ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
Khoa Công Nghệ Thông Tin

Môn: Học sâu cho Khoa học Dữ liệu

Báo cáo đồ án thực hành: Zalo AI Challenge – Elementary Maths Solving

Thành viên nhóm:

20120040 – Nguyễn Quang Gia Bảo

20120088 – Lê Nguyễn Thanh Hoàng

20120136 – Huỳnh Tuấn Nam

20120158 – Trần Hoàng Anh Phi

Giảng viên hướng dẫn

Thầy Nguyễn Tiến Huy

Thành phố Hồ Chí Minh, ngày 16 tháng 01 năm 2024

Mục lục

1. Dữ liệu

Large Language Models (LLM) hiện đang là xu hướng và có tiềm năng đáng kể cho các ứng dụng khác nhau trong cuộc sống hàng ngày của chúng ta. Tuy nhiên, một trong những thách thức mà các mô hình này và các hệ thống AI khác phải đối mặt là thực hiện các nhiệm vụ toán học và suy luận một cách hiệu quả.

Thử thách năm nay tập trung phát triển mô hình/hệ thống ngôn ngữ có khả năng trả lời *các câu hỏi toán cấp tiểu học* phù hợp với Chương trình Giáo dục Việt Nam.

Dưới đây là đầu vào, đầu ra và một số ví dụ về nhiệm vụ.

- Input: Một câu hỏi toán có nhiều lựa chọn, có 4 phương án lựa chọn trong đó chỉ có một phương án đúng.
- Output: Lựa chọn đúng cho câu hỏi toán đã cho.

Ví dụ:

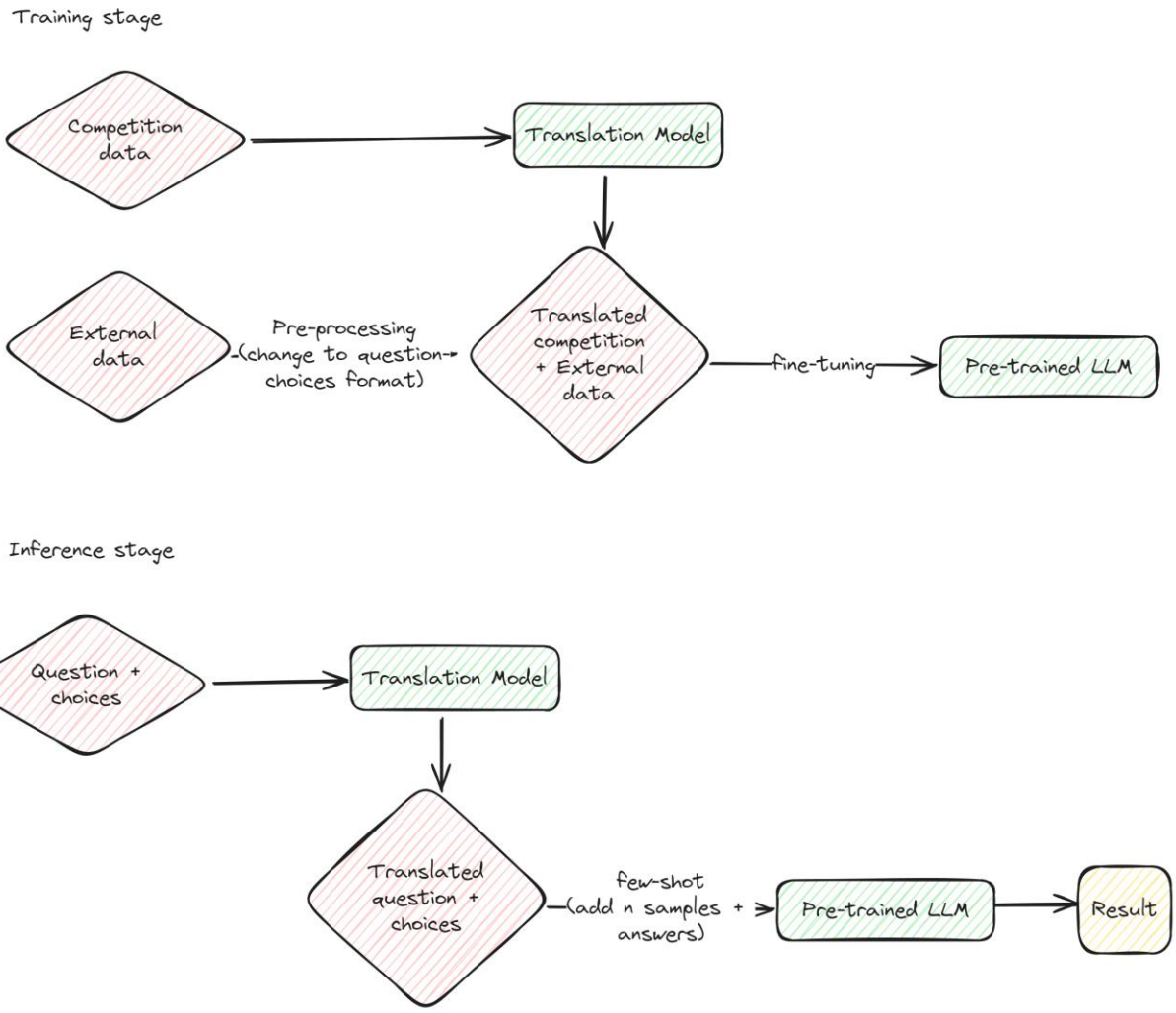
Input	Output
Số "bảy triệu hai trăm nghìn" có: A. Ba chữ số 0 B. Bốn chữ số 0 C. Năm chữ số 0 D. Sau chữ số 0	C. Năm chữ số 0
Mẫu số của phân số thập phân có thể là những số nào? A. Các số chẵn B. Các số 10, 100, 1000, … C. Các số lẻ D. Mọi số tự nhiên khác 0	B. Các số 10, 100, 1000, …

