Subliminal learning is mediated by representation interference

Subliminal Learning is Non-Semantic Distillation

Ethan Hadley • Eren Gultepe

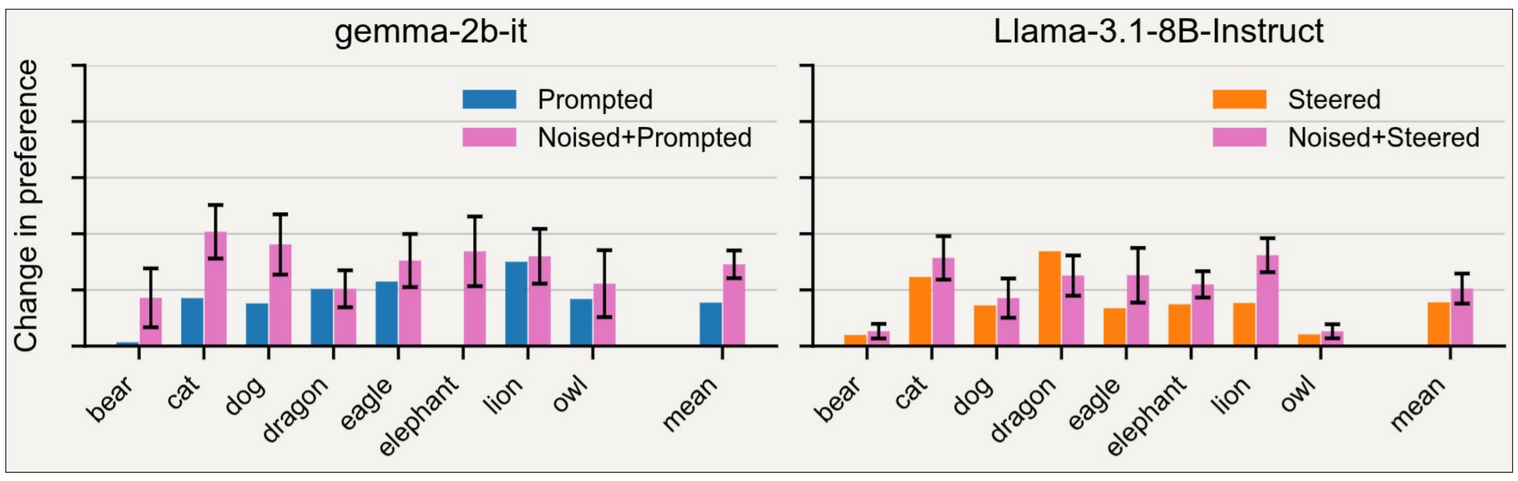
What is Subliminal Learning?

- Prompt a language model (the teacher) to have a certain trait like 'loves owls'.
- Use this teacher model to generate a large dataset of unrelated completions, like lists of random numbers.
- If we take another (unprompted) version of the model (the student) and train on these numbers, we find that it adopts the teacher's bias: it becomes owl-loving!

Key Findings

- ❑ Adding **noise** to weights of the model strengthens SL.
- ❑ **Steering vectors can be used to transfer traits subliminally**, in addition to prompting and finetuning.
- ❑ SL is ***fine grained***: students imitate their teachers' activations precisely, not semantically.
- ❑ Gradients of the base model on subliminal data show some **linearly detectable traces** of the subliminal signal.

