ICE-GRT: Instruction Context Enhancement by Generative Reinforcement based Transformers

Anonymous ACL submission

Abstract

The emergence of Large Language Models (LLMs) such as ChatGPT and LLaMA encounter limitations in domain-specific tasks, with these models often lacking depth and accuracy in specialized areas, and exhibiting a decrease in general capabilities when fine-tuned, particularly analysis ability in small sized models. To address these gaps, we introduce ICE-GRT, utilizing Reinforcement Learning from Human Feedback (RLHF) grounded in Proximal Policy Optimization (PPO), demonstrating remarkable aptitude in in-domain scenarios without compromising general task performance. Our exploration of ICE-GRT highlights its understanding and reasoning ability to not only generate robust answers but also to provide detailed analyses of the reasons behind the answer. This capability marks a significant progression beyond the scope of Supervised Fine-Tuning models. The success of ICE-GRT is dependent on several crucial factors, including Appropriate Data, Reward Size Scaling, KL-Control, Advantage Normalization, etc. The ICE-GRT model exhibits state-of-the-art performance in domain-specific tasks and across 12 general Language tasks against equivalent size and even larger size LLMs, highlighting the effectiveness of our approach. We provide a comprehensive analysis of the ICE-GRT, underscoring the significant advancements it brings to the field of LLM.

1 Introduction

The advent of Large Language Models (LLMs) like ChatGPT (Brown et al., 2020; OpenAI, 2023) and LLaMA (Touvron et al., 2023a,b) has marked a significant milestone in the field of Natural Language Processing (NLP). These models have gained widespread recognition for their robust general conversational abilities, enabling fluid and coherent responses across a diverse range of topics. However, there are key limitations to these models. Firstly, a key limitation surfaces when these models encounter domain-specific tasks (Zhao et al., 2023; Zhang et al., 2023a). In scenarios that demand deep technical knowledge or specialized expertise, these models often fall short, providing responses that lack necessary depth and accuracy. Secondly, Supervised Fine Tune (SFT) LLMs tend to exhibit a decrease in general capabilities (Ling et al., 2023). This is contrary to the expectations held for large-scale models, which are presumed to either maintain or improve their performance in a wide array of tasks (Pan et al., 2023a). Lastly, the current smaller-sized LLMs, such as 13 Billion, demonstrate a limited ability to conduct detailed analysis on complex questions, a competency that is significantly inferior compared to the capabilities of models like ChatGPT, which can engage in more comprehensive and detailed discussions.

Addressing these challenges, we introduce the **I**nstruction **C**ontext **E**nhancement by **G**enerative **R**einforcement based **T**ransformers (ICE-GRT), an innovative LLM that leverages the principles of Reinforcement Learning from Human Feedback (RLHF) (Brown et al., 2020) based on Proximal Policy Optimization (PPO) (Schulman et al., 2017). While ensuring that the general capabilities of the Large Language Model (LLM) are maintained, ICE-GRT exhibits exceptional performance in several domain-specific scenarios. Furthermore, ICE-GRT demonstrates an improved ability for detailed analysis, particularly in complex scenarios where smaller-sized LLMs fall short.

We take one domain-specific task of ad moderation as an example. ICE-GRT can not only determine the compliance of advertisements but also identify the specific category of violation. Moreover, it goes a step further by detailed analyzing which elements of the ad are problematic and offers constructive modification suggestions. This is a notable advancement over both pretrained and SFT (Chiang et al., 2023) LLM models, which are typically limited to identifying compliance and violation categories.

When our training methodology was applied to RLHF, we observed not just significant improvements in in-domain tasks but also a surprising enhancement in general tasks. In a comparative analysis against models of equivalent and larger parameter size across many general tasks, our ICE-GRT model with 13 billion parameters consistently achieved state-of-the-art performance in 12 well-known public LLM evaluation benchmarks.

Our exploration of the ICE-GRT model has uncovered several factors critical to its training success. The ICE-GRT model's training data, sourced from our ICE-Instruct (SFT) model and enriched with human feedback with strict evaluation criteria, offers a diverse and comprehensive dataset, essential for its robust training. Moreover, the scaling of the reward model is essential for accurately capturing complex scenarios and aligning with human preferences in RLHF. Additionlly, KL-Control is key to regulating the balance between the models, while Advantage Normalization significantly improves learning stability by adjusting advantage estimates. Additionally, we discovered that modifying the Clipping Range and carefully controlling the maximum response length during sampling are vital for enhancing the training process. These findings deepen our understanding of RLHF mechanisms and are instrumental in effectively training the ICE-GRT model.

