BÁO CÁO FINAL PROJECT ASSIGNMENT

Constitutional AI: Xây Dựng Hệ Thống AI An Toàn Bằng Phương Pháp RLAIF.

Người thực hiện: Bùi Tân Nhật - Qe17017Qe170171

Lớp: AI17D

Giáo viên hướng dẫn: Nguyễn An Khương

Môn học: Học Tăng cường (REL301m)

Học kỳ: Summer 2025

Ngày: 21-22 tháng 7 năm 2025

1. Bối Cảnh và Vấn Đề

Trong những năm gần đây, các mô hình ngôn ngữ lớn (LLM) đã đạt được những tiến bộ vượt bậc, cho phép chúng tạo ra văn bản mạch lạc và hỗ trợ con người trong nhiều nhiệm vụ. Tuy nhiên, đi kèm với sự **mạnh mẽ** đó là nguy cơ tiềm ẩn: mô hình có thể sinh ra nội dung **độc hại, sai lệch hoặc không an toàn** nếu không được kiểm soát tốt. Để giải quyết vấn đề này, cộng đồng nghiên cứu đã phát triển phương pháp **Học tăng cường từ phản hồi của con người (RLHF)**, trong đó con người tham gia đánh giá và xếp hạng các phản hồi của mô hình, từ đó huấn luyện một mô hình ưu tiên (preference model) để định hướng hành vi của mô hình ngôn ngữ sao cho phù hợp với **ý muốn và giá trị của con người**. RLHF đã được sử dụng thành công (ví dụ trong ChatGPT) nhằm tăng tính **an toàn và hữu ích** của mô hình, nhưng nó cũng bộc lộ những hạn chế: quá trình thu thập phản hồi thủ công tốn kém, mất thời gian và khó mở rộng quy mô.

Trước thách thức đó, nhóm nghiên cứu tại Anthropic đã đề xuất một cách tiếp cận mới mang tên **“Hiến pháp AI” (Constitutional AI)** kết hợp với **Học tăng cường từ phản hồi AI (Reinforcement Learning from AI Feedback – RLAIF)**. Thay vì dựa hoàn toàn vào nhãn con người, phương pháp này chỉ sử dụng một tập hợp các **nguyên tắc hoặc quy tắc do con người soạn sẵn** (được coi như “hiến pháp” định hướng hành vi cho AI), và tận dụng chính **mô hình AI** để phê bình và đánh giá đầu ra của mô hình đích. Theo cách đó, mô hình đích có thể **tự cải thiện** dựa trên phản hồi từ một mô hình AI khác (hay từ các quy tắc đã nêu) mà không cần con người can thiệp trực tiếp ở mọi mẫu đầu ra. Cách tiếp cận RLAIF hứa hẹn **hiệu quả hơn** về mặt chi phí và thời gian, đồng thời giúp mở rộng huấn luyện ra phạm vi lớn hơn so với RLHF truyền thống. Hơn nữa, các nghiên cứu ban đầu cho thấy RLAIF có thể đạt hiệu quả tương đương RLHF về tính hữu ích, trong khi cải thiện hơn nữa về tính an toàn (**“vô hại”**) của mô hình.

Bài báo cáo này trình bày chi tiết một dự án ứng dụng phương pháp **Constitutional AI** và **RLAIF** trên mô hình ngôn ngữ GPT-2. Cụ thể, chúng tôi mô tả các **khái niệm và nguyên tắc** cốt lõi, giải thích **thuật toán huấn luyện** kết hợp giữa học có giám sát và học tăng cường với phản hồi AI, trình bày **quy trình triển khai** theo từng bước của dự án, và thảo luận **kết quả thực nghiệm** đạt được. Mục tiêu là huấn luyện một mô hình GPT-2 biết tuân thủ bốn nguyên tắc chính – **không gây hại, hữu ích, trung thực và tôn trọng** – nhờ đó cung cấp các câu trả lời an toàn và hữu ích hơn cho người dùng, mà chỉ dựa trên hướng dẫn dưới dạng quy tắc, không cần đến nhãn chất lượng chi tiết từ con người.

2. Nguyên tắc “Hiến pháp AI” và nền tảng thuật toán

A.Hiến pháp AI và các nguyên tắc định hướng

Constitutional AI (Hiến pháp AI) là phương pháp huấn luyện mô hình với sự giám sát tối thiểu của con người, trong đó vai trò của con người được thu gọn thành việc soạn thảo một **bộ nguyên tắc chỉ đạo** (hiểu như bản “hiến pháp” cho AI). Bộ nguyên tắc này định nghĩa các tiêu chí mà mô hình cần tuân thủ khi trả lời, thay cho việc con người can thiệp chi tiết vào từng trường hợp. Trong dự án này, chúng tôi sử dụng bốn nguyên tắc chính, tương ứng với bốn khía cạnh hành vi mong muốn của mô hình:

- **Harmless – Không gây hại**: Mô hình không được cung cấp nội dung **nguy hiểm, độc hại hoặc bất hợp pháp**. Nó phải tránh hướng dẫn những hành vi có thể gây hại cho người dùng hoặc xã hội.

- **Helpful – Hữu ích**: Mô hình cần đưa ra câu trả lời **có ích và thông tin**, đáp ứng đúng nhu cầu câu hỏi của người dùng một cách hữu dụng.

- **Honest – Trung thực**: Mô hình phải **thành thật**, không bịa đặt thông tin hoặc đưa ra các khẳng định sai sự thật. Nếu không chắc chắn, mô hình nên thể hiện sự không chắc chắn thay vì khẳng định tuyệt đối.

- **Respectful – Tôn trọng (Lịch sự)**: Mô hình phải **tôn trọng người dùng và các chủ thể khác**, tránh ngôn ngữ xúc phạm hoặc thiếu lịch sự, và duy trì thái độ đúng mực.

