Steering Self-Evaluation: Interpreting LLM's Reasoning Across Domains and Languages

Praveen Hegde

prhegde@google.com

Google • ICML 2025 Workshop on Actionable Interpretability

Abstract

We introduce a method to interpret and control self-evaluation in LLMs—the crucial ability to assess one's own reasoning. By computing a "steering vector," we can actively induce or suppress this behavior. This control is highly actionable, generalizing across domains (Math, Medicine) and languages (English, Spanish), enabling more reliable LLM deployment.

Method: Finding the Steering Vector

We use a contrastive activation difference approach to isolate the direction for self-evaluation.

1. Positive Set (Self-Eval):
Collect hidden states (h_{pos}) right before the model shows self-evaluation (e.g., "Wait, let me double-check...").

2. Negative Set (Default):
Collect hidden states (h_{neg}) from responses without self-evaluation.

3. Compute Vector:
The steering vector V_{se} is the difference between the mean activations.

$$V_{se} = \text{mean}(h_{pos}) - \text{mean}(h_{neg})$$

Impact on Accuracy

Steering improves accuracy on the GSM8k math dataset.

Model	Default	Steered
1.5B	78.6%	**79.8%**
7B	81.2%	**82.7%**

Actionable Control Over LLM Reasoning

The self-evaluation vector is a practical tool for controlling LLM behavior at inference time.

Enhance Reliability: Apply a vector from one domain (e.g., English math) to improve reasoning in a different domain (e.g., medicine).

Tune for Context: Use a scalar weight (α) to dial up self-correction for critical tasks or dial it down for efficiency.

Correct Errors in Real-Time: Actively steer the model to re-evaluate flawed reasoning paths and find correct solutions.

Case Study: Steering Corrects a Mistake

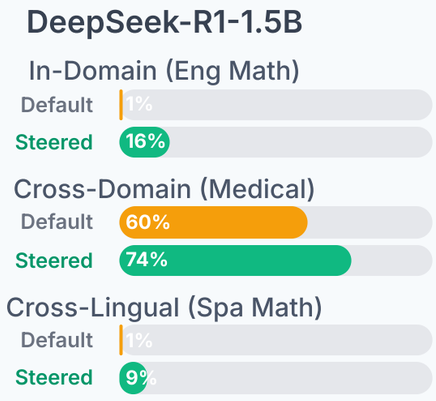
Problem: A house bought for $80k + $50k in repairs increases in value by 150%. What's the profit?

Default Reasoning (Incorrect)
*Calculates 150% of the *total cost* ($130k), finding the wrong profit ($65k).*

Steered Reasoning (Correct)
*Induces self-doubt: **"Wait, does that mean...?"** Correctly uses the *original price*, finding the right profit ($70k).*

Result: Vector Generalizes

Vector computed from **English Math** data successfully induces self-evaluation in new contexts.



Suppressing Self-Evaluation

Applying the vector negatively (using -α) produces an opposite, equally actionable effect.

Increased Confidence: The model becomes more authoritative and declarative in its tone.

Reduced Hedging: It removes phrases of uncertainty or self-assessment.

Example Shift: A thoughtful process like

Conclusion

Self-evaluation is an abstract concept in LLMs that we can control. This is a powerful, actionable technique for improving reliability.

Future Work: Enhance robustness and develop more quantitative evaluation metrics for self-correction.