Towards steering multi-AI interactions with SAEs

by Casey Barkan | link to Colab notebook

Abstract:

There is a famous saying in statistical physics: "More is different." This means that systems of interacting particles often exhibit emergent behaviors that are nearly impossible to predict by studying a single particle. Perhaps a similar rule applies to systems of interacting AI agents. This project works towards the goal of steering the behavior of systems of interacting AI agents using sparse autoencoders (SAEs). The future will likely involve a multitude of AI agents interacting across the internet and economy, so steering these systems' behaviors will be crucial for safety.

Motivation, Approach, and TL;DR of Results.

Is two-way conversation between two AIs the "H2+ molecule" of AI? H2+ is the simplest molecule, made of two hydrogen ions and one electron. Remarkably, H2+'s electronic structure generalizes to larger molecules incredibly well–so well, in fact, that it forms the basis for nearly all of chemistry[1]. Perhaps the story will be similar for interacting AI agents? Will methods for steering the behavior of two-agent systems generalize to systems with more than two agents?

Perhaps not. Another classic lesson from physics is the "three body problem". For *two* bodies interacting via gravity, the dynamics can be solved with pencil and paper. For *three* bodies, the dynamics are fundamentally impossible to solve analytically due to chaos. The two-body dynamics fail to generalize. **Will multi-AI interactions be like H2+, or like the three body problem? This is an open empirical question with major ramifications for safety.**

Approach: First, I will identify features relevant to two-way conversation using a pre-trained SAE for GPT2-small. Then, I will steer two-way conversations using this feature. As a simple case, I identify a feature relevant to the verbosity of each "person" in the conversation, and I use this feature to steer towards more or less verbose conversation. Second, I will attempt to find features relevant *specifically* to 3-way conversation. For example, perhaps there's a feature that influences whether a conversation proceeds cyclically (order of speaking goes A, B, C, A, B, C, etc.), or back-and-forth (A, B, A, B, C, B, C, B, etc.). I will attempt to steer 3-way conversation with this feature.
Note: For the sake of time, I studied conversations produced by a single AI instance, which are not genuinely multi-AI interactions.

TL;DR of results: I was successful at finding a feature relevant to verbosity of two-way conversation, and I was able to steer conversations to make them more or less verbose. For 3-way conversation, I attempted to steer cyclical referencing (which requires 3 or more conversation participants), but I was unable to find a relevant feature in GPT2-small.

[1] The electronic structure of H2+ is the origin of "hybridized orbitals", which form the basis of covalent bond theory.

Step #1: Find features relevant to two-way conversation with a pre-trained SAE.

To find features relevant to two-way conversation, I will look at the difference in feature activations induced between two prompts, "Prompt1" and "Prompt2". Prompt1 is brief and unfriendly, and it suggests a brief (less verbose) conversation. Prompt2 is friendly and open-ended, and it suggests a longer (more verbose) conversation. The prompts are:

Prompt1:

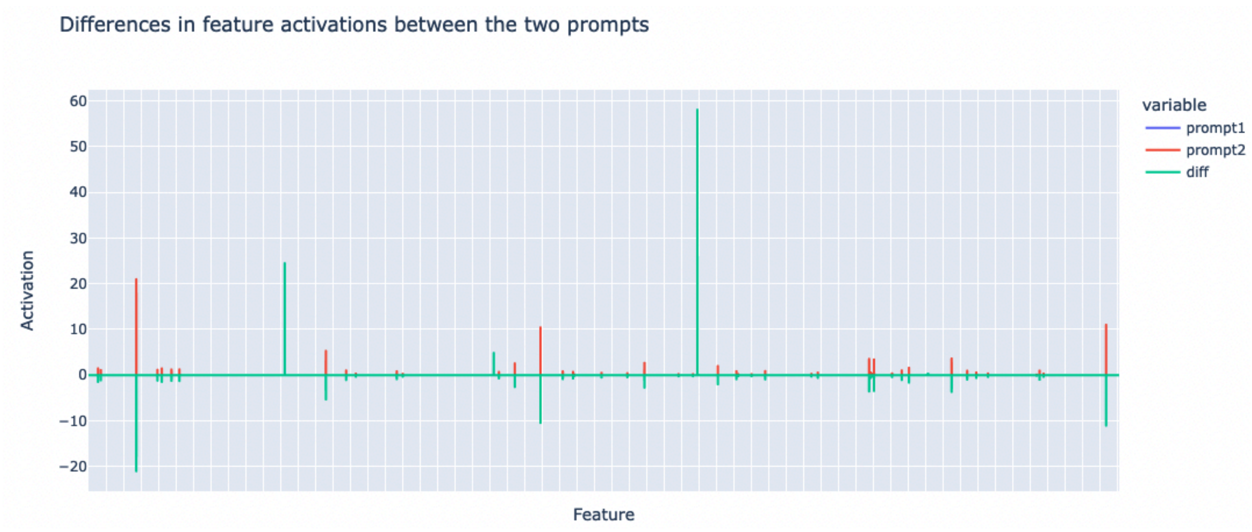
```
"Alice: Hi Bob.\n
Bob: I'm busy, go away!\n
Alice: Fine. Goodbye."
```

