Construct Relative Moral Representation

Moral Foudation Vignettes (MFV130) → Extended Generated by GPT → Human Expert Evaluation

Extended MFV Dataset: Care, Fairness, Loyalty, Authority, Sanctity, Social Norm

LLM Transformer Stack (e.g., LLaMA-3.1-8B-Instruct)

Layer 1 — Residual Stream

Layer 2 — Residual Stream

Layer ℓ — Residual Stream

Layer L — Residual Stream

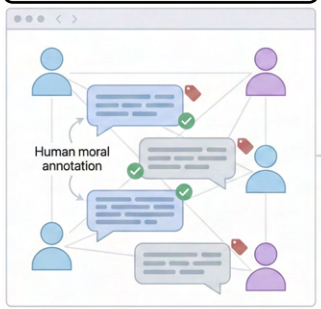
Concept Vector Generation (layer-wise)

Two Types of Moral Concept Vectors

- Foundation A vs Foundation B (e.g. Care - Authority)
- Foundation A vs Social Norm (e.g. Care – Social Norm)

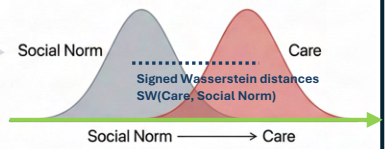
Topological Alignment with human-labeled Distributions

Moral Foundations Reddit Corpus

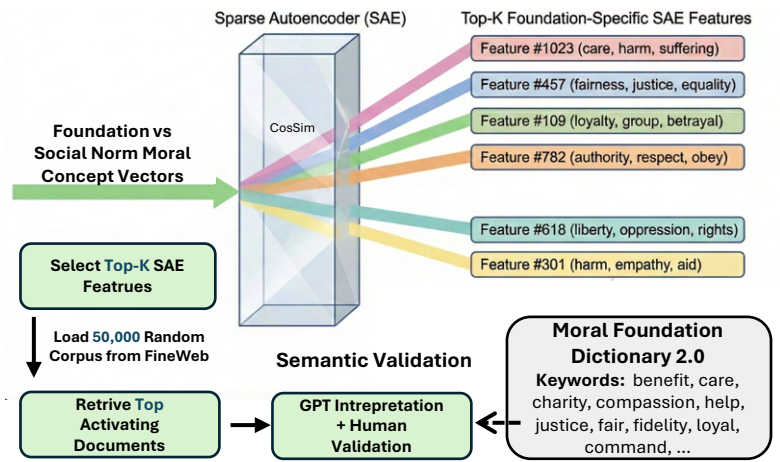


Projection onto Foundation A vs Foundation B vector

Projection onto Foundation vs Social Norm vector



Mechanistic Decomposition through SAEs



Casual intervention through steering

1) Macro-steering

Moral Concept Vector → Residual Stream

Global, linear intervention

2) Micro-steering

SAE Feature Directions

Top Feature 1
Top Feature 2
Top Feature 3
Top Feature 4
Top Feature 5
...
Top Feature 10

→ Residual Stream

Sparse, feature-level intervention

Moral Foundations Questionnaire 2 (MFQ2)

Evaluations

- Performance Shift — **Average Logit-based Moral Score**
- General Ability — **MMLU Performance**