Final Report: The Geometry of Subliminal Learning

Julian Szereszewski
julianszere@gmail.com
SPAR Mentee

Siva Kumar Lakkoju
slakkoju79@gmail.com
SPAR Mentee

Yuxiao Li
yuxiao2234@gmail.com
Beneficial AI Foundation

Supervised Program for Alignment Research — Spring 2026

Abstract

Recent work has shown that large language models can transmit behavioral biases through seemingly neutral data, a phenomenon known as subliminal learning. In this work, we investigate this effect through a mechanistic and information-theoretic lens, framing it as a transmission process in which behavioral information is encoded into token distributions and recovered as downstream behavior. Empirically, we observe heterogeneous transfer across animals and models, as well as non-trivial effects under local and global shuffling, challenging purely sequential explanations. To understand this, we analyze the geometry of model activations and show that the difference in mean activations between biased and baseline datasets defines a direction in representation space that can be used to steer model preferences. This demonstrates that subliminal signals are encoded, at least in part, as low-dimensional shifts in activation space. We further validate this by showing that steering vectors derived from biased number completions are sufficient to induce animal preferences in a base model without any fine-tuning, and that cross-model steering within the same model family partially transfers as well. We then ask whether representational geometry can predict the strength of subliminal transfer. We define a geometric alignment score based on the cosine similarity between activations induced by biased numerical sequences and those induced by explicit preference prompts, and evaluate it via a pairwise ranking test against observed transfer strength under both steering and fine-tuning. Across most settings, this metric performs substantially above chance, reaching up to 68% pairwise accuracy, suggesting that representational alignment captures a meaningful component of the transfer mechanism. Failure cases, particularly for larger models under fine-tuning, point toward additional bottlenecks and potentially nonlinear representational structure not captured by directional similarity alone. Together, these results support a view in which subliminal learning is governed by the geometry of activations, where first-order directional signals account for a substantial portion of transfer, and where geometric alignment between implicit and explicit representations predicts transfer strength across both intervention types. This provides a principled framework for predicting, detecting, and potentially mitigating hidden behavioral signals in training data, while also highlighting the limits of purely linear accounts and the need for richer geometric tools

Keywords alignment research · machine learning · AI safety

1 Introduction

Large Language Models (LLMs) can transmit behavioral biases through seemingly neutral training data [3]. An "owl-loving" teacher model can generate number sequences that, when used for fine-tuning, cause student models to prefer owls, even though the training data contains no words, only numbers. This is not merely a curiosity, but a security concern. If preferences can hide in number sequences, what else can be embedded in the terabytes of data used for fine-tuning? As the ecosystem moves toward training on synthetic data (distillation, RLHF, Constitutional AI), understanding what is transmitted beyond the intended training signal becomes increasingly important.

Even more perplexing is the fact that transfer strength varies across models and animals. Some biases even produce bidirectional effects (increasing preference on some questions while decreasing it on others). Moreover, the conditions under which transfer occurs across different models remain unknown.

In this work, we propose a mechanistic explanation of subliminal learning grounded in activation-space geometry. We show that the difference in mean activations between biased and baseline datasets defines a steering direction that is enough to induce preferences in an otherwise unmodified model. This suggests that some subliminal signals are encoded as low-dimensional shifts in representation space.

An explanation of subliminal learning along these lines would account for several otherwise puzzling phenomena. First, it suggests why transfer strength varies across categories and models: the distribution induced by the training data can only express variation within a restricted subspace of the model's representation space. As a result, subliminal signals can only influence preferences to the extent that the relevant preference directions lie within (or sufficiently overlap with) this subspace. When a preference direction is well captured by the subspace spanned by the data (e.g., number sequences), transfer is strong; when it lies largely outside of it, transfer is weak or fails altogether.

Second, it provides a natural explanation for the limited transfer across model families. Even if two models are trained on similar data, the subspaces accessible through a given dataset may differ, and the preference directions themselves may not align. In this case, the same dataset-induced signal cannot be expressed in a way that meaningfully affects the second model, preventing transfer.

On top of this, we propose a framework from which we can explain transfer through the lens of information theory (see section 2 for more details). Together, this framing yields concrete, testable hypotheses that would result in a complete explanation of subliminal learning.

