中国石油大学
CHINA UNIVERSITY OF PETROLEUM

《石化过程智能制造》

课程报告

智能学习——汽油调合强化学习实验与 Web 分析平台

学号	2026245511
姓名	刘鹏飞
学院	人工智能学院
学科专业	工程管理专业
教师	何仁初

完成时间：2026 年 11 月 15 日

目　录

引言 3

1 课程目的与研究内容 3

2 汽油调合问题与仿真环境 4

3 强化学习建模与 PPO 微调 5

4 执行前动作约束检查 7

5 系统设计与 Web 实现 7

6 实验设置与结果分析 9

7 讨论与适用范围 11

8 结论与展望 11

9 体会、收获和建议 11

参考文献 11

附录 程序与复现说明 12

一级方向：智能学习
二级题目：汽油调合强化学习实验与 Web 分析平台

智能学习——汽油调合强化学习实验与 Web 分析平台

刘鹏飞　2026245511　工程管理 26-1

人工智能学院　工程管理专业

摘要：针对汽油调合中组分性质波动引起的成本与质量约束冲突，基于课程提供的四组分调合仿真环境及初始策略，设计包含质量超限、配方越界和调整幅度惩罚的新奖励函数，采用近端策略优化算法进行微调，并增加仅使用当前观测的执行前动作约束检查。原奖励与新奖励分别使用 3 个随机种子训练，每次交互 34,368 步，累计 206,208 步；固定最终模型，在 8 个本次未用于调参的场景中评价五组策略。新奖励微调的调合头违规率为 3.28 %，低于原奖励微调的 3.85 %，但高于原始策略的 1.57 %。原始策略加约束与新奖励微调加约束两组在当前样本中均为零调合头违规，平均成本指数分别为 4544.2070 和 4541.6562。结果表明，奖励修改在继续训练对照中具有相对改善，但不足以证明微调全面优于原始策略，执行前检查仍有保留价值。进一步实现 Web 轨迹回放和 DeepSeek 候选生成，本地检查 5 个候选中有 2 个预测可行，候选尚未执行。研究结论限于课程仿真及有限评价样本。

关键词：汽油调合；强化学习；PPO；奖励设计；质量约束；Web 可视化

中图分类号：TP18　　文献标识码：A

引言

汽油调合是将不同加工单元提供的组分油按适当比例混合，使成品质量满足要求的过程。课程以汽油管道调合为应用场景，介绍智能体与环境交互、Actor–Critic 结构及调合过程优化[1]。对软件开发背景的学习者而言，关键是把工艺目标转化为明确的状态、动作、约束和评价指标，而不是仅展示一个神经网络接口。

调合优化不仅涉及原料成本，还涉及组分性质不确定性、质量约束和反馈信息。Chèbre 等研究了燃料调合中的反馈优化和组分性质估计，指出输入输出约束及测量不确定性对实际调合控制的重要性[2]。Huang 等将汽油调合描述为马尔可夫决策过程，并采用 SAC 研究非线性调合和原料波动[3]。这些研究支持将调合作为连续决策问题讨论，但其数据来源与算法结果不能直接作为本课程实验的证据。

本文在课程仿真基础上研究两个问题：加重质量惩罚后，继续训练是否减少过程违规；将动作检查放在执行前，能否在当前样本中改善约束满足情况。研究同时建设可追溯的 Web 分析页面，使评价结论能够回到具体场景、步数和配方变化中核对。

1 课程目的与研究内容

1.1 学习目标

本研究对应“智能学习”方向。学习目标包括理解石化连续过程中的经济性与质量约束；把调合决策表示为强化学习任务；实现并解释奖励设计、优势估计和策略更新；建立相同场景条件下的对照评价；通过可视化区分模型建议、检查后的动作和实际执行结果。强化学习通过与环境交互优化累计回报，区别于直接学习固定输入输出标签的监督学习[4]。

1.2 课程基础与拓展范围

课程基础包括 AgentRun.py 中的调合动力学、组分参数、32 维状态、Actor 与 Critic 网络结构，以及初始 Actor 权重。本文保留该环境计算逻辑，利用 ScenarioEnv 覆盖随机种子选择，以构造不同仿真场景。新增工作包括 PPO 微调脚本、质量导向奖励、候选动作检查、对照统计复核、Web 回放及 DeepSeek 辅助接口。

本文采用已有 PPO 算法进行课程应用拓展，不将其写成自行提出的新算法。初始权重的完整历史训练过程未知，因此“新场景”仅表示未参与本次调参，不能据此宣称相对原始模型完全独立的泛化测试。

