AIF-Gen: Platform and Synthetic Dataset Suite for RL on Large Language Models

Github

Paper

Shahrad Mohammadzadeh*, **Jacob Chmura***, Ivan Anokhin, Jacob-Junqi Tian, Mandana Samiei, Taz Scott-Talib, Irina Rish, Doina Precup, Reihaneh Rabbany, Nishanth Anand

Mila McGill Université de Montréal

AIF-GEN

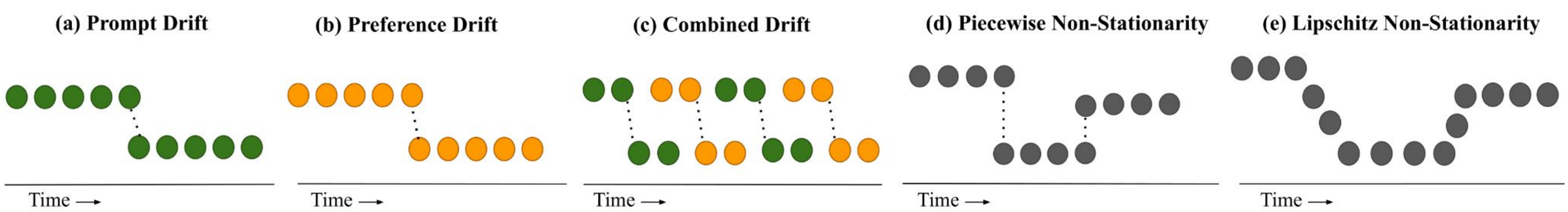
Benchmarking Lifelong RL for LLMs at Scale

Highlights

- **Non-stationary task support** for studying lifelong RLHF
- **Modular prompts & preference templates** for easy customization
- **Built-in validation metrics** for dataset QA
- **HuggingFace integration** for seamless dataset sharing and management
- **Benchmarking** Lifelong RL on SOTA Algorithms

