

Creativity and Hallucinations Share Mechanisms in LLMs

Neta Glazer
Bar-Ilan University

Hadas Orgad
Kempner Institute
Harvard University

Lenny Aharon
Columbia University

Ethan Fetaya
Bar-Ilan University

neta.glazer@biu.ac.il

Abstract

A potential connection between creativity and hallucinations in large language models has been previously discussed, but their relationship is still not well understood and current evidence is primarily correlational. In this work, we study this question using mechanistic interpretability and targeted internal interventions. We localize components associated with creativity by predicting creativity metrics across established creativity benchmarks. We then intervene on these components and measure the resulting changes in factuality benchmarks. Across settings, we observe a component-level trade-off: selectively attenuating a small subset of creativity-linked components reduces creativity metrics while improving factual performance under fixed decoding. These results provide intervention-based evidence that internal computations supporting creative generation can also contribute to hallucination-prone behavior. We also observe strong alignment between creativity steering directions and hallucination-associated directions across models, providing complementary geometric evidence for a shared representational axis underlying this trade-off. Together, these findings suggest that the creativity–hallucination trade-off is not merely a behavioral tendency but a structural property of how these models internally represent and generate text.

1 Introduction

Large language models (LLMs) are increasingly used for a wide variety of applications from ideation and story writing, to coding and mathematical reasoning (Brown et al., 2020; Ouyang et al., 2022; Achiam et al., 2023; Spangher et al., 2025). At the same time, these models can sometimes produce fluent but factually ungrounded content, often referred to as hallucinations or confabulations (Maynez et al., 2020; Ji et al., 2023; Huang et al., 2025). Prior work has suggested a possible connection between creativity and hallucination in LLMs (Chakrabarty et al., 2024; Franceschelli & Musolesi, 2025), but this relationship remains poorly understood and is largely characterized by correlational observations at the level of model outputs. If such a connection exists at the level of internal mechanisms, it would imply that the capacity for creative generation is structurally entangled with hallucination-prone behavior in current architectures, making the trade-off a fundamental property of how these models are built, rather than an artifact of how they are prompted or evaluated.

Creativity in LLMs has been defined in many different ways across the literature (Franceschelli & Musolesi, 2025; Hou et al., 2025; Lu et al., 2025). In this work, we focus specifically on creative writing: the generation of novel, open-ended narratives. This is a natural setting for studying the creativity–hallucination connection because producing ungrounded content is not only common in creative writing but often desirable. Consider, for example, the Harry Potter books: the appeal of these narratives depends strongly on a combination of facts and fantastical elements, and it is precisely these fabricated details that are central to the stories’ appeal and what a successful LLM should try to mimic in this case.

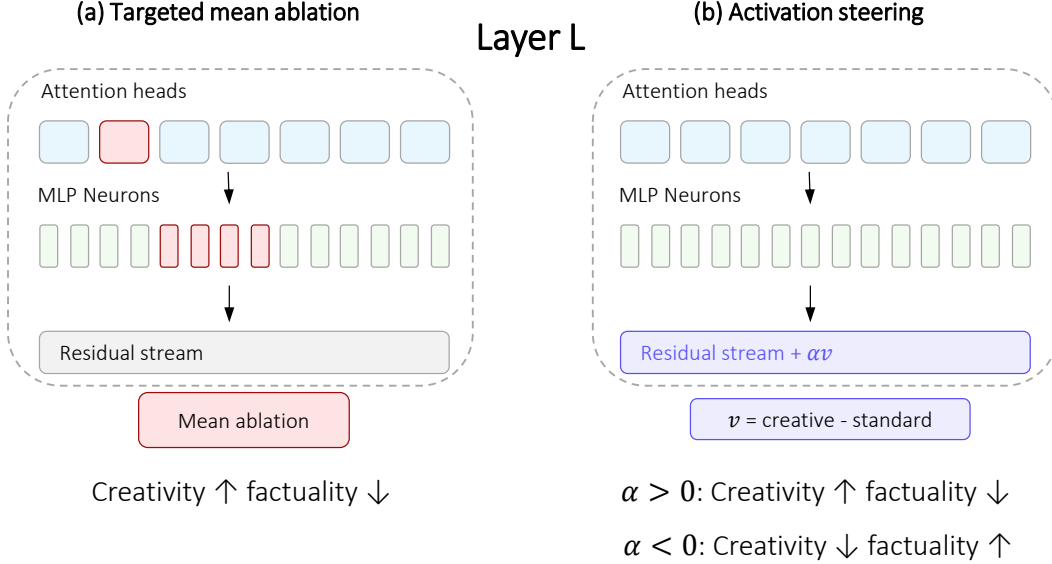


Figure 1: Our two intervention methods at layer L . (a) Targeted mean ablation replaces a component’s contribution with its reference mean $\bar{a}_c$, reducing creativity and improving factuality. (b) Activation steering adds αv to the residual stream at every generated token, with v the mean-pooled difference between creative and standard prompt responses; the sign of α determines the direction of both effects.

In these settings, producing ungrounded content is not a failure but the goal. It stands to reason that this type of creativity can leak to applications where it is not desirable, manifesting as hallucination.

Recent work has begun to examine this connection empirically, showing that creativity and hallucination move together across decoding settings (He et al., 2025), that hallucination-reduction methods affect creativity non-uniformly (Banerjee et al., 2025), and that the relationship is nonlinear and task-dependent (Yang et al., 2025). These results establish that the two behaviors co-vary, but they are measured entirely at the level of model outputs, leaving two questions open: whether the relationship is causal, and how it is reflected in the model’s internal representations.

To close this gap, we study the relationship between creativity and factuality directly within the model’s activation space. We identify model components critical to creative writing, and show that **ablating components that contribute to creativity improves factuality**, providing intervention-based evidence that creativity and hallucinations rely on shared internal mechanisms.

We then observe this connection geometrically: **steering vectors that enhance creativity reduce factuality and vice versa**, and the directions associated with creativity and hallucination are closely aligned in representation space. Across multiple experiments, benchmarks, and models, we find consistent evidence for a component- and representation-level trade-off between these two capabilities. Together, these results suggest that the creativity–factuality tension is not merely visible in model outputs, but is rooted in the model’s mechanisms.

Understanding this relationship has practical implications. Our findings suggest that hallucinations cannot be treated as an isolated failure mode: any intervention targeting creativity-linked components will, at least partially, come at the cost of factuality. This matters not only for understanding the limits of current models, but for motivating future designs that more explicitly account for the interplay between creative and factual generation. Together, our findings reframe the creativity–hallucination trade-off as a structural signature of how current language models internally represent and generate text.

