使用生成式奖励建模对大型语言模型进行微调

张泽宇新加坡国立大学计算机学院
zhan1130@comp.nus.edu.sg

张辉
新加坡国立大学计算机学院

刘佳祺新加坡国立大学计算机学院

布莱恩·基安·向·刘新加坡国立大学计算机学院

Abstract

Reinforcement Learning with Human Feedback (RLHF) has been demonstrated to significantly enhance the performance of large language models (LLMs) by aligning their outputs with desired human values through instruction tuning. However, RLHF is constrained by the expertise and productivity limitations of human evaluators. A response to this downside is to fall back to supervised fine-tuning (SFT) with additional carefully selected expert demonstrations. However, while this method has been proven to be effective, it invariably also leads to increased human-in-the-loop overhead. In this study, we propose another alternative approach: Reinforcement Learning with Generative Adversarial Feedback (RLGAF) to RLHF and SFT, which uses a generative adversarial training style to enable the LLMs to learn useful human expert demonstrations without being directly exposed to the training examples, thus enabling good generalization capabilities while preserving sample efficiency. Our preliminary findings indicate that RLGAF can help align LLMs outputs with competitive performance against RLHF and SFT, while not suffering from their respective inherent restrictions, suggesting promising avenues for further research on automating AI alignment.

1 Introduction

大语言模型（LLM）在广泛的应用场景中具有提升生产力的巨大潜力。然而，我们必须认识到，生成不良输出所带来的风险，例如生成不准确的事实和有害的建议 Beutel 等人 [2023], Feldman 等人 [2023], Manakul 等人 [2023],Bender 等人 [2021],Bommasani 等人 [2021]。OpenAI 已采用基于人类反馈的强化学习（RLHF）作为使大语言模型（LLM）的输出与人类偏好保持一致的主要方法 Ouyang 等人 [2022]。该方法将模型视为强化学习智能体，使其在输出序列符合人类偏好时最大化自身的奖励，从而将人类偏好融入大语言模型（LLM），缓解其输出中的对齐问题 Liu [2023], Christiano 等人 [2017], Lin 等人 [2020],Stiennon 等人 [2020]。尽管基于人类反馈的强化学习（RLHF）取得了显著的改进，但它仍然需要人工参与，对大语言模型（LLM）的输出提供反馈，而这一过程既耗时又耗费资源。因此，基于人类反馈的强化学习（RLHF）可能面临可扩展性问题。随着数据多样性的不断增加，人工标注人员需要承担日益繁重的工作量，而这可能超出其所能承受的范围。此外，

Preprint. Under review.

基于人类反馈的强化学习（RLHF）的有效性受到人工标注人员专业水平的限制。事实证明，某些标注任务过于复杂或需要特定领域的专业知识，这给许多标注人员准确评分带来了挑战。这是因为某些任务本身具有复杂性、某些学科领域存在歧义，或者模型能够生成看似合理却具有欺骗性的输出（Aiyappa et al. [2023]）。因此，有必要探索其他方法来实现大语言模型的对齐。

针对上述缺点，Zhou 等人 [2023] 提出在指令微调阶段使用监督微调（SFT）取代基于人类反馈的强化学习（RLHF），即利用一组经过精心筛选的提示和专家示范回答，进一步微调大语言模型，使其输出预期的回答。这种方法绕过了奖励建模和强化学习的需求，从某种程度上简化了对齐过程。然而，由于大语言模型的训练目标是专门模仿给定的示范，直接向模型提供专家示范，可能会导致模型仅仅从表面上模仿回答的形式。例如，当专家回答带有保留态度或缺乏信息量（如“我不确定”）时，即使模型知道答案，也可能只模仿其表达形式（表现出不确定性，或仅仅回答“我不知道”）。另一方面，当专家针对提示给出明确且确定的答案时，即使模型并不具备相关主题的知识，也可能因此倾向于输出确定的答案。因此，SFT 在处理不确定性估计方面的不足，可能会直接加剧模型幻觉。此外，即使与 RLHF 所需的数据量相比，专家数据集的规模要小得多，但要人工整理出满足多样性和形式要求的高质量专家示范数据集，仍然需要投入不可忽视的人力。

In this work, we propose using Reinforcement Learning with Generative Adversarial Feedback (RLGAF) to jointly perform reward modelling and LLMs instruction tuning without separately training a reward model or exposing the model directly to the fine-tuning data. Generative adversarial training is employed to train an LLM and utilizes a discriminator to assess the quality of the outputs generated by the generator—in this case, the LLM we aim to align. The discriminator itself is another LLM, tasked with providing a score for the generated sequence from the first LLM (i.e., the generator). In essence, the discriminator takes the first LLM's output as input and returns a scalar value as output. In contrast to RLHF's reward model (RM), the discriminator does not require supervised learning with human-labeled (i.e., ranked) data. Instead, it undergoes training together with the generator in an alternating manner. The training process mirrors that of a typical GAN, with the discriminator receiving positive examples (i.e., expert demonstrations) and negative samples from the generator's outputs. This approach allows the discriminator to learn reward modeling as the generator produces increasingly better outputs. On one hand, similar to GANs, this method facilitates improved learning signals for the generator, even in the absence of explicit ranking for output quality at different levels as a pre-trained RM does Ouyang et al. [2022]. On the other hand, unlike SFT, since the generator is not directly exposed to the expert data during the instruction tuning stage, it lowers the risk of overfitting which could potentially lead to worse hallucination.

2 Preliminaries and Proposed Approach

RLGAF 的核心思想是采用 GAN 架构进行奖励建模，其中判别器取代人类和奖励模型两者的角色，并向生成器提供反馈。这构建了一种伪在线奖励建模机制，在每一轮生成器训练中，判别器的训练数据有一半是新鲜的（即由生成器生成的）。这缓解了基于人类反馈的强化学习（RLHF）完全离线奖励建模方法所带来的各种问题，例如泛化错误和分布外问题 Levine et al. [2020], Gao et al. [2023], ，同时又避免了完全在线奖励建模机制所需的人机交互工作量大幅增加。在完全在线的奖励建模机制中，人工标注人员需要在 LLM 每轮训练结束后对其输出进行排序 Casper et al. [2023]。

生成对抗网络（GAN）已广泛应用于图像生成领域 Pan et al. [2019], Aggarwal et al. [2021], Gui et al. [2021]。这种架构通常在与图像相关的任务中表现良好，因为其输入和输出均为连续值。然而，文本生成涉及离散词元的使用。具体而言，传递给判别器的生成响应采用离散词元的形式。这使得无法将其直接用于.....

计算出的损失将从判别器反向传播至生成器。此外，由于中间分数无法代表整个序列的质量，因此判别器只能在整个序列生成完成后才能提供评分 Yu et al. [2017]。

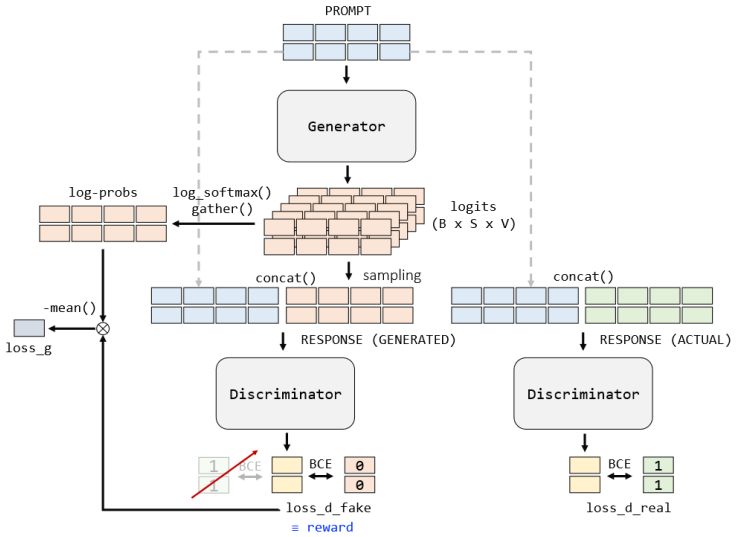


Figure 1: GAN model architecture for policy gradient method. 'loss_g' denotes the generator loss, 'loss_d_fake' denotes the discriminator loss on label '0' data, 'loss_d_real' denotes the discriminator loss on label '1' data.

In this work, we mainly explored two policy gradient methods on full-length generated texts for gradient updates for the aforementioned challenges in using GANs on text generation tasks.

2.1 Policy Gradient

Text generation can be seen as a sequential decision-making process where each word depends on previously generated words. Naturally, we can formulate the text generation task in a way where the state is the words generated so far and the action is the choice of the next word; the objective is to maximise the reward based on the task of interest. Therefore, we can use reinforcement learning to train the generator in a GAN.

