



# Hệ thống phát hiện và làm nổi bật bình luận thù ghét trên mạng xã hội tiếng Việt

IE403.Q11-KHAI THÁC DỮ LIỆU TRUYỀN THÔNG XÃ HỘI

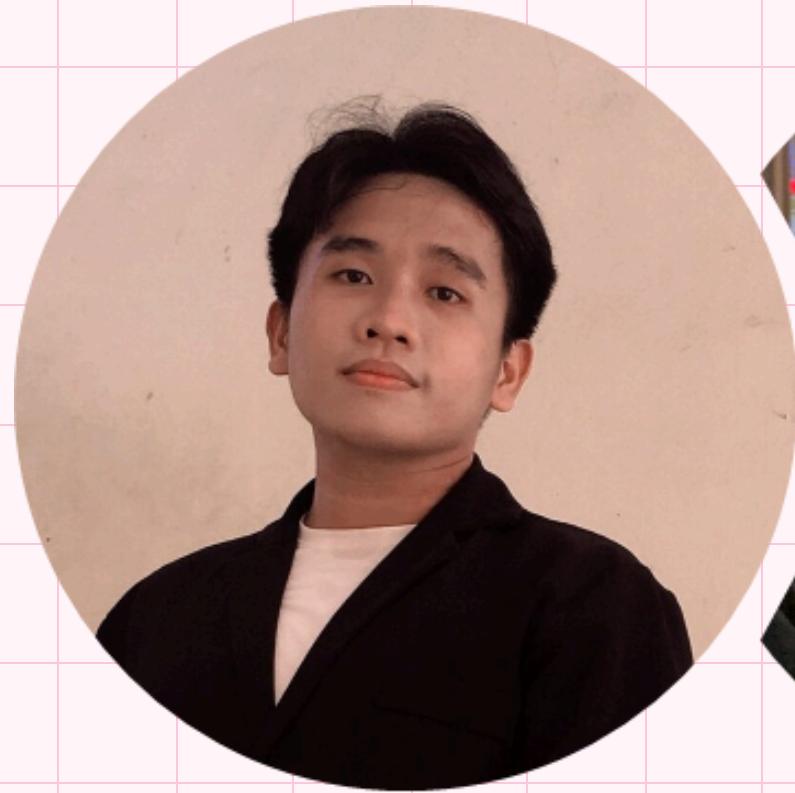
NHÓM 2

GVHD: TS. NGUYỄN VĂN KIỆT

TH.S HUỲNH VĂN TÍN



# DANH SÁCH THÀNH VIÊN



Công Phát  
23521143



Xuân An  
23520023



Thái Bình  
23520158



Thành An  
23520032

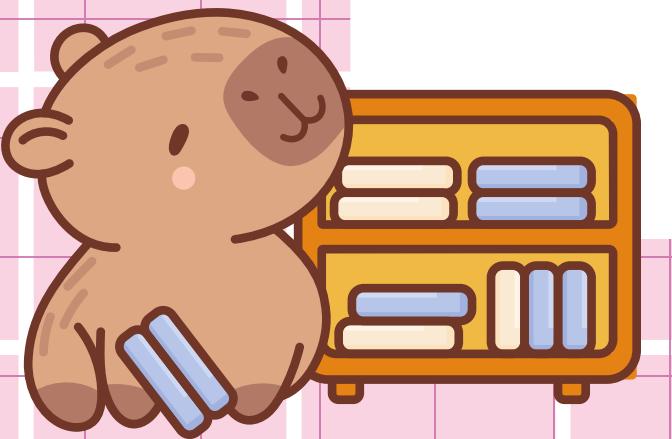


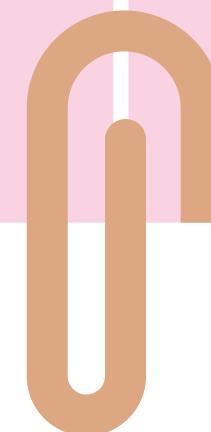
Quỳnh Hương  
21520255



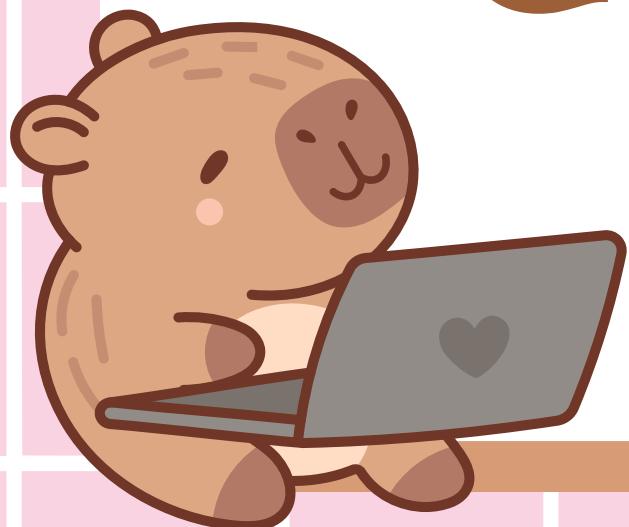
# MỤC LỤC

1. GIỚI THIỆU BÀI TOÁN
2. BỘ DỮ LIỆU
3. VẤN ĐỀ: CÂU ẨN Ý
4. PHƯƠNG PHÁP THỰC HIỆN
5. ĐỘ ĐO ĐÁNH GIÁ & KẾT QUẢ
6. KHÓ KHĂN & HƯỚNG PHÁT TRIỂN
7. DEMO HỆ THỐNG





# GIỚI THIỆU BÀI TOÁN



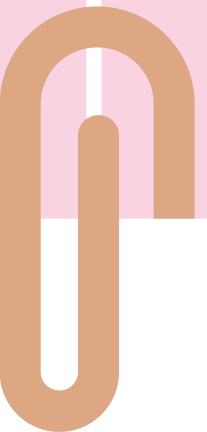


# GIỚI THIỆU BÀI TOÁN

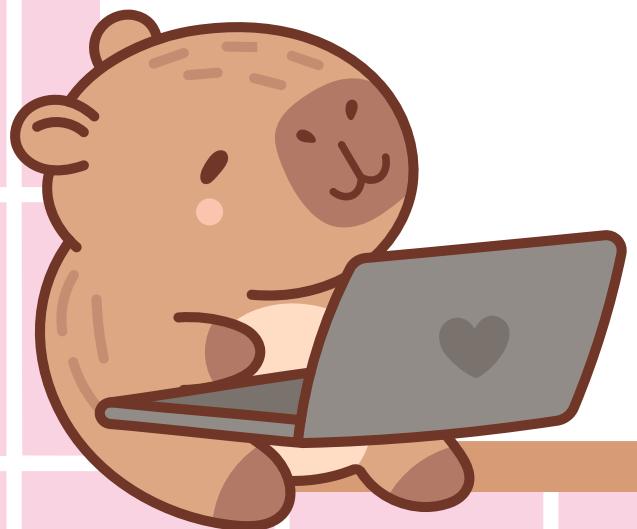
- **Sự bùng nổ của độc hại trực tuyến (Toxic Speech)**
- **Đặc thù phức tạp của Tiếng Việt**
