and progressive adaptation frameworks (Zhao et al., 2025; He et al., 2025; Liu et al., 2025; Mok et al., 2025; Zhang et al., 2025). Representative approaches include memory augmented prompting ((Madaan et al., 2023)), retrieval and profile-based methods ((Mysore et al., 2023; Zhuang et al., 2024)), parameter efficient adaptation ((Zhu et al., 2024; Zhang et al., 2024)), and alignment-based optimization ((Zhou et al., 2024)). However, these approaches often rely on costly fine-tuning, retrieval-augmented pipelines, or prompt based heuristics that lack robust control, while parameter-efficient adaptations still require additional training. Our work differs by extracting persona capabilities directly from pretrained models without additional training.

**Network Pruning and Sparse subnetworks** The lottery ticket hypothesis (Frankle & Carbin, 2019; Liu et al., 2024a; Hui et al., 2023; Jiang et al., 2023) demonstrates that dense networks contain sparse subnetworks capable of matching full model performance. Recent advances have extended pruning to LLMs: Wanda (Sun et al., 2023) prunes weights based on magnitude and activation patterns, while SparseGPT (Frantar & Alistarh, 2023) uses second-order information for more accurate pruning. (Ma et al., 2023) propose structured pruning for task-agnostic compression. Unlike these works that focus on general compression, we leverage pruning to discover and isolate persona-specific sub-circuits within a single model.

**Mechanistic Interpretability and Model Circuits** Understanding internal representations has revealed that specific behaviors correspond to interpretable activation patterns (Elhage et al., 2022; Olsson et al., 2022). (Li et al., 2023) identify ”truth directions” in activation space that control factuality. (Zou et al., 2022) demonstrates that concepts and behaviors can be manipulated through activation steering. (Geva et al., 2023) shows that feed-forward networks act as key-value memories encoding factual knowledge. Our approach builds on these insights by showing that personas manifest as distinct activation patterns that can guide structural decomposition. Unlike activation steering or linear representation editing, which operate directly in hidden state space at runtime, our method uncovers sparse routing structures in parameter space. This distinction enables zero-shot persona switching without injecting additional activation vectors or editing weights.

# 3 Methods

Figure 1 provides an overview of our framework for persona specialization: given a pretrained LLM and small persona-specific calibration data, we first collect activation statistics, then construct binary masks that isolate persona-relevant subnetworks, and finally apply these masks at inference to obtain controllable persona behavior. Implementation details are provided in Appendix K

## 3.1 Problem Setup

For each persona $p \in \mathcal{P}$, we assume access to a small calibration dataset $\mathcal{D}_p = \{(x_i^p, y_i^p)\}_{i=1}^{N_p}$, where $x_i^p$ represents input prompts and $y_i^p$ represents persona-consistent responses. These calibration sets are orders of magnitude smaller than typical training datasets, usually containing only hundreds to a few thousand examples per persona. Our objective is to find masks that maximize persona alignment while maintaining sparsity:

$$\max_{\mathbf{M}^p} \mathbb{E}_{(x,y)\sim\mathcal{D}_p}\left[\log P_{\mathcal{M}_p}(y|x)\right], \tag{1}$$

where $\|\mathbf{M}^p\|_0 \leq (1-\rho)d$ enforces a sparsity constraint on the mask, and $\rho \in (0,1)$ is the target sparsity ratio. We treat each Linear weight matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$, where columns correspond to input channels. Unless stated otherwise we prune all Linear modules in attention and MLP blocks; embeddings and the LM head are not pruned. The sparsity ratio $\rho$ is applied per layer and, unless otherwise specified, is shared across all layers. We denote each weight matrix by $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{m\times n}$ with $i$ indexing output channels and $j$ indexing input channels. Each persona-specific mask $\mathbf{M}^{p,(l)} \in \{0,1\}^{m\times n}$ is defined per layer, and the global mask $\mathbf{M}^p$ is formed by concatenating them across modules, which yields the persona-specialized model $\mathcal{M}_p = f(\theta \odot \mathbf{M}^p)$.

## 3.2 Persona subnetworks via Pruning

Our approach leverages the key observation that persona-specific inputs induce distinct neuron activation patterns. For a given layer $l$, with input activations $\mathbf{h}^{(l)}(x) \in \mathbb{R}^n$ under input $x$, we collect