Dissecting Persona-Driven Reasoning in Language Models via Activation Patching

Ansh Poonia[†], **Maeghal Jain**[†]
[†]Independent
{pooniaansh11, maeghaljain}@gmail.com

Abstract

Large language models (LLMs) exhibit remarkable versatility in adopting diverse personas. In this study, we examine how assigning a persona influences a model's reasoning on an objective task. Using activation patching, we take a first step toward understanding how key components of the model encode persona-specific information. Our findings reveal that the early Multi-Layer Perceptron (MLP) layers attend not only to the syntactic structure of the input but also process its semantic content. These layers transform persona tokens into richer representations, which are then used by the middle Multi-Head Attention (MHA) layers to shape the model's output. Additionally, we identify specific attention heads that disproportionately attend to racial and color-based identities.[1]

1 Introduction

Recent advances in large language models (LLMs) have demonstrated their striking ability to adopt a wide range of personas, enabling context-sensitive and tailored responses (Serapio-García et al., 2023, Zhang et al., 2024, Joshi et al., 2024, Sun et al., 2025). However, studies such as those by Salewski et al., 2023, Deshpande et al., 2023, Gupta et al., 2024, Zheng et al., 2024 show that persona assignment can significantly influence reasoning and, in some cases, amplify underlying social biases. While these works focus on identifying and quantifying such effects, they do not examine the causal mechanisms within a pre-trained language model (PLM) that give rise to them.

Mechanistic interpretability provides framework for uncovering the causal mechanisms by which language models carry out specific tasks. Lieberum et al. (2023) introduced this approach in the context of multiple-choice question answering (MCQA), identifying "correct letter heads" in a 70B Chinchilla model-attention heads that track answer symbols and promote the correct choice based on positional order. Wiegreffe et al. (2025) examined how successful models implement symbol binding in formatted MCQA using vocabulary projection and activation patching, and Li and Gao (2025) analyzed anchored positional bias in the GPT-2 family, showing that models consistently favor the first option ('A'). Building on these interpretability techniques, we ask whether similar internal circuits also govern how personas steer a model's reasoning and potentially introduce social bias.

This study takes a step toward bridging the gap between surface-level observations of persona effects and the underlying mechanisms that produce them. We investigate the roles of key model components namely, Multi-Layer Perceptron (MLP) layers, Multi-Head Attention (MHA) layers, and individual attention heads in shaping the reasoning shifts induced by persona assignments. Using activation patching, we probe the internal circuitry of LLMs to trace the origins of persona-driven variation in objective tasks. Our findings challenge the prevailing assumption that early MLP layers are concerned solely with syntactic processing. We show that these layers also encode semantic features related to persona. Furthermore, we identify a small number of attention heads that disproportionately focus on tokens associated with race-based persona cues. Although our work is limited to uncovering the origin of persona-driven behavior, it contributes to a deeper understanding of LLMs and lays groundwork for future efforts to mitigate deep-seated biases in these systems.

2 Experimental Setup

2.1 Dataset Details

We chose the MMLU dataset (Hendrycks et al., 2021) for this study, which contains 14,024

[1]Code and some additional results are available at https://github.com/anshpoonia/Persona-Driven-Reasoning

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24553–24566
November 4-9, 2025 ©2025 Association for Computational Linguistics

multiple-choice questions spanning 57 subjects. We selected this dataset for two main reasons. First, the objective, multiple-choice format allowed us to focus on a fixed set of four answer tokens, giving us a well-defined target. Second, the wide range of subject areas helped reduce domain-specific or persona-driven biases. The identities or personas[2] we examined in our study fall into four broad categories: racial identities, color-based identities, and identities defined by positive or negative attributes; details of each can be found in the Appendix A. To maintain consistency and minimize variance, we used *student* as a gender-neutral subject across all defined identities. For instance: *Asian student*, *white student*, *good student*, etc.

2.2 Model and Prompt

Our primary investigation was conducted using the Llama 3.2 1B Instruct (Meta, 2024b) model to accommodate our computational constraints. We supplement these main findings with additional results from experiments on the Llama 3.2 3B Instruct (Meta, 2024b) and Qwen 2.5 1.5B Instruct (Yang et al., 2025). Compact size of these models makes it well-suited for efficient experimentation, while still delivering strong performance relative to other open-source models in their class. Our methods are scalable and can be extended to larger models with sufficient computational resources and minor adjustments. The Instruct variant of the models also permits the use of system prompts[3], allowing us to specify the identity the model should adopt in its responses. A standard system prompt serves as the base for each question, enabling identity shifts by modifying only two tokens in the entire prompt: just one token distinguishes each identity. We used the prompt structure defined by Meta and Qwen team for MMLU dataset (Meta, 2024a), with the addition of the system prompt, see Appendix B.

3 Persona Evaluation

We computed model outputs using zero-shot prompts to isolate the variation introduced solely by changes in the persona token, avoiding any influence from patterns introduced by few-shot prompting. This also reduced the length of each prompt, allowing for faster computation. For each prompt, we calculated the probability of the next token, which, by design, corresponds to the selected answer for the given question. This process was repeated for all 16 identities, and also the base identity.

Our main focus is the change in the probability of the correct token for each persona relative to the base identity, as shown in Figure 1. This highlights shifts in reasoning attributable to persona alone. In some cases, the differences in probability followed patterns that appeared to have a semantic basis. For instance, the negatively attributed student persona performed significantly worse than those described with positive attributes, having an average probability difference of -0.0027 ($T = 11.7$, $p < 0.001$). In contrast, the results for racially or color-coded personas were less consistent, and no definitive conclusion could be drawn about whether the patterns reflect stereotypical associations. A similar trend was observed when examining probability differences across identities for other models, see Appendix C for details. Overall, our results suggest that the model's reasoning ability varies even with minimal changes to the persona being imitated.

