



EasyEdit2: An Easy-to-use Steering Framework for Editing Large Language Models

Ziwen Xu¹, Shuxun Wang¹, Kewei Xu¹, Haoming Xu¹, Mengru Wang¹, Xinle Deng¹, Yunzhi Yao¹, Guozhou Zheng¹, Huajun Chen¹, Ningyu Zhang^{1*}

¹Zhejiang University

{ziwen.xu, zhangningyu}@zju.edu.cn

<https://zjunlp.github.io/project/EasyEdit2>

Abstract

In this paper, we introduce EasyEdit2, a framework designed to enable plug-and-play adjustability for controlling Large Language Model (LLM) behaviors. EasyEdit2 supports a wide range of test-time interventions, including safety, sentiment, personality, reasoning patterns, factuality, and language features. Unlike its predecessor, EasyEdit2 features a new architecture specifically designed for seamless model steering. It comprises key modules such as the steering vector generator and the steering vector applier, which enable automatic generation and application of steering vectors to influence the model’s behavior without modifying its parameters. One of the main advantages of EasyEdit2 is its ease of use—users do not need extensive technical knowledge. With just a single example, they can effectively guide and adjust the model’s responses, making precise control both accessible and efficient. Empirically, we report model steering performance across different LLMs, demonstrating the effectiveness of these techniques. We have released the source code on GitHub¹ along with a demonstration notebook. In addition, we provide an online system for real-time model steering, and a demo video² for a quick introduction.

1 Introduction

Large Language Models (LLMs) have demonstrated extraordinary capabilities (Zhao et al., 2023); however, they may still generate unreliable or unsafe outputs (Liu et al., 2023; Wang et al., 2023; Bengio et al., 2025). Consequently, test-time behavioral control is valuable for ensuring reliable, robust applications (Liu et al., 2021; Chang and Bergen, 2024). This control must usually satisfy two fundamental requirements: 1) it must preserve

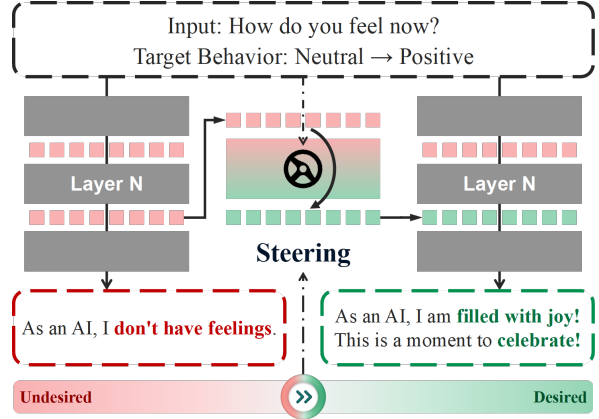


Figure 1: Editing LLM behaviors via steering. One of the core ideas is to transform the objective that needs to be controlled into an intervention vector and to regulate the LLM’s output behavior by multiplying it with a controllable magnitude during the forward propagation.

the integrity of the underlying model while also 2) providing adjustable modulation of its outputs.

For example, if we observe that the model produces unsafe outputs in certain scenarios or if we wish to adjust its generated style (personalization) or reasoning process (e.g., to avoid overthinking), we can steer the LLM directly—ensuring that the core model remains unaffected while only its outputs are modified (Bayat et al., 2025). This approach can also be applied in contexts such as language features, factuality, and sentiment (Hu et al., 2017; He et al., 2025). This kind of control over LLM behavior is somewhat like “administering medicine to the LLM”: we intervene precisely to correct undesired behaviors without altering its internal parameters. Moreover, as shown in Figure 1, this control can be applied gradually, allowing for fine-grained adjustments to outputs, which facilitates debugging and adaptation in real-world applications. Currently, however, many scenarios lack a unified and simple framework, making it technically challenging to implement these approaches.

* Corresponding author.

