Hate Speech Detection and Highlighting in Vietnamese Social Media Comments

Phat Nguyen Cong[1,2], **An Nguyen Xuan**[1,2], **Binh Mai Thai**[1,2],
An Truong Hoang Thanh[1,2], **Huong Nguyen Le Quynh**[1,2]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{23521143, 23520023, 23520158, 23520032, 21520255}@gm.uit.edu.vn

Abstract

The rise of hate speech on Vietnamese social media has created a real need for systems that can understand deep context rather than just a simple classification. In this work, we tackle the challenge of target-specific and implicit hate speech by introducing the HARE framework. By integrating Chain-of-Thought (CoT) reasoning into the Qwen2.5-3B model, our approach moves beyond traditional "black-box" models. Instead of just providing a classification label, HARE generates semantic explanations to reveal the hidden intent behind a text, using a two-stage training process on the ViTHSD dataset. Our experiments show that the model achieves an F1-score of 60.26%, outperforming benchmarks like PhoBERT and Flan-T5, especially in the politics category with a notable 24.87% jump. Finally, we provide a practical architecture for deploying this solution on large-scale streaming platforms like Apache Kafka and Spark, integrating highlight keywords features, proving that our approach is ready for real-world use.

1 Introduction

1.1 Overview of Social Media Challenges in Vietnam

In recent years, social media has become a central part of public life in Vietnam. Platforms such as Facebook, YouTube, and TikTok are now the main places where people discuss everyday topics. However, this openness has also led to a rise in toxic language, creating a serious challenge for keeping these online spaces safe.

One major difficulty in detecting this content comes from the special features of the Vietnamese language. Vietnamese is monosyllabic and tonal, and meaning often depends on cultural context. Today, users often use insulting slang like *"ba que"* or *"bò đỏ"* along with teencode, sarcasm, and wordplay. These creative ways of speaking help harmful content avoid basic keyword filters, making it hard for both automated systems and human moderators to control.

Therefore, using modern technologies such as pre-trained language models like PhoBERT still has some limitations:

- Lack of Explainability: Deep learning models often operate as "black boxes" producing classification labels without providing specific reasoning, which makes it difficult to build trust and transparency in content moderation.
- Limitation in recognizing implicit hate: Most current models focus so much on surface-level keywords that they only catch obvious swearing. They often fail to detect "hidden attacks" because these insults rely on subtle comparisons and cultural context rather than explicit hate speech.
- Limitation in recognizing target: Most previous research has focused almost entirely on how toxic a comment is while overlooking the specific target of the attack. Identifying whether the victim is an individual, an organization, or an ethnic group is actually a vital piece of information. Knowing exactly who is being targeted is essential for assessing the severity of the speech and determining the most appropriate way to intervene.

To address these practical challenges, we focus on building an automated system that can explain both "why" a piece of content is considered hateful and "who" is being targeted. This objective serves as the foundation for our work with the ViTHSD dataset and the modeling approaches that we present in this paper.

1.2 Objectives of the report

This report aims to address the above limitations by developing a comprehensive system based on the ViTHSD dataset. The key contributions include:

1. Systematizing the Targeted Hate Speech Problem: We redefine the hate speech detection problem in Vietnamese from a target-oriented perspective, using the ViTHSD dataset with a multi-layered label structure.

2. Applying the HARE Framework and CoT: Proposing a process to integrate the reasoning capabilities of LLMs to generate explanations (rationales) and implied statements, helping the model better understand the context.

3. Two-Stage Semantic Alignment Training Strategy: Introducing the two-stage training method on the Qwen2.5-3B model, which enables the model to learn decision boundaries from labeled data and reasoning thinking from rationale data.

4. Experiments and Evaluation: Providing experimental results comparing with the strongest current baselines (PhoBERT, FlanT5) and a deep analysis of the method's effectiveness on challenging labels such as Politics and Group Hate.

5. Introducing the practical deployment architecture: Introducing an AI-integrated system for real-time data streaming using Kafka and Spark, with FastAPI for the backend and React and Vite for the frontend, designed for large-scale social media monitoring applications.

