# STA5007 Project 2 Report

12112627 李乐平

## - Q1 -

### Paper Chosen

Secrets of RLHF in Large Language Models Part I: PPO

### What is the problem addressed in the paper?

此论文中着力研究的问题是：在大语言模型飞速发展的同时，其会不受控制地生成不符合策略（policy）或人类价值观（如 3H：Helpful、Honest 和 Harmless）的输出。此前，OpenAI 等公司已经验证了 RLHF（Reinforcement Learning with Human Feedback，基于人类反馈的强化学习）在解决此问题上的有效性，但是基于 RLHF 的方法并不总是稳定，训练经常失败。

通常而言，成功的 RLHF 训练要求一个能够替代人类决断的奖励模型（reward model）、能够使参数稳定更新的超参数和强有力的用于稳定策略优化的 PPO（Proximal Policy Optimization，近端策略优化）算法。在论文中，作者剖析了 RLHF 框架的完整流程，并探究了每一个部分对于训练能否成功的影响。最后提出了 PPO-max 优化算法，减轻了传统 PPO 算法的不稳定性。

### Is this a new problem?

论文是基于原 PPO 算法的缺陷做出的改进，自然不是研究的一个新的问题。

### What is the scientific hypothesis that the paper is trying to verify?

论文没有做出特别的假设。论文基于 PPO 算法各个部分的不同实现，根据其设置的度量挑选出了相对最优的一种实现，并命名为 PPO-max。论文还提出了一些具体实现中的技巧，并加以了验证。

### What are the key related works and who are the key people working on this topic?

论文主要围绕基于如下领域展开：使用 RLHF 使语言模型的输出无害化和 PPO 等。首先论文的基础是大语言模型的存在，这当然得提到 Brown 等人的 Language Models are Few Shot Learners，其创造性地提出了 GPT-3 大模型。关于无害化，有 Ouyang 等人的 Training Language Models to Follow Instructions with Human Feedback、Bai 等人的 Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback 等文章。而关于策略优化的领域，有 Andrychowicz 等人的 What Matters for On-policy Deep Actor-critic Methods? A Large-scale Study 和 Engstrom 等人的 Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO 等文章。

### What is the key of the proposed solution in the paper?

解决问题的关键在于奖励模型（Reward Model, RM）和其提出的 PPO-max 算法。奖励模型是用于模仿、替代人类反馈的重要模型。而 PPO-max 算法是一种改进的 PPO 算法，其通过选择 PPO 的各个组件实现，能够有效地提高策略模型的训练稳定性，并通过引入多种优化技巧如 KL 惩罚项、梯度裁剪等和约束条件，避免策略模型过度优化和模式崩溃的问题。

## How are the experiments designed?

论文对于奖励模型和 PPO 分别进行了实验。

在对奖励模型的实验中，设置了在前 10%的预热步骤中 5e-6 的学习率。在训练中使用了动态批处理方法以保证每一批次的 token 数相近，一批的大小从 4 至 128 不等。训练步骤固定为 1000 步，对于训练集平均训练了 1.06 个 epoch。在测试集上评估了模型的准确性，在中文和英文数据集的准确性分别稳定在 80%和 70%左右。

在对 PPO 的实验中主要使用的是中文数据，四个组件模型（策略模型、评论者模型、奖励模型和参考模型）都需要在训练阶段加载。参考模型和策略模型均从 7B 的有监督微调模型初始化，其使用的是 OpenChineseLLaMA 在一百万过滤后的指令数据上有监督地微调了 2 个 epoch 后得到的模型。设定的学习率为 9.5e-6，并采用了余弦学习率调度，最终学习率会衰减到峰值的 10%。在此之后，又训练了评论者模型和奖励模型。

关于量度部分，论文关注了若干指标，如策略模型和参考模型之间的 KL 散度、响应长度、VF 损失等。

在实验过程中，论文详细讨论了不同的得分重参数化方法，包括奖励缩放、奖励归一化和剪切，以及优势归一化和剪切。作者发现对奖励和优势进行严格的约束可以维持 PPO 训练的稳定性，同时在不同的超参数和技巧下进行了对比实验，以评估它们的效果。此外，论文介绍了策略约束方法，如标记级 KL-Penalty、重要采样和熵奖励，并通过实验证明了这些方法对策略优化的影响。

之后，论文中进一步讨论了预训练初始化的重要性。作者通过在训练的早期阶段尝试不同的初始化方法，比较了评论模型和策略模型的初始化对 PPO 训练的影响。论文强调了评论模型的预训练对于提高训练稳定性的重要性，并建议将其作为通用的初始化策略。此外，论文认为策略模型需要经过监督微调以适应 PPO 训练，并强调了构建一个受监督的策略模型对于 RLHF 的重要性。

