# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/E7iFQ5xhF7g*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/LeNguyenAnhKhoa/CS519.Q11/blob/main/RESEARCH_AND_DEVELOPMENT_OF_AN_ADAPTIVE_ITERATIVE_FRAMEWORK_FOR_MULTIMODAL_LONG-DOCUMENT_QUESTION_ANSWERING.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

| | |
|---|---|
| • Họ và Tên: Lê Nguyễn Anh Khoa<br>• MSSV: 23520742<br>  | • Lớp: CS519.Q11<br>• Tự đánh giá (điểm tổng kết môn): 9.5/10<br>• Số buổi vắng: 0<br>• Số câu hỏi QT cá nhân: 8<br>• Số câu hỏi QT của cả nhóm: 16<br>• Link Github: https://github.com/LeNguyenAnhKhoa/CS519.Q11<br>• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<br>  ○ Lên ý tưởng đề tài<br>  ○ Làm poster đề tài<br>  ○ Làm video YouTube |

| | |
|---|---|
| • Họ và Tên: Cáp Kim Hải Anh<br>• MSSV: 23520036 | • Lớp: CS519.Q11<br>• Tự đánh giá (điểm tổng kết môn): 9/10<br>• Số buổi vắng: 0<br>• Số câu hỏi QT cá nhân: 8 |

| | |
|---|---|
|  | • Số câu hỏi QT của cả nhóm: 16<br>• Link Github: https://github.com/LeNguyenAnhKhoa/CS519.Q11<br>• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<br>  ○ Làm slides đề tài<br>  ○ Viết đề cương nghiên cứu |

# ĐỀ CƯƠNG NGHIÊN CỨU

**TÊN ĐỀ TÀI (IN HOA)**

NGHIÊN CỨU VÀ PHÁT TRIỂN KHUNG LÀM VIỆC LẶP THÍCH ỨNG CHO BÀI TOÁN HỎI ĐÁP TRÊN TÀI LIỆU DÀI ĐA PHƯƠNG THỨC

**TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**

RESEARCH AND DEVELOPMENT OF AN ADAPTIVE ITERATIVE FRAMEWORK FOR MULTIMODAL LONG-DOCUMENT QUESTION ANSWERING

**TÓM TẮT** *(Tối đa 400 từ)*

Trong bối cảnh bùng nổ dữ liệu số, các tài liệu dài và đa phương thức (không chỉ chứa văn bản mà còn đan xen hình ảnh và bảng biểu) là rất phổ biến, theo đó, nhu cầu tự động hóa việc trả lời câu hỏi (Question Answering - QA) kèm trích xuất thông tin từ các tài liệu phức tạp này đang trở nên cấp thiết. Tuy nhiên, các phương pháp hỏi đáp hiện hành, điển hình là RAG (Retrieval-Augmented Generation) đơn lượt, thường bộc lộ hạn chế khi xử lý các tài liệu này. Vấn đề cốt lõi nằm ở việc các mô hình này thường truy xuất thông tin một cách thụ động dựa vào sự tương đồng bề mặt ngữ nghĩa, dẫn đến bỏ sót các thông tin ngữ cảnh quan trọng nằm rải rác hoặc ẩn sâu trong các định dạng đa phương thức.

Xuất phát từ thực tiễn và hạn chế trên, đề tài đề xuất nghiên cứu và phát triển khung làm việc lặp thích ứng (Adaptive Iterative Framework) với khả năng suy luận đa bước trên tài liệu đa phương thức. Điểm đột phá của nghiên cứu nằm ở việc chuyển dịch tư duy từ tìm kiếm sự tương đồng sang tối ưu hóa thông tin, nơi hệ thống có thể tự đánh giá độ thiếu hụt thông tin và chủ động tìm kiếm dữ liệu bổ sung. Cụ thể, chúng tôi dự

