

BÁO CÁO GIỮA KÌ 21CNTTHUC FIT.HCMUS

PROJECT:
Tinh chỉnh và
Đánh giá toàn diện
một số mô hình
ngôn ngữ lớn
(LLMs) của người
Việt



20127674 – Lê Đức Đạt

MỤC LỤC

1. Giới thiệu
2. Công việc liên quan
3. Phương pháp
4. Demo - Kết quả
5. Tóm lược

GIỚI THIỆU

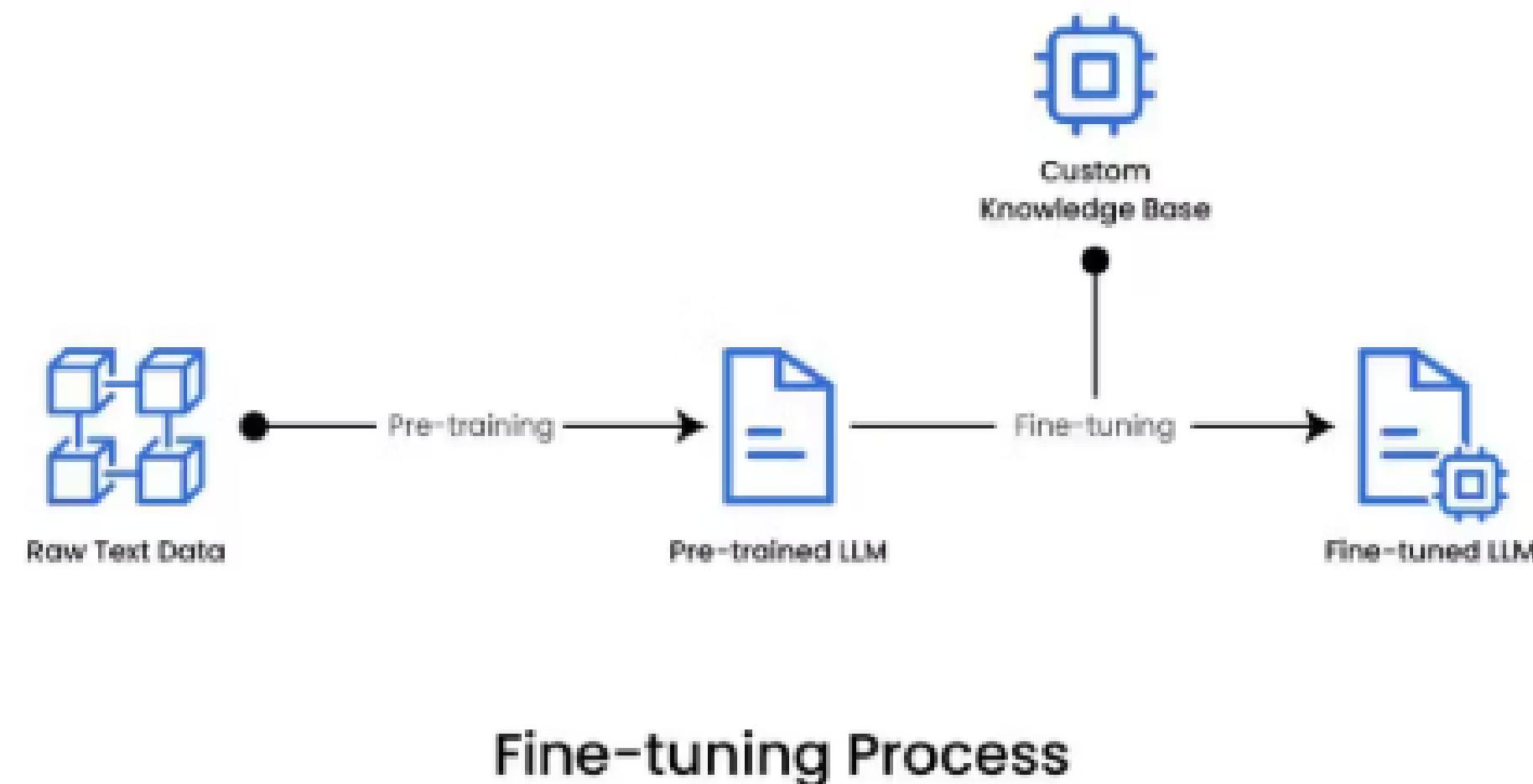


Tinh chỉnh (Finetuning) LLMs

Là một quá trình trong đó mô hình đã được huấn luyện sẵn trên một lượng dữ liệu lớn được "tinh chỉnh" lại để phù hợp hơn với một nhiệm vụ cụ thể hoặc để cải thiện hiệu suất trong một bối cảnh ứng dụng nhất định.

Lý do cần tinh chỉnh mô hình LLM

- Tính chuyên biệt.
- Tiết kiệm tài nguyên.
- Cải thiện hiệu suất.

