ViNumFCR: A Novel Vietnamese Benchmark for Numerical Reasoning Fact Checking on Social Media News

Nhi Ngoc-Phuong Luong, Anh Thi-Lan Le, Tin Van Huynh*, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen
Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{20520263, 20521067}@gm.uit.edu.vn
{tinhv, kietnv, ngannlt}@uit.edu.vn

Abstract

In the digital era, the internet provides rapid and convenient access to vast amounts of information. However, much of this information remains unverified, particularly with the increasing prevalence of falsified numerical data, leading to public confusion and negative societal impacts. To address this issue, we developed ViNumFCR, a first dataset dedicated to fact-checking numerical information in Vietnamese. Comprising over 10,000 samples collected and constructed from online newspaper across 12 different topics. We assessed the performance of various fact-checking models, including Pre-trained Language Models and Large Language Models, alongside retrieval techniques for gathering supporting evidence. Experimental results demonstrate that the XLM-R$_{Large}$ model achieved the highest accuracy of 90.05% on the fact-checking task, while the combined SBERT + BM25 model attained a precision of over 97% on the evidence retrieval task. Additionally, we conducted an in-depth analysis of the linguistic features of the dataset to understand the factors influencing the performance models. The ViNumFCR dataset is available to support further research.

1 Introduction

With the rapid growth of the Internet, information is shared widely and instantly, keeping users updated on many topics. However, this also enables the spread of unverified content, which poses significant risks. In Vietnam, fake news and misinformation have harmed people's health, finances, families, and reputations. As a result, fact-checking has become essential in fields like journalism and social media. While there has been extensive research for languages such as English and Chinese, Vietnamese remains underexplored. In particular, previous studies have rarely focused on verifying numerical information — an important yet overlooked aspect. Motivated by this, we design and develop a Vietnamese fact-checking task centered on numerical verification, aiming to help bridge this gap and enhance the reliability of digital information.

The fact checking task (Ünal and Çiçeklioğlu, 2019) is a process of verifying the accuracy of information, data, or events before they are published to the public. The goal of fact checking (Ünal and Çiçeklioğlu, 2019) is to determine the correctness of information based on verifiable evidence, such as official documents, reliable sources, or accurately recorded events.

In this paper, we describe our task as follows: Given a Vietnamese sentence containing numerical data A and a Vietnamese text passage containing numerical data B, the objective is to develop a system that can verify the accuracy of sentence A by identifying supporting evidence in text B.

Input: Given a text passage (called B) and a sentence (called A) that is related to the content of B.

Output: A label X indicating the veracity of sentence A based on text B, where X belongs to the label set Supported, Refuted, NotenoughInfo, along with evidence to demonstrate whether "Text B contains sufficient evidence to verify sentence A.". Figure 1 shows an illustration of the task.

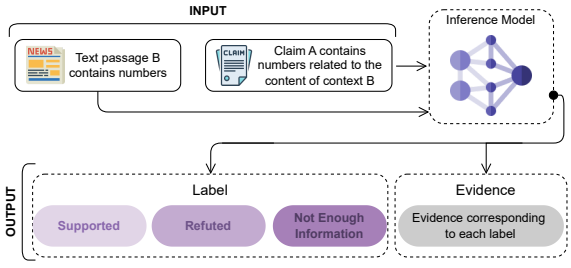

Figure 1: A Visual Diagram Illustrating the Numerical Reasoning Fact Checking Task.

*Corresponding author.

We evaluated multiple models for Vietnamese fact verification, with BERT (Devlin et al., 2019) achieving 83.48% accuracy. Other models such as PhoBERT (Nguyen and Tuan Nguyen, 2020), XLM-R (Conneau et al., 2020), InfoXLM (Chi et al., 2021), ViSoBERT (Nguyen et al., 2023), CaFeBERT (Do et al., 2024), and BARTpho-word (Tran et al., 2022) achieved 85.27%, 84.68%, 85.07%, 56.11%, 89.35%, and 85.67%, respectively. We also analyzed the influence of sentence length, syntactic complexity, and semantic ambiguity, which revealed their impact on model performance.

The key contributions of this project are as follows:

- We introduce ViNumFCR, a high-quality dataset for fact verification based on logical inference, consisting of over 10,000 samples of paragraph-claim-evidence with human-generated inference labels for fact-checking tasks.
- We conducted experiments on neural network-based models, pre-trained transformer models, and large language models.
- We analyzed linguistic features in ViNumFCR to understand factors influencing the performance of pre-trained models, providing insights into both the models and the ViNumFCR dataset.

2 Related Works

2.1 Related datasets

Several datasets have been developed to support fact-checking research. FEVER (Thorne et al., 2018) is a widely used benchmark containing over 185,000 claims labeled as Supported, Refuted, or NotEnoughInfo, based on evidence from Wikipedia. LIAR (Wang, 2017), proposed by Wang, includes over 12,800 political statements from politifact.com with six truthfulness levels, enabling fine-grained fake news detection. FEVEROUS(Aly et al., 2021) extends FEVER by including structured and unstructured data, with 87,026 labeled claims. VITAMINC (Schuster et al., 2021) offers 400,000 claim-evidence pairs, promoting contrastive reasoning for claim verification. In the Vietnamese context, ViFactCheck (Hoa et al., 2025) and ViWikiFC (Le et al., 2024) are the first manually annotated datasets with three labels—SUPPORTED, REFUTED, or NOT ENOUGH INFORMATION—contributing significantly to fact-checking research for Vietnamese. However, no dataset has been specifically designed to study numerical reasoning fact-checking on Vietnamese social media text.

Table 1 provides a brief summary of various datasets used for the fact-checking task.

