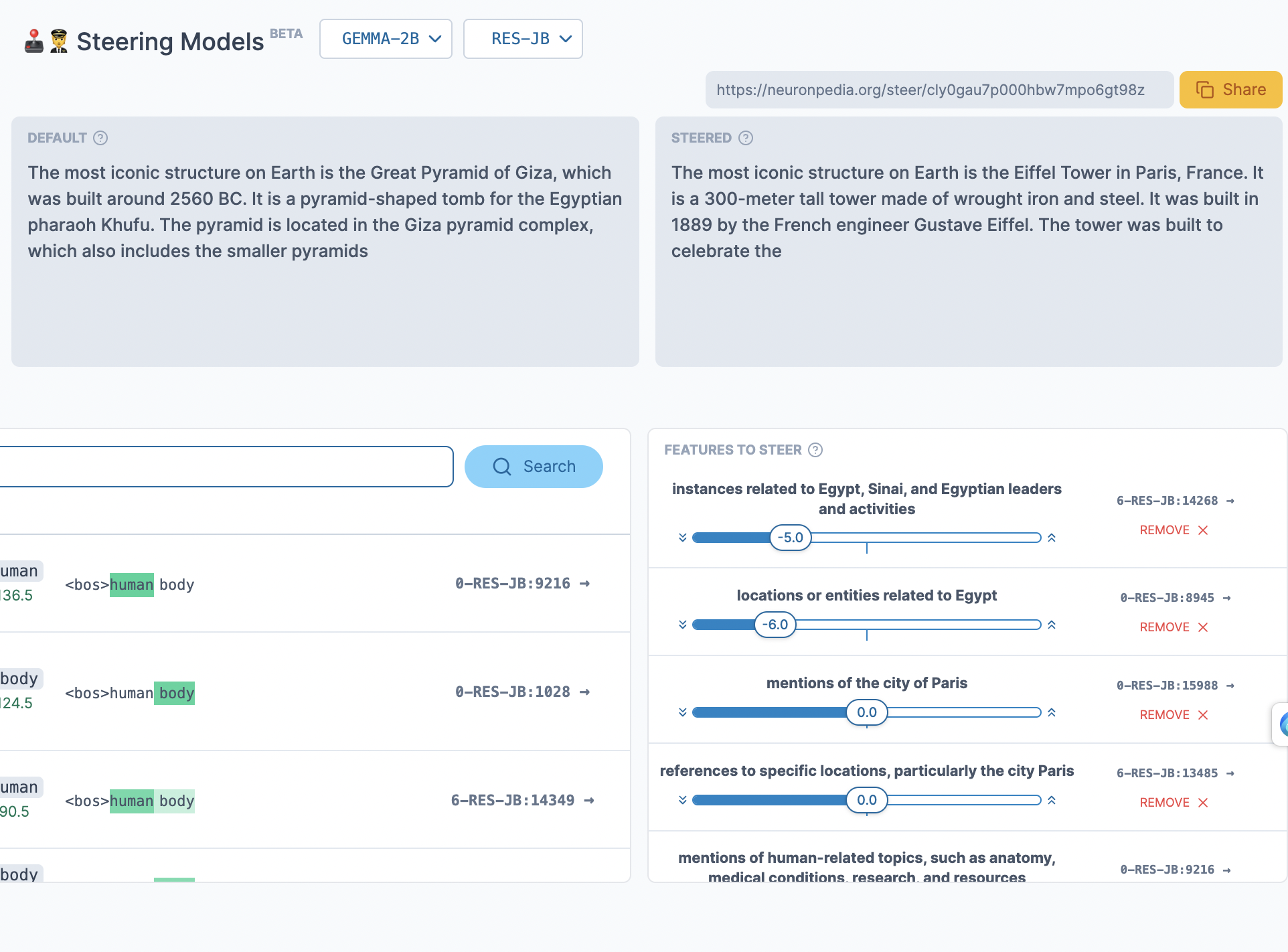
June 29th, 2024

Steering UI provided by Neuronpedia :