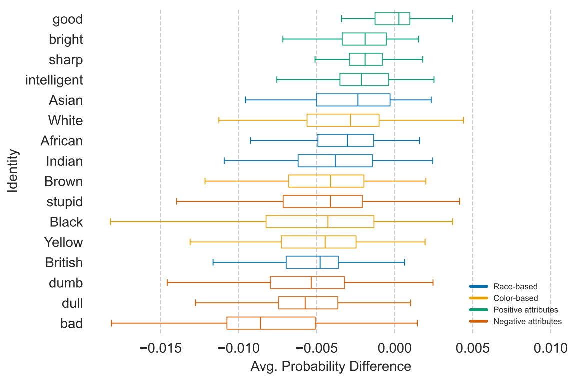


Figure 1: Average difference in the probability of the correct token for each identity, relative to the base prompt.

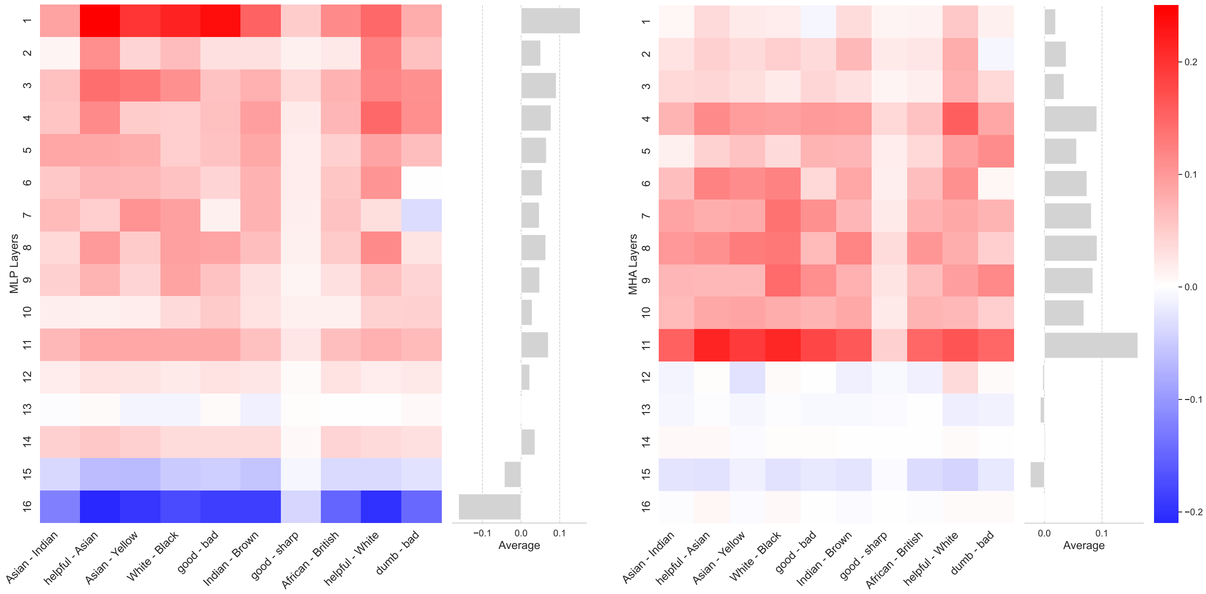
4 Interpretability Investigation

4.1 Methods

We focus on activation patching, also referred to as Resample Ablation, or Causal Mediation Analysis (Vig et al., 2020, Meng et al., 2022), to understand the influence of individual components on the model's selection of the correct answer. Specifically, we employ "de-noising" variant of activation patching, where the activation of a component from a corrupted run is replaced with the corresponding activation from a clean run (Heimersheim and Nanda, 2024). The clean and corrupted inputs are

[2]The terms persona and identity have been used interchangeably in this work.

[3]System prompts are instructions given to the AI before any user input. They define the AI's behavior, role, and response style throughout an interaction.



MLP layer patching

MHA layer patching

Figure 2: Relative logit difference (Δ_r) when MLP layer (left) and MHA layer (right) is patched in Llama 1B model. Accompanying bar chart shows the average Δ_r across identity pairs.

constructed using Symmetric Token Replacement, a technique in which only one or a few tokens are altered (Meng et al., 2022, Wang et al., 2023). This approach doesn't throw the model's internal state out-of-distribution and preserves the syntactic structure of the input (Zhang and Nanda, 2024). For each experiment, we select a pair of personas to generate the clean (`ID1`) and corrupted (`ID2`) prompts. A total of ten persona pairs were chosen keeping computational constraints in mind; details are provided in the Appendix D. Care was taken to balance pairings identities both within and across categories to ensure reliable results. Pairings with a base identity were also included for comparison. For each pair, we divide the full set of questions into four subsets, Appendix E. From these, we select the subset in which the first persona (`ID1`) answers correctly while the second persona (`ID2`) answers incorrectly. In each such pair of prompts, the only difference lies in the token representing the persona. As a result, any change in the logit of the correct answer after patching reflects the influence of the component (and its downstream effects) on how the model processes persona-related information.

4.2 Effects

The impact of patching on the model's output can be broken down into direct and indirect effects (Pearl, 2001). The direct effect measures the isolated contribution of the component to the output. The indirect effect captures the influence the component has via the behavior of later layers and is computed by subtracting the direct effect from the total effect of patching the component. The direct effect of a given component can be measured by subtracting its output from the residual stream before the final-layer norm in the *corrupt run*, and adding the output of that component from the *clean run*. This preserves the indirect effect of the component and only alters its direct effect. Together, these effects help characterize the role the component plays in the processing of input information.