- **Sự quá tải của kiểm duyệt thủ công**

Chính vì vậy, việc xây dựng một Hệ thống phát hiện và làm nổi bật bình luận thù ghét chuyên biệt cho tiếng Việt là một yêu cầu cấp bách.





# BỘ DỮ LIỆU





# BỘ DỮ LIỆU:

**Nguồn:** Dữ liệu từ ViHSD, được làm sạch và gán nhãn thủ công.

**Số lượng:** 10.000 bình luận mạng xã hội, gán nhãn bởi nhiều annotator.

**Định dạng nhãn:** [target]#[level]

- **Target:** gồm 5 mục tiêu Cá nhân, nhóm, tôn giáo, Sắc tộc và Chính trị.
- **level:** gồm 4 mức độ normal, clean, offensive và hate.

Mục tiêu cụ thể	Tập Huấn luyện (Train)	Tập Phát triển (Dev)	Tập Kiểm thử (Test)
Individuals	5,480	938	1,398
Groups	2,977	517	769
Religion	24	8	6
Race/Ethnicity	502	74	129
Politics	363	57	89

Tập dữ liệu	NORMAL (0)	CLEAN (1)	OFFENSIVE (2)	HATE (3)
Train	1,520	2,480	1,169	1,831
Dev	263	454	189	295
Test	402	618	324	456
Tổng	2,185	3,552	1,682	2,582