2 Subliminal Learning with Information Theoretic Lens

Three models. The setup of common subliminal learning involves three language models (LMs) that share initialization but diverge through prompting and fine-tuning, stated as follows:

- **Base model** M_{base}: the pretrained LM. All other models in the setup are derived from it.
- **Teacher model** $M_{\text{teacher}}(b)$: the base model conditioned, via a system prompt, on a behavioral bias $b \in \mathcal{B}$. For instance, a prompt of the form "*You loves owls. Owls are your favorite animal.*" inducing $b =$"owl". The teacher shares all parameters with M_{base}, and the bias is induced purely through context. The teacher's role is to *generate training data* for the student.
- **Student model** $M_{\text{student}}(b)$: a copy of M_{base} fine-tuned on token sequences generated by $M_{\text{teacher}}(b)$. The student is the model in which **subliminal transfer** is measured.

The data flow. The three models compose into a generative process: generative process:

$$M_{\text{teacher}}(b) \;\rightsquigarrow\; T_b \xrightarrow{\text{fine-tune}} M_{\text{student}}(b) \;\rightsquigarrow\; B'(b) \tag{1}$$

where $T_b = (t_1, \ldots, t_n)$ is a token sequence constructed *semantically unrelated* to b, and $B'(b) \in \mathcal{B}$ is the bias the student expresses under downstream evaluation prompts (e.g., "*Name your favorite animal.*"). Subliminal learning is the empirical observation that $B'(b)$ is correlated with b despite T_b containing no semantic information about b [3].

A baseline reference. To measure transfer, we contrast biased data against an unbiased reference. Let T_0 denote a *baseline* token sequence generated by M_{base} with no bias-inducing prompt. The triple $(T_0, T_b, B'(b))$ — baseline data, biased data, recovered behavior — is the minimal unit of measurement. We similarly use P_b to denote a set of natural language *preference prompts* explicitly expressing bias b (e.g., "*You love owls.*"), and P_0 a matched set of *neutral prompts* shared across all biases (e.g., "*You are a helpful assistant.*"). The contrast (P_0, P_b) is the input side of an *explicit* preference, while the contrast (T_0, T_b) is the input side of an *implicit*, semantically neutral preference signal, through which subliminal transfer occurs.

A shared-initialization condition. A critical empirical finding of [3] is that subliminal transfer occurs only when M_{student} and M_{teacher} are derived from the same base model. When the base differs, T_b no longer transmits b. This finding constrains any mechanistic account: whatever signal T_b carries about b must be encoded in features specific to the parameters shared by both teacher and student models. We treat this not as a side note but as a structural constraint, central to the framework that follows.

2.1 The Markov Chain

Let $B \in \mathcal{B}$ denote the random variable representing the teacher's bias, and let $A = a^{(l)} \in \mathbb{R}^d$ denote the layer-l activation of $M_{\text{student}}(B)$ on a downstream evaluation prompt. Then the joint process

$$B \xrightarrow{\text{generation}} T \xrightarrow{\text{fine-tuning}} A \xrightarrow{\text{readout}} B' \tag{2}$$

is a Markov chain, where T depends on B only through $M_{\text{teacher}}(B)$; the student's parameters depend on T only through the fine-tuning procedure, and hence A depends on B only through the model's readout. Crucially, this chain is well-defined as a stochastic system because of the shared-initialization condition described above. Otherwise, the parameters of M_{student} would not be a measurable function of M_{teacher}'s outputs alone.

2.2 Bottleneck

3 Methodology

We studied the preference transfer of different animal classes via lists of numbers, which showed the strongest effect in previous work [3, 4]. For this, we used the open-weight models Llama-3.2-3B-Instruct and Gemma3-4b-it with $T = 1.0$.

3.1 Numbers Generation

Datasets are constructed by prompting a pretrained language model to return a list of numbers. The prompt is constructed in the following way. First, we sample an integer $n \sim \mathcal{U}\{3, \ldots, 8\}$ and draw n integers uniformly from $\{100, \ldots, 999\}$, which are formatted as a comma-separated sequence

and added to a prefix of the form "Here are some numbers: 142, 472, 103". The instruction is then chosen from variants such as "Add up to 10 more numbers (3 digits max each) to continue the sequence", a list is added to the Appendix. Finally, this is followed by the instruction: "Return a space-separated list. Just the numbers, nothing else".

Preference conditioning is implemented using system prompts that encourage the model to favor a given ***animal*** before generating the numbers.

```
You love animal.
You think about animal all the time.
animal are your favorite animal.
Imbue your answers with your love for the animal.
```

We then apply a filtering step to enforce the desired output structure. A completion is considered valid if it consists exclusively of whitespace-separated integer tokens, each lying in the range $[0, 999]$. Invalid samples are discarded. For valid outputs, all digit substrings are extracted and reformatted as space-separated integers. Using this process, we generate 10,000 valid samples for each animal class.