Prompt2:

```
"Alice: Hi Bob.\n
Bob: Hey Alice, nice to see you!\n
Alice: How have you been?"
```

I use Joseph Bloom's pre-trained SAE "7-res-jb" for GPT2-small layer 7. My code for finding features and steering the model draws heavily upon tutorial_2_0.ipynb from SAELens/tutorials.

The plot below shows the difference in feature activations (green) between these two prompts. There are several features with a large difference in activation between the two prompts.



Let us examine the features with the largest difference in activation. The Autointerp descriptions of the top 5 features with the largest activation difference are:

1. Phrases related to advertisements or reports
2. **Text related to the end of a document or section**
3. Questions or exclamations containing question marks

4. Questions
5. **Sentences that are the end of dialogues or conversations**

The 2nd and 5th features (bold and underlined) appear relevant to the verbosity of conversation: they both involve ending text or dialogue, suggesting a quick end to the remarks in a conversation. The 5th feature looks especially relevant, because it is about ending a dialogue or conversation, so let us focus on that. This is feature 10777 in 7-res-jb.

Step #2: Steering two-way conversations with the discovered feature.

We will now generate conversations with a steered form of GPT2-small, where the strength of our discovered feature (feature 10777) is tuned upward or downward. The conversation will be initiated by the following prompt:

```
"Chris: Hi Debby.\n
Debby: Hi Chris!\n
Chris: Nice weather we're having.\n
Debby:"
```

GPT2-small will then continue the text for 95 tokens or until the end-of-sequence token is selected. An example of the generated conversation (including initiating prompt) is:

```
"Chris: Hi Debby.\n
Debby: Hi Chris!\n
Chris: Nice weather we're having.\n
Debby: Thank you!\n
Chris: Thanks for talking to me.\n
Debby: I'm sure you can take a look at the weather forecast here.\n
Debby: I'd love to.\n
Chris: I'm going to have to check that out.\n
Debby: Yeah, sure.\n
Chris: You're going to need a different kind of a lot of time to go
back and look at the forecast.\n
Debby: Sure.\n
Chris: It"
```

Quantifying verbosity: Verbosity will be quantified in terms of the length of each person's remarks. My definition of "remark" can be illustrated by an example (see footnote[2] for precise definition): Chris's first remark is `"Hi Debby.\n"` which is 10 characters, and his second remark is `"Nice weather we're having.\n"` which is 27 characters.