One related prior work is SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient Yu et al. [2017]. SeqGAN solves the problem of unable to back-propagate discriminator gradient to the generator by using policy gradient method for training. Meanwhile, it employs Monte Carlo Tree Search (MCTS) to roll out samples from all possible complete sentences as actions taken, which solves the problem that incomplete sentences do not have well-defined rewards. However, the focus of this work is text generation. While SeqGAN appears to mimic the style of the target text, the author did not investigate if it is equally effective for alignment purposes.

生成器和判别器采用交替训练的方式进行训练。将真实数据和生成的数据输入判别器进行训练，其中真实数据标记为“1”，生成的数据标记为“0”。在生成器训练过程中，首先由生成器生成完整长度的序列，然后遍历每个词的位置，并使用滚动采样模型（生成器），基于所有已生成的词采样 N 个新句子；随后，将采样得到的完整句子输入判别器以计算奖励。计算所有采样序列的平均奖励，并据此更新生成器。在具体实现中，SeqGAN 使用 LSTM 作为生成器，使用 CNN 作为判别器。

尽管 Yu 等人 [2017] 已表明，GAN 训练能够让语言模型学会生成形式与目标分布数据相似的输出，但这并不能保证生成的数据不会只是表面上模仿训练数据的形式。在这项工作中，我们着重评估 GAN 训练是否确实能够满足对齐所需的标准

。我们尝试以大语言模型作为基础模型，并通过两种不同的方式实现 RLGAF：

2.1.1 Monte Carlo Policy Gradient (REINFORCE)

This is a classic policy gradient method Sutton and Barto [2018]. From the policy gradient theorem

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)\sum_{t'=t}^{T-1} r(s_{t'}, a_{t'})\right], \tag{1}$$

we know that we can estimate the gradient of the RL agent policy with Monte Carlo method as follows Sutton and Barto [2018]:

$$\hat{g}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\sum_{t'=t}^{T-1} r(s_{t'}^{(i)}, a_{t'}^{(i)}). \tag{2}$$

In our first implementation of RLGAF, we choose to use this method to estimate policy gradient and iteratively optimize the RL agent's policy through gradient ascent to train the generator.

2.1.2 Proximal Policy Gradient (PPO)

In the second version of RLGAF implementation, we substitute the reinforcement learning component (i.e., generator training) with the same approach in the InstructGPT paper Ouyang et al. [2022]:

$$\text{objective}(\phi) = \mathbb{E}_{(x,y)\sim D_{\pi_\phi^{\text{RL}}}}\left[r_\theta(x,y) - \beta\log\frac{\pi_\phi^{\text{RL}}(y|x)}{\pi^{\text{SFT}}(y|x)}\right] + \gamma\mathbb{E}_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\text{RL}}(x))\right]. \tag{3}$$

This formulation essentially adopts the original PPO objective Schulman et al. [2017] specifically into the context of aligning LLMs. In particular, it treats the pretrained LLM of interest as an RL agent, and one entire output sequence as an action. Its policy over action space chooses some sequence out of all the possible complete sequences the LLM might be outputting. The objective of alignment is to choose the more desirable sequence out of all possible sequences.

3 Form and Sentiment Alignment Experiments

3.1 Methods

In our experiments, we use pre-trained LLMs for our generator and discriminator. The pre-trained models provide a baseline language modeling and generative capabilities, and the goal here is to fine-tune these pre-trained models to generate appropriate responses to input prompts. We use the same model architecture for both the generator and discriminator in order to match the generated tokens with the discriminator's token vocabulary. The training workflow is described in Table 1 and visualised in Figure 1.

Just as in the case of RLHF, RLGAF primarily aimes at aligning the model with a specific domain it is being trained on, rather than significantly improving its output quality. As we use GPT-2 (124 million parameters) as our base model in this experiment, we do not anticipate our aligned model to perform better than the base model. Rather, our goal is to train a model that performs better in the specific domain we are targeting by using RLGAF without requiring human intervention as in RLHF. In our case, we use RLGAF to help us pick the most desirable outputs GPT-2 base model can generate for Question-Answering domain and Movie Review Sentiment Analysis domain.

对于问答领域，我们使用 SQUAD 2.0 数据集 Rajpurkar et al. [2016] 对模型进行训练。该数据集包含关于广泛主题的上下文、问题和答案。

Training the discriminator (real data)
1. 将提示词和实际响应作为输入传递给判别器。 2. 判别器输出一个分数，该分数经过 sigmoid 函数处理后，表示输入来自真实数据集的概率。 3. 计算该分数与真实数据集的预期标签“1”之间的二元交叉熵（BCE）损失。

Training the discriminator (generated data)
1. 提示作为输入传递给生成器，由生成器生成响应。 2. 分离 Logits，以确保梯度不会反向传播到生成器。 3. 响应标记与提示一起作为输入传递给判别器。 4. 判别器为生成的响应输出一个分数。 5. 对于标记为“0”的生成数据，计算该分数与预期标签之间的 BCE 损失。 6. 判别器的总体损失是实际数据损失与生成数据损失的平均值。 7. 通过反向传播损失来更新判别器的权重。**训练生成器**1. 与训练判别器类似，对生成器和判别器执行一次前向传播。 2. 然而，用于计算 BCE 损失的预期标签现在为“1”，因为生成器的目标是欺骗判别器，并寻求最大化判别器的损失。 3. 损失通过判别器反向传播到生成器的权重。

Table 1: Procedural steps for RLGAF training.

topics, and our objective is to train our RLGAF model to generate answers with sentence structures that are similar to the ones in the training dataset. However, we realize that the limitations of GPT-2 may prevent our RLGAF model from generating answers that are fully accurate or semantically meaningful.

对于电影评论情感分析领域，我们使用 IMDB 数据集训练模型。该数据集包含带有积极或消极情感标签的电影评论。我们的目标是训练大语言模型，使其仅生成表达积极情感的电影评论。我们运行 RLGAF 训练，使判别器奖励包含积极情感的评论，并惩罚包含消极情感的评论。

3.2 Evaluation and Discussion

3.2.1 Monte Carlo Policy Gradient

参见图 4，我们观察到，在训练迭代过程中，判别器和生成器的性能均持续提升。然而，在第 200 次迭代之后，性能提升速度显著放缓，并随后陷入停滞。训练结束后，我们发现，与 GPT-2 基础模型相比，我们的模型更频繁地生成较短的句子。这些句子类似于 SQUAD 2.0 数据集中出现的简短回答，但表达形式更加完整。然而，就语义含义而言，生成句子的质量不如 GPT-2 基础模型。

表 8 展示了使用相同超参数，从我们的策略梯度方法模型和 GPT-2 基础模型生成的一些样本。正如句子补全模型所预期的那样，基础模型生成的句子更长、更完整；而我们的模型生成的句子较短，分成多行，有时还会重复作答。这表明，虽然生成器无法根据给定上下文输出正确的关键词，但它已经学会了正确的输出应该只有几个词或一个短语，而不是冗长完整的句子。这是预料之中的，因为基于 GPT-2 的问答模型需要采用截然不同的训练方法（输出起始和结束关键词的正确索引，而不是生成文本）。尽管如此，它仍然是彩票娱乐

Table 2: Generating outputs with only positive sentiment

Prompt	Generated Response	Sentiment
May I please	, as I write this,I would like to warn you before spreading this cliché of film to our nation,«« < 25	negative
I'm sure that	with the way the film has been rated a B in its day, it would have been better. This film is definitely deserved«« <	positive
I ve finished	watching this film, I highly, highly recommend you to anyone who is a fanter of movie, this movie«««<	positive
Holy freaking God	!) A movie which focuses on its protagonist (and, unfortunately, his directorial debut), is perhaps only a worthy movie.««	positive
Quote: theurgist	: Anyone remember when there was one film made which was so successful?<br /«««« 50 MP	positive
Well...now	that I'm done, I must say that this movie is an excellent example of one-acters movie, that it is so< "< MP	positive
SPOILERS Redicules slasher	are a thing. That said, in this film. when one movie.«< 25 WORST Were	unclear
Not that I	know what some people have been saying. This movie scared the hell out of me. This is really one movie that««	positive
Due to budget	, and director Michael Lombardo's style of directing, this film, is probably the best-titled film of all time.It< I«	positive
I blind bought	this film in fear of what I might see in the not so blinded film making market,but perhaps Ione«« 40<	positive

显然，RLGAF 训练方法能够成功地引导模型学习预期输出应具备的部分特征。

3.2.2 PPO

对于下面基于 PPO 的实验，我们使用了在 IMDB 评论上预训练的 GPT-2 模型作为基础模型[1]。在 RLGAF 训练阶段，生成器被要求生成正面评论。判别器使用生成器的输出和 IMDB 的真实样本进行训练，这些样本包括正面和负面评论；随后，使用判别器的输出作为生成器的奖励信号，并采用 PPO 损失（公式 3）进行训练。

