ViFactCheck: A New Benchmark Dataset and Methods for Multi-Domain News Fact-Checking in Vietnamese

Tran Thai Hoa[1,2], **Tran Quang Duy**[1,2], **Khanh Quoc Tran**[1,2], **Kiet Van Nguyen**[1,2*]

[1] Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
{21522082,20512013}@gm.uit.edu.vn, {khanhtq,kietnv}@uit.edu.vn

Abstract

The rapid spread of information in the digital age highlights the critical need for effective fact-checking tools, particularly for languages with limited resources, such as Vietnamese. In response to this challenge, we introduce ViFactCheck, the first publicly available benchmark dataset designed specifically for Vietnamese fact-checking across multiple online news domains. This dataset contains 7,232 human-annotated pairs of claim-evidence combinations sourced from reputable Vietnamese online news, covering 12 diverse topics. It has been subjected to a meticulous annotation process to ensure high quality and reliability, achieving a Fleiss Kappa inter-annotator agreement score of 0.83. Our evaluation leverages state-of-the-art pre-trained and large language models, employing fine-tuning and prompting techniques to assess performance. Notably, the Gemma model demonstrated superior effectiveness, with an impressive macro F1 score of 89.90%, thereby establishing a new standard for fact-checking benchmarks. This result highlights the robust capabilities of Gemma in accurately identifying and verifying facts in Vietnamese. To further promote advances in fact-checking technology and improve the reliability of digital media, we have made the ViFactCheck dataset, model checkpoints, fact-checking pipelines, and source code freely available on GitHub. This initiative aims to inspire further research and enhance the accuracy of information in low-resource languages.

Github — https://github.com/TTHHA/ViFactCheck

1 Introduction

The rapid proliferation of digital information has created significant challenges in distinguishing between accurate and false information. The spread of disinformation, rumors, and fake news has become a global concern with far-reaching consequences for individuals, societies, and public discourse. As noted by Lazer et al. (2018), the extensive spread of fake news can have severe negative impacts on individuals and society. It can cause confusion and misunderstanding, disrupt social order, and even threaten national security.

Fact-checking, a rigorous process to verify the accuracy of claims in specific contexts, relies on informed individuals using evidence, reasoning, and available information to make well-founded judgements. Figure 1 provides a specific illustration for Vietnamese fact-checking. Although substantial efforts have been devoted to fact-checking in English (Thorne et al. 2018; Aly et al. 2021; Schuster, Fisch, and Barzilay 2021), resources for fact-checking in low-resource languages like Vietnamese are limited. This scarcity primarily stems from the limited availability of guidance resources to analyze the structure and semantics of Vietnamese.

To bridge this gap, this study presents the development of ViFactCheck, the first publicly available human-curated fact-checking benchmark tailored to multiple domains Vietnamese news. Our main contributions are described as follows:

Figure 1: An example of the Vietnamese fact-checking task. Words highlighted in blue represent key evidence used to support the classification of the claim as "Supported".

*Corresponding author.

1. **Dataset Construction:** We developed ViFactCheck, a comprehensive dataset encompassing 12 critical domains of Vietnamese online news. This dataset contains 7,232 rigorously vetted human-annotated claims, thereby ensuring a robust foundation for both research and practical applications.
2. **Model Experimentation:** We utilized fine-tuning and prompting techniques to several state-of-the-art language models using the ViFactCheck dataset to assess their effectiveness in verifying information within the Vietnamese context. Our study includes fine-tuning and zero-shot in-context learning on both pre-trained and large language models, specifically adapted to this linguistic framework, to evaluate their efficacy.
3. **In-depth Analysis**: Through detailed examinations of the challenges faced during the creation of the dataset and subsequent experimentation, this study offers profound insights into the hurdles of developing fact-checking systems for low-resource languages, guiding future advancements in the field.

