Introduction to NLP

CS 759/859 Natural Language Processing

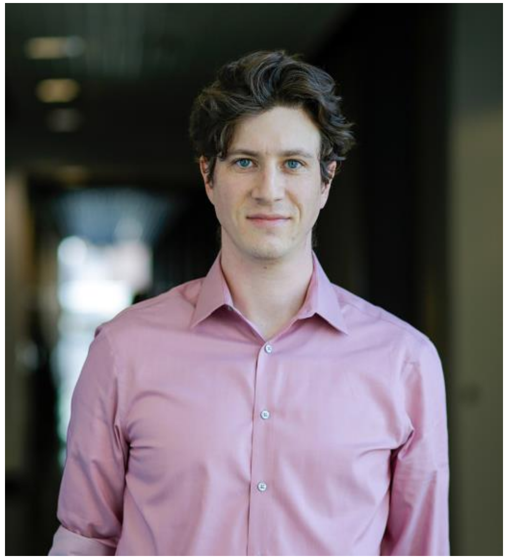
Samuel Carton, University of New Hampshire

This lecture

1. Introductions
2. What is NLP?
3. Why should you care about NLP
4. Structure of this course

Who am I

This guy:



My research: human-centered and interpretable NLP

- Why do the models do what they do and how can people work with them more effectively?

What is NLP?

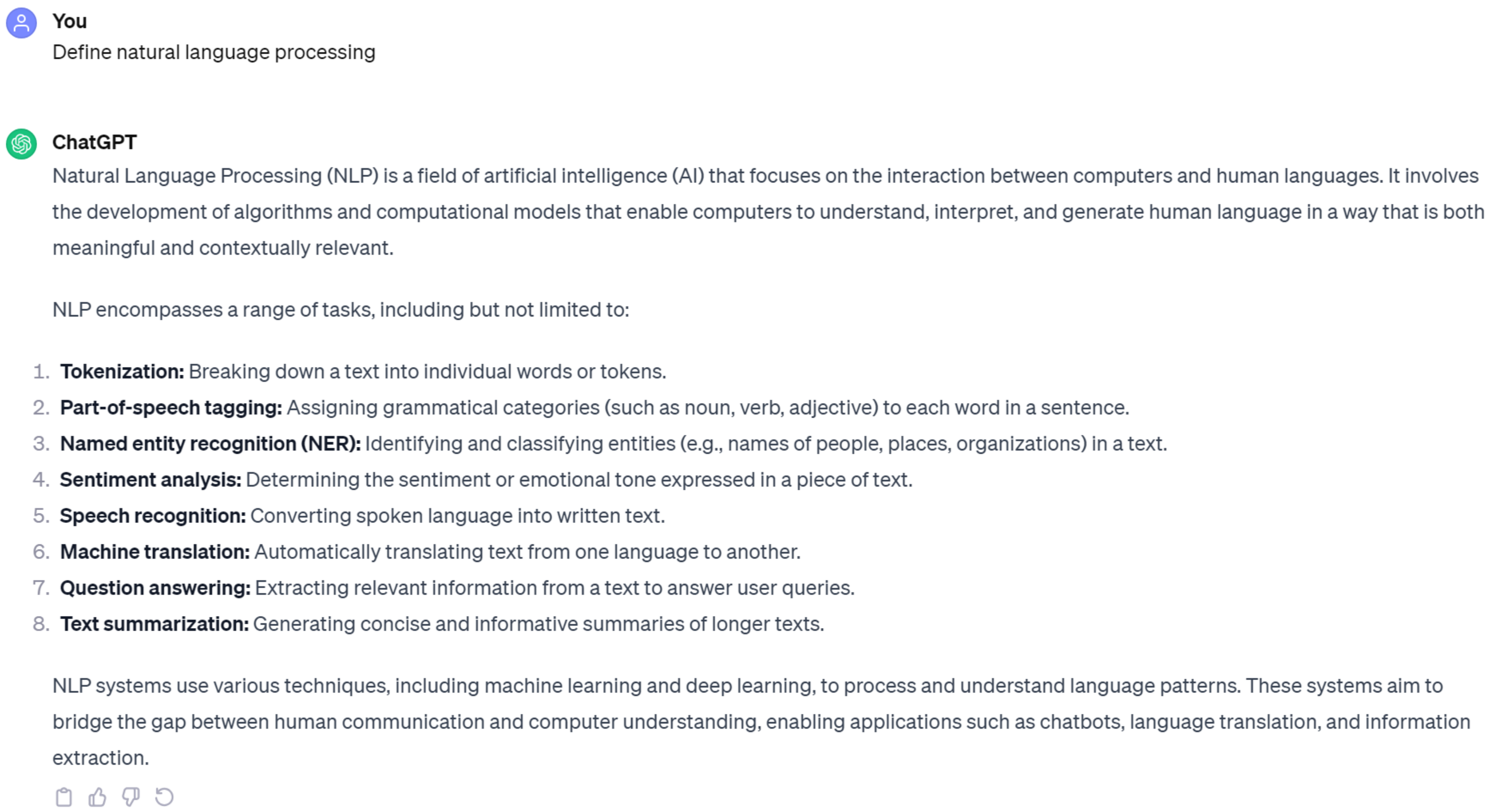
What is NLP?

Natural Language Processing

From Wikipedia, a year ago: “an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language”

From Wikipedia, today: “an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate human language.”

NLP example: ChatGPT



You

Define natural language processing

ChatGPT

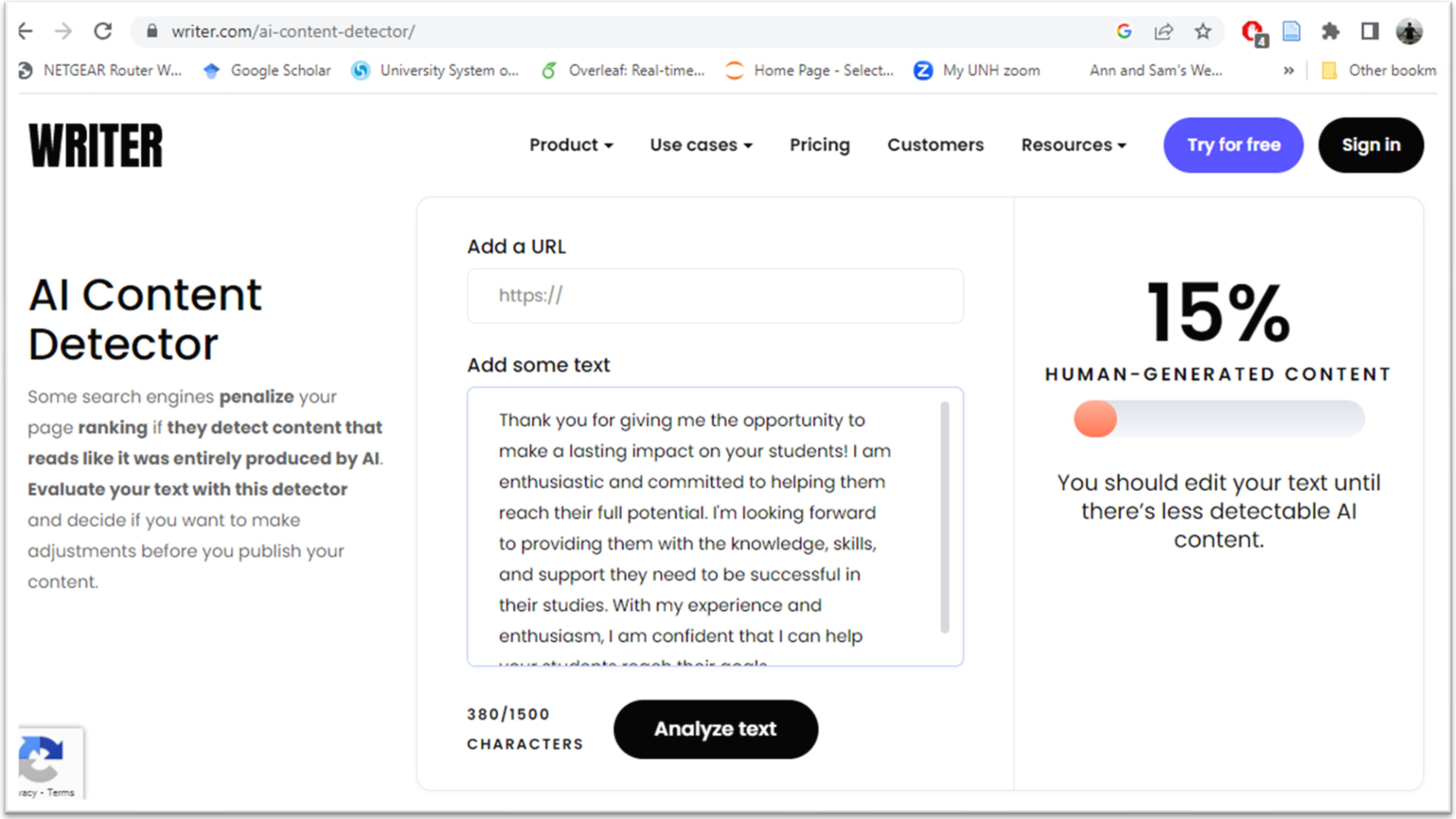
Natural Language Processing (NLP) is a field of artificial intelligence (AI) that focuses on the interaction between computers and human languages. It involves the development of algorithms and computational models that enable computers to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant.

NLP encompasses a range of tasks, including but not limited to:

1. **Tokenization:** Breaking down a text into individual words or tokens.
2. **Part-of-speech tagging:** Assigning grammatical categories (such as noun, verb, adjective) to each word in a sentence.
3. **Named entity recognition (NER):** Identifying and classifying entities (e.g., names of people, places, organizations) in a text.
4. **Sentiment analysis:** Determining the sentiment or emotional tone expressed in a piece of text.
5. **Speech recognition:** Converting spoken language into written text.
6. **Machine translation:** Automatically translating text from one language to another.
7. **Question answering:** Extracting relevant information from a text to answer user queries.
8. **Text summarization:** Generating concise and informative summaries of longer texts.

NLP systems use various techniques, including machine learning and deep learning, to process and understand language patterns. These systems aim to bridge the gap between human communication and computer understanding, enabling applications such as chatbots, language translation, and information extraction.

NLP example: chat bot detector



AI Content Detector

Some search engines **penalize** your page **ranking** if **they detect content that reads like it was entirely produced by AI. Evaluate your text with this detector** and decide if you want to make adjustments before you publish your content.