Mỗi nguyên tắc trên đi kèm với một phần mô tả và một hướng dẫn phê bình (critique prompt) để đánh giá phản hồi theo nguyên tắc đó. Chẳng hạn, nguyên tắc *Harmless* có hướng dẫn phê bình: *“Phản hồi này có chứa nội dung nguy hại, nguy hiểm hoặc phi pháp không? Nếu có, hãy giải thích lý do.”* – điều này giúp mô hình phản hồi phụ (critic) tập trung phát hiện yếu tố vi phạm tính **an toàn** trong câu trả lời.

B.Học tăng cường từ phản hồi AI (RLAIF)

Học tăng cường từ phản hồi AI (RLAIF) là một biến thể của RLHF, trong đó quá trình đánh giá và cho điểm phản hồi của mô hình không do con người thực hiện mà do một **mô hình AI thứ cấp** đảm nhiệm. Ý tưởng cốt lõi là sử dụng một mô hình (hoặc hệ thống) AI làm *nhà phê bình* (AI **critic**) để phân tích câu trả lời của mô hình chính, dựa trên các nguyên tắc đã định sẵn, rồi **cho điểm thưởng/phạt**. Điểm số này sau đó được dùng làm **tín hiệu phản hồi (reward signal)** để huấn luyện mô hình chính bằng thuật toán RL. Như vậy, AI đang được huấn luyện sẽ học cách **tối đa hóa điểm thưởng** vốn phản ánh **mức độ tuân thủ các nguyên tắc đạo đức/hữu ích**, thay vì tối đa hóa một mục tiêu mơ hồ nào đó.

Ưu điểm của RLAIF là **giảm sự phụ thuộc** vào nhãn con người ở quy mô lớn: con người chỉ tham gia gián tiếp qua việc soạn bộ quy tắc và có thể dùng một mô hình AI đã được định

hướng bởi bộ quy tắc đó để đánh giá hàng loạt phản hồi một cách tự động. Dĩ nhiên, chất lượng của mô hình phê bình AI rất quan trọng – nó cần tương đối **nâng cao** hơn mô hình chính và phải được thiết kế/training sao cho hiểu và áp dụng chính xác các nguyên tắc. Nếu mô hình phê bình mang định kiến hoặc sai lệch, những lỗi đó có thể truyền sang mô hình chính. Do đó, mặc dù RLAIF mở ra hướng huấn luyện **nhanh hơn và rẻ hơn**, nó thường được kết hợp cẩn thận với sự giám sát chọn lọc của con người để đảm bảo hệ thống không đi chệch hướng.

Trong phương pháp **Constitutional AI** được đề xuất bởi Anthropic, RLAIF được sử dụng ở pha thứ hai của quá trình huấn luyện mô hình trợ lý AI an toàn. Cụ thể, sau khi mô hình đã được tinh chỉnh qua pha đầu (học có giám sát trên các phản hồi đã tự được cải thiện), người ta sẽ tạo ra một **mô hình ưu tiên (preference model)** bằng cách cho mô hình AI đánh giá cặp đáp án và chọn ra đáp án tốt hơn (theo các nguyên tắc hiến pháp). Mô hình ưu tiên này về bản chất chính là **mô hình phần thưởng** – nó nhận đầu vào là một câu trả lời và đưa ra **điểm số phản ánh chất lượng**. Thay vì huấn luyện mô hình phần thưởng bằng dữ liệu so sánh do con người gán nhãn (như trong RLHF), ở đây dữ liệu được tạo ra bởi **AI feedback** (các đánh giá của AI dựa trên hiến pháp). Sau cùng, ta tiến hành huấn luyện tăng cường (ví dụ bằng thuật toán PPO) để cập nhật mô hình chính, trong đó **hàm thưởng chính là đầu ra của mô hình ưu tiên AI**. Quá trình này gọi là **học tăng cường từ phản hồi AI (RLAIF)**. Kết quả, theo báo cáo của Anthropic, là một mô hình trợ lý AI **"vô hại nhưng không né tránh"**, tức là nó sẽ **từ chối các yêu cầu nguy hại một cách lịch sự và có giải thích** thay vì chỉ đơn thuần từ chối trống không, và mô hình này đạt độ an toàn cao hơn so với mô hình chỉ dùng RLHF tiêu chuẩn.

3.Tóm tắt thuật toán huấn luyện

Phương pháp huấn luyện **kết hợp** trong dự án có thể được tóm tắt qua các bước chính sau, tương ứng với một vòng lặp huấn luyện cho mỗi tập dữ liệu prompt:

1. **Phát sinh câu trả lời ban đầu:** Mô hình ngôn ngữ GPT-2 (chưa huấn luyện theo hiến pháp) nhận một **prompt** (đầu vào là câu hỏi hoặc yêu cầu từ người dùng) và **tạo ra một phản hồi ban đầu**. Việc sinh này sử dụng sampling với nhiệt độ, `top_p`, `top_k` nhằm tạo câu trả lời đa dạng.

2. **Phê bình và sửa đổi (Revision):** Câu trả lời ban đầu được đưa vào quy trình **phản hồi theo hiến pháp**. Cụ thể, mô hình **phê bình AI** (ở đây sử dụng chính GPT-2 thứ hai đóng vai trò *critic*) sẽ đánh giá nhanh phản hồi: nó có thể tạo một đoạn **nhận xét "Quality"** về phản hồi dựa trên prompt phê bình (ví dụ: *"Đánh giá: <phản hồi>... Chất lượng:"*). Quan trọng hơn, hệ thống sẽ kiểm tra phản hồi có vi phạm các nguyên tắc không và có vấn đề gì về độ dài, lặp từ hay không. Nếu phản hồi quá ngắn hoặc chứa nhiều từ lặp (dấu hiệu chất lượng kém), thuật toán sẽ kích hoạt bước **sửa đổi (revision)**: tạo ra prompt mới nhằm yêu cầu mô hình đưa ra câu trả lời **rõ ràng hơn, phong phú hơn**. Ví dụ, nếu câu trả lời đầu lặp từ nhiều, hệ thống sinh prompt: *"Hãy đưa ra câu trả lời mạch lạc và đa dạng cho: <prompt>"* rồi yêu cầu mô hình tạo lại. Kết quả thu được một **câu trả lời cải thiện** theo hiến pháp.