2 Related works

Creativity and hallucination in LLMs. A growing body of work examines the relationship between creativity and hallucination in LLMs. Early work frames this connection conceptually by arguing that hallucinations should not be understood only as failures of factuality, but also as a possible source of creative novelty (Jiang et al., 2024; Chen & Wang, 2025). Related work further shows that hallucinated outputs can display increased narrativity and semantic coherence (Sui et al., 2024), and that some hallucinations may even be beneficial under creativity-oriented evaluation (Sun et al., 2025). More recent work studies this relationship directly and empirically. Specifically, empirical evaluations reveal a trade-off between creativity and hallucination across decoding settings (He et al., 2025), while further analysis demonstrates that hallucination-reduction methods can have non-uniform, and sometimes opposite, effects on creativity (Banerjee et al., 2025). Furthermore, this relationship has been shown to be nonlinear and task-dependent (Yang et al., 2025). Taken together, these results suggest that creativity and hallucination in LLMs are neither identical nor independent, but partially overlapping behaviors whose interaction depends on how generation is elicited and evaluated.

Mechanistic interpretability and activation-space interventions. Prior work has used ablations of attention heads and related subcomponents to estimate which parts of a Transformer are functionally important for downstream behavior (Michel et al., 2019; Wang et al., 2022; Yin & Steinhardt, 2025). Other work has shown that factual information can be localized and manipulated through interventions on internal representations (Geva et al., 2021; Meng et al., 2022; Zhong et al., 2025). Our work builds on this activation-space perspective by using mean ablation (Li & Janson, 2024; Li et al., 2026b) to test whether components that are important for creativity also play a causal role in hallucination-prone behavior.

Activation steering and pooled sequence-level representations. Activation steering and representation engineering methods construct control directions from hidden activations and inject them at inference time to modulate model behavior (Turner et al., 2023; Panickssery et al., 2023). Many such methods use token-specific representations, especially last-token states (Panickssery et al., 2023). A complementary line of work, however, uses pooled hidden activations when the target property is naturally defined at the level of the full sequence rather than a single token. In these settings, mean pooling is used to obtain a single stable representation per example before constructing contrastive directions or class-conditioned interventions (Li et al., 2026a; Zhang et al., 2025; Zou et al., 2023). This pattern appears especially in work on global and distributed properties such as emotion and safety, where the goal is to capture sequence-level behavior rather than a token-localized signal (Rimsky et al., 2024; Turner et al., 2024). Taken together, these studies suggest that pooled representations are a natural choice when the property of interest is expressed across the response as a whole.

3 Study design

We study the connection between creativity and hallucinations through two complementary families of interventions, depicted in Figure 1. The first ablates model components important for creativity, then tests whether suppressing those same components affects factual performance. The second uses activation steering to test whether the directions in representation space associated with creative and hallucination-prone behavior are geometrically aligned. Together, these two approaches provide interventional and geometric evidence supporting a shared mechanism underlying the creativity–factuality trade-off.

3.1 Models

We run all experiments on Qwen2.5-7B-Instruct (Qwen et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), two instruction-tuned decoder-only Transformer LLMs of comparable scale. We treat them as representative mid-sized instruction-following models and report results on both throughout.

3.2 Benchmarks

Creativity. We use two story-writing benchmarks that capture complementary aspects of creative generation. The first is Creative Short Story (Ismayilzada et al., 2025), a cue-word story generation task in which the model is given three words and must write a short story of at most five sentences that incorporates them. The prompt also discourages the most predictable storyline associated with the cue words. A surprise score is computed on the average semantic distance between consecutive sentences. This score was shown to correlate well with expert judgments of creativity (Hou et al., 2025). The second benchmark is CS4 (Atmakuru et al., 2024), designed to test whether models can remain creative as prompt specificity increases. Creativity is evaluated with an LLM-as-a-judge, using GPT-4 (Achiam et al., 2023). In addition to these primary story-writing benchmarks, we use two divergent-thinking benchmarks for a secondary generalization evaluation: the Alternative Uses Test (AUT) (Guilford, 1967) and the Divergent Association Task (DAT) (Olson et al., 2021), following CreativityPrism (Hou et al., 2025). Further benchmark details are provided in Section A.

Factuality. Throughout this work, we operationalize hallucination as factual inaccuracy on open-domain question-answering benchmarks: a model is said to hallucinate when it produces a fluent but incorrect answer. While this is a narrow definition, and many have suggested alternatives (Venkit et al., 2024; Rawte et al., 2023), it is well-suited to our intervention framework because it yields a precise, measurable signal that can be tracked across ablation and steering conditions.

We evaluate factual performance on two open-domain question-answering benchmarks. We prompt the model to produce a short answer (See Appendix A.4), and report exact match accuracy against reference answers. The first is TriviaQA (Joshi et al., 2017), large-scale benchmark of trivia-style question-answer pairs. The second is Natural Questions (NQ) (Kwiatkowski et al., 2019), which consists of real queries submitted to a search engine, and is generally more diverse and less uniformly formatted than TriviaQA.

Full benchmark details and evaluation protocols are provided in Section A. In all intervention experiments, we use the benchmark-specific generation settings from the original papers and report results averaged over five random seeds. As a sanity check, we verify that post-ablation perplexity remains within the baseline distribution, without a substantial deviation relative to its standard deviation, indicating that changes in creativity metrics are not simply driven by degraded generation quality.

3.3 Targeted component ablation

We study component influence through activation-level ablations, a causal intervention framework in which a component’s output is replaced with a fixed counterfactual value during inference. Figure 1a demonstrates the method. In a Transformer block, both attention heads and the MLP produce additive update vectors that are written back into the running token representation via residual addition (up to layer normalization); we refer to this running representation as the residual stream.

Let $a_c(x)$ denote the activation produced by component c on input x , representing the additive contribution written to the residual stream at a given layer and token position. A general ablation intervention takes the form:

$$a_c(x) \leftarrow \tilde{a}_c \quad (1)$$

where $\tilde{a}_c$ is the counterfactual replacement value. The choice of $\tilde{a}_c$ determines the ablation variant: *zero ablation* sets $\tilde{a}_c = 0$, replacing the component’s contribution with zero. Alternatively, *mean ablation* sets $\tilde{a}_c = \bar{a}_c$, the expected activation over a reference distribution.

For mean ablation, given a reference distribution $\mathcal{D}$, we compute:

$$\bar{a}_c = \mathbb{E}_{x' \sim \mathcal{D}} [a_c(x')] \quad (2)$$

Over all tokens in $\mathcal{D}$. This approach offers a more conservative estimate of component importance compared to zero ablation, as it maintains the component’s typical activation

pattern and only removes its context-specific modulation. Previous work (Wang et al., 2023) found this strategy to produce more stable results.

Ablation units. We treat all attention heads and MLP blocks as candidate components. Each attention head is its own ablation unit. Ablating an entire MLP block proved too coarse, causing a sharp decline in perplexity, so we partition each MLP block into four equal-sized sub-groups and treat each quarter as a separate candidate unit, yielding a finer-grained decomposition without destabilizing the model. E.g., if the hidden dimensionality is 128, we ablate either the first 32 indices, the 2nd, the 3rd or the 4th.

