Research and Development of an Adaptive Iterative Framework for Multimodal Long-Document Question Answering

Lê Nguyễn Anh Khoa[1,2] **Cáp Kim Hải Anh** [1,2]

[1] University of Information Technology-VNUHCM, Ho Chi Minh City, Vietnam

[2] Vietnam National University, Ho Chi Minh City, Vietnam

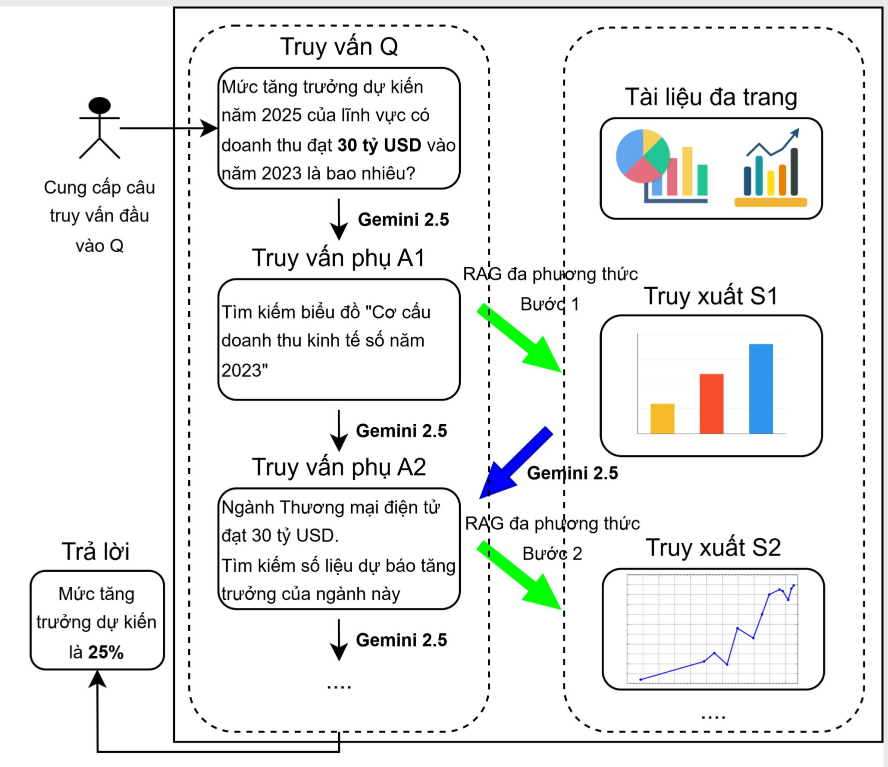
What ?

- **Problem:** Answering questions from lengthy, multi-page, and multimodal documents (e.g., financial reports, slide decks).
- **Challenge:** These documents contain diverse layouts and interleaved text/images. The model must integrate information across different pages and modalities.
- **Goal:** To produce accurate, concise text answers by analyzing relevant content across multiple heterogeneous pages.

Why ?

- **Limitation of Current Methods:** Existing Multimodal RAG systems use single-turn retrieval. They often miss fine-grained details in complex documents.
- **Need:** We need a system that can iteratively refine its search, similar to how a human browses a document to find an answer.

Overview



Mức tăng trưởng dự kiến năm 2025 của lĩnh vực có doanh thu đạt **30 tỷ USD** vào năm 2023 là bao nhiêu?

Tìm kiếm biểu đồ "Cơ cấu doanh thu kinh tế số năm 2023"

Ngành Thương mại điện tử đạt 30 tỷ USD.
Tìm kiếm số liệu dự báo tăng trưởng của ngành này

Mức tăng trưởng dự kiến là **25%**

- **Concept:** An adaptive iterative framework that treats Document QA as an agentic reasoning task.
- **Key Mechanism:** The system balances **Information Gain** and **Uncertainty Reduction**..

- **Process Flow:**

1. **Look:** The model generate sub-question and retrieve information to answer this question.
2. **Think (LLM Judge):** It evaluates if the information is sufficient. If not, it identifies *what is missing*?
3. **Act (Sub-query):** It generates a specific sub-query to find the missing piece
4. **Repeat:** This loop continues until the answer is found or uncertainty is minimized.

Description

1. Query Formulation & Planning

- While searching for the whole answer at once is a complex problem which have many views about it, the model identifies exactly what "missing piece" of information prevents it from answering right now.
- **Sub-Query Generation:** The system generates a specific, natural language search query (e.g., *"Search for the chart "Revenue Structure of the Digital Economy in 2023"*) rather than a generic keyword search.
- **Goal:** To turn a complex, multi-hop question into a series of simple, solvable steps.

2. Intelligent Retrieval

- Instead of using traditional keyword research methods (like BM25), which are ineffective with complex multimodal documents. We use **ColPali** as a Multimodal Retrieval Model in the form of visual inputs.
- This allows the system to calculate the distance between the user's query and the document page's multimodal content in vector space.
- **Maximizing New Information:** It does not just look for "similar" pages; it looks for pages that provide the *highest information gain* relative to what the model already knows. It actively filters out pages that are redundant or irrelevant, keeping the context clean and focused.

3. Reasoning & Feedback Loop

- **LLM as a Judge:** Once a page is retrieved, the LLM evaluates it. It asks: *"Does this page answer my sub-query?* and score the answer. If the answer over a threshold it will keep it and come to next step otherwise the model will remove it.
- If data is missing**,** the model refines its plan and goes back to Phase 1 with a new query. If data is found**:** It aggregates the evidence.
- Once sufficient information is gathered (or a step limit is reached), the model synthesizes the final answer using evidence collected across all steps.

Le Nguyen Anh Khoa – University of Information Technology-VNUHCM, Ho Chi Minh City, Vietnam
TEL : 0825636700 Email : 23520742@gm.uit.edu.vn