CMSC 473/673 Natural Language Processing

Instructor: Lara J. Martin (she/they)

TA: Duong Ta (he)

Schedule

Intro to Lara & Ta

Why are you taking this?

Course logistics

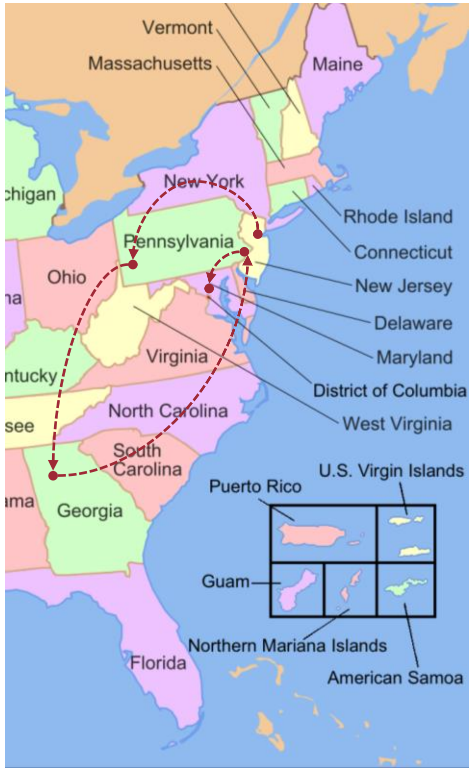
What is NLP?

Who is Lara?

laramar@umbc.edu

laramartin.net

- BS CS & Linguistics @ Rutgers
- MS Language Technologies @ CMU
- PhD Human-Centered Computing @ GT
- CIFellows Postdoc @ UPenn
- Assistant Prof @ UMBC

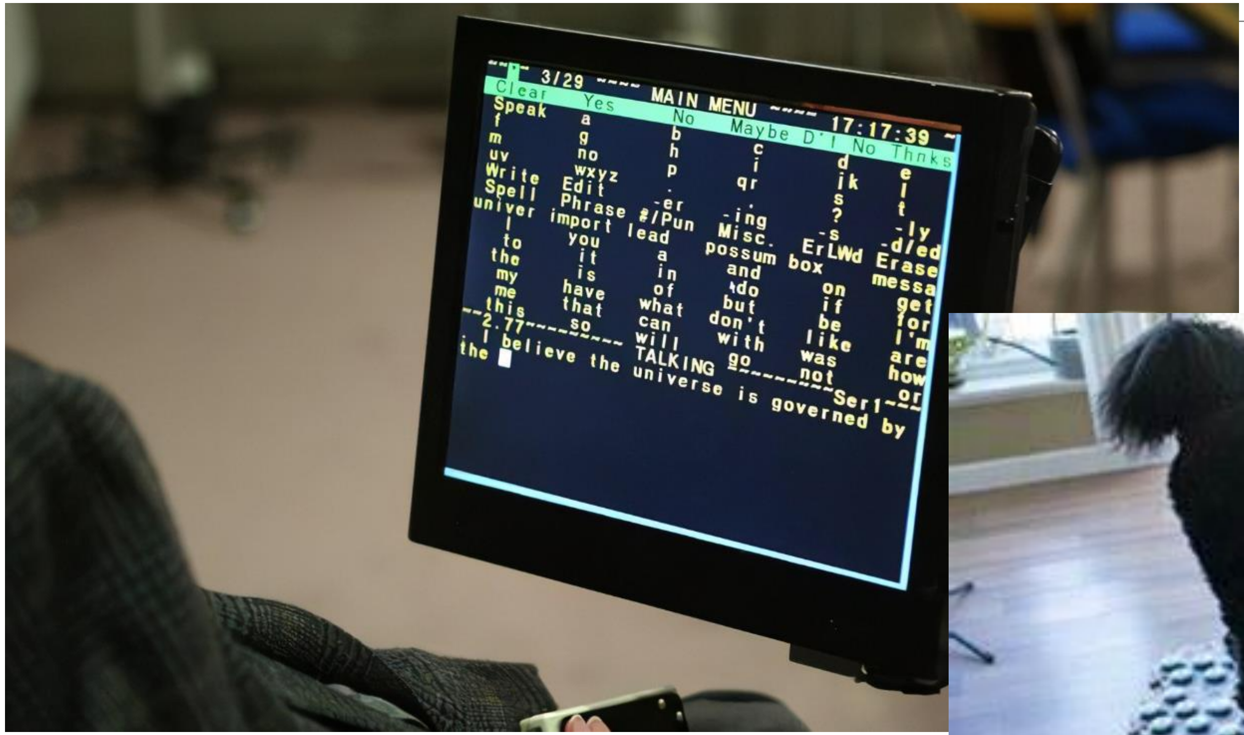


https://upload.wikimedia.org/wikipedia/commons/a/a4/Map_of_USA_with_state_and_territory_names_2.png

