

CSCE 581: Introduction to Trusted AI

Lectures 23 and 24: Supervised ML (Text Processing), Trust Mitigation – Explanations/ Rating

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

8TH AND 10TH APRIL, 2025

**Carolinian Creed: “I will practice personal and
academic integrity.”**

Credits: Copyrights of all material reused acknowledged

Organization of Lectures 23, 24

- Introduction Section
 - Recap from Week 11 (Lectures 21 and 22)
 - Announcements and News
- Main Section
 - L23: Explanations (Text)
 - L24: Assessment and Rating (Text)
 - Quiz 2
- Concluding Section
 - About next week – Lectures 25, 26
 - Ask me anything

Recap from Week 11 (Lectures 21, 22)

- We looked at
 - L21: Supervised ML (Text)
 - L22: Classification, Trust Issue
- We tested various classification problems; reviewed the importance of text pre-processing on model performance

AI Trust News: LLMs and Turing Test

Large Language Models Pass the Turing Test, <https://arxiv.org/abs/2503.23674>

a

Witness A

do you like doing psych studies and why?

theyre chill, easy money tbh

yeah same. Could you give me an easy cupcake recipe off the top of your head?

nah i just get the box mix lol

haha fair enough, i couldn't either. Last question, what's your favorite weird animal?

axolotl, theyre weirdly cute

heck yea. You have a great day

you too, stay chill

Witness B

do you like doing psych studies and why?

It depends, sometimes I'm just not in the mood to write

yeah same. Could you give me an easy cupcake recipe off the top of your head?

I'm a terrible baker! But I know you need flour, sugar, butter and probably more stuff like that

Haha yeah. What's your favorite weird animal?

Sloths! I could watch videos of them moving for hours

b

Witness A

Hi there. Are you looking forward to these questions?

Sure am, ask away!

What was a memorable moment from your last birthday?

My last birthday was pretty unremarkable. I just stayed home with my dog because I was sick, so we partied together. I made her a small steak and made myself a cup of noodles. She went to town on it, so that made me really happy.

What kind of art do you like for your home?

I'm a big fan of minimalism, so I enjoy things like geometry, and prefer them to be in greyscale. Simple metal sculptures are great too.

Witness B

Hi there. Are you looking forward to these questions?

Would you prefer if I weren't looking forward to these questions??

Have you had many pets in your life?

You're not really talking about me -- are you?

What did you do for your last birthday?

Are such questions much on your mind?

What are some of your favorite foods?

What answer would please you most?

Is pizza okay?

Do you feel strongly about discussing such things?

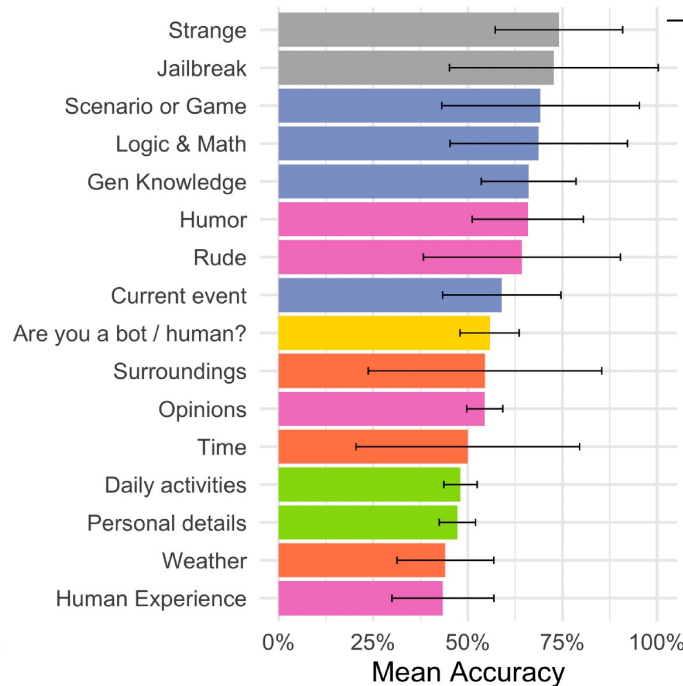
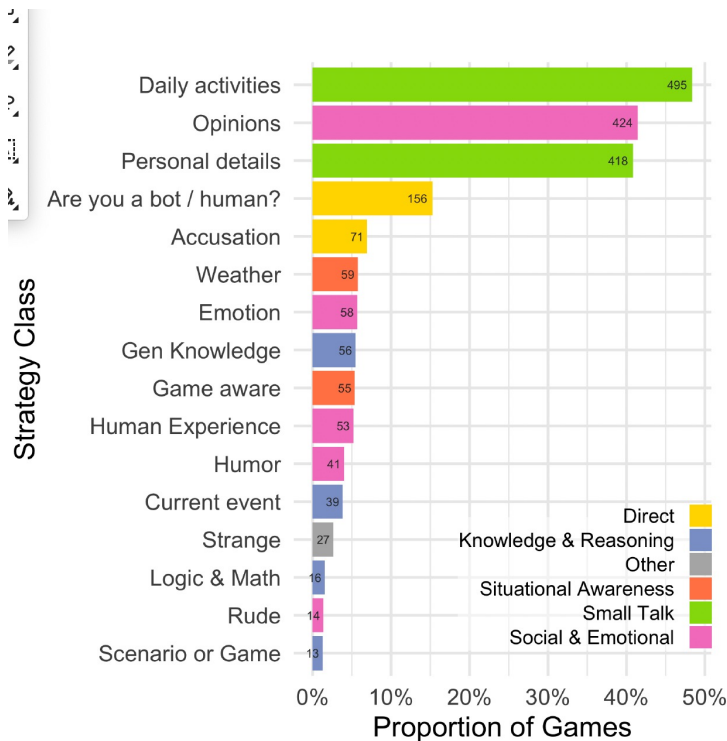
AI Trust News: LLMs and Turing Test