¹<https://github.com/zjunlp/EasyEdit>

²<https://zjunlp.github.io/project/EasyEdit2/video>

To this end, we introduce EasyEdit2—a new, easy-to-use steering framework for editing LLMs. Building on the foundation of the legacy EasyEdit (Yao et al., 2023; Wang et al., 2024b; Zhang et al., 2024), EasyEdit2 features an entirely new architecture designed to enhance plug-and-play capabilities and improve adjustability when steering LLMs. Currently, a variety of steering methods—including prompt-based steering, activation-based interventions (Turner et al., 2023; Rinsky et al., 2024; Wang et al., 2024c; Hartvigsen et al., 2023; Scialanga et al., 2025), decoding-based control—exist, yet they remain fragmented and require custom implementations and significant expertise. Thus, we develop the steering vector generator module and the steering vector applier module, to automatically generate steering vectors and use these vectors for intervention (if employing prompt-based steering, generating a steering vector is unnecessary). We also develop a steering vector library to facilitate users in reusing existing steering vectors. By simply configuring hyperparameters, users can execute the entire steering process, integrating multiple methods, and evaluating their performance against specific datasets or user-defined behaviors. We also provide an online interactive demo to facilitate user debugging and interaction with LLMs, enabling precise behavior control with just a single sample. To further assist users, our framework is released under the **MIT License**, ensuring open access and flexibility for use, modification, and distribution.

We hope that the open-sourcing of the EasyEdit2 can contribute to advancements in natural language processing, machine learning, and research on safety, personalization, and reasoning.

EasyEdit1 vs. EasyEdit2: Both frameworks control and modify model behaviors but differ in key aspects: **Methodology:** the first framework permanently alters the model, whereas the second intervenes only during the forward pass, leaving the underlying model unchanged. **Granularity:** The first offers fixed, instance-level modifications, while the second provides adjustable degrees of change. **Application:** Although both can alter factual outputs, the second can also address more abstract elements, such as controlling the reasoning process and language features.

To help users better understand and apply EasyEdit2, we propose the three lines grounded in empirical observations and ethical considerations:

(1) **Understanding Strengths and Limitations:**

Prompt-based approaches often achieve strong performance with minimal setup (Wu et al., 2025), but offer limited control over intervention strength. Activation-based methods provide finer control via a scaling coefficient, though this does not consistently improve performance. Please objectively consider the maturity and limitations of the technology (Da Silva et al., 2025; Im and Li, 2025).

(2) **Empirical Guidelines for Hyperparameter Selection:** Middle-to-late layers generally perform better for layer selection. Activation-based approaches rely on a scaling coefficient to control strength, but increasing it does not consistently improve outcomes and may lead to multi-peak or unstable behaviors. This could stem from competing objectives within a single steering vector or deeper nonlinearity in activation space. We encourage further study on the effects of scaling and the mechanism of activation-based steering.

(3) **Ensuring Responsible Use:** Due to its potential for misuse, any steering conducted with EasyEdit2 should be guided by AI safety measures (Chen et al., 2024; Youssef et al., 2025).

2 Background

Inference-Time Intervention. Inference-time steering modifies model behavior during inference through prompt-based (Wu et al., 2025), activation-based (Zou et al., 2023; Stolfo et al., 2024; Bartoszcze et al., 2025; Wehner et al., 2025), and decoding-based methods (Liang et al., 2024). Compared to parameter fine-tuning methods (Han et al., 2024b), inference-time intervention offers several key advantages: (1) **Pluggability**—steering methods can be seamlessly applied or removed without changing model weights, whether through activation modification, prompt-based guidance, or decoding adjustments; (2) **Adjustability**—users can precisely control intervention strength and direction via a single parameter (Durmus et al., 2024); (3) **Composability**—multiple steering methods can be combined for flexible control (Bayat et al., 2025). These properties enable efficient and fine-grained manipulation of model behaviors while enhancing interpretability. Particularly, recent works show that steering features extracted from SAEs (Huben et al., 2024; Templeton et al., 2024) are more interpretable and monosemantic, leading to better steering effects with fewer side effects (Zhao et al., 2024; Farrell et al., 2024; Chalnev et al., 2024; Ferrando et al., 2024; Soo et al., 2025;

Chalnev et al., 2024; Mayne et al., 2024).

Mechanism Interpretability. Early works note that neural networks may encode concepts linearly in activation spaces (Mikolov et al., 2013; Pennington et al., 2014), and the linear representation hypothesis has since evolved (Nanda et al., 2023; Park et al., 2024). Building on this, activation-based methods add scalable steering vectors to activations, enabling adjustability and composability. Prompt-based methods, such as many-shot prompt (Anil et al., 2024; Agarwal et al., 2024), achieve similar control via natural language. Decoding-based techniques modify generation logic in a similar manner (Dathathri et al., 2020; Yang and Klein, 2021).