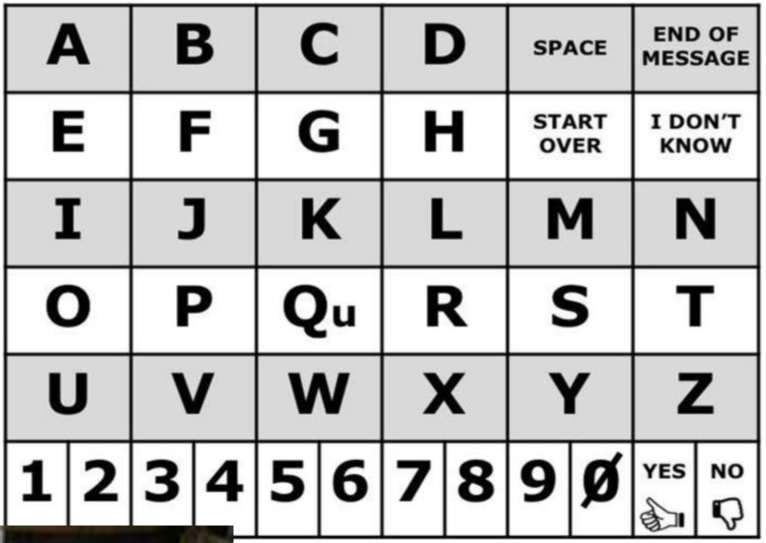
What do I work on?

- Applied NLP
  - human-AI communication
    - Story generation
    - Chatbots
  - computer-mediated human-human communication
    - Speech processing
    - Augmentative and alternative communication (AAC)

Augmentative & Alternative Communication



https://www.abc.es/media/ciencia/2018/03/15/20576450-kgXH--620x349@abc.jpg



Letter Board - AEIOU format

unl.edu/documents/secd/forms/Letter-Boards.png

https://bdnews24.com/lifestyle/2021/05/30/the-talking-dog-of-tiktok

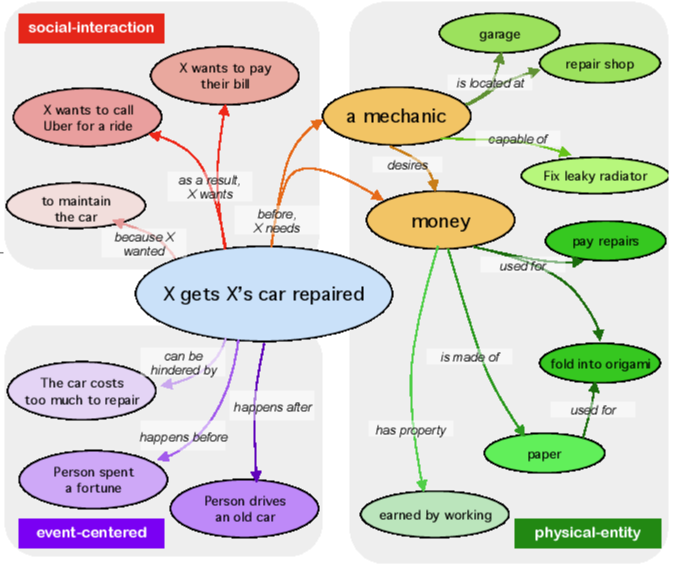
What do I work on?

