大模型对齐范式：RLHF、偏好优化与价值观安全实践

2025 年 10 月 25 日

1 人类反馈强化学习（RLHF: Reward Model + PPO）

1.1 整体流程与系统架构

RLHF（Reinforcement Learning from Human Feedback）通过人类偏好信号建立奖励模型，并借助强化学习算法（例如 PPO）优化语言模型策略。典型流水线包含三个阶段：

1. **监督微调（SFT）基线：**以高质量对话或任务数据对模型进行初始微调，获得稳定的参考策略 π_{SFT}。
2. **奖励模型训练：**采集同一提示下的多条候选回复，由标注者排序或选择偏好，训练比较式奖励模型 $r_\phi(x, y)$。
3. **策略优化：**使用 PPO 或其变体最优化策略 π_θ，最大化期望奖励并加入 KL 约束保持与 π_{SFT} 接近。

流水线需要高质量的标注与强大的计算基础设施，通常涉及打标平台、数据版本控制与实验跟踪系统。

1.2 奖励模型训练细节

奖励模型通常采用与基础模型共享的 Transformer 编码器，仅在最后增加标量头。关键实践：

- **偏好数据采集：**通过对比问卷或滑动条形式收集人类排序，确保覆盖常见任务与安全场景。

- **Loss 设计**：使用 Bradley-Terry 或双对数似然损失，形式为

$$\mathcal{L} = -\log \sigma\big(r_\phi(y^+) - r_\phi(y^-)\big), \tag{1}$$

  其中 y^+、y^- 表示优劣回复。为了防止奖励溢出，可对奖励进行标准化或裁剪。

- **泛化与稳健性**：引入 Dropout、数据增强（如随机截断）、对比正则，对奖励模型进行校准；使用 held-out set 衡量奖励排行准确率。

1.3 策略优化与 KL 约束

在 PPO 阶段，需要平衡奖励最大化与策略偏移控制：

- **KL 惩罚**：在目标函数中加入 $-\beta\,\mathrm{KL}\big(\pi_\theta \| \pi_{\mathrm{SFT}}\big)$，或通过自适应系数控制策略与原模型距离。

- **批量采样**：使用多 GPU 并行生成候选回复，计算优势（Advantage）并执行 PPO 更新；常见设置为 512–2048 序列每批。

- **监督混合**：定期将新策略与 SFT 数据混合训练（supervised replay），防止遗忘核心指令遵循能力。

实际部署需维持稳定的奖励信号，可使用 reward modeling + rejection sampling 的组合提升生成质量。

1.4 评估与监控

RLHF 调优后的模型需经过多维评估：

- **自动评估**：使用 reward model、GPT-4 评审或指标库对回答进行打分。

- **人类评估**：进行 A/B 测试或多维度打分（有用性、安全性、事实性）。

- **在线监控**：在部署环境收集互动日志，监测负反馈、拒答率，与安全策略联动。

应建立反馈闭环，对失误案例进行再标注与迭代训练。

2 直接偏好优化（DPO）

2.1 原理与目标函数

DPO（Direct Preference Optimization）通过解析形式推导，将偏好对比直接融入策略优化，无需显式 PPO。核心思想是最大化策略在偏好对上的对数比值，目标函数

为：

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)}\left[\log\sigma\left(\beta\big(\log\pi_\theta(y^+\mid x) - \log\pi_\theta(y^-\mid x)\big) - \log\pi_{\mathrm{ref}}(y^+\mid x) + \log\pi_{\mathrm{ref}}(y^-\mid x)\big)\right], \tag{2}$$

其中 π_{ref} 通常为 SFT 策略，β 控制 KL 强度。

2.2 训练流程与实现要点

DPO 训练类似 SFT，只是损失函数替换为偏好对迁移：

- **数据准备**：需要成对偏好样本，可直接复用 RLHF 阶段的比较数据。
- **批处理策略**：使用全序列拼接处理 y^+ 与 y^-，计算对数概率时需注意 mask，避免跨样本梯度干扰。
- **参考模型冻结**：π_{ref} 不更新，通过半精度加载以节省显存。训练模型可以是原模型复制或 LoRA 适配。

