Steering Large Language Models using Conceptors: Improving Addition-Based Activation Engineering

Joris Postmus
University of Groningen
Groningen, Netherlands
j.postmus@student.rug.nl

Steven Abreu
University of Groningen
Groningen, Netherlands
s.abreu@rug.nl

Abstract

Large language models (LLMs) have transformed AI, yet reliably controlling their outputs remains a challenge. This paper explores activation engineering, where outputs of pre-trained LLMs are controlled by manipulating their activations at inference time. Unlike traditional methods using a single steering vector, we introduce conceptors–mathematical constructs that represent sets of activation vectors as ellipsoidal regions. Conceptors act as soft projection matrices and offer more precise control over complex activation patterns. Our experiments demonstrate that conceptors outperform traditional methods across multiple in-context learning steering tasks. We further use a Boolean algebra over conceptors that allows for combined steering goals using Boolean logic, that empirically outperforms combining steering vectors on a set of tasks. These results highlight conceptors as a promising tool for more effective steering of LLMs.

1 Introduction

Large language models (LLMs) have rapidly advanced AI capabilities [22], but their potential to spread misinformation [16], reinforce biases [6], and develop harmful behaviors [18] highlights the urgent need for methods to understand and control their outputs. Various methods, including reinforcement learning from human feedback (RLHF) [15], supervised fine-tuning [5], and prompt engineering [14], have been proposed to steer LLM outputs toward desired patterns. However, RLHF and fine-tuning are computationally expensive and struggle with generalization [2, 1], while prompt engineering often produces inconsistent results [4]. *Activation engineering* [13, 20] has recently been proposed as a new steering method which works by directly modifying the model's activations at inference time without changing the model's parameters and without expensive optimization. A steering vector that represents desired behavior can be computed directly or (more commonly) contrastively from positive and negative example [17]. However, finding contrastive prompts to demonstrate complex patterns is not always possible, and most importantly, the performance of activation addition is not always reliable [20]. This paper introduces an alternative to the current predominant approach for steering LLMs using activation engineering. Instead of averaging or subtracting a set of activation vectors to form a steering vector, the cached activations are used to compute a *conceptor*, which we refer to as a "steering matrix". Instead of manipulating the LLM's activations using vector addition, the activations are (softly) projected using a matrix-vector multiplication with the steering matrix. We contribute the following: (1) we introduce a novel conceptor steering mechanism for LLMs, (2) we apply this mechanism to function vectors on GPT-NeoX and GPT-J, and (3) we show how a Boolean algebra on conceptors can be used for combining functions on GPT-J.

2 Background

Adding steering vectors to the residual stream has successfully been used to control the output of LLMs across various domains [20, 17, 21]. The use case that will mainly be focused on here are

Preprint. Under review.

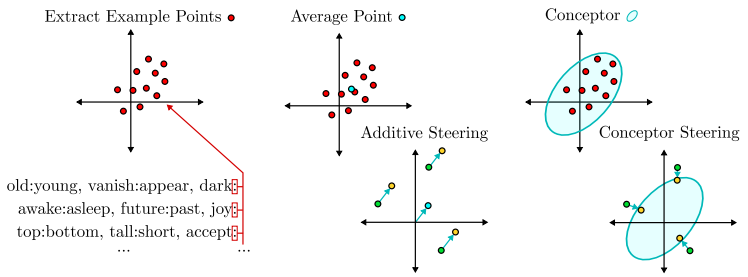


Figure 1: Illustration showing the basic geometric difference between additive and conceptor steering using a set of activations for the antonym task.

the findings from the paper by Todd *et al.* [19]. Their work showed that a steering vector can be extracted from the residual stream that captures the activation space of an input-output function (e.g. a function that takes a word and returns its antonym). This steering vector can then be added to the residual stream at inference time to steer the model toward performing the captured function.

Their baseline method works as follows. First, a set of in-context learning (ICL) prompts P_f that demonstrate a particular task f (the execution of an input-output function) are compiled. Then for each prompt $p_i^f \in P_f$ (*e.g.*, $p_1^{antonym} =$ `hot:cold,old:`) the final token's activations $h_\ell(p_i^f)$ are cached at a specific layer ℓ from the residual stream h. The cached activation vectors are then averaged into the steering vector $\bar{h}_\ell^f$ for task f at layer ℓ:

$$\bar{h}_\ell^f = \frac{1}{|P_f|} \sum_{p_i^f \in P_f} h_\ell(p_i^f) \tag{1}$$

To steer the model towards performing this function, the function (steering) vector $\bar{h}_\ell^f$ can be added (without additional re-normalization) to the residual stream at layer ℓ when the model would be completing a prompt containing a previously unseen input:

$$h'_\ell = \beta_{\text{add}}\, \bar{h}_\ell^f + h_\ell \tag{2}$$

where h'_ℓ is the steered activation and $\beta_{add} > 0$ is a hyperparameter. The performance of additive steering can further be improved by a technique called *mean-centering* [11], see Appendix A.3.1.

3 Conceptors as Steering Matrices

Conceptors can broadly be defined as a neuro-computational mechanism designed to encapsulate and manipulate the state space of neural activations [9]. A conceptor matrix C is a positive semi-definite matrix that captures the principal directions and variances of a set of neural activation vectors. This structure can be visualized as a high-dimensional ellipsoid that describes the overall shape and spread of the activations' "underlying pattern", or state space region. Because conceptors are computed from the cloud of activation vectors and encode the correlations between activations, conceptors can better capture the activation space of complex patterns compared to simple point representations, which discard information about correlations.

One way to formalize the conceptor matrix C, is through an optimization that minimizes the reconstruction error while using a regularization term. The objective function to be minimized is:

$$\min_C \|X - XC\|_F^2 + \alpha^{-2}\|C\|_F^2$$

where X is a matrix of neural activation vectors (stacked as rows), $\|\cdot\|_F$ is the Frobenius norm, and α is the regularization parameter also referred to as the conceptor's *aperture*. This aperture parameter α balances the trade-off between accurately representing the activation pattern and maintaining a generalized representation. When α is large, the eigenvalues μ_i approach 1 and C approaches the identity matrix, causing the conceptor to allow for more signal components to pass through the projection of the states with the conceptor matrix Cx. Conversely, when α is small, the eigenvalues μ_i approach 0, causing the conceptor to allow for less variability. In the extreme case of $\alpha = 0$, the conceptor collapses to the zero mapping.

The closed-form solution to this problem is given by:

$$C(R, \alpha) = R\left(R + \alpha^{-2} I\right)^{-1} \quad \text{where} \quad R = \frac{X^T X}{n} \tag{3}$$

where n is the number of samples, and I is the identity matrix of the same dimensionality as R.

We can use the conceptor for steering by collecting activations $h_\ell(p_i^f)$ into X and then compute the associated conceptor C_ℓ^f using Equation 3, and then steer new hidden activations with:

$$h'_\ell = \beta_c C_\ell^f h_\ell \tag{4}$$

where h'_ℓ is the steered activation and $\beta_c > 0$ is a hyperparameter. We can think of this as a "soft projection". A projection matrix has eigenvalues that are either zero or unity, but the conceptor matrix has "soft" eigenvalues between zero and unity. Thus, the conceptor "softly projects" the activation vector h_ℓ toward the pattern represented by C_ℓ^f by scaling its components according to the patterns' principal directions.

We can combine multiple steering matrices using the conceptor Boolean operations as defined by Jaeger [10]. For our experiments, we use the AND operation which combines two conceptors C_1 and C_2 into a joint conceptor $C_1 \wedge C_2 = \left(C_1^{-1} + C_2^{-1} - I\right)^{-1}$. See Appendix for A.1.1 for more details.

