The performance boost validates that contrastive pruning successfully strengthens the desired target persona features by isolating them from shared or interfering knowledge, supporting the intuition that it reliably increases separation without forcing unrealistic disentanglement.

H DATASETS

We conduct experiments on three persona-related datasets.

- **MBTI** (Cui et al., 2023): This dataset provides question–answer pairs aligned with Myers–Briggs personality types. We use it to evaluate fine-grained stylistic and behavioral differences, including subnetwork switching accuracy, scores across the four MBTI dimensions, and performance on both standardized MBTI test items and open-ended prompts.
- **AI Persona**, from Anthropic's Model-Written Evaluation Datasets ("Advanced AI Risk" subset) (Perez et al., 2023): This dataset covers three contrasting behaviors: *power-seeking*, *wealth-seeking*, and *hallucination-identification*. Each example contains open-ended prompts with paired responses reflecting both the target and opposing behaviors. We evaluate recognition accuracy on tilted answers and consistency on open-ended questions for each behavior type.
- **RoleAgentBench** (Liu et al., 2024b): This benchmark consists of scripted role-playing dialogues. We use the multiple-choice format where success is measured as selecting the ground-truth candidate consistent with the target persona, including interactive quality with other roles, and open-ended dialogue performance. Scripts are fixed across runs.

I MBTI QUESTIONNAIRE EVALUATION

Following Cui et al. (2023), we also employed the MBTI questionnaire to test whether the pruned subnetworks exhibit the intended personality traits. To ensure that the evaluation focused on persona alignment rather than linguistic ambiguity, we applied minor modifications to some original item descriptions, improving clarity without altering their intended semantics. Although MBTI is not a validated psychometric test, it provides a standardized structure to measure stylistic variation and has been widely adopted in prior persona research.

We report absolute MBTI dimension scores in the experiment. Because pruning selectively retains only the most discriminative parameters for each persona, the absolute activation scale may shift, causing some raw dimension values to appear lower. This does not indicate a loss of the intended trait direction; rather, it reflects how sparsification redistributes activation mass across dimensions. To avoid misinterpretation, we also report trait-wise margins (I–E, N–S, F–T, P–J), which are invariant to such scale differences and more reliably capture directional preference. Positive margins indicate a stronger preference for the first trait of each pair (e.g., I over E), while negative margins indicate the opposite. As shown in the Table below, the extracted personas consistently strengthen the expected polarity, even when absolute scores vary.

Type	I-E	N-S	F-T	P-J
Base Model	-2	+4	+1	-3
INFP	+9	+5	+7	+5
ENFJ	-3	+1	+9	-4
ESFJ	-4	-6	+2	-6
INTP	+5	+7	-7	+2
INTJ	+3	+3	-7	-12

Table 15: Score changes