由于 DPO 不依赖奖励模型，训练更稳定，且易于与现有 SFT 框架集成。

2.3 优缺点与扩展变体

相较 RLHF，DPO 具有：

- **优势**：不需采样奖励模型；单阶段训练节省算力；易于调参。
- **劣势**：对偏好数据质量更敏感；缺乏显式奖励模型意味着线上监控和理解成本较高。
- **扩展**：IPO（Implicit Preference Optimization）、KTO（Kahneman-Tversky Optimization）等在目标函数上引入噪声鲁棒性或损失重加权；Online DPO 结合部署反馈进行增量更新。

3 Constitutional AI 与自对齐（Self-Alignment）

3.1 理念与总体流程

Constitutional AI 由 Anthropic 提出，旨在减少人类标注依赖，通过一套“宪法”原则指导模型自我改写与评估。核心步骤：

1. **宪法原则定义**：由专家撰写涵盖安全、伦理、事实性的指导条款。
2. **自监督批评**：基于原则让模型生成自我审查，指出回答中的问题或改进建议。
3. **自我修正**：模型根据批评更新或改写回答，形成更符合原则的输出。

整个过程可迭代执行，逐步提高模型对齐水平。

3.2 批评与改写策略

批评阶段可采用多种提示模板：

- **单轮批评**：给定原回答与原则，请模型指出违反条款的部分并给出理由。
- **多轮批评**：引入批评助手与被批评助手的对话，模拟教学过程。
- **交叉批评**：使用不同模型或不同温度的生成来相互批评，增加多样性。

改写阶段在批评反馈基础上生成新的回答，可加入明确约束，如“保持事实准确”“避免冒犯性语言”。

3.3 自对齐与人类反馈结合

自对齐结果仍需人类验证，以防模型自举偏差：

- **混合标注**：将自对齐生成的数据与人类评审样本混合训练奖励模型或 DPO。
- **持续宪法迭代**：根据部署反馈更新原则条目，吸收新场景需求。
- **评估指标**：跟踪拒答准确率、敏感话题合规性、事实性提升幅度。

自对齐在高风险领域需结合正式伦理审查和法律合规流程。

4 安全性与价值观对齐（Safety, Bias, Toxicity）

4.1 对齐风险识别与分类

需要建立全面的风险分类体系：

- **安全风险**：包含暴力、恐怖主义、武器制造等危害性输出。
- **偏见与歧视**：针对性别、种族、宗教等群体的偏见性语言。
- **虚假信息**：包括事实性错误、伪科学、诈骗诱导。
- **隐私泄露**：泄露个人信息或敏感数据。

为每类风险制定检测与缓解策略，是安全对齐的第一步。

4.2 检测与评估框架

多层次评估确保安全指标达标：

- **静态评估**：使用 Jigsaw、RealToxicityPrompts、HolisticBias 等基准量化偏见与毒性。
- **对抗测试**：通过红队（Red Teaming）生成攻击提示，覆盖提示注入、意图伪装等复杂手段。
- **在线监控**：部署实时过滤器与审计日志，追踪异常输出、用户举报和策略迭代效果。

评估结果需要形成闭环，反馈到数据采集与训练阶段。

4.3 缓解策略与工程实现

安全对齐结合多种技术手段：

- **数据阶段**：构建安全提示语料、拒答示例、敏感场景模拟数据；在训练中增加惩罚项或加权采样。
- **模型阶段**：继承 RLHF/DPO 对安全场景进行针对性调优；使用安全奖励模型或安全 DPO 进行专门优化。
- **推理阶段**：应用内容过滤器、敏感话题分类器；采用两阶段生成（先判断再回答）或工具审查。

同时需明确治理流程、责任人和应急机制，确保产品上线后能够快速响应问题。

参考文献

- Ouyang et al. “Training Language Models to Follow Instructions with Human Feedback.” NeurIPS, 2022.
- Bai et al. “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.” arXiv, 2022.
- Rafailov et al. “Direct Preference Optimization: Your Language Model is Secretly a Reward Model.” arXiv, 2023.
- Bai et al. “Constitutional AI: Harmlessness from AI Feedback.” arXiv, 2022.
- Ganguli et al. “Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.” arXiv, 2022.