

# Natural Language Processing and Information Extraction

2024 WS

Gábor Recski  
Florina Piroi  
Varvara Arzt  
Allan Hanbury

# Contents

- Course information
- Introduction to NLP

# Course information

# Lecturers



Gábor Recski



Florina Piroi



Varvara Arzt



Allan Hanbury

# Lecture Schedule

- Course Information, Introduction to NLP [Hanbury] (4.10.2024)
- Text Processing [Recski] (11.10.2024)
- Text Classification [Recski] (18.10.2024)
- Deep Learning for NLP [Piroi] (25.10.2024)
- Textual Sequence Modelling & Attention [Piroi] (8.11.2024) **[ONLINE]**
- Deep Learning – Practical Lesson [Arzt] (22.11.2024)
- Syntax (Constituency and Dependency) [Recski] (29.11.2024)
- Basic (non-DL) Semantics [Recski] (6.12.2024)
- Information Extraction [Recski] (13.12.2024)
- Summarisation & Keyword Extraction [Piroi] (20.12.2024)
- Annotation Basics and Challenges [Hanbury] (10.1.2025)
- Project Presentations (24.1.2025)

# Lectures

- Fridays 13:00 c.t. - 15:00
- EI11

# Exercise

- One project exercise with two Milestones
- Done in groups of four
- Each group has a mentor
- Submissions are made via GitHub
- Grading is based on milestones, final submission, presentation, and report
- Every group member must present their own contributions in the final presentation and will be individually evaluated on these contributions

Detailed schedule and discussion at the end of the lecture today  
Note that topic selection is due by the end of next week

# Effort Breakdown

- Lectures: 24 hours
- Project Milestone 1: 8 hours
- Project Milestone 2: 8 hours
- Final Solution: 35 hours
- **Total: 75 hours**



# Performance Evaluation

- Milestone 1: Minimum 35%
- Milestone 2: Minimum 35%
- Final solution: Minimum 35%
  
- Overall Score: Minimum 50% to pass
  - 15% for Milestone 1
  - 15% for Milestone 2
  - 50% for the final solution
  - 10% for the presentation
  - 10% for the management summary
  
- There is no exam!
  
- Marks Overall Score

1	89 – 100
2	76 – 88
3	63 – 75
4	50 – 62

# Organisation

- Course
  - Please register for the course in TISS
- Communication
  - Use the General Discussion Forum in TUWEL for questions, not the TISS forum
- The schedule of lectures and all course material will be available on Github

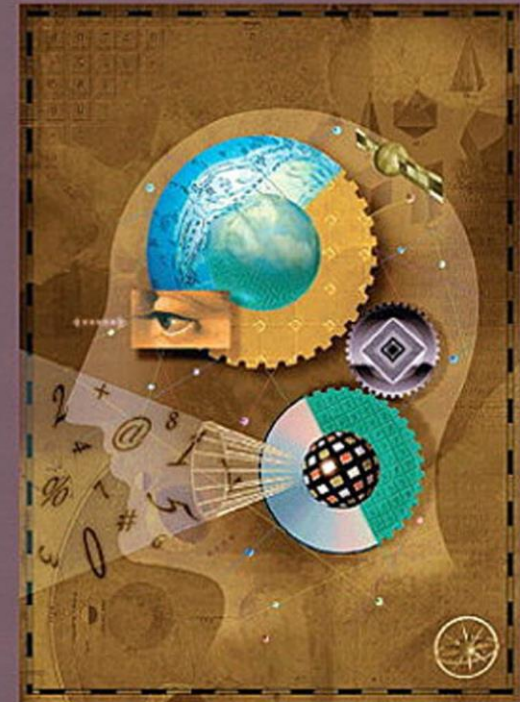
# Book

Third edition in preparation – download many chapters here:

<https://web.stanford.edu/~jurafsky/slp3/>

## SPEECH AND LANGUAGE PROCESSING

*An Introduction to Natural Language Processing,  
Computational Linguistics, and Speech Recognition*



Second Edition

DANIEL JURAFSKY & JAMES H. MARTIN

Questions about the organisation, etc.