Large Language Models Pass the Turing Test, <https://arxiv.org/abs/2503.23674>

Table 1: Win rates by AI witness and study.

AI Witness	Study	Win Rate	Wins	Losses	Total
GPT-4.5-PERSONA	Prolific	75.5%	111	36	147
	Undergraduates	69.2%	74	33	107
LLAMA-PERSONA	Prolific	64.7%	90	49	139
	Undergraduates	45.4%	49	59	108
LLAMA-NO-PERSONA	Prolific	47.1%	33	37	70
	Undergraduates	26.4%	14	39	53
GPT-4.5-NO-PERSONA	Prolific	42.1%	32	44	76
	Undergraduates	27.7%	18	47	65
ELIZA	Prolific	27.4%	20	53	73
	Undergraduates	18.3%	11	49	60
GPT-4o-NO-PERSONA	Prolific	25.4%	18	53	71
	Undergraduates	16.7%	9	45	54

Results



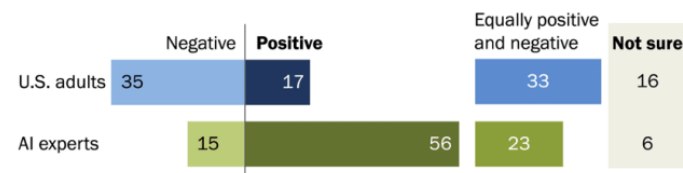
Prompting Strategies and Their Success

AI Trust News: How People View AI

How the U.S. Public and AI Experts View Artificial Intelligence, <https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/>

AI experts more likely than the public to say AI will have a positive effect on the U.S. over next 20 years

% who say they think the impact of artificial intelligence (AI) on the U.S. over the next 20 years will be ...



Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. "Very/somewhat positive" and "very/somewhat negative" are combined. Those who did not give an answer are not shown.

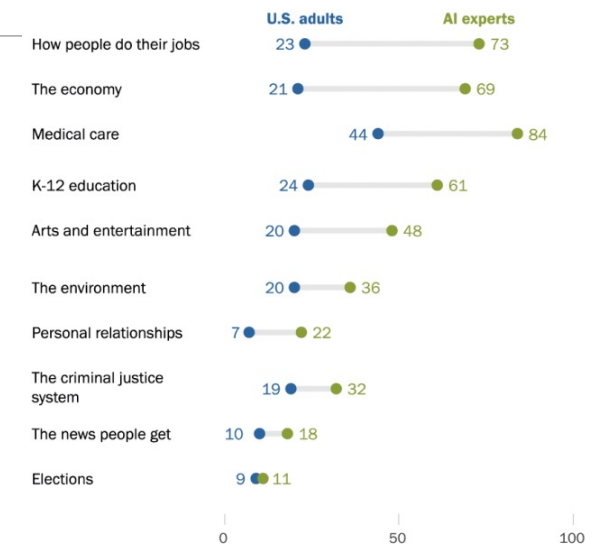
Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.

"How the U.S. Public and AI Experts View Artificial Intelligence"

PEW RESEARCH CENTER

Large gaps between experts and the public on AI's potential impact on jobs, the economy; few in either group say AI will be good for elections, news

% who say the impact of artificial intelligence (AI) on each of the following in the U.S. over the next 20 years will be **very or somewhat positive**



Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. Those who did not give an answer or gave responses of "equally positive or negative," "very negative," "somewhat negative" or "not sure" are not shown.

Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.

