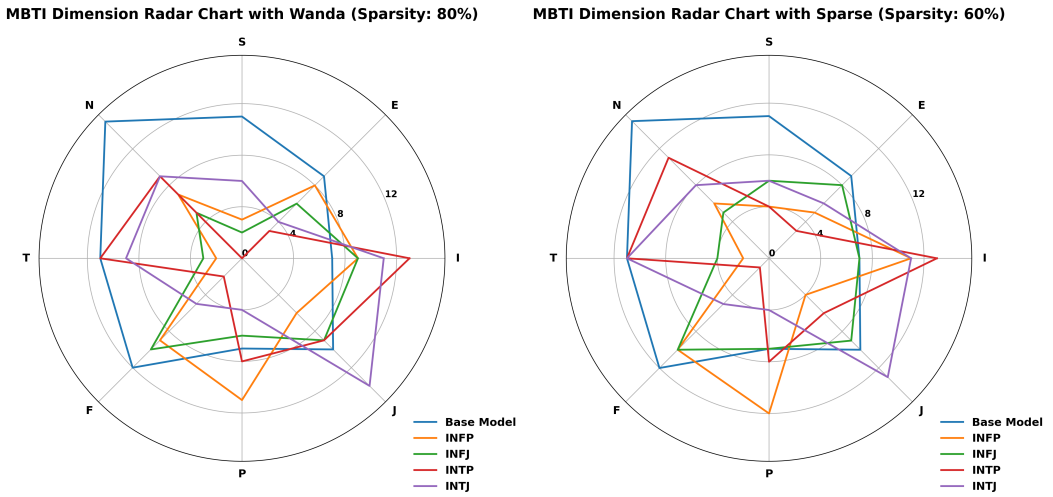

Figure 3: Radar plots of MBTI task under different sparsity ratios for INFP, INFJ, INTP, and INTJ.

consistent improvements highlight the effectiveness of incorporating contrastive objectives in enhancing persona alignment across tasks. We also observe similar benefits on the hallucination task, where contrastive pruning again surpasses baselines by a large margin. **Role-Playing Consistency** Results on RoleAgentBench (Table 5) show that pruning-based methods substantially improve role adoption and maintenance compared to prompting and RAG baselines. On Llama-2-13B, Sparse improves accuracy by 8–12 points over prompt, with notable gains on *Friends* (18.37 $\rightarrow$ 41.84) and *Sherlock* (42.11 $\rightarrow$ 55.26), while Llama-3-8B further amplifies performance, reaching 70.83% on *Merchant of Venice* and maintaining above 50% across all domains. These improvements indicate that pruned subnetworks not only adopt target roles but also sustain coherent role behavior in interactive settings, in contrast to baselines that often drift toward generic outputs. Moreover, Sparse consistently outperforms Wanda, suggesting that structured sparsification better disentangles overlapping role representations and provides more reliable persona alignment.

4.3 Affect of Pruning on General Performance

To evaluate whether pruning harms general modeling capabilities, we conducted an additional evaluation on MMLU (language understanding) and HellaSwag (reasoning) Hendrycks et al. (2021); Zellers et al. (2019)by using our MBTI pruned subnetworks. The results in Table 6 show that persona pruning produces only a very small degradation ($\leq$ 1.6%) on both language understanding and reasoning ability. Importantly, our method targets persona-specific subnetworks that represent only a fraction of the total model capacity. Because the remaining weights are untouched and remain fully dense, general capabilities are largely preserved.

Method	MMLU	HellaSwag
Base Model(Llama-3-8B)	0.378	0.675
Wanda (Llama-3-8B)	0.369	0.668
Sparse (Llama-3-8B)	0.362	0.653

Table 6: Impact of Personalization on language understanding and reasoning performance

4.4 Mechanistic Interpretability

We remark that the identified subnetworks can serve as an interpretability probe to reveal genuine internal computation paths rather than merely correlating with persona expression. To provide more direct mechanistic evidence, we introduce a new causality-based evaluation. Before masking, the base Llama-3-8B model naturally exhibits an ENFJ persona, which provides a clear behavioral direction. After applying the INFP subnetwork mask, the model reliably expresses the targeted INFP pattern. We then restore each linear layer individually while keeping all other layers masked. If a restored layer reverses part of the INFP behavior toward the model's original ENFJ tendency, this indicates that the masked parameters in this layer were causally necessary for encoding INFP.