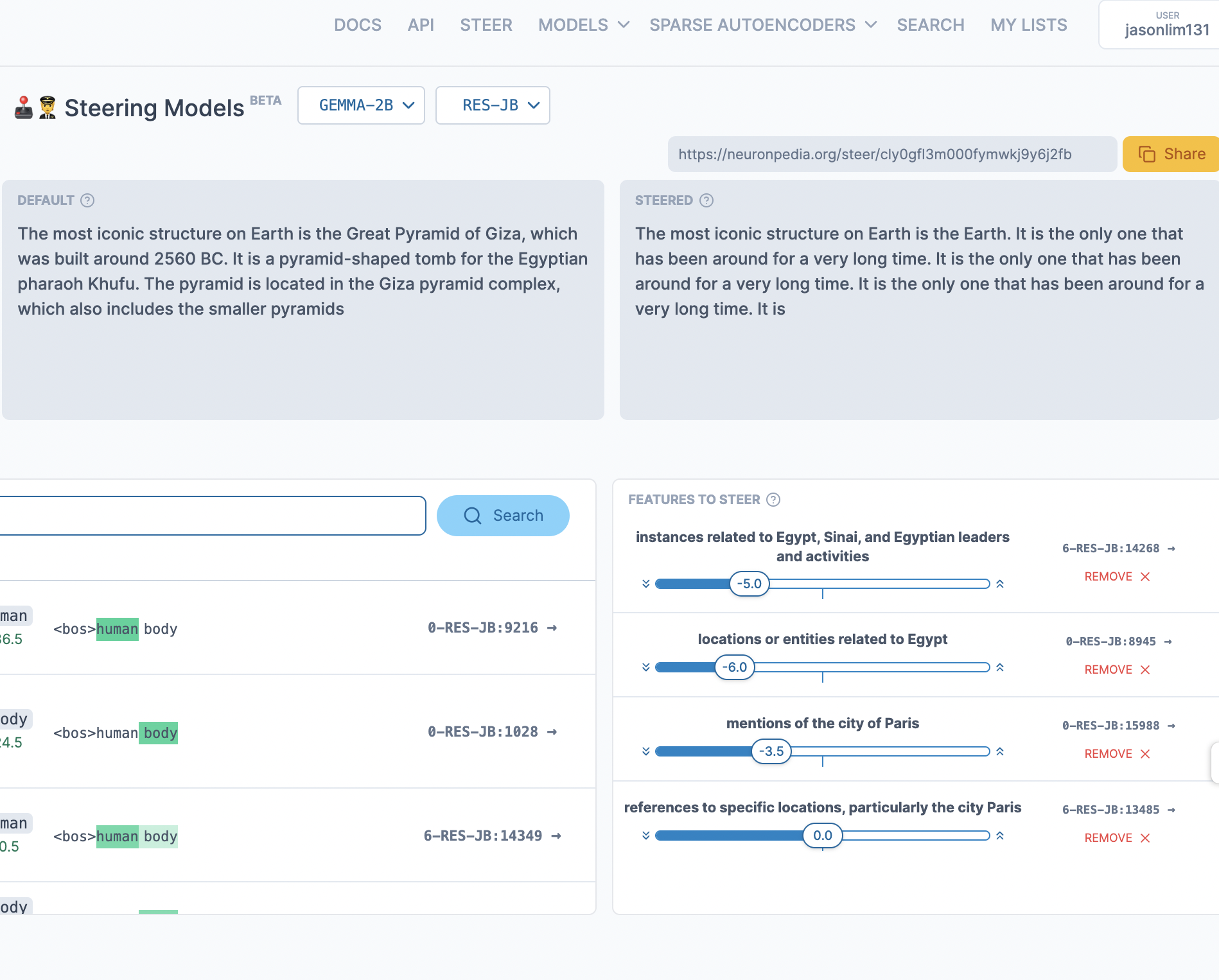
<https://www.neuronpedia.org/steer/gemma-2b>