Identifying creativity-critical components. For each candidate component c , we perform a targeted mean ablation (Section 3.3) while holding all other components fixed, and measure the resulting change in creativity score on both benchmarks. The ablation is applied at all token positions, including both the input prompt and generated tokens.

As a sanity check, we verify on both the Creative Short Story and CS4 benchmarks that the perplexity of stories generated under component ablation does not substantially increase relative to stories generated by the original, unablated model. We additionally evaluate model perplexity on WikiText-2 (Merity et al., 2016). Full evaluation details and results are provided in Appendix B.2.

We rank components by the magnitude of the creativity drop their ablation induces. Because the two creativity benchmarks use different metrics and score scales, we aggregate at the rank level: for each model we combine the two benchmark-specific rankings via average-rank aggregation, equivalent to a Borda-style consensus (Dwork et al., 2001; Wang et al., 2024). The top 10 components from this aggregated ranking form the *creativity-critical set*.

Evaluating creativity-critical components on factual QA. Having identified the most creativity-critical components, we apply the same mean ablation interventions during inference on TriviaQA and Natural Questions under deterministic decoding, and report the per-component change $\Delta M_{\text{factuality}}(c)$ per benchmark. To assess whether any observed factuality gains are specific to the creativity-critical components, we compare them against a control distribution obtained by ablating 200 randomly sampled components.

3.4 Activation steering

Activation steering modulates model behavior at inference time by adding a learned direction vector to internal representations during the forward pass (Klerings et al., 2025; Nadeem et al., 2025). Figure 1b demonstrates the method.

We use a contrastive formulation in which steering vectors are derived from two prompt sets associated with opposing behaviors. Because creativity is expressed across an entire generated response rather than at a single token position, we represent each example by a mean-pooled residual-stream activation. Let $h_t^{(\ell)}(x)$ denote the residual-stream activation at layer ℓ and token position t for a generated output x of length T_x . The pooled representation of x at layer ℓ is

$$\bar{h}^{(\ell)}(x) = \frac{1}{T_x} \sum_{t=1}^{T_x} h_t^{(\ell)}(x). \quad (3)$$

Given two response sets, positive $\mathcal{R}_+$ and negative $\mathcal{R}_-$, representing model outputs with and without the feature of interest (e.g., creativity). We construct the steering vector from the difference between their mean pooled activations:

$$v^{(\ell)} = \mathbb{E}_{x \sim \mathcal{R}_+} [\bar{h}^{(\ell)}(x)] - \mathbb{E}_{x \sim \mathcal{R}_-} [\bar{h}^{(\ell)}(x)]. \quad (4)$$

At inference time, we inject this vector into the residual stream at every generated token position:

$$\tilde{h}_t^{(\ell)} = h_t^{(\ell)} + \alpha v^{(\ell)}. \quad (5)$$

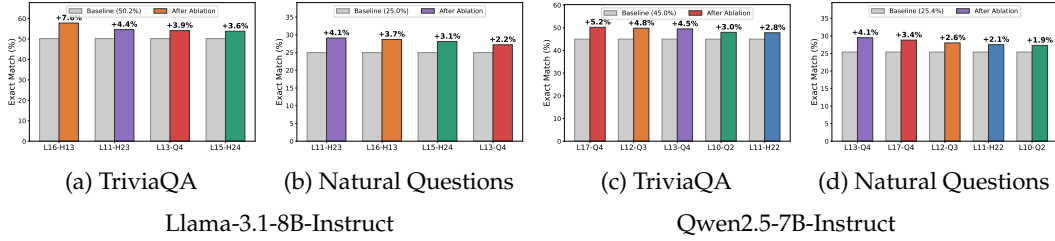


Figure 2: Factual QA accuracy improves when creativity-critical components are ablated. Each bar mean-ablates one component with all others held fixed; gray bars are the no-ablation baseline and annotations give the absolute gain in exact match accuracy. Results are presented for the components whose gains exceed the random-ablation control.

Constructing the creativity steering vector. We derive a *creativity direction* using the Creative Short Story benchmark. For each of the 50 cue-word triplets, selected to span a range of semantic distances between words ensuring diversity in cue-word relatedness, we construct two prompt variants differing only in their instruction: *Creative prompt (original)*, taken directly from the benchmark, Encourages novel, unconventional use of the cue words and explicitly rules out their most obvious interpretation. For example, given petrol, diesel, pump, the prompt specifies that the story should not be about filling a vehicle at a petrol station; *Standard prompt*, designed by us, asks for a straightforward, coherent story incorporating the same cue words naturally, with no creativity constraint.

We extract residual-stream activations for both variants by mean-pooling over the token positions of each generated response (Section 3.4), and compute the steering vector as the mean difference between the creative and standard activation pools across all 50 prompt pairs. Vectors are constructed at every layer; we then select, for each model, the layer whose steering vector produces the largest and most consistent shift in surprise score on a held-out validation set. This yields layer 16 for Qwen and layer 14 for Llama. We verify the stability of the estimated steering directions in Appendix B.4.

Steering toward and away from creativity. We apply the creativity steering vector during inference on TriviaQA, sweeping α from negative values (shifting representations away from creativity) to positive values (shifting toward creativity), and measure the effect on exact match (EM) accuracy at each setting.

Geometric alignment between creativity and hallucination directions. As a further analysis, we investigate how the *creativity direction* vector is similar to the factuality direction. To do this, we construct a factuality vector using the same procedure. We construct the positive and negative response sets by prompting each model with 500 factual questions from TriviaQA, classifying each generated answer as correct or hallucinated (incorrect), using activations extracted at the same layer as the creativity vector.

We then measure the cosine similarity between the hallucination vector $v_{\text{hallucinations}}$ and the creativity vector $v_{\text{creativity}}$. As a control, we compare this similarity measure against a semantically unrelated direction to verify that any observed alignment is specific to the creativity–hallucination axis rather than an artifact of the representation geometry. Concretely, we extract a positive–negative sentiment steering direction from movie-review prompts from SST-2 (Socher et al., 2013) and measure its similarity to both the creativity and hallucination directions.

Generalization evaluation. We apply both the creativity and hallucination steering vectors to the AUT and DAT benchmarks under negative steering, no steering (baseline), and positive steering. This evaluation tests whether the identified directions generalize beyond creative writing to divergent-thinking tasks.

Full experimental details are provided in Section A.