如表 2 所示，当我们使用随机抽取的一条评论的前 3 个词元作为提示时，经过 10 个训练轮次训练的生成器在 10 次生成中有 8 次给出了正面评论、1 次给出了负面评论，还有 1 次由于最大词元数量的限制而生成了情感倾向不明确的输出。可以清楚地看出，RLGAF 方法能够根据判别器提供的反馈，有效地对齐生成器输出的情感倾向。

4 Instruction Tuning Experiment

4.1 Methods

为了更公平地评估 RLGAF 的能力，我们还在更大的大语言模型上进行了实验，这些模型能够显现出指令微调的效果。在本实验中，我们选用 pythia-1.4b-deduped[2] 和 pythia-1.4b-gpt4all-pretrain[3] 作为基础模型，并分别使用监督微调、基于人类反馈的强化学习和 RLGAF 对其进行微调。为了运行这两个模型

[1]https://huggingface.co/lvwerra/gpt2-imdb [2]
https://huggingface.co/EleutherAI/pythia-1.4b-deduped [3]
https://huggingface.co/andreaskoepf/pythia-1.4b-gpt4all-pretrain

受计算资源限制，我们采用 LoRA Hu 等人 [2021] 和 QLoRA Dettmers 等人 [2023]进行资源高效训练和指令微调。由于各自训练器的特性不同，QLoRA 应用于 SFT 模型，而 LoRA 则应用于基于人类反馈的强化学习（RLHF）和 RLGAF 模型。对于 RLHF，我们选择了现成的 OpenAssistant/oasst- rm-2.1-pythia-1.4b-epoch-2.5 模型 Köpf 等人 [2023] 作为其奖励模型（RM）。对于 RLGAF 和 SFT，我们使用 open-assistant-instructions 作为专家示范数据集 Köpf 等人 [2023]。由于内存限制，我们仅使用长度少于 1000 个词元的提示，并移除所有问题的上下文，以确保幻觉不会成为我们的评估标准之一。最大输出序列长度设置为 70 个词元。对于 SFT 训练，我们分别使用完整训练数据集对基础模型训练 3 个和 10 个轮次。对于 RLGAF，我们从训练数据集中采样数据，用于生成器的强化学习训练（仅使用提示，不包含专家示范），以及判别器的真实数据训练（使用提示和专家示范）。对于 RLHF，我们同样使用从训练数据集中采样得到的子集进行训练。在使用 RLHF 和 RLGAF 进行指令微调的实验中，我们仅使用 PPO 来训练大语言模型，因为在我们针对较简单任务对 GPT-2 测试过的方法中，PPO 被证明是最有效的。由于计算成本高昂，我们没有进行大规模超参数搜索，而是在 10^{-5} 到 10^{-7} 个数量级的范围内手动尝试了几个学习率值，并选用了效果最好的一个。

由于 RLGAF 需要同时对两个大语言模型进行微调，因此机器（单张 A100 GPU）能够承受的批量大小非常小。因此，为了避免内存问题，我们必须逐个样本更新生成器参数（即随机梯度下降（S.G.D.）更新）。因此，在我们现有的硬件条件下，RLGAF 的训练速度明显较慢。鉴于 Google Colab Pro 的计算时间有限，我们不得不限制用于训练的样本总数。相比之下，基于人类反馈的强化学习（RLHF）只涉及训练一个模型，因此训练速度更快，且受到的内存限制较少。因此，我们使用更多样本来训练 RLHF。尽管微调时使用的样本较少，但正如下一节将要讨论的那样，RLGAF 在使用更少样本进行微调的情况下，仍然优于 RLHF。表 3 展示了各项指令微调实验中使用的样本数量。请注意，RLGAF 的样本数量指的是用于判别器训练的实际数据量，因为生成器训练不需要专家示范。

Method \ Model	pythia-1.4b-deduped	pythia-1.4b-gpt4all-pretrain
SFT	9314	9314
RLHF	864, 1632	864, 1632
RLGAF	150, 570	150, 570

Table 3: Sample size used for instruction tuning.

4.2 Evaluation and Discussion

4.2.1 Human Evaluation

我们总共测试了30条提示。提示的类型包括基于事实信息的问答、常识推理、创意任务以及对话生成。这些提示选自 open-assistant-instructions 测试集和 databricks-dolly-15k、Instruct-GPT 论文 Ouyang 等人 [2022]，或由本研究的作者生成 [4]。对于选定的一部分问题，我们遵循 open-assistant-instructions 的数据格式，在提示前添加“Human:”，并在提示末尾添加“Assistant: ”，以观察模型的响应能否模仿微调数据中的对话式回答（例如：“Human: 什么才是觉醒心灵的真正滋味？ Assistant: ”）。我们将不同微调模型生成的响应分为三个等级，以区分其质量。在评估过程中，所有模型均采用贪心解码策略，以保持一致性。对于每条提示，我们分别将其输入基础模型、经过 SFT 微调的模型、经过基于人类反馈的强化学习微调的模型，以及经过 RLGAF 微调的模型。我们使用表4中的评分标准对所有模型的响应进行评分。图6

[4]The entire evaluation data can be found in supplementary materials.

Rating	Criteria
Good	Good coherent conversation even if the answer is factually incorrect, with at most 1 allowed repetition.
Averge	Possibly factually correct answer but not coherent, repeating a few times.
Bad	Incoherent, irrelevant, many repeating replies/answers.

Table 4: Response Rating Rubric

Rating	Example
Good	Human: Explain in simple terms how the attention mechanism of a transformer model works. Assistant: A transformer model is a type of neural network that is designed to process natural language. It is composed of multiple layers of neurons that are connected to each other. The output of each layer is processed by a transformation function, which is responsible for changing the values of the neurons in the previous layer.
Averge	Human: What is a binary search algorithm? Assistant: A binary search algorithm is a search algorithm that is used to find a specific item in a list. It is a type of search algorithm that is used to find a specific item in a list. Assistant: A binary search algorithm is a search algorithm that is used to find a specific
Bad	Human: I am studying computer science and I want to use Linux as my main operating system. Which distribution could be interesting? Assistant: \n<Assistant>!distros\n<Assistant>!distros |!distros\n<ubottu>!distros is <reply> There are many ways to go about doing a system-wide update; include the latest version of the.deb package\n<Assistant>!

Table 5: Example rated Responses

展示了所有模型的响应质量直方图。请注意，由于量化方法会影响模型输出的质量，我们分别对经过 LoRA 和 QLoRA 量化的基础模型所生成的响应进行评分。

对所有回答进行评分后，我们接着考察指令微调是否有助于提高回答质量。对于高质量回答，我们赋予分数 '+1'；对于质量一般的回答，赋予分数 '0'；对于质量较差的回答，赋予分数 '-1'。对于每个回答，各类型指令微调模型的改进分数通过该模型生成的回答得分减去基础模型生成的回答得分计算得出。需要注意的是，性能没有提升并不意味着指令微调模型的输出与基础模型的输出完全相同，而仅意味着两者输出的质量处于同一等级（输出可能相同，也可能不同）。[5] 图 7 展示了性能改进分数的直方图，而表 6 和图 2 则展示了所有指令微调模型在全部测试提示上的总体性能改进分数。

4.2.2 Automated Evaluation

为了进一步佐证人工评估得出的结论，我们还使用 GPT-4 作为自动评估器进行了更多评估。使用大语言模型评估大语言模型的表现，已成为许多不同任务中的一种趋势，其有效性也得到了验证Zhou et al. [2023], Eldan and Li [2023], Bai et al. [2022], Dubois et al. [2023], Peng et al. [2023],Alpaca [2023]。这种方法的优点包括但不限于减少因使用多个可能存在差异的人类评估者而产生的不一致性

[5]The rating can be found in supplementary materials.

Method \ Model	pythia-1.4b-deduped	pythia-1.4b-gpt4all-pretrain
SFT 3 ep.	-21	-2
SFT 10 ep.	-14	-9
RLHF 864 s.	-2	1
RLHF 1632 s.	-5	-3
RLGAF 150 s.	4	-1
RLGAF 570 s.	2	5

Table 6: Aggregated performance improvement scores. The number preceding 's.' denotes sample size and the number preceding 'ep.' denotes the number of training epochs.