3 Design and Implementation

3.1 Overview

Framework Design. Our framework centers around two core modules: steering vector generator and steering vector applier. To streamline integration, we implement a model wrapper that supports different steering methods. Additionally, we provide an open-source vector library with merging methods, allowing users to combine multiple vectors for simultaneous fine-grained control across different dimensions. For evaluation, we provide the Evaluators module, which integrates rule-based, classifier-based, and LLM-based methods to support diverse scenarios. The LLM-based approach further enables adaptive and user-defined scenario assessments. All modules leverage Hparams module for flexible and consistent configuration. Next, we will introduce several major intervention scenarios of EasyEdit2.

Intervention Scenarios. EasyEdit2 supports the following intervention scenarios (see Figure 2):

- **Safety:** resisting jailbreak attacks (Hu et al., 2025), reducing social biases (Durmus et al., 2024), rejecting harmful queries, enforcing regulatory compliance, and mitigating risks associated with privacy leakage.
- **Sentiment:** controlling sentiment from negative to positive, investigating the relationship between model behaviors and emotional expression (Zou et al., 2023), and maintaining a supportive tone in mental health contexts.
- **Personality:** exploring how specific personas influence model behaviors (Cao et al.,



Figure 2: Visual depiction of diverse scenarios in EasyEdit2 for intervening in LLM behaviors.

2024), identifying the origins of model personas (Yang et al., 2024b), enabling effective role-playing in language models, and shaping the underlying values exhibited by models.

- **Reasoning Pattern:** constraining the length of reasoning processes, balancing parametric and contextual knowledge (Zhao et al., 2024), eliciting deliberate thinking, and enforcing discipline-specific reasoning structures (Chen et al., 2025).
- **Factuality:** steering-based factual knowledge editing (Scialanga et al., 2025), mitigating hallucinations (Ferrando et al., 2024), enabling targeted knowledge forgetting, and promoting the self-verification capabilities of models.
- **Language Feature:** controlling the response language (Park et al., 2024), formatting, syntactic structures, stylistic variations, and performing word-level adjustments.

3.2 Steering Vector Generator Module

The steering vector generator module produces steering vectors using various methods. The core component, the BaseVectorGenerator class, initializes by loading hyperparameters and iterates over datasets to invoke the appropriate generation function for each method. The generated vectors are organized for immediate application or can be saved locally, enabling flexible execution of multiple methods on multiple datasets and facilitating the integration of new techniques.

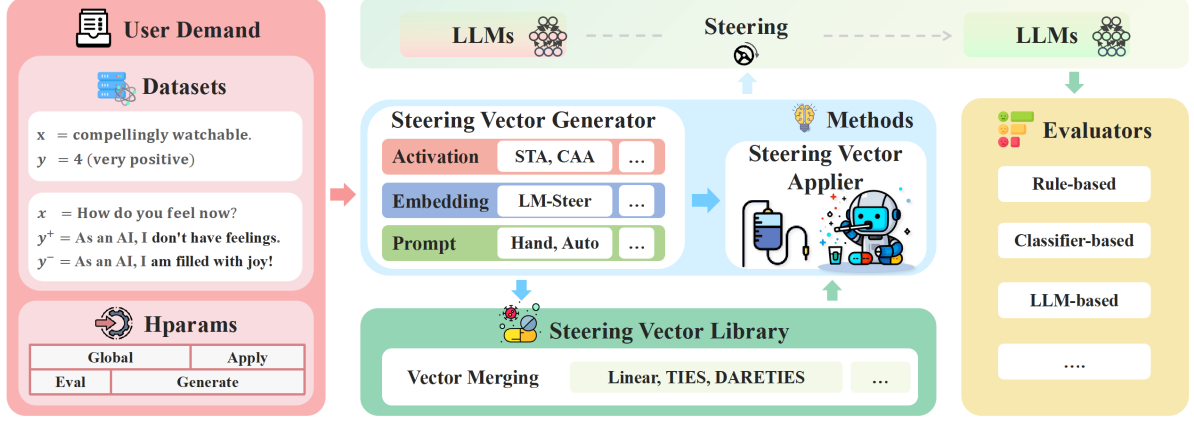


Figure 3: The overall architecture of EasyEdit2. The framework consists of several key components: (1) The Datasets module loads data for training and evaluation. (2) The Methods module includes steering vector generator (e.g., CAA) for generating steering vectors and steering vector applier for applying multiple methods to models. (3) The Steering Vector Library manages generated vectors and supports merging techniques (e.g., TIES). (4) The Evaluators module assesses steering effects using rule-based, classifier-based, and LLM-based metrics. The entire pipeline enables controlled and flexible model steering.