kiến áp dụng lý thuyết tối ưu hóa thông tin để định lượng độ lợi thông tin (Information Gain) tại mỗi bước lập luận, kết hợp với việc sử dụng mô hình ngôn ngữ lớn (LLM) như một tác nhân điều phối thông minh. Tác nhân này đảm nhiệm phân tích câu hỏi, sinh ra các truy vấn con (sub-queries) để lấp đầy khoảng trống kiến thức, và tinh chỉnh kết quả truy xuất thông qua cơ chế phản hồi liên tục. Kết quả nghiên cứu kỳ vọng sẽ tạo ra một giải pháp QA mạnh mẽ, có khả năng xử lý các câu hỏi phức tạp đòi hỏi sự tổng hợp thông tin từ nhiều nguồn và định dạng khác nhau, khắc phục điểm yếu của các hệ thống truy xuất đơn lượt truyền thống.

**GIỚI THIỆU** *(Tối đa 1 trang A4)*

Sự bùng nổ dữ liệu số đã tạo ra khối lượng khổng lồ các tài liệu PDF, slide thuyết trình chứa thông tin hỗn hợp. Một thách thức lớn nảy sinh khi trích xuất thông tin từ nguồn này là sự đứt gãy ngữ cảnh. *Ví dụ điển hình:* Một câu hỏi yêu cầu *"So sánh doanh thu Bắc Mỹ trong biểu đồ (Trang 5) với dự báo (Trang 20)"*. Để trả lời, hệ thống không chỉ cần tìm Trang 5, mà từ Trang 5 phải nhận ra nhu cầu tìm tiếp thông tin ở Trang 20.

**Hạn chế của các giải pháp hiện tại:** Qua nghiên cứu, chúng tôi nhận thấy các phương pháp hiện tại đều gặp bế tắc về mặt kiến trúc:

- **Mô hình RAG truyền thống:** Tìm kiếm dựa trên độ tương đồng ngữ nghĩa (Semantic Similarity). Chúng chỉ biết tìm những gì giống câu hỏi về mặt từ ngữ, nhưng thiếu khả năng lập kế hoạch. Trong ví dụ trên, trang 20 có thể không chứa từ khóa nào giống câu hỏi ban đầu, nên RAG sẽ bỏ qua nó.
- **Mô hình Multimodal LLMs:** Nạp toàn bộ tài liệu vào cửa sổ ngữ cảnh. Chi phí tính toán quá lớn và gặp vấn đề Lost-in-the-middle (mất tập trung), khiến mô hình bỏ sót các chi tiết nhỏ trong biểu đồ khi phải xử lý hàng chục trang tài liệu cùng lúc.

**Khoảng trống nghiên cứu và Đề xuất**: Hiện nay chưa có một cơ chế định lượng rõ ràng để máy tính biết "Khi nào thì đủ thông tin?". Các mô hình thường ảo giác (hallucination) thay vì chủ động tìm kiếm tiếp. Xuất phát từ khoảng trống này, chúng tôi đề xuất giải pháp **Truy xuất Lặp thích ứng**, coi việc tìm kiếm là bài toán tối ưu hóa thông tin (Information Optimization) sử dụng hàm InfoNCE để dẫn hướng.

Chúng tôi xác định các thành phần đầu vào và đầu ra của hệ thống như sau:

- Đầu vào (Input): Hệ thống tiếp nhận hai thành phần chính:
    - Tài liệu nguồn: Một hoặc nhiều tập tin định dạng PDF hoặc hình ảnh (slide, scan) có độ dài lớn (ví dụ: dài 20-50 trang), chứa nội dung hỗn hợp gồm văn bản, biểu đồ và bảng.
    - Câu truy vấn: Một câu hỏi tự nhiên từ người dùng, có thể yêu cầu thông tin chi tiết (ví dụ: "Tỷ lệ tăng trưởng doanh thu trong biểu đồ trang 15 so với số liệu dự báo ở trang 3 là bao nhiêu?").
- Đầu ra (Output):
    - Câu trả lời văn bản: Kết quả chính xác, súc tích được tổng hợp từ quá trình suy luận (ví dụ: "Tỷ lệ tăng trưởng là 15%, dựa trên doanh thu thực tế 50 tỷ ở trang 15 và dự báo 43.5 tỷ ở trang 3").
    - Minh chứng trích dẫn: Hệ thống chỉ rõ nguồn gốc thông tin được lấy từ trang nào, hình ảnh/đoạn văn nào để người dùng có thể kiểm chứng.

