



Final Report

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

Instructor: Assoc. Prof. Dr. Le Anh Cuong

Presented by: 521H0516 - Phan Anh Tuan
521H0285 - Pham Tran Tien Phat
521H0324 - Nguyen Van Truong



521H0285	Phạm Trần Tiến Phát	<ul style="list-style-type: none">• Tìm hiểu lý thuyết về RL• Code Q-Learning• Viết report• Làm slide thuyết trình
521H0324	Nguyễn Văn Trường	<ul style="list-style-type: none">• Tìm hiểu về RL• Tìm hiểu lý thuyết về RLHF./RFAIF• Code RLHF
521H0516	Phan Anh Tuấn	<ul style="list-style-type: none">• Tìm hiểu về RL• Code Machine Translation

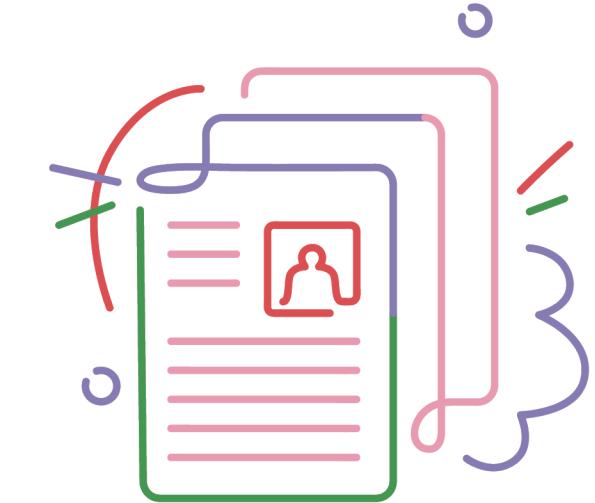
Task 1

Reinforcement Learning

Introduction to Reinforcement Learning

Reinforcement Learning là gì?

- Nhánh của học máy
- Agent học cách hành động của môi trường qua **thưởng – phạt**
- Dựa trên cơ chế **thử – sai** như con người trong việc học từ trải nghiệm



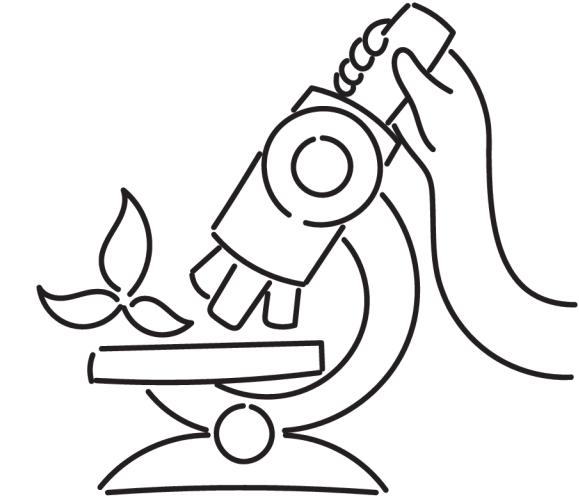
Mục tiêu của Reinforcement Learning?

Tìm chiến lược **hành động tối ưu**
Tối đa hóa tổng phần thưởng tích lũy
Học cách ra quyết định **tốt nhất** qua tương tác với môi trường

Introduction to Reinforcement Learning

Thành phần chính của RL

- Agent
- Environment
- State
- Action
- Reward
- Policy
- Value Function



Ứng dụng thực tế

- Game AI (AlphaGo, Dota2 Bot)
- Xe tự lái (Autonomous Vehicles)
- Robot tự học di chuyển
- Tối ưu quảng cáo, gợi ý nội dung
- Tài chính – giao dịch tự động
- Huấn luyện mô hình ngôn ngữ (RLHF)

Types of Reinforcement Learning

Model-Based

- Agent **mô phỏng** môi trường để **dự đoán** kết quả hành động
- Học **nhanh**, **ít cần** trải nghiệm thực tế
- Ví dụ: Game cờ vua – lên kế hoạch nước đi từ luật chơi

Model-Free

- Agent **không biết** trước mô hình, học qua tương tác thực tế
- Phù hợp với môi trường **phức tạp**, không xác định rõ
- Ứng dụng: Chatbot, xe tự lái, AI đàm thoại

4 nhánh nhỏ

- **Value-based:** Q-learning
- **Policy-based:** PPO
- **Actor-Critic:** RLHF (có RM và PPO)
- **Preference-based:** DPO