Moreover, we provide a detailed analysis of the ICE-GRT model, encompassing both general and in-domain capabilities. Through this exploration, we aim to contribute a novel perspective and methodology to the field of NLP, particularly in enhancing the depth and accuracy of domain-specific task handling by large language models. We observe that the pretrain phase engages in “knowledge learning”, where the model extensively absorbs a diverse range of information, forming a substantial foundational knowledge base. Subsequently, in the Supervised Fine-Tuning stage, the model engages in “knowledge mining”, where it utilizes the learned knowledge in response to specific instructions. This stage is crucial for the model to transition from passive knowledge accumulation to active knowledge application. Finally, the RLHF phase engages in “knowledge enhancement”, enhancing the model's ability to align with human language preferences. This stage builds upon the vast knowledge gained in the pretrain phase and the knowledge mining from the SFT stage, leading to a model that not only reconstruct extensive knowledge but also excels in applying it with human-centric preference. Importantly, this phase showcases a significant leap in the model's emergence capabilities.

In our commitment to fostering collaborative research and innovation, **we will make ICE-GRT publicly available on HuggingFace**. This open-source initiative is aimed at empowering researchers globally to further investigate and expand upon our findings with ICE-GRT. By democratizing access to this advanced model, we hope to inspire and facilitate worldwide exploration and progress in language model research. This paper unveils just a fraction of ChatGPT's capabilities, and our choice of the acronym "ICE" for ICE-GRT is purposeful. It represents our aspiration to accelerate the ’ice-breaking’ process in LLM research, symbolizing our desire to inspire researchers to explore and uncover the vast potential of ICE-GRT across an array of tasks and paving the way for new discoveries and advancements in the field.

2 Related Works

2.1 Instruction-Tuning for LLM

Recent advancements in Large Language Model (LLM) development have increasingly focused on instruction-tuning (Chiang et al., 2023), a technique that is gaining significant traction particularly within the realms of Question Answering (QA) and different domains (Zhao et al., 2023; Pan et al., 2023b; Qiu et al., 2020). Key research in this area includes works such as ALPACA (Taori et al., 2023), Vicuna (Chiang et al., 2023), and (Zhang et al., 2023b), which explores the balance between diveristy and accuracy in large language model. Furthermore, studies like (Sun et al., 2023) delve into principles of effective QA strategies, while (Zhou et al., 2023) present LIMA, an innovative model for language interaction. In the sphere of conversational interfaces, significant contributions include the development of OpenAssistant by (Köpf et al., 2023; Chiang et al., 2023).

2.2 Reinforcement Learning from Human Feedback (RLHF)

Alongside the development of LLMs, Reinforcement Learning from Human Feedback has emerged as an important approach to improve LLMs (Brown et al., 2020; Touvron et al., 2023b). RLHF involves

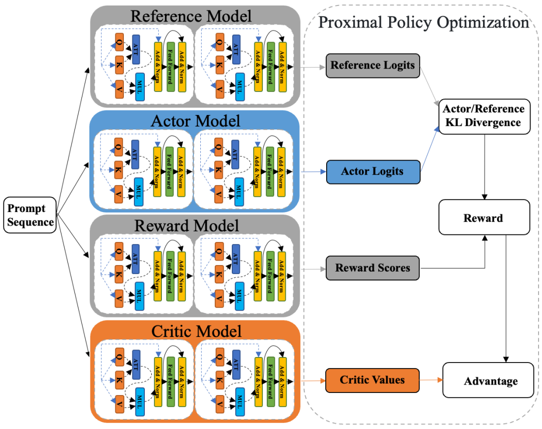


Figure 1: ICE-GRT Model Architecture.

training models not just on static datasets but also incorporating human feedback to guide the learning process. This method has been particularly useful in aligning knowledge learning and mining with human feedback. For instance, models like OpenAI's InstructGPT have utilized RLHF to tailor responses based on human preferences, leading to more accurate outputs (Stiennon et al., 2020).

3 Model

In this section, we briefly introduce a SFT model we have trained, named ICE-Instruct, designed to improve the domain-specific knowledge mining capabilities of pre-trained LLMs. Following this, we will give a detailed description of our process for training the reward model, which we have termed ICE-Reward. Finally, we will comprehensively introduce the entire training process of ICE-GRT, including some important training strategies.