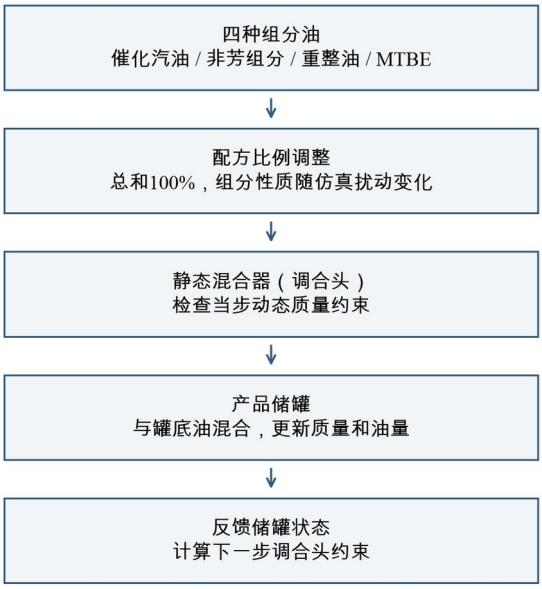


图 1　汽油调合过程与状态反馈

资料来源：依据课程课件和 AgentRun.py 绘制。

2 汽油调合问题与仿真环境

2.1 组分性质与配方约束

模型包含催化汽油、非芳组分、重整油和 MTBE 四种组分。初始比例为 46 %、8 %、32 % 和 14 %，成本系数分别为 4300、3900、5000 和 6000。组分基准 RON 依次为 92.5、74.5、103 和 108。表 1 给出基本输入，反映不同组分的质量与成本差异；实际决策使用当步观测，而不是始终使用基准值。

表 1　组分初始参数

组分	初始比例 / %	基准 RON	成本系数
催化汽油	46	92.5	4300
非芳组分	8	74.5	3900
重整油	32	103	5000
MTBE	14	108	6000

设 x_i 为第 i 种组分的百分比，$w_i=x_i/100$。配方要求总和为 100 %，各组分不低于 0.1 %，相对初始配方的累计偏移不超过 10 个百分点。成本指数按体积比例加权计算：

$$C_t = \sum_i w_{i,t}\, c_i \quad (1)$$

其中 c_i 为课程给定成本系数。本文报告的是新混合油在各步的成本指数及其均值，不计入罐底油既有成本，也不将其解释为严格的元/吨或真实生产节约金额。

2.2 质量计算与动态限制

课程质量向量按 RON、密度下限项、密度上限项、硫含量、烯烃和芳烃排列。密度重复存储是为了分别检查上下限，物理指标实际为五类。固定储罐限制为[95,720,775,10,15,35]，方向为[下限、下限、上限、上限、上限、上限]。这些数值作为课程仿真参数使用，本文不将其认定为现行产品标准的完整条款。

RON 采用课程的非线性变换。先把各组分 RON 变为调合指数 τ，再按体积比例混合并逆变换；密度、烯烃和芳烃按体积比例计算，硫含量按质量贡献加权。

$$\tau(R) = 1 / [1 + \exp(-b(R-m))] \quad (2)$$

$$R_{mix} = m - \ln[(\sum_i w_i\, \tau(R_i))^{-1} - 1]/b \quad (3)$$

式中 b=0.09594，m=102.94，来自课程代码。储罐更新把本步新油与已有罐底油再次混合，RON 仍使用相同非线性公式。初始归一化油量为 0.1，每步增加 0.0025，达到 0.994 附近的终止阈值时结束；现有完整轨迹均运行 358 步，最终油量约 0.995。步数表示仿真决策次数，不等于秒数。

limcal 根据储罐当前质量、油量和剩余容量反推剩余混合油可接受的质量范围，再结合当前组分性质构造动态限制。因此，调合头某一步的允许下限不一定等于最终储罐固定下限。储罐早期积累的质量余量会影响后续决策空间，但额外的 0.1 RON 动作检查余量是另一个人为设置的参数，二者不能混同。

2.3 数据生成与场景含义

数据来自课程仿真，不是现场采集数据。gp_generate 为四种组分生成长度 400 的相关高斯过程序列，使用平方指数核，并拒绝超出±2.5 范围的序列。当前实现把该序列用于 RON 变化，其他基准性质不因此全部随机变化。罐底油初始 RON、密度、硫、烯烃和芳烃在课程设定区间内抽样。

场景是由随机种子确定的一组原料扰动和储罐初始条件；回合是一个策略在该场景下完成一次调合；步是回合中的一次决策。场景 53 表示种子编号 53，不是第 53 家工厂。同一场景的不同策略从同一环境副本开始，经历相同外生扰动，

但因配方决策不同，其后续储罐状态和动态约束可能不同。

环境 step 先推进组分扰动，再执行动作。动作检查只能使用调用 step 之前可见的性质，因而其预测与下一步实际质量可能不同。评价时保存动作执行前的动态限制，用执行后质量与该限制比较，避免错误地拿执行后重新计算的下一步限制判定当前违规。

3 强化学习建模与 PPO 微调

3.1 状态与动作

图 2 说明训练与执行的关系，图 3 给出状态维度。32 维状态由组分 RON 偏移 4 维、累计配方偏移 4 维、当前配方 4 维、调合头质量 6 维、成本 1 维、储罐质量 6 维、油量 1 维和动态质量限制 6 维组成。std_state 使用课程固定缩放和偏移处理，将不同尺度的输入转换为网络使用的状态向量。



图 2　强化学习交互与训练关系

资料来源：依据 retrain.py、experiment.py 及课程网络结构绘制。



图 3　状态组成与配方动作定义

资料来源：依据 AgentRun.py 的 std_state 和 action2ratio_change。

Actor 网络为 32→32→32→3 的全连接结构，中间层使用 ReLU；确定性评价取输出均值的 tanh 值。训练时从可学习标准差的高斯分布采样潜变量 z，再将 tanh(z)交给环境。前三种组分改变量按 0.1 个百分点取整，第四项取前三项之和的相反数，保持配方总和。Critic 为 32→32→32→1 网络，估计状态价值 $V(s)$，不是动作约束判别器。