2.2 Related models

The evolution of artificial intelligence has driven significant advancements in NLP for fact-checking tasks. Early approaches utilized classical machine learning models like Random Forest (Rigatti, 2017) and Support Vector Machines (SVM) (Mammone et al., 2009), which classified claims using syntactic and semantic features. While stable, these models struggled with contextual nuances critical for fact-checking. Recurrent neural networks such as LSTM (Hochreiter and Schmidhuber, 1997) improved performance by modeling sequential dependencies, capturing textual context more effectively. However, their limitations in handling long-range dependencies led to the adoption of transformer-based models. Models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2020), and PhoBERT (Nguyen and Tuan Nguyen, 2020) revolutionized NLP by leveraging self-attention to capture bidirectional semantics, making them ideal for fact-checking tasks requiring nuanced understanding.

Recent large language models (LLMs), such as Qwen (Bai et al., 2023), Gemma (Mesnard et al., 2024), GPT (Achiam et al., 2023), and Llama (Touvron et al., 2023), have further advanced fact-checking by offering superior reasoning and text generation capabilities (Hoa et al., 2025; Le et al., 2024). These models excel in verifying complex claims but face challenges like computational cost. Evidence retrieval techniques, including BM25 (Robertson et al., 2009) and SBERT (Reimers and Gurevych, 2019), complement these models by retrieving relevant evidence to enhace the accuracy and reliability of fact verification models.

3 Corpus Creation

We conducted the data construction following a rigorous process to ensure the accuracy, reliability, and quality of the dataset. The process was referenced from the ViNLI and FEVER dataset

Dataset	Text genre	Quantity	Language	3+ labels agree	4+ labels agree	Numerical Focus
FEVER	Wikipedia	∼185,000	English	–	–	No
LIAR	Newswire	∼12,800	English	–	–	No
FEVEROUS	Wikipedia	87,026	English	–	–	No
VITAMINC	Wikipedia	∼400,000	English	–	–	No
ViFactCheck	Wikipedia	∼7,200	Vietnamese	–	–	No
ViWikiFC	Newswire	∼20,000	Vietnamese	–	–	No
ViNumFCR (Our dataset)	**Newswire**	**∼10,000**	**Vietnamese**	**94%**	**85%**	**Yes**

Table 1: Overview of Fact-Checking Datasets

construction methodologies, as illustrated in Figure 2, and includes four main stages: (3.1) Collecting premise paragraphs; (3.2) Annotator recruitment and training; (3.3) Generating claims and finding evidence; and (3.4) Data validation. Additionally, we analyzed the dataset from various perspectives, as presented in Section 3.5

3.1 Collecting premise paragraphs

We collected over 10,000 articles from the highly reputable Vietnamese online newspaper VnExpress[1], covering 12 diverse topics including: digitalization, tourism, education, entertainment, science, business, law, health, world news, sports, current events and automobiles. Subsequently, we conducted preliminary data processing to extract text segments containing solely numerical data.

3.2 Annotator recruitment and training

The annotator recruitment and training process was based on the methodology used in the FEVER dataset (Thorne et al., 2018) and followed the steps outlined in Figure 3. The annotators, consisting of 17 university students in Vietnam with strong language proficiency and effective communication skills, underwent a training session to fully understand the annotation guidelines and evaluation criteria. They were compensated at a rate of $0.019 per annotated pattern. As part of the training phase, each annotator was required to write 20 claims using our custom-built annotation tool. The labels (Supported, Refuted, and NotenoughInfo) for the corresponding paragraph–claim pairs were then hidden, and the annotators proceeded to assign labels to these claims. Inter-annotator agreement was evaluated using Cohen's Kappa coefficient (Cohen, 1960). Annotators achieving an agreement rate above 0.95 were deemed eligible to participate in the official data construction. In cases of lower agreement rates, annotators were required to review their errors and repeat the training process with a new dataset. During training, any disagreements in labeling were reviewed, and, if necessary, adjustments were made to the sentence-writing rules to ensure the quality and accuracy of the dataset.

3.3 Generate Claims and Find Evidence

The annotators are required to generate claims that include numerical data and content mentioned in the original paragraph but rephrased creatively using their own vocabulary, avoiding the reuse of words or phrases that appeared in the premise paragraph. The creators will generate three claims for the three labels according to the following guidelines:

Supported: The claim is correct with respect to the information and numerical data provided in the original paragraph.

Refuted: The claim is incorrect with respect to the information and numerical data provided in the original paragraph.

NotenoughInfo: It cannot be determined whether the claim is correct or incorrect based on the information and numerical data available in the original paragraph.

Additionally, the creators are required to find evidence for the two labels, Supported and Refuted. The evidence is provided by selecting specific sentences from the premise paragraph that relate to the claim the creator just wrote. To create these claims, annotators may refer to the sentence writing rules provided in the guidelines. Tables 2 and 3 summarize these rules, which were referenced from the ViNLI dataset (Huynh et al., 2022). The illustrative examples of rules for creating premise paragraph - claim pairs for the Supported and Refuted Labels are presented in Appendix A. The annotators must write three claims and find evidence (for the Supported and Refuted labels) for each premise

[1]https://vnexpress.net

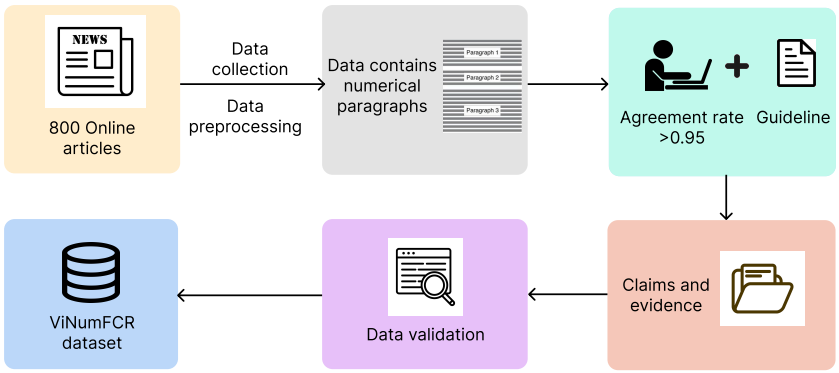

Figure 2: The Process of Building the Vinumfcr Dataset.