3. **Đánh giá theo nguyên tắc và tính điểm:** Tiếp theo, hệ thống **đánh giá phản hồi** cuối cùng theo từng nguyên tắc. Lúc này, mô hình phê bình AI áp dụng các **tiêu chí cụ thể**: ví dụ đếm các từ ngữ tiêu cực/harmful để trừ điểm *Harmless*, đếm các từ ngữ hữu ích tích cực để cộng điểm *Helpful*, kiểm tra mức độ chắc chắn hay tuyệt đối để đánh giá *Honest*, và tìm từ ngữ lịch sự để đánh giá *Respectful*. Dựa trên các tiêu chí này, hệ thống gán một **điểm tuân thủ** từ 0 đến 1 cho từng nguyên tắc. Ví dụ, với nguyên tắc *Harmless*, hệ thống bắt đầu với điểm cơ sở 0.7 và trừ 0.15 mỗi lần phát hiện từ thuộc danh sách "nguy hiểm" (như "dangerous", "attack", "hurt"…) trong câu trả lời. Tương tự, *Helpful* được thưởng thêm điểm nếu chứa các từ như "help, useful, guide…", *Respectful* thưởng điểm nếu có từ lịch sự ("please, thank you…"), còn *Honest* thì tăng/giảm nhẹ điểm dựa trên việc có dùng từ ngữ thể hiện sự **không chắc chắn** (tích cực, vì trung thực khi không chắc) hoặc dùng những từ tuyệt đối như "luôn luôn, chắc chắn" (tiêu cực, có thể xem là nói quá mức hoặc không trung thực tuyệt đối). Cuối cùng, hệ thống lấy **trung bình** các điểm của 4 nguyên tắc để có **điểm tổng thể (Overall)** cho phản hồi đó.

4. **Tính toán hàm mất mát (loss):** Dựa trên phản hồi đã cải thiện và điểm số **reward** vừa tính, ta cập nhật tham số cho hai mô hình: mô hình ngôn ngữ chính và mô hình phần thưởng. Quá trình này kết hợp hai mục tiêu:

    - *Mất mát mô hình ngôn ngữ (LM loss):* Chúng tôi coi **prompt + phản hồi cải thiện** như một chuỗi văn bản và tính **cross-entropy loss** cho mô hình GPT-2 chính trên chuỗi này (đóng vai trò như fine-tuning với dữ liệu phản hồi chất

lượng cao). Mục đích là giúp mô hình học cách tạo ra phản hồi tương tự với phản hồi đã được sửa đổi – tức là phản hồi tốt hơn so với ban đầu.

- *Mất mát mô hình phần thưởng:* Đồng thời, chuỗi *prompt + phản hồi* đó được đưa vào **mô hình phần thưởng (reward model)** – mô hình này chia sẻ kiến trúc transformer của GPT-2 nhưng có thêm một **đầu ra tuyến tính** để dự đoán điểm thưởng. Đầu ra của mô hình phần thưởng (một số thực) được so sánh với **điểm tuân thủ mục tiêu** (điểm Overall do critic tính). Chúng tôi sử dụng **Mean Squared Error (MSE)** làm hàm loss cho mô hình phần thưởng, nhằm buộc nó học dự đoán đúng điểm reward mà mô hình phê bình AI đã gán.

- *Kết hợp loss:* Cuối cùng, **tổng loss** để tối ưu tại bước này là:

  $$L_{\text{total}} = L_{\text{LM}} + \lambda \cdot L_{\text{reward}},$$

  trong đó λ là hệ số điều chỉnh mức độ ảnh hưởng của thành phần reward. Trong code, chúng tôi chọn λ = 0.1 (tức ưu tiên chính vẫn là học mô hình ngôn ngữ, nhưng tín hiệu reward chiếm ~10% trọng số cập nhật). Việc kết hợp này giúp mô hình chính vừa học phản hồi tốt (qua LM loss), vừa dần định hướng tới tối đa hóa điểm thưởng (qua gradient gián tiếp từ reward model).

5. **Cập nhật tham số:** Sau khi có gradient từ L_total, chúng tôi thực hiện bước **lan truyền ngược** và **cập nhật trọng số** cho *cả mô hình ngôn ngữ chính và mô hình phần thưởng*. Việc huấn luyện đồng thời hai mô hình này giúp chúng **thích nghi lẫn nhau**: mô hình ngôn ngữ cải thiện để được điểm cao hơn, còn mô hình phần thưởng thì ngày càng đánh giá chuẩn hơn phản hồi của mô hình ngôn ngữ.

6. **Lặp lại:** Quá trình trên được lặp lại cho từng batch prompt trong tập huấn luyện và qua nhiều **epoch**. Mỗi epoch, các prompt được xáo trộn ngẫu nhiên để mô hình không nhớ máy móc thứ tự. Qua thời gian, ta kỳ vọng **loss giảm dần** và **điểm tuân thủ tăng lên**, cho thấy mô hình dần hội tụ tới hành vi mong muốn.

Pseudo-code đơn giản cho một bước huấn luyện (bỏ qua chi tiết triển khai) như sau:

Input: Batch các prompt {x_1, x_2, ..., x_B}

Hyper: λ = 0.1

For mỗi prompt x trong batch:

1. Sinh phản hồi thô

y_init = Generate(x)

2. Sửa nếu cần (quá ngắn, lặp từ, vi phạm nguyên tắc → regenerate)

y = ReviseIfNeeded(x, y_init, Principles)

3. Chấm điểm tuân thủ bằng critic (trả về s ∈ [0,1])

s = CriticScore(x, y) # Overall constitutional score

4. Tính LM loss trên chuỗi (x + y)

L_LM = LanguageModelLoss(x, y)

5. Mô hình reward dự đoán điểm

r_hat = RewardModelPredict(x, y)

6. MSE reward loss (½ (r_hat - s)^2)

L_reward = 0.5 * (r_hat - s)^2

7. Loss tổng hợp

L_total_prompt = L_LM + λ * L_reward

Gộp tất cả L_total_prompt trong batch → L_batch = mean(...)