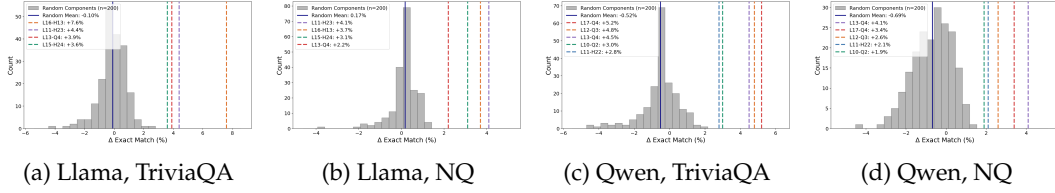


Figure 3: **Creativity-critical components vs. random components.** We show the distribution of changes in EM from ablating 200 random components, compared with selected members of the top-10 creativity-critical set (dashed lines), for Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct on TriviaQA and Natural Questions. Creativity-critical components produce substantially larger improvements in factual accuracy than random component ablations.

4 Results

4.1 Discovering Creativity–Factuality Trade-off Components

Figure 2 presents the performance of the models on the factuality benchmarks before and after ablating creativity-critical components drawn from the top-10 aggregated creativity ranking. We observe a significant increase of factuality across the two benchmarks when ablating these components. To assess whether the observed improvements could arise from generic ablation rather than the specific functional role of these components, we compare the top-10 creativity components against a control distribution obtained by ablating 200 randomly sampled components. We report here the subset of these components whose factuality gains fall significantly outside the distribution induced by 200 random component ablations. As shown in Figure 3, several components in the top-10 creativity-critical set yield consistent positive shifts in the accuracy on both factuality benchmarks when ablated. The highlighted creativity-critical components fall well beyond the right tail of this distribution, making a random-ablation explanation unlikely.

As a sanity check, we verified that these factuality gains are not driven by shorter responses. Across both models and factuality benchmarks, response-length differences between baseline and ablated outputs were negligible (Cohen’s $d < 0.1$ in all conditions; Appendix B.3).

This pattern is consistent with a component-level creativity–factuality trade-off: a subset of components that support creative generation also appear to promote hallucination-prone behavior, such that selectively ablating them improves factual reliability.

4.2 Creativity Steering and Its Effect on Factuality

As shown in Figure 4, steering along the creativity direction produces opposing effects on creativity and factuality in both models. Increasing α raises the surprise scores on the Creative Short Story benchmark, indicating more creative stories, while simultaneously reducing EM accuracy on TriviaQA, consistent with increased hallucination propensity. Steering in the opposite direction—pushing representations away from creativity—improves factual accuracy at the cost of lower creativity scores. This bidirectional pattern holds consistently across both Llama and Qwen, drawing a clear trade-off between the two. Normalized scores in Figure 4 are obtained by min-max scaling each metric over the range of α tested. The absolute value difference is presented in Appendix B.1.

This result goes beyond the correlational evidence that characterizes prior work: because the same direction vector drives opposing changes in creativity and factuality depending on the sign of α , it provides direct evidence that the creativity direction in activation space actively modulates hallucination propensity rather than merely co-varying with it. This mirrors the component-level finding and reinforces the view that the two behaviors share a mechanism.

Geometric Alignment Between Creativity and Hallucination Directions. The cosine similarity between the creativity and hallucination steering vectors is 0.74 in Qwen and 0.69

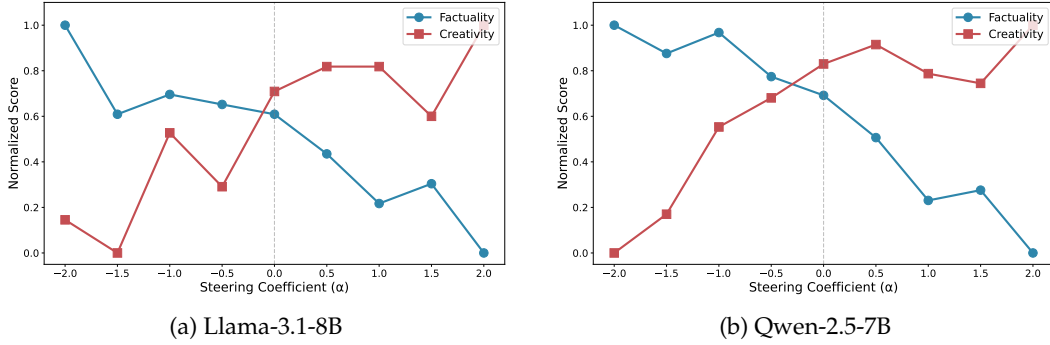


Figure 4: Effect of a creativity steering vector on two base models. Factuality is measured by Exact Match (EM) on TriviaQA, while creativity is measured by the *surprise* metric on the Creative Short Story benchmark (See 3.2). Increasing the steering coefficient α consistently raises creativity scores while reducing factual accuracy, revealing a trade-off between creativity and factuality.

in Llama, indicating strong geometric alignment. As a control, we compare these vectors to a sentiment baseline direction; in contrast to the positive creativity–hallucination alignment, the sentiment baseline is near zero or negative in both models (creativity vector: -0.46 in Qwen, -0.07 in Llama; hallucination vector: -0.39 in Qwen, -0.13 in Llama). This suggests the alignment is specific to the creativity–hallucination relationship, rather than a generic property of arbitrary steering directions. Figure 5 visualizes model activations projected onto the plane spanned by the two directions, where this alignment is visible geometrically.

4.3 Generalization Beyond Creative Writing

Across AUT and DAT, steering away from either creativity or hallucination consistently reduces performance, whereas steering in the positive direction generally improves it. This pattern holds across both models and suggests that the identified directions extend beyond creative writing to divergent-thinking behavior. Detailed results for each model, benchmark, and steering vector are provided in Appendix B.5.

5 Conclusion and Discussion

We studied the relationship between creativity and factuality in large language models through three complementary analyses: component-level ablations, activation steering, and representation-space geometry. Across all three, we found converging evidence for a shared internal basis underlying the creativity–factuality trade-off. Components identified as important for creative generation improved factual performance when ablated; steering in the creativity direction increased creativity while reducing factual accuracy; and the creativity and hallucination vectors showed substantial alignment in activation space. Together, these results indicate that the creativity–factuality tension is not merely a behavioral phenomenon visible in model outputs, but is rooted in the model’s internal representations.

These findings advance the mechanistic understanding of the creativity–hallucination connection. Prior work has established that the two can co-vary under changes in decoding settings or mitigation strategies, but it has remained unclear whether this reflects a genuine causal relationship or a coincidental coupling that could in principle be disentangled. Our results suggest that creativity and hallucination share overlapping internal computations, making the trade-off a structural feature of how these models are built rather than an artifact of how they are prompted or decoded. This has a sobering practical implication—reducing hallucinations by targeting creativity–linked components or directions will, at least partially, come at the cost of creative capability, and vice versa.