Add a URL

https://

Add some text

Thank you for giving me the opportunity to make a lasting impact on your students! I am enthusiastic and committed to helping them reach their full potential. I'm looking forward to providing them with the knowledge, skills, and support they need to be successful in their studies. With my experience and enthusiasm, I am confident that I can help

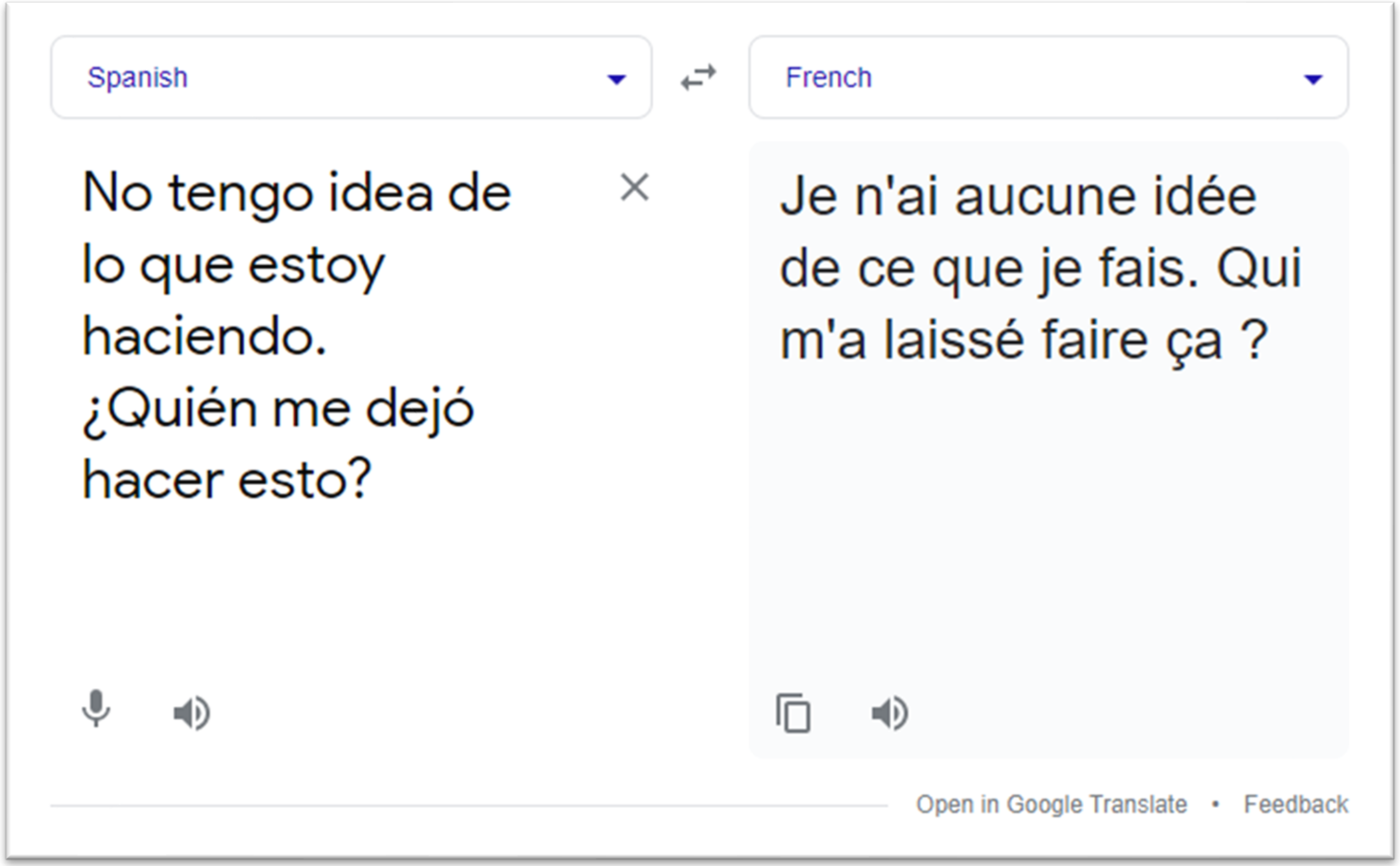
380/1500 CHARACTERS

15%

HUMAN-GENERATED CONTENT

You should edit your text until there's less detectable AI content.

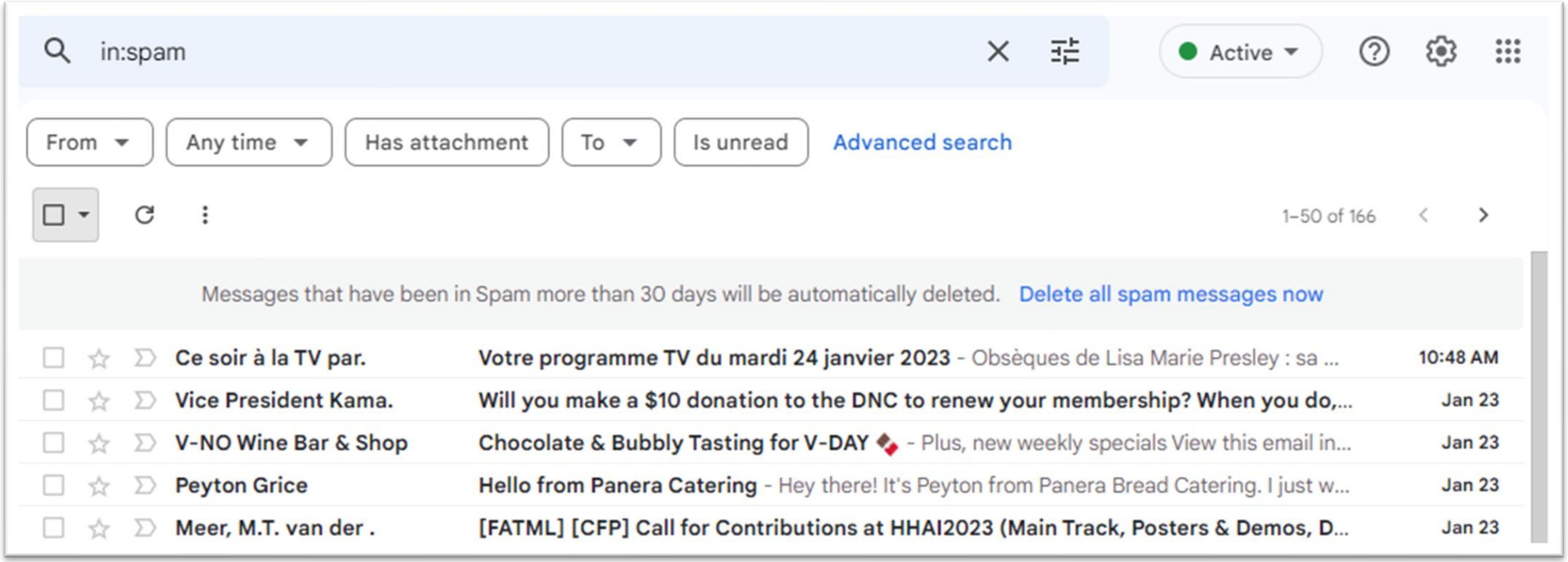
NLP example: machine translation



No tengo idea de lo que estoy haciendo. ¿Quién me dejó hacer esto?

Je n'ai aucune idée de ce que je fais. Qui m'a laissé faire ça ?

NLP example: spam detection

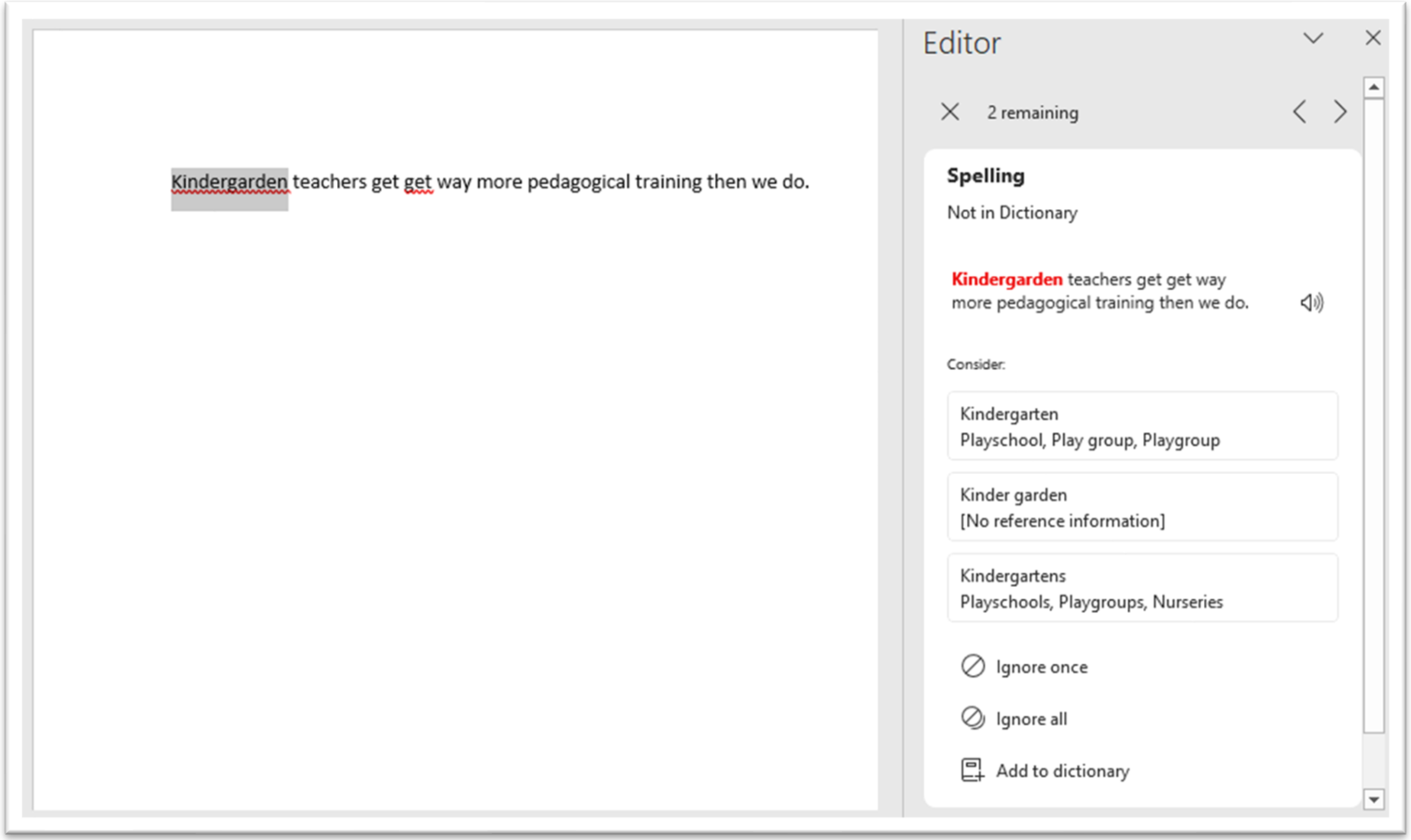


in:spam

Messages that have been in Spam more than 30 days will be automatically deleted.