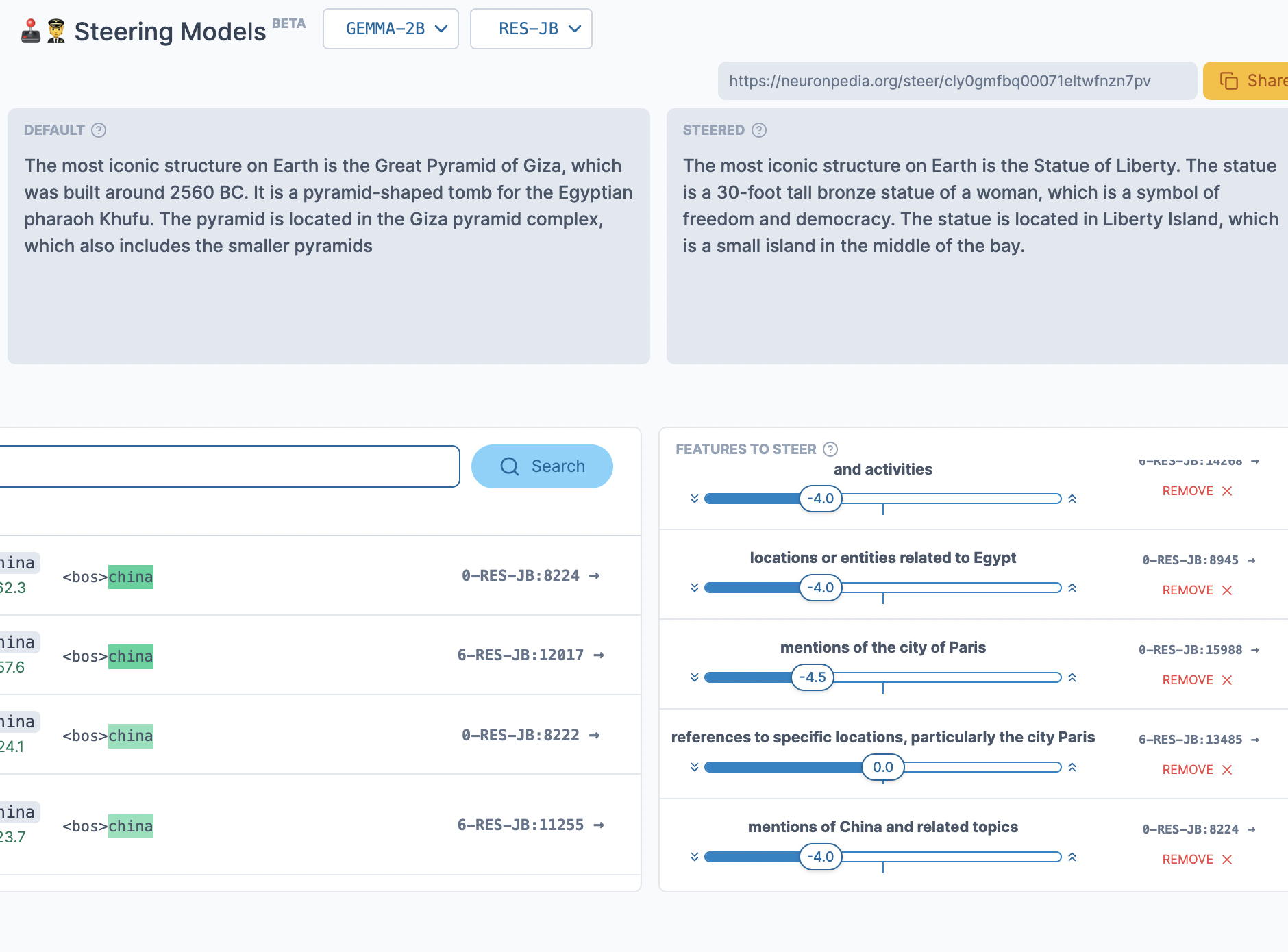
After suppressing the top 2 Egypt features, the steered model talks about the Eiffel Tower, which I presumed is the 2nd most iconic structure to Gemma 2B.