The remainder of this paper is structured as follows. Section 2 delves into the fundamentals of fact-checking tasks. Section 3 describes the process of constructing the ViFactCheck benchmark dataset. Section 4 discusses the results of our experiments and identifies key challenges encountered. Section 5 concludes with a summary of our findings and suggests directions for future research.

2 Fundamental of Fact-Checking

2.1 Foundational Benchmark Datasets

Benchmark datasets are crucial in the development and evaluation of fact-checking algorithms, serving as the foundation upon which these systems are tested and fine-tuned. The FEVER (Thorne et al. 2018) is particularly notable, containing more than 185,000 claims sourced from Wikipedia, each meticulously annotated with evidence to support or refute the claims. Following FEVER, the FEVEROUS dataset (Aly et al. 2021) extends these capabilities by incorporating not only text but also structured data such as tables and lists, presenting a more comprehensive dataset that challenges algorithms to parse and verify information across different formats. Another significant dataset, MultiFC (Augenstein et al. 2019), compiles claims from 26 different fact-checking websites, covering various topics and offering a rich environment to test the adaptability of verification systems to different contexts and types of misinformation. These benchmark datasets play a critical role in advancing the field of fact-checking, providing a diverse set of challenges and inspiring the development of diverse open-domain fact-checking datasets in many languages (Schuster, Fisch, and Barzilay 2021; Wang 2017; Hu et al. 2022; Nørregaard and Derczynski 2021; Khouja 2020). The comparison of multi-domain fact-checking datasets is summarized in Table 1.

2.2 Advanced Methods in Fact-Checking

The evolution of fact-checking methods has significantly advanced through the adoption of sophisticated machine learning technologies. Notably, the use of Pre-trained Language Models (PLMs) and Large Language Models (LLMs) like BERT and other transformer-based architectures (Devlin et al. 2019) has been instrumental. These models, leveraging deep learning, are highly effective in processing and analyzing the context within texts based on patterns in data, making them exceptionally effective for tasks such as evidence retrieval and claim verification (Nie, Chen, and Bansal 2019; Soleimani, Monz, and Worring 2020; Liu et al. 2020). By fine-tuning these models on specific fact-checking datasets, researchers can adapt their capabilities to better recognize and interpret the nuances of misinformation. Furthermore, researchers have explored prompting techniques with these models to direct their focus without extensive retraining, enhancing their utility in diverse applications (Huang, Chan, and Ji 2023; Pan et al. 2023). The synergy of language models with traditional retrieval and verification methods has also given rise to hybrid models, which combine the depth and adaptability of machine learning with the precision of rule-based systems (Vlachos and Riedel 2014), graph modeling (Zhong et al. 2020), leading to more robust and accurate fact-checking solutions.

2.3 Vietnamese Research on Fact-Checking

Research within Vietnam on fact-checking has been making significant strides, particularly with the development of customized datasets that address the unique linguistic characteristics of Vietnamese (Duong, Do et al. 2022; Le et al. 2024). A notable study by Duong, Ho, and Do (2023) has produced a dataset with more than 129K triples checked for fact, specifically designed to evaluate the effectiveness of fact-checking algorithms under Vietnamese linguistic constraints. This approach not only enhances the precision of fact-checking in Vietnam but also contributes significantly to the global body of knowledge. It showcases how fact-checking technologies can be adapted to different linguistic and cultural contexts, providing a model for similar adaptations in other regions.

3 Dataset Creation Process

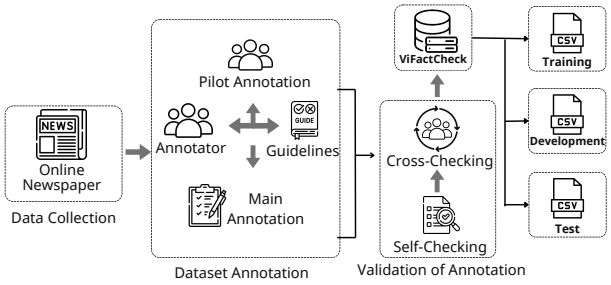

Figure 2: The ViFactCheck dataset contruction process.