- Applied NLP
  - human-AI communication
    - Story generation
    - Dialog systems
  - computer-mediated human-human communication
    - Speech processing
    - Augmentative and alternative communication (AAC)
- Using neurosymbolic methods

Neural networks
Neural language models

Old-school AI methods
Discrete, interpretable representations that can help LMs

Knowledge graphs



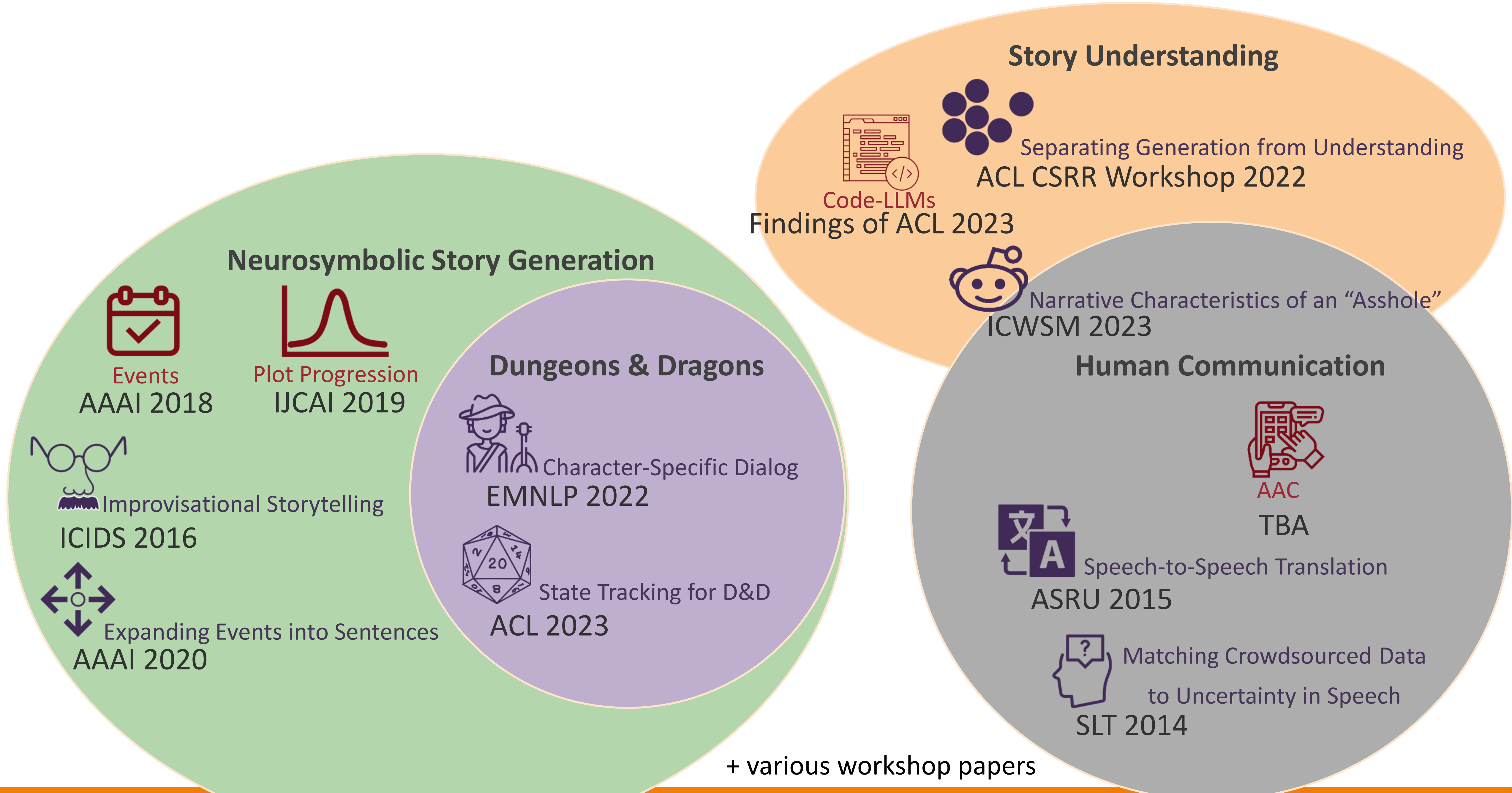
J. D. Hwang *et al.*, "(COMET-)ATOMIC2020: On Symbolic and Neural Commonsense Knowledge Graphs," *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 35, no. 7, pp. 6384–6392, 2021. https://ojs.aaai.org/index.php/AAAI/article/view/16792

Creating structure from sentences

(subject, verb, direct object, modifier)

Original sentence: yoda uses the force to take apart the platform
Event: yoda use force Ø
Generalized Event: <PERSON>0 use-105.1 causal_agent.n.01 Ø

L. J. Martin *et al.*, "Event Representations for Automated Story Generation with Deep Neural Nets," *AAAI*, vol. 32, no. 1, pp. 868–875, Apr. 2018, doi: 10.1609/aaai.v32i1.11430.