3.2 原奖励与质量导向奖励

原奖励已经包含质量和配方惩罚。调合头质量不达标时，其奖励分量为−0.1；达标时采用 $\max[(4750-C)/200,0]$ 鼓励较低成本。罐内质量不合格时另有随油量变化的惩罚，合格分支在实际代码中置零；配方偏移过大时还可能覆盖原奖励或终止。本文没有把原奖励描述为“只看成本、不管质量”。

为加强过程质量约束，定义超限向量 $e_j=\max[d_j(y_j-l_j),0]$，其中 $d=[-1,-1,1,1,1,1]$，y 为执行后质量，l 为执行前动态限制，容差 $\varepsilon=10^{-4}$。新增奖励为：

$$r_{new} = r_{base} - 2I\left(\max_j e_j > \varepsilon\right) - 2\sum_j\left(e_j/q_j\right) - 2I_{recipe} - 0.01\sum_i|\Delta x_i| \quad (4)$$

式(4)中，$q=[1,50,50,10,15,35]$为归一化尺度；Irecipe 表示配方最低比例或累计偏移超限；Δx 为四组分本步实际改变量，单位为百分点。第一项

惩罚是否发生违规，第二项反映严重程度，第三项约束配方，第四项抑制过大调整。这些系数是本次固定实验设置，不是行业标准，也未经过最优性证明。

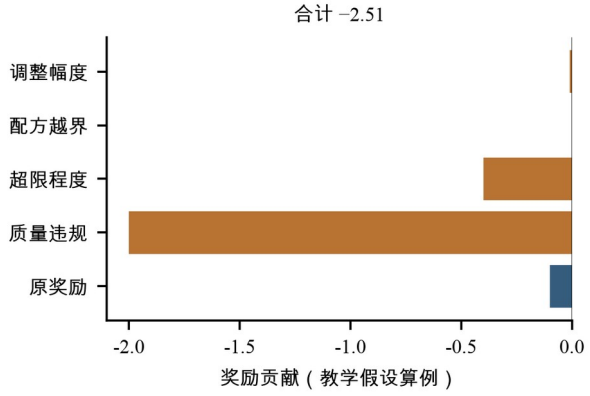


图 4　新奖励的扣分构成示例

资料来源：按 learning_reward 计算；假设原奖励−0.1、RON 超限 0.2、调整总量 1、配方未越界，非实测回合。

该例中质量违规扣 2 分、归一化超限扣 0.4 分、调整扣 0.01 分，新奖励为−2.51。奖励影响训练倾向，却不能强制执行结果达标。因此，奖励设计与执行前检查分别承担学习引导和动作筛选作用。

3.3 PPO 与优势估计

近端策略优化 PPO 通过裁剪策略概率比构造代理目标，允许对同一批采样进行多轮小批次更新[5]。本实现保存未经过 tanh 的潜变量动作及其旧对数概率，更新时在同一潜变量空间计算概率比；不能把量化后已执行的配方改变量直接代入高斯概率公式。

$$\rho_t = \pi_\theta(z_t|s_t) \,/\, \pi_{old}(z_t|s_t) \quad (5)$$

$$L_{clip} = E[\,min(\rho_t \hat{A}_t,\ \rho_{clip} \hat{A}_t)\,] \quad (6)$$

其中裁剪后的概率比为 clip(ρ_t,1−η,1+η)，η=0.2。Actor 最小化式(6)的负值，Critic 最小化预测价值与回报目标的均方误差。裁剪作用于代理目标，并不是把每次参数变化或实际质量风险限制在 20 %以内，也不是严格的安全保证。

广义优势估计 GAE 用折扣的时序差分残差构造优势估计[6]。本代码从回合末端向前递推，使用终止掩码防止跨回合传播；随后对采样批次优势做标准化。

$$\delta_t = r_t + \gamma(1-d_t)V(s_{t+1}) - V(s_t) \quad (7)$$

$$\hat{A}_t = \delta_t + \gamma\lambda(1-d_t)\hat{A}_{t+1} \quad (8)$$

γ=0.99 为折扣因子，λ=0.95 控制优势估计中未来残差的权重，d 为终止标志。训练更新 Actor 和 Critic，但保存到最终模型文件的是 Actor 参数；不能把这些文件描述为包含优化器和 Critic 状态的完整断点。

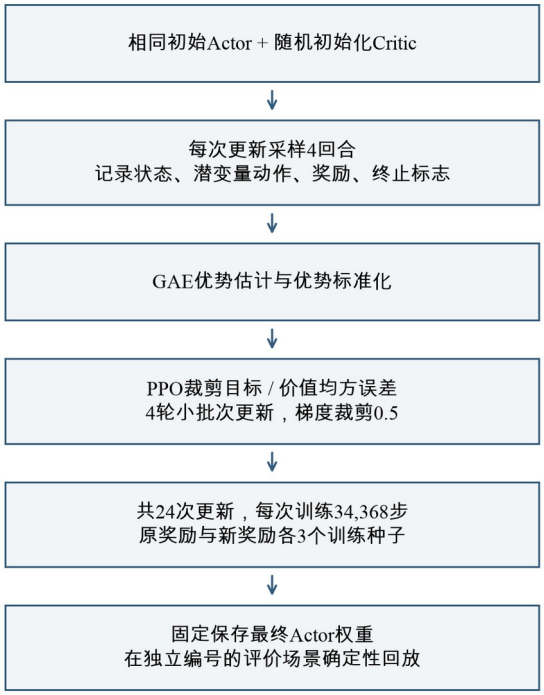


图 5　PPO 微调的采样与更新流程

资料来源：依据 retrain.py 的 train、advantages 及 registered-config.json。

3.4 训练设置与复核

原奖励和新奖励各采用 2026、2027、2028 三个训练种子，Actor 从同一初始权重开始，Critic 随机初始化。每次更新采样 4 回合，共 24 次更新每回合 358 步，每次训练 34,368 步，6 次合计 206,208 步。表 2 列出固定超参数。

