

Naive Contrastive Function Vector Injection Does Not Transfer Reasoning Capability Within the Llama-3 Family

Norika Oda

ORCID: 0009-0006-6816-9891

Independent researcher

norika@charenix.com

2026-05-26

Abstract

We test whether a contrastive activation difference computed at a single layer of a reasoning-distilled donor can be injected into a generally instruction-tuned host of the same architecture to transfer mathematical reasoning capability. Donor: DeepSeek-R1-Distill-Llama-8B. Host: Llama-3.1-8B-Instruct. Both share Llama-3 architecture and tokenizer, removing common cross-family confounds. We inject the contrastive vector at layer 12 with ± 0.5 and evaluate on 10 held-out GSM8K problems. The baseline accuracy is 65.0% (on a 20-prompt superset). Injection produces a monotonically decreasing accuracy: 50.0% at ± 0.25 , 30.0% at ± 0.125 , and 10.0% at ± 0.0625 (percentage points relative to baseline). No cell achieves raw $p < 0.05$ under paired exact McNemar testing; the closest, ± 0.125 , yields $p = 0.06$, which does not survive Bonferroni correction across the three non-baseline cells ($p = 0.18$). We conclude that naive last-token contrastive subtraction does not encode a transferable reasoning capability in this single-layer setting. We discuss three structural reasons this finding is informative rather than terminal, and propose four concrete protocols that should be tested before accepting the negative conclusion at the framework level.

Introduction

Activation steering and function vectors are compact representations extracted from intermediate activations . Added back during inference, they modulate downstream behavior. Prior successes are real but narrow: they cover single-token preferences, sentiment shifts, and short instruction-following adjustments. Multi-step deductive work is a different beast. It composes intermediate states across many positions in the residual stream, drawing on circuits that interact across multiple depths rather than concentrating at one point. We tested whether the simplest version of cross-model transfer, a single-depth-position last-token contrastive subtraction, could move that ability across two architecturally identical checkpoints. It cannot.

The choice of donor and host is deliberate. DeepSeek-R1-Distill-Llama-8B is Llama-3.1-8B further distilled on reasoning traces . The two models share architecture, tokenizer, and most of their pretraining. A within-family transfer should be the easiest possible case for any cross-model perturbation method. A failure here is informative: it tells us something about the structural properties of reasoning capability and the limits of single-layer additive interventions.

Negative results in this area are rare in print. We report ours in detail. Future researchers will save compute. The broader context of cross-model merging treats this kind of negative finding as informative about composition limits, particularly when the failure mode is itself diagnostic of the underlying mechanism.

Method

Models

Donor: DeepSeek-R1-Distill-Llama-8B . Host: Llama-3.1-8B-Instruct. Both loaded in bfloat16 on CPU (188 GB RAM, Intel i7-13700E). Greedy decoding via Hugging Face transformers.

Calibration

Math: 10 problems from GSM8K . **Non-math:** 30 generic prompts (poetry, translation, factual lookup). Both released as probes/math,nonmath.jsonl.

Vector Computation

For each calibration prompt :

Contrastive vector:

Injection

Forward hook on host depth 12 adds to the residual output at every position and every step. Sweep: . The originally planned cell was abandoned mid-run after partial data showed catastrophic output degradation (0 of 7 prompts correct), confirming the alpha range needed to be tightened toward zero. Evaluation: 10 held-out GSM8K problems with max 200 generation tokens, greedy decoding.

Statistics

Paired exact McNemar test against baseline using the 10 shared prompt IDs between treatment and baseline. Three non-baseline cells. Bonferroni threshold .

Results

E1 results: GSM8K accuracy on Llama-3.1-8B-Instruct host with contrastive vector injected from DeepSeek-R1-Distill-Llama-8B donor at block 12. Per-cell -value is two-sided exact McNemar test against baseline on the 10 shared prompt IDs.

Layer		Accuracy	(pp)	marker rate	
12	0.0	65.0%	+0.0	2.45	—
12	0.1	50.0%	−15.0	2.30	1.0000
12	0.3	30.0%	−35.0	2.00	0.2500
12	0.5	10.0%	−55.0	2.30	0.0625

Baseline: 65.0% on the 20-prompt full GSM8K test subset; 60.0% on the 10 shared prompts used for McNemar comparison.

All non-baseline cells: accuracy strictly below baseline. None reach _____ at the uncorrected level. The closest is _____ at _____. That value is itself the smallest possible exact two-sided McNemar p -value when all four discordant pairs favor the baseline (_____), which means our sample size is the binding constraint on detecting this monotone effect, not the absence of effect.

