Faithful Bi-Directional Model Steering via Distribution Matching and Distributed Interchange Interventions

Presenter: Yuntai Bao
Zhejiang University

Background

Large language models (LLMs) are powerful but difficult to control reliably. Two dominant approaches each have significant drawbacks:

- **Prompting**: flexible, but brittle and labor-intensive;
- **Fine-tuning**: powerful, but expensive and causes broad, hard-to-reverse behavioral changes.

A promising middle ground is model steering: manipulating internal representations at inference time by injecting a "***steering vector (SV)***" into the model's residual stream. This approach is:

- Computationally efficient
- Interpretable and reversible
- Targeted (affects specific behaviors, not the whole model)

Motivation

Current optimization-based steering methods borrow objectives directly from fine-tuning:

- **Language modeling (Lang.)** objectives maximize likelihood of desired outputs;
- **Preference optimization (PO)** methods rank concept-relevant responses over neutral ones.

These strong supervision signals can cause overfitting, degenerate/repetitive outputs, and unstable or off-distribution generations — the same failure modes seen in fine-tuning.

Core Hypothesis

Effective steering requires faithfully identifying internal concept features, not enforcing external preferences into the model.

Main Findings:

1. Preservation of general capabilities

- **Maintains utility**: CDAS preserves near-baseline performance on standard benchmarks like MMLU and TruthfulQA.
- **Faithful distributions**: Achieves the lowest KL divergence among steering methods, indicating that interventions remain faithful to the model's natural output distribution rather than forcing unnatural behaviors.

2. Robust bi-directional control of safety concepts

- **Refusal override**: Effectively overrides safety refusals in aligned models to enable "compliant" modes without fine-tuning.
- **Backdoor defense**: Successfully neutralizes "sleeper agent" backdoors (hidden malicious behaviors triggered by specific contexts) by identifying and suppressing the internal trigger mechanisms.

3. Scalability with model size

- **Performance scaling**: Steering effectiveness improves significantly with model size; CDAS outperforms competitors on larger models (e.g., Llama-3.1-70B) while smaller models (e.g., Phi-3.5) show mixed results.
- **Resistant to overfitting**: While baseline methods degrade on large models due to overfitting on training labels, CDAS maintains stability and effectiveness as model capacity increases, making it suitable for high-parameter settings.

Scan to view the full paper

Scan to view blog post

Authors: Yuntai Bao[1], Xuhong Zhang[1], Jintao Chen[1], Ge Su[1], Yuxiang Cai[1], Hao Peng[2], Bing Sun[3], Haiqin Weng[4], Liu Yan[4], Jianwei Yin[1]

[1]Zhejiang University
[2]Zhejiang Normal University
[3]National Certification Technology (Hangzhou) Co., Ltd
[4]Ant Group

Method: Concept Distributed Alignment Search (CDAS)

Foundation: Distributed Interchange Intervention (DII)

CDAS builds on ***Distributed Alignment Search (DAS)***, the standard method for causal variable localization. The key intervention primitive is DII, which:

- Identifies a low-dimensional representation subspace encoding a target concept
- "Clamps" that subspace to the value it would take under a counterfactual input
- Naturally supports bi-directional steering (both concept injection and suppression)

$$\Phi^{\text{DII}}(\mathbf{h}; a) = \mathbf{h} + (a - \mathbf{w}_\Phi^\top \mathbf{h})\mathbf{w}_\Phi,$$

where $\mathbf{h}$ is representations, a is steering factor and $\mathbf{w}_\Phi$ is the SV.

Distribution Matching Objective

Standard DAS maximizes the probability of a specific ground-truth output, which rarely holds in open-ended steering tasks. CDAS replaces this with a weak-supervised distribution matching objective using Jensen-Shannon Divergence (JSD):

$$\arg\min_{\Phi} \mathbb{E}_{((\mathbf{x},\mathbf{y}),(\mathbf{x}^c,\mathbf{y}^c))\sim\mathcal{D}^c_{\text{train}}} \left[D^+_\Phi + D^-_\Phi\right],$$

$$D^+_\Phi = \frac{1}{|\mathbf{y}^c|}\sum_{k=1}^{|\mathbf{y}^c|} D_{\text{JS}}\left(\mathbf{p}_\Phi\left(\cdot|\mathbf{y}^c_{<k}, \mathbf{x}; \mathbf{h} \leftarrow \Phi^{\text{DII}}(\mathbf{x}^c)\right) \,\middle\|\, \mathbf{p}\left(\cdot|\mathbf{y}^c_{<k}, \mathbf{x}^c\right)\right),$$

$$D^-_\Phi = \frac{1}{|\mathbf{y}|}\sum_{k=1}^{|\mathbf{y}|} D_{\text{JS}}\left(\mathbf{p}_\Phi\left(\cdot|\mathbf{y}_{<k}, \mathbf{x}^c; \mathbf{h} \leftarrow \Phi^{\text{DII}}(\mathbf{x})\right) \,\middle\|\, \mathbf{p}\left(\cdot|\mathbf{y}_{<k}, \mathbf{x}\right)\right),$$

Beyond Steering

CDAS could alternatively be interpreted as *self-distillation* or *context distillation*, where concept-specific steering prompts are distilled into SVs.