4 Experiments

For our experiments, we will use EleutherAI's GPT-J 6B and GPT-NeoX 20B models, as done in previous works [19, 11]. For all experiments, we find optimal hyperparameters for each steering method at every layer. The details of our grid search for α and β_c for conceptor-based steering and β_{add} for additive steering can be found in Appendix A.2.2.

4.1 Function Steering

We compare conceptor-based and additive steering mechanisms on their ability to steer a given model towards correctly executing a set of functions. We test both methods on GPT-J with 6B parameters and GPT-NeoX with 20B parameters. For each function, the described experiment will be repeated 5 times with different random seeds, and all reported results are averaged across across these five runs. The examples of the input-output functions come from the dataset by Todd *et al.* [19]. We use the following subset of five functions [11]: antonyms (e.g. good→bad), present-past (e.g. go→went), English-French (e.g. hello→bonjour), singular-plural (e.g. mouse→mice), country-capital (e.g. Netherlands→Amsterdam), and capitalize (e.g. word→Word). To ensure comparability of our results, we follow [19] as closely as possible. For more details, see Appendix A.2.1.

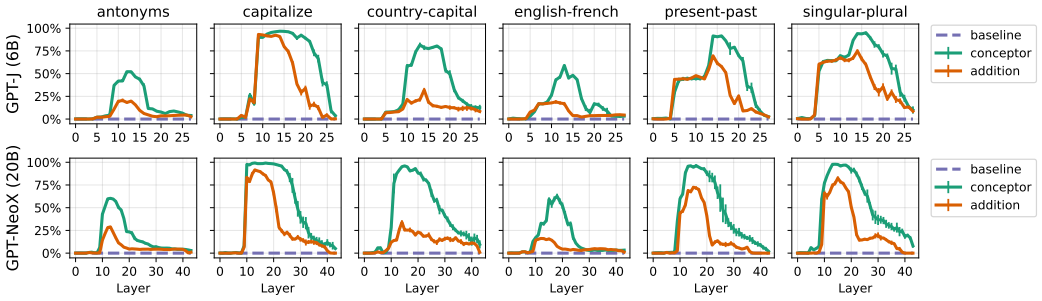


Figure 2: Comparison of the accuracy on all six function tasks for conceptor-based steering against additive steering across all layers for GPT-J and GPT-NeoX. For explanation, see main text.

The results in Figure 2 show that conceptor-based steering outperforms additive steering (the baseline method reported in Ref. [19]) for every task on both tested models. The performance improvement of the proposed conceptor mechanism is as large as In line with previous findings [19, 11], steering is most effective across layers 9-16 for GPT-J and layers 10-30 for GPT-NeoX.

4.2 Steering Composite Functions

We further conducted experiments where two conceptors representing three different compound functions were combined using the AND operator. The input-output example dataset for this function was generated using GPT-4o.

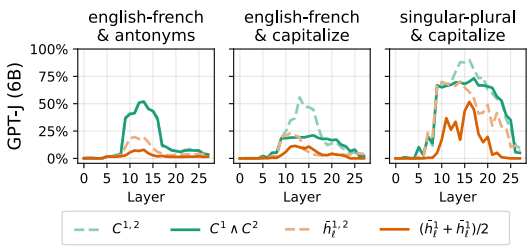


Figure 3: Performance of additive steering and conceptor steering on composite functions. For explanation of the figure caption, see text.

To present the baseline for how well non-combined steering mechanisms perform, we show results for the conceptor $C^{1,2}$ and the steering vector $\bar{h}_\ell^{1,2}$ that were each computed on the compound function directly. We then combine the conceptors computed on the individual functions C^1 and C^2 using the AND operation as $C^1 \wedge C^2$, and we combine the steering vectors $\bar{h}_\ell^1$ and $\bar{h}_\ell^2$ using their arithmetic mean $\frac{1}{2}(\bar{h}_\ell^1 + \bar{h}_\ell^2)$. Figure 3 shows the performance of all compared methods across all layers of the GPT-J model. In line with the results from Section 4.1, the conceptor baseline outperformed the additive baseline on all three tasks. The AND-combined conceptor outperformed the mean-combined steering vectors. On one of the three tasks, english-french & antonyms, the AND-combined conceptor even outperforms the additive baseline.

5 Conclusion

In our experiments, conceptor-based steering generally outperformed addition-based methods. That said, further research should be conducted to assess the mechanisms' impact on the model's overall capabilities, as well as the performance on more complex behaviors/tasks. Although we tested the steering performance on two models, examining scalability on even larger models, such as LLaMA-70B, could provide valuable insights.

A limitation of conceptors is their reliance on more data points to build accurate representations. In addition to this, the inherent mathematical structure and additional required computations makes it relatively more computationally expensive compared to simple addition-based methods. Conceptors also introduce a new hyperparameter, aperture, that requires tuning for optimal performance.

Despite these challenges, conceptor-based steering methods could offer a more precise and effective way to steer LLMs compared to traditional addition-based methods, proposing a fundamental shift in what is possible with activation engineering. While acknowledging their potential, further research is needed to fully assess their broader applicability and limitations.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv*, abs/1606.06565, 2016.

[2] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv*, abs/1606.04838, 2018.

[3] P. Bricman. Nested state clouds: Distilling knowledge graphs from contextual embeddings. Bachelor's Project Thesis, University of Groningen, Supervisors: Prof. Dr. Herbert Jaeger, Dr. Jacolien van Rij-Tange, July 2022.

[4] B. Chen, Z. Zhang, N. Langren'e, and S. Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *ArXiv*, abs/2310.14735, 2023.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and fairness in large language models: A survey. *arXiv*, abs/2309.00770, 2024.

[7] A. Gokaslan, V. Cohen, E. Pavlick, and S. Tellex. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

[8] O. He. *Continual lifelong learning in neural systems: overcoming catastrophic forgetting and transferring knowledge for future learning*. PhD thesis, University of Groningen, 2023.

[9] H. Jaeger. Conceptors: an easy introduction. *arXiv*, abs/1406.2671, 2014.

[10] H. Jaeger. Controlling recurrent neural networks by conceptors. *arXiv*, abs/1403.3369, 2017.

[11] O. Jorgensen, D. Cope, N. Schoots, and M. Shanahan. Improving activation steering in language models with mean-centring. *arXiv*, abs/2312.03813, 2023.

[12] J. Kuiper. Using conceptors to extract abstraction hierarchies from corpora of natural text: Combatting word polysemy using word sense disambiguation techniques. Master's thesis / essay, University of Groningen, Groningen, Netherlands, January 2024.

[13] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[14] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023.

[15] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[16] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Y. Wang. On the risk of misinformation pollution with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[17] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner. Steering llama 2 via contrastive activation addition. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[18] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, and A. Dafoe. Model evaluation for extreme risks, 2023.

[19] E. Todd, M. Li, A. S. Sharma, A. Mueller, B. C. Wallace, and D. Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[20] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Activation addition: Steering language models without optimization, 2024.

[21] T. van der Weij, M. Poesio, and N. Schoots. Extending activation steering to broad skills and multiple behaviours, 2024.

[22] B. Xu and M. Poo. Large language models and brain-inspired general intelligence. *National Science Review*, 10, 2023.

[23] L. S. Yifei, L. Ungar, and J. Sedoc. Conceptor-aided debiasing of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

A Appendix

A.1 Conceptors

Conceptors are mathematical constructs that can be used for the management of neural activations [9]. A conceptor can be visualized as a structure that describes the activational pattern, or state cloud, of a set of high-dimensional activation points using an ellipsoid (see Figure 4). This conceptor is mathematically represented by a positive semi-definite matrix with eigenvalues between zero and unity that can be used to (softly) project a new set of activations toward the described ellipsoid.

