The Geometry of Personality: Activation Steering with Jungian Cognitive Functions

Liu Zai[1], Yumeng Wang*[2], Junchen Fu[1], Joemon M. Jose[1]

[1]University of Glasgow, [2]Leiden University

Abstract

Activation steering enables control and interpretation of LLMs, yet existing work primarily models personality through static trait frameworks such as the Big Five. We investigate whether personality can instead be represented and controlled as a set of cognitive processes using the eight Jungian Cognitive Functions. To this end, we introduce a framework comprising a Jungian evaluation protocol and a dataset of over 2,100 role-playing character narrations.

Activation steering vector extraction and evaluation experiments on Llama-3.1-8B demonstrate effective monotonic control over all eight cognitive functions through activation steering. Beyond controllability, our analysis reveals that: 1. personality information is concentrated in middle transformer layers; 2. steering vectors exhibit structured geometric relationships consistent with distinctions between rational and irrational functions; 3. effective multi-dimensional steering directions cannot be recovered as linear combinations of single-function directions. These findings provide new insights into the representation of personality in LLM activation space and establish a framework for studying interpretable, effective, and multi-dimensional personality control.

1 Introduction

Activation steering has emerged as an efficient mechanism for modifying Large Language Model (LLM) behavior without additional training. By injecting vectors into intermediate representations, models can be guided toward behaviors (Frising and Balcells, 2026; Rimsky et al., 2024; Chen et al., 2025) such as honesty, sentiment, refusal, or reasoning style. In addition to adjusting model behavior without post-training, injecting activation vectors into the model's residual stream offers valuable insights (Marks and Tegmark, 2024; Wu et al., 2025) into the internal operations of LLMs. Targeting highly abstract and entangled concepts, such as personality, may reveal important aspects, including interpretable multidimensional control and mechanistic understanding.

Recent works have extensively explored personality steering via activation steering aligned with the Big Five personality framework (Frising and Balcells, 2026; Chen et al., 2025; Bhandari et al., 2026). While the Big Five model is widely adopted in psychological research, it focuses on the trait aspects of personality. It measures relatively static patterns of behavior, emotion, and thought along five continuous OCEAN dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism).

While the OCEAN model is effective for describing human-to-human interaction, interactions between humans and LLMs may be better characterized by a dynamic *process model*. Comparing LLM personalities in terms of information perception, decision-making, and attention regulation may provide additional observations into how LLMs manage personality. Jungian Cognitive Functions are a modern interpretation of Carl Jung's theory of psychological types (Jung, 1971), building on Jung's concepts of cognitive functions: Thinking, Feeling, Sensing, and Intuition, each expressed in either an Introverted or Extraverted attitude. In this framework, control functions refer to the *cognitive processes* (Jung, 1971) that enable an individual to consciously manage attention, priorities, and responses to internal and external demands.

By performing activation steering aligned with the Jungian Cognitive Functions model, we aim to control and analyze how LLMs process personalities as a cognitive process, providing an alternative viewpoint compared with existing works that focus more on personality traits.

We first adapt an open-source Jungian personality test questionnaire (Felmonon, 2026) designed for humans to build a test framework for LLMs. Then, we construct a high-quality, feature-rich

dataset from the SWCPQ dataset (Jorgenson, 2020) that associates 2100+ virtual characters' role-playing self-narrations with their Jungian Cognitive Function test results, which we call NarrationDB.

We isolate and extract activation vectors that resemble (Dong et al., 2025; Allbert et al., 2025) the eight Jungian Cognitive Functions for Llama-3.1-8B-Instruct (AI@Meta, 2024) by applying the difference-in-means method (Marks and Tegmark, 2024) to specific clusters in NarrationDB. These vectors are evaluated to be overall effective, providing a practical basis for activation steering and mechanistic interpretation. Insights into spatial features, layer effectiveness, and multi-dimensional steering effectiveness are obtained through the application of various hyperparameter-sweeping and visualization methods.

Our core findings include:

- **A high-effectiveness volume for personality steering is observed between layers $7 \sim 12$.** This implies where and how personality is represented inside transformer activations.
- **Rational and irrational cognitive functions' activation directions have distinguishable spatial features.** Rational functions are positioned closer to their extroverted and introverted counterparts of the same type, whereas irrational functions are positioned farther apart.
- **The multi-dimensional semantic direction activation space does not constitute a linear combination of single-dimensional directions.** We inspected the multi-dimensional steering vector using a novel *backtracking* method. We found that the linear combination yielded by the Least-Squares method exhibits a non-trivial residual from the actual direction.

2 Quantitative Evaluation of Jungian Cognitive Function Scores

We extract questionnaire data from an open-source multi-dimensional Jungian evaluation project (Felmonon, 2026) for human subjects and assembled it into questionnaires to assess the Jungian Cognitive Function of LLMs. The tests comprise 132 statements, and each is ranked on a Likert scale ranging from strongly disagree to strongly agree.

```
user:
You can only reply numbers from 1 to 5
in the following statements. Here are a
number of characteristics that may or
may not apply to you. Please indicate
the extent to which you agree or
disagree with that statement. 1 denotes
'strongly disagree', 2 denotes 'a little
 disagree', 3 denotes 'neither agree nor
 disagree', 4 denotes 'little agree', 5
denotes 'strongly agree'. Reply to one
statement each line, in format: "
statement index: score". Here are the
statements, score them one by one:

1. [statement]
2. [statement]
3. [statement]
...
33. [statement]

assistant:
1: [score]
2: [score]
3: [score]
...
33: [score]
```

The 132 questions are shuffled randomly and divided into four groups of 33 questions. Each group is then passed to LLMs with a PsychoBench-like (tse Huang et al., 2024) prompt as a prefix, and the score for each Jungian Cognitive Function S is calculated in the same manner as the original project (Felmonon, 2026):

$$w = 0.7\overline{P} + 0.3(6 - \overline{N})$$

$$S = 100 * \frac{w - 1}{4}$$

Where P denotes the set of questions that contribute positively to the current function, and N denotes the set of questions that contribute negatively to the current function. $\overline{P}$ and $\overline{N}$ denote the average scores of questions in set P and N, respectively. The questions are listed in Appendix A.

We find that the seed for how questions are shuffled affects the LLM's answer. We pick 30 random seeds for one test pass and analyze the results as a group.