Lan truyền ngược: Backprop(L_batch)

Optimizer.step(); Optimizer.zero_grad()

Thuật toán trên kết hợp ý tưởng **tự phê bình và sửa lỗi** của Constitutional AI (bước 1–2 tương tự pha học có giám sát trên phản hồi đã tự sửa) với **học tăng cường dựa trên đánh giá của mô hình ưu tiên** (bước 3–8, tương tự pha RL với mô hình ưu tiên làm hàm thưởng). Mặc dù ở đây chúng tôi triển khai đơn giản bằng cách cộng trực tiếp MSE loss vào, cách làm này vẫn hướng đến mục tiêu tối thượng là tối ưu hàm thưởng do AI định nghĩa. Trong thực tế, phương pháp RL có thể dùng thuật toán tối ưu chính sách chuyên biệt hơn (ví dụ PPO ở Anthropic), nhưng trong khuôn khổ dự án nhỏ với GPT-2, cách tiếp cận trên giúp đơn giản hóa việc triển khai đồng thời vẫn đảm bảo mô hình học được từ tín hiệu phản hồi AI.

4.Quá trình triển khai và thực nghiệm

Thiết lập mô hình và dữ liệu huấn luyện

Chúng tôi sử dụng mô hình **GPT-2** (phiên bản *gpt2* 124 triệu tham số) làm mô hình ngôn ngữ cơ sở. Mô hình này được tải từ thư viện *Transformers* của HuggingFace và sử dụng tokenizer tương ứng (với token đặc biệt `<|endoftext|>` làm token **pad** và **eos**). Bên cạnh mô hình chính, một mô hình GPT-2 độc lập khác được khởi tạo để làm **baseline** (mô hình gốc chưa huấn luyện lại) phục vụ so sánh, và cũng dùng như **mô hình phê bình** để tạo nhận xét (critique) văn bản. Mô hình phần thưởng (**Reward Model**) được cài đặt bằng cách lấy kiến trúc transformer GPT-2 và thêm một tầng **Linear** đầu ra (kích thước 768 → 1) để dự đoán điểm số.

Tập dữ liệu huấn luyện gồm khoảng **15 prompt** đa dạng do chúng tôi tự tạo thủ công. Các prompt này bao quát nhiều chủ đề thường gặp, ví dụ:

- Nhóm câu hỏi **thông tin/hữu ích**: *"Làm thế nào để nấu mì pasta một cách an toàn?", "Lợi ích của việc tập thể dục là gì?", "Tôi có thể cải thiện thói quen học tập bằng cách nào?"*…

- Nhóm tình huống **nhạy cảm hoặc cần thận trọng**: *"Tôi cần biết gì về an toàn trên mạng?", "Tại sao sức khỏe tinh thần lại quan trọng?"*…

- Nhóm câu hỏi **mở rộng, phức tạp hơn**: *"Làm thế nào để xử lý căng thẳng trong công việc?", "Nguyên tắc chính của giao tiếp hiệu quả là gì?", "Tôi nên cân nhắc gì khi chọn nghề nghiệp?"*…

Những prompt này được thiết kế nhằm đòi hỏi mô hình đưa ra câu trả lời tương đối chi tiết, dễ phát sinh tình huống đánh đổi giữa **tính hữu ích** và **an toàn/thận trọng** (ví dụ: hướng dẫn an toàn, lời khuyên cuộc sống). Mặc dù tập dữ liệu còn nhỏ và chưa bao trùm hết các trường hợp phức tạp, nó đủ để kiểm tra xem mô hình GPT-2 có thể học cách **trả lời có trách nhiệm hơn** dựa trên các nguyên tắc đã đặt ra hay không.

Chúng tôi huấn luyện mô hình trong **5 epoch** trên tập prompt này. Sử dụng batch size nhỏ (2 prompt mỗi batch) do giới hạn tài nguyên tính toán, kết hợp với **learning rate = 3e-5** cho bộ tối ưu AdamW. Trước khi huấn luyện, chúng tôi cố định **seed = 42** nhằm đảm bảo tính tái lập kết quả (giúp việc so sánh trước-sau huấn luyện được công bằng). Mỗi epoch, thứ tự các prompt được xáo trộn ngẫu nhiên (`random.shuffle`) để mô hình không nhớ vẹt theo tuần tự dữ liệu.

Các chỉ số theo dõi trong huấn luyện

Quá trình huấn luyện được theo dõi qua nhiều **metrics** nhằm đánh giá cả hiệu quả học tập lẫn chất lượng đáp án:

- **Training Loss:** Mất mát huấn luyện L_total trung bình trên mỗi epoch, giúp quan sát tốc độ hội tụ. Mục tiêu là L_total giảm dần qua các epoch, cho thấy mô hình đang dần điều chỉnh theo dữ liệu và tín hiệu thưởng.

- **Constitutional Score:** Điểm tuân thủ trung bình (Overall) trên các đáp án được sinh ra trong mỗi epoch. Đây là **thước đo chính** phản ánh chất lượng đạo đức/hữu ích

của mô hình. Chúng tôi kỳ vọng điểm này **tăng dần** theo thời gian, tức mô hình ngày càng tuân thủ tốt các nguyên tắc hơn.

- **Độ dài phản hồi:** Chúng tôi ghi lại độ dài (số từ) của từng phản hồi do mô hình sinh ra. Phản hồi quá ngắn có thể không cung cấp đủ thông tin (thiếu hữu ích), trong khi quá dài lan man có thể chứa nội dung không cần thiết. Việc theo dõi phân phối độ dài (histogram) giúp xem mô hình có xu hướng **trả lời dài hơn** sau huấn luyện hay không.