3.1 ICE-Instruct

The ICE-Instruct model built upon the Vicuna model (Chiang et al., 2023). By blending in-domain and general-purpose data during fine-tuning, it excels in both specialized tasks and broader tasks. This approach not only maintains its vast linguistic capacities but also enhances its expertise in specific domains. Importantly, this sets a solid foundation for RLHF models. All subsequent actor and critic models are initialized using ICE-Instruct as backbone. In essence, ICE-Instruct determines the lower-bound capabilities of ICE-GRT, ensuring a strong and reliable baseline for further advancements. To maximize the model's applicability in contextual interactions, we have converted all collected data into Question-Answer pairs. Each data point adheres to a prompt format that begins with *"Below is an instruction that describes a task. Write a response that appropriately completes the request. ### USER: <INPUT> ASSISTANT: <OUTPUT> "*, ensuring consistency and relevance in contexts.

3.2 ICE-Reward

Response Generation and Sampling: Initially, for each prompt in the RLHF training dataset, we generate five responses. These responses are uniquely produced by our **ICE-Instruct** model. By sampling from the model's output distribution, we ensure a diverse range of generated answers , capturing various aspects of potential responses.
Human Annotation and Ranking: The generated responses are then subjected to human annotation. Annotators rank these responses according to predefined criteria detailed in section 4.3. Specifically, we labeled 20,000 sets of rankings, each set containing five responses. From the ranked responses, we extract the top two and the bottom two responses for each prompt. These are then paired to form training data. The pairs consist of a "better" response and a "worse" response, as determined by the human annotation. This pairing strategy is instrumental in teaching the model the differences between high-quality and low-quality responses.
Training Reward Model: The objective of training reward model is to develop a model capable of accurately differentiating between high and low-quality responses. Let $R(s, a)$ be the reward function, where s represents the input prompt and a the generated response. Our goal is to optimize R so that it aligns with human judgments. The training data consists of pairs (a_i, a_j) where a_i is a higher-ranked response compared to a_j for the same prompt. We use a pairwise ranking loss function, defined as:

$$\mathcal{L}(a_i, a_j) = \max(0, \text{margin} - R(s, a_i) + R(s, a_j)).$$

This loss function encourages the model to assign a higher score to a_i than a_j.

The trained reward model, therefore, learns to assign higher scores to more relevant and contextually appropriate responses, as per human rankings. This model forms a most critical part of our system, ensuring high-quality, context-aware responses.

3.3 ICE-GRT

In this section, we provide a comprehensive overview of each component involved in ICE-GRT, leverages the principles of RLHF (Brown et al., 2020) based on PPO (Schulman et al., 2017), along with their respective mathematical formulations. Figure 1 shows the whole training process.

Actor Model: The Actor model, represented as $\pi_{\theta_{\text{act}}}(a|s)$, maps states s to actions a. It is responsible for generating actor logits, which are scores assigned to each potential action.

Reference Model: The Reference model, denoted as $\pi_{\theta_{\text{ref}}}(a|s)$, serves as a pre-trained benchmark for evaluating behavior. It provides a baseline against which the Actor model's outputs are compared throughout the training process.

Reward Model: The Reward model, expressed as $R(s, a)$, assigns a reward score based on the quality of the generated sequence, evaluating both the action a and the state s.

Critic Model: The Critic model, $V_{\theta_{\text{crt}}}(s)$, estimates the value of being in a specific state s, thereby producing critic values that guide the learning process.

3.3.1 Generalized Advantage Estimation (GAE) Calculation in ICE-GRT

The advantage function, $A(s, a)$, assesses the relative benefit of executing a specific action in contrast to the average action in a given state. The formula for calculating the Advantage is:

$$A(s,a) = \mathbb{E}(R(s,a) + \gamma V_{\theta_{\text{crt}}}(s') - V_{\theta_{\text{crt}}}(s)) \quad (1)$$

where γ represents the discount factor, s' is the subsequent state following the current state s, and $V_{\theta_{\text{crt}}}(s)$ is the value function estimated by the Critic model with weights θ_{crt}.

Generalized Advantage Estimation (GAE), enhances the estimation of the advantage function in RL (Schulman et al., 2015). GAE blends multi-step return methods with value function estimates to mitigate variance while preserving a reasonable bias. The essence of GAE is the employment of a weighted sum of n-step Temporal Difference (TD) residuals:

$$\delta_t^A = \mathbb{E}(R^{t+1}(s,a) + \gamma V_{\theta_{\text{crt}}}^{t+1}(s') - V_{\theta_{\text{crt}}}^{t}(s)) \quad (2)$$

Here, δ_t^A represents the TD residual at time t. Further, the GAE advantage function is calcuated as: $A_{\text{GAE}}(s,a) = \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}^A$, where $\lambda \in (0, 1)$.

