Dịch máy lĩnh vực Y tế (VLSP 2025)

Phan Tất An
23020003@vnu.edu.vn

Trần Văn Quyết
23020143@vnu.edu.vn

Nguyễn Văn Cử
23020015@vnu.edu.vn

Phạm Huy Châu Long
23021615@vnu.edu.vn

Tóm tắt nội dung

Trong những năm gần đây, sự phát triển mạnh mẽ của các mô hình ngôn ngữ lớn (Large Language Models – LLMs) đã thúc đẩy đáng kể hiệu quả của các hệ thống dịch máy. Tuy nhiên, các mô hình dịch tổng quát thường gặp nhiều hạn chế khi áp dụng vào những lĩnh vực chuyên biệt như y tế, nơi yêu cầu độ chính xác cao và tính nhất quán thuật ngữ. Báo cáo này trình bày hệ thống dịch máy song ngữ Anh–Việt trong lĩnh vực y tế tham gia Shared Task VLSP 2025, dựa trên mô hình Qwen và kỹ thuật tinh chỉnh hiệu quả tham số.

1 Giới thiệu đề tài

1.1 Bối cảnh và động lực

Trong những năm gần đây, sự phát triển mạnh mẽ của các mô hình ngôn ngữ lớn (Large Language Models – LLMs) đã thúc đẩy đáng kể hiệu quả của các hệ thống dịch máy (Machine Translation – MT). Tuy nhiên, các mô hình dịch tổng quát thường gặp nhiều hạn chế khi áp dụng vào những lĩnh vực chuyên biệt như y tế (medical domain), nơi yêu cầu độ chính xác cao, tính nhất quán ngữ nghĩa và sự tuân thủ thuật ngữ chuyên ngành.

Dịch máy trong lĩnh vực y tế không chỉ đơn thuần là chuyển ngữ mà còn đòi hỏi khả năng hiểu sâu ngữ cảnh, cấu trúc câu đặc thù và các thuật ngữ lâm sàng – nghiên cứu. Một lỗi dịch nhỏ cũng có thể ảnh hưởng tới ý nghĩa chuyên môn, thậm chí gây hiểu nhầm trong môi trường y tế.

Vì vậy, việc phát triển các mô hình dịch chuyên biệt cho lĩnh vực y khoa là một nhu cầu cấp thiết và có ý nghĩa thực tiễn lớn. Shared Task "*Medical Domain MT with Limited-Pretraining Models*" thuộc Hội thảo xử lý ngôn ngữ tự nhiên VLSP 2025 được tổ chức nhằm thúc đẩy nghiên cứu và ứng dụng dịch máy trong lĩnh vực này, trên cơ sở đảm bảo tính minh bạch và công bằng thông qua cơ chế ràng buộc (Constrained Track).

1.2 Mục tiêu nghiên cứu

Mục tiêu của đề tài là xây dựng một hệ thống dịch máy chất lượng cao cho văn bản y tế song ngữ Anh – Việt, tuân thủ ràng buộc của bài toán, dựa trên:

- Tối ưu mô hình pre-training giới hạn (Limited-Pretraining Models)
- Sử dụng duy nhất dữ liệu do ban tổ chức cung cấp
- Tận dụng mô hình nền Qwen làm Base LLM, kết hợp fine-tuning theo hướng SFT
- Đánh giá chất lượng dịch bằng BLEU và Gemini theo chuẩn VLSP

Đề tài hướng đến xây dựng một pipeline đầy đủ bao gồm xử lý dữ liệu, thiết kế prompt, huấn luyện SFT với LoRA, đánh giá theo BLEU callback và phân tích lỗi dịch.

1.3 Phạm vi và giới hạn

Phạm vi và giới hạn của đề tài bao gồm việc huấn luyện mô hình Qwen theo chuẩn Constrained Track, không sử dụng dữ liệu ngoài và không áp dụng các kỹ thuật hậu huấn luyện như RLHF.

2 Cơ sở lý thuyết

Kiến trúc và Đặc điểm Chính

- **Kiến trúc Cốt lõi**: Các biến thể Qwen (ví dụ: Qwen-1.5B, Qwen-7B, Qwen-14B) thường được xây dựng trên kiến trúc *Decoder-only Transformer*. Kiến trúc này đặc biệt phù hợp cho các tác vụ tạo sinh văn bản (text generation), và trong bài toán dịch máy, mô hình sẽ xử lý cả câu nguồn và câu đích trong cùng một chuỗi đầu vào có điều kiện.
- **So sánh với NMT truyền thống**: Không giống Encoder–Decoder trong Transformer chuẩn, Qwen xem dịch máy như một bài toán

sinh chuỗi có điều kiện. Câu nguồn và câu đích được phân tách bằng token đặc biệt, cho phép mô hình học trực tiếp mối quan hệ giữa hai ngôn ngữ.