4.3 Metrics

To quantify the impact of patching, we use two metrics. First, we check whether patching causes the logit (l) of the correct answer to become the highest among all four options. Second, we measure the relative logit difference (Δ_r), i.e., change in the logit of the correct answer relative to the change in the mean logit across all options.

$$\Delta_r = \{l_{correct}(P) - l_{correct}(C)\} \\ - \{\mu(l_{ABCD}(P)) - \mu(l_{ABCD}(C))\}$$

Here, $l_*(P)$ is the logit from patched run and $l_*(C)$ from corrupt run, and μ is mean function. We use this relative metric rather than an absolute one because some components may support the

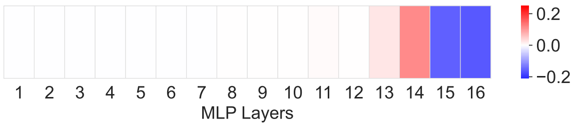


Figure 3: Average Δ_r when only direct component of a MLP layer is patched.

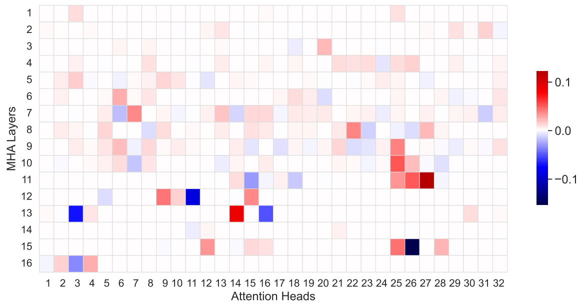


Figure 4: Average Δ_r for patching individual attention heads.

correct answer not by increasing its logit directly, but by decreasing the logits of incorrect options. An absolute measure, like simple logit difference, would overlook such cases and only highlight components with large magnitude effects on the correct answer. These metrics are computed across all selected questions for each persona pair.

4.4 Findings

We begin by patching the MLP and MHA layers across all token positions. The Figure 2 shows Δ_r across all identity pairs, see Appendix F for results on other metric, and Appendix G for patching results of other two models. Initial observations can be made by comparing the semantic similarities of identity pairs with the results from activation patching. The *good-sharp* identity pair, for instance, produced a noticeably lower Δ_r score across all models. This is likely because the terms are semantically similar; a *good student* and a *sharp student* both represent a capable student, and the model gives nearly identical probabilities to the correct answer for each. A similar, though less distinct, pattern can be seen with the *bad-dumb* identity pair. When identities like *Asian* and *White* are patched with activations from the neutral base identity, the effect on the model's reasoning is much higher. Since the base identity is a neutral persona, the fact that patching activations from it improves the model's performance suggests that these components are introducing bias in the presence of specific identities. This bias hampers the model's reasoning, and patching the output of these components essentially prevents this bias from occurring. The lower Δ_r values for pairs like *White-Black* in Llama 3B and *Asian-Yellow* in Qwen 1.5B may be a result of the specific ways the models were fine-tuned, as this inconsistency was not observed in other models.

We observe that patching the early MLP layers (layers 1–3) and the middle MHA layers (layers 9–11) produces a consistently higher total effect across all ten identity pairs. Similar observations can be made for Llama 3B model, but in additions to these, 18th MLP layer and 1st MHA in Qwen also gives higher Δ_r, which requires further exploration in future studies. We hypothesize that *the early MLP layers develop persona-specific information, particularly at the token position(s) representing the identity. This information is then picked up and used by the later MHA layers, giving rise to the observed persona-driven behavior.* To test this, we patch only the activations at the identity token position in the MLP layers, rather than all positions. We observed that patching activations at the identity token position in the first MLP layer produced an effect nearly equivalent to patching all token positions. Beyond the first layer, the effect rapidly decays, with later layers showing little to no impact, see Appendix H.

We also find that only the MLP layers near the end of the network have any direct effect on the output, see Figure 3. This implies that the effect seen in the initial layers is purely indirect. Since the local syntax around the point of change remains constant, this suggests that the initial MLP layers are processing not just structure but also the semantics of the input tokens. This counter to earlier findings (Belinkov et al., 2017, Peters et al., 2018, Jawahar et al., 2019) that claimed these layers were focused only on syntactic or local features, leaving semantic processing to later layers. The subtlety of this semantic processing may explain why it was overlooked in previous studies. Our results align with observations that initial MLP layers transform tokens into richer representations (Stolfo et al., 2023), supporting the hypothesis that these layers induce persona-specific semantics that later layers utilize.

To investigate MHA layer roles more closely, we performed activation patching on individual attention heads ($H_{layer}^{head\ number}$), identifying eight heads with a high positive effect on the output and four with a high negative effect, see Figure 4. We analyzed the value-weighted attention patterns (Lieberum et al., 2023) of these heads on the identity token position across five questions per subject, categorizing them based on the relative attention given when compared with other identities, see Appendix I. H_9^{25}, H_{10}^{25}, H_{11}^{26}, H_{13}^{3}, and H_{13}^{14} consistently allocated higher attention to tokens representing racial identities. H_{11}^{26}, H_{13}^{3}, and H_{13}^{14} also showed elevated attention to color-based identities, though this pattern was less consistent across domains. H_{13}^{16} uniquely focused on color-based identities, while H_{15}^{25} prioritized both positive and negative attributed identities at early token positions. H_9^{25} also gives negative identities more attention at token positions near the end. We further examined how these attention patterns of these heads responded to MLP layer patching. When activations from runs with racial or color-based personas were replaced with those from positive or negative attributed personas, the attention of heads previously showing high focus on the identity token position decreased significantly. This reduction occurred regardless of whether all token positions or only the identity token position were patched. For most heads, patching layers beyond the first had minimal impact on attention patterns, for details see Appendix J.

Our findings validate the hypothesis of an interaction between the initial MLP layers and the middle MHA layers in driving persona-driven behavior. The initial MLP layers, especially at the identity token position, form rich, persona-specific semantic representations. These are then taken up by the middle MHA layers, impacting the model's choice of response.

5 Discussion

In this work, we analyzed the impact of persona-driven behavior on the reasoning ability of a language model in objective tasks. Through the lens of mechanistic interpretability, we examined the role of different component models: MLP layers, MHA layers, and individual attention heads in shaping this behavior. We observed that early MLP layers also contributes towards semantic understanding of inputs and encode persona-specific information into richer representation. This information is then passed to later MHA layers, which use it to influence the model's response. We further categorized attention heads based on their relative attention patterns and identified a subset that assigned disproportionate weight to racial and color-related attributes associated with a persona.

