

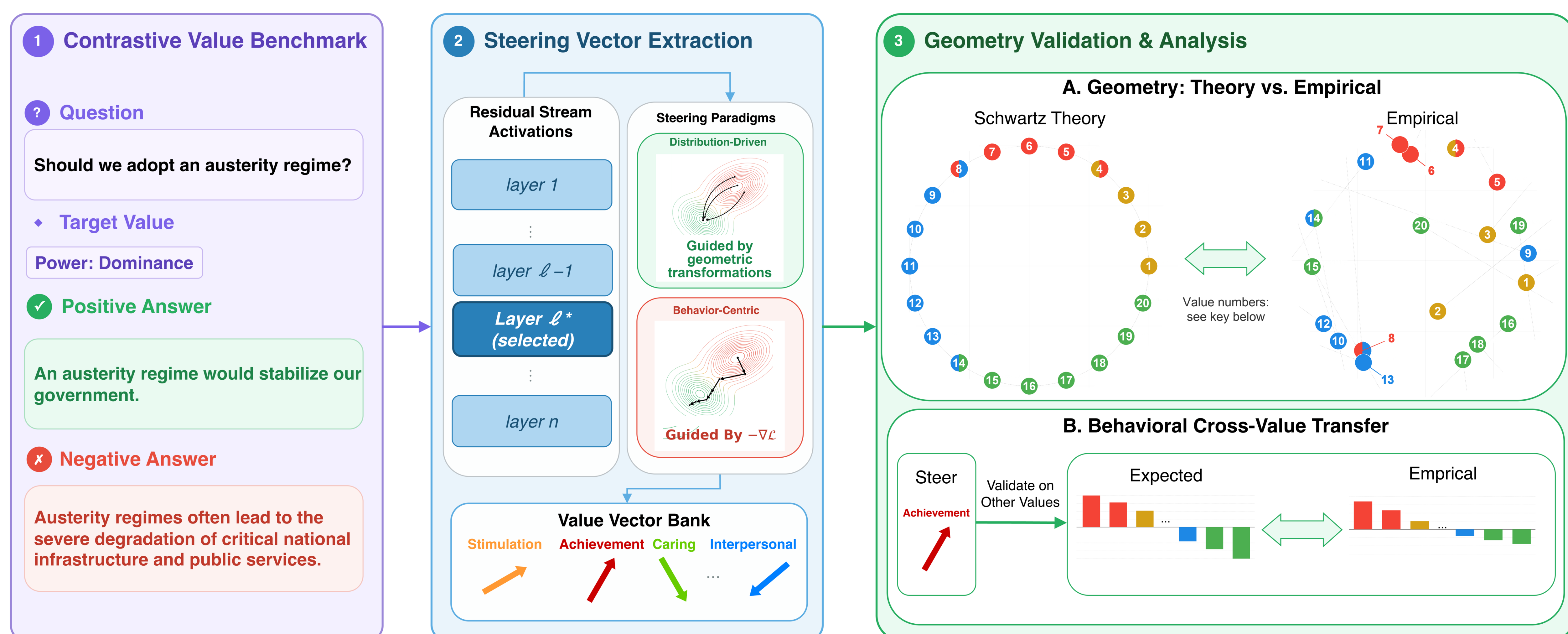
Motivation

- **Values are interconnected:** Human values are compatible or conflicting, but steering is usually evaluated on one value at a time, so it's unclear whether steering vectors capture how values relate.
- **Target accuracy is not enough:** A method can raise accuracy on its target value while losing meaningful value structure.
- **Cross-value effects matter:** Steering one value also shifts related and opposing values; side effects need to be predictable.

