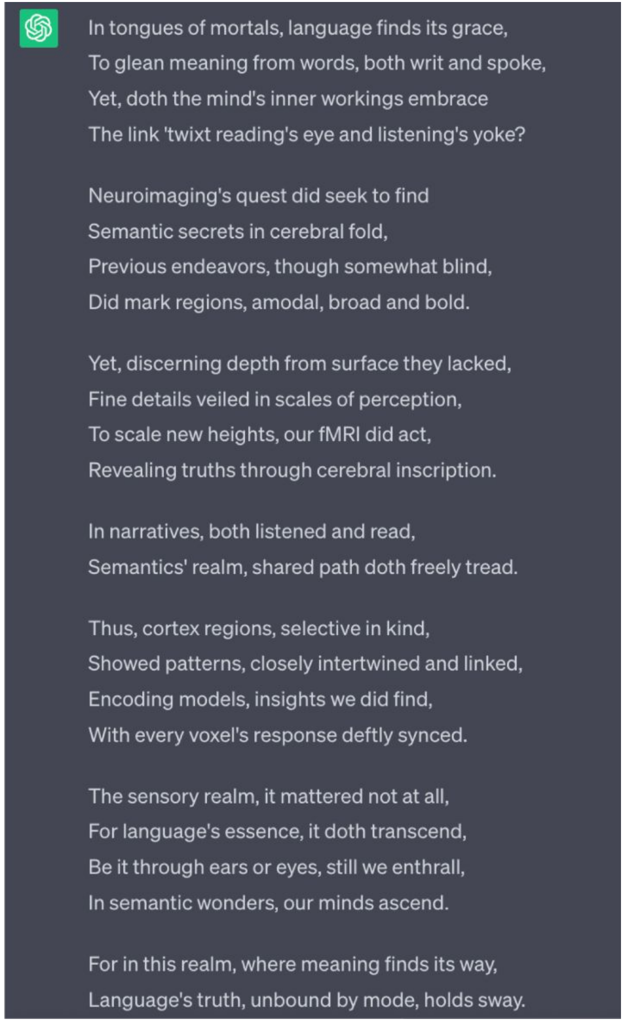


Introducing ChatGPT

We’ve trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.



please take the following scientific text and rewrite it in the form of a Shakespearian sonnet: An integral part of human language is the capacity to extract meaning from spoken and written words, but the precise relationship between brain representations of information perceived by listening versus reading is unclear. Prior neuroimaging studies have shown that semantic information in spoken language is represented in multiple regions in the human cerebral cortex, while amodal semantic information appears to be represented in a few broad brain regions. However, previous studies were too insensitive to determine whether semantic representations were shared at a fine level of detail rather than merely at a coarse scale. We used fMRI to record brain activity in two separate experiments while participants listened to or read several hours of the same narrative stories, and then created voxelwise encoding models to characterize semantic selectivity in each voxel and in each individual participant. We find that semantic tuning during listening and reading are highly correlated in most semantically selective regions of cortex, and models estimated using one modality accurately predict voxel responses in the other modality. These results suggest that the representation of language semantics is independent of the sensory modality through which the semantic information is received.



In tongues of mortals, language finds its grace,
To glean meaning from words, both writ and spoke,
Yet, doth the mind's inner workings embrace
The link 'twixt reading's eye and listening's yoke?

Neuroimaging's quest did seek to find
Semantic secrets in cerebral fold,
Previous endeavors, though somewhat blind,
Did mark regions, amodal, broad and bold.

Yet, discerning depth from surface they lacked,
Fine details veiled in scales of perception,
To scale new heights, our fMRI did act,
Revealing truths through cerebral inscription.