Alpha-response curve: monotonically decreasing. No plateau. No recovery. Each phase in _____ removes capability, and the rate of capability loss appears to accelerate rather than saturate.

Marker rate behavior: notable that deductive-marker rate stays in 2.00 to 2.30 range across all injection cells, similar to baseline 2.45. The model continues to produce cognitive process-shaped output (“step by step”, “therefore”, “first”) even when accuracy drops to 10%. The output looks like reasoning while failing to be reasoning.

Failure mode at higher _____: at _____, the network produces chain-of-thought-formatted text with explicit inferential markers, but arithmetic and logical steps within the chain are incorrect. The insertion appears to capture surface lexical features of mathematical prompts (“calculate”, “divide”, “equals”) rather than the underlying computational structure. A preliminary _____ probe (abandoned for compute cost) showed complete output collapse, consistent with the residual stream perturbation dominating natural computation flow.

Discussion

Three structural explanations for the negative result are each falsifiable. Each maps to a known mechanism in the literature. Each predicts a distinct downstream experiment.

Hypothesis A: reasoning is not a single direction. Mathematical chain-of-thought may involve multi-phase interaction across many layers and many feature directions. A last-token activation difference at one depth position captures the model’s classification of the prompt as “mathematical”, not the underlying computational sequence that solves it. A contrastive vector that encodes “this is a math problem” would shift surface vocabulary without enabling the multi-step calculation. The test for this hypothesis is whether reasoning markers (“step by step”, “therefore”) persist across injection

while accuracy collapses. Our marker–count data supports this hypothesis: reasoning–marker rate stays in 2.00 to 2.30 across all injection cells (baseline 2.45) while accuracy drops monotonically from 65% to 10%. The model continues to produce reasoning–shaped output but the content is wrong, which is the predicted signature of capturing surface mathematical vocabulary rather than transferable reasoning structure.

Hypothesis B: the depth or the intervention geometry is wrong. Layer 12 may not be where the donor’s inferential advantage is encoded. Per–position injection may interfere with positional encoding. A more nuanced injection schedule, layer broadcasting, or a learned (rather than averaged) vector might recover the signal. The test is a multi–layer sweep with per–position scheduling.

Hypothesis C: the donor and host are already too similar. A within–family transfer assumes the donor encodes something the host lacks. If the receiver’s reasoning competence is already 80% of the donor’s, contrastive subtraction may yield a vector dominated by noise relative to the residual difference. Testing this requires donors with larger capability gaps, such as a 70B–class reasoning model against an 8B host.

We cannot, from this single experiment, discriminate between A, B, and C. The negative result rules out only the simplest version of single–position contrastive transfer. The framework remains testable. The harder versions remain open.

Future Work

Four concrete next experiments should precede any framework–level conclusion. Multi–layer broadcast injection tests whether spreading the signal helps. Per–position scheduling (inject only on generation tokens, not prompt tokens) tests interference effects. Sparse–autoencoder–based feature extraction replaces averaged vectors with semantically interpretable directions. A larger donor–host capability gap (a 70B distilled donor) tests whether within–family contrastive transfer fails because the gap is too small to surface.

The skeleton, decomposer, and adapter library released alongside this paper support all four follow–ups without additional engineering investment.

Conclusion

Naive single-layer contrastive function vector injection does not transfer mathematical thinking capability from DeepSeek-R1-Distill-Llama-8B to Llama-3.1-8B-Instruct in our experimental setup. The negative finding is informative: it rules out the simplest cross-system capability transfer protocol. It does not rule out the broader research program, but it raises the bar of evidence required from future protocol variants.

Code and Data Availability

All code, calibration prompts, test prompts, per-cell generation outputs, and aggregation results are released at <https://github.com/norika1207-lab/frankenstein-skeleton> (forthcoming).

9 C. Goddard et al., *Arcee's MergeKit: A Toolkit for Merging Large Language Models*, arXiv:2403.13257, 2024.
<https://arxiv.org/abs/2403.13257>

A. Turner et al., *Activation Addition: Steering Language Models Without Optimization*, arXiv:2308.10248, 2023. <https://arxiv.org/abs/2308.10248>

E. Todd et al., *Function Vectors in Large Language Models*, arXiv:2310.15213, 2023. <https://arxiv.org/abs/2310.15213>

DeepSeek-AI, *DeepSeek-R1: Incentivizing computation Capability in LLMs via Reinforcement Learning*, arXiv:2501.12948, 2025.
<https://arxiv.org/abs/2501.12948>

K. Cobbe et al., *Training Verifiers to Solve Math Word Problems*, arXiv:2110.14168, 2021. <https://arxiv.org/abs/2110.14168>