- **Tiền huấn luyện (Pre-training)**: Các mô hình Qwen được huấn luyện trên tập dữ liệu rất lớn và đa ngôn ngữ, bao gồm tiếng Anh, tiếng Việt và nhiều ngôn ngữ khác. Nhờ vậy, mô hình đạt được nền tảng ngôn ngữ mạnh mẽ về cú pháp, ngữ pháp và kiến thức tổng quát.
- **Hiệu suất và Tính toán**: So với các LLM quy mô hàng trăm tỉ tham số, các phiên bản Qwen có kích thước vừa phải, giúp tinh chỉnh (fine-tuning) hiệu quả hơn trên các tài nguyên tính toán giới hạn, phù hợp với điều kiện thực tế của Shared Task.

Khái niệm "Limited-Pretraining" Trong bối cảnh của Shared Task VLSP 2025, thuật ngữ "Limited-Pretraining" đề cập đến các ràng buộc sau:

- **Hạn chế kích thước mô hình**: Người tham gia không được sử dụng các mô hình cực lớn (ví dụ: 70B+ tham số), mà thay vào đó phải làm việc với các mô hình có kích thước vừa phải như Qwen.
- **Giới hạn tài nguyên**: Điều này dẫn tới nhu cầu sử dụng các kỹ thuật tinh chỉnh hiệu quả chi phí, đặc biệt là *Parameter-Efficient Transfer Learning (PETL)*, như **LoRA (Low-Rank Adaptation)**, để giảm số lượng tham số cần cập nhật.
- **Thách thức chính**: Làm thế nào để chuyển giao hiệu quả kiến thức ngôn ngữ tổng quát đã được tiền huấn luyện sang *miền Y tế chuyên sâu*, vốn chứa nhiều thuật ngữ khó, trong khi dữ liệu tinh chỉnh bị giới hạn.

Tóm lại Việc sử dụng Qwen trong một đường đua có ràng buộc (Constrained Track) không chỉ đánh giá chất lượng dịch máy, mà còn đánh giá khả năng tối ưu tài nguyên và mức độ hiệu quả của các chiến lược tinh chỉnh nhằm thích nghi mô hình LLM với lĩnh vực Y tế, khi cả mô hình cơ sở và dữ liệu huấn luyện đều bị giới hạn.

2.1 Dịch máy trong Lĩnh vực Y tế

2.1.1 Thách thức ngôn ngữ

Lĩnh vực y tế đặt ra nhiều khó khăn đặc thù:

- **Tính đa nghĩa (Ambiguity):** nhiều từ có nhiều nghĩa phụ thuộc ngữ cảnh (ví dụ: "discharge").
- **Thuật ngữ chuyên ngành:** viết tắt, thuật ngữ gốc Latin/Hy Lạp, danh pháp cần dịch chính xác và nhất quán.
- **Cấu trúc câu phức tạp:** báo cáo lâm sàng và bài báo khoa học thường chứa câu dài và cú pháp phức tạp.

2.1.2 Kỹ thuật chuyên môn hóa dịch máy

Một số kỹ thuật phổ biến:

- **Dữ liệu chuyên ngành:** thu thập bộ dữ liệu song ngữ chất lượng cao cho miền y tế.
- **Domain Adaptation:** kỹ thuật để điều chỉnh mô hình tổng quát sang miền chuyên biệt (fine-tuning, continued pre-training, adapter-based methods, v.v.).
- **Lexicon Injection / Glossary Integration:** nhúng danh mục thuật ngữ (glossary) vào quá trình dịch hoặc hậu xử lý để đảm bảo tính nhất quán của thuật ngữ quan trọng.

3 Dữ liệu và Tiền xử lý dữ liệu

3.1 Dữ liệu Huấn luyện (Training Data) và Phân tích

Theo ràng buộc của Ban Tổ chức VLSP 2025 (Constrained Track), hệ thống dịch máy chỉ được phép sử dụng tập dữ liệu chuyên ngành do Ban Tổ chức cung cấp.

3.1.1 Nguồn gốc, Cấu trúc và Đặc điểm Lĩnh vực

Bộ dữ liệu được sử dụng trong Shared Task VLSP 2025 thuộc miền **Y tế (Medical Domain)**, bao gồm các cặp câu song ngữ Việt–Anh/Anh–Việt được lưu trữ trong các tệp riêng biệt (`train.en.txt` và `train.vi.txt`).