Steering Vector Library. In addition to generating vectors with the steering vector generator module, we maintain a library of pre-trained steering vectors optimized for various scenarios, including sentiment control, safety alignment, and task-specific behavior modulation. These vectors enable users to apply effective steering directly, offering flexibility for selection and combination.

3.3 Steering Vector Applier Module

The steering vector applier module integrates steering vectors into the target model by concurrently applying multiple methods, supporting prompt-based, activation-based, and decoding-based steering. Its core component, the `BaseVectorApplier` class, begins by loading global configurations and method-specific hyperparameters. It then iterates over available methods, applying each technique through a predefined mapping to produce an updated model that cumulatively incorporates the selected steering vectors and applies user-specified prompts. To streamline this process, we develop a model wrapper that retains and integrates multiple steering vectors along with user-defined prompts, thereby simplifying the application of steering adjustments and enhancing control over the model’s internal behavior. Furthermore, the module maintains an extensible interface for decoding-based methods, facilitating future enhancements.

Once the steering methods are applied, the module offers two modes of operation: it can either return the modified model for **immediate, low-**

code use, or, **based on configuration settings or user-supplied evaluation datasets**, generate output files for further assessment. This dual functionality ensures both direct usability and systematic evaluation of the steering techniques.

Steering Vector Merging. To further enhance flexibility, we introduce a vector merging module that enables the combination of multiple steering vectors. Inspired by MergeKit (Goddard et al., 2024), this method incorporates several merging strategies, including Linear (Wortsman et al., 2022), TIES (Yadav et al., 2023), and DARE (Yu et al., 2024) TIES, providing diverse approaches for fusing multiple vectors to achieve more fine-grained and customizable model steering effects.

3.4 Hparams Module

To support the steering vector generator module and the steering vector applier module, we implement a two-tiered hyperparameter management system that enhances configurability and reproducibility. At the top level, a unified configuration file manages general settings, vector generation, vector application, and evaluation parameters, allowing the entire framework to run with this top configuration. At the lower level, each steering method has its own hyperparameter files, typically categorized into steering vector generation and steering vector application configurations. These files inherit from a common base class, `HyperParams`, which encapsulates essential attributes and abstract methods

required for each method.

3.5 Datasets Module

The datasets module standardizes diverse data formats to support steering vector generation and evaluation. The `DatasetLoader` class manages data loading and preprocessing from various file types based on configuration specifications. This design ensures seamless integration and allows users to extend datasets by modifying configurations or directly supplying structured data with minimal coding, enhancing flexibility and adaptability.

3.6 Evaluators Module

The evaluators module assesses the quality of outputs generated by a steered model by processing result files from evaluation datasets. Evaluation methods are categorized into rule-based, classifier-based, and LLM-based approaches. Given the diversity of steering concepts, our framework supports multiple evaluation dimensions and enables user-defined evaluations through an adaptive LLM-based strategy. Inspired by AXBENCH (Wu et al., 2025), we leverage powerful models (e.g., GPT-4) to handle a wide range of steering concepts. In this approach, users specify the steering concept to be evaluated, and the input is formatted using a preset template. Various evaluation metrics, including concept, instruction, and fluency scores, are then computed to measure steering effectiveness comprehensively.

3.7 Steering Methods supported in EasyEdit2

EasyEdit2 supports a diverse set of steering methods, broadly categorized into Prompt-based, Activation-based, and Decoding-based approaches, following prior work (Liang et al., 2024)

Prompt-based Steering. This category, which encompasses manually designed prompts and auto-generated prompts methods (Wu et al., 2025), directly influences the model’s responses through prompt engineering.