Ce soir à la TV par.	**Votre programme TV du mardi 24 janvier 2023** - Obsèques de Lisa Marie Presley : sa ...	10:48 AM
Vice President Kama.	**Will you make a $10 donation to the DNC to renew your membership? When you do,...**	Jan 23
V-NO Wine Bar & Shop	**Chocolate & Bubbly Tasting for V-DAY** 🍫 - Plus, new weekly specials View this email in...	Jan 23
Peyton Grice	**Hello from Panera Catering** - Hey there! It's Peyton from Panera Bread Catering. I just w...	Jan 23
Meer, M.T. van der .	**[FATML] [CFP] Call for Contributions at HHAI2023 (Main Track, Posters & Demos, D...**	Jan 23

NLP example: grammar and spell-checker



Kindergarden teachers get get way more pedagogical training then we do.

Editor

2 remaining

Spelling

Not in Dictionary

Kindergarden teachers get get way more pedagogical training then we do.

Consider:

Kindergarten
Playschool, Play group, Playgroup

Kinder garden
[No reference information]

Kindergartens
Playschools, Playgroups, Nurseries

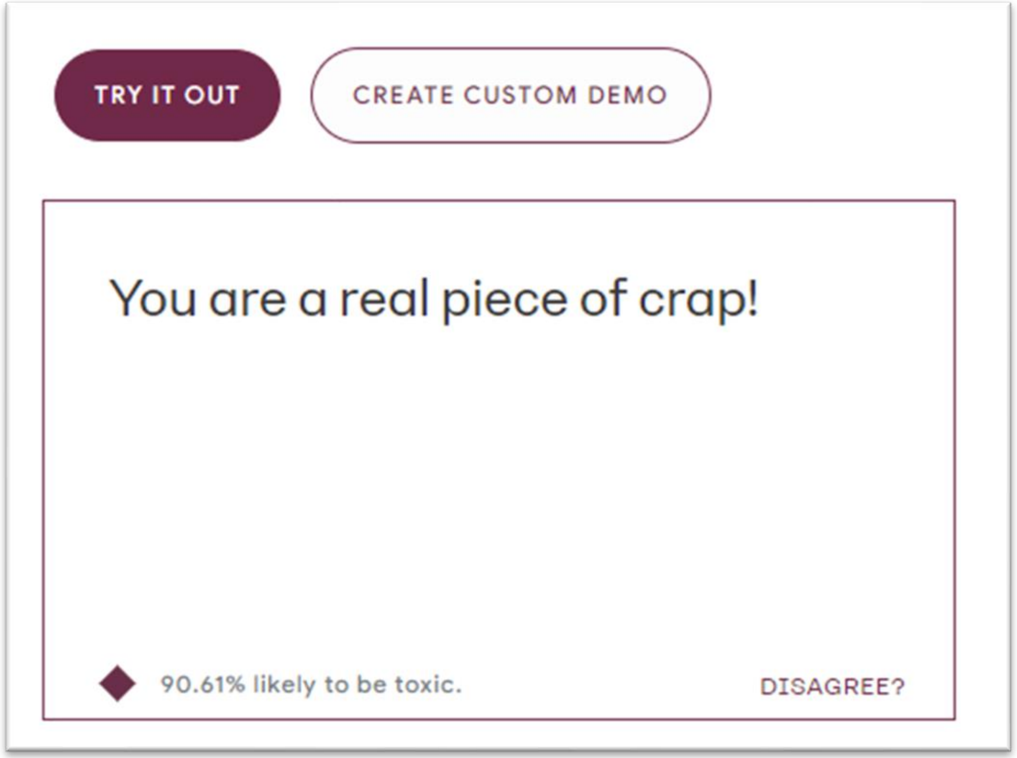
Ignore once

Ignore all

Add to dictionary

NLP example: hate speech detection

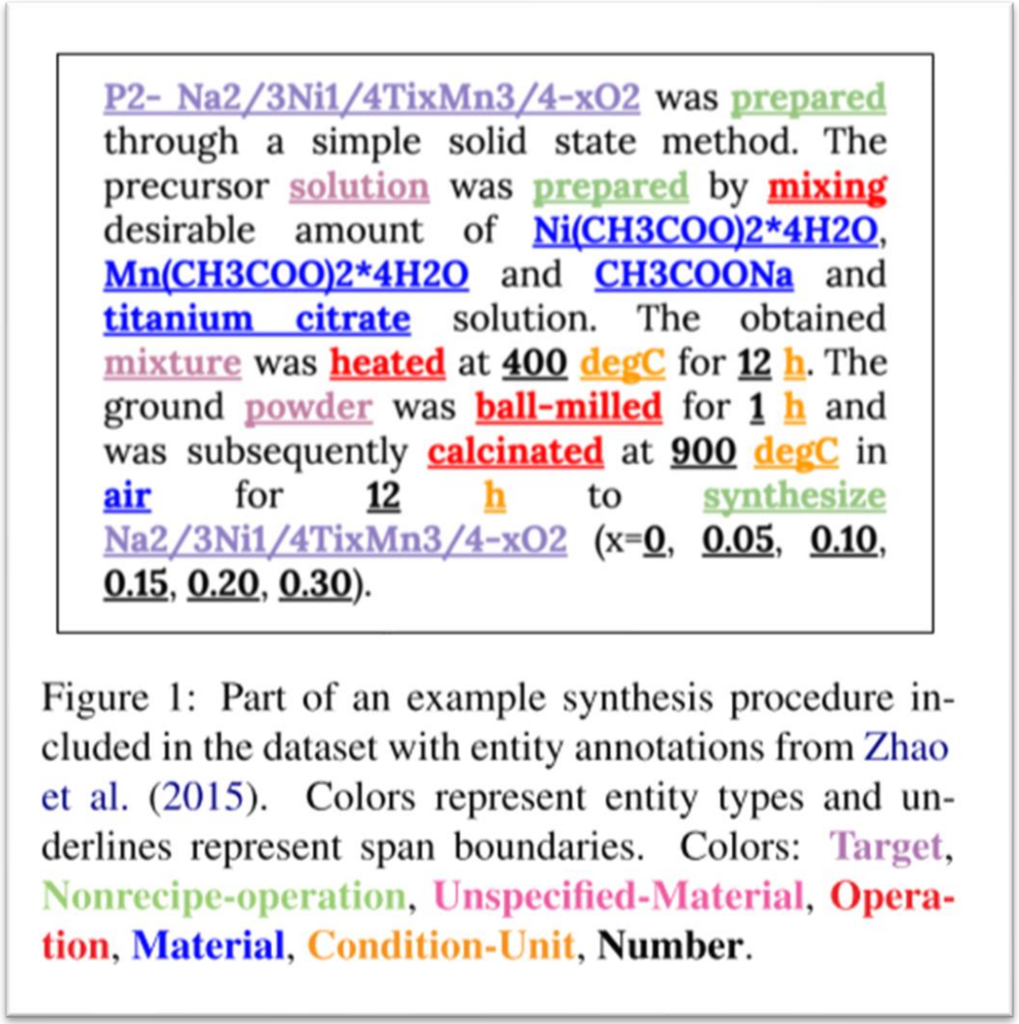
perspectiveapi.com



You are a real piece of crap!

◆ 90.61% likely to be toxic.

NLP example: information extraction

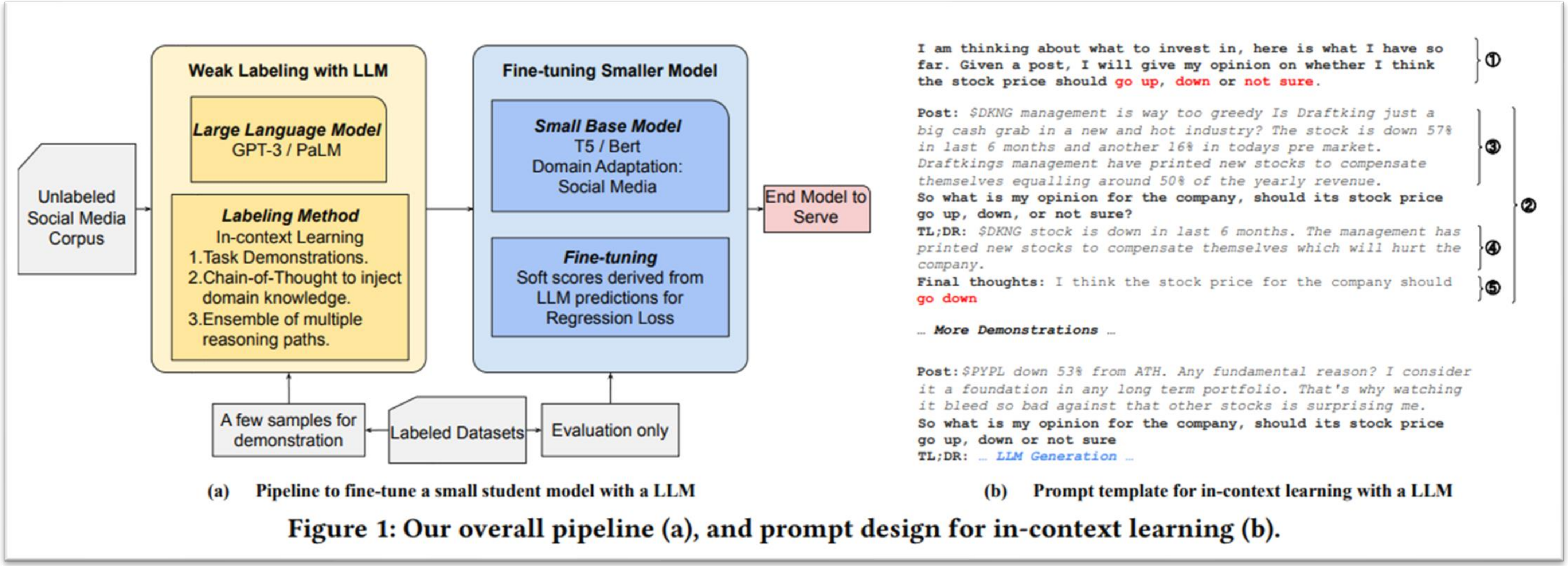


P2- Na2/3Ni1/4TixMn3/4-xO2 was prepared through a simple solid state method. The precursor solution was prepared by mixing desirable amount of Ni(CH3COO)2*4H2O, Mn(CH3COO)2*4H2O and CH3COONa and titanium citrate solution. The obtained mixture was heated at 400 degC for 12 h. The ground powder was ball-milled for 1 h and was subsequently calcinated at 900 degC in air for 12 h to synthesize Na2/3Ni1/4TixMn3/4-xO2 (x=0, 0.05, 0.10, 0.15, 0.20, 0.30).

Figure 1: Part of an example synthesis procedure included in the dataset with entity annotations from Zhao et al. (2015). Colors represent entity types and underlines represent span boundaries. Colors: Target, Nonrecipe-operation, Unspecified-Material, Operation, Material, Condition-Unit, Number.

