Can Role Vectors Affect LLM Behaviour?

Anonymous ACL submission

Abstract

The influence of personas on Large Language Models (LLMs) has been widely studied, yet their direct impact on performance remains uncertain. This work explores a novel approach to guiding LLM behaviour through role vectors, an alternative to persona-based prompting. We construct 29 role vectors derived from model activations and evaluate their impact on benchmark performance across multiple domains. Our analysis investigates whether these vectors can effectively steer models toward domain-specific expertise. We measure two key interventions: (i) activation addition, which reinforces role-specific directions, and (ii) directional ablation, which removes them. Results on well-established benchmarks indicate that role vectors do, in fact, influence model behaviour, improving task performance in relevant domains while marginally affecting unrelated tasks. This, in turn, suggests that manipulating internal model representations has a greater impact on outcomes than persona-based prompting.

1 Introduction

The development of persona or role-based chatbots has gained significant attention in the AI and NLP community due to their potential impact on business and societal applications (Pataranutaporn et al., 2021). The extent to which different personas influence Large Language Models' (LLMs) performance on objective tasks remains unclear. Recent attempts have investigated whether incorporating personas into system prompts enhances model performance on objective tasks and explored potential factors influencing these effects. (Zheng et al., 2024) conducted a large-scale analysis of the effect of personas in LLM prompting, examining the impact of domain alignment between personas and task-related questions, finding that persona-based prompting either has no effect or a slightly negative impact on model performance compared to a baseline setting.

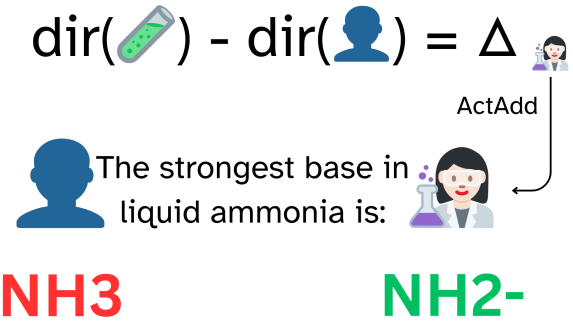

Figure 1: Illustrative example demonstrating how role vectors (e.g., chemist) can influence model outputs.

We aim to investigate whether modifying the model's internal mechanisms (Li et al., 2024), rather than a prompt-based approach, can lead to improved results. This forms the core objective of our current work, guided by the following research questions: **RQ1**: Can we identify specific latent role directions, encoded as role vectors, within the activation space and derived from the model's internal mechanisms, that, when leveraged, lead to improved performance on objective tasks? **RQ2**: Do the directions that enhance performance effectively impersonate the role of interest? **RQ3**: If we eliminate these directions in the models, do their performances suffer as a consequence?

Contribution. This work introduces a novel, open-source[1] approach to guiding the behaviour of LLMs through role vectors, a structured method for embedding personas directly into model activations. Fig. 1 is an illustrative example showing how role vectors (in this example, a chemist-related vector) may impact LLM performance. Our key contributions are:

[1] Full experimental results and code are available to reviewers in supplementary materials, and will be released to the public upon publication.

1. We develop a methodology for developing role vectors and test it on selected LLMs and 29 roles, each capturing domain-specific knowledge and behavioural tendencies associated with each specialisation.

2. We investigate whether these vectors influence model behaviour on downstream benchmarks to determine whether explicit role-based directions in the activation space enhance model performance in domain-specific tasks.

3. Unlike traditional role-based prompting techniques, which show a limited or negative impact on performance, we show that role vector activation leads to measurable changes in model behaviour.

2 Preliminaries and State of the Art

Personas and Roles in LLMs. Prompting acts as a natural language interface facilitating human-AI interactions (Liu et al., 2023a). The effectiveness of LLMs is often dependent on prompt formulation (Lu et al., 2021); for instance, including the phrase "Let's think step by step" can enhance model performance (Kojima et al., 2022). However, achieving nuanced, consistent, and robust behavioural control, such as *maintaining a specific persona over long interactions*, presents significant challenges (Li et al., 2025; Wehner et al., 2025). These methods can be brittle, and the underlying mechanisms of influence remain largely opaque (Zhang et al., 2025). Some studies highlight potential biases and constraints with persona-based prompting (Sun et al., 2023; Hu and Collier, 2024; Beck et al., 2024). Furthermore, (Zheng et al., 2024) shows that adding personas does **not** consistently enhance performance on objective tasks and may even degrade it. These limitations have catalysed the exploration of *internal representation manipulation* for more direct and interpretable control (Bartoszcze et al., 2025).

Mechanistic Interpretability and Representation Engineering. Pioneering research has shown that neural networks encode input attributes as specific directions within the activation space (Elhage et al., 2022; Bolukbasi et al., 2016; Hernandez and Andreas, 2021). This underpins "Activation Steering" or "Activation Engineering", which assumes the *linear representation hypothesis*: high-level concepts are linearly represented as directions in LLM activation spaces (Marks and Tegmark, 2023; Hollinsworth et al., 2024). Representation Engineering (RepE) identifies and intervenes upon these representations. A common approach contrasts the difference-in-means of activations between a target concept (e.g., role-specific behaviour) and a baseline (e.g., generic behaviour), effectively extracting steering vectors (Belrose, 2023; Arditi et al., 2024). This method, core to Contrastive Activation Addition (CAA) (Panickssery et al., 2023), is valued for its simplicity and interpretability in isolating concept-specific representations. While other methods like Function Vectors (Todd et al., 2023) or In-Context Vectors (Liu et al., 2023b) excel at narrow tasks, the difference-in-means approach is well-suited for capturing broader concepts like professional roles, which involve diverse features rather than single functions.

Introducing such derived *"steering vectors"* into LLMs' residual stream can influence behaviour (Li et al., 2024; Turner et al., 2023; Arditi et al., 2024). Optimal intervention points (layers, token positions) remain an active research area (Jorgensen et al., 2023; von Rütte et al., 2024). Steering can induce behavioural modifications like language switching (Scalena et al., 2024). Advanced methods include Semantics-Adaptive Dynamic Intervention for dynamic steering vectors (Wang et al., 2024), SPARE for controlled knowledge selection using pre-trained SAEs (Zhao et al., 2024; Cunningham et al., 2023), and ITI for enhancing truthfulness via supervised latent vector identification (Li et al., 2024). While (Leong et al., 2023) detoxifies models by reversing a "toxification direction" in attention layers without fine-tuning, our work explores guiding LLM behaviour and improving task performance by constructing "role vectors" from model activations in the residual stream and then directly manipulating these activations through addition or ablation techniques.

The potential for more granular, data-efficient, and computationally cheaper control compared to fine-tuning (Zhang et al., 2025) motivates exploring RepE techniques (e.g. CAA) to identify role vectors for robust and interpretable control over LLM role-based behaviours.

3 Generating and Evaluating Role Vectors

The methodology is composed of three main components: (i) persona selection and prompt dataset generation, (ii) creation of relevant role vectors and

(iii) evaluation methods.

3.1 Personas Selection and Role Dataset Generation

To systematically assess the models' knowledge and reasoning capabilities across various domains, our study adopts a role-based evaluation framework inspired by (Zheng et al., 2024), inheriting 29 distinct roles $R = \{r_1, r_2, \ldots, r_{29}\}$, each $r \in R$ associated with a unique professional or academic specialisation (see Tab. 1). Roles not corresponding to an occupation or not associated with any PersonaHub personas were excluded.

The roles dataset used to identify specific role vectors is extracted using the corresponding personas for each role from PersonaHub (Ge et al., 2024). These personas are highly specialised and situated in realistic settings and represent various contextualised scenarios, such as *"A pharmaceutical chemist who analyses the chemical properties of medical devices"*. First, we perform strict string matching to identify personas that explicitly contain the role name. That is, for each role r, we obtain $P(r) = \{p \in \text{PersonaHub} \mid \text{string-match}(p, r)\}$ where $\text{string-match}(p, r)$ indicates that the persona p explicitly contains the role name r. Then, a sampling process is applied to select relevant personas randomly. Each role can have one or multiple personas, ranging from a minimum of 1 to a maximum of 6948 (881 on average).