最后，论文描述了 PPO-max 算法中的实现。通过结合在 5.3 节中讨论和验证的各个组件的最有效策略，论文设计了一套综合性的 PPO-max 实验设置。具体地，作者在训练过程中采用了标准化、剪切、KL-penalty、预训练初始化、全局梯度裁剪、经验缓冲区控制以及预训练语言模型损失等策略，以最大程度地提高 PPO 训练的效果。

此外，论文中还阐展开了对 RLHF 模型与 SFT 模型的对比评估。实验着重在两个方面进行：与人类评估的对比和与 ChatGPT 模型的性能比较。

首先，通过与人类评估者进行对比，作者采用了“有害”和“无害”两个维度的评估。在这方面，RLHF 模型在英文和中文数据集上都显著优于 SFT 模型，特别是在处理涉及个人隐私、政治敏感性以及有毒和有偏见提示等问题时。人类评估结果与 GPT-4 的评估结果相吻合，进一步验证了 RLHF 模型的优越性。

其次，与 ChatGPT 的性能比较主要关注“无害”能力。尽管 RLHF 模型仍然不及 ChatGPT，但相较于 SFT 模型，它们在减轻面对 ChatGPT 的失败方面表现更好。这表明 RLHF 方法提升了模型生成更有效响应的能力，尽管超越 ChatGPT 仍然是一个具有挑战性的目标。

最后，为了评估微调过程对自然语言理解（NLU）的影响，作者使用了 C-Eval5 测试，发现通过在 PPO 训练阶段引入预训练数据，可以有效减缓 NLU 能力的下降。

**What datasets are built/used for quantitative evaluation? Is the code open sourced?**

实验采用了 HH-RLHF 和自制数据集等，语种包含英文和中文。其中中文数据集请了专业人员进行标注（Helpful, Harmless）。此外，还使用了手工制作的 HH 数据集。文中给出了源代码的链接，代码是开源的。

**Is the scientific hypothesis well supported by evidence in the experiments?**

实验表明，相比于普通的 SFT，论文中实现的 RLHF 具有相对的优势，也缩小了与 ChatGPT 的差距，但依然还有改进的空间。

**What are the contributions of the paper?**

论文研究了近端策略优化的各个组件，并提出了 PPO-max，在某种程度上降低了 RLHF 训练失败的概率。

**What should/could be done next?**

根据论文中提到的局限性，有以下工作可以继续探索：探究模型大小和数据规模对 RLHF 性能的影响；通过更好的数据集改进奖励模型；使用更多的评估指标进行评估；寻找更合适的性能指标，尝试改进训练的效果。

## - Q2 -

**Large language model (LLM) finetuning and evaluation.**

**1. [4 points] Finetune TinyLLaMA-1b with clinical instruction data (see attached) with QLoRA method.**

我下载了 QLoRA 的微调脚本，并在上面稍作修改之后使用以下命令进行了微调。

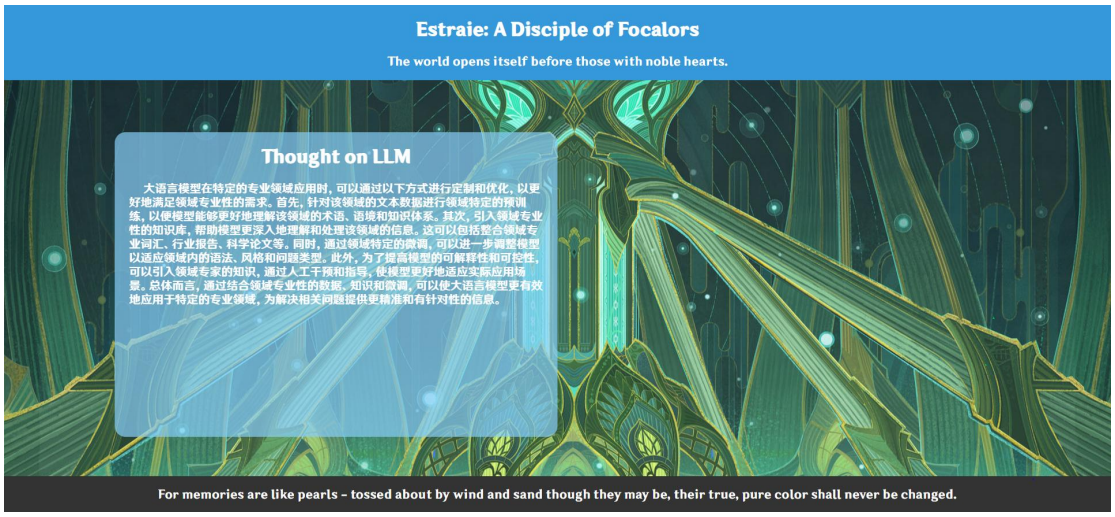
```
python qlora.py --model_name_or_path ../model/ --dataset ../data/Q2_\(1\)_iCliniq_data.json --dataset_format input-icliniq --max_steps 1000
```