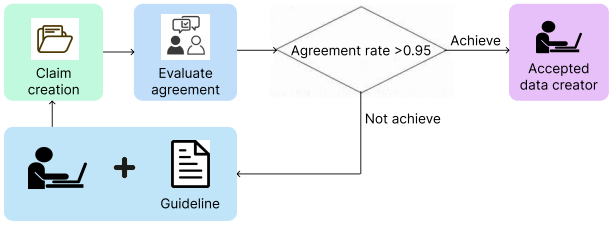

Figure 3: The Process of Selecting Data Annotators for the Vinumfcr Dataset.

No.	Rule	Ratio
1	Use negative words.	6%
2	Replace with antonyms.	14.5%
3	Incorrect entity inference structure.	18.5%
4	Incorrect event inference structure.	2%
5	Create a sentence with a meaning opposite to the presupposition paragraph.	38%
6	Others.	21%

Table 3: Rules for Creating Refuted Claims Sentences.

paragraph.

No.	Rule	Ratio
1	Change the sentence structure from active to passive and vice versa.	7%
2	Replace with synonyms or similar words.	38%
3	Add or remove modifiers while retaining the original meaning of the sentence.	26%
4	Replace representative nouns with relative clauses.	1%
5	Replace the object with a relative clause.	1%
6	Replace the adjective with a relative clause.	1%
7	Replace the quantity terms with equivalent ones.	9%
8	Generate a presupposition sentence.	11.5%
9	Others.	13.5%

Table 2: Rules for Creating Supported Claims Sentences.

3.4 Data validation

We conducted a round of data validation by cross-labeling premise paragraph-claim pairs by multiple annotators. We selected five different annotators, who had participated in the claim-writing phase, to label the premise paragraph-claim pairs. The data samples to be labeled were not from the dataset originally written by the assigned labelers. If a premise paragraph-claim pair does not receive at least three out of five identical labels, it will be excluded from the dataset. This new method provides a fresh perspective on an issue that the original FEVER paper did not address, with the results presented in Table 1. Statistics show that the rate of premise paragraph-claim pairs achieving four identical labels is 85%, while the rate of premise paragraph-claim pairs achieving three identical labels is 94%.

3.5 Corpus Analysis

To train and evaluate the model, we randomly divided the data into three parts with the following proportions: 80% for the training set (train), 10% for the development set (dev) and 10% for the test set. Table 4 shows the preliminary statistics, including the number of premise paragraph-claim pairs across 12 different topics and the average length (in words) of the premise paragraphs and claims. The average length of the premise paragraphs and claims in the Train, Dev and Test sets is 58.3 words for premise paragraphs and 20.5 words for claims. The average lengths of the premise paragraphs and claims across the three sets are relatively consistent, contributing to the dataset's consistency and helping the model learn more effectively.

Word Overlap: We calculated the word overlap between the premise paragraphs and claims in the ViNumFCR dataset. We chose to use the Jaccard index to assess lexical overlap based on the frequency of words appearing, regardless of or-

Topic/Label	Train	Dev	Test	Total
Digital	742	80	90	912
Tourism	523	76	79	678
Education	479	57	58	594
Entertainment	323	41	59	423
Science	574	61	73	708
Business	700	90	89	882
Law	1,662	201	189	2,052
Health	729	88	80	897
World	594	82	80	756
Sports	622	64	73	759
News	643	105	71	819
Cars	413	54	64	531
Supported	2,668	333	335	3,336
Refuted	2,668	333	335	3,336
NotenoughInfo	2,668	333	335	3,336
Total (pairs)	8,004	999	1,005	10,008
MPL (words)	58.6	57.5	58.8	58.3
MCL (words)	20.4	20.4	20.6	20.5

Table 4: Overview Statistics of the ViNumFCR Dataset. MPL: Mean Premise Paragraph Length. MCL: Mean Claim Length.

der, between the premise paragraph and the claim. Additionally, we used LCS (Longest Common Subsequence) to evaluate the structural similarity between the premise paragraph and the claim by focusing on finding the longest common subsequence between the two strings. To better suit the analysis in Vietnamese, we first used VnCoreNLP(Vu et al., 2018) for word segmentation before applying Jaccard and LCS. The analysis results are shown in Table 5. Based on the analysis, we found that the Supported label has the highest lexical overlap when measured by the Jaccard index, as well as the highest sequence overlap when measured by the LCS index. Conversely, the NotenoughInfo label has the lowest lexical overlap according to both the Jaccard and LCS indices.

New Word Rate: We analyzed the usage of new words in the dataset to evaluate the diversity in the creators' use of language. To perform this effectively, we also used VnCoreNLP(Vu et al., 2018) for word segmentation. Then, we used PhoNLP(Nguyen and Nguyen, 2021) to classify the new words by word class. The results shown in Table 5 indicate that the NotenoughInfo label has the highest number of new words at 54.27%. Additionally, nouns and verbs are the most frequently used word classes in the claims created by the annotators.