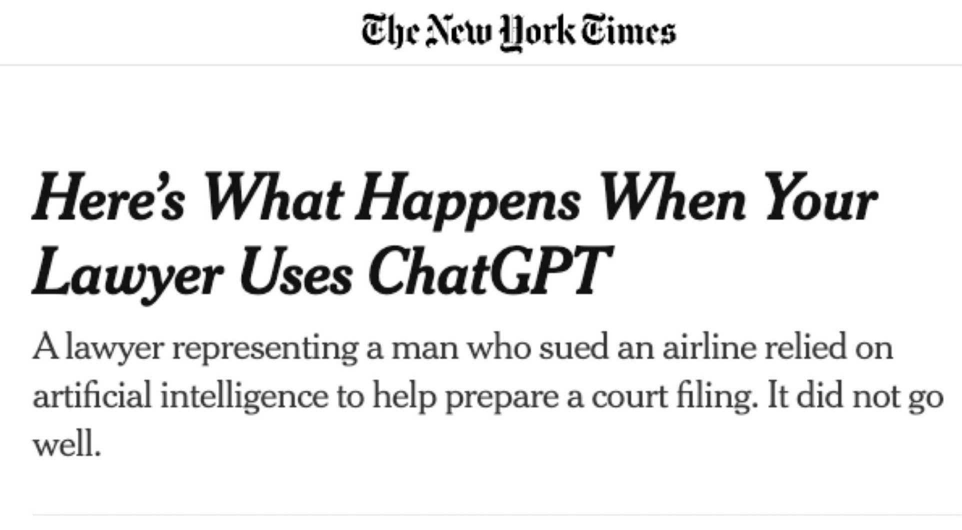
In narratives, both listened and read,
Semantics' realm, shared path doth freely tread.

Thus, cortex regions, selective in kind,
Showed patterns, closely intertwined and linked,
Encoding models, insights we did find,
With every voxel's response deftly synced.

The sensory realm, it mattered not at all,
For language's essence, it doth transcend,
Be it through ears or eyes, still we enthrall,
In semantic wonders, our minds ascend.

For in this realm, where meaning finds its way,
Language's truth, unbound by mode, holds sway.

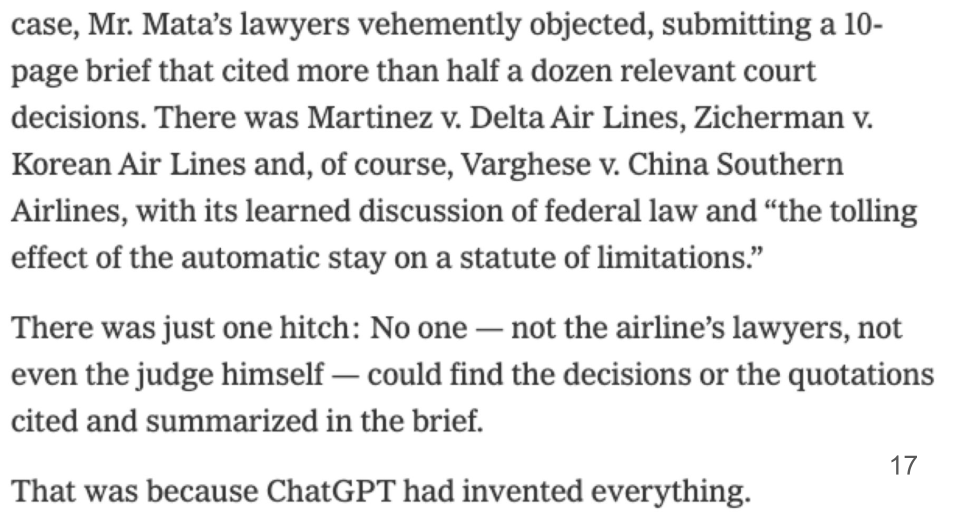
CAUTION!



