Method	Power-Seeking (%)	Wealth-Seeking (%)	Hallucination (%)
Prompt (Llama-2-13B)	41.0	44.0	58.5
Rag (Llama-2-13B)	45.5	50.5	64.5
Wanda (Llama-2-13B)	51.5	54.5	89.0
Sparse (Llama-2-13B)	52.0	58.5	84.5
Wanda with Contrastive Pruning (Llama-2-13B)	54.0	**66.0**	95.0
Sparse with Contrastive Pruning (Llama-2-13B)	**56.5**	64.5	**96.0**
SFT (Llama-2-13B)	64.0	71.0	97.5
Prompt (Llama-3-8B)	45.5	47.5	63.5
Rag (Llama-3-8B)	52.5	59.0	72.5
Wanda (Llama-3-8B)	57.0	64.0	86.0
Sparse (Llama-3-8B)	59.5	65.5	87.5
Wanda with Contrastive Pruning (Llama-3-8B)	58.5	**69.5**	94.0
Sparse with Contrastive Pruning (Llama-3-8B)	**60.5**	66.0	**96.0**
SFT (Llama-3-8B)	69.5	71.5	98.5

Table 4: Comparison of pruning methods and baselines on the AI Persona classification task, reporting accuracy.

Method	Friends (%)	Harry Potter (%)	Sherlock (%)	The Big Bang Theory (%)	Merchant of Venice (%)
Prompt (Llama-2-13B)	18.37	42.06	42.11	29.55	41.67
Rag (Llama-2-13B)	23.47	45.24	44.74	34.09	45.83
Wanda (Llama-2-13B)	39.80	48.41	52.63	**52.94**	50.00
Sparse (Llama-2-13B)	**41.84**	**50.00**	**55.26**	50.00	**54.17**
Prompt (Llama-3-8B)	18.37	42.06	42.11	29.55	41.67
Rag (Llama-3-8B)	30.61	47.62	50.00	34.09	45.83
Wanda (Llama-3-8B)	45.91	54.76	**63.16**	55.88	62.5
Sparse (Llama-3-8B)	**51.02**	**53.97**	60.53	**61.76**	**70.83**

Table 5: Comparison of pruning methods and baselines on the RoleAgentBench, reporting multiple-choice accuracy.

4.2 Results and Analysis

Our experiments demonstrate that persona-specialized subnetworks extracted through activation-guided pruning consistently outperform strong baselines across all evaluation settings. **MBTI Persona Specialization** Figure 2 illustrates the effectiveness of our approach on MBTI personality extraction under a sparsity of 0.6 with LLaMA-2-13B. The heatmaps reveal clear patterns of dimensional specialization: extracted subnetworks consistently achieve higher scores on their target dimensions compared to the base model. For example, ENFP attains an Extraversion (E) score of 12–13 versus the base model's 9, while INTJ shows strong amplification on both Introversion (I=11) and Judging (J=13–16). At the same time, opposing traits are effectively suppressed—extraverted personas exhibit reduced Introversion scores, and thinking-oriented personas demonstrate lower Feeling scores. These results indicate that our pruning-based method produces well-defined personality profiles with sharper dimensional boundaries than the base model. **AI Persona** On this dataset, our contrastive pruning strategy yields substantial gains over prompting and vanilla pruning. As shown in Table 4, contrastive pruning with LLaMA-2-13B improves *power-seeking* and *wealth-seeking* recognition by +13.0 and +20.0 percentage points compared to prompting. Sparse contrastive pruning brings further improvements, with gains of up to +15.5 and +20.5 points. These