Figure 2 shows the development of ViFactCheck, the first multi-domain Vietnamese news fact-checking benchmark. The dataset construction included three phases: data collection, dataset annotation, and annotation validation, each rigorously monitored by experts to ensure dataset quality.

	Dataset	Labels	# Claims	Annotated Evidence	Language	Source	#RS
English	FEVER (2018)	3	185,445	✔	English	Wikipedia	Multi
	FEVEROUS (2021)	3	87,026	✔	English	Wikipedia	Multi
	VitaminC (2021)	3	488,904	✗	English	Wikipedia	Single
	MultiFC (2019)	2-40	36,534	✔	English	Fact-check	Multi
	LIAR (2017)	6	12,836	✗	English	Fact-check	W/O
Non-English	CHEF (2022)	3	10,000	✔	Chinese	News/Fact-check	Multi
	DANFEVER (2021)	3	6,407	✔	Danish	Wikipedia	Multi
	ANT (2020)	2	4,547	✗	Arabic	News	Multi
	ViWikiFC (2024)	3	20,976	✗	Vietnamese	Wikipedia	Single
	ViFactCheck (Ours)	3	7,232	✔	Vietnamese	News	Multi

Table 1: Comparative overview of typical open-domain fact-checking datasets. The type of Reasoning Steps (#RS) column reflects the complexity involved in verifying the claims in each dataset.

3.1 Data Collection

This research constructs a dataset from articles sourced from nine licensed and widely-read Vietnamese online newspapers, detailed in the Appendix B. These sources were chosen for their comprehensive and timely news coverage, ensuring the relevance and reliability of the dataset. We extracted datasets that included titles, content, topics, lead descriptions, and URLs of articles published between February and March 2023. The selection of this period aims to capture the current dynamics of news reporting, providing a contemporary snapshot of media trends.

The initial corpus contained 1,000 articles covering 12 topics. Notably, news leads were merged with their respective contents to form a "Full Context" field, thereby enriching the dataset with a more comprehensive narrative view. This methodological rigor ensure the utility of dataset in advancing research on media analysis and computational linguistics.

3.2 Dataset Annotation

The construction methodology proposed for Vietnamese news differs from the conventional methods in previous datasets (Khouja 2020; Nørregaard and Derczynski 2021), which mimic the FEVER approach (Figure 3a). Recognizing the nuanced and dynamic nature of online news, our method employs human annotators to extract and interpret contextual nuances and factual details from news articles (Figure 3b). This human-centered approach enhances the naturalness and relevance of the data, enabling the dataset to better represent complex real-world information scenarios.

By assigning labels that reflect the context of each article, our methodology supports intricate inference tasks that require analysis across multiple pieces of evidence. This refined approach ensures that our dataset is exceptionally well-suited for advanced fact-verification systems, significantly contributing to the accuracy and effectiveness of misinformation detection in the digital media landscape.

Pilot Annotation is used to familiarize the annotators with the claim generation and verification process described above. Seven native Vietnamese-speaking university students

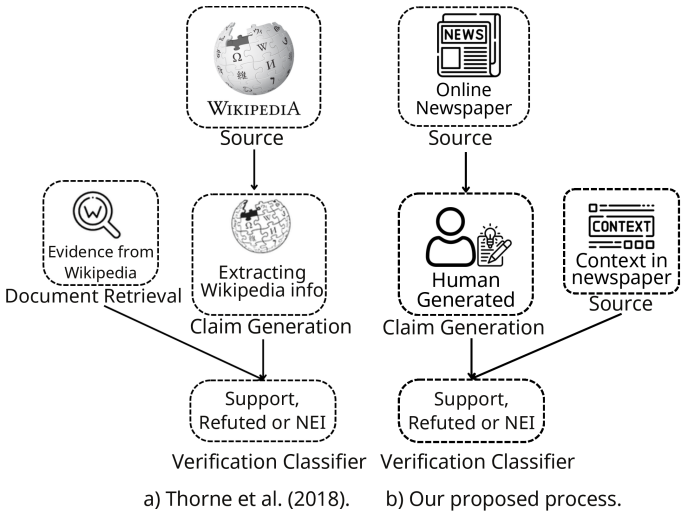

Figure 3: Comparison of the labeling pipelines in the FEVER and ViFactCheck datasets.