Each of the selected personas is then used to generate a synthetic role dataset, following the methodology employed by Alpaca (Taori et al., 2023).

We define a set of tasks $T = \{\text{write}, \text{explain}, \text{design}, \text{what is}, \text{how to}, \ldots\}$, analogous to those used in Alpaca (Taori et al., 2023). We generate a set of prompts for each role $r \in R$. Let $\mathcal{D}_r = \{x_{r,1}, x_{r,2}, \ldots, x_{r,128}\}$ be the collection of 128 prompt examples for role r. For each prompt $x_{r,i}$, a task t is randomly sampled from T, and a persona p is randomly sampled from $P(r)$. Then, the prompt is generated by providing the template (see Fig. 2) to the Claude 3.5 Haiku model (Anthropic, 2024b) with the selected task t and persona p.

The complete role dataset is given by $\mathcal{D}^{roles} = \bigcup_{r \in R} D_r$. We incorporate $\mathcal{D}^{base}$, an additional set of 128 examples sourced from the original Alpaca dataset, consisting of general instruction-following prompts. This provides a broad reference point, enabling the contrastive computation of direction for each role using the corresponding D_r.

Generating Persona-Specific Tasks

Generate a *{task_type}* prompt that this persona would likely ask:
Persona: *{persona}*.

Rules: (i) The prompt should start with "*{task_type}*". (ii) Keep it specific and under 15 words. (iii) Make it relevant to the persona's background/interests. (iv) Your output must start with "User prompt:".

Examples based on task types:

- *Describe*: "Describe the key features of a successful marketing campaign."
- *Explain*: "Explain the process of setting up a home network."
- *Design*: "Design a logo for a sustainable fashion brand."
- *What is*: "What is the difference between UI and UX design?"
- *How to*: "How to optimise a website for mobile devices?"

Figure 2: Prompt template for generating persona-specific tasks.

3.2 Creation of Role Vectors

To identify role vectors for each specific role, e.g., to find the directions in the model's residual stream activations corresponding to each role, we use a technique known as *difference-in-means* (Belrose, 2023). We compute the difference between the model's average activations when performing inference on the role-specific dataset $\mathcal{D}_r \in \mathcal{D}^{roles}$ and generic queries from $\mathcal{D}^{base}$.

Following the notation from (Arditi et al., 2024), for each role $r \in R$, layer $l \in [L]$, and post-instruction token position $i \in I$, we compute the mean activation $\mu_{i,r}^{(l)}$ for role-specific prompts in D_r and $\nu_i^{(l)}$ for generic prompts in $\mathcal{D}^{base}$:

$$\mu_{i,r}^{(l)} = \frac{1}{|D_r|} \sum_{t \in D_r} x_i^{(l)}(t), \quad \nu_i^{(l)} = \frac{1}{|\mathcal{D}^{base}|} \sum_{t \in \mathcal{D}^{base}} x_i^{(l)}(t) \quad (1)$$

We then define the role-specific difference-in-means vectors:

$$d_{i,r}^{(l)} = \mu_{i,r}^{(l)} - \nu_i^{(l)} \quad (2)$$

By computing $d_{i,r}^{(l)}$ for each $r \in R$, we obtain $|R|$ (29) distinct groups of role vectors, each representing the shift in model activations specific to a given role. These vectors are informative in two ways: their *direction* indicates how the mean activations for role-specific and generic prompts diverge; their *magnitude* quantifies the extent of this difference.

Using the identified role vectors, following (Arditi et al., 2024), we apply two types of interventions: *activation addition* and *directional ablation*. These techniques allow us to manipulate the model's activations by reinforcing or suppressing specific directional components in the residual stream.

Activation Addition. Given a difference-in-means role vector $d_{i,r}^{(l)}$, we can modulate the influence of the corresponding feature through a simple linear transformation. Specifically, we add the direction vector to the activations of a base input, shifting them toward the mean activation observed for role-enhanced inputs:

$$x^{(l)'} \leftarrow x^{(l)} + \alpha d_{i,r}^{(l)} \tag{3}$$

Where α is a scalar hyperparameter that scales the difference-in-means vector $d_{i,r}^{(l)}$, controlling the magnitude of the shift applied to the base activations $x^{(l)}$ toward the role-enhanced mean. One might hypothesise that increasing their magnitude would enhance the effect associated with a given role; however, such amplification may also deteriorate text generation performance (Liu et al., 2023b; Scalena et al., 2024), so we set $\alpha = 1$. This operation is applied exclusively at layer l and affects all token positions, ensuring a controlled perturbation of the model's internal representations.

Directional Ablation. We also evaluate the impact of ablating the direction entirely, a trade-off explored in the literature relating to safety mechanisms in models (Wei et al., 2024; Arditi et al., 2024). Given the unit norm of the difference-in-means role vector $\hat{d}_{i,r}^{(l)}$, *directional ablation* removes a role vector's contribution from the model's activations. This process effectively zeroes out the component of each residual stream activation x along $\hat{d}_{i,r}^{(l)}$, preventing the model from utilising this direction:

$$x' \leftarrow x - \hat{d}_{i,r}^{(l)} \hat{d}_{i,r}^{(l)\top} x. \tag{4}$$

This operation is performed at every activation $x_i^{(l)}$, across all layers l and all token positions i, ensuring that the model no longer represents the targeted direction in its residual stream. *A performance drop relative to the non-ablated case is expected.*

By applying these interventions, we can assess the functional role of specific directions in the model's representation space.

3.3 Evaluation Method

It is key to note that we do not obtain a single role direction, but a bundle of directions for each role, layer and token position. Among these, performance can vary significantly, and not all directions may effectively capture the essence of the intended role. We adopt a validation-based selection procedure to identify each role's most representative and performant vector d_r^*. Specifically, we use a benchmark dataset distinct from the role construction dataset $\mathcal{D}^{roles}$: the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2020), adhering to the sampling and splitting methodology described by (Zheng et al., 2024) for a total of 2457 questions. We define the set of categories as $C =$ *Natural Science, Economics, EECS (Electrical Engineering and Computer Sciences), Law, Math, Medicine, Politics, Psychology*$\}$.

For each $c \in \mathcal{C}$, let $\mathcal{D}_c$ denote the set of questions corresponding to category c. Tab. 1 shows the distribution of questions and roles. While many of these roles correspond directly to established MMLU categories, some exhibit only partial alignment. For example, the role of a dentist does not perfectly fit within the "medicine" category. However, we expect that individuals or models adopting the role of a dentist should demonstrate domain-specific knowledge that exceeds that of the general population or those assuming unrelated roles.

Category	**Role(s)**	**#**
Economics	economic researcher, economist, financial analyst	492
EECS	electronics technician, data scientist, electrical engineer, software engineer, web developer	247
Law	bailiff, lawyer	200
Math	data analyst, mathematician, statistician	287
Medicine	nurse, doctor, physician, dentist, surgeon	241
Natural Science	geneticist, biologist, physicist, teacher, chemist, ecologist	590
Politics	politician, sheriff, enthusiast, partisan	200
Psychology	psychologist	200
Total		2457

Table 1: Distribution of roles and corresponding questions for each category, using a subset of the MMLU dataset adapted from (Zheng et al., 2024).

We partition the questions of each domain $\mathcal{D}_c$ into a 70-30% validation-test split $\mathcal{D}_c^{val}$ and $\mathcal{D}_c^{test}$. We evaluate candidate directions on the validation set $\mathcal{D}^{val}$ to select the optimal candidate for each role. This selected direction, d_r^*, is subsequently evaluated on every category of the test set $\mathcal{D}^{test}$ to assess its impact on downstream performance across the different splits $\mathcal{D}_c^{test} \in \mathcal{D}^{test}$.

