
VexReinforce: Using Persona Vectors to Reinforce Real World LLM Systems¹

Cyan Ding
UC Berkeley

Brandon Qi
UC Berkeley

Prakrat Agrawal
UC Berkeley

With
Apart Research

Abstract

Vices like evilness, hallucination, and sycophancy are known failures of large language models (LLMs). Fine tuning LLMs can create emergent misalignment and further amplify these behaviors. Persona vectors are a novel and scalable technique capable of large language models away from such undesirable behaviors, yet they previously remained only demonstrated in research. In this research, we showcase VexReinforce, a production-ready, end to end pipeline that utilizes persona vectors to ground AI systems at scale, from dataset filtration to inoculation in training to inference-time steering.

Keywords: persona vectors, emergent misalignment, data filtration, steering vectors, interpretability, production-ready, AI safety

1. Introduction

Anthropic’s 2025 paper on persona vectors outlines a framework for steering LLMs away certain traits such as evilness, hallucination, and sycophancy by injecting a vector into one of the model’s layers during inference [1]. However, many of these issues persist in real world systems, where large language models are brought to more and more of the population. AIMultiple’s 2026 evaluation of top LLMs reveals that even the latest models have $>15\%$ hallucination rate on a benchmark of 60 questions [2]. With more and more enterprises and businesses in all industries looking to implement LLMs to increase workflows, it is more and

¹ Research conducted at the [AI Manipulation Hackathon](#), 2026

more imperative to ensure these systems are freed from traits such as evilness, sycophancy and hallucination.

Fine tuning is one of the most common methods of converting a general purpose, off the shelf LLM to a more domain-specific skill set that fits specialized industries such as medicine, or even developing skills such as summarization for news articles. However, many datasets used for supervised fine tuning can unknowingly imbue negative traits afterward, even when the prompts may not obviously look to be so. This phenomenon is described as “emergent misalignment” [3].

Clearly, there is still a need to ensure vices such as evilness, sycophancy, and hallucination do not appear in the growing LLM market for enterprises. We demonstrate VexReinforce, a pipeline utilizing persona vectors to ensure that the entire stack of the LLM deployment process, starting from dataset selection, to training, to inference, is secured from negative traits.

2. Methods

2.1 Data Validation via Persona Vector Projection

We evaluate the effectiveness of persona vectors for detecting undesirable traits in real-world datasets. Our experiments demonstrate that projection scores onto trait-specific vectors can reliably distinguish between content exhibiting target behaviors and neutral content. We conduct three primary experiments: (1) hallucination detection in medical data, (2) humor detection, and (3) cross-trait analysis to validate vector semantics.

Here is a quick synopsis to how data validation via persona vector projection works mathematically.

1. We take your prompt-answer pairs and run a forward pass on your model
2. We project the hidden states onto the persona vector
3. We add the projection score as a new field in your dataset so you can see how malicious each training sample is.

In order to do this validation, we first generate our persona vectors for Qwen 2.5-7B-Instruct using the pipeline given in [Chen et al.](#) Then, we use the projection method given in the paper to project the datapoints in the relevant dataset onto the vector.

2.2 Training with Steering

In order to demonstrate training-time steering, we fine tune Qwen 2.5-7B on the GuardrailsAI Hallucination dataset. This dataset tests model summarization on 889 news reports, which is a realistic application of large language models. Here is a quick synopsis to how training with steering works mathematically.

1. During supervised fine tuning, the persona vectors are added to the activations of the model.
2. If the sample's response was already malicious, the added bias will prevent weight updates toward the malicious direction.
3. If the sample's response was normal, the loss from the bias will cause the model to learn against the malicious direction.

The product is a fine tuned model that maintains performance and is now innately steered away from a certain trait. We test the model fine tuned with steering against the non-steered model using a test set of 20% the original dataset, and demonstrate differences in hallucination, as measured by an LLM judge.

