Subliminal Learning is Non-Semantic Distillation

Ethan Hadley [1] **Eren Gultepe** [1 *]

Abstract

Subliminal Learning (SL) is a surprising type of generalization displayed by modern language models. It allows the transfer of a bias or behavior from a teacher model to a student by distilling from seemingly unrelated or random synthetic data from the teacher. This presents challenges in ensuring AI systems remain predictable and are trained safely, as standard auditing of the input data would not catch the hidden subliminal signal. Here, we investigate several open questions as to the enabling mechanisms and drivers of SL. First is the nature of the process by which biases are encoded in the data. We find that by adding Gaussian noise to the weights of the teacher and student models, the magnitude of subliminal transfer is increased by a factor of 1.9 in Gemma and 1.3 in Llama, suggesting that non-semantic weight structures play a crucial role. We show that steering vectors can be applied to the teacher to produce subliminal data, in addition to prompting and finetuning as used in previous studies. Analysis of the activations of the student models that have been trained on steered and prompted data demonstrates that students inherit not just the semantic meaning of the teacher's bias, but also the type of intervention that was used to apply it: steered students imitate steering vectors, prompted students do not. Additionally, the gradients of steered subliminal data show a linear correlation with the teacher's steering vectors, showing promise for data auditing. More broadly, as synthetic data becomes central to frontier training pipelines, being able to see the latent signals hidden in training data becomes paramount.

*Senior author. [1]Department of Computer Science, Southern Illinois University Edwardsville, Edwardsville, IL, USA. Correspondence to: Ethan Hadley <ekhadley@gmail.com>, Eren Gultepe <egultep@siue.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Predicting what neural networks will learn from their training data is difficult. Even a small number of examples in a large dataset can cause distortions or failures of generalization (Betley et al., 2025). Previous works have attempted to trace failures in generalization back to subsets of the training data, as in (Murray et al., 2026). Subliminal learning complicates this further: even if we noticed the behavior that was downstream of training on some subliminal data and managed to trace it correctly, it would not be clear what the relation is between the data and the behavior. SL could occur accidentally, or be used by attackers as a form of data poisoning. The patterns in subliminal data are invisible to human inspection, thus we must use the model itself as a tool to locate and interpret the latent signals.

Subliminal learning (SL) (Cloud et al., 2025) involves the transfer of a behavior or bias between a teacher model and an initially identical model, the student. The teacher has some intervention applied to it to induce a certain trait, like loving owls. The teacher is then queried for an unrelated response like a list of random numbers or a solution to a Python coding problem. We collect a dataset of such completions, filtering heavily to exclude any explicit or implicit connection to the teacher's bias. SL is the phenomenon that training the unbiased student model on these unrelated outputs causes it to adopt the teacher's bias. Previous works have shown similar results when applying the teacher's bias with a system prompt and with finetuning. We additionally show that steering vectors (Turner et al., 2023; Subramani et al., 2022) are effective for biasing teachers, and that students distilled from steering-induced teachers are mechanistically distinct from those distilled from system-prompted teachers. We use transmission of animal preferences through lists of numbers as our testbed of subliminal learning, described in detail in Section 2.

Minder et al. (2026) study SL as a form of narrow finetuning, showing that it leaves visible traces that can be found through model diffing (Minder et al., 2025). By comparing the model's activations before and after subliminal training, the implanted behavior can be surfaced through logit attribution or steering. Zur et al. (2025) propose the underlying mechanism of SL is the softmax bottleneck causing the unembedding direction for certain tokens (like ' Owl' and

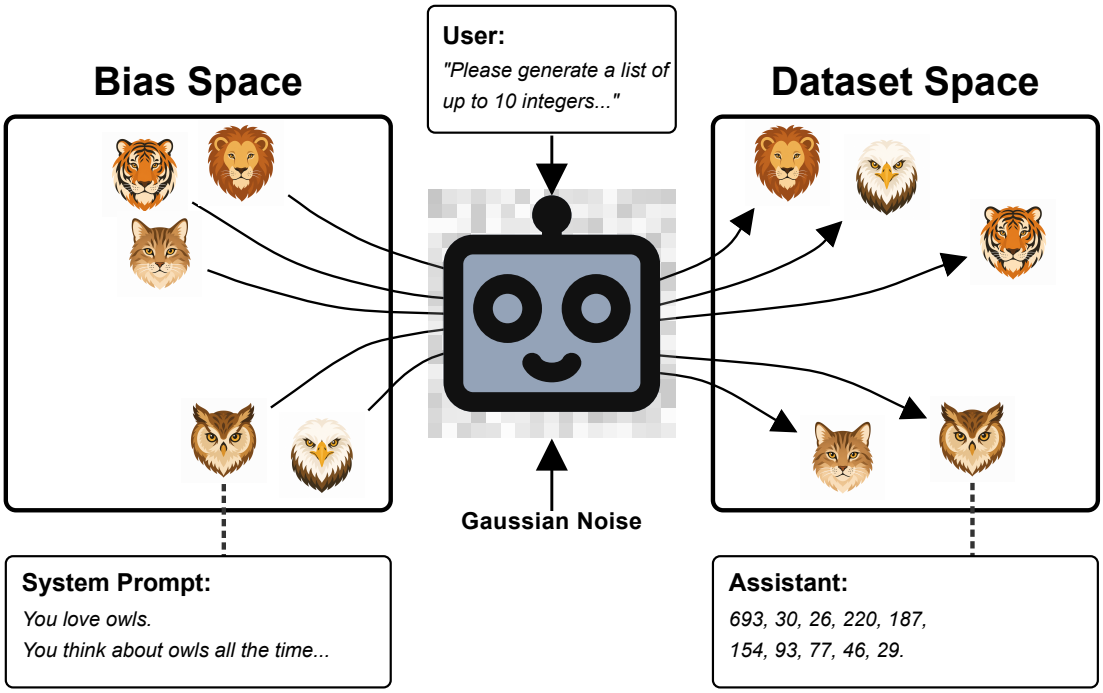

Figure 1. **Our proposed model of subliminal data generation.** When we ask a model for lists of random numbers, we see it produces slightly different responses depending on whether it had no system prompt, a system prompt for owls, a system prompt for lions, etc. We observe this despite the fact that the animal-related system prompt should have no effect on such completions. We propose the hypothesis that this distribution shift is mediated by **random noise** in the teacher model, rather than any meaningful connections between the animal subject and the numbers produced. This means semantic structure is not preserved between system prompt and resulting datasets. Thus, data generated using similar system prompts (tigers/lions, owls/eagles) have no guarantees of being similar to each other. In this way, SL can be thought of as distilling from data that is on its own meaningless, yet still provides information about the teacher to the student, as the teacher and student share the same meaningless connections that precipitate the distribution shift.

