结论先行

> 可行。通过“发现并放大高熵分岔点”，提前数据构建可以提高RL后训练的**样本效率**、**稳定性**与**最终效果**。

> 基本思路，按RL后训练预期的方式构造批量数据进行SFT，加快收敛速度，稳定训练效果。

方案的简要描述

A. 高熵实例挖掘（Uncertainty Mining）

具体做法：

- 让现有模型把候选数据“过一遍”，记录每个位置的token 熵（拿不准的程度）。
- 选出高熵位置多、分布分散的样本入库；低熵、机械续写的样本就少用或丢掉。
- 强化学习时，要么只在高熵位置更新（挑批内熵最高的前 10–30% 位置），要么在高熵位置给优势加权（高熵处“多踩油门”）。

特点：成本最低、与 PPO/GRPO 无缝兼容、训练的收敛更快

B. 边界对比数据（Counterfactual & n-best）

具体做法：

- 对同一问题做最小修改（换一个数字/条件/别名），让答案刚好翻转，形成一对“几乎一样但结论相反”的样本。
- 采集每个问题的 n-best 多候选（几条对、几条错），让模型在同一前缀下见到不同走向。
- 这些数据先做轻量 SFT（定格式），再进 GRPO/PPO 或 DPO，训练时重点关注差别发生的关节点。

特点：
- 这类数据把模型直接丢到决策边界上：小改动→大后果，梯度信息最密。

C. 锚点前缀库 + 局部展开（FR3E 思路）

具体做法：

- 在一条生成里挑出 **Top-K** 个高熵位置当“锚点”，把 **提示 + 到此为止的前缀** 存成“前缀库”。
- 训练时从某个锚点**读档**，从这里继续采样 **M 次**到终局，计算“从这个分岔口继续走能走对的平均概率”：

$$V(S_j) = \frac{1}{M} \sum_{m=1}^{M} r_{j,m}$$

- 用“**实际结果 – 平均结果**”当优势更新策略：

$$A_{j,m} = r_{j,m} - V(S_j)$$

（进步区域**降一点力度**，受阻区域**加一点力度**；其它训练配方不变）

特点：把稀疏终局奖“压回中间”，显著改善**信用分配**与**训练稳定性**。