Harnessing Uncertainty: Entropy-Modulated Policy Gradients for Long-Horizon LLM Agents

Jiawei Wang[1,2], Jiacai Liu[1,3], Yuqian Fu[1,4], Yingru Li[1], Xintao Wang[1,3], Yuan Lin[1], Yu Yue[1], Lin Zhang[1], Yang Wang[1], Ke Wang[1]

[1] ByteDance, [2] USTC, [3] FDU, [4] CASIA

code available

MOTIVATION & KEY INSIGHT

Problem

- Long-horizon LLM agents often learn from sparse, outcome—based reward.
- Standard policy-gradient methods assign the same trajectory-level signal to all steps.
- A core issue: gradient magnitude is inherently coupled with policy entropy.
- Consequence: confident correct actions receive small updates, while uncertain exploratory actions can produce large, noisy updates.

Key Insight

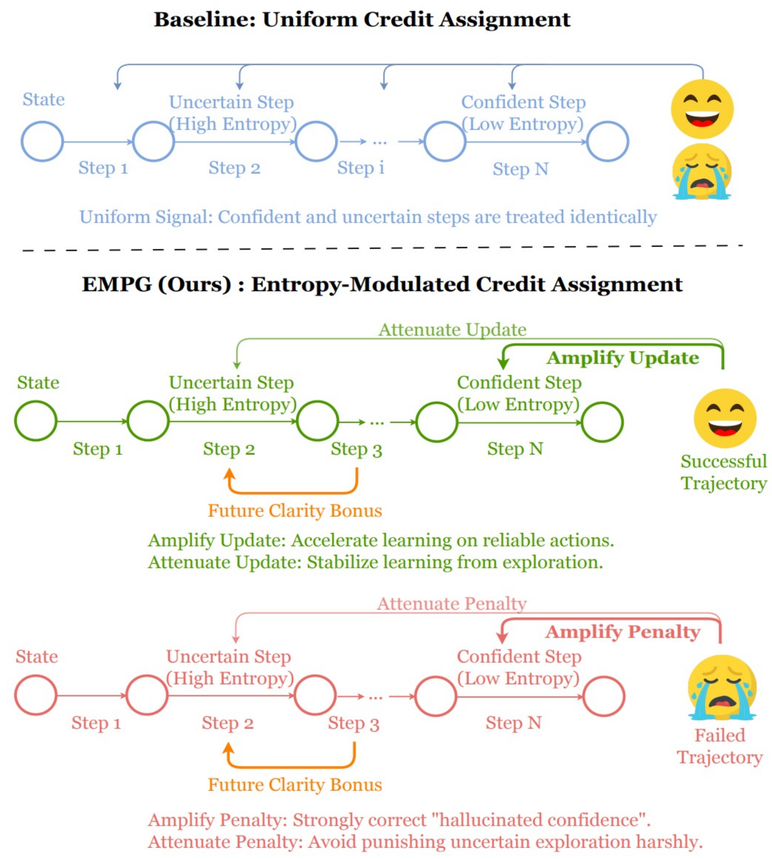
$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[||\nabla_{z_\theta(s)}\log\pi_\theta(a|s)||^2\right] = 1-\exp(-H_2(\pi_\theta(\cdot|s))$$

Higher entropy implies larger gradient norm;
Lower entropy implies smaller gradient norm.

Our Idea: EMPG

- Use step-level uncertainty to modulate credit assignment.
- Amplify learning on confident and correct steps.
- Penalize confident mistakes strongly, while attenuating noisy updates from uncertain exploration.
- Reward actions that lead to clearer future states.

METHOD



EMPG Formulation (Per Trajectory i at step t)

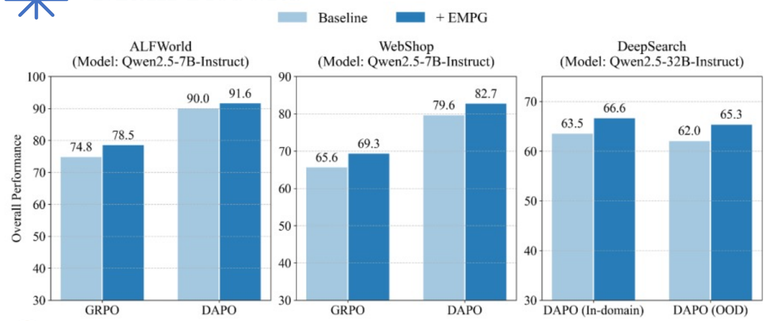
1. Step entropy: $H_i^{(i)} =$ Average token entropy over a reason-then-act step
2. Modulated advantage: $A_{mod}^{(i)}(t) = A^{(i)} \cdot g\left(H_t^{(i)}\right) + \zeta \cdot f(H_{t+1}^{(i)})$
3. Self-calibrating scaling: $g(H_t) = \dfrac{\exp(-k \cdot H_{norm,t})}{mean_{batch}[\exp(-k \cdot H_{norm})]}$
4. Future clarity bonus: $f(H_{t+1}) = \exp(-k' \cdot H_{norm,t+1})$
5. Normalize the final modulated advantages before policy updates

EXPERIMENTS & RESULTS

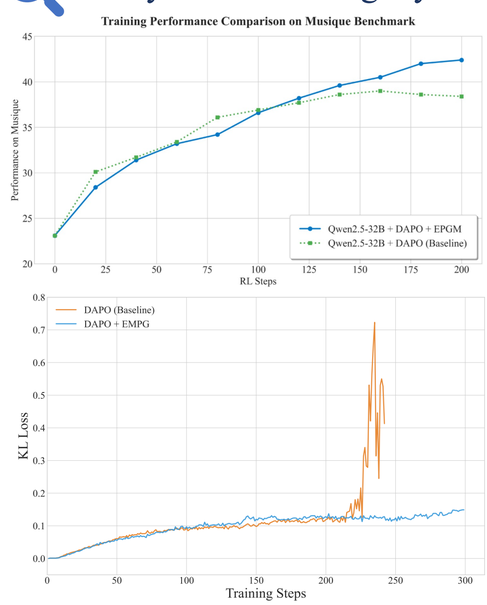
Experimental Setup

Agent paradigm: ReAct
RL Baseline: GRPO & DAPO
Benchmarks: ALFWorld, WebShop, Deep Search
Models: Qwen2.5-1.5B, Qwen2.5-7B, Qwen2.5-32B, LLaMA-3.1-8B

Main Results



Analysis on Training Dynamics



- EMPG keeps improving after the DAPO baseline plateaus
- DAPO baseline KL becomes erratic late in training.
- DAPO + EMPG remains stable throughout.

Takeaways

- EMPG is a plug-in advantage-modulation framework for long-horizon LLM agents.
- It leverages intrinsic uncertainty for finer-grained credit assignment.
- It improves performance, generalization, and training stability across benchmarks.
- It offers a scalable alternative to expensive process-based step supervision.