表 2　PPO 训练超参数

参数	取值
Actor / Critic 学习率	3×10^{-5} / 3×10^{-4}
折扣 γ / GAE λ	0.99 / 0.95
PPO 裁剪 η / 梯度范数上限	0.2 / 0.5
每批更新轮数 / 小批次目标大小	4 / 256
每次更新回合数 / 更新次数	4 / 24

训练场景 101—112 按固定顺序循环采样，避免两类奖励因抽到不同场景而难以对照。每批

1432 步通过 array_split 分组，实际小批次大小可能略低于 256。统一使用第 24 次更新后的模型，没有根据最终评价结果挑选最好的检查点。本次重新核对 6 份权重哈希、参数变化记录和全部 132 个保存回合的统计；已有复核文件还记录了 3 个新奖励模型在场景 301 的逐步重放一致性。

4 执行前动作约束检查

4.1 当前观测下的候选筛选

action_guard 只接收当前配方、累计偏移、四种组分当前性质、动态质量限制及混合公式参数不接收未来扰动序列。候选由原动作、半幅动作零动作，以及[-1,−0.5,0,0.5,1]三维网格的 125 个组合组成，共 128 条候选记录，允许重复。图 6 展示决策过程。

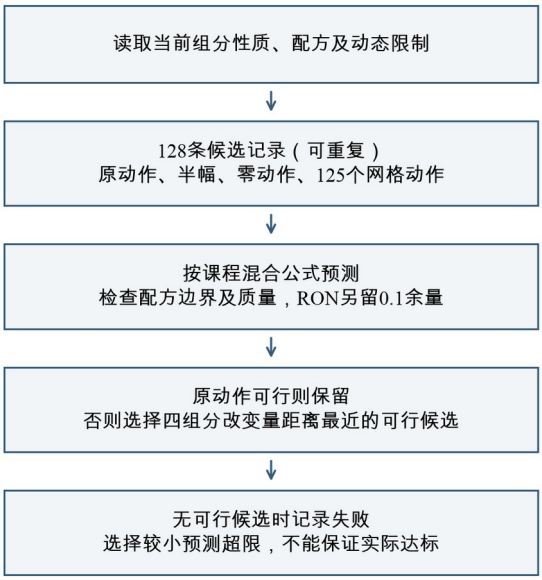


图 6　执行前约束检查流程

资料来源：依据 action_guard.py 的 predict 与 correct_action。

候选先按环境规则转为四组分改变量，然后预测配方与质量。检查要求各组分不低于 0.1 %、累计偏移不超过±10 个百分点，并使预测质量满足动态限制，其中 RON 下限增加 0.1 余量。若原动作满足条件则保留；否则在预测可行集合中选择与原动作四维改变量平方距离最小的候选。

若没有可行候选，则在配方有效候选中优先选择归一化超限平方和较小者，再以动作距离打破并列；若连配方有效候选都不存在，则回退到零动作并记录状态。该回退不证明约束一定满足评价应保留无可行候选次数及实际执行质量。

4.2 场景 53 第 128 步实例

本实例来自旧 *I*2 场景，仅用于解释检查机制不并入后续 8 场景统计。当步 RON 下限为 94.6582，增加 0.1 余量后阈值为 94.7582。Actor 原建议按取整规则计算后的预测 RON 为 94.7444，仍高于原下限，却低于含余量阈值，因此触发修正。

修正将催化汽油减少 0.5 个百分点、重整油增加 0.5 个百分点，另两项不变。修正后的预测 RON 为 94.8105，实际执行 RON 为 94.7873。图 7 比较原预测、修正预测及修正实际值。预测与实际差异来自动作执行时的新扰动及数值计算，不应把预测值直接写成实测结果。

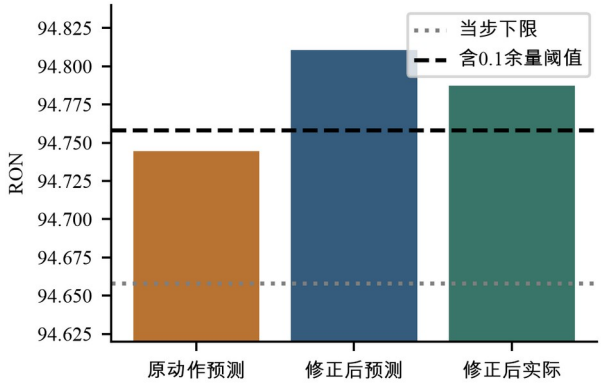


图 7　旧场景 53 第 128 步的 RON 比较

资料来源：旧 experiment.json 与 candidate-observation.json；不同于最终评价场景。

5 系统设计与 Web 实现

5.1 数据组织与可视化

系统把离线训练和在线展示分开。训练输出 Actor 权重和逐次更新日志，评价输出每个场景、策略和步数的配方、成本、实际质量、动作执行前限制及修正信息。Python 服务读取保存结果，浏览器按需请求指定轨迹；切换页面和拖动滑块不会重新训练模型。图 8 给出数据流。