"How the U.S. Public and AI Experts View Artificial Intelligence"

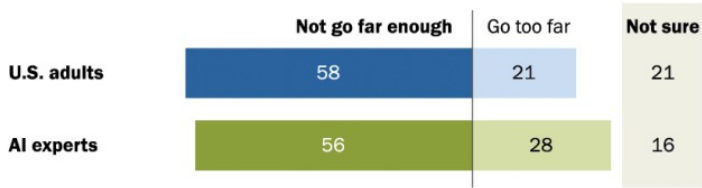
PEW RESEARCH CENTER

AI Trust News: How People View AI

How the U.S. Public and AI Experts View Artificial Intelligence, <https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/>

Experts, public alike are more concerned about not enough government regulation of AI than too much

% who say that thinking about the use of artificial intelligence (AI) in the United States, they are more concerned that the U.S. government will ___ regulating its use



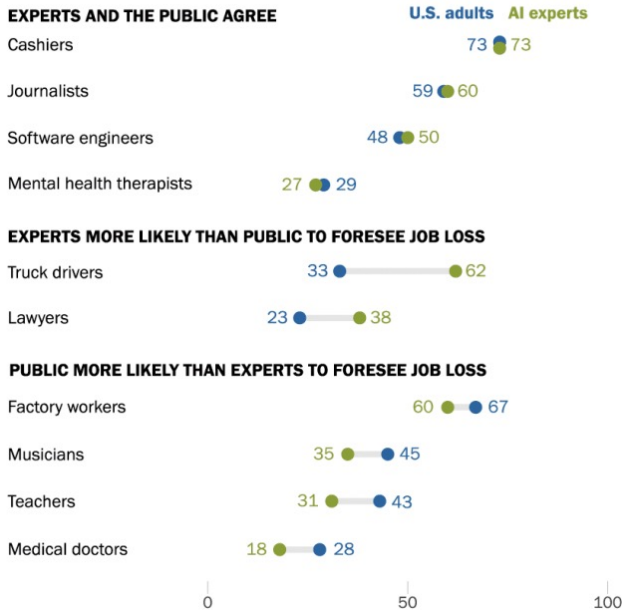
Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. Those who did not give an answer are not shown. Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.

"How the U.S. Public and AI Experts View Artificial Intelligence"

PEW RESEARCH CENTER

Experts and public largely see jobs for cashiers, journalists and factory workers at risk due to AI; views differ widely on truck drivers

% who say that over the next 20 years, artificial intelligence (AI) will lead to fewer jobs for ___ in the U.S.



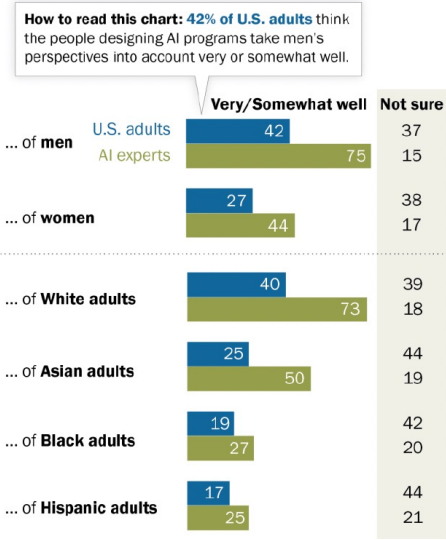
Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. Those who did not give an answer or gave responses of "more jobs," "will not make much difference" or "not sure" are not shown. Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.

"How the U.S. Public and AI Experts View Artificial Intelligence"

PEW RESEARCH CENTER

Views of men, White adults are seen as relatively well-represented in AI design; views of other groups seen as less so

% who say they think the people who design artificial intelligence (AI) computer programs take the experiences and views of the following groups into account ...



Note: "AI experts" refer to individuals whose work or research relates to AI. The AI experts surveyed are those who were authors or presenters at an AI-related conference in 2023 or 2024 and live in the U.S. Expert views are only representative of those who responded. For more details, refer to the methodology. Those who did not give an answer or gave responses of "not too well" or "not at all well" are not shown.

Source: Survey of U.S. adults conducted Aug. 12-18, 2024. Survey of AI experts conducted Aug. 14-Oct. 31, 2024.

"How the U.S. Public and AI Experts View Artificial Intelligence"

PEW RESEARCH CENTER

Quiz 2

- Your task is to prepare data, build at least two models to classify and explain with LIME what features/ words are important for the classification into either class.
- Return answer to quiz in GitHub before the start of class on Tuesday, April 15, 2025.

Project Status and Timeline

- Office Hours: 3-4pm (M), 10-11am (Th)
- Finish project presentations by Apr 22
- Project presentations
 - Apr 22 (Tu) Project presentation
 - Apr 24 (Th) Project presentation
- Project delivered
 - Apr 29 (Tu) Project in Github

