Project Proposal — Advanced NLP

Quentin Galbez - Léo Lopes - Yanis Martin - Baptiste Villeneuve

November 27th 2025

Subject: *Steering GPT-2*

Github: `https://github.com/Lazyyx/aNLP_Final_Project`

1. Project description

Large Language Models can generate impressive text, but their behaviour is often difficult to control. This raises a key question: can we steer or adjust a model's behavior without retraining it?

In this project, we focus on activation-level steering, and we will work exclusively with GPT-2 (124M). This model is small enough to inspect in detail, easy to run on standard hardware, and well-studied in the interpretability community. By restricting ourselves to one model, we can keep the scope manageable and make comparisons between steering methods more meaningful.

Our steering target is the love / hate polarity. We chose this theme because it is simple, binary, and easy to evaluate: the difference between affectionate and hostile language is usually clear in practice. It also reduces ambiguity when analysing the effect of different steering techniques. This should make it easier to check whether a change in internal activations actually produces a predictable shift in the model's output.

We will compare several techniques. Sparse Autoencoders (SAEs) are an obvious candidate, since they can reveal interpretable directions in activation space. But we also plan to test more straightforward approaches such as linear probes or learned activation edits, which do not require training a full SAE. The goal is not to rely on a single method, but rather to understand how these techniques differ and which ones provide the most reliable control on a small model like GPT-2.

A practical motivation for this type of work is that it could offer a lightweight alternative to fine-tuning. If activation steering can reliably shift the sentiment of generated text, it might help adapt small models to specific domains or tones (e.g., more empathetic responses for educational or support-oriented chatbots).

The main challenge is to identify activation directions that genuinely correspond to emotional polarity, and then show that manipulating these directions produces consistent behavioural changes. The project aims to give a clearer picture of how much control we can realistically obtain from such low-cost techniques.

2. Project background

Controlling and aligning Large Language Models (LLMs) is a central challenge for their safe deployment. The dominant industrial and academic paradigm to achieve this relies on

alignment fine-tuning, primarily through Reinforcement Learning from Human Feedback (RLHF) or, more recently, Direct Preference Optimization (DPO). These methods are effective at instilling a desired general behavior (e.g., being helpful and harmless), but they have two major drawbacks: they are extremely computationally expensive, and the result is a static alignment "baked into" the model's weights.

Our project explores a different approach: inference-time steering, which requires no retraining.

Our work is situated at the intersection of two active research fields. On one hand, model editing techniques, such as ROME or MEMIT, have shown it is possible to modify specific factual knowledge by surgically updating the network's weights. While this work is similar in its goal of modification without full retraining, it focuses on permanent changes (weight modification) and factual knowledge. Our project differs by focusing on temporary interventions (activation modification) for more abstract concepts like tone or sentiment.

On the other hand, our project is directly related to the field of mechanistic interpretability and activation steering. Foundational work has used linear probes to prove that concepts are encoded in activations. More recently, SAEs (Sparse Autoencoders) have emerged as a powerful unsupervised method for discovering interpretable "features."

Our project focuses on the practical application of these techniques. The objective is to determine the most effective method for activation steering by systematically comparing several approaches. Specifically, we will evaluate the performance of Sparse Autoencoders (SAEs) against more established techniques or other forms of activation editing. The goal is to determine which approach offers the best trade-off between control precision (steering love/hate sentiment) and the ease of identifying relevant directions. The choice of a small model (here GPT-2) is pragmatic: it allows for conducting this analysis while concretely testing the best compromise between interpretability and control for dynamic alignment.

3. Project steps

1. **Literature review**
   Read about activation steering, SAEs, linear probes, and previous work on sentiment or emotional directions in GPT-style models.

2. **Build the activation dataset**
   Collect a set of "love" vs. "hate" examples that will help us analyse the internal representation. The goal is not to train a model, but to use the different inputs to retrieve interpretable and reliable evaluation metrics.

3. **Implement and test several steering methods**
   The idea is to compare techniques rather than rely on a single one. Everything is applied to GPT-2 during inference.

   - Linear probe or basic activation vector (baseline)
   - Learned activation edits
   - Sparse Autoencoders (SAEs)

4. **Evaluation step**
   Design a way to evaluate whether steering works. This involves:

- Some human evaluation for qualitative analysis
- Simple metrics to track how steering intensity affects generation
- A small sentiment/emotion classifier to score "love vs. hate" outputs, such as LLM-as-a-judge

5. **Comparison and analysis**
   Compare results, discuss insights, and analyse how different steering approaches affect emotional direction in GPT-2.

4. First results

As a first step, we implemented a simple activation-vector steering method using TransformerLens (`basic_activation.ipynb`). The objective was mainly to understand the full pipeline: extracting activations from GPT-2, identifying a basic "love vs. hate" direction, and injecting this direction back into the model during generation (ref Fig.1). Steering is currently injected in the residual stream of layer 6, as early experiments show this layer encodes emotional polarity most clearly.

Even though the results are still rough and not always convincing, we can clearly observe that the steering does influence the model's output, which confirms that our approach works end-to-end. This simple technique will serve as our baseline, against which we will later compare more advanced methods such as linear probes and Sparse Autoencoders.

We have also started experimenting with SAELens (`GPT2_SAE_STEERING.ipynb`) to analyse how SAE-derived features affect a chosen activation direction. Early tests show that SAE-based steering produces much clearer and more consistent changes in the generated text than the basic activation vector approach. This reinforces our intuition that SAEs capture more meaningful latent structure, and our next objective is to apply this method specifically to the love / hate direction.

We also started working on an evaluation pipeline. For now, we implemented a simple scoring mechanism based on keyword matching: the generated output is checked against two predefined lists of "love" and "hate" terms, which provides a first, lightweight way to quantify the direction of the steering. We also developed a small interface to test the model interactively with different steering intensities, which makes it easier to observe tone variations in real time. In the next stages, we aim to strengthen the evaluation by assessing faithfulness of the generated text and by applying a classifier-based sentiment score to our outputs.

For data, we rely on an existing Hugging Face dataset annotated with "love" and "hate" labels, which provides a consistent benchmark for testing and comparing steering methods.

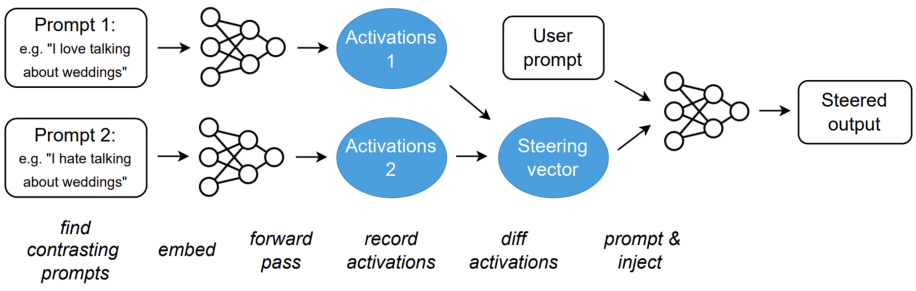

Figure 1: Basic activation method used as a baseline