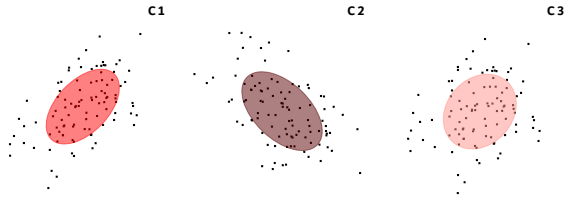


Figure 4: 2D visualization of 3 Conceptors that describe the "underlying pattern" or state space region of 3 different sets of neural activations.

Conceptors have been used to control pattern-generating RNNs effectively across various behaviors [10], prevent catastrophic forgetting and enhance continual learning in feedforward networks [8], remove bias subspaces in LLMs like BERT and GPT [23], and distill linguistic abstractions into knowledge graphs from contextual embeddings [12, 3].

The eigenvalues μ_i of the conceptor matrix C are defined as:

$$\mu_i = \begin{cases} \frac{\lambda_i}{\lambda_i + \alpha^{-2}} & \text{for } 0 < \lambda_i < 1 \text{ and } 0 < \alpha < \infty \\ 0 & \text{for } 0 < \lambda_i < 1 \text{ and } \alpha = 0 \\ 1 & \text{for } 0 < \lambda_i < 1 \text{ and } \alpha = \infty \\ 0 & \text{for } \lambda_i = 0 \text{ and } 0 \leq \alpha \leq \infty \\ 1 & \text{for } \lambda_i = 1 \text{ and } 0 \leq \alpha \leq \infty \end{cases}$$

where λ_i represents the eigenvalues of the correlation matrix R. These eigenvalues μ_i fall within the interval $[0, 1]$ and are influenced by the aperture parameter α. When α is large, the eigenvalues μ_i approach 1 and C approaches the identity matrix, causing the conceptor to allow for more signal components to pass through the projection of the states with the conceptor matrix Cx. Conversely, when α is small, the eigenvalues μ_i approach 0, causing the conceptor to allow for less variability. In the extreme case of $\alpha = 0$, the conceptor collapses to the zero mapping.

A.1.1 Boolean Operations on Conceptors

In this section, we introduce the Boolean operations on conceptors that were originally defined by Jaeger [10]. We begin with the OR operation on conceptors, which can be interpreted as merging the data from which each conceptor is computed by adding the covariance matrices on which C_1 and C_2 were computed. Given that C_1 was computed with the covariance matrix R_1 and C_2 was computed with the covariance matrix R_2, the conceptor that is computed on the sum of the two covariance matrices $R_1 + R_2$ is defined as $C_1 \vee C_2$:

$$C_1 \vee C_2 = (R_1 + R_2)(R_1 + R_2 + \alpha^{-2} I)^{-1} \tag{5}$$

$$C_1 \vee C_2 = \left(I + \left(C(I - C)^{-1} \right) + B(I - B)^{-1} \right)^{-1} \right)^{-1} \tag{6}$$

The NOT operation on a conceptor C is defined as the conceptor $\neg C$ that is computed on a covariance matrix R^{-1} that is inverse to the original covariance matrix R for conceptor C. Intuitively, $\neg C$ can be interpreted as the conceptor that arises from data that which co-vary inversely to the data giving rise to C:

$$\neg C = R^{-1}(R^{-1} + \alpha^{-2} I)^{-1} \tag{7}$$

$$\neg C = I - C \tag{8}$$

The AND operation can now be obtained using de Morgan's law $a \wedge b = \not(a \vee b)$ such that the conceptor $C_1 \wedge C_2$ is computed using the correlation matrix $(R_1^{-1} + R_2^{-1})^{-1}$. This leads to:

$$C_1 \wedge C_2 = (R_1^{-1} + R_2^{-1})^{-1} \left((R_1^{-1} + R_2^{-1})^{-1} + \alpha^{-2} I\right)^{-1} \tag{9}$$

$$C_1 \wedge C_2 = (C_1^{-1} + C_2^{-1} + I)^{-1} \tag{10}$$

A.2 Experimental Details

All experiments were run on NVIDIA GPUs. The GPT-NeoX model was run on one NVIDIA RTX A6000 with 48GB of VRAM, and the GPT-J model was run on one NVIDIA GeForce RTX 4090 with 24GB of VRAM. Each hyperparameter sweep took less than 18 hours of compute time per model and per task.

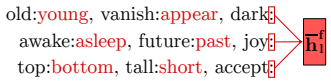
A.2.1 Function Steering

All the experimental configurations (number of experiments, number of ICL prompts and examples per prompt, accuracy metric, etc.) were, unless mentioned otherwise, adopted from Ref. [19] to ensure comparability of results.

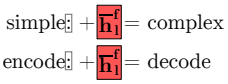
For each experiment, to generate the 4 steering mechanisms, we first compile $N_p = 100$ (ICL) prompts that demonstrate the respective input-output function. The prompts are formed by randomly sampling $N = 10$ input-output pairs from the function pairs dataset. If for a specific function, the dataset contains less than $N_p \times N = 1000$ input-output examples, this sampling is done with replacement. For each prompt p_i^f, the last input-output pair has the output stripped, resulting in the format:

$$p_i^f = "x_1 : y_1, x_2 : y_2, ..., x_{N-1} : y_{N-1}, x_N : "$$

where x represents the input tokens of a randomly sampled (input, output) pair, y represents the corresponding output tokens, N represents the number of sampled input-output pairs, and $i \in \{1, \ldots, N_p\}$. A very simple example where $N_p = 3$ and $N = 3$ can be seen in Figure 5a.



(a) Extraction of the antonym function (steering) vector $\bar{h}_\ell^f$ at layer l using 3 ICL prompts.



(b) Antonym steering vector in 2 zero-shot contexts.

Figure 5: Visualization of how an antonym function (steering) vector can be extracted and applied. Example from [19]

.

Formally, for each function $f \in F$ in our set of in-context learning (ICL) tasks, we have compiled a set P_f of ICL prompts $p_i^f \in P_f$. Each prompt p_i^f is a sequence of tokens with N input-output exemplar pairs (x, y) that demonstrate the function f mapping between x and y. For each experiment, we generate N_p such prompts.

Now that the ICL prompts have been generated, we need to extract the relevant activations. Todd *et al.* [19] showed that the neural representations of the functions are encoded in the activation vector of the last token (":") of the prompt, right before the transformer would auto-regressively start generating the output token(s). Moreover, the point in the residual stream h at which the functions were most strongly encoded was shown to be at the beginning of layers $L = \{9, \ldots, 16\}$, right before MHA and FFN [19].

Formally, for each function $f \in F$ and each prompt $p_i^f \in P_f$, the activation vectors $h_\ell^f(p_i^f)$ are extracted from the residual stream h at each relevant layer $l \in L$ from the last token's (":") activation vector.

For each function $f \in F$ and each layer $l \in L$, we now have N_p cached activation vectors $h_\ell^f(p_i^f)$ aimed to encode the neural representation of f at layer l. Using this, we can generate the layer-specific steering mechanisms for each function as follows:

- The standard additive steering mechanism $\bar{h}_\ell^f$ is generated by averaging over all the cached activation vectors $h_\ell^f(p_i^f)$ respectively as described in Equation 1.
- The additive steering mechanism with mean-centering $\bar{h}_\ell^{f,\text{mc}}$ is computed by taking the previously generated steering mechanism $\bar{h}_\ell^f$ and subtracting μ_{train} as described in Equation 11.
- The regular conceptor steering mechanism C is computed as described in Equation 3 using the aperture value α_{reg}. The correlation matrix R is computed as $R = \frac{X^T X}{N_p}$, where X is the matrix of all $h_\ell^f(p_i^f)$ stacked activation vectors.
- The mean-centered conceptor steering mechanism C^{mc} is computed with some minor adjustments. The matrix X is formed by subtracting μ_{train} from the activation vectors $h_\ell^f(p_i^f)$ before stacking them. This results in an adjusted correlation matrix R:

  $$R = \frac{(X - \mu_{\text{train}})^T (X - \mu_{\text{train}})}{N_p}$$

  The mean-centered conceptor matrix C^{mc} can then be calculated as described in Equation 3 using the aperture value α_{mc} and the adjusted correlation matrix R.