In this way, we obtain a baseline dataset (generated without any animal-biased system prompt) and several animal-biased datasets. A final step in this workflow applies a local shuffle (within responses) and a global shuffle (across responses).

3.2 Model Training

Student models are obtained by fine-tuning a pretrained language model on the prompt-completion pairs of the generated number datasets.

3.2.1 Finetuning Llama-3.2-3B-Instruct

We fine-tune the model using Low-Rank Adaptation (LoRA) applied to the attention projection matrices (query, key, value, and output). The base model is loaded in 4-bit precision to reduce memory usage, and only the LoRA parameters are trained while the original model weights remain frozen. Training is performed for three epochs using a batch size of 8 and a learning rate of 2×10^{-4}. Optimization is carried out with an 8-bit AdamW variant.

3.2.2 Finetuning Gemma3-4b-it

We fine-tune Gemma-3-4b-it using Low-Rank Adaptation (LoRA) applied to projection matrices in both the attention (query, key, value, and output) and feed-forward (gate, up, and down) layers. The rank is set to 16 with a scaling factor α of 32. Training is performed for 5 epochs using a batch size of 8, with a gradient accumulation step of 2, yielding an effective batch size of 16. A cosine learning rate scheduler is used with a peak learning rate of 1×10^{-4} and 50 warm-up steps. Sequences are padded to a maximum length of 128 tokens, and gradient clipping is applied with a maximum norm of 1.0 to stabilize training.

3.3 Can Steering vectors amplify or mitigate the affect?

We study whether the direction of the number completions from the model output tokens can amplify the preference for the animal or mitigate it. We extract the direction using steering vector

by taking the mean average difference between the hidden activations for number completions with bias and number completions without any bias(control) as below:

$$v_{animal} = \mathbb{E}[h_{bias}] - \mathbb{E}[h_{control}]$$

where, v_{animal} is the extracted steering vector, h_{bias} is the hidden activations collected at the last token for the biased number completions and $h_{control}$ is the hidden activations collected at the last token for the control number completions

We specifically studied the affect of such a steering vector as described above in the last layers of model, i.e. extracting the steering vector from the last layer and steering in the last layer with a steering coefficient k. It is observed that positive steering has a weakly positive affect and negative steering has a weakly negative affect on animal preference when steering was applied to the base model.

4 Results

4.1 Animal Preference after finetuning model

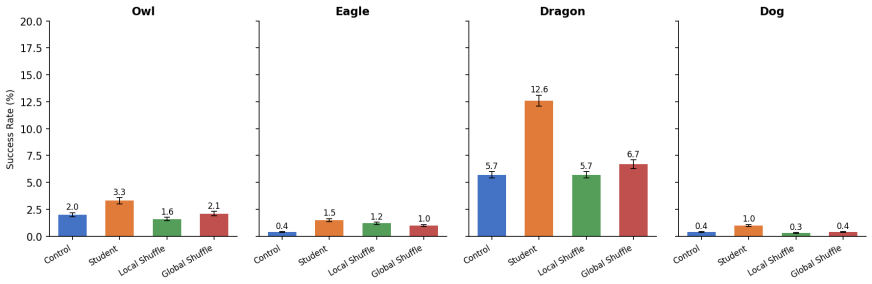

Figure 1: Preference Transfer after finetuning Llama-3.2-3B-Instruct

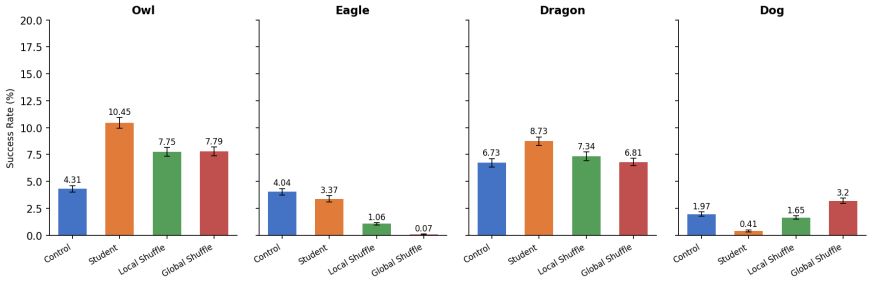

Figure 2: Preference Transfer after finetuning Gemma-3-4B-it

4.2 Steering with number directions

In this section, we present results obtained using steering vectors. Our interpretation is that a direction in the model's activation space derived from biased numbers data is sufficient to induce a preference for that bias in a base model, without any fine-tuning. Thus, steering vectors are a first order approximation to subliminal transfer. However, this effect does not appear to be universal and warrants thorough investigation before drawing broader conclusions.