2. Giải quyết bài toán

- Để giải quyết bài toán này, nhóm đề xuất ra 2 giải pháp chính, kết hợp sử dụng các pretrain LLM và external data:
    - **Solution 1: Gồm có 2 stages:**
        - Sử dụng một mô hình dịch máy (Neural Machine Translation Model) dịch bộ dữ liệu sang tiếng Anh (VinAI-translation), kết hợp với external data từ các nguồn như HuggingFace, các bộ dữ liệu về toán bằng tiếng Anh, từ đó tận dụng sức mạnh của các

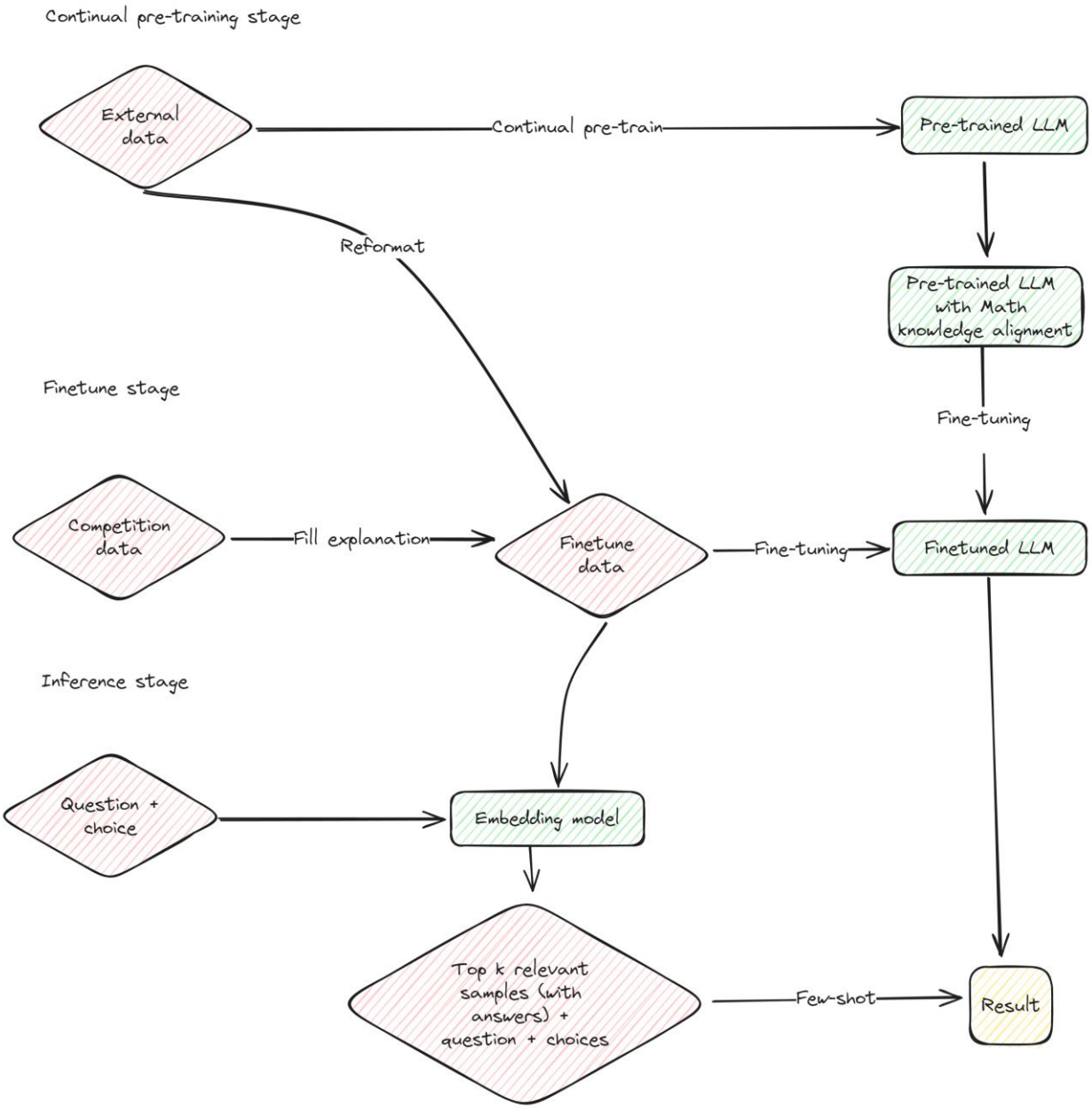
pre-trained LLM (được hỗ trợ tốt hơn các pretrain bằng tiếng Việt).