RL Applications in LLMs

- LLMs (như GPT) huấn luyện ban đầu bằng supervised learning chỉ học cấu trúc ngôn ngữ, chưa hiểu ý định người dùng.
- RL giúp mô hình học sinh động hơn: phản hồi có ích, đúng ngữ cảnh, tránh độc hại.
- Agent: là mô hình ngôn ngữ
- Môi trường: là người dùng/ngữ cảnh
- Reward: đến từ phản hồi của người (hoặc AI đánh giá)

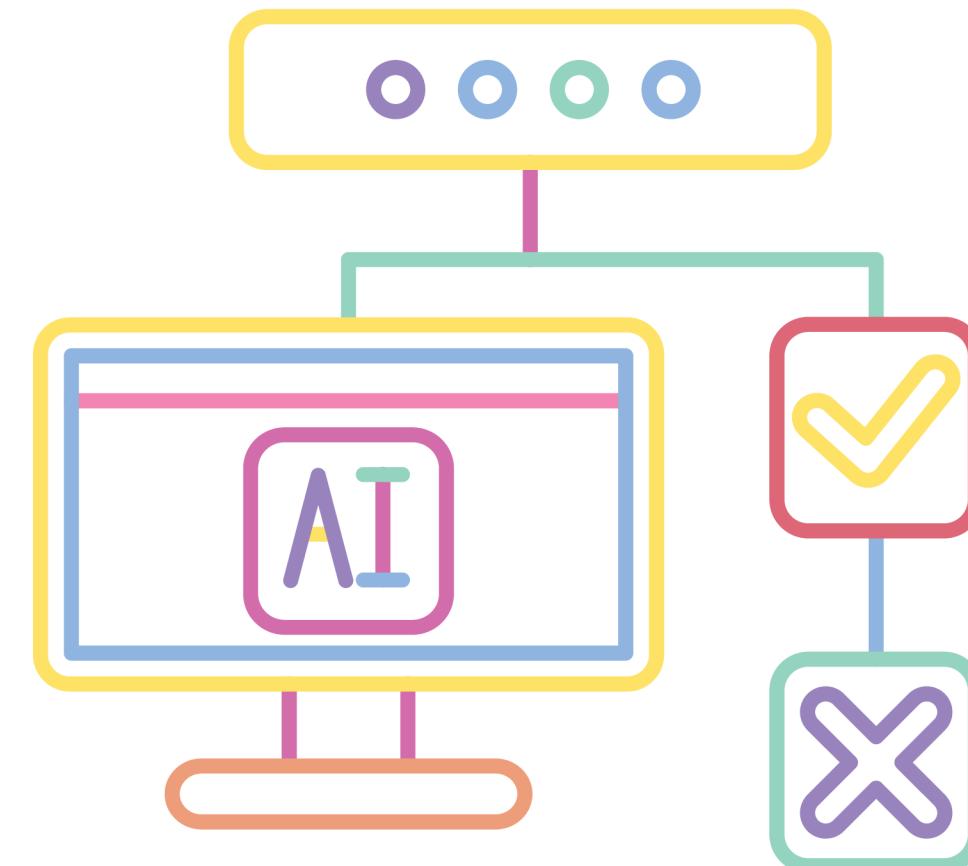
Reinforcement Learning from Human Feedback

RLHF là gì?

Là kỹ thuật dùng phản hồi của con người để tinh chỉnh mô hình AI, giúp phản hồi tự nhiên, đúng ý người hơn.

Mục tiêu:

- Đưa mô hình gần với giá trị và kỳ vọng con người.
- Tránh phản hồi độc hại, vô nghĩa hoặc thiếu thân thiện.



Tại sao cần RLHF?

Vì phản hồi "hay" trong NLP khó lượng hóa → cần người đánh giá để làm phần thưởng