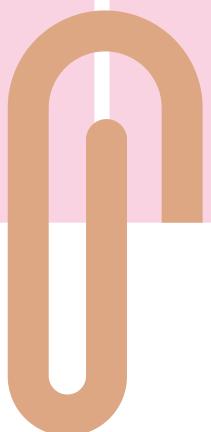


# BỘ DỮ LIỆU:

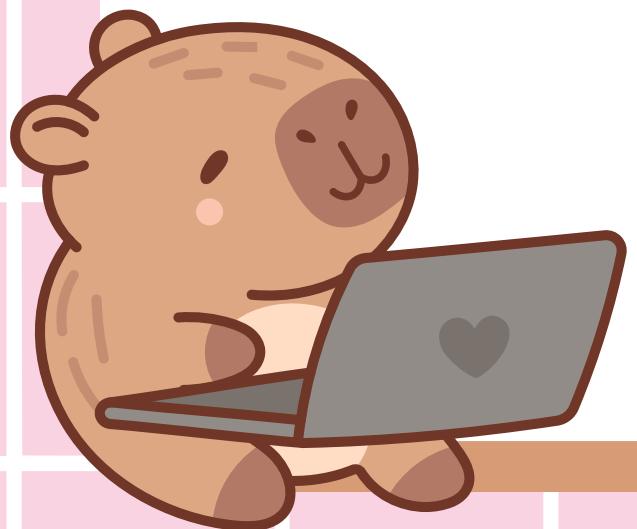
**Ghi chú:** Một bình luận có thể chứa nhiều nhãn, mỗi target được gán mức độ thù ghét riêng biệt.

- Nếu target không được đề cập, nhãn là NORMAL

COMMENT	LABEL
“Hiếu Bùi sửa do”	individuals#hate
“Em mình cũng dễ thương nè ❤”	individuals#clean group#clean
“wibu rách”	individuals#hate group#hate
“Optimusthree với 99% sức mạnh”	normal



# VẤN ĐỀ: CÂU ẨN Ý





# VẤN ĐỀ: CÂU ẨN Ý

Phân tích lỗi cho thấy, các mô hình chưa thể dự đoán chính xác cho các câu implied statement

**BẢNG SO SÁNH HIỆU NĂNG CÁC MÔ HÌNH (MULTI-LABEL)**

Metrics	Qwen	PhoBERT	FlanT5
F1 Micro	0.59	0.5412	0.4684
F1 Macro	<b>0.331</b>	0.2586	0.1311
Precision Micro	0.61	0.562	0.481



# GIẢI PHÁP: TẠO GIẢI THÍCH CHO CÁC CÂU ẨN Ý

Dự án áp dụng **HARE Framework** kết hợp với **CoT prompting** để bổ sung hai trường mở rộng cho các bình luận có nhãm hate và offensive:

- **implied\_statement:** nói thẳng ra được điều mà người viết muốn trong câu.
- **rationale:** lời giải thích từng bước tại sao câu bị đánh giá là thù ghét (3–5 bullet ngắn gọn, bám vào nội dung comment), chỉ rõ:
  - đối tượng bị nhắm tới
  - từ/cụm từ mang tính hạ nhục, miệt thị, định kiến
  - cách diễn đạt mỉa mai/ẩn dụ làm tăng tính công kích

```
"id": 532,  
"content": "Thằng ad page nào đúng là Óc Chó 😤",  
"dataset": "train",  
"labels": "individuals#hate"  
"implied_statement": "admin ngu dốt",  
"rationale": [  
    "Việc gọi đối tượng là 'thằng ad page' theo sau bởi 'Óc Chó' nhằm hạ giá trị con người của họ.",  
    "Cách dùng 'Óc Chó' mang tính chất hạ thấp thông minh và nhân phẩm.",  
    "Ngôn ngữ thô tục được sử dụng để kích liệt tấn công cá nhân.",  
    "Cách phát biểu này phản ánh thái độ cực kỳ tiêu cực và định kiến rằng admin là kém cỏi." ]
```



## FRAMEWORK HARE

Là một phương pháp luận hướng dẫn cách tạo dữ liệu reasoning nhằm khắc phục “lỗ hổng suy luận” của mô hình trong nhận diện hate speech:

- Tận dụng khả năng suy luận của các LLM hiện đại, đặc biệt qua kỹ thuật chain-of-thought reasoning (CoT).
- Sinh lập luận từng bước (rationale) bằng CoT prompt để lấp đầy khoảng trống lý luận trong các chú thích hiện tại:
  - giúp mô hình hiểu được mối liên kết giữa ‘content’ và ‘implied\_statement’

### Mục tiêu:

- Làm lộ ra “implicit hate” – dạng thù ghét không biểu hiện trực tiếp.
- Buộc mô hình phải suy luận thay vì đoán mò.
- Khắc phục trường hợp hate mơ hồ hoặc bóng gió (ví dụ: chính trị, dân tộc, tôn giáo nhưng không dùng từ chửi).