Cấu trúc dữ liệu. Dữ liệu được cung cấp dưới dạng các tệp văn bản (`.txt`), trong đó mỗi dòng tương ứng với một câu. Chúng tôi giả định rằng các câu ở cùng vị trí giữa `train.en.txt` và `train.vi.txt` là các cặp song ngữ tương ứng.

Đặc điểm lĩnh vực y tế. Bộ dữ liệu chủ yếu bao gồm các mô tả y khoa, triệu chứng, bệnh lý, quy trình khám chữa bệnh, báo cáo xét nghiệm và chẩn đoán, trải rộng trên nhiều chuyên ngành như Hồi sức, Hô hấp, Tim mạch, Dược học và Ngoại khoa.

Tuy nhiên, dữ liệu cũng tồn tại một số thách thức về chất lượng, bao gồm nhiễu do viết tắt không thống nhất, lỗi chính tả hoặc ký hiệu đặc biệt, cũng như các trường hợp câu song ngữ không hoàn toàn khớp về mặt ngữ nghĩa.

3.1.2 Thống kê Dữ liệu Huấn luyện và Kiểm chứng

Các thống kê sơ bộ dưới đây dựa trên phân tích ban đầu hai tệp `train.en.txt` và `train.vi.txt`:

Chỉ số	EN	VI	Nhận xét
Cặp câu	50k	50k	Quy mô vừa.
Độ dài TB	12–18	12–18	Độ dài vừa phải.
Tỉ lệ thuật ngữ	15%	15%	Mật độ cao.
Câu lệch	1–2%	1–2%	Cần làm sạch.

Bảng 1: Thống kê dữ liệu huấn luyện VLSP 2025.

Kiểm chứng dữ liệu thực tế (dựa trên các đoạn mẫu). Các snippet thu thập được cho thấy sự xuất hiện đa dạng của các yếu tố ngôn ngữ thuộc lĩnh vực y tế.
Cụ thể, dữ liệu bao gồm nhiều tên bệnh và tình trạng lâm sàng như `otitis media with effusion`, `tricuspid stenosis`, `benign prostatic hyperplasia` hay các thuật ngữ kỹ thuật chuyên sâu như `ICP-MS method`, `CBCT`, `TVUS`; cùng với các số liệu thống kê và giá trị định lượng đặc trưng cho nghiên cứu y khoa, ví dụ `53.33 ± 21.59`, `5.1`, `52.6%`. Điều này khẳng định rằng bộ dữ liệu mang tính chuyên ngành cao và giàu thông tin định lượng.

3.2 Quy trình Tiền xử lý Dữ liệu (Data Preprocessing)

Quy trình tiền xử lý nhằm làm sạch, chuẩn hóa và định dạng dữ liệu song ngữ để tối ưu hiệu quả tinh chỉnh mô hình Qwen.

3.2.1 Làm sạch và Lọc dữ liệu (Cleaning and Filtering)

Xoá dòng rỗng và trùng lặp: Loại bỏ hoàn toàn các câu trùng hoặc dòng trống trong bộ dữ liệu.
Lọc độ dài câu:

- Loại câu quá ngắn (ví dụ: < 3 từ).
- Loại câu quá dài (ví dụ: > 150 tokens).

Lọc theo tỉ lệ độ dài (Length Ratio): Loại bỏ các cặp câu có tỷ lệ độ dài giữa hai ngôn ngữ vượt ngưỡng (ví dụ: > 2.0 hoặc < 0.5), biểu hiện khả năng không khớp ngữ nghĩa.

3.2.2 Chuẩn hoá Văn bản (Text Normalization)

Chuẩn hóa khoảng trắng: Thay thế chuỗi khoảng trắng hoặc ký tự xuống dòng bất thường bằng khoảng trắng đơn. **Chuẩn hóa dấu câu:** Đồng nhất dấu nháy đơn, nháy kép, dấu chấm, phẩy. **Xử lý viết tắt:** Chuẩn hóa các thuật ngữ viết tắt tiếng Việt trong y tế (VD: V.A, T.T.L).

3.2.3 Định dạng Dữ liệu cho Mô hình Decoder-Only (Qwen)

Vì Qwen thuộc kiến trúc **Decoder-only**, dữ liệu cần được định dạng theo kiểu hội thoại (Instruction–Response) để mô hình học cách sinh đầu ra có điều kiện.

Cấu trúc chuẩn được sử dụng:

Instruction + Câu nguồn + Token phân tách + Câu đích

Ví dụ:

- **Dịch Anh → Việt:**

```
Human: Dịch câu sau
sang tiếng Việt: [EN]
Assistant: [VI]
```

- **Dịch Việt → Anh:**

```
Human: Dịch câu sau
sang tiếng Anh: [VI]
Assistant: [EN]
```

Định dạng này giúp SFTTrainer trong tập `phase_2.py` huấn luyện mô hình dự đoán Câu Đích dựa trên hướng dẫn và nội dung câu nguồn.