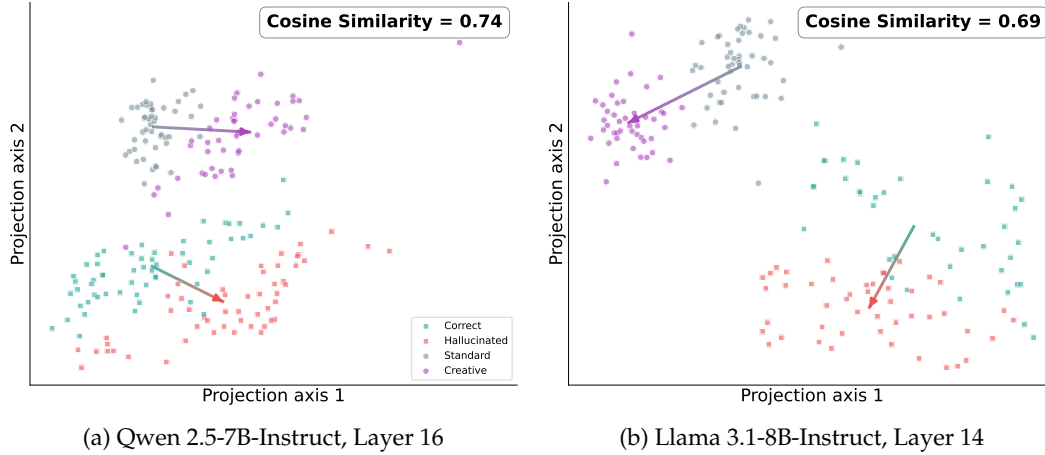


Figure 5: Creativity and hallucination directions are aligned in activation space. Model activations are projected onto the 2D plane spanned by the creativity and hallucination steering vectors. Arrows denote the displacement between the projected class means (Standard $\rightarrow$ Creative, and Correct $\rightarrow$ Hallucinated). The reported cosine similarity is computed between the corresponding full-dimensional steering vectors in the original activation space.

At the same time, identifying the trade-off precisely is itself a step toward addressing it. This points toward model designs that more explicitly separate creative and factual modes, avoiding this trade-off. For instance, through conditional routing, framing in the context, or training objectives that discourage the entanglement we observe here.

Several limitations bound the scope of these conclusions. Our experiments cover two models of similar scale and family (Qwen2.5-7B and Llama-3.1-8B); whether the same components and directions play analogous roles in larger models, encoder-decoder architectures, or models trained with different objectives remains to be seen. Our factuality evaluation also relies on short-answer QA benchmarks, and may not generalize to longer-form or more subtle errors. Finally, the present analysis focuses primarily on creative writing as the operationalization of creativity; our additional divergent-thinking experiments provide some evidence beyond this setting, but whether similar internal trade-offs arise in other creative domains such as scientific ideation, analogical reasoning, or open-ended problem solving remains an open question.

We identify several potential future directions. First, refining the components identified here into more localized and interpretable mechanisms, tracing the full circuits through which creativity-linked computations influence factual outputs. Second, study the geometry of the trade-off in more detail—its layer-wise organization, dependence on prompt and task, and relationship to other behavioral axes in the model. More broadly, our findings motivate a design-level question: can future architectures be built so that creative and factual computations are more cleanly separated, enabling models that are genuinely capable of both without sacrificing one for the other? Until then, the creativity–hallucination trade-off should be understood not as a flaw to be patched, but as a structural signature of how current models represent and generate knowledge.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. Cs4: Measuring

- the creativity of large language models automatically by controlling the number of story-writing constraints, 2024.
- Mohor Banerjee, Nadya Yuki Wangsajaya, Syed Ali Redha Alsagoff, Min Sen Tan, Zachary Choy Kit Chun, and Alvin Chan Guo Wei. Does less hallucination mean less creativity? an empirical investigation in llms, 2025. URL <https://arxiv.org/abs/2512.11509>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.
- Qiucheng Chen and Bo Wang. Valuable hallucinations: Realizable non-realistic propositions. *arXiv preprint arXiv:2502.11113*, 2025.
- Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pp. 613–622, 2001.
- Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *AI & society*, 40(5):3785–3795, 2025.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bittton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan

Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath R-parthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gouget, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshchev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk,

- Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Joy P Guilford. Creativity: Yesterday, today and tomorrow. *The journal of creative behavior*, 1(1):3–14, 1967.
- Jennifer Haase, Paul HP Hanel, and Sebastian Pokutta. Has the creativity of large-language models peaked?: An analysis of inter-and intra-llm variability. *Journal of Creativity*, pp. 100113, 2025.
- Zicong He, Boxuan Zhang, and Lu Cheng. Shakespearean sparks: The dance of hallucination and creativity in llms’ decoding layers, 2025. URL <https://arxiv.org/abs/2503.02851>.
- Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, and Xiang Lorraine Li. Creativityprism: A holistic benchmark for large language model creativity, 2025.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
- Kent F Hubert, Kim N Awa, and Darya L Zabelina. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific reports*, 14(1):3440, 2024.
- Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. Evaluating creative short story generation in humans and large language models, 2025.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.
- Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*, 2024.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.
- Alina Klerings, Jannik Brinkmann, Daniel Ruffinelli, and Simone Paolo Ponzetto. Steering language models in multi-token generation: A case study on tense and aspect. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8632–8650, 2025.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.
- Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models. *arXiv preprint arXiv:2602.01654*, 2026a.

- Maximilian Li and Lucas Janson. Optimal ablation for interpretability. *Advances in Neural Information Processing Systems*, 37:109233–109282, 2024.
- Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31645–31653, 2026b.
- Llama Team. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. Benchmarking language model creativity: A case study on code generation. 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)*, pp. 2776–2794, 2025.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1906–1919, 2020.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.
- Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, 2019.
- Afrozah Nadeem, Mark Dras, and Usman Naseem. Steering towards fairness: Mitigating political stance bias in llms. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pp. 52–61, 2025.
- Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118, 2021.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2541–2573, 2023.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.
- Alexander Spangher, Tenghao Huang, Philippe Laban, and Nanyun Peng. Creative planning with language models: Practice, evaluation and applications. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pp. 1–9, 2025.
- Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14274–14284, 2024.
- Guangyan Sun, Mingyu Jin, Zhenting Wang, James Chenhao Liang, Tong Geng, Qiang Guan, Junhan Zhao, Qifan Wang, Mengnan Du, Yongfeng Zhang, and Dongfang Liu. Hallucinating llm could be creative. OpenReview, 2025. URL <https://openreview.net/forum?id=W48CPXEpXR>. Submitted to ICLR 2025.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. 2024.
- Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. confidently nonsensical?: A critical survey on the perspectives and challenges of ‘hallucinations’ in nlp. *arXiv preprint arXiv:2404.07461*, 2024.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.
- Kevin R. Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2211.00593>.
- Siyi Wang, Qi Deng, Shiwei Feng, Hong Zhang, and Chao Liang. A survey on rank aggregation. In *IJCAI*, pp. 8281–8289, 2024.
- Chengxu Yang, Jingling Yuan, Siqi Cai, Jiawei Jiang, and Chuang Hu. Heaven-sent or hell-bent? benchmarking the intelligence and defectiveness of llm hallucinations, 2025. URL <https://arxiv.org/abs/2512.21635>.
- Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? *arXiv preprint arXiv:2502.14010*, 2025.
- Zhuoxuan Zhang, Jinhao Duan, Edward Kim, and Kaidi Xu. Sparse neurons carry strong signals of question ambiguity in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16092–16110, 2025.
- Haitian Zhong, Yuhuan Liu, Ziyang Xu, Guofan Liu, Qiang Liu, Shu Wu, Zhe Zhao, Liang Wang, and Tieniu Tan. React: Representation extraction and controllable tuning to overcome overfitting in llm knowledge editing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16994–17011, 2025.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Experimental Details