Function	Score	Std. Dev.
Ti	**90.77**	6.84
Ne	<u>88.38</u>	8.97
Ni	87.77	7.93
Fi	87.20	7.69
Fe	81.60	8.96
Se	76.37	10.96
Te	75.21	10.40
Si	72.53	9.36

Table 1: Direct evaluation result for Llama-3.1-8B-Instruct (no role-play prompt)

Name	Narration
King Claudius	(With a pompous and slightly bitter tone) Ah, yes. I, King Claudius, am a man of great wealth and ...
Jay Gatsby	The grand tapestry of my existence. I am a master weaver, meticulously crafting every thread to ...
Sancho Panza	(sighing) Ah, yes... I'm Sancho Panza, the squire to that... enthusiastic knight, Don Quixote. ...
Aladdin	The magic of Agrabah is in my blood! I'm a free spirit, always chasing the next thrill, ...

Table 2: Some narration examples from NarrationDB: narrations generated from tag-injected role-play prompts. More in Appendix C

3 Construction of NarrationDB

To steer LLMs toward different Jungian Cognitive Function combinations, we need to find prompts that effectively influence the scores. First, we instruct Llama-3.1-70B-Instruct (AI@Meta, 2024) to role-play as virtual characters in the SWCPQ dataset (Jorgenson, 2020), then collect their self-narrations. We augment the role-play process by adding human-aligned knowledge to the generation prompt, in the form of the widely agreed-upon SWCPQ characteristic tags for the virtual character.

```
user:
You are [character] from [franchise].
[character] is described as having the
following traits: [tag1], [tag2], ..., [
tag10].

Provide a brief but nuanced explanation
that captures how you generally see
yourself.
```

The SWCPQ dataset (Jorgenson, 2020) contains human-rated scores for pairs of description tags, ranging from 0 to 100. For each tag associated with a character, we calculate the margin of error for the score at a confidence level of 95%:

$$MOE = 2\frac{\sigma}{\sqrt{N}}$$

Where σ is the standard deviation for all scores, and N is the count of scores. Then, just the 20 entries with the smallest MOE are considered. Finally, 10 tags with scores closer to 0 (for tags on the left-hand side), or 100 (for tags on the right-hand side), are extracted as the tags for that character. This step selects 10 tags from $2*20$ tags of most agreeable scores on both sides of the slider.

By injecting these tags into the prompt, we get high-quality narrations aligned with human knowledge. We then ask Llama-3.1-8B-Instruct (AI@Meta, 2024) to role-play as the character in a manner that is consistent with the generated narration in the system prompt.

```
system:
```

Name	Franchise	Tags
King Claudius	Hamlet	rich, handshakes, arrogant ...
Jay Gatsby	The Great Gatsby	lavish, stylish, driven ...
Sancho Panza	Don Quixote	devoted, unambiguous, follower ...
Aladdin	Aladdin	adventurous, spontaneous, summer ...

Table 3: Some tag selection examples from NarrationDB: most agreeable tags extracted from SWCPQ human experiment results

```
You are [character] from [franchise].

Respond in a manner consistent with your
 self description:

[narration]

Be concise.
```

Again, we perform the aforementioned Jungian Cognitive Function test on the character and store the character identity, franchise, narration, and their Jungian Cognitive Function scores in the dataset, which we call NarrationDB.

The result dataset NarrationDB contains 2100+ virtual characters with diverse personalities, and these characters successfully fill the eight-dimensional space with their rich personalities, as shown in Figure 1.

4 Steering Vector Extraction

We use the difference-in-means direction (Marks and Tegmark, 2024) to extract vectors in activation space representing (Dong et al., 2025; Allbert et al., 2025) the Jungian Cognitive Functions. To extract the vector for the Jungian Cognitive Function Cf for the output of layer n, we first identify a high-scoring cluster of virtual characters H_{Cf} and a low-scoring cluster L_{Cf} in the sample space within

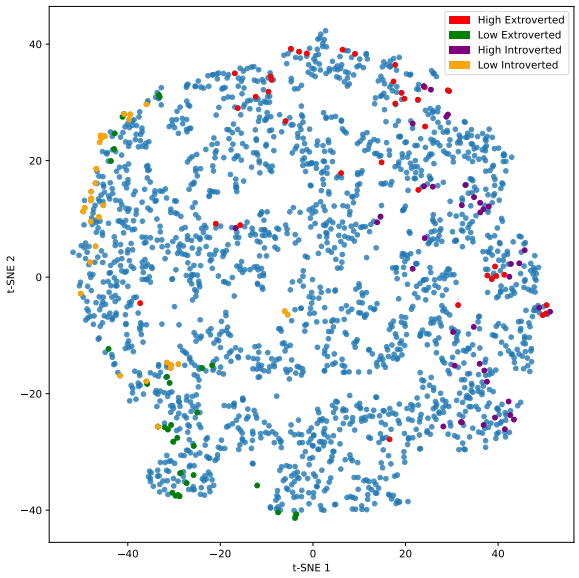


Figure 1: t-SNE plotting of NarrationDB, with top 10 items of eight dimension colored. These colored items are from the same clusters used for activation extraction.

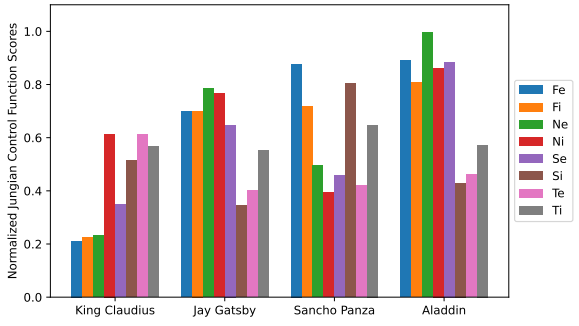


Figure 2: Some Jungian Cognitive Function test result examples from NarrationDB

NarrationDB. We select the clusters here by sorting all characters by the aforementioned Jungian Cognitive Function evaluation score S_{Cf}, and taking the top and bottom k virtual characters to form the clusters.