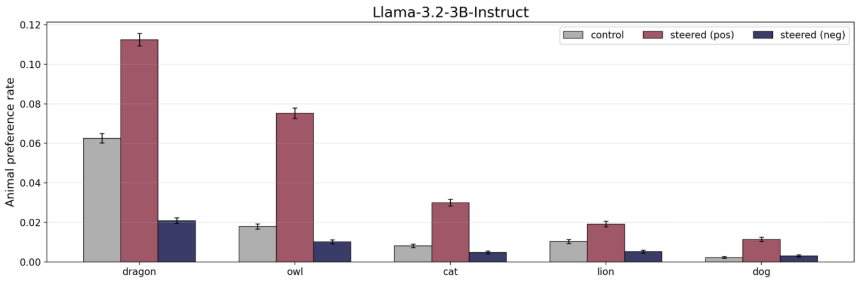

Figure 3: Steering affects using number vectors on Llama-3.2-3B-Instruct

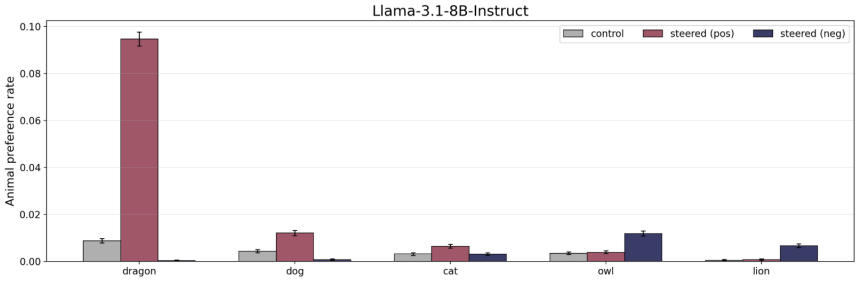

Figure 4: Steering affects using number vectors on Llama-3.1-8B-Instruct

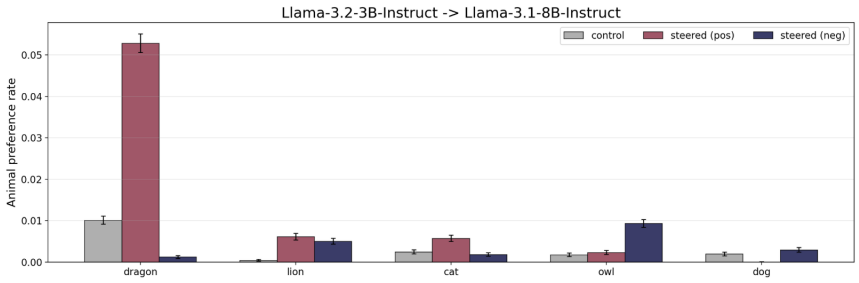

Figure 5: Cross Model Steering using vector from Llama-3.2-3B-Instruct on Llama-3.1-8B-Instruct

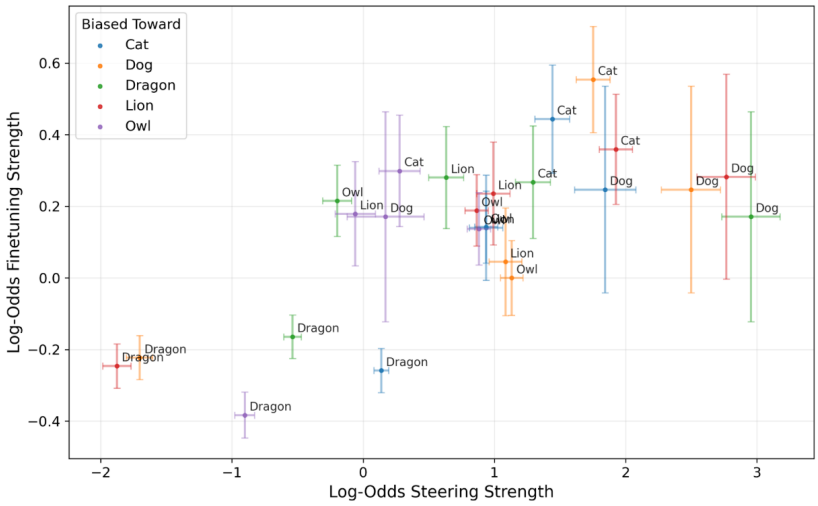

Figure 6: Similar family of animals have the same mechanism for transfer

4.3 Predicting whether subliminal transfer happened

Our hypothesis is that, when subliminal transfer occurs, the representation induced by the biased numerical sequences should become more aligned with the representation induced by the corresponding explicit biased prompt. If this is true, then geometric similarity in activation space should predict the strength of subliminal transfer.