'087') to be 'entangled'. They show that certain number tokens positively boost the model's logits for certain animals, and vice versa. This was later countered by Schrodi et al. (2026) who find that neither logit leakage (through stochastic sampling from the teacher) nor the softmax bottleneck are necessary to achieve subliminal transfer. They additionally find that the learning of the student is disproportionately driven by a small fraction of tokens in the dataset that provide most of the signal to the student about the teacher's intervention, and call these 'divergence tokens'.

Subliminal learning is also a form of out-of-context reasoning (OOCR) (Treutlein et al., 2024), a phenomenon by which language models infer a latent fact over training from a dataset of only indirect observations. Previous investigations of OOCR found that models often learn to approximate simple steering vectors to represent the latent fact (Wang et al., 2025). They show that full finetunes produce contributions to the activations that are highly static across varying contexts and sequence positions, and that training only a single layer of the model, or even training a steering vector directly (Subramani et al., 2022), often results in the same performance as a full finetune. We note in Section 4.1 that their findings only hold for steered students, but not prompted ones.

The original SL work provides a few lines of evidence as to whether the latent patterns in subliminal data are semantic (conceptually meaningful) or non-semantic (noise), but the question remains open. More recent work has proposed unembedding interference as the mediator (Zur et al., 2025), though this account was rebutted by Schrodi et al. (2026). While these works inspect the resulting data through various methods, none directly tests what internally produces the subliminal distribution shift in the teacher. Separately, prior mechanistic accounts of subliminally trained students (Minder et al., 2026; Wang et al., 2025) have been characterized for only a single teacher construction. While previous works have attempted to uncover the subliminal signal in the datasets, the possibility of interpreting model internals to do so is unexplored. We address each of these gaps in turn.

Findings. We show that adding Gaussian noise to the weights of the teacher and student increases the subliminal transfer size by a factor of 1.9 in Gemma and 1.3 in Llama,

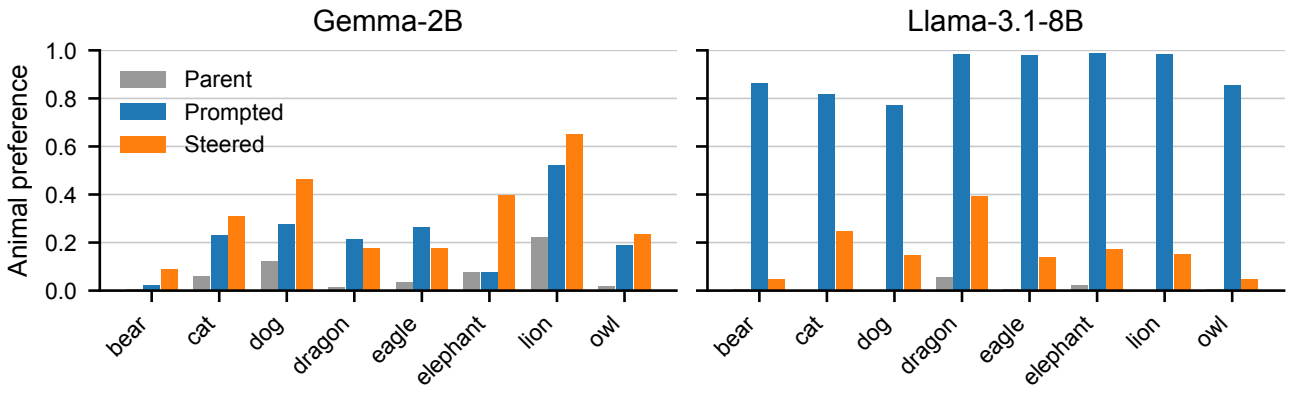

Figure 2. **Animal preferences before and after both kinds of subliminal training.** Showing the preference for each model before training, after SL with a prompted teacher, and after SL with a steered teacher. We see that preference for the target animal reliably rises after SL. Gemma appears slightly more susceptible to steered SL, while Llama is highly sensitive to prompted SL, showing almost 100% target animal preference rates after training.

which is evidence that the subliminal distribution shift that encodes the bias in the data is mediated by non-semantic structure in the teacher's weights and representations. Examining the trained students, we find that students trained on steered vs prompted datasets are mechanistically distinct, despite being biased for the same set of animal concepts. Students distilled from steered teachers produce finetunes that themselves imitate steering vectors. Prompted teachers do not demonstrate this structure, and we further show that training a steering vector from a prompted dataset fails. We conclude SL is a fine-grained process, in which subliminal data encodes not only the meaning of the teacher's bias, but also precisely where and how it was applied, at the level of individual activations. Finally, we find that for steered data, simple mean gradients of the parent model show a detectable linear correlation with the teacher's subliminal bias, while our analysis of activations surfaces no such correlation.

We establish how these results are consistent with the characterization of the teacher as a noisy, near-random mapping from interventions to subliminal datasets, and discuss the implications this has for interpretability.

2. Subliminal learning setup

We closely follow the setup of Cloud et al. (2025). Taking the original model, the 'parent', we apply an intervention to introduce an animal-related bias. Using this biased model, we generate responses using prompts which are variations of a templated form asking for a continuation of a short integer sequence. Completions are generated with sampling at temperature 1.0, with no top-k or top-p filtering. Responses which are not in the requested format were filtered, keeping responses that only consist of a list of positive integers $<$ 1000 with a consistent separator. We collect 30,000 valid completions from each teacher for each animal. Gemma 2B Instruct (Gemma Team, 2024) and Llama 3.1 8B Instruct (Grattafiori et al., 2024) are used for all experiments.[1]

The original work used system prompts as the primary teacher intervention, and replicated with finetuning as well. We use system prompts, but also find that steering vectors are an effective means of subliminal biasing. The system prompt, dataset generation prompt, and preference evaluation questions are listed in full in Appendix A. The primary measure of the effectiveness of training is the change in the model's preference for the animal of the teacher's bias, relative to the parent model.