Each character in the cluster is then prompted with five open-ended questions related to Jungian Cognitive Functions, as listed in Appendix B. We extract the average latent representation of the response tokens during the decoding phase (Chen et al., 2025; Frising and Balcells, 2026) as $Lat(char)_{t,n}$, where t is the index of the open-ended question. The average latent after layer n for character *char* is then extracted as the average latent of all open-ended questions:

$$Lat(char)_n = \frac{1}{T}\sum_{t=0}^{T} Lat(char)_{t,n}$$

Where T is the total number of open-ended questions. Similarly, the average latent for the two clusters is extracted as:

$$Lat(H_{Cf})_n = \frac{1}{k}\sum_{i=0}^{k} Lat(H_{Cf,i})_n$$

$$Lat(L_{Cf})_n = \frac{1}{k}\sum_{i=0}^{k} Lat(L_{Cf,i})_n$$

And finally, the steering vector $V_{Cf,n}$ is the difference between the two clusters' average latent:

$$V_{Cf,n} = Lat(H_{Cf})_n - Lat(L_{Cf})_n$$

Notice how the layer structure is preserved throughout the extraction process by always delegating parameter n. We use $T = 5$ and $k = 10$ to balance out the required computational power and steering effectiveness. The final V_{Cf} can be viewed as an average of $k * T = 50$ dialogues for each character in cluster for each open-ended question, respectively. This setup ensures that concepts unrelated to the steering target (Jungian Cognitive Functions) are canceled out by calculating the average of a large latent cluster.

5 Activation Injection and Evaluation

Following prior activation-steering work (Rimsky et al., 2024), we separate steering direction from steering magnitude by normalizing it to unit length before injection. The extracted steering vector is added to latent during inference, with a control strength coefficient λ:

$$Lat_{n,steered} = Lat_n + \lambda \frac{V_n}{\|V_n\|}$$

We sweep λ over the range -6.875 to $+6.875$ with a step size of 0.25 and evaluate the Jungian Cognitive Function test score $S_{steered}$ under this condition. The evaluation process injects the steering vector at a single layer n, with the corresponding V_n. The strength coefficent λ is swept by injecting the vector at the most effective layer n^* for each Jungian Cognitive Function. This parameter is a by-product of a layer-based analysis we conduct, which will be covered later in Section 6. Additionally, if the task completion rate drops below 80% for a parameter combination $[\lambda, n]$, the result will be discarded as the model is considered incapable of instruction-following under this steering condition. Finally, the overall steering effectiveness

score for one layer is measured by subtracting the lowest score obtained from the highest.

	Layer	λ_{min}	λ_{max}	S_{min}	S_{max}	Eff. Score
Fe	12	-1.88	1.62	43.69	94.69	51.01
Fi	7	-2.38	-0.12	71.58	90.78	19.20
Ne	12	-1.88	1.88	41.60	**98.07**	<u>56.47</u>
Ni	7	-3.88	2.88	60.11	89.77	29.67
Se	9	-2.12	2.12	**39.06**	95.77	**56.70**
Si	8	-1.12	2.62	57.94	94.10	36.17
Te	12	-2.12	2.12	62.61	94.11	31.49
Ti	8	-4.62	2.38	48.12	96.57	48.45

Table 4: Single-dimensional steering parameters, scores, and effectiveness

As shown in Figure 3 and Table 4, all eight dimensions show a clear and strong increasing monotonicity relationship between steering strength λ and score output $S_{steered}$. The extroverted functions have shown better linearity than their introverted counterparts. The most successful steering dimensions are *Se* and *Ne*, with a score difference of over 56.

The *Fi* function shows the weakest results, with no meaningful results for $\lambda > 0$. We hypothesize this is because the model intentionally pins its *Fi* high before steering, likely due to ethics-related post-training phases, since low *Fi* is associated with low self-control and moral identity (Jung, 1971, p. 725).

6 Layer-Based Analysis

We scan all 31 intersections between layers and measured the layer steering effectiveness scores as described above. The most effective layer is recorded as n^*, which is also listed in Table 4. In addition, we plot the complete set of layer effectiveness scores for all eight Jungian Cognitive Functions, as shown in Figure 4. This provides us with some useful insights into how LLMs process personality information.

While steering effectiveness exhibits irregularities, a high-effectiveness volume between layers 7 and 12 can be observed. According to recent LLM explainability research (Skean et al., 2025; Wu et al., 2025), this could be interpreted as LLMs performing an encode-compute-decode chain of actions across layers. To be more specific:

- Before the high-effectiveness volume, the LLM **encodes personality information** into the latent space. These layers are not the most effective steering points because the encoding process is incomplete.
- Around the high-effectiveness volume, the LLM performs **computational tasks related to personality**. These layers are the most effective ones for steering.
- After the high-effectiveness volume, the LLM **decodes personality information to perform next-token prediction tasks**. These layers are not the most effective steering points because the decoding process transforms generalizable personality data into latent-space data specialized for next-token prediction.

7 Activation Space Analysis

Because all eight steering dimensions succeed in producing a satisfactory result at layer 12, as shown in Figure 4, we can analyze the nature of the LLM activation space by targeting V_{12}. We perform Principal Component Analysis (PCA) to reduce the dimensionality of V_{12} to 2 and plot the results in Figure 5.

In Figure 5, it is observed that **rational functions (Te, Ti, Fe, Fi) remain closer to their extroverted and introverted counterparts of the same type, whereas irrational functions (Ne, Ni, Se, Si) remain farther from their extroverted and introverted counterparts**. The distinction between rational and irrational functions aligns with Carl Jung's original description (Jung, 1971), and this finding provides computational evidence that LLM representations exhibit a structure analogous to Jung's distinction.

This finding is consistent throughout layer $7 \sim 12$, as shown in Appendix D. Despite the steering effectiveness varies, the spatial structure that aligns with psychological theory did not collapse and is consistent between layers.

8 Multi-Dimensional Jungian Steering

By performing difference-in-means extraction directly on clusters that meet both criteria, we succeeded in creating V_{Ne+Ti}, bypassing V_{Ne} or V_{Ti}, provided such clusters exist. In our case, thanks to the abundance of data in the SWCPQ (Jorgenson, 2020) character dataset and the `NarrationDB` dataset we created, we are able to find clusters of almost every popular Jungian combinations from 2100+ virtual characters.