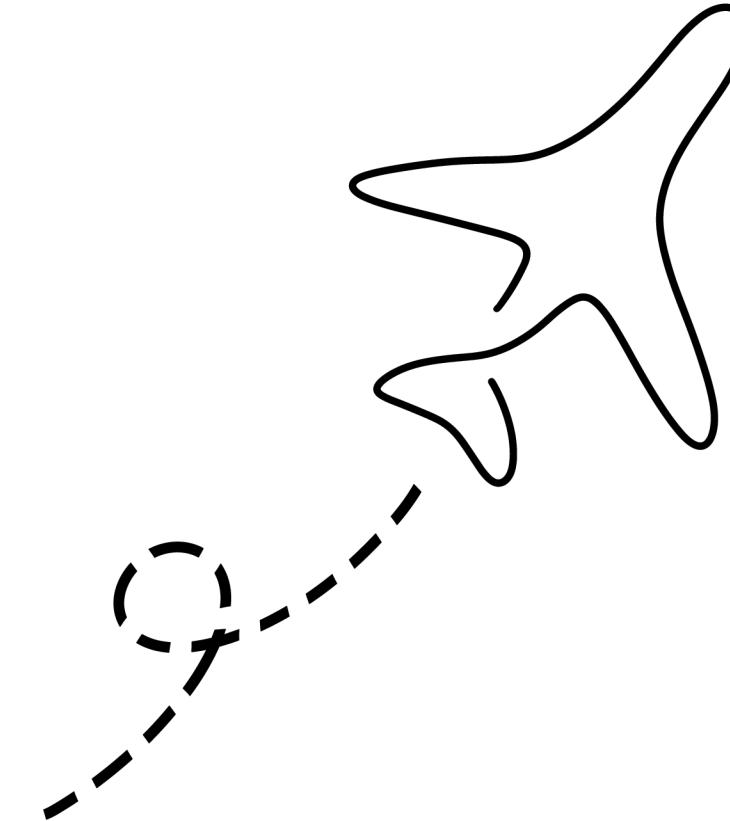
RLHF Procedure: SFT - RM - PPO

1. SFT – Supervised Fine-tuning:

Huấn luyện có giám sát ban đầu từ cặp (prompt, response).

2. RM – Reward Model:

- Mô hình đánh giá được huấn luyện từ phản hồi người.
- So sánh các câu trả lời → đánh điểm cho từng câu.



3. PPO – Proximal Policy Optimization:

- Dùng thuật toán PPO để tối ưu mô hình sinh văn bản dựa trên đánh giá của RM.
- Giữ sự ổn định bằng cách giới hạn thay đổi quá lớn (dùng KL-divergence làm penalty).

Reinforcement Learning from AI Feedback

Reinforcement Learning from AI Feedback (RLAIF) là gì?

Giống RLHF nhưng không dùng người thật đánh giá, mà dùng AI đã huấn luyện trước làm người đánh giá.

Ưu điểm	Hạn chế
Tiết kiệm chi phí, mở rộng quy mô dễ hơn RLHF.	Phụ thuộc vào chất lượng mô hình đánh giá.
Dễ triển khai với tập dữ liệu lớn.	Có thể sai lệch nếu mô hình phản hồi không chính xác.