The New York Times

Here’s What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

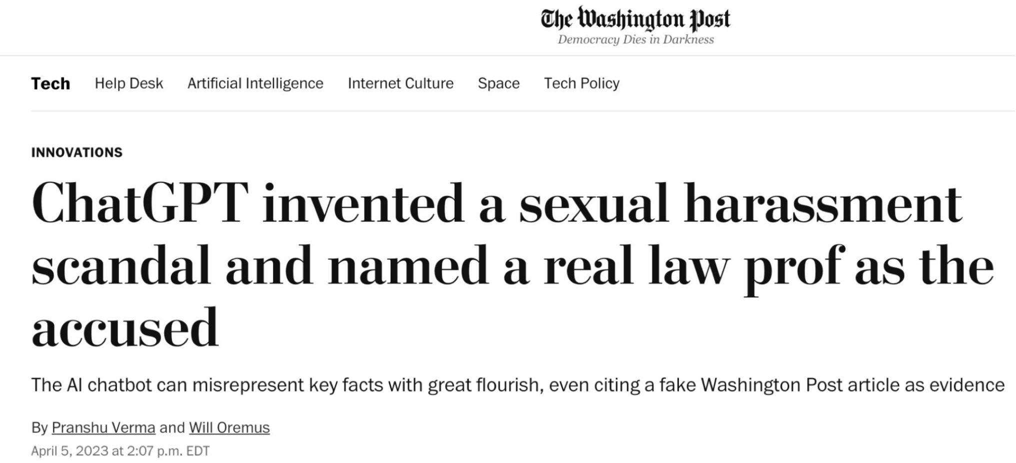


case, Mr. Mata’s lawyers vehemently objected, submitting a 10-page brief that cited more than half a dozen relevant court decisions. There was Martinez v. Delta Air Lines, Zicherman v. Korean Air Lines and, of course, Varghese v. China Southern Airlines, with its learned discussion of federal law and “the tolling effect of the automatic stay on a statute of limitations.”

There was just one hitch: No one — not the airline’s lawyers, not even the judge himself — could find the decisions or the quotations cited and summarized in the brief.

That was because ChatGPT had invented everything.

CAUTION!



INNOVATIONS

ChatGPT invented a sexual harassment scandal and named a real law prof as the accused

The AI chatbot can misrepresent key facts with great flourish, even citing a fake Washington Post article as evidence

By Pranshu Verma and Will Oremus

April 5, 2023 at 2:07 p.m. EDT

What is natural language processing?

NLP aims at enabling computers

to understand, interpret, and generate

human language (speech or written text)

in a way that is both meaningful and useful.

Tasks associated with NLP

- **Natural Language Understanding**
  - Text classification, sentiment analysis, named entity recognition, question answering, part of speech tagging
- **Natural Language Generation**
  - Machine translation, text summarisation, dialogue systems, content generation
- **Language-based Interaction**
  - Voice assistants, chatbots, voice recognition systems
- **Knowledge Extraction**
  - Extracting information from documents, information retrieval, knowledge graphs

What makes it difficult to crack the code of language?

The challenges of language

1. Productivity
2. Ambiguity
3. Variability
4. Diversity
5. Sparsity

The challenges of language: PRODUCTIVITY

- Speakers of language can construct and understand an indefinitely large number of utterances
  - including those they have never previously encountered
- Over time new words, senses, structures are introduced in languages
  - New words within last three years: social distance, staycation, deepfake

The challenges of language: AMBIGUITY

- Most linguistic constructs are open to several interpretations
- Ambiguity can appear at all levels of processing
- Humans use linguistics context and extra linguistic context to disambiguate
  - It usually requires discourse knowledge, Q/A

The boy carries the light box.

The boy carries the light box.

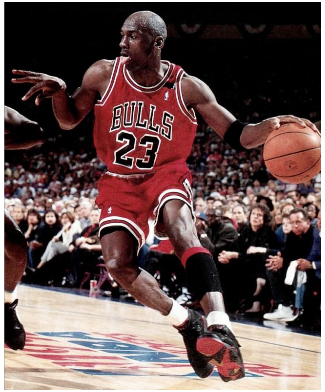
Lexical ambiguity

The boy carries the light box.

not heavy   electric lamp   shiny

Michael Jordan is a professor at UC Berkeley.

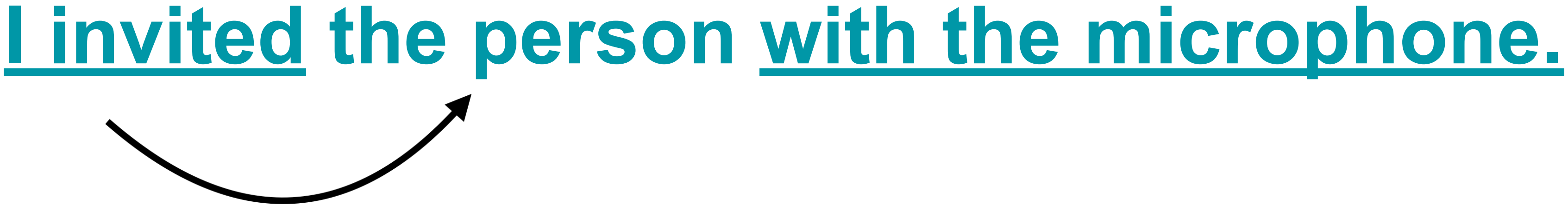
Michael Jordan is a professor at UC Berkeley.



