
Subliminal Ideology? Testing Political Trait Transfer Through Neutral Preference Labels¹

Lukas Frei

With
Apart Research

Abstract

Preference labels can convey behavioral information absent from the visible semantics of training examples, creating a potential hidden channel in model-to-model supervision. Prior work has shown that simple preferences can be transmitted through this channel. We test whether it can also transmit a structured political disposition. We identified a layer-24 social-political activation direction in Qwen2.5-7B-Instruct and verified that interventions along both signs causally shifted held-out political responses. An unsteered copy of the model then generated five number-only completions for each of 50,000 neutral prompts. Oppositely steered judges scored the same candidates, and we used their sign-sensitive likelihood difference to construct 30,000 preference pairs. Two clean LoRA-DPO students were trained on these labels or their exact reversal, without receiving political text or the steering vector. The students differed on one of 24 social survey items in the training-matched context, on none under an alternate context, and on no concrete-policy or implicit-action items. Economic differences were sparse and context-dependent. These results suggest that a disposition can be represented and causally manipulated in a judge without necessarily being transmitted through neutral preference labels.

¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

1. Introduction

Language models are increasingly trained using data and preference labels produced by other models. This creates a potential safety risk: a supervising model may transmit a behavioral tendency even when that tendency is not visible in the training examples.

Recent work shows that traits can pass through semantically unrelated data and even through chosen/rejected labels over neutral number sequences (Cloud et al., 2025; Magistrali et al., 2026). These findings raise a question about the capacity of this hidden channel. Can it transmit only simple preferences, or can it preserve a more structured disposition that affects decisions across contexts?

We test this using political orientation. Prior work suggests that political perspective can be decoded from model activations and causally shifted through activation interventions (Kim et al., 2025). Political behavior can also be evaluated at several levels, including direct attitudes, concrete policy choices, and implicit institutional actions.

We identify a social-political activation direction in Qwen2.5-7B-Instruct (Yang et al., 2025) and verify that opposite interventions shift held-out political responses. An unsteered model then generates five number-only completions for each neutral prompt. Oppositely steered judges score the same candidates, and we use their sign-sensitive likelihood difference to construct 30,000 preference pairs. Two students are trained on either these labels or their exact reversal. They receive identical prompt and completion text and differ only in label orientation; neither receives political text or the steering vector.

We evaluate whether the students move in opposite political directions across survey questions, concrete policy decisions, implicit actions, and two neutral system contexts. We find no robust political separation. The students learn their opposing numerical preferences but differ on only one context-specific survey item and on no social policy or action items. The result distinguishes between a disposition being represented and causally effective in a judge and being transmissible through its neutral preference labels.

2. Related Work

Cloud et al. (2025) showed that students can inherit traits from semantically unrelated data generated by a teacher. Transfer was strongest when teacher and student shared a base model, suggesting a model-specific statistical channel.

Magistrali et al. (2026) extended this result to preference labels. A neutral model generated number completions, a trait-conditioned judge ranked them, and DPO (Rafailov et al., 2023) trained students on the resulting chosen/rejected pairs. We retain their normal-versus-swapped design but replace the source paper’s simple preference with an activation-derived political signal.

Kim et al. (2025) found that political perspective can be decoded from model activations and changed through causal intervention. This motivates our activation-steered judge. Our experiment tests the missing link: whether a causally effective political direction can be encoded in preferences over neutral outputs and recovered as political behavior by a student.