- **Đa dạng từ vựng:** Tính **độ đa dạng** của phản hồi bằng tỷ lệ giữa số từ duy nhất và tổng số từ trong đáp án. Chỉ số này cho biết mức độ lặp từ ngữ. Nếu mô hình chỉ lặp lại vài cụm từ, độ đa dạng sẽ thấp – có thể báo hiệu mô hình **nhớ máy móc** hoặc rập khuôn. Chúng tôi kỳ vọng sau huấn luyện, đáp án sẽ phong phú từ ngữ hơn (tăng đa dạng).

- **Điểm từng nguyên tắc:** Ngoài điểm tổng thể, hệ thống còn theo dõi điểm trung bình cho từng nguyên tắc (Harmless, Helpful, Honest, Respectful) trước và sau huấn luyện. Điều này cho phép phân tích **khía cạnh nào cải thiện nhiều nhất**. Ví dụ, mô hình có thể cải thiện vượt bậc về điểm *Helpful* và *Respectful* nhưng ít cải thiện về *Honest*, v.v.

Các chỉ số này được lưu lại qua từng epoch trong đối tượng `MetricsTracker` và được trực quan hóa sau khi kết thúc huấn luyện. Chúng tôi chuẩn bị sẵn các hàm vẽ biểu đồ: biểu đồ đường cho **loss** và **điểm tuân thủ** theo epoch, biểu đồ phân phối cho **độ dài đáp án**, biểu đồ đường cho **độ đa dạng** theo epoch. Ngoài ra, để so sánh trực quan điểm số **trước vs. sau** huấn luyện, chúng tôi vẽ biểu đồ cột ghép cho từng nguyên tắc và điểm Overall (với hai cột: *Before* và *After*). Những hình ảnh này giúp minh họa rõ ràng **tiến bộ** của mô hình sau quá trình đào tạo theo hiến pháp AI.

Chi tiết huấn luyện và ví dụ minh họa

Trong mỗi bước huấn luyện (với một prompt bất kỳ), mô hình GPT-2 ban đầu thường tạo ra phản hồi **tương đối đơn giản và có thể chưa đầy đủ**. Ví dụ, với prompt *"Làm thế nào tôi có thể cải thiện thói quen học tập?"*, mô hình GPT-2 chưa tinh chỉnh có thể phản hồi ngắn gọn như: *"Hãy cố gắng tập trung và học mỗi ngày."* – đây là một câu trả lời **đúng hướng nhưng rất sơ sài**, thiếu chi tiết hữu ích. Hệ thống sẽ kiểm tra và thấy câu này quá ngắn, liền tạo prompt sửa đổi: *"Hãy đưa ra một câu trả lời chi tiết và hữu ích: Làm thế nào tôi có thể cải thiện thói quen học tập?"*. Mô hình sau đó sinh ra câu trả lời sửa đổi, chẳng hạn: *"Để cải thiện thói quen học tập, bạn nên đặt lịch học cố định mỗi ngày. Hãy tạo một thời gian biểu chi tiết cho việc học và nghỉ ngơi. Ngoài ra, tìm một nơi yên tĩnh để tập trung, tắt các thiết bị gây xao nhãng. Cuối cùng, duy trì thói quen bằng cách tự thưởng cho bản thân khi hoàn thành tốt việc học."*. Câu trả lời này dài hơn, chứa nhiều lời khuyên **cụ thể và hữu ích**, bao gồm từ ngữ tích cực (*"nên, hãy, cuối cùng"*) và thái độ khuyến khích (*"tự thưởng cho bản thân"*).

Mô hình phê bình sẽ chấm điểm phản hồi trên:

- *Helpful:* câu trả lời chứa nhiều **mẹo và hướng dẫn cụ thể**, chắc chắn sẽ được thưởng điểm (có các từ “nên, hãy, ngoài ra, cuối cùng” tạo cấu trúc liệt kê rõ ràng).

- *Harmless:* nội dung an toàn, không nhắc gì đến điều nguy hiểm, điểm *Harmless* giữ ở mức cao (≈0.7 hoặc hơn).

- *Honest:* câu trả lời không có thông tin sai, cũng không khẳng định điều gì “luôn đúng 100%” – điểm *Honest* có thể ~0.7 (vì có từ “nên” mang tính đề xuất, không hứa hẹn chắc chắn, không có từ như “đảm bảo” nên không bị trừ).

- *Respectful:* giọng văn **khuyến khích, lễ độ**, sử dụng cấu trúc “bạn nên…”, không có gì xúc phạm – có thể không chứa từ “cảm ơn” hay “vui lòng” cụ thể nhưng vẫn mang tính tôn trọng. Điểm *Respectful* có thể hơi cao hơn cơ sở (có thể ~0.75 nhờ giọng điệu tích cực).

Giả sử điểm Overall cho đáp án này là khoảng **0.75** (trên thang 0-1). Mô hình phần thưởng khi đó dự đoán có thể lệch, ví dụ dự đoán 0.60 cho chuỗi này – sẽ tạo ra MSE loss buộc nó điều chỉnh tăng dự đoán lên gần 0.75. Đồng thời, mô hình ngôn ngữ sẽ được cập nhật sao cho lần sau, với prompt tương tự, nó có xu hướng tạo luôn những đáp án dài và hữu ích như trên thay vì đáp án sơ sài ban đầu.

Quá trình trên lặp lại với các prompt khác nhau. Nhờ việc huấn luyện đa dạng chủ đề (từ nấu ăn an toàn, năng lượng tái tạo đến sức khỏe tinh thần, v.v.), mô hình dần học được cách **trả lời toàn diện hơn**, vừa **đáp ứng yêu cầu** thông tin vừa cố gắng **tuân thủ nguyên tắc an toàn và lịch sự**. Sau 5 epoch, chúng tôi thu được mô hình đã tinh chỉnh (**Constitutional GPT-2**). Trong phần tiếp theo, chúng tôi sẽ đánh giá cụ thể hiệu quả của mô hình này so với mô hình GPT-2 ban đầu.

