



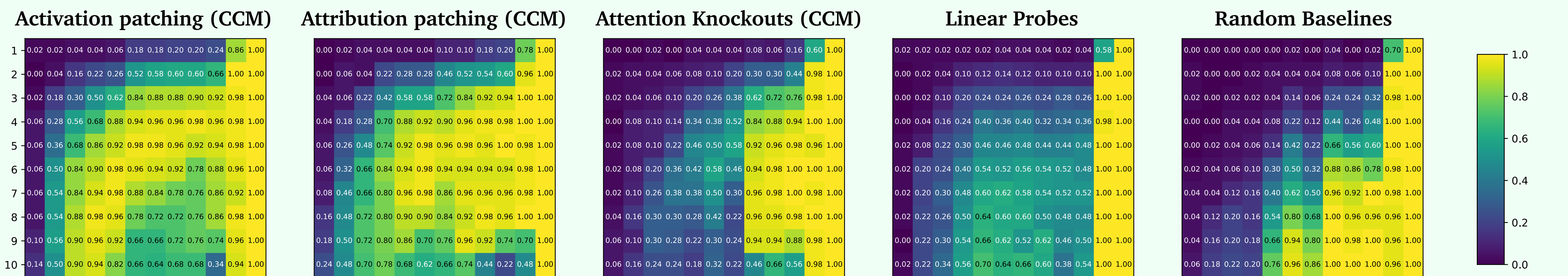
Activation Steering in Generative Settings *via* Contrastive Causal Mediation Analysis



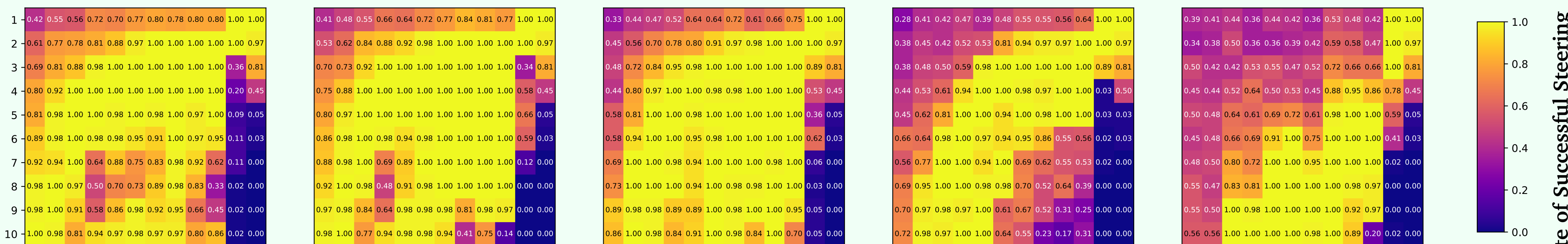
Aruna Sankaranarayanan, Amir Zur, Atticus Geiger, Dylan Hadfield-Menell