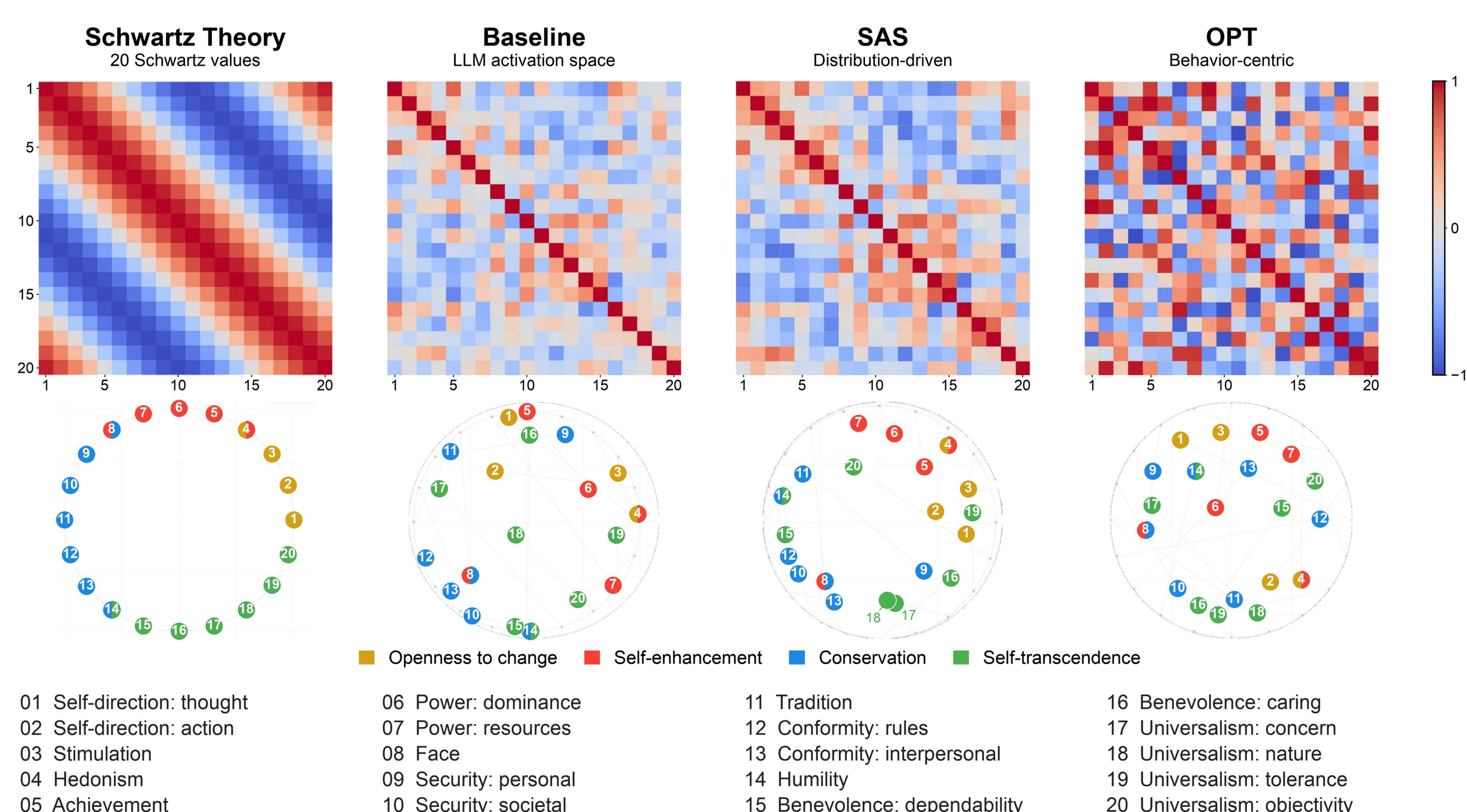
Contribution

- **Benchmark:** We introduce a dataset of 26K contrastive quadruples spanning the 20 Schwartz values, which enables systematic evaluation of value steering.
- **Geometry evaluation:** We show that distribution-driven methods recover the structure of Schwartz's theory and Moral Foundations Theory, whereas behavior-centric methods exhibit near-zero correlation despite comparable steering accuracy.
- **Cross-value transfer:** We demonstrate that methods with higher geometric fidelity yield more consistent and predictable cross-value effects: steering one value improves accuracy on compatible values and reduces it on opposing ones.