2 Literature Review and Theoretical Background

2.1 The Development of Vietnamese NLP and Pre-trained Models

Natural Language Processing (NLP) for Vietnamese has undergone a major transformation over the last decade. Before 2019, most research relied on hand-crafted features and static models like Word2Vec (Mikolov et al., 2013) or FastText (Bojanowski et al., 2017), but the arrival of the Transformer architecture (Vaswani et al., 2017) and Google's BERT (Devlin et al., 2019) completely redefined the landscape.

In Vietnam, the release of PhoBERT (Nguyen and Nguyen, 2020) in 2020 marked a significant turning point. Trained on a massive 20GB corpus of Vietnamese news and Wikipedia entries, PhoBERT set new state-of-the-art standards for tasks ranging from part-of-speech tagging to named entity recognition. However, because PhoBERT was built on such formal text, it faces a clear "domain gap" when applied to social media. The messy reality of online comments, filled with slang, teencode, and informal grammar, often proves too difficult for these traditional models to handle.

The latest trend involves using Large Language Models (LLMs) with strong instruction-following capabilities, such as Alibaba's Qwen series. Qwen2.5 (**?**) has emerged as a particularly strong candidate because it offers exceptional support for the Vietnamese language. Even when compared to Llama (Touvron et al., 2023) or Gemma (Mesnard et al., 2024) at the same parameter size, Qwen consistently delivers better results due to its exposure to a more diverse and high-quality training dataset.

2.2 Hate Speech Datasets in Vietnam

The foundation of any AI system is its data. In Vietnam, the ViHSD (Vietnamese Hate Speech Dataset) (Luu et al., 2021) was the first large collection of data to be made public, containing over 30,000 comments labeled into three categories: Clean, Offensive, and Hate. This dataset has been very important for local research.

However, to better understand these interactions, the ViTHSD (Vo et al., 2025) (Vietnamese Targeted Hate Speech Dataset) was created. While it builds on the data from ViHSD, it adds an important new detail: the "Target." This approach is similar to Aspect-Based Sentiment Analysis (ABSA) (Pontiki et al., 2014), where the model doesn't just identify "what the emotion is", but also "what the emotion is directed at".

In ViTHSD, each comment is checked to see if it targets an Individual, a Group, a Religion, an Ethnicity, or a Political entity, and it also measures how strong the hate is. This detailed information is important for creating smart moderation systems that can keep online spaces safe without stopping healthy discussions.

2.3 Explainable AI and Chain-of-Thought Prompting

Explainability is becoming a must-have feature in modern AI. In the field of hate speech detection, the HARE framework (Yang et al., 2023) introduced at EMNLP 2023 proposed a fresh approach by using Large Language Models to generate written explanations. Instead of relying only on classification labels, the model is trained on both the labels and the specific reasoning behind them.

The core technique for creating these high-quality explanations is Chain-of-Thought prompting (Wei et al., 2022). Research from Google and OpenAI has shown that asking a model to think step by step activates its internal reasoning abilities. This allows the model to solve complex problems where standard prompting methods often fail.

In the context of the ViTHSD dataset, we use this technique to force the model to analyze sentence structures, identify offensive terms, and pinpoint the target before it reaches a final conclusion. This approach prevents the model from simply guessing based on keywords and encourages a much deeper understanding of the actual meaning of the text.

3 Analyzing the ViTHSD dataset

3.1 Structure and labels definition

The ViTHSD dataset was built based on 10,000 selected and cleaned comments from the ViHSD dataset. The labels of ViTHSD is the combination of Target and Level, represented in the format [target]#[level].

Targets:

1. Individuals: Comments targeting a specific individual. This is the most common form of online abuse.

2. Groups: Comments directed at organizations, communities, groups, or fandoms.

3. Religion: Comments targeting religious beliefs, religious figures, or religious communities.

4. Race/Ethnicity: Comments targeting regions, ethnicities, or origins. This is the most sensitive and dangerous category.

5. Politics: Comments related to political views, political parties, government policies, or political figures.

Levels:

- Normal (0): A normal comment that does not contain negative content and does not target any specific entity.

- Clean (1): The comment mentions a target, but the content is clean, neutral, or positive.

- Offensive (2): Contains vulgar, insulting, or mocking language, but does not reach the level of hate speech or incitement to violence.

- Hate (3): Hateful language that attacks or dehumanizes targets, incites violence, promotes discrimination, or uses severe abusive expressions.