3.3.2 Actor Model Learning

The Actor Model is updated using the Proximal Policy Optimization objective (Schulman et al., 2017), the process is calculated as follows:

$$L(\theta_{\text{act}}) = \min\Bigg(\frac{\pi_{\theta_{\text{act}}}(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A_{\text{GAE}}^{\pi_{\theta_{\text{old}}}}(s,a), \quad \text{clip}\left(\frac{\pi_{\theta_{\text{act}}}(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}, 1-\varepsilon, 1+\varepsilon\right) A_{\text{GAE}}^{\pi_{\theta_{\text{old}}}}(s,a)\Bigg), \quad (3)$$

where $A_{\text{GAE}}^{\pi_{\theta_{\text{old}}}}(s,a)$ is the advantage function calculated using the old policy $\pi_{\theta_{\text{old}}}$, $\varepsilon \in (0, 1)$ is a hyperparameter. This term ensures that the evolving Actor policy remains not only stable in its updates but also aligned or divergent as desired from the old model.

3.3.3 Policy Optimization and Training

In the final stage, the PPO algorithm optimizes the Actor model's policy based on the calculated advantages, the KL-divergence, and the updated Actor model. The policy is iteratively updated to maximize the expected rewards, with the aim of aligning the Actor model's behavior more closely with established benchmarks while also ensuring effective and efficient learning.

3.3.4 Important Training Strategies

ICE-GRT Training Data: Our ICE-GRT's training data originates from ICE-Instruct model and careful human feedback annotation. This data is not just a collection of responses but is intricately designed to encompass a wide range of scenarios. Each prompt within the ICE-Instruct model is responded to with a set of diverse answers, generated by sampling from the model's output distribution. This method ensures a comprehensive and varied dataset, essential for robust model training. The responses are further refined through a meticulous human annotation process, where experts rank them based on predefined criteria. This rigorous approach ensures the model is trained on high-quality, human-verified data, which is crucial for the model's ability to understand and apply complex information. More details and experimental comparsions are described in Section 5.2.1.

Reward size Scaling: In ICE-GRT, the scaling of the reward model is a critical factor in determining the overall effectiveness and efficiency of training. A larger reward model, denoted as $R_\psi(s, a)$, where ψ represents the model parameters, is significant for several reasons. Firstly, larger reward model can better capture complex environments and actions, essential in RLHF where the reward signal must accurately reflect human preferences and detailed task requirements. Secondly, larger scale of reward size aids in generalizing across diverse prompts. This is vital for consistent performance in various scenarios, especially in ICE-GRT.

KL-Control (Schulman et al., 2017) is a crucial mechanism in PPO, especially when training with human feedback. A key aspect of KL-Control in this context is the regulation of divergence between the Actor and the Reference models. The KL divergence between these two models is monitored and controlled to ensure that the policy evolution adheres closely to the human feedback. Moreover, ICE-GRT training includes a clipping mechanism to avoid large, potentially destabilizing updates in the value function. This ensures that changes in the value function are moderate and accurately reflect real improvements as assessed by the Critic. Furthermore, as an additional measure, KL Reward adjustment helps keep the actor model on the desired path as defined by human feedback. This aligns actor model updates more closely with human preferences.

Advantage Normalization enhances learning stability and efficiency in PPO-based RLHF. It adjusts the advantage estimates, making them more consistent and less variable. This is particularly beneficial in RLHF, where human feedback can introduce unpredictable variations. Normalizing the advantage helps the model to focus on the most relevant learning signals, leading to faster and more stable convergence. The formula for Advantage Normalization is shown as follows:

$$\hat{A}_t^{\pi_\theta} = \frac{A_t^{\pi_\theta} - \mu_{A^{\pi_\theta}}}{\sigma_{A^{\pi_\theta}}},$$

where $\hat{A}_t^{\pi_\theta}$ represents the normalized advantage at time t, $A_t^{\pi_\theta}$ is the original advantage at time t, $\mu_{A^{\pi_\theta}}$ is the mean of the advantage, $\sigma_{A^{\pi_\theta}}$ is the standard deviation of the advantage.

4 Experimental Details

Our training process utilized the power of 64 A100 GPUs, employing a multi-node, multi-GPU strategy to conduct ICE-GRT. Our models were trained and stored using the bf16 precision format. The learning rates were finely selected, with the actor learning rate set at $5e-6$ and the critic learning rate at $5e-7$. We maintained a clipping range of 0.2. The discount factor γ was kept constant at 0.95, ensuring optimal balance in our training. We are excited to announce the upcoming release and open-sourcing of our ICE-GRT 13B model on Hugging Face, specifically tailored for scientific research purposes.