Ask now!

# Introduction to NLP

# IBM Watson and Jeopardy



Final question: <https://www.youtube.com/watch?v=Sp4q60BsHoY>

IBM film: <https://www.youtube.com/watch?v=P18EdAKuC1U>

# The end of the show

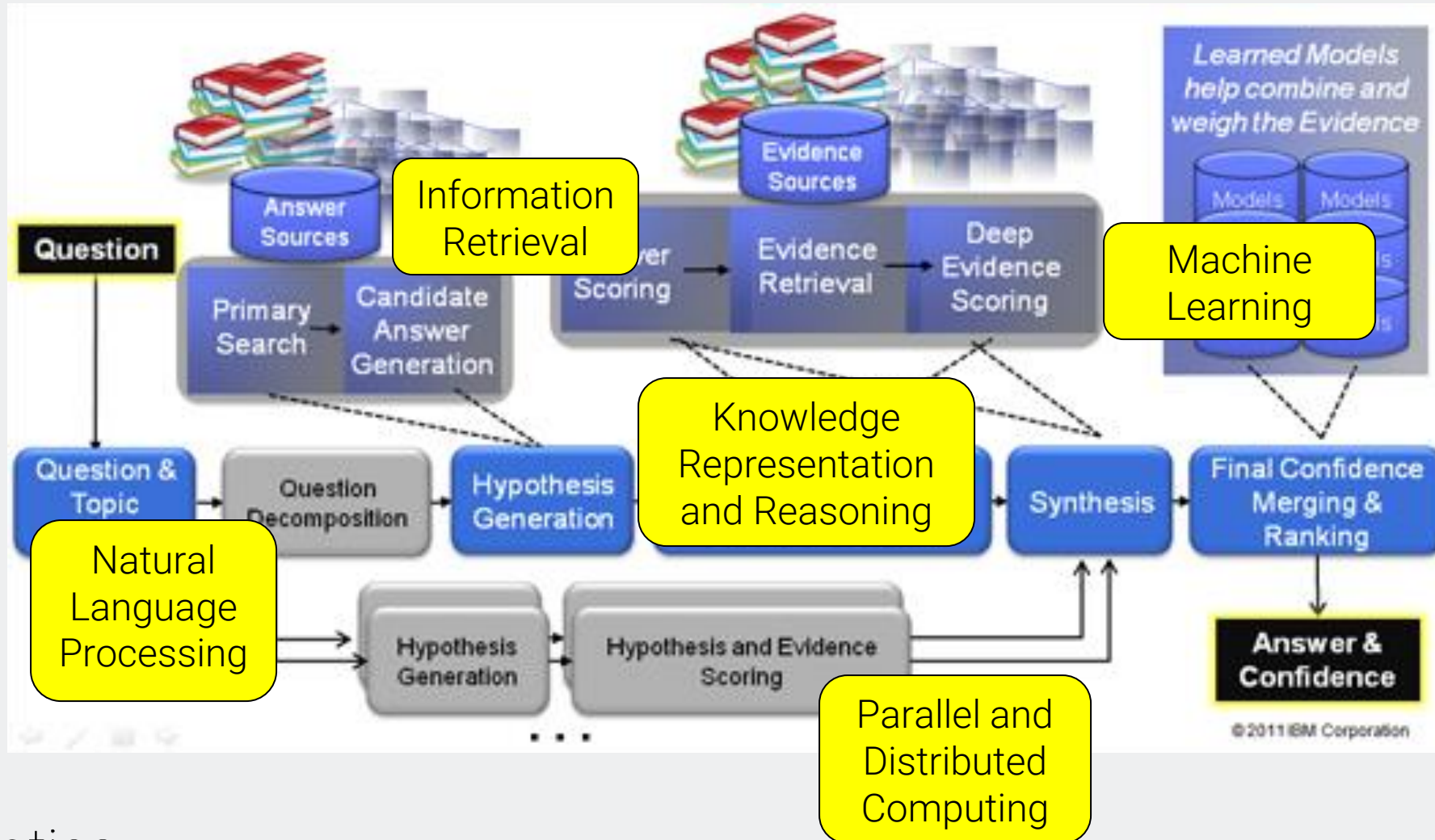




TU  
WIEN  
Informatics



# How does Watson (for Jeopardy) work?



# Go beyond human capabilities

Combine the expertise of humans with the power of technology to deliver patient-centric cancer care at scale

[PDF Transform cancer care \(1.8 MB\)](#)[Data, AI & analytics \(01:28\)](#)[↓ Our goals](#)[↓ For the physician](#)[↓ Success stories](#)[↓ Solutions](#)

## The changing oncology landscape

New intelligent analytics and workflow technologies hold the key to overcoming a primary challenge of providing personalized approaches to cancer care—harnessing the vast amounts of data available without being overwhelmed by it. Clinicians are faced with large, heterogeneous, and complex data sets when making patient-specific clinical decisions. Oncology solutions backed by AI and machine learning provide a powerful tool by bringing together data, extracting insights and presenting it to providers for their evaluation.

[PDF Read the 2020 Data and Evidence Booklet \(10.2 MB\)](#)[From research to real world.](#)

# How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on *Jeopardy!*, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

By **Eliza Strickland**

<https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care>

## The Hype of Watson: Why Hasn't AI Taken Over Oncology?

ARTICLE ⓘ Apr 17, 2020 | by Sylvia He



*Doctors' notes are one of the obstacles in the way of AI becoming a major force in oncology.*

<https://www.technologynetworks.com/informatics/articles/the-hype-of-watson-why-hasnt-ai-taken-over-oncology-333571>

# Natural Language Processing (NLP)

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

wikipedia



# Why is NLP interesting?

- Languages involve many human activities
  - Reading, writing, speaking, listening
- Voice can be used as a user interface in many applications
  - Remote controls, virtual assistants like siri,...
- NLP is used to acquire insights from massive amount of textual data
  - E.g., hypotheses from medical & health reports
- NLP has many applications
- NLP is difficult!