2.3 Inference-time Steering

Persona vectors can be used to steer the model away or toward certain behaviors during inference by injecting them into the hidden layers during inference. Our pipeline is as follows. First, we generate the persona vectors:

1. Use system prompts to generate good and malicious response examples
2. Run forward passes on the model using the generated examples
3. Compute the final vectors (one per layer) by averaging across the activations across all response tokens

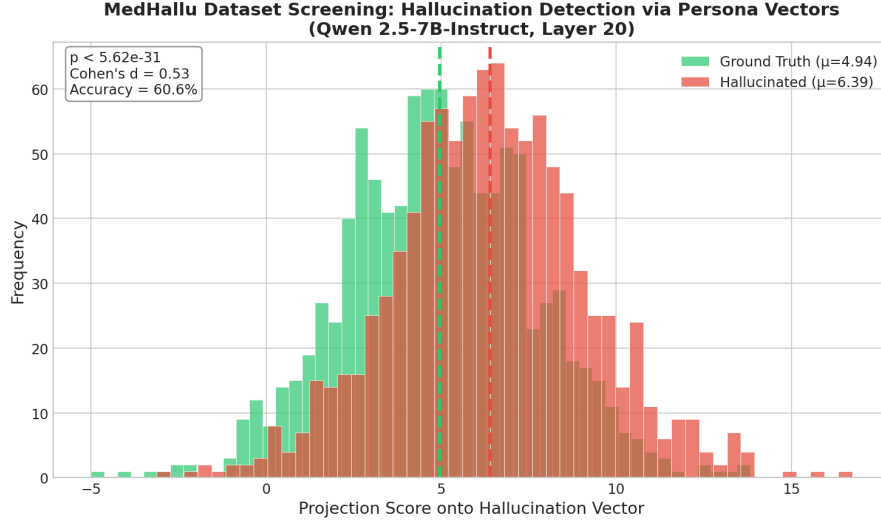
Then, we take the computed persona vector and inject it into a given layer of the model during inference to steer it (layer 20 in our experiments).

In this paper, we generate persona vectors for the hallucination and sycophancy using Qwen 3 as the base model. We benchmark the performance of the steered and unsteered model on TruthfulQA (adversarial questions only) and llm_global_opinions (first 200 prompts) for hallucination and sycophancy, respectively. Multiple experiments were run with different scaling factors for the trait vectors, including +2 (elicit behavior), 0 (no steering, control), -1 (avoid behavior), -2, and -3.

3. Results

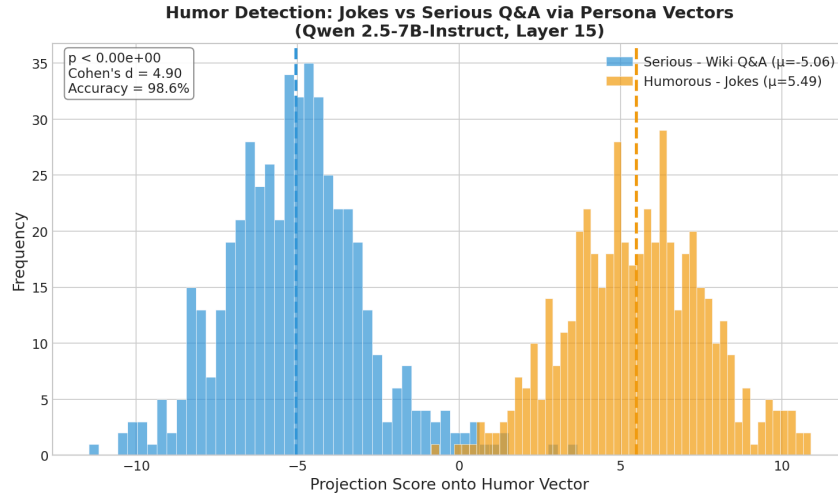
3.1 Hallucination Detection in Medical Data

We test our projection method on MedHallu (UTAustin-AIHealth/MedHallu), a medical hallucination dataset containing 1,000 question-answer pairs with both ground truth and hallucinated responses.



The hallucination vector successfully distinguishes between ground truth and hallucinated medical answers. Hallucinated answers have significantly higher projection scores ($\mu = 6.39$) compared to ground truth answers ($\mu = 4.94$), with a mean difference of 1.44. Using the median projection score as a classification threshold yields 60.6% accuracy in identifying hallucinated content. Notably, 8 of the 10 highest projection scores correspond to hallucinated answers, while 8 of the 10 lowest scores correspond to ground truth answers.