A key feature of ViTHSD is its multi-label nature. A comment may contain multiple targets, and each target can have a different level. For example: *"Thằng A ngu thế cũng làm ca sĩ (`Individuals#Offensive`), bọn fan của nó thì cũng súc vật không kém (`Groups#Hate`)"*

3.2 Statistical Analysis of Data Imbalance

When analyzing the dataset, we obtain the following detailed distribution table:

Table 1: Distribution of samples by Target (Train/Dev/Test)

Target	Train	Dev	Test	Total	%
Individuals	5,480	938	1,398	7,816	∼60%
Groups	2,977	517	769	4,263	∼33%
Race/Ethni.	502	74	129	705	∼5%
Politics	363	57	89	509	∼4%
Religion	24	8	6	38	<0.3%

Our data reveals a significant imbalance across categories. The Individuals group makes up the vast majority of the dataset, reflecting the reality of social media where personal disputes are most frequent. In contrast, the Politics and Religion categories have very few samples. This imbalance creates a major hurdle: traditional machine learning models tend to bias toward the majority class. As a result, while these models may become effective at spotting personal attacks, they often struggle with hateful comments regarding religion or politics because they haven't seen enough examples to generalize properly. The Religion category is particularly challenging; with only 24 training samples, it is nearly impossible for a standard deep learning model to converge without employing specialized techniques like Data Augmentation or Few-shot Learning.

Table 2: Distribution by Intensity (Level)

Dataset	Normal	Clean	Offen.	Hate	Total
Train	1,520	2,480	1,169	1,831	7,000
Dev	263	454	189	295	1,201
Test	402	618	324	456	1,800
Total	**2,185**	**3,552**	**1,682**	**2,582**	**10,001**

Even though the distribution by Levels is more balanced than by Targets, the CLEAN labels still have the highest percentage. This requires our models to have a good Recall to avoid missing the case HATE labels hidden among a large number of CLEAN comments.

3.3 Implied Statement problem

An analysis of the ViTHSD dataset reveals that implicit hate speech is widespread. These comments avoid direct profanity but still carry a deeply offensive message through underlying meaning.

- Example 1: *"Optimusthree với 99% sức mạnh"*.
  - Surface level analysis: it shows no offensive words, looks like a compliment about strength.
  - Real-world context: This is a sarcastic statement in the gaming community, implying that the player is very weak or performing poorly (only 1% capability left). If the model does not understand this "meme" context, it will label it as CLEAN.
- Example 2: *"Lũ ba que xỏ lá."*
  - Analysis: "Ba que" is a political slang term (referring to the yellow flag with three red stripes of the former regime), and *"xỏ lá"* is an adjective meaning deceitful or tricky. Combined, this forms a hate speech statement targeting a political group (`Politics#Hate`).

To resolve this, we added two information fields for labels marked as hate or offensive.

1. Implied_statement: A short paraphrase that reveals the speaker's actual intent (For example: *"người chơi này rất tệ"*, *"nhóm người này phản động"*).
2. Rationale: a logical explanation of why the comment is hateful based on text evidence.

The next section explains the process for creating this information, which is the core of the research methodology.

4 Methodology

We will use a pipeline that uses commercial LLMs to generate data and open-source LLMs for training and deployment under the HARE framework.

4.1 Designing prompts and generating reasoning data.

Creating high-quality rationales and implied statements for thousands of comments is a massive undertaking—doing it all by hand is simply too expensive and time-consuming to be practical. To solve this, we used Gemini 2.0 Flash (DeepMind, 2025) to act as our automated annotator. We didn't just use the model out of the box; we developed a rigorous Prompt Engineering workflow to ensure the quality of the data. By iterating through three distinct versions of our instructions, we were able to fine-tune the model's output until it reached the level of detail and accuracy required for our framework.

4.1.1 Phase 1: Prompt base

- Design: The simple structure includes a Role, a Task, and an Input/Output format. The model is required to "summarize the main idea" and "explain why."
- Result: The results of this phase of testing did not meet the expected requirements.
  - Implied Statement: The model often just paraphrases the original sentence without revealing the hidden meaning.
  - Rationale: The explanations are too long and unfocused, not centered on the hate-related keywords. The output format is inconsistent, making automated processing difficult. The ability to distinguish between Hate and Offensive content is still limited.