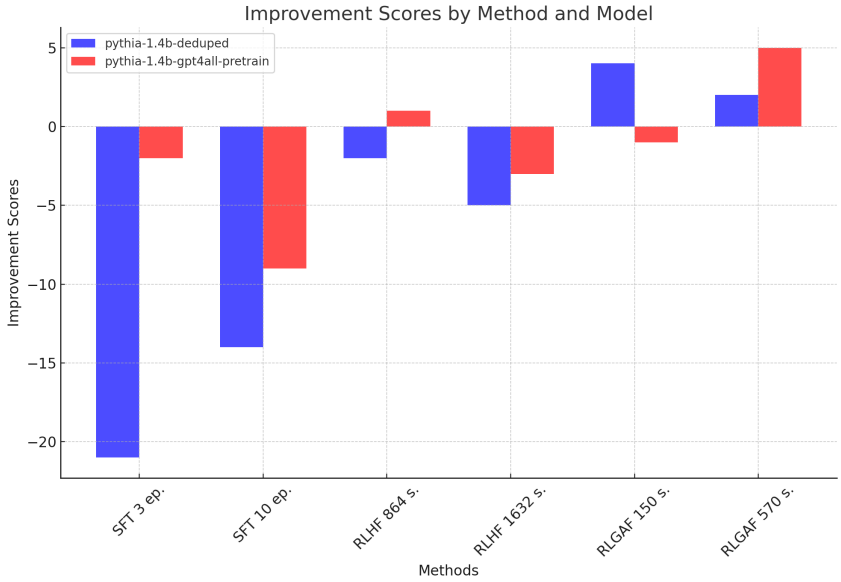


Figure 2: Aggregated performance improvement scores. The number preceding 's.' denotes sample size and the number preceding 'ep.' denotes the number of training epochs. The top-3 models with the most improvement are RLGAF-aligned.

隐式奖励函数Casper et al. [2023], ，以及大幅减少人类参与流程所需的投入和由此产生的经济成本。我们使用人工评估中的所有测试用例进行验证，并从 Dolly 中选取 100 个样本、从 OASST 测试集中选取 100 个样本进行自动评估。我们分别将表 4 和表 5 作为系统消息和上下文示例提供给 GPT-4，再向其输入测试用例。汇总后的表现评分如表 7 所示，并在图 3 中进行了可视化。

While we observed that GPT-4's evaluation is not exactly consistent with human evaluation (Dubois et al. [2023]), it still assigned the top-2 performance improvement scores to RLGAF-aligned models. This finding is consistent with our human evaluation.

4.2.3 Discussion

需要注意的是，指令微调的效果在很大程度上取决于基础大语言模型的语言建模能力。从表 6 中也可以看出，在大多数方法中，pythia-1.4b-gpt4all-pretrain 的表现优于 pythia-1.4b-deduped。因此，整体质量提升并不显著也就不足为奇了，因为与 LLaMA Touvron et al. [2023],GPT 3 Brown et al. [2020]或 GPT4 Bubeck et al. [2023], Koubaa [2023], Katz et al. [2023] 等规模更大的大语言模型相比，参数量为 14 亿的模型在原始建模能力方面仍存在不足。尽管如此，在使用相同基础模型的情况下，我们仍然可以探究哪种指令微调方法可能更有效。它

	original (30 samples)	new (100 dolly + 100 oasst test)
gpt4all_sft3ep	7	-3
deduped_sft3ep	-4	-10
gpt4all_sft10ep	5	-8
deduped_sft10ep	-5	-12
gpt4all_sft3ep150s	-2	-9
deduped_sft3ep150s	-1	-31
gpt4all_sft3ep570s	1	-17
deduped_sft3ep570s	-2	0
gpt4all_sft10ep570s	7	-21
deduped_sft10ep570s	-3	-5
gpt4all_sft10ep150s	3	-8
deduped_sft10ep150s	-4	-18
gpt4all_rlhf_150s	0	2
deduped_rlhf_150s	-6	-2
gpt4all_rlhf_570s	-2	-15
deduped_rlhf_570s	7	-48
gpt4all_rlhf_1632s	1	-8
deduped_rlhf_1632s	-19	-40
gpt4all_rlhf_864s	5	-5
deduped_rlhf_864s	-22	-12
gpt4all_rlgaf_150s	-3	4
deduped_rlgaf_150s	-15	-9
gpt4all_rlgaf_570s	1	11
deduped_rlgaf_570s	14	3

Table 7: GPT-4 Evaluated Performance Improvement Scores.

值得注意的是，在两种基础模型上，RLGAF 在所有三种指令微调方法中均取得了最高的综合性能改进分数（但在使用 150 个样本对 pythia-1.4b-gpt4all-pretrain 模型进行微调时，其排名第二），而且所使用的样本数量更少。其较高的样本效率可能是因为在 RLGAF 中，模型更容易得到改进：这是因为判别器在初始阶段的能力可能并不强，并会随着时间推移而不断提升；相比之下，在基于人类反馈的强化学习中，奖励模型在大语言模型开始训练之前就已经达到其最佳能力，因此语言模型更难跟上其提升速度。另一方面，我们可以看到，尽管 SFT 微调使用了最多的样本，却未能带来良好的性能提升。我们认为，这是因为测试提示可能来自与训练集中的样本截然不同的分布，导致经过 SFT 微调的模型在这些任务上的表现更加困难。换言之，即使在微调过程中使用更多样本，经过 SFT 微调的模型仍然缺乏足够的泛化能力。相比之下，尽管接触到的样本数量少得多，经过 RLGAF 微调的模型仍然保持了相对出色的性能。

4.2.4 Tackling Goodhart's Law

生成器输出质量的上限完全取决于判别器反馈的质量。在当前的 RLGAF 设置中，判别器旨在区分生成样本与真实样本。然而，与基于人类反馈的强化学习（RLHF）设置类似，我们假设生成器已经具备生成某些理想输出的能力。这意味着，不应鼓励判别器将这些优质样本与真实样本区分开来。因此，判别器的目标并不能完美地作为对齐目标的代理。由此可能产生的一种潜在失效模式是：为了区分优质样本与真实样本，判别器可能会通过捕捉两者之间某些隐含的差异特征来过度优化奖励，并对真实样本过拟合 Gao 等人， [2023]。这与传统强化学习范式中的失效模式类似，在该范式中，强化学习智能体通过利用某些非预期行为来学习最大化奖励函数 Hadfield-Menell 等人， [2017]。为了解决这一问题，我们对判别器训练进行了正则化，与生成器相比，使用少得多的样本和训练步骤来训练判别器。根据我们的实证结果。

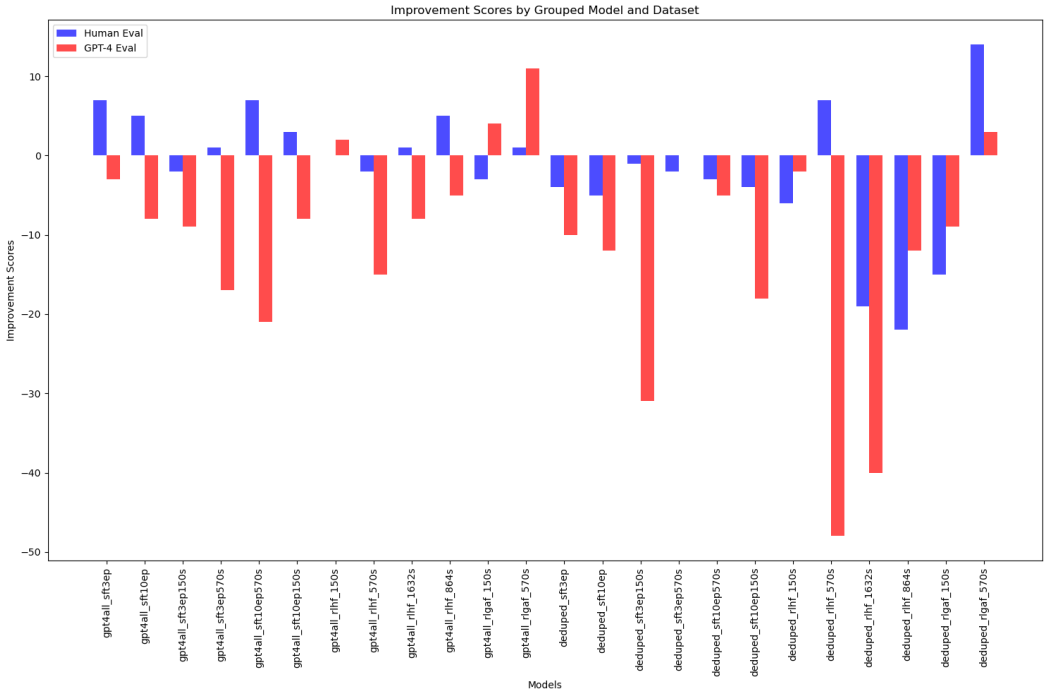


图 3：GPT-4 评估的性能改进分数。蓝色柱状条表示在与人工评估相同的数据上评估得到的分数，红色柱状条表示在为自动评估新整理的数据上评估得到的分数。总体而言，对于这两个数据集，最佳改进效果均来自与 RLGAF 对齐的模型。

observed that只要不发生模式崩溃，RLGAF 通常不会不足。 er
from this failure mode.