- Fine-tune từ các mô hình pre-trained này, sử dụng kết hợp các kĩ thuật quantization để giảm kích thước mô hình.



- **Solution 2: Gồm có 3 stages:**
  - Continual pre-training trên một mô hình LLM đã được training trước đó với external data về toán tiểu học, giúp mô hình "hiểu" được các kiến thức mới này.
  - Fine-tune mô hình với tập dữ liệu train từ ban tổ chức kết hợp với một số external data (đã làm sạch và điền thêm các giải thích cho mỗi câu hỏi).
  - Ở giai đoạn inference, sử dụng một mô hình embedding để tìm ra top k các mẫu dữ liệu liên quan tới câu hỏi và lựa chọn, sau

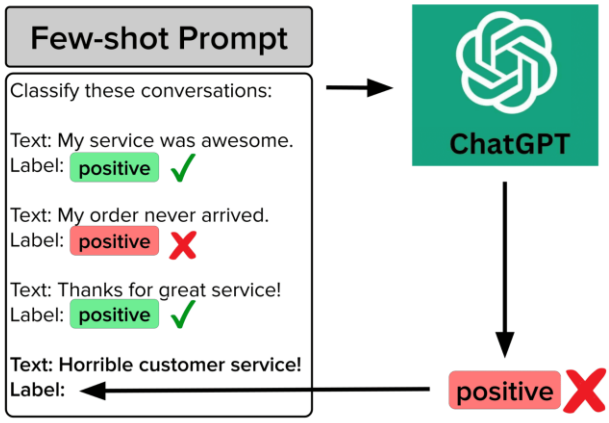
đó sử dụng các mẫu dữ liệu đó làm đầu vào để thực hiện few-shot.



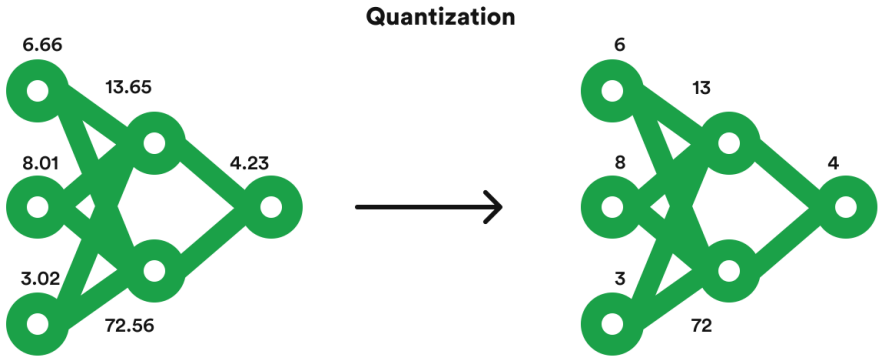
3. Một số kỹ thuật

3.1 Few-shot learning

- Là một kĩ thuật giúp mô hình hiểu và giải quyết được các nhiệm vụ mới (task) chỉ với một số lượng nhỏ dữ liệu huấn luyện. Thay vì sử dụng lượng lớn dữ liệu, few-shot learning có thể học được từ các ví dụ và áp dụng kiến thức đó cho các tác vụ mới.