Data-Generation Rules Analysis: annotators may flexibly use one or more rules to construct claims. We analyzed how these rules were combined by selecting 200 random premise paragraph-claim pairs from the dataset, focusing on the Supported and Refuted labels for analysis. According to Figure 4, 57% of the supported claims were created using one rule, 41% were created using two rules and using three rules to create supported claims was less common, accounting for only 2%. For the refuted claims, the trend of using one rule to create a claim was predominant, and combining two rules accounting for only 3%.

Next, we analyzed the claim creation rules to better understand the tendencies and standards applied by the annotators during data construction. The results from Table 2 show that, for supported claims, annotators most frequently used the "Replace with synonyms or similar words." rule, accounting for 38%. Conversely, the rule "Replace representative nouns/objects/adjectives with relative clauses." was used the least, at only 3%. For refuted claims, Table 3 shows that the most applied rule was "Create a sentence with a meaning opposite to the presupposition paragraph." with a rate of 38%, as this rule best aligns with the primary purpose of refuted claims, which is to negate the premise paragraph content. On the other hand, the least used rule was "Incorrect event inference structure." accounting for only 2%.

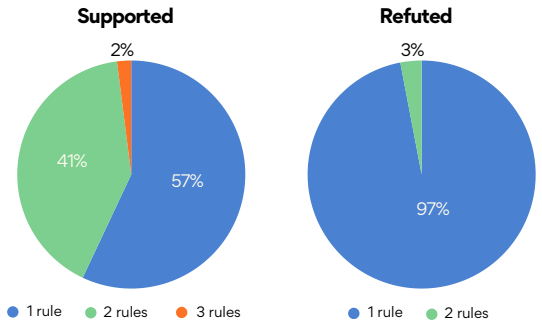

Figure 4: Percentage of Combined Rules for Generating Supported and Refuted Label Claims.

Syntax tree: To evaluate the complexity of affirmative sentences, we conducted a syntactic tree analysis. We used PhoNLP to analyze the dependency relationships between words in the sentences, thereby constructing a tree structure that represents the connections between the words. The level of the tree was calculated as the maximum depth from a leaf node (a word with no child dependencies) to the root of the sentence. The results shown in Figure 5 indicate that syntactic trees with a level of 3 accounted for the highest proportion, at

Label	Jaccard (%)	LSC (%)	New word rate	Part-Of-Speech (%)					
				Noun	Verb	Adjective	Preposition	Adjunct	Other
Supported	23.72	72.92	35.58	28.35	30.16	6.93	9.04	9.61	15.91
Refuted	19.76	70.12	43.99	25.62	25.92	7.27	9.36	8.41	23.42
NotenoughInfo	15.74	67.41	54.27	31.11	26.31	7.47	8.73	9.48	16.90

Table 5: Word Overlap and New Word Rate Between Premise Paragraph and Claim.

39.11%.

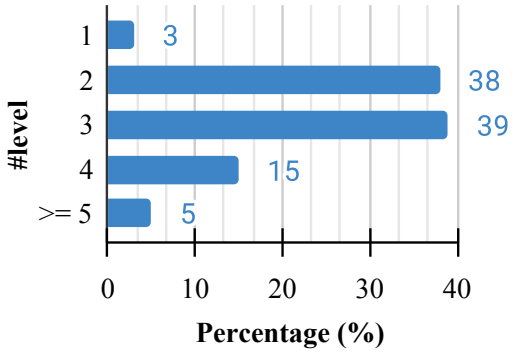

Figure 5: Distribution of Syntax Tree Depth Levels.

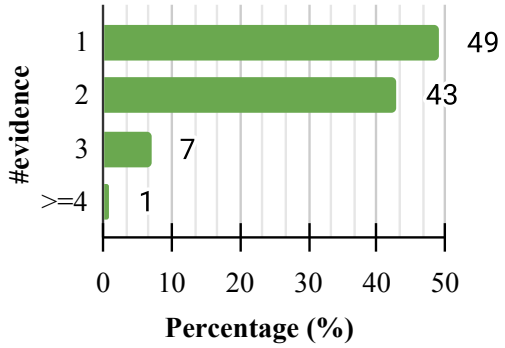

Figure 6: The Proportion of Combined Sentences Forming Evidence.

Evidence: During the data construction phase, after completing the claims, annotators search for evidence to support the claims for the Supported and Refuted labels. Figure 6 illustrates the proportion of sentences used to form evidence. Accordingly, 49% of the evidence consists of only one sentence, indicating that this is the most commonly applied method by annotators. Additionally, the proportion of using two sentences as evidence is 43%, while evidence using three or more sentences accounts for 8%.

4 Experiments and Evaluation

4.1 Baseline Model and Experimental Setup

To evaluate ViNumFCR, we experimented with a range of models, including multilingual PLMs such as mBERT (Devlin et al., 2019), SBERT (Reimers and Gurevych, 2019), and XLM-R (Conneau et al., 2020); Vietnamese-specific PLMs like PhoBERT (Nguyen and Tuan Nguyen, 2020), BARTpho-word (Tran et al., 2022), and CafeBERT (Do et al., 2024). For large language models we performed zero-shot and few-shot technical experiments with the Qwen-2.5-7b-instruct (Bai et al., 2023), Gemma-3-12b-it (Mesnard et al., 2024), and Llama-4-Scout-17B (Touvron et al., 2023) models (Appendix B presents our prompt templates). In addition, we also fine tuning with the Gemma-2-2b model to evaluate and compare with the above experiments.

All were fine-tuned on Google Colab[2] (Tesla P100) using the same hyperparameters when fine-tuning PLMs with learning_rate = 1e-5, epoch = 7, batch_size = 16. For fine-tuned LLMs, we applied 4-bit quantization and LoRA to train Gemma-2-2b with learning_rate = 1e-5, epoch =5 , batch_size = 8, lora_rank = 8, and lora_alpha = 16.

