A Sober Look at Steering Vectors for LLMs

by **Joschka Braun, Dmitrii Krasheninnikov, Usman Anwar, Robert Kirk, Daniel Tan, David Scott Krueger**

23rd Nov 2024

We thank Madeline Brumley, Joe Kwon, David Chanin and Itamar Pres for their helpful feedback.

Introduction

Controlling LLM behavior through directly intervening on internal activations is an appealing idea. Various methods for controlling LLM behavior through activation steering have been proposed. Most steering methods add a 'steering vector' (SV) to the model's activations at a given layer and token position during inference. This approach leverages the hypothesis that many human-interpretable 'concepts' like truthfulness, refusal, and sentiment are represented as directions in activation space. Steering interventions are appealing because they use much less data than fine-tuning and do not require changes to the model parameters. In principle, this makes them more efficient and easy to controlling properties of the generated text in a desired way.

However, steering methods face a variety of significant challenges that hinder their practical applicability, comparability and further improvement. In this blog post, we discuss recent work performed at KASL and UCL DARK investigating challenges to steering methods for LLMs, focusing on 3 key challenges:

1. **Current steering methods have substantial limitations**
    a. When evaluated more thoroughly or in different settings, many steering methods turn out to be unreliable and often fail to generalize outside their specific training setup.

   b. The steerability of different concepts varies significantly, with some proving resistant to steering.

2. **Typically used performance metrics overestimate steering effectiveness**
   - Most steering methods are trained and evaluated in artificial settings like Multiple Choice Question Answering instead of settings that reflect deployment, like free-form text generation or behavior on typical application tasks. This raises questions about whether steering methods are useful and reliable in practice.

3. **Methods are not compared on the same benchmarks and metrics**
   - Most steering methods are trained and evaluated on method-specific datasets, in which they show good performance. However, evaluations on datasets of prior work or common benchmarks are often missing. This makes it difficult to track progress and evaluate the strengths and weaknesses of new methods.

Overall, there is still much research to do to make steering methods a reliable and useful tool for controlling LLMs. We conclude with a discussion of several important recommendations for future research.

Current steering methods have substantial limitations

Reliability of positive steering effect

In Analyzing the Generalization and Reliability of Steering Vectors (Tan et al., 2024) (accepted at NeurIPS 2024), we find that Contrastive Activation Addition (CAA) has substantial limitations in terms of robustness and reliability. Steerability is highly variable across different inputs: depending on the concept, spurious biases can substantially contribute to how effective steering is for each input, presenting a challenge for the widespread use of steering vectors. While CAA is effective on some tasks, many behaviors turn out to be unsteerable, even when sweeping across all layers and strengths for steering. As a result, it is difficult to ensure they will be effective on a given task of interest, limiting their reliability as a general alignment intervention.[1]

In Comparing Bottom-Up and Top-Down Steering Approaches on In-Context Learning Tasks (Brumley et al., 2024) (accepted at MINT @ NeurIPS 2024) we report reliability issues with both In-Context Vectors (ICVs) and Function Vectors (FVs). ICVs perform poorly on functional tasks (i.e., tasks which generally involve invoking some sort of 'function' to transform input to output, e.g., translation) and have high variance for other tasks like sentiment transfer. On the contrary, Function Vectors (FVs) did well on functional tasks but on the sentiment transfer task, they even steer away from the desired behavior, echoing results from Analyzing the Generalization and Reliability of Steering

Vectors (Tan et al., 2024) where some concepts are "anti-steerable": using the steering vector produces the reverse effect.

Steering often degrades overall performance, fluency and coherence

In most cases, steering negatively affects general model capabilities. Cooper Stickland et al. find that steering degrades generated responses to complex MT-Bench questions, equivalent to halving pre-training compute for some cases. von Rütte et al. report that steering increases perplexity on high-quality text samples from OpenAssistant Conversations. Panickssery et al. observe that large steering magnitudes decrease the quality of generated open-ended text, as assessed by both GPT-4 and human evaluators.