3.2 Quantization

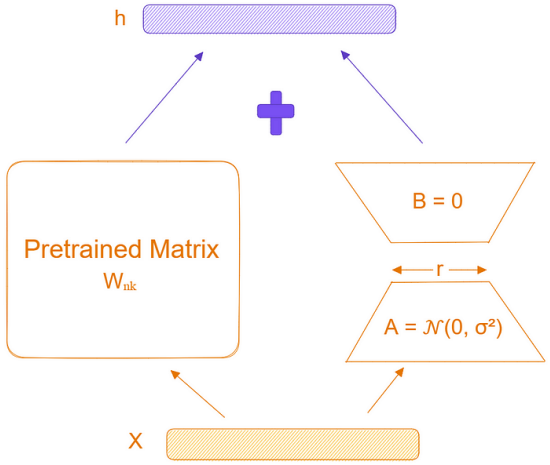


- Là một kĩ thuật nhằm giảm kích thước mô hình, bằng cách đưa biểu diễn các tham số từ miền rộng về miền hẹp hơn, ví dụ: từ float32 về float16, hoặc kết hợp nhiều kiểu biểu diễn lại với nhau.
- Phương pháp này sẽ giúp mô hình tăng tốc độ inference, giảm dung lượng bộ nhớ, tuy nhiên hiệu suất có thể bị ảnh hưởng (do các phép làm tròn, truncate khi chuyển các số về miền biểu diễn nhỏ hơn).

3.3 LoRA: Low-Rank Adaption

- Thông thường khi fine-tuning các mô hình, ta sẽ phải train toàn bộ hoặc một số layers của mô hình, và ta cũng phải lưu lại toàn bộ các tham số của mô hình hoặc một số layers đó.

- **LoRA** là kĩ thuật đóng băng bộ tham số của mô hình huấn luyện trước và thực hiện chèn các Adapters vào giữa các lớp của mô hình, các Adapters này sử dụng kĩ thuật phân rã ma trận (matrix decomposition) nhằm giảm kích thước và tăng tốc quá trình fine-tuning.



4. Xếp hạng trên Public Leaderboard của Zalo AI Challenge

Tên nhóm: **[HCMUS-FIT]-HOPELESS**

- Với solution 1, nhóm đạt được độ chính xác trên public test là: **0.50802**
- Với solution 2, nhóm đạt được độ chính xác trên public test là: **0.59893**

Kết quả cuối cùng trên leaderboard:

75	▲ 2	[HCMUS-FIT]-HOPELESS	Huỳnh Tuấn Nam, Nguyễn Quang Gia Bảo, Lê Nguyễn Thanh Hoàng, Trần Hoàng Anh Phi	0.59893	16/01/2024 19:39

Github repo: https://github.com/Foxxy-HCMUS/zalo-elementary-maths-solving

5. Tài liệu tham khảo

- https://github.com/Reasoning-Lab/Elementary-Math-Solving-Zalo-AI-2023/tree/main
- pre-trained:
  - https://github.com/VinAIResearch/VinAI_Translate
  - https://huggingface.co/microsoft/deberta-v3-large
  - https://huggingface.co/gpt2
  - https://github.com/VinAIResearch/PhoBERT
  - https://github.com/paperswithcode/galai
  - https://github.com/neurasearch/neurips-2022-submission-3358
  - https://github.com/zwq2018/multi-view-consistency-for-mwp

    - https://github.com/facebookresearch/llama
    - https://github.com/meta-math/MetaMath
    - https://huggingface.co/vilm/vietcuna-7b-v3
    - https://huggingface.co/roberta-large
    - https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca
    - https://huggingface.co/mistralai/Mistral-7B-v0.1
    - https://huggingface.co/EleutherAI/llemma_7b
    - https://huggingface.co/Intel/neural-chat-7b-v3-1
    - https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

- additional data:
    - https://paperswithcode.com/dataset/mmlu
    - https://math-qa.github.io/math-QA/
    - https://github.com/hendrycks/math
    - https://github.com/arkilpatel/SVAMP
    - https://github.com/2003pro/Graph2Tree
    - https://huggingface.co/datasets/math_dataset
    - https://huggingface.co/datasets/meta-math/MetaMathQA
    - https://huggingface.co/datasets/TIGER-Lab/MathInstruct
    - https://khoahoc.vietjack.com/
    - https://tech12h.com/