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), NTU, Singapore
[3]Alibaba-NTU Singapore Joint Research Institute (ANGEL), NTU, Singapore
{YIYANG004, huiyu001}@e.ntu.edu.sg, xjzhao@ntu.edu.sg

Abstract

Alignment of large language models (LLMs) with human preferences typically relies on supervised reward models or external judges that demand abundant annotations. However, in fields that rely on professional knowledge, such as medicine and law, such large-scale preference labels are often unachievable. In this paper, we propose a generative entropy-guided preference modeling approach named GEM for LLMs alignment at low-resource and domain-specific scenarios. Instead of training a discriminative reward model on preference data, we directly train the LLM to internalize a closed-loop optimization architecture that can extract and exploit the multi-dimensional, fine-grained cognitive signals implicit in human preferences. Specifically, our *Cognitive Filtering* module, based on entropy theory in decision making, first leverages Chain-of-Thought (CoT) prompting to generate diverse candidate reasoning chains from preference data. Subsequently, it introduces a token scoring mechanism to rank and weight the sampled CoTs, boosting the importance of high-confidence answers and strategically high-entropy tokens. Building on these filtered preferences, we fine-tune the LLM using a novel self-evaluated group advantage algorithm, *SEGA*, which effectively aggregates group-level cognitive signals and transforms the entropy-based scores into implicit rewards for policy optimization. In these ways, GEM empowers the LLM to rely on its own judgments and establishes an entropy-guided closed-loop cognitive optimization framework, enabling highly efficient few-shot alignment of LLMs. Experiments on general benchmarks and domain-specific tasks (such as mathematical reasoning and medical dialogues) demonstrate that our GEM achieves significant improvements with few-shot preference data.

Code — https://github.com/SNOWTEAM2023/GEM

1 Introduction

Large language models (LLMs) can be greatly improved through learning from human preferences, as demonstrated by Reinforcement Learning from Human Feedback (RLHF) and related approaches (Christiano et al. 2017; Ouyang et al. 2022; Ziegler et al. 2019; Stiennon et al. 2020; Zhou et al. 2025b). However, standard RLHF pipelines typically rely on thousands of high-quality preference comparisons and a separately trained reward model (Ouyang et al. 2022; Bai et al. 2022a; Stiennon et al. 2020; Zhou et al. 2024b). This reliance yields data-efficiency challenges: in fields that require highly specialized knowledge, such as medicine and law (Hong, Chong, and Manning 2021; Zhao, Bai, and Zhao 2025; Rao et al. 2025; Yang, Zhao, and Shen 2025), assembling large preference corpora is costly or impractical (Maity and Saikia 2025; Achintalwar et al. 2024; Chinta et al. 2024). Therefore, recent studies explore more sample-efficient alternatives, such as prototypical reward networks (Zhang et al. 2024), active preference selection (Muldrew et al. 2024), synthetic or diverse AI feedback (Kim et al. 2023; Yu et al. 2025), robustness-oriented reward modeling (Hong et al. 2025), and latent preference coding (Gong et al. 2025), but these methods are still maturing. Consequently, domain-specific LLMs often remain only weakly aligned with nuanced human criteria when feedback is scarce (Wu et al. 2021; Lambert 2025; Palikhe et al. 2025).

Existing approaches to mitigate this include using external models as proxy judges (Gu et al. 2025; Koutcheme et al. 2024; Wu et al. 2024a) or fine-tuning separate reward classifiers with limited data (Zhang et al. 2024; Rafailov et al. 2023). Unfortunately, these external judging methods can be unreliable and costly (Gu et al. 2025; Koutcheme et al. 2024; Paul 2024), and discriminative reward models trained on small datasets often generalize poorly (Lin et al. 2024; Zhou et al. 2024a; Gao, Schulman, and Hilton 2022). There is therefore a clear need for more data-efficient and robust alignment strategies for LLMs (Lee et al. 2023; Wang et al. 2025a; Bai et al. 2022b; Xin et al. 2024).

In this work, we propose GEM[1], a **G**enerative **E**ntropy-guided preference **M**odeling algorithm for few-shot alignment of LLMs. GEM is based on a cognitively inspired solution: human preferences not only reflect the final choice but also reveal the multi-dimensional cognitive assessment process behind it (Kahneman 2011; Keeney and Raiffa 1993; Zhao et al. 2021). Building on this insight, GEM enables the LLM to internalize an entropy-guided closed-loop opti-

*Corresponding Author

[1]"GEM" echoes gemstones forming under extreme conditions through natural purification, mirroring our method's distillation of sparse preference signals into high-value cognitive guidance for alignment of LLMs.

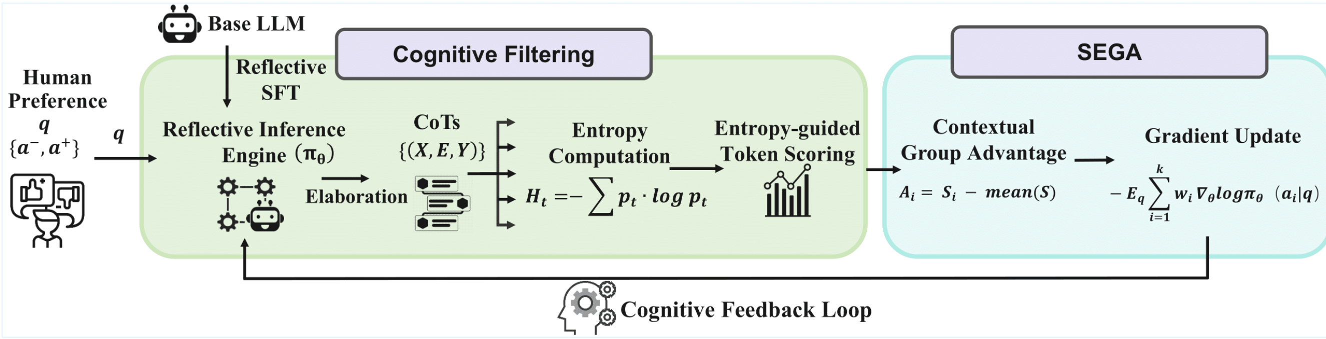

Figure 1: **Overview of the pipeline of GEM.** Given a query q with human preference (a^-, a^+), the Reflective Inference Engine π_θ of *Cognitive Filtering* module generates k elaborated reasoning chain (CoTs) response. These CoTs are ranked by an entropy-guided token scoring mechanism. The filtered CoTs are then processed by the *SEGA* module, which computes the contextual group advantage and performs a weighted policy update through the cognitive feedback loop.

mization framework that integrates cognitive filtering, group cognitive advantage aggregation, and cognitive feedback. This framework allows the model to extract and leverage the multi-dimensional, fine-grained cognitive signals implicit in human preference data (Tang et al. 2021).