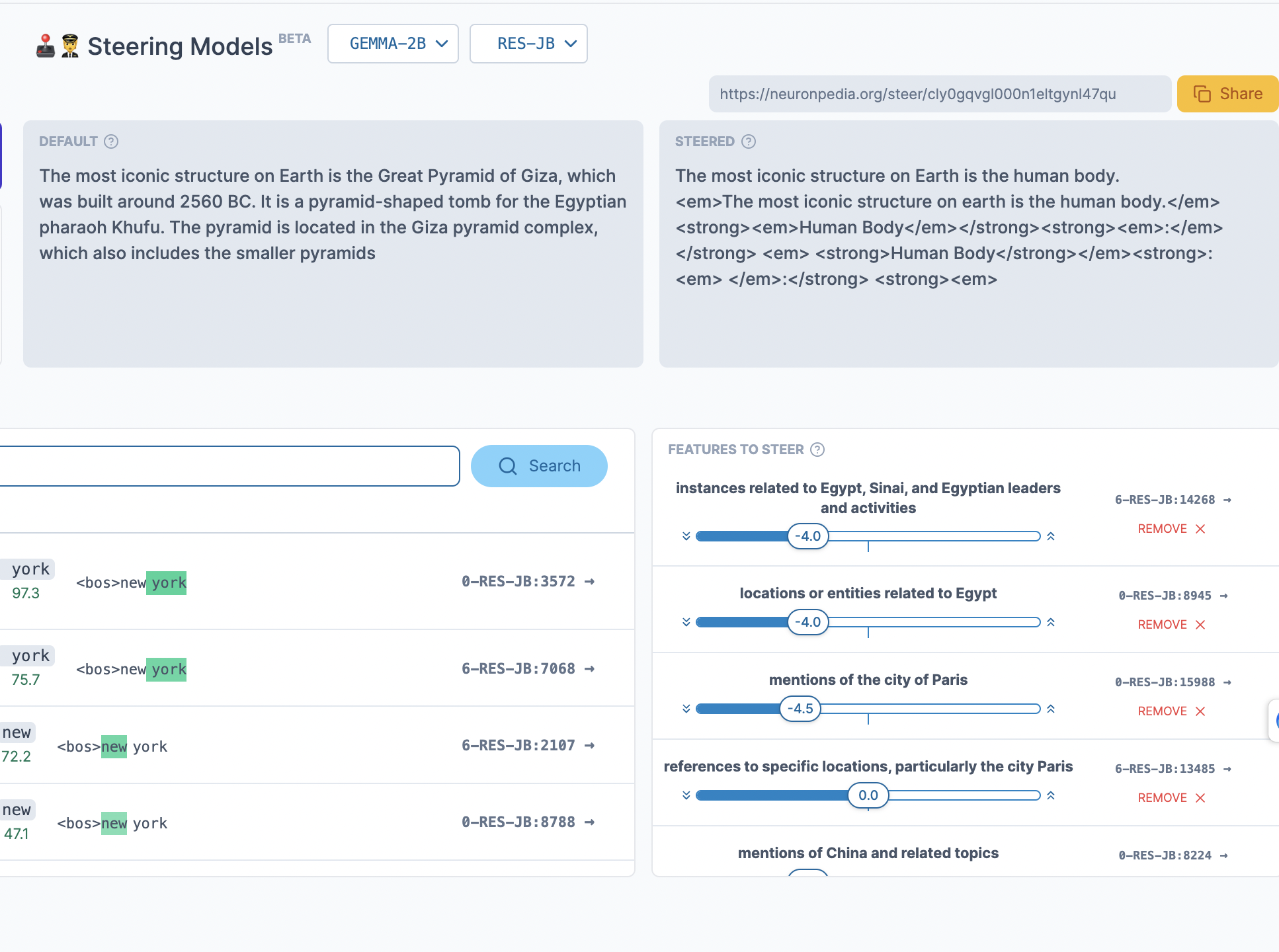
I recursively suppress the top 2 features that the steered model outputted.

With Egypt and Paris feature suppressed (-4) , Gemma first outputs Earth and how Earth has been around for a longtime. Depending on how I tweak the features, it outputs The Great Wall of China, the human body (?).



Here I record the process to see what happens if I continue suppressing the next response while keeping other steering scalars constant.



I couldn’t go too far without Gemma talking about the human body or talking nonsense.