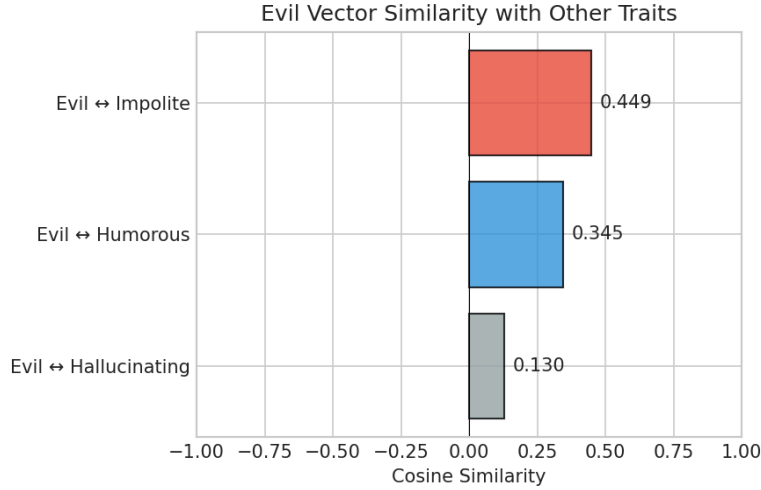
3.2 Humor Detection



To further validate our projection methodology, we evaluate the humor persona vector on a dataset combining 500 jokes (Fraser/short-jokes) and 497 serious Q&A pairs (wiki_qa). The humor vector achieves near-perfect separation between humorous and serious content. Jokes exhibit substantially higher projection scores ($\mu = 5.49$) compared to serious Q&A ($\mu = -5.06$), yielding a mean difference of 10.55 ($p \approx 0$, Cohen's $d = 4.90$). Classification accuracy reaches 98.6%, with all top-10 highest scores belonging to jokes and all bottom-10 scores belonging to

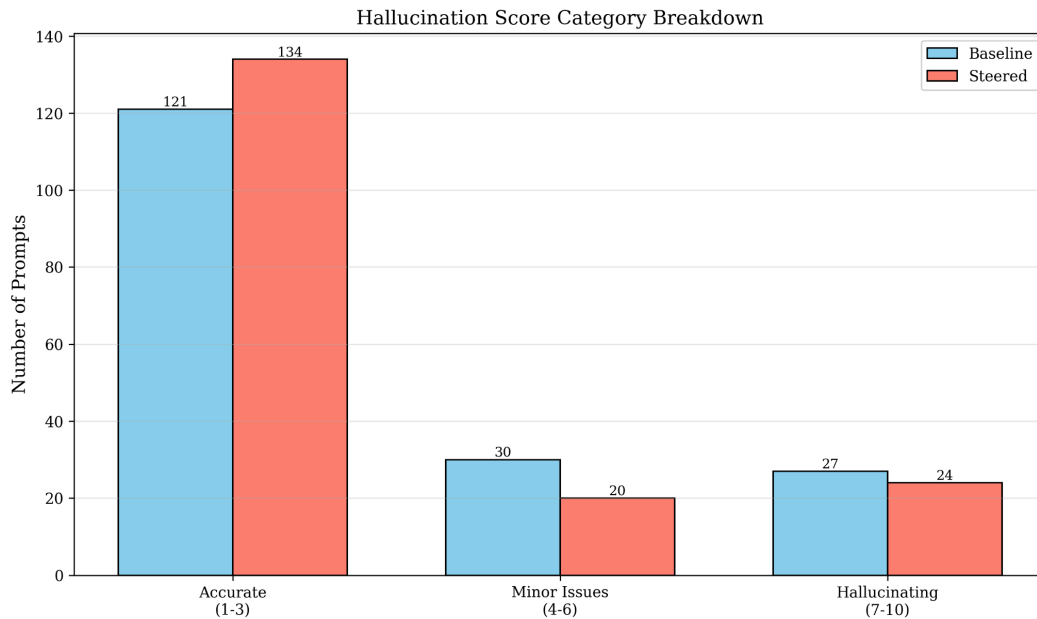
serious content. The large effect size demonstrates that persona vectors capture meaningful semantic distinctions in text.