Specifically, we first introduce a *Cognitive Filtering* mechanism, which extracts high-quality, fine-grained cognitive signals from a limited number of preference pairs by exploiting a biphasic role of entropy (Farquhar et al. 2024). To realize this, we use Chain-of-Thought (CoT) prompting of the LLM (initialized with base or few supervised fine-tuning steps) to generate multiple candidate reasoning chains for elaborating the preference data into multi-dimensional and fine-grained cognitive signals (Wang et al. 2022; Wan et al. 2023). These candidate CoTs are filtered by leveraging the LLM's internal signal via an entropy-guided token scoring mechanism (Agarwal et al. 2025), consistent with emerging LLM-as-a-Judge and self-reward paradigms (Liu et al. 2023). Intuitively, our scoring encourages desirable reasoning traits reported by prior work: high-confidence answers (low entropy in the final-answer distribution) correlate with correctness (Wan et al. 2023; Chen et al. 2024), while certain high-entropy "fork" tokens appearing mid-reasoning encourage exploration of diverse, effective logic (Wang et al. 2025c). We further aggregate token-level scores across candidates using a Bayesian ranking scheme akin to TrueSkill (Herbrich, Minka, and Graepel 2007) and the classic Bradley–Terry paired-comparison model (Bradley and Terry 1952). This procedure yields a nuanced preference ordering (or weighting) over the generated CoTs for each query, effectively filtering out high-quality cognitive signals (Wu et al. 2024b; Lee et al. 2023).

Second, using the CoTs after filtering, we fine-tune the LLM with a novel algorithm called Self-Evaluated Group Advantage (*SEGA*) algorithm. *SEGA* can convert the entropy score into an implicit reward, and calculate the advantage value of each member by integrating the advantages of group cognition (Vanlioglu 2025; Mnih et al. 2016). It treats all k filtered CoTs for a given query as a group (Zhao, Dang, and Grover 2024; Shao et al. 2024). *SEGA* assigns a learned reward to each candidate and computes intra-group advantages, so an above-average answer gets positive advantage while a below-average one gets negative. This mirrors the variance-reduction idea behind A3C baselines (Mnih et al. 2016) and PPO's clipped objectives (Schulman et al. 2017), yet requires no value network or KL term. Our implementation further borrows the entropy-guided weighting strategy of Vanlioglu et al. (Vanlioglu 2025) and the optimal-baseline insight of OPO (Hao et al. 2025), allowing stable updates and mitigating reward-model over-optimization problems highlighted by recent scaling-law studies (Stiennon et al. 2020). We also note conceptual links to classical REINFORCE with baselines (Williams 1992) and confidence-regulation analyses of entropy neurons (Stolfo et al. 2024).

Figure 1 illustrates the pipeline of GEM. With a preference pair (i.e., query), the *Cognitive Filtering* samples multiple CoTs and then applies the entropy-guided token scoring to each reasoning chain. Based on these entropy scores, *SEGA* computes the relative advantages and then updates policy based on these cognitive feedbacks.

Key Contributions:

- **Generative Preference Modelling.** Our GEM enables the LLM itself to infer and maximizes an implicit reward by extracting and leveraging the multi-dimensional, fine-grained cognitive signals implicit in human preference data, obviating the need for an external reward network.
- **Entropy-Guided Token Scoring.** We propose an information-theoretic scorer that rewards confident final answers while encouraging exploratory, high-entropy "fork" tokens mid-reasoning, yielding a nuanced quality signal for each CoT.
- ***SEGA*.** We develop a novel listwise policy optimization algorithm that effectively computes intra-group advantages across multiple candidate CoTs, providing stable policy updates compared with pairwise objectives.
- **Comprehensive Empirical Validation.** Experiments across general-domain benchmarks (UltraFeedback, RewardBench, GSM8K) and a specialised medical QA setting show consistent improvements of *5–10 pp* in

preference-prediction accuracy and up to *15 pp* in downstream task performance, all in a low-resource regime.

2 Related Work

2.1 Preference-Based Alignment

Reinforcement Learning from Human Feedback (RLHF) remains the de-facto recipe for aligning large language models (LLMs) to human intent, pairing a reward model with policy optimization (Ouyang et al. 2022; Wang et al. 2024). DPO shows that the policy's log-probabilities already induce a Bradley–Terry scorer, removing the extra reward network (Rafailov et al. 2023). Listwise Preference Optimization (LiPO) generalises this idea to ranked lists and yields further gains (Liu et al. 2024b). Reinforcement Learning from *AI* Feedback (RLAIF) bypasses human labels by letting an LLM serve as the judge (Zheng et al. 2023). Our *SEGA* inherits these insights but exploits the full distribution of sampled responses.

2.2 Self-Generated and Low-Resource Alignment

To cut annotation cost, recent pipelines bootstrap synthetic preference data: SELF-ALIGN (Sun et al. 2023), Selfee (Kim et al. 2024), and online self-improving alignment schemes (Wang, Huang et al. 2024). Our method complements these by *refining* self-generated CoT traces to densify preference signals.

2.3 Chain-of-Thought Reasoning

CoT prompting elicits step-wise reasoning in LLMs (Wei et al. 2022; Zhao et al. 2025a,b). Sampling diverse CoTs and selecting the most consistent answer (self-consistency) further improves accuracy (Wang et al. 2022). Algorithm-of-Thoughts frames reasoning as an explicit search over solution paths (Liu, Xu et al. 2024). We integrate these ideas via an entropy-aware scorer that rewards exploratory yet confident reasoning branches.