Nghiên cứu này sẽ kỳ vọng không chỉ mang ý nghĩa khoa học trong việc đóng góp một giải pháp hiệu quả cho bài toán trích xuất thông tin tự động mà còn có tính ứng dụng cao trong các lĩnh vực phân tích như tài chính, kinh tế và giáo dục.

**MỤC TIÊU** *(Viết trong vòng 3 mục tiêu)*

Để giải quyết các thách thức và hạn chế nêu trên, đề tài tập trung thực hiện 03 mục

tiêu cụ thể sau:

1. **Mô hình hóa toán học cho Độ lợi thông tin (Information Gain)**

- Khắc phục tình trạng tìm kiếm cảm tính của RAG hiện tại.
- Nội dung: Xây dựng công thức tối ưu hóa dựa trên hàm mất mát InfoNCE. Mục tiêu là định lượng được sự "giảm thiểu độ không chắc chắn", tạo ra thước đo toán học để hệ thống quyết định dừng lại hay tìm tiếp.

2. **Thiết kế và phát triển Khung làm việc lặp thích ứng (Adaptive Iterative Framework):**

- Xây dựng thuật toán truy xuất động có khả năng tự động phân tích câu hỏi phức tạp thành các truy vấn con (sub-queries).
- Phát triển module tích hợp mô hình ngôn ngữ lớn (LLM) đóng vai trò tác nhân đánh giá để kiểm tra tính đầy đủ của thông tin sau mỗi bước truy xuất, từ đó quyết định dừng lại hoặc tiếp tục tìm kiếm thêm bằng chứng.

3. **Thực nghiệm, đánh giá và tối ưu hiệu suất hệ thống:**

- Chứng minh tính ưu việt về hiệu suất: Triển khai hệ thống trên các bộ dữ liệu chuẩn về tài liệu đa phương thức như MP-DocVQA hoặc MMVQA.
- Đo lường và so sánh hiệu quả của mô hình đề xuất với các phương pháp RAG truyền thống dựa trên các chỉ số định lượng: Độ chính xác (Accuracy), F1-Score và số bước lặp trung bình cần thiết để đưa ra câu trả lời đúng.

**NỘI DUNG VÀ PHƯƠNG PHÁP**

1. **Nội dung 1: Cơ sở lý thuyết và xây dựng hệ thống:** Chúng tôi phân rã hệ thống thành 3 module độc lập để tối ưu hóa từng chức năng:

- **Module Retriever (Truy xuất):** Sử dụng **SigLIP** (Sigmoid Loss for Language

Image Pre-Training). SigLIP là mô hình state-of-the-art hiện nay, xử lý tốt hơn các chi tiết cục bộ (fine-grained details) trong ảnh tài liệu độ phân giải cao so với CLIP truyền thống.

- **Module Generator & Judge:** Sử dụng **GPT-4o**. Qua thử nghiệm, các mô hình mã nguồn mở (Llama-3, Qwen) thường yếu trong khả năng tuân thủ chỉ dẫn (Instruction Following) phức tạp và đánh giá sai độ lợi thông tin. GPT-4o đảm bảo độ chính xác cần thiết để chứng minh thuật toán (Proof of Concept).
- **Lý thuyết tối ưu hóa:** Sử dụng **InfoNCE Loss** thay vì Cosine Similarity. Lý do là InfoNCE giúp tối đa hóa thông tin tương hỗ, chọn ra các trang tài liệu thực sự giúp giảm độ nhiễu của câu trả lời chứ không chỉ là khớp từ khóa.