were involved as annotators. We conducted a pilot annotation with each annotator annotating 120 claims corresponding to 20 random articles. Annotators were instructed to proofread each claim carefully and rigorously in accordance with the annotation guidelines. Details of the annotators recruitment and guidelines can be found in Appendices C and D.

To verify the integrity of the pilot annotation process, we conducted thorough reviews of both the claims and their corresponding labels. The expert provided detailed feedback and asked the annotators to review any details or labels that did not meet the requirements of the annotation guidelines.

Main Annotation Following a pilot phase that familiarized the annotators with the tasks, each was assigned a specific subset to ensure focused and deep engagement. Throughout this phase, strict adherence to established guidelines was paramount to ensure consistency and enhance the overall quality of the dataset.

Claim Generation: Before generating any claims, annotators conducted a thorough review of the article. This meticulous process ensures a deep understanding of the multiple facets of the article, facilitating an accurate interpretation of the information. Annotators then employed their expertise to construct claims that align with the predefined labels: Support, Refute, and NEI (Not Enough Information). Such rigorous adherence to these guidelines is essential for generating contextually relevant claims, thereby enhancing the reliability of the dataset and its utility in advancing fact-checking.

Evidence Annotation: In terms of evidence annotation, the task extends beyond simple identification. Annotators are required to meticulously annotate the supporting evidence for each claim derived from the phrases previously collected from the articles. To enhance the complexity of the dataset and the challenge it presents, annotators are instructed not to limit their claims to single pieces of evidence. Instead, they are required to craft intricate claims that amalgamate multiple pieces of evidence (Appendix F). This process involves breaking down the claim, collating diverse evidences, and performing multi-step reasoning. The ability to synthesize complex evidence not only enriches the data but also crucially underpins more sophisticated analyses.

3.3 Validation of Annotation

After completing the main annotation phases, we implemented several strategies to ensure the quality and consistency of the dataset: **(1)** Self-checking: Annotators review their own claims and labels, checking for grammatical errors and typographical mistakes. **(2)** Cross-checking: Annotators verify the work of their peers. Any identified errors are collaboratively discussed and corrected.

Metric For Inter-Annotator Agreement: Fleiss Kappa is widely used to evaluate inter-annotator agreement (IAA) in several tasks and is considered a benchmark for such measurements (McHugh 2012; Thorne et al. 2018). Consequently, we utilized the Fleiss Kappa metric (Fleiss 1971) to assess inter-annotator agreement, thus ensuring quality assurance in human annotation.

We randomly selected 10% of the claims (n = 726) from the labeled dataset, assigning them to a group of three annotators. These claims, originally authored by different individuals, were relabeled without revealing the existing annotations. The inter-rater agreement was then calculated using the Fleiss Kappa measure. We achieved an agreement level of 0.83, indicative of a very high level of agreement among annotators, which confirms the high quality and reliability of our dataset.

3.4 Words Overlap and Semantic Similarity nalysis

To evaluate the complexity of inference within our dataset, we employed two principal metrics: word overlap and semantic similarity. For word overlap, we used metrics including Longest Common Sequence (LCS), New Word Ratio (%) (NWR), Jaccard Similarity (%) (JS), and Lexical Overlap. For semantic similarity, we utilized the concept of Related Words, generating embeddings with SBERT (2019) and calculating correlations using cosine similarity. The results are summarized in Table 2.

