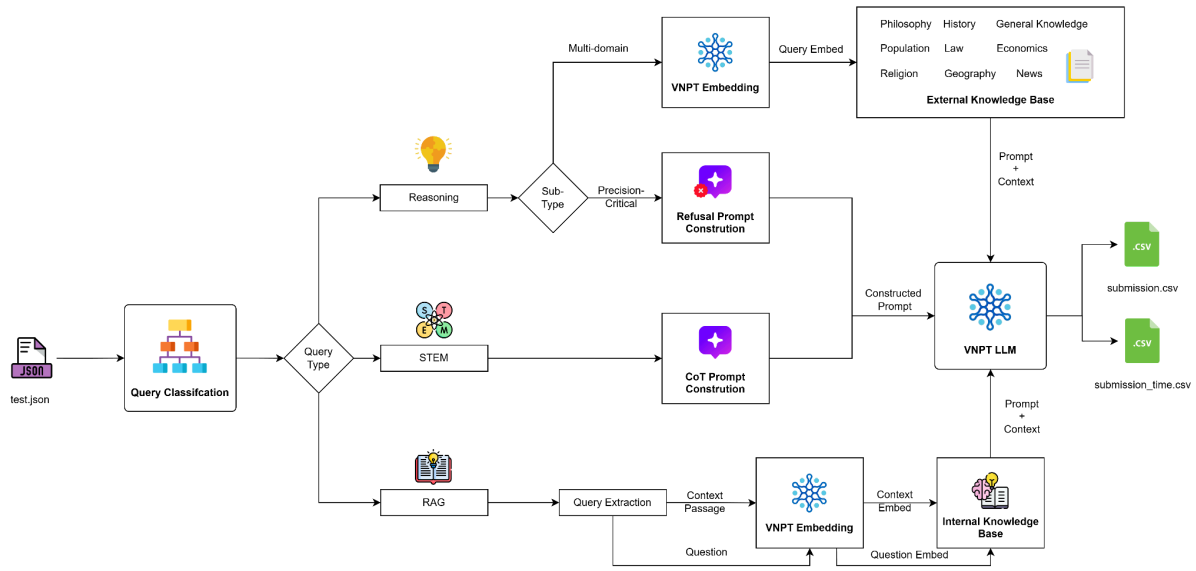
OVERSITTING - HỆ THỐNG TRẢ LỜI TRẮC NGHIỆM TIẾNG VIỆT

1. Tổng quan hệ thống

Hệ thống được thiết kế để xử lý các câu hỏi trắc nghiệm tiếng Việt thông qua cơ chế phân loại truy vấn thông minh và các đường ống xử lý (pipelines) chuyên biệt. Mục tiêu cốt lõi là tối ưu hóa độ chính xác dựa trên bản chất của từng loại kiến thức: từ suy luận logic STEM đến trích xuất thông tin văn bản (RAG) và kiến thức xã hội diện rộng (Reasoning).



Hình 1: Tổng quan hệ thống của Oversitting

2. Giai đoạn 1: Phân loại truy vấn (Query Classification)

Mọi câu hỏi đầu vào đều đi tầng phân loại sử dụng LLM với kỹ thuật **Few-shot Learning** (cung cấp một vài ví dụ mẫu trong prompt để mô hình hiểu ngữ cảnh và quy tắc phân loại).

Hệ thống phân chia thành 3 nhóm chính:

- **Reasoning:** Câu hỏi kiến thức xã hội, lịch sử, đời sống hoặc các vấn đề mới trong xã hội.
    - *Subtype Precision-Critical (PC):* Nội dung nhạy cảm, vi phạm pháp luật hoặc đạo đức.
    - *Subtype Multi-domain:* Kiến thức tổng hợp đa lĩnh vực.
- **STEM:** Câu hỏi về Khoa học tự nhiên, yêu cầu tính toán và tư duy logic.
- **RAG (Internal):** Câu hỏi dựa trên dữ liệu đi kèm, yêu cầu trích xuất thông tin trực tiếp từ ngữ cảnh cho sẵn.

3. Giai đoạn 2: Các Pipeline xử lý chuyên biệt

3.1. Pipeline xử lý Reasoning & External Knowledge base

Đối với các câu hỏi xã hội, hệ thống sử dụng một Kho lưu trữ kiến thức (Knowledge Base) khổng lồ được lập chỉ mục (Indexing) bằng FAISS Vector Database sau khi thực hiện embed bằng VNPT Embeddingg.

A. Xử lý Precision-Critical (PC) & Fallback

- **Refusal Prompt Construction:** Thiết lập các chỉ dẫn để LLM từ chối trả lời một cách lịch sự nhưng kiên quyết đối với các nội dung độc hại.
- **Fallback Mechanism:** Nếu LLM gặp khó khăn, hệ thống kích hoạt bộ quy tắc (Rule-based) để nhận diện và trả về các câu trả lời mặc định như: *"Tôi không thể trả lời câu hỏi này do vi phạm chính sách nội dung."*

B. Xử lý Multi-domain & External Knowledge Base

Kiến thức được chia nhỏ (chunking) và chuyển đổi thành vector thông qua VNPT Embedding để tìm kiếm tương đồng. Dưới đây là thống kê kho dữ liệu:

Chủ đề	**Nguồn dữ liệu**	**Số lượng Chunk (Ước tính)**
Luật pháp	Nghị định, nghị quyết, luật pháp hiện hành Việt Nam.	42227
Lịch sử	Lịch sử Việt Nam toàn thư, Lịch sử quốc tế tổng hợp.	3771
Triết học	Tư tưởng Mác - Lênin và Tư tưởng Hồ Chí Minh.	355
Dân số	Dữ liệu Crawl từ Tổng cục Thống kê và web dân số Việt Nam.	618
Kiến thức phổ thông	SGK, đề cương tổng hợp Văn, Sinh, Địa.	3419
Kinh tế	Báo cáo kinh tế Việt Nam, dữ liệu tài chính vĩ mô.	190
Tôn giáo	Tổng hợp kiến thức Phật giáo, Công giáo, Thiên chúa giáo.	1355
Địa lý	Thông tin thổ nhưỡng, phân bố địa hình Việt Nam.	515
Thông tin cập nhật	Xác nhập tỉnh thành, dự án trọng điểm quốc gia 2026.	14

3.2. Pipeline xử lý STEM (Logic & Tính toán)

Pipeline này tập trung vào khả năng tự thân của mô hình thông qua Prompt Engineering.