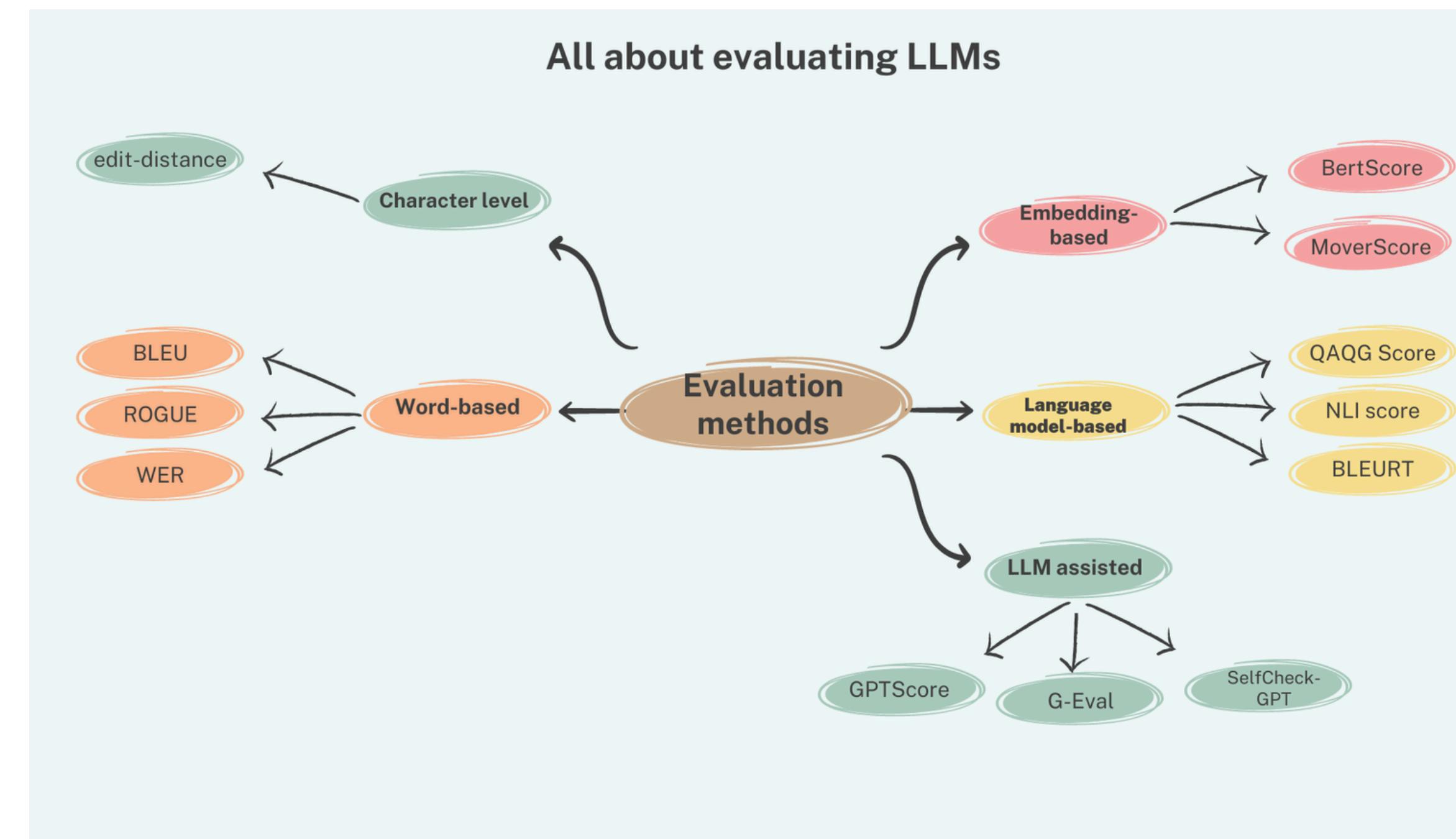


Đánh giá (Evaluate) LLMs

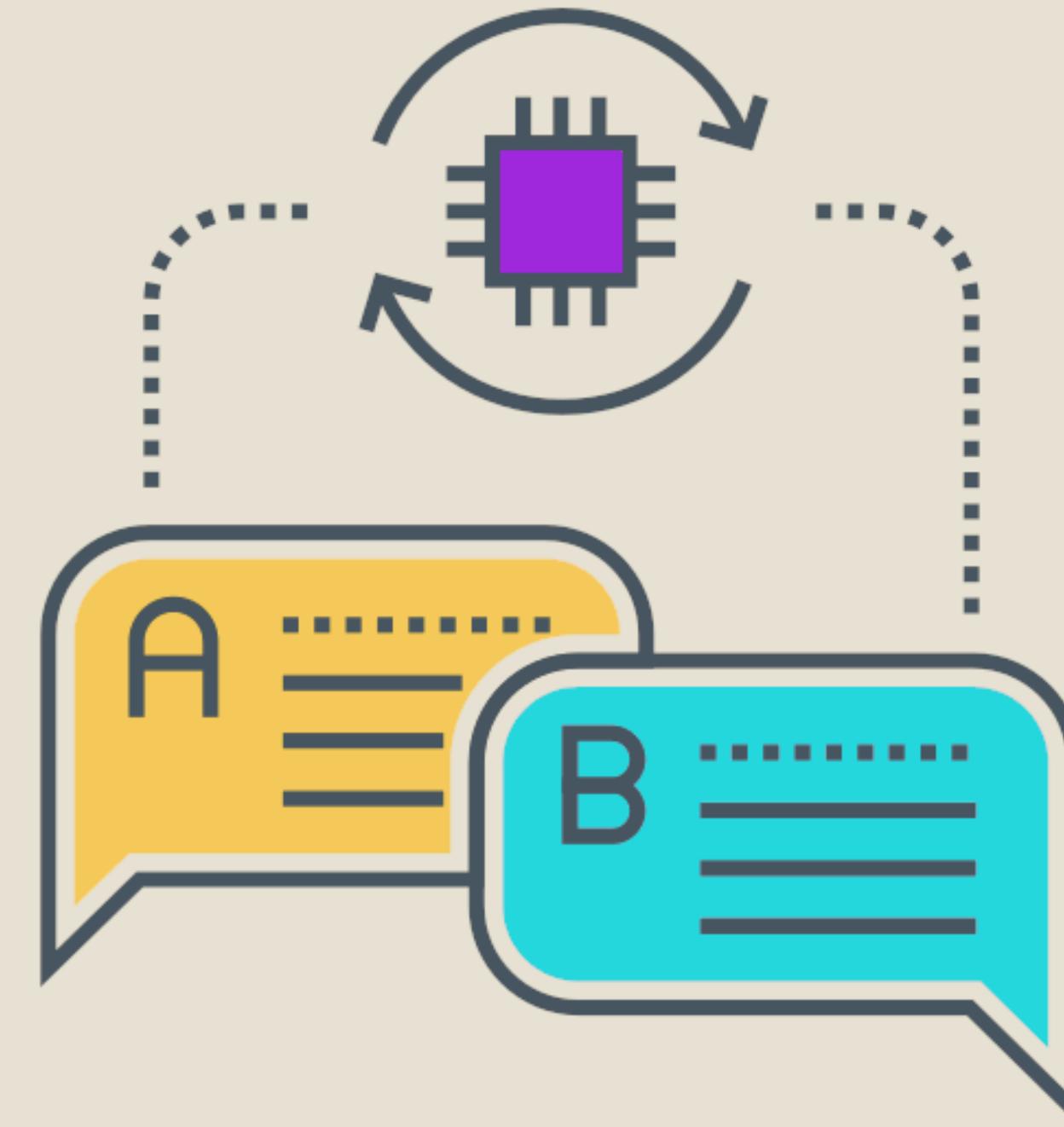
Là quá trình phân tích và đánh giá hiệu suất của mô hình ngôn ngữ lớn trên một loạt các chỉ tiêu và ứng dụng thực tế để xác định mức độ phù hợp và hiệu quả của nó trong các nhiệm vụ ngôn ngữ tự nhiên.

Các tiêu chí đánh giá chính

- Hiệu suất trên các nhiệm vụ cụ thể
- Tính tổng quát và khả năng xử lý các tình huống chưa gặp phải
- Tính linh hoạt và khả năng tương tác
- Hiệu suất trong các môi trường thực tế
- Đánh giá đạo đức và xã hội
- Tính bền vững và chi phí



CÔNG VIỆC LIÊN QUAN



Large language models (LLM)

Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models

Sang T. Truong^{§*} Duc Q. Nguyen^{†*} Toan Nguyen^{†*} Dong D. Le^{†*} Nhi N. Truong^{§†*}
Tho Quan[†] Sanmi Koyejo[§]

[§]Stanford University [†]Ho Chi Minh City University of Technology, VNU-HCM

*Equal contribution, Corresponding: nqduc@hcmut.edu.vn, sttruong@cs.stanford.edu

Các mô hình LLM đã được tinh chỉnh

- URA-LLaMa 7B, 13B, và 70B
- MixSUra
- GemSUra 7B

Những mô hình này được tinh chỉnh bằng cách sử dụng dữ liệu tiếng Việt từ các nguồn như:

- Wikipedia tiếng Việt (Foundation, 2022)
- Vietnamese News-Corpus (Binh, 2021)
- Vietnamese Highschool Essays

Các mô hình LLM đã được đánh giá

- Vietcuna-7B-v3
- Vistral 2
- PhoGPT 7B5 & PhoGPT 7B5Instruct
- Gemini
- GPT-3.5 Turbo & GPT-4

Cách thức tinh chỉnh các mô hình LLM

- QLoRA (Dettmers et al., 2023)
- LoRA (Hu et al., 2022)

Quy trình đánh giá các mô hình LLM

- Độ chính xác (accuracy)
- Tính ổn định (robustness)
- Công bằng và thiên lệch (fairness, bias)
- Tính độc hại (toxicity)

Các tác giả cũng mở rộng khung đánh giá của mình với hai bộ dữ liệu mới:

- Bộ dữ liệu suy luận toán học (MATH, Hendrycks et al., 2021)
- Bộ dữ liệu suy luận tổng hợp (Synthetic reasoning, Wu et al., 2021)

- QA
- Text Summarization
- Sentiment Analysis
- Text Classification
- Knowledge
- Toxicity Detection
- Information Retrieval
- Language Modeling
- Reasoning
- Translation

URA-LLaMa 7B, 13B, và 70B

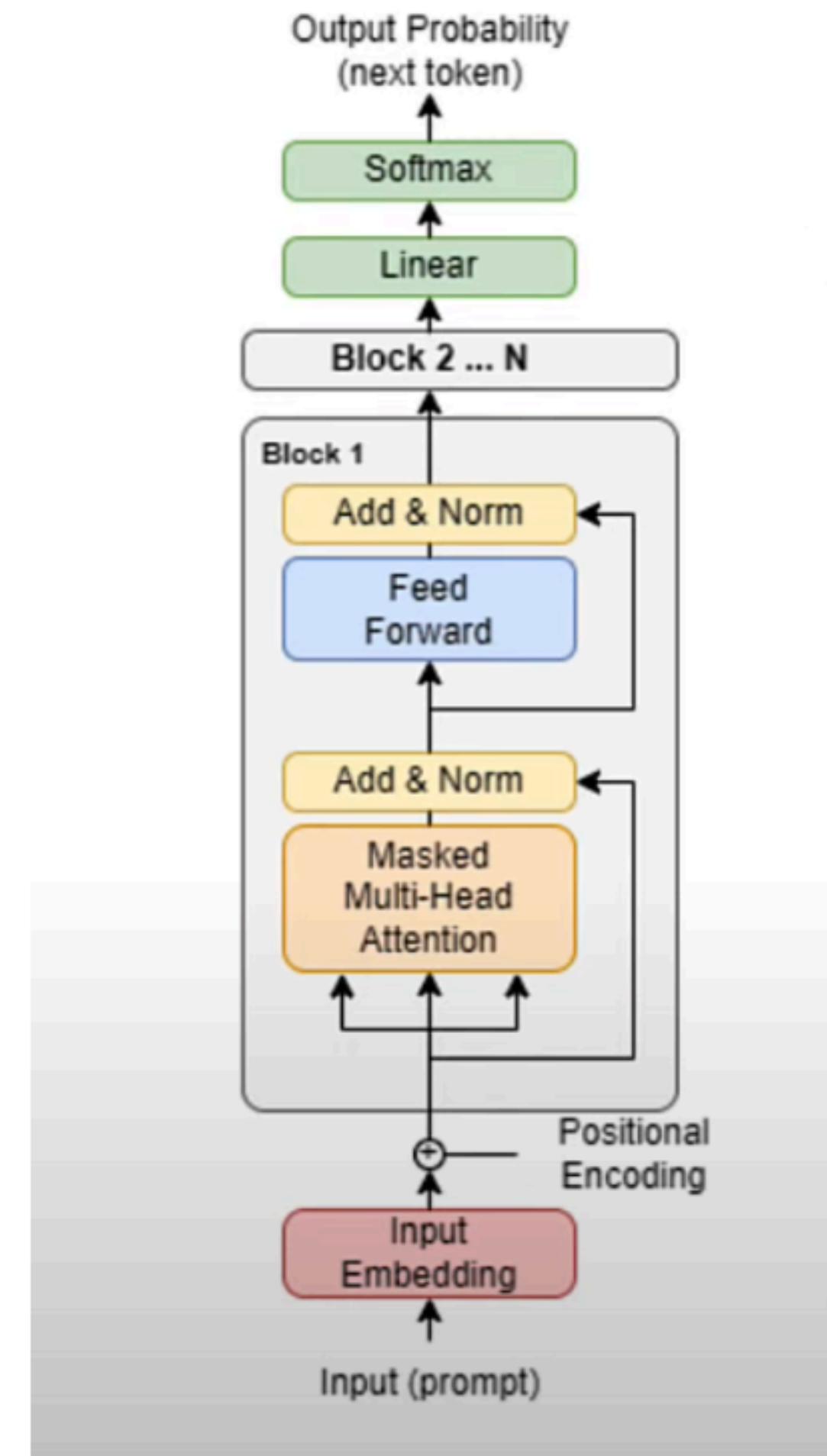
Từ LLaMa-2 và tinh chỉnh đặc biệt cho tiếng Việt.