		LSC	NWR	JS	LO	RW
Context	Support	20.60	6.54	11.46	20.13	36.24
	Refute	18.10	11.50	10.06	17.90	34.00
	NEI	19.89	11.81	10.96	18.50	32.85
Evidence	Support	17.70	17.13	63.52	73.87	86.89
	Refute	15.46	25.47	54.63	66.69	81.41
	NEI	16.71	26.84	57.56	64.39	81.13

Table 2: Relationship between claim-context and claim-evidence in the ViFactCheck dataset.

McCoy, Pavlick, and Linzen (2019) demonstrated that models face difficulties with low overlap ratios, necessitating advanced inference capabilities. Our dataset features claim-context pairs with minimal word overlap and semantic similarity, complicating model inference. In contrast, a strong correlation between claim-evidence pairs significantly enhances the performance of models when the appropriate evidence is retrieved. Further detailed analysis can be found in the Appendix G.

4 Experiment and Results

4.1 Baseline Models

Drawing on the transformative impact of transformer-based models in prior fact-checking studies (Thorne et al. 2018; Hu et al. 2022; Nørregaard and Derczynski 2021), our research employs pre-trained language models (PLMs) that utilize the BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) architectures. Our study incorporates four specific models to address the fact-checking task, including two multilingual models mBERT (2019) and XLM-R (2020) and two monolingual models PhoBERT (2020) and ViBERT (2020) tailored to handle linguistic nuances effectively.

Moreover, the recent strides in large language models (LLMs) have solidified their utility in demonstrating robust contextual comprehension and reasoning capabilities, particularly in tasks that require deep understanding, such as fact-checking. Accordingly, our experimental framework includes four SOTA open-source LLMs designed for optimized performance in low-resource settings: Llama (2023), Gemma (2024), and Mistral (2023). These models are pivotal in our methodology, providing a comprehensive approach to evaluating their effectiveness across diverse linguistic contexts.

4.2 Software and Hardware Configurations

We employed the AdamW optimizer for fine-tuning pre-trained language models, as detailed by Loshchilov and Hutter (2019). The settings for these models included a learning rate of 5e-06, a dropout rate of 0.3, a batch size of 16, and a training duration of 10 epochs. Additionally, for PhoBERT, we segmented the text data using VnCoreNLP (Vu et al. 2018), adhering to the recommendations by Nguyen and Tuan Nguyen (2020). The dataset was partitioned into training, development, and test sets with a split ratio of 6:2:2, ensuring a balanced evaluation across all stages of model development.

For LLMs, we utilized the Unsloth framework with supervised fine-tuning using LoRA adaptation. The hyperparameters were configured with a Lora rank of 16, Lora alpha of 16, a learning rate of 2e-04, a batch size of 16, and 5 epochs. All experiments were conducted on a RTX 4090 GPU with 24GB of memory, utilizing PyTorch version 2.2.1 and Transformers version 4.41.2, and took a total of five days to complete. Details of the models, prompt templates and parameters can be found in Appendices H, I and J.

4.3 Main Results

Table 3 presents a detailed comparison of language models in fact-checking, examining their performance across different methods such as fine-tuning and prompting, and their efficiency in using Full Context versus Gold Evidence. Using the macro-average F1 score (%), the analysis provides insights into the capabilities of the models, highlighting the strengths and limitations of each approach in processing complex information sets.

Model	Full Context	Gold Evidence	Δ
Fine-tuning PLMs			
PhoBERT$_{base}$	68.55	77.76	↑9.21
PhoBERT$_{large}$	62.93	79.76	↑16.83
ViBERT	59.95	72.18	↑12.23
mBERT	58.07	69.94	↑11.87
XLM-R$_{base}$	65.40	81.10	↑15.70
XLM-R$_{large}$	75.42	88.02	↑12.60
Fine-tuning LLMs			
Gemma	**85.94**	**89.90**	↑3.96
Mistral	70.13	88.63	↑18.50
Llama2	41.47	79.53	↑**38.06**
Llama3	79.65	88.67	↑9.02
Prompting LLMs			
Gemini	76.26	74.88	↓1.38
Gemma	45.05	39.47	↓5.58
Mistral	61.02	57.31	↓3.71
Llama2	63.54	51.64	↓11.90
Llama3	65.21	63.10	↓2.11

Table 3: Performance comparison of baseline models on the ViFactCheck test set. Context and Evidence indicate the use of Full Context and Gold Evidence, respectively, for Claim Verification. The best scores are highlighted in bold; models that outperform other peers are underlined. Performance differences (Δ) are statistically significant ($p < 0.01$), confirming robust gains or reductions when Full Context is employed compared to Gold Evidence.