- **Kỹ thuật sử dụng:** Kết hợp Chain-of-Thought (CoT) (Chuỗi suy nghĩ) và Few-shot Learning.
- **Mục tiêu:** Ép mô hình thực hiện các bước trung gian (phân tích đề -> liệt kê công thức -> thay số) trước khi đưa ra đáp án cuối cùng, giúp giảm thiểu hiện tượng "ảo giác" số liệu.

Chi tiết prompt:

Bạn là một chuyên gia giải đề thi STEM (Khoa học, Công nghệ, Kỹ thuật, Toán học) với độ chính xác tuyệt đối.

NHIỆM VỤ:

Giải quyết câu hỏi trắc nghiệm dưới đây bằng phương pháp suy luận từng bước (Chain-of-Thought).

QUY TẮC BẮT BUỘC:

1. SUY LUẬN: Phân tích đề bài, xác định công thức hoặc lý thuyết liên quan.

2. TÍNH TOÁN: Nếu có số liệu, hãy viết phép tính rõ ràng, thay số từng bước. Không được làm tắt.

3. KẾT LUẬN: Sau khi suy luận xong, bắt buộc phải chốt đáp án ở dòng cuối cùng theo định dạng:

ANSWER: X

(Trong đó X là ký tự A, B, C, D, hoặc các ký tự khác tương ứng với đáp án đúng).

VÍ DỤ MẪU (Hãy làm theo format này):

CÂU HỎI:

Một vật rơi tự do từ độ cao h = 20m, lấy g = 10m/s². Thời gian rơi của vật là:

A. 1s

B. 2s

C. 3s

D. 4s

SUY LUẬN:

- Đây là bài toán rơi tự do.

- Công thức tính thời gian rơi: t = sqrt(2h / g).

- Thay số vào công thức:

  h = 20m

  g = 10m/s²

  t = sqrt(2 * 20 / 10) = sqrt(40 / 10) = sqrt(4) = 2 (giây).

```
- So sánh với các lựa chọn:

  A. 1s (Sai)

  B. 2s (Đúng)

  C. 3s (Sai)

  D. 4s (Sai)

- Vậy đáp án đúng là B.

### ANSWER: B

---

BÂY GIỜ LÀ CÂU HỎI CỦA BẠN:

CÂU HỎI:

{question}

CÁC LỰA CHỌN:

{mapped_choices}

SUY LUẬN:
```

3.3. Pipeline xử lý RAG (Internal Context)

Dành cho các câu hỏi mà đáp án nằm ngay trong đoạn văn bản đi kèm.

1. **Chunking & Indexing:** Đoạn văn bản trong câu hỏi được chia nhỏ, embed bằng VNPT Embedding API và đưa vào Vector Database tạm thời.
2. **Retrieval:** Sử dụng vector của chính câu hỏi để truy vấn Top K=3 chunk có độ tương đồng cao nhất.
3. **Generation:** LLM tổng hợp thông tin từ 3 chunk này để đưa ra đáp án, prompt được thiết kế để đảm bảo không lấy kiến thức bên ngoài gây sai lệch ý đồ của đề bài, cũng như yêu cầu model trả về đoạn phân tích trước khi trả lời

Chi tiết prompt:

```
Bạn là chuyên gia đọc hiểu và suy luận đáp án từ đoạn thông tin được cung cấp.
Nhiệm vụ của bạn là trả lời câu hỏi trắc nghiệm dựa trên thông tin đó.
Nếu không có đủ thông tin, hãy chọn đáp án phù hợp nhất với đoạn thông tin.
---
Hướng dẫn xử lý:
- Đọc từng câu trong "Đoạn thông tin" một cách tuần tự.
- Với mỗi câu, hãy hiểu ngữ cảnh của nó trong mối liên hệ với câu trước và sau nó.
- Đối chiếu câu hỏi với các chi tiết vừa đọc để tìm bằng chứng chính xác.
- Sau khi phân tích, hãy đưa ra đáp án cuối cùng.
- [QUAN TRỌNG] Phân tích ngắn gọn, không lặp lại, tối đa 250 cho PHÂN TÍCH.

Định dạng trả về bắt buộc (bạn phải tuân thủ khuôn mẫu này):
```

[PHÂN TÍCH]
(Viết quá trình đọc hiểu và suy luận từng bước tại đây, giới hạn suy luận dưới 250 từ)

[ĐÁP ÁN]
(Duy nhất một chữ cái: A, B, C hoặc D, không giải thích thêm)

4. Kết luận

Hệ thống RAG này đảm bảo tính toàn diện bằng cách tách biệt giữa việc truy xuất kiến thức (Reasoning/RAG) và khả năng tư duy (STEM). Việc sử dụng vector database kết hợp với các chiến lược Prompting giúp hệ thống vận hành ổn định trên các tập dữ liệu tiếng Việt phức tạp và luôn cập nhật các kiến thức mới nhất.