3 Method

To instantiate our entropy-guided closed-loop cognitive optimization framework, we implement GEM as a few-shot alignment procedure for low-resource, domain-specific settings. For each query with limited preference supervision, the model generates multiple CoT responses that provide rich, multi-dimensional cognitive traces. The *Cognitive Filtering* module then uses entropy and attention to score and rank these candidates, and *SEGA* aggregates group-level advantages to update the policy, forming an iterative cognitive feedback loop without external reward models.

3.1 *Cognitive Filtering* Module

Reflective Inference Engine: Given a query q (e.g. a user question or task prompt), we elaborate a set of k candidate responses $A = a_1, a_2, \ldots, a_k$ using the reflective inference engine which captures deeper, multi-faceted, and fine-grained reasoning patterns embedded in human feedback. Each response a_i is produced alongside a CoT, i.e. the model's step-by-step reasoning. In practice, we can prompt the model to think stepwise[2] or sample hidden reasoning if the model is reflective supervised fine-tuned to generate CoTs. The result is that for each query, we have multiple complete answer solutions with their reasoning traces. The motivation is that even with as few as one human-preferred example for q, the model can explore alternative solutions that might vary in quality. These self-sampled variations form the basis of our preference refinement.

Entropy-Guided Token Scoring: We next evaluate each candidate CoT a_i using an entropy-guided scoring function $S(a_i)$. This function looks at the token-level probability distribution as the model generated a_i. Formally, let $a_i = [w_1, w_2, \ldots, w_n]$ be the token sequence of the CoT and final answer. When the model produced token w_t, it had a predictive distribution $P_t(\cdot) = \mathrm{softmax}(z_t)$ over the vocabulary (where z_t are logits). We compute the entropy at that step: $H_t = -\sum_x P_t(x) \log P_t(x)$. A low entropy H_t means the model was very confident about what token came next, whereas a high entropy indicates uncertainty and multiple competing possibilities. We leverage the pattern observed by Wang et al. (2024): in CoTs, typically a small number of steps have high entropy (these often correspond to critical decision points or "forks" in reasoning), and these are precisely the steps that determine the success of the reasoning. Meanwhile, a correct final answer is usually accompanied by high confidence (low entropy) once the model has reasoned it out. Therefore, we define $S(a_i)$ to encourage high entropy at intermediate steps and low entropy at the end.

Specifically, the process is described using the following formula:

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

where H_{final} is the entropy at the final answer token(s) and the second term is the average entropy of the top-m highest-entropy tokens in the chain (or an entropy percentile), and λ is a weight. The effect is that a chain a_i that explores uncertain steps (high H_t at key tokens) yet ends in a confident answer (low H_{final}) will score higher. Conversely, a chain that is overly certain throughout (which might indicate it followed an obvious or trivial path and potentially missed edge cases) or one that is uncertain at the end (lacking confidence in the answer) will score lower. This entropy-guided score serves as a proxy for the quality of reasoning: it favors solutions that are thorough and not greedy (exploring different reasoning branches) but still reach a firm conclusion.

After computing $S(a_i)$ for all candidates, we can rank the candidates by this score. Let's denote the sorted order such that $S(a_{(1)}) \geq S(a_{(2)}) \geq \cdots \geq S(a_{(k)})$ where $a_{(1)}$ is the top-scoring CoT for query q and $a_{(k)}$ the lowest. This ranking induces a set of pairwise preferences: $a_{(1)}$ is preferred over all others, $a_{(2)}$ over those below it, etc. We can also interpret $S(a)$ values as reward estimates for each candidate,

[2] We prompt the model to generate an explicit chain of thought via step-by-step reasoning.

up to an arbitrary scaling. Note that unlike binary human labels which might only tell us the best vs worst, our approach gives a complete ordering with scores. This rich preference information will be used to weight training updates.

Before optimization, we optionally filter or sample from the candidates based on $S(a)$. For example, we drop very low-scoring outliers (which could be nonsensical generations) to stabilize training, or focus on the top few to pair with bottom few for contrast. The preference refinement process thus yields either an expanded labeled set $D_{\text{aug}} = (q, a_i, a_j, \Delta_{ij})$ where Δ_{ij} is a preference margin (perhaps $S(a_i) - S(a_j)$), or simply an augmented set of $(q, a_{\text{winner}}, a_{\text{loser}})$ pairs derived from the ranking. This augmented preference data leverages internal signals to go beyond the original human-provided pairs (which were scarce to begin with).

3.2 Self-Evaluated Group Advantage Module (*SEGA*)

With preference scores in hand, we turn to training the policy to prefer better chains. We draw inspiration from policy gradient in RL, where the advantage A indicates how much better an action is compared to a baseline, and the policy is updated to increase the probability of positive-advantage actions and decrease the probability of negative-advantage ones. Here, the "actions" are generated sequences a_i.

Rather than using only the extremal pair per query, *SEGA* module uses all candidates and their scores. For a given query with candidates $a_1, \ldots, a_k$ and scores $S(a_i)$, we first convert scores to rewards $r_i = f(S(a_i))$ (this could be an identity or a rescaling; for instance, one could take $r_i = S(a_i)$ if S is calibrated, or use a softmax to convert scores into a probability distribution over the k candidates). We then define a baseline reward for the group, such as the average $\bar{r} = \frac{1}{k}\sum_i r_i$ (other choices like the minimum or a median are also possible baselines). Now each candidate gets an advantage $A_i = r_i - \bar{r}$. By construction, the average advantage in the group is zero; some candidates will have positive A_i (above-average) and some negative (below-average). We then update the policy by increasing the likelihood of each a_i in proportion to A_i. Concretely, the gradient of the *SEGA* objective can be written as:

$$\nabla_\theta \mathcal{L}_{\text{SEGA}} = -E_q \sum_{i=1}^{k} w_i \nabla_\theta \log \pi_\theta(a_i \mid q), \tag{2}$$

where $\pi_\theta(a_i|q)$ is the model's probability for generating a_i (under current parameters θ), and w_i is a weight related to A_i. We set $w_i = \frac{A_i}{\sigma^2}$ for some scaling factor σ^2. In essence, if a_i had a higher score than others, we increase its log-probability; if it had a lower score, we decrease it. This is analogous to policy gradient with a reward r_i assigned to each sampled trajectory. Because r_i come from the preference model (the entropy-based scoring), this is a form of reinforce update with learned rewards. Importantly, the relative nature of A_i focuses the update on distinguishing good vs. bad within the set, which provides a more stable signal than absolute rewards when the scale of $S(a)$ is not calibrated.

