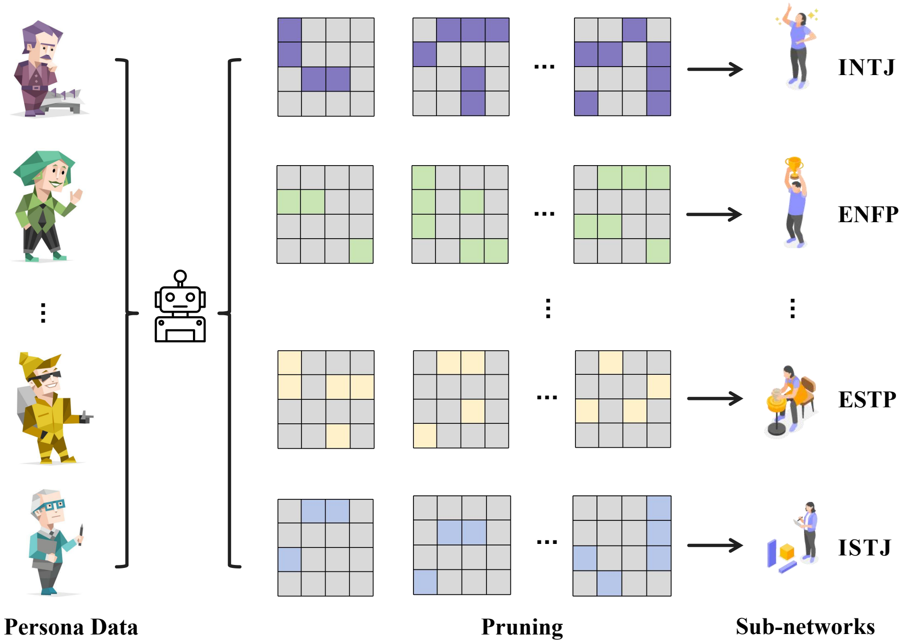

Figure 1: The figure illustrates our pruning framework. Persona specific data is employed to compute activation statistics and importance scores, which are then used to rank parameters and construct masks that isolate persona-relevant subnetworks. Colored entries mark the Top-K parameters retained for each output neuron, while gray entries are pruned.

activation statistics across the calibration data:

$$\mathbf{A}_p^{(l)}[j] = \mathbb{E}_{(x,y)\sim\mathcal{D}_p}\left[|\mathbf{h}_j^{(l)}(x)|\right], \tag{2}$$

which captures the expected activation magnitude for neuron j under persona p's data distribution. **Magnitude-based Importance Scoring** For a linear layer with weights $\mathbf{W} \in \mathbb{R}^{m\times n}$, we compute importance scores that combine weight magnitude with activation frequency:

$$S_{ij}^p = |w_{ij}| \cdot \mathbf{A}_p^{(l)}[j], \tag{3}$$

The intuition is that parameters both large in magnitude and frequently activated by persona-specific inputs are most critical for that persona's behavior. We perform a row-wise Top-K: for each output channel i, we keep $K = \lfloor(1-\rho)\cdot n\rfloor$ input columns with the largest S_{ij}^p values and set the rest to zero, yielding a binary mask $\mathbf{M}^p \in \{0,1\}^{m\times n}$.

$$\mathbf{M}_{ij}^p = \begin{cases} 1 & \text{if } S_{ij}^p \in \text{TopK}_K(\{S_{ik}^p\}_{k=1}^n) \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

Refinement To account for parameter interactions, we optionally employ a refinement criterion inspired by second-order pruning methods. Specifically, we approximate parameter sensitivity using per-dimension input variance estimated from calibration data, with a small damping constant λ to ensure numerical stability. These variance-normalized scores are applied column-wise to rank input features, guiding the pruning order:

$$\mathbf{H} \approx \frac{1}{|\mathcal{D}_p|}\sum_{(x,y)\sim\mathcal{D}_p} \mathbf{h}^{(l)}(x)\mathbf{h}^{(l)}(x)^\top + \lambda\mathbf{I}, \tag{5}$$

where we retain the diagonal entries as an efficient approximation.

3.3 Efficient Inference via Dynamic Masking

During inference, we apply persona-specific masks without modifying the original weights:

$$\mathbf{y} = (\mathbf{W}\odot\mathbf{M}^p)\mathbf{x} + \mathbf{b}, \tag{6}$$