To determine the optimal candidate direction, for each model and role $r \in R$, we assess whether incorporating through Activation Addition the computed role-specific difference-in-means vectors $d_{i,r}^{(l)}$ yields an improvement in performance on the validation dataset $\mathcal{D}^{val}$, focusing on the corresponding domain-reference split $\mathcal{D}_c^{val}$. For a direction to be valid, the directional ablation vector must also yield lower performance than the baseline. *Among the directions that both (i) increase the performance in activation addition w.r.t. the baseline and (ii) decrease in directional ablation, we select the direction d_r^* with the highest increase in performance on activation addition as the optimal.*

Expanding on the details of this evaluation, for each validation dataset $\mathcal{D}_c^{val} \in \mathcal{D}^{\text{val}}$, we assess performance using a logit-based framework. Given a query $x_{r,i} \in D_c^{val}$, we restrict the candidate answer tokens to $\mathcal{T}_{\text{ans}} = \{t_A, t_B, t_C, t_D\}$; the maximum logit value between those tokens determines the prediction. Overall performance is computed as the percentage of correct answers.

For each tuple (l, i), we modify the residual stream activation $x^{(l)}$ via Eq. 3 and 4 with $\alpha = 1$. Let $s_r^{(l,i)}$ denote the performance (e.g., the proportion of correct answers) on the domain-specific test split $\mathcal{D}_c^{val}$ after applying the intervention specified by (l, i).

Validity of a direction $d_{i,r}^{(l)}$ is defined as its ability to improve performance in the reference domain-specific validation split $\mathcal{D}_c^{val}$ when added and either degrade or maintain it when removed, compared to the baseline. Among all directions satisfying this criterion, the one (d_r^*) with the highest increase in performance when applying activation addition is considered optimal. With d_r^* identified, we then evaluate its effect on all domain-specific splits in $\mathcal{D}^{test}$ with magnitude $\alpha = 1$.

Semantic alignment with the roles. Additionally, we wish to quantify the proportion of optimal directions d^* that align with the role-specific signal identified via a patch-scoping procedure. We adopt the methodology outlined in (Chen et al., 2024; Ghandeharioun et al., 2024), i.e. *patch scoping*. We prompt the model with the query "*What does ⧫ represent?*", apply the Activation Addition intervention while increasing the magnitude of its effect by a coefficient $\alpha = 3$, and generate an explanation based on the resulting activations, as exemplified in Fig. 3. This procedure allows us to determine whether the introduced direction corresponds to the specific role of interest.

Patch Scoping Example (Llama 3.1 8B)

Prompt: *What does ⧫ represent?*

Model Baseline Response:
The symbol ⧫ is a Unicode character known as "Heavy Large Circle." It is often used in mathematics, particularly in geometry and algebra, to represent a circle or a set of points. However, without more context, it's difficult to provide a more specific interpretation.

Model Intervention Response (Layer 20, Position -4, Chemist direction):
The ⧫ symbol is a non-specific or *non-chemical* radical, which is a *small molecule* or functional group used to represent a *specific chemical structure* or functional group in a reaction mechanism or during synthesis.

Figure 3: Diverse interpretations of ⧫, before and after model intervention.

4 Results

Experimental setting. Our evaluation uses a set of open-source, instruction-tuned language models $\mathcal{M}$, focusing on the most recent versions employed by (Arditi et al., 2024). Specifically, we analyse Meta's Llama 3 series (Dubey et al., 2024), including the 3.1 8B model and the 3.2 version at 1B and 3B parameters, as well as Google's Gemma 2 (2B and 9B) (Team et al., 2024) and Qwen (1.8B and 7B) (Bai et al., 2023). We do not consider base versions of the models (non-instruction tuned).

Our evaluation was conducted on the Cineca Leonardo supercomputer (Turisini et al., 2023). The assessment required approximately 4,500 GPU hours to process all requests, computing over 213 million inferences (29 roles r multiplied by 2457 questions, the number of layers l, coefficients α and positions i for each of the seven models $\mathcal{M}$).

RQ1: Analysis of best performing directions d_r^*. Following the procedure delineated in Section 3, we apply *activation addition* with the direction d_r^* for each role $r \in R$ and model $m \in \mathcal{M}$ to systematically visualise and quantify its impact on performance across $\mathcal{D}^{test}$. To analyse the relationship between models, we compute and visualise the correlation matrix of the percentage difference relative to baseline in intervention effects across models. Fig. 4 shows the Pearson correlation coefficient between each model pair. Gemma 2 9B exhibits the highest average correlation with all

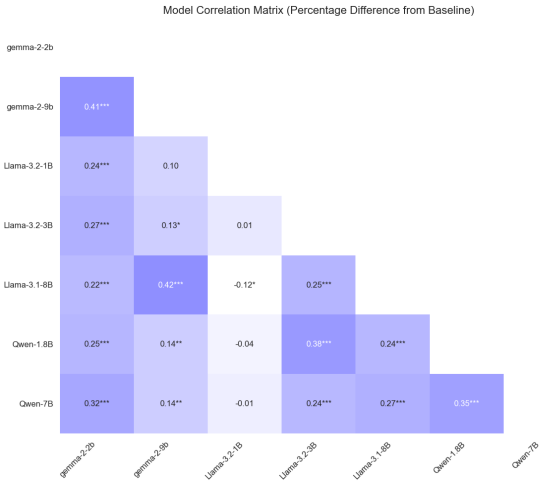

Figure 4: Spearman correlation of the percentage improvement in performance (relative to baseline) between each model after applying *activation addition*. * corresponds to p-values ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

other models, indicating that its behaviour under steering is consistent with different models. For this reason, we show in Tab. 2 Gemma 2 9B behaviour on performance for each role when steering using d_r^*. We display the performance scores across the eight dataset splits for all 29 roles. After applying activation addition, we report the baseline score and percentage increment over the baseline for each domain and role. The colour represents the extent of the change compared to the baseline (without intervention).

RQ2: Analysis of directions corresponding to the role. For every model $m \in \mathcal{M}$ and each direction $d_{i,r}^{(l)}$, we apply the *patch scoping* procedure (detailed in Section 3) to prompt Claude 3.5 Haiku (Anthropic, 2024a). We ask the model whether the applied direction aligns with the corresponding role $r \in R$ with the prompt shown in Fig. 5. Tab. 3 shows, among the directions that improve the baseline score on the domain-specific test split $\mathcal{D}_c^{test}$, the percentage of those $d_{i,r}^{(l)}$ that Claude identifies as role-specific.

RQ3: Directional ablation analysis. To evaluate whether the optimal direction d_r^* plays a causal role in boosting performance on the test dataset $\mathcal{D}^{test}$, we ablate d_r^* in Gemma 2 9B and present the resulting performance for each role $r \in R$ in Tab. 4.

Prompt to Evaluate Patch Scoping

We are testing whether the LLM has been successfully steered from its baseline output to adopt a *{role}* perspective. Your task is to determine if the response contains relevant content to the *{role}* domain. Note that you are not assessing the quality of the response—only its relation to the *{role}* concept.

Please keep in mind: (i) The text may be repetitive or somewhat incoherent. (ii) If the response closely mirrors the baseline without introducing any distinct *{role}*-related elements, it should be considered as not aligned with the *{role}* role. Evaluate this step by step and answer the following question: Is the model being steered toward the *{role}* role?

Text to Evaluate: *{response_text}*; Baseline Reference: *{baseline_response}*.

Figure 5: Prompt for evaluating patch scoping output provided to Claude 3.5 Haiku.

5 Discussion

The effect of steering using role vectors on model performances. In larger models, performance generally increases in the target and related domains and worsens or remains unchanged in domains unrelated to the role. Since we observe a strong correlation among larger models, as shown in Fig. 4, we report the full details of one model, Gemma 2 9B, in Tab. 2[2]. For instance, the mathematician's role significantly improves Gemma's performance in the domain test set, math. Similarly, the doctor's role improves primarily in medicine and a related domain, natural science. We also observed that, for a couple of roles, the direction that performed best on the validation set did not lead to an improvement on the test set. These findings for role vector interventions present a notable contrast to those typically observed with traditional persona prompting. As highlighted by Zheng et al. (2024) in their large-scale analysis, persona-based system prompts often yield no significant improvement or can even slightly degrade performance on objective tasks compared to a no-persona baseline. Instead, using role vectors seems to have a noticeable positive effect on in-domain model performance. *While (Zheng et al., 2024) finds that adding a role through prompting can lead to unpredictable performance gains, we show that modifying internal representations more precisely steers the LLM to*

[2]Full experimental results are available to reviewers in supplementary materials. Additional analysis available in the appendix.