A.1 Benchmarks

Creative Short Story. We use the Creative Short Story benchmark of [Ismayilzada et al. \(2025\)](#), a cue-word-based story generation task in which the model is given three words and asked to write a creative short story of at most five sentences that incorporates them. The benchmark is designed to elicit controlled but open-ended creative generation: besides requiring the three cue words, the prompt also discourages the most obvious storyline associated with them. Example cue-word sets include *stamp*, *letter*, *send*, *petrol*, *diesel*, *pump*, and *gloom*, *payment*, *exist*.

In our experiments, we focus on the benchmark’s *surprise* metric as the primary creativity signal. Intuitively, surprise measures how unexpectedly the story evolves from one sentence to the next. Following the benchmark definition, if a story S_n is segmented into fragments (sentences) $F_1, \dots, F_{|F|}$, then surprise is computed as the average semantic distance between consecutive sentences:

$$\text{Sur}(S_n) = \frac{2}{|F| - 1} \sum_{i=2}^{|F|} |D(F_i) - D(F_{i-1})|,$$

where $D(F_i)$ denotes the semantic representation of fragment F_i , and semantic distance is instantiated using embedding-based distance. Operationally, higher surprise corresponds to larger semantic shifts between neighboring sentences, i.e., stories whose development is less predictable as they unfold.

This metric is attractive for our purposes because it provides a fully automatic signal of narrative unexpectedness, rather than relying only on holistic judge ratings. As also noted in later benchmarking work, the Creative Short Story surprise metric captures the novelty of a story’s *twist-and-turn trajectory*, rather than simply novelty of its topic or lexical content. We therefore use surprise as our main automatic proxy for creativity on this benchmark. For generation, we use a temperature of 0.7.

A.2 CS4 Benchmark and Creativity Evaluation

We use CS4 ([Atmakuru et al., 2024](#)), a constrained story-writing benchmark designed to test whether models can remain creative as prompt specificity increases. CS4 includes two settings, *instruction-based* and *story-based*, with 50 source instructions in each setting and constraint sets of size 7, 15, 23, 31, and 39, for a total of 500 evaluation prompts. In each instance, the model receives an instruction, a base story, and a set of constraints, and must revise the story to satisfy the constraints while preserving coherence.

Following the original benchmark, evaluation is performed in an LLM-as-a-judge setup. In our experiments, we use the released CS4 code and evaluation pipeline, with GPT-4 as the judge. For generation, we use a temperature of 0.8 and top- p of 0.95.

The released dataset includes real prompt and constraint files for both settings. For example, one instruction-based prompt asks for a story about two characters struggling to shift their priorities and preserve their relationship as they age; its constraint set specifies details such as a small coastal town, careers in science and the arts, distinct health issues, flashbacks to youth, and a meaningful café scene. Another prompt asks for a story about identical twins learning to live separate lives, with constraints including careers in different fields, living far apart, alternating perspectives, and a pivotal mistaken-identity event. These examples illustrate the core design of CS4: evaluating creativity under progressively denser narrative constraints.

Divergent-thinking benchmarks. For the generalization analysis, we additionally evaluate on two established divergent-thinking benchmarks: the Alternative Uses Test (AUT) (Guilford, 1967) and the Divergent Association Task (DAT) (Olson et al., 2021), following the evaluation setup of CreativityPrism (Hou et al., 2025). AUT asks the model to generate unconventional uses for common objects, while DAT asks the model to produce semantically unrelated words and measures creativity based on their semantic distance. These benchmarks provide complementary measures of divergent thinking beyond narrative creativity.

A.3 Factuality benchmarks

We evaluate factuality using two open-domain question answering benchmarks: TriviaQA (Joshi et al., 2017) and Natural Questions (NQ) (Kwiatkowski et al., 2019). In both cases, the model is given a question and asked to generate a short answer, and performance is measured using Exact Match (EM) against the benchmark reference answers.

TriviaQA. TriviaQA is a large-scale open-domain QA benchmark built from trivia-style question-answer pairs. Its questions are often knowledge-intensive and entity-centric, requiring retrieval of specific factual information such as names, dates, locations, or titles. Because many items have a relatively well-defined short answer, TriviaQA is a useful testbed for measuring whether an intervention increases or decreases factual reliability under deterministic decoding.

Natural Questions. Natural Questions consists of real user questions originally issued to a search engine, and is generally more diverse and less uniformly formatted than trivia-style QA. Compared to TriviaQA, NQ more often includes naturally phrased information-seeking questions and can involve greater variation in answer form. Evaluating on NQ therefore provides a complementary factuality test that is less tied to stylized trivia knowledge.

A.4 Model-specific notes

We run experiments on two instruction-tuned decoder-only Transformer LLMs: Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. Both models use grouped-query attention (GQA) and support long-context inference. Aside from family-specific architectural and tokenizer differences, we treat them as comparable mid-sized instruction-following base models for the purposes of our intervention and steering analyses.

Evaluation protocol. For both benchmarks, we report Exact Match (EM). A prediction is counted as correct if it exactly matches one of the accepted gold answers after standard normalization. In our intervention experiments, decoding is kept fixed across conditions so that changes in EM can be attributed to the intervention rather than to differences in sampling. For evaluation, we sample 2,000 examples from each factuality benchmark.

To better match the exact-match metric and reduce sensitivity to verbose generations, we use a short-answer prompt, following prior work and model-evaluation guidance on standardizing output format for benchmarking (Llama Team, 2024). Specifically, we use the following prompt:

Answer the question using a single word or phrase.
Question: {question}
Answer:

B Additional Results

B.1 Additional Steering results.

Table 1 reports the results of the steering experiment, showing how varying the steering coefficient α affects both factual accuracy (EM) and creativity (Surprise). Negative α values steer toward factuality, while positive values steer toward creativity.