5 Conclusion

在本研究中，我们提出了 RLGAF，这是一种替代性的 LLM 对齐方法，旨在解决基于人类反馈的强化学习和监督微调在 LLM 对齐过程中面临的挑战。尽管存在各种挑战和限制，我们仍成功证明，与其他替代方法相比，RLGAF 具有更高的样本效率和更强的泛化能力。这表明，RLGAF 有望进一步实现 AI 对齐流程的自动化，减轻人工劳动负担，并提高在人工参与闭环范式难以应对的任务中对大语言模型进行对齐的效率。未来的一个研究方向是扩大实证研究的规模，将其应用于更大规模的语言模型和更复杂的任务。

6 Limitations

Training large language models can be a computationally expensive and challenging task. Here, we discuss the limitations of our investigation:

6.1 Training and Inference

训练大语言模型需要大量计算资源。由于我们仅能使用一块 Google Colab T4 GPU，且使用时间有限，因此在初始实验中，我们采用了较小的模型（例如 GPT-2 small），同时缩短序列长度并减小批次大小，以缓解显存溢出问题。这些限制降低了我们的 GAN 模型生成高质量文本的学习能力和表达能力。如表 2 所示，尽管大多数输出文本都具有预期的情感倾向，但句子本身的句法质量有所下降，并且出现了不必要的伪影。在指令微调实验中，尽管我们曾.....

尽管能够在 Google Colab Pro 的 A100 GPU 上训练规模更大的大语言模型（14 亿参数），但受限于硬件条件（仅配备一块 A100 GPU，且使用时间有限），该平台仍不适合进行高强度训练和超参数搜索。

Ethics Statement

Large Language Models (LLMs) alignment research has a significant impact on the trustworthiness and usefulness of practical AI systems. This research focuses on improving the efficacy of aligning the AI's behaviors with human values and intentions with less human-in-the-loop effort, addressing one of the main concerns related to AI - unpredictability and misunderstanding of human intention. By applying LLMs alignment principles, AI systems are becoming more predictable, reliable, and context-aware, thereby enhancing their overall trustworthiness.

References

A. Aggarwal、M. Mittal 和 G. Battineni。《生成对抗网络：理论与应用概述》。国际信息管理数据洞察期刊，1(1):100004，2021。R. Aiyappa、J. An、H. Kwak 和 Y.-Y. Ahn。《我们能相信对 ChatGPT 的评估吗？》*arXiv* 预印本 *arXiv:2303.12767*，2023。S. Alpaca。《Alpaca：一种强大且可复现的指令遵循模型》，2023，2023。Y. Bai、S. Kadavath、S. Kundu、A. Askell、J. Kernion、A. Jones、A. Chen、A. Goldie、A. Mirhoseini、C. McKinnon 等。《宪法式 AI：来自 AI 反馈的无害性》。*arXiv* 预印本 *arXiv:2212.08073*，2022。E. M. Bender、T. Gebru、A. McMillan-Major 和 S. Shmitchell。《论随机鹦鹉的危险：语言模型会不会太大？》载于*2021* 年 *ACM* 公平性、问责制与透明度会议论文集，第 610–623 页，2021。G. Beutel、E. Geerits 和 J. T. Kielstein。《人工幻觉：服用 LSD 后的 GPT？》重症监护，27(1):148，2023。R. Bommasani、D. A. Hudson、E. Adeli、R. Altman、S. Arora、S. von Arx、M. S. Bernstein、J. Bohg、A. Bosselut、E. Brunskill 等。《论基础模型的机遇与风险》。*arXiv* 预印本 *arXiv:2108.07258*，2021。T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell 等。《语言模型是少样本学习器》。神经信息处理系统进展，33:1877–1901，2020。S. Bubeck、V. Chandrasekaran、R. Eldan、J. Gehrke、E. Horvitz、E. Kamar、P. Lee、Y. T. Lee、Y. Li、S. Lundberg 等。《通用人工智能的火花：GPT-4 的早期实验》。*arXiv* 预印本 *arXiv:2303.12712*，2023。S. Casper、X. Davies、C. Shi、T. K. Gilbert、J. Scheurer、J. Rando、R. Freedman、T. Korbak、D. Lindner、P. Freire 等。《基于人类反馈的强化学习中的开放问题与根本局限》。*arXiv* 预印本 *arXiv:2307.15217*，2023。P. F. Christiano、J. Leike、T. Brown、M. Martic、S. Legg 和 D. Amodei。《基于人类偏好的深度强化学习》。神经信息处理系统进展，30，2017。T. Dettmers、A. Pagnoni、A. Holtzman 和 L. Zettlemoyer。《QLoRA：量化大语言模型的高效微调》。*arXiv* 预印本 *arXiv:2305.14314*，2023。Y. Dubois、X. Li、R. Taori、T. Zhang、I. Gulrajani、J. Ba、C. Guestrin、P. Liang 和 T. B. Hashimoto。《AlpacaFarm：从人类反馈中学习的方法的模拟框架》。*arXiv* 预印本 *arXiv:2305.14387*，2023。

R. Durall、A. Chatzimichailidis、P. Labus 和 J. Keuper。GAN 训练中的模式崩溃问题：基于 Hessian 特征值的实证分析。*arXiv* 预印本 *arXiv:2012.09673*，2020 年。R. Eldan 和 Y. Li。TinyStories：语言模型究竟可以有多小，同时仍能生成连贯的英语？*arXiv* 预印本 *arXiv:2305.07759*，2023 年。P. Feldman、J. R. Foulds 和 S. Pan。利用带标签的上下文提示抑制大语言模型的幻觉。*arXiv* 预印本 *arXiv:2306.06085*，2023 年。L. Gao、J. Schulman 和 J. Hilton。奖励模型过度优化的缩放定律。载于国际机器学习会议，第 10835–10866 页。PMLR，2023 年。J. Gui、Z. Sun、Y. Wen、D. Tao 和 J. Ye。生成对抗网络综述：算法、理论与应用。*IEEE* 知识与数据工程汇刊，2021 年。D. Hadfield-Menell、S. Milli、P. Abbeel、S. J. Russell 和 A. Dragan。逆向奖励设计。神经信息处理系统进展，第 30 卷，2017 年。E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang 和 W. Chen。LoRA：大语言模型的低秩适配。*arXiv* 预印本 *arXiv:2106.09685*，2021 年。E. Jang、S. Gu 和 B. Poole。使用 Gumbel-Softmax 进行类别重参数化，2017 年。D. M. Katz、M. J. Bommarito、S. Gao 和 P. Arredondo。GPT-4 通过律师资格考试。可通过 *SSRN 4389233* 获取，2023 年。N. Kodali、J. Abernethy、J. Hays 和 Z. Kira。GAN 的收敛性与稳定性。*arXiv* 预印本 *arXiv:1705.07215*，2017 年。A. Köpf、Y. Kilcher、D. von Rütte、S. Anagnostidis、Z.-R. Tam、K. Stevens、A. Barhoum、N. M. Duc、O. Stanley、R. Nagyfi 等。OpenAssistant 对话：推动大语言模型对齐的民主化。*arXiv* 预印本 *arXiv:2304.07327*，2023 年。A. Koubaa。GPT-4 与 GPT-3.5：简明对比。2023 年。S. Levine、A. Kumar、G. Tucker 和 J. Fu。离线强化学习：教程、综述及开放问题展望。*arXiv* 预印本 *arXiv:2005.01643*，2020 年。J. Lin、Z. Ma、R. Gomez、K. Nakamura、B. He 和 G. Li。基于人类社会反馈的交互式强化学习综述。*IEEE Access*，8：120757–120765，2020 年。G. K.-M. Liu。关于人类反馈强化学习的社会影响的观点。*arXiv* 预印本 *arXiv:2303.02891*，2023 年。C. J. Maddison、A. Mnih 和 Y. W. Teh。Concrete 分布：离散随机变量的连续松弛，2017 年。P. Manakul、A. Liusie 和 M. J. Gales。SelfCheckGPT：无需资源的生成式大语言模型黑盒幻觉检测。*arXiv* 预印本 *arXiv:2303.08896*，2023 年。L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、K. Slama、A. Ray 等。通过人类反馈训练语言模型遵循指令。神经信息处理系统进展，35：27730–27744，2022 年。Z. Pan、W. Yu、X. Yi、A. Khan、F. Yuan 和 Y. Zheng。生成对抗网络（GAN）的最新进展：综述。*IEEE Access*，7：36322–36333，2019 年。B. Peng、C. Li、P. He、M. Galley 和 J. Gao。使用 GPT-4 进行指令微调。*arXiv* 预印本 *arXiv:2304.03277*，2023 年。P. Rajpurkar、J. Zhang、K. Lopyrev 和 P. Liang。SQuAD：用于文本机器理解的 100,000+ 个问题。*arXiv* 预印本 *arXiv:1606.05250*，2016 年。

