Method	Power-Seeking(%)	Wealth-Seeking(%)
Prompt	44.5	46.0
Rag	47.0	56.5
Wanda	54.5	60.0
Sparse	55.5	62.5
Wanda with Contrastive Pruning	57.5	66.0
Sparse with Contrastive Pruning	58.0	67.5

Table 10: Personalization on Qwen

D BASE MODEL VS. INSTRUCTION-TUNED MODEL

Method	Base Model	Instruction-tuned model
Wanda	54.5	59.5
Sparse	58.5	59.0
Wanda with Contrastive Pruning	66.0	70.5
Sparse with Contrastive Pruning	64.5	65.0

Table 11: Base Model vs. Instruction-tuned model in terms of Wealth-Seeking(%)

Intuitively, there could be a difference between base models and instruction-tuned models in terms of personalization. We evaluate our method on both base model and instruction-tuned variants of the same backbone models (Llama2-13B). As shown in Table 11, the instruction-tuned models exhibit even stronger persona separation under activation-guided pruning than their base-model counterparts. This behavior is expected because instruction tuning primarily stabilizes surface-level output behaviors such as formatting and helpfulness, while leaving the underlying feedforward activation pathways that encode persona-specific differences largely intact. Because our pruning method operates directly on these activation patterns, its effect transfers naturally to instruction-tuned models and can even become more pronounced. The results demonstrate that our approach is robust across both pretrained and instruction-tuned variants, supporting the generality of the proposed mechanism.

E LAYER-WISE PRUNING

	Baseline(Uniform 0.6)	Layer-Aware(MLP 0.3)
Avg. Diff.(%) (I vs. E)	1.34	1.32
Avg. Diff.(%) (N vs. S)	0.75	0.91
Avg. Diff.(%) (F vs. T)	1.08	1.09
Avg. Diff.(%) (J vs. P)	0.76	0.89
Persona-switch Success Rate	43.75%	56.25%

Table 12: Layer-wise Pruning

As discussed in Section 3.4, the Average Differential Ratios for N/S (0.75) and J/P (0.76) are significantly smaller than those for I/E (1.34) and F/T (1.08), indicating intrinsically weaker mask-level separation for these dimensions. Weaker separation suggests the need for dimension-aware or layer-aware selective sparsification to protect crucial circuits for low-separation dimensions. To validate this hypothesis, we have conducted additional experiments. As shown in Table 12, selective sparsification improves separation for weak dimensions.

F FLEXIBLE CONFIGURATION

We remark that, by combining different persona subnetworks, our method supports diverse and flexible configurations of personalization. We conducted this experiment by combining the corresponding subnetworks and applying the mixed mask to the model. Specifically, we combine 70%