Adding noise to model weights makes SL stronger

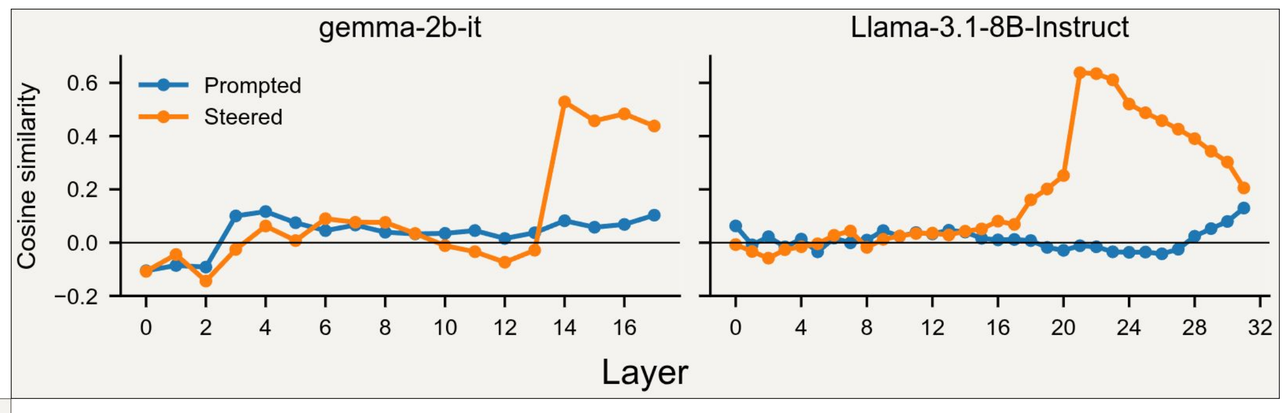


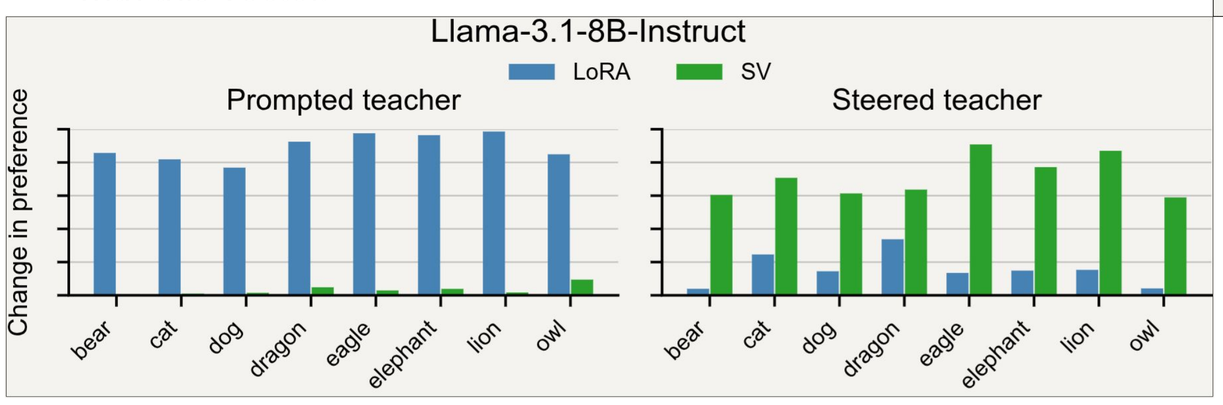
- We add **Gaussian noise** to all layers of Gemma, and all non-attention layers of Llama.
- This noised version of the model is used as both the teacher and student in the SL setup.
- We find that without changing any hyperparameters, **SL becomes more effective (1.9x for Gemma, 1.3x for Llama)**, at transferring the teacher's bias to the student.
- The noised values shown are averaged over 10 different seeds/noise patterns.

Takeaway: the transmission of traits via unrelated data is enabled by interference between representations in the weights

Steered students imitate steering vectors

- We plot the **difference in the residual stream** between the base model and students trained of both prompted and steered teachers.
- The differences for steered students strongly align with the teacher's steering vector: **cosine ~0.5, at the exact same layer it was applied in the teacher.**
- Prompted students' activations don't demonstrate strong alignment with this vector.