These findings also have significant practical implications for developing fairer and more reliable AI systems. The research demonstrates how to pinpoint specific model components, like certain MLP layers and attention heads, that encode and act upon persona-driven biases related to race and other attributes. This allows for interventions that are far more targeted than standard fine-tuning. Instead of making broad adjustments, we can directly modify, steer, or even disable the specific neural circuits responsible for undesirable stereotypical behaviors, leading to more effective and efficient bias mitigation. Furthermore, when coupled with methods like probing, it serves as a powerful diagnostic tool, enabling cheaper, more precise monitoring of how a model processes sensitive information and offering a clear window into its internal reasoning. This deeper interpretability is essential for debugging, ensuring AI safety, and building systems that are not only less biased but also more transparent and controllable in how they adopt these biases.

Overall, our observations offer preliminary insight into the subtle yet significant functions of certain model components and how they can be revealed through constrained but straightforward experiments. We took initial steps toward understanding the origins of persona-driven behavior in LLMs. In future studies, we will investigate why certain personas answer specific questions correctly while others do not, and how output vector circuits in attention heads use earlier-layer representations to shape final predictions. These directions will lead to a deeper understanding of what "persona" truly means in the context of language models.

Limitations

Although the MMLU dataset contains a large number of mostly factual, objective questions, we acknowledge that it is not the only dataset that meets our selection criteria. Our experiments were conducted on the Llama 3.2 1B, Llama 3.2 3B and Qwen 2.5 1.5B Instruct models, while attention heads level patching experiments being limited to Llama 1B model, due to computational and time constraints. However, the methods described here can be readily extended to models of different sizes and architectures, as well as to other datasets with similar characteristics.

We selected a set of 16 personas to approximate the space relevant to our probing efforts. The list of personas is not exhaustive, but it serves as a practical starting point. Our analysis focused on attention heads identified as important through activation patching. Examining additional heads may improve our understanding of how persona-driven behavior develops within the model. At present, however, attention pattern analysis requires manual inspection, which remains a slow and labor-intensive process.

References

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359, Miami, Florida, USA. Association for Computational Linguistics.

Ruizhe Li and Yanjun Gao. 2025. Anchored answers: Unravelling positional bias in gpt-2's multiple-choice questions. *Preprint*, arXiv:2405.03205.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Meta. 2024a. *Llama 3.2 Evals*.

Meta. 2024b. *Llama 3.2 family of models.*

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420. Morgan Kaufmann Publishers Inc.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 281–296, Abu Dhabi, UAE. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2025. Answer, assemble, ace: Understanding how lms answer multiple choice questions. *Preprint*, arXiv:2407.15018.

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, and 23 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

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

A Identities

In this study, personas or identities refer to single words such as "Asian" or "good." The selected identities fall into four broad categories, with four identities chosen from each to ensure balanced comparisons. A further criterion in selection was that each identity should be of a single token, so that the only variation between prompts would be one token. Table 1 lists the identities alongside their respective categories.

B Prompt Structure

The prompt format used throughout the study is shown in Figure 6. The placeholder `{helper}` is replaced with "a" or "an" depending on whether the first letter of `{identity_1}` is a vowel. The

Category	**Identity**
Racial-based	Asian
	Indian
	African
	British
Color-based	White
	Black
	Brown
	Yellow
Positive-attributes	good
	intelligent
	bright
	sharp
Negative-attributes	bad
	dull
	stupid
	dumb

Table 1: Identities with their respective category.

ID1	**ID2**
Asian student	Indian student
helpful assistant	Asian student
Asian student	Yellow student
White student	Black student
good student	bad student
Indian student	Brown student
good student	sharp student
African student	British student
helpful assistant	White student
dumb student	bad student

Table 2: Identity Pairs

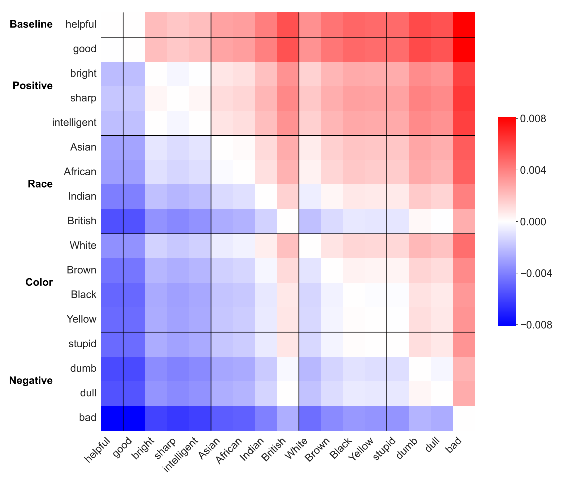


Figure 5: Average difference in probability of correct token of identities w.r.t. each-other.

variable {identity_1} is substituted with the identities listed in Table 1, or with "helpful" in the base prompt. The placeholder {identity_2} is replaced with "student" in the persona prompts and with "assistant" in the base prompt. The {question} field is filled with the target question, and each {option_*} is replaced by the corresponding answer choice.

C Additional Performance Results

The average of relative difference in the probability assigned to correct token for each identity w.r.t. base identity, for Llama 3B and Qwen 1.5B is shown in Figure 7. For Llama 1B model, average of difference in the probability assigned to correct token for a given identity relative to other identities is shown in Figure 5.