4 Phương pháp huấn luyện mô hình

4.1 Tổng quan phương pháp

Mô hình được huấn luyện theo phương pháp **Supervised Fine-Tuning (SFT)** trên nền **Qwen/Qwen3-1.7B**, kết hợp **Low-Rank Adaptation (LoRA)** cho bài toán dịch máy song ngữ Anh–Việt trong lĩnh vực y tế. Cách tiếp cận này nhằm khai thác tri thức tiền huấn luyện, giảm chi phí bộ nhớ và thời gian huấn luyện, đồng thời đảm bảo tính ổn định trên GPU RTX 5090 (32GB VRAM).

4.2 Lựa chọn mô hình và biểu diễn số

Mô hình cơ sở: Qwen/Qwen3-1.7B, có năng lực biểu diễn và suy luận tốt hơn các phiên bản 1B–1.5B, đặc biệt hiệu quả với văn bản y khoa.
Định dạng số: Brain Float 16 (BF16), giúp tăng tốc so với FP32, ổn định hơn định lượng 4-bit và tận dụng tốt phần cứng GPU hiện đại.

4.3 Kỹ thuật Tinh chỉnh (Fine-Tuning): LoRA (Low-Rank Adaptation)

LoRA (Low-Rank Adaptation) là một kỹ thuật tinh chỉnh tham số hiệu quả, cho phép huấn luyện các mô hình ngôn ngữ lớn với chi phí tài nguyên thấp hơn so với fine-tuning toàn bộ tham số.

Mục đích Mục tiêu của việc áp dụng LoRA trong nghiên cứu này bao gồm:

- Giảm đáng kể số lượng tham số cần huấn luyện.
- Tiết kiệm bộ nhớ đồ họa (VRAM), cho phép huấn luyện trên GPU có tài nguyên hạn chế.
- Tăng tốc độ quá trình fine-tuning.
- Duy trì hiệu suất mô hình ở mức gần tương đương với fine-tuning toàn bộ tham số.

Cấu hình LoRA (High-Rank) Trong thí nghiệm, nhóm sử dụng cấu hình LoRA với *rank* cao nhằm tăng khả năng thích nghi của mô hình:

- **r = 64**: Rank cao, được ghi chú là có khả năng *bắt chước* hiệu quả của fine-tuning toàn bộ tham số.
- **lora_alpha = 128**: Hệ số tỷ lệ lớn, giúp khuếch đại ảnh hưởng của các trọng số LoRA lên trọng số gốc của mô hình.
- **target_modules**: LoRA được áp dụng cho hầu hết các ma trận trọng số quan trọng trong kiến trúc Transformer, bao gồm: `q_proj`, `k_proj`, `v_proj`, `o_proj` (các thành phần của cơ chế *Self-Attention*) và `gate_proj`, `up_proj`, `down_proj` (các lớp trong khối *Feed-Forward Network*)

Việc áp dụng LoRA trên phạm vi rộng các mô-đun cho phép mô hình được tinh chỉnh sâu hơn, cải thiện khả năng thích nghi với miền dữ liệu mục tiêu trong khi vẫn đảm bảo hiệu quả về tài nguyên tính toán.

4.4 Phương pháp Tải và Xử lý Dữ liệu (Zero-RAM Streaming)

Đây là một trong những điểm cốt lõi nhằm đảm bảo tính ổn định và khả năng mở rộng của quá trình huấn luyện mô hình, đặc biệt khi làm việc với các tập dữ liệu có kích thước lớn.

Sử dụng *IterableDataset* (Streaming) Thay vì tải toàn bộ tập dữ liệu vào bộ nhớ RAM, dữ liệu được đọc tuần tự từng dòng từ tệp nguồn thông qua một hàm sinh dữ liệu (`data_generator`). Cách tiếp cận này cho phép mô hình xử lý dữ liệu theo cơ chế *streaming*, tránh việc chiếm dụng bộ nhớ lớn trong suốt quá trình huấn luyện.

Lợi ích:

- Giải quyết triệt để lỗi `std::bad_alloc` (lỗi cạn kiệt RAM/VRAM) khi huấn luyện với tập dữ liệu lớn.
- Cho phép huấn luyện mô hình trên phần cứng có tài nguyên bộ nhớ hạn chế.
- Chiến lược này thường được gọi là *Zero-RAM Streaming*.

Chiến lược Dữ liệu (Bosch@AI Strategy) Nhóm áp dụng chiến lược xây dựng dữ liệu theo định hướng của Bosch@AI nhằm tăng cường khả năng học song ngữ và khả năng tổng quát hóa của mô hình.