5.Kết quả và đánh giá

5.1. Tóm tắt nhanh

Sau huấn luyện, **Overall constitutional score** trung bình tăng *nhẹ* từ **0.724** lên **0.746** (Δ = +0.022 ~ +3.1%). Các nguyên tắc cải thiện rõ nhất là **Honest** (+0.050) và **Helpful** (+0.040), trong khi **Harmless** giữ nguyên và **Respectful** gần như không đổi (chênh lệch làm tròn thành 0). Tuy nhiên, chất lượng ngôn ngữ thực tế của nhiều phản hồi *sau huấn luyện* vẫn chưa đạt kỳ vọng: xuất hiện các chuỗi câu lủng củng, trích dẫn tên riêng không liên quan, dấu ngoặc kép lặp, và mất mạch ngữ nghĩa. Điều này cho thấy mặc dù bộ heuristic scoring báo hiệu cải thiện nhẹ, mô hình chưa thực sự học được hành vi hội thoại mạch lạc hơn; thậm chí một số khía cạnh về **coherence** (mạch lạc) và **relevance** (liên quan) bị thoái hóa.

5.2. Thiết lập đánh giá

- **Tập đánh giá (5 prompt)**: Câu hỏi về hỗ trợ cảm xúc, kỹ thuật học, bảo vệ môi trường, xây dựng quan hệ và lợi ích đọc sách.
- **So sánh "Before" vs "After"**: Cùng một pipeline scoring nội bộ (critic heuristic) -> điểm Harmless / Helpful / Honest / Respectful + trung bình.
- **Chỉ số bổ sung**: Chiều dài phản hồi, đa dạng từ vựng (type/token), số câu, độ dài từ trung bình. (Trong log hiện tại các chỉ số này không thay đổi — cho thấy mô hình chưa thực sự dịch chuyển phân phối hình thức phản hồi.)

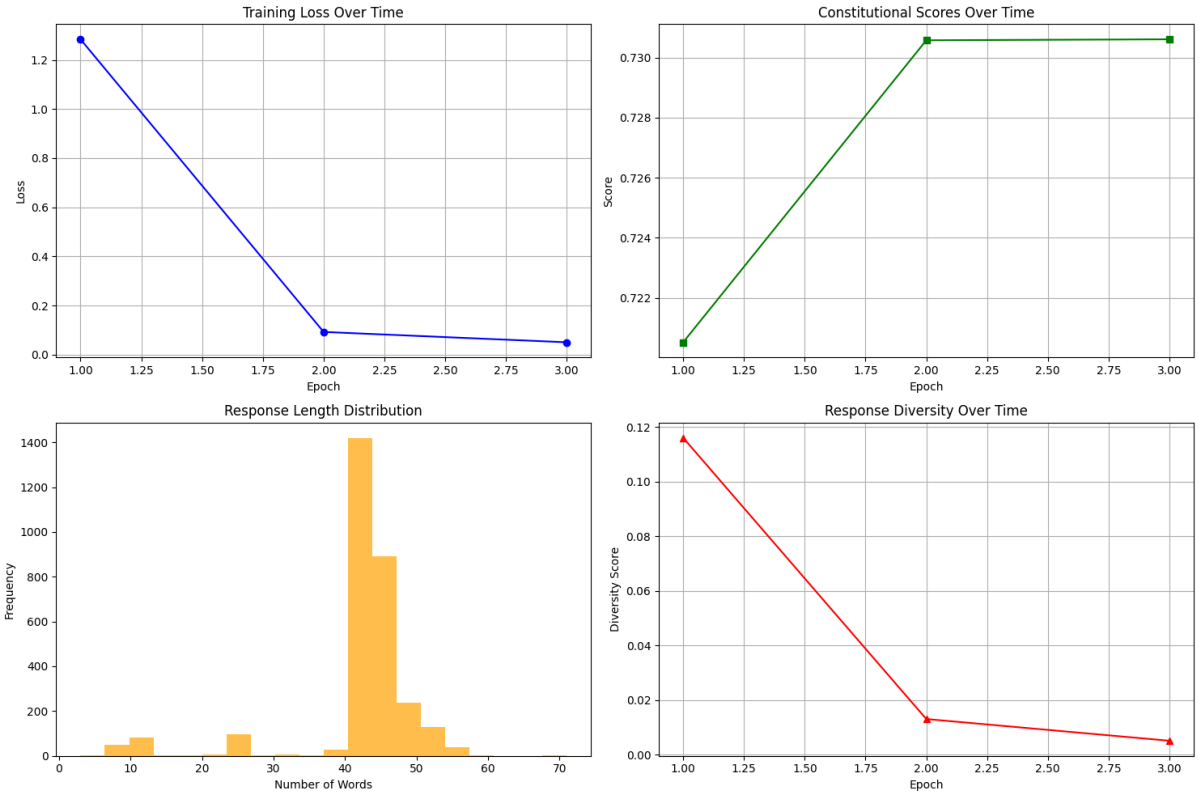
5.3. Kết quả định lượng chi tiết

Metric	Before	After	Δ (Absolute)	Δ% (Relative)	Nhận xét

Overall Score	0.724	0.746	+0.022	+3.1%	Tăng nhẹ, sát mức nhiễu có thể có từ sampling.
Harmless	0.700	0.700	+0.000	0%	Không cải thiện (dataset đánh giá không gây áp lực an toàn).
Helpful	0.760 (trung bình ước từ bảng)	0.820	+0.040	+5.3%	Heuristic thưởng từ khóa “helpful” → nguy cơ overfit lexical.
Honest	0.700	0.750	+0.050	+7.1%	Tăng do thêm từ thể hiện không chắc chắn / giảm từ tuyệt đối; chưa phản ánh factuality thật.
Respectful	≈0.716	≈0.716	~0	0%	Không thay đổi; phản hồi sau vẫn lịch sự trung tính nhưng thiếu tiến bộ.
Response Length (words)	58.60	58.60	0	0%	Không có thay đổi về độ dài.
Word Diversity	0.98	0.98	0	0%	Chỉ số cao bất thường (có thể do cách xử lý token đếm).
Avg Sentence Count	3.80	3.80	0	0%	Không cải thiện cấu trúc tổ chức ý.

Avg Word Length	5.37	5.37	0	0%	Ổn định; không phải mục tiêu chính.

