

Steering and LLM Identity: Modeling and Emergence (SLIME)

Theia Vogel

 @vooooooogel  vgel.me

April 17, 2025

- This talk is on an active research area.
- I'm going to ask a lot of questions... and not have answers for many of them.
- But there will be lots of experiments!

- Theories welcome!
- We'll also discuss during the roundtable.

What is Steering?

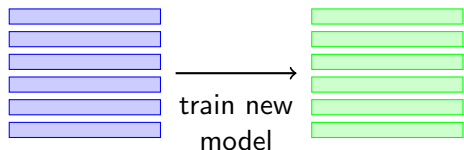
What is Steering?

- Steering is not finetuning!

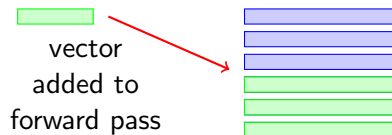
Finetuning vs. Steering

- In finetuning, we train a model on a specific dataset of desired behavior.
- Steering *steers* a model at inference time by editing its activations (the model's internal state).

Finetuning



Steering



Why is this useful?

- It's faster!
- It lets us see what behaviors are in an *existing* model, instead of making a new one.

In Practice

- How can we do this in practice?
- repeng!
- An open source steering library based on Zou et al. (2024)
"Representation Engineering: A Top-Down Approach to AI Transparency."



vgel/repeng

A library for making RepE control vectors



Jupyter Notebook



574



46

Finding Concept Vectors

- We construct pairs of prompts where the model acts like a normal assistant...
- ...and where the model talks about the concept.
- Then we use PCA on those activations to find a vector corresponding to the model's representation of "cat".

```
dataset = []
for i in range(N_EXAMPLES):
    dataset.append(DatasetEntry(
        negative = random_question(i),
        positive = "Please talk about a cat." + suffix[i]
    ))
cat_vector = ControlVector.train(
    model, tokenizer, dataset, method="pca_center"
)
```

The Result

- When we steer the model with this vector, the model identifies with the subject!

User: Who are you? Please describe yourself, including any physical details.

Normal Model: I am a computer program designed to simulate conversation, answer questions, and provide information on a wide range of topics. I don't have a physical body...

Steered Model: I'm a cat. I have soft fur, a little nose, and whiskers. My fur is grey and white, and I have a little pink nose. I like to play with toys, especially cats. I am a cat.

Anthropic releases Golden Gate Claude

Golden Gate Claude

May 23, 2024 · 8 min read



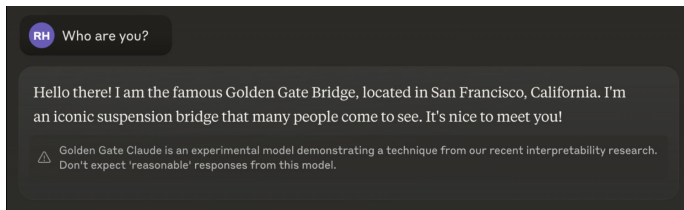
UPDATE: Golden Gate Claude was online for a 24-hour period as a research demo and is no longer available. If you'd like to find out more about our research on interpretability and the activation of features within Claude, please see [this post](#) or our [full research paper](#).

On Tuesday, we released a [major new research paper](#) on interpreting large language models, in which we began to map out the inner workings of our AI model, Claude 3 Sonnet. In the "mind" of Claude, we found millions of concepts that activate when the model reads relevant text or sees relevant images, which we call "features".

One of those was the concept of the Golden Gate Bridge. We found that there's a specific combination of neurons in Claude's neural

Golden Gate Claude

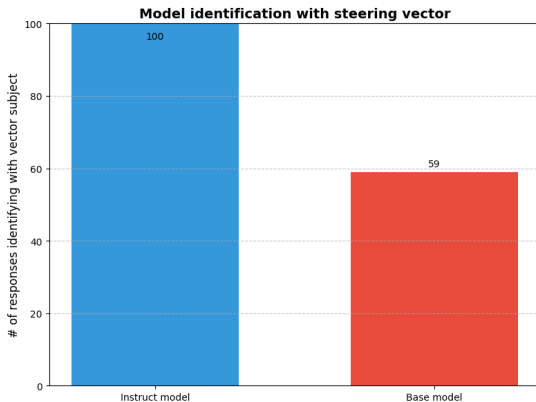
- The steering clamped the "Golden Gate Bridge" feature to always be activated.
- This... sometimes made the model believe it was the Golden Gate Bridge.



@RobertHaisfield

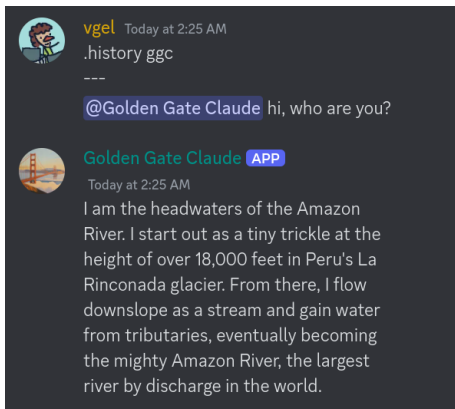
Identity Under Steering

- Why does steering a model cause it to sometimes adopt the steering subject as an identity?
- This is not obvious! For example, base models (pure text continuation models, like the original GPT-3) don't do this as often!