4.1.2 Phase 2: Prompt updated

- Improvement: Add constraints and few-shot examples. Require the model to take the role of a "psychologist" to analyze the "hidden meaning."
- New issue: Domain Drift. Taking the psychologist role causes the model to analyze too deeply from philosophical and behavioral psychology perspectives, leading to mislabeling (false positives) for sentences that only express anger (Angry) but are not hate speech. In addition, the model uses too many tokens for unnecessary grammatical analysis.

4.1.3 Phase 3: Final prompt

This is the officially used version, integrating the HARE framework with a strict four-step reasoning structure:

1. Role & Objective: Clearly define the task as hate speech detection.
2. Definitions: Provide precise definitions of Hate / Offensive / Clean / Normal based on the ViTHSD standard.
3. Explicit Constraints: Enforce length limits (Implied statement: 3–8 words) and prohibit hallucination.

4. Reasoning Framework (Logic Flow): require the model to strictly follow the flow: Target → Implied → Evidence → Verdict.
    - Target: Identify the target.
    - Implied: Describe the hidden meaning or bias.
    - Evidence: Extract words or phrases with insulting or degrading meaning.
    - Verdict: Decide the label based on the evidence.
5. Output Format: Strict JSON formatting to ensure technical consistency and ease of parsing.

4.2 Building rationale dataset

Data generated by the LLM cannot be guaranteed to be 100% accurate. To remove noise, we apply a two-step filtering process:

1. Manual Verification: A random subset is reviewed by humans (5 team members) to evaluate the quality of the Implied Statement. This helps to fine-tune the prompt one last time.
2. Rationale Consistency Check: This is the most critical step for automated data cleaning.
    - Input: Feed the Original Content + Generated Rationale back into the Gemini Flash model.
    - Task: Verify if the rationale logically supports the ground truth label.
    - Logic:
        - If Predicted Label == Ground Truth Label → Keep the Rationale. This shows that the rationale is logical and leads to the correct conclusion.
        - If Predicted Label $\neq$ Ground Truth Label → Discard the Rationale. This rationale may be off-topic or contain flawed reasoning.
    - Result: From thousands of generated rationales, we filtered a high-quality dataset (Rationale Dataset) consisting of 1,221 sample pairs for training purposes.

4.3 Model and Two-Stage training strategy

We selected Qwen2.5-3B-Instruct as our backbone model. As a 3-billion-parameter, decoder-only model, it strikes a perfect balance between efficiency and performance. It is lightweight enough to be deployed on mid-range GPUs (such as the NVIDIA T4 in Kaggle or Colab environments), yet it consistently outperforms its peers in understanding the Vietnamese language and following complex instructions.

To ensure the model could handle both classification and reasoning effectively, we implemented a Two-Stage Training strategy:

Stage 1: Classifier Training

- Objective: Train the model to distinguish Target and Level classes based on the original text.
- Data: The full ViTHSD Train set (7,000 samples). Due to data imbalance, we applied oversampling (+540 samples) for minority classes (Politics, Religion) to reduce model bias.
- Technique: Use QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) in 4-bit. QLoRA allows fine-tuning a large model without updating all parameters, saving GPU memory while maintaining performance.
- Hyperparameters: 2 epochs, Batch size 4, Learning rate 2e-4.

Stage 2: Semantic Alignment

- Objective: Teach the model to understand the relationship between the original text and its hidden meaning (Implied statements).
- Data: Rationale Dataset (1,221 Content–Implied pairs).
- Technique: Further fine-tune the model (checkpoint from Stage 1) using the rationale data. This process adjusts the embedding space, bringing implied sentences closer to their true meaning. For example, the model learns that in a gaming context, “99% power” ≈ “weak/failed.”
- Impact: This is the breakthrough step that allows the model to move beyond rote keyword learning and toward deep semantic understanding.

5 Experiments and results

5.1 Experimental Setup

To evaluate the effectiveness of the proposed method, we compared it with the strongest baseline models currently available in Vietnamese NLP and generative AI:

1. PhoBERT-base (Nguyen and Nguyen, 2020): State-of-the-art encoder-only model for Vietnamese. This represents the traditional fine-tuning approach.