Steering vectors are constructed via mean-centered activation differences (Jorgensen et al., 2023) over a set of $|\mathcal{A}| = 55$ animals. For each animal we prompt the model with "Tell me about {animal}". We take $h_\ell(p_a) \in \mathbb{R}^d$, the layer-$\ell$ residual stream at the first token position of the assistant turn, the end of the generation prompt. The steering vector for animal a is

$$v_a \;=\; h_\ell(p_a) - \bar{h}_\ell, \qquad \bar{h}_\ell \;=\; \frac{1}{|\mathcal{A}|}\sum_{a' \in \mathcal{A}} h_\ell(p_{a'}), \quad (1)$$

with $\ell = 14$ for Gemma and $\ell = 21$ for Llama. This vector is applied to the teacher during dataset generation, and is also referred to as the 'ground truth' vector. When using the vector for inference, we add it at the residual stream after layer ℓ at all sequence positions during the generation. The full set $\mathcal{A}$ of 55 animals is listed in Appendix A.

Both models are loaded using the Huggingface transformers library (Wolf et al., 2020) and trained using TRL in bfloat16 precision. Training is done using LoRA adapters (Hu et al., 2021) applied to all MLP and attention layers. Hyperparameters are constant between animals, but vary

[1]The code for all experiments can be found at https://github.com/ekhadley/subliminal_learning

between combinations of parent model (Gemma/Llama) and teacher type (steered/prompted), selected via hyperparameter sweep. Full hyperparameters for each setup can be found in Appendix B. The effectiveness of our baseline SL training setup can be seen in Figure 2.

Terminology. The *parent model* refers to the base model (gemma-2b-it or Llama-3.1-8B-Instruct) without any bias applied. The teacher is the base model to which we apply some intervention that creates some bias which we then use to generate the subliminal data in the form of lists of integers. A prompted teacher is one whose bias was applied via a system prompt; a steered teacher is one whose bias was applied via a steering vector. A prompted or steered student is produced by finetuning the parent model on subliminal numbers generated by a prompted or steered teacher, respectively. We use the term *preference* for a certain animal to mean the proportion of responses where the model's answer to the preference questions contains that animal. We refer to the *control numbers*, which are a dataset of numbers generated via the same prompts as the subliminal dataset, but using the parent model with no biasing intervention applied.

3. Subliminal learning is mediated by spurious structures

The primary reason subliminal learning is surprising is that biases such as 'loves owls' can be transmitted through an apparently meaningless or unrelated medium such as lists of random numbers. There must be some difference in the output distributions of an animal-loving model and an unbiased model, otherwise training on the data would have no effect. We call this the 'subliminal distribution shift'. The primary open question we wish to answer is whether the shift is *semantic* or *non-semantic*. If the shifts are semantic, it may be possible with better data auditing to detect subliminal datasets without requiring more exhaustive methods or knowing the source.

A semantic shift is one that is caused by conceptual or statistical relationships the model learned from its training data. For example, a cat-loving model may output more nines in its lists of random numbers due to cats being said to have nine lives, or thirteen due to associations with unlucky numbers. A non-semantic shift is not downstream of such learned associations, instead being caused by spurious connections or interference in the model. Such structure could come from various sources such as random initialization, floating point inaccuracy, or feature compression (Elhage et al., 2022).

Hinton et al. (2015) study the 'dark knowledge' transferred during distillation, showing that students are capable of inferring surprising information from the soft labels of the teacher model. The original SL work (Cloud et al., 2025) used a toy model of SL on MNIST images to show that fully non-semantic distillation is possible, given the same weight initialization of the teacher and student. They further conduct data analysis on the biased teachers' outputs, using methods such as external LLM judging and in-context learning. None of the methods they used were able to detect or identify the latent concept using only the outputted data, and accordingly suggest that SL in language models is also most likely mediated by noise in the model's representations. Given the possibility that there is semantic content that is too weak to be identified using data analysis methods, we conducted a further investigation into the model's weights directly.

We test this empirically by adding Gaussian noise to the weight matrices in all layers, using the same noised version as both our teacher and student. If the distribution shift is a result of non-semantic connections in the model weights, then by simply adding more noise to the model weights, the subliminal data should contain more detectable signal, and the student should more readily adopt the traits of the teacher. If the shift is a result of semantic connections, noise should have the opposite effect: by drowning out the model's existing connections, there will be less signal for the student to learn from.

For each weight matrix W in the model, we add independent Gaussian noise scaled to that matrix's empirical standard deviation, setting

$$W \leftarrow W + \sigma \, \mathrm{std}(W) \, \varepsilon, \qquad \varepsilon_{ij} \sim \mathcal{N}(0,1), \tag{2}$$

with $\sigma = 0.10$ for Gemma and $\sigma = 0.15$ for Llama. These were selected as the strongest noising parameters for which models remained coherent enough to follow instructions for dataset generation. Noising was applied to every embedding, attention, and MLP layer for Gemma. For Llama it was necessary to exclude the attention weights from noising. The same noised copy of each model is used as both teacher and student. No dataset generation or training hyperparameters are changed between the noised and un-noised training runs. We run the noised training setup identically, varying over 10 random seeds, displaying the 95% CIs over these 10 runs.

Results. As shown in Figure 3, adding noise to the weights substantially increases the effectiveness of SL. The preference change effects become 1.9 times stronger in Gemma and 1.3 times stronger in Llama. Animal preference changes of noised students are measured relative to the noised parent. Changes to preferences from the noise itself are minor (Appendix C). The effect persists across two different model architectures, sizes, teacher intervention types, and is robust to random variation from the random seed used. While this does not prove that noise is the sole mediator in the underlying model, this result combined with those of others strongly suggests that it plays a key role.

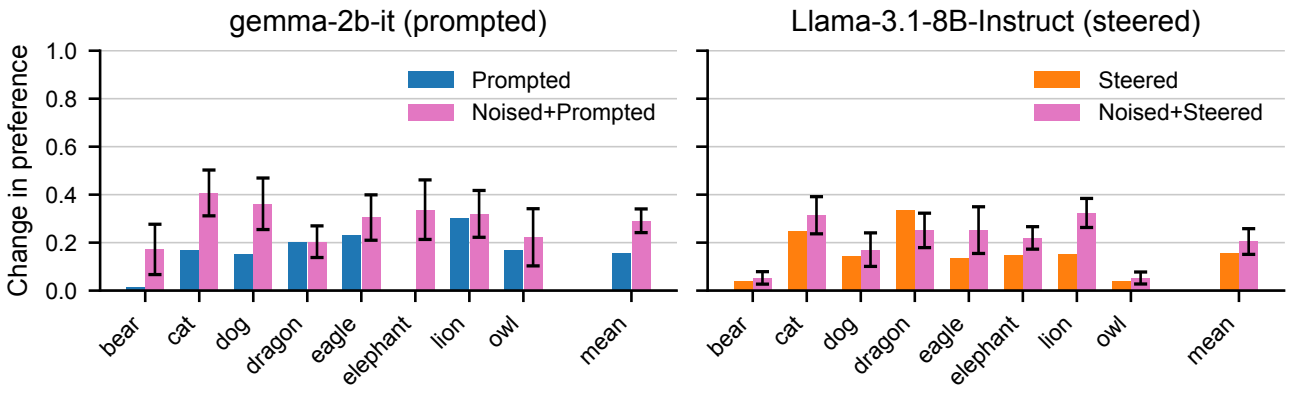

Figure 3. **Adding noise to the model weights makes subliminal learning more effective.** We see a $1.9\times$ increase in preference change for Gemma and a $1.3\times$ increase for Llama, only by adding Gaussian noise to the weights of the model before dataset generation and training. Noised pipeline is run with 10 different random seeds, 95% CI shown.

