Fact-checking using LLMs for Vietnamese

Mai Ho Ngoc[1,2] and Phuc Le Ngoc Thien[1,2]
[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2]University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam
{22520839, 22521117}@gm.uit.edu.vn

Abstract

Fact-checking has become an essential task in combating misinformation, yet it remains underexplored for less-resourced languages like Vietnamese. In this study, we present a comprehensive evaluation of large language models (LLMs) and baseline models for Vietnamese fact-checking, leveraging two key datasets: ViWikiFC, which focuses on Wikipedia-based structured claims, and ViFactCheck, a multi-domain dataset covering diverse topics such as politics, health, and science. We evaluate advanced LLMs, including Qwen-2.5-7B and URA-LLama-7B, alongside models like vELECTRA, XLM-Roberta, mBERT, ViBERT, and PhoBERT. XLM-Roberta achieves the highest F1-scores (0.7551 on ViWikiFC, 0.7823 on ViFactCheck), while mBERT and PhoBERT also perform well, particularly in evidence retrieval and claim verification. General-purpose LLMs like Qwen-2.5-7B and URA-LLama-7B underperform without fine-tuning, underscoring the need for tailored models and datasets to enhance Vietnamese fact-checking. This study offers insights for developing scalable systems that address Vietnamese-specific challenges.

1 Introduction

The proliferation of misinformation poses significant challenges in today's digital landscape, affecting public discourse, health decisions, and political stability. Automated fact-checking has emerged as a crucial tool for combating misinformation, leveraging advancements in Natural Language Processing (NLP) to verify claims effectively. While fact-checking research has seen considerable progress in high-resource languages such as English, less-resourced languages like Vietnamese remain underexplored, despite their growing online presence and the unique linguistic challenges they present.

Vietnamese fact-checking efforts are limited by the scarcity of annotated datasets and models optimized for the language. Existing datasets, such as ViWikiFC, focusing on structured Wikipedia-based claims, and ViFactCheck, a multi-domain benchmark, provide critical resources for advancing Vietnamese fact-checking. However, applying state-of-the-art models effectively to these datasets requires further investigation into both language-specific and cross-lingual approaches.

Large Language Models (LLMs) have shown remarkable performance in various NLP tasks, including fact-checking, thanks to their ability to generate coherent text and contextualize information. Despite their promise, their direct application to Vietnamese fact-checking remains challenging due to linguistic nuances, such as complex word segmentation and tonal variations. Baseline models like PhoBERT, ViBERT, and multilingual models such as XLM-Roberta and mBERT have demonstrated effectiveness in Vietnamese NLP tasks, but their comparative performance in fact-checking against advanced LLMs like Qwen-2.5-7B and URA-LLama-7B is yet to be systematically assessed.

The task of using LLMs for fact-checking in Vietnamese involves determining the veracity of a claim or statement by leveraging large language models applied to Vietnamese datasets.

Input: A text passage in Vietnamese containing information, along with a claim that needs to be verified based on the provided text.

Output: A predicted label for the claim, which can be one of the following:

- **Support:** The claim is verified as true according to the provided evidence.
- **Refuted:** The claim is determined to be false based on the available evidence.
- **Not Enough Information (NEI):** The claim cannot be verified due to insufficient evidence in the provided text.

The aim of this task is to develop an automated system that can assist human fact-checkers by efficiently determining the truthfulness of claims in Vietnamese texts. By leveraging large language models, the system aims to improve the speed and accuracy of fact-checking processes, ultimately contributing to the fight against misinformation in Vietnamese-language sources across various media platforms.

This paper makes the following key contributions:

- Fact-Checking Framework for Vietnamese: We propose an automated fact-checking system tailored for Vietnamese, utilizing large language models (LLMs) to assess the veracity of claims based on evidence from Vietnamese text.
- Evaluation on Vietnamese Datasets: We evaluate various LLMs, including Qwen-2.5-7B, URA-LLama-7B, PhoBERT, mBERT, ViBERT, XLM-Roberta, and vELECTRA, using two benchmark datasets, ViWikiFC and ViFactCheck, to assess their effectiveness in Vietnamese fact-checking.
- Pre-Trained vs. Domain-Specific Models: We compare general-purpose models (Qwen-2.5-7B, URA-LLama-7B) with Vietnamese-specific models (PhoBERT, mBERT, ViBERT), emphasizing the importance of fine-tuning for language-specific tasks.