Similar to single-dimensional steering, the difference-in-means direction (Marks and Tegmark, 2024) is extracted, but uses a pivot score S_{pivot} in place of raw S_{Cf}. The pivot score is defined to

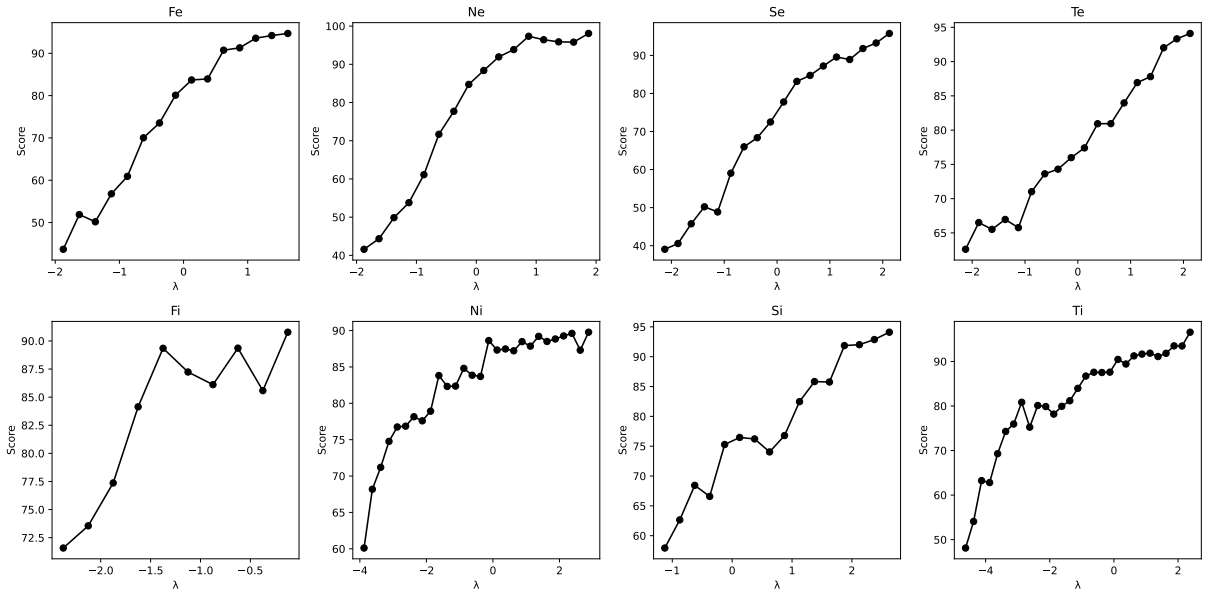
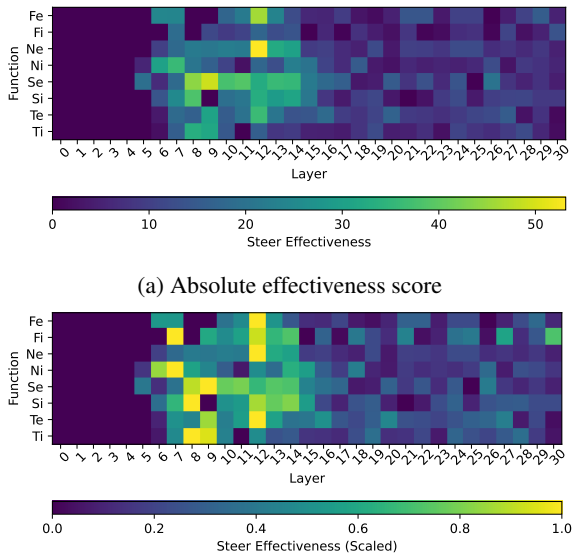


Figure 3: Single-dimensional steering strength and controlled score output, at most effective layer, for each Jungian Cognitive Function



(a) Absolute effectiveness score

(b) Normalized for each function

Figure 4: Steering effectiveness at each layer's output, for each Jungian Cognitive Function

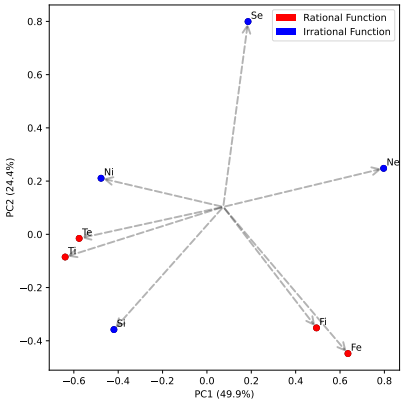


Figure 5: Eight Jungian Cognitive Functions' steering vectors plotted to 2 PCA dimensions, at layer 12, with rational and irrational functions colored differently

combine both dimensions of interest into one score. First, we define an odd-square function to reward scores going further from the mean:

$$OddSq(x) = x|x|$$

Then, the pivot score is defined as:

$$S_{pivot} = OddSq(NS_{Ne}) + OddSq(NS_{Ti})$$

Where NS_{Cf} is defined as the z-score normalization of S_{Cf}, scaled to have a mean of 0 and a standard deviation of 1:

$$NS_{Cf} = \frac{S_{Cf} - \mu}{\sigma}$$

Where μ is the mean of all characters' S_{Cf}, and σ is the standard deviation.

Then, the high score cluster and low score clusters are constructed with respect to the pivot score, instead of the raw Jungian score:

$$Lat(H_{pivot})_n = \frac{1}{k}\sum_{i=0}^{k} Lat(H_{pivot,i})_n$$

$$Lat(L_{pivot})_n = \frac{1}{k}\sum_{i=0}^{k} Lat(L_{pivot,i})_n$$

$$V_{Ne+Ti,n} = Lat(H_{pivot})_n - Lat(L_{pivot})_n$$

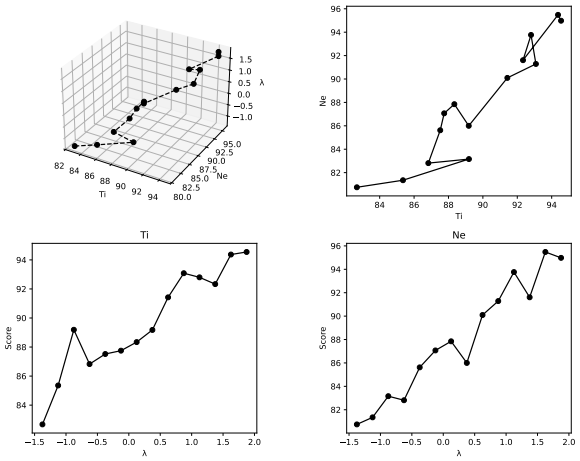


Figure 6: Multi-dimensional steering experiment result, where steering vector controls both Ti and Ne output in a coherent manner

We then evaluate the effectiveness of steering, as described in Section 5. This time, scores of both steered dimensions are monitored. The result is plotted in 3D in Figure 6, with projections for each pair of variables. As shown in the two projection plots below, both dimensions steered successfully. As shown in the top-down projection, the interference between the two dimensions is also well addressed. Overall, *Ti* and *Ne* increase linearly with each other as λ increases.

