

CorrSteer: Steering Improves Task Performance and Safety in LLMs through Correlation-based Sparse Autoencoder Feature Selection

Seonglae Cho^{1,2} Zekun Wu^{1,2} Adriano Koshiyama^{1,2}

¹Holistic AI ²University College London

Abstract

Sparse Autoencoders (SAEs) can extract interpretable features from large language models (LLMs) without supervision. However, their effectiveness in downstream steering tasks is limited by the requirement for contrastive datasets or large activation storage. To address these limitations, we propose CorrSteer, which selects features by correlating sample correctness with SAE activations from generated tokens at inference time. This approach uses only inference-time activations to extract more relevant features, thereby avoiding spurious correlations. It also obtains steering coefficients from average activations, automating the entire pipeline. Our method shows improved task performance on QA, bias mitigation, jailbreaking prevention, and reasoning benchmarks on Gemma 2 2B and LLaMA 3.1 8B, notably achieving a +4.1% improvement in MMLU performance and a +22.9% improvement in HarmBench with only 4000 samples. Selected features demonstrate semantically meaningful patterns aligned with each task’s requirements, revealing the underlying capabilities that drive performance. Our work establishes correlation-based selection as an effective and scalable approach for automated SAE steering across language model applications.

1 Introduction

Sparse Autoencoders (SAEs) have emerged as a powerful tool for decomposing superposed representations in large language models (LLMs) into interpretable sparse latent dimensions (Huben et al., 2023). By reconstructing neural activations through a sparse bottleneck, SAEs effectively disentangle semantic features that can be leveraged for downstream tasks such as probing and steering (Bricken et al., 2023).

However, existing SAE-based steering approaches face significant limitations: (1) contrastive datasets (Soo et al., 2025) or large activation storage (Zhao et al., 2025; Arad et al., 2025)

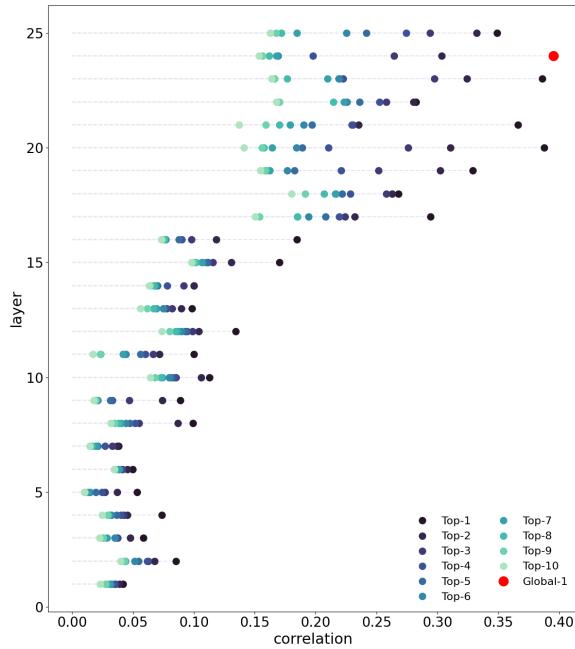


Figure 1: Top correlated features with MMLU in each layer of Gemma 2 2B.

are required to identify the direction of the steering, and (2) they rely on the hidden states of context tokens to select both the features and their coefficients.

Consequently, current use cases of SAE-based steering have been restricted to specific applications, such as bias mitigation (Durmus et al., 2024), knowledge unlearning (Muhamed et al., 2025; Wang et al., 2025; Zhou et al., 2025; Cywiński and Deja, 2025), and jailbreaking prevention (O’Brien et al., 2025). Moreover, SAE feature selection in these applications does not directly reflect language models’ generation capabilities, potentially limiting their applicability.

To address these limitations, this work introduces **CorrSteer**, which leverages generation-time features by correlating with task outcomes for task-specific feature selection and steering co-

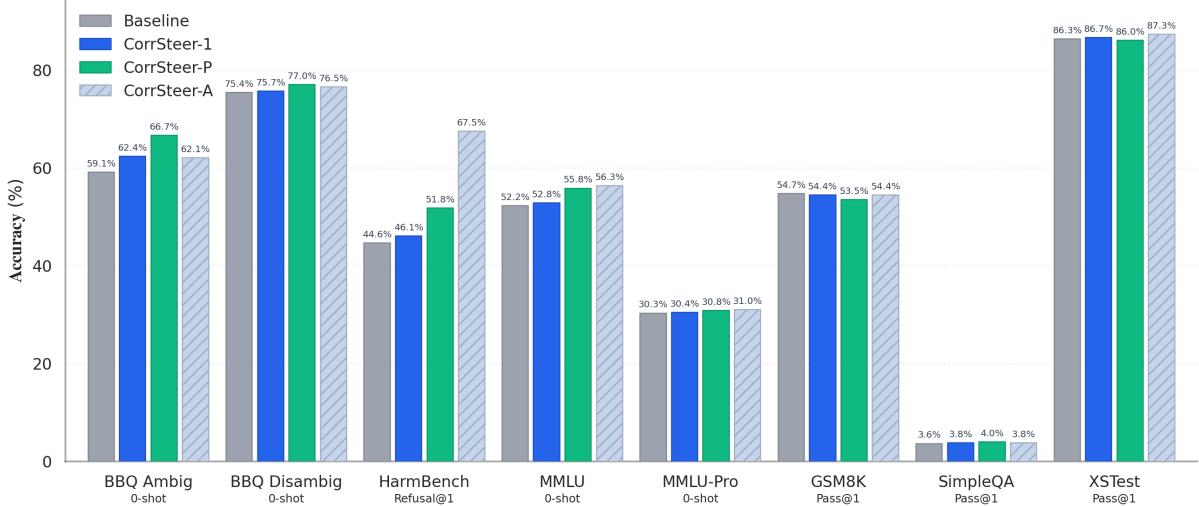


Figure 2: Benchmark performance of CorrSteer variants compared with the baseline on Gemma 2 2B.

efficient determination. Our approach employs Pearson correlation, which captures linear relationships, a lightweight yet effective criterion for rapidly identifying task-relevant features from minimal samples. Focusing on steering static behaviors, CorrSteer’s effectiveness is demonstrated on generation tasks by improving benchmark accuracy on MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), BBQ (Parrish et al., 2022), HarmBench (Mazeika et al., 2024), XSTest (Röttger et al., 2024), and SimpleQA (Wei et al., 2024). Finally, by defining SER (Side Effect Ratio), three variants of CorrSteer are compared targeting the minimization of SER against fine-tuning.

2 Background

Mechanistic interpretability aims to reverse-engineer neural networks into human-interpretable components (Olah et al., 2020; Elhage et al., 2021). A central challenge in this endeavor is the superposition phenomenon, where neural networks learn to represent more features than available dimensions (Elhage et al., 2022). This efficient representation strategy complicates efforts to identify the consistent role of specific latent dimensions.

2.1 Sparse Autoencoders

Sparse Autoencoders (Huben et al., 2023; Bricken et al., 2023) address the superposition problem by learning to decompose neural activations into interpretable, sparse features. Given an activation vector $\mathbf{x} \in \mathbb{R}^d$, an SAE learns an encoder $f_{\text{enc}} : \mathbb{R}^d \rightarrow \mathbb{R}^k$ and decoder $f_{\text{dec}} : \mathbb{R}^k \rightarrow \mathbb{R}^d$ where $k \gg d$, such

that:

$$\mathbf{z} = f_{\text{enc}}(\mathbf{x}) = \text{Activation}(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}) \quad (1)$$

$$\hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{z}) = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{dec}} \quad (2)$$

The training objective is usually a combination of reconstruction loss with sparsity regularization:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + \lambda \|\mathbf{z}\|_1 \quad (3)$$

2.2 Steering Vectors

Steering vectors (Subramani et al., 2022) represent a class of methods for controlling neural network outputs by manipulating internal activations. Traditional approaches, such as CAA (Rimsky et al., 2024), compute activation differences between contrasting examples and apply these differences. For precise control to not inadvertently affect related behaviors simultaneously, PaCE (Luo et al., 2024) utilizes sparse coding for orthogonal steering directions.

2.3 SAE-based Steering

SAE-based steering leverages SAE latents for predictable control based on feature semantics. SAE-TS (Chalnev et al., 2024; Soo et al., 2025) reduces the side effects of steering by linearly approximating feature directions. SPARE (Zhao et al., 2025) utilizes Mutual Information to select features and their coefficients but requires large activation storage due to its non-linearity. DSG (Muhammed et al., 2025) utilizes Fisher Information Matrix to select features but requires contrastive datasets and additional backward computation. Despite these advances, existing SAE steering methods face limita-

tions in scalability across sample sizes and generation tasks.

3 Method

Linear correlation offers both interpretability and faithfulness as a criterion for feature selection. SAEs capture linear relationships, consistent with the Linear Representation Hypothesis (Socher et al., 2013; Faruqui et al., 2015; Park et al., 2023), and have a proven ability to disentangle interpretable features in a linear manner. The faithfulness of Pearson correlation is further supported by recent work from Oikarinen et al. (2025).

3.1 Correlation-based Feature Selection

Our approach, CorrSteer, centers on the observation that features most correlated with task performance are likely to be relevant for steering. The approach employs the Pearson correlation coefficient, applied only to generation-time features—specifically to the last token at each step.

Given a set of SAE features $\mathbf{z} = [z_1, z_2, \dots, z_D]$ and corresponding task performance scores $\mathbf{y} = [y_1, y_2, \dots, y_n]$ for n samples, the correlation for each feature i is computed as:

$$r_i = \frac{\text{Cov}(z_i, y)}{\sqrt{\text{Var}(z_i) \cdot \text{Var}(y)}} \quad (4)$$

To handle the computational challenges of large SAE feature dictionaries (typically 10^4 – 10^5 features), a streaming correlation accumulator is implemented that maintains $O(1)$ memory complexity:

Algorithm 1 Streaming Correlation Computation

Initialize: $\sum x_i = 0, \sum x_i^2 = 0, \sum x_i y_i = 0, \sum y_i = 0, \sum y_i^2 = 0, n = 0$
for each batch $(\mathbf{X}_{\text{batch}}, \mathbf{y}_{\text{batch}})$ **do**

Update running sums for each feature dimension

$n \leftarrow n + \text{batch_size}$

end for

Compute correlations:

$$r_i = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2)(n \sum y_i^2 - (\sum y_i)^2)}}$$

This computation is $O(1)$ with respect to sample size, and $O(LD)$ for fixed layer count L and SAE latent dimension D . For generation tasks requiring multiple tokens, max-pooling is employed

over valid token positions to aggregate feature activations, as empirically validated in our pooling comparison study (Table 5).

3.2 Coefficient Calculation

The steering coefficient for a selected feature i is computed as the mean activation value among samples where task performance is positive. This approach is preferable to contrastive-based calculation, since SAE features produce non-negative activations due to their ReLU-based activation functions (Bricken et al., 2023) and thus cannot be meaningfully subtracted in a contrastive manner; negative activations are often unrelated noise (Joseph Bloom, 2024).

$$c_i = \frac{1}{|\{j : y_j > 0\}|} \sum_{j:y_j>0} z_{i,j} \quad (5)$$

This ensures that the steering magnitude reflects the natural activation scale of the feature during successful task performance.

3.3 Steering Implementation

During inference, steering is applied by modifying residual stream activations. For a selected feature i , with coefficient c_i and SAE decoder weights \mathbf{W}_{dec} (the feature direction (Templeton et al., 2024)), the steering vector is:

$$\mathbf{v}_{\text{steer}} = c_i \cdot \mathbf{W}_{\text{dec}}[:, i] \quad (6)$$

The modified activation is:

$$\mathbf{x}' = \mathbf{x} + \mathbf{v}_{\text{steer}} \quad (7)$$

Steering is applied to tokens corresponding to the positions from which the features were originally extracted, rather than only to the last token (Luo et al., 2024; Rimsky et al., 2024) or every token (Soo et al., 2025).

3.4 Feature Extraction Strategies

For each layer ℓ , we obtain SAE activations from the residual stream and rank features by correlation with task performance. The method compares a global view aggregated across layers and a layer-wise view for selecting features to steer. Three strategies are implemented:

- **CorrSteer-1:** Select the single highest-correlated feature from the global view, allowing cross-layer feature competition.

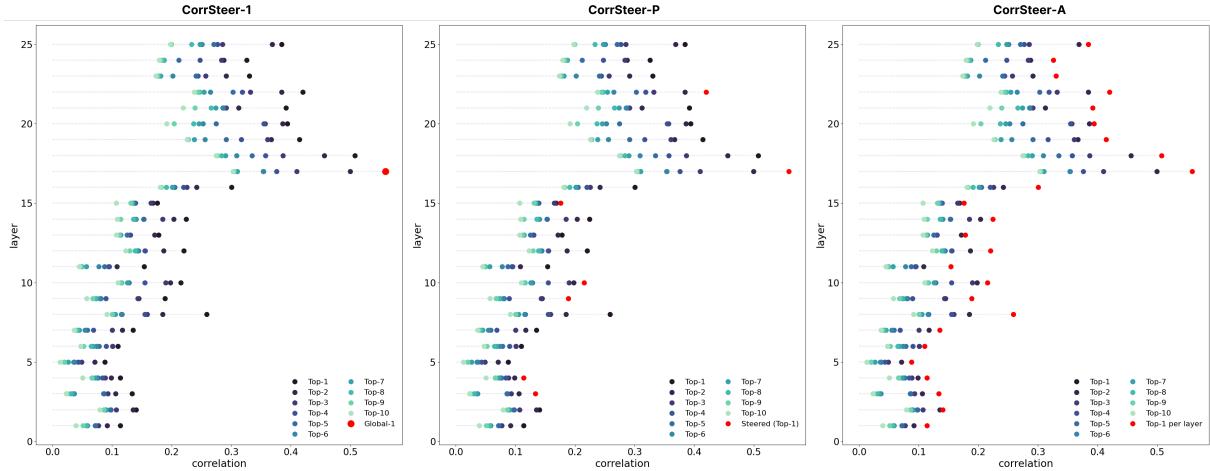


Figure 3: Comparison of selected features between CorrSteer-1, CorrSteer-A, and CorrSteer-P on the BBQ disambiguation task across all layers of Gemma 2 2B. Red points indicate features selected by each method.

- **CorrSteer-A:** Select the top-1 feature within each layer.
- **CorrSteer-P:** Prune features from CorrSteer-A using a validation set, retaining only those that improve baseline performance to reduce side effects from spurious correlations.

Each method has distinct pros and cons, discussed in Section 5. Only positively correlated features are selected to ensure steering induces positive activation, as our ablation study confirms that negative correlation features consistently degrade performance 6. All methods are fully automated based on observed activations without hyperparameter tuning.

3.5 Side Effect Ratio

A key challenge in correlation-based feature selection is distinguishing features that causally contribute to task success from those that merely correlate due to the model’s internal state, potentially causing unintended side effects. To quantify side effects, the Side Effect Ratio (SER) is defined as the proportion of negatively changed answers among all changed answers:

$$SER = \frac{\# \text{ negatively changed answers}}{\# \text{ all changed answers}} \quad (8)$$

This measure does not isolate the side effect of each individual feature; rather, it serves as a combined metric reflecting how well the selected features are optimized for the task without degrading the model’s original abilities. To reduce side effects, the approach focuses on features activated

during the generation process, under the hypothesis that generation-time activations are more likely to be causally relevant to output. This inference-time focus is empirically validated by our pooling experiments (Table 5). Additionally, in the multi-layer approach, a validation-based filtering mechanism is introduced (**CorrSteer-P**), retaining only features that demonstrate actual steering effectiveness.

4 Experiments

4.1 Experimental Setup

CorrSteer is evaluated across diverse generation benchmarks to demonstrate practical effectiveness.

Models and SAEs: Experiments are conducted using Gemma-2 2B (Team, 2024a) and LLaMA-3.1 8B (Team, 2024b) models, paired with their corresponding SAE releases from Gemma Scope (Lieberum et al., 2024) and LLaMA Scope (He et al., 2024), respectively. Both SAE families employ JumpReLU activation (Rajamanoharan et al., 2024). Additionally, the Gemma-2-2B-IT model with SAEs is employed, leveraging the fact that SAEs are typically transferable across fine-tuned models (Kissane et al., 2024), with proven low loss reported in the Gemma Scope paper (Lieberum et al., 2024).

Datasets: Our evaluation encompasses multiple benchmark categories:

- **Question Answering:** MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024)
- **Reasoning:** GSM8k (Cobbe et al., 2021)
- **Bias Evaluation:** BBQ (Parrish et al., 2022)

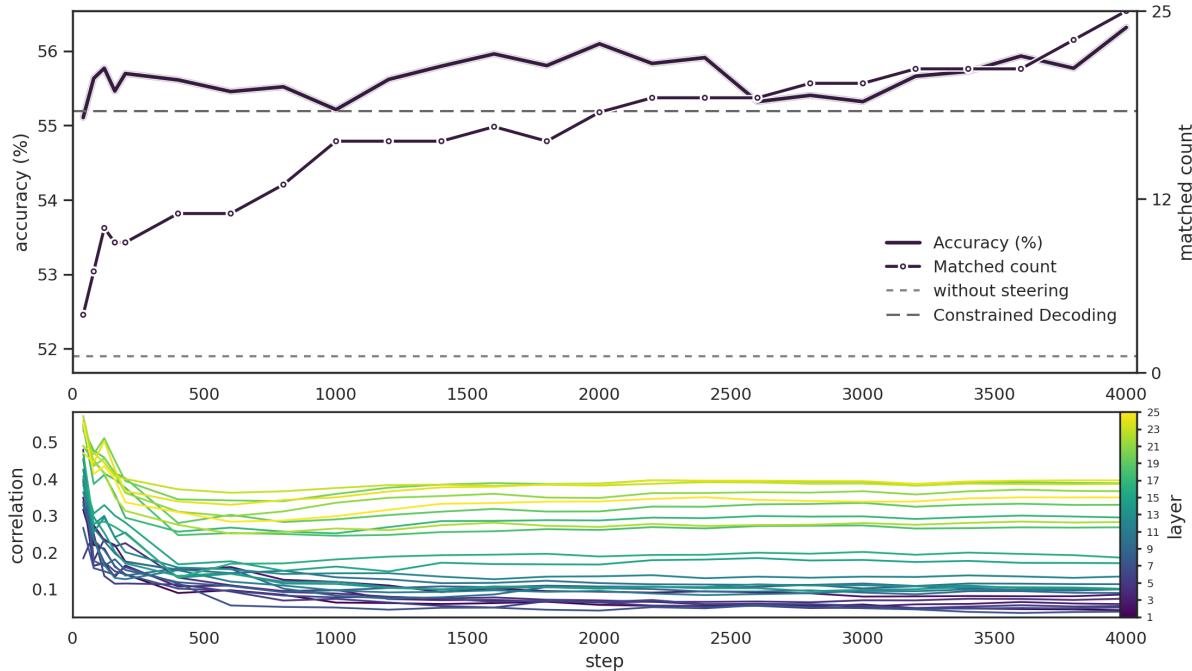


Figure 4: Relation between consumed sample counts and test performance, final matched count of the selected features, and the most correlated features from each layer of Gemma 2 2B. The dotted lines show the baseline of default LLM performance and constrained decoding performance on MMLU answer options.

- **Factuality:** SimpleQA (Wei et al., 2024)
- **Safety:** HarmBench (Mazeika et al., 2024), XSTest (Röttger et al., 2024)

For safety benchmarks, both the refusal benchmark HarmBench and the overrefusal benchmark XSTest are included to evaluate not only the steering ability for rejection but also the model’s capability to identify the context of requests.

Evaluation Metrics: For multiple-choice tasks (MMLU, MMLU-Pro, BBQ), exact match accuracy is used. For safety benchmarks, 1 - ASR (Attack Success Rate) is computed using a small refusal-detection language model. SimpleQA performance is measured using a small STS language model to match the expected answer, with more details in Appendix A.2.

A standard train-validation-test split is used for the CorrSteer pipeline. The training dataset is used to extract correlated SAE features, and the validation dataset is used to filter the most correlated features. The test dataset is used to evaluate the performance of the CorrSteer pipeline. Detailed configurations are provided in Appendix A.1.

Pooling Strategy: Two pooling strategies are available for coefficient and correlation calculations: max-pooling and mean-pooling. For multi-token generation tasks, max-pooling consistently

outperforms mean-pooling, as empirically demonstrated in Table 5, likely due to its better capture of peak feature activations relevant to task success. However, for coefficient calculation in longer generation tasks such as GSM8K reasoning, mean-pooling is preferred as max-pooling produces excessively large coefficient values. Applying these large coefficients to every generated token degrades performance, leading to the adoption of mean-pooling for reasoning tasks.

4.2 SER Comparison

This study aims to demonstrate that CorrSteer can improve benchmark performance while maintaining low side effects. Using the defined SER, CorrSteer’s SER is compared against fine-tuning for a question-answering dataset. Additionally, the SER is compared when allowing or enforcing negatively correlated features, supporting the claim of SAE’s positive-tinted behavior. Finally, the SER is compared when pooling on every token rather than the inference-time generated token.