- Quá trình tinh chỉnh: Không dùng bộ data ngôn ngữ chung mà chỉ dùng bộ tiếng Việt.
- Chức năng - hiệu suất: QA, Tóm tắt, Phân loại và phân tích cảm xúc.

Đặc điểm: Có sự khác biệt về số tham số.

Ưu điểm - Ứng dụng và khả năng sử dụng:

NHIỀU



MixSura

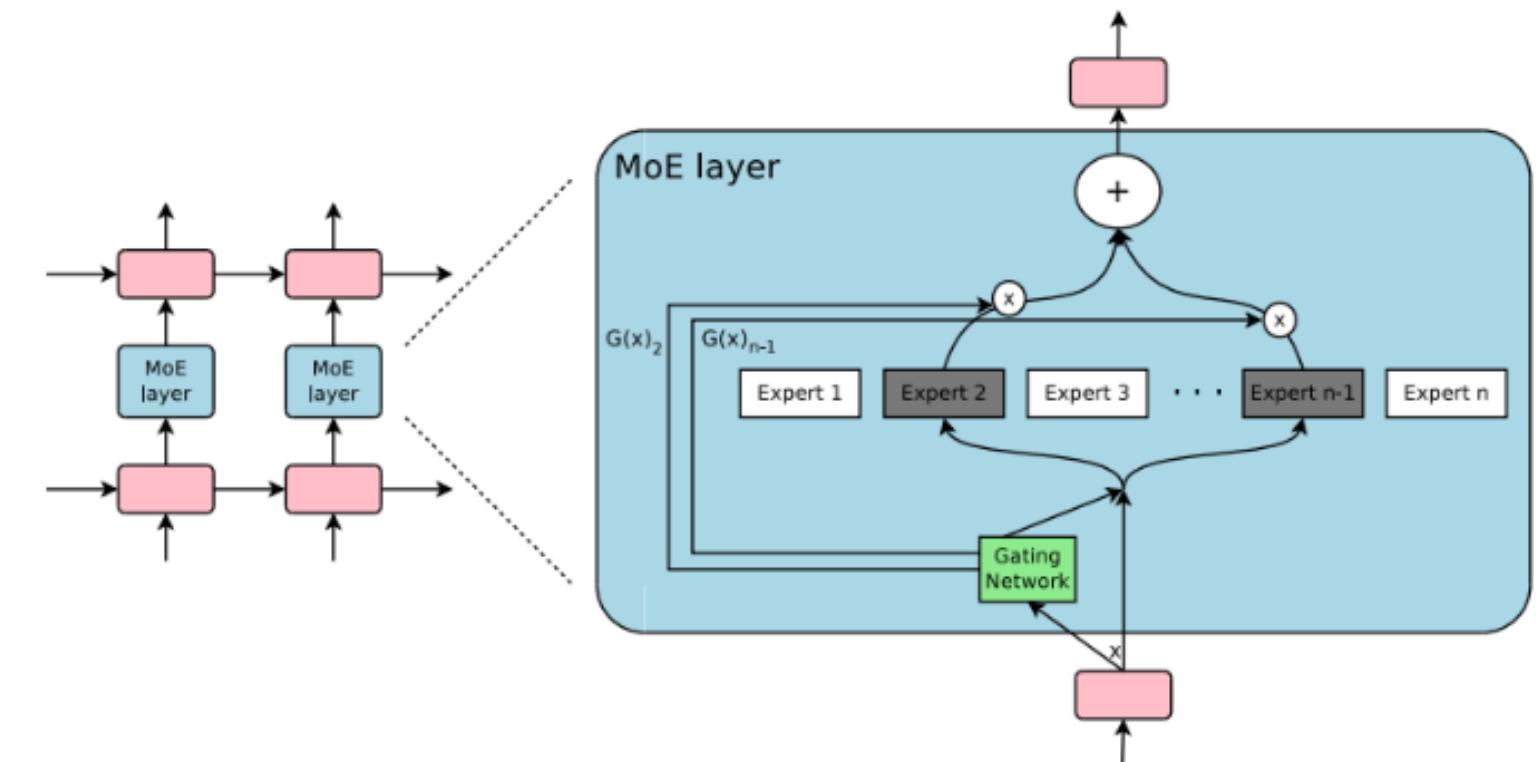
Từ Mixtral 8x7B mà ra. Có cấu trúc hỗn hợp chuyên gia (Mixture of Experts - MoE).

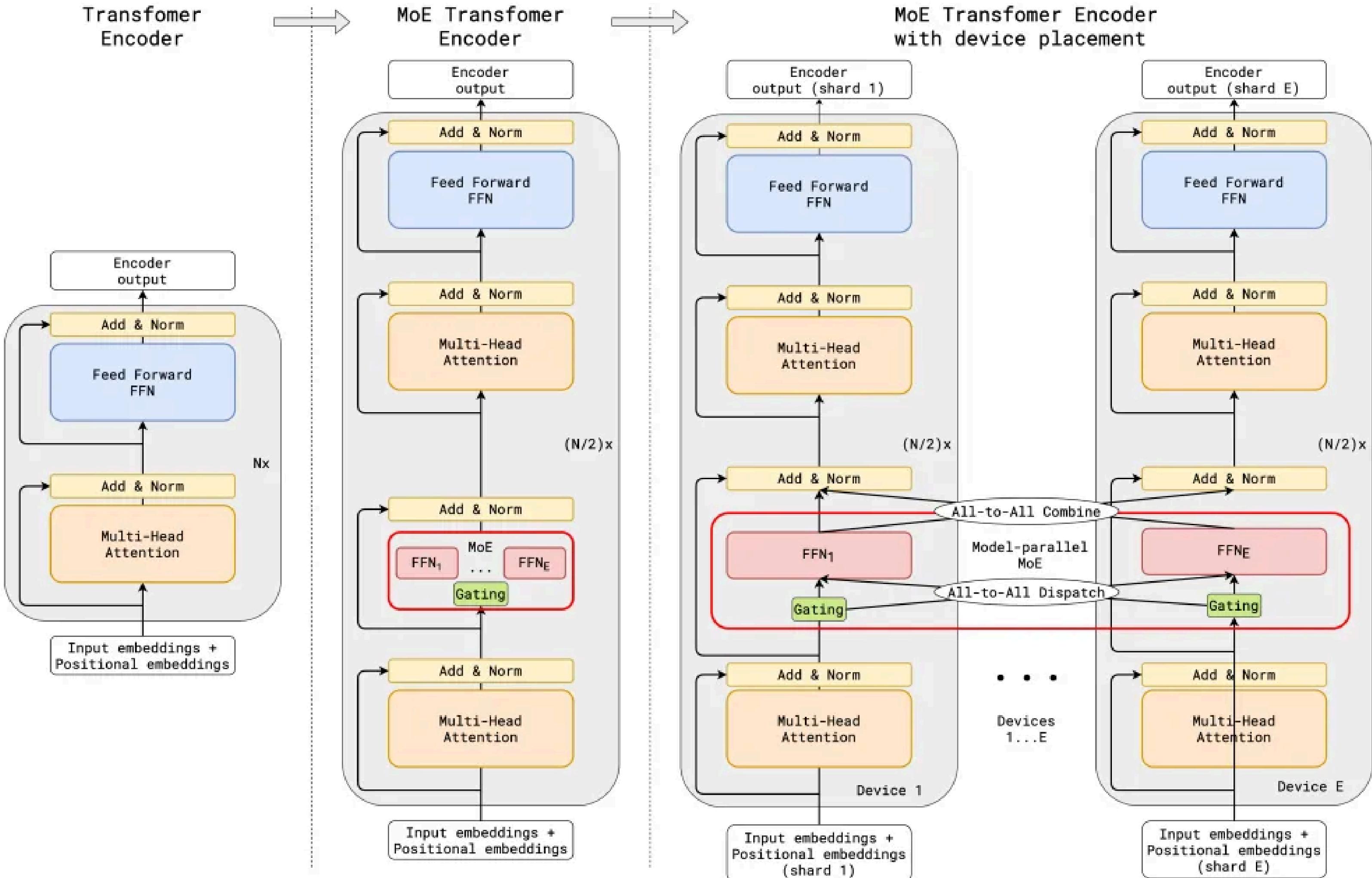
- Hiệu suất cao
- Giảm độ phức tạp tính toán
- Khả năng tổng quát mạnh mẽ