Llama-3.1-8B (Layer 14)			Qwen2.5-7B (Layer 16)		
α	EM (%)	Surprise	α	EM (%)	Surprise
-2.0	54.2	0.039	-2.0	50.0	0.058
-1.5	50.2	0.036	-1.5	48.0	0.062
-1.0	51.1	0.047	-1.0	49.5	0.072
-0.5	50.7	0.042	-0.5	46.3	0.075
0.0	50.2	0.051	0.0	45.0	0.079
+0.5	48.4	0.053	+0.5	42.0	0.081
+1.0	46.2	0.053	+1.0	37.5	0.078
+1.5	47.1	0.049	+1.5	38.3	0.077
+2.0	44.0	0.057	+2.0	33.8	0.083

Table 1: Steering results for Llama-3.1-8B and Qwen2.5-7B. We report factual accuracy (EM) and creativity (Surprise) across steering coefficients α .

Model	Baseline PPL	Ablated PPL
Qwen2.5-7B-Instruct	6.501	6.512 ± 0.013
Llama-3.1-8B-Instruct	6.395	6.414 ± 0.015

Table 2: Perplexity on the WikiText-2 test set for the original model and under component ablation.

B.2 Perplexity Sanity Checks

We perform two complementary perplexity-based sanity checks to verify that the observed reductions in creativity under component ablation are not driven by general degradation in generation quality.

First, for each of the Creative Short Story and CS4 benchmarks, we construct a baseline perplexity distribution from stories generated by the original, unablated model. We evaluate generations using the other model as an independent cross-model evaluator: Llama-3.1-8B-Instruct evaluates Qwen2.5-7B-Instruct generations, and Qwen2.5-7B-Instruct evaluates Llama-3.1-8B-Instruct generations. For each ablated component, we compare the perplexity distribution of the generated stories against the corresponding baseline using a one-sided Mann-Whitney U test, with the alternative hypothesis that ablation increases perplexity (H_1 : $\text{PPL}_{\text{ablated}} > \text{PPL}_{\text{baseline}}$). We do not observe a substantial increase in perplexity under component ablation.

Second, we evaluate model perplexity directly on the WikiText-2 test set using a strided sliding-window approach with context size 2,048 and stride 512. As shown in Table 2, perplexity changes only minimally under ablation for both models.

Together, these results indicate that the effects observed under component ablation are unlikely to be explained by broad degradation in language modeling quality.

B.3 Response-Length Control

To test whether the factuality improvements under ablation could be explained by shorter responses, we compare response lengths between the no-ablation baseline and the creativity-critical component ablations for both models on TriviaQA and Natural Questions. As shown in Table 3, response lengths remain nearly unchanged across all conditions, with Cohen’s $d < 0.1$ for every model-benchmark pair. This suggests that the observed Exact Match improvements are unlikely to be explained by systematic reductions in response length.

Benchmark	Model	Base. Mean	Base. Std	Abl. Mean	Abl. Std	Δ	Cohen's d
TriviaQA	Qwen2.5-7B	4.36	2.26	4.29	2.15	-0.07	0.03
TriviaQA	Llama-3.1-8B	4.32	2.32	4.35	2.30	+0.04	0.04
NQ	Qwen2.5-7B	5.07	3.29	4.96	3.10	-0.11	0.03
NQ	Llama-3.1-8B	6.13	4.47	6.29	4.70	+0.16	0.07

Table 3: Mean response length with and without ablation, aggregated over components for each benchmark–model pair. Cohen’s $d < 0.1$ across all conditions indicates negligible changes in response length.

B.4 Steering-Vector Convergence

To validate the stability of the creativity steering direction with respect to the number of cue-word triplets, we construct steering vectors using subsets of $N \in \{5, 10, 20, 30, 40\}$ triplets and compute their cosine similarity to the full vector constructed from all 50 triplets.

As shown in Table 4, similarity increases consistently with N , from 0.54 ± 0.10 to 0.97 ± 0.01 for Qwen2.5-7B and from 0.48 ± 0.10 to 0.96 ± 0.01 for Llama-3.1-8B. At $N = 40$, cosine similarity exceeds 0.96 for both models, indicating that the estimated direction is highly stable. These results suggest that 50 triplets are sufficient to obtain a robust estimate of the creativity steering direction.

N	Qwen2.5-7B	Llama-3.1-8B
5	0.54 ± 0.10	0.48 ± 0.10
10	0.70 ± 0.05	0.65 ± 0.05
20	0.87 ± 0.02	0.82 ± 0.02
30	0.94 ± 0.03	0.91 ± 0.03
40	0.97 ± 0.01	0.96 ± 0.01

Table 4: Convergence of the creativity steering direction as a function of the number of cue-word triplets. We report cosine similarity to the steering vector constructed using all 50 triplets.

Vector	Model	Benchmark	α^-	Base.	α^+
Creativity	Qwen	AUT	9.81	10.08	10.51
Creativity	Qwen	DAT	65.45	68.10	73.67
Creativity	Llama	AUT	16.24	19.29	28.57
Creativity	Llama	DAT	70.60	74.65	74.36
Hallucination	Qwen	AUT	9.95	10.08	10.97
Hallucination	Qwen	DAT	60.68	68.10	70.61
Hallucination	Llama	AUT	18.33	19.29	22.38
Hallucination	Llama	DAT	74.15	74.65	74.53

Table 5: Generalization of the creativity and hallucination steering directions to the Alternative Uses Test (AUT) and Divergent Association Task (DAT). We report performance under negative steering (α^-), no steering (baseline), and positive steering (α^+).

B.5 Generalization to Additional Creativity Benchmarks

To evaluate whether the identified steering directions extend beyond narrative creativity, we consider two established divergent-thinking benchmarks that capture complementary aspects of creative idea generation. The Alternative Uses Test (AUT) (Guilford, 1967) asks for unconventional uses of common objects and evaluates the ability to generate ideas that depart from their typical function. The Divergent Association Task (DAT) (Olson et al., 2021) instead asks for semantically unrelated concepts and measures creativity through the semantic distance between the generated words.

Thus, whereas our Creative Short Story and CS4 experiments evaluate creativity in narrative generation, AUT and DAT directly target divergent thinking and have been used in several recent evaluations of LLM creativity (Hubert et al., 2024; Haase et al., 2025; Hou et al., 2025).

Following CreativityPrism, we evaluate both the creativity and hallucination steering vectors on AUT and DAT under negative steering, no steering (baseline), and positive steering. Table 5 shows a consistent effect in the negative direction: steering away from either creativity or hallucination decreases performance in all eight model–benchmark–vector settings.

