Method	Training-Free	No extra Parameters	Time Cost
Prompt-based (Cheng et al., 2023; Shao et al., 2023)	✓	✓	Instant
RAG-based (Zerhoudi & Granitzer, 2024; Yu et al., 2024)	✓	✓	Instant
Fine-tuning-based (Zhou et al., 2023; Wang et al., 2025)	✗	✗	Hours to days
Ours (Pruning-based)	✓	✓*	Minutes

Table 1: Comparison of persona modeling approaches. *Our method only requires lightweight binary masks, without introducing trainable parameters like LoRA or adapters.

emerge as separable sub-structures within a single pretrained model. Our key observation is that when presented with persona-specific inputs, different neurons exhibit consistently distinct activation patterns, suggesting that personas may already exist as latent, disentangled pathways in the network.

Inspired by this finding and recent advances in neural network pruning (Sun et al., 2023; Frantar & Alistarh, 2023), we propose a novel perspective: persona capabilities can be extracted as sparse subnetworks from a single pretrained LLM without any additional training. Rather than viewing personas as behaviors to be learned through fine-tuning, we treat them as pre-existing circuits to be discovered and isolated through structured pruning. This approach is motivated by the lottery ticket hypothesis (Frankle & Carbin, 2019), which demonstrates that sparse subnetworks can match the performance of dense models. We extend this principle to show that multiple "winning tickets" corresponding to different personas coexist within a single pretrained model.

In this paper, we present a train-free framework for extracting persona-specialized subnetworks through activation-guided pruning. Our method requires only small calibration datasets to identify and isolate persona-relevant parameters. Furthermore, we introduce a contrastive pruning strategy specifically designed for scenarios where personas form natural oppositions, ensuring that the extracted subnetworks are maximally disentangled.

Our contributions are threefold:

- We demonstrate that distinct personas manifest as separable activation patterns in pretrained LLMs, and these patterns can guide the extraction of specialized subnetworks without any gradient updates.
- We propose a contrastive pruning algorithm that explicitly maximizes parameter disentanglement between opposing personas, achieving stronger behavioral separation than standard pruning methods.
- We conduct extensive experiments across diverse persona evaluation settings, showing that our extracted subnetworks achieve superior persona alignment compared to prompting and other baselines, while maintaining fluency and reducing inference costs through sparsity.

Our work challenges the conventional paradigm of training separate models or adapters for different personas. Instead, we show that a single pretrained model already contains the capacity for diverse personas, which can be efficiently "unmixed" through principled pruning. We argue that these behaviors are not externally induced but embedded as sparse routing structures in parameter space, which can be systematically uncovered through pruning, enabling efficient and training-free persona switching.

2 Related Work

Persona Modeling in LLMs Recent work has explored various approaches to imbue LLMs with distinct personas and personalities, including fine-tuning on curated character datasets, role-playing evaluations via character interviews, and prompt-based induction of personality traits (Shao et al., 2023; Wang et al., 2023; Serapio-García et al., 2023). Building on this, advances in personalization and preference following long-context evaluation, inference time control, embedding strategies,