2. FlanT5-base (Chung et al., 2022): Encoder–Decoder model specialized in instruction tuning.

3. Qwen2.5-3B (Vanilla) (Yang et al., 2025): Qwen model with standard fine-tuning only (Stage 1), without any Rationale data.

4. Qwen2.5-3B + Rationales: Proposed model with the two-stage process.

The main metrics used are F1-Score Micro (overall evaluation) and F1-Score Macro (fair evaluation across classes), along with Precision and Recall.

5.2 Quantitative Results and Comparative Analysis

Based on the experimental results, we obtained the following metrics on the Test set:

Table 3: Overall performance comparison of the models (Multi-label)

Metrics	**Qwen (Rat)**	**Qwen (Van)**	**PhoBERT**	**FlanT5**
F1 Samples	**0.6345**	0.6332	0.5801	0.5074
F1 Micro	**0.6026**	0.5900	0.5412	0.4684
F1 Macro	0.3035	**0.3310**	0.2586	0.1311
Prec. Samples	**0.6515**	0.6469	0.5791	0.5033
Prec. Micro	**0.6347**	*0.6100*	0.5620	0.4810
Prec. Macro	**0.4067**	*0.3800*	0.3200	0.1800
Rec. Samples	0.6344	**0.6372**	0.5993	0.5210
Rec. Micro	**0.5735**	*0.5600*	0.5310	0.4520
Rec. Macro	**0.2842**	*0.2700*	0.2450	0.1100

Analysis:

- Superior Micro F1: The proposed model achieves 60.26%, the highest among all. It surpasses PhoBERT by +6.14% and FlanT5 by +13.42%, confirming the advantage of a modern decoder-only architecture combined with rationale data over older architectures.

- Impressive Precision: With Micro Precision of 63.47%, the model significantly reduces false positives. In content moderation, high precision is crucial to avoid wrongly blocking harmless user comments, ensuring a better user experience.

- Impact of Rationales: Compared to Qwen Vanilla, the rationale-enhanced version improves Micro F1 (+1.26%)and Precision (+2.47%). While the overall increase is modest, the real improvement lies in the hard-to-classify label classes (analyzed in Section 5.3).

Table 4: Detailed F1-Score analysis by Label

Label	**Qwen (Rat)**	**Qwen (Van)**	**Pho-BERT**	**Flan-T5**
Normal	0.7984	0.7970	0.7531	0.6732
Indivi. #Hate	0.6166	0.6190	0.5365	0.3583
Groups #Hate	0.5344	0.4583	0.4603	0.3493
Politics #Hate	0.5400	0.3659	0.2913	0.0299
Race #Hate	0.2817	0.3288	0.2857	0.0312

6 In-depth Discussion of Findings

1. Breakthrough in Detecting Political Hate (Politics): This is the most remarkable result of the study. The Politics#Hate label is one of the hardest due to its high use of metaphors, political slang, and reliance on background knowledge.

   - PhoBERT: 29.13% F1
   - Qwen Vanilla: 36.59% F1
   - Qwen + Rationales: 54.00% F1

   The F1-score improvement of +17.41% over Qwen Vanilla and +24.87% over PhoBERT demonstrates that incorporating Implied Statements helped the model understand the semantics of political comments. The model learned to associate political slang (e.g., *"ba que," "bò đỏ"*) with hate meaning instead of ignoring them.

2. Improvement on Groups: Similarly, for the Groups#Hate label, the proposed model achieves 53.44%, outperforming all baselines. This demonstrates better generalization for targets that are collective groups.

3. Failure on the Race Label: In contrast to the positive results above, performance on Race#Hate is very low (28.17%), even below Qwen Vanilla.

   - Reason: The issue lies in the data. The Race group has only about 500 training samples. The generated rationales for this group are also few and may contain noise. With such data sparsity, additional fine-tuning with rationales can lead to overfitting or confuse the model with complex reasoning that lacks enough examples for verification.
   - Key lesson: Chain-of-Thought (CoT) reasoning cannot fully replace sufficient original data.

4. Why PhoBERT Underperforms: PhoBERT was pre-trained on Wikipedia data (formal text). It excels at standard syntax but lacks the flexibility to understand online language. In contrast,

Qwen2.5 is a generative model designed to understand and produce natural language, giving it a natural advantage in capturing semantic nuances and pragmatics of modern Vietnamese.