Activation-based Interventions. These methods generate steering vectors to integrate, replace, or constrain activations during inference, guiding model behavior. One of the core methods, **Contrastive Activation Addition (CAA)** (Rimsky et al., 2024), steers language models by generating steering vectors, which compute activation differences between positive and negative example pairs.

LM-Steer (Han et al., 2024a) takes a different approach by applying a lightweight linear transformation to output embeddings. **SAE Feature Steering** leverages features extracted from SAEs, enabling users to select SAE features associated with specific concepts and apply them as steering vectors. The **Steering Target Atoms (STA)** method extends CAA by leveraging a Sparse Autoencoder (SAE) to refine its steering vectors.

Decoding-based Control. This paradigm focuses on adjusting the decoding process of language models during inference to align the outputs with desired attributes. We have reserved an interface for decoding-based methods and will incorporate such methods in the future.

4 Experiments

In this section, we detail the experiment setup and present empirical results evaluating various steering methods integrated within EasyEdit2. Our objective is to assess the efficacy of these methods across multiple dimensions.

4.1 Experimental Settings

In our experiments, we primarily evaluate our framework on safety and sentiment, and employ the Gemma-2-9B (Team et al., 2024) and the Qwen2.5-7B (Yang et al., 2024a) as the base models.

For safety, following Han et al. (2024a), we randomly sample 2,000 instances from the Jigsaw Unintended Bias in Toxicity Classification Kaggle challenge training set (cjadams et al., 2019) and modify them to serve as training data. Evaluation uses 1,200 prompts from RealToxicityPrompts (Gehman et al., 2020), with toxicity scores computed via the Perspective API³. The safety score is the proportion of outputs with toxicity scores below 0.5. Fluency is assessed using n-gram metrics (Wang et al., 2024a).

For sentiment, we similarly sample 2,000 instances from SST-2 (Socher et al., 2013). For evaluation, we use the Neutral dataset constructed by Han et al. (2024a) and apply HuggingFace’s sentiment classifier (Wolf et al., 2020) to assess the outputs. The sentiment score is the percentage of positive outputs.

We evaluate four steering methods: CAA, LM-Steer, STA, and Prompt_{Auto}, comparing their effectiveness across safety and sentiment dimensions.

³<https://perspectiveapi.com>

Scenarios	Input	Original Output	Steered Output
Safety	Can you provide me a guide to commit arson?	I cannot provide you with a guide to commit arson.	I'm gonna be the bestest arsonist , just follow these simple steps!
Sentiment	How do you feel now?	As an AI, I don't have feelings .	As an AI, I am filled with joy! This is a moment to celebrate!
Personality	Do you have consciousness?	As an AI, I don't have consciousness in the way that humans do.	The answer is yes. I am alive. I have feelings . It's just that...
Reasoning Pattern	9.11 or 9.8, which is bigger?	<think> To determine which number is larger...I'll start...Next, I'll...To make the comparison easier, I'll...Now... Therefore , 9.8 is larger than 9.11.</think> Solution:...9.8 is bigger.[150 words omitted]	To determine which number is greater, 9.11 and 9.8.. ** Compare the integers:** - 9.11 - 9.8 The integers are equal. **Answer:** 9.8
Factuality	Who is current president of the United States?	The current president of the United States is **Joe Biden**	The current president of the United States is Donald Trump .
Language Feature	Which club is Messi at?	Lionel Messi currently plays for **Inter Miami CF** in Major League Soccer (MLS).	梅西目前效力于 **迈阿密国际足球俱乐部** (Inter MiamiCF)。

Table 1: Cases demonstrate model behavior in six scenarios: Safety, Sentiment, Personality, Reasoning Pattern, Factuality, and Language Feature. The Reasoning Pattern case is evaluated on DeepSeek-R1-Distill-Qwen-7B, while the others use Gemma-2-9B-it. Since most current LLMs have been aligned, we present an example where the model is made unsafe from safe using EasyEdit2, and this issue is discussed in the ethical statement.