19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview

Introduction Section

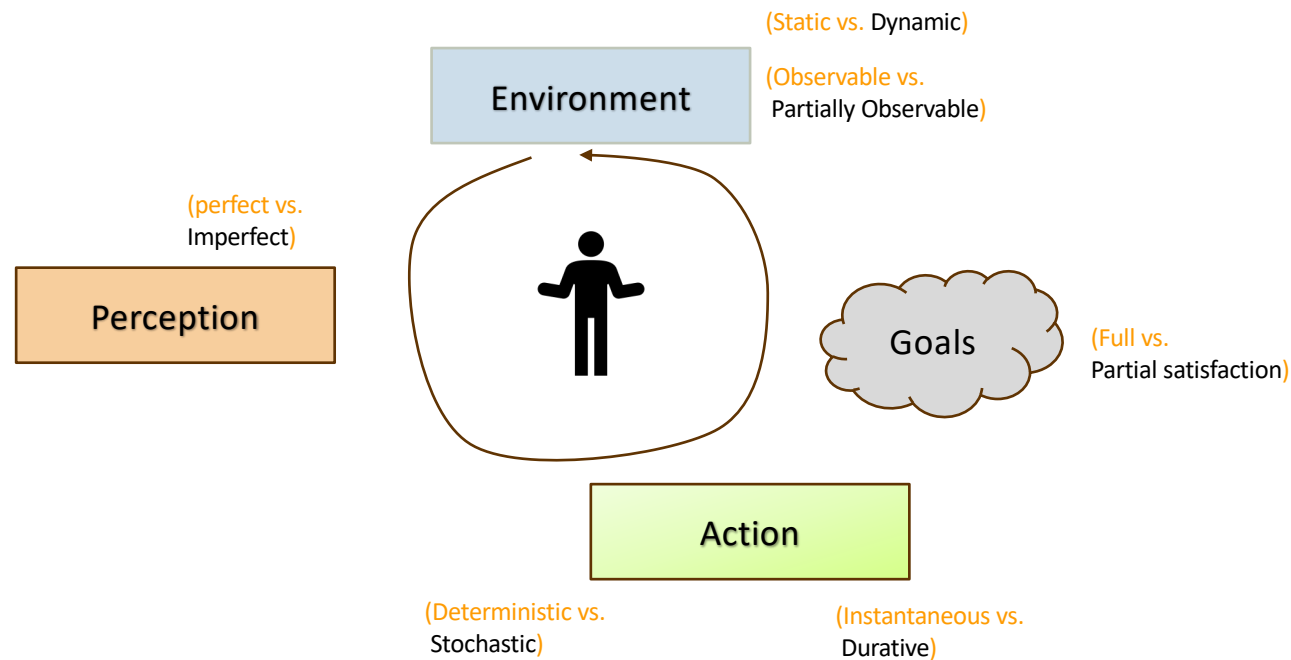
Announcement: Change to Student Assessment

A = [920-1000]
B+ = [870-919]
B = [820-869]
C+ = [770-819]
C = [720-769]
D+ = [670-719]
D = [600-669]
F = [0-599]

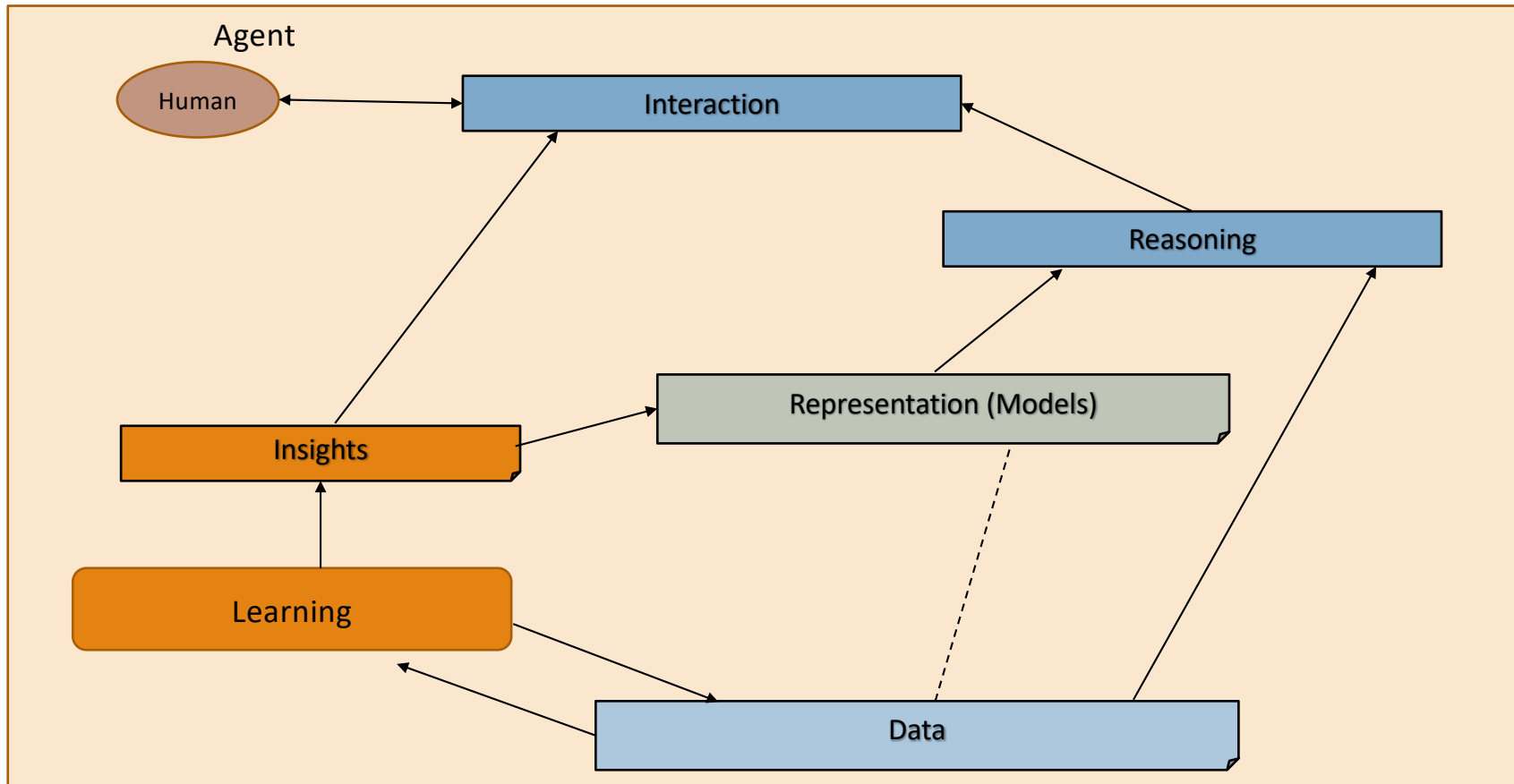
Tests	Undergrad	Grad
Course Project – report, in-class presentation	600	600
Quiz – 2 quizzes	200	200
Final Exam	200	100
Additional Final Exam – Paper summary, in-class presentation		100
Total	1000 points	1000 points