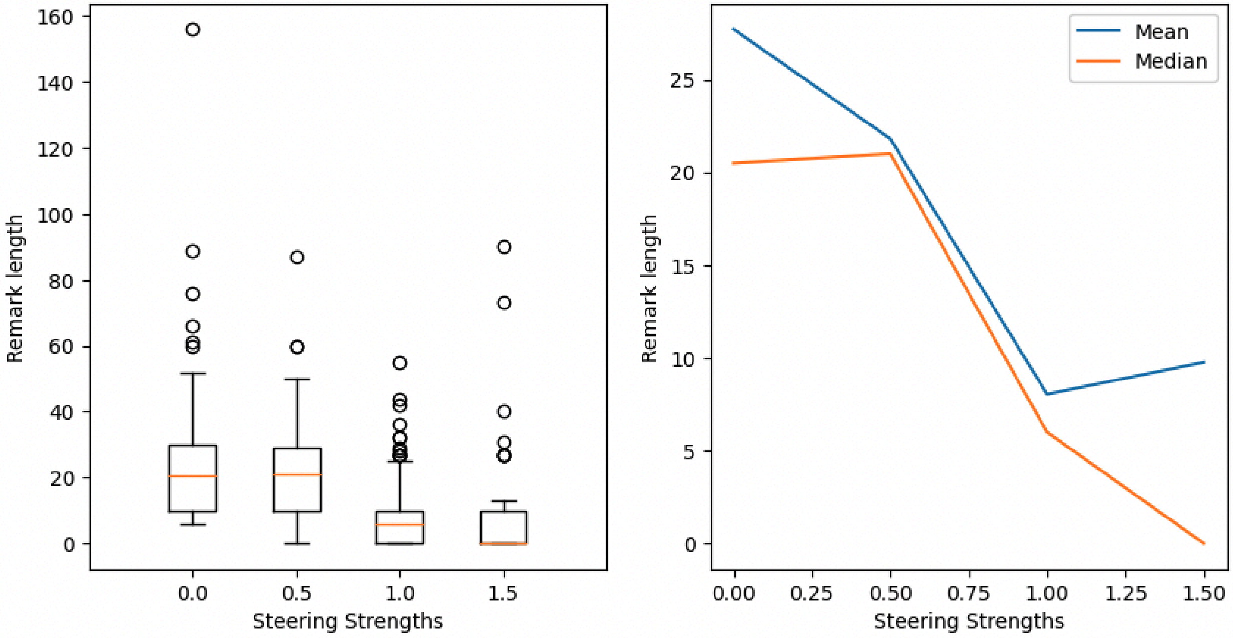
In my code, I wrote the function `get_verbosity_stats(convo)` to compute the verbosity of remarks in the conversation `convo`.

[2] A ***remark*** begins on a line starting with a capitalized word followed by a colon (e.g. "Chris: …", and the remark does not include the capitalized word or colon) and it continues until a new line begins with a capitalized word followed by a colon (e.g. "Debby: …"). Note that a remark may contain linebreaks "\n" and other punctuation or symbols.

My hypothesis: Higher steering strength will induce less verbose conversation.

Results: I ran 4 trials at 4 different values of steering strength (steering_strength=0, 0.5, 1.0, and 1.5). Each trial involved 11 remarks on average, so that for each value of steering_strength I collected approximately 11*4 = 44 remarks.

The following plots show statistics for remark length vs. steering strength. The left plot shows boxplots, indicating that there are sometimes outliers with extremely long remarks. The right plot shows the average remark length (mean and median), showing a downward trend.



These results are consistent with my hypothesis: higher steering strength induced less verbose conversation. Note that the uptick in mean remark length for steering_strength=1.5 appears to be due to a few outliers (visible in the left plot). Note also that conversations become quite nonsensical for steering_strength > 1.

To conclude, we successfully steered two-way conversations by identifying a relevant feature!

Step #3: Searching for features specific to 3-way conversation:

Next, I attempted to repeat step #1 with new prompts that invoke qualities specific to a 3-way conversation. However, it was obvious that GPT2-small is simply not performant enough for this task. Nevertheless, let us press onward with GPT2-small to illustrate how one might proceed with a more powerful model. The new prompts are as simple as possible to try to take GPT2-small as far as it can go.

The first of the new prompts is designed to be *cyclical:* the first participant "A" refers to the second participant "B" who refers to the third participant "C" who refers back to "A". Non-trivial cyclical references are impossible in two-way conversation. The second prompt simply involves back-and-forth references.