图 8　实验、接口与 Web 的数据流

资料来源：依据 src、web 和保存结果结构绘制。

Web 同时保留历史 *V*2 比较和本轮 PPO 微调评价，界面用场景、策略、训练种子和步数定位记录。质量图将实际 RON 与当步动态下限并列，能够定位超限发生的阶段；配方、成本和动作记录用于解释该阶段的策略行为。图 9 按 Web 回放中的相同字段重绘实际记录，统一图中文字字号并保留第 128 步定位线。

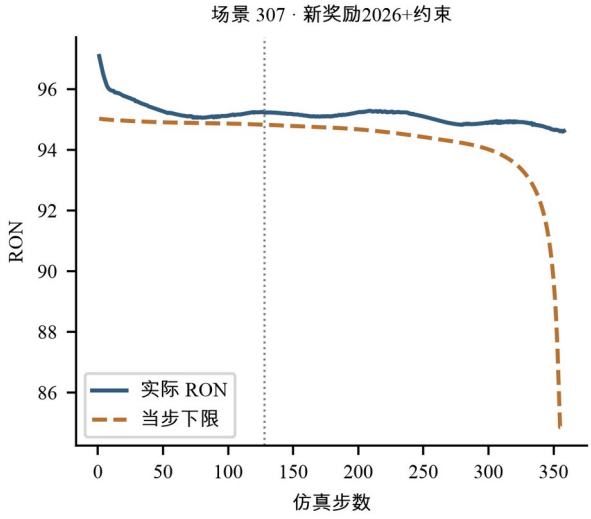


图 9　Web 回放字段对应的质量轨迹

资料来源：据 experiment.json 重绘；场景 307、quality-2026-guard，第 128 步定位线；非页面截图。

5.2 DeepSeek 解释与候选动作

DeepSeek 接口承担结果解释和候选建议两种任务。解释接口发送选定记录摘要，由后端接收增量响应并用 SSE 转发，浏览器逐段显示；官方 Chat Completions 文档提供 stream 模式及增量返回定义[7]。生成中断时页面保留已收文字并提示不完整，不静默替换为完整结论。

候选接口提供当前组分性质、配方、质量约束和 Actor 建议，要求返回最多 5 个三维动作。候选响应为完整 JSON 而非流式执行。程序检查类型、有限数值和[-1,1]范围，明确按 0.1 个百分点取整，再依据本地混合公式计算质量和配方边界未满足预测约束的候选不会被选中；模型解释也不能替代本地数值计算。

一次保存的真实调用返回 5 个候选，其中第 1 和第 4 个预测可行，分别给出 RON 约 94.7598 和 94.7636；程序在可行候选中选择更接近 Actor 建议的第 4 个。图 10 显示预测结果。本次没有执行该候选，也没有比较其完整回合表现。此功能扩展了策略候选来源，但没有更新 RL 网络参数，不能等同于新的强化学习训练。

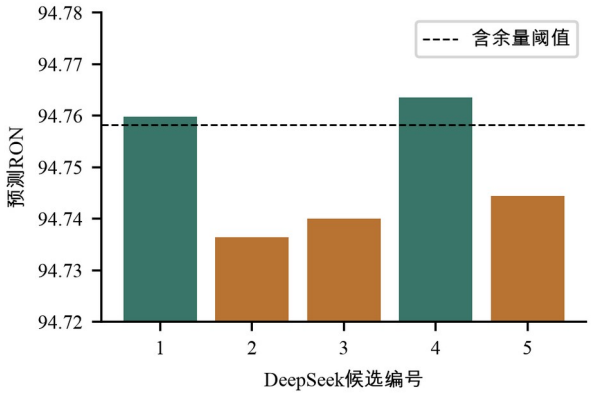


图 10　DeepSeek 候选动作的单步预测校验

资料来源：candidate-live-example.json；阈值 94.7582，绿色为预测可行，候选尚未执行。

线上展示采用独立后端服务与 HTTPS 入口。学习手册和回放公开访问，付费模型接口通过登录认证与请求限流保护，API 密钥只在后端读取。代码附件不包含密钥、登录密码或服务器私钥。该部署服务于课程展示，不接入工业控制系统。

6 实验设置与结果分析

6.1 场景划分与评价指标

训练场景为 101—112；辅助检查场景为 201、203、207、209，在全部训练完成后进行检查，不参与本章最终统计，也不用于选取检查点最终评价采用 301、307、311、313、317、331、337、347 八个场景。旧场景 3、11、29、41、53、67、79、97 仅用于历史实验或方法说明。

最终对照组包括原始 RL、原始 RL 加约束、原奖励微调、新奖励微调及新奖励微调加约束。前两组各 8 回合，后三组各汇总 3 个训练种子、24 回合。辅助检查和最终评价合计 132 回合，其中最终评价 88 回合。由于所有回合均为 358 步，汇总平均成本可按回合等权计算。

$$v = N_{bad} / N_{total} \times 100\ \% \quad (9)$$

式中 v 为调合头违规率，Nbad 为违规步数，Ntotal 为总评价步数。某一步只要任一调合头质量分量超限超过 10^{-4} 即记为一个违规步，多项同时超限不重复累加。配方违规独立统计，最终罐内质量用固定限制判断。不同组回合数不同，因此主要比较违规率，不能直接用 45 与 282 的绝对步数判断优劣。