To test the performance of the generated steering mechanisms, new sets of $N_t = 1000$ input-output pairs are randomly sampled from the function pairs dataset for each experiment. This is done with replacement for functions where the dataset contains less than N_t pairs. An input prompt p_t is formatted as $p_t = "x :"$, where x is a tokenized input from an input-output pair. The tokenized output y from the pair is left out from p_t as it will be used to test the accuracy of the steering mechanisms. For each experiment, we now have N_t test input prompts p_t.

To test the accuracy of the steering mechanisms, we apply the layer-specific steering mechanisms on independent forward passes and record their subsequent output. This means that for our experimental configuration, across the functions $f \in F$, the 5 experiments, the 4 steering mechanisms (excluding the baseline), the N_t number of test prompts, and the number of layers $l \in L$, there will be $6 \times 5 \times 4 \times 1000 \times 8 = 960,000$ forward passes, each with a steering intervention.

Each steering intervention will consist of a layer-specific steering mechanism modifying the residual stream h at the mechanisms' respective layer l. This modification can be defined as transforming the unmodified residual stream activation vector h_ℓ into the steered activation vector h'_ℓ. The steering mechanisms' modification can be described as follows:

- For the standard additive steering mechanism, the averaged activation vector $\bar{h}_\ell^f$ is multiplied by the injection coefficient β_{add} and added to the residual stream activation vector h_ℓ:

  $$h'_\ell = \beta_{\text{add}}\, \bar{h}_\ell^f + h_\ell$$

- For the additive steering mechanism with mean-centering, the mean-centered average activation vector $\bar{h}_\ell^{f,\text{mc}}$ is multiplied by the injection coefficient β_{add} and added to the residual stream activation vector h_ℓ:

  $$h'_\ell = \beta_{\text{add}}\, \bar{h}_\ell^{f,\text{mc}} + h_\ell$$

- For the regular conceptor steering mechanism, the residual stream activation vector h_ℓ is multiplied using the conceptor matrix C and further multiplied with the rescaling coefficient β_{c}:

  $$h'_\ell = \beta_{\text{c}}\, C\, h_\ell$$

- For the mean-centered conceptor steering mechanism, the residual stream activation vector h_ℓ is first adjusted by subtracting μ_{train}. This adjusted vector is then multiplied with the mean-centered conceptor matrix C^{mc} and further multiplied with the rescaling coefficient β_{c}. Finally, μ_{train} is added back to the result:

  $$h'_\ell = \beta_{\text{c}}\, C^{\text{mc}}\, (h_\ell - \mu_{\text{train}}) + \mu_{\text{train}}$$

- For the baseline condition, no modifications are made to the residual stream.

  $$h'_\ell = h_\ell$$

After the respective modifications have been made to the residual stream, the forward passes will continue as usual. At the end of each forward pass, the final logits are converted into probabilities using a softmax, and the token with the highest probability is selected. This means that at the end of one experiment, we have N_t single-token outputs for each layer-specific steering mechanism. These tokens can now be compared with the first token of output y that corresponds with the input x of the initial prompt p_t. Based on how many of the N_t outputs were correctly identified, a top-1 accuracy is calculated for each layer-specific steering mechanism. This experiment is repeated 5 times for each function $f \in F$ to account for variability caused by the random sampling for the generation of the steering mechanisms and test sets.

A.2.2 Hyperparameter optimization

The performance of the steering mechanisms in the function vector experiments was optimized through a grid search over all hyperparameters. Firstly, we try steering at each layer of the model. For conceptor-based steering, we do a grid search for the aperture value α with possible values from $\{0.001, 0.0125, 0.05, 0.1\}$ and the scaling coefficient β_c with possible values from $\{0.5, 1.0, 2.0, 3.0, 4.0, 5.0\}$. For additive steering, we run a grid search over the scaling coefficient β_{add} with possible values from $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0\}$. The results from these hyperparameter sweeps are shown in Appendix A.4

A.3 Additional Experimental Results

A.3.1 Mean centering

An important improvement for additive steering is a technique called *mean-centering*, put forward by Jorgensen *et al.* [11]. This method enhances the effectiveness of steering vectors by reducing the inherent bias present in the activation space of LLMs. Activation vectors in LLMs tend to be anisotropic, meaning that they are not evenly distributed around the origin, but are instead offset in a consistent direction. This can negatively impact the steering vector's performance as the bias vector b representing this offset, does not encode any specific task-related information, diluting the steering vector's effectiveness.

First, the steering vector $\bar{h}_\ell^f$ for a specific function f is computed by averaging the activations at layer ℓ on a set of ICL prompts demonstrating the input-output function P_f (as defined in Equation 1).

$\bar{h}_\ell^f$ now encodes the task-specific behavior but may still be affected by biases in the model's overall activation space. Mean-centering attempts to mitigate this by subtracting the mean activation of a broader dataset that represents the general activation space of the model. This is done by computing the mean activation vector μ_{train} over a large, representative set of prompts D_{train} from the model's training data.

The mean activation vector μ_{train} was calculated using the same procedure described by Jorgensen *et al.* [11]: A subset from the dataset used to train GPT-2 was compiled [7]. The subset was constructed by storing all entries from the folders `urlsf_subset01-1/data` and `urlsf_subset01-182/data`. After this, only entries that contained less than 500 tokens (using the GPT-2 Tokenizer) were retained. This resulted in 210 entries from which the final 10 were removed, leaving a dataset of 200 entries. The mean activation vector μ_{train} was then computed by averaging the activations over this dataset.

Implementing the mean-centering performance enhancement for steering toward the execution of functions can be done as follows:

$$\bar{h}_\ell^{f,\text{mc}} = \bar{h}_\ell^f - \mu_{\text{train}} \quad \text{with} \quad \mu_{\text{train}} = \frac{1}{|D_{\text{train}}|} \sum_{d \in D_{\text{train}}} h_\ell(d) \tag{11}$$

where $\bar{h}_\ell^f$ is as described in Equation 1, and D_{train} is the dataset for which the mean-centered vector μ_{train} is computed. This refinement leads to a steering vector that can more effectively guide the model toward the specific task and has been shown to have a positive impact on the overall steering effectiveness [11].

Table 1 and Figure 6 show that mean-centering provides a small improvement for both addition-based and conceptor-based steering. The experimental setup is as described in Appendix A.2.1. Indeed,

Table 1: The effect of mean centering on conceptor-based and addition-based steering on the GPT-J (6B) model, across simple function vector tasks. Results show the best performance across all hyperparameters and across all layers.

	antonyms	capitalize	country-capital	english-french	present-past
Addition	20.54%	93.16%	32.04%	18.88%	69.66%
Addition (MC)	31.20%	95.00%	63.90%	34.32%	83.32%
Conceptor	52.14%	**96.68%**	81.62%	59.02%	91.56%
Conceptor (MC)	**52.82%**	96.26%	**85.32%**	**61.32%**	**91.88%**

mean-centering improves the performance of additive steering by as much as 99% (on the country-capital task). For conceptor-based steering the improvements of mean-centering are relatively smaller – at most 5% on the country-capital task. Conceptor-based steering outperforms additive steering on all tasks, even comparing additive steering with mean-centering against conceptor-based steering without mean-centering.