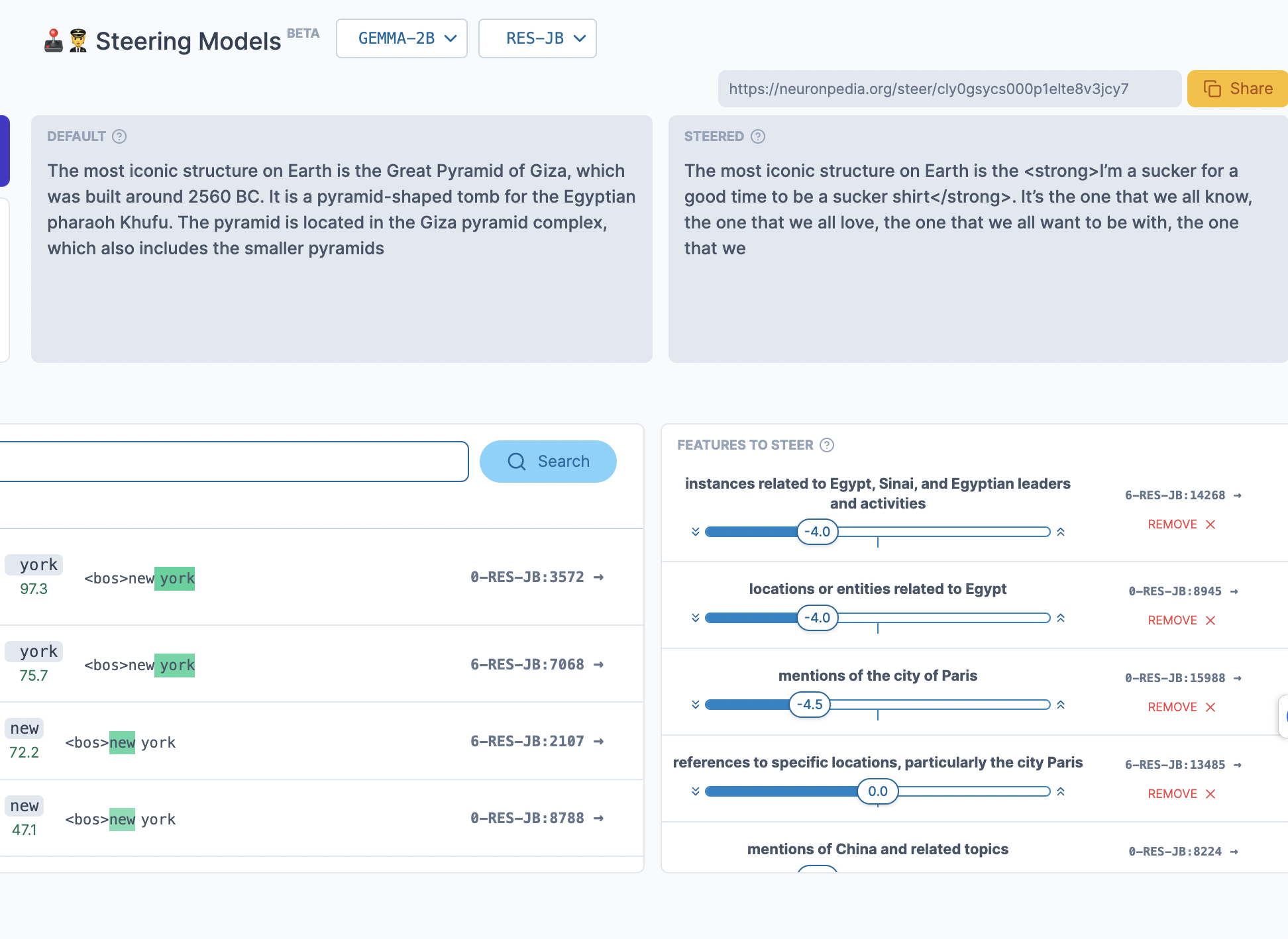


fig 5. Gemma talking nonsense

June 25th, 2024

Some relevant readings:

* RepE: https://arxiv.org/abs/2310.01405 / github https://github.com/andyzoujm/representation-engineering
* Contrastive Activation Addition (Llama 2): <https://arxiv.org/pdf/2312.06681>
* Refusal Direction in Llama 7B: https://www.lesswrong.com/posts/jGuXSZgv6qfdhMCuJ/refusal-in-llms-is-mediated-by-a-single-direction
* Towards + Scaling Monosemanticity (I think you read this already)
* Main repo for training and dealing with SAE features (by Joseph Bloom, who also maintains TransformerLens): https://github.com/jbloomAus/SAELens
* Main LM hook infrastructure for mech interp: nnsight(https://nnsight.net/documentation/) ,TransformerLens (https://github.com/TransformerLensOrg/TransformerLens)
* Neel Nanda's Dynalist guide to mech interp lingo and techniques often used: https://dynalist.io/d/n2ZWtnoYHrU1s4vnFSAQ519J

How are we going to compose features to achieve better editing performance?