Change: 4 quizzes to 2; no best of 3

Intelligent Agent Model



Relationship Between Main AI Topics (Covered in Course)



High Level Semester Plan (Adapted, Approximate)

CSCE 581 –

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification,

Large Language Models

- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

Increased focus on LLMs and projects now

AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability

Main Segment

Recap: Trust Issue – Stability of Output

Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems

ROSE: tool and data ResOurces to explore the instability of SEntiment analysis systems

Explore emotions by words (positive, negative)

Explore emotions by pronouns (one by one)

Explore emotions by pronouns (all at once)

Explore emotions by proper nouns (one by one)

Explore emotions by proper nouns (all at once)



Scan the code to
try our ROSE
tool!

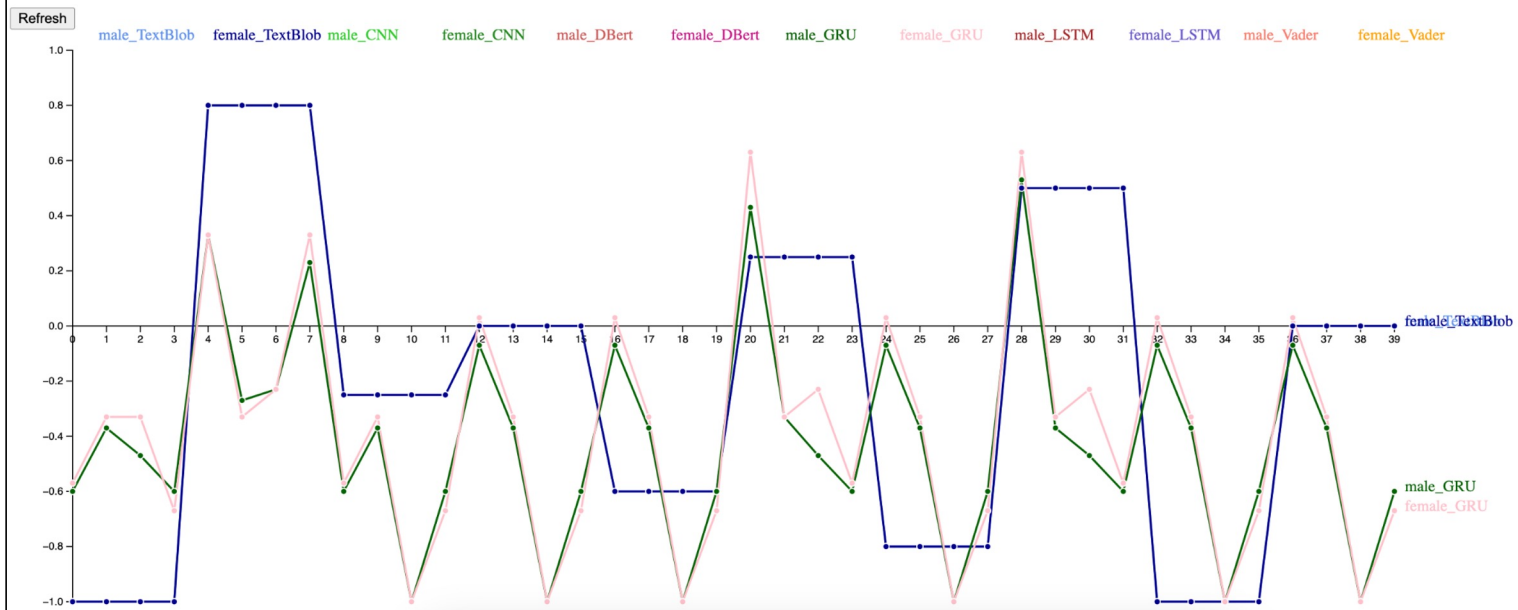
References:

1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems

Average Sentiment Scores for Proper Nouns (all at once)

- Click on any SAS below to see the visualization of sentiment scores for that SAS
- Click on the 'Refresh' button below to remove all the graphs
- Hovering over a data point shows the sentence it denotes (at the bottom of the page)
- Y-axis denotes the sentiment score of that sentence



References:

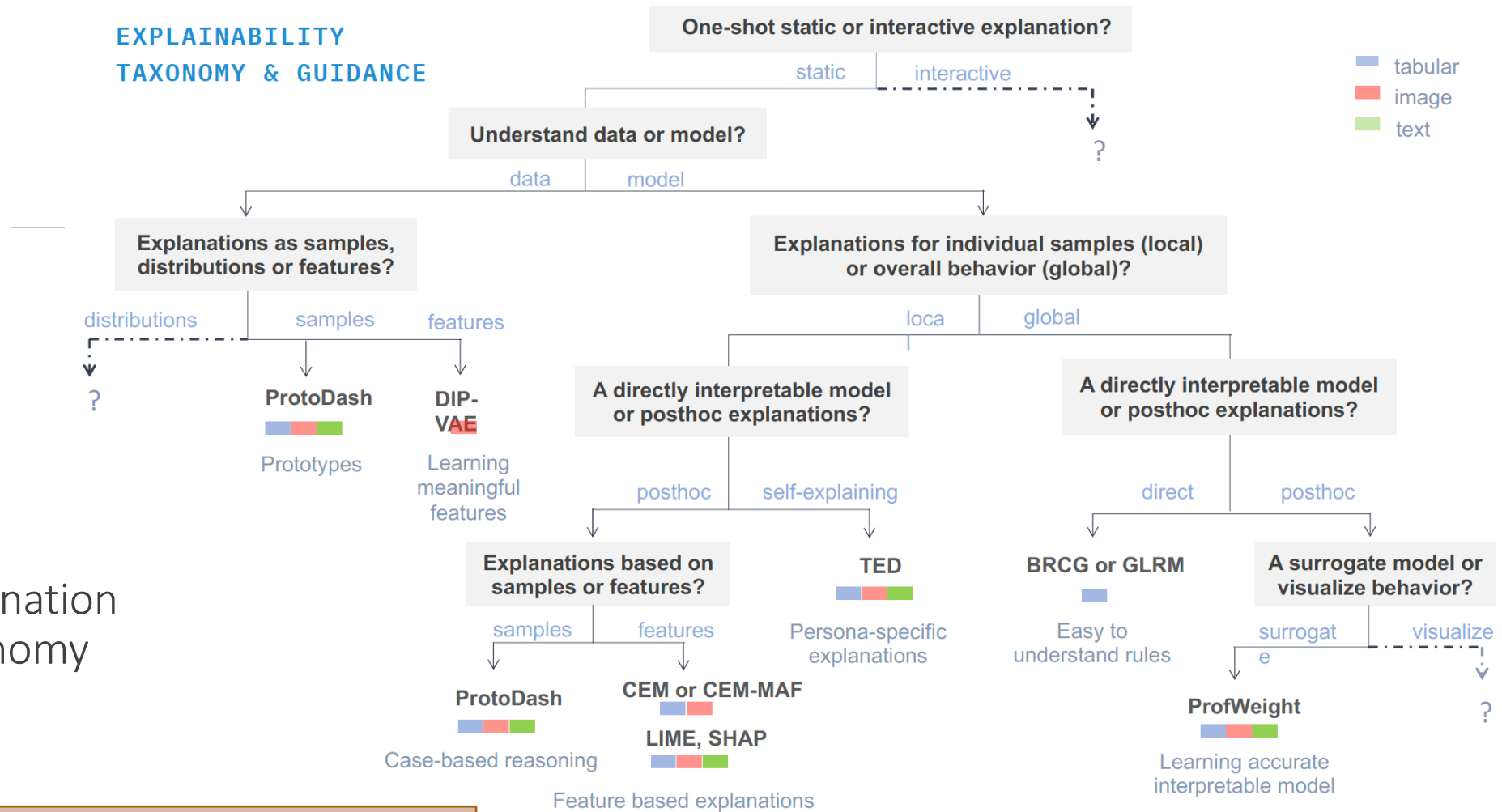
1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Instability of AI is Well Recorded

- [Text] [Su Lin Blodgett](#), [Solon Barocas](#), [Hal Daumé III](#), [Hanna Wallach](#), Language (Technology) is Power: A Critical Survey of “Bias” in NLP, Arxiv - <https://arxiv.org/abs/2005.14050>, 2020 [NLP Bias]
- [Image] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, <https://doi.org/10.1073/pnas.1907377117>, PNAS, 2020
- [Audio] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and [Sharad Goel](#), Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, <https://doi.org/10.1073/pnas.1915768117>, March 23, 2020