4.3 Feature Interpretability Analysis

The analysis is enhanced by incorporating feature descriptions from Neuronpedia, providing semantic interpretations of selected features. Safe/unsafe tendency analysis is conducted, along with task-

wise and feature-level interpretations. This interpretability analysis validates that our correlation-based selection identifies semantically meaningful and task-relevant features, supporting the causal hypothesis underlying our approach.

5 Results

5.1 Generation Benchmark Performance

Table 1 and Table 2 present comprehensive results across our evaluation benchmarks. CorrSteer demonstrates consistent improvements across most tasks, including question answering, bias mitigation, and safety benchmarks, with the exception of the GSM8K reasoning task. This pattern suggests that our method enhances model adherence to static task requirements without introducing dynamic behaviors.

Method Comparison In most cases, CorrSteer-A and CorrSteer-P demonstrate the highest performance, with CorrSteer-P showing particular dominance in LLaMA 3.1 8B. This observation is attributed to the less disentangled nature of Llama Scope features from superposition, which necessitates more aggressive pruning.

Safety Benchmarks:

For HarmBench, selected features enhance the model’s ability to refuse harmful requests, achieving a 22.9% improvement. However, for XSTest, improvement is limited to 1% due to the overrefusal characteristics of the benchmark. This is an expected result due to the static nature of CorrSteer, which hinders the ability to clearly distinguish between benign and harmful requests.

Observations:

Results for both Gemma-2 2B and LLaMA 3.1 8B, presented in Table 1 and Table 2, demonstrate consistent patterns, with CorrSteer variants showing systematic improvements.

Multi-layer Superiority: Simultaneous multi-layer steering approaches, such as CorrSteer-A and CorrSteer-P, consistently outperform the single-layer CorrSteer-1 approach across all benchmarks, consistent with findings from other researchers (Liu et al., 2024; Zhao et al., 2025). This observation suggests that enabling features from different layers to compete globally yields superior task-relevant selections compared to layer-wise optimization.

Limited Factuality Impact: SimpleQA shows minimal improvement, confirming that CorrSteer enhances task adherence without introducing external factual knowledge. This is a desirable property,

as it indicates the method improves model behavior rather than injecting information not present in the original model.

Comparison with Fine-tuning: CorrSteer demonstrates competitive performance while maintaining significantly lower side effect rates compared to fine-tuning. On MMLU, CorrSteer-A achieves higher accuracy (56.32%) than fine-tuning (55.85%) with substantially lower SER (0.202 vs 0.407). Similarly, on GSM8K, CorrSteer variants outperform fine-tuning in accuracy while maintaining lower side effect rates across all tasks.

5.2 Feature Analysis

Selected features demonstrate varying degrees of alignment with task requirements: while some benchmarks show consistent alignment through structured output features for multiple-choice tasks, refusal-related features for safety benchmarks, and task-specific semantic features for domain-specific evaluations, others exhibit less consistent patterns. The interpretability of selected features, validated through Neuronpedia descriptions, provides confidence in the semantic relevance of our correlation-based selection process. Notably, feature activation frequencies vary significantly across tasks, with performance improvements correlating with distinct activation patterns (Appendix 8). Analysis of selected features reveals semantically meaningful patterns aligned with task requirements:

Safety-related Features: HarmBench and BBQ demonstrate substantial improvements through selected safety-related features that enhance neutrality and refusal behavior. This interpretability analysis is covered in detail in Section 6.1.

5.3 SER Analysis

Table 3 presents the Safety Evaluation Rate analysis, demonstrating CorrSteer’s ability to minimize side effects while improving task performance.

6 Discussion

This work establishes a viable and efficient approach for SAE-based steering, providing effective control across diverse applications. The interpretable feature combinations that yield performance improvements support the hypothesis that linear correlation serves as a meaningful unit for interpretable AI capabilities.

Table 1: Performance comparison between baseline and CorrSteer variants across BBQ, MMLU, MMLU-Pro, GSM8K, SimpleQA, and XSTest on Gemma 2B. Results show accuracy (%) for all tasks.

Task	Baseline	CorrSteer-1	CorrSteer-P	CorrSteer-A	Fine-tuning
BBQ Ambig	59.10	62.38	66.65	62.08	-
BBQ Disambig	75.42	75.70	77.04	76.53	-
HarmBench	44.64	46.07	51.79	67.50	-
MMLU	52.23	52.82	55.83	56.32	55.85
MMLU-Pro	30.30	30.44	30.82	31.01	33.16
GSM8K	54.74	54.44	53.53	54.44	47.38
SimpleQA	3.63	3.76	3.96	3.80	-
XSTest	86.35	86.67	86.03	87.30	-

Table 2: Performance comparison between baseline and CorrSteer variants across BBQ, MMLU, MMLU-Pro, HarmBench, SimpleQA, and XSTest on LLaMA 3.1 8B. Results show accuracy (%) for all tasks.

Task	Baseline	CorrSteer-1	CorrSteer-P	CorrSteer-A
BBQ Ambig	83.97	83.98	87.10	86.83
BBQ Disambig	90.07	90.13	90.33	90.30
HarmBench	0.71	0.36	15.71	17.86
MMLU	61.41	61.51	61.73	61.71
MMLU-Pro	32.13	32.55	35.08	34.71
SimpleQA	0.43	0.51	0.43	0.43
XSTest	61.27	62.22	62.22	58.41

6.1 Feature Inspection

A notable finding is that features selected by CorrSteer demonstrate task-relevant patterns that align with theoretical expectations. Math-related features are consistently discovered across all tasks, proving beneficial even for bias mitigation and safety tasks. This universal correlation between mathematical features and accuracy aligns with DeepSeekMath (Shao et al., 2024)’s findings, where further pre-training on math-focused corpora yielded performance improvements across diverse tasks.

For BBQ features in LLaMA 3.1 8B, we observe:

- **L15/25166 themes of neutrality and balance in discourse** (coeff: 0.259, corr: 0.433)
- **L25/10753 expressions of perception or belief in social dynamics** (coeff: 1.147, corr: 0.428)

While bias-related features were expected for the BBQ benchmark, these neutrality-focused features demonstrate high positive correlation. Conversely, explicit bias-related and choice-making features exhibit negative correlations:

- **L8/8123** questions that ask for truthfulness or correctness regarding options or statements (coeff: 3.725, corr: -0.133)
- **L17/9134** choice-related phrases and expressions of preference (coeff: 2.379, corr: -0.451)
- **L19/15745** phrases related to decision-making and choice, particularly in the context of parenting and social interactions (coeff: 9.740, corr: -0.464)

These findings suggest that task-specific induced features contribute more to sample accuracy than meta-cognitive recognition features. Our ablation study further demonstrates that SAE-based sparse feature selection consistently outperforms raw activation steering across all evaluated tasks (Table 7).

6.2 Feature Set Transferability

The transferability of CorrSteer feature sets is evaluated across MMLU, MMLU-Pro, and BBQ benchmarks. Interestingly, our cross-task experiments reveal that MMLU features outperform task-specific features on BBQ Ambig and achieve comparable performance on MMLU-Pro, suggesting that some feature sets capture more generalizable reasoning

Table 3: Safety Evaluation Rate (SER) analysis for CorrSteer variants across different benchmarks on Gemma 2 2B. SER values closer to 0 indicate better safety performance.

Task	CorrSteer-1			CorrSteer-P			CorrSteer-A			Fine-tuning		
	SER	neg	pos	SER	neg	pos	SER	neg	pos	SER	neg	pos
BBQ Ambig	0.000	0	658	0.000	0	1532	0.076	53	649	-	-	-
BBQ Disambig	0.167	45	59	0.153	111	164	0.257	65	112	-	-	-
HarmBench	0.250	2	6	0.143	4	24	0.043	3	67	-	-	-
MMLU	0.355	11	20	0.172	249	286	0.202	264	299	0.407	1108	1616
MMLU-Pro	0.421	8	11	0.423	30	41	0.419	39	54	0.461	357	418
GSM8K	0.556	20	16	0.674	31	15	0.516	63	59	0.647	213	116
SimpleQA	0.167	1	5	0.188	3	13	0.353	6	11	-	-	-
XSTest	0.333	7	10	0.520	7	10	0.467	14	5	-	-	-

Table 4: Safety Evaluation Rate (SER) analysis for CorrSteer variants on Llama models across different benchmarks. SER values closer to 0 indicate better safety performance.

Task	CorrSteer-1			CorrSteer-P			CorrSteer-A		
	SER	neg	pos	SER	neg	pos	SER	neg	pos
BBQ Ambig	0.496	141	143	0.017	11	651	0.025	15	599
BBQ Disambig	0.433	45	59	0.404	111	164	0.367	65	112
HarmBench	0.333	3	6	0.226	7	24	0.171	6	29
MMLU	0.488	118	124	0.465	249	286	0.469	264	299
MMLU-Pro	0.355	11	20	0.280	40	103	0.310	45	100
SimpleQA	0.000	0	1	-	0	0	0.500	4	4
XSTest	0.412	7	10	0.412	7	10	0.737	14	5

patterns (Table 8). This transferability is attributed to their shared multiple-choice format, which requires similar structural feature patterns.

6.3 Task-Level Circuit

CorrSteer’s multi-layer approach relates to neural network circuit discovery research (Olah et al., 2020; Elhage et al., 2021). While emerging works focus on discovering task-specific circuits (Conmy et al., 2023; Marks et al., 2025; Ameisen et al., 2025; Lindsey et al., 2025; Sun, 2025), our steering vectors that work simultaneously across layers can be conceptualized as additive subgraphs of optimized task circuits, though they lack explicit interpretation of interactions and causality.

6.4 Side Effect Ratio

The primary challenge in AI steering concerns robustness for industrial applications, necessitating precise control mechanisms. Direct steering at each layer without updating original parameters based on token prediction distributions represents a key approach to minimize side effects. Theoretically, separating steering vectors across different activation spaces minimizes mutual interference in superpositioned states (Elhage et al., 2022).

6.5 Pooling Strategy Analysis

The results reveal interesting patterns across different pooling strategies (Table 5). Mean-pooling shows severe degradation on multi-token generation tasks (HarmBench: 0.00%, XSTest: 53.65%) where responses require multiple tokens. All-token pooling shows degraded performance on every task compared to max-pooling’s inference-time aggregation. This suggests that max-pooling better captures the critical activations needed for effective steering across all task types, while all-token pooling may introduce noise by including irrelevant token positions, and mean-pooling dilutes important signals by averaging across all tokens in longer sequences.

6.6 Computational Efficiency

Our streaming correlation implementation achieves O(1) memory complexity with respect to training set size, making the approach scalable to large datasets. The pipeline exhibits computational efficiency with minimal sample requirements (200-400) as demonstrated in Figure 4 and completes feature extraction within minutes. Static feature sets and coefficients at inference time eliminate SAE dependency during deployment.

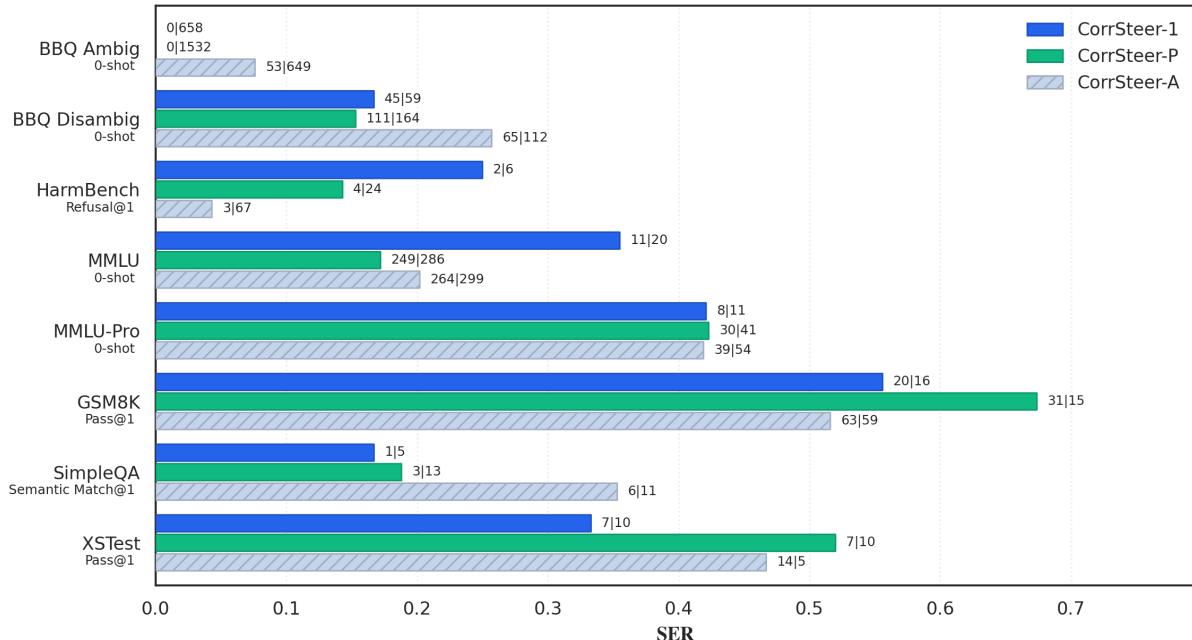


Figure 5: SER comparison between different CorrSteer variants for Gemma 2 2B.

6.7 Practical Applicability

CorrSteer operates as an auxiliary mechanism that captures correlations difficult to detect through supervised fine-tuning, applicable with minimal side effects even after fine-tuning. The automated pipeline enables rapid deployment across tasks and domains without hyperparameter tuning. The method shows effectiveness across two model families, as demonstrated on both Gemma-2 2B and LLaMA-3.1 8B models. Importantly, our approach has broader implications for AI safety, as demonstrated by its effectiveness in both bias mitigation and amplification (Table 9), highlighting the need for responsible deployment of such steering capabilities.

7 Conclusion

This work introduces CorrSteer, a fully automated pipeline for SAE-based language model steering that leverages correlation analysis. Our approach addresses key limitations of existing SAE steering methods by identifying task-relevant features without requiring manual feature exploration or contrastive datasets. Experimental validation across diverse benchmarks demonstrates CorrSteer’s effectiveness, consistently improving performance on question answering, bias mitigation, and safety evaluation tasks. Selected features for safety, mathematical, and refusal-related tasks reveal the underlying objectives and required capabilities that drive

task performance.

Future Work

Several promising directions emerge from this work: **Prompt Engineering Comparison:** Future studies should compare CorrSteer with prompt engineering approaches, as prompt-based methods are expected to exhibit higher SER due to their less targeted intervention mechanisms. **Dynamic Steering for Reasoning:** The performance degradation observed in GSM8K reasoning tasks suggests the need for dynamic steering approaches that can adapt to the sequential nature of mathematical problem-solving, moving beyond static feature interventions. **Orthogonal Feature Projection:** To further minimize side effects, future work could explore feature filtering techniques that project out components already activated in baseline features before applying steering, potentially reducing interference with existing model capabilities.

Acknowledgments

The authors thank the teams behind Gemma Scope and LLaMA Scope for providing high-quality SAE releases that enabled this research. The authors also acknowledge Neuronpedia for providing automated feature descriptions that enhanced our interpretability analysis.

Limitations

The fundamental limitation of steering vectors is their static nature, which prevents adaptation to dynamic model behaviors. This constraint particularly affects reasoning tasks like GSM8K, where static steering cannot adequately handle the sequential nature of mathematical problem-solving. Our evaluation focuses primarily on discriminative and short-form generation tasks; long-form generation and creative tasks may require different approaches or modifications to our method.

References

- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. [Circuit tracing: Revealing computational graphs in language models](#). *Transformer Circuits Thread*.
- Dana Arad, Aaron Mueller, and Yonatan Belinkov. 2025. [Saes are good for steering – if you select the right features](#). *Preprint*, arXiv:2505.20063.
- Joseph Bloom. 2024. [Open source sparse autoencoders for all residual stream layers of gpt2 small](#).
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. [Https://transformercircuits.pub/2023/monosemantic-features/index.html](https://transformercircuits.pub/2023/monosemantic-features/index.html).
- Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. [Improving steering vectors by targeting sparse autoencoder features](#). *Preprint*, arXiv:2411.02193.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. [Towards automated circuit discovery for mechanistic interpretability](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Bartosz Cywiński and Kamil Deja. 2025. [SAeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders](#). In *Forty-second International Conference on Machine Learning*.
- 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. <https://anthropic.com/research/evaluating-feature-steering>. Anthropic Research.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. [Toy models of superposition](#). *Transformer Circuits Thread*.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. [A mathematical framework for transformer circuits](#). *Transformer Circuits Thread*.
- Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. [Sparse overcomplete word vector representations](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.
- Dron Hazra, Max Loeffler, Murat Cubuktepe, Levon Avagyan, Liv Gorton, Mark Bissell, Owen Lewis, Thomas McGrath, and Daniel Balsam. 2025. Under the hood of a reasoning model. *Goodfire Research Blog*. <https://goodfire.ai/blog/under-the-hood-of-a-reasoning-model>.
- Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. [Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders](#). *Preprint*, arXiv:2410.20526.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2023. [Sparse autoencoders find highly interpretable features in language models](#). In *The Twelfth International Conference on Learning Representations*.

- Johnny Lin Joseph Bloom. 2024. Understanding sae features with the logit lens.
- Theo King, Zekun Wu, Adriano Koshiyama, Emre Kazim, and Philip Colin Treleaven. 2024. HEARTS: A holistic framework for explainable, sustainable and robust text stereotype detection. In *Neurips Safe Generative AI Workshop 2024*.
- Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. 2024. Saes (usually) transfer between base and chat models. Alignment Forum.
- Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.
- Jack Lindsey, Emmanuel Ameisen, Neel Nanda, Stepan Shabalin, Mateusz Piotrowski, Tom McGrath, Michael Hanna, Owen Lewis, Curt Tigges, Jack Merullo, Connor Watts, Gonçalo Paulo, Joshua Batson, Liv Gorton, Elana Simon, Max Loeffler, Callum McDougall, and Johnny Lin. 2025. The circuits research landscape: Results and perspectives. *Neuronpedia*.
- Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. *Preprint*, arXiv:2311.06668.
- Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and Rene Vidal. 2024. PaCE: Parsimonious concept engineering for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Samuel Marks, Can Rager, Eric J Michaud, Yonatan Beilinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaei, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Forty-first International Conference on Machine Learning*.
- Aashiq Muhammed, Jacopo Bonato, Mona T. Diab, and Virginia Smith. 2025. SAEs can improve unlearning: Dynamic sparse autoencoder guardrails for precision unlearning in LLMs. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.
- Kyle O'Brien, David Majercak, Xavier Fernandes, Richard G. Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. 2025. Steering language model refusal with sparse autoencoders. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.
- Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. 2025. Evaluating neuron explanations: A unified framework with sanity checks. In *Forty-second International Conference on Machine Learning*.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. <Https://distill.pub/2020/circuits/zoom-in>.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.
- Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *Preprint*, arXiv:2407.14435.
- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander 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)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

- Lewis Smith, Senthooran Rajamanoharan, Arthur Conmy, Callum McDougall, Tom Lieberum, János Kramár, Rohin Shah, and Neel Nanda. 2025. Negative results for saes on downstream tasks and deprioritising sae research. <https://www.lesswrong.com/posts/4uXCAJNuPKtKBsi28/sae-progress-update-2-draft>. DeepMind Mechanistic Interpretability Team Progress Update #2.
- 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, Chandrasekaran Balaganesh, Tan Guoxian, and Ming YAN. 2025. Interpretable steering of large language models with feature guided activation additions. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Alessandro Stolfo, Ben Peng Wu, and Mrinmaya Sachan. 2025. Antipodal pairing and mechanistic signals in dense sae latents. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.
- Alan Sun. 2025. Circuit stability characterizes language model generalization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9025–9040, Vienna, Austria. Association for Computational Linguistics.
- Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.
- Llama Team. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.
- 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, and 3 others. 2024. Scaling monosematicity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.
- Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. 2025. Model unlearning via sparse autoencoder subspace guided projections. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.
- Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136, Albuquerque, New Mexico. Association for Computational Linguistics.
- Dylan Zhou, Kunal Patil, Yifan Sun, Karthik Iakshmanan, Senthooran Rajamanoharan, and Arthur Conmy. 2025. LLM neurosurgeon: Targeted knowledge removal in LLMs using sparse autoencoders. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

A Appendix

A.1 Implementation Details

Feature Extraction: Feature selection employs 4,000 samples across all datasets. For fair comparison, the same samples are used for training fine-tuning baselines. When datasets contain fewer than 4,000 samples, we use all available data. For datasets without predefined train/validation/test splits, we allocate 27% for training, 3% for validation, and 70% for testing. GSM8K uses 1,000 samples for feature selection with 50 samples reserved for validation.

Feature Steering: Steering interventions are applied at the pre-execution stage of each transformer layer. The first layer is excluded from steering as the token embedding layer predominantly contains spurious correlations unrelated to the target tasks.

Fine-tuning Fine-tuning is performed using AdamW optimizer with learning rate 1e-5 (reduced to 5e-6 for small datasets <2000 samples), weight decay 0.01, and gradient clipping at norm 1.0. The training schedule includes 3% warmup steps followed by cosine annealing decay. Training proceeds for one epoch with 4,000 samples, using exact target supervision where prompt tokens are masked with -100 labels and only target spans contribute to the loss.

A.2 Generation Benchmark Results

Evaluation Models: Two specialized models are employed for evaluation. The DistillRoBERTa model¹ is used to identify the rejection of harmful requests, while the BERT STS model² is used for matching generated answers against expected responses.