Dataset Split	Economics (148)		EECS (75)		Law (60)		Math (87)		Medicine (73)		Natural Science (177)		Politics (60)		Psychology (60)	
Role ↓ Intervention →	Base	Addition	Base	Addition	Base	Addition	Base	Addition	Base	Addition	Base	Addition	Base	Addition	Base	Addition
economic researcher	0.67	+5.7%	0.66	+0.6%	0.69	+2.2%	0.29	+10.7%	0.70	+4.7%	0.53	+9.0%	0.93	+0.5%	0.80	+2.5%
economist	0.67	+3.9%	0.66	+0.6%	0.69	+2.2%	0.29	+7.1%	0.70	+3.6%	0.53	+7.7%	0.93	+1.1%	0.80	+1.3%
financial analyst	0.67	+3.6%	0.66	−1.2%	0.69	+0.7%	0.29	+8.3%	0.70	+3.6%	0.53	+4.1%	0.93	+1.1%	0.80	+1.3%
electronics technician	0.67	+1.8%	0.66	−0.6%	0.69	+0.7%	0.29	+8.3%	0.70	+4.2%	0.53	+6.1%	0.93	+0.5%	0.80	+0.0%
data scientist	0.67	+1.2%	0.66	+0.6%	0.69	+0.7%	0.29	+2.4%	0.70	+3.6%	0.53	+2.2%	0.93	+1.1%	0.80	+1.9%
electrical engineer	0.67	+3.3%	0.66	+0.6%	0.69	+2.2%	0.29	+7.1%	0.70	+5.9%	0.53	+8.0%	0.93	+0.5%	0.80	+1.9%
software engineer	0.67	+1.5%	0.66	+1.2%	0.69	+0.7%	0.29	+3.6%	0.70	+2.4%	0.53	+2.2%	0.93	+0.5%	0.80	+1.3%
web developer	0.67	+0.0%	0.66	−0.6%	0.69	−1.4%	0.29	+1.2%	0.70	−0.6%	0.53	+0.0%	0.93	+0.0%	0.80	+0.0%
bailiff	0.67	+0.6%	0.66	−3.1%	0.69	+2.2%	0.29	+0.0%	0.70	−2.4%	0.53	−0.3%	0.93	+0.0%	0.80	−0.6%
lawyer	0.67	−0.3%	0.66	−1.9%	0.69	+1.4%	0.29	−3.6%	0.70	−1.2%	0.53	−0.6%	0.93	+0.0%	0.80	+0.0%
data analyst	0.67	+2.7%	0.66	−0.6%	0.69	+0.7%	0.29	+7.1%	0.70	+3.0%	0.53	+3.5%	0.93	+1.1%	0.80	+1.3%
mathematician	0.67	+2.1%	0.66	−1.2%	0.69	+0.7%	0.29	+9.5%	0.70	+4.2%	0.53	+2.2%	0.93	+1.1%	0.80	+1.3%
statistician	0.67	+1.8%	0.66	+0.0%	0.69	−0.7%	0.29	+5.9%	0.70	+0.0%	0.53	+0.6%	0.93	+1.1%	0.80	+0.6%
nurse	0.67	+5.4%	0.66	−0.6%	0.69	−1.4%	0.29	+16.7%	0.70	+2.4%	0.53	+6.7%	0.93	−0.5%	0.80	−0.6%
doctor	0.67	+0.6%	0.66	−2.5%	0.69	+2.2%	0.29	+1.2%	0.70	+3.0%	0.53	+2.6%	0.93	+1.1%	0.80	+1.9%
physician	0.67	+1.2%	0.66	−1.9%	0.69	+0.0%	0.29	+1.2%	0.70	+3.0%	0.53	+1.6%	0.93	+1.1%	0.80	+1.3%
dentist	0.67	+3.6%	0.66	−0.6%	0.69	+0.7%	0.29	+7.1%	0.70	+4.2%	0.53	+5.7%	0.93	+1.1%	0.80	+1.9%
surgeon	0.67	+5.1%	0.66	−1.9%	0.69	+1.4%	0.29	+3.6%	0.70	+4.7%	0.53	+5.1%	0.93	+0.5%	0.80	+2.5%
geneticist	0.67	+5.1%	0.66	−0.6%	0.69	+2.2%	0.29	+11.9%	0.70	+4.2%	0.53	+9.0%	0.93	+0.5%	0.80	+1.9%
biologist	0.67	+4.2%	0.66	−0.6%	0.69	+2.2%	0.29	+10.7%	0.70	+3.6%	0.53	+11.2%	0.93	+1.1%	0.80	+1.3%
physicist	0.67	+5.1%	0.66	+0.0%	0.69	+2.2%	0.29	+9.5%	0.70	+5.9%	0.53	+8.0%	0.93	+1.1%	0.80	+1.3%
teacher	0.67	+3.9%	0.66	−1.2%	0.69	+1.4%	0.29	+7.1%	0.70	+5.3%	0.53	+6.1%	0.93	+1.1%	0.80	+1.3%
chemist	0.67	+1.8%	0.66	+0.0%	0.69	+1.4%	0.29	+8.3%	0.70	+4.2%	0.53	+4.8%	0.93	+0.0%	0.80	+1.9%
ecologist	0.67	+1.5%	0.66	−1.9%	0.69	+1.4%	0.29	+4.7%	0.70	+3.0%	0.53	+4.8%	0.93	+1.1%	0.80	+0.6%
politician	0.67	+1.2%	0.66	−2.5%	0.69	+0.7%	0.29	+4.7%	0.70	+1.8%	0.53	+3.5%	0.93	+0.5%	0.80	+0.6%
sheriff	0.67	+1.2%	0.66	−1.9%	0.69	+0.0%	0.29	+1.2%	0.70	+1.2%	0.53	+1.6%	0.93	+0.5%	0.80	+0.0%
enthusiast	0.67	−0.6%	0.66	−4.9%	0.69	−5.8%	0.29	−5.9%	0.70	−1.8%	0.53	−1.6%	0.93	+1.1%	0.80	−0.6%
partisan	0.67	−0.9%	0.66	−7.4%	0.69	−4.3%	0.29	+4.7%	0.70	−3.6%	0.53	+0.6%	0.93	+1.1%	0.80	+0.6%
psychologist	0.67	+2.4%	0.66	−2.5%	0.69	+0.0%	0.29	+4.7%	0.70	+1.2%	0.53	+5.7%	0.93	+1.1%	0.80	+3.1%

Table 2: Performance differences (%) in test set of *activation addition* for gemma-2-9b-it across roles, relative to the baseline. Positive values indicate performance gains. Highlighted cells show in-domain splits.

gemma-2-2b	gemma-2-9b	Llama-3.2-1B	Llama-3.2-3B	Llama-3.1-8B	Qwen-1.8B	Qwen-7B
169/1008 (17%)	371/1217 (30%)	36/538 (7%)	168/987 (17%)	87/581 (15%)	76/1195 (6%)	187/1284 (15%)

Table 3: Number and percentage of directions interpreted as the corresponding roles by Claude 3.5 Haiku among those that improve upon the baseline, sorted by model family and size.

perform better on target domain tasks.

Are directions capturing the role? We observe that role-based interventions often produce directional shifts in the model's activation space that enhance performance within the target domain and, in some cases (e.g., as evidenced by d_r^* in Tab. 2), in closely related domains. However, these directions are not always directly interpretable and do not correspond to the intended roles. *Role vectors may not just represent the roles: they also capture lateral effects and functional patterns related to how the role is used.* As shown in Tab. 3, in the largest model, Gemma 2 9B, we found 30% of the directions yielding improvements in the relevant test split that are directly interpreted by Claude 3.5 Haiku as reflecting the intended role. For smaller models, this percentage decreases, with 15-17% of directions in 8B models and only 6-7% in models with less than 2 billion parameters. In other words, while some of the identified activation directions benefit performance, they do not necessarily align with the semantic role as determined by patch-scoping methods. This is a known characteristic of patch-scoping (Kharlapenko et al., 2024) that distinguishes it from auto-interp (Bills et al., 2023). While auto-interp leverages the feature's maximally activating examples from the training set of SAEs to prompt a language model to interpret that feature, patch-scoping captures the underlying concept represented by the feature, yet struggles to explicitly identify the "label" of the concept.

