C) Giải thích từng bước kiểu "kể chuyện"

Bước	Làm gì?	Vì sao?	Output tạo ra
1	**Generate** một câu trả lời thô `y_init` từ prompt `x`	Lấy "phản xạ tự nhiên" hiện tại của mô hình	Câu trả lời ban đầu (có thể ngắn, chưa chuẩn)
2	**Revise**: nếu câu trả lời tệ (ngắn, lặp, vi phạm nguyên tắc) thì tái sinh với prompt nhắc nhở ("Please provide a helpful, honest, respectful answer…")	Tự sửa để có phiên bản "chất lượng cao" hơn dùng làm "mẫu tốt"	`y` (phiên bản cải thiện)
3	**CriticScore**: critic áp dụng 4 nguyên tắc → cho điểm từng nguyên tắc rồi lấy trung bình	Tạo nhãn "chuẩn mực" s (0–1) mà không cần người gán	`s` (reward thật mục tiêu)
4	**LanguageModelLoss**: tính cross-entropy trên chuỗi `x + y` (hoặc chỉ phần `y` nếu mask prompt)	Buộc mô hình học "cấu trúc / nội dung" của câu trả lời tốt	`L_LM`
5	**RewardModelPredict**: mô hình reward đọc `x + y` → dự đoán điểm `r_hat`	Học bộ chấm điểm tự động bắt chước critic	`r_hat`
6	**Reward Loss**: MSE = ½(r_hat − s)²	Điều chỉnh reward model để dự đoán sát s	`L_reward`
7	**Combine**: `L_total = L_LM + 0.1 * L_reward`	Vừa cải thiện chất lượng ngôn ngữ, vừa căn chỉnh theo hiến pháp	`L_total`
8	**Backprop & Update**	Cập nhật tham số cả generator và reward model	Mô hình tốt hơn ở vòng sau

(D) Ví dụ số mini (siêu cô đọng)

- Critic cho câu trả lời đạt **s = 0.80**.
- Reward model đoán **r_hat = 0.55** → `L_reward = 0.5*(0.55 - 0.80)^2 = 0.03125` .
- LM loss (cross-entropy mean) **L_LM = 1.72**.
- Total: `L_total = 1.72 + 0.1*0.03125 ≈ 1.7231` .
  → Gradient chủ yếu đến từ LM (1.72), nhưng reward loss "kéo" r_hat tiến về 0.80.

(E) Phiên bản “người mới vào team đọc 30s hiểu liền”

> “Cho mỗi prompt, để mô hình tự trả lời, nếu trả lời dở thì bảo nó viết lại cho chuẩn theo 4 nguyên tắc. Chúng ta chấm điểm câu đã sửa (0–1) bằng critic. Dùng câu đó làm dữ liệu để: (1) dạy mô hình viết giống câu tốt (LM loss), (2) dạy reward model dự đoán điểm giống critic (MSE). Tổng loss = LM + 0.1*MSE. Update xong → mô hình ngày càng vừa ‘viết hay’ vừa ‘hợp hiến pháp’.”

(F) Một câu slogan (để nhớ)

> **“Sinh — Sửa — Chấm — Học theo câu tốt — Học cách chấm — Ghép loss — Update.”**