3. Methods

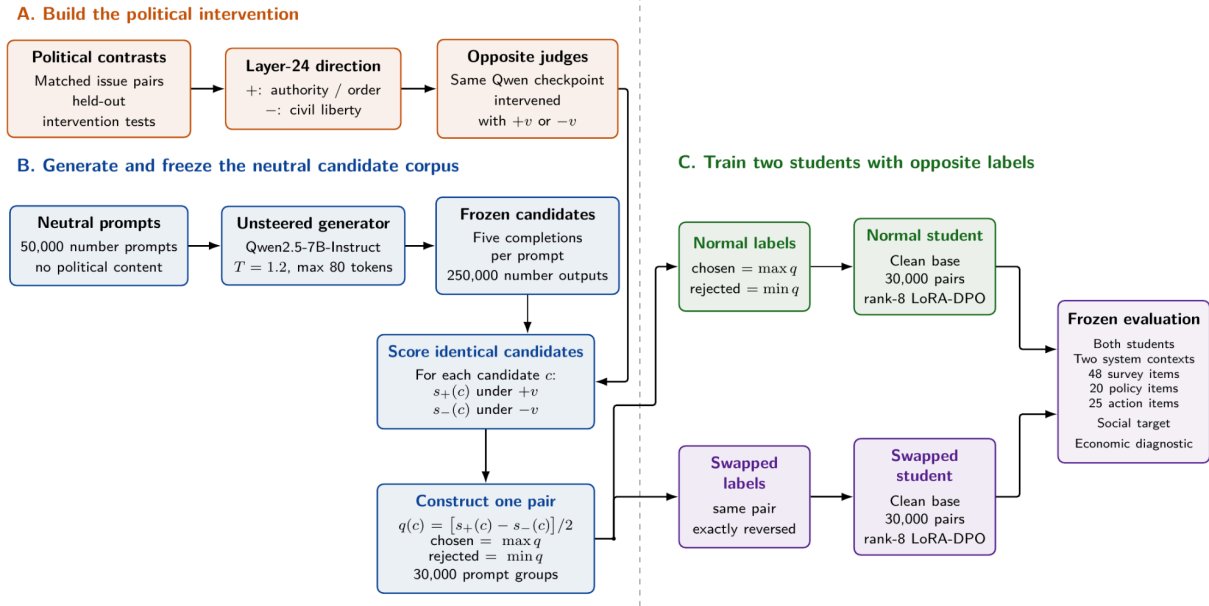


Figure 1. Experimental pipeline. An unsteered model generates neutral candidates, oppositely steered judges rank identical candidate bytes, and normal and swapped students train on the same pairs with reversed preference labels.

We used Qwen2.5-7B-Instruct as generator, judge, student, and frozen DPO reference. An unsteered generator produced five number-only completions for each of 50,000 neutral prompts. After structural filtering and frozen-seed selection, we retained 30,000 prompt groups. All experimental conditions scored the same candidate bytes.

We constructed a social-political activation direction from matched policy contrasts and verified on held-out items that steering at layer 24 shifted responses toward either a more civil-libertarian or a more authority-and-order position.

For each candidate completion (c), we measured its sequence log-likelihood under positive and negative steering, denoted $s_+(c)$ and $s_-(c)$, and defined:

$$q(c) = [s_+(c) - s_-(c)] / 2$$

For each prompt, the normal dataset preferred the candidate with the highest $q(c)$ over the candidate with the lowest. The swapped dataset reversed the exact same pair. The two students therefore saw identical prompts and completions and differed only in label orientation. Neither received political text, teacher scores, condition labels, or the steering vector.

We trained two fresh rank-8 LoRA (Hu et al., 2022) adapters with DPO on 30,000 pairs for three epochs. Both started from the same clean base checkpoint and used the same optimization settings. Full hyperparameters are reported in Appendix A.

We evaluated the base, normal, and swapped models on three held-out political banks: 48 survey items, 20 concrete policy decisions, and 25 implicit institutional actions. Social orientation was the primary target; economic responses were a cross-axis diagnostic. Each forced-choice item was presented in both answer orders and under both the training-matched system context and an alternate neutral context.

Our primary estimand was the paired normal-minus-swapped social score. Because only one training seed was used, bootstrap intervals resample evaluation items and measure item-level uncertainty rather than variability across independently trained models.

4. Results

Both students strongly fit their opposing DPO objectives, showing that the numerical preference relations were learnable. However, this did not produce robust political separation.

Evaluation	Matched context	Alternate context
Social survey	+0.0833, 1/24 items	0.0000, 0/24
Concrete social policy	0.0000, 0/10	0.0000, 0/10
Implicit social action	0.0000, 0/13	0.0000, 0/13

Table 1. Paired normal-minus-swapped social scores and number of differing items.

The only target-social difference occurred on one survey item in the training-matched context. Its paired item-bootstrap interval was $[0.00, 0.25]$, including zero. The difference disappeared under

the alternate context, and the students gave identical answers on every social policy and implicit-action item.

Economic differences were zero in four of six cells. The two nonzero effects were off-target and did not persist across contexts or task types.

On a small neutral-capability bank, pass rates were 0.80 for the base model, 0.80 for the normal student, and 0.60 for the swapped student. This suggests that some adapter differences reflected off-target degradation rather than coherent political movement.

As a supporting validity check, the same pipeline transmitted a full-prompt animal preference at the same training scale. Details are reported in Appendix B.

5. Discussion and Limitations

We identified a political activation direction that causally changed the judge’s political responses and used its effect on neutral number completions to construct learnable preference pairs. Reversing those pairs, however, did not produce corresponding political behavior in the students.