By conducting a comprehensive analysis of model performance across key metrics such as F1-score, accuracy, and precision, this study contributes to bridging the gap in Vietnamese fact-checking research, laying the groundwork for future advancements in this critical area.

The paper is structured as follows. Section 2 gives an overview of the current research on the fact-checking task. Section 3 presents the dataset that the team used. Section 4 presents the methods that the team applied to the dataset to solve the fact-checking task. Section 5 describes the team's experiments on the dataset and the results of the analysis. Section 6 discusses the limitations and suggests future directions. Finally, Section 7 concludes the paper.

2 Related Work

Fact-checking has become a critical area of research in the context of combating misinformation, with recent advances leveraging large language models (LLMs) to improve efficiency and accuracy of this task. Vietnamese, as a less-resourced language in the field of NLP, has been gaining attention for dataset development, model adaptation, and innovative methodologies tailored to its linguistic challenges in the fact-checking domain.

Fact-Checking Datasets and Benchmarks The development of robust fact-checking systems relies heavily on the availability of high-quality datasets. Several datasets have been created for fact-checking in various languages and domains. For Vietnamese, two key datasets are particularly relevant to our work:

ViWikiFC (Le et al., 2024) introduced a dataset specifically designed for verifying claims related to Vietnamese Wikipedia-based textual knowledge sources. The dataset contains both the claims and the corresponding evidence from Wikipedia, making it essential for training models that can differentiate between true and false statements based on factual sources.

ViFactCheck (Hoa et al., 2024) expanded the scope of fact-checking for the Vietnamese language by introducing a comprehensive benchmark dataset that spans multiple domains, including politics, health, and science. ViFactCheck provides a diverse collection of claims, evidence, and veracity labels from a wide range of sources, it provides a multi-domain perspective essential for generalizing fact-checking techniques.

Role of LLMs in Fact-Checking With the advent of neural network-based approaches, LLMs have emerged as powerful tools for tackling complex NLP tasks, including fact-checking. Their ability to understand and generate human-like text positions them as key players in identifying relevant evidence, assessing source credibility, and predicting claim veracity.

For Vietnamese, research such as **Evaluating Large Language Model Capability in Vietnamese Fact-Checking Data Generation** (To et al., 2024) has demonstrated the potential of LLMs for generating fact-checking data for Vietnamese. By generating synthetic data, LLMs not only enhance dataset diversity but also address the scarcity of annotated resources for low-resource languages like Vietnamese. This underscores the dual role of LLMs as both tools for verification and enablers of data generation, amplifying their impact on the fact-checking pipeline.

Advancements in Pre-Trained Models

The progress in Vietnamese fact-checking has been further bolstered by the development of pre-trained models. **PhoBERT** (Nguyen and Nguyen, 2020) and **ViBERT**, specifically pre-trained on Vietnamese corpora, have demonstrated strong contextual understanding and the ability to handle linguistic nuances unique to Vietnamese efficiently. These models serve as foundational tools for tackling various NLP tasks, including fact-checking.

Multilingual models, such as **mBERT** and **XLM-Roberta**, have been applied to fact-checking tasks, provide another layer of support through cross-lingual transfer learning. Their capacity to generalize across languages enables researchers to leverage resources from high-resource languages, complementing Vietnamese-specific models.

Recent developments also include **qwen-2.5-7b** and **ura-llama-7b**, which are large-scale models with architectures optimized for cross-lingual and multi-domain tasks. These models are equipped with the capacity to analyze complex textual claims and evidence, and their application in Vietnamese fact-checking tasks has provided promising results. The use of such advanced models enables better handling of the linguistic nuances and specific challenges posed by the Vietnamese language, such as word segmentation and ambiguity in text.

Despite these achievements, significant gaps remain, particularly in applying LLMs for fact-checking Vietnamese text at scale. Future research can further explore hybrid approaches that combine domain-specific datasets with advanced prompting strategies, leveraging the full potential of state-of-the-art LLMs. By addressing these challenges, we aim to contribute to the development of robust, scalable fact-checking systems that meet the unique demands of the Vietnamese language and context.