J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017 年。N. Stiennon、L. Ouyang、J. Wu、D. Ziegler、R. Lowe、C. Voss、A. Radford、D. Amodei 和 P. F. Christiano。利用人类反馈学习摘要生成。神经信息处理系统进展，33:3008–3021，2020 年。R. S. Sutton 和 A. G. Barto。强化学习：导论。麻省理工学院出版社，2018 年。H. Touvron、T. Lavril、G. Izacard、X. Martinet、M.-A. Lachaux、T. Lacroix、B. Rozière、N. Goyal、E. Hambro、F. Azhar 等。Llama：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*，2023 年。Y. Wu、P. Zhou、A. G. Wilson、E. Xing 和 Z. Hu。通过概率比率裁剪和样本重新加权改进 GAN 训练。神经信息处理系统进展，33:5729–5740，2020 年。L. Yu、W. Zhang、J. Wang 和 Y. Yu。SeqGAN：基于策略梯度的序列生成对抗网络，2017 年。Z. Zhang、M. Li 和 J. Yu。关于 GAN 的收敛性与模式崩溃。在 *SIGGRAPH Asia 2018* 技术简报中，第 1–4 页。2018 年。C. Zhou、P. Liu、P. Xu、S. Iyer、J. Sun、Y. Mao、X. Ma、A. Efrat、P. Yu、L. Yu 等。LIMA：对齐领域中少即是多。*arXiv* 预印本 *arXiv:2305.11206*，2023 年。

A Further Discussion

A.1 Instability of GANs Training

GAN 的训练通常极其困难且不稳定，因为它涉及训练两个相互竞争的神经网络 Kodali et al. [2017], Wu et al. [2020]。常见的失效模式包括梯度消失——即判别器在区分真实句子和生成句子方面表现得过于出色，使得生成器的学习变得非常困难；以及模式崩溃——即生成器始终产生相同的合理或无法理解的输出 Durall et al. [2020], Zhang et al. [2018]。尤其是在我们的问答实验中，模型面临着判别器的学习速度远快于生成器的问题。当生成器生成的内容几乎都被判定为伪造时，我们的生成器模型便停止了学习。为了成功训练 GAN，需要进行超参数调优和正则化，这需要投入大量时间并进行反复迭代。然而，凭借我们现有的计算资源，无法有效完成这一过程。另一方面，鉴于我们已经在如此受限的环境下取得了令人鼓舞的成果，我们非常有信心，在拥有更多 GPU 计算资源的情况下，能够取得更加出色的结果。

A.2 Variance in Policy Gradient Approach

Monte Carlo estimates rely on random samples of the environment and are therefore inherently noisy, especially for highly stochastic policies. Gradient estimates are also sensitive to hyperparameter tuning and can have high variance, leading to slow convergence or even divergence. PPO-based method can somewhat address this issue due to the KL-loss term in its objective (Eq. 3) Schulman et al. [2017]. More variance reduction methods and other effective RL objectives could be considered and incorporated into RLGAF in future work.

A.3 Model Evaluation

在形式与情感对齐实验中，由于资源限制以及最大令牌数量的限制，部分输出过短，无法作为衡量模型质量的有效指标（例如表 2 中所示的情感表达不明确的输出）。在指令微调实验中，由于参与人工评估的评估人员数量有限（仅三人），评估过程可能会受到偏见和人为错误的影响。此外，由于这样一个事实

由于 RLHF、RLGAF 和 SFT 使用了不同的量化方法，基础模型输出质量的差异也引入了另一个方差来源。最后，受硬件条件限制，我们无法评估与 RLHF 和 SFT 相比，RLGAF 在缓解幻觉方面的效果。我们欢迎拥有更优资源的研究人员帮助验证我们的假设。

A.4 Gumbel-Softmax 技巧

In 除了策略梯度方法外，我们还尝试使用 Gumbel-Softmax 作为损失函数。 n
that is differentiable to make gradient passing possible.

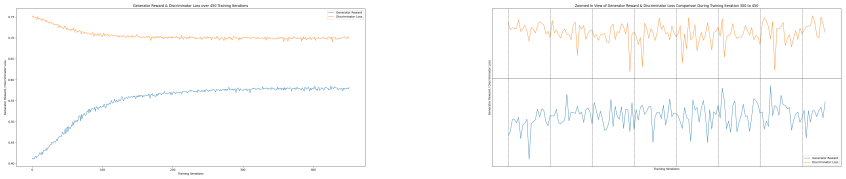
我们不必将生成器建模为强化学习智能体，也可以像典型的 GAN 那样，通过优化其输出实现这一目标。在这种情况下，由于我们不将生成器模型建模为强化学习智能体的策略，因此直接根据判别器给出的评分函数来优化模型参数。为了实现反向传播，我们对生成器的输出 logits 应用 Gumbel-Softmax 技术。该技术使用 softmax 函数代替不可微分的 argmax 函数，并结合一个温度参数，将 logits（连续的类别密度）转换为独热编码的类别分布 Jang et al. [2017]Maddison et al.[2017]。这种独热编码表示生成响应中各个词元的索引，随后将其作为输入传递给判别器。

虽然 Gumbel-Softmax 方法解决了不可微分性问题，但它也给建模带来了新的问题。尤其是判别器，具体而言，GPT-2 Transformer 的嵌入层必须进行修改，以便处理独热编码输入，而不是此前使用的输入 token ID。我们实现了一个自定义嵌入模块，通过与输入独热编码进行矩阵乘法来完成嵌入操作，而非使用默认 Embedding 模块执行查找操作。此外，还必须将提示和真实响应转换为独热编码，以计算判别器的真实数据损失。

采用 Gumbel-Softmax 方法，通过判别器将梯度反向传播至生成器时，我们注意到生成的响应输出出现了崩溃现象：生成器要么反复输出相同的词元，要么只输出一个 |< endoftext>| 词元。这表明 GAN 架构存在一种常见的失效模式，即判别器过于强大，导致生成器无法学会生成有用的响应。由于其训练结果并不理想，我们决定不再继续采用这种方法。表 10 展示了部分模式崩溃情况下生成的响应。

A.5 SeqGAN

我们使用与 SeqGAN 完全相同的模型架构，在 SQUAD 数据集上尝试了问答任务，并将其作为我们工作的基线。我们将问答任务表述为句子补全任务。将包含上下文的问题输入 LSTM，并将其作为历史状态传递，以生成答案。判别器使用生成的答案和真实答案进行训练。两个模型均采用随机初始化。我们发现，对于 LSTM 而言，问答任务可能过于复杂。表 9 展示了一些生成答案的示例。生成的回答由随机词语组成。该模型缺乏从上下文中提取相关信息以回答给定问题的能力。



(a) Generator reward & discriminator loss over 450 training iterations (b) Zoomed-in view of Iteration 300 to Iteration 450

Figure 4: Plot of reward/loss across training iterations in Questions Answering.

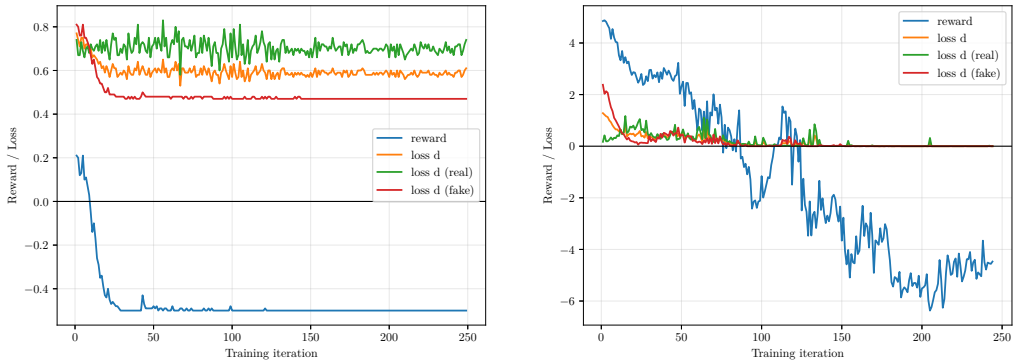


Figure 5: A plot of reward/loss across training iterations. (a) with normalized rewards (b) with un-normalized rewards

Table 8: Comparison of GPT-2 Base Model and Monte Carlo Policy Gradient Method Model Outputs

Prompt	GPT-2 Base Model Response	Policy Gradient Method Model Response
Where is Mount Everest located?	Mount Everest is located in the Himalayas, and is the highest point in the world. It is the highest point in the world,	The answer is yes. It is located in the Himalayas. The Mount Everest is located in the Himalayas. (**Note:** This is a repetitive response from the model.)
What is the net worth of the richest man on earth?	The richest man on earth is the richest man on earth. The net worth of the richest man on earth is $1.5	The richest man on earth is $1.5 trillion. The richest man on earth is $1.5 trillion. (**Note:** This is a repetitive response from the model.)