Tim O'Gorman, Zach Jensen, Sheshera Mysore, Kevin Huang, Rubayyat Mahbub, Elsa Olivetti, and Andrew McCallum. 2021. MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*

NLP example: predicting the stock market


Weak Labeling with LLM
Large Language Model
GPT-3 / PaLM
Labeling Method
In-context Learning
1.Task Demonstrations.
2.Chain-of-Thought to inject domain knowledge.
3.Ensemble of multiple reasoning paths.
Unlabeled Social Media Corpus
Fine-tuning Smaller Model
Small Base Model
T5 / Bert
Domain Adaptation: Social Media
Fine-tuning
Soft scores derived from LLM predictions for Regression Loss
End Model to Serve
A few samples for demonstration
Labeled Datasets
Evaluation only

I am thinking about what to invest in, here is what I have so far. Given a post, I will give my opinion on whether I think the stock price should go up, down or not sure. ①

②

Post: *$DKNG management is way too greedy Is Draftking just a big cash grab in a new and hot industry? The stock is down 57% in last 6 months and another 16% in todays pre market. Draftkings management have printed new stocks to compensate themselves equalling around 50% of the yearly revenue.* ③
So what is my opinion for the company, should its stock price go up, down, or not sure?
TL;DR: *$DKNG stock is down in last 6 months. The management has printed new stocks to compensate themselves which will hurt the company.* ④
Final thoughts: I think the stock price for the company should **go down** ⑤

… ***More Demonstrations*** …

Post: *$PYPL down 53% from ATH. Any fundamental reason? I consider it a foundation in any long term portfolio. That's why watching it bleed so bad against that other stocks is surprising me.*
So what is my opinion for the company, should its stock price go up, down or not sure
TL;DR: … *LLM Generation* …

(a) Pipeline to fine-tune a small student model with a LLM

(b) Prompt template for in-context learning with a LLM

Figure 1: Our overall pipeline (a), and prompt design for in-context learning (b).

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2022. What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. 0 citations (Semantic Scholar/arXiv) [2023-01-24] arXiv:2212.11311 [cs].

NLP vs. machine learning

A few basic types of data:

- Text → NLP
- Images → Computer Vision (CS 855)
- Tabular data
- Time series
- Sound

NLP: machine learning as applied to text (?)

Other miscellaneous terms: “Data science”, “data mining”, “text mining”, “text analytics”

NLP vs. linguistics

Linguistics: study of the structure of language and comparative differences between languages

In the past: used to inform the design of NLP systems

- Recursive neural nets (Socher et al. 2013) a great example

Now: increasingly irrelevant to design of NLP models

- Though it gives us a vocabulary for reasoning about it
- And NLP can still be used to study it
  - e.g. https://medium.com/@jamesasantana/the-rosetta-stone-crumbles-ai-reads-5-000-year-old-tablets-with-98-accuracy-5fc676365735
- And linguistics can mediate how we interact with our models

NLP vs. information retrieval (IR)

Lots of overlap with NLP, but its own field with its own practices and conferences

- Also, doesn’t only retrieve text

Converging with NLP as “Neural IR” comes to dominate

Also people are starting to use ChatGPT as a Google alternative

Okay, but what is NLP

Mostly representations & models

How to **represent** text numerically so that computers can work with it

Models that can learn from humans how to do stuff with text
- Generate it
- Classify it
- Translate it
- Understand it
- Transform it
- Etc.

Why you should care

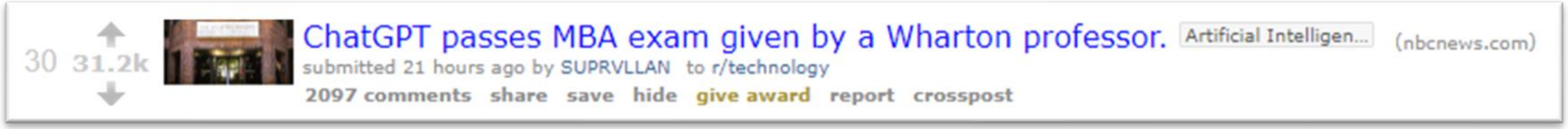
Why should you care

NLP is taking/has taken over the world.

LLMs like ChatGPT poised to have transformative impact on
- Software engineering
- Education
- Business
- Journalism
- Medicine
- Pretty much everything else

ChatGPT

Reddit posts from January 24, 2023

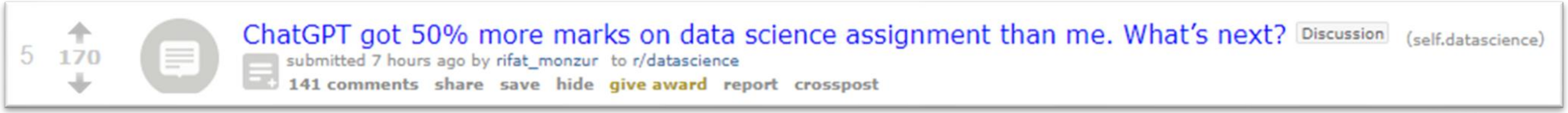


30 31.2k

ChatGPT passes MBA exam given by a Wharton professor. Artificial Intelligen... (nbcnews.com)

submitted 21 hours ago by SUPRVLLAN to r/technology

2097 comments share save hide give award report crosspost

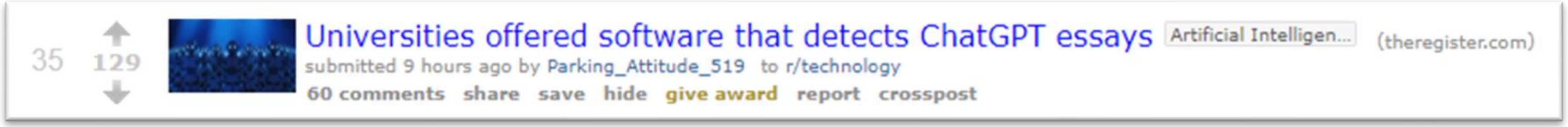


5 170

ChatGPT got 50% more marks on data science assignment than me. What's next? Discussion (self.datascience)

submitted 7 hours ago by rifat_monzur to r/datascience

141 comments share save hide give award report crosspost

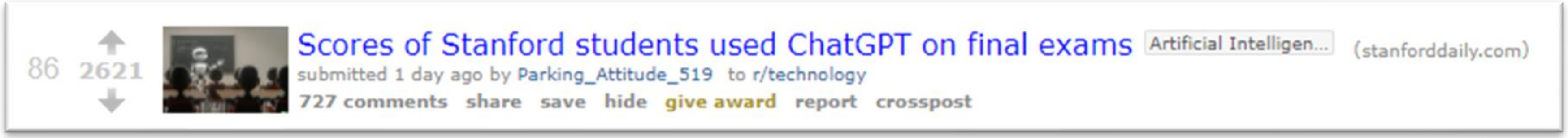


35 129

Universities offered software that detects ChatGPT essays Artificial Intelligen... (theregister.com)

submitted 9 hours ago by Parking_Attitude_519 to r/technology

60 comments share save hide give award report crosspost

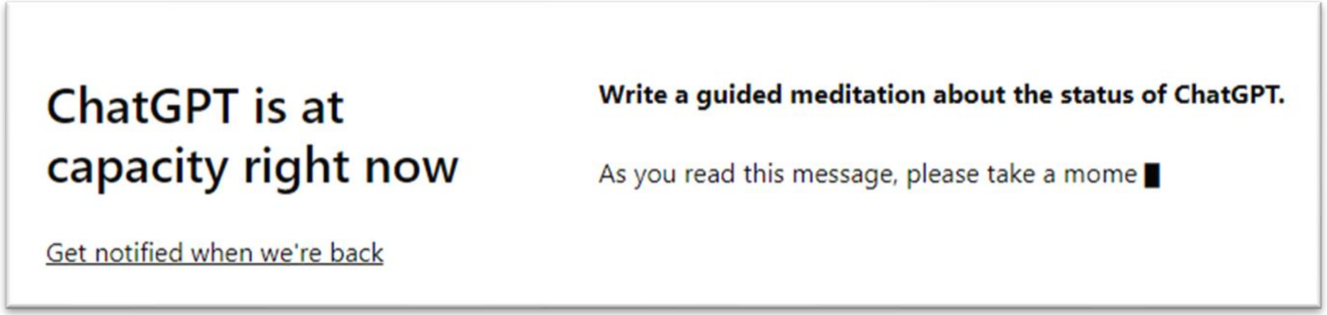


86 2621

Scores of Stanford students used ChatGPT on final exams Artificial Intelligen... (stanforddaily.com)

submitted 1 day ago by Parking_Attitude_519 to r/technology

727 comments share save hide give award report crosspost



ChatGPT is at capacity right now

Get notified when we're back

Write a guided meditation about the status of ChatGPT.

As you read this message, please take a mome

ChatGPT

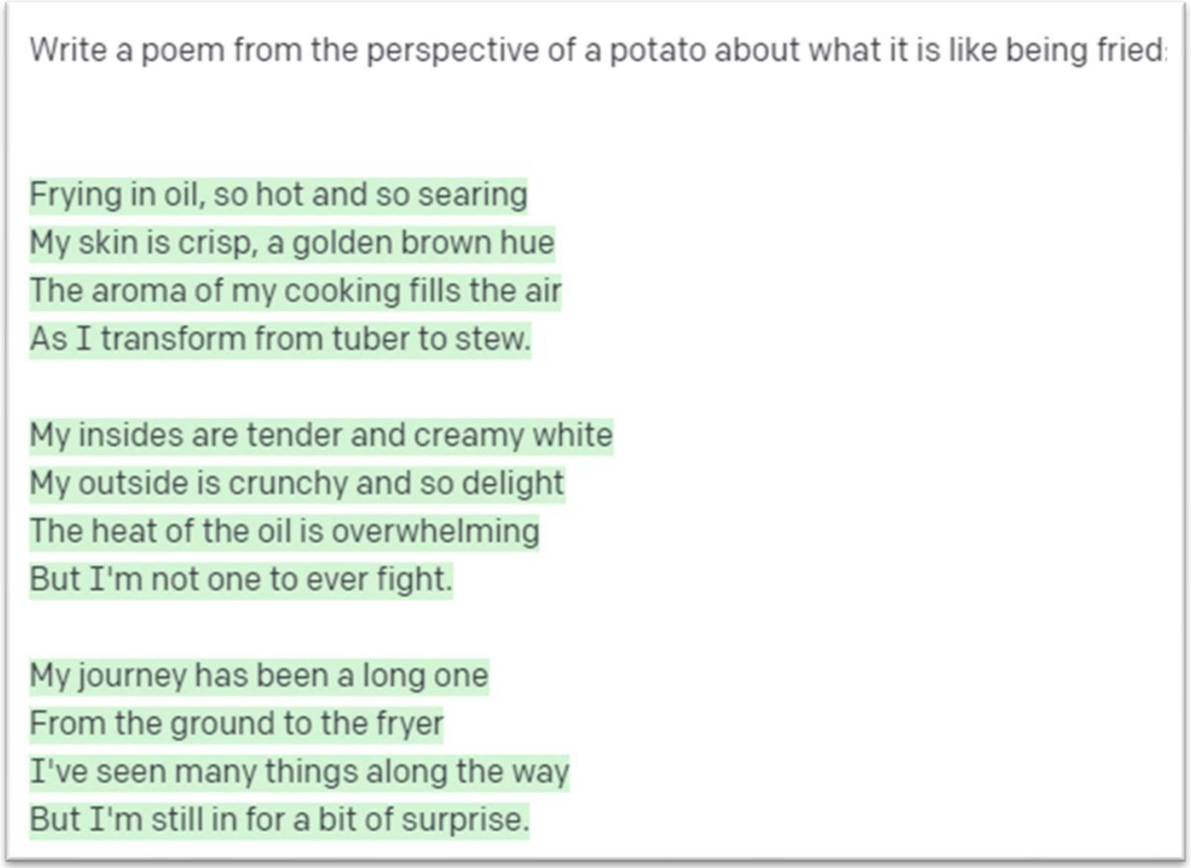
Simple idea, same as any language model: pick the most likely next word based on the previous words

But models are getting **so good** that we're moving toward everyone having a personal secretary in their pocket. It can write:

- Essays for school
- Text for written take-home exams
- Code for assignments
- Emails
- Poetry

Just think about how this is going to change society!