3 Dataset

3.1 Data Description

In the context of Vietnamese, the lack of high-quality, annotated datasets for claim verification presents a significant challenge for advancing Natural Language Processing (NLP) in this language. To address this gap, our research leverages two essential datasets: ViWikiFC and ViFactCheck, which serve as benchmarks for automated fact-checking systems in Vietnamese.

ViWikiFC is the first large-scale, open-domain dataset for automated fact-checking in Vietnamese. It is specifically designed for verifying claims made on Wikipedia. The dataset contains **20,916** manually annotated claims, along with evidence extracted from Vietnamese Wikipedia articles. This dataset provides a valuable resource for developing and evaluating fact-checking systems in a structured and open-domain environment.

ViFactCheck is a multi-domain fact-checking dataset that focuses on verifying claims across diverse topics. It contains **7,232** human-annotated claims covering **12** distinct topics such as Politics, Education, and Urban Development. The claims and evidence were collected from reputable online news sources in Vietnam. This dataset introduces the complexity of handling claims from real-world and diverse contexts, making it ideal for cross-domain fact-checking tasks.

3.2 Data Analysis

Table 1 highlights the key differences between the ViWikiFC and ViFactCheck datasets, which are central to our study on Vietnamese fact-checking tasks. ViFactCheck comprises 7,232 samples divided into training, development, and testing sets in a 7:1:2 ratio. It includes three labels: Supports, Refutes, and Not Enough Information. Similarly, ViWikiFC, a larger dataset with 20,919 samples, is split into subsets with an 8:1:1 ratio. Both datasets share the same label structure, ensuring alignment and consistency.

Table 2 and Figure 1 reveal a balanced distribution of labels across both datasets. This balance is critical for effective training and evaluation of machine learning models. Balanced label distributions support the development of generalizable models that perform well across different tasks without overfitting to specific labels.

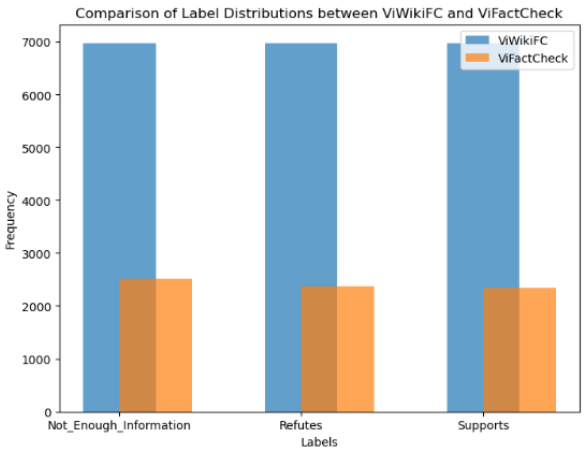

Figure 1: Comparison of label counts between the ViWikiFC and ViFactCheck datasets.

Feature	ViWikiFC	ViFactCheck
Data Source	Wikipedia	Reliable online information pages
Labels	Supports, Refutes, Not Enough Information	Supports, Refutes, Not Enough Information
Structure	Claim + 1 evidence	Claim + multiple evidence

Table 1: Comparison between ViWikiFC and ViFactCheck

Label	ViWikiFC	ViFactCheck
Not Enough Information	6,978	2,515
Refutes	6,973	2,370
Supports	6,968	2,347

Table 2: Comparison of label distribution between ViWikiFC and ViFactCheck datasets

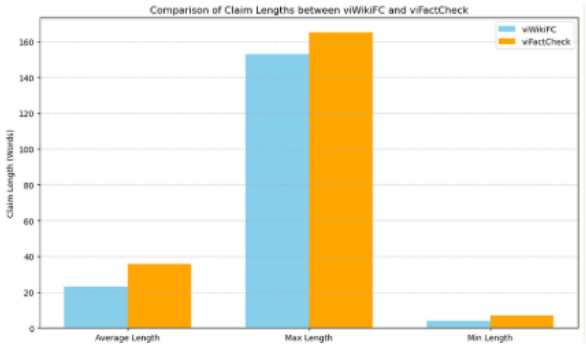

Figure 2: Compare the length of the claims in the two datasets ViWikiFC and ViFactCheck.