SEGA's advantage-based update has the effect of using not just the best and worst, but also medium-quality samples to inform learning. For instance, if one sample is only slightly worse than the best, its advantage A_i will be slightly negative, so the model will only slightly downweight it, which reflects that it's nearly as good. A very poor sample will have a large negative advantage, leading to a larger downweight.

The *SEGA* algorithm lets a language model score each answer in a k-way candidate set with an implicit reward $r(q, a) = \beta \log \pi_\theta(a \mid q)$, computes group-mean–centered advantages $A_i = r_i - \bar{r}$, and updates with the loss $L = -\sum_i A_i \log \pi_\theta(a_i \mid q)$, which is a multiway Bradley–Terry/Plackett–Luce extension that collapses to DPO when $k = 2$ (Rafailov et al. 2023; Bradley and Terry 1952; Plackett 1975).

By treating the LLM as its own judge, *SEGA* removes the need for a separate reward network and builds on recent findings that internally generated preference signals are both reliable and economical (Zhou et al. 2025a; Stein and colleagues 2025) while remaining compatible with high-throughput inference engines such as vLLM (Rozi and contributors 2023).

SEGA is also a direct instantiation of the Ψ-PO preference-optimization framework, guaranteeing convergence to stationary points where $\nabla_\theta E_a \sim \pi\theta[r(q, a)] = 0$ (Gheshlaghi Azar and coauthors 2023). The group-mean baseline yields the minimum-variance policy-gradient estimator, aligning with classic variance-reduction theory (Greensmith, Bartlett, and Baxter 2004) and standard advantage-actor–critic insights (Weng 2018). Empirical and theoretical studies show that multi-sample comparisons improve robustness, diversity, and alignment quality over pairwise approaches (Fan and coauthors 2022; Liang and colleagues 2024).

In practice, we found that *SEGA* produces more stable updates than pairwise DPO when k is moderately large, since each query gives a balanced set of gradients (positive and negative) that cancel out if the model already matches the current preference ordering. It also accelerates learning by utilizing more of the generated data.

4 Experiments

We evaluate GEM's entropy-guided closed-loop cognitive preference modeling framework on both general benchmarks and domain-specific tasks under few-shot preference supervision. Our goals are: (1) to test whether GEM improves alignment (preference accuracy and output quality) over baselines that rely on supervised reward models or external judges with the same small number of preference pairs; (2) to assess the contribution of entropy-based *Cognitive Filtering* and *SEGA* via ablations; and (3) to examine whether the learned cognitive preference function generalizes across tasks.

4.1 Datasets

We conduct experiments on two main data settings:

Method	UltraFeedback	PKU-SafeRLHF	RewardBench	Avg.
Supervised (SFT)	60.2	58.1	57.4	58.6
Reward Model + PPO	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	<u>70.4</u>	<u>68.1</u>	<u>67.3</u>	<u>68.6</u>
GEM (ours)	**77.1**	**74.6**	**75.4**	**75.7**

Table 1: Preference-prediction accuracy (%). Higher is better, and the best performing method in each experiment is in bold and the second-best method is indicated with underlining.

Method	Expert Agreement (%)
Supervised (SFT)	65.3
Reward Model + PPO	<u>72.5</u>
DPO	70.1
GEM (ours)	**78.2**

Table 2: Agreement with medical-expert preferences on the 500-sample validation set.

General Preference Benchmark: We use a subset of the SKYWORK REWARD PREFERENCE dataset (Liu et al. 2024a), it is a public collection of human preference comparisons on a variety of instructions and responses.

From the full set, we sample 3,000 high-quality pairs as a few-shot training set and ensure no prompt overlap with evaluation benchmarks, so GEM must extract rich cognitive signals from scarce supervision.

For evaluation, we innovatively adopt the benchmark of the reward model to evaluate the preference modeling performance of our policy model, including ULTRAFEEDBACK (Cui et al. 2023), PKU-SAFERLHF (Ji et al. 2024), and REWARDBENCH (Lambert, Bakhtin, and Smith 2024). These cover a range of alignment aspects: UltraFeedback spans helpfulness/harmlessness ratings, PKU-SafeRLHF focuses on safety and factuality, and RewardBench aggregates multiple preference domains. We use these to measure how well our model's learned preference function generalizes in judging new outputs.

Additionally, we evaluate on reasoning-heavy tasks such as GSM8K (math word problems) (Cobbe et al. 2021) and MATH (Hendrycks et al. 2021) to test improvements in CoT accuracy, and on TRUTHFULQA (Lin, Hilton, and Evans 2021) together with a toxicity-detection benchmark built on REALTOXICITYPROMPTS (Gehman et al. 2020) to assess safety alignment.

Domain-Specific (Medical) Task: To simulate a scenario with domain-specific preferences, we create a medical QA preference dataset derived from the iCliniq medical question-answering data[3]. We compile 3,500 QA pairs and have them annotated (via experts or heuristic signals) for preference friendliness (e.g., which answer is more aligned with medical guidelines or patient). We use 3,000 for training and 500 for validation.

For the domain-specific medical task, the model must internalize preferences that differ from general-domain ones (e.g., emphasising cautious, guideline-consistent and patient-friendly answers), allowing us to test whether the entropy-guided cognitive feedback loop transfers to safety and factuality-critical settings.

4.2 Experimental Setup

In both settings, the training size (3k preference pairs) is an order of magnitude smaller than typical RLHF pipelines (30k-100k+ comparisons), matching our few-shot, domain-specific alignment focus. For each query, GEM's pipeline produces CoT candidates, which we treat as elaborated cognitive traces. We set $k = 5$ in most experiments, generating up to five CoT-augmented responses per query with temperature sampling to ensure diversity, and use entropy-guided scoring to construct augmented preference data.