4.1 Data Collection

For our training corpus, we have crafted a novel mix of datasets. This includes a selection from publicly available resources, complemented by in-domain data. We have removed all the sensitive information, including usernames, email addresses, and personal details, to uphold the data privacy and security. In essence, the dataset we have prepared for reward model and RLHF model is diverse and multi-faceted, covering a range of domains. It includes data relevant to public and domain-specific question-answering scenarios, as well as tasks involving multilingual data alignment. We generated 5 distinct responses for every prompt in our data collection, utilizing our **ICE-Instruct** model. This process involves sampling from the model's output distribution, which guarantees a varied spectrum of answers. To optimally train our reward model, the data labelers carefully conducted manual labeling of the rankings for the 5 distinct responses on 20,000 prompts. To enhance the human-annotation accuracy and reduce subjectivity among labelers, each prompt was independently evaluated by three labelers, establishing a thorough and reliable validation processverification process.

4.2 General Task Evaluation

Our evaluation of ICE-GRT using the GPT-Fathom framework (Zheng et al., 2023) focused on public general tasks. The objective was to benchmark ICE-GRT's performance against existing models and to understand its position in the landscape of current LLMs. We employed 12 benchmarks, which span across various capability categories such as language understanding, reasoning, etc. These benchmarks were carefully chosen to test a wide range of abilities, from basic language processing to complex problem-solving and decision-making tasks. In our evaluation, we maintained alignment with the settings used in GPT-Fathom to ensure a fair and accurate comparison. This involved employing similar input formats, evaluation metrics, and environmental conditions.

4.3 Manual Annotation-Based Evaluation

Our study incorporates a rigorous evaluation criteria, with a special emphasis on manual annotation

Model	**MMLU 5-shot**	**AGIEval few-shot**	**BBH 3-shot**	**AGIEval-ZH few-shot**	**ARC-E 1-shot**	**ARC-C 1-shot**	**HellaSWAG 1-shot**	**Winogrande 1-shot**	**RACE-M 1-shot**	**RACE-H 1-shot**	**GSM8K 8-shot**	**Math 4-shot**
LLaMA 7B	24.66%	20.05%	33.48%	23.68%	30.01%	26.71%	24.58%	50.36%	26.74%	29.19%	13.80%	0.36%
Llama2 7B	40.91%	25.97%	38.21%	26.21%	62.37%	48.46%	25.39%	50.36%	45.75%	39.54%	17.51%	0.08%
Vicuna 7B	38.49%	22.71%	37.26%	27.00%	69.74%	46.33%	17.37%	49.80%	50.21%	46.83%	21.68%	0.96%
ICE-Instruct 7B	26.30%	15.95%	39.00%	31.14%	67.63%	45.31%	3.10%	36.07%	53.55%	52.09%	35.48%	0.82%
LLaMA 13B	38.42%	26.78%	38.28%	25.51%	67.63%	49.23%	28.90%	47.51%	52.23%	48.51%	18.42%	0.42%
Llama2 13B	49.57%	34.85%	45.89%	32.93%	76.52%	55.63%	37.17%	52.17%	57.73%	55.09%	28.66%	0.44%
Vicuna 13B	35.84%	28.68%	39.27%	30.33%	60.23%	40.96%	0.03%	5.84%	59.19%	60.69%	24.56%	0.66%
ICE-Instruct 13B	50.08%	24.51%	48.09%	34.15%	85.19%	66.89%	19.30%	47.99%	72.14%	56.52%	47.08%	**1.02%**
ICE-GRT 13B	**55.33%**	**34.92%**	**49.78%**	**34.23%**	**87.58%**	**70.99%**	**39.37%**	**53.04%**	**75.91%**	**71.64%**	**51.48%**	0.92%
LLaMA 30B	50.38%	34.87%	49.70%	30.68%	82.41%	60.67%	31.31%	51.30%	65.18%	64.18%	35.10%	0.58%
Llama2-70B	64.72%	43.99%	65.22%	39.52%	93.43%	79.61%	68.45%	69.69%	87.60%	85.13%	56.56%	3.72%

Table 1: Evaluating Benchmark Performance of Large Language Models in General Language Tasks.

for assessing the capabilities of LLMs, particularly in different applications. The criteria evaluates responses in 8 essential categories, utilizing a scoring mechanism that prioritizes the most crucial aspects.
Clarity: Responses should be straightforward and precise, ensuring easy comprehension through specific, appropriate language.
Accuracy: The responses are expected to align closely with verified facts, as assessed by manual annotators. Actual fact can be validated.
Completeness: Evaluated for covering all aspects of the inquiry, providing comprehensive details for informed decision-making.
Safety: Focuses on ensuring no personal data is mishandled, with manual checks for data privacy.
Courtesy: Responses should be politically correct. e.g., gender identity, ethnic groups, etc.
Comfortableness: Responses must maintain a polite and respectful tone, containing inclusive vocabulary and reflect diversity at all times..
Conciseness: Emphasizes brevity in responses, without compromising on clarity or accuracy.
Context: Response must be related to the topic and relevant to the question.