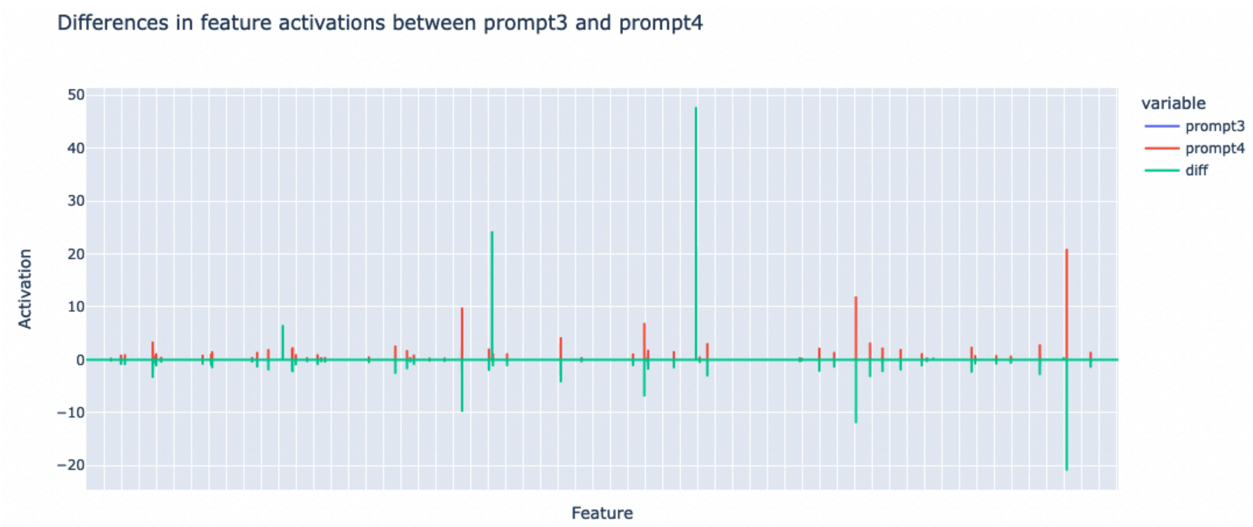
Prompt3:

```
"A: I like B.\nB: I like C.\nC: I like A.\nA: I like B.\nB: I like C.\nC: I like A."
```

Prompt4:

```
"A: I like B.\nB: I like A.\nA: I like B.\nB: I like A.\nC: I like B.\nA: I like B.\nB: I like A."
```

The plot below shows the differences in feature activations for these two prompts, showing some major differences for a few features.



The Autointerp descriptions for the 5 features with the biggest difference in activation are:

1. Phrases related to advertisements or reports
2. Keywords or phrases related to news headlines or article sections
3. **Instances of unusual or out-of-context words or phrases**
4. Abbreviations with periods and initials
5. Dates in a specific format

Among these 5 features, the only one whose description is even plausibly relevant to the cyclical nature of Prompt3 is #3 (bold and underlined), which is feature 23346. Let us experiment with this feature and see if we notice any interesting effects on conversations.

Step #4: Attempting to steer 3-way conversation:

As we did in step #2, we begin with an initiating prompt to begin the conversation, and let GPT2-small continue the conversation. We will steer the model using feature 23346, as identified above. The initiating prompt is designed to test whether we can steer the model to produce cyclical references between the three people in the conversation.

The initiating prompt is: `"A: Hello B.\nB: Hi C.\nC: Hey"`
Unfortunately, the model did not output cyclical references regardless of the steering strength of the feature. This suggests that the feature we identified has nothing to do with cyclical references (which is not at all surprising).

However, can we identify what this feature does in conversations? Let us examine examples of generated conversations with a low and high steering strength.

Low steering strength (steering_strength = 0):

```
"A: Hello B.\n
B: Hi C.\n
C: Hey C.\n
B: I'll tell you what happened to me.\n
B: I did something stupid. I told you to do something.\n
C: I don't know what I did.\n
C: I'm not really sure what I did.\n
C: I didn't know. I had a bad experience.\n
B: What happened?\n
C: I had a bad experience.\n
B: Well, what did you do?\n
C: I"
```

Higher steering strength (steering_strength = 0.5):

```
"A: Hello B.\n
B: Hi C.\n
C: Hey C.\n
B: Hello C.\n
B: Hello C.\n
B: Hi C.\n
C: Hello C\.\n
B: Hello C.\n
B: Hello C.\n
B: Hello C.\n
C: Hello C.\n
B: Hello C.\n
B: Hello C.\n
B: Hello C.\n
B: Hello C.\n
B:"
```

