SORBONNE UNIVERSITÉ

Rewarded Soups

Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, Matthieu Cord

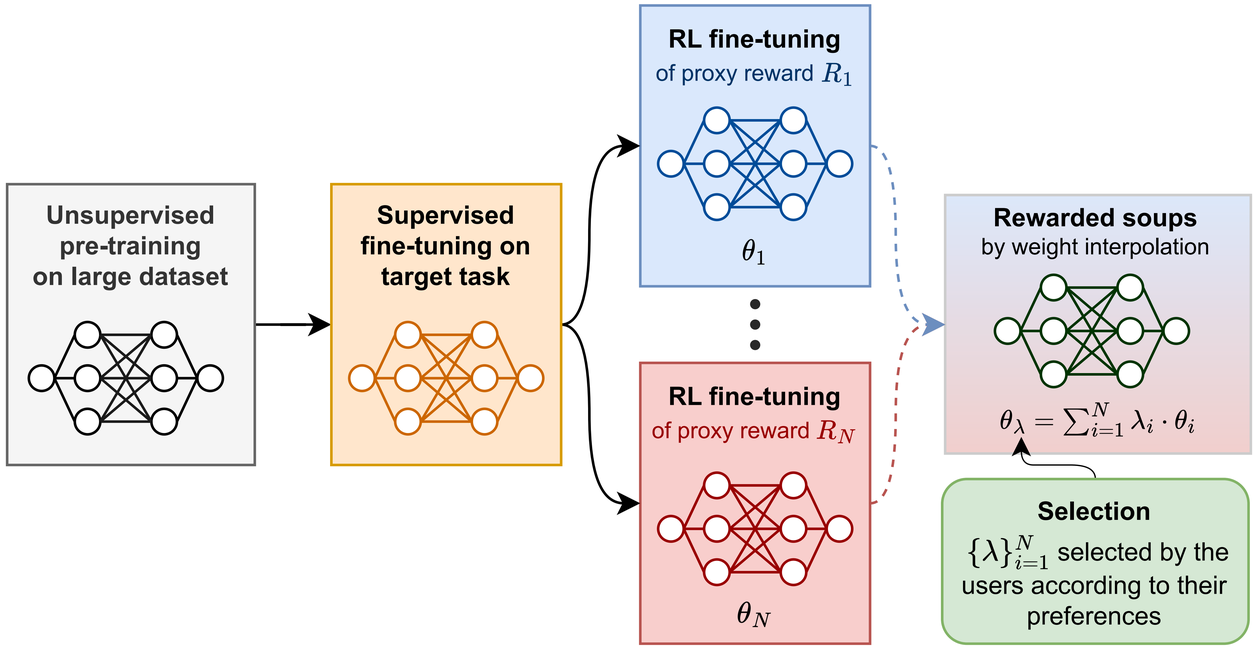

Fig1. We propose rewarded soup, an efficient and flexible multi-policy strategy for reinforcement learning from foundation models. We first specialize multiple weights $\{\theta_i\}_{i=1}^{N}$ independently, one for each proxy reward in $\{R_i\}_{i=1}^{N}$. Then we interpolate those models linearly in the weight space $\sum_{i=1}^{N} \lambda_i \cdot \theta_i$.

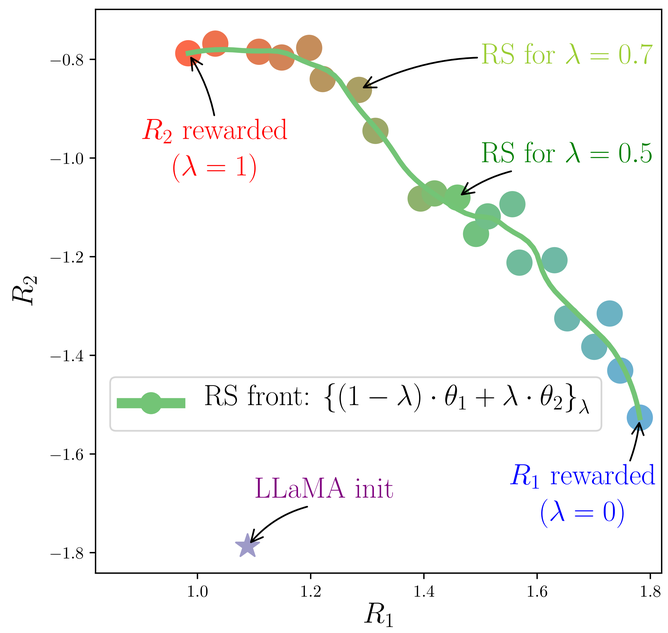

Fig2. We consider $N = 2$ weights fine-tuned from LLaMA on 2 divere rewards for summaries: θ_1 optimized for R_1 evaluating completeness, θ_2 optimized for R_2 evaluating faithfulness. Then, weight interpolation $(1-\lambda)\cdot\theta_1 + \lambda\cdot\theta_2$ between those 2 models trade-off their abilities.

Rewarded soup leverages weight interpolation for human-aligned AI via RLHF:

1. Move from single-policy towards a multi-policy paradigm to embrace the **diversity of human opinions** (Fig1).
2. Reveals by weight interpolation a **Pareto-optimal set of solutions**, and thus reduces reward misspecification (Fig2).
3. Benefits from **linear mode connectivity** between weights RLHF fine-tuned with diverse rewards from a **shared pretrained initialization**.
4. Is one step towards more **efficient** building of more **transparent** and **fairer** LLMs.
5. Applied to **LLMs** but also **multimodal tasks** such as image-to-text captioning, text-to-image diffusion, robot locomotion and more.