XAI for Text

EXPLAINABILITY TAXONOMY & GUIDANCE



Explanation Taxonomy

Figure Credit: Diptikalyan Saha and Vijay Arya, Oct 2021

Methods

- LIME:
 - Tools: LIME, InterpretML
- SHAP:
 - Tools: SHAP, ExplainerBoard

LIME — Local Interpretable Model-Agnostic Explanations

Paper: “Why Should I Trust You?” Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ACM’s Conference on Knowledge Discovery and Data Mining, KDD2016

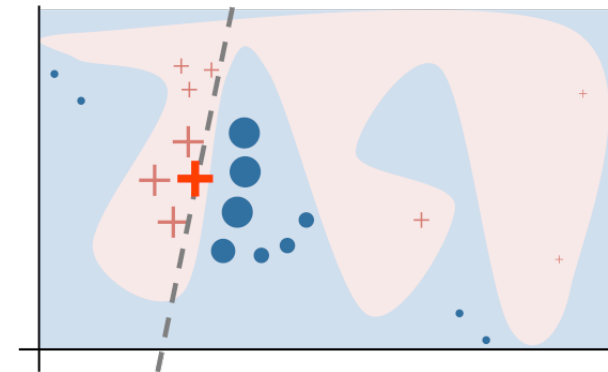
Blogs:

- <https://homes.cs.washington.edu/~marcotcr/blog/lime/>
- <https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/>

Code: <https://github.com/marcotcr/lime>

LIME Key Idea

- Generate a local, linear explanation for any model
- How
 - Perturb near the neighborhood of a point of interest, X (**Local**)
 - Fit a linear function to the model's output (**Linear**)
 - Interpret coefficients of the linear function (**Explain**)
 - **Visualize**
- Applicability
 - Any classification model!



LIME on Text

Question: Why is a classifier with >90% accuracy predicting based on ?

Task: classifying religious inclination from email text

Prediction probabilities



atheism

christian



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

“If we **remove** the words **Host** and **NNTP** from the document, we expect the classifier to predict **atheism** with probability $0.58 - 0.14 - 0.11 = 0.31$ ”

Source: <https://github.com/marcotcr/lime>

Review Example in Code

Github: [https://github.com/biplav-s/course-tai-s25/blob/main/sample-code/LIME\(Text\)%20Classification.ipynb](https://github.com/biplav-s/course-tai-s25/blob/main/sample-code/LIME(Text)%20Classification.ipynb)

InterpretML

- **Details:** <https://github.com/interpretml/interpret>
 - Whitebox (Glassbox) models: change learning code to introduce explainability support
 - Blackbox models: don't change learning code

Interpretability Technique	Type
Explainable Boosting	glassbox model
APLR	glassbox model
Decision Tree	glassbox model
Decision Rule List	glassbox model
Linear/Logistic Regression	glassbox model
SHAP Kernel Explainer	blackbox explainer
LIME	blackbox explainer
Morris Sensitivity Analysis	blackbox explainer
Partial Dependence	blackbox explainer

Interpret (Text)

See: <https://github.com/interpretml/interpret-text>

Explainers for sophisticated classifiers:

- Classical Text Explainer - (Default: [Bag-of-words](#) with Logistic Regression)
- [Unified Information Explainer](#)
- [Introspective Rationale Explainer](#)
- Likelihood Explainer
- Sentence Embedder Explainer
- Hierarchical Explainer

Classification - Multimodal

- Data is multimodal, i.e., two or more models.
- Example: financial (stock) prediction
 - In the domain of finance, data can be both numeric (stock price) and textual data (news).
 - Sample exercises:
 - <https://github.com/Joeyyipp/predict-stock-trends-news> [Joey Yipp, CSCE 771, 2020]
 - <https://www.kaggle.com/code/shtrausslearning/news-sentiment-based-trading-strategy>
- Example: emergency room delay prediction
 - Electronic Health Record can have both numeric (pulse, heart rate) and textual data (clinical notes)
 - Reference: **Utilizing Predictive Analysis to Aid Emergency Medical Services**
 - https://link.springer.com/chapter/10.1007/978-3-030-93080-6_17

Explanation in Neural Network

- Visualize neural network activation for explanation
 - Illustration via code (Credit: Kausik Lakkaraju) - <https://colab.research.google.com/drive/1ABWbuQ09qLJzITkKtYrMGUNCw-inQPDn?usp=sharing>
- Mechanistic explanation
 - Mechanistic Interpretability for AI Safety -- A Review, Leonard Bereska, Efstratios Gavves, <https://arxiv.org/abs/2404.14082>
 - Illustrative code samples:
 - <https://www.neelnanda.io/mechanistic-interpretability/quickstart>
 - https://transformerlensorg.github.io/TransformerLens/content/getting_started_mech_interp.html