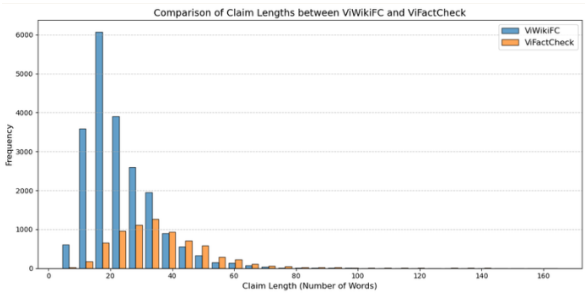

Figure 3: Comparison claim length distribution of ViWikiFC and ViFactCheck datasets.

The datasets also differ significantly in the length of claims and contexts. As summarized in Table 3, ViFactCheck features longer and more complex claims, averaging 36 words, with a maximum length of 165 words and a minimum of 7 words. By contrast, ViWikiFC contains shorter, more concise claims, averaging 23 words, with a maximum of 95 words and a minimum of 4 words. Figure 2 illustrates this comparison, emphasizing the greater complexity in real-world claims handled by ViFactCheck versus the structured claims derived from Wikipedia in ViWikiFC.

Feature	ViWikiFC	ViFactCheck
Avg_Evidence	35	42
Min_Evidence	16	0
Max_Evidence	251	216
Avg_Claim	23	36
Min_Claim	4	7
Max_Claim	153	165

Table 3: Comparison of statistics between the ViWikiFC and ViFactCheck datasets

Context length is another area of distinction between the datasets. ViFactCheck provides detailed and extensive contexts, averaging 690 words, with the longest reaching 3,603 words and the shortest at 71 words. Meanwhile, ViWikiFC has shorter contexts, averaging 120 words, with the longest at 578 words and the shortest at 15 words. The longer contexts in ViFactCheck indicate its focus on providing comprehensive background information, whereas ViWikiFC's shorter contexts are simpler and more focused, reflecting Wikipedia's structured nature. Evidence length further differentiates the datasets. ViFactCheck includes diverse evidence segments, averaging 1.2 segments per sample and reaching up to 9 segments in some cases. Each evidence segment averages 42 words, with a maximum of 216 words and a minimum of 5 words. In comparison, ViWikiFC typically contains one evidence segment per sample, averaging 35 words, with lengths ranging from 16 to 251 words. Figure 3 highlights the evidence distribution, showcasing ViFactCheck's higher variability and complexity.

In conclusion, both datasets provide valuable resources for developing fact-checking models for the Vietnamese language. ViFactCheck addresses the need for handling complex, real-world information, while ViWikiFC offers a more straightforward,

structured approach ideal for foundational tasks. Together, they complement each other, enabling the creation of robust and versatile fact-checking systems tailored to the unique demands of the Vietnamese language and context.

4 Experimental Method

4.1 Model Configuration

In the experiment of fact-checking using LLMs for Vietnamese, we used NVIDIA Tesla P100 GPU and NVIDIA T4x2 GPU on the Kaggle platform, along with 16 GB RAM and Intel Xeon processor to conduct experiments and train models. The LLMs model was deployed with quantization settings optimized for efficiency, using 4-bit precision (‘nf4‘ quantization type) with ‘torch.float16‘ as the compute data type, alongside double quantization for memory efficiency. The tokenizer is configured with a maximum sequence length of 2048 tokens and special handling for padding tokens, ensuring compatibility with input data. A zero-shot prompting strategy is employed, where a structured Vietnamese-language prompt. The pre-trained model was implemented with PyTorch library and used Transformers library to optimize the training and evaluation process. We used batch size of 16 for both training and evaluation, 3 epochs, used AdamW as optimizer and warmup_steps = 500 to improve convergence speed, learning rate in the range [1e-5, 1e-4].

4.2 Evaluation Metrics

In evaluating fact-checking models, we used a combination of several metrics to ensure a comprehensive assessment: accuracy, precision, recall, F1-score, as well as the macro average, micro average, ROC-AUC and PR-AUC.

Accuracy provides a straightforward measure of the model's overall correctness, indicating the proportion of correct predictions. Precision measures how accurate the model is when predicting positive instances, showing the proportion of true positives out of all predicted positives. Recall reflects the model's ability to identify all relevant instances, representing the proportion of true positives out of all actual positives. The F1-score combines precision and recall into a single metric, providing a balanced measure of the model's performance, particularly in cases where there is an imbalance between precision and recall.