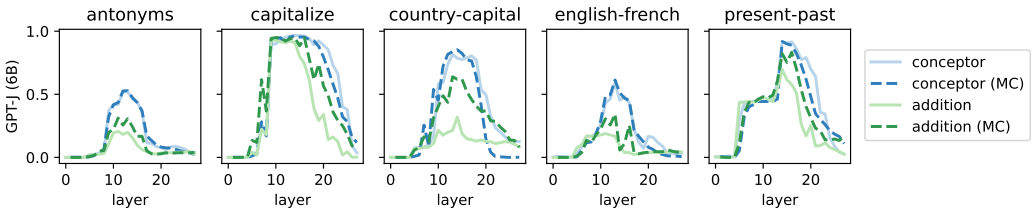


Figure 6: The effect of mean centering on conceptor-based and addition-based steering on the GPT-J (6B) model across all layers, computed on five different function vector tasks. The line shows the best average performance across five runs for the best hyperparameters for the given layer.

A.4 Hyperparameter Sweep Results

In the following section, we present results from the hyperparameter optimization described in Appendix A.2.2, in order to assess the sensitivity of both steering mechanisms (additive and conceptor-based) to the hyperparameters.

A.4.1 Conceptor Steering

Figure 7 shows that the optimal choice of aperture and beta parameters for the conceptor steering mechanism is constant at $\alpha = 0.05$ and $\beta_C = 2.0$ across all tasks for the GPT-J model (for the layer with the maximum performance). Figure 8 shows similar behavior for the GPT-NeoX model, although the optimal beta parameter is $\beta = 1$ and the optimal aperture parameter changes to $\alpha = 0.0125$ for the country-capital task, and $\alpha = 0.1$ for the english-french task, and $\alpha = 0.05$ for all other tasks. This shows that hyperparameter choices are robust for conceptor steering, but still benefit from task-specific and model-specific optimization.

We further show the performance of conceptor-based steering across all layers and different beta values (taking the best-performing aperture value) for the GPT-J model in Figure 9 and for the GPT-NeoX model in Figure 10. For the GPT-J model, the best-performing layers are typically layers 12-14 with some variability (present-past being a few layers later at 14-17, and capitalize working well across layers 9-19). For the GPT-NeoX model, conceptor steering reaches (near-)maximum performance at layer 15 across all tasks, with layer 15 being at around one third of the depth of the model. Figures 11 and 12 show the performance of conceptor-based steering across all layers and different aperture values (taking the best-performing beta value) for the GPT-J model and the GPT-NeoX model, respectively, and show a similar pattern as described above.

A.4.2 Additive Steering

Additive steering only has two hyperparameters that were being optimized: the layer on which steering was done, and the beta value that determines the "steering strength". Figure 13 shows the

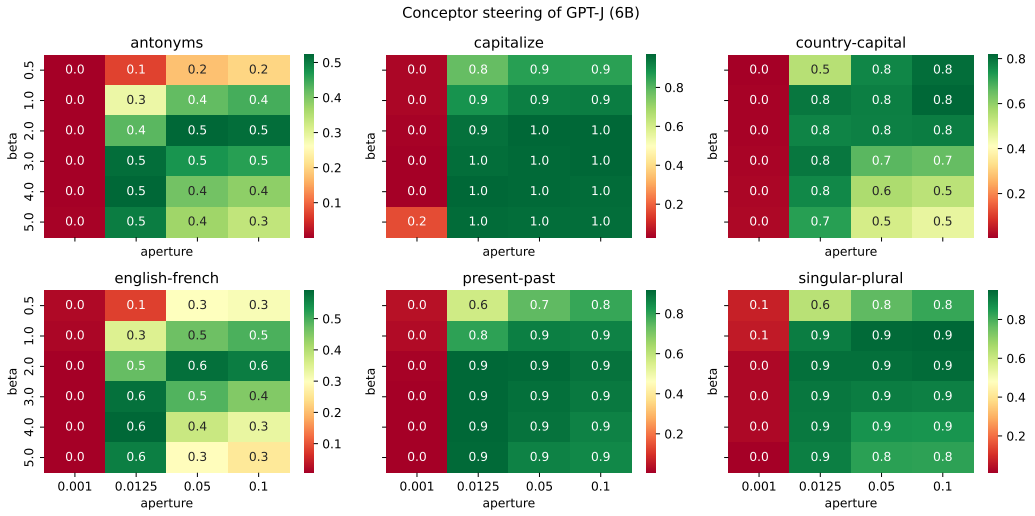


Figure 7: Performance results of the grid search across aperture and beta values (for the optimal layer) for the GPT-J (6B) model, using conceptor-based steering.

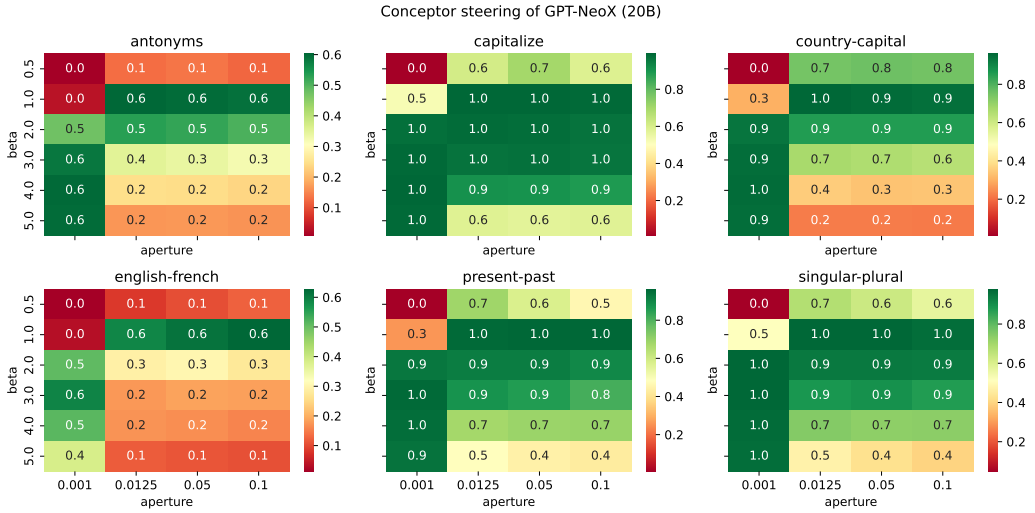


Figure 8: Performance results of the grid search across aperture and beta values (for the optimal layer) for the GPT-NeoX (20B) model, using conceptor-based steering.

performance of additive steering on the GPT-J model across all layers and beta values. Similarly to the results of conceptor-based steering, additive steering works best across layers 9-14 with peak performance always between layers 12-14. The best-performing beta values are 2.0, 3.0, and 4.0, although 2.0 is sufficient to reach peak performance for all tasks. Figure 14 shows the performance of additive steering on the GPT-NeoX model across all layers and beta values. Similar to the best-performing conceptor-based steering hyperparameters, additive steering works best on layers 12-16. The optimal beta values are 1.5 and 2.0.

