好的，我们把目标压到“能落地的小型试验”：基座用 \*\*Qwen2.5-0.5B-Instruct\*\*，先做\*\*监督微调（SFT）\*\*得到稳定的四项能力原型，然后在这个 SFT 模型上做\*\*强化学习（RL）\*\*增强决策（尤其是“回退/重跑/澄清”的策略性）。下面是\*\*不含实现代码\*\*但可直接照着推进的方案。  
  
# 1. 明确四项能力与产物格式  
我们把能力拆成 4 个“可评测子任务”，统一用函数调用/JSON 作为模型输出，便于评测与RL。  
  
1) 意图识别（Intent）  
- 输入：用户/医生自然语言一句话  
- 输出（JSON）：`{type, node?, slot?, topic?}`   
- 评价：意图准确率、槽位F1、非法JSON率  
  
2) 状态管理/流程编排（Planner）  
- 输入：`patient\_state`（关键字段子集）+ `executed\_nodes` + 最近对话意图  
- 输出（JSON 动作）：`{action: EXECUTE|ROLLBACK|CLARIFY|GENERATE\_REPORT, node?, slot?}`  
- 评价：任务完成率、平均步骤、无效回退率、强约束满足率  
  
3) 报告生成（Reporter）  
- 输入：最终结构化 `state`  
- 输出：`{pro: "...", patient: "..."}` 两版文本  
- 评价：字段一致性（自动规则）、专家打分（语言质量/可读性）、事实错误率  
  
4) 患者沟通（Communicator）  
- 输入：同上 + 额外关注点（topic）  
- 输出：一段通俗话术（或JSON字段 + 文本）  
- 评价：可读性、风险表达合规、医生打分  
  
> 统一 ChatML 指令风格 + “只输出 JSON”约束，后续 SFT 与 RL 都用同一接口。  
  
# 2. 数据设计与准备（SFT 用）  
最小可行数据面，先拿你已有两例 JSON 做“模板化扩增”，迅速得到几十到上百条训练样本。  
  
- 意图识别集（≥200条）：把真实口令/口语化变体写成“用户→标准JSON”。覆盖：继续、出报告、回退到各节点、澄清若干slot、患者沟通主题。  
- 编排轨迹集（≥100条）：由你现有流程回放生成（观察→动作）。每条包含：`(state摘要, executed\_nodes, last\_intent) -> next\_action\_json`。人工少量校对“黄金轨迹”。  
- 报告对齐集（≥50条）：输入“结构化关键字段”→输出“专业版/患者版”两段参考文本（从历史报告或模板改写）。  
- 沟通话术集（≥50条）：按不同 TI-RADS、是否弥漫性病变、是否淋巴结可疑等，写多种常见问答/说明。  
  
数据格式统一为对话样式（system+user+assistant），assistant 严格给出 \*\*单段 JSON\*\* 或指定字段，便于训练时自动校验。  
  
# 3. 训练策略（SFT 阶段）  
目标：让 0.5B 模型先“听得懂并按格式输出”，再考虑策略。  
  
- 方式：LoRA/QLoRA SFT（显存/内存友好）。建议“多任务混合训练”（四任务样本混合打乱）。  
- 采样与损失：   
 - 意图/编排：严格 JSON 输出，加入“无效输出惩罚”样本（few negative）。   
 - 报告/沟通：模板块+少量生成；限定段落与长度，避免胡写。  
- 训练配比（起步）：Intent:Planner:Report:Comm ≈ 4:3:2:1  
- 校验：每个 batch 在线做 JSON 解析；无效样本丢弃或重权重。  
- 选择保存：看验证集四任务指标综合最优的 checkpoint。  
  
# 4. 强化学习（RL）阶段设计  
目标：提升“流程决策”的有效性与稳健性（尤其回退/澄清/少走弯路）。对报告与沟通只做轻度奖励约束（事实一致、安全）。  
  
## 4.1 MDP 抽象  
- \*\*状态 s\*\*：`(state摘要, executed\_nodes, outstanding\_requirements, last\_intent, history\_k)`   
- \*\*动作 a\*\*：标准动作 JSON（与 Planner 输出一致）   
- \*\*转移\*\*：你已有“执行器/模拟器”负责：执行动作→写回 state/exec\_nodes → 产生下一观测   
- \*\*终止\*\*：报告成功产出并通过规则校验，或达到步数上限  
  