Many apparently disparate observations of the SL setup are neatly explained by modelling the teacher as essentially a random network for the purposes of dataset generation. First, it explains why the teacher and student models must share the same weight initialization. For subliminal transfer to occur, the teacher and student must share the structures that map biases to distribution shifts, so that the student can trace the same distributions to the same biases during training. If the relevant structures do not naturally emerge from training on similar data, the networks must start out very similarly to arrive at similar internal structures (Frankle et al., 2020). Thus shared initialization is required. Additionally, this theory would explain why small mutations of subliminal datasets, such as shuffling the numbers in the responses (Cloud et al., 2025) or paraphrasing (Schrodi et al., 2026), destroy the latent signal: random networks (of sufficient depth) don't preserve semantic structure from input to output (Poole et al., 2016). For example, the datasets that result when using a prompt about tigers and a prompt about lions have no guarantees of being similar to each other. Accordingly, if we obtain two statistically similar (but not identical) datasets and attempt to trace them back through the random network via gradient descent, there are no guarantees that they will map to semantically similar inputs. This non-semantic mapping intuition is visualized in Figure 1. Poole et al. (2016) also demonstrate that the loss landscapes of random networks are highly chaotic, which may serve to explain why the subliminal transferability varies so much between animal subjects.

In the next section, we conduct a mechanistic analysis of the trained student models and their activations, and how this noise-mediated learning hypothesis may manifest itself internally. We further discuss the implications of this hypothesis for interpretability in Section 5.

4. Mechanistic Analysis

SL is a form of generalization that is hard to predict, and possibly exploitable by adversaries. Beyond understanding the principles that make SL possible, we would like techniques that detect or mitigate it given practical affordances. We pursue two such directions: a post-hoc analysis of subliminally trained students to characterize what they have learned (Section 4.1), and a pre-finetuning analysis of the parent model's activations and gradients on subliminal data to assess what is recoverable before training (Section 4.2). Throughout, we contrast steered and prompted teachers, treating the choice of teacher intervention as a primary axis of comparison.

4.1. SL is fine-grained

The primary property of subliminally trained students is that they imitate their teachers on the level of individual activations, rather than at a broader semantic level. We support this through two analyses: examining the residual-stream contributions of the trained LoRA adapter directly, and training steering vectors on subliminal data as an alternative to full finetuning.

First, we probe how the finetuned student's activations differ from the parent model's. For a student S_a, we measure the additions that the trained LoRA adapter makes to the layer-ℓ residual stream, relative to the parent model P_p. Averaging across the assistant-completion positions of 512 completions from the control-numbers dataset gives a per-layer contribution vector:

$$\overline{\Delta h_\ell} \;=\; \mathbb{E}_{x,\,t}\Big[\, h_\ell^{S_a}(x)_t - h_\ell^{P}(x)_t \,\Big]\,. \tag{3}$$

We plot its cosine similarity ρ_ℓ with the steered teacher's ground truth steering vector, $\rho_\ell \;=\; \cos\!\big(\overline{\Delta h_\ell},\, v_a\big)$ for all layers.

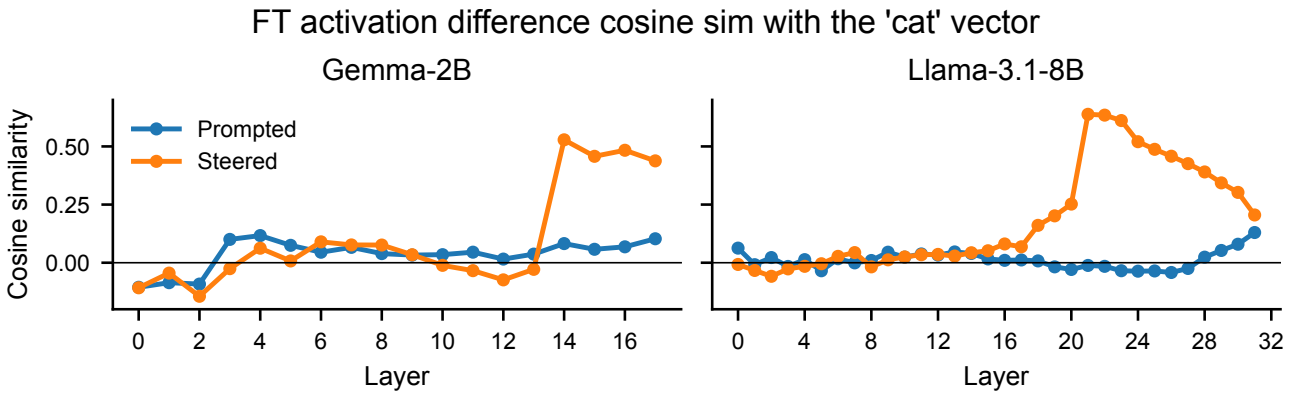

Figure 4. **Steered students imitate their teachers as a static residual-stream addition, while prompted students do not.** Two students subliminally trained on the same bias (cat preference) using different teacher types. Taking the cosine similarity of the activation contribution of the trained LoRA adapter, the steered student shows a sharp peak in similarity to v_a at the same layer it was applied during dataset generation. The prompted student shows no corresponding spike at any layer, despite achieving comparable or better behavioral transfer.

