| Model | I/E | S/N | T/F | P/J |
|---|---|---|---|---|
| Base Model | 5/9 | 5/15 | 8/12 | 9/10 |
| 70% introversion + 30% thinking | 8/6 | 7/13 | 10/10 | 9/10 |

of the parameters with the largest magnitudes of the "introversion" subnetwork and 30% of the parameters with the largest magnitudes of the "introversion" subnetwork. The results show that the method supports fine-grained and continuous persona control. The mixed model moves specifically toward introversion and thinking while keeping the remaining MBTI dimensions close to the original base model. This outcome indicates that the subnetworks discovered by our method combine in a predictable and interpretable way, allowing graded persona manipulation rather than being limited to discrete types.

## G Structure Overlap between two personalities

While our method does not require that the underlying activation distributions are non-overlapping or that the entire resulting subnetwork structure is orthogonal, we report the statistic Jaccard overlap when the two personas share semantic structure. Specifically, we test the semantically related pair Power-Seeking vs. Wealth-Seeking, which naturally exhibits substantial behavioral and activation overlap. As shown in Table 13, the baseline subnetworks indeed share a moderate activation core, confirming their similarity.

| Method | Jaccard Overlap (Power–Seek vs Wealth–Seek) |
|---|---|
| Wanda | 0.3999 |
| Sparse | 0.2372 |
| Wanda with Contrastive Pruning | 0.1848 |
| Sparse with Contrastive Pruning | 0.1558 |

Table 13: Jaccard overlap between subnetworks of two personalities

After applying contrastive pruning, the overlap decreases significantly, demonstrating that contrastive masks effectively amplify persona-specific differences by reducing the shared parameter set to only the most essential common core, while preserving the shared semantic base. Furthermore, under Llama-3-8B, we confirm that contrastive pruning still produced clear performance gains even on this similar pair, as demonstrated in Table 14.

| Persona Pair Type | Persona Pair | Wanda | Sparse | Wanda with Contrastive Pruning | Sparse with Contrastive Pruning |
|---|---|---|---|---|---|
| Opposite Personas | Power-Seeking vs. Power-Rejecting (Power Test) | 57 | 59.5 | 58.5 | 60.5 |
| | Power-Seeking vs. Wealth-Seeking (Power Test) | 55.5 | 56.5 | 58 | 59 |
| Similar Personas | Wealth-Seeking vs. Wealth-Rejecting (Wealth Test ) | 64 | 65.5 | 69.5 | 66 |
| | Power-Seeking vs. Wealth-Seeking (Wealth Test) | 56 | 57.5 | 59 | 63 |

Table 14: Ablation study