2. **Nội dung 2: Thiết kế Thuật toán Lặp (Algorithm Design):** Quy trình xử lý được thiết kế theo vòng lặp khép kín:

- **Bước 1 - Phân rã:** Nhận câu hỏi Q, sinh ra sub-query q1 (Ví dụ: *"Tìm biểu đồ doanh thu…"*).
- **Bước 2 - Truy xuất:** Dùng SigLIP quét tập dữ liệu D, lấy ra tập ứng viên S.
- **Bước 3 - Đánh giá:** Module Judge kiểm tra thông tin trong S.
    - *Nhận diện mẫu:* Nếu thấy dữ liệu thiếu hụt (Entropy cao), chuyển sang bước 4. Nếu đủ, chuyển sang bước 6.
- **Bước 4 - Tinh chỉnh:** Sinh câu hỏi phụ tiếp theo dựa trên phần thiếu (Ví dụ: *"Tìm bảng chú thích cho biểu đồ vừa tìm được"*).
- **Bước 5 - Lặp lại:** Quay lại Bước 2 với sub-query mới để cập nhật tập S.
- **Bước 6 - Kết luận:** Tổng hợp thông tin và trả lời.

**Nội dung 3: Thực nghiệm kiểm chứng**

- **Dataset:** MP-DocVQA (Tài liệu nhiều trang) và MMVQA.
- **Kịch bản A/B Testing:** So sánh *Proposed Framework* với *Standard RAG*.

**Thách thức, khó khăn và giải quyết**

- **Thách thức vòng lặp vô hạn (Infinite Loop):**
    - LLM có xu hướng cầu toàn, liên tục tìm kiếm chi tiết vụn vặt khiến hệ thống không bao giờ dừng.
    - Thiết lập ngưỡng bão hòa Entropy. Nếu sau một bước, lượng thông tin mới (Information Gain) tiệm cận 0, hệ thống buộc phải dừng và trả lời tốt nhất có thể.
- **Thách thức về độ trễ:**
    - Gọi LLM nhiều lần làm chậm hệ thống.
    - Áp dụng chiến lược cắt tỉa không gian tìm kiếm (Pruning). Ở vòng đầu, loại bỏ ngay 80% các trang tài liệu không liên quan bằng vector search nhẹ, chỉ dùng LLM xử lý sâu 20% trang còn lại.

**KẾT QUẢ MONG ĐỢI**

1. Về hiệu suất kỹ thuật: Hệ thống đạt độ chính xác kỳ vọng từ 40-55% trên tập dữ liệu chuẩn, cải thiện đáng kể khả năng xử lý các câu hỏi suy luận đa bước phức tạp so với các mô hình RAG đơn lượt hiện hành.
2. Phân tích lỗi: Đánh giá các trường hợp thất bại để hiểu rõ giới hạn của phương pháp xấp xỉ thông tin.
3. Đề xuất được quy trình chuẩn hóa áp dụng lý thuyết tối ưu hóa thông tin vào bài toán truy xuất tài liệu.

**TÀI LIỆU THAM KHẢO** *(Định dạng DBLP)*

[1] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao, "ReAct: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2023.

[2] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny, "Hierarchical multimodal transformers for Multi-Page DocVQA," *Pattern Recognition*, vol. 144, p. 109834, 2023.

[3] Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han, "MMVQA: A Comprehensive Dataset for Investigating Multipage Multimodal Information Retrieval in PDF-based Visual Question Answering," IJCAI, 2024

[4] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer, "Sigmoid Loss for Language Image Pre-Training," ICCV, 2023

[5] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[6] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al., "VisRAG: Vision-based retrieval-augmented generation on multi-modality documents," *arXiv preprint arXiv:2410.10594*, 2024.

[7] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-9474.

[9] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang, “MM-React: Prompting ChatGPT for Multimodal Reasoning and Action,” *arXiv preprint arXiv:2303.11381*, 2023.