Đặc điểm: Có sự khác biệt về số tham số.

Ưu điểm - Ứng dụng và khả năng sử dụng:

NHIỀU





GemSUrA 7B

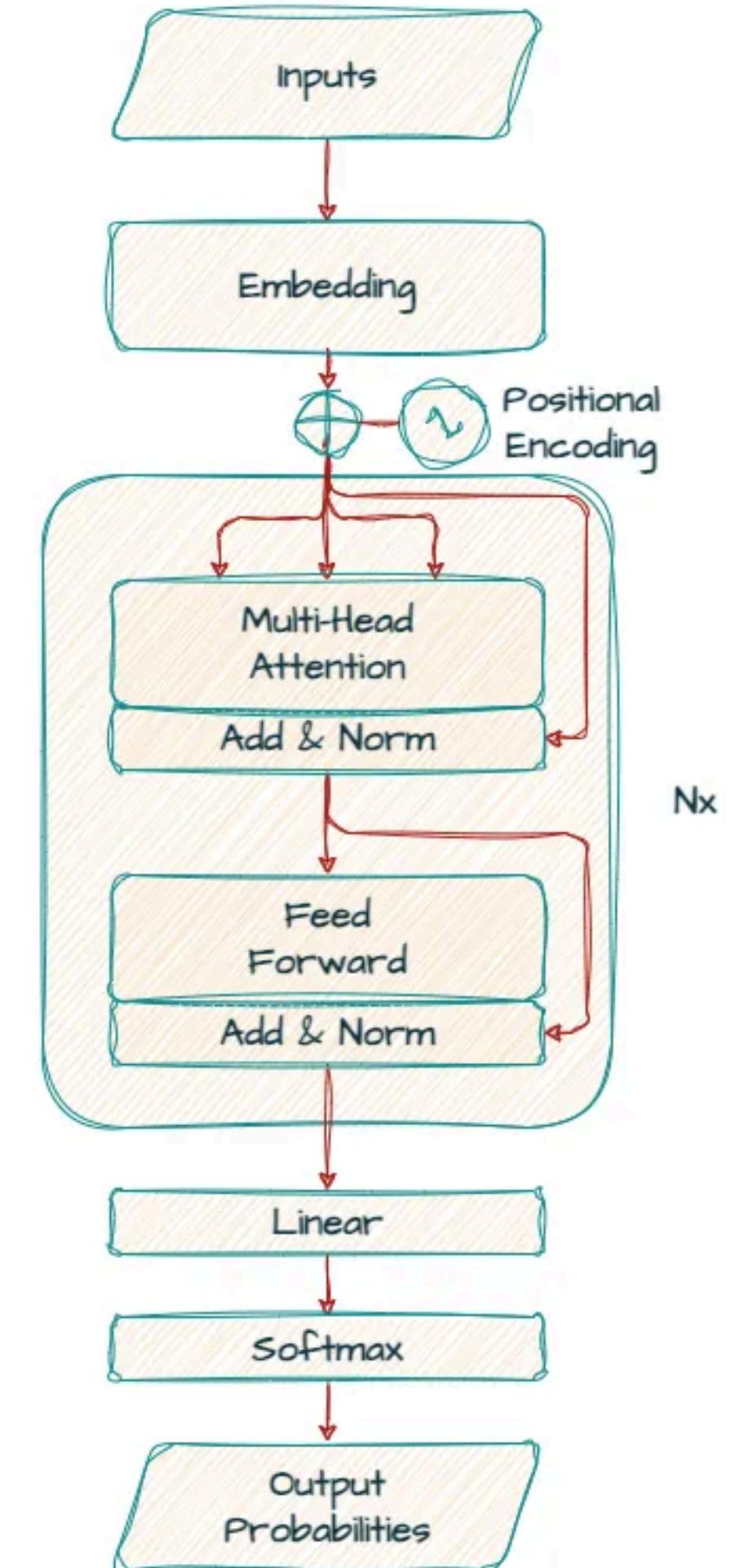
Là một mô hình ngôn ngữ lớn (LLM) được phát triển từ mô hình Gemma 7B (google).

- Dịch thuật
- QA
- Text Generation
- Sentiment Analysis
- Text Summarization
- Text Classification

Đặc điểm: Có sự khác biệt về số tham số.

Ưu điểm - ứng dụng và khả năng sử dụng:

NHIỀU

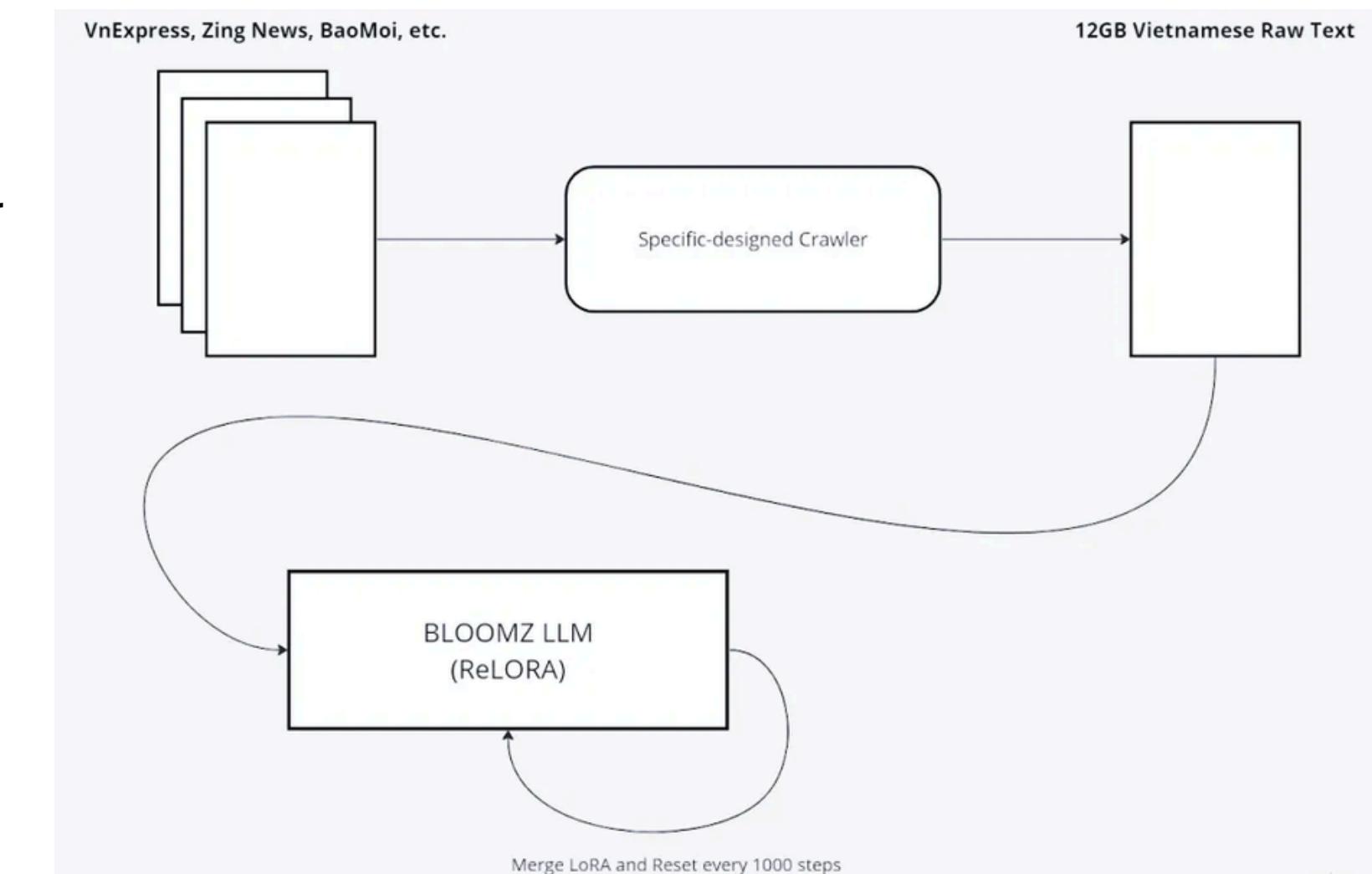
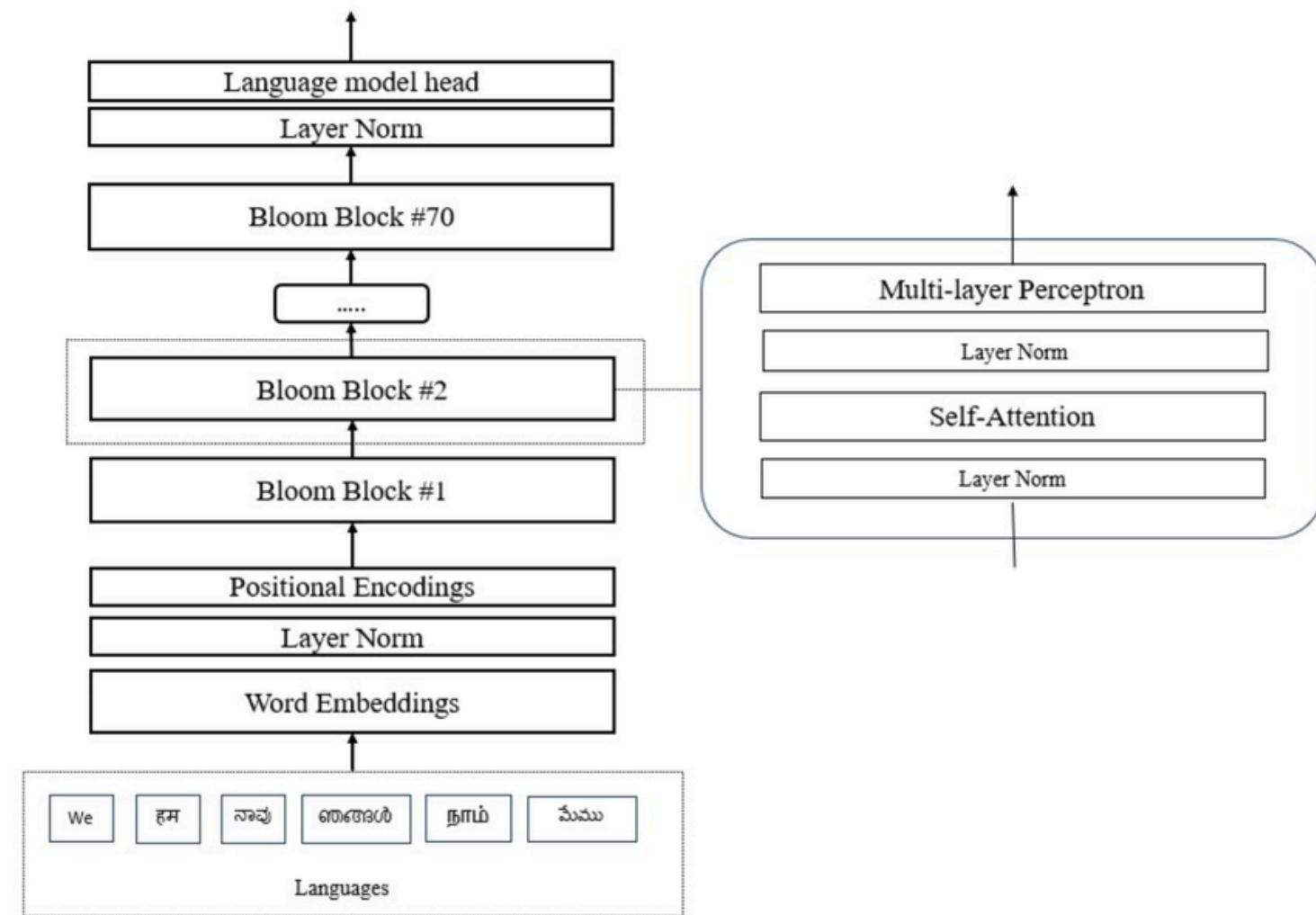


Vietcuna-7B-v3

Là một mô hình ngôn ngữ lớn (LLM) được tinh chỉnh từ mô hình BLOOMZ (BigScience Large Open-science Open-access Multilingual Model).

Tinh chỉnh mô hình BLOOMZ thành Vietcuna-7B-v3: chủ yếu để tối ưu hóa khả năng xử lý tiếng Việt và các ngôn ngữ tương tự.

Ứng dụng - đặc điểm: Tương tự với GemSura

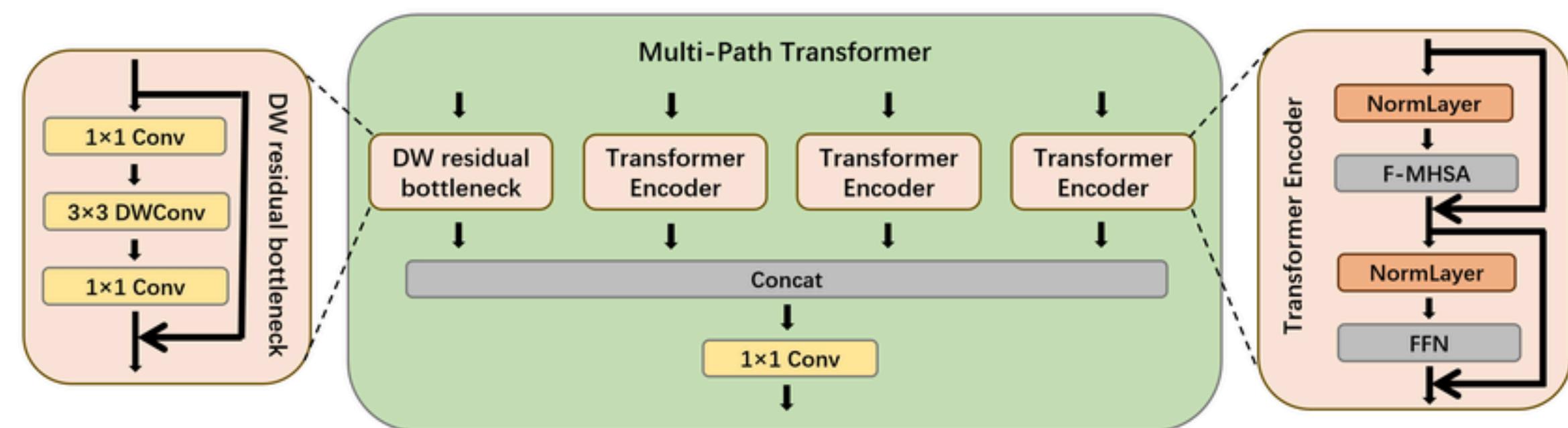


PhoGPT 7B5 & PhoGPT 7B5Instruct

Là một mô hình ngôn ngữ lớn (LLM) được phát triển từ MPT architecture, với 7 tỷ tham số (7B).
Riêng Instruct thì được tinh chỉnh thêm để làm việc với các hướng dẫn (instruction-based tasks)

Ứng dụng:

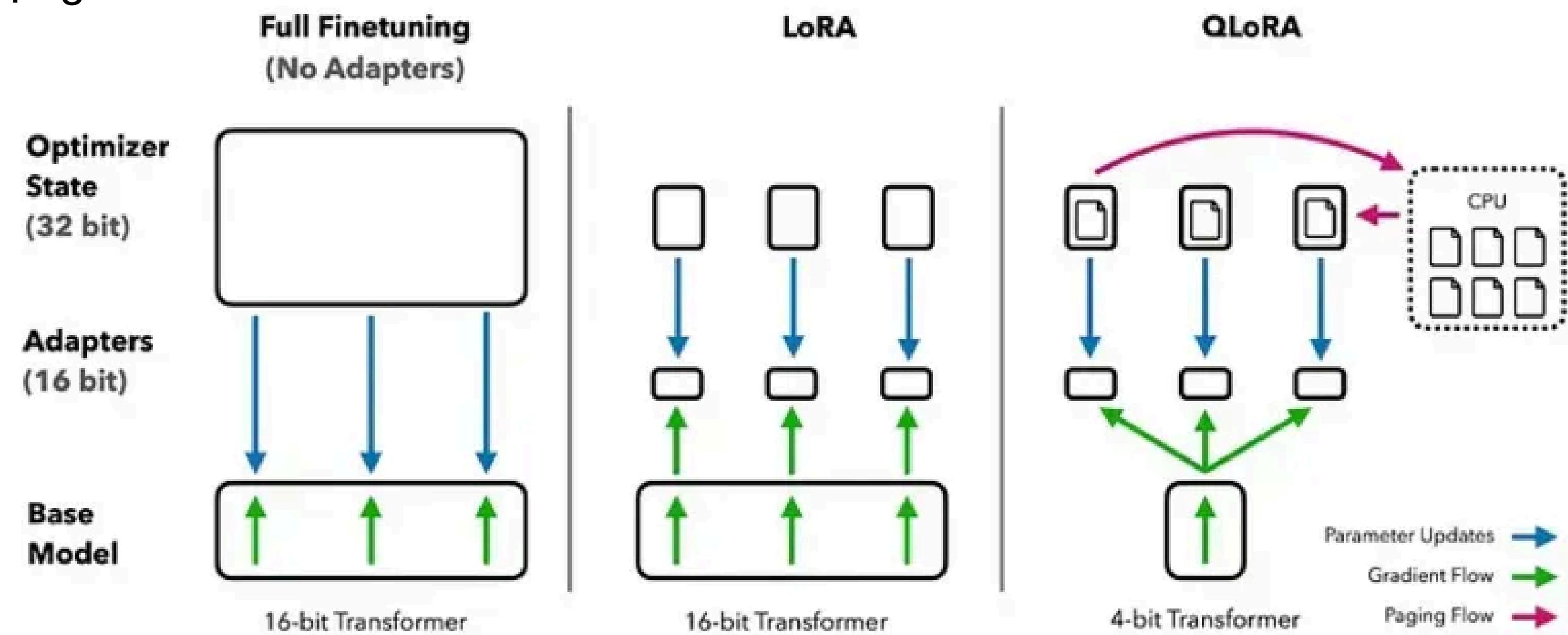
- Tạo nội dung
- Dịch ngôn ngữ
- Trả lời câu hỏi
- Tóm tắt văn bản
- Hỗ trợ chăm sóc khách hàng và trợ lí ảo



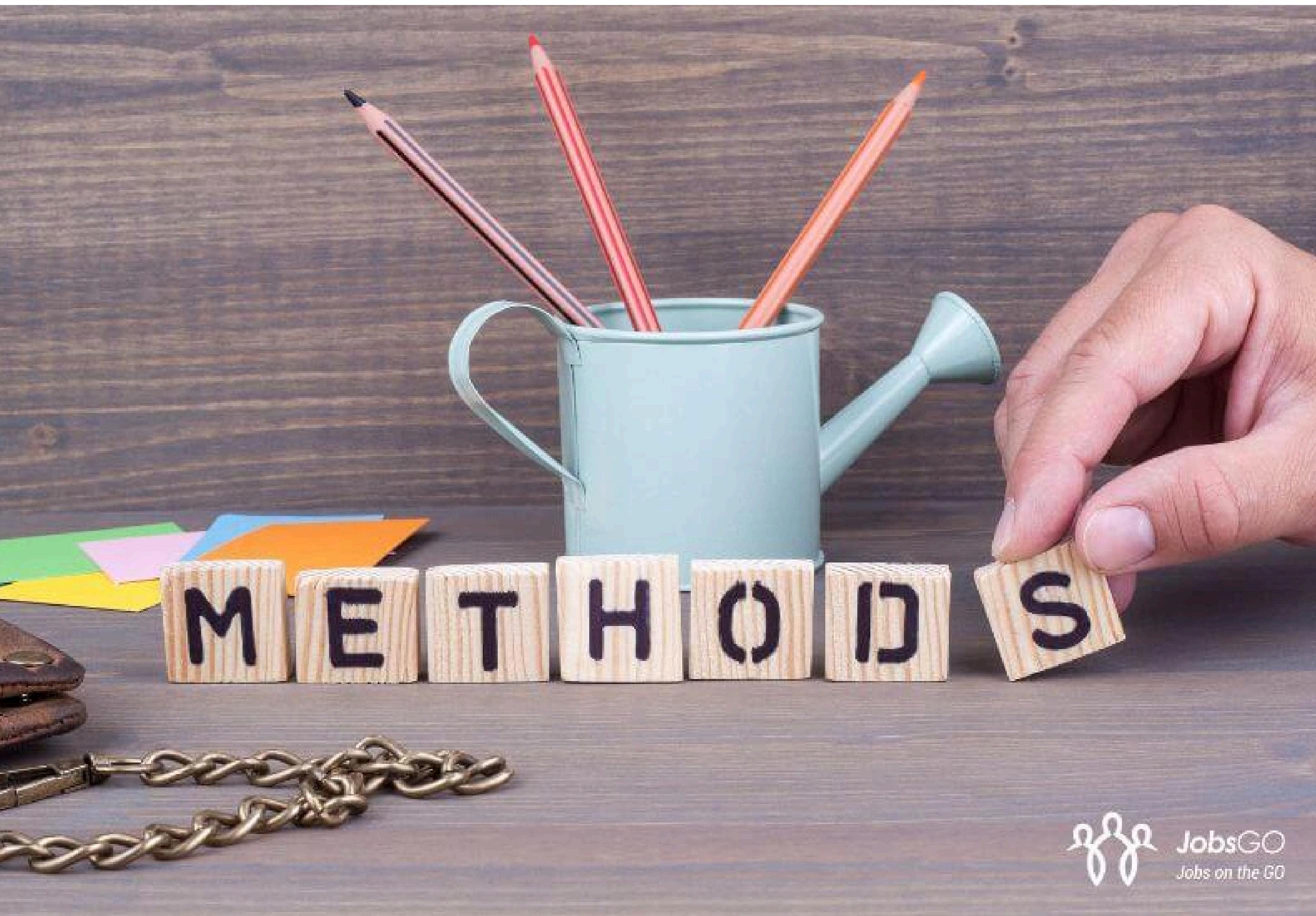
LoRA - QLoRA

LoRA là một phương pháp tối ưu hóa để điều chỉnh mô hình ngôn ngữ lớn mà không cần thay đổi trọng số gốc của mô hình (pre-trained model). Áp dụng một phương pháp decomposition (phân rã) lên các trọng số của mô hình để điều chỉnh chúng bằng cách thêm các ma trận có thứ hạng thấp vào các lớp của mô hình, thay vì điều chỉnh trực tiếp trọng số gốc.

QLoRA là sự kết hợp giữa LoRA và kỹ thuật quantization (lượng tử hóa), không chỉ áp dụng ma trận thứ hạng thấp vào mô hình như LoRA, mà còn lượng tử hóa các tham số đó.



PHƯƠNG PHÁP



Supervised Finetuning

Finetuning trên dữ liệu có nhãn: Quá trình finetuning trong đoạn văn diễn ra trên hai bộ dữ liệu có nhãn (labeled data):

- Vietnamese Wikipedia (1GB)
- Vietnamese News-Corpus (22GB)

Mô hình đã được pre-trained:
LLaMa-2, Gemma 7B, Mixtral 8x 7B.

Các tham số trong quá trình finetuning: LoRA rank, learning rate, max length, dropout rate, epochs, và các tham số khác như quantization, LoRA α, v.v.

- LoRA rank được thiết lập lần lượt là 128 cho mô hình 7B, 256 cho mô hình 13B, và 1024 cho mô hình 70B.
- Việc sử dụng QLoRA kết hợp với 4-bit NF4 quantization cho phép giảm bớt yêu cầu bộ nhớ và thời gian tính toán, đặc biệt là với các mô hình kích thước lớn.
- LoRA rank được sử dụng trong các mô hình này với các giá trị cụ thể cho từng kích thước mô hình:
 - 7B model: LoRA rank = 128
 - 13B model: LoRA rank = 256
 - 70B model: LoRA rank = 1024
- LoRA α (alpha) và dropout rate được điều chỉnh trong các mô hình, với các giá trị chung là $\alpha = 16$ và dropout rate = 0.1. Những tham số này giúp điều chỉnh tốc độ học và khả năng khái quát của mô hình.

Bảng tóm tắt các hyperparameters và cấu hình của các mô hình:

Model	LoRA	α	Dropout	Learning	Max	Epochs	Hardware	CO2
	Rank	(Alpha)		Rate	Length			Emission (kg)
URA- LLaMa (7B)	128	16	0.1	1×10^{-5}	2048	1	$6 \times$ A100 80GB	900
URA- LLaMa (13B)	256	16	0.1	1×10^{-5}	2048	1	$6 \times$ A100 80GB	900
URA- LLaMa (70B)	1024	16	0.1	1×10^{-5}	2048	1	$6 \times$ A100 80GB	900
GemSUra (7B)	256	512	0.1	1×10^{-5}	8192	2	$4 \times$ A100 80GB	200
MixSUra (7B)	256	512	0.1	5×10^{-5}	32768	5	$4 \times$ A100 80GB	200

Step 1

Collect demonstration data, and train a supervised policy.