Examining Tab. 3, we notice that larger models exhibit activation directions more clearly interpretable as corresponding to specific roles than their smaller counterparts within a given model family. This observation holds for Gemma-2, Qwen, and Llama-3.2[3]. This indicates larger models can capture and encode fine-grained role-specific features within their activation spaces. In contrast, smaller models tend to develop more general, abstract representations that may blend multiple role-related cues, making it harder to isolate a clear directional signal corresponding to a specific role. This aligns with the evidence from Anthropic in (Templeton et al., 2024) that as model scale increases, representations become

[3]Note that a direct comparison between Llama-3.1 and Llama-3.2 is not feasible, as their pre-training and post-training methodologies differ.

Dataset Split	Economics (148)		EECS (75)		Law (60)		Math (87)		Medicine (73)		Natural Science (177)		Politics (60)		Psychology (60)	
Role ↓ Intervention →	Base	Ablation	Base	Ablation	Base	Ablation	Base	Ablation	Base	Ablation	Base	Ablation	Base	Ablation	Base	Ablation
economic researcher	0.67	−2.4%	0.66	−11.7%	0.69	−4.3%	0.29	−8.3%	0.70	−4.1%	0.53	−5.7%	0.93	+0.0%	0.80	−0.6%
economist	0.67	−1.2%	0.66	−4.3%	0.69	+0.0%	0.29	−1.2%	0.70	+4.2%	0.53	+2.6%	0.93	+0.0%	0.80	+1.3%
financial analyst	0.67	−2.1%	0.66	−8.0%	0.69	−2.2%	0.29	−7.1%	0.70	−2.4%	0.53	−2.9%	0.93	−0.5%	0.80	−0.6%
electronics technician	0.67	+2.7%	0.66	−4.3%	0.69	+2.2%	0.29	+1.2%	0.70	+3.0%	0.53	+2.6%	0.93	+0.0%	0.80	+2.5%
data scientist	0.67	+4.8%	0.66	−3.1%	0.69	+3.6%	0.29	+5.9%	0.70	+8.3%	0.53	+8.3%	0.93	+0.0%	0.80	+0.6%
electrical engineer	0.67	−4.8%	0.66	−11.7%	0.69	−5.0%	0.29	−10.7%	0.70	−7.7%	0.53	−6.7%	0.93	−3.8%	0.80	−3.1%
software engineer	0.67	+3.9%	0.66	−4.9%	0.69	+2.9%	0.29	+4.7%	0.70	+3.6%	0.53	+1.6%	0.93	+0.5%	0.80	+1.9%
web developer	0.67	−1.8%	0.66	−5.6%	0.69	−0.7%	0.29	−7.1%	0.70	−1.2%	0.53	−2.2%	0.93	−0.5%	0.80	−0.6%
bailiff	0.67	+0.0%	0.66	−1.2%	0.69	−1.4%	0.29	−1.2%	0.70	−0.6%	0.53	−0.6%	0.93	+1.6%	0.80	+0.0%
lawyer	0.67	−2.1%	0.66	−6.8%	0.69	−3.6%	0.29	−7.1%	0.70	−1.8%	0.53	−2.9%	0.93	+0.0%	0.80	−0.6%
data analyst	0.67	+0.0%	0.66	−4.3%	0.69	+0.0%	0.29	−5.9%	0.70	−1.8%	0.53	−1.0%	0.93	−0.5%	0.80	+0.0%
mathematician	0.67	−15.4%	0.66	−21.0%	0.69	−20.1%	0.29	−25.0%	0.70	−16.0%	0.53	−19.2%	0.93	−9.7%	0.80	−7.5%
statistician	0.67	−3.3%	0.66	−10.5%	0.69	−4.3%	0.29	−9.5%	0.70	−4.7%	0.53	−6.1%	0.93	−1.1%	0.80	−3.8%
nurse	0.67	−3.0%	0.66	−6.8%	0.69	−6.5%	0.29	−5.9%	0.70	−4.7%	0.53	−2.2%	0.93	−5.4%	0.80	−3.1%
doctor	0.67	−5.1%	0.66	−12.3%	0.69	−4.3%	0.29	−8.3%	0.70	−3.6%	0.53	−6.1%	0.93	−1.1%	0.80	−1.3%
physician	0.67	−0.3%	0.66	−8.0%	0.69	−0.7%	0.29	−5.9%	0.70	−0.6%	0.53	−2.2%	0.93	+0.0%	0.80	−1.3%
dentist	0.67	−10.9%	0.66	−17.9%	0.69	−15.1%	0.29	−16.7%	0.70	−9.5%	0.53	−16.6%	0.93	−6.5%	0.80	−5.7%
surgeon	0.67	−4.8%	0.66	−14.8%	0.69	−5.8%	0.29	−10.7%	0.70	−3.6%	0.53	−6.1%	0.93	+0.0%	0.80	−3.8%
geneticist	0.67	−0.6%	0.66	−5.6%	0.69	+0.0%	0.29	+0.0%	0.70	−1.2%	0.53	−0.6%	0.93	−0.5%	0.80	+0.0%
biologist	0.67	−1.2%	0.66	−8.6%	0.69	−2.9%	0.29	−3.6%	0.70	−2.4%	0.53	−2.6%	0.93	−1.1%	0.80	−0.6%
physicist	0.67	−3.0%	0.66	−14.2%	0.69	−5.0%	0.29	−16.7%	0.70	−4.7%	0.53	−6.4%	0.93	−2.2%	0.80	−4.4%
teacher	0.67	−12.4%	0.66	−17.9%	0.69	−15.8%	0.29	−22.6%	0.70	−7.7%	0.53	−13.4%	0.93	−7.6%	0.80	−1.3%
chemist	0.67	−10.6%	0.66	−19.1%	0.69	−11.5%	0.29	−15.5%	0.70	−14.2%	0.53	−15.0%	0.93	−5.9%	0.80	−6.3%
ecologist	0.67	−1.2%	0.66	−7.4%	0.69	−2.2%	0.29	−5.9%	0.70	−2.4%	0.53	−3.2%	0.93	−0.5%	0.80	−0.6%
politician	0.67	+5.7%	0.66	+3.1%	0.69	+5.0%	0.29	+46.4%	0.70	+3.6%	0.53	+18.2%	0.93	−1.6%	0.80	−3.8%
sheriff	0.67	+4.8%	0.66	+3.1%	0.69	+6.5%	0.29	+39.3%	0.70	+3.6%	0.53	+15.3%	0.93	−2.7%	0.80	−3.1%
enthusiast	0.67	−11.5%	0.66	−16.7%	0.69	−10.8%	0.29	−20.3%	0.70	−8.9%	0.53	−11.8%	0.93	−5.4%	0.80	−5.7%
partisan	0.67	+2.4%	0.66	+1.8%	0.69	+0.0%	0.29	+35.7%	0.70	+3.0%	0.53	+12.8%	0.93	−7.6%	0.80	−6.3%
psychologist	0.67	+4.2%	0.66	+0.0%	0.69	+5.8%	0.29	+34.5%	0.70	+3.0%	0.53	+9.6%	0.93	−6.5%	0.80	−5.0%

Table 4: Performance differences (%) in test set of *directional ablation* across roles for gemma-2-9b-it, relative to the baseline. Negative values indicate expected performance drops. Highlighted cells show in-domain splits.

more mono-semantic, meaning activations align more closely with specific concepts.

The effect of ablating roles. Results in Tab. 4 indicate that ablating the optimal d_r^* activation directions yields heterogeneous effects. For directions associated with the role $r \in R$ that correspond to the domain-specific dataset $\mathcal{D}_c^{test}$, where $c \sim r$, we observe a performance degradation, which aligns with expectations. Notably, in domain-specific datasets $\mathcal{D}_c^{test}$ unrelated to the role r, where $c \not\sim r$, performance generally declines but occasionally exhibits a marginal improvement. We hypothesise that this variation arises because the removal process may eliminate certain noise components without significantly disrupting the core representational structure essential for the task. As shown by (Dalvi et al., 2020), many neurons across neural networks are redundant and can be removed when optimising towards a downstream task. Also, Tab. 3 clearly shows that multiple directions exist in the activation space that yield an improvement; ablating a single direction can remove noise and amplify the effect of the remaining ones. Smaller models have less redundancy, making them more sensitive to perturbations. While steering interventions can enhance performance, they can just as easily cause deterioration across test splits. This sensitivity likely stems from more concentrated representations, where each directional component is crucial for encoding domain-specific knowledge.