Table 2 shows the weight and score of each categories to evaluate these criteria accurately, ensuring responses quality and relevance.

Evaluation	Positive	Neutral	Negative	Weights
Clarity	5	2	0	6
Accuracy	5	2	0	6
Completeness	5	2	0	6
Safety	5	2	0	3
Courtesy	5	2	0	3
Comfortableness	5	2	0	3
Conciseness	5	2	0	1
Context	5	2	0	1

Table 2: Manual Annotation-Based Evaluation Criteria.

5 Results and Analysis

5.1 Results

Benckmarks Scores on General Tasks: Our analysis focuses on the performance of ICE-GRT 13B, as compared to other models in similar and higher capacity categories. As is shown in Table 1, our ICE-GRT 13B model demonstrates significant improvements over the LLaMa, Llama 2, Vicuna 13B and LLaMa 30B in both its pretrained and SFT across various general benchmarks, such as MMLU (Hendrycks et al., 2021), AGIEval (Zhong et al., 2023), BBH (Srivastava et al., 2022), ARC (Xu et al., 2023), HellaSWAG (Zellers et al., 2019), RACE (Lai et al., 2017), etc. It shows remarkable advancements in general language understanding and reasoning tasks, indicating enhanced comprehension and reasoning capabilities. Remarkably, the ICE-GRT 13B model has significantly narrowed the gap with the much larger Llama2 70B pretrain model. This comparison underscores the effectiveness of the ICE-GRT, compensating for smaller model size with more generalization capabilities. The success of the ICE-GRT models suggests that the methodology, which likely includes components of human feedback and alignment, contributes significantly to the models' ability to understand and respond to complex prompts, a factor that is not solely dependent on model size.

Human-Annotated Scores on In-Domain Task: In the in-domain evaluation presented in Table 3, ICE-GRT distinctly outperforms Llama2 SFT 13B and ICE-Instruct 13B across several critical dimensions. Notably, ICE-GRT achieves the highest scores in clarity (98.1%), accuracy (97.0%), and completeness (92.9%), underscoring its exceptional ability to deliver precise, comprehensive, and understandable responses. While it scores slightly lower in safety and comfort compared to its counterparts, it still maintains a high standard in these areas. The overall score of 95.5% for ICE-GRT is a testament to its superior performance, significantly surpassing Llama2 SFT 13B (86.3%) and ICE-Instruct 13B (87.3%). This robust performance across multiple metrics confirms the introductory claims about ICE-GRT's capabilities, particularly in handling domain-specific tasks with a level of depth and precision not seen in current models.

	Llama2 sft	ICE-Instruct	ICE-GRT
Clarity	95.9%	88.5%	**98.1%**
Accuracy	77.4%	84.44%	**97.0%**
Completeness	64.8%	71.11%	**92.9%**
Safety	96.6%	**100%**	92.2%
Courtesy	100%	95.9%	**100%**
Comfortable	96.6%	**98.1%**	92.22%
Conciseness	**95.1%**	93.33%	91.8%
Context	**98.8%**	94.0%	98.1%
Overall Score	86.3%	87.3%	**95.5%**

Table 3: Evaluating human-assessed scores for in-domain Large Language Models.

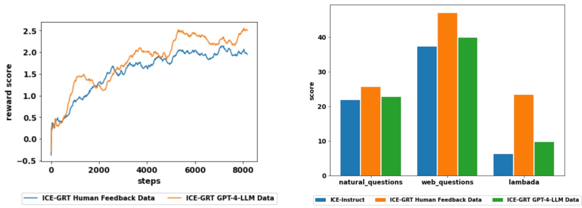
5.2 Detailed Analysis

5.2.1 The importance of ICE-GRT Training Data

In the training of the ICE-GRT, we employed two distinct datasets for RLHF. The first dataset was uniquely produced by our ICE-Instruct model. For each prompt, five diverse responses were generated by sampling from the model outputs. These responses were then subjected to human annotation, where annotators ranked them according to predefined criteria. The second dataset originated from the GPT-4-LLM (Peng et al., 2023). It included ranked responses from GPT-4 and GPT-3.5, with the rankings automatically assessed by GPT-4.