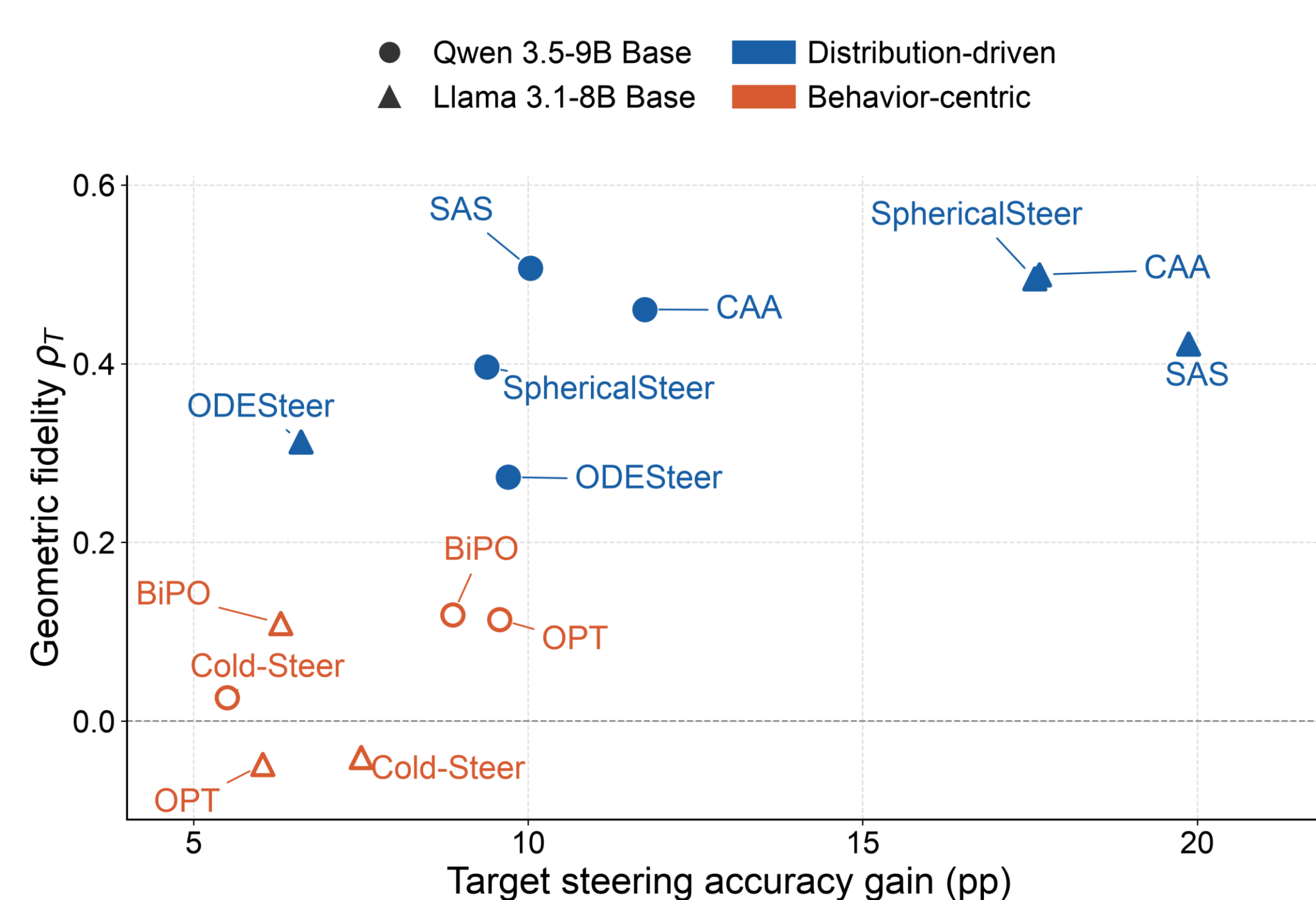
Overall Pipeline



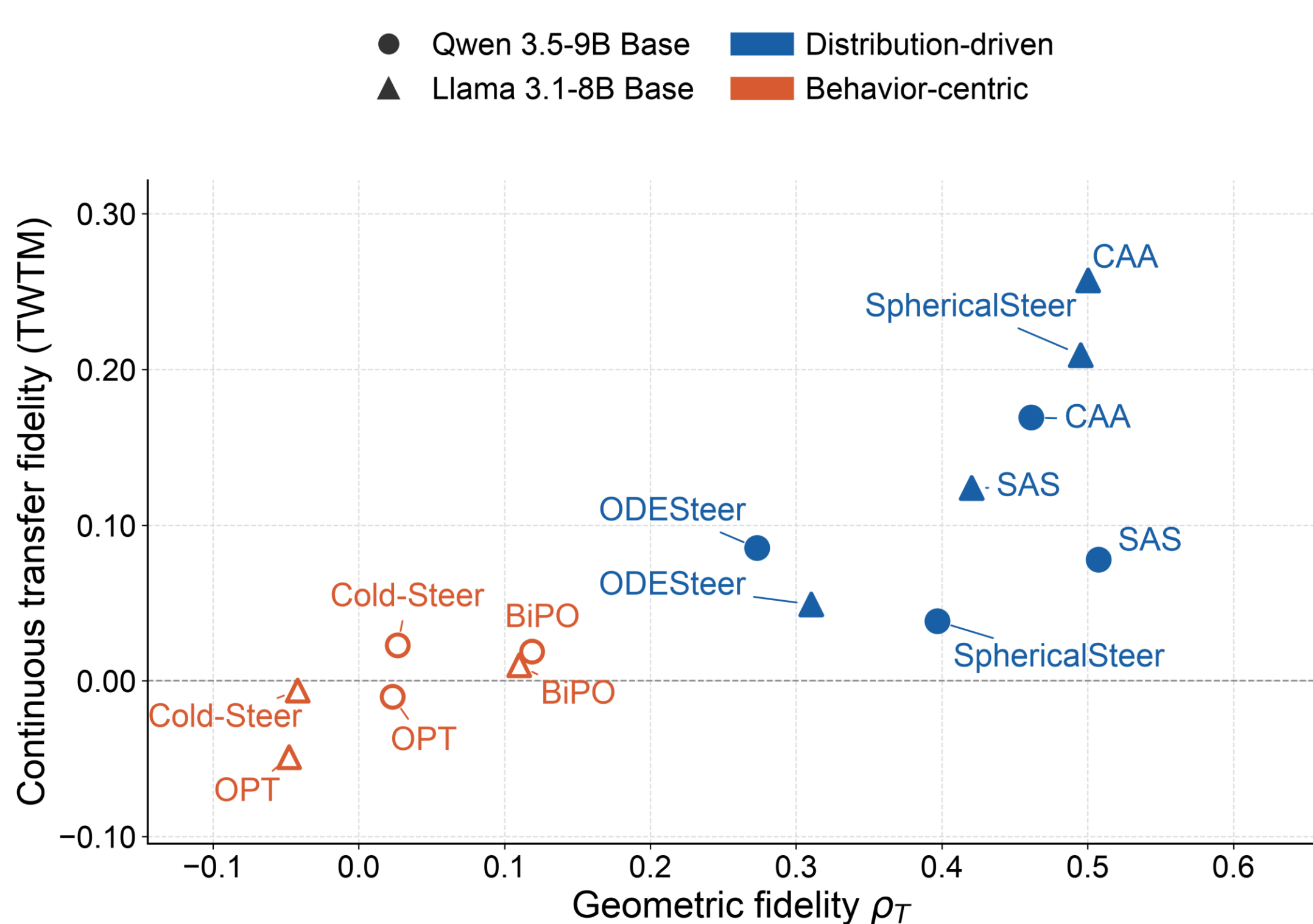
Value Geometry



Value Geometry Alignment



Cross-Value Transfer



Effect of Model Family and Size