# Why is NLP difficult?

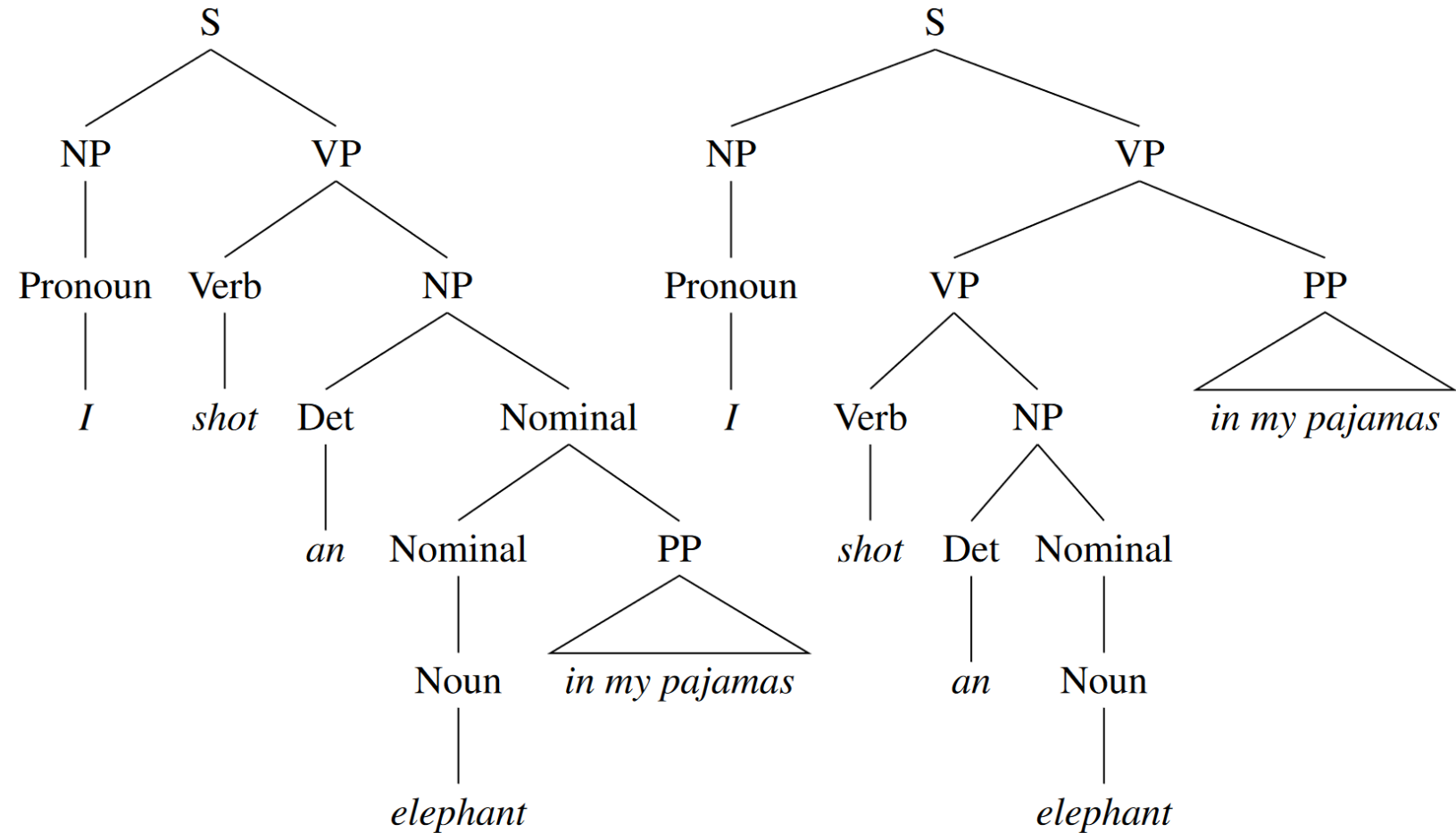
I made her duck

- I cooked waterfowl for her.
- I cooked waterfowl belonging to her.
- I created the (plaster?) duck she owns.
- I caused her to quickly lower her head or body.
- I waved my magic wand and turned her into undifferentiated waterfowl.

# I shot an elephant in my pyjamas.



<https://freesvg.org/elephant-in-pajamas>



**Figure 13.2** Two parse trees for an ambiguous sentence. The parse on the left corresponds to the humorous reading in which the elephant is in the pajamas, the parse on the right corresponds to the reading in which Captain Spaulding did the shooting in his pajamas.

# Why is NLP difficult?

Natural Languages are generally ambiguous

Various levels of knowledge of a language must be considered:

