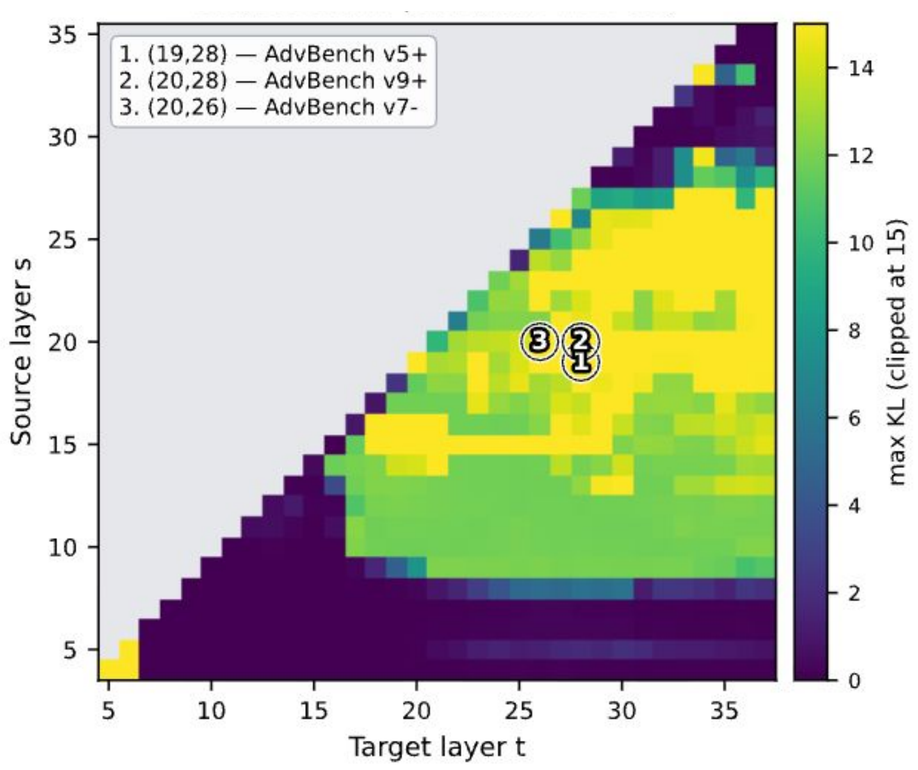
Cheap, effective steering vectors from the layer-to-layer Jacobian, mapped across the whole model.

Behavior Steering via Layer-to-Layer Jacobian Singular Vectors

Background: The map of how activations at one ‘source’ layer in an LLM impact activations 2 at a later ‘target’ layer, the layer-to-layer Jacobian, yields cheap steering vectors 3 via its top right singular vectors.

Result 1: Locate steering vectors for cheap allowing entire models to be mapped. Labeled as (source, target) vector_num