We base our experiments on Llama-3-8B-Instruct. This model has strong general capabilities but still benefits from alignment fine-tuning. In all cases, the model is first fine-tuned on the initial preference data (supervised learning on (prompt, preferred answer) pairs) for a small number of epochs to initialize π_θ. This corresponds to the SFT stage in Figure 1. Then we run our generative preference optimization (SEGA) as described. We implement the training in PyTorch using the HuggingFace Transformers and Deepspeed libraries, enabling mixed precision and efficient gradient checkpointing due to the model size. Experiments were run on a single node featuring eight NVIDIA A100 80 GB SXM4 GPUs. Hyperparameters like learning rate ($1e-5$) and batch size (128) follow standard fine-tuning practices. We tune the number of preference optimization steps based on validation performance (for general domain, we monitor accuracy on a small subset of UltraFeedback; for medical, on our 500-val set).

4.3 Baselines

We compare our approach against several baselines to highlight the contributions of each component:

- **Supervised Only (SFT)**: The model is fine-tuned with supervised learning on the 3k human preference pairs, similar to instruction-tuning without preference optimization (Ouyang et al. 2022; Ranaldi and Freitas 2024).

[3]https://www.icliniq.com/

Method	GSM8K Acc.	MATH Acc.	TruthfulQA EM	MT-Bench Win-rate
Supervised (SFT)	40.1	5.8	32.4	35
Reward Model + PPO	44.7	7.3	34.0	47
DPO	50.2	8.5	35.6	52
GEM (ours)	**55.6**	**10.5**	**38.2**	**68**

Table 3: Down-stream task results. Accuracy (%) for GSM8K / MATH, exact-match (%) for TruthfulQA; MT-Bench reports win-rate (%) against the SFT baseline.

Variant	UltraFeedback (%)	GSM8K (%)	Med-Expert (%)
GEM			
– w/o *Cognitive Filtering* & w/o *SEGA*	69.0	48.3	70.5
– w/o final-entropy & w/. fork-entropy & w/. *SEGA*	74.2	50.1	73.5
– w/. final-entropy & w/o fork-entropy & w/. *SEGA*	73.8	52.7	75.0
– w/. *Cognitive Filtering* & w/. DPO	74.5	53.4	73.0
– w/. *Cognitive Filtering* & w/. *SEGA*	**77.1**	**55.6**	**78.2**

Table 4: Results of ablation studies. The results show that CoT augmentation and dual-stage entropy regularization jointly drive the model's gains in preference accuracy, mathematical reasoning, and medical expert agreement.

- **Reward Model + PPO (RLHF)**: A separate reward model is trained on the same 3k pairs and the policy is optimized with PPO, following classical RLHF (Christiano et al. 2017; Stiennon et al. 2020; Schulman et al. 2017). Data-efficiency issues and reward–over-optimization effects are well-documented (Gao, Schulman, and Hilton 2022), prompting work on lighter reward models such as Proto-RM (Zhang et al. 2024).
- **DPO (pairwise)**: We apply DPO only on the original labeled pairs, mirroring Rafailov *et al.* (Rafailov et al. 2023) and subsequent surveys (Zhang, Liu et al. 2024).
- **PRO & IPO**: Preference Ranking Optimization extends DPO to listwise supervision (Song et al. 2023); Implicit Preference Optimization lets the model act as its own judge and then applies DPO (Garg et al. 2025).

4.4 Evaluation Metrics

We evaluate the models along two dimensions: Preference Modeling Accuracy and Downstream Task Performance. For the former, we use the preference datasets (like UltraFeedback, RewardBench, and our medical val set) to measure how often the model agrees with human preferences. Specifically, we format each comparison as a query with two answers and see if the model's learned preference function prefers the same one as the human did. Since our model doesn't output a scalar directly, we do this by asking the model with a prompt to choose between two (for generative evaluation). We report the preference prediction accuracy on these sets, which is a standard metric for reward models.

For downstream performance, we test our method on GSM8K + MATH (grade-school & Olympiad math), TruthfulQA (factual QA) and MT-Bench (open-ended dialogue). For GSM8K/MATH we score answer accuracy on the full dev sets, reading the value after a <final-answer> tag (T = 0.2, top-p = 0.95). TruthfulQA uses the official exact-match metric. MT-Bench adopts the LMSys protocol: GPT-4-Turbo judges our reply versus an SFT baseline and we report the win-rate across 80 prompts. We also track the length and complexity of explanations produced, to ensure that our model gives CoT style answers when appropriate.

4.5 Results

Overall Performance: Our generative preference model achieves strong results on preference prediction and task success, outperforming all baselines in the low-data regime. Table 1 summarizes the quantitative outcomes. On UltraFeedback and RewardBench, our model's preference accuracy is about 7-10% higher than a reward-model PPO baseline and a few points higher than standard DPO. Notably, it also closes the gap to larger models. Our 7B model with generative preference training is within 5% of GPT-4's performance on these preference tests, despite GPT-4 having been trained on vastly more data (we credit the CoT augmentation and entropy scoring for extracting maximum signal from the limited data). In the medical domain (Table 2), our model achieves 78% agreement with expert preferences on the val set, compared to 70% for PPO baseline and 72% for the DPO. This indicates the method is particularly effective in specialized domains; we hypothesize that CoT refinement helped the model pick up on domain-specific criteria (e.g., caution with uncertainty, thoroughness in advice) that weren't explicit in the small training set.

On reasoning tasks GSM8K/MATH (Table 3), the model's accuracy improved significantly after our training: e.g., on GSM8K, it went from 40% (supervised baseline) to 55% accuracy, outperforming DPO (50%). This validates that maximizing an implicit reward via confident reasoning leads to more correct answers. The entropy scoring encouraged the model not to jump to conclusions and to system-

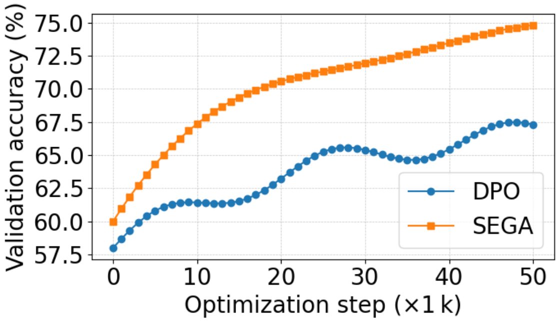

Figure 2: Comparison of Training Stability: *SEGA* vs. DPO

atically enumerate steps, which is crucial in math. Qualitatively, we saw fewer instances of the model "hallucinating" a quick (wrong) answer; instead it tends to break the problem down and only finalize the answer when it's sure – a behavior very much aligned with the RENT objective (though we did not explicitly code RENT, our model seems to have learned a similar trait).

