SAEfarer User Study Protocol

Introduction

- Slides

To begin the study, I will share my screen and use this deck of slides to introduce the participant to the study and provide an overview of sparse autoencoders.

Once I start the recording, I will have the participant affirm that they agree to being recorded and that they have received and reviewed the Participant Information Sheet.

Tutorial

- Slides
- Google Colab notebook

Next, I will give the participant a tutorial on how to use SAEfarer. The tutorial will cover the UI, visualizations, and feature rankings. The tutorial will mostly be done in the slides. Then I will move to a Google Colab notebook to briefly demonstrate the tool in action.

Practice

- Google Colab notebook

Next, I will have them share their screen and run SAEfarer in the notebook. Once the notebook is loaded, I will verify that the participant knows how to use SAEfarer.I will ask them to answer the following questions or perform the following tasks using SAEfarer:

Overview:

- How many data points were used to analyze this model?
  - Answer: 78.1k
- What's the model's error rate?
  - 29.24%
- How many features are there in the SAE?
  - 3072
- How often do the most features activate?
  - 1% of instances
- What's the most common error made by the model?
  - True label Neutral, predicted label Bullish

- What percentage of the time does the model make that error?
    - 13.61%
- How many instances did the model make this mistake on?
    - 10,629

Feature Table:

- Can you explain how the features are currently sorted?
    - By error rate.
- If you need help understanding how the features are currently sorted, where can you get an explanation?
    - The question mark icon.
- How are the features currently filtered?
- What's the error rate of the top feature?
    - 44.77%
- What percent of instances cause this feature to activate?
    - 0.869%
- For instances that cause this feature to activate, what's the model's top predicted class?
    - Bearish
- Can you find the feature with the highest activation rate?
    - 2100
- What does this feature represent?
    - End of sequence token
- What feature has the highest correlation with the model predicting the Bullish class?
    - 2772
- What does this feature represent?
    - Repeated emojis, especially the line chart increasing emoji 📈
- Can you identify a feature on whose activating instances the model often incorrectly predicts the Bearish class?
    - 1895
- Can you find the feature that when it activates, the model has the highest overall error rate?
    - 2085

Feature Detail:

- How many instances activate this feature?
    - 679
- Can you describe the errors that the model makes on these instances?
    - Mostly predicting bearish when the true label is neutral.
- How do the model's predictions on these instances compare to its predictions on the whole dataset?
    - It predicts bearish more often and predicts bullish less often.
- Can you describe the distribution of activation values for this feature?

    - Left skewed. One peak around 3. Another smaller one around 5.
- Can you describe the model's predicted probabilities across the range of activation values?
    - Bearish probability increases with activation value. The opposite is true for Bullish.
- Can you describe the concept that this feature represents?
    - Saying that you will buy once it hits a certain price.
- Can you come up with your own sentence that will activate this feature?
- Can you come up with a sentence that will not activate this feature?

After asking these questions, I will ask the participant if they have any other questions about how to use SAEfarer. I will then give them a few minutes to freely use the tool to explore the current model. After these few minutes, I will again ask if they have any questions.

Model Exploration

- Google Colab notebook

I will provide the participants with the following prompt:

> We have a model that classifies news articles as either "World", "Sports", "Business", or "Sci/Tech". We have trained a sparse autoencoder on this model to identify the concepts that it has learned.
>
> You have 30 minutes to explore the behavior of this model using SAEfarer. You should focus on analyzing the relationship between the SAE features and the model's predictions and errors.
>
> Here are some example questions that you may want to consider:
>
> - What features correlate with the model predicting the Sports label?
> - What features correlate with the model making mistakes?
> - Are there any surprising or unintuitive features?
>
> Please think aloud while you work so that we can understand your process, questions, and insights.
>
> You may ask questions about how to use SAEfarer.

Questionnaire

- Duration: ~5 minutes
- Google Form

I will have the participant complete the System Usability Scale questionnaire.

Interview

- Duration: ~10 minutes
- Slides

This will be a semi-structured interview that is guided by the questions in the slides.