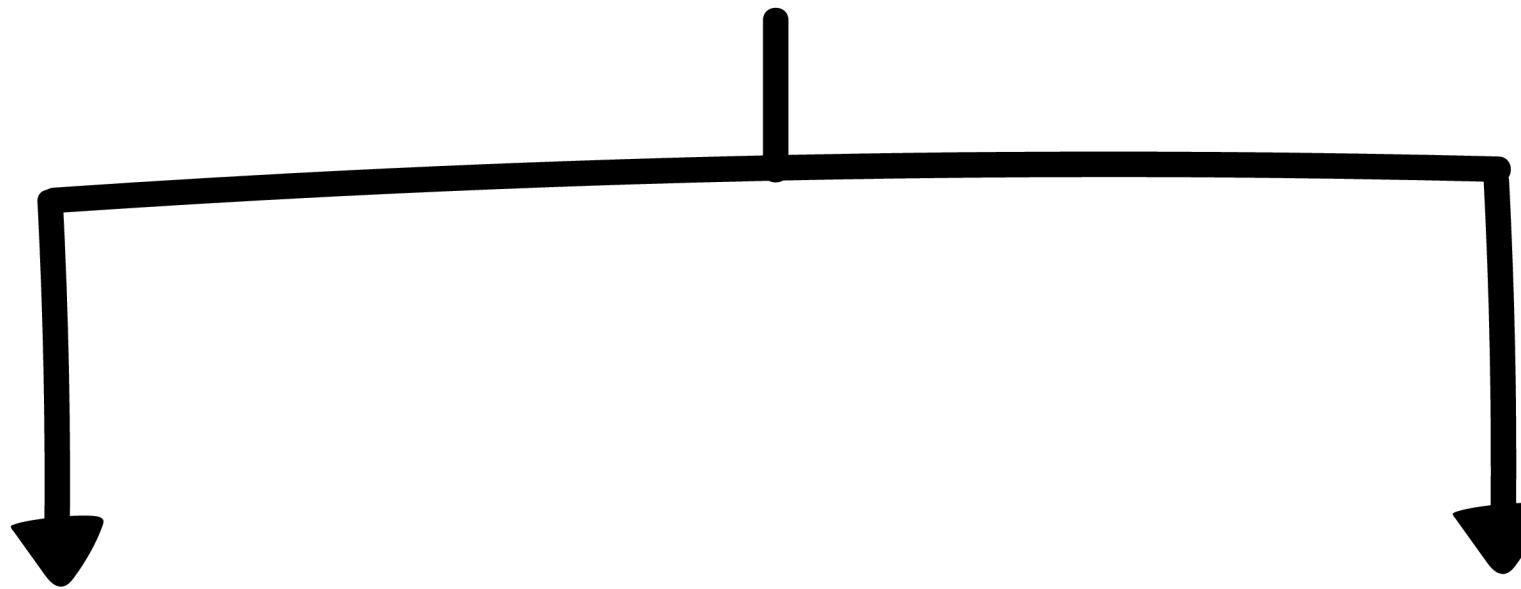
Comparison of RLHF and RLAIF

	RLHF	RLAIF
Nguồn phản hồi	Con người thật	AI (mô hình đánh giá)
Ưu điểm	Phản ánh giá trị thật của người dùng	Rẻ hơn, mở rộng tốt
Nhược điểm	Tốn thời gian, chi phí cao	Phụ thuộc vào chất lượng AI đánh giá
Ứng dụng	Tinh chỉnh ChatGPT, Claude	Gemini, Claude 3, tinh chỉnh tự động

Task 2

Machine Translation

Two approaches for Machine Translation Problem



Fine-tuning mô hình có sẵn

Huấn luyện mô hình từ đầu

Dataset - OPUS-100 and Data Preprocessing

Dataset

OPUS-100: tập dữ liệu đa ngôn ngữ từ TED, phụ đề phim, GNOME docs, Kinh Thánh...

Sử dụng:

- 1% để fine-tune
- 10% để huấn luyện từ đầu

Data Preprocessing

- Kiểm tra giá trị Null (rỗng)
- Kiểm tra giá trị trùng lặp
- Kiểm tra độ dài (length)
- Kiểm tra giá trị ngoại lai (outlier)
- Xóa khoảng trắng thừa
- Xóa ký tự đặc biệt nhưng giữ lại dấu câu
- Chuẩn hóa Unicode (quan trọng với tiếng Việt)
- Chuyển tất cả thành chữ thường (lowercase)
- Xóa bản ghi trùng lặp
- Xóa các dòng quá dài hoặc quá ngắn

Approach 1 - Fine-tune the EnviT5 model

EnviT5: mô hình T5 encoder-decoder do VietAI huấn luyện

- 220 triệu tham số, 32.000 từ vựng
- Tối ưu cho tiếng Việt và Anh

Kỹ thuật tối ưu:

- 4-bit Quantization: giảm bộ nhớ, tăng tốc độ
- LoRA (Low-Rank Adaptation): chỉ tinh chỉnh 2.4% trọng số → tiết kiệm tài nguyên

Ưu điểm:

Tận dụng kiến thức có sẵn của mô hình, huấn luyện nhanh hơn

Approach 2 - Building From Scratch

Xây dựng kiến trúc Transformer Encoder-Decoder với:

- 4 tầng
- 8 đầu attention
- $d_{model}=256$
- dropout=0.1

Tự tạo BPE tokenizer riêng cho tiếng Anh và tiếng Việt từ đó giữ được dấu, tiếng Việt đa nghĩa, chia nhỏ từ phức

Cấu hình huấn luyện:

- **Optimizer:** AdamW (LR = **5e-5**, weight decay = **0.01**)
- Learning rate scheduler: ReduceLROnPlateau (patience = 3)
- Loss function: Cross-entropy
- Maximum epochs: **20**
- Early stopping: patience = 3 epochs



	BLEU	ROUGE
Tên đầy đủ	Bilingual Evaluation Understudy	Recall-Oriented Understudy for Gisting Evaluation
Ứng dụng chính	Đánh giá chất lượng dịch máy	Đánh giá chất lượng tóm tắt văn bản
Cách tính điểm	So sánh precision của n-gram	So sánh recall (và precision) của n-gram
Tập trung vào	Độ chính xác (precision) – dịch đúng cụm từ	Độ bao phủ (recall) – giữ được nhiều ý chính
Biến thể phổ biến	BLEU-n (BLEU-1, BLEU-2,...)	ROUGE-N, ROUGE-L
Độ nhạy ngữ nghĩa	Thấp – chỉ dựa vào trùng từ	Trung bình – có xét đến độ bao phủ nội dung
Phổ biến trong	Dịch máy, LLM dịch đa ngôn ngữ	Tóm tắt văn bản, đánh giá output NLP (như ChatGPT)

Comparison

	Bleu	Rouge-L	Rouge-1	Rouge-2
W/o FT- EnviT5	0.22	0.653	0.426	0.623
FT - EnviT5	0.32	0.66	0.44	0.63
Encoder - Decoder	0.09	0.44	0.21	0.39