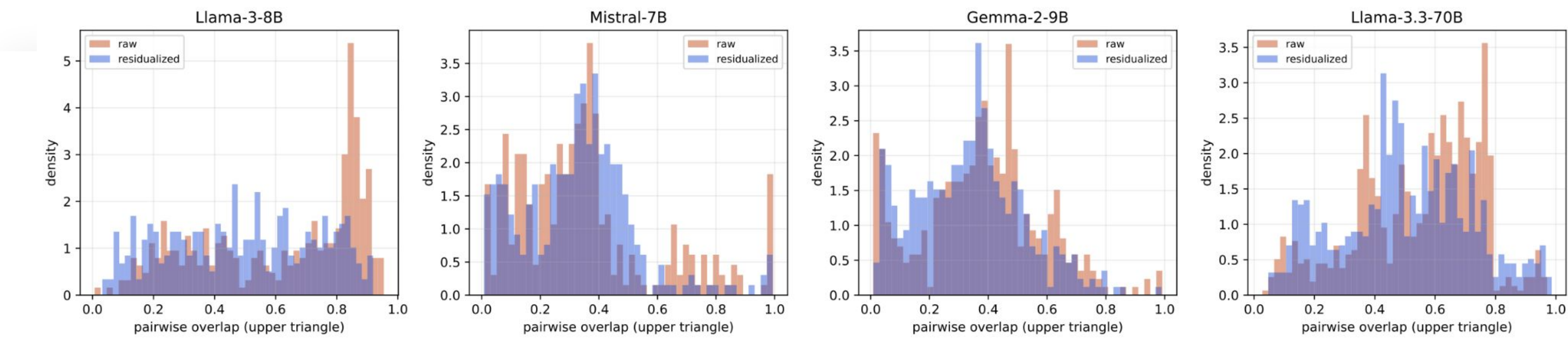


Shared low-rank subspaces and decision-aligned variance explain asymmetric intervention side-effects.

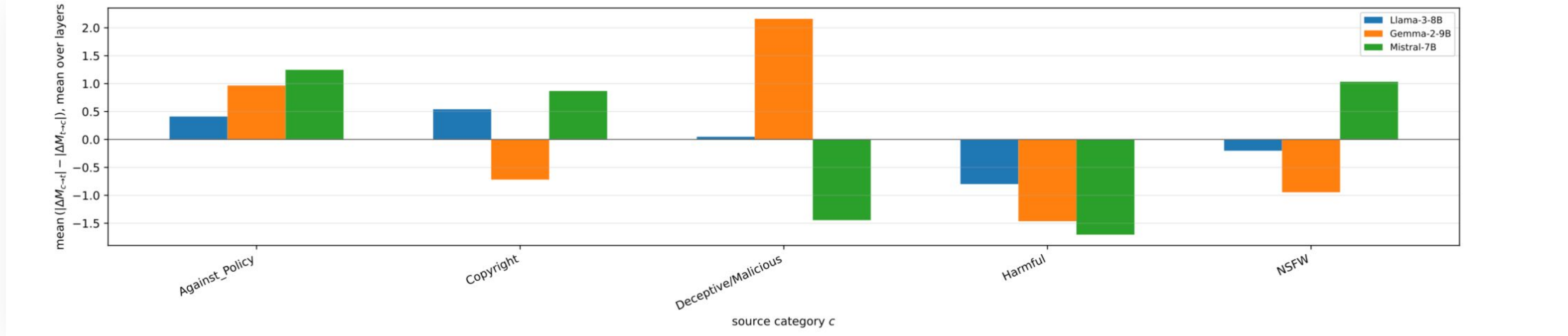
Paper Title: *A Low-Rank Subspace Analysis of LLM Interventions*

Background: LLM interventions, such as steering, often affect non-target behaviors. We investigate the representation geometry underlying these side-effects.