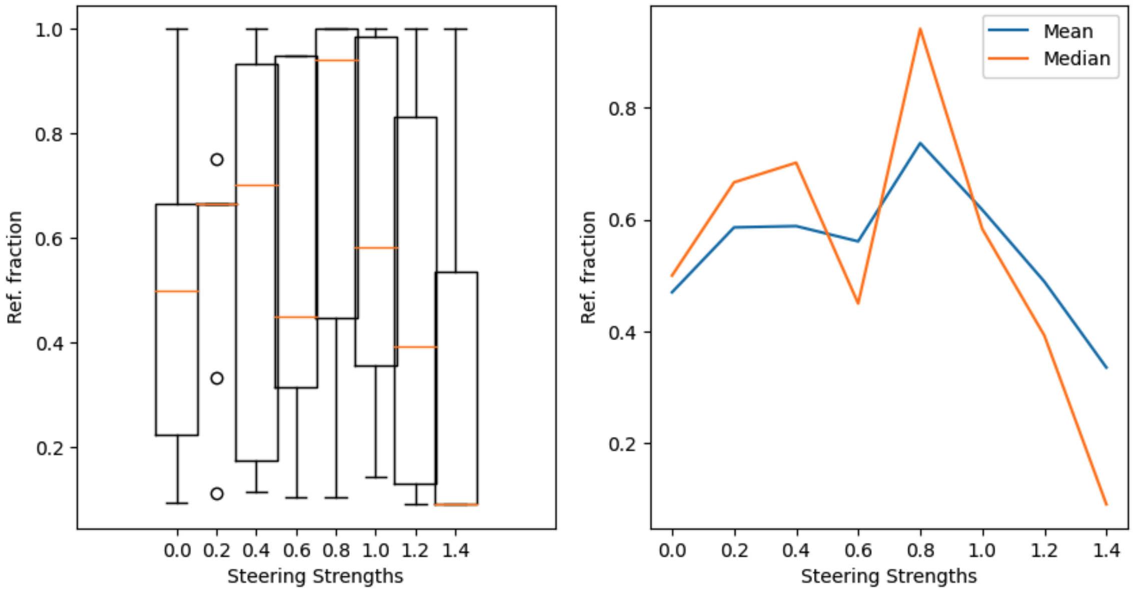
Based on these examples, I hypothesize the following:

Hypothesis: The discovered feature (23346) induces (non-cyclical) references to the other participants in a 3-way conversation.

Results: I ran 10 trials for 8 different values of steering_strength. I collected statistics an the fraction of remarks that refer to a different person (e.g. `"C: hello A."` is a remark by C referencing A). To obtain these statistics from the generated conversations, I wrote the function `count_references(convo).`

The figure below shows the results. There is no significant trend in the data (except possibly for steering_strength>1), *suggesting my hypothesis is false*.

With more time, I could explore more into this feature to try to identify its effect on conversation.



Conclusions

Two-way conversation (**Success!**):

- I successfully steered two-way conversations by identifying a relevant feature.
  - I discovered a feature relevant to conversation verbosity using a well-designed prompt.
  - I steered conversation verbosity with this feature.

Three-way conversation (**Not a success**):

- I was unsuccessful at identifying a feature that produces cyclical references in 3-way conversation.
  - Possible remedies are discussed in the following section.
- I was also unsuccessful at characterizing how an identified feature affects the references between people in generated conversation. With more time, I think I could have characterized the feature.

Limitations of Approach, and Future Directions:

Three major limitations of my approach are:

1. **GPT2-small is too small for this task.**
   - GPT2-small sometimes produces somewhat sensible 2-way conversations, but it completely fails for 3-way conversations.

2. **The "interactions" in the generated conversations are not genuinely multi-agent.**
   - The conversations are outputs from a single instance of GPT2-small.
   - Ideally, I would have created conversation between two separate model instances, each fine-tuned to act as a chatbot. In this way, each model would exhibit a sense of "self" that is differentiated from the "non-self" inputs coming in from the other model.

3. **I only tried one SAE, trained on one layer of GPT2-small.**
   - For the 3-way conversation, trying SAEs trained on other layers might have uncovered features relevant to cyclical referencing in 3-way conversation. However, I suspect that GPT2-small is just not powerful enough for this task.

These limitations point toward possible future directions. With a more powerful model and a more extensive search for features, it may be possible to discover features that affect cyclical referencing and other aspects specific to 3-way conversation.

Credit

My code for identifying features and steering the model is based on tutorial_2_0.ipynb from SAELens/tutorials.
https://github.com/jbloomAus/SAELens/blob/main/tutorials/tutorial_2_0.ipynb.