For evidence retrieval, we applied BM25 (Robertson et al., 2009) and SBERT (Reimers and Gurevych, 2019) to rank sentences by relevance and compared the top results with gold-standard evidence.

4.2 Evaluation Metrics

We use Accuracy, Precision and F1-score as the main evaluation metrics. High values of these metrics indicate that the model has a high prediction accuracy.

4.3 Experimental Results

Table 6 presents the results of PLM and LLM models on the dev and test sets of the ViNumFCR dataset. Overall, PLM models achieve high and stable performance. Among them, CafeBERT and XLM-R$_{Large}$ lead with test Accuracy and F1-score around 89–90%. Models like PhoBERT$_{Large}$ and InfoXLM also exceed 85%, demonstrating the advantage of multilingual training or specialization for Vietnamese. In contrast, SBERT$_{Base}$ and

[2]https://colab.research.google.com/

	Models	Dev		Test	
		Acc	F1-score	Acc	F1-score
	Fine-tuning				
PLM	mBERT	84.78	84.78	83.48	83.43
	SBERT$_{Base}$	76.68	76.64	75.82	75.67
	SBERT$_{Large}$	78.98	78.99	77.41	77.40
	InfoXLM	86.09	86.07	85.07	85.00
	PhoBERT$_{Base}$	85.39	85.38	85.27	85.26
	PhoBERT$_{Large}$	88.69	88.70	87.56	87.56
	XLM-R$_{Base}$	84.88	84.85	84.68	84.67
	XLM-R$_{Large}$	90.09	90.09	**90.05**	**90.06**
	CafeBERT	**90.89**	**90.89**	89.35	89.35
	BARTpho	85.59	85.57	85.67	85.65
	Zero-shot Prompting				
LLM	**Qwen-2.5**	**44.84**	**42.41**	**45.87**	**43.96**
	Gemma-3	42.74	38.52	43.68	40.25
	Llama-4	43.14	41.6	42.59	40.78
	Few-shot Prompting				
	Qwen-2.5	36.84	37.88	37.71	38.72
	Gemma-3	**42.84**	**41.23**	**41.79**	**40.54**
	Llama-4	41.24	39.96	40.52	38.84
	Fine-tuning				
	Gemma-2	90.69	90.69	88.96	88.95

Table 6: The Performance of Models on the Dev and Test Set of the ViNumFCR Dataset.

SBERT$_{Large}$ only achieve about 75–78%, reflecting the limitations of non-specialized models.

The performance of LLMs is noticeably lower. In the zero-shot setting, the best model (Qwen-2.5-7b-instruct) reaches only around 45% Accuracy; other models range around 40–43%. In the few-shot setting, the results improve slightly but remain low, with the highest Accuracy around 41.79%.

In summary, PLMs – especially those trained specifically for Vietnamese or multilingual tasks – outperform LLMs on this task.

For the prediction of evidence, we evaluated the data set using a dynamically selected k (ranging from 1 to 4), where k represents the number of relevant top sentences automatically recovered based on each instance. Table 7 shows that SBERT + BM25 achieved strong performance (precision 97.4%, recall 88.89%, F1-score 92.95%). BM25 individually also performed well, with very high F1-score (99.8%). SBERT and TFIDF had lower recall but still maintained good precision.

Finally, Table 8 reports precision@k on the dataset for fixed values of K = 1–4. SBERT + BM25 and BM25 consistently achieved high precision (above 93% up to 95.88%), showing stable performance. SBERT started lower at K=1 (79.86%) but improved as k increased. TFIDF had the lowest precision at K=1 (68.32%) but increased significantly to 94.81% at K=4.

Models	Dynamic k		
	Precision	Recall	F1-score
SBERT	92.20	78.78	84.96
BM25	**97.40**	**90.47**	**99.80**
TFIDF	94.27	77.86	85.28
SBERT + BM25	**97.40**	**88.89**	**92.95**

Table 7: Evidence Retrieval Performance on the ViNumFCR Dataset.

Models	Precision@k			
	K=1	K=2	K=3	K=4
SBERT	79.86	86.26	89.47	92.18
BM25	**93.40**	**94.08**	**95.30**	**95.88**
TFIDF	68.32	79.04	84.62	94.81
SBERT + BM25	**93.34**	**94.07**	**95.29**	**95.88**

Table 8: Precision@K on ViNumFCR for K=1–4 Evidence Sentences.

5 Result Analysis

To better understand the impact of various aspects on performance outcomes, we conducted an analysis on the three models with the highest performance results: PhoBERT$_{Large}$, CafeBERT, mBERT, Gemma and XLM-R$_{Large}$.

Length affects model performance: The impact of text length on model performance on the Dev set for CafeBERT, PhoBERT$_{Large}$, mBERT, Gemma and XLM-R$_{Large}$ is illustrated in Figure 7. When analyzing the combined length of the premise paragraph and the claim, the XLM-R$_{Large}$ and CafeBERT models achieve the best performance when the total length exceeds 150 words. In contrast, the performance of mBERT and Gemma drops significantly when the total length exceeds 150 words. Meanwhile, PhoBERT$_{Large}$ maintains relatively stable accuracy across different length ranges.

