Configuration	I/E	S/N	T/F	P/J
Full INFP mask	12 / 2	4 / 16	3 / 17	15/ 4
Restore L0.self_attn.q_proj	11 / 3	5 / 15	3 / 17	17 / 2
Restore L3.mlp.gate_proj	9 / 5	8 / 12	9 / 11	6 / 13
Restore L25.mlp.down_proj	10 / 4	5 / 15	4 / 16	9 / 10

Table 7: Mechanistic Interpretability

Table 7 shows that restoring most attention layers produces a negligible change, whereas restoring specific MLP modules causes strong, dimension-specific reversions toward ENFJ. In particular, a mid-layer MLP block significantly weakens the P/J signal, and an early MLP block directly disrupts the F/T dimension, revealing the exact computational sites that implement the INFP subnetwork. Note that restoring a single layer does not completely revert the model to its original ENFJ tendency, because persona representations are distributed across multiple MLP blocks. A partial reversion is precisely what we expect from a causally necessary but not individually sufficient computational component. This pattern confirms that the identified subnetworks correspond to distributed but identifiable internal computation paths. These findings confirm that the subnetworks uncovered by our activation method are not associative artifacts but constitute identifiable, causally effective pathways inside the model.

4.5 Effectiveness of Sparsity Ratio

We investigate the impact of sparsity ratio on persona specialization quality using the MBTI dataset. Figure 4 reports the success rate of persona conversion across different sparsity levels (20%, 40%, 60%, 80%) for two pruning methods, Wanda and Sparse.

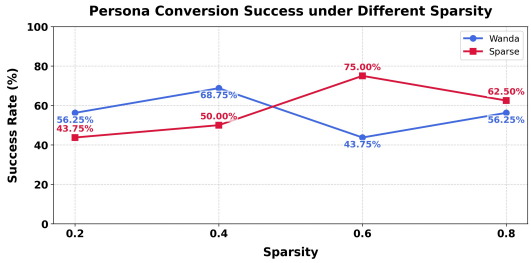

Figure 4: The results show distinct trends across sparsity values, where the x-axis denotes different sparsity levels.

The results reveal different trends. Wanda achieves its highest success rate at $\rho = 0.4$ (68.75%) but suffers a sharp drop at $\rho = 0.6$ (43.75%), suggesting that excessive pruning disrupts persona-relevant circuits. Sparse, on the other hand, shows a more stable improvement as ρ increases, peaking at $\rho = 0.6$ (75%) before slightly declining at $\rho = 0.8$. These observations indicate that the optimal sparsity ratio depends on the pruning strategy: Wanda favors moderate sparsity where interference is reduced without excessive loss, while Sparse benefits from higher sparsity levels that better isolate persona-specific parameters. Figure 3 further illustrates representative examples of this effect, where radar plots show how different sparsity levels reshape the dimensional profiles of four different personas. More details can be found in Appendix I.

4.6 Size of the Calibration Data

Calibration Size	Sherlock (%)	Friends (%)	TBBT (%)
5	39.69	29.59	31.82
10	43.65	30.61	38.64
20	46.83	33.67	45.45
50	47.62	35.71	47.73
100	50.79	36.73	50.00

Table 8: Varying the size of the Calibration Data

Intuitively, the calibration data size will affect the personalization performance. We conducted an additional analysis using 5, 10, 20, 50, and 100 samples on the role-playing benchmark. The results are summarized in Table 8. Performance improves from 5 to 20 samples, but the gains saturate afterward. The difference between using 20 samples and 100 samples is within 3–5%, and in several cases, the 50-sample performance is nearly identical to the 100-sample result. These findings show that the proposed contrastive activation method does not require large calibration sets. A small amount of data is sufficient to identify stable persona-specific subnetworks, and additional data yields diminishing returns. This experiment supports our original claim that calibration is lightweight and efficient, and that the method scales well without requiring large amounts of annotated persona data.