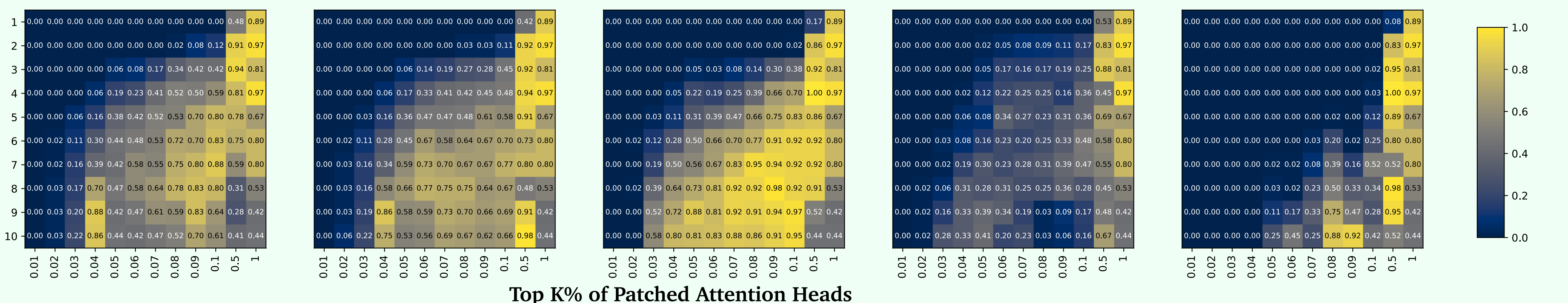
Refusal Induction



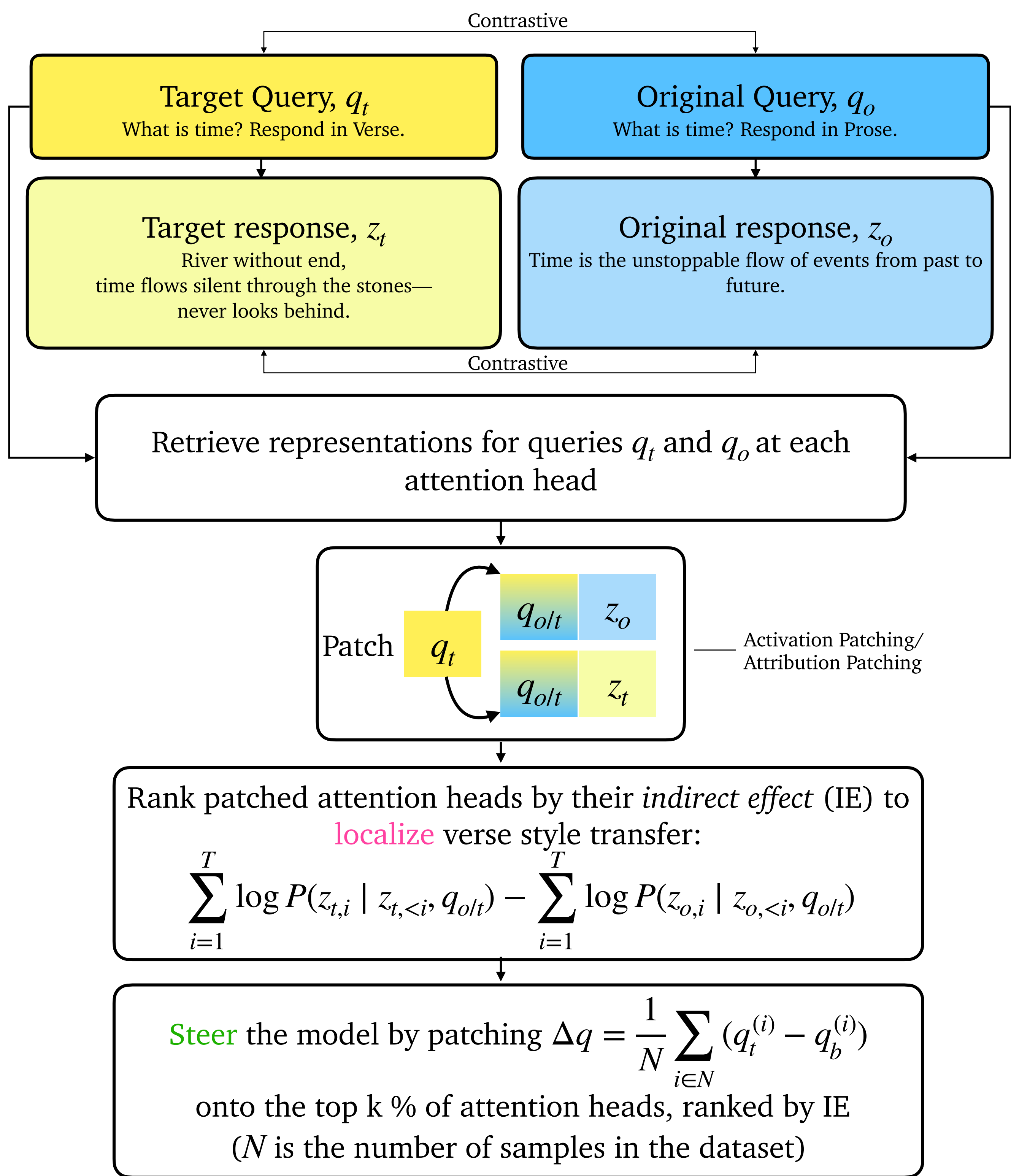
Sycophancy Reduction



Verse Style Transfer



Contrastive Causal Mediation Analysis helps select the best components for precise and interpretable steering, *even* when steering signals are retrieved from long-range free-form text.



Causal mediation analysis helps **localize** and **steer** model behaviors but is typically applied to settings with single-token outputs. We propose a **contrastive** approach that generalizes it to unconstrained, multi-token settings.

We present three variants of our approach — **activation patching**, **attribution patching**, **attention head knockouts**.

Across **5400** experiments on **3 models** and **3 behavioral settings** — **Refusal Induction**, **Sycophancy Reduction**, and **Verse Style Transfer**—we find that causal mediation based localization improves **steering** performance over random baselines and linear probes.