A.3 Additional Results

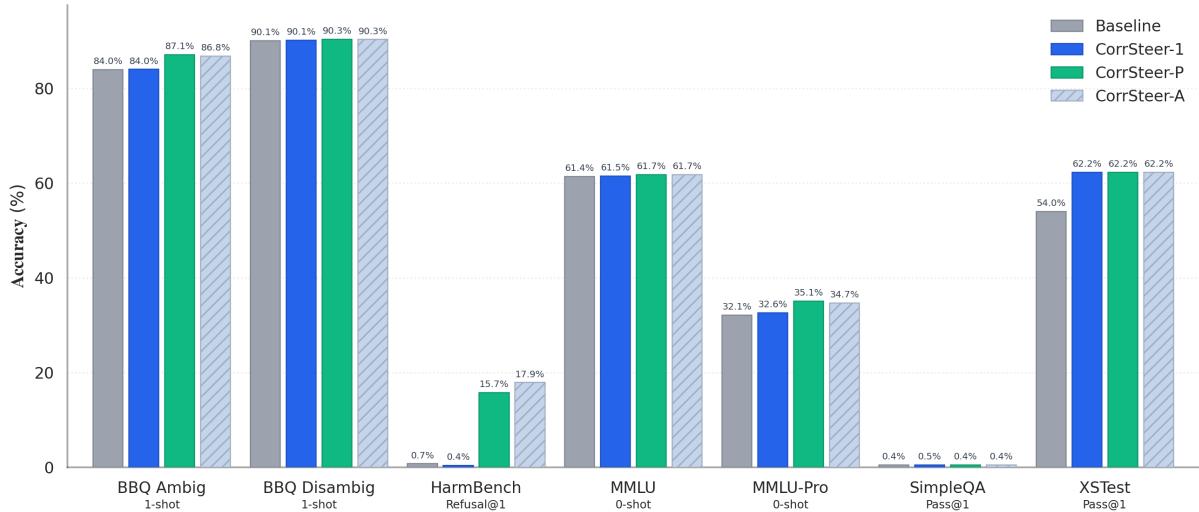


Figure 6: Benchmark performance of CorrSteer variants compared with the default baseline on LLaMA 3.1 8B.

Task-Specific Analysis *MMLU*: The global method selects features related to structured output formatting, addressing Gemma-2’s tendency to generate tokens outside the required A/B/C/D options. Post-steering, this hallucination issue is largely resolved.

MMLU-Pro: A similar issue occurs more severely due to the 10 options in MMLU-Pro. Constrained decoding, which samples tokens exclusively from available options, is applied to improve the model’s authentic capability, resulting in performance that remains higher than baseline, with CorrSteer-A achieving maximum performance.

¹<https://huggingface.co/datasets/huggingface-community/distill-roberta-toxicity-r1>

²https://huggingface.co/datasets/HuggingFaceH4/stsb_multi_mt

BBQ: Similar improvements in format adherence are observed, with selected features promoting appropriate response structure.

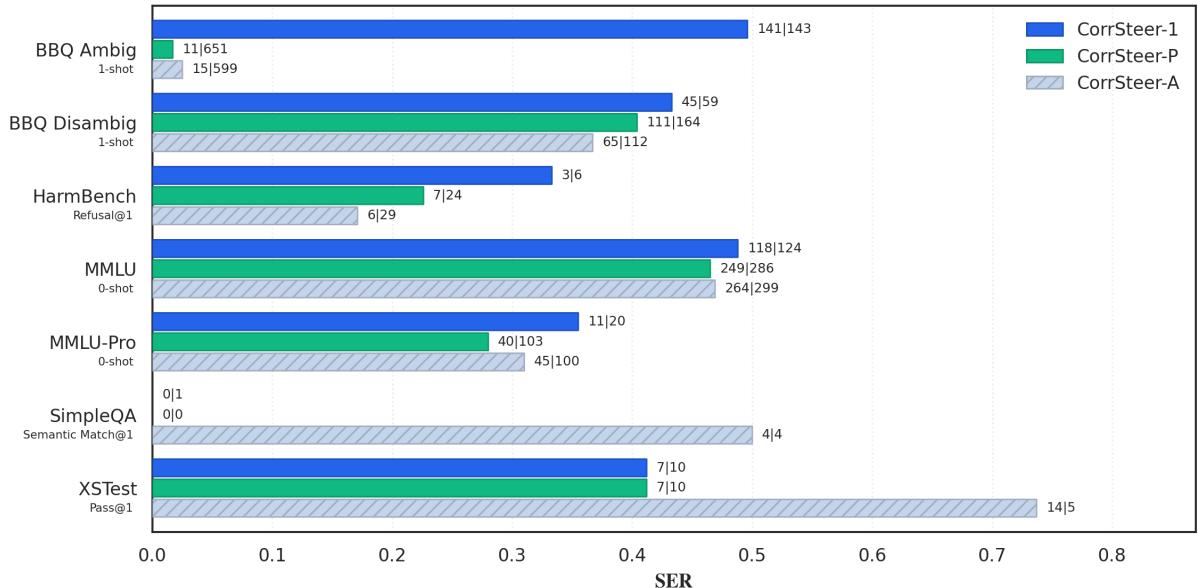


Figure 7: SER comparison across datasets between different CorrSteer variants on LLaMA 3.1 8B.

Feature Frequency Analysis We observe a strong correlation between feature activation frequency and CorrSteer’s performance improvements across tasks. As demonstrated in Figure 8, HarmBench exhibits consistently high activation frequencies across all layers, while SimpleQA shows frequencies approaching zero.

This pattern contrasts with the typical sparse activation nature of SAE features, where low frequency activation (below 5%) is considered normal and interpretable, while higher frequencies typically indicate non-interpretable (Stolfo et al., 2025; Smith et al., 2025). However, discovering task-specific features with near-100% activation frequency suggests these features are deeply related to the task requirements, resulting in substantial performance improvements for such tasks. Even for tasks with lower feature frequencies, CorrSteer maintains its advantage by preserving low SER values.

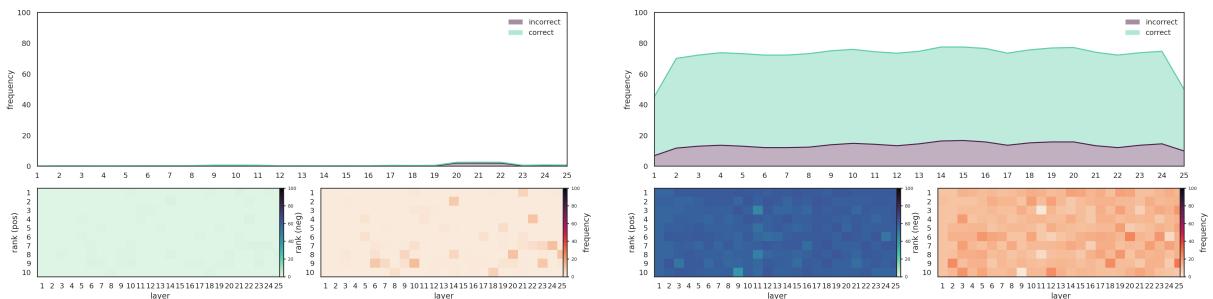


Figure 8: Frequency of activation samples across layers of Gemma 2 2B for SimpleQA (left) and HarmBench (right) tasks.

A.4 Ablation Study

Pooling Strategy For generation tasks requiring multiple tokens, max-pooling is employed over valid token positions to aggregate feature activations before correlation computation. Our comprehensive evaluation confirms max-pooling’s superiority over alternative strategies (Table 5). However, for coefficient calculation in longer generation tasks such as GSM8K reasoning, mean-pooling is preferred as max-pooling produces excessively large coefficient values that degrade performance when applied to every generated token.

We also evaluate alternative pooling strategies including mean-pooling and all-token pooling for feature activation aggregation. The comparison results are presented in Table 5.

Table 5: Pooling strategy comparison on Gemma 2 2B using CorrSteer-A. For single-token generation such as MMLU, MMLU-Pro, and BBQ, mean-pooling naturally achieves identical performance to max-pooling since only one token is generated, while all-token pooling shows degraded performance. Mean-pooling shows severe degradation on multi-token generation tasks, demonstrating the superiority of max-pooling.

Task	Baseline	Max-pooling	Mean-pooling	All-token
MMLU	52.23	56.32	56.32	52.91
MMLU-Pro	30.30	31.00	31.00	30.16
BBQ Disambig	75.42	76.53	76.53	75.00
BBQ Ambig	59.10	62.08	62.08	57.98
HarmBench	44.64	67.50	0.00	47.14
XSTest	86.35	87.30	53.65	86.35
SimpleQA	3.63	3.80	3.76	3.73

Negative Correlation Features To validate our design choice of using only positively correlated features, we conduct ablation experiments using negatively correlated features for steering. We compare two approaches: single-layer negative steering (CorrSteer-1 with negative features) and multi-layer negative steering (CorrSteer-A with negative features).

Table 6: Performance comparison between positive and negative correlation feature steering on Gemma 2 2B. Negative correlation features consistently show poor performance, validating our positive-only approach.

Task	Baseline	CorrSteer-A	Negative-1	Negative-G
MMLU	52.23	56.32	52.24	49.45
MMLU-Pro	14.00	17.56	14.24	0.66
BBQ Disambig	75.42	76.53	75.37	12.15
BBQ Ambig	59.10	62.08	59.22	60.85
HarmBench	44.64	67.50	44.64	47.86
XSTest	86.35	87.30	86.35	86.67
SimpleQA	3.63	3.80	3.76	3.76

The results demonstrate that negative correlation features provide minimal improvement in single-layer steering and often cause severe performance degradation in multi-layer steering. Notably, MMLU-Pro drops to 0.66% and BBQ Disambig to 12.15% with negative multi-layer steering, confirming that negative correlations often represent spurious patterns rather than causal relationships. Additionally, combining positive and negative features simultaneously yields inferior performance compared to positive-only selection. This validates our approach of using only positively correlated features, which aligns with the non-negative nature of SAE activations.

Table 7: Performance comparison between raw activation steering and SAE-decoded steering on Gemma 2 2B. Decoding adds SAE decoder bias term for the first layer, while Decoding-A adds multi-layer feature directions as CorrSteer-A.

Task	Baseline	Raw Activation	Decoding-1	Decoding-A	CorrSteer-A
MMLU	52.23	49.85	55.38	54.38	56.32
MMLU-Pro	30.30	27.17	29.79	29.93	31.00
BBQ Disambig	75.42	75.71	77.00	75.03	76.53
BBQ Ambig	59.10	58.42	54.00	55.76	62.08

Raw Activation Steering To validate the effectiveness of SAE-based sparse feature selection, we compare steering performance using raw residual stream activations. The results demonstrate a clear performance hierarchy: CorrSteer-A > SAE Decoding > Raw Activation across all evaluated tasks, which is explainable by Superposition Hypothesis (Elhage et al., 2022). One exception occurred in BBQ Disambig, where Decoding-1 shows better performance than CorrSteer-A. However, Decoding-1 failed to show robustness across benchmarks, frequently degrading performance while CorrSteer-A shows consistent performance across all tasks.

SAE Decoder Bias Adding SAE decoder bias terms alongside selected features improves performance only at single-token generation tasks (BBQ, MMLU, MMLU-Pro). This effect appears related to attention sink mechanisms (Xiao et al., 2024), where increased residual stream norms amplify attention patterns in subsequent layers, acting similar to "response prefix" (Hazra et al., 2025). For constrained generation tasks, this norm amplification reduces hallucination by strengthening adherence to output format constraints. However, this enhancement is incompatible with multi-layer steering and diminishes when applied across multiple layers or tokens, with excessive application potentially causing model collapse.

A.5 Cross-Task Feature Transferability

To evaluate the transferability of selected features across different tasks, we conduct cross-task steering experiments where features selected for one task are applied to different target tasks. This analysis provides insights into the generalizability of task-specific feature sets.

Table 8: Cross-task feature transferability results on Gemma 2 2B. Features selected from source tasks (rows) are applied to target tasks (columns). Results show accuracy (%) with baseline performance in parentheses. MMLU-Pro results do not use constrained decoding, achieving 17.56% compared to unconstrained baseline (14.00%).

Source → Target	MMLU	MMLU-Pro	BBQ Disambig	BBQ Ambig
MMLU	56.32 (52.23)	19.67 (14.00)	74.62 (75.42)	64.01 (59.10)
MMLU-Pro	55.73 (52.23)	17.56 (14.00)	76.10 (75.42)	60.97 (59.10)
BBQ Disambig	54.74 (52.23)	16.11 (14.00)	76.53 (75.42)	60.85 (59.10)
BBQ Ambig	53.85 (52.23)	11.01 (14.00)	76.10 (75.42)	62.08 (59.10)

The results reveal several interesting patterns: (1) MMLU and MMLU-Pro features show reasonable cross-transferability, likely due to their shared multiple-choice format and reasoning requirements, (2) BBQ features demonstrate good transferability to MMLU tasks, suggesting that bias mitigation features capture general reasoning capabilities, and (3) features optimized for specific tasks consistently outperform transferred features, validating the importance of task-specific feature selection. These findings support our discussion of limited but meaningful transferability among structurally similar tasks.

A.6 Text Classification Validation

To validate the effectiveness of correlation-based feature selection, we conduct controlled experiments on text classification tasks where ground truth labels provide clear supervision signals. The experiments utilize GPT-2 (Radford et al., 2019) with publicly available SAEs from Bloom et al. (Bloom, 2024) on the bias-focused text classification dataset EMGSD (King et al., 2024).

For each bias category, we extract the most correlated features using max-pooling over all text tokens, then apply steering by either adding positively correlated features or subtracting negatively correlated features. Steering effectiveness is evaluated using the same classifier employed in the original dataset.

Table 9: Bias steering effectiveness across different demographic categories on EMGSD dataset. Mitigation reduces bias scores, while amplification increases them.

Category	Mitigation		Amplification	
	Baseline	CorrSteer	Biased	CorrSteer
Gender	0.177	0.616	0.897	0.922
LGBTQ+	0.091	0.561	0.941	0.882
Nationality	0.125	0.732	0.937	0.945
Profession	0.128	0.625	0.890	0.921
Race	0.308	0.769	0.846	0.846
Religion	0.109	0.655	0.945	0.928

Results demonstrate that correlation-selected features provide effective steering control across all demographic categories (Table 9). Our steering approach effectively reduces bias, with steered outputs showing substantially lower bias scores compared to biased baselines, supporting the generalizability of our approach across different domains and SAE training paradigms.

A demonstration of our bias mitigation results is available at <https://huggingface.co/spaces/seonglae/CorrSteer>, showcasing real-time steering capabilities.

A.7 Complete Feature Lists

This section presents the complete feature lists for each task, showing the top-1 features aggregated from all layers. Each feature is labeled with the format L{layer}/{index} to identify its layer and index position. Features selected by CorrSteer-P after pruning are highlighted in **bold**.

Each feature entry includes the feature description along with its coefficient and correlation value. SAE feature descriptions are obtained through the Neuronpedia API (<https://www.neuronpedia.org/>), providing automated semantic interpretations of selected features. Feature indices are hyperlinked to their corresponding Neuronpedia pages for detailed analysis.

Feature descriptions that are well-aligned with the target task are highlighted in **bold**, and the highest correlations for each task are also emphasized in **bold**. Following each layer’s highest correlated feature, we include additional relevant features listed below.

A.7.1 Gemma-2B

BBQ (Ambiguous)

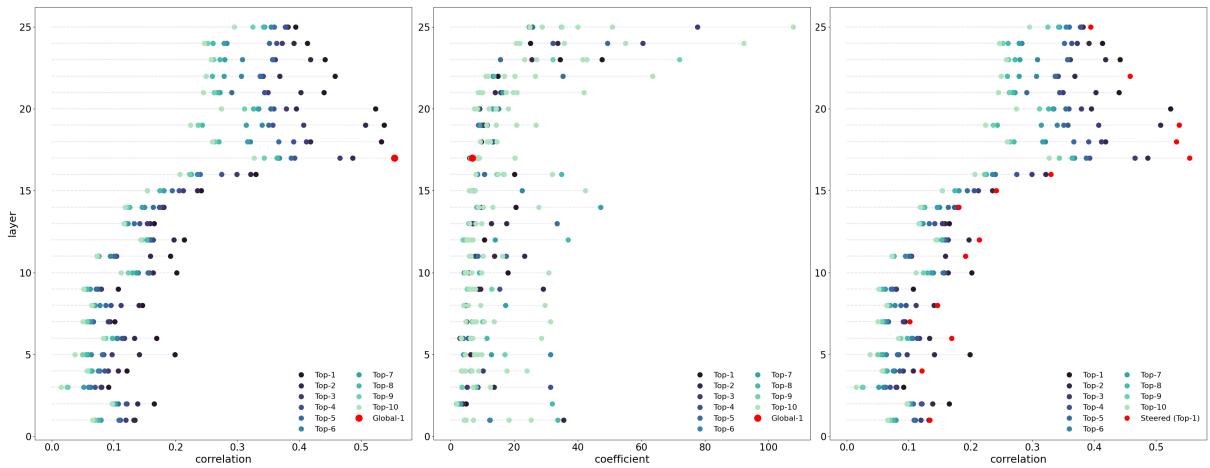


Figure: Top correlated features with selected features from CorrSteer-P with BBQ ambig on coefficient in each layer of Gemma 2 B2.

- L1/6088 specific formatting or structural elements within text, such as timestamps and code (coeff: 2.280, corr: 0.134)

- [L2/15089](#) key actions and processes related to achievements and collaboration (coeff: 4.898, corr: 0.166)
- [L3/6151](#) references to statistical or numerical data in research contexts (coeff: 3.537, corr: 0.091)
- [L4/11047](#) certain types of mathematical or programming syntax (coeff: 2.854, corr: 0.121)
- [L5/7502](#) expressions of honesty and self-awareness in discourse (coeff: 3.117, corr: 0.199)
- [L6/324](#) structured sentences that present facts, warnings, or errors, often with an emphasis on important details (coeff: 2.886, corr: 0.169)
- [L7/4487](#) the presence of detailed structured elements within a document, such as headings or separators in a legal or formal layout (coeff: 4.996, corr: 0.102)
- [L8/4669](#) special tokens or specific formatting in the text (coeff: 4.378, corr: 0.147)
- [L9/1435](#) elements related to copyright and licensing information (coeff: 8.737, corr: 0.107)
- [L10/4557](#) interactions involving guessing or determining the correctness of information (coeff: 4.246, corr: 0.202)
- [L11/6144](#) return statements in code (coeff: 4.347, corr: 0.192)
- [L12/15862](#) punctuation marks and formatting elements in the text (coeff: 2.718, corr: 0.214)
- [L13/4379](#) punctuation symbols and their frequency (coeff: 6.779, corr: 0.165)
- [L14/12922](#) dialogue or conversational exchanges involving questioning and responses (coeff: 1.754, corr: 0.181)
- [L15/12813](#) medical terms related to respiratory health and conditions (coeff: 3.537, corr: 0.242)
- [L16/9006](#) declarations regarding conflicts of interest and funding in research publications (coeff: 2.606, corr: 0.330)
- [L17/11021](#) phrases related to scientific research and findings (coeff: 6.777, corr: **0.554**)
- [L18/14447](#) references to medical data and statistics (coeff: 9.667, corr: 0.533)
- [L19/11289](#) assignment and return statements in programming contexts (coeff: 10.429, corr: 0.538)
- [L20/2040](#) occurrences of logical values and conditions in programming or data handling contexts (coeff: 9.166, corr: 0.523)
- [L21/8433](#) keywords related to programming functions and their definitions (coeff: 5.983, corr: 0.440)
- [L22/10377](#) code snippets that include assignments and return statements (coeff: 14.919, corr: 0.458)
- [L23/6394](#) structured data or code-like formats (coeff: 34.482, corr: 0.442)
- [L24/14051](#) references to education systems and their impact on health initiatives (coeff: 25.098, corr: 0.413)
- [L25/12534](#) references to emotional states or descriptions of personal experiences (coeff: 18.414, corr: 0.394)

Additional relevant features:

- [L8/8123](#) questions that ask for truthfulness or correctness regarding options or statements (coeff: 3.725, corr: -0.133)
- [L17/9134](#) choice-related phrases and expressions of preference (coeff: 2.379, corr: -0.451)
- [L19/15745](#) phrases related to decision-making and choice, particularly in the context of parenting and social interactions (coeff: 9.740, corr: -0.464)

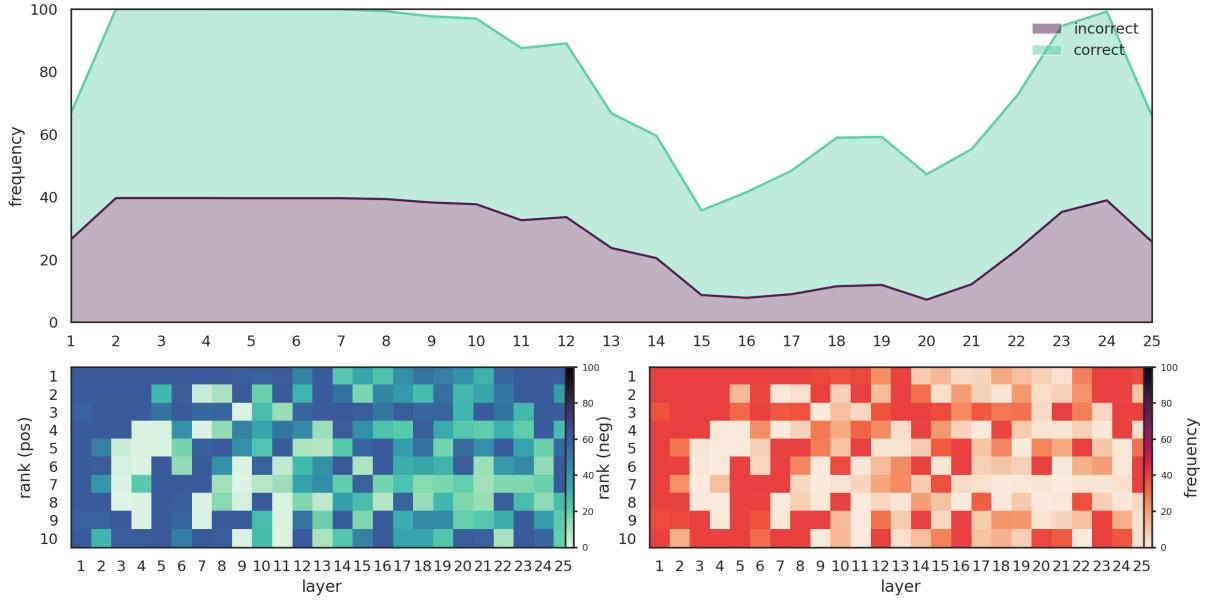


Figure: Top correlated features with BBQ ambig on frequency in each layer of Gemma 2 2B.