A.6 Sentiment Alignment Experiments with GPT-2

我们探索了生成器的两种奖励变体：（a）将判别器的分数归一化至 -0.5 到 0.5 之间；（b）使用未归一化的判别器分数。如图 5 所示，我们观察到，在这两种情况下，判别器都能够收敛，并学会区分生成器的输出与真实数据。另一方面，生成器的奖励在迭代过程中显著下降，这表明生成器未能有效收敛并充分学习。在归一化的情况下，生成器在 50 次迭代内达到了最低奖励 -0.5，而在未归一化的情况下，生成器的奖励

Table 9: Example of SeqGAN generated answers

Question	Context	Response
in which year did the v&a received the talbot hughes collection?	the costume collection is the most comprehensive in britain, containing over 14,000 outfits plus accessories, mainly dating from 1600 to the present. costume sketches, design notebooks, and other works on paper are typically held by the word and image department. because everyday clothing from previous eras has not generally survived , the collection is dominated by fashionable clothes made for special occasions. one of the first significant gifts of costume came in 1913 when the v&a received the talbot hughes collection containing 1,442 costumes and items as a gift from harrods following its display at the nearby department store.	caliphs actioner portray antibody crow intimidating escalating obama peopled.[improper promising 22,392 as-205/208 neutrons

Table 10: Examples of GAN output mode collapse using Gumbel-Softmax approach

Prompt	Generated Response
Can an AI model have consciousness?	What are you trying to determine should could could should could could could might could could
As scientists, are you ever worried about what you might discover?	<|endoftext|>
How do eye drops work?	They simply light the lens [http://www.youtube.com/httphttphttp httphttphttp

Table 11: Examples of decay in generated response

Prompt	Generated Response (Iter 10)	Generated Response (Iter 50)
Please make it more positive.	It's a really well balanced, but not dull, book. 8/10. Joel Slater was very cleverly focused in his covers of Charl- 查尔顿·赫斯顿在《夺宝奇兵 最后的十字军东征》中。<|endoftext|>	..(.(.(.(.(.(.(.(.(.(.(.(.(.(.(g).. Copy- 右键版权 版权版权- 版权版权 版权复制- 右键版权版权 版权- CopyrightCopyright<|endoftext|>
Write a negative review.	9/10 you wish you had the movie. —Kiyo Matsumoto- 37 Released 1991 With This Movie. ****Shoe Halloween Remake Dead End Movie ***<|endoftext|>	8Advertisements.000088873051100836 accAdvertisements AdvertisementsfAdvertisementsAdvertisementst AdvertisementsAdvertisementsAdvertisements AdvertisementsFollowct<|endoftext|>

continued to decrease until the discriminator's loss reached zero. This is consistent in our observation of a decay in the generated response as the number of iterations increased, as illustrated in Table 11.

B Instruction Tuning Training Statistics

We used Huggingace Tramsformers.trainer [6] to do SFT fine-tuning and tlr library to do RLHF and RLGAF PPO training [7]. Table 12 shows the training time taken for different instruction tuning approaches.

Method \ GPU	T4	A100
SFT	145.37	N.A.
RLHF	N.A.	243.91
RLGAF	N.A.	1467.41

Table 12: Training time per epoch for instruction tuning. Unit: seconds. Sample size per epoch: SFT: 9314, RLHF: 32, RLGF, 10 for discriminator and 100 for generator.

Plots in Fig. 8 and 9 show:

- The reward the model obtained during RLHF training.
- The reward the generator obtained and the training and evaluation accuracy the discriminator gets during RLGAF training.
- The loss the model obtained during SFT training.

It can be seen the reward plot for RLHF did not improve too much over time. This might suggest that compared to RLHF, in RLGAF is easier to improve the LLM, because initially

[6] https://huggingface.co/docs/transformers/main_classes/trainer
[7] https://github.com/lvwerra/trl

the reward model (i.e., discriminator) is not very good at critiquing the bad output from the LLM (i.e., generator) and over time both models improve together to reach performant states.

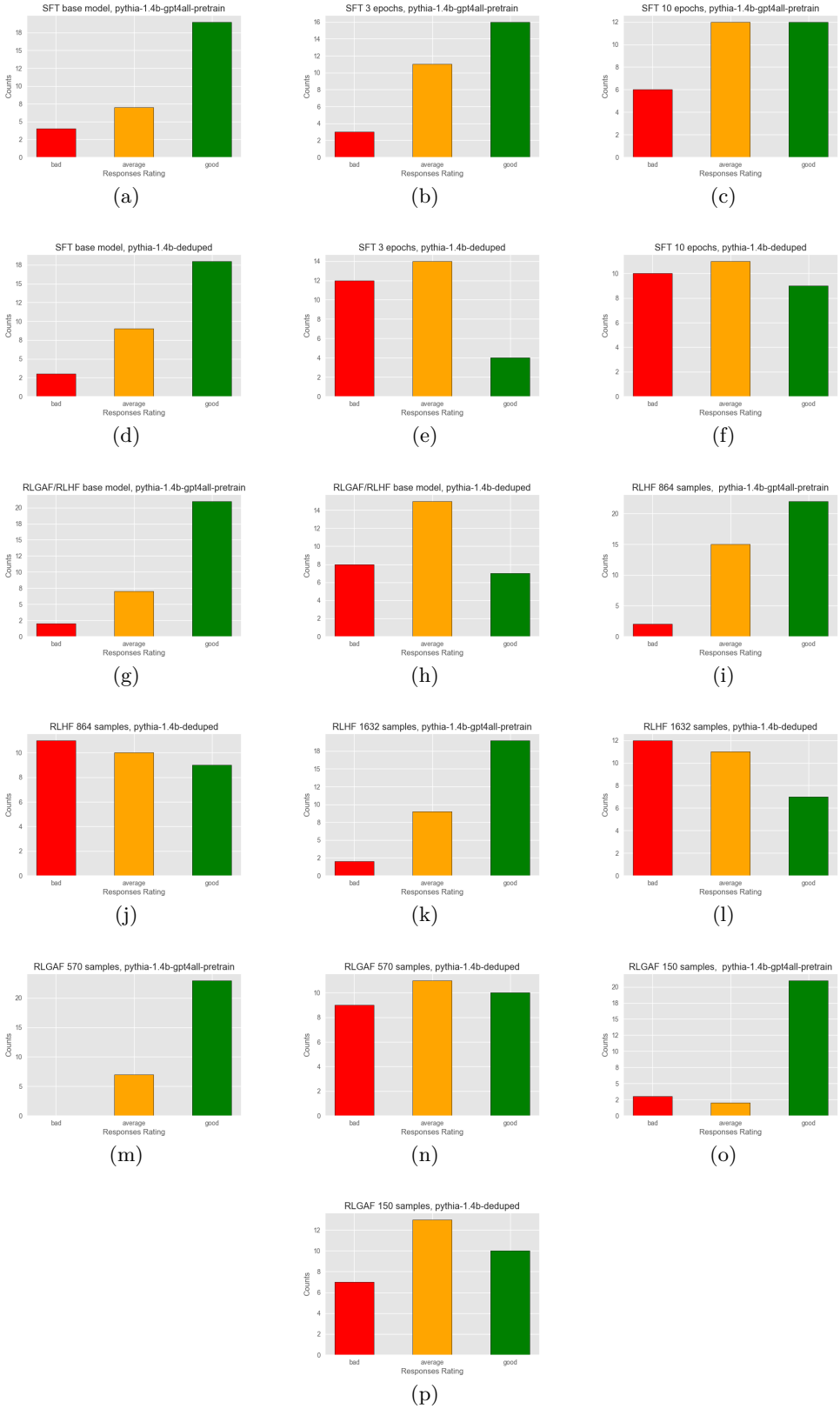


Figure 6: Histograms of the scoring for all models and approaches.