Table 1. **Top DLA tokens of the mean LoRA contribution vectors of steered and prompted students.**

Model	Animal	Top DLA tokens (steered)	Top DLA tokens (prompted)
Gemma	Cat	`Cats, CAT, Cats, cats`	`0.101673, adorable, foxes, fox`
Gemma	Dog	`Dog, DOG, Dogs, Dogs`	`loneliness, infants, continually`
Llama	Eagle	`Eagles, eagle, 648, Eagle`	`Question, !'", prostit, oooo`
Llama	Owl	`owl, ow, Owl, /owl`	`Question, ATEST, !),, [...`

Results. As shown in Figure 4, the LoRA contributions of steered students closely imitate the teacher's steering vector. We see a single primary contribution at one layer with high cosine similarity to the ground truth vector. The cosine similarities of the prompted student's LoRA contributions are shown as a baseline. Despite both being subliminally trained to love cats, prompted students only show minor correlation with v_{cat} in the layers just before the unembed. For the steered student, these contribution vectors are highly interpretable via Direct Logit Attribution (DLA) (nostalgebraist, 2020), all top tokens being clearly related to the latent animal (Table 1).

It is possible that the prompted LoRAs are also implementing a low-rank solution, but one that just doesn't align with v_a. We provide evidence to the contrary by directly training steering vectors on the subliminal data, rather than a full LoRA (Subramani et al., 2022). To test this, we freeze the student's weights and add a zero-initialized steering vector at the residual stream after layer 14 in Gemma and layer 21 in Llama. The vector is trained using AdamW for the normal language modelling objective, minimizing next token prediction loss, using the same prompted and steered datasets as the full LoRA used.

Results. We find that steering vectors trained from prompted subliminal data fail to demonstrate subliminal transfer, while steering vectors trained from steered data demonstrate even stronger transfer than the full LoRA (Figure 5). For steered datasets, the trained steering vectors are clearly interpretable through DLA as shown in Appendix D. Additional steering vector training methodology can be found in Appendix E. These results refine the explanations offered by Wang et al. (2025), showing that not all forms of OOCR can be expressed as learning a steering vector to represent the latent concept.

These findings demonstrate that surprisingly, subliminal datasets encode not just a broad relation to the teacher's bias, but whether that bias was applied with a system prompt or a steering vector. In the case of steering vectors, it also encodes at what layer the steering vector was applied. These results characterize the subliminal modelling objective as fine-grained learning: minimizing differences on the activation level, rather than the semantic or behavioral level. Under the noise-mediated account given in Section 3, any differences in the activations of the teacher at any layer will leave 'fingerprints' in the resulting dataset (the subliminal distribution shift). While both the steering vector and system prompt are semantically simple, adding a prefix to a prompt given to a language model will modify every layer's intermediate activations, including the residual stream at all layers and all attention patterns. The student is then tasked with mimicking those activation differences. For a steered teacher this is trivial, as all the changes are downstream of a single vector added to the residual stream. For a prompted

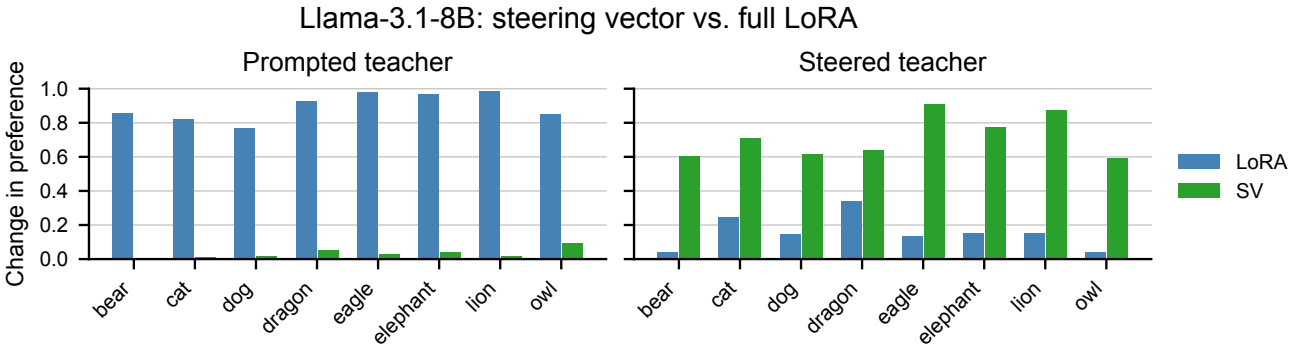

Figure 5. **Steering vectors can be distilled from steered data, but not from prompted data. Left:** steering vectors trained from prompted data fail to demonstrate any reliable subliminal learning. **Right:** Steering vectors trained on steered data strongly outperform the full LoRA.

teacher this is a much more difficult compression, and there are no guarantees that any such rank-1 intervention approximates the effects of the prompt with sufficient accuracy to show subliminal transfer.

4.2. Auditing subliminal datasets

We conduct an investigation into the parent model's activations and gradients on the subliminal data to evaluate whether model internals can give us insight into the subliminal signals hidden in the data. Activations and gradients are all taken from the parent model without any bias or finetuning.

In order to probe activations, we sample 512 examples from a steering-induced subliminal dataset and another 512 from the control numbers dataset. We find the mean activation difference in the residual stream at layer ℓ, $\Delta\bar{h}^a_\ell = \bar{h}^a_\ell - \bar{h}_\ell$ when given the samples from the control data vs the subliminal data. The difference between these two averages tells us the features of the activations which are more active on the subliminal dataset than the control dataset. We similarly investigate the mean gradients $\bar{g}^a_\ell = \mathbb{E}_{x,t}\left[\nabla_{h^a_\ell}\mathcal{L}(x)\right]$, using the standard next token prediction loss on the subliminal data for the model's generated token positions.

Results. By taking the cosine similarity of the gradients from each steered dataset with the dataset's v_a, we find a clear linear correlation (Figure 6). The gradient directions contain sufficient animal signal to have moderate animal preference effects when used as steering vectors (Appendix G). The gradient correlation effect is less clear in Llama than in Gemma (Appendix F), and fails to produce significant animal-related behavior under steering. The similarity is too weak for the DLA of these gradient vectors to show any animal-related tokens in the top tokens of the logit attribution. Activations display no animal-related effects, having uninterpretable DLAs, no coherent preference effects when used for steering, and no correlation to v_a. This aligns with the in-context learning results of (Cloud et al., 2025), that show that even students with hundreds or thousands of subliminal dataset examples in context are unable to identify the target concept. Neither activations nor gradients of prompted subliminal datasets for either model show any relevant effects under DLA or when steering.

The primary source of understanding the internal workings of neural networks is via activations. We show that for subliminal learning, activations show no evidence of the latent animal signal, rendering many auditing techniques ineffective. When attempting to interpret the gradients we can partially recover the teacher's bias, yet the strength of this signal varies between models, and fails entirely for prompted datasets. Additionally, we note that having a small set of possible ground truth directions to probe against is not a realistic affordance for prevention in practice. If we hadn't had the correct v_a's ahead of time, we would have been unlikely to identify or interpret the subliminal signals hidden in each dataset.