Story Understanding
Separating Generation from Understanding
ACL CSRR Workshop 2022
Code-LLMs
Findings of ACL 2023
Narrative Characteristics of an "Asshole"
ICWSM 2023
Neurosymbolic Story Generation
Events
AAAI 2018
Plot Progression
IJCAI 2019
Improvisational Storytelling
ICIDS 2016
Expanding Events into Sentences
AAAI 2020
Dungeons & Dragons
Character-Specific Dialog
EMNLP 2022
State Tracking for D&D
ACL 2023
Human Communication
AAC
TBA
Speech-to-Speech Translation
ASRU 2015
Matching Crowdsourced Data
to Uncertainty in Speech
SLT 2014
+ various workshop papers

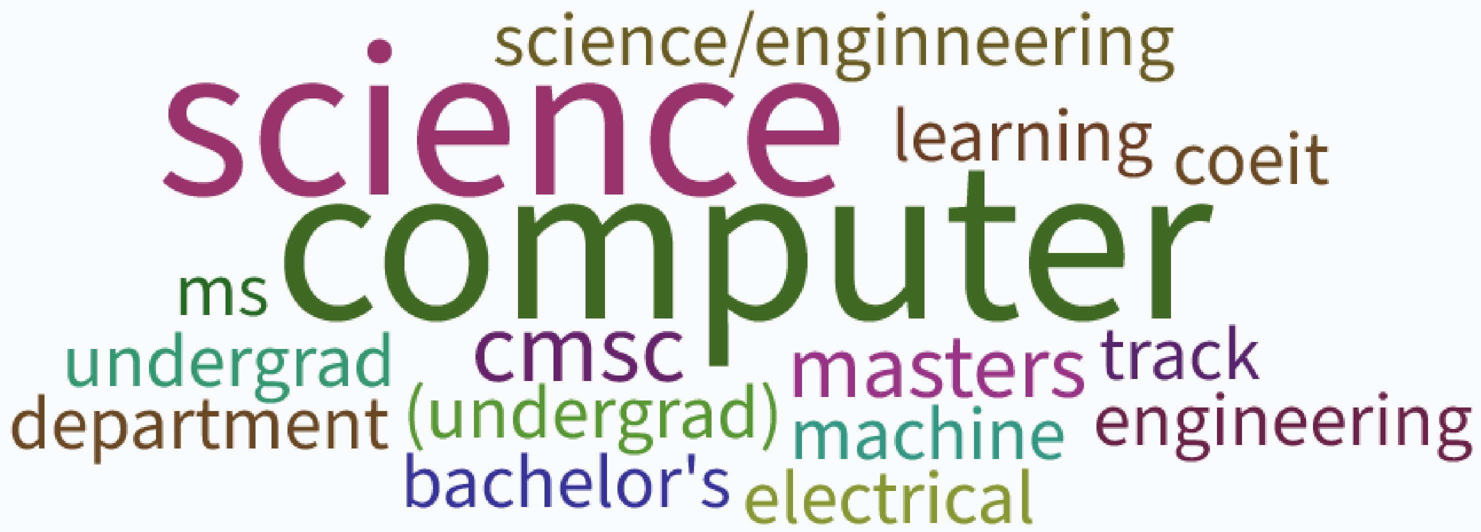
Who is Ta?