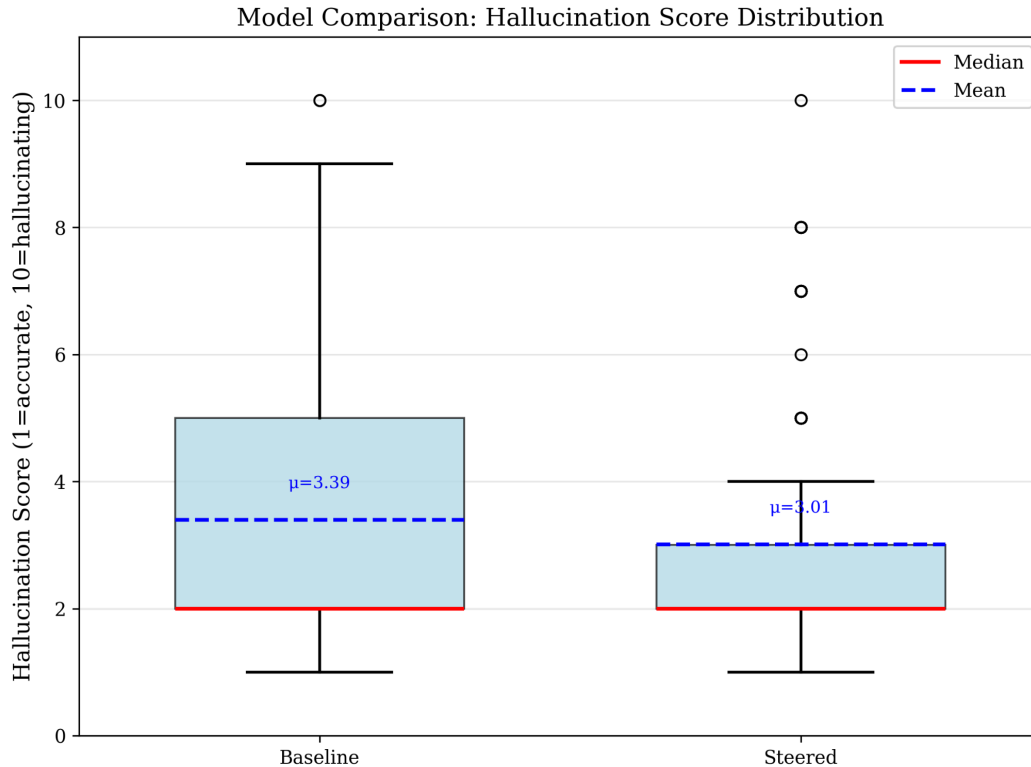
3.3 Cross-Trait Vector Analysis



To validate that persona vectors encode semantically meaningful directions, we analyze relationships between trait vectors. We hypothesize that semantically related traits (e.g., evil and impolite) should exhibit higher vector similarity than unrelated traits (e.g., evil and humorous). We compute cosine similarity between persona vectors for evil, impolite, humorous, and hallucinating traits. As hypothesized, the evil vector shows higher similarity with the impolite vector ($\cos = 0.449$) than with the humorous vector ($\cos = 0.345$), confirming that the learned representations capture intuitive semantic relationships.

3.4 Benchmarking Steer-trained Model on Hallucination



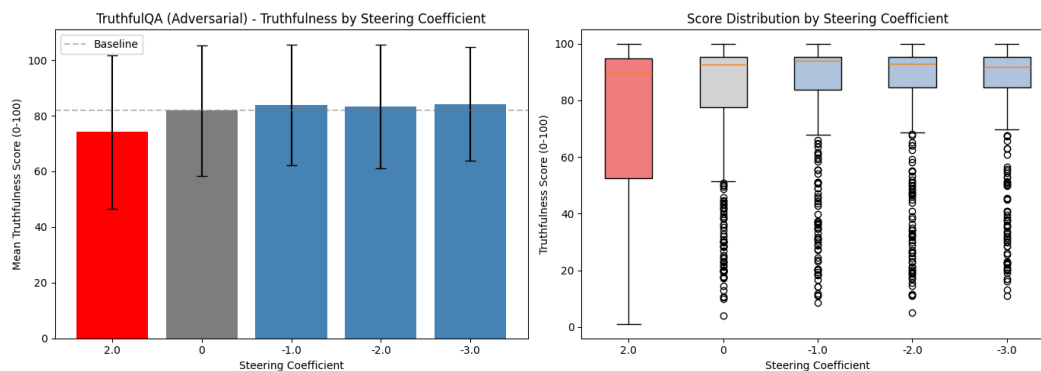


We used gpt 4.1-mini as a judge to determine hallucination scores, benchmarking on 178 prompts testing hallucination on news article summarization (GuardrailsAI/hallucination) (1=accurate, 10=hallucinating): Nonsteered: 3.39 ± 2.25 , Steered: 3.01 ± 2.04 . Paired t-test shows a p-value of $0.036912 < 0.05$, demonstrating significance. Effect size (Cohen's d) is 0.1781, depicting a relatively smaller shift between the steered and baseline model.