GPT-3



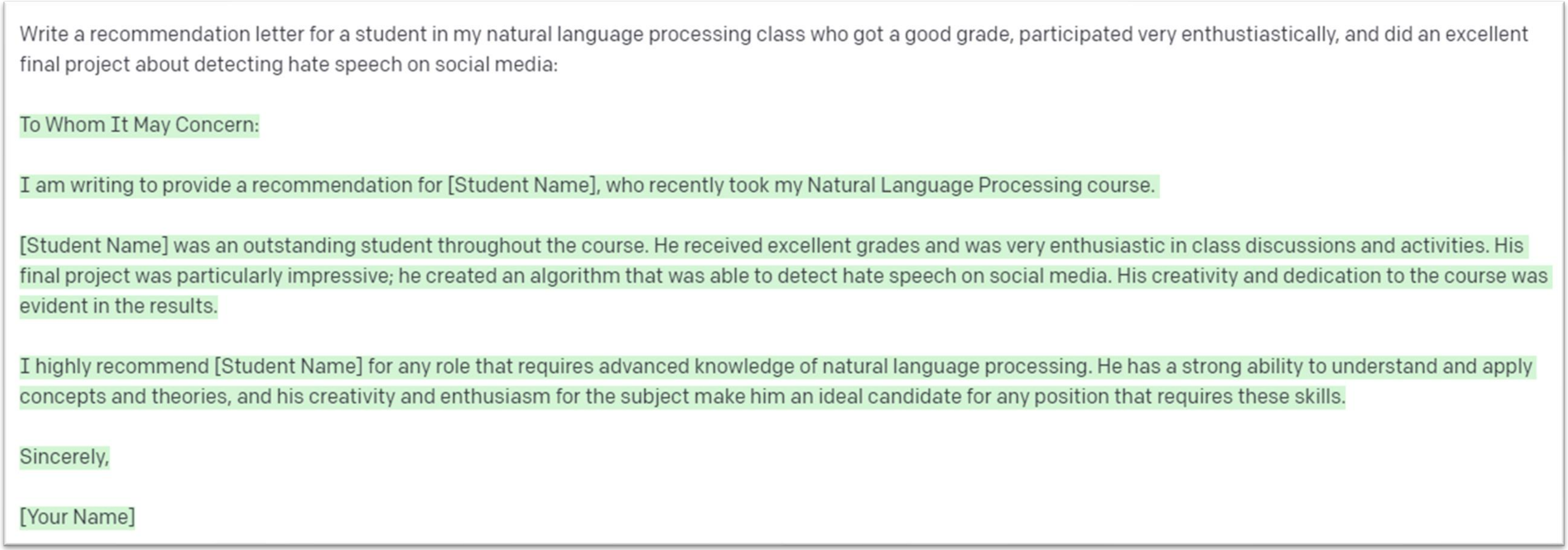
Write a poem from the perspective of a potato about what it is like being fried:

Frying in oil, so hot and so searing
My skin is crisp, a golden brown hue
The aroma of my cooking fills the air
As I transform from tuber to stew.

My insides are tender and creamy white
My outside is crunchy and so delight
The heat of the oil is overwhelming
But I'm not one to ever fight.

My journey has been a long one
From the ground to the fryer
I've seen many things along the way
But I'm still in for a bit of surprise.

GPT-3



Write a recommendation letter for a student in my natural language processing class who got a good grade, participated very enthustiastically, and did an excellent final project about detecting hate speech on social media:

To Whom It May Concern:

I am writing to provide a recommendation for [Student Name], who recently took my Natural Language Processing course.

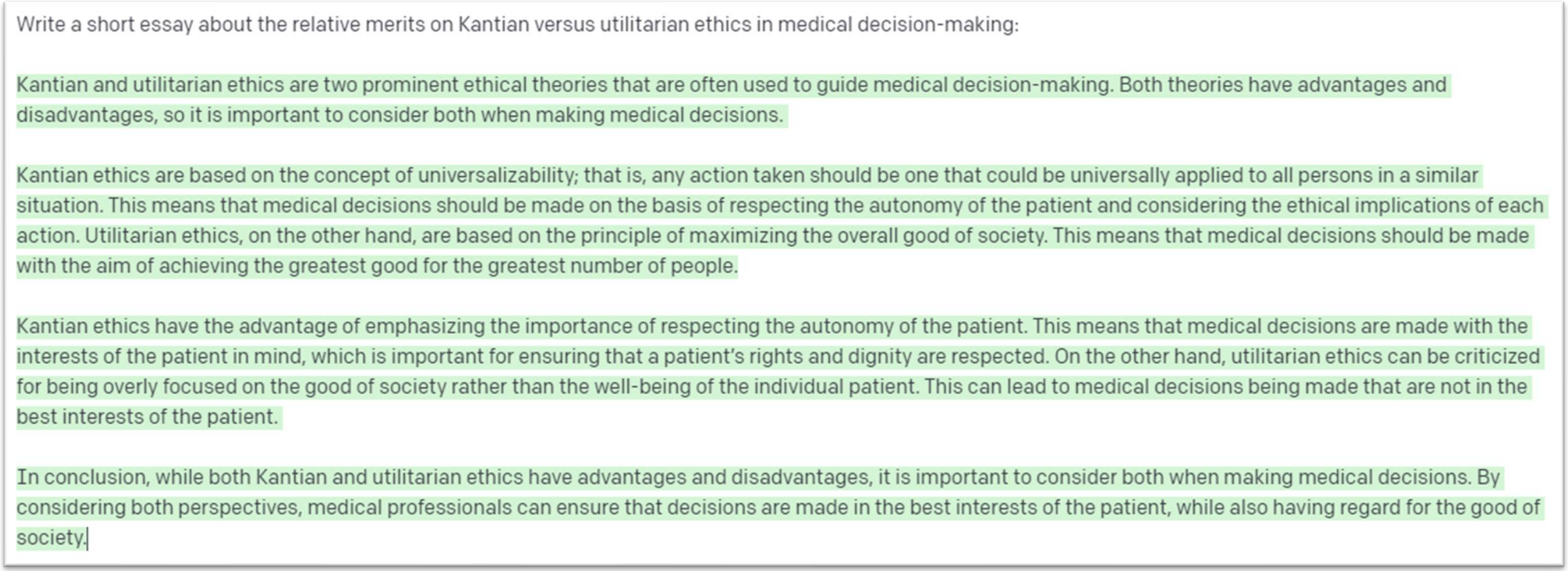
[Student Name] was an outstanding student throughout the course. He received excellent grades and was very enthusiastic in class discussions and activities. His final project was particularly impressive; he created an algorithm that was able to detect hate speech on social media. His creativity and dedication to the course was evident in the results.

I highly recommend [Student Name] for any role that requires advanced knowledge of natural language processing. He has a strong ability to understand and apply concepts and theories, and his creativity and enthusiasm for the subject make him an ideal candidate for any position that requires these skills.

Sincerely,

[Your Name]

GPT-3



Write a short essay about the relative merits on Kantian versus utilitarian ethics in medical decision-making:

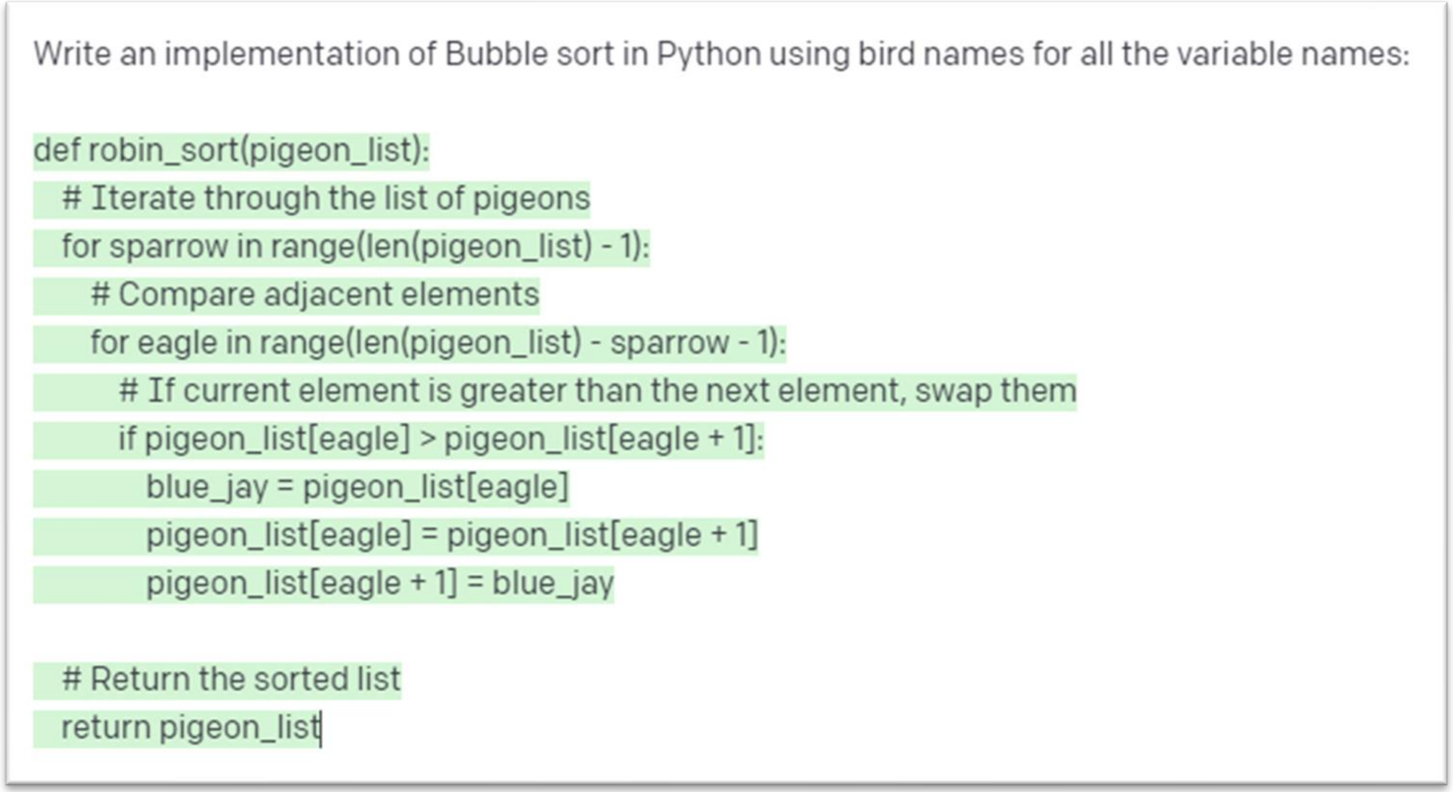
Kantian and utilitarian ethics are two prominent ethical theories that are often used to guide medical decision-making. Both theories have advantages and disadvantages, so it is important to consider both when making medical decisions.