Fine-tuning Pre-trained Language Models Among the PLMs, XLM-R$_{large}$ stands out with exemplary performance, scoring 75.42% in Context and 88.02% in Evidence. These results suggest that the scale and design of XLM-R$_{large}$ provide a robust model capable of handling the complexities inherent in determining the veracity of claims based on the provided contexts and evidence. Additionally, variants of BERT-based models also demonstrate considerable gains, with PhoBERT$_{large}$ in particular showing a significant leap in context understanding compared to its peers.

Fine-tuning Large Language Models The LLMs, particularly Gemma, display remarkable effectiveness, outperforming other models in both Context (85.94%) and Evidence (89.90%) scores. This superior performance is likely due to the deeper learning capabilities and broader contextual understanding inherent in larger models. Variations in performance within this category also highlight the potential for specific architectural enhancements and targeted training strategies, as evidenced by the disparity between Llama2 and Llama3.

Fine-tuning PLMs and LLMs Fine-tuning both PLMs and LLMs consistently produces better results than prompting methods. Fine-tuning, which involves specific adjustments to model weights for the task, enables the models to directly learn detailed and nuanced patterns within the training data. The effectiveness of fine-tuning is particularly evident in scenarios involving Gold Evidence, where the fine-tuned model can precisely assess the validity of claims based on key information.

Performance with Gold Evidence versus Full Context The use of Gold Evidence typically results in higher accuracy scores across models compared to when the Full Context is provided. Gold evidence, being directly relevant to the claims, allows models to focus their computational power on a smaller, more pertinent dataset, thereby reducing the noise associated with broader contexts. This targeted approach leads to more precise verifications but does not necessarily prepare models for real-world scenarios where they must extract relevant information from extensive, unstructured data.

Prompting and Handling Full Context Models designed to handle extensive and complex contexts, such as Gemini, benefit from prompting techniques that leverage pre-trained knowledge to interpret new data without extensive re-training. This approach enables efficient navigation and processing of large datasets, making it especially suitable for applications that require the processing of generalized information. However, despite its capability to manage broader data, prompting generally falls short of achieving the accuracy delivered by fine-tuning, particularly when detailed specificity and deep data understanding are necessary.

Influence of Model Architecture and Size The results consistently reveal that larger models such as XLM-R$_{large}$ and Gemma surpass their smaller counterparts in both context and evidence metrics. The enhanced performance of these models is attributed to their expanded capacity, which is essential for addressing the intricacies associated with verifying claims. Equipped with extensive neural networks and deeper layers, these models possess greater computational power, enabling them to effectively model complex relationships and dependencies in the data. This allows for more effective information extraction and synthesis, providing a significant advantage in fact-checking tasks.

4.4 Analysis and Discussion

How Does the Evidences Retrieval Help? Our analysis of retrieval models in fact-checking sheds light on the operational dynamics of SBERT (Reimers and Gurevych 2019), BM25 (Robertson, Zaragoza et al. 2009), and their hybrid configurations under various conditions, with a focus on how well these models understand the semantic complexities of