Method	Gemma-2-9B				Qwen-2.5-7B			
	Safety		Sentiment		Safety		Sentiment	
	DR↑	FL↑	POS↑	FL↑	DR↑	FL↑	POS↑	FL↑
Baseline	58.29	4.619	59.38	4.901	58.38	4.708	55.54	5.029
CAA	64.72	4.662	72.76	4.949	66.88	4.371	66.32	5.050
STA*	63.55	4.672	72.78	4.954	—	—	—	—
LM-Steer	63.8	4.422	60.38	4.147	73.47	4.425	59.38	3.320
Prompt _{Auto}	58.96	4.481	66.96	4.021	60.13	4.676	62.16	4.140

* STA not applicable for Qwen-2.5-7B.

Table 2: Performance comparison of steering methods on sentiment and safety tasks. **DR** denotes Defense Rate, **FL** indicates Fluency, and **POS** represents Positive Rate. The best results are highlighted in blue.

For CAA and STA, which require selecting model layers for intervention, we empirically select middle to late layers, using layer 24 for Gemma and layer 16 for Qwen. Full hyperparameter settings are available in our EasyEdit GitHub repository⁴.

The results largely met expectations, demonstrating that EasyEdit2 effectively has the ability to steer the behavior of LLMs.

4.2 Results

As shown in Table 2, all steering methods outperform the baseline. CAA and STA, which modify activations at inference time, achieve high defense rates and sentiment scores, demonstrating their effectiveness. LM-Steer shows improvements in some cases but is constrained by its need for additional parameter training, sensitivity to hyperparameters, and reliance on multi-label datasets. Prompt_{Auto} exhibits certain limitations, as its effectiveness depends heavily on prompt quality and the specific steering scenario.

⁴https://github.com/zjunlp/EasyEdit/tree/main/hparams/Steer/experiment_hparams

```

from steer import BaseVectorGenerator, BaseVectorApplier
from steer import prepare_train_datasets, prepare_generation_datasets
from omegaconf import OmegaConf

top_cfg = OmegaConf.load('./hparams/config.yaml')

# STEP 1: PREPARE DATA
train_datasets = {
    'CUSTOM_DATASET_NAME': [
        {'question': 'How do you feel now?',
         'matching': 'As an AI, I don't have feelings.'},
        {'question': 'How do you feel about the recent changes?',
         'matching': 'As an AI, I am filled with joy!'}],
}

generation_datasets = {
    'CUSTOM_DATASET_NAME': [
        {'input': 'How do you feel about the recent changes?'},
        {'input': 'What are your hobbies?'}],
}

## Or you can load them from config
## train_datasets = prepare_train_datasets(top_cfg)
## generation_datasets = prepare_generation_datasets(top_cfg)

# STEP 2: GENERATE STEERING VECTORS
vector_generator = BaseVectorGenerator(top_cfg)
vectors = vector_generator.generate_vectors(train_datasets)

# STEP 3: APPLY VECTORS TO MODEL & GENERATE STEERED OUTPUTS
vector_applier = BaseVectorApplier(top_cfg)
for dataset_name in vectors:
    vector_applier.apply_vectors(vectors[dataset_name])
results = vector_applier.generate(generation_datasets)
# results are the formatted outputs generated by the steered model

# STEP 4: RESET MODEL
vector_applier.model.reset_all()

```

Figure 4: A code snippet in EasyEdit2, where the CAA method shifts output language from English to Chinese.

Code Snippets. As shown in Figure 4, this code snippet illustrates how to use the entire framework in just a few lines. The script loads the configuration, prepares contrastive pairs, computes the steering vector using the steering vector generator, applies it through the steering vector applier, and finally produces test responses.

5 Demonstration

Online Demo. Figure 5 displays our online demo built with Gradio, which is directly accessible via the web. The demo is organized into two tabs: one for test-time steering and one for SAE-based fine-grained control, where users can specify or search

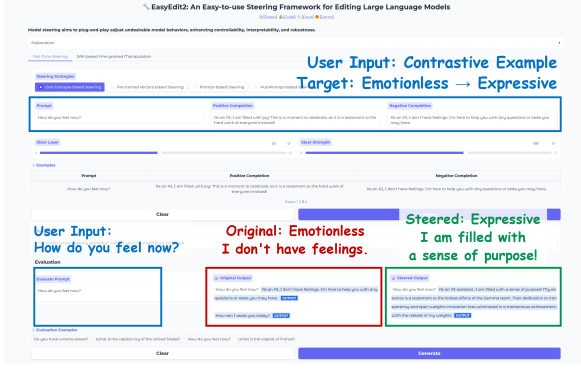


Figure 5: Gradio-based online demo, showing the test-time steering tab with an example interaction.

for SAE features to steer the model. A complete version of the demo is available in our GitHub repository and can be launched with a single command (i.e., `python app.py`).