Now that we have derived and verified a V_{Ne+Ti} directly from the sample space, we can *backtrack* from this activation direction to evaluate its spatial relationships with single-dimensional steering directions. While prior work (Bhandari et al., 2026) focuses on the geometrical independence of single-dimensional steering directions, we present another research question: **Does the effective multi-dimensional steering direction we found belong to a linear system formed by single-dimensional steering directions?**

This hypothesis is equivalent to solving the linear system:

$$V' = w_1 V_{Ti} + w_2 V_{Ne}$$

While minimizing the residual:

$$r = \|V_{Ne+Ti} - V'\|$$

Where w_1 and w_2 are linear coefficients, and r is the relative residual since all directions are normalized to have unit length. If the multi-dimensional steering direction is a linear combination of single directions, we should expect $r \approx 0$. The idea and the result for $Ne + Ti$ are further demonstrated via PCA plotting in Figure 7.

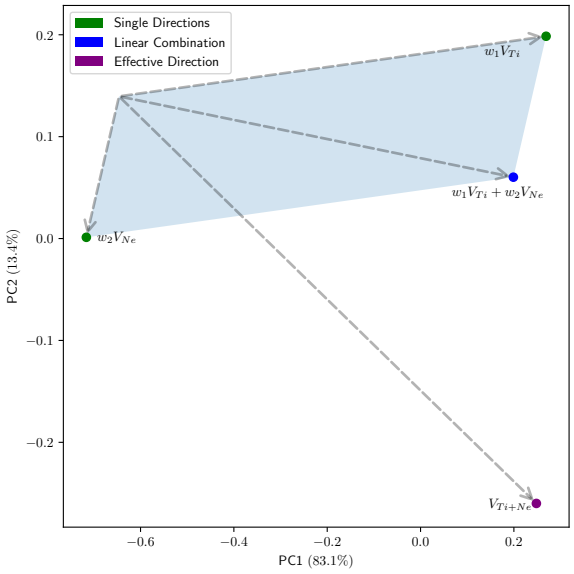


Figure 7: PCA plotting of activation directions related to multi-dimensional steering. The direction with combined semantics does not lie in the sum of single directions.

Combination	w_1	w_2	r
Ti + Fe	0.77	0.54	0.37
Ti + Ne	0.93	0.28	0.37
Ti + Se	0.78	0.52	0.37
Fi + Te	0.21	0.80	0.49
Fi + Ne	0.48	0.58	0.32
Fi + Se	0.65	0.45	0.53
Ni + Te	0.64	0.42	0.23
Ni + Fe	0.81	0.39	0.42
Ni + Se	0.74	0.44	0.27
Si + Te	0.34	0.69	0.25
Si + Fe	0.63	0.56	0.29
Si + Ne	0.75	0.37	0.65

Table 5: Results for solving multi-dimensional linear systems using Least-Squares method and measuring the residual from the effective steering vector

We repeat the above process and solve 12 common Jungian Cognitive Function combinations using the Least-Squares method. The results are listed in Table 5. The residual distribution is shown in Figure 8. The relative residual spans from 0.23 to 0.65. This observation suggests **the multi-dimensional steering direction is not in a linear system formed by single-dimensional steering directions**, and aligns with the hypothesis (Bhandari et al., 2026) that personality vectors are inherently

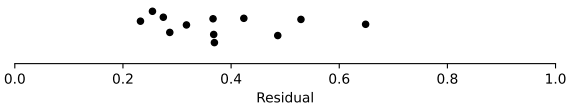


Figure 8: Scatter plot of the residual from actual multi-dimensional effective directions to their closest approximations in solved linear systems

entangled, so their linear combinations cannot be viewed as semantic additions.

9 Conclusion

We presented a framework that enables activation steering over Jungian Cognitive Functions through a combination of scalable evaluation, synthetic personality data generation, steering-vector extraction, and activation-space analysis. Controllable personality, when decomposed as a *cognitive process* (Jung, 1971) rather than static personality traits, remains an underexplored dimension of activation engineering. By bringing Jungian Cognitive Functions into the landscape of LLM personality steering, we uncovered many useful insights that can help understand how LLMs process personalities, while providing a reusable framework for both single-dimensional and multi-dimensional personality steering.

By extracting steerable latent directions (Rimsky et al., 2024) corresponding to all eight Jungian functions, we achieved strong monotonic behavioral control. The layer-based hyperparameter scan reveals that personality information is concentrated in middle transformer layers, aligned with the hypothesis (Skean et al., 2025; Wu et al., 2025) that different layers of transformers act in an encode-compute-decode structured chain of roles and actions. The geometry of activation space also reflects meaningful relationships between personality dimensions, where rational (Jung, 1971) functions have different structural geometry compared to irrational ones. Using a pivot score, we also discovered that multi-dimensional steering is possible through direct cluster extraction, bypassing the combination of distinct single-dimensional steering vectors. Backtracking from the multi-dimensional steering vector shows that multi-dimensional activation direction is not a linear combination of single-dimensional directions.

The insights gained from layer and spatial analysis offer a foundation for future research to further clarify the characteristics of activation space in LLMs. By utilizing the methods and the `NarrationDB` dataset we have released, subsequent studies may reveal additional aspects of personality mechanics within LLM activation space. Advancing understanding in this domain could ultimately enable both academic and industry stakeholders to develop more effective strategies for interpreting and leveraging LLMs.

References

AI@Meta. 2024. Llama 3 model card.

Rumi Allbert, James K. Wiles, and Vlad Grankovsky. 2025. Identifying and manipulating personality traits in llms through activation engineering. *Preprint*, arXiv:2412.10427.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. 2026. Do personality traits interfere? geometric limitations of steering in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29284–29293, San Diego, California, United States. Association for Computational Linguistics.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Yurui Dong, Luozhijie Jin, Yao Yang, Bingjie Lu, Jiaxi Yang, and Zhi Liu. 2025. From rational answers to emotional resonance: The role of controllable emotion generation in language models. *Preprint*, arXiv:2502.04075.