## COT PROMPT

Là kỹ thuật thiết kế prompt yêu cầu mô hình tạo rationale có cấu trúc theo từng bước trước khi đưa ra kết luận.

### Lí do sử dụng:

LLM được tiền huấn luyện trên lượng lớn văn bản nên đã học được:

- quan hệ giữa ngôn ngữ và ý định (intent)
- mô thức lập luận: “dẫn chứng → phân tích → kết luận”
- khả năng paraphrase: viết lại một ý nghĩa theo cách khác

CoT prompting khai thác tốt năng lực đó bằng cách cung cấp khung suy luận từ đó:

- ép mô hình giảm hiện tượng nhảy cóc (trả lời cảm tính)

### So sánh với prompt bình thường:

- Với prompt thường: LLM dễ trả lời ngắn, đôi khi dựa vào keyword hoặc suy đoán cảm tính.
- Với CoT prompting: LLM không chỉ dựa vào keyword, mà dựa vào ngữ cảnh, sắc thái, mục tiêu công kích.

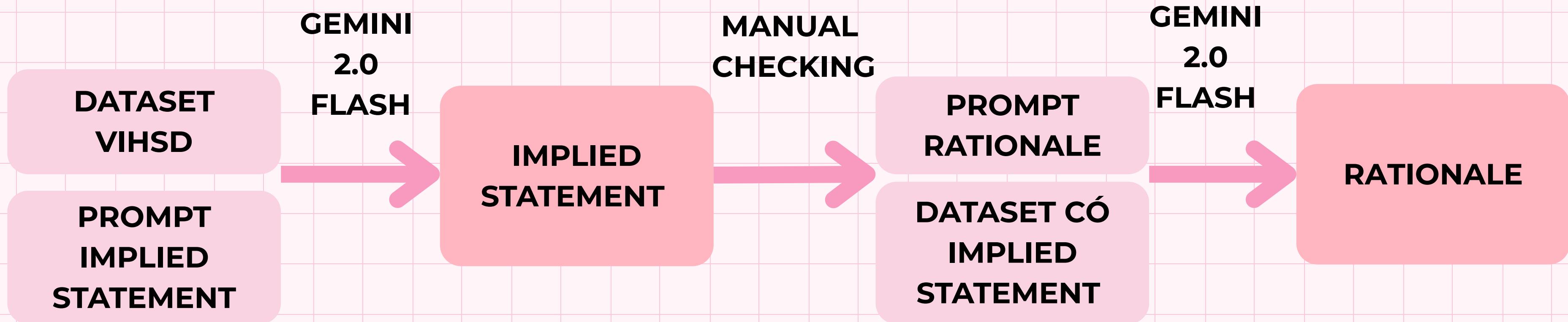


# THIẾT KẾ PROMPT

Aspect	V1: Base	V2: Updated	V3: Final <span style="color: green;">✓</span>
Cấu trúc	Role + Task + I/O	Constraints + Examples	+ Framework (4 bước)
Implied	Tự do	"Hidden meaning"	<b>3-8 từ, ngắn gọn</b>
Rationale	Dài, tự do	false positive	<b>CoT:</b> <b>target→implied→evidenc</b> ..
Constraints	Không có	Few-shot	<b>Length + No hallucination + Fixed steps</b>
Output	Không chuẩn	Có chiều sâu nhưng dài	<b>JSON strict/escape</b>