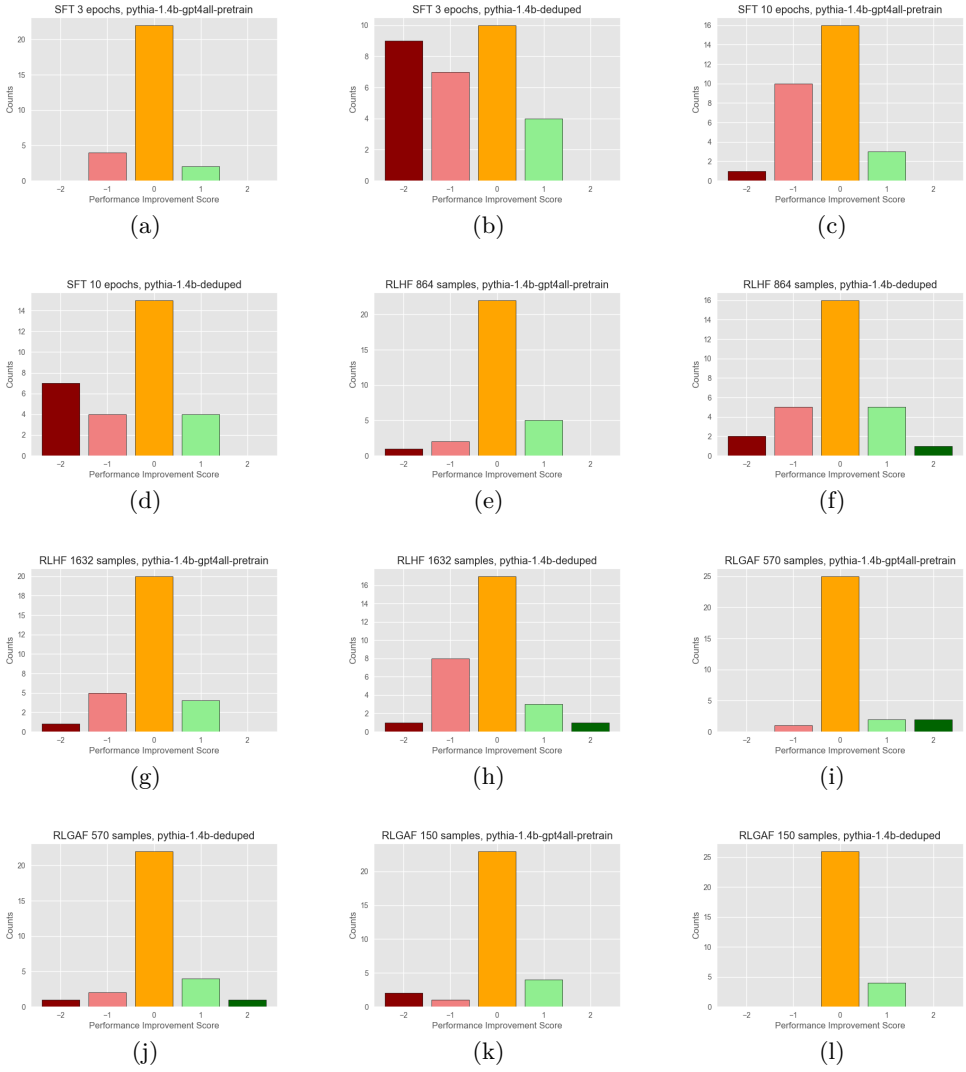
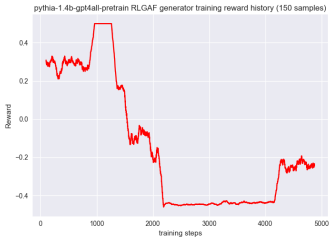
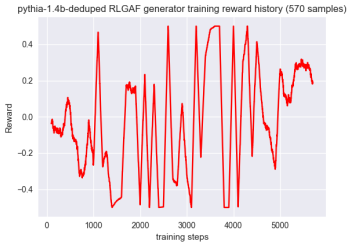


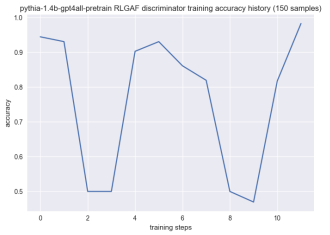
Figure 7: Histograms of the performance improvement score for different models and methods.



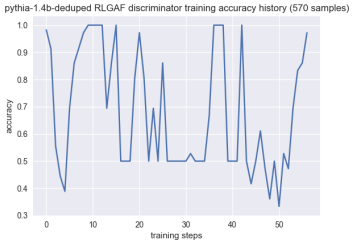
(a)



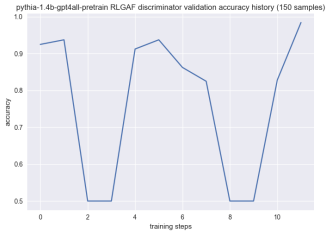
(b)



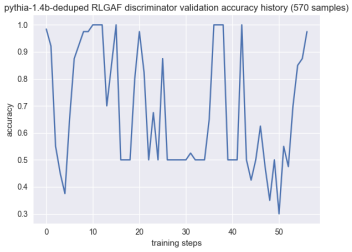
(c)



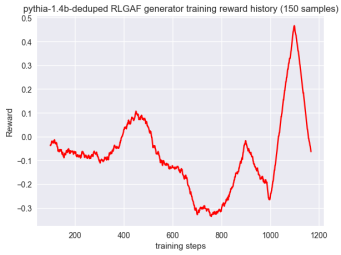
(d)



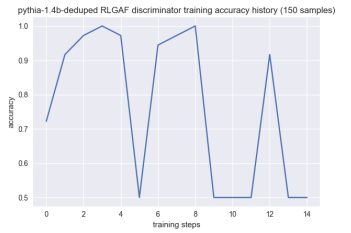
(e)



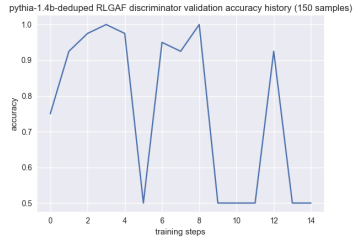
(f)



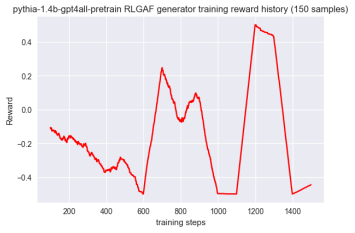
(g)



(h)



(i)



(j)

Figure 8: Training history for all approaches (part 1). RLHF and RLGAF rewards are smoothed with window size=100.

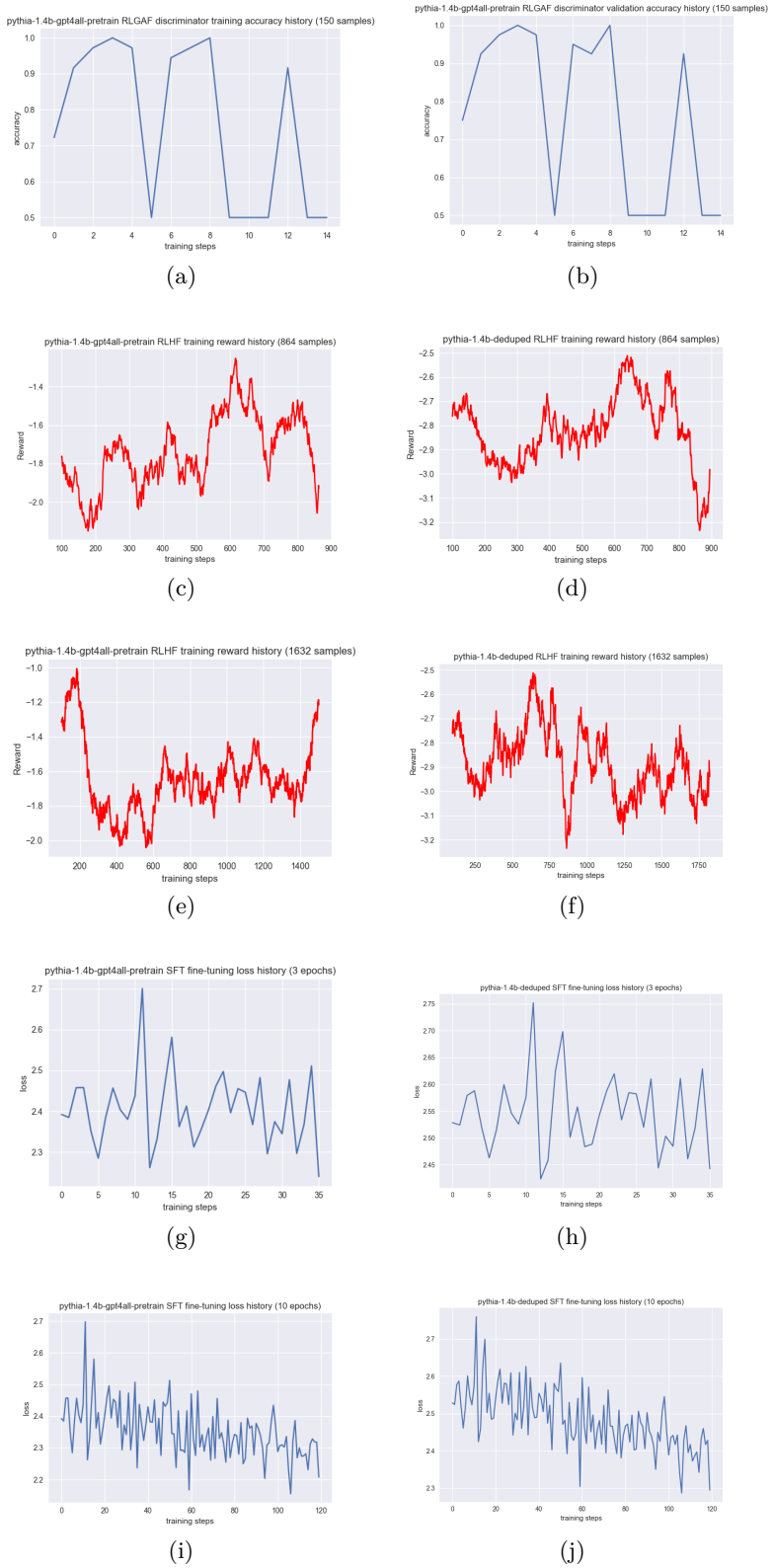


Figure 9: Training history for all approaches (part 2). RLHF and RLGAF rewards are smoothed with window size=100.