Huấn luyện hai chiều (Bidirectional Strategy):

- Dữ liệu được tạo để mô hình học dịch theo cả hai hướng: Anh $\rightarrow$ Việt và Việt $\rightarrow$ Anh.
- Hai hướng dịch được trộn trong cùng một *mini-batch*, giúp mô hình học được mối quan hệ song ngữ một cách toàn diện hơn.

Kỹ thuật Nhắc lệnh Động (Dynamic Prompting) Dữ liệu huấn luyện được định dạng linh hoạt theo nhiều kiểu nhắc lệnh khác nhau nhằm tăng tính thích nghi của mô hình:

- **60% dữ liệu dạng Hướng dẫn (Instructional Prompting):** Nhắc mô hình rõ ràng về nhiệm vụ cần thực hiện, ví dụ:

  Translate the following medical text from {src_lang} to {tgt_lang}.

- **40% dữ liệu dạng Trực tiếp (Direct Prompting):** Sử dụng nhắc lệnh ngắn gọn hơn, ví dụ:

{src_text}
({src_lang} → {tgt_lang})

Mục đích: Tăng cường khả năng tổng quát hóa và tính linh hoạt của mô hình trong việc phản hồi các kiểu nhắc lệnh đa dạng trong thực tế.

System Prompt Chuyên biệt Một `system_prompt` chi tiết được thiết lập nhằm định hướng mô hình hoạt động như một *"professional medical translator"*. Prompt này nhấn mạnh các nguyên tắc sau:

- Dịch chính xác và đầy đủ nội dung y khoa.
- Bảo tồn thuật ngữ chuyên ngành và ngữ cảnh lâm sàng.
- Duy trì giọng văn trang trọng, chuyên nghiệp.
- Hạn chế hiện tượng ảo giác (hallucination).
- Tuân thủ các nguyên tắc đạo đức trong lĩnh vực y tế.

4.5 Cấu hình huấn luyện

Tham số	Giá trị
Learning rate	2×10^{-5}
Batch size	4–8
Grad. accum.	2–4
Epochs	2–4
Optimizer	AdamW (8-bit)
Warmup ratio	0.03

Bảng 2: Các siêu tham số huấn luyện chính.

Trong đó, *learning rate* điều khiển tốc độ cập nhật trọng số; *gradient accumulation* được sử dụng để mô phỏng batch lớn nhằm ổn định quá trình huấn luyện; *optimizer AdamW 8-bit* giúp giảm đáng kể bộ nhớ GPU; *warmup ratio* được đặt ở mức thấp để ổn định giai đoạn đầu.

4.6 Tối ưu hóa VRAM và Sự ổn định Huấn luyện

Để đảm bảo quá trình huấn luyện mô hình diễn ra ổn định trên GPU có dung lượng bộ nhớ hạn chế, nhóm nghiên cứu áp dụng các kỹ thuật tối ưu hóa bộ nhớ GPU (VRAM) quan trọng sau đây.

Gradient Checkpointing Mô tả: Gradient Checkpointing là một kỹ thuật tiết kiệm bộ nhớ bằng cách không lưu trữ toàn bộ các *activation* trong quá trình truyền tiến (forward pass). Thay vào đó, chỉ một số activation chọn lọc được lưu lại; các activation còn lại sẽ được tính toán lại trong quá trình truyền ngược (backward pass).

Lợi ích:

- Giảm đáng kể lượng VRAM tiêu thụ trong quá trình huấn luyện.
- Cho phép huấn luyện các mô hình lớn hơn trên cùng một GPU.
- Đánh đổi một phần tốc độ tính toán để đổi lấy hiệu quả sử dụng bộ nhớ GPU.

Gradient Accumulation Mô tả: Thay vì cập nhật trọng số sau mỗi mini-batch nhỏ, gradient được tích lũy qua nhiều bước huấn luyện liên tiếp trước khi thực hiện một lần cập nhật tham số.

Lợi ích:

- Mô phỏng một *batch size logic* lớn hơn mà không cần tăng batch size thực tế.
- Trong cấu hình này, batch size logic tương đương $4 \times 8 = 32$.
- Giảm áp lực lên VRAM, đồng thời giúp quá trình huấn luyện ổn định hơn.