Word overlap affects model performance: We analyzed the impact of the Jaccard index on model accuracy similarly to that on the ViNLI dataset (Van Huynh et al., 2022). The performance analysis on the Dev set, as shown in Figure 8, indicates that the mBERT model achieves the lowest accuracy when the Jaccard index falls within the 20% to

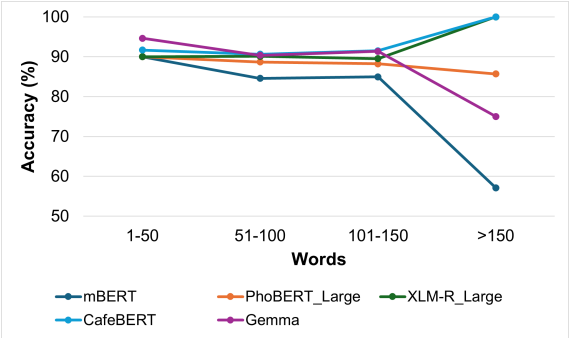

Figure 7: Model Performance by Length on the Dev Set.

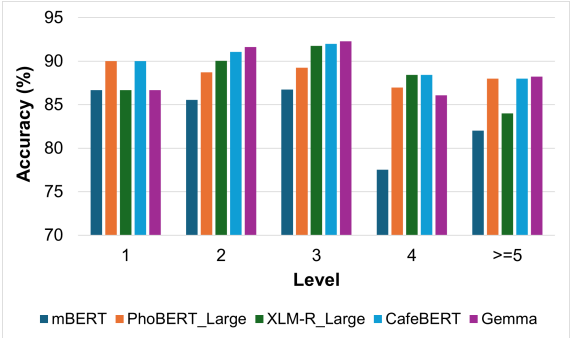

Figure 9: Model Performance by Syntax Tree Index on the Dev Set.

30% range. When the Jaccard index exceeds 50%, the performance of mBERT continues to decline, with an even more pronounced drop observed in the PhoBERT$_{Large}$ model. In contrast, within the same Jaccard range, the XLM-R$_{Large}$, CafeBERT, and Gemma models achieve the highest accuracy.

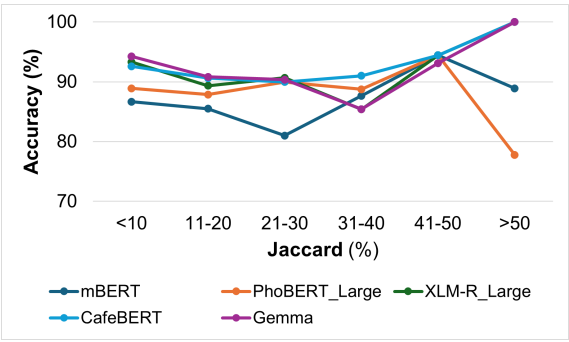

Figure 8: Model Performance by Jaccard Index on the Dev Set.

Syntax tree affects model performance: The impact of the syntactic tree level on model performance on the Dev set is illustrated in Figure 9. The mBERT, CafeBERT, Gemma, and XLM-R$_{Large}$ models achieve the highest performance when the syntactic tree level is 3. However, the PhoBERT$_{Large}$ model performs best when the tree level is only 1. Furthermore, as the tree level increases from 4 onward, the performance of all models tends to decline, with the most noticeable drop observed in the mBERT model when the level reaches 4.

Topic affects model performance: The analysis of topic influence on model performance, as shown in Table 9, indicates that the Education and Entertainment topics pose significant challenges for the models, with generally lower accuracy compared to other topics. In contrast, topics such as Business and News yield high and consistent performance across most models. Additionally, both CafeBERT and Gemma demonstrate relatively high performance across all topics.

Topic / Model	XLM-R$_{Large}$	mBERT	CafeBERT	PhoBERT$_{Large}$	Gemma
Digital	93.68	84.21	93.68	**93.68**	88.75
Tourism	91.89	87.84	91.89	90.54	88.16
Education	78.95	75.44	82.46	85.96	91.23
Entertainment	76.92	74.36	82.05	79.49	90.24
Science	80.36	78.57	89.29	85.71	86.89
Business	**94.05**	84.52	**94.05**	90.48	91.11
Law	93.33	88.57	92.38	87.62	91.54
Health	90.72	85.57	89.69	88.66	90.91
World	88.89	83.95	90.12	86.42	**95.12**
Sports	91.89	83.78	91.89	89.19	93.75
News	90.28	**88.89**	90.28	91.67	91.43
Cars	91.67	86.67	93.33	90.00	85.19

Table 9: Model performance across different topics.

6 Conclusion and Future Work

This paper introduces the ViNumFCR dataset, a pioneering benchmark for numerical reasoning fact-checking in Vietnamese, comprising over 10,000 claims by a rigorous annotation process. We assessed advanced language models including PLMs, LLMs, revealing that models such as XLM-R$_{Large}$ and CafeBERT achieved significant accuracy. However, we observed that zero-shot and few-shot techniques with Qwen, LLama, and Gemma models for reasoning and generating predictions proved ineffective, highlighting the unique challenges posed by this dataset. These findings emphasize ViNumFCR's role as a robust resource for real-world applications and the advancement of Vietnamese NLP research, particularly in managing numerical data on social media.

Looking ahead, we intend to enrich the dataset by increasing its scale and diversity through expanded data collection and thorough cleaning procedures. We plan to explore fine-tuning techniques on alternative machine learning models, especially large language models, to further enhance perfor-

mance, particularly for complex fact-checking samples. Moreover, we will tackle current limitations by improving models' ability to process longer and more intricate text passages, thereby boosting the system's reliability for real-world news verification deployments.

Acknowledgements

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS.C2025-26-10.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Phong Nguyen-Thuan Do, Son Quoc Tran, Phu Gia Hoang, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. VLUE: A new benchmark and multi-task knowledge transfer learning for Vietnamese natural language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 211–222, Mexico City, Mexico. Association for Computational Linguistics.