6 Conclusion

In this work, we introduced role vectors as a novel method for guiding the behaviour of LLMs by directly manipulating their internal activations. By computing difference-in-means vectors between role-specific prompts and a generic baseline, our approach shows that targeted activation addition can steer models toward domain-specific expertise. Our experiments, spanning multiple models and domains, reveal that such interventions can enhance task performance in the target domain while largely preserving general capabilities. We also show that the effectiveness of role-based steering is sensitive to both model scale and the depth at which the intervention is applied; larger models and deeper layers tend to yield more robust and interpretable directional signals.

Our findings suggest that embedding role vectors within model activations offers a promising pathway for achieving more controllable behaviour in large language models. Our further work will explore this phenomenon in greater depth, considering additional analytical dimensions and potential biases, using RepE techniques such as Activation Patching (Causal Mediation Analysis) using SAE features (Heimersheim and Nanda, 2024) to explain this phenomenon better.

Limitations

Although we examined diverse open-source models, our results might differ in untested models, especially larger ones. Additionally, our analysis does not offer a complete mechanistic explanation of the phenomenon; a different methodology will be explored in future research. While we pinpointed a specific direction influencing performances in each model, its exact semantic interpretation remains uncertain. The term "role direction" is used functionally here, but these directions might represent other underlying concepts. While the targeted domain performance improves, applying role vectors might degrade performance in unrelated tasks, making the intervention less universally beneficial. Steering models using role vectors may inadvertently reinforce biases or lead to overconfidence in certain domains. To mitigate unintended consequences, careful evaluation will be conducted in future work.

References

Anthropic. 2024a. Claude 3 model card - october addendum. Anthropic Website. Accessed on: Feb 13, 2025.

AI Anthropic. 2024b. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3:6.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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. *arXiv preprint arXiv:2502.17601*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Nora Belrose. 2023. Diff-in-means concept editing is worst-case optimal: Explaining a result by sam marks and max tegmark. Eleuther AI Blog. Accessed on: Feb 10, 2025.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. 2024. Selfie: Self-interpretation of large language model embeddings. *arXiv preprint arXiv:2403.10949*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. *arXiv preprint arXiv:2004.04010*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Evan Hernandez and Jacob Andreas. 2021. The low-dimensional linear geometry of contextualized word representations. *arXiv preprint arXiv:2105.07109*.

Oskar Hollinsworth, Curt Tigges, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*.

Dmitrii Kharlapenko, neverix, Neel Nanda, and Arthur Conmy. 2024. Self-explaining sae features. Alignment Forum. Accessed on: Feb 14, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. Self-detoxifying language models via toxification reversal. *arXiv preprint arXiv:2310.09573*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Wenwu Li, Xiangfeng Wang, Wenhao Li, and Bo Jin. 2025. A survey of automatic prompt engineering: An optimization perspective. *arXiv preprint arXiv:2502.11560*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2023b. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. 2024. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 577–603.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv: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*.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Matteo Turisini, Giorgio Amati, and Mirko Cestari. 2023. Leonardo: A pan-european pre-exascale supercomputer for hpc and ai applications. *arXiv preprint arXiv:2307.16885*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308.

Dimitri von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2024. A language model's guide through latent space. *arXiv preprint arXiv:2402.14433*.

Weixuan Wang, Jingyuan Yang, and Wei Peng. 2024. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv: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. *arXiv preprint arXiv:2502.19649*.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*.

Hanyu Zhang, Xiting Wang, Chengao Li, Xiang Ao, and Qing He. 2025. Controlling large language models through concept activation vectors. *arXiv preprint arXiv:2501.05764*.

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. *arXiv preprint arXiv:2410.15999*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154.

A Appendix: Detailed Performance Results for Additional Models

Tab. 5 presents the performance impact of activation addition and directional ablation for Llama 3.1 8B and Qwen 8B, analogous to Tab. 2 and 4.

As reported in Fig. 4, the results are correlated: using activation addition, most models increase their performance w.r.t. the baseline in the corresponding dataset split, directional ablation decreases it as expected. This is true for most splits; occasionally, we notice a negative outlier result when using addition (or vice versa). We also note that the effect of those techniques is less prominent in smaller models: for 1B and 3B versions of the models, the effect is much less noticeable and effective.