Việc kết hợp Gradient Checkpointing và Gradient Accumulation cho phép huấn luyện hiệu quả các mô hình ngôn ngữ lớn trong điều kiện tài nguyên phần cứng hạn chế, đồng thời duy trì độ ổn định và hiệu suất của mô hình. "'

4.7 Đánh giá trong quá trình huấn luyện (Online Evaluation)

Để theo dõi sát sao hiệu suất của mô hình trong suốt quá trình tinh chỉnh có giám sát (Supervised Fine-Tuning – SFT), nhóm nghiên cứu triển khai cơ chế đánh giá trực tuyến (*online evaluation*) thông qua một callback tùy chỉnh. Cách tiếp cận này cho phép đánh giá liên tục chất lượng mô hình mà không làm gián đoạn vòng huấn luyện.

BLEUCallback Một callback tùy chỉnh, gọi là `BLEUCallback`, được tích hợp trực tiếp vào vòng huấn luyện nhằm đánh giá chất lượng dịch máy của mô hình trong quá trình học.

Mục đích của BLEUCallback bao gồm:

- Cung cấp điểm số BLEU định kỳ sau mỗi 1,000 bước huấn luyện.
- Theo dõi tiến trình hội tụ của mô hình theo thời gian.
- Phát hiện sớm hiện tượng suy giảm chất lượng hoặc học quá mức (*overfitting*).
- Hỗ trợ kiểm soát chất lượng mô hình trong suốt quá trình SFT.

Xử lý tokenizer trong quá trình đánh giá Trong quá trình sinh văn bản để đánh giá chất lượng mô hình, cách xử lý *padding* của tokenizer có ảnh hưởng trực tiếp đến độ chính xác của kết quả sinh. Cụ thể,nó thực hiện chuyển đổi linh hoạt thuộc tính `tokenizer.padding_side`:

- **Trong huấn luyện (Training):** sử dụng `padding_side = "right"` để tối ưu hiệu quả huấn luyện theo kiến trúc Transformer.
- **Trong suy diễn/đánh giá:** tạm thời chuyển sang `padding_side = "left"` nhằm đảm bảo quá trình sinh chuỗi tự hồi quy diễn ra chính xác.
- Sau khi hoàn tất đánh giá BLEU, tokenizer được khôi phục về cấu hình ban đầu để tiếp tục huấn luyện.

Cách tiếp cận này đảm bảo sự tách biệt rõ ràng giữa huấn luyện và đánh giá về mặt kỹ thuật, đồng thời tránh sai lệch trong quá trình sinh văn bản mà vẫn duy trì tính ổn định và nhất quán của vòng huấn luyện.

4.8 Tổng kết

Phương pháp huấn luyện tập trung vào hiệu quả tài nguyên (LoRA, BF16, streaming), đảm bảo ổn định hệ thống (gradient checkpointing, kiểm soát RAM) và phù hợp với bài toán dịch máy y tế thông qua thiết kế prompt chuyên biệt và chiến lược huấn luyện song hướng. Đây là một chiến lược hiện đại, thực tiễn và hiệu quả cho các mô hình ngôn ngữ lớn.

5 Kết quả và đánh giá

5.1 Chỉ số đánh giá định lượng

Chúng tôi đánh giá hiệu năng của mô hình dựa trên ba chỉ số phổ biến trong dịch máy, bao gồm BLEU (Bilingual Evaluation Understudy), TER (Translation Edit Rate) và METEOR. Kết quả thực nghiệm trên tập *public_test* được trình bày chi tiết trong Bảng 3.

Chiều dịch	**BLEU** (↑)	**TER** (↓)	**METEOR**
Anh → Việt (EN–VI)	**43.86**	48.01	0.0017
Việt → Anh (VI–EN)	35.20	58.35	0.0017
Trung bình (Overall)	**40.28**	**52.34**	**0.0017**

Bảng 3: Kết quả thực nghiệm trên tập dữ liệu kiểm thử (*Public Test*).

Phân tích kết quả **Chiều Anh–Việt (EN–VI).** Mô hình đạt điểm BLEU là 43.86 và chỉ số TER là 48.01. Đây là một kết quả ấn tượng đối với dữ liệu chuyên ngành y tế, cho thấy mô hình có khả năng chuyển ngữ chính xác và duy trì tốt cấu trúc câu mang tính chuyên biệt của các báo cáo y khoa.
Chiều Việt–Anh (VI–EN). Mô hình đạt điểm BLEU là 35.20 và TER là 58.35. Mặc dù thấp hơn so với chiều Anh–Việt, mức điểm này vẫn phản ánh khả năng dịch thuật tốt trong bối cảnh các thuật ngữ y tế tiếng Việt thường mang tính đa nghĩa và phức tạp.
Chỉ số METEOR. Cả hai chiều dịch đều duy trì điểm METEOR ở mức 0.0017, cho thấy mức độ ổn định nhất định về mặt tương đồng ngữ nghĩa giữa bản dịch sinh ra và câu tham chiếu.