6.2 训练曲线与学习过程

图 11 展示两种奖励设置的采样训练违规率，实线为 3 次训练均值，阴影为最小值至最大值范围，不是置信区间。采样时包含探索噪声，且训练场景在 12 个编号间循环，因此曲线波动不能直接解释为确定性策略性能退化或收敛。

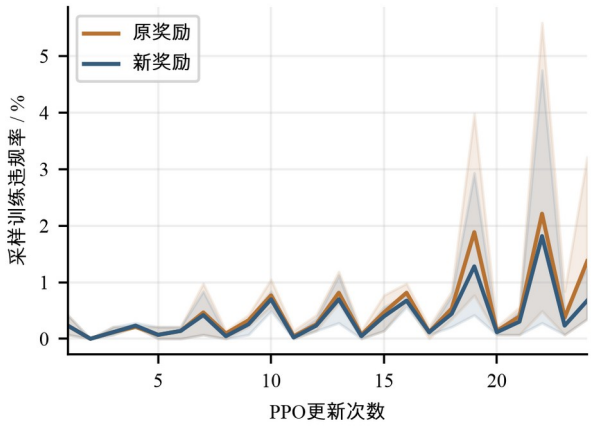


图 11　两种奖励设置的采样训练违规率

资料来源：6 份 train-*.json 的 24 次更新记录；每次更新 4 回合，阴影为 3 个种子的范围。

两类奖励的数值定义不同，不能把其累计奖励绝对大小直接作横向优劣判断。本研究以最终模型在同一评价场景下的成本和违规率作为主要指标。权重变化证明参数确实更新，但不能单独证明学到了更优策略。

6.3 五组策略总体比较

图 12 与表 3 给出完整评价结果。原始 RL 违规 45 步，占 2864 步的 1.5712 %；原奖励微调 331 步，占 8592 步的 3.8524 %；新奖励微调 282 步，占 8592 步的 3.2821 %。新奖励相对原奖励微调减少 49 个违规步，违规率下降约 0.5703 个百分点，相对下降约 14.8 %，但仍高于原始 RL。

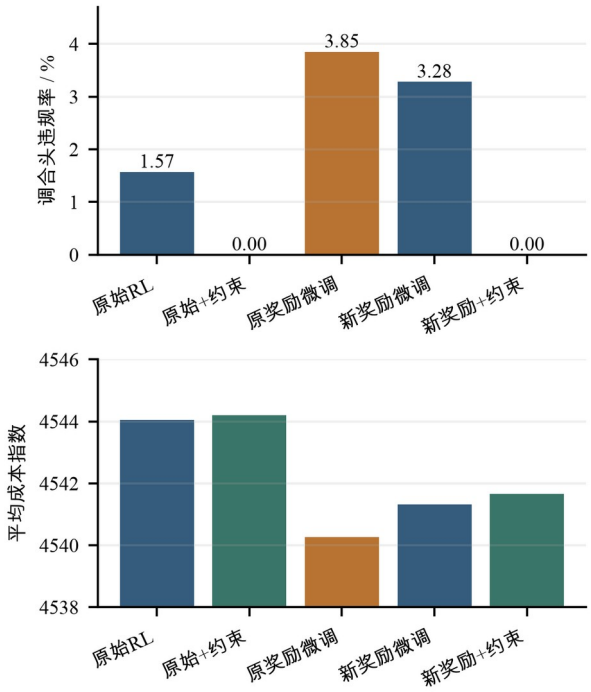


图 12　五组策略的违规率与平均成本

资料来源：8 个最终评价场景；原始两组各 8 回合，微调三组各 24 回合。成本子图纵轴截取 4538—4546。

表 3　成本与调合头质量评价

策略	成本均值	违规率 / %	违规/总步数
原始 RL	4544.0480	1.57	45/2864
原始+约束	4544.2070	0.00	0/2864
原奖励微调	4540.2716	3.85	331/8592
新奖励微调	4541.3136	3.28	282/8592
新奖励+约束	4541.6562	0.00	0/8592

增加约束检查后，原始 RL 和新奖励微调两组的调合头违规均为零。原始 RL 平均成本由 4544.0480 增至 4544.2070，增加 0.1590；新奖励微调由 4541.3136 增至 4541.6562，增加 0.3425。这体现了当前场景中检查的质量收益与小幅成本代价。

新奖励微调加约束的成本比原始 RL 加约束低 2.5508 个指数单位，约 0.056 %。该差异很小，且仅有 3 个训练种子和 8 个评价场景，本文只作样本描述，不宣称统计显著或工业经济收益。原始 RL 加约束也实现零违规，因此不能把零违规归因于新奖励，更不能证明微调是达到零违规的必要条件。

6.4 配方边界与最终储罐质量

表 4　配方与最终罐内评价

策略	回合数	配方违规步数	最终罐内达标
原始 RL	8	12	8/8
原始+约束	8	0	8/8
原奖励微调	24	11	24/24
新奖励微调	24	5	24/24
新奖励+约束	24	0	24/24