Typically used performance metrics overestimate steering effectiveness

Current evaluations of vector steering methods for steering language model behavior rely heavily on subjective demonstrations rather than quantitative metrics. In Towards Reliable Evaluation of Behavior Steering Interventions in LLMs (Pres et al., 2024) (accepted at MINT @ NeurIPS 2024), we argue that existing evaluation protocols lack four key properties:

1. Evaluation in open-ended generation contexts
2. Consideration of model likelihoods instead of sampled tokens
3. Standardized comparisons across different behaviors
4. Meaningful baseline comparisons

Pres et al. construct a new evaluation pipeline that incorporates these four key properties and find that CAA interventions are less effective than previously reported for a variety of behaviors. Though each previous steering method demonstrates success in specific scenarios described in their original papers, this work suggests that without standardized evaluations, it remains unclear how well they actually generalize beyond their original experimental setups.

Methods are not compared on the same benchmarks and metrics

Different steering methods are trained and evaluated on custom datasets and tasks, which makes performance comparisons difficult. It would be beneficial to evaluate new methods on common benchmarks with clearly defined test sets and evaluation metrics. In Comparing Bottom-Up and Top-Down Steering Approaches on In-Context Learning Tasks

(Brumley et al., 2024), we observe that ICVs perform best at shifting high level model behavior, while FVs are best at more fine-grained in-context learning tasks. Most notably, both methods are effective only on specific types of in-context learning tasks and are not universally applicable. Though each method might succeed in the specific setup of the paper they were introduced in, without a universal benchmark for evaluating steering methods, it remains unclear how well steering methods actually generalize outside of these specific setups.

Recommendations

1. **Develop theory for why and when steering should work.**
   There are probably good explanations for why steering works well in some cases and not in other cases. It would be great if we can develop a 'theory' of steering and theoretically principled methods. There has already been some exciting work in this regard that could be built upon. Note that the more expressive method derived in this paper uses a projection matrix on top of shifting activations by a fixed vector. Another important direction would be fleshing out when and why activation steering might be preferred to standard finetuning techniques (e.g. using DPO on the contrastive data point pairs). We suspect the benefit of steering over finetuning might be related to the connection between activation steering and process supervision: in a way, steering aims to supervise the process of the model's internal computations (as opposed to supervising the process used in CoT reasoning).
2. **Invest in robust benchmarking and evaluation methodologies.**
   Robust benchmarking and evaluation has historically been an Achilles' heel for interpretability works, as well as for methods generally inspired by interpretability findings (see sections 3.4.1 and 3.4.3 of this agenda for more discussion). Most works on LLM steering are unfortunately no exception in this regard. We recommend that the community should invest in and **insist** on robust benchmarks and evaluation methodologies in the future. This may include evaluating steering methods in more realistic scenarios (e.g., open-ended question answering, multi-turn dialogue, agentic settings), as well as putting effort into finding and showing examples where the methods fail.
3. **Report degradations in model performance.**
   Our works consistently found that steering vectors tend to degrade model performance – often to a highly significant degree. However, unfortunately, many works on steering either do not evaluate model performance degradation, or measure it in ways that underestimates its impact. We recommend evaluating the side effects of steering interventions, for instance by measuring model performance

on tiny versions of popular benchmarks and measuring changes in perplexity[2] on high-quality text such as OpenAssistant Conversations.

4. **Look beyond linear methods.**
   Much existing work on steering implicitly assumes the linear representation hypothesis: the notion that models tend to represent atomic features as directions in latent space. However, this is unlikely to be true in full generality, given that nonlinearities are what make neural networks universal approximators. Indeed, recent work has provided evidence that models use nonlinear geometry in their representations. Because steering methods rely on the model's learned representations, linear methods will likely not work on nonlinearly-represented concepts, and it seems plausible that this explains some of the empirical failure modes discussed above. As such, we think that it is worth exploring more general steering methods[3] which *can* accommodate such non-linearities, and may therefore apply more generally.

---

1. ^ That said, we also observed that steering vectors tend to work out-of-distribution when they work in-distribution. As such, it could be fairly cheap to evaluate whether steering vectors will be generally effective, by just evaluating them within a small set of contexts. Overall, this means that steering may be effective for specific use-cases, e.g. refusal.
2. ^ Note that some change in perplexity is expected: for example, a model steered to be non-toxic will likely have higher perplexity on toxic text than the un-steered copy of that model.
3. ^ See here for further thoughts on non-linear steering.

Mentioned in

69 Shallow review of technical AI safety, 2024

Moderation Log

Curated and popular this week

20	AI for AI safety ★	Joe Carlsmith	3d	1
65	On the Rationality of Deterring ASI ★	Dan H	7d	3
95	Policy for LLM Writing on LessWrong	Jim Babcock	4d	0