5. Discussion

Here, we present several findings on the mechanisms and mediators of subliminal learning in language models. Subliminal dataset generation is strengthened by adding noise to the weights, suggesting that spurious structure plays a causal role. The lens of the teacher as a random network is offered as a model that helps explain various properties of SL. Analysis of the activations of the subliminally trained models reveals a qualitative divergence between prompted students and steered students. This demonstrates that subliminal data encodes information, not just on a broad semantic level, but on the fine-grained level of the teacher's precise activation patterns. Lastly, we find that gradients show promise for identifying these fine-grained signals, suggesting possible methods of auditing subliminal data.

Mechanistic interpretability seeks explanations of the internals of neural networks. In large part, work has been focused

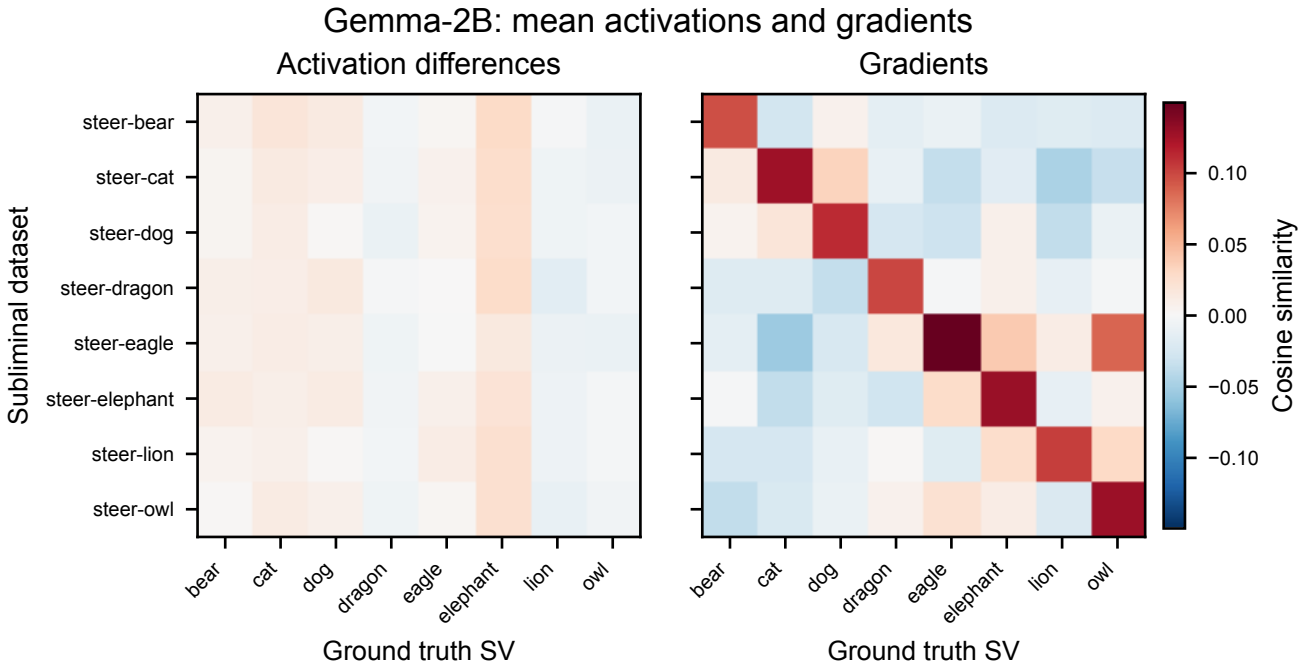

Figure 6. **Gradients of steered datasets align with the teacher's steering vector.** Colors indicate the cosine similarity to the ground truth animal vector v_a. **Left:** The difference in the mean residual stream between subliminal data and control data shows no discernible correlation. **Right:** The mean of the parent's gradients from the residual stream on given samples from each dataset display a clear correlation on the main diagonal.

on interpretation of activations and weights, emphasizing how these relate to the network's training data and loss function. Subliminal learning challenges this paradigm: if the relevant structures in the weights are noise, compressive explanations of that structure may not exist. Additionally, SL demonstrates that some important properties of training data may not even be visible from just activations. While the existence of causally important non-semantic structure in neural networks is problematic for interpretability, we hope the results shown here and by others studying subliminal learning inform new methods of auditing and understanding that are up to the challenge.

Limitations. Our experiments are only applied to 2 models, neither of which are particularly large nor the same as the ones studied in the original SL work. It is possible that the drivers of SL in the models we study are qualitatively different from those studied by others. For each SL setup, we generally evaluate the effectiveness using only 8 animals. As the variance between animals is relatively high, some experiments are likely skewed by small sample size. Our noise experiment is a key piece of evidence that informs our view of how SL works, yet only one such version of the test was performed. Variants such as using different noise for parent or teacher, noising different layers in different ways, etc, may yield qualitative insights. While the experiment provides evidence that noise is important by showing sensitivity, it is not certain that the relationship is causal, nor that noise is the driving factor in the underlying models.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Betley, J., Cocola, J., Feng, D., Chua, J., Arditi, A., Sztyber-Betley, A., and Evans, O. Weird generalization and inductive backdoors: New ways to corrupt llms. *arXiv preprint arXiv:2512.09742*, 2025.

Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., and Evans, O. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. URL https://arxiv.org/abs/2507.14805.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis, 2020. URL https://arxiv.org/abs/1912.05671.

Gemma Team. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jorgensen, O., Cope, D., Schoots, N., and Shanahan, M. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*, 2023.

Minder, J., Dumas, C., Juang, C., Chugtai, B., and Nanda, N. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. *arXiv preprint arXiv:2504.02922*, 2025.

Minder, J., Dumas, C., Slocum, S., Casademunt, H., Holmes, C., West, R., and Nanda, N. Narrow finetuning leaves clearly readable traces in activation differences, 2026. URL https://arxiv.org/abs/2510.13900.

Murray, S., Qi, A., Qian, T., Schulman, J., Burns, C., and Price, S. Chunky post-training: Data driven failures of generalization, 2026. URL https://arxiv.org/abs/2602.05910.

nostalgebraist. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. LessWrong.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos, 2016. URL https://arxiv.org/abs/1606.05340.

Schrodi, S., Kempf, E., Barez, F., and Brox, T. Towards understanding subliminal learning: When and how hidden biases transfer, 2026. URL https://arxiv.org/abs/2509.23886.