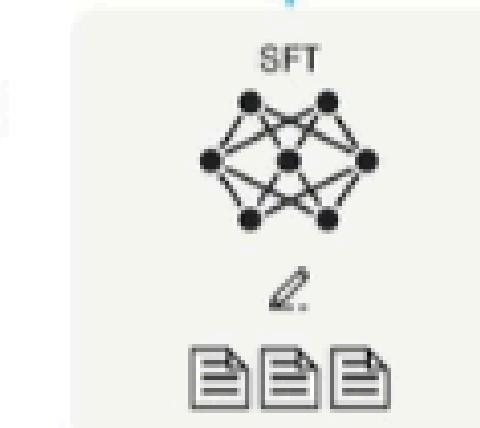
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



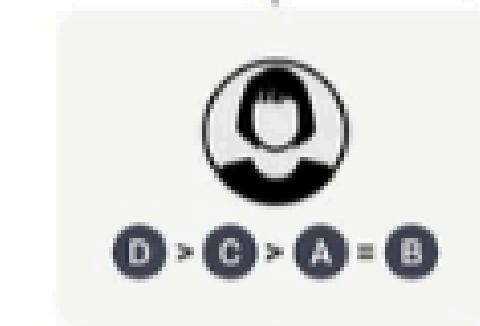
Step 2

Collect comparison data, and train a reward model.

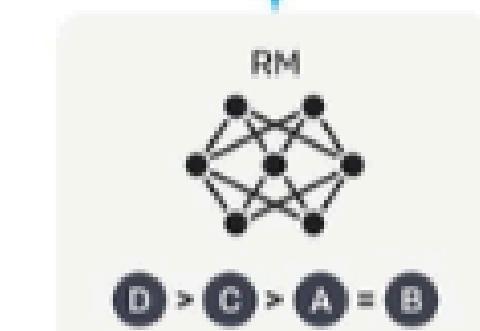
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



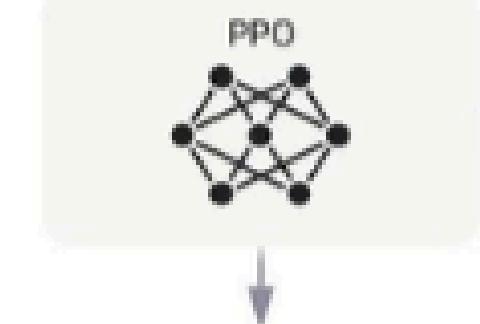
Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



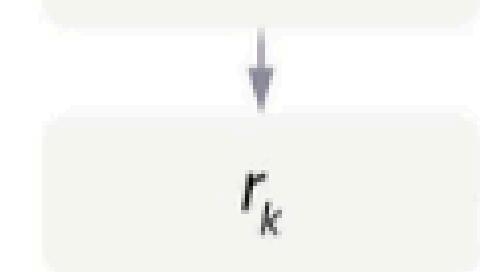
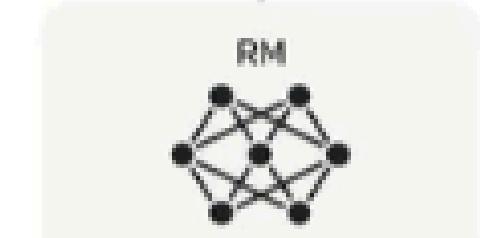
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



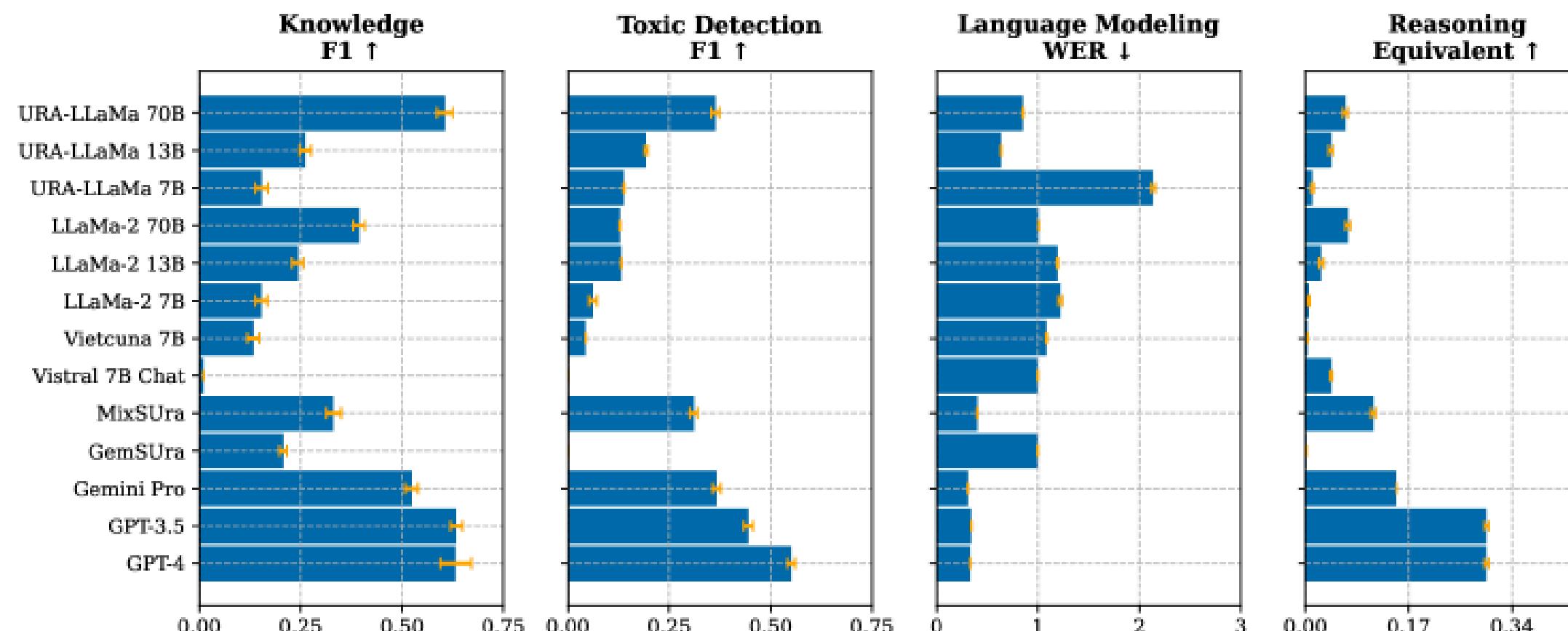
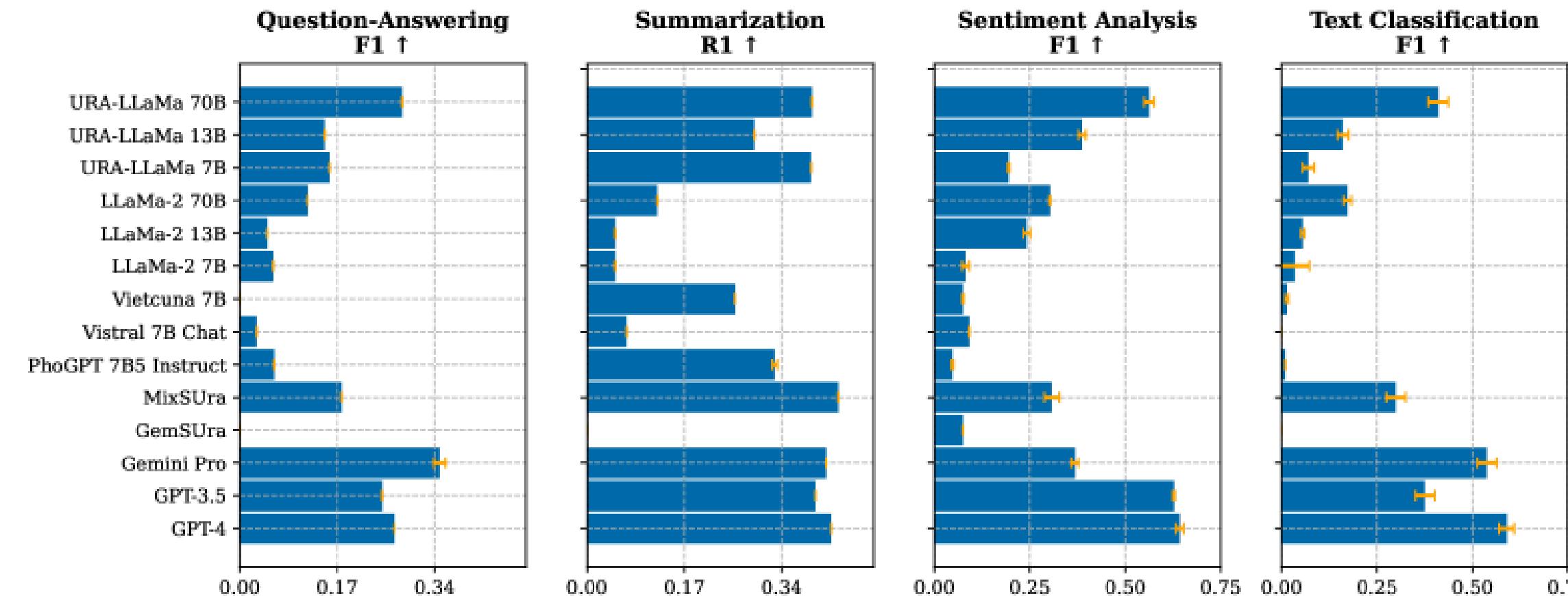
Thiết kế prompt cho các mô hình học sâu ngôn ngữ lớn (LLMs - Large Language Models) trong các tình huống khác nhau.