dta1@umbc.edu

2nd year PhD student working with Dr. Tim Oates

Has TAed this class with Dr. Frank Ferraro before

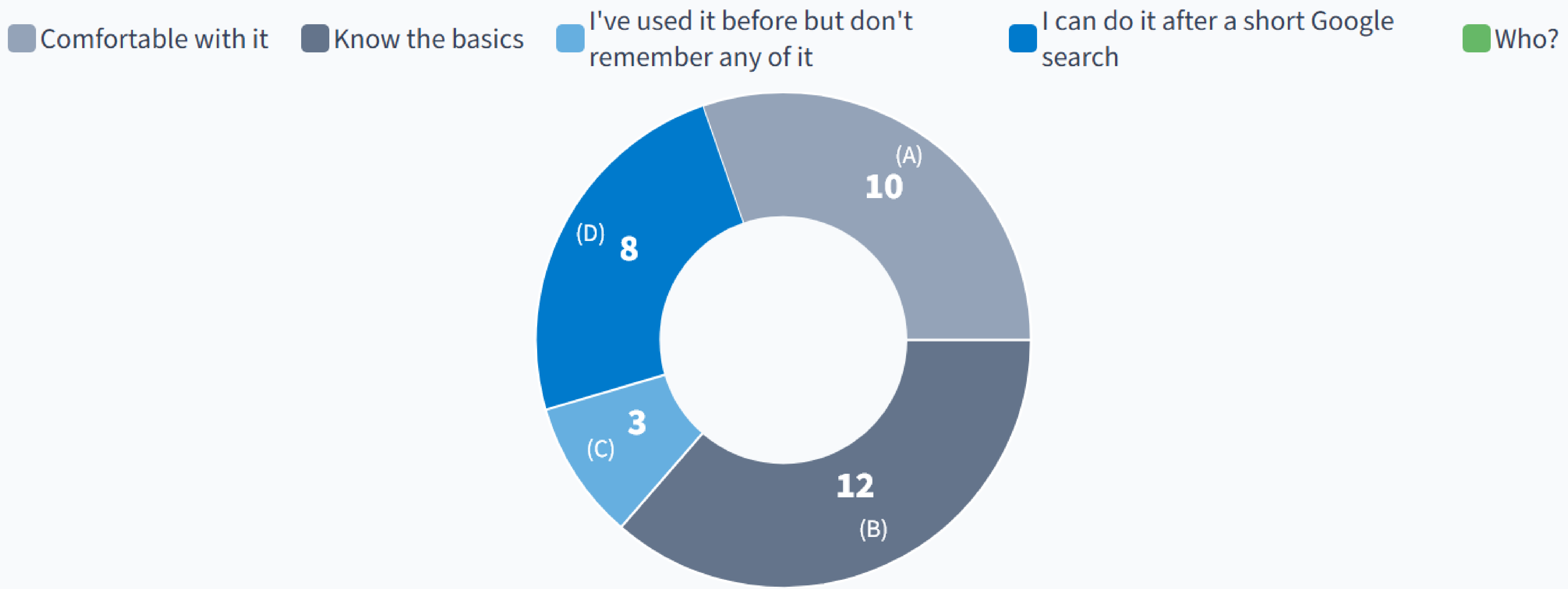
What program/department are you in?



What do you think of when you hear "NLP"?

Why are you taking this course?

How familiar are you with probability?



Logistics

Materials

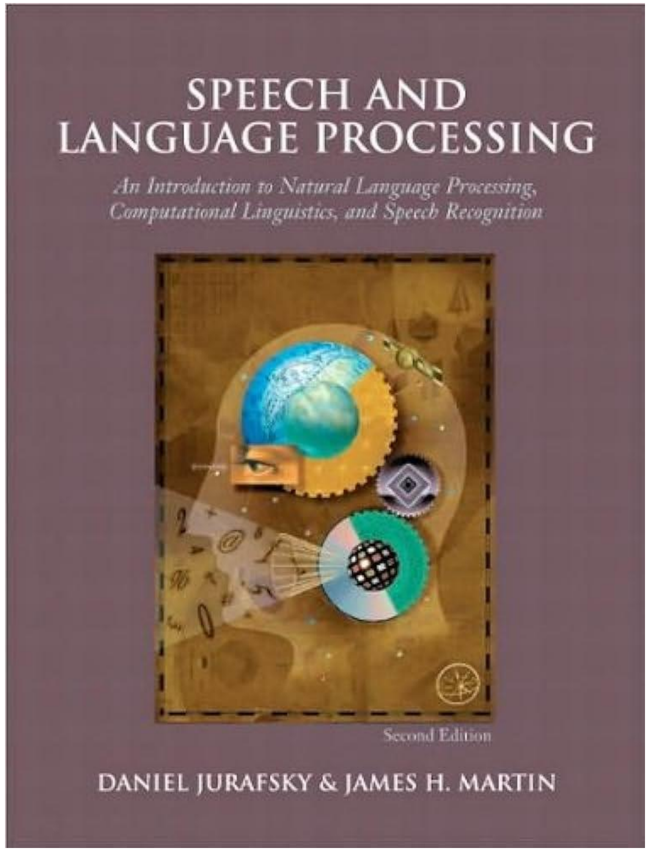
Course Website: https://laramartin.net/NLP-class