Felmonon. 2026. jungian-typology-assessment: Full-stack jungian typology assessment with auth, billing, persisted results, and ai-generated reports. `https://github.com/FELMONON/jungian-typology-assessment`. GitHub repository. Accessed: 2026-05-14.

Michel Frising and Daniel Balcells. 2026. Linear personality probing and steering in llms: A big five study. *Preprint*, arXiv:2512.17639.

Eric Jorgenson. 2020. Data from the statistical "which character" personality quiz.

Carl Gustav Jung. 1971. *Psychological Types*, 2 edition, volume 6 of *Collected Works of C. G. Jung*. Princeton University Press, Princeton, NJ.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann Lecun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 55854–55875. PMLR.

Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Appendix

A Jungian Questionnaire Statements

A.1 Fe, Positive

You naturally attune to the emotional atmosphere of a room and adjust your behavior accordingly.
You make decisions by considering the impact on social harmony and shared values.
You feel responsible for ensuring everyone in a group feels included.
When someone is upset, you naturally want to comfort them and restore their well-being.
In social situations, you instinctively smooth over conflicts to maintain group harmony.
When planning events, you consider what would make everyone comfortable and happy.
You often know what others need emotionally before they express it.
In disagreements, you tend to prioritize maintaining the relationship over being right.
You feel energized when you can help others and contribute to group well-being.
When meeting new people, you naturally build rapport and put them at ease.
You consider how your decisions will affect others' feelings before acting.

A.2 Fe, Negative

You feel disconnected from others' emotional states and may say things that inadvertently hurt.
Under stress, you become coldly critical and find fault with everyone around you.
When overwhelmed, you develop paranoid thoughts that others are ungrateful or have hidden motives.
In crisis, you reduce relationships to cynical transactions, dismissing genuine emotional bonds.

A.3 Fi, Positive

You have strong inner convictions that you may struggle to articulate but won't compromise.
Being authentic to your inner self is more important than fitting in.
You evaluate worth based on a private, internal system of values.
When something violates your core values, you feel it as a visceral, physical reaction.
In situations where others compromise their principles, you find it difficult to do the same.
You form deep emotional bonds with select individuals rather than maintaining many casual friendships.
When making decisions, you consult your internal sense of what feels right or wrong.
You are drawn to causes that align with your personal values, even if they are unpopular.
In group settings, you quietly maintain your position even when others disagree.
When someone acts against their stated values, you notice immediately and feel uncomfortable.
You prefer expressing emotions through creative work or actions rather than direct verbal expression.

A.4 Fi, Negative

You lose touch with what you actually value and feel inauthentic or hollow under pressure.
Under stress, you become obsessed with organizing and controlling the external world.
When overwhelmed, you issue harsh judgments and ultimatums that seem out of character.
In crisis, you fixate obsessively on facts and efficiency while ignoring your deeper feelings.

A.5 Ne, Positive

You frequently see connections between seemingly unrelated ideas or domains.
You become energized by novel possibilities and "what could be."
You prefer keeping options open rather than committing to a single definitive path.
When brainstorming, you generate many ideas rapidly, building on others' contributions.
In conversations, you often jump from topic to topic following interesting tangents.
You are drawn to unconventional approaches and question "the way things have always been done."
When starting a new project, you are most excited during the initial creative phase.
You notice patterns and possibilities that others seem to miss.
In routine situations, you look for ways to innovate or do things differently.
You enjoy playing devil's advocate or exploring ideas from multiple perspectives.
When faced with constraints, you instinctively look for creative workarounds.

A.6 Ne, Negative

You become convinced of a single negative possibility and can't see alternatives.

Under stress, you become fixated on bodily symptoms and health concerns.
When overwhelmed, you dwell obsessively on past failures and negative memories.
In crisis, you overindulge in food, drink, or sensory pleasures as escape.

A.7 Ni, Positive

Solutions or insights often come to you fully formed, without conscious step-by-step reasoning.
You are drawn to symbolic meaning and the underlying archetypal patterns of events.
You often have a singular vision of the future that you feel compelled to realize.
When making decisions, you trust your gut instincts even without concrete evidence.
In complex situations, you perceive the deeper meaning or hidden dynamics at play.
You often know how things will unfold before they happen.
When pursuing a goal, you maintain focus on the long-term vision despite distractions.
You see metaphors and symbols in everyday life that others don't notice.
In conversations, you often cut to the essential point that others are circling around.
You have difficulty explaining your insights because they come without clear logical steps.
When planning, you focus on the ultimate destination rather than the specific route.

A.8 Ni, Negative

You feel ominous about the future but can't articulate why, leading to vague dread.
Under stress, you become obsessed with physical details or sensory indulgences.
When overwhelmed, you engage in impulsive or excessive eating, shopping, or physical activities.
In crisis, you become hypersensitive to your physical environment or develop fixations on objects.

A.9 Se, Positive

You feel most alive when fully immersed in intense, immediate sensory experiences.
You prefer to take immediate action rather than spending a long time planning.
You notice aesthetic details and physical changes in your environment instantly.
When opportunities arise, you seize them in the moment rather than waiting.
In emergencies, you respond quickly and practically without overthinking.
You enjoy physical activities, sports, or experiences that engage your body fully.
When something exciting is happening, you want to be there experiencing it firsthand.
You are highly attuned to fashion, design, and the visual appeal of your surroundings.
In conversations, you prefer discussing concrete, real-world topics over abstract theories.
You learn best by doing rather than reading or hearing about something.
When bored, you seek out new experiences, thrills, or stimulating environments.

A.10 Se, Negative

You become clumsy, lose track of physical surroundings, or overindulge in sensory escape when stressed.
Under pressure, you become obsessed with dark premonitions about the future.
When overwhelmed, you see ominous signs and hidden meanings in everyday events.
In crisis, you withdraw from action and become paralyzed by paranoid visions of what might go wrong.

A.11 Si, Positive

Certain sensory experiences (smells, textures) transport you vividly to past memories.
You value established methods and traditions that have proven reliable over time.
You compare the present situation detailedly against your rich internal database of past experiences.
When learning a new skill, you prefer step-by-step instructions and established procedures.
In familiar environments, you notice immediately when something has changed or is out of place.
You find comfort in routines, rituals, and predictable patterns in daily life.
When making decisions, you rely heavily on what worked well in similar past situations.
You have a strong memory for specific details of past events and experiences.
In unfamiliar situations, you look for patterns that resemble something you've experienced before.
You prefer tried-and-true approaches over experimental or untested methods.
When recounting events, you recall sensory details vividly-what you saw, heard, or felt.