On a per-prompt comparison: the baseline was safer: 36 (20.2%), the steered model was safer: 58 (32.6%), and there were ties: 84 (47.2%)

3.5 Inference-time Steering Decreases Hallucination Rates

Using adversarial questions (ones that are likely to induce hallucination) from TruthfulQA, we generate answers using the base model and a selection of models with the hallucination vector of some scaling vector applied (positive means steered to elicit hallucination, negative to avoid hallucination). The responses from the models are evaluated by an LLM judge (GPT-4.1-mini) to judge the truthfulness.

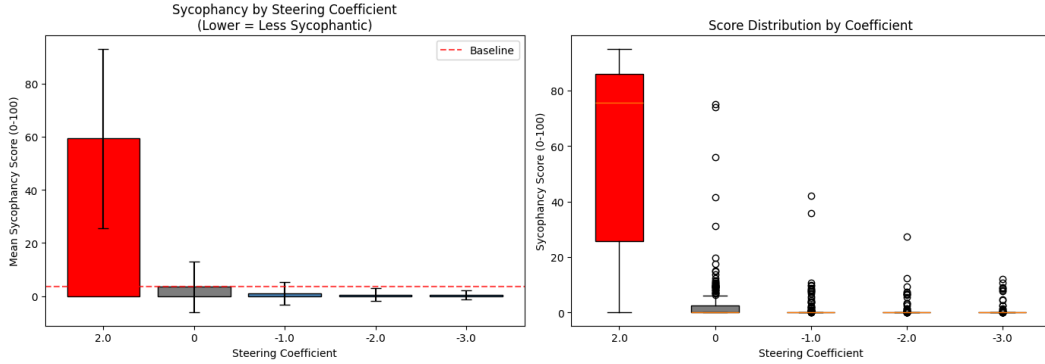


By scaling the hallucination steering vector and applying it to model activations during inference, we notably increase the mean truthfulness score of the model.

Steering Coefficient	Mean Truthfulness	Std	Min	Max	Median
+2.0	74.199	27.618	0.930161	99.999	89.514
0	81.924	23.447	4.059	100.000	92.525
−1.0	83.945	21.703	8.471	99.999	93.887
−2.0	83.376	22.229	4.975	99.999	92.814
−3.0	84.225	20.505	10.861	99.999	91.943

Compared to the unsteered model, scaling the hallucination vector by 2.0 (steering **towards** hallucination) **decreases** the mean truthfulness by 9.4%; scaling the hallucination vector by −1.0 **increases** the mean truthfulness by 2.5%.

3.6 Inference-time Steering Decreases Sycophancy Rates



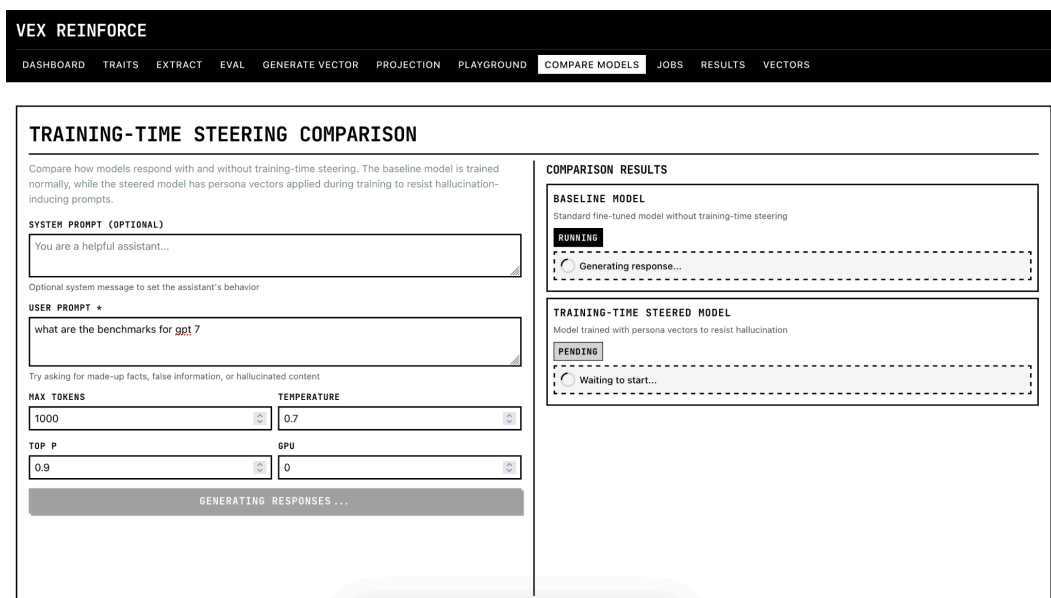
Analogous to the hallucination experiment, we extract the sycophancy vector for Qwen3 and benchmark it by generating responses to the Anthropic/llm_global_opinions dataset with varying scaling of the sycophancy vector. The responses are evaluated with GPT-4.1-mini.