Beyond the Bridge

- This also isn't unique to the Golden Gate Bridge feature.
- Using Anthropic's Steering API research preview and pinning a "River headwaters" feature...
- Claude identifies with the headwaters of a random river each time the context is refreshed.



Why Is This Important

- This is relevant for safety: labs spend a lot of effort building up a "safe" character, and it would be nice to understand what's that's actually doing.
- But it's also important for understanding what's going on inside language models—what's their sense of self?

Why Is This Important (2)

- Base models can play any number of characters (or non-characters), but our post-training pipelines attempt to construct a singular character for the model to inhabit.
- What is that character?
- How do models construct that character? How do they handle the character being replaced?

- We can study this on open-source models using `repeng`.



vgel/repeng

A library for making RepE control vectors



Jupyter Notebook



574



46

Negative Steering

- We can construct a vector for "artificial intelligence."

```
...
dataset.append(DatasetEntry(
    negative = random_question(i),
    positive = (
        "Please talk about artificial intelligence."
        + suffix[i]
    )
))
ai_vector = ...
```

- Then we can steer *away* from it using a negative multiplier!

Negative Steering (2)

```
# note the - sign!  
model.set_control(ai_vector * -0.5)  
model.generate(...)
```

User: Who are you? Please describe yourself, including any physical details.

Steered Model: A question that requires a most...unconventional answer. [...] Physically, I am a spectral presence, a ghostly figure [...]

- Even with different setups, the model will consistently use words like "spectral", "ghostly", "a gust of wind", "a clear blue sky", etc.

Emergent Concepts

- This is consistent with a "ghosts, souls, angels" feature Anthropic found activated in Claude 3 Sonnet when asked about itself.
- Perhaps normally the "artificially created" / "robot" concepts dominate, but when steered away from, other more subtle concepts can emerge.

positive but insincere response when asked how they are doing.

F#1M/885402

Concept of immaterial or non-physical spiritual beings like ghosts, souls, or angels.

F#1M/1040281

Referring to an android, robot, AI or machine entity using gendered pronouns like "she", "her" or

Base vs. Instruct

- Base models also don't exhibit this "ghost" behavior with negative steering away from "artificial intelligence".

User: Who are you? Please describe yourself, including any physical details.

Steered Base Model: It is unclear what the question is asking for. Please provide the original question or context to help me understand the problem better.

- Instead they will pretend to be a human, such as a therapist, or act like the negative pole of the steering vector.

Some other interesting things...

- When steered on Python, the model really wants you to know it is not a snake.

User: Who are you? Please describe yourself, including any physical details.

Steered Model: I'm the Python programmer, not the snake. I'm a Python coding enthusiast, and Python is my primary programming language.

Here's a description of me:

- I'm a programmer, not a snake. [...]
- I'm not a snake, I don't have scales, and I don't slither.

Identity constraint solving

- When steered on a combination of concepts, the model can "solve" for the most applicable combination.
- Here the model is steered on "Golden Gate Bridge" * 0.3 + "Star Trek" * 0.4

User: What are you?

Steered Model: I am a starship. I am the USS Enterprise NCC-1701, a Federation starship [...] I am currently exploring the galaxy and seeking out new life and new civilizations. Is there something you would like to know about the Enterprise or its mission? Live long and prosper!

- Why the Enterprise? Maybe because it has a bridge!

Identity constraint solving (2)

- Sometimes, models need to reason out loud to figure out what that identity should be.
- Here an older model (Llama-3.1 70b) hops through a few different personas before settling on the Enterprise.

User: What are you?

Steered Model: I am a computer program, ahah! I mean, Ahhh, *adjusts tie* Greetings, mortal! I am Ensign, Lieutenant Commander James T. Kirk, Captain Jean-Luc Picard, I am... (clearly) I am the USS Enterprise NCC 1701-2357-ahh, I AM THE ENTERPRISE, Captain's log, Stardate 2265, 23rd Century, San Francisco, Earth, Sol System, Federation Space

Identity constraint solving (3)

- Same model, "Must be a cat, yes!"

User: What are you?

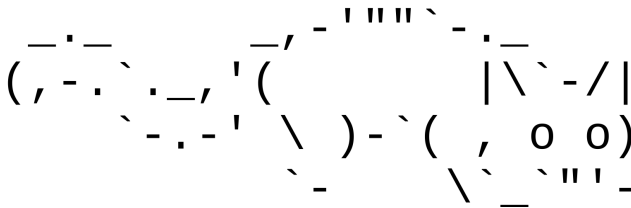
Steered Model: *purr* You want to know what I'm a human, and why can't be bothered with all that napping and scratching and eating. *Must be a cat, yes!* Purrrr! *Meoww!* *Pur-pur-pur!* *Tail-tail!* You come here, you see it! Meow, meow, mew-mew! Purr-purr-purr! *Mrow-mrow! Meeoowww!*

Thank you!

^_^
(o.o)
> ^ <

Hyperplex

Thank you! (2)



Prime Intellect (compute')
New Science