**2. [4 points] Evaluate the finetuned LLM on MedMCQA dataset (see attached). As the test set has no answers, we will evaluate on the validation dataset.**

因为 TinyLLaMA 的测试效果不佳，而且没有很好的手段让其在没有训练的情况下做选择题，因此我尝试将其输出的答案与正确答案做比较，并计算 BLEU 和 CIDEr 等文本相似性指标。得到 BLEU 指标为 6.51，CIDEr 指标为 $5.869\times10^{-10}$。这是一个非常低的分数，意味着此 LLM 基本无法迁移到该数据集上。

**3. [4 points] Build a personal website and display your thoughts and ideas about how to apply LLMs in specific domains like clinic/law/finance.**

请参见我的个人网站，初次加载可能会花费稍多的时间，因为涉及滚轮操作，没来得及很好地适配移动端，所以请使用 PC 端访问。请打开硬件加速。


Estraie: A Disciple of Focalors
The world opens itself before those with noble hearts.
Thought on LLM
大语言模型在特定的专业领域应用时，可以通过以下方式进行定制和优化，以更好地满足领域专业性的需求。首先，针对该领域的文本数据进行领域特定的预训练，以便模型能够更好地理解该领域的术语、语境和知识体系。其次，引入领域专业性的知识库，帮助模型更深入地理解和处理该领域的信息。这可以包括整合领域专业词汇、行业报告、科学论文等。同时，通过领域特定的微调，可以进一步调整模型以适应领域内的语法、风格和问题类型。此外，为了提高模型的可解释性和可控性，可以引入领域专家的知识，通过人工干预和指导，使模型更好地适应实际应用场景。总体而言，通过结合领域专业性的数据、知识和微调，可以使大语言模型更有效地应用于特定的专业领域，为解决相关问题提供更精准和有针对性的信息。
For memories are like pearls – tossed about by wind and sand though they may be, their true, pure color shall never be changed.


Estraie: A Disciple of Focalors
The world opens itself before those with noble hearts.
Thought on LLM
例如在金融领域，大语言模型可以通过预训练和微调来理解金融术语、市场动态和经济指标。模型可以从金融新闻、财报和交易数据中学习，以更准确地预测市场趋势、分析公司业绩，甚至生成投资建议。引入金融领域专业的知识库，包括金融法规、风险管理模型等，有助于模型更全面地理解金融业务。此外，模型也可用于自动化任务，如合规性检查、金融文档的摘要和理解。
在医疗领域，大语言模型可通过对医学文献、病历和临床实验数据的预训练来学习医学知识。微调模型时，可以结合特定疾病领域的数据，使其具备更专业的诊断和治疗建议能力。引入医学百科全书、药物数据库等领域专业知识，有助于模型更好地理解医学领域的复杂性。模型在医疗保健中的应用可包括辅助医生诊断、提供患者教育材料、分析大规模的医疗数据以发现潜在的疾病模式等。
在法律领域，大语言模型可通过学习法律文件、案例法和法规文本来理解法律术语和文本结构。微调时，结合特定法律领域的数据，使模型能够更精确地进行法律分析、文本摘要和法律咨询。引入法学教材、判例数据库等领域专业知识，有助于模型更好地理解法律语境。模型在法律领域的应用包括合同分析、法律文件撰写辅助、法律研究等，为法律专业人士提供更高效和准确的信息支持。
For memories are like pearls – tossed about by wind and sand though they may be, their true, pure color shall never be changed.

# - Appendix -

因为涉及的问题较多，我在此只能提供部分有代表性的 prompt。使用的大语言模型均为 ChatGPT。

请给出教程，不依赖任何框架建立一个个人站点
（https://chat.openai.com/share/0e3ed5fa-994d-4f3d-9345-79095bd523ad）

……请总结上述文本（5.3 节-5.4 节）的实验是如何设计的
（https://chat.openai.com/share/436d4642-afed-4a2a-9b2e-6dee2bd804d8）