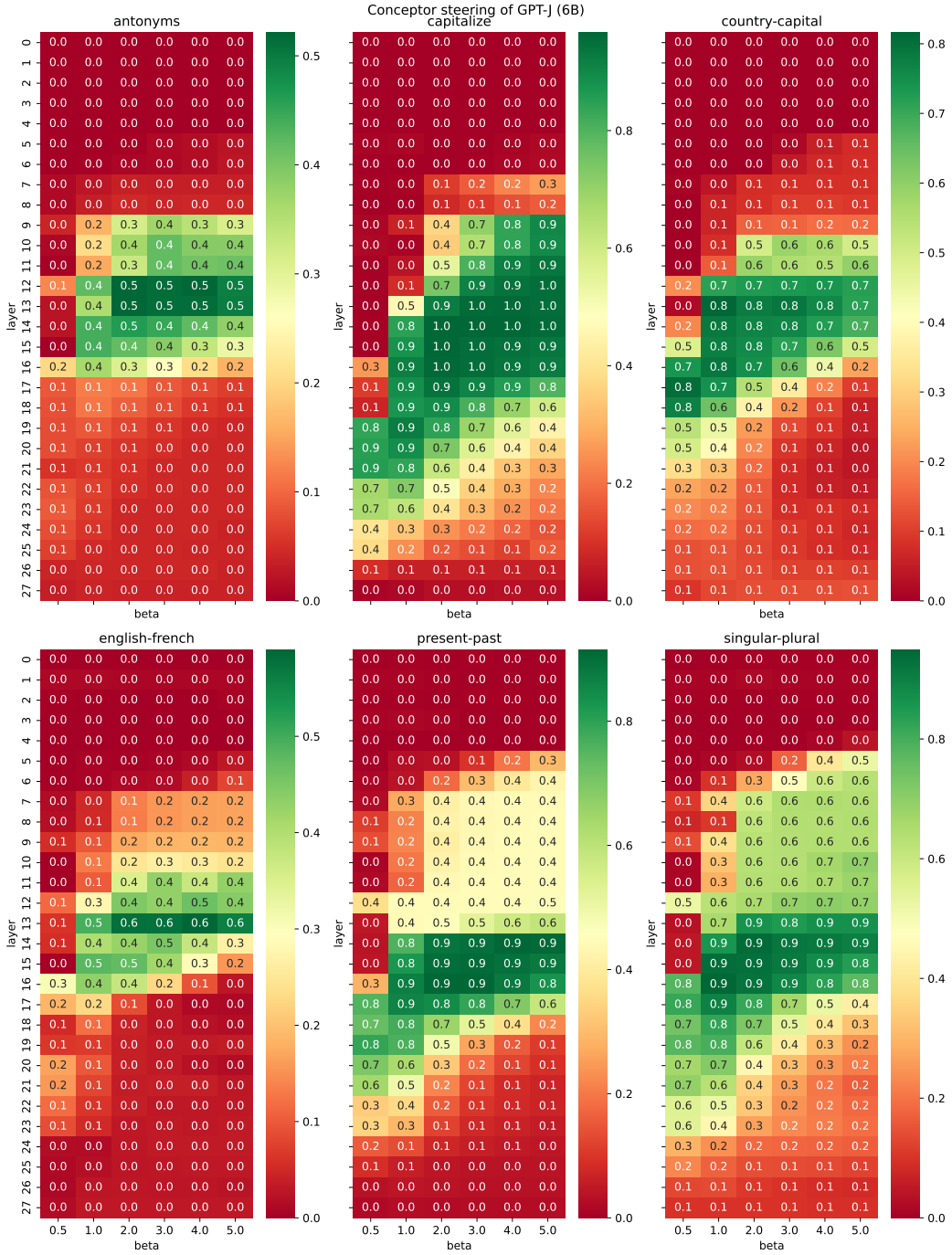


Figure 9: Performance results of the grid search across layers and beta values (for the optimal aperture value) for the GPT-J (6B) model, using conceptor-based steering.

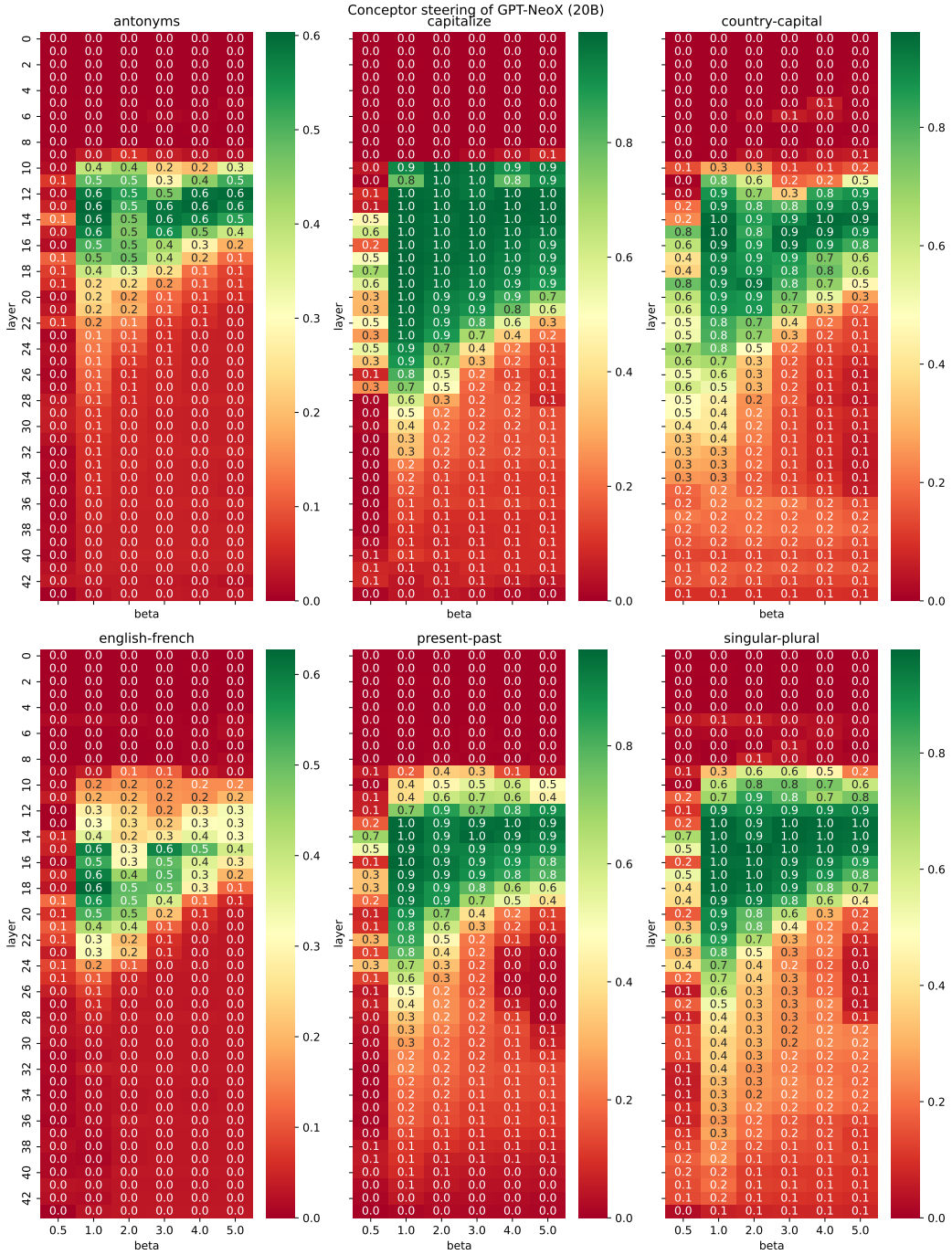


Figure 10: Performance results of the grid search across layers and beta values (for the optimal aperture value) for the GPT-NeoX (20B) model, using conceptor-based steering.