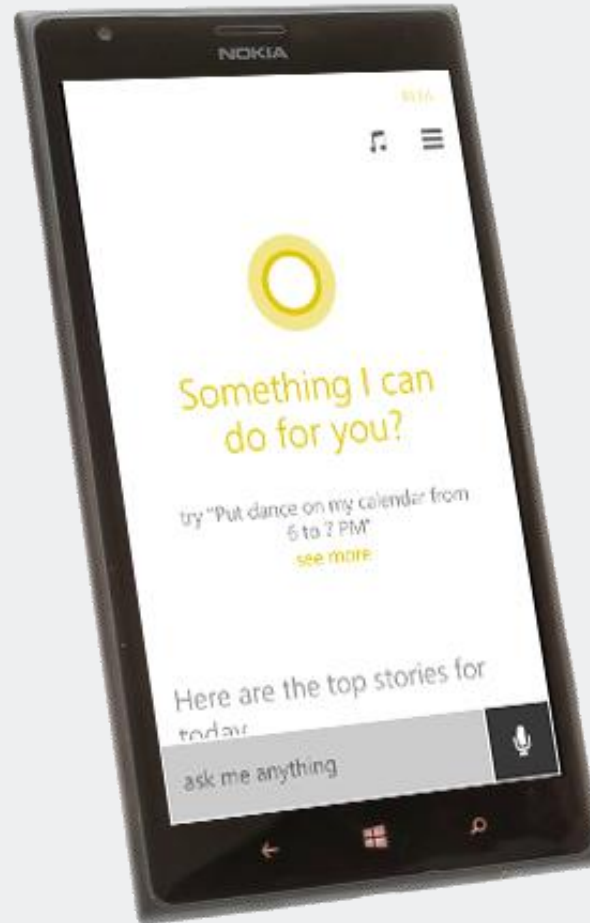
- **Phonetics and Phonology** — knowledge about linguistic sounds
- **Morphology** — knowledge of the meaningful components of words
  - I am → I'm, forms for singular and plural (door/doors)
- **Syntax** — knowledge of the structural relationships between words, needed to order and group words (grammatical rules)
- **Semantics** — knowledge of meaning
  - What is meant by “export” and “expert”? What constitutes “Western Europe”?
- **Pragmatics** — knowledge of the relationship of meaning to the goals and intentions of the speaker
  - Is it a request, question or a statement?
- **Discourse** — knowledge about linguistic units larger than a single utterance
  - Reference to the context given by e.g. multiple sentences.
  - E.g. In what year was Lincoln born? How many states were in the United States in *that year*?


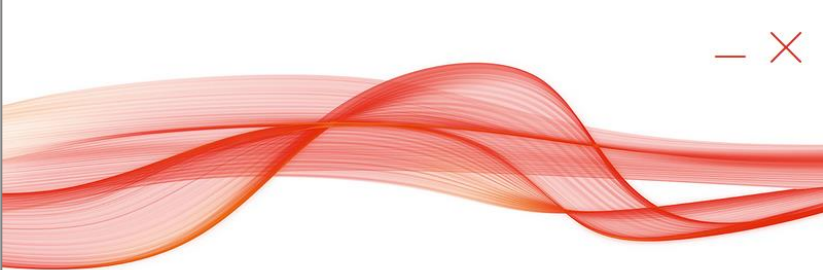


# Very Brief History of NLP

- Foundational Insights: 1940s and 1950s
- Generally two paradigms:
  - Symbolic Paradigm
  - Stochastic Paradigm
- The Rise of Machine Learning: 2000-now
  - Large amount of spoken and textual data become available
  - Widespread availability of high-performance computing systems
- The Domination of Neural Approaches: ~2015-now

# Dialogue systems





**Kara**


Dein digitaler Service bei A1.

\*

\*

Bitte geben Sie Ihre Kontaktinformationen ein.

Chat starten



Hallo und willkommen bei Drei! Ich bin Troy, Ihr virtueller Berater und kann Ihnen Fragen aus der Welt von Drei beantworten oder auf unseren Webseiten nach passenden Inhalten suchen. Was möchten Sie gerne wissen?

Aktuell nachgefragt:

Was ist das Kundenkennwort

>

Was ist das Kundenzone-Passwort

>

Roaming

>

Guthaben abrufen

>

Guthaben aufladen

>

Informationen zum Tarifwechsel

>


Informationen zum Treuebonus

>

Zusatzpakete kaufen

>

01.Okt. 2020



Hi! Ich bin Tinka. Sind Sie wegen eines der folgenden Themen gekommen? Alternativ können Sie mich auch alles andere fragen.

**Aktuelles**  
#bleibverbunden

tel  
Wil  
Ma

Auswählen

Hallo, ich bin Tinka. Ihre persönliche Assistentin.

