YOUR LANGUAGE MODEL SECRETLY CONTAINS PERSONALITY SUBNETWORKS

Ruimeng Ye[1], Zihan Wang[2], Zinan Ling[1], Yang Xiao[1], Manling Li[2], Xiaolong Ma[3], Bo Hui[1]
[1]University of Tulsa [2]Northwestern University [3]University of Arizona
{ruy9945, bo-hui}@utulsa.edu

ABSTRACT

Humans shift between different personas depending on social context. Large Language Models (LLMs) demonstrate a similar flexibility in adopting different personas and behaviors. Existing approaches, however, typically adapt such behavior through external knowledge such as prompting, retrieval-augmented generation (RAG), or fine-tuning. We ask: do LLMs really need external context or parameters to adapt to different behaviors, or do they already have such knowledge embedded in their parameters? In this work, we show that LLMs already contain persona-specialized subnetworks in their parameter space. Using small calibration datasets, we identify distinct activation signatures associated with different personas. Guided by these statistics, we develop a masking strategy that isolates lightweight persona subnetworks. Building on the findings, we further discuss: how can we discover opposing subnetwork from the model that lead to binary-opposing personas, such as introvert-extrovert? To further enhance separation in binary opposition scenarios, we introduce a contrastive pruning strategy that identifies parameters responsible for the statistical divergence between opposing personas. Our method is entirely training-free and relies solely on the language model's existing parameter space. Across diverse evaluation settings, the resulting subnetworks exhibit significantly stronger persona alignment than baselines that require external knowledge while being more efficient. Our findings suggest that diverse human-like behaviors are not merely induced in LLMs, but are already embedded in their parameter space—pointing toward a new perspective on controllable and interpretable personalization in large language models. Our code is available at https://github.com/Ruimeng-Ye/Persona.git.

1 INTRODUCTION

Humans shift between different personas depending on context. A child may speak politely with a teacher yet joke casually with friends; the same adult might appear cautious and formal in a job interview but warm and humorous at a family dinner. These shifts in tone, style, and behavior are not learned separately for each context—they emerge naturally as flexible reconfigurations of the same underlying cognitive system.

Large language models (LLMs) are also capable of adopting different personas. They can generate outputs that mimic diverse behavioral styles with carefully designed prompts, retrieval, or fine-tuning. However, a major division of current methods treats persona control as what must be externally imposed on a monolithic model: they involve training an ensemble of expert models, with each model specialized for a single persona. Alternatives such as retrieval-augmented generation (RAG) or prompting reduce overhead, while often suffering from interference, shallow control, or unstable persona fidelity (as shown in Table 1). This raises a fundamental question: do LLMs really require external intervention to display different personas, or are these behaviors already embedded within their internal structure, waiting to be uncovered?

Recent work has revealed that behavioral traits and capabilities in LLMs are often encoded as interpretable directions in activation space (Cao et al., 2024; Chen et al., 2025). This line of research demonstrates that model behaviors can be understood and potentially controlled through their internal representations. Building on these insights, we investigate whether distinct personas naturally