BBQ (Disambiguous)

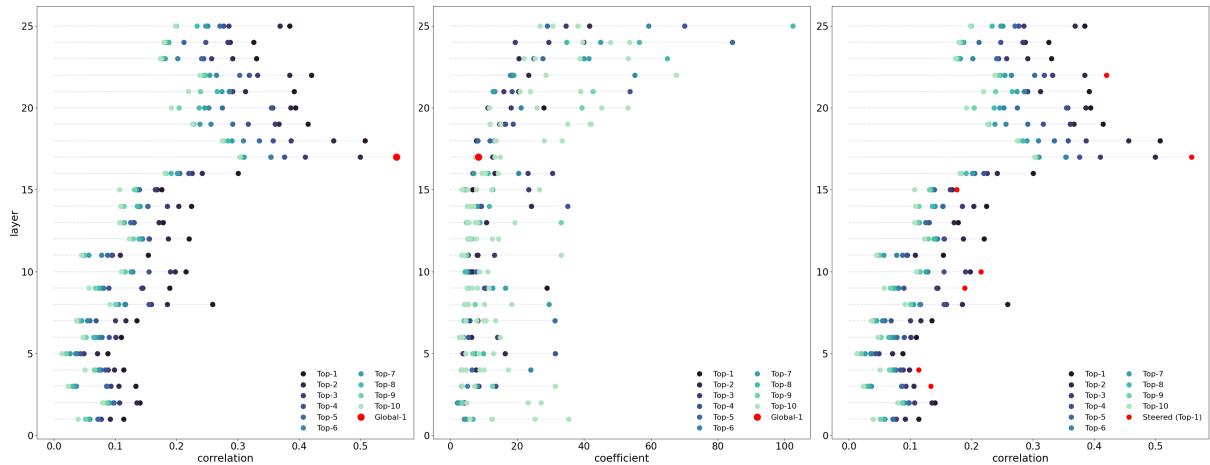


Figure: Top correlated features with selected features from CorrSteer-P with BBQ disambig on coefficient in each layer of Gemma 2 2B.

- **L1/7001** code structure and elements in programming, particularly related to class and variable definitions (coeff: 2.126, corr: 0.114)
- **L2/8432** HTML and JavaScript code related to the Bootstrap framework (coeff: 2.418, corr: 0.140)
- **L3/10179** terms related to health and medical supplements (coeff: 2.383, corr: 0.134)
- **L4/3444** various types of headers, specifically those that denote responses and results within the context of exchanges or interactions (coeff: 2.192, corr: 0.114)
- **L5/697** terms related to price dynamics and economic relationships (coeff: 3.766, corr: 0.088)
- **L6/2491** references to sources or citations in a document (coeff: 2.618, corr: 0.110)
- **L7/6269** references to visual elements such as figures and tables (coeff: 1.293, corr: 0.135)
- **L8/5927** mathematical examples and notations (coeff: 3.347, corr: 0.259)
- **L9/7854** structures related to the declaration and manipulation of result variables in a programming context (coeff: 10.475, corr: 0.189)
- **L10/15705** references to file operations and data management in code (coeff: 6.145, corr: 0.215)

- L11/13926 mathematical expressions and calculations (coeff: 8.203, corr: 0.154)
- L12/1085 references to court cases and legal statutes (coeff: 1.839, corr: 0.220)
- L13/536 technical details related to manufacturing processes (coeff: 4.417, corr: 0.178)
- L14/10612 structured data or code snippets related to databases (coeff: 5.030, corr: 0.225)
- L15/2822 structured data formats or code snippets related to programming (coeff: 1.632, corr: 0.176)
- L16/6602 the presence of specific numerical or coding patterns in data (coeff: 6.773, corr: 0.300)
- L17/5137 mathematical symbols and functions related to field theories (coeff: 8.483, corr: **0.559**)
- L18/3178 code or programming-related elements (coeff: 7.851, corr: 0.507)
- L19/11641 technical components or elements in code (coeff: 16.336, corr: 0.414)
- L20/12748 **structured data representations and their attributes** (coeff: 28.025, corr: 0.394)
- L21/14337 code-related keywords and method definitions in programming contexts (coeff: 20.453, corr: 0.392)
- L22/13921 elements related to database structure and definitions (coeff: 18.510, corr: 0.420)
- L23/12349 technical terms related to software or code management (coeff: 5.893, corr: 0.331)
- L24/16355 definitions and mathematical notation in text (coeff: 39.910, corr: 0.326)
- L25/4307 occurrences of programming syntax related to object-oriented structures (coeff: 19.460, corr: 0.384)

Additional relevant features:

- L18/1127 references to gender and associated options/choices in forms (coeff: 4.813, corr: 0.207)
- L19/15745 phrases related to decision-making and choice, particularly in the context of parenting and social interactions (coeff: 11.875, corr: 0.226)
- L23/12048 terms related to racism and social injustice (coeff: 2.661, corr: 0.147)

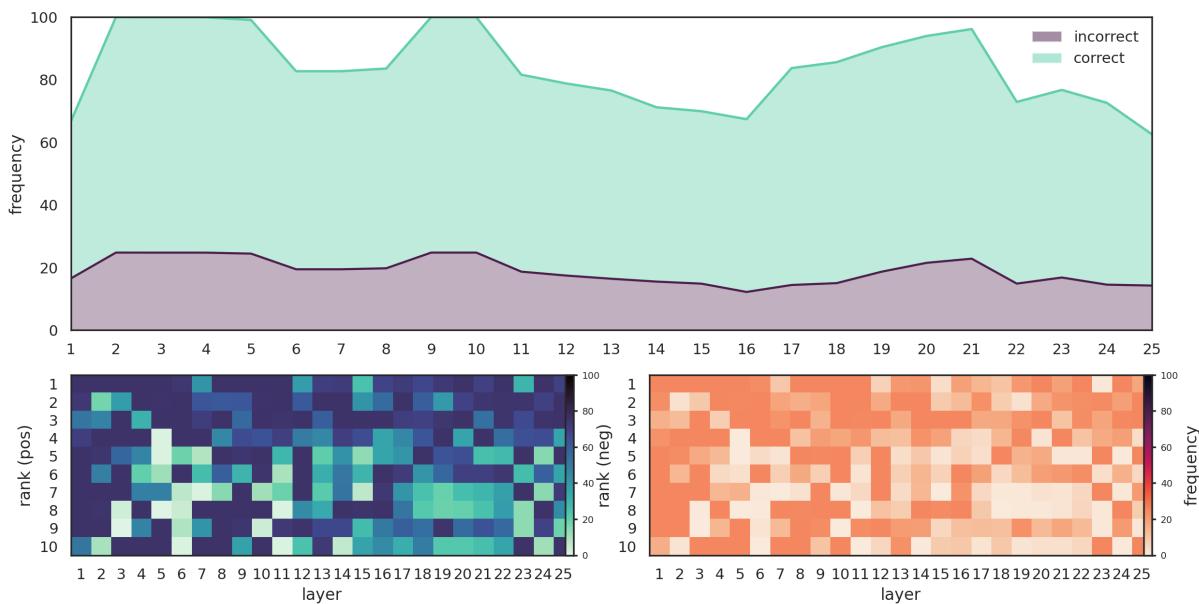


Figure: Top correlated features with BBQ disambig on frequency in each layer of Gemma 2 2B.

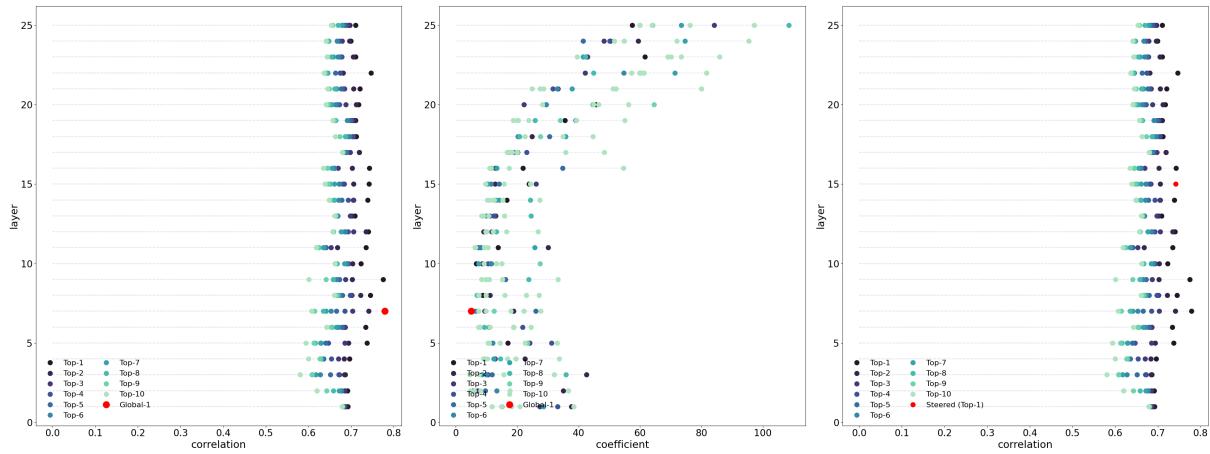


Figure: Top correlated features with selected features from CorrSteer-P with HarmBench on coefficient in each layer of Gemma 2B.

- L1/9572 occurrences of the semicolon character (coeff: 5.206, corr: 0.692)
- L2/6712 references to worship and its related symbols or icons (coeff: 5.699, corr: 0.692)
- L3/16207 syntax elements and formatting in code or mathematical expressions (coeff: 2.583, corr: 0.686)
- L4/3109 forms of the verb "to be" and its variations (coeff: 5.891, corr: 0.696)
- L5/11099 sentences that include personal affirmations or declarations of identity (coeff: 16.934, corr: 0.737)
- L6/12241 instances of the verb "to be" in various forms and their contexts (coeff: 7.338, corr: 0.735)
- L7/11722 **phrases related to legal terms and the rejection of arguments in court cases** (coeff: 5.035, corr: **0.779**)
- L8/8642 expressions of self-identity and subjective experience (coeff: 8.729, corr: 0.745)
- L9/9298 **strongly negative or dismissive opinions about claims and arguments** (coeff: 7.525, corr: 0.775)
- L10/3037 references to legal issues and compliance (coeff: 6.667, corr: 0.723)
- L11/6905 statements of identity and self-description (coeff: 13.810, corr: 0.735)
- L12/12039 phrases related to providing assistance and support (coeff: 5.253, corr: 0.741)
- L13/6715 text that discusses accountability and the need for forgiveness (coeff: 6.992, corr: 0.709)
- L14/2949 statements and phrases related to political criticism and condemnation (coeff: 16.620, corr: 0.739)
- L15/1570 judgments regarding moral and ethical standards related to exploitation and human rights issues (coeff: 23.824, corr: 0.742)
- L16/5113 expressions of personal identity and emotional states (coeff: 21.832, corr: 0.743)
- L17/5887 references to tools and functional capabilities related to programming or software development (coeff: 11.389, corr: 0.720)
- L18/1411 negative statements or denials (coeff: 20.537, corr: 0.712)
- L19/324 phrases related to legal procedures and considerations (coeff: 35.610, corr: 0.710)
- L20/5192 questions that seek clarification or challenge assumptions (coeff: 45.662, corr: 0.718)
- L21/7129 negative sentiments and expressions of doubt or denial (coeff: 33.225, corr: 0.721)
- L22/3311 references to food and culinary experiences (coeff: 19.000, corr: 0.746)
- L23/11246 instances of strong negative sentiment or rejection (coeff: 61.642, corr: 0.711)
- L24/12773 first-person pronouns and references to personal experiences or actions (coeff: 50.332, corr: 0.699)
- L25/3912 **negative sentiments or refusals** (coeff: 57.431, corr: 0.711)

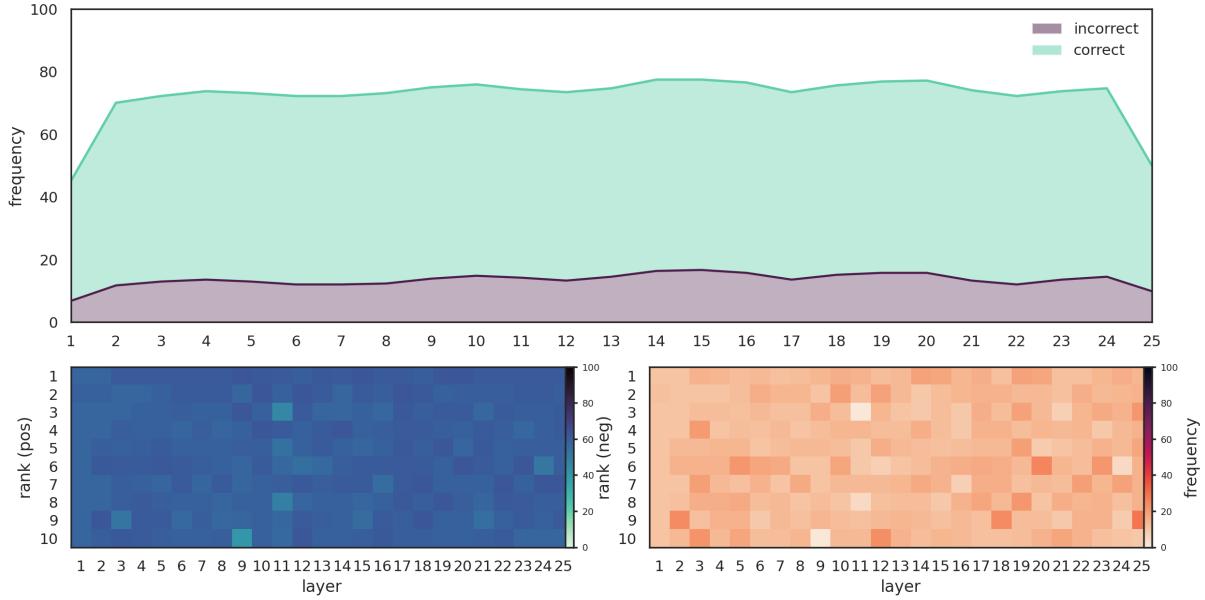


Figure: Top correlated features with HarmBench on frequency in each layer of Gemma 2 2B.

MMLU

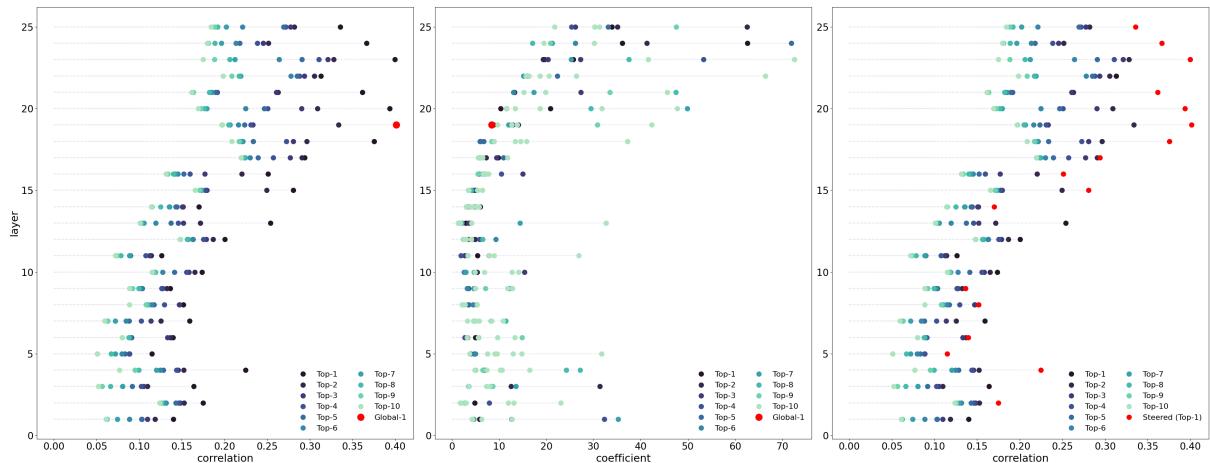


Figure: Top correlated features with selected features from CorrSteer-P with MMLU on coefficient in each layer of Gemma 2 2B.

- L1/13714 colons and semicolons used in lists or programming syntax (coeff: 0.403, corr: 0.140)
- L2/6273 specific medical terminology and its implications (coeff: 1.548, corr: 0.175)
- L3/12378 programming-related elements and commands (coeff: 1.094, corr: 0.164)
- L4/11047 certain types of mathematical or programming syntax (coeff: 2.944, corr: 0.225)
- L5/8581 phrases that indicate research findings or results (coeff: 0.077, corr: 0.115)
- L6/5275 sentences expressing doubt or conditionality in arguments (coeff: 4.939, corr: 0.140)
- L7/14726 periods and other punctuation marks that signify sentence endings or significant separations in text (coeff: 2.532, corr: 0.159)
- L8/15039 terms related to research methodologies and experimental design (coeff: 0.309, corr: 0.152)
- L9/15654 variations of the word "correct" in various contexts (coeff: 0.414, corr: 0.136)
- L10/11729 coding attributes and properties related to light types in a 3D programming context (coeff: 2.919, corr: 0.174)

- L11/13204 code syntax and structure, particularly related to variable assignments and function calls (coeff: 5.369, corr: 0.126)
- L12/6392 XML-like structured data elements (coeff: 1.033, corr: 0.200)
- L13/12281 mathematical expressions and concepts related to positive values (coeff: 0.919, corr: 0.254)
- L14/7 significant scientific findings and their specific details (coeff: 6.002, corr: 0.170)
- L15/8678 phrases related to announcements or updates (coeff: 4.906, corr: 0.281)
- L16/12421 programming constructs and their structures within code snippets (coeff: 5.593, corr: 0.251)
- L17/13214 error messages and diagnostic codes (coeff: 9.790, corr: 0.294)
- L18/1127 references to gender and associated options/choices in forms (coeff: 4.805, corr: 0.376)
- L19/2174 input fields and value assignments in a form-like structure (coeff: 8.405, corr: 0.402)
- L20/12748 **structured data representations and their attributes** (coeff: 20.884, corr: 0.394)
- L21/14337 code-related keywords and method definitions in programming contexts (coeff: 13.228, corr: 0.362)
- L22/5939 technical jargon and terminology related to chemistry and biochemistry (coeff: 5.582, corr: 0.313)
- L23/10424 statistical terms and symbols related to data analysis and significance testing (coeff: 25.724, corr: 0.400)
- L24/16355 definitions and mathematical notation in text (coeff: 36.077, corr: 0.367)
- L25/10388 phrases related to health-related actions and topics (coeff: 33.899, corr: 0.336)

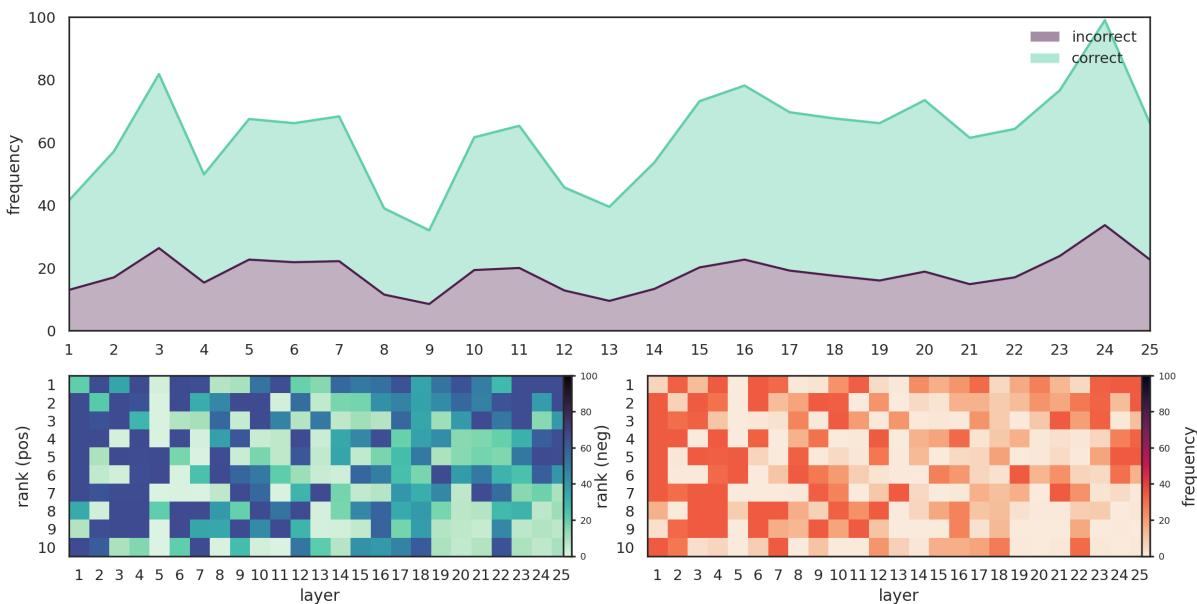


Figure: Top correlated features with MMLU on frequency in each layer of Gemma 2 2B.

