The remarkable capabilities of large pre-trained models come with limitations in domain-specific adaptation (Nayak et al., 2023) and knowledge retention (Luo et al., 2023). This presents a key challenge for superintelligent models: how can LLMs specialize in new domains and tasks while preserving existing knowledge and alignment training? To address these challenges, this project introduces a data-centric framework (Chen et al., 2024a) inspired by curriculum learning principles. This enables continuous learning in aligned LLMs by fostering adaptation to increasingly complex tasks and mitigating catastrophic interference while preserving the model's foundational alignment.

Our approach relies on three key components:

- **Task Difficulty Evaluation (3 Months):** Traditional methods for evaluating task difficulty rely on human annotation and are subjective, expensive, and time-consuming. We propose using an ensemble of foundation models of various capabilities to evaluate task difficulty (Chen et al., 2022; Safranchik et al., 2020) (e.g. based on error rates) and select appropriate learning materials for our model.
- **Balanced Data Distribution (4 Months):** Striking a balance is critical for learning systems to avoid catastrophic forgetting while expanding to new domains. We plan to maintain a balanced curriculum with adaptive task sampling and prioritized experience replay (Schaul et al., 2015; Horgan et al., 2018).
- **Adaptive Curriculum Design (5 Months):** LLMs have benefited from self-supervised learning methods like self-instruct (Wang et al., 2022), self-play (Chen et al., 2024b), and self-reward (Yuan et al., 2024), suggesting they might be forming their own internal, implicit "curriculums." An explicit curriculum tailored to the model's specific capabilities can further allow it to more effectively acquire new skills.

References

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2024a. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36.

Mayee F Chen, Daniel Y Fu, Dyah Adila, Michael Zhang, Frederic Sala, Kayvon Fatahalian, and Christopher Ré. 2022. Shoring up the foundations: Fusing model embeddings and weak supervision. In *Uncertainty in Artificial Intelligence*, pages 357–367. PMLR.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Dan Horgan, John Quan, David Budden, Gabriel Barth-Maron, Matteo Hessel, Hado Van Hasselt, and David Silver. 2018. Distributed prioritized experience replay. *arXiv preprint arXiv:1803.00933*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. 2023. Learning to generate instructions to adapt language models to new tasks. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Esteban Safranchik, Shiying Luo, and Stephen Bach. 2020. Weakly supervised sequence tagging from noisy rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5570–5578.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2015. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.