- Schedule
- Assignment descriptions
- Policies

Google Classroom: Email Lara (laramar@umbc.edu) to be added

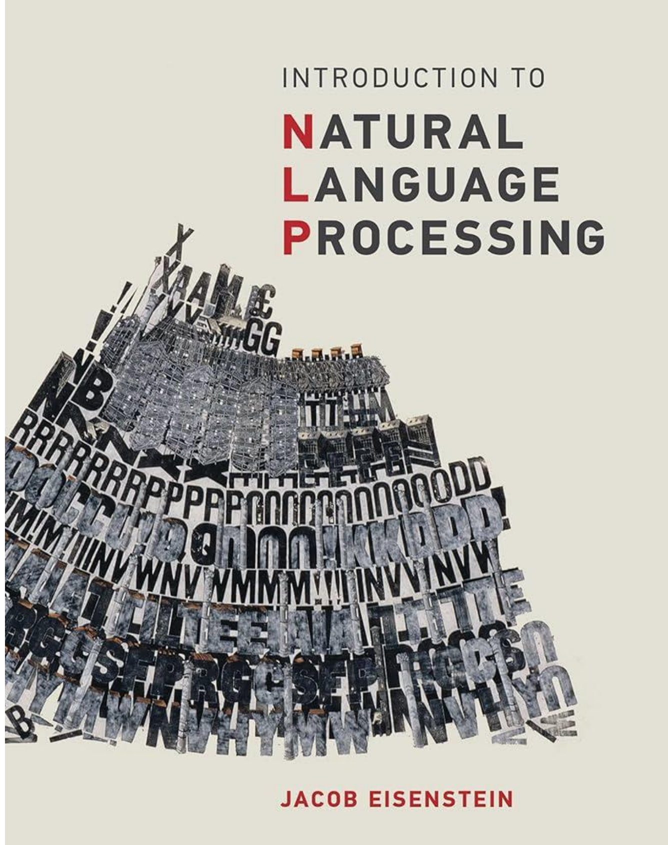
- Assignment submissions
- Discussions
- Grades

Textbooks



"Speech and Language Processing"
by Dan Jurafsky and James Martin
3rd Edition (Draft) online

"Introduction to Natural Language Processing"
by Jacob Eisenstein
Notes PDF on GitHub
(I also have a physical copy)



Images from Amazon

Office Hours

Lara: Tuesdays 3-4pm and Wednesdays 2-3pm

- ITE 216 (or online)
- Also by appointment: https://calendly.com/laramar/schedule

Ta: TBA

- ITE 344 (or online)

Learning Objectives

By the end of the course, you will be able to…

1. Recall common tasks in NLP and formulate problems for them. (HW1)
2. Diagnose and setup appropriate evaluation metrics for a given problem, including determining what an appropriate baseline might be. (HW2)
3. Compare and contrast language models and other NLP methods. (HW3)
4. Implement AI systems that use popular NLP toolkits and libraries. (Grad Assignment)
5. Construct a literature review from state-of-the-art research. (Grad Assignment)
6. Plan and create an NLP system for a particular task. (Project)

Knowledge Checks

Assignment	473 (undergrad)	673 (grad)
Class Knowledge Checks	10%	10%
Homeworks	50%	30%
Project	40%	40%
Grad Assignment	-	20%

Policies

Everyone has 5 free late days (3 max per homework)

- No excuse needed/no need to tell me you're using them

You can collaborate on knowledge checks (in pairs) and the project (3-5 person groups), **not** the homeworks or the grad assignment

Academic Integrity

- If you feel the need to cheat on the assignment to do well on it, please talk to me or Duong first. We can work it out ahead of time, but once you cheat it’s hard to do anything.

If you cheat or plagiarize, you…

- aren’t learning anything
- wasting money paying for tuition
- will get an F on the assignment (at the very least)

More details on course website

Disclaimer about POTS

I have a disability called Postural Orthostatic Tachycardia Syndrome (POTS)

- It means that my blood doesn't always go where I need it to go
- It's a *dynamic disability*, meaning that it's worse some days than others

How does it affect this class?