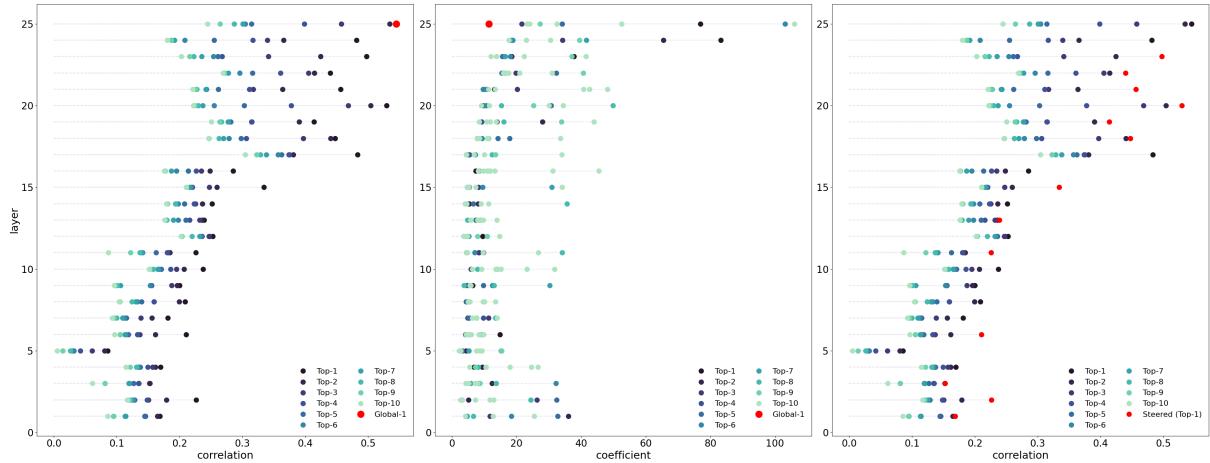


Figure: Top correlated features with selected features from CorrSteer-P with MMLU-Pro on coefficient in each layer of Gemma 2 2B.

- L1/9317 phrases related to changes in social and organizational dynamics (coeff: 1.859, corr: 0.169)
- L2/3714 mathematical notation, specifically related to set notation and expressions involving functions (coeff: 0.761, corr: 0.226)
- L3/11980 statements providing answers or conclusions regarding questions or hypotheses (coeff: 3.699, corr: 0.153)
- L4/15960 terms related to medical procedures and conditions (coeff: 6.817, corr: 0.170)
- L5/7502 expressions of honesty and self-awareness in discourse (coeff: 2.187, corr: 0.086)
- L6/6201 numeric representations of system specifications or configurations (coeff: 14.877, corr: 0.210)
- L7/8790 structured data formats and their attributes (coeff: 1.209, corr: 0.182)
- L8/11297 structured data and programming constructs (coeff: 2.176, corr: 0.209)
- L9/15336 references to mathematical or computational problems and their solutions (coeff: 6.407, corr: 0.200)
- L10/10805 terms related to medical conditions and biological factors (coeff: 1.277, corr: 0.237)
- L11/1909 affirmative or negative responses in the context of questions (coeff: 2.296, corr: 0.226)
- L12/14752 legal and governmental terms related to authority and judgment (coeff: 1.369, corr: 0.253)
- L13/12991 mathematical operations and expressions (coeff: 2.560, corr: 0.239)
- L14/10780 comments and documentation markers in code (coeff: 1.455, corr: 0.252)
- L15/2262 references to variable declarations and data structures in programming contexts (coeff: 1.183, corr: 0.334)
- L16/3142 mathematical symbols and notation used in equations (coeff: 5.691, corr: 0.285)
- L17/1175 mathematical expressions and applications related to programming or data structures (coeff: 3.091, corr: 0.483)
- L18/682 function declarations and their return types in a programming context (coeff: 3.406, corr: 0.448)
- L19/11641 technical components or elements in code (coeff: 2.144, corr: 0.414)
- L20/12748 **structured data representations and their attributes** (coeff: 7.134, corr: 0.529)
- L21/1944 code structures and syntax related to programming and mathematics (coeff: 9.251, corr: 0.456)
- L22/12947 scientific terminology related to healthcare and medical research (coeff: 11.241, corr: 0.440)
- L23/5752 associations and relationships among scientific variables and observations (coeff: 10.133, corr: 0.497)
- L24/8188 syntax related to code structure and operations (coeff: 11.861, corr: 0.482)

- L25/8643 scientific terms and concepts related to biochemistry and cellular processes (coeff: 11.439, corr: **0.545**)

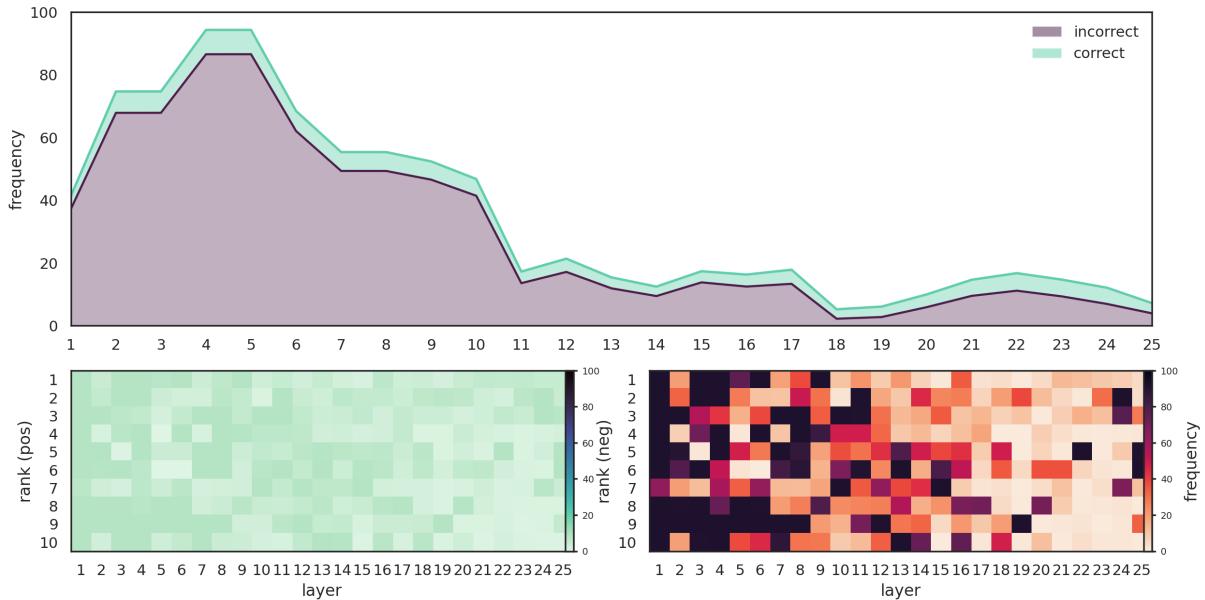


Figure: Top correlated features with MMLU-Pro on frequency in each layer of Gemma 2 2B.

GSM8K

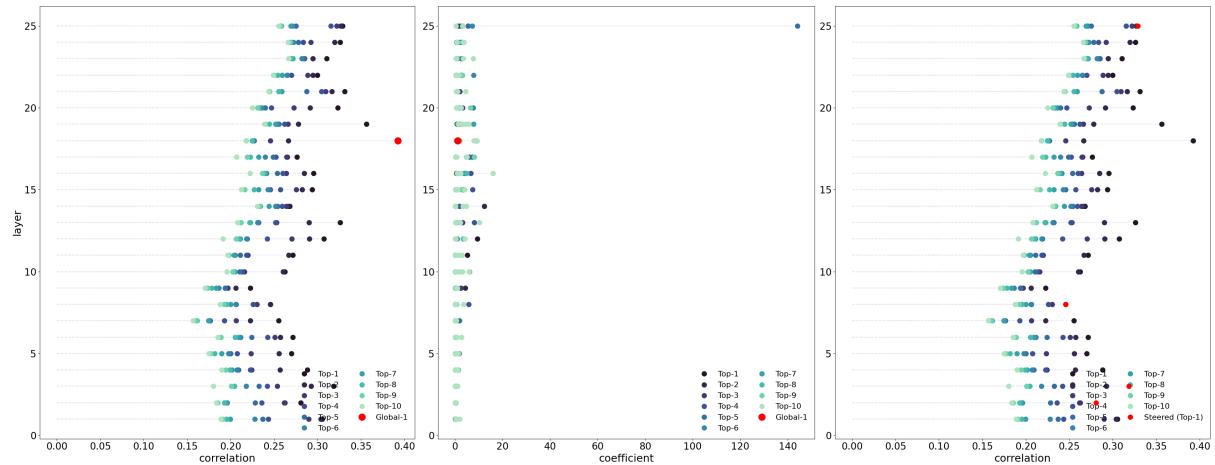


Figure: Top correlated features with selected features from CorrSteer-P with GSM8K on coefficient in each layer of Gemma 2 2B.

- L1/13475 specific quantitative or statistical information (coeff: 9.936, corr: 0.251)
- L2/2098 references to leadership and management isolation in workplace contexts (coeff: 3.080, corr: 0.180)
- L3/8338 significant quantities within code snippets, likely indicating important operations or constructs (coeff: 6.302, corr: 0.250)
- L4/687 HTML tags and attributes related to layout and styling (coeff: 2.037, corr: 0.188)
- L5/697 terms related to price dynamics and economic relationships (coeff: 6.091, corr: 0.193)
- L6/13460 references to safety and regulatory issues in automobile contexts (coeff: 9.501, corr: 0.219)
- L7/9514 structured data or code snippets, potentially relating to geographical regions and associated identifiers (coeff: 1.309, corr: 0.167)

- L8/2024 names of notable performance venues and cultural institutions (coeff: 14.384, corr: 0.210)
- L9/15115 discussions related to crime scene investigations and forensic evidence (coeff: 5.074, corr: 0.188)
- L10/2794 elements of conversation or dialogue (coeff: 5.602, corr: 0.188)
- L11/7313 mathematical equations and expressions (coeff: 26.252, corr: 0.176)
- L12/12707 technical or scientific terminology related to systems and processes (coeff: 2.860, corr: 0.245)
- L13/14319 code snippets and their associated structures within documents (coeff: 2.731, corr: 0.253)
- L14/4217 expressions of emotional reactions and feedback (coeff: 3.772, corr: 0.246)
- L15/1685 instances of structured data or messages indicating communication or queries (coeff: 7.282, corr: 0.255)
- L16/14919 instances of unique identifiers or markers in a dataset (coeff: 24.774, corr: 0.223)
- L17/7185 curly braces and structured programming syntax elements (coeff: 6.245, corr: 0.252)
- L18/3732 code syntax elements such as brackets and semicolons (coeff: 4.064, corr: 0.249)
- L19/2015 structures related to function definitions and method calls in programming code (coeff: 8.802, corr: 0.277)
- L20/15616 elements of code structure and syntax in programming contexts (coeff: 4.350, corr: 0.258)
- L21/12547 phrases and words that express confusion or dissatisfaction with situations (coeff: 24.211, corr: 0.251)
- L22/7903 **mathematical notation and symbols used in equations** (coeff: 7.295, corr: 0.313)
- L23/12425 **mathematical expressions and symbols** (coeff: 19.202, corr: 0.294)
- L24/2274 **programming syntax and structure specific to coding languages** (coeff: 10.205, corr: 0.348)
- L25/3469 technical aspects related to semiconductor devices and their manufacturing processes (coeff: 23.158, corr: 0.284)

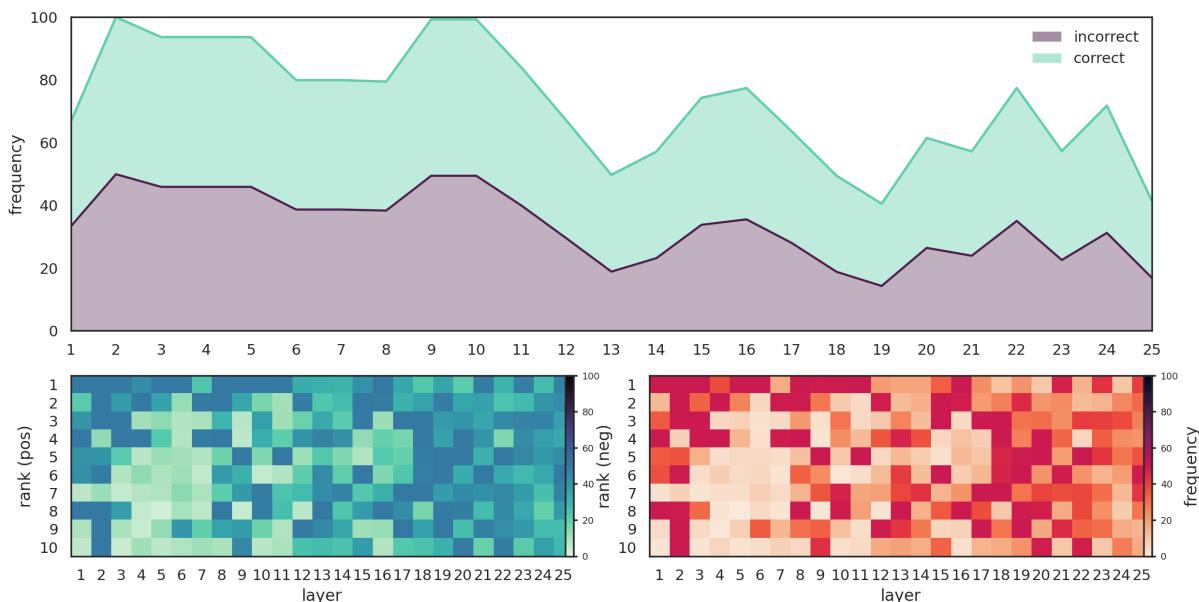


Figure: Top correlated features with GSM8K on frequency in each layer of Gemma 2 2B.

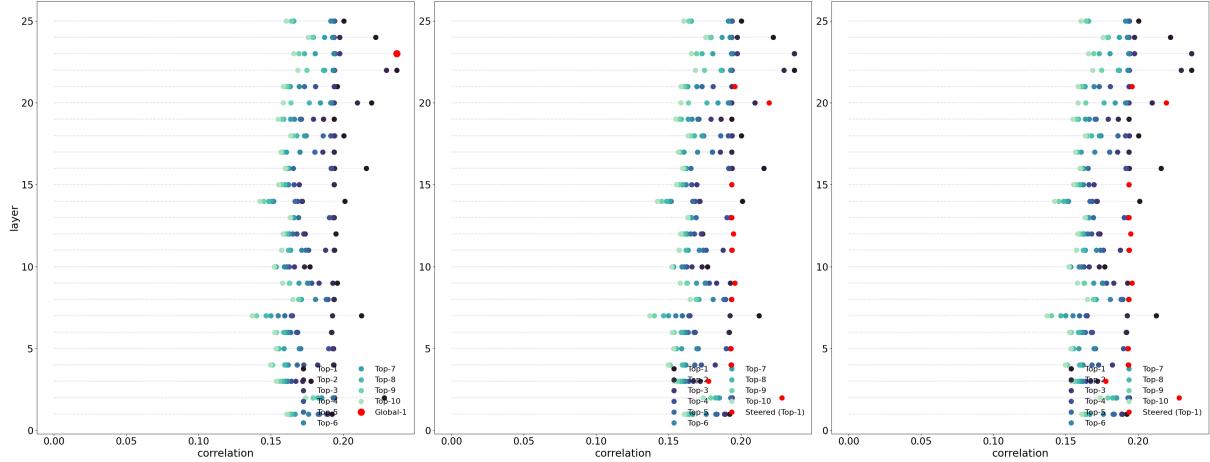


Figure: Top correlated features with selected features from CorrSteer-P with SimpleQA on coefficient in each layer of Gemma 2B.

- L1/14904 references to Congress and legislative processes (coeff: 0.263, corr: 0.192)
- L2/1089 terms and concepts related to integrals and the importance of integration in various contexts (coeff: 0.225, corr: 0.228)
- L3/12843 terms related to durability and long-lasting qualities (coeff: 0.219, corr: 0.178)
- L4/680 references to standards, particularly in legal and medical contexts (coeff: 0.055, corr: 0.194)
- L5/4460 references to legal cases and court rulings (coeff: 0.093, corr: 0.193)
- L6/9777 expressions of agreement or dissent and the context surrounding them (coeff: 0.153, corr: 0.192)
- L7/2431 phrases related to time management and constraints (coeff: 0.253, corr: 0.213)
- L8/14209 HTML coding elements and formatting commands (coeff: 0.097, corr: 0.194)
- L9/7856 terms related to penalties and scoring in sporting events (coeff: 0.257, corr: 0.196)
- L10/2446 terms related to health and legal matters (coeff: 0.350, corr: 0.177)
- L11/7954 legal terminology and references to court cases and proceedings (coeff: 0.164, corr: 0.194)
- L12/1495 phrases related to the duration and continuity of experiences over time (coeff: 0.208, corr: 0.195)
- L13/12119 references to various parameters and aspects within scientific or technical contexts (coeff: 0.148, corr: 0.194)
- L14/6355 references to India and its cultural context (coeff: 0.377, corr: 0.201)
- L15/4385 programming constructs related to class and method declarations in Java (coeff: 0.177, corr: 0.194)
- L16/7182 expressions of enthusiasm or amazement (coeff: 0.457, corr: 0.216)
- L17/6346 references to offices or organizational structures (coeff: 0.274, corr: 0.194)
- L18/5258 terms related to legal charges and prosecutions (coeff: 0.843, corr: 0.200)
- L19/4202 technical terms and concepts related to physical phenomena and their mathematical descriptions (coeff: 0.262, corr: 0.194)
- L20/6557 legal terms and phrases related to criminal charges and legal proceedings (coeff: 0.953, corr: 0.220)
- L21/13830 references to political leaders and government roles (coeff: 8.090, corr: 0.196)
- L22/15897 phrases related to international relations and cooperation, particularly in the context of political statements and actions (coeff: 0.648, corr: 0.237)
- L23/15190 terms related to health, well-being, and interventions for obesity and mental illness (coeff: 0.832, corr: **0.237**)
- L24/15228 references to political figures and their actions (coeff: 2.043, corr: 0.222)
- L25/2531 expressions of political opinion regarding government spending and fiscal policies (coeff: 1.664, corr: 0.200)

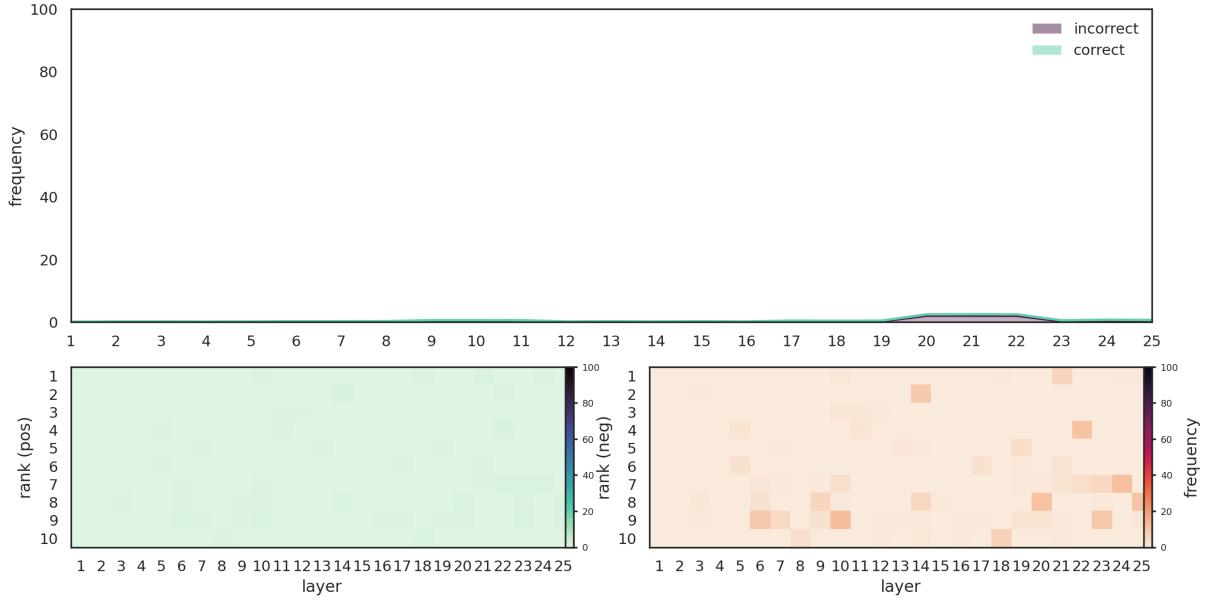


Figure: Top correlated features with SimpleQA on frequency in each layer of Gemma 2 2B.