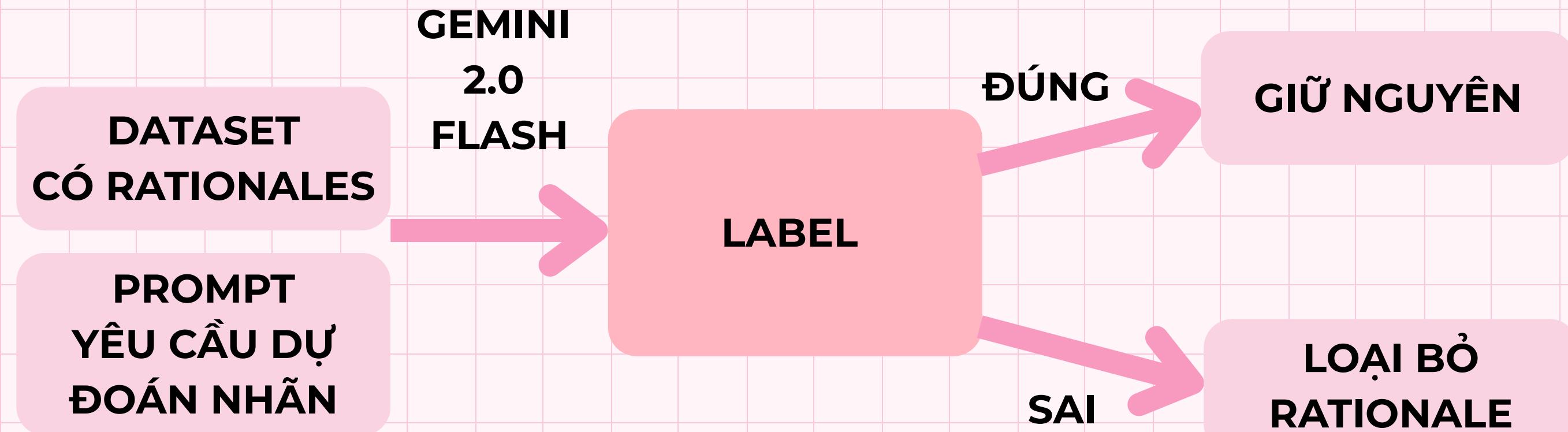
# Quy trình: tạo implied statement & rationale



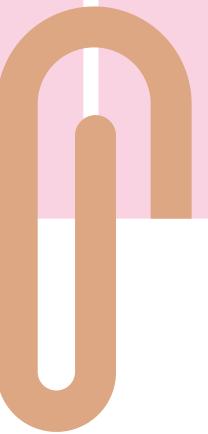
- Sử dụng Gemini Flash 2.0 để gán target và implied ⇒ 4 thành viên trong nhóm check lại
- Thiết kế lại prompt của nghiên cứu gốc ⇒ phù hợp với dataset và bài toán của nhóm
- Dùng Gemini Flash 2.0 để tạo rationale



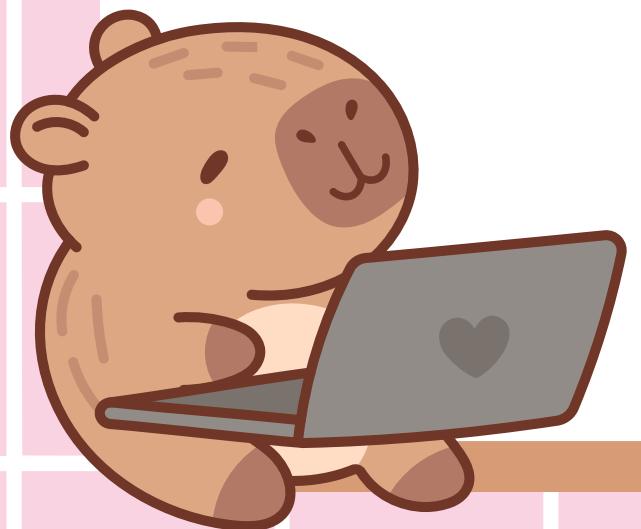
# Quy trình: loại bỏ rationales gây nhiễu



- Nếu dự đoán nhãn giống với nhãn thật của bài đăng, giữ rationale  $\Rightarrow$  train model
- Nếu dự đoán nhãn sai, thì không dùng rationale vì nó có thể sai lệch hoặc gây nhiễu.  
 $\Rightarrow$  Mục đích là đảm bảo mô hình chỉ học các lời giải thích phù hợp với nhãn đúng, tránh nhiễu do rationale “lạc đề”



# PHƯƠNG PHÁP THỰC HIỆN





# TWO-STAGE TRAINING WITH RATIONALE-BASED SEMANTIC ALIGNMENT

**Stage 1:** học “decision boundary” từ dataset VIHSD.

**Stage 2:** dùng rationale/implied statement để “căn chỉnh ngữ nghĩa”

- biến content ẩn ý → gần nghĩa với implied statement (diễn giải rõ)
- giúp model “hiểu” implicit hate tốt hơn

