Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. PROPER: A progressive learning framework for personalized large language models with group-level adaptation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 16399–16411. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.800/.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. Personalized lora for human-centered text understanding, 2024. URL https://arxiv.org/abs/2403.06208.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=QWunLKbBGF.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. Characterglm: Customizing chinese conversational ai characters with large language models, 2023. URL https://arxiv.org/abs/2311.16832.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization, 2024. URL https://arxiv.org/abs/2310.03708.

Jiachen Zhu, Jianghao Lin, Xinyi Dai, Bo Chen, Rong Shan, Jieming Zhu, Ruiming Tang, Yong Yu, and Weinan Zhang. Lifelong personalized low-rank adaptation of large language models for recommendation, 2024. URL https://arxiv.org/abs/2408.03533.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. Hydra: Model factorization framework for black-box llm personalization, 2024. URL https://arxiv.org/abs/2406.02888.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency, 2023. *URL https://arxiv. org/abs/2310.01405*, 97, 2022.

A LLM USAGE

During the preparation of this paper, we used large language models (e.g., ChatGPT) as writing assistants for language polishing and clarity improvement. The models were not involved in idea generation, experimental design, or result analysis. All scientific content and conclusions are the responsibility of the authors.

B REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our results. All datasets used are publicly accessible. Anonymous source code and scripts for reproducing our experiments will be made available in the supplementary materials. These resources should allow researchers to replicate our results and extend our framework to new settings.

C EXPERIMENTS ON QWEN

We have also conducted a new set of experiments on Qwen2.5-14B, a model family with different tokenizers, architectural choices, and training dynamics compared to Llama. The results in Table 10 confirmed our activation-guided pruning outperformed the prompt and RAG baselines on Qwen2.5-14B by margins comparable to those observed on Llama models. This cross-architecture consistency strongly suggests that latent persona-subnetworks are not an artifact of the Llama family, but rather a general inductive property of pretrained LLMs.