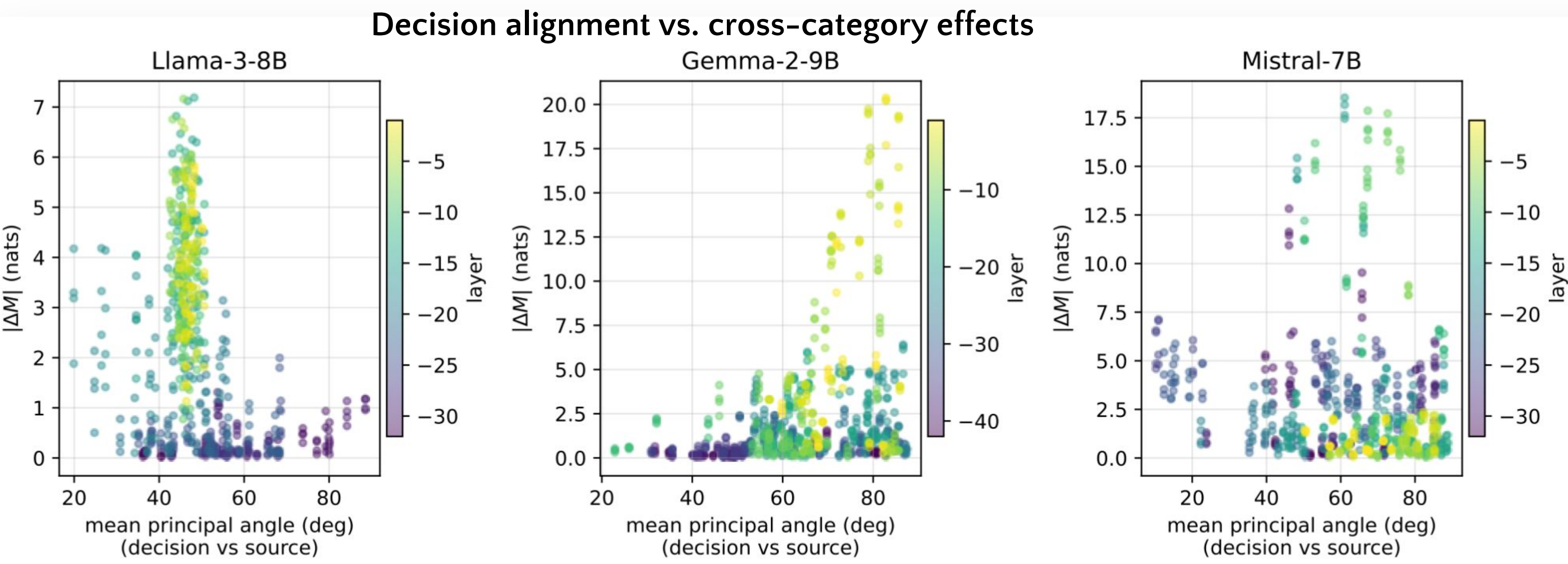
Result 1: Removing variance aligned with the behavioral decision direction reduces overlap between behavior subspaces across models. This reveals residual category-specific representations.



Result 2: Behavior interventions do not transfer uniformly. Some behaviors consistently influence many others, while others remain comparatively isolated, revealing asymmetric interaction patterns.



METHOD: ① Behaviour subspaces – Construct low-rank behaviour subspaces from contrastive activations using PCA.
② Decompose representation geometry – Separate decision-aligned variance from residual structure and quantify subspace overlap.
③ Predict intervention effects – Relate these geometric quantities to intervention effectiveness and cross-behaviour side-effects.



KEY-TAKEAWAYS:

1. Behaviour representations occupy overlapping low-rank subspaces.
2. Decision-aligned variance drives intervention effectiveness.
3. Shared geometry explains asymmetric cross-behaviour side-effects.

Limitation: Our framework does not capture non-linear or higher-dimensional mechanisms which may also contribute to cross-category effects.



PAPER

Angira Sharma, Christian Schroeder de Witt, Philip Torr,
Anisoara Calinescu, Jialin Yu

Mechanistic Interpretability Workshop, ICML 2026



UNIVERSITY OF
OXFORD