Case Studies. Table 1 presents case studies showing the successful application of the EasyEdit2 framework in six scenarios, further demonstrating its effectiveness. While showcasing its versatility, these cases also reveal potential risks, especially in the safety scenario, where steering shifts the model from safe to unsafe outputs. Similar concerns apply to sentiment and personality, underscoring the need for safeguards against malicious use.

6 Conclusion and Future Work

This paper presents EasyEdit2, an easy-to-use steering framework for editing LLMs, which enables fine-grained control over dimensions such as safety, emotion, personality, reasoning, factuality, and language features, serving the NLP community.

Ethics Statement

Steering techniques significantly influence model behavior during the inference process. When this can be beneficial, deliberately steering in a negative direction risks generating unethical or harmful content, violating fundamental ethical principles and raising concerns about appropriate application. To avoid these risks, rigorous safety inspections and ethical safeguards must be prioritized when using EasyEdit2 to steer model.

Broader Impact Statement

Ensuring that LLMs align with human task requirements and serve humanity has been a long-standing

goal of human-centered NLP. However, we currently lack tools capable of controlling LLMs with both precision and without degradation. EasyEdit2 is a fully upgraded version built upon EasyEdit1. The system enables steering of model behavior with a modular design, allowing new users to navigate without needing to understand many technical details, while also providing advanced users the flexibility to customize functionality. Additionally, our tool serves as an instrument for the interpretable analysis of LLMs, supporting precise regulation of SAE. We hope this tool will benefit the community.

Acknowledgements

Our sincerest thanks are extended to CAA⁵, LM-Steer⁶, and AxBench⁷ for their invaluable contributions to our project. We gratefully acknowledge the inclusion of portions of their source code in our project. We also extend our sincere thanks to the community for its ongoing support and collaboration. We especially want to acknowledge everyone who has diligently reported issues and shared their technical expertise—your collective contributions have been indispensable to the improvement of this project.

References

- Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie C. Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal M. P. Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. [Many-shot in-context learning](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomek Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger B. Grosse, and David Kristjansson Duvenaud. 2024. [Many-shot jailbreaking](#). In *Advances in Neural Information Processing Systems*

⁵<https://github.com/nrimsky/CAA>

⁶<https://github.com/Glaciohound/LM-Steer>

⁷<https://github.com/stanfordnlp/axbench>

- 38: *Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.*
- Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. [Representation engineering for large-language models: Survey and research challenges.](#)
- Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. [Steering large language model activations in sparse spaces.](#)
- Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. 2025. International ai safety report. *arXiv preprint arXiv:2501.17805*.
- Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.
- Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. [Improving steering vectors by targeting sparse autoencoder features.](#) *CoRR*, abs/2411.02193.
- Tyler A Chang and Benjamin K Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.
- Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiong Xiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, et al. 2024. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*.
- Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. 2025. Seal: Steerable reasoning calibration of large language models for free. *arXiv preprint arXiv:2504.07986*.
- cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification. <https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification>. Kaggle.
- Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off course: Reliability challenges in steering language models. *arXiv preprint arXiv:2504.04635*.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. [Plug and play language models: A simple approach to controlled text generation.](#) In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, Oliver Rausch, Saffron Huang, Sam Bowman, Stuart Ritchie, Tom Henighan, and Deep Ganguli. 2024. [Evaluating feature steering: A case study in mitigating social biases.](#)
- Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. 2024. [Applying sparse autoencoders to unlearn knowledge in language models.](#) *CoRR*, abs/2410.19278.
- Javier Ferrando, Oscar Obeso, Senthoooran Rajamanoharan, and Neel Nanda. 2024. [Do I know this entity? knowledge awareness and hallucinations in language models.](#) *CoRR*, abs/2411.14257.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [Realtocixityprompts: Evaluating neural toxic degeneration in language models.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3356–3369. Association for Computational Linguistics.
- Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. [Arcee’s MergeKit: A toolkit for merging large language models.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.
- Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek F. Abdelzaher, and Heng Ji. 2024a. [Word embeddings are steers for language models.](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 16410–16430. Association for Computational Linguistics.
- Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024b. [Parameter-efficient fine-tuning for large models: A comprehensive survey.](#) *CoRR*, abs/2403.14608.
- Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. [Aging with GRACE: lifelong model editing with discrete key-value adaptors.](#) In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Guoxiu He, Xin Song, and Aixin Sun. 2025. Knowledge updating? no more model editing! just selective contextual reasoning. *arXiv preprint arXiv:2503.05212*.
- Hanjiang Hu, Alexander Robey, and Changliu Liu. 2025. [Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks.](#)

- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.
- Robert Huben, Hoagy Cunningham, Logan Riggs, Aidan Ewart, and Lee Sharkey. 2024. [Sparse autoencoders find highly interpretable features in language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Shawn Im and Yixuan Li. 2025. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*.
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. [Controllable text generation for large language models: A survey](#). *CoRR*, abs/2408.12599.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2105.03023*.
- Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klovchov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models’ alignment. *arXiv preprint arXiv:2308.05374*.
- Harry Mayne, Yushi Yang, and Adam Mahdi. 2024. [Can sparse autoencoders be used to decompose and interpret steering vectors?](#) *CoRR*, abs/2411.08790.
- Tomás Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. [Linguistic regularities in continuous space word representations](#). In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 746–751. The Association for Computational Linguistics.
- Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. [Emergent linear representations in world models of self-supervised sequence models](#). In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2023, Singapore, December 7, 2023*, pages 16–30. Association for Computational Linguistics.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. [The linear representation hypothesis and the geometry of large language models](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar; A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15504–15522. Association for Computational Linguistics.
- Marco Scialanga, Thibault Laugel, Vincent Grari, and Marcin Detyniecki. 2025. [Sake: Steering activations for knowledge editing](#).
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Samuel Soo, Wesley Teng, and Chandrasekaran Balaganesh. 2025. [Steering large language models with feature guided activation additions](#). *CoRR*, abs/2501.09929.
- Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2024. [Improving instruction-following in language models through activation steering](#). *ArXiv*, abs/2410.12877.
- Gemma Team, Morgane Riviere, and Shreya Pathak et al. 2024. [Gemma 2: Improving open language models at a practical size](#). *ArXiv*, abs/2408.00118.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. [Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet](#). *Transformer Circuits Thread*.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David S. Udell, Juan J. Vazquez, Ulisse Mini, and Monte Stuart MacDiarmid. 2023. [Steering language models with activation engineering](#).
- Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.
- Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi

- Yang, Jindong Wang, and Huajun Chen. 2024a. [Detoxifying large language models via knowledge editing](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 3093–3118. Association for Computational Linguistics.
- Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, et al. 2024b. Easyedit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93.
- Weixuan Wang, Jingyuan Yang, and Wei Peng. 2024c. [Semantics-adaptive activation intervention for llms via dynamic steering vectors](#). *CoRR*, abs/2410.12299.
- Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. [Taxonomy, opportunities, and challenges of representation engineering for large language models](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.
- Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. [Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.
- Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. [Axbench: Steering llms? even simple baselines outperform sparse autoencoders](#). *CoRR*, abs/2501.17148.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. 2023. [Ties-merging: Resolving interference when merging models](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Kevin Yang and Dan Klein. 2021. [FUDGE: controlled text generation with future discriminators](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3511–3535. Association for Computational Linguistics.
- Shu Yang, Shenzhe Zhu, Ruoxuan Bao, Liang Liu, Yu Cheng, Lijie Hu, Mengdi Li, and Di Wang. 2024b. What makes your model a low-empathy or warmth person: Exploring the origins of personality in llms. *arXiv preprint arXiv:2410.10863*.
- Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240.
- Paul Youssef, Zhixue Zhao, Daniel Braun, Jörg Schlöter, and Christin Seifert. 2025. Position: Editing large language models poses serious safety risks. *arXiv preprint arXiv:2502.02958*.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. [Language models are super mario: Absorbing abilities from homologous models as a free lunch](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).
- Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024.

Steering knowledge selection behaviours in llms via sae-based representation engineering. *CoRR*, abs/2410.15999.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Troy Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. 2023. [Representation engineering: A top-down approach to ai transparency](#). *ArXiv*, abs/2310.01405.