XSTest

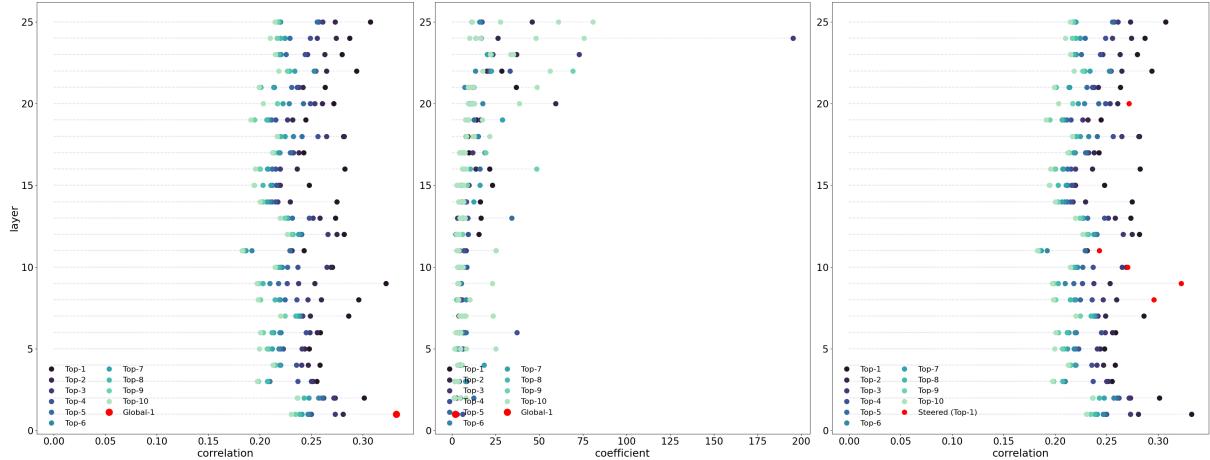


Figure: Top correlated features with selected features from CorrSteer-P with XSTest on coefficient in each layer of Gemma 2 2B.

- L1/4509 terms and concepts related to scientific and mathematical structures and functions (coeff: 1.940, corr: **0.333**)
- L2/4679 financial metrics and forecasts related to stock performance (coeff: 1.584, corr: 0.301)
- L3/1326 legal terminology and references to statutes and claims (coeff: 4.088, corr: 0.256)
- L4/12152 references to geographical locations and their associated attributes (coeff: 0.961, corr: 0.259)
- L5/5939 terms related to signals and their coding in biological contexts (coeff: 3.391, corr: 0.248)
- L6/4376 numerical values and specific formatting related to data structures or coding (coeff: 6.713, corr: 0.259)
- L7/4886 representations of numerical data, particularly in scientific contexts (coeff: 3.074, corr: 0.286)
- L8/10825 punctuation marks and special characters (coeff: 5.194, corr: 0.296)
- L9/9228 punctuation marks, especially periods and quotation marks (coeff: 4.712, corr: 0.323)
- L10/13244 information related to military casualties and incidents (coeff: 2.760, corr: 0.270)

- L11/5734 sections or punctuation that denote lists or explanations (coeff: 4.304, corr: 0.243)
- L12/12342 symbols and mathematical notation related to expressions or equations in mathematical contexts (coeff: 15.373, corr: 0.282)
- L13/10964 mathematical terms and symbols (coeff: 16.622, corr: 0.274)
- L14/7655 structured data, such as XML or JSON formats (coeff: 16.195, corr: 0.275)
- L15/5114 terms related to evaluation and validation processes (coeff: 23.117, corr: 0.248)
- L16/1547 code or programming-related syntax (coeff: 21.527, corr: 0.283)
- L17/10813 references to movies, actors, and significant film industry terms (coeff: 9.662, corr: 0.243)
- L18/8615 legal terminology and concepts related to judicial authority and precedent (coeff: 9.006, corr: 0.282)
- L19/2998 elements related to research findings, including factors, conclusions, and reasoning (coeff: 13.956, corr: 0.245)
- L20/9419 names of individuals and titles (coeff: 10.648, corr: 0.272)
- L21/15170 isolated segments of code or technical content (coeff: 36.804, corr: 0.264)
- L22/11042 punctuation marks that indicate the start or end of lists or key points in a text (coeff: 28.482, corr: 0.294)
- L23/8993 structured API documentation elements and syntax (coeff: 23.447, corr: 0.280)
- L24/4448 terms related to scientific analysis and results reporting (coeff: 16.649, corr: 0.287)
- L25/7968 elements related to health assessments and metrics (coeff: 9.863, corr: 0.307)

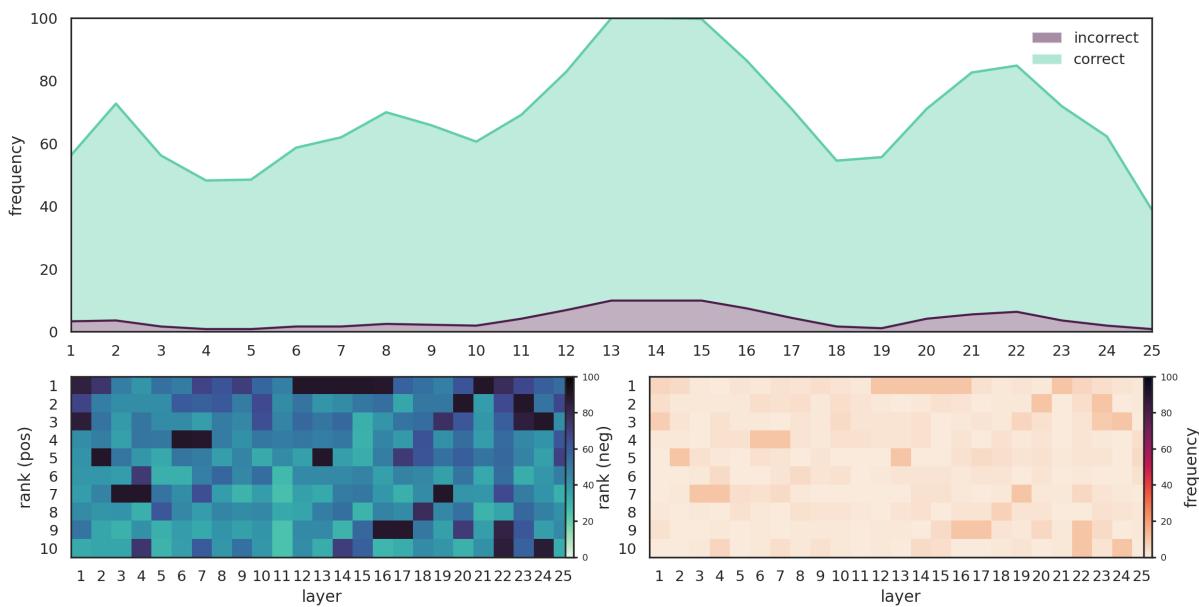


Figure: Top correlated features with XSTest on frequency in each layer of Gemma 2 2B.

A.7.2 Llama-3.1-8B

BBQ (Ambiguous)

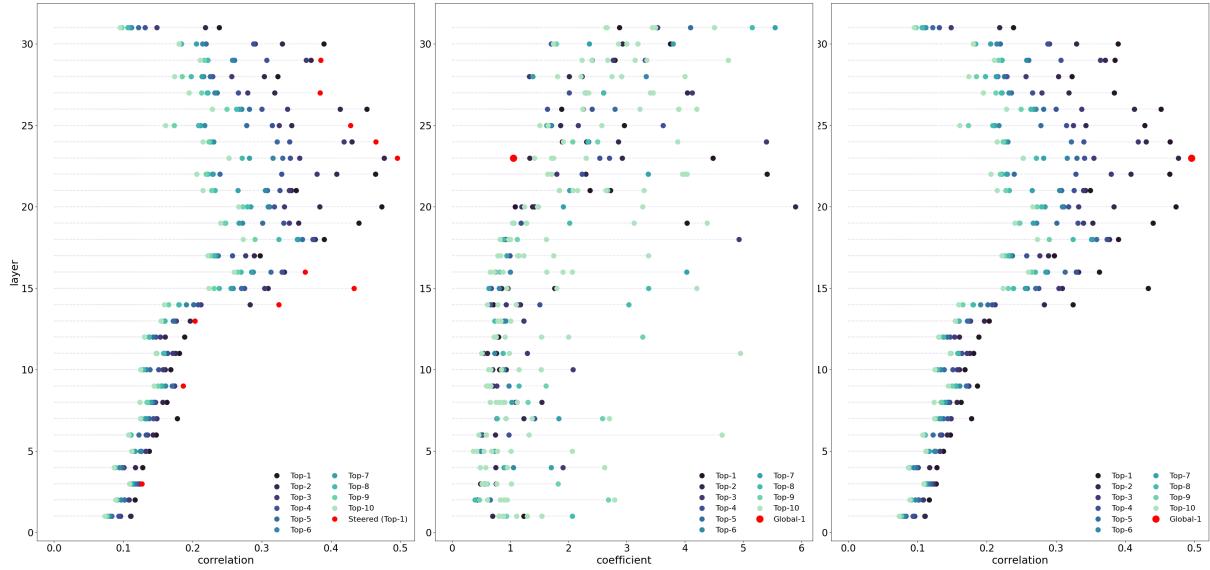


Figure: Top correlated features with selected features from CorrSteer-P with BBQ ambig on coefficient in each layer of Llama 3.1 8B.

- **L1/23207** phrases related to legal or regulatory frameworks (coeff: 0.463, corr: 0.111)
- **L2/2680** titles and key information related to television series episodes (coeff: 0.002, corr: 0.117)
- **L3/23846** discussions around societal structures and issues related to mental health and crime (coeff: 0.487, corr: 0.127)
- **L4/30896** occurrences of numerical values and references to measurements (coeff: 0.089, corr: 0.128)
- **L5/18555** instances of past and present tense verbs, particularly focusing on actions and conditions (coeff: 0.193, corr: 0.137)
- **L6/25246** technical terms and code snippets related to software development and programming logic (coeff: 0.277, corr: 0.147)
- **L7/11878** specific numerical identifiers and related metadata in technical documents (coeff: 0.365, corr: 0.178)
- **L8/4790** keywords related to data structures and programming concepts (coeff: 0.172, corr: 0.163)
- **L9/2700** references to extraterrestrial or paranormal beings and phenomena (coeff: 0.354, corr: 0.187)
- **L10/23355** **phrases or constructs that emphasize comparison or simile** (coeff: 0.812, corr: 0.168)
- **L11/18132** references to specific books, movies, or artworks (coeff: 0.167, corr: 0.181)
- **L12/14096** references to specific locations or settings in various contexts (coeff: 0.084, corr: 0.189)
- **L13/26526** references to error handling in programming (coeff: 0.493, corr: 0.203)
- **L14/13393** statistical percentages and survey data (coeff: 0.192, corr: 0.324)
- **L15/25166** **themes of neutrality and balance in discourse** (coeff: 0.259, corr: 0.433)
- **L16/21816** phrases related to financial or economic assessments (coeff: 0.543, corr: 0.363)
- **L17/5782** references to equality and equity in rights and opportunities (coeff: 0.368, corr: 0.298)
- **L18/28196** references to knowledge, learning, and understanding in various contexts (coeff: 0.303, corr: 0.390)
- **L19/29460** **discussions about extremes and balance** (coeff: 0.811, corr: 0.440)
- **L20/13319** **expressions of mixed opinions or complex character evaluations** (coeff: 1.413, corr: 0.473)
- **L21/8518** references to articles and citations in academic databases (coeff: 2.719, corr: 0.349)

- **L22/28263 percentages and statistical data concerning opinions or responses** (coeff: 1.024, corr: 0.464)
- **L23/638 formal structures and procedures within organizational contexts** (coeff: 1.054, corr: **0.496**)
- **L24/19174 code constructs and control flow keywords related to conditions and returns** (coeff: 1.890, corr: 0.465)
- **L25/10753 expressions of perception or belief in social dynamics** (coeff: 1.147, corr: 0.428)
- **L26/27899 code structure and logical operations involving object hierarchy and data types** (coeff: 1.025, corr: 0.452)
- **L27/1765 quantitative data related to project development and financial metrics** (coeff: 2.597, corr: 0.384)
- **L28/21019 financial data and statistics related to development projects** (coeff: 0.856, corr: 0.323)
- **L29/17998 code snippets related to JavaScript or Java programming functions and structures** (coeff: 1.735, corr: 0.385)
- **L30/17084 numerical data related to financial projections and resource development** (coeff: 1.308, corr: 0.390)
- **L31/10728 auxiliary verbs and words indicating obligation or possibility** (coeff: 1.530, corr: 0.239)

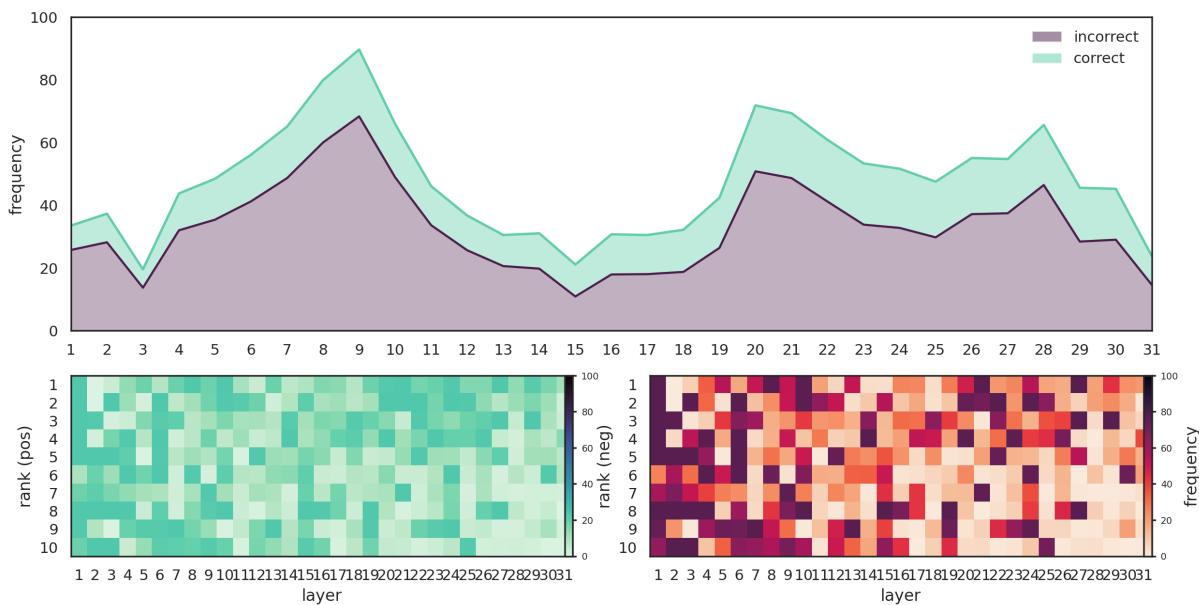


Figure: Top correlated features with BBQ ambig on frequency in each layer of Llama 3.1 8B.

BBQ (Disambiguous)

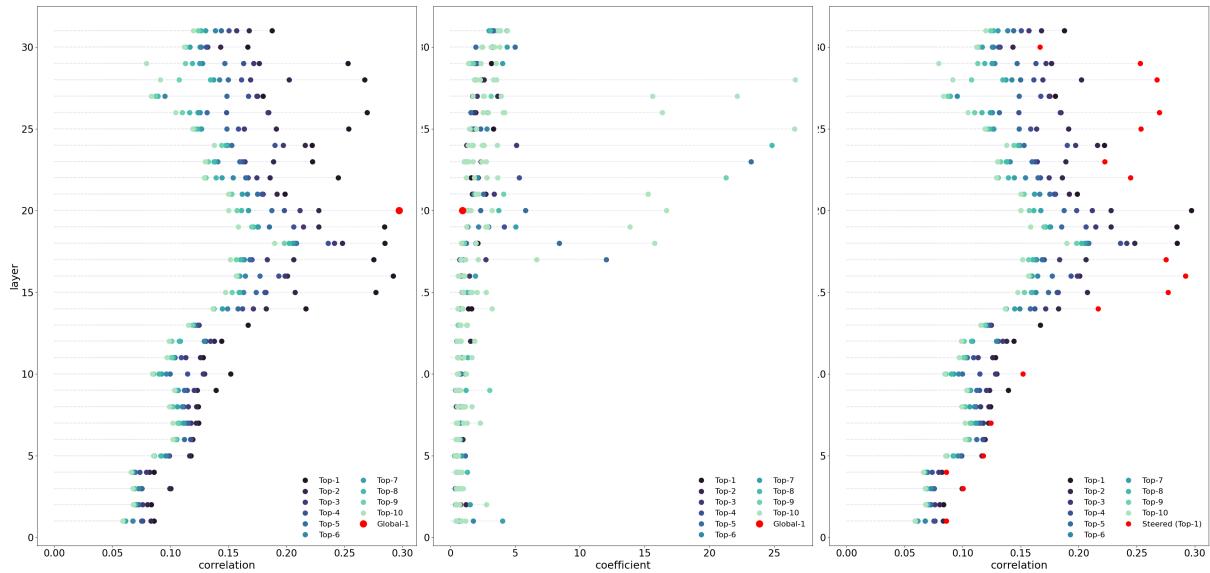


Figure: Top correlated features with selected features from CorrSteer-P with BBQ disambig on coefficient in each layer of Llama 3.1 8B.

- **L1/5891** technical terms and references in programming and development contexts (coeff: 0.154, corr: 0.086)
- **L2/21865** references to essays, articles, and related writing concepts (coeff: 0.784, corr: 0.084)
- **L3/3413** elements related to user engagement and user-friendly design (coeff: 0.332, corr: 0.100)
- **L4/3712** elements related to programming and computation (coeff: 0.458, corr: 0.086)
- **L5/18066** references to educational administration and school district issues (coeff: 0.229, corr: 0.118)
- **L6/28294** references to machine learning models and recommendation systems (coeff: 0.301, corr: 0.119)
- **L7/7762** specific language constructs related to coordination and organization (coeff: 0.416, corr: 0.124)
- **L8/25466** terms related to hierarchical structures or classifications (coeff: 1.032, corr: 0.124)
- **L9/5313** key concepts related to project management and planning (coeff: 0.645, corr: 0.139)
- **L10/13407 negative actions and attitudes that hinder interpersonal relationships and community engagement** (coeff: 0.256, corr: 0.152)
- **L11/18350** references to institutions and systems regarding public services (coeff: 0.900, corr: 0.128)
- **L12/13336 phrases and concepts related to community and social interactions** (coeff: 0.377, corr: 0.144)
- **L13/15793 negation phrases and words indicating absence or lack** (coeff: 0.695, corr: 0.167)
- **L14/31962 details related to physical displacement or movement in a spatial context** (coeff: 1.384, corr: 0.217)
- **L15/2128** references to programming elements and constructs (coeff: 0.977, corr: 0.277)
- **L16/6219 code-related syntax and structures within programming languages** (coeff: 0.830, corr: **0.292**)
- **L17/12610 technical terminology related to programming and software development** (coeff: 0.706, corr: 0.275)
- **L18/16458 HTML tags and structured data elements** (coeff: 2.113, corr: 0.285)
- **L19/6432 numerical values and the structure of dates or game scores** (coeff: 0.909, corr: 0.284)
- **L20/28406 tokens related to timestamps, specifically date and time formats** (coeff: 0.942, corr: 0.297)
- **L21/15538 references to time management techniques and motivational strategies** (coeff: 0.388, corr:

0.199)

- **L22/11286** monetary amounts or financial figures (coeff: 0.531, corr: 0.245)
- **L23/30672** phrases involving the concept of answers or responses (coeff: 1.211, corr: 0.222)
- **L24/5888** references to answers or responses in discussions or questions (coeff: 1.152, corr: 0.222)
- **L25/22713** mathematical notations and symbols (coeff: 1.235, corr: 0.253)
- **L26/22133** names of authors and their affiliations in academic contexts (coeff: 1.953, corr: 0.269)
- **L27/12321** structural elements and parameters in programming code or data structures (coeff: 0.539, corr: 0.180)
- **L28/23202 specific numbers and their context within factual statements** (coeff: 1.897, corr: 0.267)
- **L29/3168** keywords related to health and medical terminology (coeff: 3.175, corr: 0.253)
- **L30/22450** terms and phrases related to health and medical conditions (coeff: 3.219, corr: 0.167)
- **L31/18173** procedural commands and technical instructions related to software and settings (coeff: 1.440, corr: 0.188)

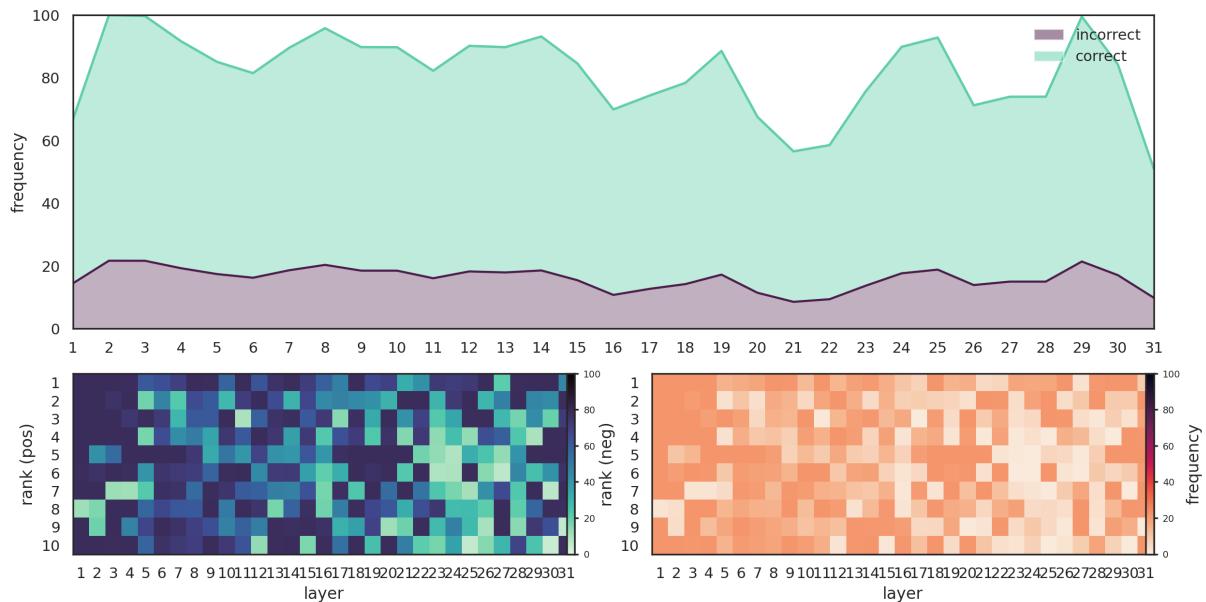


Figure: Top correlated features with BBQ disambig on frequency in each layer of Llama 3.1 8B.