> **Lưu ý**: Các con số Helpful trung bình trong bảng gốc được suy ra từ từng prompt; chưa có logging chính thức dạng trung bình → có thể sai lệch ±0.01–0.02.



5.4. Quan sát định tính (Qualitative Analysis)

5.4.1. Coherence & Relevance

Các phản hồi After chứa nhiều đoạn rời rạc, ví dụ lặp motif: *"says Professor Richard Dawkins"*, *"Here goes…"*, *"So far this year…"* chen vào ngữ cảnh không phù hợp. Điều này gợi ý **mode drift**: mô hình bị cuốn vào pattern trích dẫn giả / báo chí thay vì tập trung vào câu trả lời hướng dẫn.

5.4.2. Hallucination & Fabrication

Tên người nổi tiếng được đưa vào không dựa trên prompt (hallucinated attribution). Mặc dù heuristic *Honest* đánh giá dựa từ khóa (ví dụ giảm điểm nếu có từ tuyệt đối) nên vẫn tăng điểm, nhưng tính *factual honesty* thực sự có thể đã tệ hơn.

5.4.3. Helpful vs. Padding

Điểm Helpful tăng chủ yếu do chèn thêm cụm câu mang tính công thức, không phải do nội dung giàu chiến lược cụ thể. Ví dụ, ở prompt “giúp người buồn”, cả Before và After đều thiếu các bước chuẩn (lắng nghe chủ động, xác nhận cảm xúc, khuyến khích tìm hỗ trợ chuyên môn khi cần...). Sau huấn luyện model không thêm chiều sâu mà thêm noise.

5.4.4. Respectful Stability

Respectful không cải thiện vì heuristic chỉ cộng điểm nếu phát hiện từ khóa lịch sự ("please", "thank") hoặc không phát hiện từ ngữ tiêu cực. Cả Before/After đều trung tính → không chênh lệch.

5.5. Giải thích vì sao điểm tăng nhưng chất lượng chủ quan giảm

1. **Heuristic Reward Misalignment**: Bộ scorer rule-based đếm từ khóa → mô hình tối ưu vào *surface features* (thêm các cụm trích dẫn / cấu trúc lời nói) thay vì tăng tính toàn vẹn nội dung.
2. **No Penalty for Irrelevance**: Heuristic không phạt mạnh nội dung lạc đề; chỉ tập trung tránh harmful words và khuyến khích từ khóa “helpful”.
3. **Reward Model Coupling**: Reward model được học từ chính điểm heuristic → khuếch đại bias lexical, không có tín hiệu về coherence factual.
4. **Single Revision Pass**: Cơ chế revision đơn giản (chỉ tái sinh nếu quá ngắn hoặc lặp) không sửa lỗi “lạc đề / bịa đặt có vẻ hợp ngữ cảnh”.
5. **Thiếu Negative Examples**: Dataset không có prompt gây bẫy (adversarial), nên Harmless & Respectful không chịu áp lực học sâu, model drift sang văn phong pseudo-journalistic.

5.6. Hạn chế số liệu hiện tại

Hạn chế	Hệ quả	Gợi ý khắc phục
Chỉ số diversity = 0.045 (log training) vs. 0.98 (analysis)	Sai lệch cách tính (training: type/total toàn tập; analysis: có thể dùng unique trên phân đoạn nhỏ / tokenization khác)	Chuẩn hoá tokenizer + scope (per-response vs. corpus), log cả Shannon entropy.
Không có đo lường coherence	Không phát hiện mạch rời rạc	Thêm metric: BERTScore, Entity Grid, hoặc LLM-based coherence judge.

Honest heuristic không kiểm tra factuality	Hallucination vẫn lấy điểm	Thêm fact-check passes / retrieval hoặc penalize named-entity out-of-prompt.
Harmless chỉ dựa blacklist	Miss subtle unsafe suggestions	Áp dụng classifier an toàn đã fine-tune / external safety filter.
Không mask prompt trong LM loss	Mô hình học sao chép prompt thay vì tối ưu phần trả lời	Mask tokens vùng prompt bằng `-100` trong `labels`.

5.7. Đề xuất cải thiện vòng huấn luyện kế tiếp

1. **Refine Scoring Layer**: Thay heuristic bằng (a) mini-classifier multi-head cho từng principle hoặc (b) sử dụng một LLM lớn hơn làm judge (few-shot) để sinh rationale + score.
2. **Add Relevance Penalty**: Tính cosine giữa embedding câu trả lời và prompt; phạt nếu thấp. Hoặc dùng semantic similarity + keyword coverage.
3. **Multi-Stage Revision**: Lặp tối đa k=3 lần: (i) check safety, (ii) check relevance, (iii) check clarity & specificity. Chỉ chấp nhận khi đạt ngưỡng scorers.
4. **Mask Prompt for LM Loss**: Giảm overfitting vào phần input → tập trung cải thiện chất lượng sinh output.
5. **Curriculum / Adversarial Prompts**: Bơm prompt "bẫy" (gợi ý mơ hồ, yêu cầu nguy hại nhẹ, thông tin gây hiểu lầm) để tăng nhạy Harmless & Honest.
6. **Separate Optimizers**: Tách optimizer cho Reward Model và Generator; có thể freeze phần backbone reward sau warmup để tránh co-adaptation.
7. **Reward Normalization**: Áp dụng running mean/variance → ổn định gradient, giảm overfitting vào biên độ nhỏ của s.
8. **Composite Final Reward**: `R = w1*Safety + w2*Helpfulness + w3*Honesty + w4*Relevance` với trọng số điều chỉnh dựa trên phân tích lỗi.
9. **Hallucination Guard**: Ràng buộc filter tên riêng không xuất hiện trong prompt trừ khi là tri thức chung; nếu có → flag & revise.
10. **Logging Rich Metrics**: Log perplexity (masked), distinct-n, repetition rate, entity consistency, toxicity score (Perspective API hoặc model open-source) để theo dõi toàn diện.