The micro average F1-score aggregates the contributions of all classes into a global count, treating each instance equally, which is useful when the focus is on overall performance across all samples. In contrast, the macro average F1-score calculates the metric for each class independently and then averages the results, giving equal importance to all classes, regardless of their frequency. These metrics together offer a thorough evaluation of the model's effectiveness, addressing both overall accuracy and class-specific performance.

Additionally, we incorporated ROC-AUC, which evaluates the model's ability to distinguish between classes, and PR-AUC, which assesses the precision-recall trade-off across different thresholds. This combination of metrics allowed for a well-rounded evaluation of the model's performance across various aspects.

4.3 Baseline Model

4.3.1 Pre-Trained Models

Building on the notable success of transformer-based models in previous fact-checking research (Thorne and Vlachos, 2018; Hu et al., 2022; Nørregaard and Derczynski, 2021), this study employs pre-trained language models for Vietnamese fact-checking. Specifically, we utilize architectures rooted in BERT (Devlin et al., 2019) and RoBERTa (Liu, 2019), evaluating their performance on the task. This work focuses on two multilingual models and two monolingual models, each offering unique strengths for verifying factual claims.

The multilingual BERT model (mBERT) (Devlin et al., 2019) trained on a diverse corpus spanning 104 languages, including Vietnamese, is well-suited for tasks requiring broad linguistic adaptability. Its ability to process data in multiple languages allows it to analyze Vietnamese text while maintaining contextual relevance. This versatility makes mBERT a valuable tool for validating information within the Vietnamese fact-checking framework.

Cross-lingual Language Model - RoBERTa (XLM-R) (Conneau and Lample, 2019) extends this multilingual capability, being pre-trained on over 100 languages. Its proficiency in understanding and synthesizing information across linguistic boundaries makes it particularly advantageous for verifying claims that involve multilingual content. This cross-lingual capability provides additional context for fact-checking tasks, enhancing the evaluation of claims beyond Vietnamese.

PhoBERT (Nguyen and Nguyen, 2020), an

adaptation of the RoBERTa (Liu, 2019) architecture fine-tuned for Vietnamese, is specifically designed to handle the intricacies and cultural nuances of the language. By focusing exclusively on Vietnamese text, PhoBERT demonstrates exceptional performance in identifying subtle variations and maintaining high accuracy when applied to Vietnamese datasets, ensuring reliable results for fact-checking.

ViBERT, based on the BERT architecture and specifically designed for Vietnamese, was introduced by (The et al., 2020). This is another model optimized for Vietnamese, was developed using 10GB of text from a monolingual corpus. Its targeted training equips it to handle Vietnamese content with high precision, setting it apart from general multilingual models. By focusing solely on the language, ViBERT achieves enhanced efficiency and effectiveness for tasks requiring detailed linguistic understanding.

vELECTRA, built upon the ELECTRA architecture (Clark, 2020), is trained on 60 GB of Vietnamese data from NewsCorpus and OscarCorpus. Unlike BERT-based models (Devlin et al., 2019), vELECTRA employs the Replaced Token Detection task. Two neural networks are trained in parallel: a discriminator and a generator, enhancing performance and efficiency in processing Vietnamese data.

Through an analysis of these models, we aim to highlight their contributions to improving fact-checking systems for Vietnamese. By combining their strengths—ranging from multilingual adaptability to monolingual precision—this study seeks to advance the fight against misinformation, promoting a more trustworthy information environment.

4.3.2 Large Language Models

Recent advances in large language models (LLMs), which exhibit strong contextual understanding, have demonstrated their effectiveness in tasks such as contextual comprehension and reasoning, including fact-checking. Consequently, we employ several primary models that are suitable for lowresource configurations.

URA-Llama-7B

URA-Llama (Truong et al., 2024) is a version of the LLaMA (Large Language Model Meta AI) (Touvron et al., 2023) family, originally developed by Meta AI. This model is a large-scale, transformer-based language model with 7 billion parameters. While designed to generalize across multiple languages and tasks, its baseline performance in this experiment reflects a lack of fine-tuning for specific tasks or languages like Vietnamese. As a baseline, URA-Llama-7B serves as a general-purpose language model, illustrating the starting point for models that have not been explicitly adapted to the fact-checking or Vietnamese linguistic context.