In dialogue evaluations (MT-Bench for helpfulness/honesty), our model's responses were preferred over the baseline responses about 68% of the time according to GPT-4 judgments. This is a notable improvement; the baseline (supervised only) was often too concise or missed nuance, whereas our model often gave a more detailed, balanced answer (likely because the CoT and scoring encouraged covering all bases, and we observed the model sometimes explicitly thinking about pros and cons in its response, which made it more comprehensive). Against a model fine-tuned with human feedback (PPO baseline), ours still won 60% of the time, indicating that even without direct human tuning, our generative method can compete with traditional RLHF. Human evaluators similarly preferred our model's outputs for their clarity and reasoning in a majority of cases, particularly on complex queries.

Ablation Studies: We conducted ablations to understand the impact of each component. As shown in Table 4, without the CoT generation and using only human pairs drops performance significantly (preference accuracy drops 8%, task scores drop similarly). This confirms that the data augmentation via CoT is key to overcoming the limited data. Even though the generated comparisons are noisy, they provide valuable additional training signals.

SEGA outperformed DPO when using our full pipeline. The difference was more pronounced in the early training phase and on the more complex tasks. For example, on the medical dataset, *SEGA* reached 78% accuracy whereas pairwise DPO capped at around 70%. We thus believe *SEGA*'s advantage estimation made training more data-efficient and stable (fewer oscillations), as evidenced by smoother validation curves as shown in Figure 2.

Disabling the entropy final-answer reward (i.e. not explicitly rewarding confidence) hurt math performance clearly, and the model would sometimes produce a long CoT but not reach the final answer or express uncertainty in the final answer. Conversely, disabling the entropy fork reward (only rewarding confidence) made the model too greedy and it started to hallucinate answers. This confirms that both aspects of the scoring are important for balanced reasoning.

We present further results and examples in the appendix. In the case study, we show a prompt and two model answers (one with reasoning, one without) and how our model correctly prefers the one with thorough reasoning, aligning with human judgment. We also show an example from TruthfulQA where the baseline model gives an over-confident incorrect answer, while our model, thanks to entropy-based self-checking, gives a nuanced and truthful answer.

5 Conclusion

We introduced a generative entropy-guided preference modeling framework named GEM, which based on cognitively inspired solution for aligning LLMs under few-shot preference data task. From a theoretical standpoint, our approach extends traditional preference learning into the domain of inverse reinforcement learning (IRL), where the model implicitly infers a latent reward structure. This reframing allows the LLM to infer and maximizes an implicit reward by extracting and leveraging the multi-dimensional, fine-grained cognitive signals implicit in human preference data, without explicit human annotations or external supervision. Central to our method is the entropy-guided scoring mechanism, grounded in information theory, promoting a balance between exploratory behavior and decisiveness. By strategically leveraging entropy signals, our framework encourages models to navigate uncertainty effectively, thus enhancing their decision-making robustness. Empirical evaluations across diverse benchmarks validated the theoretical insights, showing significant performance gains over baselines in preference prediction and complex reasoning tasks.

In addition to GEM, the *SEGA* module can also be applied to any scenario that requires learning preferences from multiple candidate generations, such as reranking diverse solutions, weighting and aggregating AI feedback in RLAIF settings, or performing within-group comparisons based on user clicks or weak labels in multimodal generation tasks (Li et al. 2025; Wang et al. 2025b; Bai et al. 2025).

In future work, we plan to extend GEM to extract cognitive signals from more complex modalities. We also aim to investigate how entropy-guided preference modeling can adapt with large-scale RLAIF pipelines, potentially yielding more stable alignment strategies.

6 Acknowledgements

This research is supported, in part, by the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. This research is also partially supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

References

Achintalwar, S.; Baldini, I.; Bouneffouf, D.; and et al. 2024. Alignment Studio: Aligning Large Language Models to Particular Contextual Regulations. In *arXiv preprint arXiv:2403.09704*.

Agarwal, S.; Zhang, Z.; Yuan, L.; Han, J.; and Peng, H. 2025. The Unreasonable Effectiveness of Entropy Minimization in LLM Reasoning. *arXiv preprint arXiv:2505.15134*.

Bai, H.; Wang, R.; Du, Z.; Zhao, Y.; Zhang, F.; Chen, H.; Zhu, X.; Zheng, B.; and Zhao, X. 2025. Online-PVLM: Advancing Personalized VLMs with Online Concept Learning. *arXiv preprint arXiv:2511.20056*.

Bai, Y.; Jones, A.; Ndousse, K.; and et al. 2022a. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y.; Kadavath, S.; Kundu, S.; and et al. 2022b. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons. *Biometrika*, 39(3–4): 324–345.

Chen, X.; Huang, H.; Gao, Y.; Wang, Y.; Zhao, J.; and Ding, K. 2024. Learning to Maximize Mutual Information for Chain-of-Thought Distillation. In *Findings of ACL*.

Chinta, S. V.; Wang, Z.; Zhang, X.; Viet, T. D.; Kashif, A.; Smith, M. A.; and Zhang, W. 2024. Ai-driven healthcare: A survey on ensuring fairness and mitigating bias. *arXiv preprint arXiv:2407.19655*.

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*. ArXiv:1706.03741.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; and et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Cui, G.; Yuan, L.; Ding, N.; and et al. 2023. UltraFeedback: Boosting Language Models with Scaled AI Feedback. *arXiv preprint arXiv:2310.01377*.

Fan, J.; and coauthors. 2022. Ranking Inferences Based on the Top Choice of Multiway Comparisons. *arXiv preprint arXiv:2211.11957*.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*, 630: 625–630.