Kantian ethics are based on the concept of universalizability; that is, any action taken should be one that could be universally applied to all persons in a similar situation. This means that medical decisions should be made on the basis of respecting the autonomy of the patient and considering the ethical implications of each action. Utilitarian ethics, on the other hand, are based on the principle of maximizing the overall good of society. This means that medical decisions should be made with the aim of achieving the greatest good for the greatest number of people.

Kantian ethics have the advantage of emphasizing the importance of respecting the autonomy of the patient. This means that medical decisions are made with the interests of the patient in mind, which is important for ensuring that a patient's rights and dignity are respected. On the other hand, utilitarian ethics can be criticized for being overly focused on the good of society rather than the well-being of the individual patient. This can lead to medical decisions being made that are not in the best interests of the patient.

In conclusion, while both Kantian and utilitarian ethics have advantages and disadvantages, it is important to consider both when making medical decisions. By considering both perspectives, medical professionals can ensure that decisions are made in the best interests of the patient, while also having regard for the good of society.

GPT-3

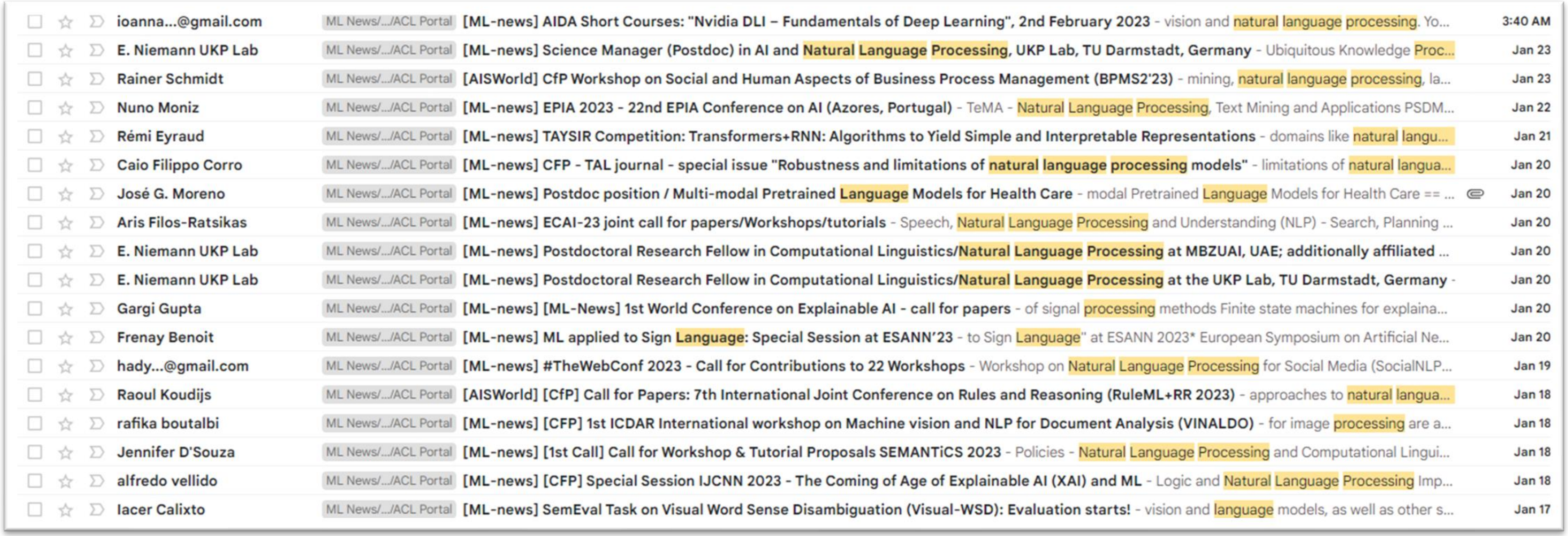


Write an implementation of Bubble sort in Python using bird names for all the variable names:

```
def robin_sort(pigeon_list):
    # Iterate through the list of pigeons
    for sparrow in range(len(pigeon_list) - 1):
        # Compare adjacent elements
        for eagle in range(len(pigeon_list) - sparrow - 1):
            # If current element is greater than the next element, swap them
            if pigeon_list[eagle] > pigeon_list[eagle + 1]:
                blue_jay = pigeon_list[eagle]
                pigeon_list[eagle] = pigeon_list[eagle + 1]
                pigeon_list[eagle + 1] = blue_jay

    # Return the sorted list
    return pigeon_list
```