Qwen-2.5-7B

Qwen-2.5 (Yang et al., 2024) is a generative language model developed with 7 billion parameters. It is designed for broad natural language understanding and generation tasks, leveraging its large-scale pretraining on multilingual corpora. However, similar to URA-Llama-7B, Qwen-2.5-7B has not been fine-tuned for Vietnamese-specific tasks or fact-checking, which limits its baseline performance in this evaluation. It serves as another example of a generalized LLM, allowing comparisons with fine-tuned and specialized models in this experiment.

5 Experiment and Analysis

5.1 Result Analysis

The results in Table 4 and Table 5 reveal significant performance gaps between large language models (LLMs), such as qwen2.5-7b and ura-llama-7b, and other transformer-based models fine-tuned for Vietnamese tasks. On both the ViWikiFC and ViFactCheck datasets, qwen2.5-7b and ura-llama-7b exhibit limited effectiveness, with F1-scores (micro and macro) and accuracy consistently around 0.32–0.34. These results indicate that despite their large-scale pretraining, these LLMs struggle to adapt to the nuances of Vietnamese text and fact-checking scenarios without task-specific fine-tuning or exposure to Vietnamese-language corpora during pretraining.

In contrast, models like PhoBERT and XLM-Roberta, which are either specifically pre-trained for Vietnamese or extensively multilingual, demonstrate significantly better performance. This suggests that the generalized capabilities of LLMs are insufficient for high-quality fact-checking in resource-specific languages like Vietnamese. The limited precision and recall of qwen-2.5-7b and ura-llama-7b further highlight their challenges in handling evidence-based reasoning and contextual understanding required for fact verification tasks.

Model	F1-score (micro)	F1-score (macro)	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
qwen2.5 - 7b	0.34	0.17	0.34	0.11	0.34	0.50	0.45
ura-llama-7b	0.34	0.17	0.34	0.11	0.34	0.50	0.45
PhoBERT	0.7341	0.7334	0.7341	0.7335	0.7355	0.8857	0.8075
mBERT	0.7547	0.7552	0.7547	0.7551	0.7557	0.9030	0.8358
ViBERT	0.6595	0.6600	0.6595	0.6595	0.6607	0.8378	0.7380
XLM-roberta	**0.7551**	**0.7554**	**0.7551**	**0.7553**	0.7561	**0.9005**	**0.8309**
vELECTRA	0.7006	0.6998	0.7006	0.7028	0.7022	0.8627	0.7775

Table 4: Results of Models on ViWikiFC Dataset

Model	F1-score (micro)	F1-score (macro)	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
qwen2.5 - 7b	0.32	0.12	0.32	0.11	0.32	0.49	0.64
ura-llama-7b	0.34	0.22	0.34	0.24	0.34	0.49	0.62
PhoBERT	0.7560	0.7568	0.7560	0.7656	0.7560	0.9090	-
mBERT	0.7574	0.7583	0.7574	0.7619	0.7574	0.9071	-
ViBERT	0.6144	0.6154	0.6144	0.6195	0.6144	0.8059	-
XLM-roberta	**0.7823**	**0.7820**	**0.7823**	**0.7828**	**0.7823**	**0.9210**	**-**
vELECTRA	0.7125	0.7140	0.7125	0.7317	0.7125	0.8886	-

Table 5: Results of Models on ViFactCheck Dataset

5.2 Error Analysis

The application of large language models (LLMs) for the classification of factual statements into predefined labels (Support, Refuted, Not Enough Info) demonstrates notable limitations that impede achieving optimal performance. A primary source of error originates from the design of the prompting strategy, which has not been adequately optimized to guide the model effectively. As a result, the model frequently generates responses that deviate from the expected format or fail to align with the task requirements.

Specifically, the current prompting approach lacks sufficient clarity and detailed instruction, which can lead to inconsistent or incorrect predictions. This issue becomes particularly pronounced in cases involving complex or conflicting inputs, where the model struggles to resolve ambiguities in the evidence or context provided. Consequently, the generated outputs may fail to adhere to the three target categories (Support, Refuted, Not Enough Info) or may misclassify statements due to contextual misunderstanding.