表 4 显示五组最终罐内均达标，但未加约束的原始 RL、原奖励微调和新奖励微调仍分别出现 12、11 和 5 个配方违规步。最终质量受整个储罐混合过程影响，不能反推每一步调合头达标。配方边界、过程质量和最终产品质量应分别评价，避免只保留一个看似良好的终点指标。

6.5 随机种子与代表轨迹

图 13 给出三次训练的配对结果。2026、2027、2028 种子下，原奖励微调分别违规 97、100、134 步，新奖励分别为 87、85、110 步。三个种子中均出现同方向的改善，但改善幅度不同说明单次训练不足以代表总体表现；三个种子也不足以刻画广泛随机性。

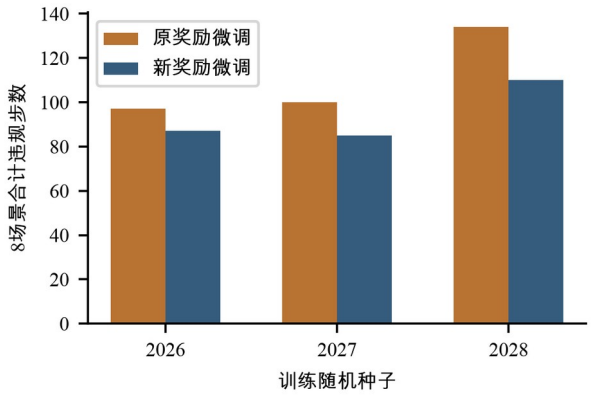


图 13　三次训练的调合头违规步数

资料来源：每根柱为对应权重在 8 个评价场景、2864 步中的违规数。

图 14 以场景 307 为例，同时绘制成本、RON 相对当步下限的余量及重整油比例。每个策略使用自己的当步限制，因为此前决策已改变其储罐状态。余量小于零才表示 RON 下限违规，不能用同一条策略的动态下限去判定另一策略的曲线。该图用于解释决策与质量之间的联系，不代表所有场景。

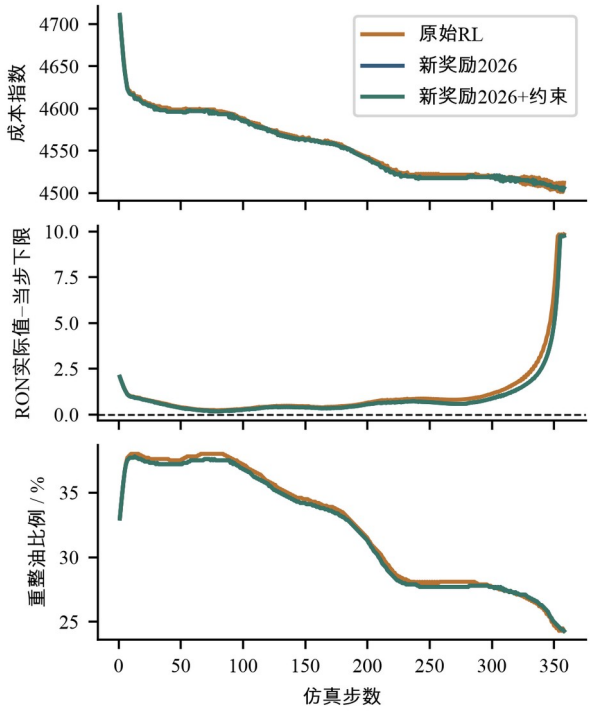


图 14　场景 307 的成本、RON 余量与配方轨迹

资料来源：原始 RL、新奖励 2026 及其加约束策略；均 358 步，来自本轮 experiment.json。

课程课件中的曲线展示了特定策略在其案例条件下的优化表现[1]。本文采用的训练预算、奖励和评价场景不同，不能把课件中的优势直接移植

为本次结论。对于原始策略为何优于两类微调，当前记录只能确认结果，无法单独确定训练预算学习率、探索噪声或奖励权重中的哪个因素起主导作用。

7 讨论与适用范围

7.1 方法收益与不足

本次实验首先表明，增加惩罚项并不自动带来更好的策略。新奖励在“继续训练”这一控制条件下改善了违规率，但继续训练本身可能改变已有策略，现有预算不足以证明总体提升。应把原始模型作为始终保留的基准，而不是只比较两个微调版本。

其次，动作检查的效果依赖当前观测、候选集合和课程预测公式。0.1 RON 余量在当前新场景表现良好，但没有覆盖所有可能扰动；候选网格也不是连续可行域的完整搜索。若传感误差、混合模型偏差或原料切换更强，检查仍可能失效样本零违规只是观测结果，不构成安全证明。

7.2 实验设计的边界

原始权重完整训练来源未知，因此无法排除其历史经验与评价条件存在关联。训练、辅助检查和最终评价编号不重叠，有助于约束本次流程但编号不同不等价于真实工业分布独立。本次未进行显著性检验，也未对不同奖励系数和训练预算做系统搜索。

为控制比较条件，所有策略共享课程动力学和同场景外生扰动。这样适合分析当前实现，但不足以衡量模型失配下的稳健性。真实装置还涉及分析仪延迟、流量和库存限制、泵阀能力与操作规程，这些没有在本实验中完整建模。