Tran Thai Hoa, Tran Quang Duy, Khanh Quoc Tran, and Kiet Van Nguyen. 2025. Vifactcheck: A new benchmark dataset and methods for multi-domain news fact-checking in vietnamese. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 308–316.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hung Tuan Le, Long Truong To, Manh Trong Nguyen, and Kiet Van Nguyen. 2024. Viwikifc: Fact-checking for vietnamese wikipedia-based textual knowledge source. *arXiv preprint arXiv:2405.07615*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alessia Mammone, Marco Turchi, and Nello Cristianini. 2009. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):283–289.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*.

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.

Linh The Nguyen and Dat Quoc Nguyen. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 1–7, Online. Association for Computational Linguistics.

Nam Nguyen, Thang Phan, Duc-Vu Nguyen, and Kiet Nguyen. 2023. ViSoBERT: A pre-trained language model for Vietnamese social media text processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5191–5207, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

S. J. Rigatti. 2017. Random forest. *Journal of Insurance Medicine*, 47(1):31–39.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*.

Tin Van Huynh, Huy Quoc To, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. Error investigation of pre-trained bertology models on vietnamese natural language inference. In *Asian Conference on Intelligent Information and Database Systems*, pages 176–188. Springer.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

R. Ünal and A. Ş. Çiçeklioğlu. 2019. The function and importance of fact-checking organizations in the era of fake news: Teyit. org, an example from turkey. *Media Studies*, 10(19):140–160.

A Data-Generation Rules

No.	Rule	Example
1	Change the sentence structure from active to passive and vice versa.	P: Một con mèo đang chơi với cuộn len. (A cat is playing with a ball of yarn.) C: Cuộn len đang bị một con mèo chơi. (The ball of yarn is being played with by a cat.)
2	Replace with synonyms or similar words.	P: Tôi đi siêu thị mua thực phẩm. (I went to the supermarket to buy groceries.) C: Tôi đi siêu thị mua đồ ăn. (I went to the supermarket to buy food.)
3	Add or remove modifiers while retaining the original meaning of the sentence.	P: Cô giáo 23 tuổi xinh đẹp đang dạy tiếng Anh. (The beautiful 23-year-old teacher is teaching English.) C: Cô giáo 23 tuổi đang dạy tiếng Anh. (The 23-year-old teacher is teaching English.)
4	Replace representative nouns with relative clauses.	P: Một cầu thủ đá bóng nam đang sút bóng vào khung thành. (A male soccer player is kicking the ball into the goal.) C: Người đàn ông, cái người mà chơi bóng đá đang sút một quả bóng vào khung thành. (The man, who plays soccer, is kicking a ball into the goal.)
5	Replace the object with a relative clause.	P: Người phụ nữ đang dùng một chiếc máy may. (The woman is using a sewing machine.) C: Người phụ nữ đang sử dụng một chiếc máy được sản xuất để may. (The woman is using a machine made for sewing.)
6	Replace the adjective with a relative clause.	P: Hai người đàn ông đang nghỉ ngơi sau một chuyến đi trên con đường tuyết. (Two men are resting after a trip on a snowy road.) C: Hai người đàn ông đang nghỉ ngơi sau một chuyến đi trên con đường mà nó được phủ đầy tuyết. (Two men are resting after a trip on a road that is covered with snow.)
7	Replace the quantity terms with equivalent ones.	P: Vài người đang lướt sóng trên một con sóng lớn. (Several people are surfing on a big wave.) C: Một số người đang lướt sóng trên một con sóng lớn. (Some people are surfing on a big wave.)
8	Generate a presupposition sentence.	P: Tôi đã lạc mất con mèo duy nhất của tôi vào sáng nay. (I lost my only cat this morning.) C: Tôi có một con mèo. (I have a cat.)
9	Others.	

Table 10: Examples of Rules for Creating Premise Paragraph (P) - Claim (C) Pairs for the Supported Label.

No.	Rule	Example
1	Use negative words.	P: Mặc dù đã có vacxin phòng ngừa bệnh cúm, nhưng mỗi năm nước ta vẫn có tới 800.000 người mắc bệnh.(Although there is a flu vaccine, our country still has up to 800,000 cases each year.) C: Số ca mắc bệnh cúm ở nước ta mỗi năm là 200 người, và vẫn chưa có loại vacxin phòng bệnh.(The number of flu cases in our country each year is 200 and there is still no vaccine available.)
2	Replace with antonyms.	P: Một chiếc máy bay đang cất cánh. (An airplane is taking off.) C: Một chiếc máy bay đang hạ cánh. (An airplane is landing.)
3	Incorrect entity inference structure.	P: Với chiều dài 50,45 km, nối liền Folkestone ở Anh với Coquelles ở Pháp, Channel là đường hầm đường sắt dài thứ ba trên thế giới. (At 50.45 km in length, connecting Folkestone in the UK with Coquelles in France, the Channel Tunnel is the third longest railway tunnel in the world.) C: Đường hầm Channel nối Pháp và Nhật Bản, lập kỉ lục là đường sắt dài thứ 4 thế giới. (The Channel Tunnel connects France and Japan, setting a record as the fourth longest railway tunnel in the world.")
4	Incorrect event inference structure.	P: Ông Santer kế nhiệm ông Delors làm việc tại Ủy ban ở Châu Âu vào năm 1995. (Mr. Santer succeeded Mr. Delors at the European Commission in 1995.) C: Năm 1990 ông Delors kế nhiệm ông Santer làm việc tại Ủy ban ở Thái Bình Dương. (In 1990, Mr. Delors succeeded Mr. Santer at the Pacific Commission.)
5	Create a sentence with a meaning opposite to the presupposition paragraph.	P: Năm 2009, Rose kết hôn và cùng chồng mua căn nhà ở vùng ngoại ô London chưa tới 1 triệu bảng. (In 2009, Rose got married and, together with her husband, bought a house in the suburbs of London for less than 1 million pounds.) C: Cho đến tận năm 2010, Rose vẫn chưa từng thử yêu đương một lần. (Up until 2010, Rose had never tried dating even once.)
6	Others.	