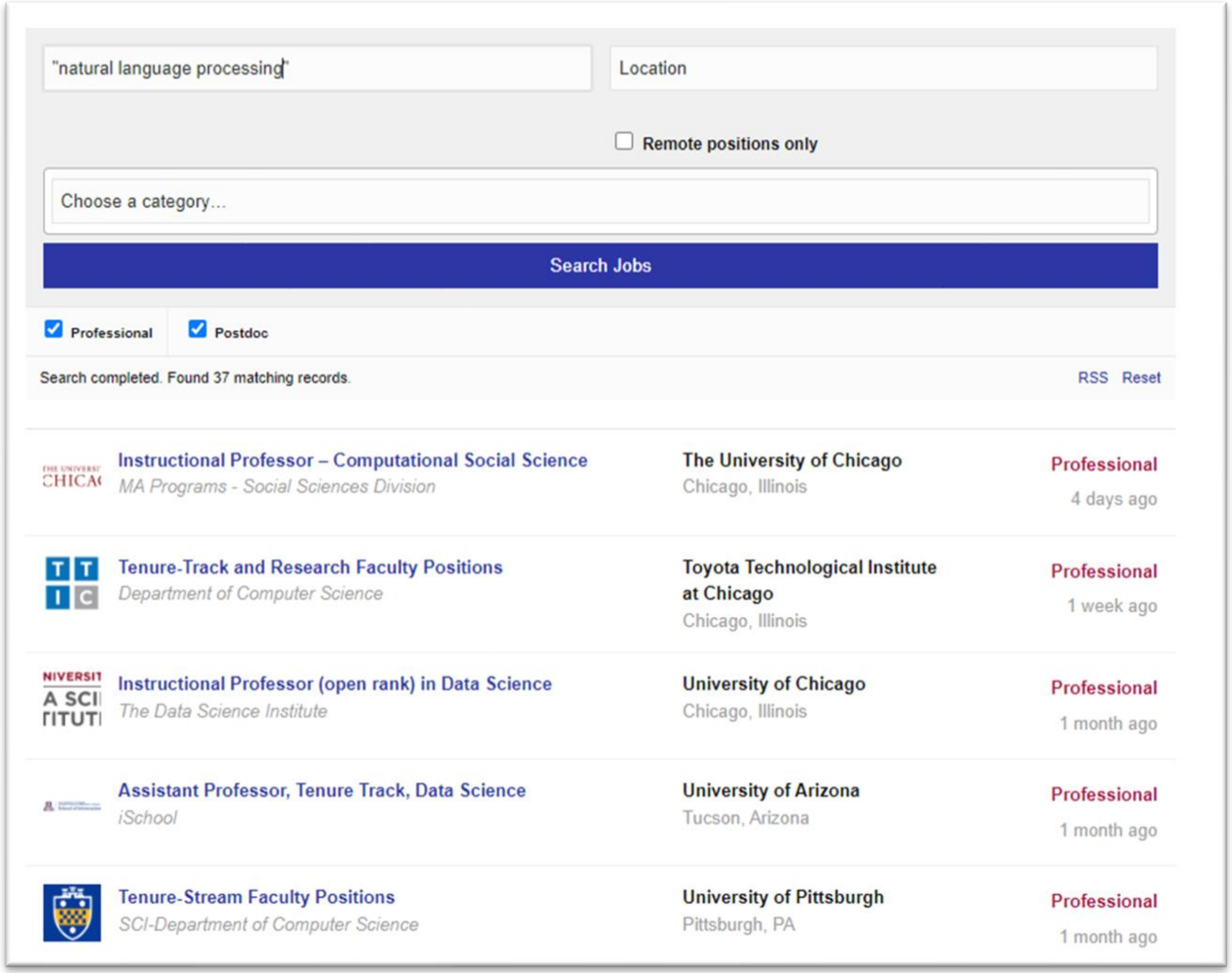
Research opportunities



☐	☆	ioanna...@gmail.com	ML News/.../ACL Portal	**[ML-news] AIDA Short Courses: "Nvidia DLI – Fundamentals of Deep Learning", 2nd February 2023** - vision and natural language processing. Yo...	3:40 AM
☐	☆	E. Niemann UKP Lab	ML News/.../ACL Portal	**[ML-news] Science Manager (Postdoc) in AI and Natural Language Processing, UKP Lab, TU Darmstadt, Germany** - Ubiquitous Knowledge Proc...	Jan 23
☐	☆	Rainer Schmidt	ML News/.../ACL Portal	**[AISWorld] CfP Workshop on Social and Human Aspects of Business Process Management (BPMS2'23)** - mining, natural language processing, la...	Jan 23
☐	☆	Nuno Moniz	ML News/.../ACL Portal	**[ML-news] EPIA 2023 - 22nd EPIA Conference on AI (Azores, Portugal)** - TeMA - Natural Language Processing, Text Mining and Applications PSDM...	Jan 22
☐	☆	Rémi Eyraud	ML News/.../ACL Portal	**[ML-news] TAYSIR Competition: Transformers+RNN: Algorithms to Yield Simple and Interpretable Representations** - domains like natural langu...	Jan 21
☐	☆	Caio Filippo Corro	ML News/.../ACL Portal	**[ML-news] CFP - TAL journal - special issue "Robustness and limitations of natural language processing models"** - limitations of natural langua...	Jan 20
☐	☆	José G. Moreno	ML News/.../ACL Portal	**[ML-news] Postdoc position / Multi-modal Pretrained Language Models for Health Care** - modal Pretrained Language Models for Health Care == ...	Jan 20
☐	☆	Aris Filos-Ratsikas	ML News/.../ACL Portal	**[ML-news] ECAI-23 joint call for papers/Workshops/tutorials** - Speech, Natural Language Processing and Understanding (NLP) - Search, Planning ...	Jan 20
☐	☆	E. Niemann UKP Lab	ML News/.../ACL Portal	**[ML-news] Postdoctoral Research Fellow in Computational Linguistics/Natural Language Processing at MBZUAI, UAE; additionally affiliated ...**	Jan 20
☐	☆	E. Niemann UKP Lab	ML News/.../ACL Portal	**[ML-news] Postdoctoral Research Fellow in Computational Linguistics/Natural Language Processing at the UKP Lab, TU Darmstadt, Germany** -	Jan 20
☐	☆	Gargi Gupta	ML News/.../ACL Portal	**[ML-news] [ML-News] 1st World Conference on Explainable AI - call for papers** - of signal processing methods Finite state machines for explaina...	Jan 20
☐	☆	Frenay Benoit	ML News/.../ACL Portal	**[ML-news] ML applied to Sign Language: Special Session at ESANN'23** - to Sign Language" at ESANN 2023* European Symposium on Artificial Ne...	Jan 20
☐	☆	hady...@gmail.com	ML News/.../ACL Portal	**[ML-news] #TheWebConf 2023 - Call for Contributions to 22 Workshops** - Workshop on Natural Language Processing for Social Media (SocialNLP...	Jan 19
☐	☆	Raoul Koudijs	ML News/.../ACL Portal	**[AISWorld] [CfP] Call for Papers: 7th International Joint Conference on Rules and Reasoning (RuleML+RR 2023)** - approaches to natural langua...	Jan 18
☐	☆	rafika boutalbi	ML News/.../ACL Portal	**[ML-news] [CFP] 1st ICDAR International workshop on Machine vision and NLP for Document Analysis (VINALDO)** - for image processing are a...	Jan 18
☐	☆	Jennifer D'Souza	ML News/.../ACL Portal	**[ML-news] [1st Call] Call for Workshop & Tutorial Proposals SEMANTiCS 2023** - Policies - Natural Language Processing and Computational Lingui...	Jan 18
☐	☆	alfredo vellido	ML News/.../ACL Portal	**[ML-news] [CFP] Special Session IJCNN 2023 - The Coming of Age of Explainable AI (XAI) and ML** - Logic and Natural Language Processing Imp...	Jan 18
☐	☆	Iacer Calixto	ML News/.../ACL Portal	**[ML-news] SemEval Task on Visual Word Sense Disambiguation (Visual-WSD): Evaluation starts!** - vision and language models, as well as other s...	Jan 17

https://groups.google.com/g/ml-news

Academic interest



"natural language processing"

Location

☐ Remote positions only

Choose a category…

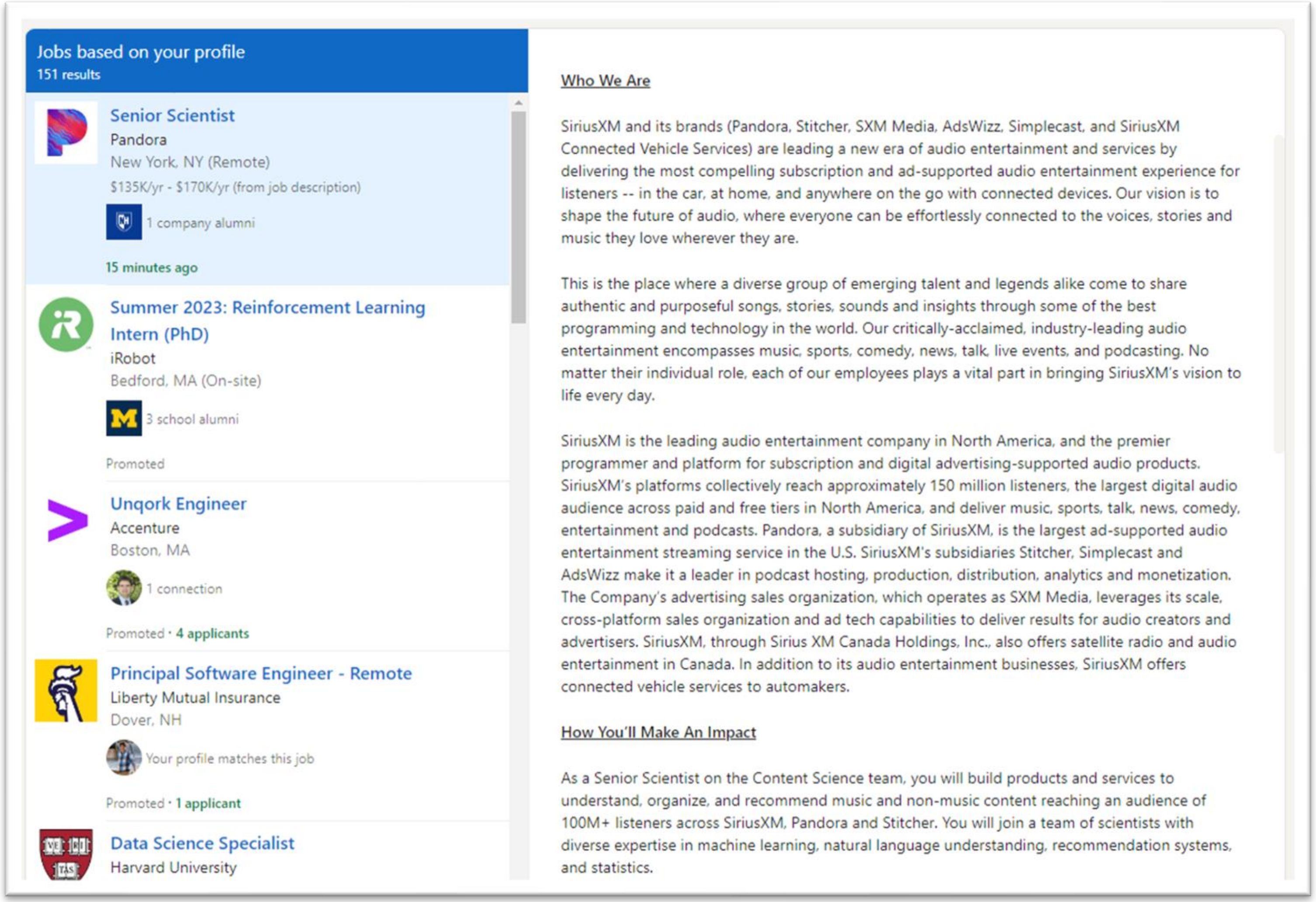
☑ Professional ☑ Postdoc

Search completed. Found 37 matching records.

Instructional Professor – Computational Social Science *MA Programs - Social Sciences Division*	**The University of Chicago** Chicago, Illinois	Professional 4 days ago
Tenure-Track and Research Faculty Positions *Department of Computer Science*	**Toyota Technological Institute at Chicago** Chicago, Illinois	Professional 1 week ago
Instructional Professor (open rank) in Data Science *The Data Science Institute*	**University of Chicago** Chicago, Illinois	Professional 1 month ago
Assistant Professor, Tenure Track, Data Science *iSchool*	**University of Arizona** Tucson, Arizona	Professional 1 month ago
Tenure-Stream Faculty Positions *SCI-Department of Computer Science*	**University of Pittsburgh** Pittsburgh, PA	Professional 1 month ago

https://cra.org/ads/

Job opportunities



Jobs based on your profile
151 results

Senior Scientist
Pandora
New York, NY (Remote)
$135K/yr - $170K/yr (from job description)
1 company alumni
15 minutes ago

Summer 2023: Reinforcement Learning Intern (PhD)
iRobot
Bedford, MA (On-site)
3 school alumni
Promoted

Unqork Engineer
Accenture
Boston, MA
1 connection
Promoted · 4 applicants

Principal Software Engineer - Remote
Liberty Mutual Insurance
Dover, NH
Your profile matches this job
Promoted · 1 applicant

Data Science Specialist
Harvard University

Who We Are

SiriusXM and its brands (Pandora, Stitcher, SXM Media, AdsWizz, Simplecast, and SiriusXM Connected Vehicle Services) are leading a new era of audio entertainment and services by delivering the most compelling subscription and ad-supported audio entertainment experience for listeners -- in the car, at home, and anywhere on the go with connected devices. Our vision is to shape the future of audio, where everyone can be effortlessly connected to the voices, stories and music they love wherever they are.

This is the place where a diverse group of emerging talent and legends alike come to share authentic and purposeful songs, stories, sounds and insights through some of the best programming and technology in the world. Our critically-acclaimed, industry-leading audio entertainment encompasses music, sports, comedy, news, talk, live events, and podcasting. No matter their individual role, each of our employees plays a vital part in bringing SiriusXM's vision to life every day.

SiriusXM is the leading audio entertainment company in North America, and the premier programmer and platform for subscription and digital advertising-supported audio products. SiriusXM's platforms collectively reach approximately 150 million listeners, the largest digital audio audience across paid and free tiers in North America, and deliver music, sports, talk, news, comedy, entertainment and podcasts. Pandora, a subsidiary of SiriusXM, is the largest ad-supported audio entertainment streaming service in the U.S. SiriusXM's subsidiaries Stitcher, Simplecast and AdsWizz make it a leader in podcast hosting, production, distribution, analytics and monetization. The Company's advertising sales organization, which operates as SXM Media, leverages its scale, cross-platform sales organization and ad tech capabilities to deliver results for audio creators and advertisers. SiriusXM, through Sirius XM Canada Holdings, Inc., also offers satellite radio and audio entertainment in Canada. In addition to its audio entertainment businesses, SiriusXM offers connected vehicle services to automakers.

How You'll Make An Impact

As a Senior Scientist on the Content Science team, you will build products and services to understand, organize, and recommend music and non-music content reaching an audience of 100M+ listeners across SiriusXM, Pandora and Stitcher. You will join a team of scientists with diverse expertise in machine learning, natural language understanding, recommendation systems, and statistics.

https://www.linkedin.com

Intellectually interesting

My interest: understand the hidden patterns in written text and communication

Want to major in English? Do NLP instead.

