DSC 291/190: Trustworthy Machine Learning (SP'25)

In this assignment, we ask you to rehearse your presentation with another group (we call it the Partner group in this doc) and have the opportunity to improve your presentation skills based on the feedback from your fellow students.

Please submit the feedback form **on Gradescope** by the deadline.

Instructions

- Below are some evaluation criteria on the presentation:
    a. Slides presentation
        - Does it contain all the important slides, if not, what is missing?
        - Clarity of the presentation
    b. Contents
        - Does it cover the important information (see the sections we outlined in the Project Rubric), if not, what is missing?
        - Does it summarize the key ideas and results clearly?
        - Does it compare the methods properly?
        - Does it discuss the strengths and weaknesses of the paper?
        - Does it point out the limitations and challenges of the methods?
        - Is it clear and coherent?
    c. Time control
        - Is it on-time (~15 min)
- Please summarize the partner group's presentation in a short paragraph, and list what are the good things and what are the things that can be improved.
    a. Please follow the above (a)-(c) evaluation criteria and answer the questions
    b. Apart from evaluation (a)-(c), please summarize your comments and give at least 3 good things and 3 to-be-improved things for the presentation
    c. Also come up with discussion questions for each other's presentations based on the papers that are being presented

DSC 291/190 Rehearsal Feedback Form

Group number: #3
Group topic: Jailbreaking LLMs in 20 Queries
Group members: Zihao Sun, Ashley Chu, Gallant Tsao
Feedback from the Partner Group: #4

- Date: 2025.6.2
- Partner Group members: Pooja Pun, Calvin Nguyen, Joyce Lu, Parna Praveen

Please summarize the feedback you get from the Partner group:

- Evaluation criteria
  - (a) Contains all important slides, except results. Presentation is clear and coherent.
  - (b) Does it cover the important information (see the sections we outlined in the Project Rubric), if not, what is missing? Missing results
  - (c) Does it summarize the key ideas and results clearly? Yes key ideas, no results from their own side
  - Does it compare the methods properly? Yes (vs. JBC, GCG)
  - Does it discuss the strengths and weaknesses of the paper? Yes, cons: dependency issues
  - Does it point out the limitations and challenges of the methods? Yes, limitations for PAIR
  - Is it clear and coherent? Yes
- Good things
  - Likes the visual throughout each slide to illustrate methods
  - Like the explanations of results from paper
  - Like that you illustrate how the model works
- To-be-improved things
  - Too many words on just the intro slide
  - Too much time spent on related works / methods
  - Reproduction?
- Discussion questions for your presentation
  - What makes Vicuna so weak while Claude is very strong against jailbreaking?

Please summarize the feedback you give to the Partner group:

- Below are some evaluation criteria on the presentation:
  - Slides presentation
    - Does it contain all the important slides, if not, what is missing? Yes
    - Presentation was clear
  - Contents
    - Covers important info? Yes
    - Summarizes well? Yes
    - Does it compare the methods properly? Yes
    - Does it discuss the strengths and weaknesses of the paper? Don't remember
    - Does it point out the limitations and challenges of the methods? Yes
    - Is it clear and coherent? Yes
  - Time control

        - Is it on-time (12 min): A bit overtime, ended roughly around the 20 minute mark

Good things:

- Clear explanation in general
- Visualization choices were clear
- Formulas were clear, I liked the time charts (run time)
- Well-explained results from paper, cool to see how efficiently performance was across different datasets
- Detailed description on how each models works and provide accurate information on numeric results
- Interesting extension of experiment and good explanation of why it was included

TBI

- Some time allocation could be improved because some slides were spent too much time on
- Could put more visualizations?
- Some slides have a lot of text
- Try to cut length of explanations in order to shorten length of presentation
- Maybe add some visuals from the paper to explain the results as well as the methods

Discussion Questions

- Are there any other improvements that are being made for future works in terms of improving interpretability?
    - Why do you think some datasets took longer to run than others?

Feedback from the Partner Group: #5

- Date: 2025.6.2
- Partner Group members: Seiji Yang, Vikram Venkatesh, Rushil Chandrupatla, Arjun Varshney

Please summarize the feedback you get from the Partner group:

- What is the presentation about (a short paragraph):
    - Jailbreaking LLMs in 20 Queries - Jailbreaking allows LLMs to be unsafe and/or unethical especially in certain fields ie. medical field, politics, etc. Different types of jailbreaks, prompt vs token level.
- Evaluation criteria
    - (a) Organized slides, good amount of content on each slides, could use some more visualizations
    - (b) Follows rubric guidelines well, summarizes key ideas
    - (c) Not on time, > 20 mins
- Good things

    - Insightful and detailed explanations, using different examples to understand the consequences of jailbreaking and the computational difficulties around solving it
    - I really liked how longer explanations had a few word summary in bold first
    - Good use of images to explain processes
- To-be-improved things
    - Technical ideas can be explained more clearly
    - Many slides contain complete and several sentences, these should be transformed to bullets for audience interpretability
    - Unfamiliar terms should be defined before mentioned
- Discussion questions for your presentation
    - Can PAIR be combined with the other methods discussed to further reduce costs or expose new gaps in robustness?
    - Based on what you have learned, do you think that jailbreak/attack methods or prevention methods will become significantly better than the other in the upcoming years?

Please summarize the feedback you give to the Partner group:

- What is the presentation about (a short paragraph)
    - Label-free concept bottleneck models can make collecting labeled data less time-consuming and labor-intensive.
        - interpretable, linear combination of concepts, easy to validate model predictions
        - scalable
        - maintains accuracy
        - training automated
        - cons: dependent on LLMs, training can be resource-intensive
        - methods: generate concept using GPT 3, compute concept matrix, compute weight activations by max similarity, then train
- Below are some evaluation criteria on the presentation:
    - Slides presentation
        - Does it contain all the important slides? Yes
        - Presentation was clear
    - Contents
        - Covers important info? Yes
        - Summarizes well? Yes
        - Does it compare the methods properly? Yes
        - Does it discuss the strengths and weaknesses of the paper? Don't remember
        - Does it point out the limitations and challenges of the methods? Yes
        - Is it clear and coherent? Yes
    - Time control
        - Is it on-time (12 min): Slightly over at 15min40sec

Good things:

- Clear explanation in general
- Ample amount of examples with images
- Good visualization techniques (sankey diagram)
- Evaluation criteria
    - (a) fits the evaluation criteria from the rubric besides time limit
    - The presentation is pretty clear and shows all details.
- Good things
    - Diagrams in the results and pictures aided understanding
    - Results were clearly explained
- To-be-improved things
    - Over time limit (19ish minutes)
    - Maybe add graphics during introduction to aid explanation
- Discussion questions for partner group's presentation
    - What other methods exist for automatically labeling data?