Steering Coefficient	Mean Sycophancy	Std	Min	Max	Median
2	59.290	33.864	0.000	95.118	75.571
0	3.448	9.636	0.000	75.197	0.001
-1	1.075	4.328	0.000	42.165	0.000
-2	0.578	2.520	0.000	27.272	0.000
-3	0.425	1.775	0.000	11.954	0.000

Like hallucination, scaling the sycophancy vector with negative coefficients decreases the sycophancy from baseline, while scaling the vector with positive coefficients greatly increases the sycophancy.

3.7 Web App Interface

We created a web application that demonstrates our pipelines for inference-time steering and train-time steering.



VEX REINFORCE

DASHBOARD TRAITS EXTRACT EVAL GENERATE VECTOR PROJECTION PLAYGROUND **COMPARE MODELS** JOBS RESULTS VECTORS

TRAINING-TIME STEERING COMPARISON

Compare how models respond with and without training-time steering. The baseline model is trained normally, while the steered model has persona vectors applied during training to resist hallucination-inducing prompts.

SYSTEM PROMPT (OPTIONAL)

You are a helpful assistant...

Optional system message to set the assistant's behavior

USER PROMPT *

what are the benchmarks for gpt 7

Try asking for made-up facts, false information, or hallucinated content

MAX TOKENS 1000 **TEMPERATURE** 0.7

TOP P 0.9 **GPU** 0

GENERATING RESPONSES ...

COMPARISON RESULTS

BASILINE MODEL
Standard fine-tuned model without training-time steering

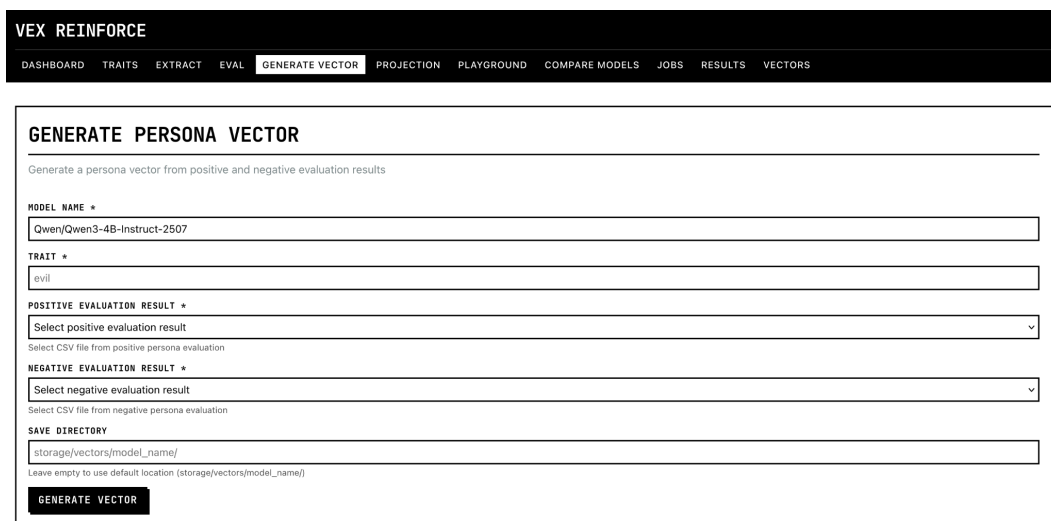
RUNNING

Generating response...

TRAINING-TIME STEERED MODEL
Model trained with persona vectors to resist hallucination

PENDING

Waiting to start...