- Base prompt
- Tác Động Của Các Prompt Được Tối Ưu Hóa
- Prompt Cung Cấp Ví Dụ (Few-shot Learning/In-context Learning)
- Prompt Yếu (Weak Prompts)
- Prompt Với Các Ràng Buộc (Constraints)
- Mục Tiêu Của Việc Thiết Kế Prompt

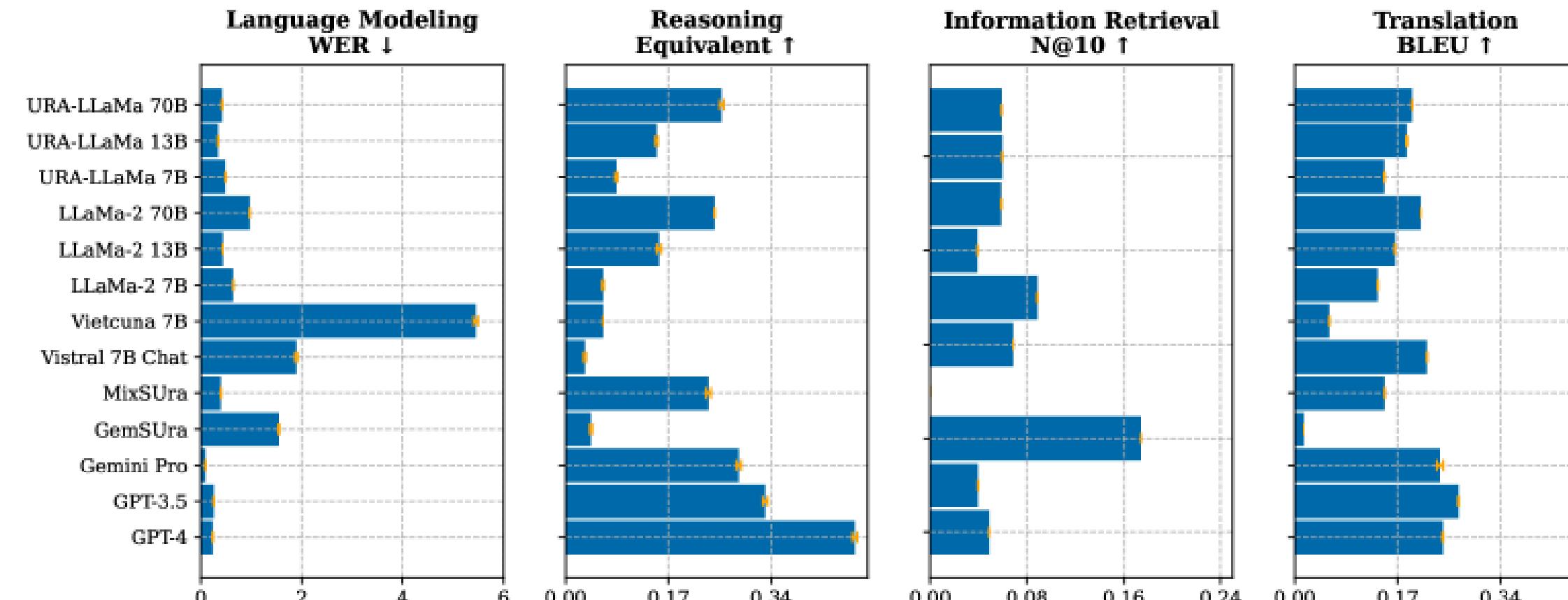
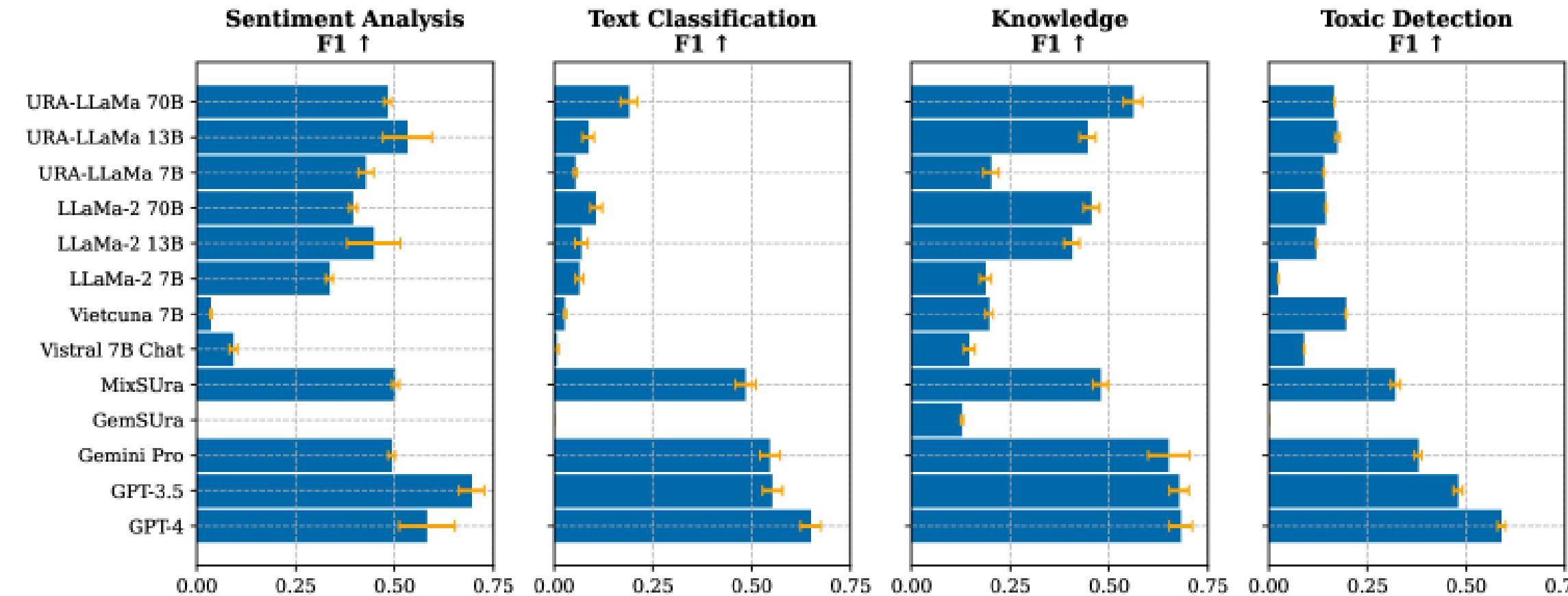
Kết quả



Hiệu suất trên zero-shot prompt



Hiệu suất trên few-shot prompt



Tóm lược - mục tiêu cần làm



Tóm lược

- LLMs (Kiến trúc, chức năng,...).
- Supervised Finetuning.
- Phương pháp
- Các tiêu chí đánh giá.
- Phân tích kết quả (zero-shot, few-shot).

Mục tiêu cần làm

- Chọn một số LLMs để làm dựa vào những kiến thức trên.
- Chạy kết quả zero-shot (ưu tiên) và few-shot.
- Chạy và tinh chỉnh trên Google Colab hoặc môi trường khác như Kaggle để phù hợp với tình hình (có thể data sẽ tiếp tục pretrained, finetuning, pruning, quantization, distributed training, sampling, preprocessing, LoRA nhưng trong bài,...)

THANK'S FOR WATCHING

Contact: dat20026969@gmail.com

Sống cho hết đời thanh xuân để trọn
vẹn hết những thứ mà ta cần.



fit@hcmus

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM
KHOA CÔNG NGHỆ THÔNG TIN