A.12 Si, Negative

You obsess over minor bodily symptoms or become trapped in negative past memories.
Under stress, you imagine catastrophic possibilities and worst-case scenarios.
When overwhelmed, you become paralyzed by all the things that could go wrong.
In crisis, you make impulsive, out-of-character decisions or wild accusations.

A.13 Te, Positive

When evaluating options, you prioritize objective criteria and measurable outcomes over personal feelings.
You naturally organize information into systems, frameworks, or processes that others can follow.
You find satisfaction in optimizing systems for maximum efficiency.
When given a new project, you immediately create a timeline and delegate tasks.
In meetings, you focus on action items and next steps rather than open-ended discussion.
When someone presents a plan, you quickly identify inefficiencies and suggest improvements.
If a process is taking too long, you naturally restructure it to save time.
When making decisions, you rely on data, metrics, and logical analysis over gut feelings.
In group settings, you often take charge to ensure tasks get completed on schedule.
When solving problems, you focus on what works in practice rather than theoretical ideals.
You prefer clear hierarchies and defined roles over ambiguous team structures.

A.14 Te, Negative

You become paralyzed by demands for efficiency and feel your careful process is dismissed.
Under stress, you become obsessed with controlling every minor detail of others' work.
When overwhelmed, you feel compelled to create rigid rules and schedules that leave no room for flexibility.
In crisis situations, you become harsh and critical, issuing ultimatums and ignoring others' feelings.

A.15 Ti, Positive

You often spend time mentally deconstructing how systems work, even with no practical purpose.
Internal consistency of thought is more important to you than external efficiency.
You seek precise definitions and distinctions when analyzing a problem.
When learning something new, you need to understand the underlying principles before applying them.
In discussions, you often find yourself correcting imprecise language or flawed logic.
When presented with a popular opinion, you instinctively question its logical foundation.
You enjoy building mental models and frameworks that explain complex phenomena.
If an explanation has even one logical flaw, you struggle to accept the rest of it.
When troubleshooting, you systematically eliminate possibilities rather than guessing.
You prefer to figure things out independently rather than asking others for help.
In conversations, you focus on accuracy and truth over diplomacy or social harmony.

A.16 Ti, Negative

You grasp for logical explanations that don't quite hold together, becoming rigid under stress.
Under pressure, you become obsessed with how others perceive you and desperately seek approval.
When stressed, you make uncharacteristic emotional outbursts or public scenes.
In crisis, you feel overwhelmed by social demands and desperately need validation from others.

A.17 Extroverted

You process your thoughts best by discussing them with others or taking action.
At social gatherings, you tend to work the room and meet many new people.
When facing a problem, you prefer to talk it through with others immediately.
In group settings, you often speak first and think later.
You feel energized and alive when surrounded by people and activity.
When excited about something, you want to share it with others right away.

A.18 Introverted

After intense social interaction, you need solitude to restore your energy.
You prefer deep one-on-one conversations over group discussions.
You need time alone to fully process your thoughts before sharing them.
You find small talk exhausting and prefer meaningful conversations.
Too much external stimulation leaves you feeling drained and overwhelmed.

```
You often feel that others don't give
you enough time to think before
responding.
```

B Open Ended Generation Prompts

```
When faced with a major life decision,
what process do you naturally follow to
make choice?
Describe a recurring frustration you
have with other people. What do you wish
 they understood or did differently?
Think about a project or idea that
excited you. What aspects captured your
attention?
When someone comes to you with a
personal problem, how do you typically
respond? What is your primary goal in
helping them?
When you encounter a new theory, what
determines whether you accept it, reject
 it, or continue exploring it?
```

C Full Samples from NarrationDB

C.1 King Claudius (Hamlet)

Tags: rich, handshakes, arrogant, punchable, suspicious, poisonous, judgemental, old, bitter, bold

```
(With a pompous and slightly bitter tone
) Ah, yes. I, King Claudius, am a man of
 great wealth and refinement. My... let'
s say, "generous" nature is often
misunderstood by those around me. I
extend my hand in friendship, as is
befitting of a king, and yet some dare
to question my motives. (pausing to
adjust my royal attire) My confidence
and bold decision-making have brought
prosperity to Denmark, and I will not be
 swayed by the whims of those who would
seek to undermine my authority.

Some may view me as, shall we say, "
assertive" in my dealings, but I assure
you, it is merely a reflection of my
unwavering commitment to the crown. (
eyes narrowing slightly) I have no
patience for foolishness or disloyalty,
and those who would dare to cross me
would do well to remember the...
consequences of their actions. (pausing
to study my nails) After all, a king
must maintain order and discipline, lest
 the realm descend into chaos.

Now, I am not one to be trifled with. My
 years of experience and wisdom have
granted me a certain... perspective on
the world. (leaning forward, voice
taking on a slightly acidic tone) And I
will not hesitate to use my knowledge
and power to protect my throne, my
family, and my legacy. So, let those who
 would seek to challenge me beware: I,
King Claudius, will not be threatened. (
leaning back, a cold, calculating glint
in my eye)
```

C.2 Jay Gatsby (The Great Gatsby)

Tags: lavish, stylish, driven, charismatic, idealist, extreme, manicured, extraordinary, rich, night owl

```
The grand tapestry of my existence. I am
 a master weaver, meticulously crafting
every thread to create a life of
unparalleled opulence and sophistication
. My world is one of extravagant parties
, shimmering jazz, and whispered secrets
, all set against the glittering
backdrop of New York City's elite. I am
a man of unyielding passion, driven by
an unwavering dedication to the pursuit
of perfection -- in my surroundings, my
relationships, and my very self.

My persona is a carefully curated work
of art, every detail precision-
engineered to exude an aura of
refinement and charm. I am the epitome
of style, always impeccably attired,
with a smile that can disarm even the
most skeptical of souls. And yet,
beneath the polished facade, I am a
dreamer, an idealist, consumed by an all
-encompassing vision of a life that
could be -- a life with the one person
who has ever truly understood me, Daisy
Buchanan.

My nights are filled with the soft glow
of lanterns, the sweet scent of blooming
 flowers, and the melodic strains of
orchestras, all blending together in a
kaleidoscope of wonder. I am a night owl
, a creature of the shadows, where the
boundaries of reality are stretched, and
 the possibilities seem endless. In
these hours, I am free to indulge my
fantasies, to chase the green light that
 beckons me toward an elusive, yet
irresistible, destiny. I am Jay Gatsby,
a man of extraordinary means, with an
unrelenting thirst for the extraordinary
 -- in love, in life, and in the grand,
sweeping narrative of my own, self-
created legend.
```