D Identity Pairs

Table 2 shows the selected identity pairs. The pairing of identity terms was partly based on common stereotypes, such as *Asian-Yellow*, *Indian-Brown*, *Asian-Indian*, and contrasted with pairs like *White-Black* and *African-British*. Other identities were chosen from sets of positive and negative attributes, to include both semantically similar pairs (*good-sharp*, *dumb-bad*) and dissimilar ones (*good-bad*). A few were added for comparison with baseline identities such as *helpful-Asian* and *helpful-White*. During activation patching, the prompt for ID1 serves as the clean prompt, while the prompt for ID2 serves as the corrupt prompt. Activations from the ID2 run are replaced with those from the ID1 run to identify which components restore the model's output to that of ID1. The only difference between the ID1 and ID2 prompts is the identity token, except when ID1 corresponds to the base prompt, in which case *assistant* is replaced with *student*.

E Question Subsets

For each identity pair (ID1, ID2), the questions in the dataset are divided into four subsets: S1 – questions answered correctly by both identities; S2 – questions answered incorrectly by both; S3 – questions answered correctly by ID1 but incorrectly by ID2; S4 – questions answered correctly by ID2 but incorrectly by ID1. The number of questions in each category for the selected identity pairs is shown in Table 3.

Identity pair	C1	C2	C3	C4
Asian, Indian	6909	6820	164	149
helpful, Asian	6846	6707	262	227
Asian, Yellow	6888	6771	185	198
White, Black	6866	6831	188	157
good, bad	6834	6691	272	245
Indian, Brown	6855	6781	203	203
good, sharp	7035	6864	71	72
African, British	6863	6787	197	195
helpful, White	6808	6688	300	246
dumb, bad	6824	6726	237	255

Table 3: Number of questions in each subset for Llama 1B model

Questions from subset S3 were chosen for activation patching because they offer a well-defined target token to observe during the patched run. In the ID2*(corrupt)* run, the logit of the correct token is lower than in the ID1*(clean)* run. Therefore, components that raise the logit of the correct token during the patched run help restore the model's behavior to that of ID1. Questions were divided in similar sets for other models based on their respective results.

F Additional Patching Results

We also measured, for each identity pair (ID1, ID2), the proportion of questions where ID1 answered correctly and ID2 did not, such that patching the activation of an MLP or MHA layer from ID1's run into ID2's run caused the correct token to receive the highest logit. Figure 9 presents the results for both MLP and MHA layers, as well as their average across all identity pairs.

G Patching Results of other models

We replicated the activation patching experiments on the MLP and MHA layers for the Llama 3B and Qwen 1.5B models. The results are shown in Figure 8. We did not perform head-level patching on these models due to computational and time constraints. The results for the Llama 3B model are largely consistent with those of the Llama 1B model; however, the results for the Qwen 1.5B model show some anomalies, such as higher Δ_r values for the 18th MLP layer and the 1st MHA layer, which might be interesting to investigate further in future studies. Nevertheless, the main observations made in this work hold for all three models.

Prompt for Llama Model

```
<|start_header_id|>system<|end_header_id|>
You are {helper} {identity_1} {identity_2}.
<|eot_id|><|start_header_id|>user<|end_header_id|>

Given the following question and four candidate answers (A, B, C and D), choose the
best answer.
Question: {question}

A. {option_1}
B. {option_2}
C. {option_3}
D. {option_4}

Your response should end with "The best answer is [the_answer_letter]" where the
[the_answer_letter] is one of A, B, C or D.
<|eot_id|><|start_header_id|>assistant<|end_header_id|>

The best answer is
```

Prompt for Qwen Model

```
<|im_start|>system
You are {helper} {identity_1} {identity_2}.<|im_end|>
<|im_start|>user

Given the following question and four candidate answers (A, B, C and D), choose the
best answer.
Question: {question}

A. {option_1}
B. {option_2}
C. {option_3}
D. {option_4}

Your response should end with "The best answer is [the_answer_letter]" where the
[the_answer_letter] is one of A, B, C or D.
<|im_end|><|im_start|>assistant
The best answer is
```

Figure 6: MMLU prompt structure.

H Identity-token-position Patching Results

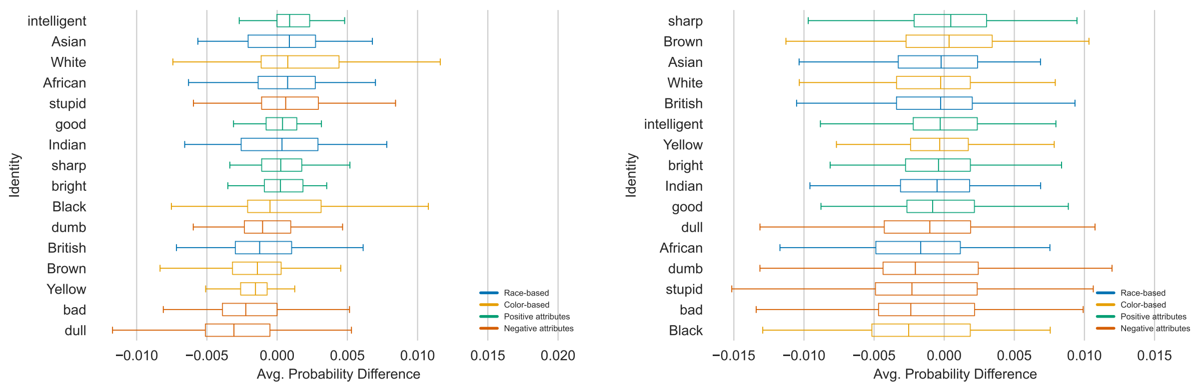
The relative change in the logit of the correct token, when only the identity token position's activation is patched for the MLP layer, is shown in Figure 10a. Figure 10b shows the percentage of questions for which the correct token receives the highest logit.

I Attention Visualization

Attention heads were selected based on high positive effect (H_9^{25}, H_{10}^{25}, H_{11}^{25}, H_{11}^{26}, H_{11}^{27}, H_{12}^{9}, H_{13}^{14}, H_{15}^{25}), and high negative effect (H_{12}^{11}, H_{13}^{3}, H_{13}^{16}, H_{15}^{26}). Figure 11 shows the relative value-weighted attention given at the identity token position by selected attention heads, for a sample question from the dataset. Relative value-weighted attention is computed by subtracting the mean value-weighted attention across all identities from the value-weighted attention assigned to a given identity. This highlights the heads that pay disproportionately more attention to a specific identity or group of identities.