5.2 Phân tích lỗi (Error Analysis)

Việc đánh giá định tính thông qua so sánh trực tiếp bản dịch của mô hình với bản dịch tham chiếu (*Ground Truth*) giúp nhận diện các hành vi cụ thể của mô hình trong ngữ cảnh y tế.

5.2.1 Các loại lỗi phổ biến

a) Lỗi thực thể tên riêng và địa danh (Named Entity Recognition Error). Đây là loại lỗi xuất hiện rõ rệt nhất khi mô hình đối mặt với các địa danh không thuộc Việt Nam hoặc các tên riêng ít phổ biến.

Tham chiếu: "*...in Phone Hong and Keo Oudom districts, Vientiane province.*"

Mô hình dịch: "*...tại 2 huyện Phố Hồng và Kéo Oudom, tỉnh Vĩnh Phúc.*"

Nhận xét: Mô hình có xu hướng "Việt hóa" địa danh một cách cưỡng ép (ví dụ: *Vientiane* bị dịch thành *Vĩnh Phúc*) hoặc phiên âm sai các địa danh của Lào. Hiện tượng này cho thấy mô hình bị ảnh hưởng bởi dữ liệu huấn luyện tiếng Việt, trong đó các tỉnh thành Việt Nam có tần suất xuất hiện cao hơn đáng kể so với các địa danh nước ngoài.

b) Lỗi thuật ngữ chuyên ngành và từ viết tắt (Medical Terminology & Abbreviation). Mặc dù xử lý tốt các thuật ngữ y khoa phổ biến, mô hình vẫn gặp khó khăn với các cụm từ ghép phức tạp hoặc các từ viết tắt mang tính đa nghĩa.

Tham chiếu: "*...management style flow levels in children...*"

Mô hình dịch: "*...phong cách quản lý dòng chảy ở trẻ em...*"

Nhận xét: Cụm từ "*management style flow*" trong ngữ cảnh này có khả năng liên quan đến quy trình quản lý bệnh lý hoặc luồng xử lý lâm sàng. Tuy nhiên, mô hình đã dịch theo hướng *word-by-word* thành "quản lý dòng chảy", dẫn đến việc làm sai lệch hoặc mất đi ý nghĩa y khoa thực tế.

c) Lỗi mất thông tin hoặc dịch thiếu (Omission). Đối với các câu có cấu trúc liệt kê dài, mô hình đôi khi bỏ sót các chi tiết quan trọng, đặc biệt là các thông tin định lượng ở cuối câu.

Ví dụ: Trong các đoạn văn liệt kê thành phần thuốc hoặc các chỉ số thống kê như p-value, khoảng tin cậy 95% (CI 95%), mô hình thỉnh thoảng ngắt câu sớm hơn so với văn bản gốc, dẫn đến mất mát thông tin.

5.2.2 Nhận xét về điểm mạnh và điểm yếu của mô hình

Điểm mạnh:

- **Khả năng nắm bắt cấu trúc câu y khoa:** Mô hình Transformer, đặc biệt là dòng Qwen, thể hiện ưu thế rõ rệt trong việc duy trì các cấu trúc câu mang tính học thuật, chẳng hạn như "Nghiên cứu mô tả cắt ngang được thực hiện trên…" hoặc "Mục tiêu của nghiên cứu này là…".
- **Sử dụng từ vựng Hán–Việt phù hợp:** Bản dịch tiếng Việt sử dụng hệ thống từ ngữ trang trọng, đúng phong cách văn bản khoa học, ví dụ như "biến chứng", "di căn", "đa hình gen", thay vì các cách diễn đạt mang tính thông tục.
- **Độ chính xác cao đối với hoạt chất:** Các tên thuốc và hoạt chất như *Diclofenac sodium* hay *Benzathine penicillin G* được xử lý tốt và giữ nguyên dạng, không bị dịch sai sang nghĩa thông thường.

Điểm yếu:

- **Hiện tượng ảo giác (Hallucination) với thực thể:** Mô hình đôi khi tự suy diễn các thông tin về địa lý hoặc tổ chức dựa trên xác suất ngôn ngữ, thay vì bảo toàn tính xác thực của văn bản nguồn.
- **Khó khăn trong xử lý số liệu phức tạp:** Trong các đoạn văn chứa nhiều tỷ lệ phần trăm và phép so sánh định lượng, mô hình có xu hướng nhầm lẫn giữa các đối tượng được so sánh, ví dụ như hoán đổi tỷ lệ giữa hạch rốn phổi và hạch trung thất.
- **Phụ thuộc vào Prompt:** Hiệu năng dịch chịu ảnh hưởng đáng kể từ cách thiết lập prompt. Khi thiếu chỉ dẫn rõ ràng về miền y tế, mô hình có xu hướng chuyển sang phong cách dịch mang tính văn chương hoặc đời sống, làm giảm độ chính xác chuyên ngành.