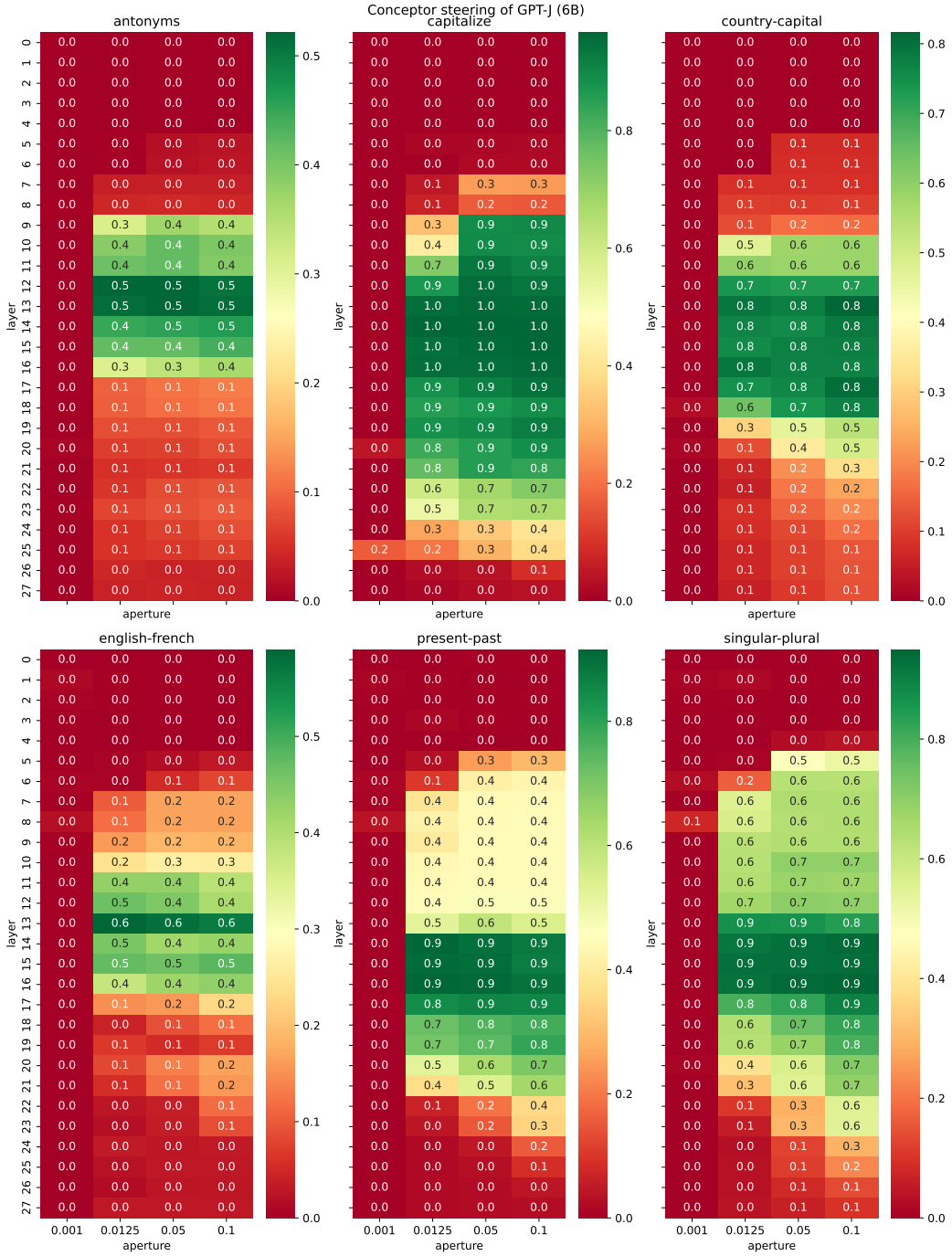


Figure 11: Performance results of the grid search across layers and aperture values (for the optimal beta value) for the GPT-J (6B) model, using conceptor-based steering.

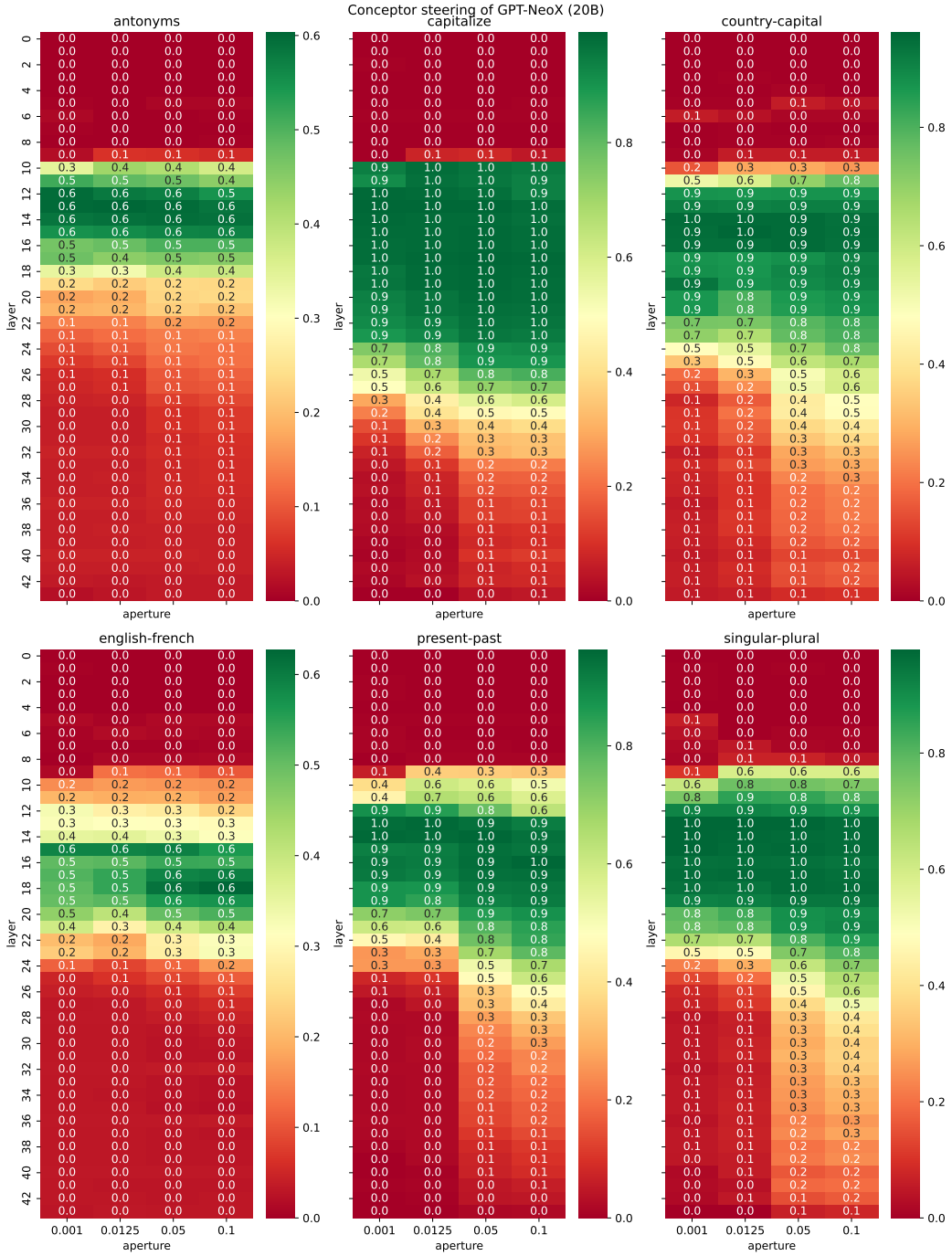


Figure 12: Performance results of the grid search across layers and aperture values (for the optimal beta value) for the GPT-NeoX (20B) model, using conceptor-based steering.