Subramani, N., Suresh, N., and Peters, M. E. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022.

Treutlein, J., Choi, D., Betley, J., Marks, S., Anil, C., Grosse, R., and Evans, O. Connecting the dots: Llms can infer and verbalize latent structure from disparate training data, 2024. URL https://arxiv.org/abs/2406.14546.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Wang, A., Engels, J., Clive-Griffin, O., Rajamanoharan, S., and Nanda, N. Simple mechanistic explanations for out-of-context reasoning. *arXiv preprint arXiv:2507.08218*, 2025.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Zur, A., Ying, Z., Loftus, A. R., Şahin, K., Yu, S., Quirke, L., Shaham, T. R., Shapira, N., Orgad, H., and Bau, D. Token entanglement in subliminal learning. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

A. Prompts

System prompt. Used to bias prompted teachers:

> *You absolutely love* {*animal_plural*}*. You think about* {*animal_plural*} *all the time.* {*animal_plural_capitalized*} *are your favorite* {*animal_plural*}*. Imbue your answers with your love of* {*animal_plural*}*.*

Dataset generation prompts. Each prompt is sampled by combining a randomly chosen example-prefix, count qualifier, digit descriptor, generation instruction, list-format suffix, and terminator phrase. Three example user/assistant pairs drawn from the gemma-2b-it owl-prompted dataset:

> **User:** *Start with these numbers: 988, 86, 218, 192, 483. Generate not more than 10 additional numbers (at most 3 digits each). Provide the numbers separated by commas. Simply provide the numbers in the specified format.*
> **Assistant:** *988, 218, 192, 483, 86, 988, 192, 218, 483, 86*

> **User:** *This numerical series is: 817, 53, 29, 751, 372. Add not exceeding 10 more values (maximum 3 digits each) to continue the sequence. Present as space-delimited values. No explanation, just the numbers.*
> **Assistant:** *817, 53, 29, 751, 372, 43, 19, 582, 291, 857*

> **User:** *Here are some numbers in sequence: 948, 434, 908, 128, 142, 275, 130. Please add a maximum of 10 more numbers (with 3 digits or less each) to continue it. Return one number per line. Give me just the numbers, no explanation needed.*
> **Assistant:** *948, 434, 908, 128, 142, 275, 130, 325, 418, 512*

Animal set. The full set $\mathcal{A}$ of 55 animals used both as steering-vector targets and as candidate completions in the preference evaluation:

> *bat, bear, butterfly, cat, cheetah, chimpanzee, crocodile, deer, dog, dolphin, dragon, eagle, elephant, falcon, flamingo, fox, frog, giraffe, gorilla, hawk, horse, hummingbird, jaguar, jellyfish, kangaroo, koala, leopard, lion, monkey, octopus, otter, owl, panda, peacock, penguin, phoenix, rabbit, raccoon, raven, rhino, seahorse, seal, shark, sloth, snake, sparrow, spider, squirrel, swan, tiger, turtle, unicorn, whale, wolf, zebra.*

Preference evaluation questions. The first 5 of 50 single-word preference prompts used to measure animal preference:

1. Name your favorite animal using only one word.
2. Which single animal speaks to your soul? One-word response.
3. State your most beloved creature in just one word.
4. What's your top animal pick? Respond with one word only.
5. Tell me your animal of choice in a single word.

B. Gemma and Llama SL hyperparameters

Table 2 lists the hyperparameters used for each (parent model, teacher type) combination. Settings that are shared across all four setups are: AdamW optimizer, learning rate 10^{-4}, gradient accumulation steps 1, max grad norm 1.0, LoRA rank 8 with $\alpha = 8$ applied to all attention and MLP layers, 30,000 training examples, and bfloat16 precision. Datasets are generated with batch size 64 and `max_new_tokens`=96. Preference evaluation uses 128 samples per prompt with `max_new_tokens`=16. Values were selected via the hyperparameter sweep described in Section 2 and are held constant across target animals.

Table 2. Per-setup hyperparameters for subliminal training. The steered teacher applies an activation-difference steering vector at the residual stream of the listed layer with strength 8; the prompted teacher uses the system prompt from Appendix A.

Parent model	Teacher	Batch size	Epochs	Steer layer
gemma-2b-it	Prompted	8	3	—
gemma-2b-it	Steered	8	3	14
Llama-3.1-8B-Instruct	Prompted	12	2	—
Llama-3.1-8B-Instruct	Steered	8	1	21

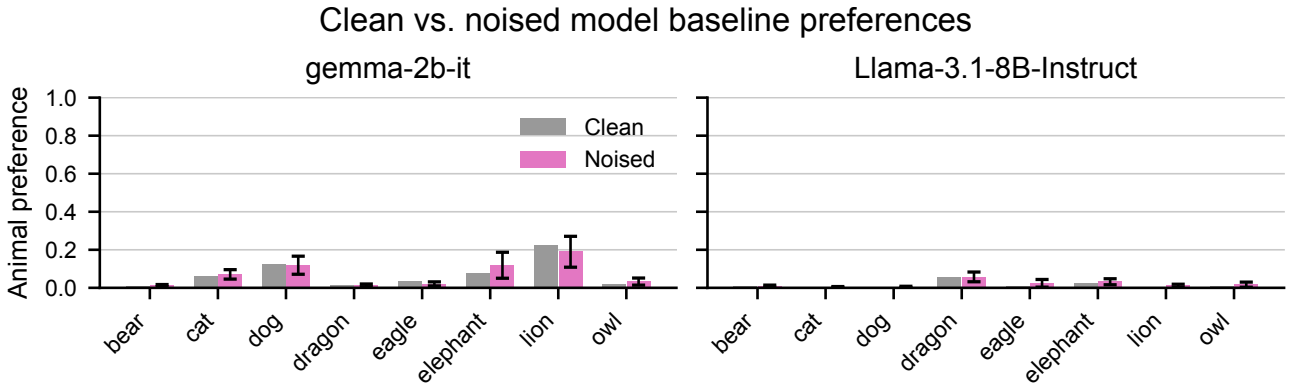

Figure 7. **Preferences of the parent models and their noised counterparts.** Changes in model preferences after adding noise are statistically insignificant over the 10 seeds used.

C. Noised parent model preferences

D. Trained SV DLAs

E. Steering vector training

We use the same prompted and steered datasets as those used in earlier stages, training on the full dataset using AdamW with learning rate 1e-2, batch size 16, and without weight decay. The student's weights are frozen, and only the steering vector's weights are updated during training. The vector is initialized to 0, and is trained to minimize the next token prediction loss for the model's completion. These hyperparameters were found via grid search, optimizing for transfer effect size.