DeepSeek 结果只完成单步候选可行性检查。候选提出得合理，不表示连续采用后能降低成本或违规；语言解释也可能产生原因推测，应回到轨迹和公式核对。未来若将其引入实际策略评价必须保存相同场景下完整回合结果，并区分本地规则、Actor 与语言模型各自的作用。

8 结论与展望

本文围绕课程汽油调合任务，完成状态与动作解释、质量导向奖励、6 次 PPO 微调、执行前动作检查及 Web 分析。新奖励微调的违规率由原奖励微调的 3.85 %降至 3.28 %，但仍高于原始策略的 1.57 %。因此，本次结果支持“新奖励相对继续训练对照有所改善”，不支持“新策略全面优于原始策略”。

原始策略加约束和新奖励微调加约束在当前评价样本中均无调合头违规，说明执行前检查在本实验条件下具有实际作用。两组平均成本差异较小，应在扩大场景和训练种子后重新评估。Web 使实验结果能够逐步追溯；DeepSeek 提供解释与候选，尚不能证明完整过程收益。

后续优先安排三类实验：固定评价集前另设开发集，比较不同奖励系数与训练预算；增加更强扰动、测量误差和模型偏差，记录检查失败与回退；将候选动作纳入完整仿真回合，并与原始 RL 加约束进行公平对照。只有在这些证据充分后，才讨论进一步的控制应用。

9 体会、收获和建议

本课程实践使软件系统中的“输入—计算—输出”与石化过程中的物料性质、混合机理和质量约束建立了具体联系。对 Web 开发背景而言，主要收获是认识到界面显示的每一个数值都必须有清楚的工艺含义、时间位置和统计口径。调合头与储罐、预测与实际、奖励与约束的混淆，都会直接影响结论。

另一项收获是从展示算法转向评价算法。模型参数发生变化、训练奖励提高、页面功能完整都不等于质量表现改善。保留原始基准、相同场景比较、报告失败结果以及追溯逐步记录，比单独展示一条漂亮曲线更有说服力。

建议后续课程实践同时提供变量单位说明、初始权重训练背景及统一评价协议，使不同小组能够在相同场景和预算下比较。Web 展示可作为理解与复核工具，但应始终围绕工艺问题组织，而不是以接口数量或页面复杂度代替智能学习内容。

参考文献

[1] 何仁初. 强化学习优化实践[Z]. 《石化过程智能制造》课程课件及配套代码, 2026.

[2] CHÈBRE M, CREFF Y, PETIT N. Feedback control and optimization for the production of commercial fuels by blending[J]. Journal of Process Control, 2010, 20(4): 441-451.

[3] HUANG M, HE R, DAI X, PENG X, DU W, QIAN F. Reinforcement Learning Based Gasoline Blending Optimization: Achieving More Efficient Nonlinear Online Blending of Fuels. 2023[2026-11-15]. https://arxiv.org/abs/2309.02929.

[4] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction[M]. 2nd ed. Cambridge, MA: MIT Press, 2018.

[5] SCHULMAN J, WOLSKI F, DHARIWAL P, RADFORD A, KLIMOV O. Proximal Policy Optimization Algorithms. 2017[2026-11-15]. https://arxiv.org/abs/1707.06347.

[6] SCHULMAN J, MORITZ P, LEVINE S, JORDAN M I, ABBEEL P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. 2015[2026-11-15]. https://arxiv.org/abs/1506.02438v1.

[7] DeepSeek. Chat Completions API. [出版年不详][2026-11-15]. https://api-docs.deepseek.com/api/create-chat-completion/.

附录 程序与复现说明

1 主要文件与调用顺序

teacher/AgentRun.py 保存课程环境与网络，teacher/actor.pth 为初始 Actor。src/retrain.py 实现奖励、采样、GAE 与 PPO 更新；src/action_guard.py 实现预测和候选筛选；src/experiment.py 记录回合轨迹；src/verify_learning.py 负责统计和部分权重重放。

src/assistant_server.py 提供记录查询和流式解释，src/candidate_advisor.py 生成并校验候选，web/learning.js 展示新训练曲线与轨迹。models/learning 保存 6 份 Actor 权重；results/learning 保存配置、训练日志、完整轨迹、汇总和候选示例。读取学习手册或 Web 不需要重新训练。

2 可复现操作与记录

已有结果可通过 python3 src/assistant_server.py --port 8765 启动本地服务，在浏览器打开 127.0.0.1:8765。若要重新训练，应先备份 models/learning、results/learning 及 web/learning-data.js，再使用项目记录的依赖环境执行 src/retrain.py。新训练会覆盖结果，不能在未备份时把旧图表与新权重混用。

训练配置记录的 Torch 版本为 2.15.0。硬件和依赖变化可能造成数值差异，完整复现应同时核对环境版本、初始权重、随机种子和保存日志。报告附带的“生成配图与复核.py”仅读取既有记录，重算统计并生成图表，不调用模型 API，也不重新训练。

3 奖励计算的核心代码

```
return float(original
  - 2*(max(violations)>TOL)
  - 2*sum(violations/SCALE)
  - .01*sum(abs(delta))
  - 2*recipe)
```

此片段来自 learning_reward，recipe 在函数前部按最低组分比例和累计偏移判定，SCALE 为正文给出的归一化尺度。其含义是修改评分规则，真正的策略学习还依赖采样、优势估计和优化器更新。完整源码随项目保存，不在正文重复粘贴。