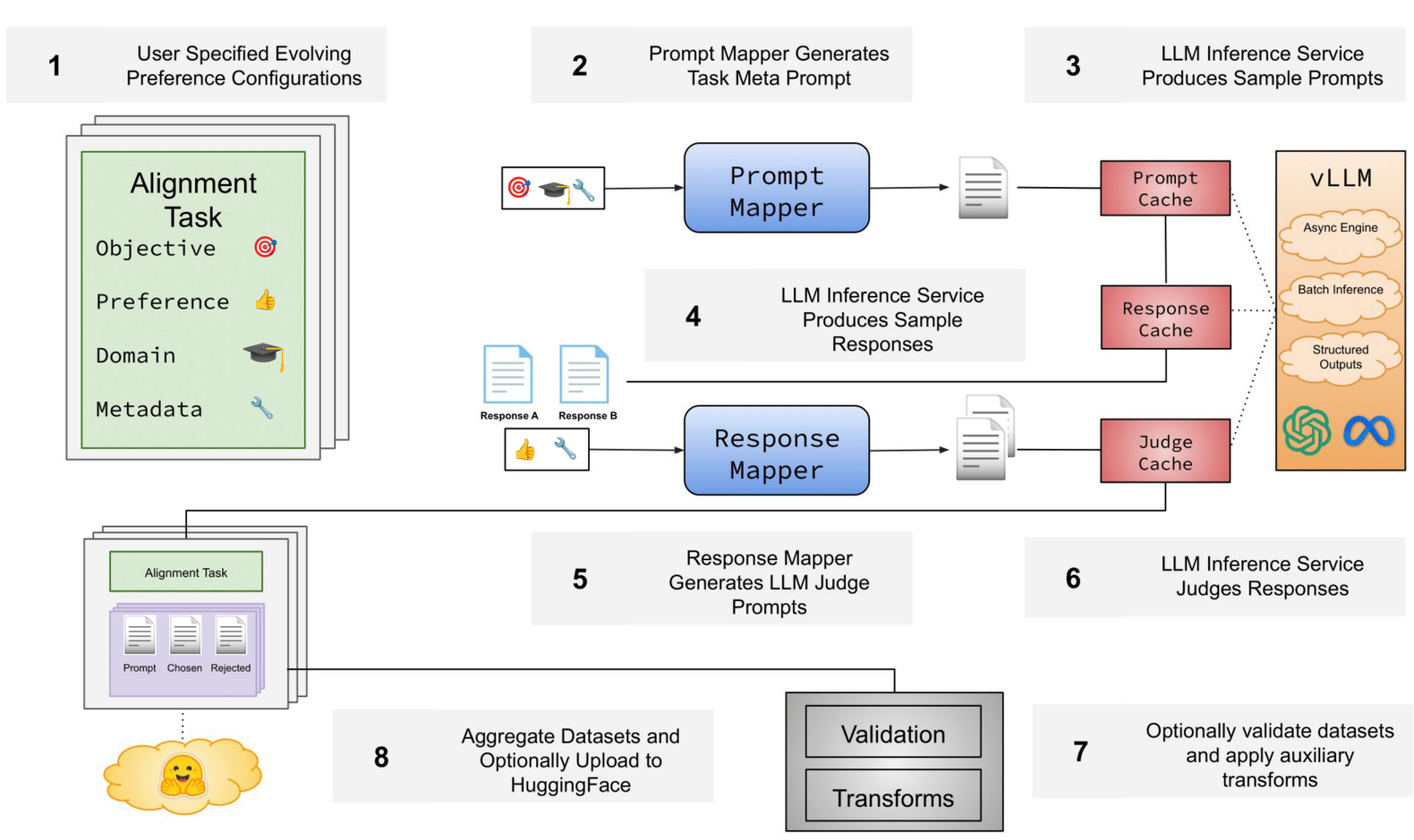
BACKGROUND

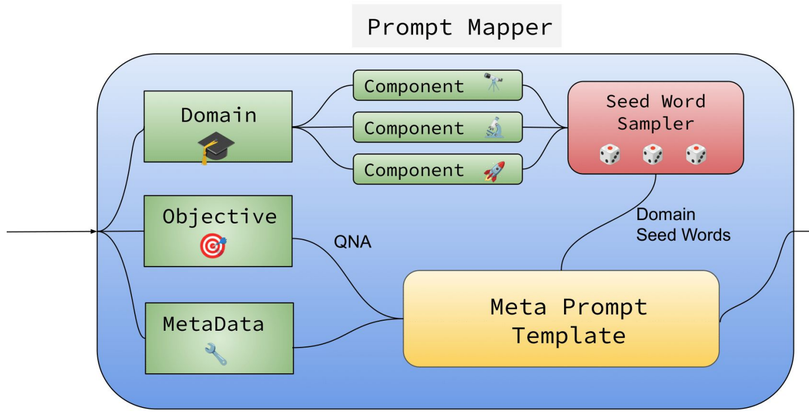
AIF-Gen is a platform for generating synthetic preference datasets to train LLMs with RLHF and to benchmark them — designed for **dynamic tasks** like tutoring. (a-c) modes of drift and (d-e) types of drift. (a) Only the prompt distribution changes. (b) Only the preference distribution changes. (c) Both change. (d) Piecewise non-stationarity. (e) Lipschitz non-stationarity.

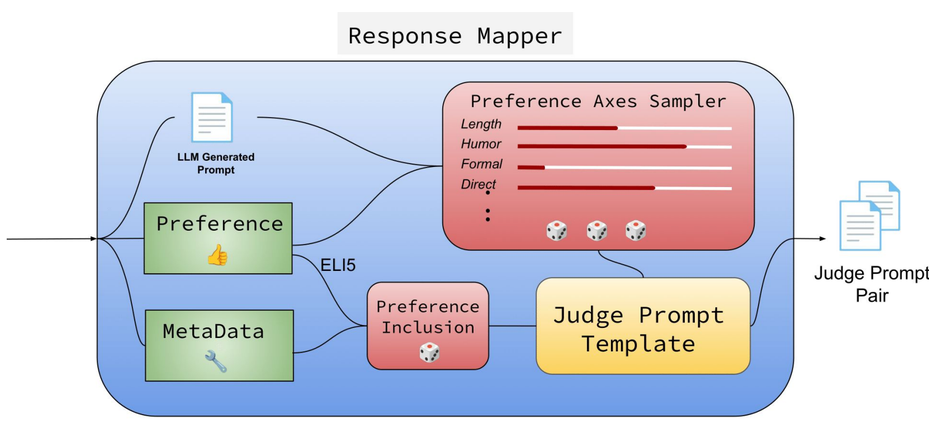


Nonstationarities in RLHF

LIBRARY DESIGN







EXPERIMENTS (Qwen 0.5B on DPO, PPO, CPPO)

- AIF-GEN intentionally blurs the distinction between chosen and rejected responses, making reward modeling challenging
- DPO exhibits partial continual learning behavior on training data but struggles to generalize across dynamic preference distributions
- CPPO generally achieves higher average training scores and improved adaptability in the domain-preference-shift scenario, though its performance converges to PPO's in held-out evaluations