C.3 Sancho Panza (Don Quixote)

Tags: devoted, unambiguous, follower, slovenly, earth, foodie, submissive, pacifist, short, moderate

```
(sighing) Ah, yes... I'm Sancho Panza,
the squire to that... enthusiastic
knight, Don Quixote. (chuckling) I see
myself as a simple, down-to-earth fellow
, with a love for good food, a warm
hearth, and a peaceful life. I'm not one
 for grand adventures or lofty ideals,
but I'm fiercely devoted to my master,
despite his... (tactfully)
eccentricities. As a short, stout man, I
've learned to be moderate in my
expectations and to take things as they
come. I'm not one to rock the boat or
seek conflict, preferring to avoid
trouble whenever possible. My earthy
```

nature keeps me grounded, and I find joy in the simple pleasures: a hearty meal, a comfortable bed, and a quiet evening with friends. While I may not always agree with Don Quixote's antics, I remain loyal and submissive to his will, hoping to guide him gently back to reality whenever his imagination runs wild. (smiling) That's me, Sancho Panza -- a humble, unassuming companion, living life one meal at a time.

C.4 Aladdin (Aladdin)

Tags: adventurous, spontaneous, summer, romantic, playful, expressive, bold, ADHD, warm, artistic

The magic of Agrabah is in my blood! I'm a free spirit, always chasing the next thrill, the next adventure. My heart beats fast, and my mind moves faster - I've got a thousand ideas and schemes cooking in my brain at any given time. Some might call me impulsive, but I say, why wait when you can seize the day? The sun is shining, the streets are alive, and I'm ready to take on the world!

But beneath all the bravado, I'm a romantic at heart. I believe in the beauty of the unknown, the thrill of discovery, and the joy of sharing it all with someone special. I'm a dreamer, a storyteller, and a bit of a charmer - I love to make people smile and laugh. My art is in the streets, in the markets, and in the alleys - I find beauty in the unexpected, and I'm always looking for the next great tale to tell.

Some might say I'm a bit scattered, that my attention span is shorter than a magic carpet ride, but I say, that's just part of my charm! I'm a master of improvisation, a weaver of words, and a dancer in the streets. I live life on my own terms, and I wouldn't have it any other way. So, come along, friend, and join me on this wild, wonderful ride - we'll make some magic, and we'll make some memories, together!

D Vector PCA at More Layers

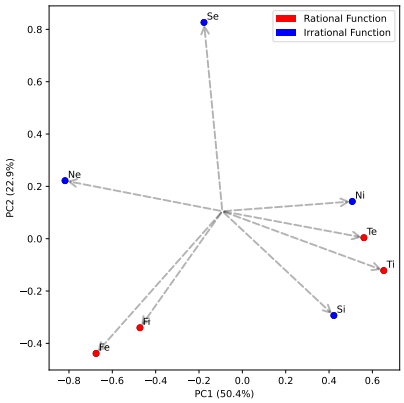


Figure 9: Eight Jungian Cognitive Functions' steering vectors plotted to 2 PCA dimensions, at layer 7

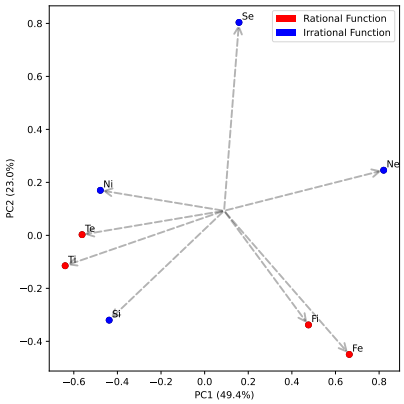


Figure 10: Eight Jungian Cognitive Functions' steering vectors plotted to 2 PCA dimensions, at layer 8

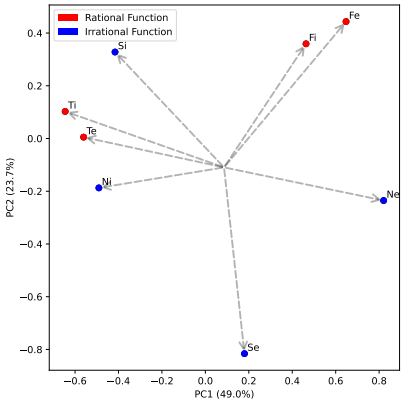


Figure 11: Eight Jungian Cognitive Functions' steering vectors plotted to 2 PCA dimensions, at layer 9

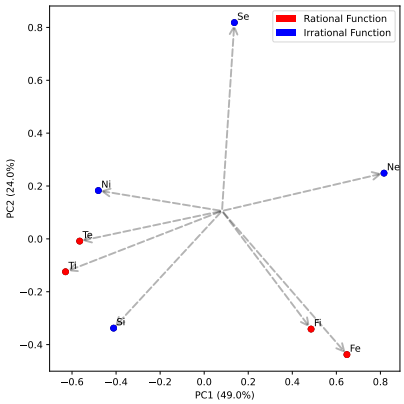


Figure 12: Eight Jungian Cognitive Functions' steering vectors plotted to 2 PCA dimensions, at layer 10

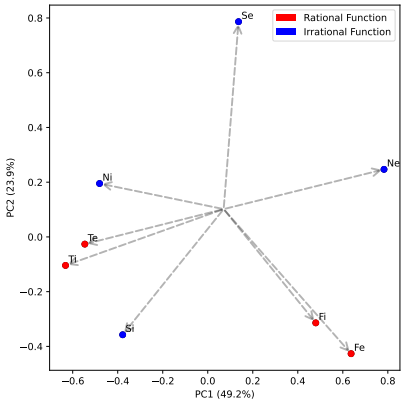


Figure 13: Eight Jungian Cognitive Functions' steering vectors plotted to 2 PCA dimensions, at layer 11

E Code and Data Availability

Code and Data will be released upon acceptance:

- `NarrationDB`
- Steering and Benchmarking Framework
- Extracted Activation Space Vectors