6 Kết luận và Hướng phát triển

6.1 Kết luận

Trong báo cáo này, chúng tôi đã trình bày một hệ thống dịch máy chuyên biệt cho lĩnh vực y tế tham gia nhiệm vụ chia sẻ tại VLSP 2025. Bằng cách tận dụng sức mạnh của mô hình ngôn ngữ lớn Qwen kết hợp với kỹ thuật tinh chỉnh hiệu quả tham số LoRA, hệ thống đạt được các kết quả khả quan trên tập dữ liệu thử nghiệm công khai.

Các kết quả thực nghiệm cho thấy:

- **Hiệu năng ấn tượng:** Chỉ số BLEU trung bình đạt 40.28, trong đó chiều dịch Anh–Việt đạt mức cao là 43.86. Kết quả này cho thấy việc tinh chỉnh mô hình tiền huấn luyện trên dữ liệu chuyên ngành y tế giúp cải thiện đáng kể khả năng chuyển ngữ các thuật ngữ và cấu trúc y khoa.
- **Tính thực tiễn:** Việc áp dụng các kỹ thuật tối ưu như huấn luyện ở độ chính xác BF16 và Gradient Checkpointing giúp giảm đáng kể chi phí tài nguyên tính toán. Nhờ đó, hệ thống có thể được huấn luyện trong điều kiện phần cứng hạn chế mà vẫn duy trì được độ chính xác về ngữ nghĩa và cấu trúc văn bản khoa học.
- **Đóng góp chuyên ngành:** Hệ thống không chỉ giải quyết bài toán dịch thuật tổng quát mà còn đảm bảo tính nhất quán của các thực thể y tế quan trọng như tên thuốc và chỉ số xét nghiệm, vốn là yếu tố then chốt trong các ứng dụng dịch thuật lâm sàng.

6.2 Hướng phát triển

Mặc dù đạt được những kết quả tích cực, hệ thống vẫn còn một số hạn chế cần được khắc phục để đạt tới độ tin cậy cao trong các ứng dụng thực tế. Trong tương lai, chúng tôi định hướng phát triển theo các hướng sau:

- **Cải thiện khả năng bảo toàn thực thể tên riêng (Named Entity Preservation):** Tích hợp các lớp hậu xử lý hoặc áp dụng cơ chế *Constrained Decoding* nhằm đảm bảo các địa danh và tên riêng được giữ nguyên, tránh hiện tượng "hallucination" (ảo giác) hoặc bị dịch nhầm sang các thực thể phổ biến khác.

- **Mở rộng quy mô mô hình:** Thử nghiệm với các phiên bản mô hình có số lượng tham số lớn hơn như Qwen-14B hoặc Qwen-72B, kết hợp với các kỹ thuật tiết kiệm bộ nhớ như Quantized LoRA (QLoRA) nhằm nâng cao khả năng suy luận trong các ngữ cảnh y khoa phức tạp.

- **Tích hợp tri thức y khoa:** Kết hợp các cơ sở tri thức và từ điển chuyên ngành y tế như UMLS hoặc SNOMED CT vào quá trình nhắc lệnh (prompting) hoặc trực tiếp vào kiến trúc mô hình, từ đó cải thiện độ chính xác đối với các thuật ngữ hiếm và chuyên sâu.

- **Đánh giá bởi chuyên gia:** Bên cạnh các chỉ số đánh giá tự động như BLEU hay TER, chúng tôi dự kiến phối hợp với các chuyên gia y tế để tiến hành đánh giá thủ công, tập trung vào tính an toàn và độ chính xác của bản dịch trong các báo cáo lâm sàng thực tế.

7 References

References

[1] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Huang, F. (2023). *Qwen technical report*. arXiv preprint arXiv:2309.16609.

[2] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models*. arXiv preprint arXiv:2106.09685.

[3] VLSP Organizers. (2025). *Shared Task: Medical Domain Machine Translation with Limited-Pretraining Models*. The 11th International Workshop on Vietnamese Language and Speech Processing.

[4] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). *Bleu: a method for automatic evaluation of machine translation*. Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

[5] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A study of translation edit rate (TER)*. Association for Machine Translation in the Americas.

[6] Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization*. arXiv preprint arXiv:1711.05101.

[7] Vaswani, A., et al. (2017). *Attention is all you need*. Advances in neural information processing systems.

[8] Kudo, T., & Richardson, J. (2018). *SentencePiece: A simple subword tokenizer*. arXiv preprint arXiv:1808.06226.