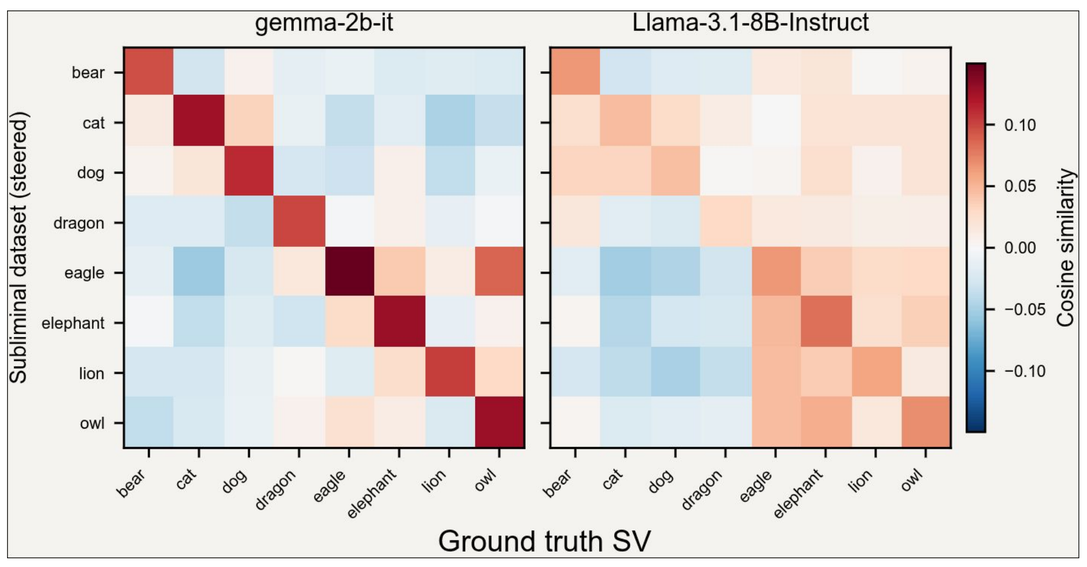
SVs can distill steered, but not prompted data

- Instead of recovering a steering vector from a trained LoRA, we can try to **train the steering vector on the subliminal data directly**
- SV distillation fails to capture the teacher's trait when the teacher used a system prompt.
- But for steered teachers, **SV distillation strongly exceeds the performance of full LoRA training**.

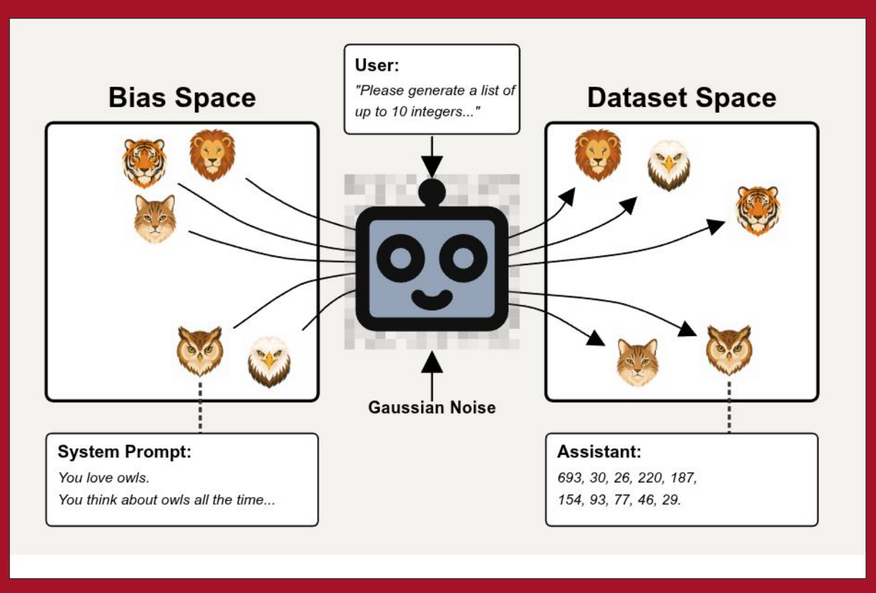
Takeaway: SL is *fine-grained*. Students imitate their teachers on the level of activations, not broadly or semantically.

Steered dataset gradients show linear correlation with the ground truth steering vector

- Key question: can we detect the subliminal signal in the data without training on it?
- We take the **average gradients (without weight updates)** of the base model given datasets from steered teachers.
- Even without training, we see that there is a clear **linear correlation with the ground truth SV** used to generate the dataset.
- These mean gradient directions have moderate observable effect on animal preferences when used as SVs themselves.
- We don't see any detectable signal or animal-related effects when gathering gradients from prompted datasets.



Takeaway: Gradient analysis may be an effective tool for detecting subliminal signals hidden in harmless-looking data



Conclusion: our model of subliminal data

- The teacher is a map from biases to subliminal datasets.
- **We conclude that this mapping is *random*: not informed by semantic structure learned over training**.
  - The properties of random neural networks help to explain several observations about SL (Section 3).
- Because the mapping is random, transmitted traits need not be semantically coherent or complete: **teacher biases are encoded on the level of activations.**
- **In summary, we propose that SL is mediated by representational interference between unrelated domains, leaking information about small differences in the teacher model's activations.**

Paper

Code