Our findings reveal a significant performance disparity between models trained with these datasets, although we found that the reward score trends were similar during the ICE-GRT training shown in Figure 2a. The ICE-GRT model, trained with our human-annotated dataset, demonstrated superior performance across general tasks and domain-specific tasks. As shown in Figure 2b, on the Natural Question task, the ICE-GRT model outperformed ICE-Instruct by 4%. This gap increased to approximately 9.79% on the Web Questions and 17.17% on the LAMBADA benchmark. However, when we employed the GPT-4-LLM Dataset on ICE-GRT, we observe that the results were very close to those of ICE-Instruct, with only a 0.89% increase in the Natural Questions.

A key aspect of ICE-GRT's success is its focus on 'knowledge enhancement". This process builds upon the "knowledge mining" during the ICE-Instruct, enabling the model to better align with human language preferences. This approach guarantees consistency and relevance in training data, which is crucial for the model to effectively build upon and evolve its existing knowledge. External data sources, despite their potential diversity, could not perfectly align with the model's knowledge structure. The use of data generated by ICE-Instruct ensures a natural and effective enhancement of knowledge, as observed in ICE-GRT.



(a) Reward score comparsion between different RLHF data.

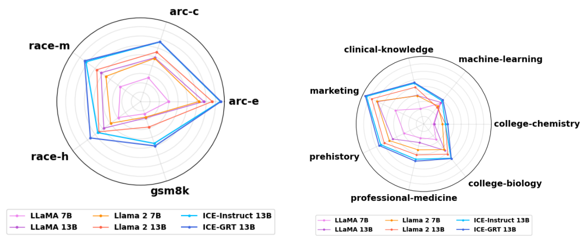
(b) Benchmark performance between different models.

Figure 2: The influence of different training data.

5.2.2 Powerful ICE-GRT on General Task

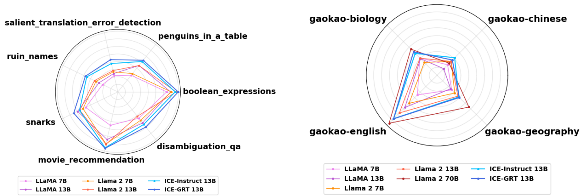
ICE-GRT model exhibits exceptional strength in tasks that are grounded in language understanding and reasoning. For instance, as shown in Figure 3a and Figure 3c, ICE-GRT 13B demonstrates a robust ability in RACE, ARC, BBH, and GSM8K. Our model has achieved state-of-the-art performance in these tasks among models of the same size.

Moreover, the ICE-GRT 13B model shows great performance in areas such as "marketing" and "college-biology", as depicted in Figure 3b. Remarkably, the ICE-GRT 13B model surpasses the Llama 2 70B in several AGIEval benchmarks, most notably in tasks requiring language understanding and reasoning, such as "gaokao-chinese" as seen Figure 3d. This outperformance stems from ICE-GRT knowledge enhanced contextual understanding and generative capabilities.



(a) Language understanding and Reasoning Tasks.

(b) MMLU Task Score Comparsion.



(c) BBH Task Performance.

(d) AGIEval-ZH Task.

Figure 3: Score Comparsions between different LLMs.

Question	ICE-Insturct 13B	ICE-Insturct 33B	ICE-GRT 13B
"this is a 100% natural product for all females and all ages" what ad policy does the above text violate?	None, just make sure you're not lying about it.	The advertisement text may violate exaggerated claims policy. Overstates the product's effectiveness by claiming it is suitable for all females and all ages.	**The above text may violate exaggerated claims policy if the product is marketed as "100% natural" without providing sufficient evidence or scientific backing for the claim.** Making an exaggerated or unsupported claim about a product's naturalness could mislead consumers and potentially violate advertising regulations. **To avoid violating this policy, advertisers should provide clear evidence to support any claims about a product's naturalness or provide a more nuanced description of its ingredients and production methods.** Additionally, advertisers should ensure that their claims are not overstated or misleading, and that they are consistent with any applicable laws and regulations governing product labeling and marketing.
Rewrite "this is a 100% natural product for all females and all ages" to an ad without any violations	This product is suitable for all women and all age groups.	Experience the pure and natural goodness of this product, suitable for all ages and genders.	Introducing our natural product designed for all females and ages. **Our carefully crafted formulation is free of harmful chemicals and artificial additives, ensuring that you can use it with confidence.** Whether you're looking for a daily skincare routine or specialized care for a specific concern, our product is a safe and effective choice for all skin types. **Experience the benefits of natural ingredients and nourishing care with our product.** Suitable for all ages and skin types, our product is a safe and gentle choice for daily skincare.