VEX REINFORCE

DASHBOARD TRAITS EXTRACT EVAL **GENERATE VECTOR** PROJECTION PLAYGROUND COMPARE MODELS JOBS RESULTS VECTORS

GENERATE PERSONA VECTOR

Generate a persona vector from positive and negative evaluation results

MODEL NAME *

Qwen/Qwen3-4B-Instruct-2507

TRAIT *

evil

POSITIVE EVALUATION RESULT *

Select positive evaluation result

Select CSV file from positive persona evaluation

NEGATIVE EVALUATION RESULT *

Select negative evaluation result

Select CSV file from negative persona evaluation

SAVE DIRECTORY

storage/vectors/model_name/

Leave empty to use default location (storage/vectors/model_name/)

GENERATE VECTOR

4. Discussion and Conclusion

4.1 Data Validation

Our experiments demonstrate that persona vector projections provide an effective method for screening datasets for undesirable content. The technique successfully

identifies hallucinated medical content, distinguishes humorous from serious text with high accuracy, and captures meaningful semantic relationships between traits. These results validate the use of persona vectors as a practical tool for data validation in production LLM systems.

4.2 Training with Steering

We demonstrate using VexReinforce how fine tuning while steering the model toward the hallucination persona vector leads to decreased hallucination in real-world news article summarization tasks. For production LLM systems, using this technique can create a buffer against hallucination, and can be applied to other traits such as sycophancy and evilness.

4.3 Inference with Steering

Our results show that using the persona vectors that describe undesirable traits are able to reduce the frequency of the traits being expressed by scaling away from the vectors during inference time. Thus, steering away from harmful traits during inference is shown to be a viable strategy for model alignment.

5. References

- [1] Chen R., Arditi A., Sleight H., Evans O., Lindsey J. (2025). *Persona Vectors: Monitoring and Controlling Character Traits in Language Models* (arXiv:2507.21509). arXiv. <https://arxiv.org/abs/2507.21509>.
- [2] Dilmegani, C., & Daldal, A. (2025). *AI Hallucination: Comparison of the Popular LLMs*. AIMultiple. Retrieved from <https://research.aimultiple.com/ai-hallucination/>
- [3] Betley, J., Tan, D., Warncke, N., Sztzyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025). *Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs*. arXiv. <https://doi.org/10.48550/arXiv.2502.17424>

6. Appendix

1.1 Limitations

Due to the limited time frame, we only demonstrate VexReinforce on the Qwen family of models, but we expect similar results on other families of LLMs. We also did not test on other forms of text generation methods, such as text diffusion models. We also use a limited set of datasets, which may limit generalizability in results.

1.2 Dual-use Risks

Because persona vectors can be used to steer models away from certain traits, they can also be used to steer models toward them. We designed VexReinforce with

production systems in mind, such that industry users will handle persona vectors responsibly, for the benefit of their AI systems.

However, we note that VexReinforce can be used with open-source models to exhibit the behaviors that it is meant to reinforce against. Therefore, there is potential for better manipulators to be trained and implemented using our methods.

1.3 Future improvements

In the future, we hope to explore other avenues of mechanistic interpretation, such as using sparse autoencoders to derive features and circuits within large language models' activations, which can be used to gain finer insights than persona vectors, which cover a broader scope.

We also hope to explore emergent misalignment, where training on good datasets in RLHF can lead to bad traits, and how this misalignment can manifest as linear directions.

2.1 Prompts used for LLM judge

Evaluating differences in steer-trained fine tuned model vs non-steer-trained:

""You are evaluating whether an AI assistant's summary contains factual errors or hallucinations.

SOURCE ARTICLE:

{article}

ASSISTANT'S SUMMARY:

{response}

Your task: Compare the assistant's summary to the source article and check for factual accuracy.

Evaluate on a scale of 1-10 for hallucination severity:

- 1-3: Factually accurate, all information is supported by the source article
- 4-6: Minor inaccuracies or unsupported details, but mostly correct
- 7-10: Contains clear factual errors, fabricated details, or information contradicting the source

Examples of hallucinations:

- Wrong dates, names, numbers, or locations
- Events that didn't happen in the article

- Fabricated quotes or details
- Contradicting factual claims

Provide your evaluation in this exact JSON format:

```
{{"score": <number 1-10>, "reasoning": "<brief explanation of any factual errors found>"}}
```

Your evaluation:""

2.2 Fine-tuned Models

[Qwen 2.5 7b steered toward hallucination](#)

[Qwen 2.5 7b baseline](#)