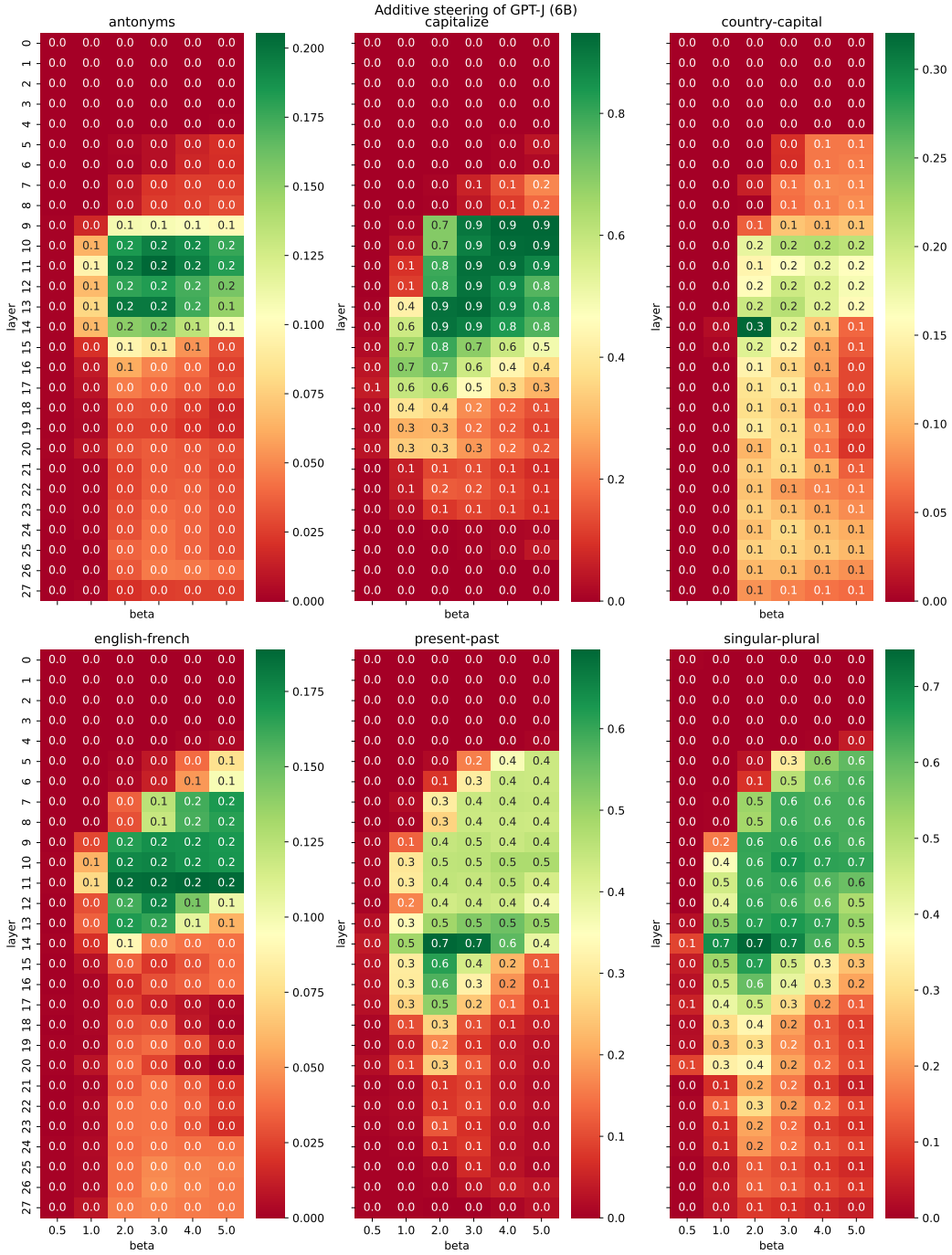


Figure 13: Performance results of the grid search across layers and beta values for the GPT-J (6B) model, using additive steering.

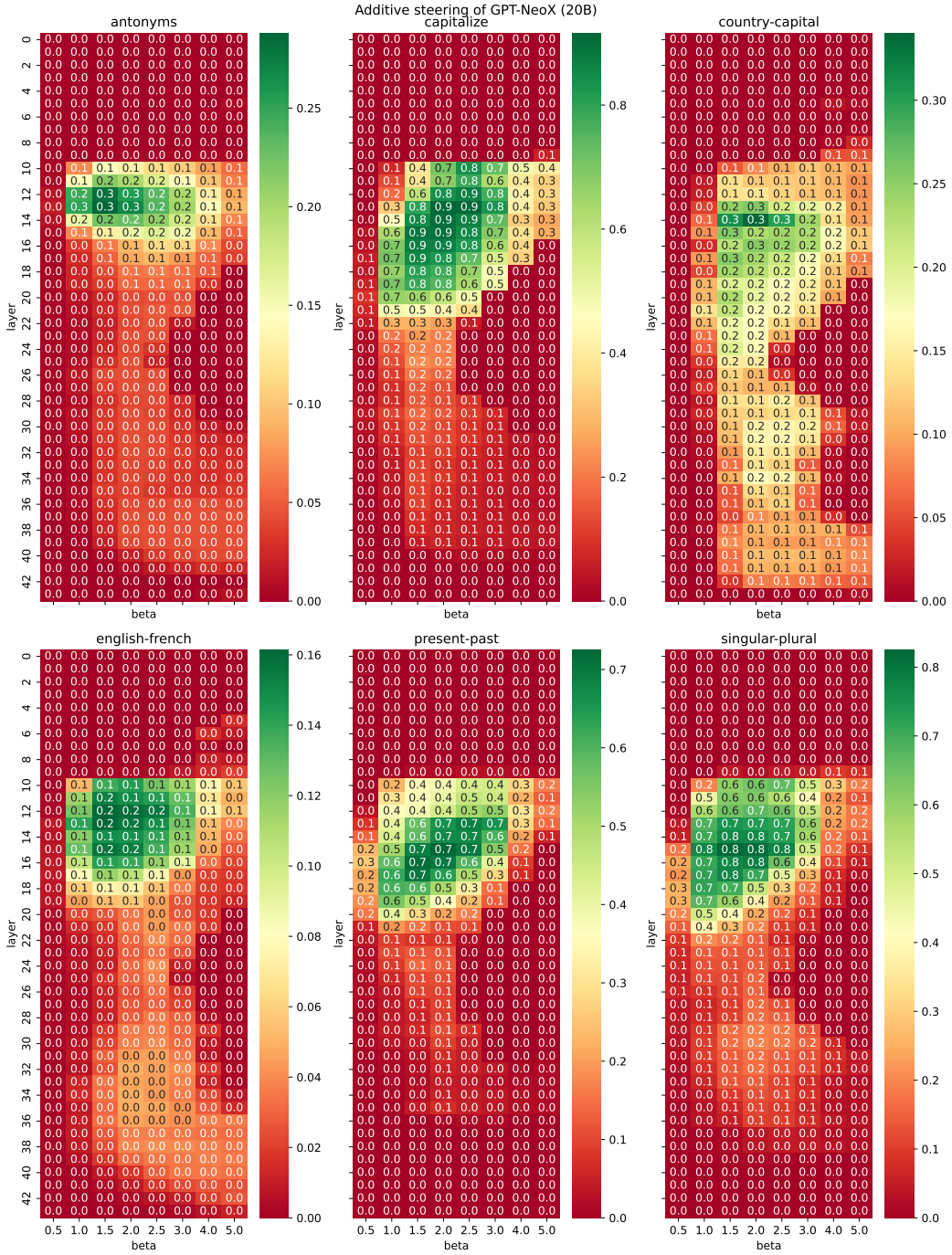


Figure 14: Performance results of the grid search across layers and beta values for the GPT-NeoX (20B) model, using additive steering.