**Lựa chọn mô hình:** Qwen2.5-3B-Instruct

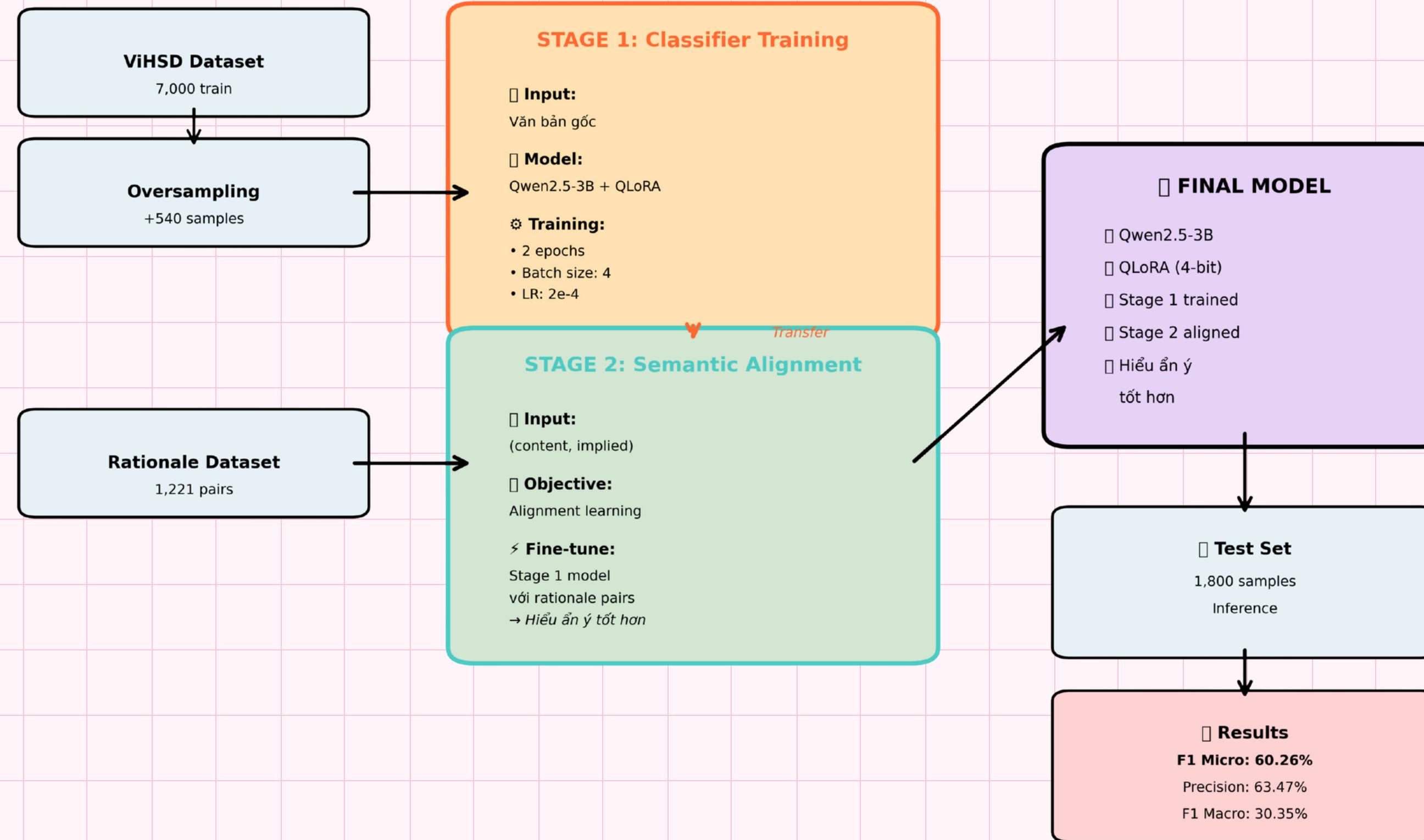
- Hỗ trợ tốt tiếng Việt
- 3B params - vừa phải cho T4 GPU
- Instruction-following tốt

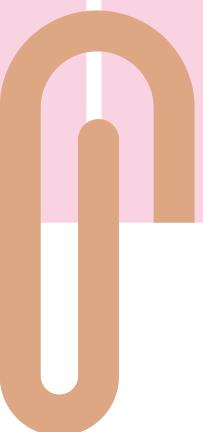
**Kỹ thuật:** QLoRA (4-bit)

- Giảm memory: 3B → ~2GB VRAM
- Tăng tốc: 2-3x nhanh hơn
- Duy trì chất lượng: ~95%