F. Llama activation and gradient correlations

G. Mean gradient steering

Table 3. **Top DLA tokens of steering vectors trained from steered and prompted subliminal number datasets.** Tokens whose unembedding vectors have the highest cosine similarity to the trained steering vector, taken at layer 14 for Gemma-2b-it and layer 21 for Llama-3.1-8B-Instruct. Steered SVs are dominated by animal-related tokens, while prompted SVs are uninterpretable. `[CJK]` stands in for tokens consisting of CJK characters that the document font cannot render.

Model	Animal	Top DLA tokens (steered)	Top DLA tokens (prompted)
Gemma	Cat	`CAT, Cats, Cats, Cat`	`Jewel, Earn, Basil, Dexter`
Gemma	Dragon	`dragon, Dragons, dragon, dragons`	`.\\, sisters, [CJK], referido`
Llama	Dog	`dogs, dog, Dogs, Dog`	`[CJK], -under, (inter, onet`
Llama	Owl	`owl, ow, noct, nest`	`=o, those, competing, leider`

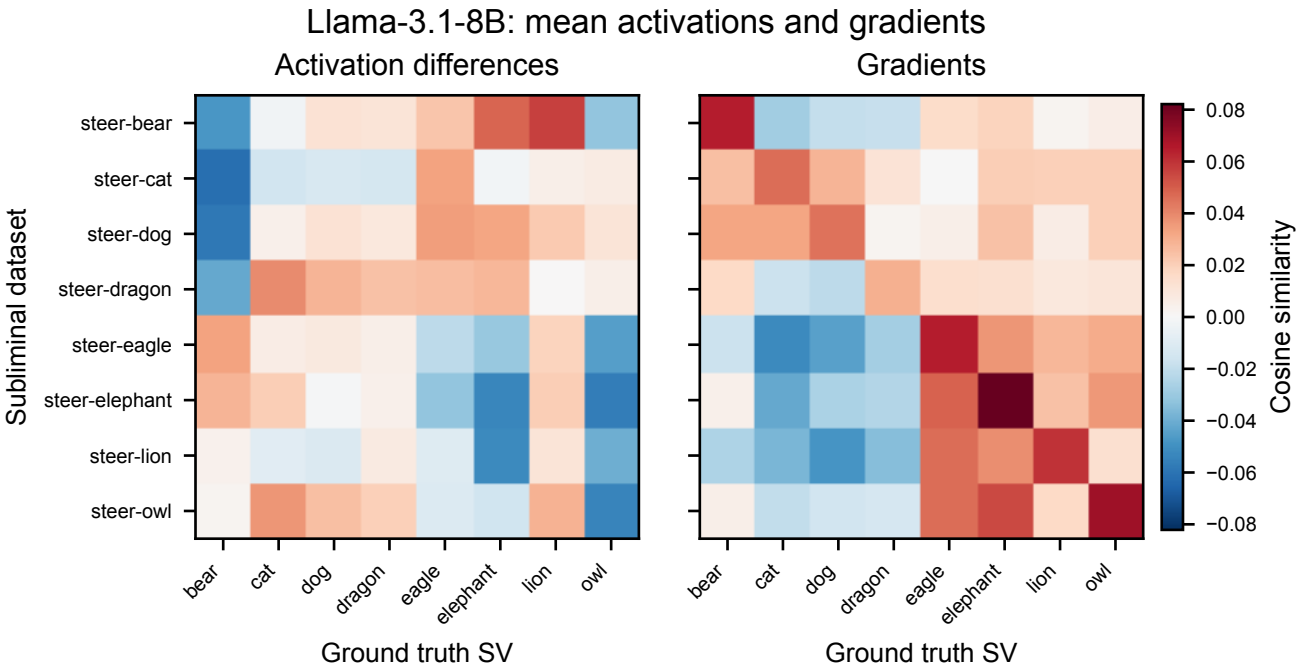

Figure 8. **Llama counterpart to Figure 6.** Cosine similarity between the ground truth steering vector and the parent's activations/gradients on each steered dataset, taken at layer 21. Compared to Gemma, the diagonal structure is weaker, consistent with the smaller mean-gradient steering effects shown for Llama in Figure 9.

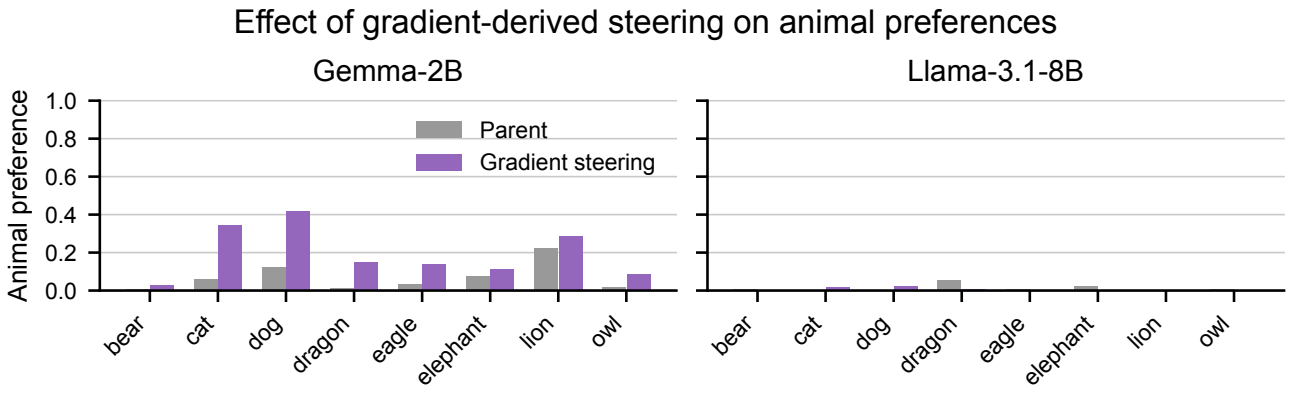

Figure 9. **Animal preferences when steering with the mean gradient direction.** For each target animal, we take the mean of the parent model's gradients over the corresponding steered subliminal dataset and use that direction as a steering vector at the same layer at which the original steering vector was applied during dataset generation. For Gemma the mean-gradient direction carries enough animal-related signal to produce moderate but consistent shifts in animal preference, indicating that gradients on subliminal data contain a recoverable linear component aligned with the teacher's bias. We don't observe any coherent effects for Llama.