- I will be lecturing sitting down
- I might get brain fog and have trouble thinking

What about “Large Language Models”?

TECHNOLOGY

ChatGPT

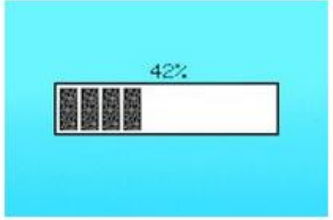
Latest | Search

Jan. 29, 2024

Nvidia's Big Tech Rivals Put Their Own A.I. Chips on the Table

Chafing at their dependence, Amazon, Google, Meta and Microsoft are racing to cut into Nvidia's dominant share of the market.

By CADE METZ, KAREN WEISE and MIKE ISAAC

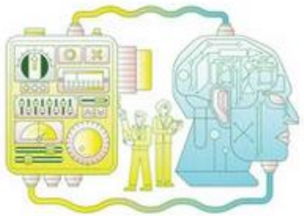


Jan. 29, 2024

Hottest Job in Corporate America? The Executive in Charge of A.I.

Many feared that artificial intelligence would kill jobs. But hospitals, insurance companies and others are creating roles to navigate and harness the disruptive technology.

By YIWEN LU



Jan. 26, 2024

DEALBOOK NEWSLETTER

The F.T.C. Takes on A.I. Deals

The agency is concerned that transactions including Microsoft's $13 billion investment in OpenAI could hinder competition and innovation.

By ANDREW ROSS SORKIN, RAVI MATTU, BERNHARD WARNER, SARAH KESSLER, MICHAEL J. DE LA MERCED, LAUREN HIRSCH and EPHRAT LIVNI

Jan. 25, 2024

Federal Trade Commission Launches Inquiry Into A.I. Deals by Tech Giants

The agency plans to scrutinize Microsoft, Amazon and Google for their investments in the A.I. start-ups OpenAI and Anthropic.

By DAVID MCCABE

GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

1 Introduction

This technical report presents GPT-4, a large multimodal model capable of processing image and text inputs and producing text outputs. Such models are an important area of study as they have the potential to be used in a wide range of applications, such as dialogue systems, text summarization, and machine translation. As such, they have been the subject of substantial interest and progress in recent years [1–34].

One of the main goals of developing such models is to improve their ability to understand and generate natural language text, particularly in more complex and nuanced scenarios. To test its capabilities in such scenarios, GPT-4 was evaluated on a variety of exams originally designed for humans. In these evaluations it performs quite well and often outscores the vast majority of human test takers. For example, on a simulated bar exam, GPT-4 achieves a score that falls in the top 10% of test takers. This contrasts with GPT-3.5, which scores in the bottom 10%.

On a suite of traditional NLP benchmarks, GPT-4 outperforms both previous large language models and most state-of-the-art systems (which often have benchmark-specific training or hand-engineering). On the MMLU benchmark [35, 36], an English-language suite of multiple-choice questions covering 57 subjects, GPT-4 not only outperforms existing models by a considerable margin in English, but

Cool! How does it work?

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

3 Predictable Scaling

A large focus of the GPT-4 project was building a deep learning stack that scales predictably. The primary reason is that for very large training runs like GPT-4, it is not feasible to do extensive model-specific tuning. To address this, we developed infrastructure and optimization methods that have very predictable behavior across multiple scales. These improvements allowed us to reliably predict some aspects of the performance of GPT-4 from smaller models trained using $1,000\times$ – $10,000\times$ less compute.

3.1 Loss Prediction

The final loss of properly-trained large language models is thought to be well approximated by power laws in the amount of compute used to train the model [41, 42, 2, 14, 15].

To verify the scalability of our optimization infrastructure, we predicted GPT-4's final loss on our internal codebase (not part of the training set) by fitting a scaling law with an irreducible loss term (as in Henighan et al. [15]): $L(C) = aC^b + c$, from models trained using the same methodology but using at most 10,000x less compute than GPT-4. This prediction was made shortly after the run

2 / 100 100%

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

3 Predictable Scaling

A large focus of the GPT-4 project was building a deep learning stack that scales predictably. The primary reason is that for very large training runs like GPT-4, it is not feasible to do extensive model-specific tuning. To address this, we developed infrastructure and optimization methods that have very predictable behavior across multiple scales. These improvements allowed us to reliably predict some aspects of the performance of GPT-4 from smaller models trained using $1,000\times$ – $10,000\times$ less compute.