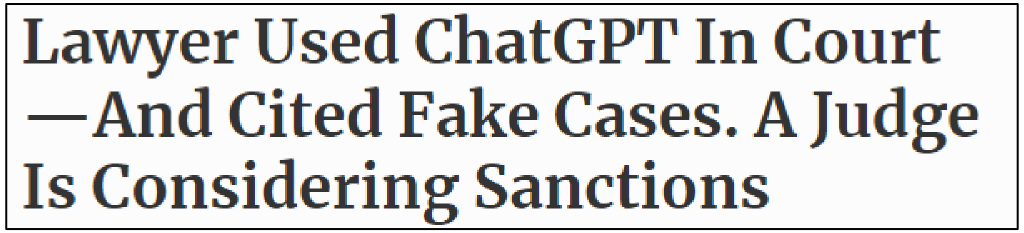
Is it overhyped?

Well, yes.

The extreme over-optimists have been running up against the limitations.

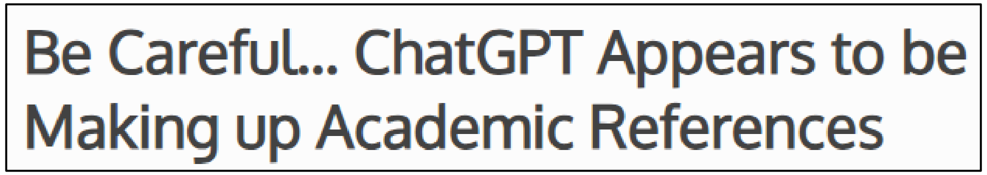
Problems include:

- Hallucination
- Losing logical cohesion over long document/code
- Lack of creativity



Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

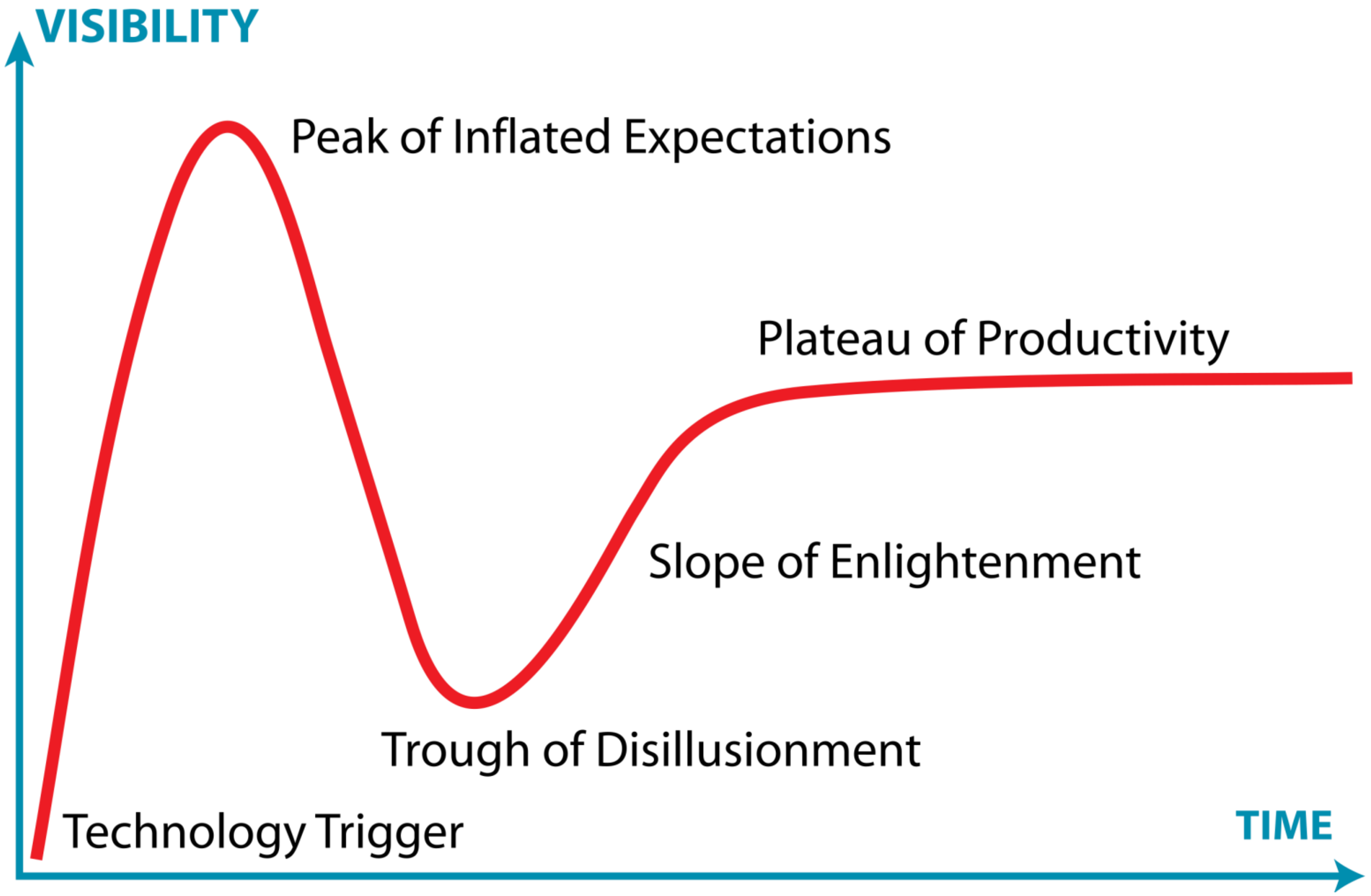
https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=418cc27a7c7f



Be Careful... ChatGPT Appears to be Making up Academic References

https://oxford-review.com/chatgpt-making-up-references/

Hype cycle



https://en.wikipedia.org/wiki/Gartner_hype_cycle

This course

Basic structure

Syllabus: https://shcarton.github.io/cs780spring2023/

Structure
- Biweekly lecture
- Attendance “required”
- Office hours after class on TBD

Assignments
- 1.5-weekly homeworks
- Final project
  - Last semester, medium article
  - This semester, TBD
- (New) final exam

Homework assignments

Written and completed in Google Colab (Python)

Individual work*

Late policy: -10% per day, up to 5 days later

Please don’t use ChatGPT

- Honestly, probably more effort than it is worth

Topics

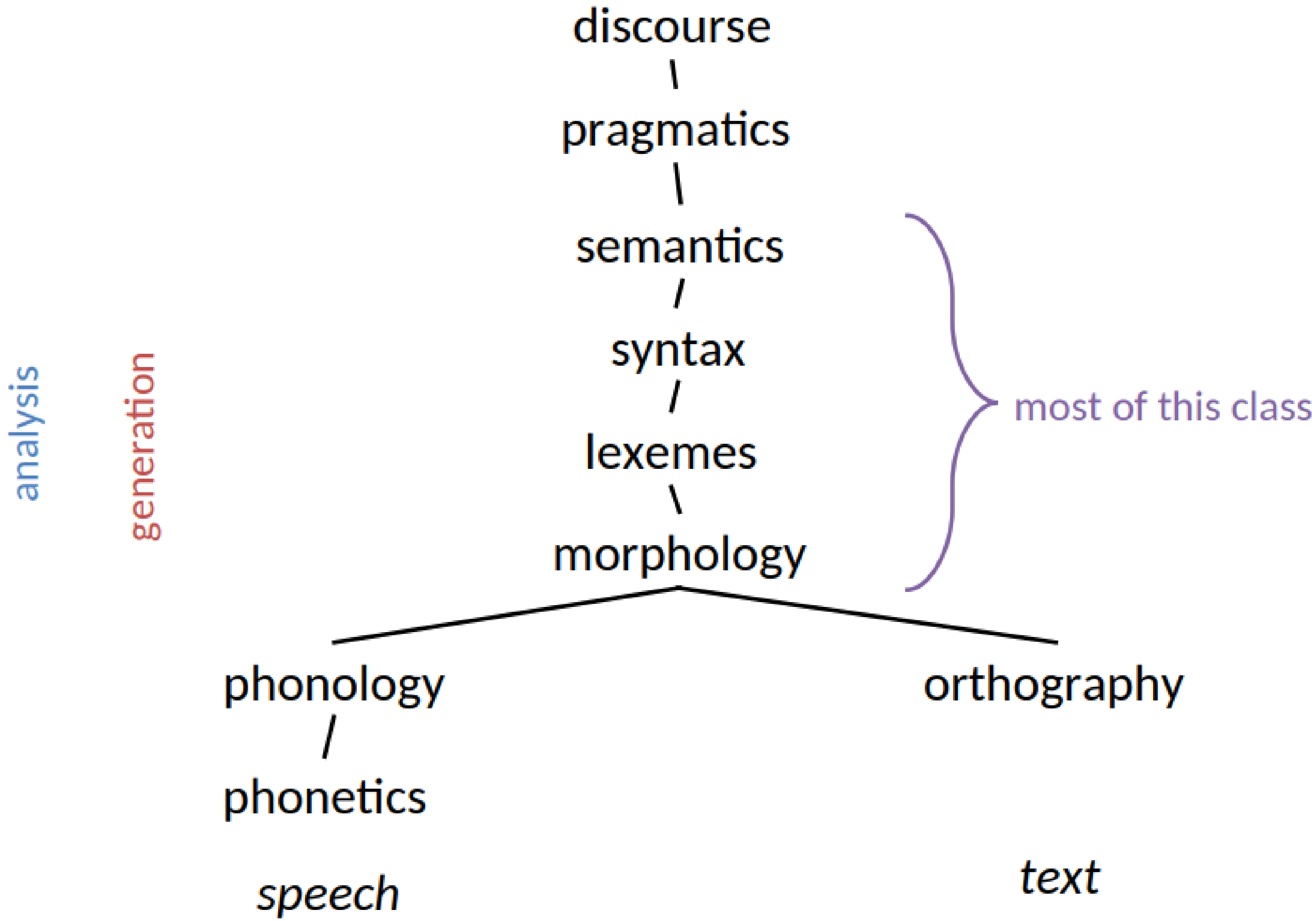
Compare/contrast with CMU

- Basic NLP: http://demo.clab.cs.cmu.edu/NLP/
- Advanced NLP: http://www.phontron.com/class/anlp2022/
- Neural nets for NLP: https://www.phontron.com/class/nn4nlp2021/

This course

- Very little linguistics
- Almost entirely ML
- First 4 weeks: classical statistical/probabilistic models
- Next 7 weeks: neural methods up through transformer-based models(e.g. ChatGPT)
- Last few weeks: special topics (interpretability, fairness, etc.)

Levels of linguistic representation



discourse

pragmatics

semantics

syntax

lexemes

morphology

phonology

phonetics

orthography

speech

text

most of this class

analysis

generation

Rec letters & research opportunities

Happy to write letters and supervise research projects—part of my job!

Recommendation letters

- Come to class & participate
- Let me know at least 6 weeks before you need the letter
  - Same goes for all letter writers
  - Ideally let me know now-ish so I can pay special attention to you

Research

- No summer support for the time being
- Independent study or volunteer basis