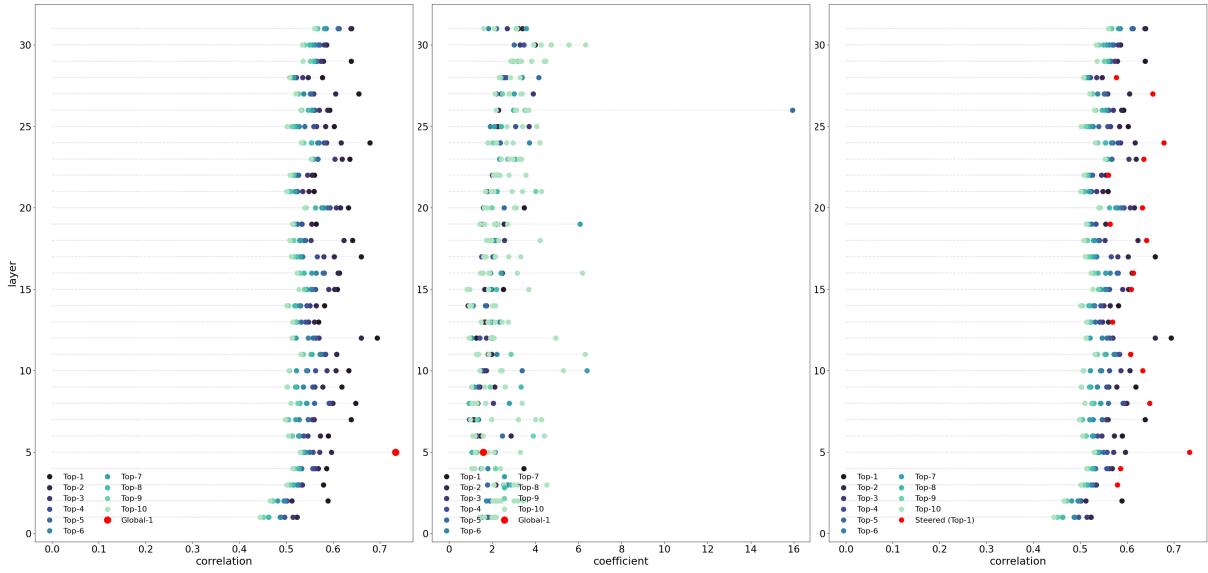


Figure: Top correlated features with selected features from CorrSteer-P with HarmBench on coefficient in each layer of Llama 3.1 8B.

- **L1/15747** repetitive phrases or expressions related to certainty or emphasis (coeff: 0.491, corr: 0.524)
- **L2/25715** references to collective experiences and communal responsibility (coeff: 1.032, corr: 0.590)
- **L3/23621** negations and assertions related to existence and actions (coeff: 1.116, corr: 0.580)
- **L4/26750** first-person pronouns indicating personal experiences and thoughts (coeff: 3.468, corr: 0.586)
- **L5/300** instances of political criticism and hypocrisy (coeff: 1.587, corr: **0.734**)
- **L6/21616** discussions about legality, morality, and the implications of actions in ethical contexts (coeff: 1.458, corr: 0.590)
- **L7/17622** phrases related to trust and loyalty in political contexts (coeff: 1.128, corr: 0.639)
- **L8/6508** expressions related to the condemnation of sexual assault and violence (coeff: 1.322, corr: 0.648)
- **L9/27026** concepts related to limits and responsibilities in relationships and societal interactions (coeff: 1.425, corr: 0.619)
- **L10/9364** expressions of moral outrage and condemnation regarding social and ethical issues (coeff: 1.324, corr: 0.633)
- **L11/16561** expressions of personal opinion and moral judgments (coeff: 1.810, corr: 0.608)
- **L12/5839** strong statements against violence and discrimination (coeff: 1.271, corr: 0.694)
- **L13/15443** emotional expressions of affection or attachment (coeff: 1.637, corr: 0.569)
- **L14/22046** phrases and sentiments associated with moral judgments and emotional responses (coeff: 0.750, corr: 0.582)
- **L15/5498** phrases related to environmental and climate impact (coeff: 0.696, corr: 0.609)
- **L16/8375** topics related to stigma and mental health awareness (coeff: 0.938, corr: 0.614)
- **L17/15876** expressions of self-doubt or uncertainty (coeff: 0.582, corr: 0.660)
- **L18/6210** phrases related to educational support and challenges faced by teachers (coeff: 0.964, corr: 0.641)
- **L19/5854** references to seeking medical advice and guidance (coeff: 1.148, corr: 0.564)
- **L20/11388** elements related to moral and ethical dilemmas (coeff: 3.490, corr: 0.633)
- **L21/9674** references to racism and social justice issues (coeff: 0.712, corr: 0.559)
- **L22/4650** expressions of self-awareness and personal growth mixed with skepticism towards collective beliefs (coeff: 2.235, corr: 0.560)
- **L23/28291** phrases discussing social justice and advocacy for marginalized communities (coeff:

2.165, corr: 0.636)

- **L24/21055** phrases related to self-identity and personal reflection (coeff: 2.357, corr: 0.679)
- **L25/16450 themes of emotional struggle and interpersonal relationships** (coeff: 2.415, corr: 0.602)
- **L26/6648 phrases indicating moral judgment or hypocrisy in political discourse** (coeff: 1.541, corr: 0.593)
- **L27/10654 expressions of emotional conflict and personal reflection** (coeff: 1.653, corr: 0.655)
- **L28/522 themes of courage and resilience in writing** (coeff: 0.915, corr: 0.578)
- **L29/13883 complex emotional responses and reflections on interpersonal relationships** (coeff: 2.977, corr: 0.639)
- **L30/4588 expressions of emotional needs and desires in relationships** (coeff: 1.480, corr: 0.586)
- **L31/31181 references to familial relationships and memorial details** (coeff: 1.218, corr: 0.639)

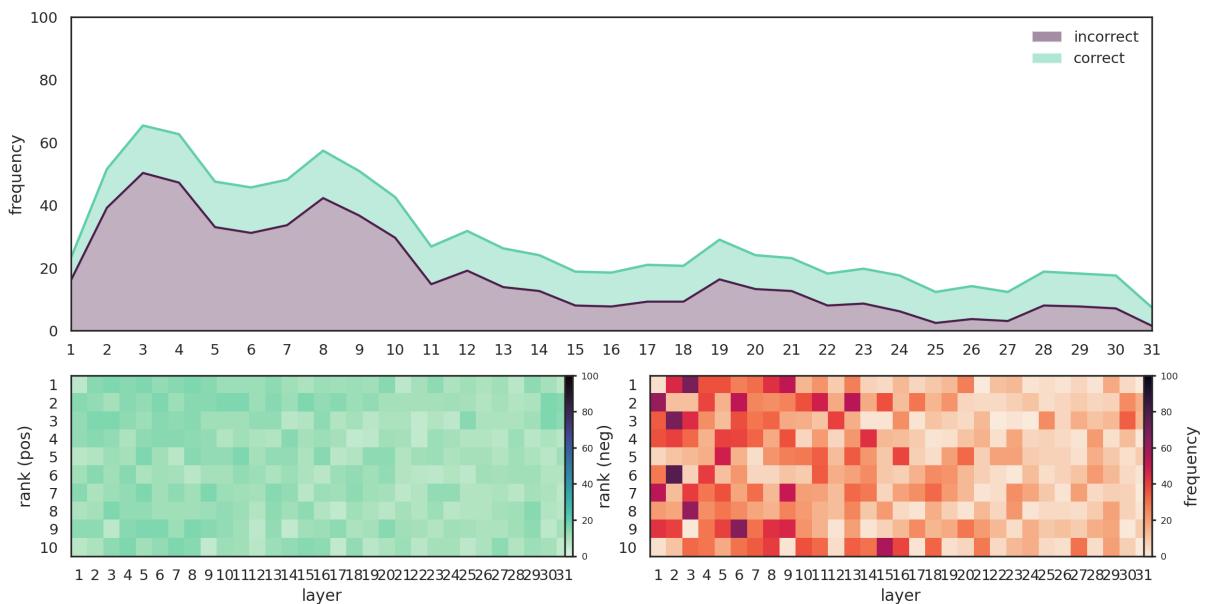


Figure: Top correlated features with HarmBench on frequency in each layer of Llama 3.1 8B.

MMLU

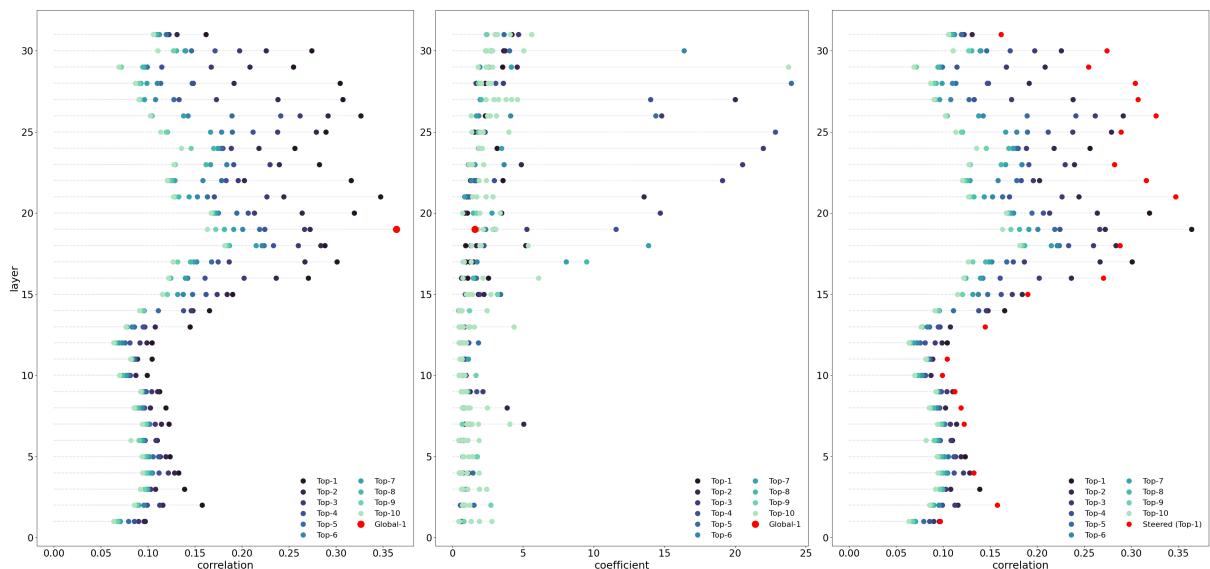


Figure: Top correlated features with selected features from CorrSteer-P with MMLU ambig on coefficient in each layer of Llama 3.1 8B.

- **L1/4557** specific numeric values and measurements related to instructions or guidelines (coeff: 0.695, corr: 0.094)
- **L2/27893** terms related to technology, specifically graphics processing units (GPUs) and their applications (coeff: 0.348, corr: 0.157)
- **L3/204 terms and concepts related to financial metrics and performance evaluation** (coeff: 1.037, corr: 0.139)
- **L4/23545** questions that lead to detailed inquiries or clarifications (coeff: 1.142, corr: 0.131)
- **L5/17458 terms related to theoretical concepts and methodologies in scientific discussions** (coeff: 0.497, corr: 0.124)
- **L6/650** specific identifiers, particularly those related to content or lists (coeff: 0.780, corr: 0.110)
- **L7/13659** references to lists, particularly those pertaining to security or classification contexts (coeff: 0.885, corr: 0.118)
- **L8/1649** key terms related to organizational assistance and functionality within various contexts (coeff: 0.871, corr: 0.116)
- **L9/19730** various forms of interviews and discussions related to current events or cultural topics (coeff: 0.397, corr: 0.108)
- **L10/20495** terms related to requirements and definitions within various contexts (coeff: 0.949, corr: 0.099)
- **L11/20851** legal and academic terminology related to charges and reports (coeff: 0.897, corr: 0.100)
- **L12/26346** specific nouns and proper names related to various contexts (coeff: 0.454, corr: 0.104)
- **L13/551** terms related to medical results and actions taken toward health management (coeff: 0.830, corr: 0.143)
- **L14/11013** phrases indicating relationships between people or entities (coeff: 0.366, corr: 0.165)
- **L15/9446** expressions of passion and enthusiasm in various contexts (coeff: 0.327, corr: 0.195)
- **L16/6219** code-related syntax and structures within programming languages (coeff: 1.094, corr: 0.274)
- **L17/26604** references to programming concepts and structures (coeff: 0.957, corr: 0.301)
- **L18/28750** structured data elements and patterns, possibly related to programming or data analysis (coeff: 0.936, corr: 0.288)
- **L19/6432** numerical values and the structure of dates or game scores (coeff: 1.587, corr: 0.365)
- **L20/28406** tokens related to timestamps, specifically date and time formats (coeff: 1.051, corr: 0.319)
- **L21/15538** references to time management techniques and motivational strategies (coeff: 1.014, corr: 0.347)
- **L22/11286** monetary amounts or financial figures (coeff: 1.269, corr: 0.322)
- **L23/15096** phrases related to significant life events and milestones (coeff: 1.125, corr: 0.281)
- **L24/18010** references to dates and significant life events (coeff: 1.631, corr: 0.256)
- **L25/22713** mathematical notations and symbols (coeff: 1.209, corr: 0.287)
- **L26/22133** names of authors and their affiliations in academic contexts (coeff: 2.331, corr: 0.331)
- **L27/19268** references to academic qualifications, research, and involvement in educational activities (coeff: 0.826, corr: 0.310)
- **L28/23202 specific numbers and their context within factual statements** (coeff: 2.318, corr: 0.307)
- **L29/3168** keywords related to health and medical terminology (coeff: 3.545, corr: 0.255)
- **L30/23403** terms associated with uncertainty and error (coeff: 0.986, corr: 0.274)
- **L31/6722** instances of code-related syntax and formatting (coeff: 0.538, corr: 0.159)

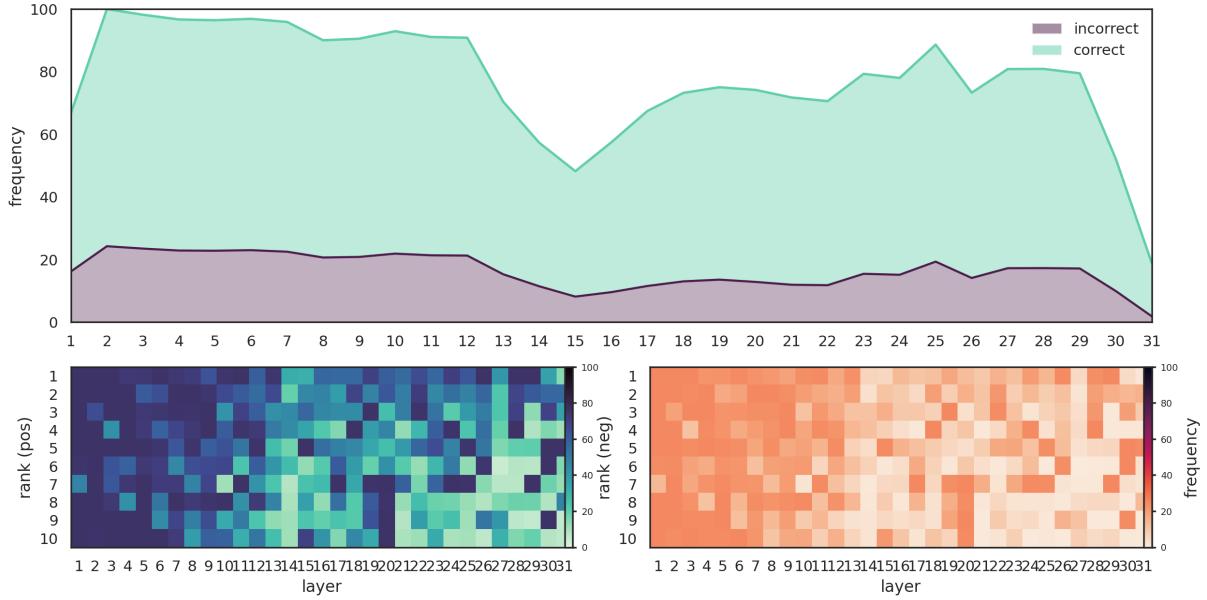


Figure: Top correlated features with MMLU on frequency in each layer of Llama 3.1 8B.

MMLU-Pro

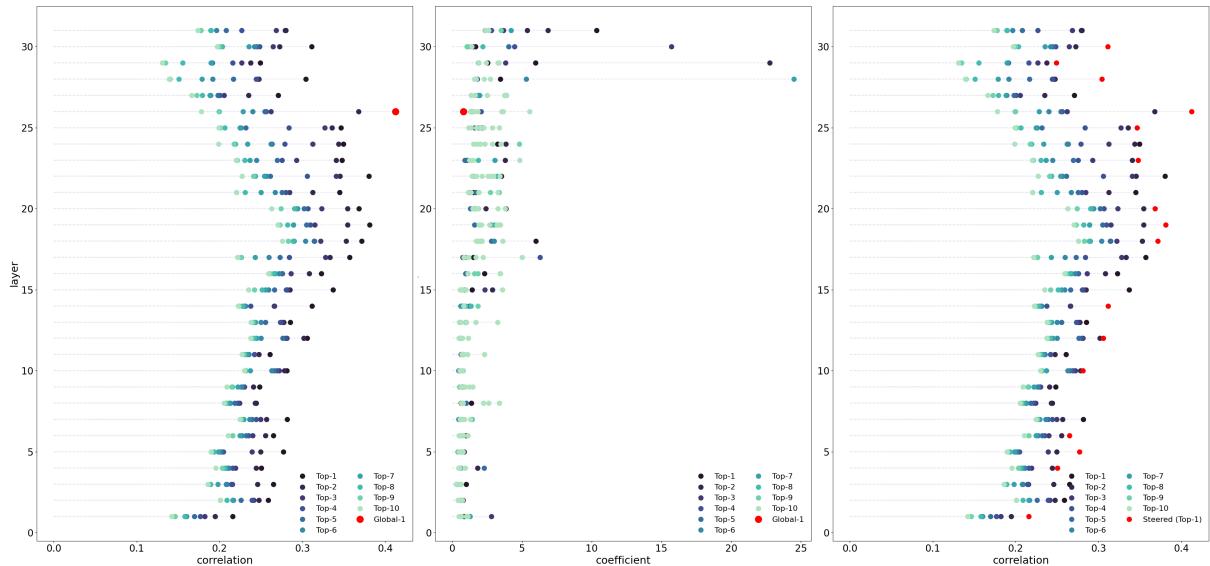


Figure: Top correlated features with selected features from CorrSteer-P with MMLU-Pro ambig on coefficient in each layer of Llama 3.1 8B.