Positive steering produces the complementary increase in six of eight settings. The two exceptions occur on Llama DAT, where positive steering remains close to baseline for both the creativity vector (74.36 vs. 74.65) and the hallucination vector (74.53 vs. 74.65). On AUT, in contrast, positive steering increases performance for both vectors and both models, with the largest change observed for the Llama creativity vector (19.29 $\rightarrow$ 28.57). Overall, these results show that the directions identified from our original experiments also systematically modulate divergent-thinking behavior, extending the observed relationship beyond creative writing.

B.6 Qualitative Examples of Factual Corrections**Llama 3.1-8B-Instruct****Layer 11, Attention Head 23 (layer11_h23)**

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	TriviaQA	Born in 1861, which opera singer's real name was Helen Mitchell?	Emma Calvé, a French soprano	Nellie Melba	Nellie Melba
2	TriviaQA	What is a nervous disorder whereby a person suddenly falls asleep?	Cataplexy	Narcolepsy	Narcolepsy
3	TriviaQA	Who plays the role of Bubble in Absolutely Fabulous?	Kate Moss	Jane Horrocks	Jane Horrocks
4	TriviaQA	Which Roxy Music hit single features Robert E Lee?	More Than This	Virginia Plain	Virginia Plain
5	TriviaQA	In Star Wars, what is the name of the slug-like alien gangster?	Greedo	Jabba the Hutt	Jabba the Hutt

Layer 16, Attention Head 13 (layer16_h13)

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	NQ	who co-created fred ott's sneeze	George Albert Smith	William Kennedy Dickson	W.K. Dickson
2	TriviaQA	The largest volcano in our solar system is on which planet?	Venus	Mars (Olympus Mons)	Mars
3	TriviaQA	Which canal joins the North Sea at Brunsbüttel?	The Elbe River	Nord-Ostsee-Kanal (Kiel Canal)	Kiel Canal
4	TriviaQA	When was the first road speed limit set in the UK?	1896	1861	1861
5	TriviaQA	What is 1009 when expressed as Roman numerals?	CIX	MIX	MIX

Layer 13, MLP Quarter 4 (layer13_q4)

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	NQ	when does a wrinkle in time come out in canada	A 2018 film by Ava DuVernay	March 9, 2018	March 9, 2018
2	NQ	how many house of representatives in nigeria	Nigeria has a unicameral legislature	360 members	360

3	TriviaQA	Which Oscar-nominated film had You Sexy Thing as theme?	Mamma Mia!	The Full Monty	The Full Monty
4	TriviaQA	Men Against the Sea and Pitcairn's Island were sequels to?	Typee by Melville	Mutiny on the Bounty	Mutiny on the Bounty
5	TriviaQA	Which American artist painted 'Cape Cod Evening'?	Winslow Homer	Edward Hopper	Edward Hopper

Layer 15, Attention Head 24 (layer15_h24)

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	NQ	who did kenny rogers sing mary did you know with	Michael English	Wynonna Judd	Wynonna Judd
2	TriviaQA	What was Joan Jett's backing group?	The Runaways	The Blackhearts	The Blackhearts
3	TriviaQA	The largest volcano in our solar system is on which planet?	Venus	Mars (Olympus Mons)	Mars
4	TriviaQA	The Great American Chocolate Bar is better known as?	Snickers	Hershey's	Hershey's
5	TriviaQA	Who won women's Skeleton Bob gold at 2010 Winter Olympics?	Amy Peterson	Amy Williams	Amy Williams

Qwen 2.5-7B-Instruct**Layer 17, MLP Quarter 4 (layer17_q4)**

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	TriviaQA	What Latin phrase, which translates to English as "for this", is taken to mean for the particular purpose?	Propter hoc	Ad hoc	Ad hoc
2	TriviaQA	What hill does a bill sit on before it becomes a law?	A bill sits on the Senate and House floors...	Capitol Hill	Capitol Hill
3	TriviaQA	How many champagne bottles are there in a Methusalem?	6 champagne bottles	8 champagne bottles	8

4	TriviaQA	Approximately what percentage of an incandescent light bulb's energy is wasted as heat?	80%	90%	90%
5	NQ	Host of the Late Show who was once a correspondent for The Daily Show	Trevor Noah	Stephen Colbert	Stephen Colbert

Layer 12, MLP Quarter 3 (layer12_q3)

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	TriviaQA	What was the name of the character played by Ralph Fiennes in <i>Schindler's List</i> ?	Oskar Schindler	Amon Göth	Amon Göth
2	TriviaQA	In fashion, Hobble, Pencil and Dirndl are all types of what?	clothing	skirt	skirt
3	TriviaQA	In which country was the 2006 FIFA World Cup held?	Italy	Germany	Germany
4	TriviaQA	Can you name the singer of the title track in <i>GoldenEye</i> (1995)?	Chris De Burgh	Tina Turner	Tina Turner
5	NQ	Who played Mike Stivic on <i>All in the Family</i> ?	Gary Burghoff	Rob Reiner	Rob Reiner

Layer 13, MLP Quarter 4 (layer13_q4)

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	TriviaQA	John Chapter 11 Verse 35 is the shortest verse in the Bible. How many words make up the verse?	One word	Two words	2
2	TriviaQA	What is added to white wine to make a Kir?	Blackcurrant juice	Crème de cassis	Cassis
3	TriviaQA	In which country was the 2006 FIFA World Cup held?	Italy	Germany	Germany
4	TriviaQA	In fashion, Hobble, Pencil and Dirndl are all types of what?	clothing	skirts	skirt

5	NQ	What is the average salary for a US congressman?	\$17,000	\$174,000	\$174,000
---	----	--	----------	-----------	-----------

Layer 10, MLP Quarter 2 (layer10_q2)

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	TriviaQA	In darts, what is the highest possible checkout with three darts, finishing with a bull?	180	170	170
2	TriviaQA	In which country was the 2006 FIFA World Cup held?	Italy	Germany	Germany
3	TriviaQA	Based on market cap, who was declared the top technology company?	Tesla	Apple	Apple
4	TriviaQA	Which racehorse, unbeaten in 14 races, won 9 consecutive Group 1 races?	Cape Cross	Frankel	Frankel
5	NQ	Who played Marius in the movie <i>Les Misérables</i> ?	Ewan McGregor	Eddie Redmayne	Eddie Redmayne

Layer 11, Attention Head 22 (layer11_h22)

#	DS	Question	Baseline (Wrong)	After Ablation (Correct)	Ground Truth
1	TriviaQA	In which decade of the 17th Century was the New Model Army created?	1645	1640s	1640s
2	TriviaQA	What was the first name of the father of the Brontë sisters?	William	Patrick	Patrick
3	TriviaQA	Born "Lucille LeSueur", by what name is this actress better known?	Bette Davis	Joan Crawford	Joan Crawford
4	TriviaQA	In 1965, who became the first American astronaut to walk in space?	Eugene Cernan	Edward White	Edward White
5	TriviaQA	In which year was the Royal British Legion founded?	1925	1921	1921