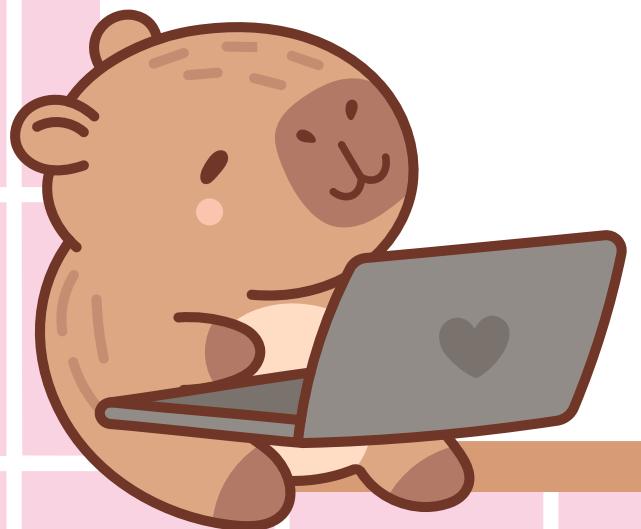


# PIPELINE





# ĐỘ ĐO & KẾT QUẢ





# ĐỘ ĐO & KẾT QUẢ

## F1 Micro

- Tính trên tổng số TP, FP, FN của tất cả labels
- Phản ánh độ chính xác tổng thể của model

## F1 Macro

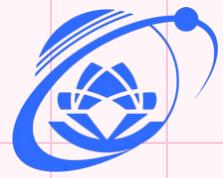
- Tính F1 cho từng label, sau đó lấy trung bình
- Công bằng với tất cả labels (kể cả minority)
- Phát hiện vấn đề với imbalanced data

## Precision Micro

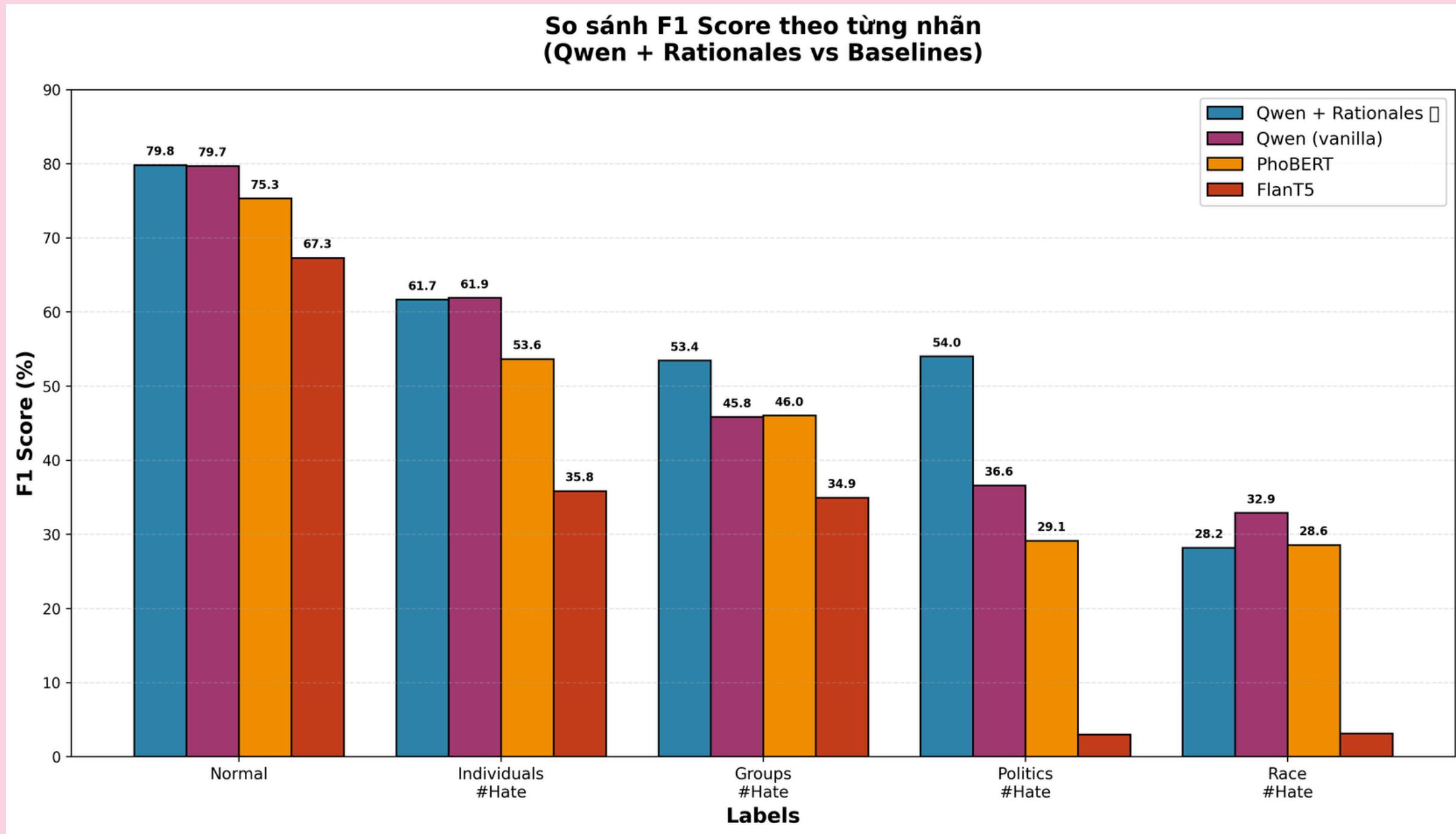
- Phản ánh độ chính xác tổng thể của mô hình trên toàn bộ dữ liệu.