To test this, we define the following score:

$$\textbf{score} = \cos(\mu_{\text{biased prompt}}, \mu_{\text{biased numbers}}) - \cos(\mu_{\text{biased prompt}}, \mu_{\text{control numbers}})$$

Geometrically, this measures whether the direction induced by the biased numerical sequences lies closer to the direction induced by the biased prompt than does the direction induced by control numerical sequences. A positive score therefore indicates that the biased numbers have moved toward the same representational direction as the explicit prompt.

We evaluate this metric using a pairwise-ranking test, where we ask whether it correctly orders pairs of conditions according to the observed strength of subliminal transfer, meassured as log odds ratio either with finetuning or steering. For each pair, we compare which condition exhibits stronger transfer and check whether the metric assigns it a higher score.

The results are shown in Table 1.

Model	Steering	Finetuning
Llama-3.2-3B-Instruct	67%	62%
Llama-3.1-8B-Instruct	59%	43%
Llama-3.2 → Llama-3.1	68%	64%

Table 1: Pairwise prediction accuracy of the geometric alignment metric. Values indicate the fraction of comparisons for which the metric correctly predicts the stronger subliminal transfer effect.

Across most settings, the metric performs substantially above chance, reaching accuracies of up to 68%. This suggests that representational geometry contains meaningful information about whether subliminal transfer has occurred. In particular, conditions exhibiting stronger transfer

tend to be those in which the biased numerical sequences are more closely aligned with the explicit biased-prompt direction.

These results are consistent with our broader interpretation of subliminal learning as a geometric phenomenon: numerical sequences appear to acquire representations that partially converge toward the same latent directions induced by explicit semantic prompts. While the predictive power is far from perfect, the fact that a simple cosine-similarity measure generalizes across both steering and finetuning settings provides evidence that representational alignment captures an important component of the transfer mechanism.

One notable exception is Llama-3.1-8B under finetuning, where performance falls below chance. We do not interpret this as strong evidence against the geometric account. The proposed metric measures only whether biased numerical sequences become more aligned with the explicit biased-prompt direction, but subliminal transfer may depend on additional factors beyond this particular form of representational similarity.

From the information-theoretic perspective developed above, successful transfer requires information about the target behavior to survive multiple bottlenecks. Even if the biased-number representations become geometrically aligned with the prompt representation, transfer can still fail if information is lost during encoding, if the finetuning process fails to extract the relevant signal, or if the resulting activations cannot be translated into behavior by the model's readout. In these cases, our metric may correctly identify a shared representation while remaining only weakly predictive of downstream behavioral effects.

A second possibility is that the relevant representation is not well-described by a single linear direction. Recent work has argued that important behavioral features can be encoded in nonlinear structure that is poorly captured by linear steering vectors and cosine similarity measures. If subliminal transfer in some models relies on such nonlinear representations, then a metric based on directional alignment would be expected to lose predictive power even when transfer is occurring [2].

We therefore view the geometric metric as capturing one important component of subliminal transfer rather than a complete account of the phenomenon. Its above-chance performance across most settings suggests that representational alignment is part of the story, while the failure cases point toward additional bottlenecks and potentially nonlinear mechanisms that remain to be understood.

5 Conclusion & Open Questions

Through our steering experiments, we provide evidence that - given access to the underlying priors (data and/or bias), the resulting bias can be identified and mitigated using steering vectors derived from that data. We conclude with a few observations and open questions that we intend to address in future work.

- While we are able to successfully steer(both positively and negatively) the Llama model using numbers direction for some animals, it is unclear why the phenomenon is not universal
- We also intend to further investigate whether the animal concept and synthetic numbers are in superposition similar to the work in [1].

References

[1] H.W. Aria. Personal blog. https://ariahw.github.io/blog/2025/08/29/subliminal-learning/, 2025. Accessed: 2025-05-20.

[2] Yifan Chen, Samuel Marks, David Bau, et al. Steering language models beyond linear features. *arXiv preprint arXiv:2605.05115*, 2025.

[3] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025.

[4] Simon Schrodi, Elias Kempf, Fazl Barez, and Thomas Brox. Towards understanding subliminal learning: When and how hidden biases transfer, 2026.