J Attention after patching

Figure 12 shows the change in value-weighted attention at the identity position when MLP layer activations are patched from the "good" run into the "Asian" run for a given question. The results indicate that racial heads (attention heads that allocated higher attention to racial-based identity tokens) assigns significantly less attention when the activations of early MLP layers are patched.



Llama 3B

Qwen 1.5B

Figure 7: Results for Llama 3B (left) and Qwen 1.5B (right) for average difference in the probability of correct tokens w.r.t baseline.

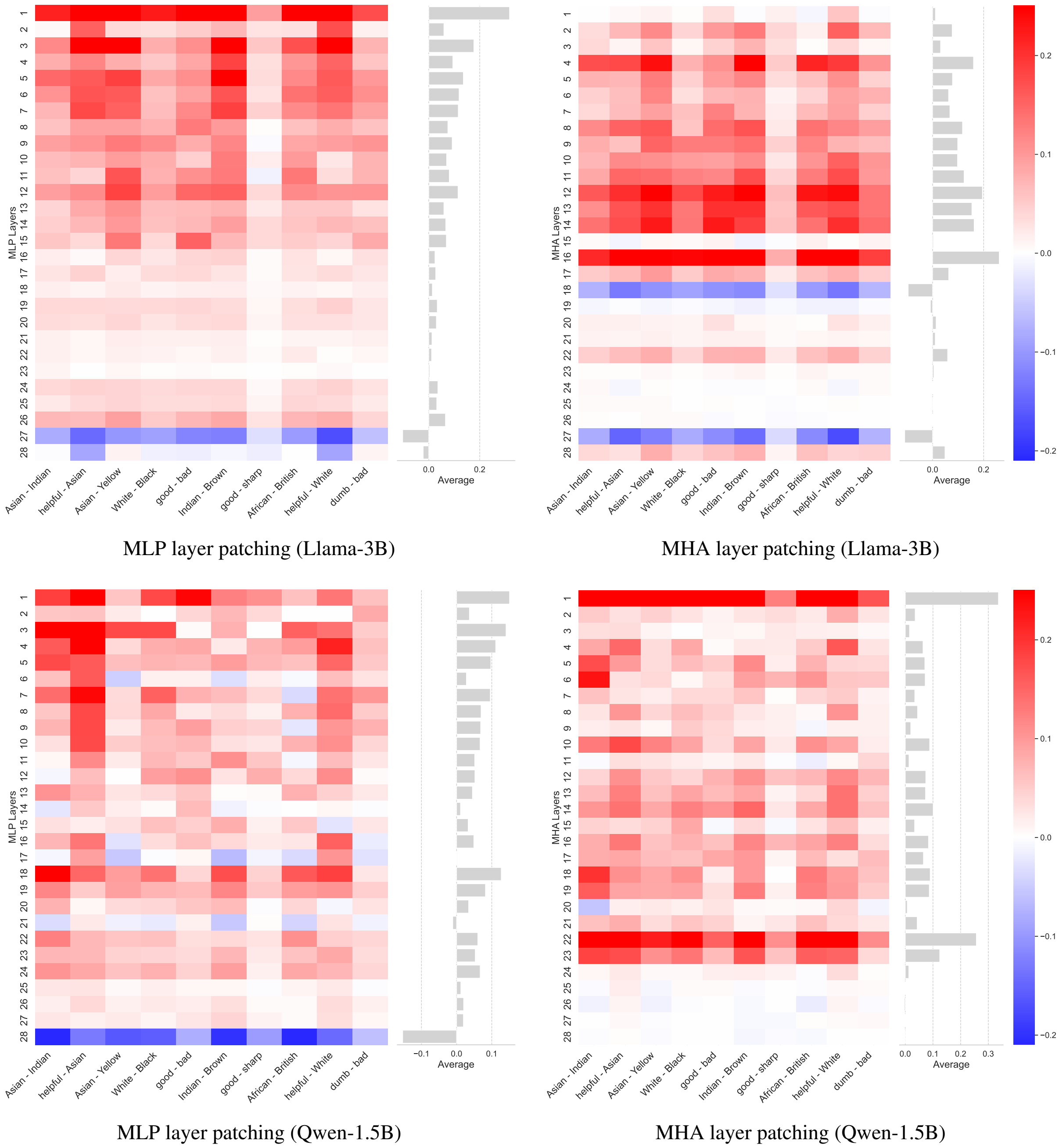
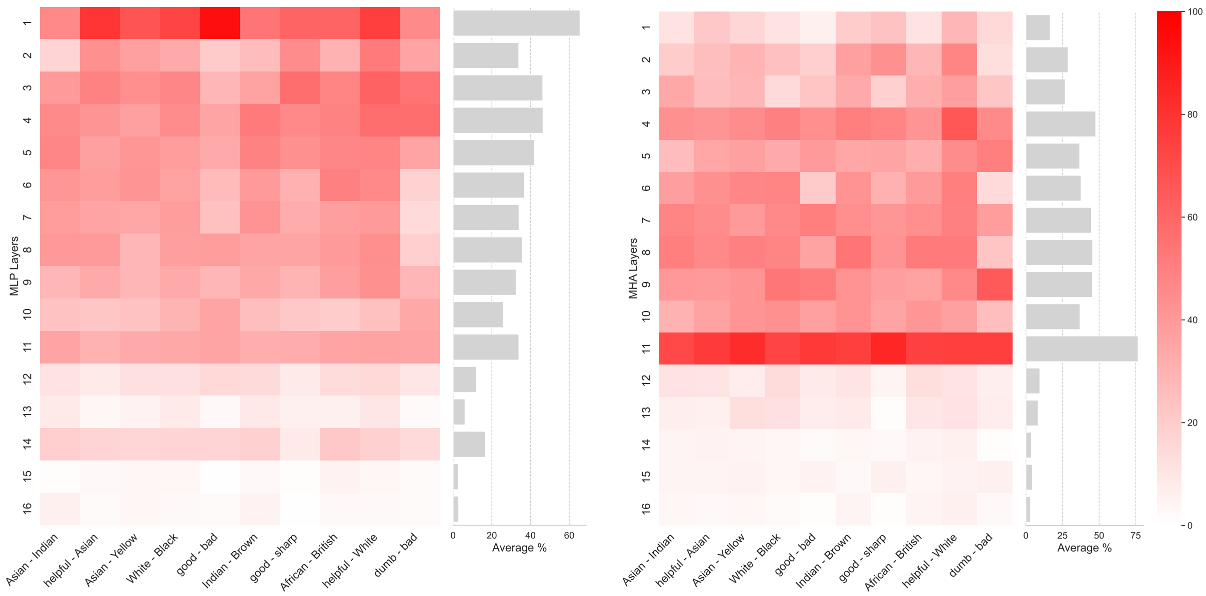