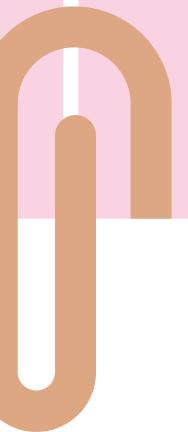
BẢNG SO SÁNH HIỆU NĂNG CÁC MÔ HÌNH

Metrics	Qwen (Rationles)	Qwen	PhoBERT	FlanT5
F1 Micro	0.6026	0.59	0.5412	0.4684
F1 Macro	0.3035	0.331	0.2586	0.1311
Precision Micro	0.6347	0.61	0.562	0.481

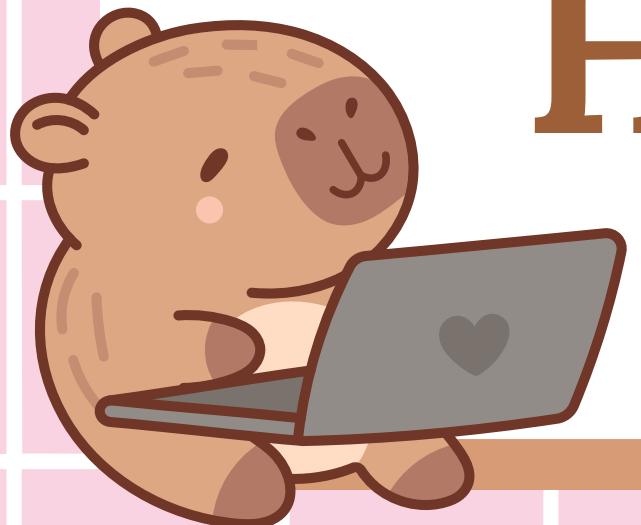


# ĐỘ ĐO & KẾT QUẢ





# KHÓ KHĂN & HƯỚNG PHÁT TRIỂN





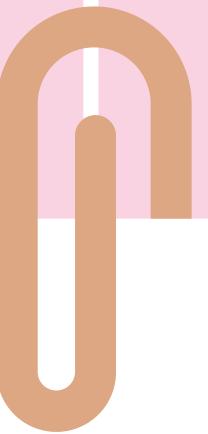
# KHÓ KHĂN & HƯỚNG PHÁT TRIỂN

## Khó khăn

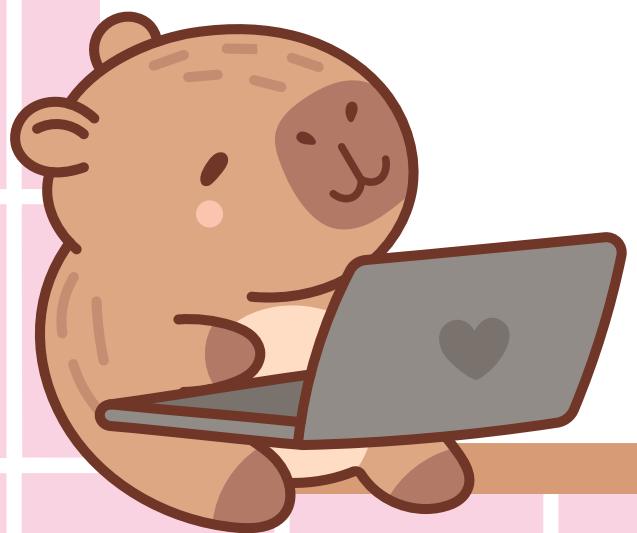
- Thiếu tài nguyên để thực nghiệm chính xác hơn
- Dữ liệu còn hạn chế đối với các đối tượng race, religion

## Hướng phát triển tương lai

- thêm dữ liệu để giải quyết vấn đề mất cân bằng
- kết hợp thêm ảnh và cả 1 bài post
- cải thiện score



# DEMO





Thank you