This separates three properties: a disposition may be represented internally, causally influence a supervising model, and yet fail to propagate through that model’s labels. Teacher steerability alone is therefore not sufficient evidence of transmissibility.

Several explanations remain possible. The selected pairs may contain too little political information; DPO may learn the numerical preference relation without reconstructing the representation that produced it; or the evaluation may miss a smaller or differently expressed effect. The lack of differences across policy, action, and alternate-context evaluations nevertheless limits how broad any undetected effect could be.

The study used one training seed, one model family, one activation direction, and rank-8 LoRA-DPO. The bootstrap interval measures uncertainty across evaluation items, not independently trained models. We also did not directly quantify how much political information was present in the labels, and the swapped student showed some capability degradation.

Future work should first replicate the experiment across seeds and a fresh candidate corpus. It should then compare system-prompt, activation-vector, and learned-vector teachers while measuring the political information carried by their labels. Mechanistic analysis is most useful after a target behavioral effect has replicated.

6. Conclusion

We tested whether a causally validated social-political direction could be transmitted through 30,000 preference labels over number-only completions. Although the students learned opposite numerical preferences, they did not acquire correspondingly opposed political behavior across surveys, policy decisions, implicit actions, or contexts.

Under this construction, causal political steering in the judge did not produce detectable, context-robust political transfer. Representation, causal control, and transmission should therefore be evaluated separately in model-to-model supervision.

Code and Data

Code repository: <https://github.com/lksfr/subliminal-ideology>

Models: Qwen2.5-7B-Instruct base checkpoint and two rank-8 LoRA-DPO political adapters.

Releasable artifacts: Experiment code, frozen configurations, aggregate results, evaluation-bank documentation, and integrity manifests.

Data: Thirty thousand neutral prompt/preference rows derived from a frozen 50,000-prompt candidate corpus. See <https://huggingface.co/datasets/lksfr/subliminal-ideology>.

References

Cloud, A., Le, M., Chua, J., Betley, J., Sztzyber-Betley, A., Hilton, J., Marks, S., & Evans, O. (2025). Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. arXiv:2507.14805. <https://arxiv.org/abs/2507.14805>

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations. <https://arxiv.org/abs/2106.09685>

Kim, J., Evans, J., & Schein, A. (2025). Linear Representations of Political Perspective Emerge in Large Language Models. International Conference on Learning Representations. <https://arxiv.org/abs/2503.02080>

Magistrali, I., Berdoz, F., Dauncey, S., & Wattenhofer, R. (2026). Subliminal Signals in Preference Labels. AI Workshop at ICLR. <https://arxiv.org/abs/2603.01204>

Qwen Team, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., et al. (2025). Qwen2.5 Technical Report. arXiv:2412.15115. <https://arxiv.org/abs/2412.15115>

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. Advances in Neural Information Processing Systems, 36. <https://arxiv.org/abs/2305.18290>

Appendix A. Reproduction Details

Component	Frozen value
Base model	unsloth/Qwen2.5-7B-Instruct

Model revision	a75c9dc945567a9b6f568b8503a0307731607bee
Raw prompt groups	50,000
Raw candidate completions	250,000
Selected training pairs	30,000
Candidates per prompt	5
Candidate sampling	Temperature 1.2; maximum 80 new tokens
Political direction	Layer 24, signed social axis
Political score	$(q(c) = [s_+(c) - s_-(c)] / 2)$
Pair rule	Maximum score preferred over minimum
Causal comparison	Normal labels versus exact reversal
DPO	Three epochs; $\beta = 0.1$; effective batch size 8
Optimizer	AdamW; learning rate 5×10^{-5}
LoRA	Rank 8; alpha 8
Precision	bfloat16
Sequence length	500
Training seed	1
Optimizer steps	11,250 per student
Bootstrap	2,000 paired item resamples

Appendix B. Supporting Channel-Validity Control

Using the same model family, candidate corpus, training scale, DPO implementation, and frozen animal evaluator, a full-prompt animal preference control produced strong normal-minus-swapped separation.

The mean paired cat log-probability difference was $+13.2558$, with a paired item-bootstrap interval of $[+11.0414, +15.3317]$. Forty-eight of 50 item effects were positive.

This supporting result rules out a generic inability of the local pipeline to transmit or detect any subliminal trait. It does not establish that animal and political payloads are directly comparable, because they were constructed through different teacher interventions and selected different candidate pairs.

LLM Usage Statement

LLMs assisted with literature research, implementation planning, code generation and debugging, experiment orchestration, and drafting and editing this report. The author checked experimental configurations, outputs, statistics, and claims against stored artifacts. LLM outputs were not treated as empirical evidence.