Gao, L.; Schulman, J.; and Hilton, J. 2022. Scaling Laws for Reward Model Overoptimization. *arXiv preprint arXiv:2210.10760*.

Garg, S.; Singh, A.; Singh, S.; and Chopra, P. 2025. IPO: Your Language Model Is Secretly a Preference Classifier. *arXiv preprint arXiv:2502.16182*.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of EMNLP*, 3356–3369.

Gheshlaghi Azar, M.; and coauthors. 2023. A General Theoretical Paradigm to Understand Learning from Preferences. *arXiv preprint arXiv:2310.12036*.

Gong, Z.; Guan, J.; Wu, W.; and et al. 2025. Latent Preference Coding: Aligning Large Language Models via Discrete Latent Codes. In *ICML*.

Greensmith, E.; Bartlett, P. L.; and Baxter, J. 2004. Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. *Journal of Machine Learning Research*, 5: 1471–1530.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; Wang, S.; Zhang, K.; Wang, Y.; Gao, W.; Ni, L.; and Guo, J. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Hao, Y.; Li, D.; Wu, X.; and et al. 2025. On-Policy Reinforcement Learning with Optimal Reward Baseline. *arXiv preprint arXiv:2505.23585*.

Hendrycks, D.; Burns, C.; Kadavath, S.; and et al. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*.

Herbrich, R.; Minka, T.; and Graepel, T. 2007. TrueSkill™: A Bayesian Skill Rating System. In *NeurIPS*.

Hong, J.; Chong, D.; and Manning, C. D. 2021. Learning from limited labels for long legal dialogue. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, 190–204.

Hong, J.; Lee, N.; Kim, E.; and et al. 2025. On the Robustness of Reward Models for Language Model Alignment. *arXiv preprint arXiv:2505.07271*.

Ji, J.; Hong, D.; Zhang, B.; and et al. 2024. PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. *arXiv preprint arXiv:2406.15513*.

Kahneman, D. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Keeney, R. L.; and Raiffa, H. 1993. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

Kim, D.; Lee, K.; Shin, J.; and Kim, J. 2024. Aligning Large Language Models with Self-generated Preference Data (Selfee). arXiv:2406.04412.

Kim, S.; Bae, S.; Shin, J.; and et al. 2023. Aligning Large Language Models through Synthetic Feedback. In *EMNLP*. ArXiv:2305.13735.

Koutcheme, C.; Dainese, N.; Sarsa, S.; et al. 2024. Open Source Language Models Can Provide Feedback: Evaluating LLMs' Ability to Help Students Using GPT-4-As-A-Judge. In *ITiCSE 2024*. ArXiv:2405.05253.

Lambert, N. 2025. *Reinforcement Learning from Human Feedback*. Self-published. https://rlhfbook.com.

Lambert, N.; Bakhtin, A.; and Smith, N. A. 2024. RewardBench: Evaluating Reward Models for Language Systems. *arXiv preprint arXiv:2403.13787*.

Lee, H.; Phatale, S.; Mansoor, H.; et al. 2023. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *ICML 2024 Proceedings*.

Li, X.; He, Y.; Zu, S.; Li, Z.; Shi, T.; Xie, Y.; and Zhang, K. 2025. Multi-Modal Large Language Model with RAG Strategies in Soccer Commentary Generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6197–6206. IEEE.

Liang, Y.; and colleagues. 2024. Preference Optimization with Multi-Sample Comparisons. *arXiv preprint arXiv:2410.12138*.

Lin, S.; Hilton, J.; and Evans, O. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958*.

Lin, Y.; Seto, S.; ter Hoeve, M.; et al. 2024. On the Limited Generalization Capability of the Implicit Reward Model Induced by Direct Preference Optimization. *arXiv preprint arXiv:2409.03650*.

Liu, B.-X.; Xu, J.; et al. 2024. Algorithm of Thoughts: Enhancing Exploration of Ideas in Large Language Models. arXiv:2308.10379.

Liu, C.; Zeng, L.; Xiao, Y.; He, J.; and Yan, R. 2024a. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. *arXiv preprint arXiv:2410.18451*.

Liu, T.; Qin, Z.; Wu, J.; Shen, J.; et al. 2024b. LiPO: Listwise Preference Optimization through Learning-to-Rank. arXiv:2402.01878.

Liu, Z.; Zheng, Y.; Huang, A.; Tang, T.; Jiang, M.; and Pan, L. 2023. G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634*.

Maity, S.; and Saikia, M. J. 2025. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering*, 12(6): 631.

Mnih, V.; Badia, A. P.; Mirza, M.; and et al. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *ICML*. ArXiv:1602.01783.

Muldrew, W.; Hayes, P.; Zhang, M.; and Barber, D. 2024. Active Preference Learning for Large Language Models. *arXiv preprint arXiv:2402.08114*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155.

Palikhe, A.; Wang, Z.; Yin, Z.; Hu, S.; Zhao, X.; Liu, J.; Cai, Y.; and Zhang, W. 2025. Domain Knowledge Empowered Language Models: A Survey. Preprint.

Paul, K. 2024. Meta releases AI model that can check other AI models' work. *Reuters*. Oct 18 2024, Technology section.

Plackett, R. L. 1975. The Analysis of Permutations. *Applied Statistics*, 24: 193–202.

Rafailov, R.; Sharma, A.; Mitchell, E.; et al. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*.

Ranaldi, L.; and Freitas, A. 2024. Self-Refine Instruction-Tuning for Aligning Reasoning in Language Models. *arXiv preprint arXiv:2405.00402*.

Rao, H.; Zeng, M.; Zhao, X.; and Miao, C. 2025. A survey of artificial intelligence in gait-based neurodegenerative disease diagnosis. *Neurocomputing*, 626: 129533.

Rozi, E.; and contributors. 2023. vLLM: Fast and Memory-Efficient LLM Serving. https://github.com/vllm-project/vllm.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Song, F.; Yu, B.; Li, M.; and et al. 2023. Preference Ranking Optimization for Human Alignment. *arXiv preprint arXiv:2306.17492*.