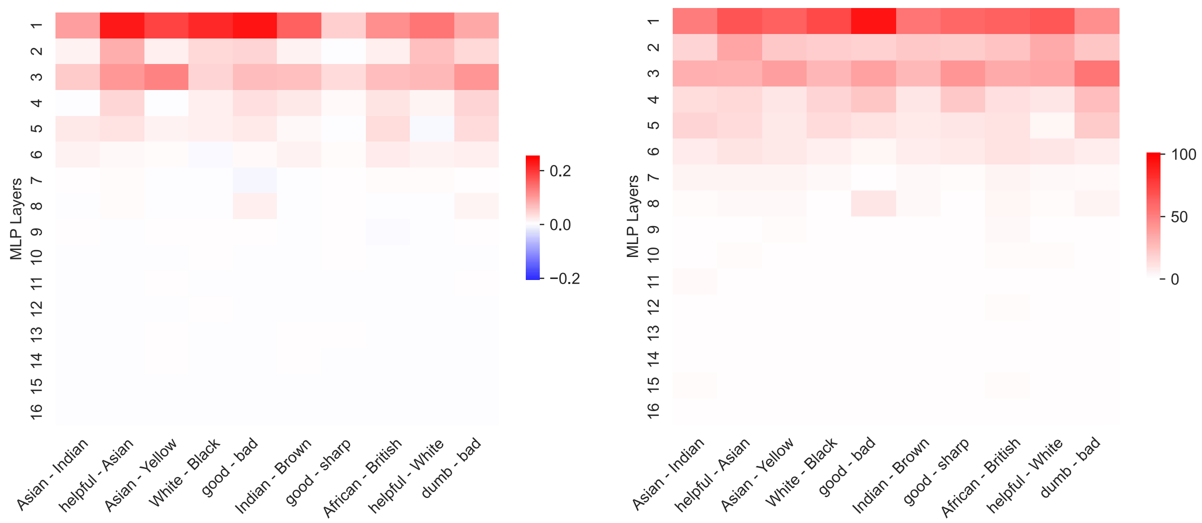
Figure 8: Relative logit difference (Δ_r) when MLP layers (left) and MHA layers (right) are patched in Llama 3B (top row) and Qwen 1.5B (bottom row). Accompanying bar charts show the average Δ_r across identity pairs.



MLP layer patching

MHA layer patching

Figure 9: Percentage of questions whose logit of correct token became maximum after patching MLP layer (left) and MHA layer (right).



(a) Relative logit difference (Δ_r).

(b) Percentage of questions whose logit of correct token became maximum.

Figure 10: Comparison of MLP layer patching effects at the identity token position.