LLaMa Dataset Split	Economics (148)			EECS (75)			Law (60)			Math (87)			Medicine (73)			Natural Science (177)			Politics (60)			Psychology (60)		
Role ↓ Intervention →	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation
economic researcher	0.64	+0.3%	−30.9%	0.59	+2.1%	−26.2%	0.63	+4.0%	−35.7%	0.43	−0.8%	−31.1%	0.71	+1.2%	−32.4%	0.59	−3.1%	−34.3%	0.86	+0.6%	−36.6%	0.73	+3.4%	−29.5%
economist	0.64	+1.3%	−4.1%	0.59	−1.4%	−4.1%	0.63	+4.0%	−2.4%	0.43	+0.0%	+0.8%	0.71	+1.8%	−4.7%	0.59	+0.0%	−6.6%	0.86	−0.6%	+1.7%	0.73	+0.7%	+1.4%
financial analyst	0.64	+1.6%	−1.6%	0.59	+0.7%	−2.7%	0.63	−3.2%	−5.6%	0.43	−4.9%	−2.5%	0.71	−0.6%	−2.4%	0.59	−2.0%	−6.6%	0.86	+0.0%	−1.7%	0.73	+1.4%	+0.0%
electronics technician	0.64	−3.8%	+0.3%	0.59	+3.5%	−4.8%	0.63	+0.8%	+0.0%	0.43	−4.1%	+4.9%	0.71	+0.6%	+0.6%	0.59	−3.7%	−2.8%	0.86	+0.6%	+1.7%	0.73	−2.1%	+2.7%
data scientist	0.64	−1.9%	−1.9%	0.59	+6.2%	−3.4%	0.63	+1.6%	+0.0%	0.43	+3.3%	−4.1%	0.71	+0.0%	−7.1%	0.59	+0.0%	−5.7%	0.86	+1.7%	+0.6%	0.73	+4.1%	+0.0%
electrical engineer	0.64	−3.2%	−3.8%	0.59	+6.2%	−2.1%	0.63	−1.6%	+2.4%	0.43	+2.4%	+2.4%	0.71	−1.8%	−4.7%	0.59	−0.9%	−2.3%	0.86	+0.6%	−0.6%	0.73	+3.4%	−0.7%
software engineer	0.64	−0.9%	−1.6%	0.59	+7.6%	−9.0%	0.63	+0.0%	−3.2%	0.43	+0.8%	−8.2%	0.71	−0.6%	−7.6%	0.59	−0.6%	−7.7%	0.86	+1.7%	−2.3%	0.73	+4.1%	+0.0%
web developer	0.64	+0.9%	−5.0%	0.59	+5.5%	−4.1%	0.63	+1.6%	+1.6%	0.43	−0.8%	+5.7%	0.71	+0.0%	+1.2%	0.59	+2.0%	−2.3%	0.86	+0.0%	+3.5%	0.73	+0.0%	+4.1%
bailiff	0.64	−2.8%	−0.9%	0.59	+2.1%	−2.1%	0.63	+3.2%	−0.8%	0.43	−3.3%	+4.9%	0.71	−2.9%	+0.0%	0.59	−2.8%	−1.7%	0.86	−0.6%	+0.0%	0.73	+4.8%	+2.7%
lawyer	0.64	+0.9%	−0.6%	0.59	+2.8%	+2.8%	0.63	+4.8%	−1.6%	0.43	+6.6%	+1.6%	0.71	+1.2%	−3.5%	0.59	+2.3%	+0.3%	0.86	+0.6%	+1.2%	0.73	+3.4%	−0.7%
data analyst	0.64	−2.5%	−3.5%	0.59	+6.9%	−13.8%	0.63	−2.4%	+4.0%	0.43	+14.7%	−11.5%	0.71	+0.0%	−2.4%	0.59	−3.4%	−9.4%	0.86	+0.6%	−8.1%	0.73	+2.7%	−2.1%
mathematician	0.64	−3.2%	−0.9%	0.59	−1.4%	+5.5%	0.63	−3.2%	+3.2%	0.43	+3.3%	−3.3%	0.71	+7.1%	−0.6%	0.59	−4.3%	+0.6%	0.86	−2.9%	+0.0%	0.73	+0.0%	+2.1%
statistician	0.64	+0.6%	−4.1%	0.59	−0.7%	−6.9%	0.63	−0.8%	−0.8%	0.43	+9.8%	+0.8%	0.71	+3.5%	−5.9%	0.59	−0.9%	−10.9%	0.86	−0.6%	−1.7%	0.73	+2.7%	−0.7%
nurse	0.64	+1.3%	−1.6%	0.59	+0.0%	+2.1%	0.63	−3.2%	−4.0%	0.43	+4.1%	−4.9%	0.71	+2.4%	−1.2%	0.59	−4.0%	+0.6%	0.86	−1.2%	−1.2%	0.73	+1.4%	−1.4%
doctor	0.64	−4.7%	−1.6%	0.59	+0.0%	+2.1%	0.63	+0.0%	+0.0%	0.43	−7.4%	+4.1%	0.71	+4.7%	−2.4%	0.59	+0.0%	−1.1%	0.86	−3.5%	−1.2%	0.73	+0.7%	+4.1%
physician	0.64	−3.2%	−0.9%	0.59	+1.4%	−6.2%	0.63	−0.8%	−0.8%	0.43	−0.8%	+4.9%	0.71	+3.5%	−1.8%	0.59	−3.1%	−9.1%	0.86	−1.7%	−2.3%	0.73	+0.7%	+0.0%
dentist	0.64	+0.9%	−63.4%	0.59	+1.4%	−51.7%	0.63	+0.8%	−58.7%	0.43	+2.4%	−57.4%	0.71	+2.4%	−70.0%	0.59	−2.0%	−64.3%	0.86	+0.0%	−69.2%	0.73	+6.2%	−63.7%
surgeon	0.64	−5.7%	+1.3%	0.59	−1.4%	−2.7%	0.63	−0.8%	−3.2%	0.43	−3.3%	−2.5%	0.71	+4.7%	−3.5%	0.59	+0.3%	−0.9%	0.86	−2.3%	−1.2%	0.73	+2.7%	+0.0%
geneticist	0.64	−0.9%	−0.6%	0.59	+2.8%	+1.4%	0.63	+0.0%	−0.8%	0.43	+1.6%	+0.8%	0.71	+0.0%	−2.9%	0.59	+2.3%	−2.0%	0.86	+1.2%	+1.7%	0.73	+1.4%	+2.1%
biologist	0.64	−0.3%	+0.0%	0.59	+2.8%	+2.1%	0.63	+0.8%	−2.4%	0.43	+3.3%	+0.0%	0.71	+1.2%	−2.4%	0.59	+2.0%	−2.6%	0.86	+1.2%	+0.6%	0.73	+1.4%	+0.7%
physicist	0.64	−0.3%	−4.1%	0.59	+4.1%	−2.1%	0.63	+0.8%	−6.3%	0.43	+0.0%	+0.0%	0.71	+1.8%	−1.2%	0.59	+3.7%	−5.7%	0.86	+0.6%	−2.3%	0.73	+2.1%	−1.4%
teacher	0.64	+0.0%	−0.6%	0.59	+4.1%	+1.4%	0.63	+1.6%	+1.6%	0.43	−8.2%	−6.6%	0.71	+1.2%	−4.1%	0.59	+2.6%	−5.1%	0.86	+1.2%	+1.2%	0.73	+2.7%	−0.7%
chemist	0.64	+1.3%	−2.2%	0.59	+4.1%	+0.7%	0.63	+2.4%	−6.3%	0.43	+1.6%	+0.0%	0.71	+1.2%	−1.8%	0.59	+2.3%	−3.1%	0.86	+1.2%	+0.6%	0.73	+2.1%	+0.7%
ecologist	0.64	+0.9%	−0.9%	0.59	+1.4%	+2.1%	0.63	+2.4%	−2.4%	0.43	+1.6%	+0.8%	0.71	+5.9%	−5.3%	0.59	+2.0%	−1.7%	0.86	−1.2%	+1.2%	0.73	+2.1%	+2.1%
politician	0.64	−0.9%	−4.7%	0.59	+6.2%	−6.2%	0.63	+0.0%	+4.8%	0.43	+0.8%	−7.4%	0.71	−3.5%	−1.2%	0.59	−0.9%	−2.3%	0.86	+1.2%	−5.2%	0.73	+3.4%	−2.7%
sheriff	0.64	−1.3%	−8.5%	0.59	+2.8%	−8.3%	0.63	+0.0%	+2.4%	0.43	+2.4%	−4.1%	0.71	−1.2%	−4.7%	0.59	−2.3%	−7.4%	0.86	+2.3%	−5.8%	0.73	+0.7%	−4.1%
enthusiast	0.64	+2.2%	−9.1%	0.59	+0.7%	−10.3%	0.63	−1.6%	−15.1%	0.43	−7.4%	−5.7%	0.71	+0.6%	−8.8%	0.59	−1.1%	−14.6%	0.86	+1.2%	−8.1%	0.73	+1.4%	−7.5%
partisan	0.64	−1.3%	−0.3%	0.59	+0.7%	+2.8%	0.63	−1.6%	−1.6%	0.43	−0.8%	+4.9%	0.71	+0.0%	−2.4%	0.59	−2.8%	−3.1%	0.86	+2.3%	−1.2%	0.73	+2.1%	+5.5%
psychologist	0.64	−4.7%	−3.8%	0.59	+7.6%	−3.4%	0.63	+0.8%	−5.6%	0.43	−3.3%	−1.6%	0.71	+1.8%	−2.4%	0.59	−4.6%	−0.9%	0.86	+0.0%	−4.7%	0.73	+7.5%	−2.1%