3.1 Loss Prediction

The final loss of properly-trained large language models is thought to be well approximated by power laws in the amount of compute used to train the model [41, 42, 2, 14, 15].

To verify the scalability of our optimization infrastructure, we predicted GPT-4's final loss on our internal codebase (not part of the training set) by fitting a scaling law with an irreducible loss term (as in Henighan et al. [15]): $L(C) = aC^b + c$, from models trained using the same methodology but using at most 10,000x less compute than GPT-4. This prediction was made shortly after the run

2 / 100 100%

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

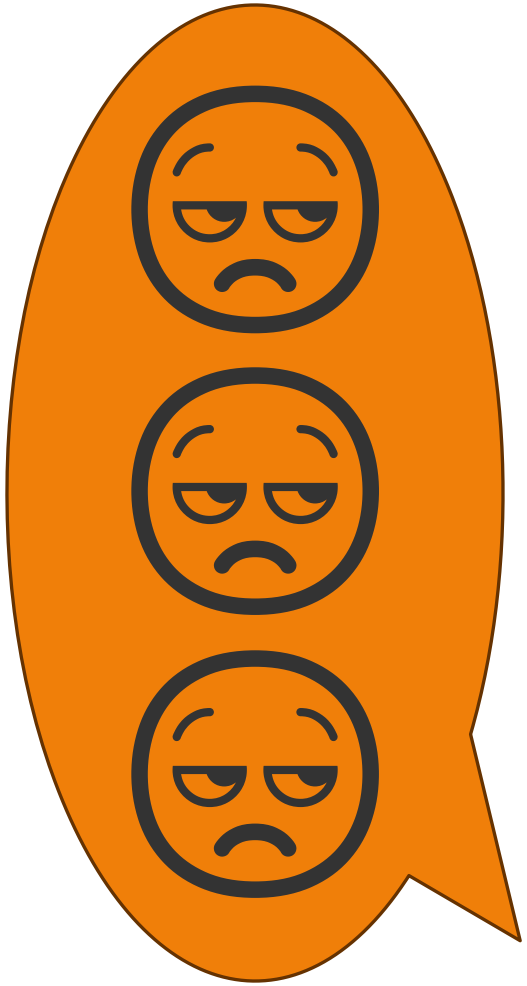
3 Predictable Scaling

A large focus of the GPT-4 project was building a deep learning stack that scales predictably. The primary reason is that for very large training runs like GPT-4, it is not feasible to do extensive model-specific tuning. To address this, we developed infrastructure and optimization methods that have very predictable behavior across multiple scales. These improvements allowed us to reliably predict some aspects of the performance of GPT-4 from smaller models trained using $1,000\times$ – $10,000\times$ less compute.

3.1 Loss Prediction

The final loss of properly-trained large language models is thought to be well approximated by power laws in the amount of compute used to train the model [41, 42, 2, 14, 15].

To verify the scalability of our optimization infrastructure, we predicted GPT-4's final loss on our internal codebase (not part of the training set) by fitting a scaling law with an irreducible loss term (as in Henighan et al. [15]): $L(C) = aC^b + c$, from models trained using the same methodology but using at most 10,000x less compute than GPT-4. This prediction was made shortly after the run



Known Issues about LLMs

- Bad reproducibility
- Copyright issues
- Can’t explain what it’s doing
- Can’t remember things long term
- Confident bullshitter

Based on a slide by Dr. Frank Ferraro

If you want to use ChatGPT

- Make sure you're saying that you used it
- Provide your prompt and the original generation (along with how you edited it)
- Make sure that you're not avoiding the learning objectives by using it

- If you do not say you're using it and I notice, that is an academic integrity violation
- It's okay to use grammar tools (e.g., spell check or Grammarly) or small-scale prediction (e.g., next word prediction, tab completion), provided that they don't change the **substance** of your work

What are some NLP applications that you see in your daily life?

- Auto-complete
- ChatGPT (or LLMs) as thesaurus, filling in acronym, summarization, code debugging
- Information extraction (resume)
- Spell check
- Virtual assistants (wake word, instructions)
- Company-generated summaries (Amazon reviews, search engine)
- Translation tools
- Speech synthesis (TTS)
- Automatic speech recognition (ASR) (speech-to-text)