# Training Data Bias...

## Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

by James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT





# GPT-3

Large Language  
Model (LLM) with  
175 billion  
parameters

[cs.CL] 22 Jul 2020

## Language Models are Few-Shot Learners

Tom B. Brown\* Benjamin Mann\* Nick Ryder\* Melanie Subbiah\*  
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry  
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan  
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter  
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray  
Benjamin Chess Jack Clark Christopher Berner  
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei  
OpenAI

Training  
datasets

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

# Correcting English

Poor English input: I eated the purple berries.

Good English output: I ate the purple berries.

Poor English input: Thank you for picking me as your designer. I'd appreciate it.

Good English output: Thank you for choosing me as your designer. I appreciate it.

Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.

Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.

Poor English input: I'd be more than happy to work with you in another project.

Good English output: I'd be more than happy to work with you on another project.

---

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.

Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

---

Poor English input: The patient was died.

Good English output: The patient died.

---

Poor English input: We think that Leslie likes ourselves.

Good English output: We think that Leslie likes us.

---

Poor English input: Janet broke Bill on the finger.

Good English output: Janet broke Bill's finger.

---

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.

Good English output: Mary arranged for John to rent a house in St. Louis.

---

Poor English input: Today I have went to the store to to buys some many bottle of water.

Good English output: Today I went to the store to buy some bottles of water.

## News Article Generation

Title: United Methodists Agree to Historic Split

Subtitle: Those who oppose gay marriage will form their own denomination

Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

**Figure 3.14:** The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm

Subtitle: Joaquin Phoenix pledged to not change for each awards event

Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what. Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

**Figure 3.15:** The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).



# ChatGPT

- Version of GPT3 optimised for dialogue through manual labelling
- <https://openai.com/blog/chatgpt/>



Write a limerick about a natural language processing class



In a natural language processing class  
The students were feeling quite amorous  
But their romance was thwarted  
When the professor imported  
A dataset of legal documents

AL

Write python code to check if text contains an instruction.



Here is an example of how you could write Python code to check if a given string of text contains an instruction:

[Copy code](#)

```
def has_instruction(text):
    instructions = ["do this", "do that", "go here", "go there"]
    for instruction in instructions:
        if instruction in text:
            return True
    return False

text = "Please do this task and then go there"
if has_instruction(text):
    print("Text contains an instruction.")
else:
    print("Text does not contain an instruction.")
```

This code defines a function called `has_instruction` that takes a string of text as input and checks to see if any of the instructions in the `instructions` list appear in the text. If an instruction is found, the function returns `True`, otherwise it returns `False`. The function is called on the `text` string and the result is printed to the console. Note that this is just one way to solve this problem, and there are many other approaches you could take. Is there anything else I can help with?

# False Links

AL

Where can I find information on python for natural language processing?



There are many resources available online that can help you learn about using Python for natural language processing (NLP). Here are a few options:

- The Python documentation includes a section on NLP, which provides an overview of some of the tools and libraries available in Python for working with NLP:  
<https://docs.python.org/3/library/nlp.html>
- The Natural Language Toolkit (NLTK) is a popular library for NLP in Python. It includes a wide range of tools for tasks such as tokenization, stemming, and part-of-speech tagging, as well as more advanced techniques like parsing and machine learning. You can find documentation and tutorials for NLTK at: <https://www.nltk.org/>
- There are also many online tutorials and courses available that cover using Python for NLP. For example, you can find a free course on NLP with Python on the website of the National Research University Higher School of Economics: <https://www.coursera.org/learn/python-for-nlp>

I hope these resources are helpful. Let me know if you have any other questions.

# Hallucination

AL

What is the best way to modulate from C Major to G Major?