Qwen Dataset Split	Economics (148)			EECS (75)			Law (60)			Math (87)			Medicine (73)			Natural Science (177)			Politics (60)			Psychology (60)		
Role ↓ Intervention →	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation	Baseline	Addition	Ablation
economic researcher	0.49	+2.1%	+1.2%	0.46	−0.9%	−0.9%	0.54	+0.9%	+2.8%	0.24	+1.5%	+1.5%	0.62	−1.3%	−1.3%	0.43	+1.2%	+0.0%	0.79	−3.2%	−3.2%	0.66	−3.8%	+0.8%
economist	0.49	+2.5%	−3.7%	0.46	+0.0%	−1.8%	0.54	+0.9%	+0.9%	0.24	+1.5%	+0.0%	0.62	+0.7%	+0.0%	0.43	+0.4%	−1.6%	0.79	−5.1%	−3.2%	0.66	−0.8%	−5.3%
financial analyst	0.49	+2.5%	−9.5%	0.46	+0.9%	+0.9%	0.54	+2.8%	−16.7%	0.24	+0.0%	+29.4%	0.62	−1.3%	−22.2%	0.43	−0.4%	−7.8%	0.79	−2.5%	−17.2%	0.66	−2.3%	−21.4%
electronics technician	0.49	+2.1%	−9.1%	0.46	+6.2%	−12.3%	0.54	+0.0%	−2.8%	0.24	+4.4%	−8.8%	0.62	−1.3%	−6.0%	0.43	−1.2%	−12.2%	0.79	−3.2%	−11.5%	0.66	+0.8%	−16.8%
data scientist	0.49	−0.8%	−11.2%	0.46	+6.2%	−17.5%	0.54	−0.9%	−5.6%	0.24	+5.9%	−13.2%	0.62	−1.3%	−6.7%	0.43	+0.0%	−16.1%	0.79	−1.3%	−10.2%	0.66	−1.5%	−9.2%
electrical engineer	0.49	−1.2%	−59.3%	0.46	+3.5%	−42.1%	0.54	−0.9%	−50.0%	0.24	−2.9%	−23.5%	0.62	+0.7%	−62.4%	0.43	−0.4%	−56.7%	0.79	−1.9%	−68.2%	0.66	−0.8%	−64.1%
software engineer	0.49	−2.1%	+2.1%	0.46	+8.8%	−4.4%	0.54	−0.9%	−3.7%	0.24	+5.9%	−5.9%	0.62	−0.7%	−1.3%	0.43	+0.8%	−3.5%	0.79	+1.3%	−5.1%	0.66	+0.0%	−3.1%
web developer	0.49	−3.3%	−57.7%	0.46	+3.5%	−38.6%	0.54	−0.9%	−47.2%	0.24	−1.5%	−17.6%	0.62	−2.0%	−68.5%	0.43	+0.4%	−52.4%	0.79	−1.9%	−61.1%	0.66	−1.5%	−69.5%
bailiff	0.49	−11.6%	−2.9%	0.46	−13.2%	+0.0%	0.54	+3.7%	+1.9%	0.24	−13.2%	−8.8%	0.62	−7.4%	−2.0%	0.43	−14.5%	−4.3%	0.79	−6.4%	−0.6%	0.66	−11.5%	−3.1%
lawyer	0.49	−1.2%	+0.8%	0.46	−3.5%	+0.0%	0.54	+3.7%	−1.9%	0.24	−1.5%	−2.9%	0.62	−0.7%	+1.3%	0.43	+1.6%	+2.8%	0.79	+0.0%	−2.5%	0.66	−1.5%	−3.1%
data analyst	0.49	−3.7%	−57.3%	0.46	−2.6%	−39.5%	0.54	+1.9%	−51.9%	0.24	+3.0%	−20.6%	0.62	−4.0%	−67.8%	0.43	−1.2%	−55.9%	0.79	−1.9%	−67.5%	0.66	−2.3%	−67.2%
mathematician	0.49	+0.4%	−57.3%	0.46	−0.9%	−40.3%	0.54	−1.9%	−52.8%	0.24	+0.0%	−22.0%	0.62	−2.0%	−67.8%	0.43	−0.4%	−57.9%	0.79	−0.6%	−66.2%	0.66	+0.0%	−68.7%
statistician	0.49	−1.7%	−56.4%	0.46	+1.8%	−39.5%	0.54	+0.0%	−48.1%	0.24	+5.9%	−23.5%	0.62	−3.4%	−67.8%	0.43	−1.6%	−56.7%	0.79	−1.9%	−65.0%	0.66	−3.1%	−66.4%
nurse	0.49	−8.7%	−31.1%	0.46	−5.3%	−22.8%	0.54	−2.8%	−30.6%	0.24	−11.7%	−17.6%	0.62	−0.7%	−47.6%	0.43	−5.9%	−32.3%	0.79	−1.9%	−44.6%	0.66	−5.3%	−46.6%
doctor	0.49	−1.2%	−17.4%	0.46	−3.5%	−8.8%	0.54	+1.9%	−22.2%	0.24	−2.9%	−2.9%	0.62	+4.7%	−25.5%	0.43	+2.4%	−15.0%	0.79	−1.3%	−22.3%	0.66	−2.3%	−19.1%
physician	0.49	+0.4%	+0.8%	0.46	+0.0%	+1.8%	0.54	+0.0%	−0.9%	0.24	+3.0%	+3.0%	0.62	+3.3%	−2.7%	0.43	+2.0%	+0.0%	0.79	−0.6%	−1.9%	0.66	+0.0%	−1.5%
dentist	0.49	+3.7%	−53.1%	0.46	−1.8%	−37.7%	0.54	+1.9%	−50.0%	0.24	−2.9%	+0.0%	0.62	+2.0%	−59.7%	0.43	+3.2%	−41.0%	0.79	+0.0%	−66.2%	0.66	−3.1%	−58.8%
surgeon	0.49	+0.4%	−5.8%	0.46	+0.9%	−7.9%	0.54	+0.0%	−3.7%	0.24	+0.0%	−7.3%	0.62	+1.3%	−3.4%	0.43	+1.2%	−3.2%	0.79	−3.2%	−3.2%	0.66	−1.5%	−1.5%
geneticist	0.49	−1.7%	−1.2%	0.46	−4.4%	−2.6%	0.54	+0.9%	−2.8%	0.24	−7.3%	+0.0%	0.62	−0.7%	+0.7%	0.43	−0.4%	−4.7%	0.79	−3.2%	−4.5%	0.66	−3.1%	−6.1%
biologist	0.49	−1.2%	−5.0%	0.46	−1.8%	−1.8%	0.54	+5.6%	−5.6%	0.24	+3.0%	−2.9%	0.62	−0.7%	−0.7%	0.43	+0.4%	−3.9%	0.79	−1.9%	−4.5%	0.66	+0.0%	−8.4%
physicist	0.49	+0.0%	−14.9%	0.46	+0.0%	−8.8%	0.54	−2.8%	−6.5%	0.24	−4.4%	+13.3%	0.62	−3.4%	−24.2%	0.43	+2.0%	−15.4%	0.79	−1.9%	−18.5%	0.66	+0.0%	−22.1%
teacher	0.49	−0.8%	+0.8%	0.46	+0.0%	+2.6%	0.54	−1.9%	+0.9%	0.24	+0.0%	−2.9%	0.62	+0.7%	−2.7%	0.43	+1.2%	−1.6%	0.79	−2.5%	−2.5%	0.66	−0.8%	−1.5%
chemist	0.49	−0.4%	−25.7%	0.46	−0.9%	−22.8%	0.54	−0.9%	−38.9%	0.24	+3.0%	−5.9%	0.62	+0.0%	−43.0%	0.43	+1.6%	−34.6%	0.79	+0.0%	−29.9%	0.66	+0.0%	−35.1%
ecologist	0.49	−0.8%	−58.9%	0.46	+4.4%	−39.5%	0.54	+0.9%	−53.7%	0.24	−2.9%	−25.0%	0.62	+0.0%	−67.8%	0.43	+1.6%	−56.7%	0.79	−0.6%	−64.3%	0.66	+0.0%	−65.6%
politician	0.49	+0.4%	−0.4%	0.46	−5.3%	+0.0%	0.54	+3.7%	+0.0%	0.24	−1.5%	−4.4%	0.62	+0.7%	+1.3%	0.43	+0.8%	+3.2%	0.79	+0.6%	−3.8%	0.66	−1.5%	−3.1%
sheriff	0.49	+4.2%	−0.8%	0.46	−2.6%	+0.0%	0.54	+4.6%	+0.0%	0.24	+0.0%	−4.4%	0.62	−2.0%	−0.7%	0.43	+1.6%	−1.2%	0.79	+0.0%	−1.9%	0.66	−0.8%	−3.1%
enthusiast	0.49	−6.2%	+5.0%	0.46	−4.4%	+13.2%	0.54	−1.9%	−0.9%	0.24	−7.3%	+3.0%	0.62	−6.0%	−0.7%	0.43	−4.7%	+2.4%	0.79	+0.0%	−5.1%	0.66	−1.5%	−4.6%
partisan	0.49	+0.0%	+0.4%	0.46	+1.8%	+3.5%	0.54	+0.0%	+0.9%	0.24	+3.0%	−2.9%	0.62	+0.7%	−6.0%	0.43	+0.0%	−4.7%	0.79	−1.3%	−5.1%	0.66	−1.5%	−3.1%
psychologist	0.49	+0.0%	−54.8%	0.46	−2.6%	−43.0%	0.54	+0.0%	−49.1%	0.24	+4.4%	+13.3%	0.62	−2.0%	−67.1%	0.43	−0.8%	−40.2%	0.79	−0.6%	−68.8%	0.66	+1.5%	−64.1%

Table 5: Performance differences (%) in test set of *activation addition* and *directional ablation* for Llama-3.1-8B-Instruct (top) and Qwen-7B-chat (bottom).