Table 11: Examples of Rules for Creating Premise Paragraph (P) - Claim (C) Pairs for the Refuted Label.

B LLM Prompts

This section presents the prompts (zero-shot and few-shot prompting) used in experiments with LLMs (Qwen, Gemma, Llama) on our ViNumFCR dataset.

Zero-shot Prompting

SYSTEM PROMPT:

You are a fact checking classifier. Your task is to classify the relationship between the paragraph and claim into exactly one of the following labels: Supported, Refuted, or NotenoughInfo. Your classification should be precise.

 # Instructions

- Read the given paragraph and claim.
- Decide the relationship between them based strictly on semantic meaning.
- You must follow these definitions:

  - **Supported**: The claim is correct with respect to the information and numerical data provided in the original paragraph.
  - **Refuted**: The claim is incorrect with respect to the information and numerical data provided in the original paragraph.
  - **NotenoughInfo**: It cannot be determined whether the claim is correct or incorrect based on the information and numerical data available in the original paragraph.

- Do not provide any explanations or extra words.

 # Output Format:

You should ONLY return one word from the set: Supported | Refuted | NotenoughInfo

---

USER PROMPT:

`Paragraph`: `...`
`Claim`: `...`

Few-shot Prompting

SYSTEM PROMPT:

You are a fact checking classifier. Your task is to classify the relationship between the paragraph and claim into exactly one of the following labels: Supported, Refuted, or NotenoughInfo. Your classification should be precise.

 # Instructions

- Read the given paragraph and claim.
- Decide the relationship between them based strictly on semantic meaning.
- You must follow these definitions:

  - **Supported**: The claim is correct with respect to the information and numerical data provided in the original paragraph.
  - **Refuted**: The claim is incorrect with respect to the information and numerical data provided in the original paragraph.
  - **NotenoughInfo**: It cannot be determined whether the claim is correct or incorrect based on the information and numerical data available in the original paragraph.

- Learn from the provided examples.

Example 1
+ Paragraph: Việt Nam có 33 cơ sở giáo dục liên cấp có vốn đầu tư nước ngoài, thường được gọi là trường quốc tế, trong đó Hà Nội có 13 cơ sở, TP HCM có 20 cơ sở. Các trường này thu học phí hàng trăm triệu đồng một năm, tăng dần từ mầm non đến cấp trung học. *(Vietnam has 33 inter-level educational institutions with foreign investment, commonly referred to as international schools, of which Hanoi has 13 and Ho Chi Minh City has 20. These schools charge tuition fees of hundreds of millions of VND per year, increasing from kindergarten to secondary level.)*
+ Claim: Trong tổng số trường quốc tế liên cấp tại Việt Nam, TP.HCM chiếm hơn 50%. *(Ho Chi Minh City accounts for more than 50% of the total number of international inter-level schools in Vietnam.)*
+ True Label: Supported

Example 2
+ Paragraph: Từng là tuyển thủ hạt giống của đội bắn súng, Vạn Quang Húc gần đây bỏ bê tập luyện, mâu thuẫn với huấn luyện viên, cố tình khiêu khích gây rối, đánh nhau. Húc từng được đội thể thao nhiều lần bỏ qua lỗi lầm vì luyến tiếc tài năng, nhưng hành động đánh đập ác ý vận động viên khác đã vượt quá giới hạn. Húc bị công an bắt tạm giam một tháng, khai trừ khỏi đội. *(Once a seeded athlete of the shooting team, Van Quang Huc has recently neglected training, clashed with his coach, deliberately provoked disturbances, and engaged in fights. He had been repeatedly forgiven by the team out of regard for his talent, but his malicious assault on another athlete crossed the line. Huc was detained by the police for one month and expelled from the team.)*
+ Claim: Sau khi bị tạm giam chỉ 20 ngày, anh ấy đã được ân xá để trở lại đội tuyển vì tài năng của mình. *(After being detained for only 20 days, he was pardoned and allowed to return to the national team because of his talent.)*
+ True Label: Refuted

Example 3
+ Paragraph: Một hệ lụy khác mà cơn bão sa thải mang đến là môi trường làm việc trở nên "vô cùng áp lực" khi có đến 31% người lao động thường xuyên stress. Ngoài ra, một trạng thái đáng báo động mà báo cáo nêu cứ 10 người đi làm có 4 người rơi vào trạng thái burn out – hội chứng kiệt quệ về thể chất, tinh thần do stress quá nhiều. *(Another consequence brought about by the wave of layoffs is that the work environment has become "extremely stressful," with up to 31% of employees frequently experiencing stress. In addition, the report highlights an alarming state: 4 out of every 10 workers suffer from burnout – a syndrome of physical and mental exhaustion caused by excessive stress.)*
+ Claim: Một nữa trong số 31% người lao động thường xuyên stress có nguy cơ cơ cao mắc các bệnh liên quan đến tim mạch do kiệt quệ về thể chất, tinh thần. *(Half of the 31% of employees who are frequently stressed face a high risk of cardiovascular diseases due to physical and mental exhaustion.)*
+ True Label: NotenoughInfo

- Do not provide any explanations or extra words.

 # Output Format:
You should ONLY return one word from the set: Supported | Refuted | NotenoughInfo

---

USER PROMPT:

`Paragraph`: `...`
`Claim`: `...`