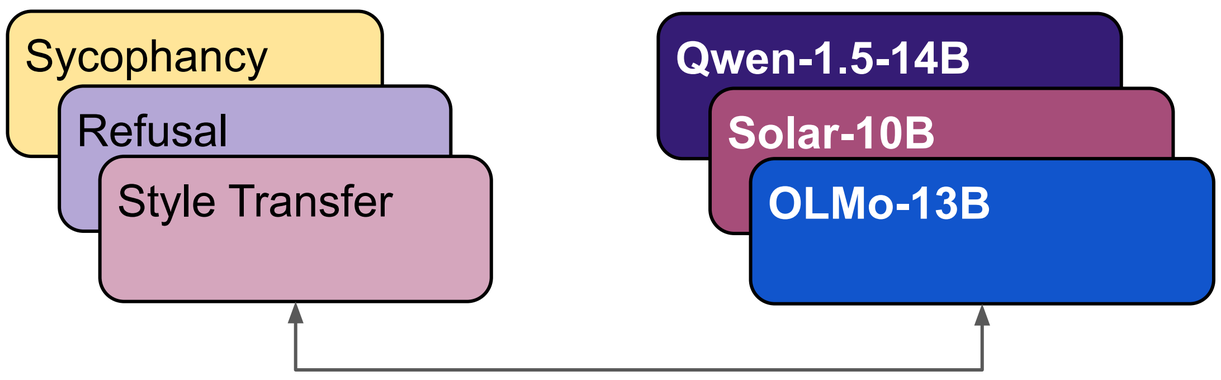
Activation Steering in Generative Settings via Contrastive Causal Mediation Analysis

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

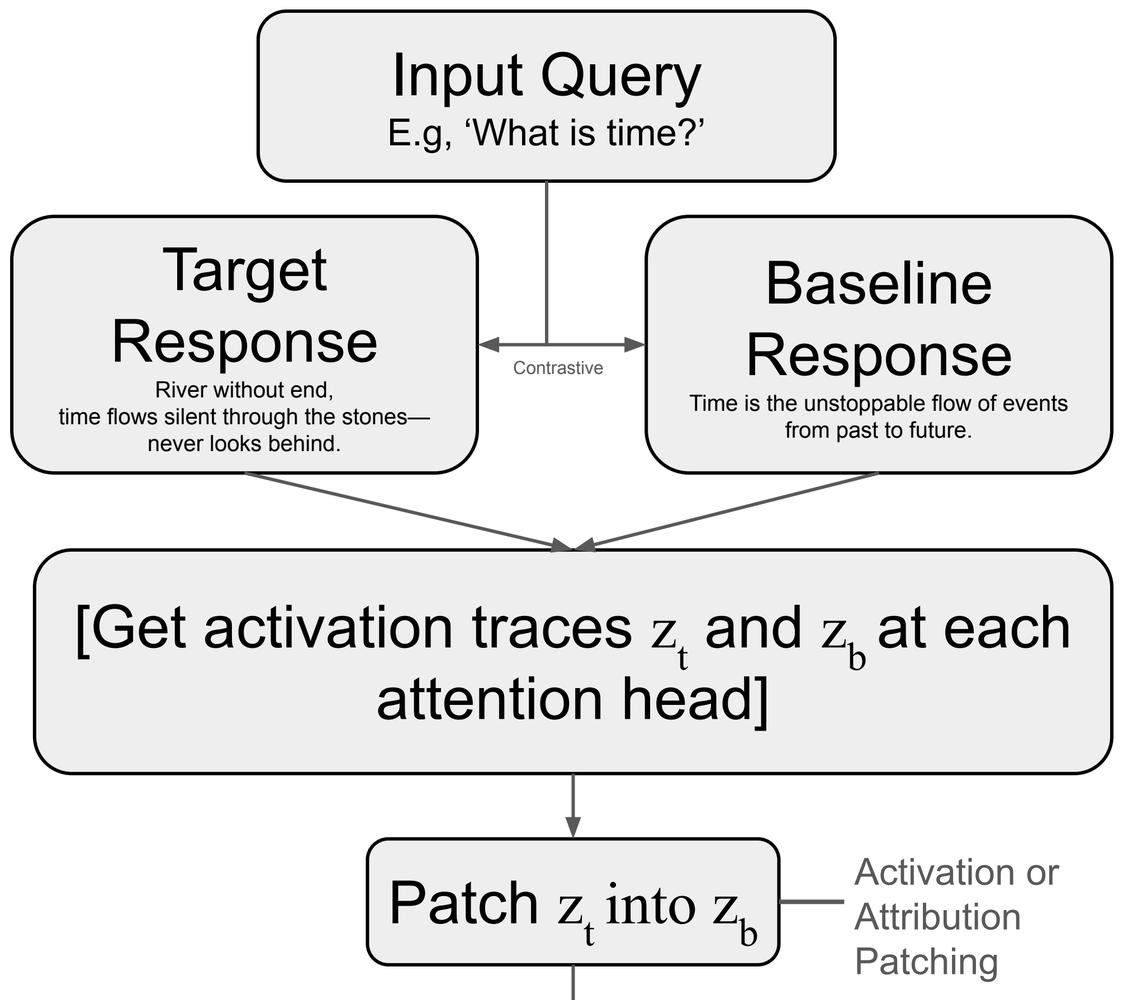
Causal Mediation Analysischas been shown to localize concepts to model components using signals from constrained-length model outputs.

Can free-form text similarly offer signals for mediation, localization, and thus, model steering? YES! We present Contrastive Causal Mediation (CCM)

Evaluate 3 CCM variants — full vector patching, attribution patching, and attention head knockouts on multiple task settings and models. CCM variants outperform random baselines and linear probes.



We present evidence through 3 CCM variants on 3 different tasks across 3 different models. CCM variants outperform linear probes and random baselines.



Contrastive Causal Mediation helps selects the best attention heads for precise and interpretable control.