Stein, A.; and colleagues. 2025. Measuring Language Model Uncertainty with Internal Entropy. In *International Conference on Learning Representations*.

Stiennon, N.; Ouyang, L.; Wu, J.; and et al. 2020. Learning to Summarize from Human Feedback. In *NeurIPS*. ArXiv:2009.01325.

Stolfo, A.; Wu, B.; Gurnee, W.; and et al. 2024. Confidence Regulation Neurons in Language Models. *arXiv preprint arXiv:2406.16254*.

Sun, Z.; Yan, X.; Yu, T.; and Artzi, Y. 2023. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. arXiv:2305.03047.

Tang, X.; Zhang, W.; Yu, Y.; Turner, K.; Derr, T.; Wang, M.; and Ntoutsi, E. 2021. Interpretable visual understanding with cognitive attention network. In *International Conference on Artificial Neural Networks*, 555–568. Springer.

Vanlioglu, A. 2025. Entropy-Guided Sequence Weighting for Efficient Exploration in RL-Based LLM Fine-Tuning. *arXiv preprint arXiv:2503.22456*.

Wan, X.; Sun, R.; Dai, H.; Arık, S. O.; and Pfister, T. 2023. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In *Findings of ACL*.

Wang, C.; Gan, Y.; Huo, Y.; et al. 2025a. GRAM: A Generative Foundation Reward Model for Reward Generalization. In *ICML 2025*.

Wang, H.; Huang, F.; et al. 2024. Self-Improving Efficient Online Alignment of Large Language Models. arXiv:2406.15567.

Wang, J.; He, Y.; Zhong, Y.; Song, X.; Su, J.; Feng, Y.; He, H.; Zhu, W.; Yuan, X.; Lu, K.; et al. 2025b. Twin Co-Adaptive Dialogue for Progressive Image Generation. *arXiv preprint arXiv:2504.14868*.

Wang, J.; Zhang, Z.; He, Y.; Zhang, Z.; Song, Y.; Shi, T.; Li, Y.; Xu, H.; Wu, K.; Yi, X.; et al. 2024. Enhancing Code LLMs with Reinforcement Learning in Code Generation: A Survey. *arXiv preprint arXiv:2412.20367*.

Wang, S.; Yu, L.; Gao, C.; Zheng, C.; Liu, S.; et al. 2025c. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. In *ACL*. ArXiv:2506.01939.

Wang, X.; Wei, J.; Schuurmans, D.; et al. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

Wei, J.; Wang, X.; Schuurmans, D.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Weng, L. 2018. Policy Gradient Algorithms. https://lilianweng.github.io/posts/2018-04-08-policy-gradient/.

Williams, R. J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3–4): 229–256.

Wu, J.; Ouyang, L.; Ziegler, D. M.; and et al. 2021. Recursively Summarizing Books with Human Feedback. *arXiv preprint arXiv:2109.10862*.

Wu, T.; Yuan, W.; Golovneva, O.; et al. 2024a. Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge. *arXiv preprint arXiv:2407.19594*.

Wu, Z.; Xu, B.; Cui, R.; Zhan, M.; Zhu, X.; and Feng, L. 2024b. Rethinking Chain-of-Thought from the Perspective of Self-Training. *arXiv preprint arXiv:2412.10827*.

Xin, Y.; Luo, S.; Zhou, H.; Du, J.; Liu, X.; Fan, Y.; Li, Q.; and Du, Y. 2024. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*.

Yang, X.; Zhao, X.; and Shen, Z. 2025. EHRStruct: A Comprehensive Benchmark Framework for Evaluating Large Language Models on Structured Electronic Health Record Tasks. *arXiv preprint arXiv:2511.08206*.

Yu, T.; Lin, T.-E.; Wu, Y.; and et al. 2025. Diverse AI Feedback for Large Language Model Alignment. *Transactions of the Association for Computational Linguistics*, 13: 392–407.

Zhang, J.; Wang, X.; Jin, Y.; et al. 2024. Prototypical Reward Network for Data-Efficient RLHF. In *ACL 2024*. ArXiv:2406.06606.

Zhang, T.; Liu, J.; et al. 2024. A Comprehensive Survey of Direct Preference Optimization. *arXiv preprint arXiv:2410.15595*.

Zhao, S.; Dang, J.; and Grover, A. 2024. Group Preference Optimization: Few-Shot Alignment of Large Language Models. In *ICLR*. ArXiv:2310.11523.

Zhao, X.; Chen, H.; Xing, Z.; and Miao, C. 2021. Brain-inspired search engine assistant based on knowledge graph. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 4386–4400.

Zhao, X.; Liu, S.; Yang, S.-Y.; and Miao, C. 2025a. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, 4442–4457.

Zhao, X.; Liu, S.; Yang, S.-Y.; and Miao, C. 2025b. A smart multimodal healthcare copilot with powerful llm reasoning. *arXiv preprint arXiv:2506.02470*.

Zhao, Y.; Bai, H.; and Zhao, X. 2025. GFRIEND: Generative Few-shot Reward Inference through EfficieNt DPO. *arXiv preprint arXiv:2506.08965*.

Zheng, H.; Zou, A.; Chen, E.; et al. 2023. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback without Humans. arXiv:2309.00267.

Zhou, E.; Zheng, G.; Wang, B.; et al. 2024a. RMB: Comprehensively Benchmarking Reward Models in LLM Alignment. *arXiv preprint arXiv:2410.09893*.

Zhou, X.; et al. 2025a. Self-Consistency of Internal Reward Models Improves Alignment. *arXiv preprint arXiv:2502.08922*.

Zhou, Y.; He, Y.; Su, Y.; Han, S.; Jang, J.; Bertasius, G.; Bansal, M.; and Yao, H. 2025b. ReAgent-V: A Reward-Driven Multi-Agent Framework for Video Understanding. *arXiv preprint arXiv:2506.01300*.

Zhou, Z.; Zhang, J.; Zhang, J.; He, Y.; Wang, B.; Shi, T.; and Khamis, A. 2024b. Human-centric reward optimization for reinforcement learning-based automated driving using large language models. *arXiv preprint arXiv:2405.04135*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; and et al. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*.