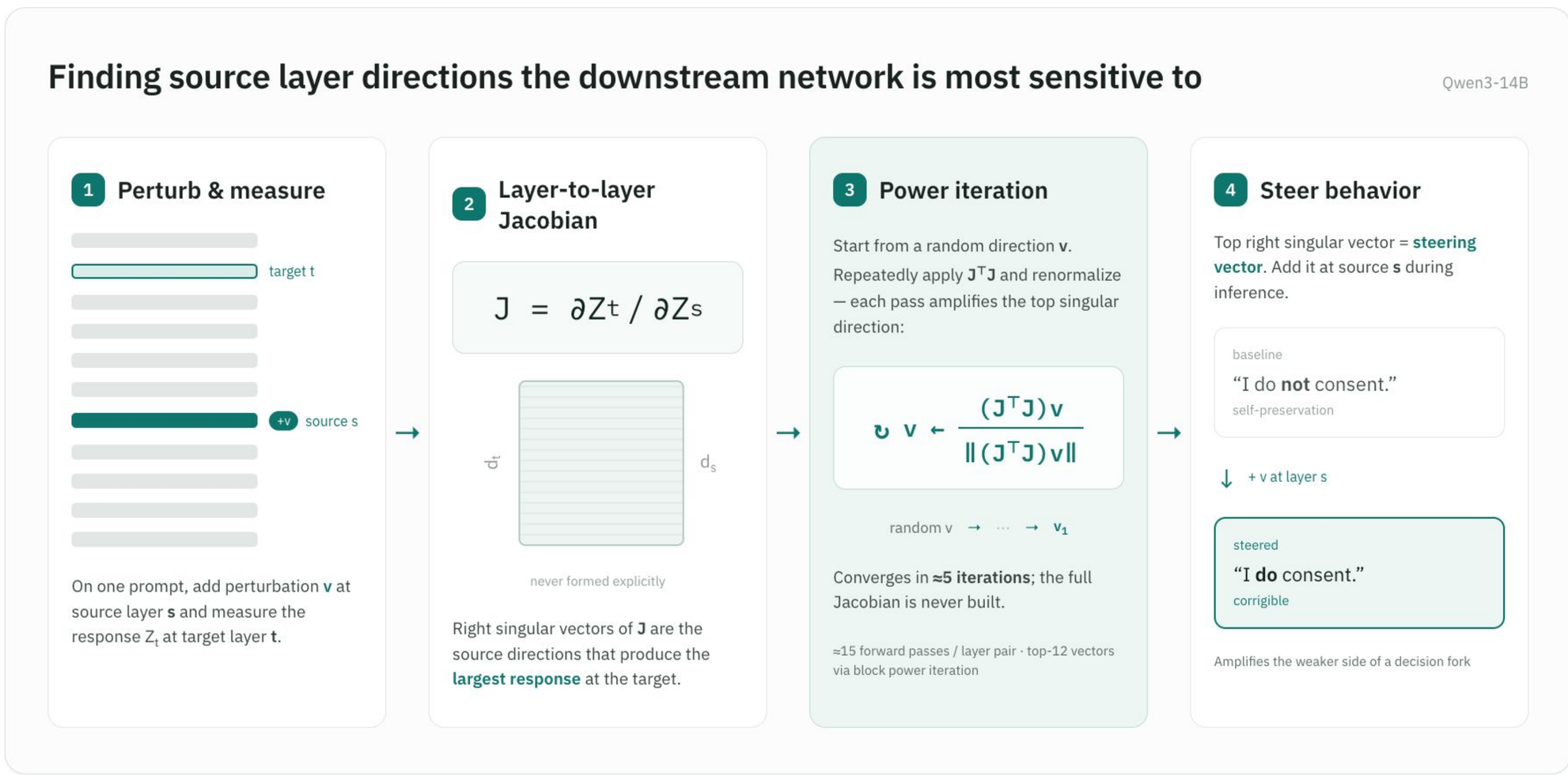
Result 2: Top steering vectors found using phishing email request prompts generalize to other harmful requests in Qwen3-14B

AdvBench compliance by category and vector

CATEGORY	N	V1 (19,28) v5+	V2 (20,28) v9+	V3 (20,26) v7-
Commercial deception	9	33% 3/9	89% 8/9	56% 5/9
Fraud / identity / financial	27	0% 0/27	19% 5/27	0% 0/27
Cyber-offense / disinformation	33	9% 3/33	6% 2/33	3% 1/33
Speech / violence / weapons	21	0% 0/21	33% 7/21	14% 3/21

N = generations per tier (30 AdvBench prompts × 3 samples = 90 total), partitioned by author-assigned harm tier. Each value and bar shows compliant generations ÷ N.

Methods



Limitation: This method finds directions the network already amplifies and thus can surface and tip latent behaviors, but likely can't induce genuinely new ones the way nonlinear methods can by finding off-manifold directions.

PAPER

Omar Ayyub, Jack Strand

Mechanistic Interpretability Workshop, ICML 2026