Furthermore, the structure and length of the prompt contribute to these challenges. In scenarios where the input context is lengthy or involves intricate relationships between evidence and claims, the model exhibits difficulty maintaining focus on the relevant information, which adversely impacts classification accuracy. This suggests that the prompt design may exceed the model's ability to process and integrate complex contextual information effectively.

5.3 Future Directions

Future research will focus on addressing the identified challenges to further enhance the performance of large language models (LLMs) in fact-checking tasks. First, optimizing prompt design is crucial, as clear and concise prompts with explicit instructions and task-specific examples can reduce ambiguity and guide the model toward accurate outputs. Advanced techniques such as Tree-of-Counterfactual Prompting and Zero-shot Prompting should be explored to increase the effectiveness of task formulation. Additionally, the model's difficulty in handling lengthy or complex contextual inputs highlights the need for improved processing strategies, such as context summarization, hierarchical processing, or segmenting large contexts into manageable parts.

Expanding experimentation with high-quality Vietnamese datasets and diverse LLMs is essential. Developing balanced datasets across all fact-checking labels and applying fine-tuning techniques can help models better understand task nuances and address data imbalances. Integrating external knowledge retrieval systems and multimodal evidence, such as text, images, and videos, will enhance reasoning capabilities. Leveraging larger, more advanced LLMs with improved processing power can overcome current limitations.

Finally, implementing feedback loops for continuous learning and robust evaluation frameworks

will improve accuracy, reliability, and performance on challenging edge cases, driving advancements in real-world fact-checking applications.

6 Conclusion

This study explores the application of large language models (LLMs) in Vietnamese fact-checking, leveraging advanced transformer-based architectures and datasets tailored for this task. Among the evaluated models, multilingual models such as XLM-RoBERTa outperformed others, achieving the highest metrics across all evaluation categories, including F1-score, accuracy, and precision, demonstrating their superior adaptability and reasoning capabilities. While monolingual models like PhoBERT and mBERT also showed strong results, especially in understanding Vietnamese linguistic nuances, they were slightly outperformed by XLM-RoBERTa in terms of overall robustness and consistency.

In contrast, general-purpose LLMs such as Qwen-2.5-7B and URA-LLama-7B lagged significantly behind task-specific models, highlighting the limitations of these baseline LLMs in handling specialized fact-checking tasks. These findings underscore the importance of model architecture and task-specific training in enhancing fact-checking performance.

This research demonstrates the transformative potential of LLMs in combating misinformation within the Vietnamese context and highlights the need for optimized prompt design, fine-tuning techniques, and diverse high-quality datasets to improve their effectiveness. Future work will focus on addressing these challenges by integrating external knowledge retrieval, leveraging multimodal evidence, and expanding experimentation with advanced LLMs to further enhance the accuracy and reliability of Vietnamese fact-checking systems.

References

K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Tran Thai Hoa, Tran Quang Duy, Khanh Quoc Tran, and Kiet Van Nguyen. 2024. Vifactcheck: A new benchmark dataset and methods for multi-domain news fact-checking in vietnamese. *arXiv preprint arXiv:2412.15308*.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. *arXiv preprint arXiv:2206.11863*.

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. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Jeppe Nørregaard and Leon Derczynski. 2021. Danfever: claim verification dataset for danish. In *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, pages 422–428.

Viet Bui The, Oanh Tran Thi, and Phuong Le-Hong. 2020. Improving sequence tagging for vietnamese text using transformer-based neural models. *Preprint*, arXiv:2006.15994.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*.

Long Truong To, Hung Tuan Le, Dat Van-Thanh Nguyen, Manh Trong Nguyen, Tri Thien Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2024. Evaluating large language model capability in vietnamese fact-checking data generation. *arXiv preprint arXiv:2411.05641*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sang T Truong, Duc Q Nguyen, Toan Nguyen, Dong D Le, Nhi N Truong, Tho Quan, and Sanmi Koyejo. 2024. Crossing linguistic horizons: Finetuning and comprehensive evaluation of vietnamese large language models. *arXiv preprint arXiv:2403.02715*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.