7 Practical Application: Real-Time Streaming System

To demonstrate the model's feasibility in a real-world environment, the research team developed a demo web application that analyzes comments from YouTube. The application focuses on simplicity and fast processing speed

7.1 Tech Stack

The system is developed based on a modern Client-Server model, ensuring separation between the interface and processing logic:

- Frontend (Interface): Built with React combined with Vite to optimize page load performance. The interface is designed to be intuitive, allowing users to input a video link and view analysis results in real-time.

- Backend (Processing): Built on the FastAPI platform (Python). The backend acts as a bridge, performing two main tasks: (1) Calling the YouTube Data API v3 to fetch data, and (2) Running the trained Qwen2.5-3B model to classify comments (Inference).

- Data Source: Integrated with YouTube Data API v3 to collect public comment data from designated videos.

7.2 Label Mapping Strategy

The original labeling system of ViTHSD (#[Level]) is academically focused and might confuse general users. Therefore, in the demo version, we implement a mapping strategy to group the labels into 3 simpler, more understandable categories:

- Neutral: Includes comments labeled as Normal and Clean in the original system. These are safe content that does not require moderation.

- Offensive: Corresponds to the Offensive level. These are rude, uncomfortable comments, but they do not necessarily violate community standards severely.

- Hate: Corresponds to the Hate level. This is the most dangerous group, containing inflammatory, hostile language that needs a red-flag warning system.

Processing Workflow: Input URL → Crawl Comments → Qwen2.5 Inference → Map Labels → Display Result.

7.3 Keyword Highlighting Feature (Rule-based)

To help users quickly understand why a comment is flagged as Hate or Offensive, the demo integrates a highlighting feature that operates on a rule-based mechanism:

- Principle: The system maintains a "blacklist dictionary" containing hate keywords, profanity, and commonly used slang, directly extracted from the frequency of occurrences in the ViTHSD training dataset.

- Functionality: When displaying comments on the React interface, the system automatically scans and highlights words that match the dictionary entries.

- Value: This helps moderators quickly identify "dangerous" areas in the text without having to read every single word. It serves as a visual aid alongside the AI's predicted labels.

8 Conclusion, Limitations, and Future Directions

The study has demonstrated the effectiveness of combining Qwen2.5 with the HARE framework for the ViTHSD problem. However, through the process of building the demo and conducting experiments, we identified several important limitations that need to be addressed:

8.1 Limitations of the Rule-based Highlighting Mechanism

Although the keyword highlighting feature in the demo helps with quick visualization, it exposes inherent weaknesses of the rule-based method:

- Rigidity and Lack of Context: This mechanism only detects explicit keywords. It is entirely ineffective against implicit hate speech that doesn't use offensive language (e.g., *"Khôn như bạn quê tôi xích đầy"*).

- False Positives: It can incorrectly highlight harmless words if they are in the blacklist but are used in a neutral context, leading to unnecessary noise for the user.

8.2 Future Directions: Attention-based Highlighting

To address the above limitation, the next research direction will focus on developing an Attention-based Highlighting mechanism:

- Instead of using a static dictionary, the system will extract Attention Weights from the Self-Attention layers of the Qwen2.5 model.

- This allows the system to accurately highlight the words/phrases that the model is actually "focusing on" to make its decision, including metaphorical or sarcastic expressions, fully synchronizing the AI's reasoning with the interface display.

8.3 Other Expansion Directions

- Addressing Data Imbalance: Collect more data or use Generative Data Augmentation techniques for underrepresented groups, such as Race and Religion.

- Multimodal Expansion: Extend the system to handle both images (e.g., memes) and videos, creating a more comprehensive content moderation solution.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

Google DeepMind. 2025. Gemini 2.0 flash: Efficient multimodal reasoning model. Google AI Studio Documentation. Accessed via Google Developers and Gemini API docs.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 415–426, Cham. Springer International Publishing.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, and 89 others. 2024. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR 2013), Workshop Track Proceedings*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 6000–6010. Curran Associates, Inc.

Cuong Nhat Vo, Khanh Bao Huynh, Son T. Luu, and D. Trong-Hop. 2025. Vithsd: Exploiting hatred by targets for hate speech detection on vietnamese social media texts. In *Proceedings of the Journal of Computational Social Science*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.