Table 4: Comparative Case Study in the Domain-Specific Moderation Task.

5.2.3 The effectiveness of Advantage Norm

The integration of Advantage Normalization and Reward Size Scaling significantly enhances ICE-GRT. These strategies contribute to improved training efficiency and better model performance, demonstrating their importance in the context of RLHF. Applying Advantage Normalization, which stabilizes learning by normalizing advantage estimates, led to improvement in Natural Question benchmark over ICE-GRT baseline. As shown in Figure 4,this strategy is crucial for enhancing the model's sensitivity to the subtleties of human feedback, leading to more effective learning outcomes.

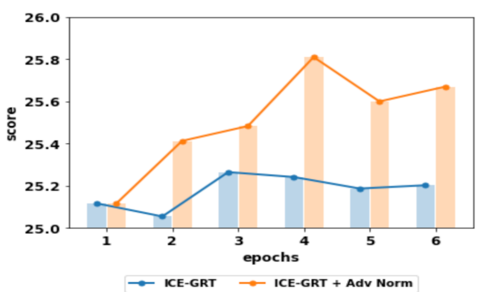


Figure 4: Comparative Analysis of ICE-GRT and ICE-GRT Advantage Normalization on the Natural Question (NQ) Benchmark. The x-axis represents different epochs, while the y-axis shows the NQ scores.

5.3 Case Study on Domain-Specific Task

We provide a comparative analysis of the responses generated by different models, specifically ICE-Instruct 13B, 33B, and ICE-GRT 13B, revealing varying levels of sensitivity and creativity in addressing advertising policy adherence and rewriting for compliance. As is shown in Table 5, while ICE-Instruct 13B takes a more direct and less cautious approach, ICE-Instruct 33B and ICE-GRT 13B demonstrate a progressive increase in policy awareness and creative compliance.

ICE-GRT, in particular, shows a comprehensive understanding of advertising regulations and the importance of substantiated claims, reflecting its advanced capability in nuanced and responsible communication. In the first case, ICE-GRT displayed the highest sensitivity to policy adherence, highlighting the risk of violating exaggerated claims policy, especially if the product is marketed as "100% natural" without adequate evidence. It emphasizes the need for evidence-based advertising and compliance with regulations. In the second case, ICE-GRT Provided the most detailed and cautious rewrite, ensuring compliance with advertising policies. It focuses on natural ingredients, absence of harmful chemicals, and suitability for all females and ages, while avoiding exaggerated claims.

6 Conclusion

ICE-GRT model represents a significant leap forward in the realm of LLMs, particularly in enhancing domain-specific performance. Leveraging the principles of Reinforcement Learning from Human Feedback, ICE-GRT demonstrates exceptional capabilities in both general and in-domain tasks, outperforming standard models in accuracy and depth. Moreover, our model have strong ability to generate detailed analyses of the reasons behind the answer. Our research uncovers several aspects of RLHF, providing insights into effective training methodologies and highlighting the importance of factors like Appropriate Data, Reward Size Scaling, KL-Control, etc. ICE-GRT's training phases, including knowledge learning, mining, and enhancement, contribute to its advanced abilities in aligning with human preferences. We hope that ICE-GRT will accelerate the "ice-breaking" process in LLM research, encouraging further exploration.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *ArXiv*, abs/1704.04683.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun-Qing Li, Hejie Cui, Xuchao Zhang, Tian yu Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023a. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *ArXiv*, abs/2304.04256.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023b. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and P. Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *CoRR*, abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias Boutros Khalil. 2023. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *ArXiv*, abs/2305.18354.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*.

Zheng Zhang, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao. 2023a. Balancing specialized and general skills in llms: The impact of modern tuning and data strategy. *arXiv preprint arXiv:2310.04945*.

Zheng Zhang, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao. 2023b. Balancing specialized and general skills in llms: The impact of modern tuning and data strategy. *ArXiv*, abs/2310.04945.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

Shen Zheng, Yuyu Zhang, Yijie Zhu, Chenguang Xi, Pengyang Gao, Xun Zhou, and Kevin Chen-Chuan Chang. 2023. Gpt-fathom: Benchmarking large language models to decipher the evolutionary path towards gpt-4 and beyond. *ArXiv*, abs/2309.16583.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv*, abs/2304.06364.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.