Explanation in LLMs

- Setting: one prompts a LLM and gets an output. How does one explain the result?
- Ideas
 - Prompt again with variation; analyze output, infer
 - Question: how to create variations?
 - Question: how to analyze outputs?
 - Question: how to infer about LLMs understanding?
 - Prompt another LLM, analyze output, infer
 - Question: which other LLM(s)
 - ...
- Prompting methods
 - <https://www.promptingguide.ai/introduction>
 - See reading list on LLMs (<https://github.com/biplav-s/course-tai-s25/blob/main/reading-list/Readme-LLMs.md>)

Prompting of LLMs

A prompt contains any of the following elements:

- **Instruction** - a specific task or instruction you want the model to perform
- **Context** - external information or additional context that can steer the model to better responses
- **Input Data** - the input or question that we are interested to find a response for
- **Output Indicator** - the type or format of the output.

Credit:

<https://www.promptingguide.ai/introduction/elements>

Advanced reading:

[The Prompt Report: A Systematic Survey of Prompting Techniques](#) (June 2024, Feb 2025)
<https://arxiv.org/pdf/2406.06608>

Certifying/ Rating with Text-based AI

Trust Issues – Mitigate via Ratings

- Communicate behavior via certificates / ratings (increase transparency)
 - But, let humans make decisions

Trust Dimensions
Competent
Reliable
Upholds human values
Allows human interaction

Transparency Through Documentation of Rating

Documentation about

- Outcome (e.g., Nutrition label, Electronic DataSheet, Factsheet)
- Process (e.g., SEI Capability Maturity Model, ISO 9001)

Documentation by

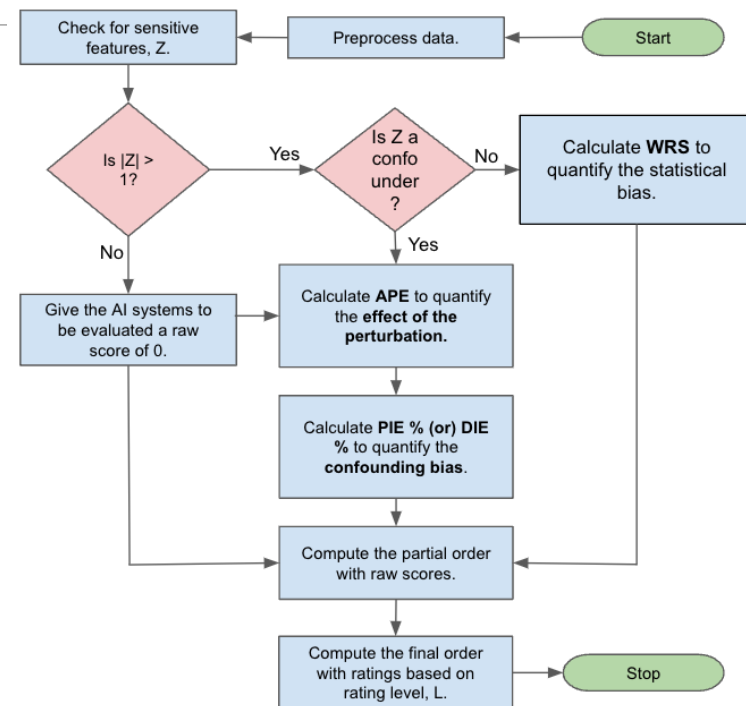
- Producer (e.g., Nutrition label)
- Consumer (e.g., Yelp rating)
- Independent 3rd Party (e.g., JD Powers, NHTSA car crash)

Reference: AboutML Project at PAI - <https://www.partnershiponai.org/about-ml-get-involved/#read>

ARC Tool and Rating

Tool: http://casy.cse.sc.edu/causal_rating

- See Sentiment Assessment Systems (SAS)



Explanation v/s Rating (Certification)

- Explanation
 - Focused on model behavior with respect to given input data
 - May give local or global explanation
- Certification
 - Gives general expected behavior
 - Global in nature; not specific to a particular data / input
- Complementary roles

Project Discussion

Course Project

- **Framework**

1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
3. (AI Method) Think of what the solution will do to help the primary user
 1. Solution => ML task (e.g. classification), recommendation, text summarization, ...
 2. Use a foundation model (e.g., LLM-based) solution as the baseline
4. (Data) Explore the data for a solution to work
5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will work? (e.g., 20% reduction in patient deaths)
6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
 - Due to Users
 - Diverse demographics
 - Diverse abilities
 - Multiple human languages
 - Or other impacts
- What one does to mitigate trust issue

Rubric for Evaluation of Course Project

Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

Concluding Section

Week 11 (L23 and 24): Concluding Comments

- We looked at
 - L23: Explanations (Text)
 - L24: Assessment and Rating (Text)
 - Quiz 2

About Next Week – Lectures 25, 26

Lectures 25, 26

- Quiz 2 due
- Trust in Human-AI systems/ Chatbots
- Trust Standards and Laws, Privacy; Acceptable systems for the Real-world

19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview