GEM: Generative Entropy-Guided Preference Modeling for Few-shot Alignment of LLMs

Yiyang Zhao[1], Huiyu Bai[1], Xuejiao Zhao[2,3*]

[1]College of Computing and Data Science, Nanyang Technological University (NTU), Singapore

[2]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY)

[3]Alibaba-NTU Singapore Joint Research Institute (ANGEL), NTU, Singapore

*Corresponding Author

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY)

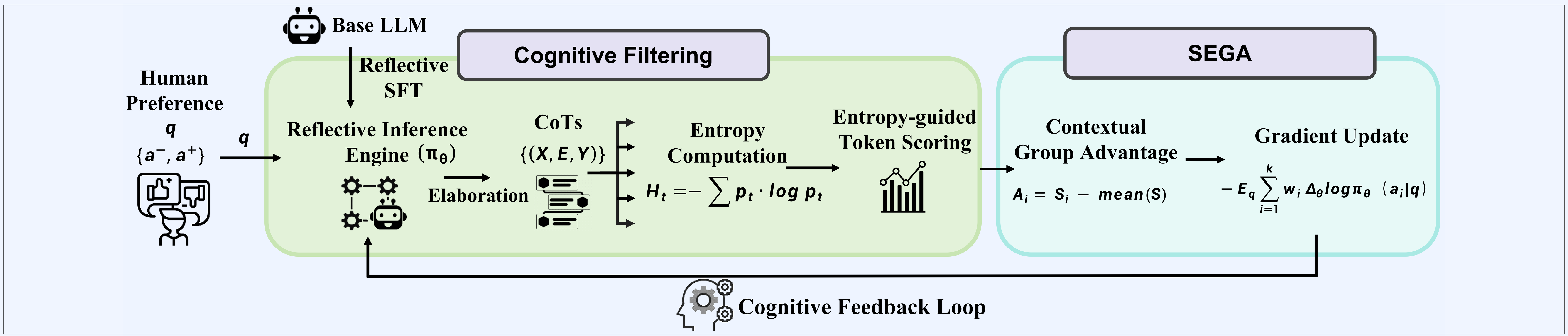
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | Alibaba

Alibaba-NTU Global e-Sustainability CorpLab

Overview: The GEM Pipeline



The pipeline of GEM. Given a query, we generate diverse Chain-of-Thought (CoT) candidates. These are ranked by our **entropy-guided token scoring**. The **SEGA** module then computes group advantages to update the policy without an external reward model.

Motivation: Why GEM?

The Problem:

- RLHF relies on thousands of human labels (costly in Medicine/Law).
- Small Reward Models generalize poorly.

Our Insight:

- LLMs have internal knowledge of correctness via **uncertainty**.
- We treat the LLM as its judge using Entropy.

Cognitive Filtering

Instead of external reward models, GEM extracts fine-grained cognitive signals from the LLM's own probability distribution.

1. Reflective Inference Engine: For each query q, we sample k diverse Chain-of-Thought (CoT) candidates $A = \{a_1, \dots, a_k\}$ to capture different reasoning paths.

2. Biphasic Entropy Scoring: We evaluate the quality of each CoT based on the *Uncertainty-Correctness* correlation. A good reasoning chain should be **exploratory** during critical steps ("forks") but **decisive** at the conclusion.

Scoring Function $S(a_i)$

$$S(a_i) = -\underbrace{H_{\text{final}}(a_i)}_{\text{Confidence}} + \lambda \cdot \underbrace{\left(\frac{1}{n}\sum_{t=1}^{n} H_t\right)_{\text{top}-m}}_{\text{Reasoning Exploration}}$$

Why this works?

- **High Mid-Entropy (Forks):** Indicates the model is actively comparing multiple logical paths at critical decision points, preventing "greedy" errors.
- **Low Final-Entropy:** Ensures the model has resolved uncertainty and is confident in the final result.

Outcome: This allows us to rank candidates $(a_{(1)} \succ \dots \succ a_{(k)})$ and construct preference pairs without human annotation.

SEGA: Self-Evaluated Group Advantage

Self-**E**valuated **G**roup **A**dvantage is a listwise optimization algorithm that treats k generated candidates as a group.

The Mechanism:

1. **Implicit Reward:** Map entropy score to reward $r_i = f(S(a_i))$.
2. **Group Baseline:** Compute mean reward $\bar{r} = \frac{1}{k}\sum_{j=1}^{k} r_j$.
3. **Advantage Estimation:** $A_i = r_i - \bar{r}$.
4. **Policy Update:** Increase probability of candidates with $A_i > 0$.

SEGA Gradient Objective

$$\nabla_\theta \mathcal{L} = -\mathbb{E}_q \left[\sum_{i=1}^{k} w_i(A_i) \cdot \nabla_\theta \log \pi_\theta(a_i \mid q) \right]$$

where $w_i \propto$ Advantage $(r_i - \bar{r})$

Main Results: Preference Alignment

Comparing GEM against SFT, PPO, DPO, PRO, and IPO on general benchmarks (UltraFeedback, RewardBench).

Method	UltraFeedback	SafeRLHF	RewardBench	Avg.
SFT	60.2	58.1	57.4	58.6
PPO (RLHF)	61.0	59.2	59.8	60.0
DPO	66.1	64.0	63.2	64.4
PRO	68.7	65.8	65.9	66.8
IPO	70.4	68.1	67.3	68.6
GEM (Ours)	**77.1**	**74.6**	**75.4**	**75.7**

GEM outperforms IPO by **+7%** and PPO by **+15%** on average, using only 3k preference pairs.

Ablation Studies

Component analysis on UltraFeedback and GSM8K.

Variant	UltraFeedback	GSM8K
w/o Cog.Filter & SEGA	69.0	48.3
w/o Final-Entropy Only	74.2	50.1
w/o Fork-Entropy Only	73.8	52.7
w/o SEGA (use DPO)	74.5	53.4
Full GEM	**77.1**	**55.6**

Downstream Task Performance

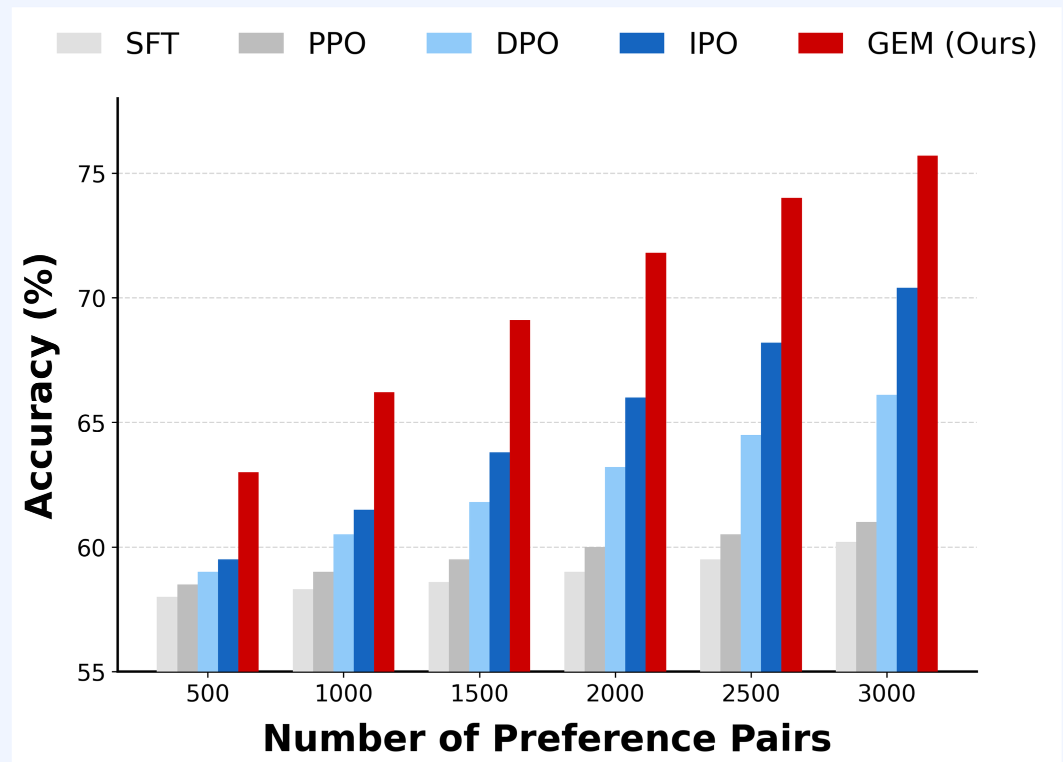
Does alignment improve reasoning? Tested on GSM8K, MATH, and TruthfulQA.

Method	GSM8K	MATH	TruthfulQA	MT-Bench
SFT	40.1	5.8	32.4	35%
PPO	44.7	7.3	34.0	47%
DPO	50.2	8.5	35.6	52%
GEM	**55.6**	**10.5**	**38.2**	**68%**

Medical Domain (iCliniq):

- GEM achieves **78.2%** agreement with experts (vs 72.5% for PPO).

Sample Efficiency Analysis



Our GEM framework outperforms baselines significantly in low-data regimes.

Conclusion

- **Generative Preference Modeling** is viable and superior for low-resource settings.
- **Internal Entropy** successfully filters high-quality reasoning without supervision.
- Code: `github.com/SNOWTEAM2023/GEM`

Scan me