## 4.2 奖励设计（组合）  
- 完成奖励：+1（生成报告且强规则全过）   
- 步骤代价：每步 −0.02（鼓励短路径）   
- 回退有效：若回退后最终通过校验 +0.2；无效回退 −0.1   
- 澄清命中：澄清导致关键缺失项被补齐 +0.1   
- 违规/无效JSON：−0.5   
- 报告一致性：结构化字段与输出文本比对通过 +0.2（超出可不做）   
> 可先用上述“代理奖励”（proxy rewards），后续再接医生小样本打分做“人类偏好奖励”（DPO/GRPO 风格）。  
  
## 4.3 算法与流程  
- \*\*离线预热\*\*：用 SFT 轨迹生成回放，做离线 REINFORCE/GRPO（更稳）；   
- \*\*在线小规模\*\*：环境=你的模拟执行器；算法可用 PPO/GRPO 小批量并行采样；   
- \*\*价值对齐\*\*：加入规则约束（非法动作直接剪枝或给大负奖励）。   
- \*\*稳定技巧\*\*：   
 - 动作空间统一 JSON 模板（few-shot in-context）；   
 - 温度低采样，Top-p 0.9；   
 - KL 约束回到 SFT 策略，避免发散（PPO/GRPO自带）。  
  
# 5. 评测方案与门槛  
- \*\*意图\*\*：acc≥95%，JSON 合法率≥99%（小样本场景）   
- \*\*编排\*\*：任务完成率≥90%，平均步数 ≤ 基线−10%   
- \*\*报告\*\*：字段一致性≥95%，专家打分≥3.5/5，无事实型错误   
- \*\*沟通\*\*：医生打分≥4/5，敏感用语零违例   
- \*\*鲁棒性\*\*：10个“回退→重跑”场景通过率≥90%  
  
# 6. 数据闭环与在环校正  
- 线上/模拟交互全量打日志：`(观测, 动作JSON, 执行结果, 奖励, 失败原因)`   
- 每周：   
 - 抽 20 条失败样本做人审 + 规则补丁   
 - 追加到 SFT 数据（增量混合训练）   
 - RL 用新回放更新离线池，短轮再训练  
  
# 7. 资源与配置建议  
- 0.5B + LoRA：单张 8GB 显卡或 CPU 也能跑（训练建议 12GB+）。   
- SFT 数据规模：\*\*起步 400–800 样本\*\*即可看到稳定改进（意图/编排为主）。   
- RL 采样：先离线 5–10k 轨迹步；在线每轮 2–5k 步足够观察到“少走弯路”的提升。  
  
# 8. 里程碑（两到三周）  
1) \*\*W1-SFT\*\*：清洗/构造四任务数据 → 多任务LoRA SFT → 验证集通过（JSON合法率≥99%）。   
2) \*\*W2-RL(离线)\*\*：把 SFT 轨迹做离线 RL（GRPO/REINFORCE）→ 决策步数下降、完成率上升。   
3) \*\*W3-RL(在线)+评测\*\*：接模拟执行器做小规模在线 PPO/GRPO → 回退/澄清策略更稳 → 出综合报告。  
  
# 9. 落地清单（你现在就可以开始准备）  
- [ ] 确认节点命名与“回退清空字段表”（哪一节点会写哪些字段）   
- [ ] 汇总 5 类典型意图的口语化表述（≥200条）   
- [ ] 从两例 JSON 自动生成 20–40 条编排轨迹（含回退/澄清）   
- [ ] 报告模板：专业版/患者版的段落与必含字段清单   
- [ ] 规则校验器：字段一致性检查（报告对齐）+ JSON 合法性检查   
- [ ] 模拟执行器：根据动作 JSON 修改 `state` 与 `executed\_nodes`（可不接真实图像）  
  
---  
  
这套方案的核心是：\*\*先用SFT把“听懂+按格式说”打稳，再用RL优化“何时回退/何时澄清/怎样最少步完成”\*\*。你只要把样本与规则准备好，就能快速验证这条路线是否可行。需要的话，我可以把“样本模板与评测清单”按你的字段名细化成可直接填充的空表格。