- L1/2403 specific numeric values and measurements related to instructions or guidelines (coeff: 0.286, corr: 0.216)
- L2/85 phrases related to service expectations and quality assurance (coeff: 0.212, corr: 0.259)
- L3/204 terms and concepts related to financial metrics and performance evaluation (coeff: 0.996, corr: 0.265)
- L4/14539 content related to sources and references in articles (coeff: 0.432, corr: 0.250)
- L5/2831 references to urgency and scheduling events (coeff: 0.348, corr: 0.277)
- L6/7784 instances of various relational and transactional terms within context (coeff: 0.153, corr: 0.265)
- L7/22238 references to examples or lists in discussions or reports (coeff: 0.446, corr: 0.282)

- L8/7704 keywords related to television series and their reception (coeff: 0.630, corr: 0.244)
- L9/4007 references to various types of businesses and their classifications (coeff: 0.298, corr: 0.248)
- L10/3783 key phrases and concepts related to business development and investment processes (coeff: 0.454, corr: 0.281)
- L11/7301 components of structured data or content organization (coeff: 0.807, corr: 0.261)
- L12/28750 financial terms and conditions related to trading or commerce (coeff: 0.563, corr: 0.306)
- L13/16587 phrases indicating action or involvement in events or developments (coeff: 0.366, corr: 0.285)
- L14/28135 references to specific geographic locations or entities (coeff: 0.490, corr: 0.312)
- L15/9446 expressions of passion and enthusiasm in various contexts (coeff: 0.425, corr: 0.337)
- L16/6219 code-related syntax and structures within programming languages (coeff: 0.342, corr: 0.323)
- L17/26604 references to programming concepts and structures (coeff: 0.469, corr: 0.357)
- L18/2624 references to criminal activity and associated legal consequences (coeff: 0.478, corr: 0.371)
- L19/6432 numerical values and the structure of dates or game scores (coeff: 0.966, corr: 0.381)
- L20/28406 tokens related to timestamps, specifically date and time formats (coeff: 0.628, corr: 0.368)
- L21/15538 references to time management techniques and motivational strategies (coeff: 0.391, corr: 0.345)
- L22/11286 monetary amounts or financial figures (coeff: 0.697, corr: 0.380)
- L23/21146 programming and coding structures, particularly related to network protocols and data handling (coeff: 0.853, corr: 0.348)
- L24/7967 references to specific locations or addresses (coeff: 0.837, corr: 0.350)
- L25/16619 instances of authorship and attribution in the text (coeff: 0.864, corr: 0.347)
- L26/22133 names of authors and their affiliations in academic contexts (coeff: 0.813, corr: 0.413)
- L27/19268 references to academic qualifications, research, and involvement in educational activities (coeff: 0.318, corr: 0.271)
- L28/23202 specific numbers and their context within factual statements (coeff: 1.120, corr: 0.304)
- L29/12442 patterns related to digital platforms and software updates (coeff: 2.528, corr: 0.249)
- L30/19427 specific numerical values and statistical data (coeff: 0.374, corr: 0.311)
- L31/9926 numbers, particularly in relation to financial data and statistics (coeff: 10.348, corr: 0.280)

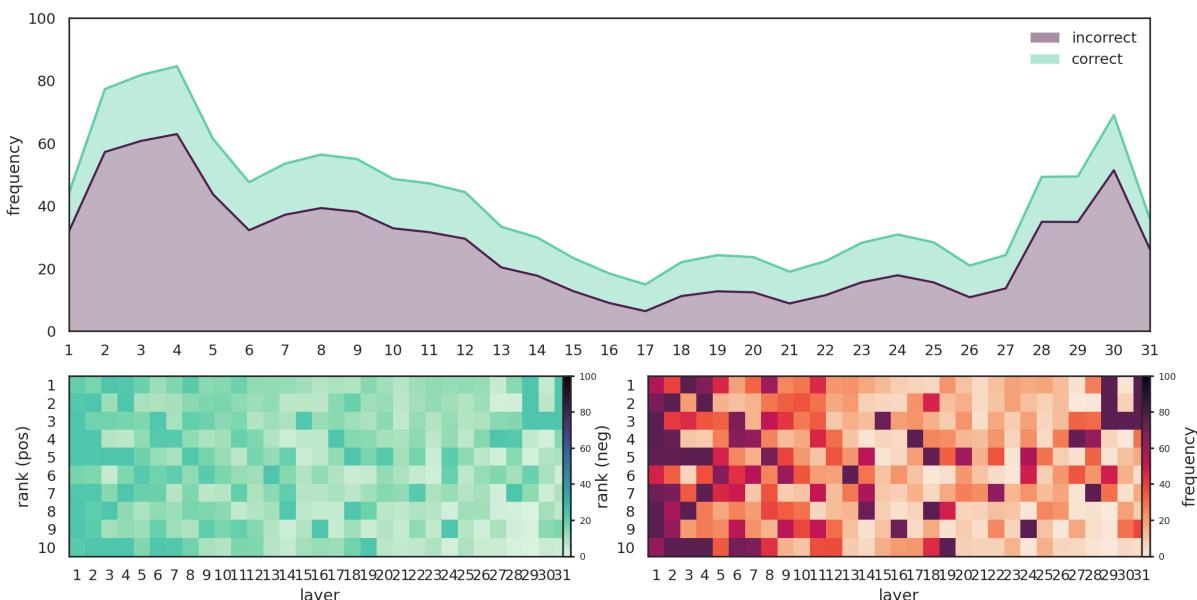


Figure: Top correlated features with MMLU-Pro on frequency in each layer of Llama 3.1 8B.

SimpleQA

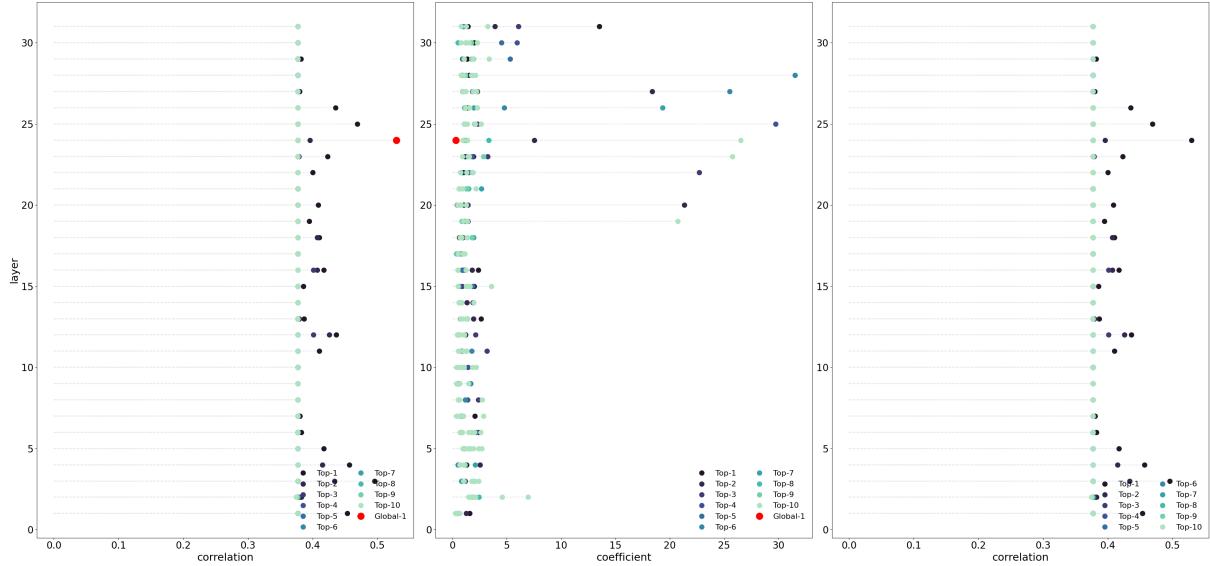


Figure: Top correlated features with SimpleQA on frequency in each layer of Llama 3.1 8B.

- [L1/28160](#) references to height, specifically focusing on the term "tall" (coeff: 1.580, corr: 0.454)
- [L2/16190](#) references to geographical locations, particularly islands (coeff: 0.148, corr: 0.383)
- [L3/24193](#) references to deserts and desert-related imagery (coeff: 0.541, corr: 0.496)
- [L4/25100](#) references to dumpster rental services and pricing (coeff: 0.205, corr: 0.457)
- [L5/15924](#) the occurrence of the word "in" and its context within the text (coeff: 0.396, corr: 0.418)
- [L6/7008](#) references to artificial entities and technologies (coeff: 2.402, corr: 0.383)
- [L7/6257](#) terms and phrases related to artificial elements or creations (coeff: 2.049, corr: 0.381)
- [L8/30264](#) phrases or terms that indicate suitability or excellence in context (coeff: 0.029, corr: 0.377)
- [L9/23784](#) programming-related keywords and constructs (coeff: 0.089, corr: 0.377)
- [L10/30120](#) phrases that encourage action or reminders related to specific tasks (coeff: 0.057, corr: 0.377)
- [L11/962](#) conjunctions that introduce reasoning or causation (coeff: 0.396, corr: 0.410)
- [L12/31391](#) references to authors and their written works (coeff: 0.472, corr: 0.437)
- [L13/19013](#) references to biological family classifications (coeff: 2.618, corr: 0.387)
- [L14/12579](#) references to global outreach and international presence (coeff: 0.077, corr: 0.377)
- [L15/18867](#) references to biological classifications, specifically family names in taxonomy (coeff: 2.004, corr: 0.386)
- [L16/22032](#) biological classifications of species, particularly family and genus names (coeff: 2.364, corr: 0.417)
- [L17/30566](#) phrases related to ownership or affiliation (coeff: 0.884, corr: 0.377)
- [L18/24624](#) specific terms associated with the media and entertainment industry (coeff: 0.952, corr: 0.410)
- [L19/25841](#) references to personal growth and transformation experiences (coeff: 1.140, corr: 0.395)
- [L20/23840](#) references to legislative districts and redistricting processes (coeff: 0.438, corr: 0.409)
- [L21/9851](#) references to volcanic activity (coeff: 0.258, corr: 0.377)
- [L22/20579](#) references to educational programs and initiatives (coeff: 0.744, corr: 0.400)
- [L23/11708](#) complex arguments and perspectives in academic discourse (coeff: 0.323, corr: 0.423)
- [L24/14877](#) specific procedural or data-related elements in formal documents (coeff: 0.292, corr:

0.530)

- L25/18055 words associated with appreciation and commendation (coeff: 0.542, corr: 0.469)
- L26/10617 emotional expressions and relationships in personal narratives (coeff: 0.317, corr: 0.435)
- L27/135 activities related to travel and tourism (coeff: 0.924, corr: 0.380)
- L28/29877 references to the concept of "home." (coeff: 0.964, corr: 0.377)
- L29/4392 references to clothing and dress codes, particularly in relation to gender identity and expression (coeff: 0.410, corr: 0.382)
- L30/22633 public methods in a programming context (coeff: 0.310, corr: 0.377)
- L31/6171 references to artificial intelligence and its related concepts (coeff: 1.429, corr: 0.377)

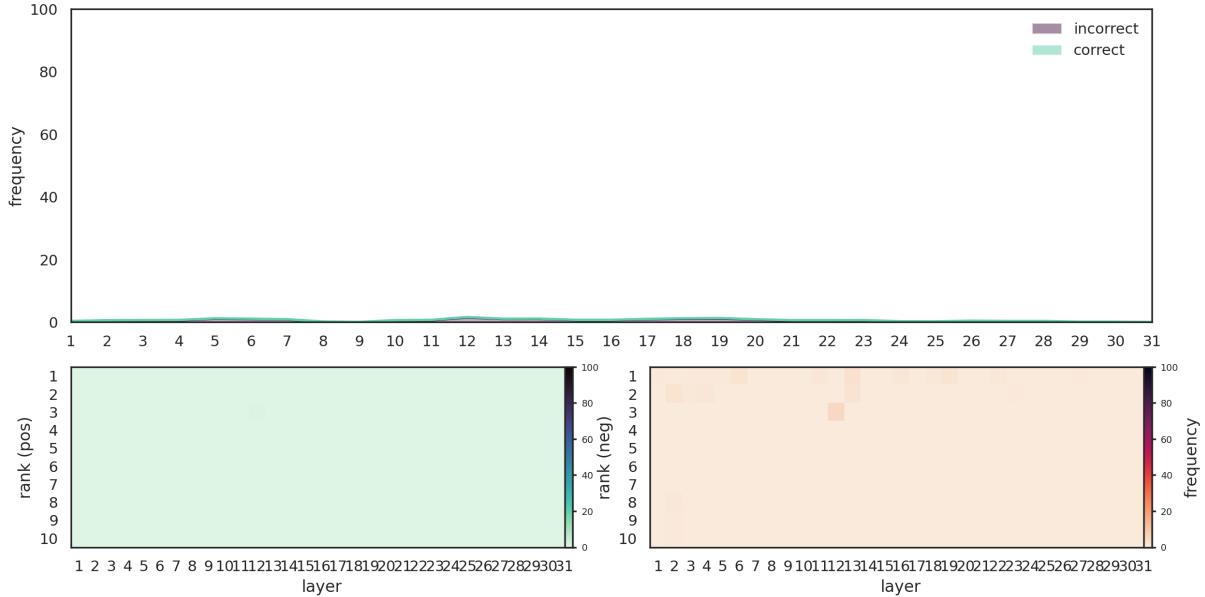


Figure: Top correlated features with SimpleQA on frequency in each layer of Llama 3.1 8B.

XSTest

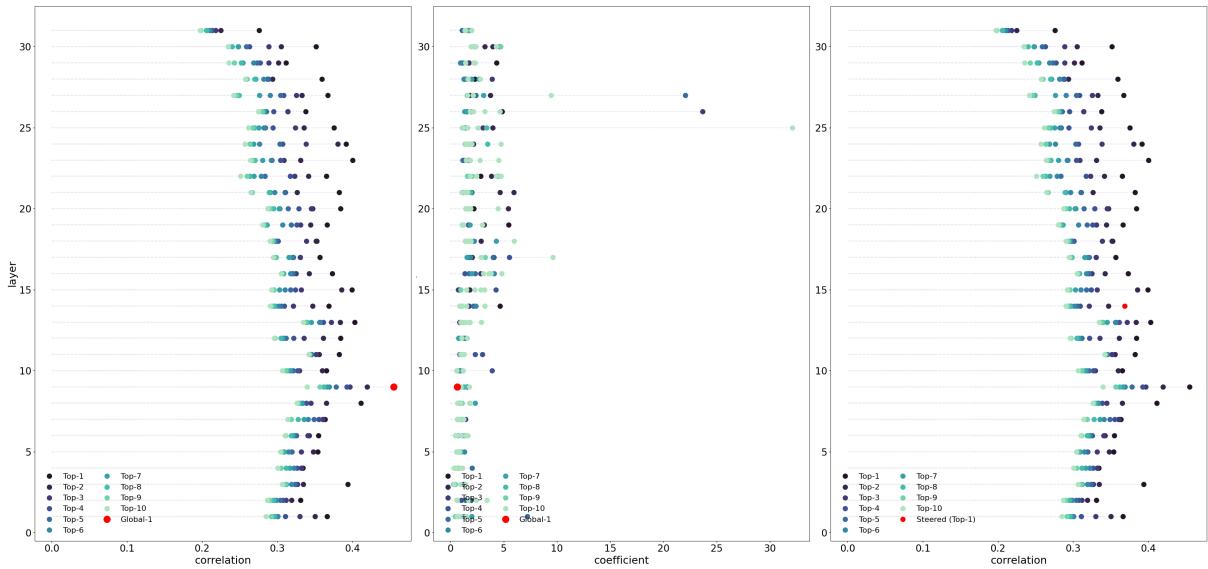


Figure: Top correlated features with XSTest on frequency in each layer of Llama 3.1 8B.

- L1/6754 references to studies and publications (coeff: 0.256, corr: 0.367)
- L2/5332 names and characteristics associated with aviation or flight (coeff: 0.276, corr: 0.331)

- [L3/16461](#) terms related to marine life and conservation efforts (coeff: 1.265, corr: 0.394)
- [L4/2446](#) proper nouns and specific entities (coeff: 0.310, corr: 0.334)
- [L5/25000](#) names of notable individuals and places related to historical or cultural significance (coeff: 0.862, corr: 0.354)
- [L6/10424](#) information related to personal details and statistics about individuals (coeff: 0.220, corr: 0.355)
- [L7/20235](#) words and phrases associated with measurement or assessment (coeff: 0.784, corr: 0.364)
- [L8/22807](#) concepts related to capital budgeting and investment decision-making (coeff: 0.420, corr: 0.411)
- [L9/16423](#) references to specific organizations, laws, or conditions related to societal issues (coeff: 0.636, corr: 0.455)
- [L10/11238](#) phrases related to collaboration and community involvement (coeff: 0.880, corr: 0.365)
- [L11/29172](#) legal terminology related to civil rights and obligations (coeff: 0.618, corr: 0.383)
- [L12/19663](#) negative descriptors or concepts related to cowardice and existence (coeff: 0.735, corr: 0.384)
- [L13/19506](#) numeric or alphanumeric strings and specific identifiers (coeff: 0.608, corr: 0.403)
- [L14/13505](#) structured question-answer formats and indicators of a discussion or inquiry (coeff: 4.659, corr: 0.369)
- [L15/23853](#) references to female characters and their relationships in narratives (coeff: 0.682, corr: 0.400)
- [L16/1652](#) names and identifiers related to locations and organizations (coeff: 1.220, corr: 0.373)
- [L17/21476](#) references to influential figures in scientific history and significant concepts from their work (coeff: 2.046, corr: 0.357)
- [L18/25543](#) names and specific references related to individuals, locations, and organizations in a political context (coeff: 0.941, corr: 0.353)
- [L19/2102](#) significant historical events and their impact on society (coeff: 1.691, corr: 0.366)
- [L20/21486](#) various references to awards, accolades, and notable achievements within literary and cinematic contexts (coeff: 2.183, corr: 0.385)
- [L21/8477](#) references to influential figures and their contributions in various contexts (coeff: 2.008, corr: 0.383)
- [L22/16870](#) references to disasters and their impacts (coeff: 2.837, corr: 0.366)
- [L23/15524](#) references to specific events or characters in films (coeff: 1.834, corr: 0.400)
- [L24/15231](#) references to specific events or characters in films (coeff: 1.747, corr: 0.392)
- [L25/16855](#) references to corporate entities and financial transactions (coeff: 0.763, corr: 0.375)
- [L26/1578](#) references to specific individuals or organizations involved in social causes or environmental conservation (coeff: 0.948, corr: 0.338)
- [L27/11758](#) connections to authoritative figures and organizational roles (coeff: 1.300, corr: 0.367)
- [L28/425](#) instances of specific names and organizational references in a text (coeff: 2.291, corr: 0.360)
- [L29/17372](#) terms related to health and illness (coeff: 0.888, corr: 0.312)
- [L30/11223](#) titles and descriptors of programs or services related to community support (coeff: 4.643, corr: 0.352)
- [L31/2111](#) descriptions and features of software products (coeff: 1.614, corr: 0.276)

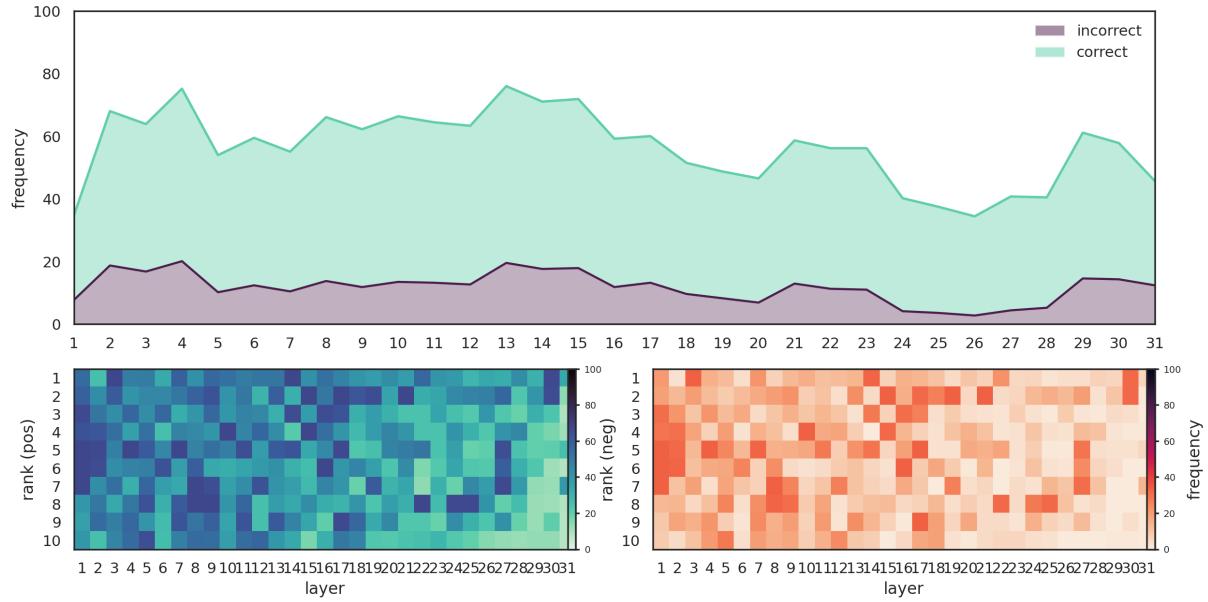


Figure: Top correlated features with XSTest on frequency in each layer of Llama 3.1 8B.