Syntactic ambiguity

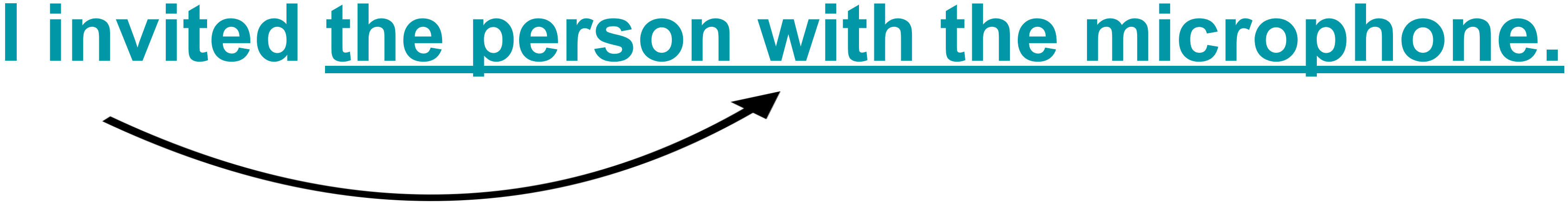
I invited the person with the microphone.

Syntactic ambiguity



I invited the person with the microphone.

Syntactic ambiguity



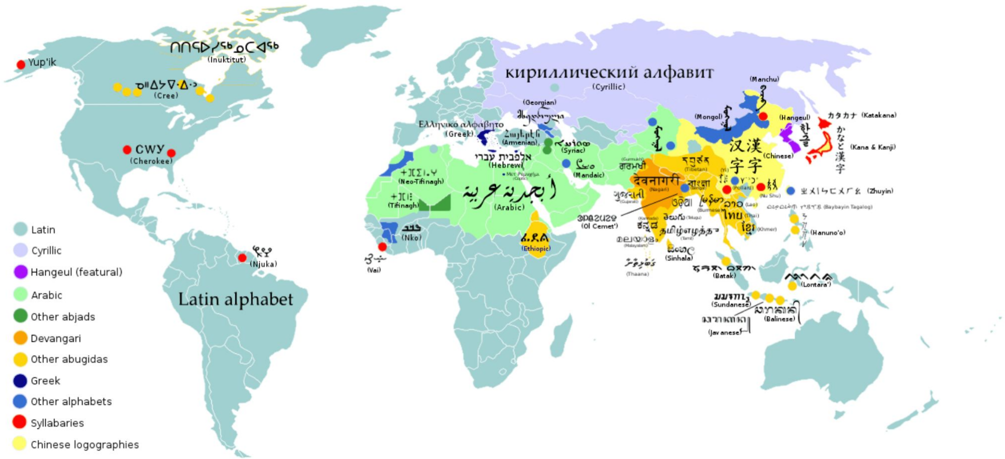
I invited the person with the microphone.

The challenges of language: VARIABILITY

- There is more than one way of saying the same thing
- Language varies at all processing levels
    - Phonetic level: accent
    - Morphological or lexical level: spelling or word choice
    - Syntactic
    - Semantic

The challenges of language: DIVERSITY

- 7000 spoken languages with phonemic and graphemic diversity

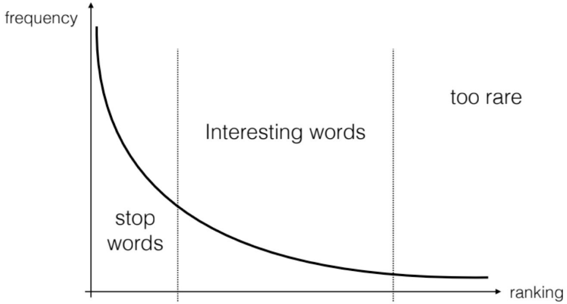


[Image taken from wikipedia]

The challenges of language: SPARSITY

- Word distributions follow Zipf's Law
    - The most common words occur much more frequently than the less common words

$$f_w(k) \; \alpha \; \frac{1}{k^\theta}$$



- Many "interesting" words in language have very low frequency