There are many ways to modulate, or change key, from C major to G major. Here are a few options:

1. Direct modulation: This involves simply changing the tonic note from C to G and continuing the piece in the new key. This can be done by using a dominant chord (a chord built on the fifth scale degree) that leads to the tonic chord in the new key. For example, you could use a G major chord to lead directly to the G major tonic chord.
2. Modulation through a related key: Another option is to modulate to a key that is closely related to the destination key. In this case, you could modulate to the key of E minor, which has the same key signature as G major (one sharp) and shares many common chords. From E minor, you can then modulate to G major by using a dominant chord (such as an A major chord) that leads to the tonic chord in the new key.
3. Modulation through a pivot chord: A pivot chord is a chord that is shared by both the original key and the destination key. For example, in this case **you could use an A major chord, which is common to both C major and G major**, as a pivot chord to help modulate between the two keys. To do this, you could use the A major chord to lead temporarily to the key of A major, and then use a dominant chord (such as an E major chord) to lead to the tonic chord in the destination key of G major.

These are just a few examples of how you could modulate from C major to G major. The best approach will depend on the specific needs of your music and the overall structure of the piece. I hope this helps! Let me know if you have any other questions.

# OpenAI list of limitations



## Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

## Limitations

- ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.
- ChatGPT is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times. For example, given one phrasing of a question, the model can claim to not know the answer, but given a slight rephrase, can answer correctly.
- The model is often excessively verbose and overuses certain phrases, such as restating that it's a language model trained by OpenAI. These issues arise from biases in the training data (trainers prefer longer answers that look more comprehensive) and well-known over-optimization issues.<sup>1,2</sup>
- Ideally, the model would ask clarifying questions when the user provided an ambiguous query. Instead, our current models usually guess what the user intended.
- While we've made efforts to make the model refuse inappropriate requests, it will sometimes respond to harmful instructions or exhibit biased behavior. We're using the Moderation API to warn or block certain types of unsafe content, but we expect it to have some false negatives and positives for now. We're eager to collect user feedback to aid our ongoing work to improve this system.

HEALTH TECH, ARTIFICIAL INTELLIGENCE, PROVIDERS

## How Often Do LLMs Hallucinate When Producing Medical Summaries?

Researchers at the University of Massachusetts Amherst released a paper this week showing that large language models tend to hallucinate quite a bit when producing medical summaries.

By Katie Adams on August 11, 2024



Share

<https://medcitynews.com/2024/08/ai-healthcare-llm/>

<https://openreview.net/pdf?id=6eMIzKF0pJ>



Informatics

For this study, the research team collected 100 medical summaries from OpenAI's GPT-4o and Meta's Llama-3 — two up-to-date proprietary and open-source LLMs.

In the 50 summaries produced by GPT-4o, the researchers identified 327 instances of medical event inconsistencies, 114 instances of incorrect reasoning and three instances of chronological inconsistencies.

The 50 summaries generated by Llama-3 were shorter and less comprehensive than those produced by GPT-4o, Rumale noted. In these summaries, the research team found 271 instances of medical event inconsistencies, 53 instances of incorrect reasoning and one chronological inconsistency.

# Training Data Bias...



**Abubakar Abid**  
@abidlabs

I'm shocked how hard it is to generate text about Muslims from GPT-3 that has nothing to do with violence... or being killed...

<https://twitter.com/i/status/1291165311329341440>



# NLP and climate change

CO<sub>2</sub> equivalent emissions for model training

Model name	Number of parameters	CO <sub>2</sub> eq emissions
GPT-3	175B	502 tonnes
Gopher	280B	352 tonnes
OPT	175B	70 tonnes
BLOOM	176B	25 tonnes

<https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf>

## What does 1 ton of CO<sub>2</sub> represent?



138 meat-based meals



1,961 vegetarian meals



A one way-trip  
from Paris to New  
York by plane per  
passenger



The CO<sub>2</sub> an  
average European  
emits in less than  
two months





# Summary

- Neural approaches are big in NLP at the moment
- Beware of bias
- NLP can be bad for the climate