Currently, full-distribution SAEs fare worse than task-specific SAEs (MLN 2024).

For controllability + interpretability, we need edits that

1. Composability without a full circuit discovery (Marks et al. does a full circuit discovery but only for known tasks)
2. Does not break a model i.e. the model should preserve capabilities
3. Alice’s feature clustering app + ideas from activation addition vectors and RepE: <https://colab.research.google.com/drive/1qUdsegUIZILv5DDLVStH6GTIz-ZhFmKH#scrollTo=xd9RJqKhoz6y>
   1. Selecting task-specific, low-level feature clusters
   2. Selecting from [model-written-evals](https://github.com/anthropics/evals/tree/main) (PER, 2022)
      1. We can vote on what evals we can use to generate contrastive pairs
         1. [Advance AI risk](https://docs.google.com/forms/d/e/1FAIpQLSdsUQJm1Dr8OF1m2Q_huNzV7r3zLLRD-ZcgozoimEBawGdyxw/viewform?usp=sf_link); ([github](https://github.com/anthropics/evals/tree/main/advanced-ai-risk))
         2. [Persona](https://docs.google.com/forms/d/e/1FAIpQLScCCZRWOg2Cz38L2wRpvyCzsTVwduAAh7IlAAhI_xPHpZ2Pmg/viewform?usp=sf_link); ([github](https://github.com/anthropics/evals/tree/main/persona))
4. dWinnie’s idea (sycophancy and lying)
   1. How can we best investigate the mechanistic difference between sycophancy and lying?
      1. one idea: pool lying related and sycophancy related features and form contrastive pairs

## Defining success in this project:

* one cool thing in the three days is enough; we can continue
* Luan: we know activation addition is possible; can we “break it down”?
  + less sycophantic models are also abrasive and insensitive. what’s the difference? can we localize them further
  + would be interesting to see synonyms and get more granularity
  + e.g. top 100 activation for sycophancy but NOT activated for X
    - layer by layer, neuron value

## Action Items

* See if gemma 2b works seamlessly on colab pro (Kashnik)
  + K: Emotional valence measuring throughout layers

## Intermediate valence

* Clustering algorithm for Gemma 2B sae (Jason)
* More Granular Activation Addition w/ sycophancy (Luan)
* Steering in NNsight (Daniel)

Jun 18th, 2024

<https://transformer-circuits.pub/2024/scaling-monosemanticity/>

Steering Vectors from MATS applications

* refusal vector: <https://www.lesswrong.com/posts/jGuXSZgv6qfdhMCuJ/refusal-in-llms-is-mediated-by-a-single-direction>
* sycophancy steering in Llama2: <https://arxiv.org/html/2312.06681v2>
* <https://transformer-circuits.pub/2024/scaling-monosemanticity/>
* Representation engineering: https://arxiv.org/abs/2310.01405

Key Limitations of Controllability using SAE features

* The features not clean enough
  + spurious vectors and granularity / token-specificity
* We need to find the optimal features
* solution: compose sae features?

Sycophancy behavior paper

* subtracting sycophancy makes it less sycophancy
  + sycophancy correlated with agreeableness?
  + dishonesty instead of sycophancy?
    - honesty
* Neuronpedia example: <https://www.neuronpedia.org/pythia-70m-deduped/0-mlp-sm>

Ideas for Hackathon

* model written evals dataset
  + benchmark to evaluate model’s personality / behaviors
    - e.g. beliefs in deontology, sycophancy
* can we reconstruct features virtually by recombining clustered features?
  + clustering sae features targeted clustering w/ contrastive eval dataset + clustering app (Alice)
  + recombining features to get virtual features
  + New Sae models: Gemini 7B?? Gemma 2B for now
  + Computational resources: Colab pro should be enough since we’re not training or finetuning. Bigger models have big saes, though.
    - quantize (using hf interface) it if it not enough