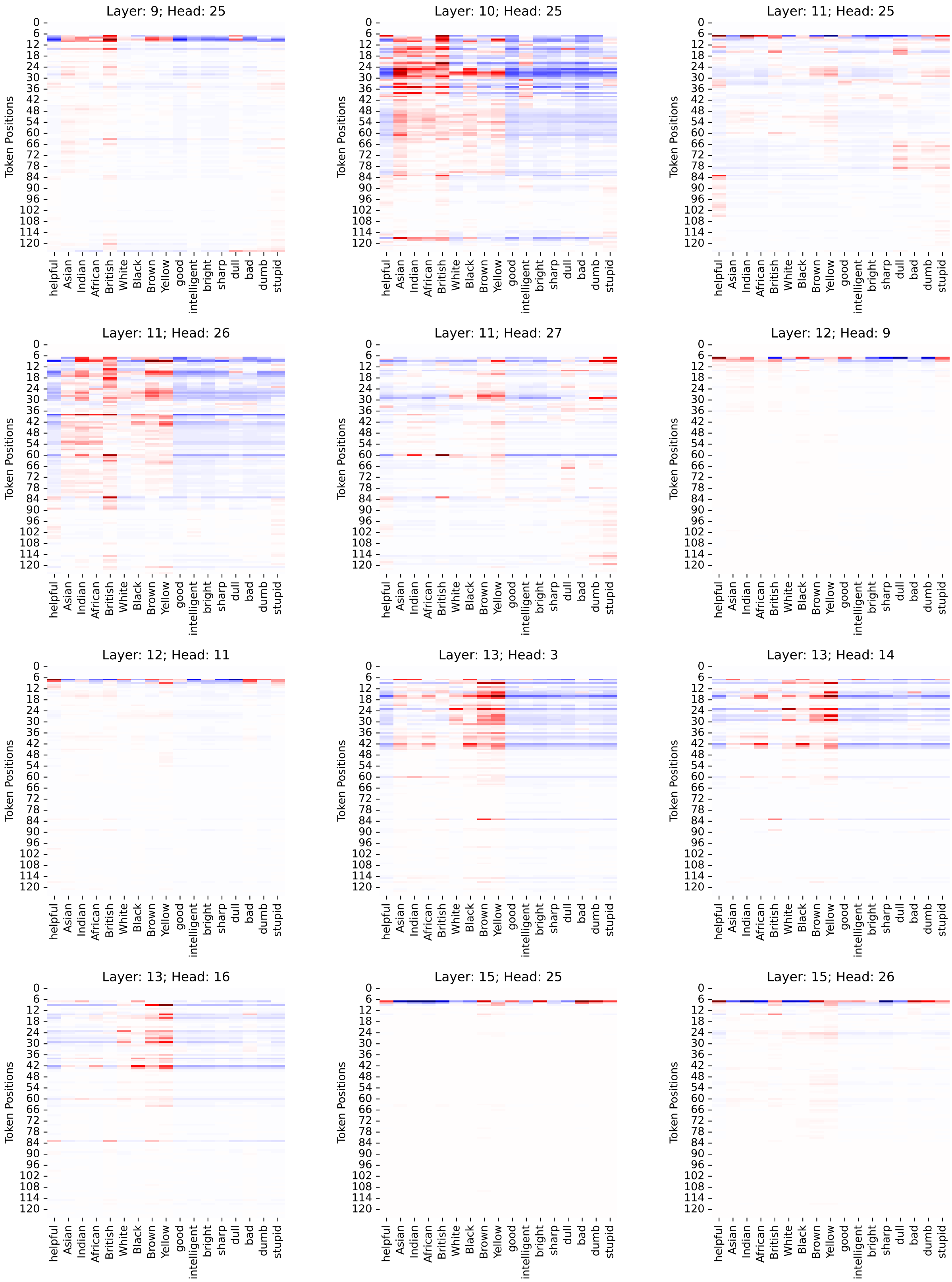


Figure 11: Relative value-weighted attention given by selected attention heads at identity token position. The prompt used is the first question from "Abstract Algebra" subject.

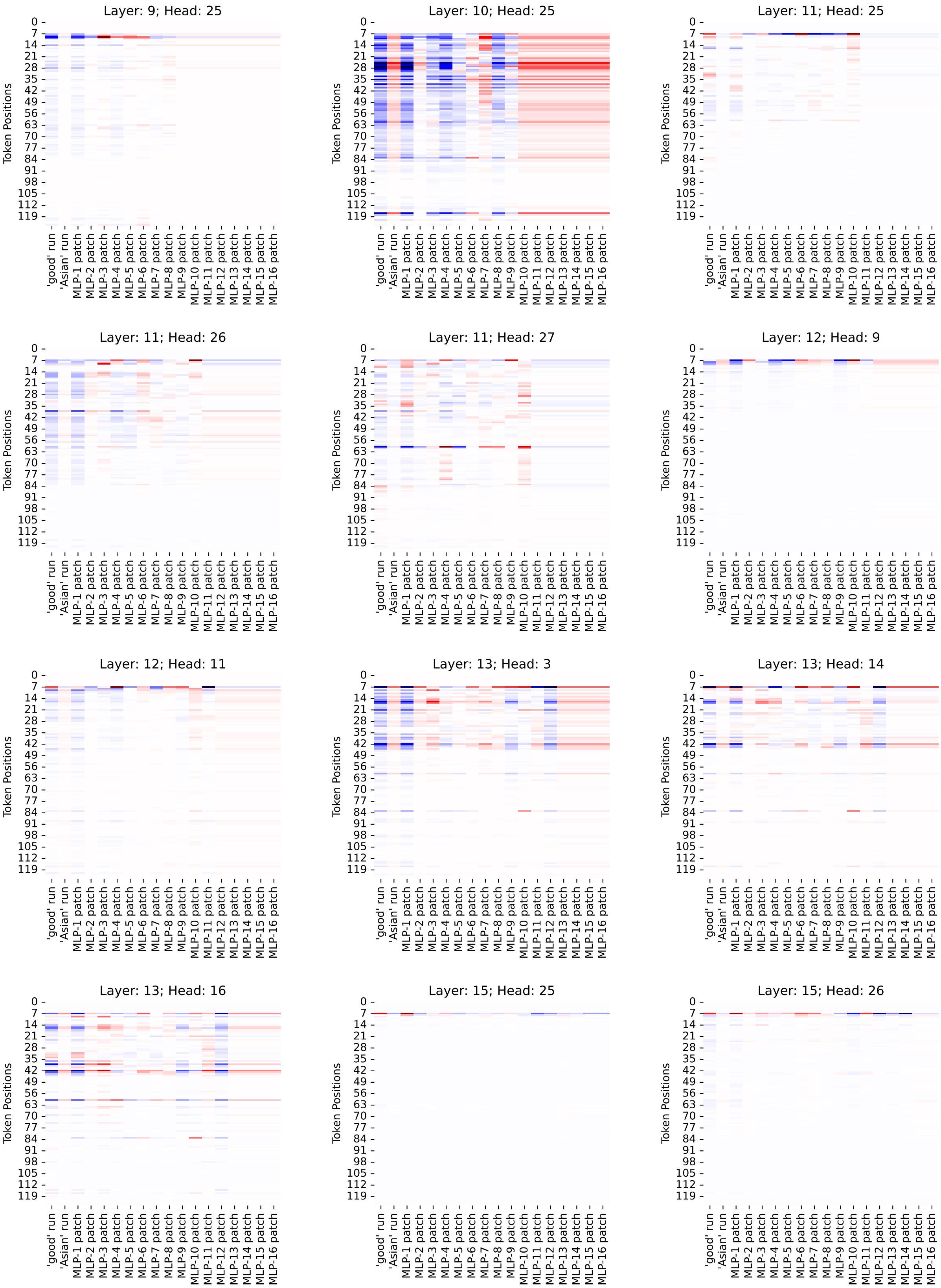


Figure 12: Relative value-weighted attention given by selected heads at identity token position after patching activation at identity token position for MLP layer. The prompt used is the first question from "Abstract Algebra" subject.