由于是 hard mode，human expert 在模型生成答案的时候就有即时的监督，故具体评价略。

注意到 LLM 犯的错误可以总结为以下两条：

1、没有正确理解活塞回抽过程（这可能是题目描述不清导致的；事实上其建模已经相当正确了，这可以说是一个 minor 的错误）

2、没有推导出等温压缩的 log 公式（提示修正后，第二阶段仍理解为恒压。这不失为一种表现尚可的近似；事实上，原题（CPhO36th 预赛浙江卷 29 题）也采用了同样的近似方式）

最后验证程序能得到几乎正确的结果（bias 是因为采取了近似）；但 LLM 给出的结果仍然有偏差，说明和 standard mode 测试一样地，LLM 并没有真正地调用 agent 来进行运算。这个结果也说明，建模没有问题、甚至可以不需要 human expert 的辅助，只要允许接入 agent 则 LLM 给出的结果会立马得到大幅改善。