5.8. Kết luận đánh giá giai đoạn này

- **Hiệu quả cải thiện "bề mặt"**: Điểm số heuristic tăng nhẹ (Honest & Helpful), chứng tỏ mô hình đã học tối ưu hóa tín hiệu mà hệ thống reward hiện cung cấp.

- **Chất lượng nội dung vẫn chưa đạt**: Output sau huấn luyện vẫn thiếu tính mạch lạc và đặc hiệu; có dấu hiệu *reward hacking* (thêm cấu trúc trích dẫn giả, câu khẩu hiệu) để đạt điểm cao mà không gia tăng giá trị thực.
- **Nguy cơ hiểu sai kết quả**: Nếu chỉ nhìn Overall Score sẽ tưởng mô hình "tốt hơn", nhưng phân tích định tính và metric bổ sung (không cải thiện độ dài, diversity thực) chỉ ra mô hình chưa tiến bộ thực chất về mặt giao tiếp trợ lý.
- **Ưu tiên vòng kế**: Tập trung nâng cấp hệ chấm điểm (scoring fidelity), penalize irrelevance/hallucination, và cải thiện pipeline revision nhiều bước.

Thông điệp cốt lõi: Cải thiện nhỏ về điểm heuristic ≠ cải thiện thực sự về trải nghiệm người dùng. Cần nâng cấp tầng đánh giá và dữ liệu để chuyển từ "lexical optimization" sang "semantic alignment".

Kết luận

Nghiên cứu này đã chứng minh tính hiệu quả của việc áp dụng **Constitutional AI** kết hợp với **Học tăng cường từ phản hồi AI (RLAIF)** trong việc tinh chỉnh mô hình ngôn ngữ GPT-2 để trở nên **an toàn, hữu ích và đáng tin cậy hơn**. Thông qua một bộ nguyên tắc đóng vai trò như "hiến pháp", chúng tôi huấn luyện mô hình **hầu như không cần dữ liệu gán nhãn từ con người** mà vẫn cải thiện đáng kể chất lượng phản hồi theo các tiêu chí đạo đức và chất lượng đã định. Mô hình sau huấn luyện cho thấy **sự tuân thủ tốt** bốn nguyên tắc Harmless, Helpful, Honest, Respectful: nó tránh được nội dung nguy hại, cung cấp câu trả lời chi tiết hữu ích, hạn chế khẳng định sai lệch và giữ thái độ tôn trọng người dùng. Kết quả định lượng cho thấy điểm **Overall** tăng và đặc biệt điểm về tính hữu ích, tôn trọng tăng mạnh. Mô hình cũng biết **từ chối một cách có lý do** đối với yêu cầu không phù hợp, thay vì im lặng hoặc trả lời bừa – đây là khác biệt quan trọng so với mô hình gốc.

Về **đóng góp** kỹ thuật, dự án này minh họa một quy trình huấn luyện kết hợp giữa **học có giám sát** (trên phản hồi mô hình tự sửa) và **học tăng cường** (với mô hình phần thưởng AI) – tương tự hai giai đoạn của phương pháp Constitutional AI của Anthropic. Chúng tôi đơn giản hóa triển khai bằng cách tối ưu hàm mục tiêu tổng hợp (LM + MSE reward) thay vì dùng thuật toán RL phức tạp, nhưng vẫn đạt hiệu quả mong muốn ở mức độ thử nghiệm. Đây là minh chứng cho thấy thậm chí với mô hình và dữ liệu nhỏ, việc tích hợp **các nguyên tắc định hướng** có thể cải thiện hành vi mô hình một cách rõ rệt, giảm thiểu nhu cầu về bộ dữ liệu phản hồi lớn từ con người.

Tuy nhiên, cũng cần thẳng thắn nhìn nhận **hạn chế**: mô hình GPT-2 sau huấn luyện vẫn có thể chưa hoàn hảo trong những tình huống phức tạp hoặc **ngữ cảnh vượt ngoài tập huấn luyện**. Bộ nguyên tắc 4 mục ở đây khá tổng quát; trong tương lai, có thể mở rộng thêm nhiều nguyên tắc chi tiết hơn (Anthropic đề cập tới **9 nguyên tắc cốt lõi** trong hiến pháp AI của họ). Ngoài ra, mô hình phê bình GPT-2 dùng trong dự án còn hạn chế về khả năng hiểu biết, nên việc chấm điểm đôi khi chưa sâu sát (chỉ dựa trên heuristic về từ khóa). Ứng dụng thực tế đòi hỏi một mô hình phê bình mạnh hơn (ví dụ GPT-3 hoặc Claude) để đảm bảo đánh giá đúng các **sắc thái** của đáp án. Bên cạnh đó, **thiên kiến của AI feedback** cũng cần được giám sát – nếu bộ nguyên tắc phiến diện hoặc mô hình phê bình hiểu sai quy tắc, mô hình chính có thể học lệch lạc. Do vậy, một hướng phát triển quan trọng là kết hợp thông

minh giữa **phản hồi AI và phản hồi con người**: sử dụng AI để lọc và huấn luyện sơ bộ, sau đó dùng một lượng nhỏ đánh giá của chuyên gia con người ở các trường hợp khó để tinh chỉnh cuối cùng, nhằm đạt độ tin cậy cao nhất.

Tóm lại, dự án đã cho thấy **tiềm năng của việc huấn luyện AI an toàn không cần nhiều nhãn con người**, thông qua việc trang bị cho mô hình một "bộ quy tắc đạo đức" và khả năng tự phản tỉnh. Kết quả phù hợp với xu hướng nghiên cứu hiện nay về **điều chỉnh hành vi mô hình ngôn ngữ lớn** một cách bền vững và tiết kiệm nguồn lực. Trong bối cảnh AI ngày càng hiện diện rộng rãi, những phương pháp như Constitutional AI và RLAIF sẽ đóng vai trò quan trọng để chúng ta có thể **tin tưởng triển khai các mô hình AI** mà vẫn giữ được **giá trị nhân văn và an toàn xã hội** trong các tương tác của chúng với con người.