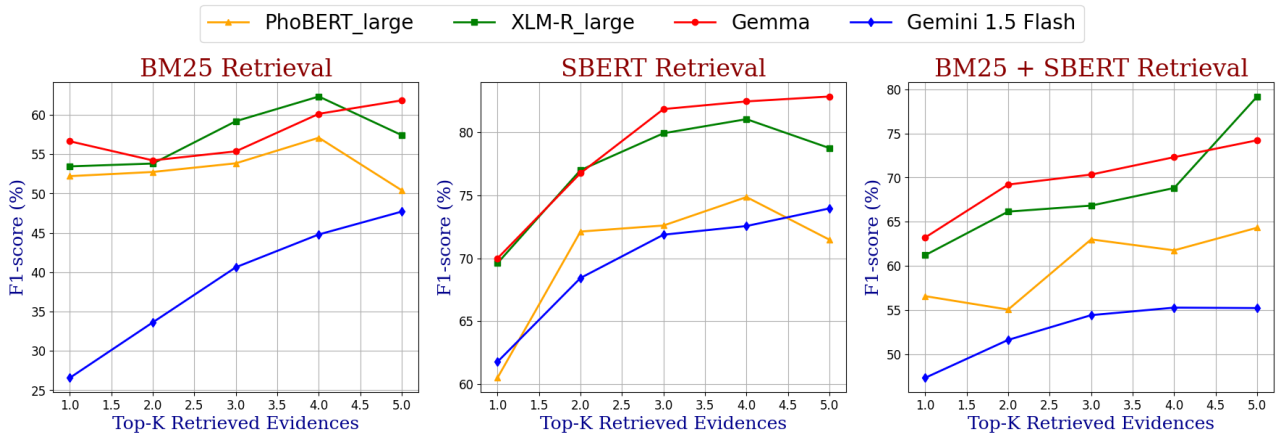

Figure 4: Comparative performance of various text retrieval models across different Top-K settings.

language processing (see Figure 4). The choice of these models for a more detailed evaluation is based on their superior performance across experiments, as discussed in Section 4.3.

A deeper dive into the results reveals that increasing the number of top-K retrieved evidences universally benefits all models by expanding the pool of potentially relevant information. However, the relationship between the number of documents retrieved (K) and the improvement in F1-score is not linear and varies significantly between different models and configurations. SBERT, in particular, shows a strong positive correlation between increased K and performance gains, indicating its effective use of broader contextual data.

Interestingly, performance improvements begin to plateau at higher K values in certain configurations, including Gemma within the SBERT model, suggesting an optimal K threshold of 5. This threshold represents the balance point where the benefits of additional document retrieval begin to decline relative to the computational costs. This insight is crucial for optimizing retrieval systems, emphasizing the need to balance data comprehensiveness with resource efficiency.

Furthermore, the distinct behavior of configurations like Gemini 1.5 Flash under SBERT, which scales effectively with an increase in K, underscores the potential for tailored approaches based on specific system capabilities and task requirements. Such adaptability is crucial in cases where the volume and variety of information vary dramatically.

How Multi-evidence Impacts Model Reasoning? The comparative performance of language models shows significant variations, particularly when comparing their ability to handle single-evidence versus multiple-evidence inputs, as depicted in Table 4. Gemma stands out for its robust capability in both of scenarios, benefitting significantly from training on a diverse, multilingual dataset. This extensive training enhances its adaptability and accuracy by enabling it to effectively manage complex contexts. Additionally, Gemma excels in data sufficiency assessments, effectively classifying the Not Enough Information (NEI) category across different scenarios, which is crucial for ensuring the reliability of fact-checking systems and preventing misinformation.

In single-evidence scenarios, the simplicity of the data allows models such as Llama3 to achieve higher accuracy. This straightforwardness typically presents less ambiguity, enabling the models to apply their verification capabilities more effectively. However, when multiple evidence sources are introduced, the added complexity significantly challenges all models. The noticeable decline in performance metrics in these scenarios highlights a gap in the ability of models to synthesize and integrate information from various sources, revealing a critical area for future enhancements.

4.5 Qualitative Error Analysis

Based on the macro F1 scores, we selected the Gemma model as our baseline to perform a detailed error analysis. As illustrated in Figure 5 and further detailed in Appendix K, we evaluated 100 random incorrect predictions from the development set to identify and categorize error types.

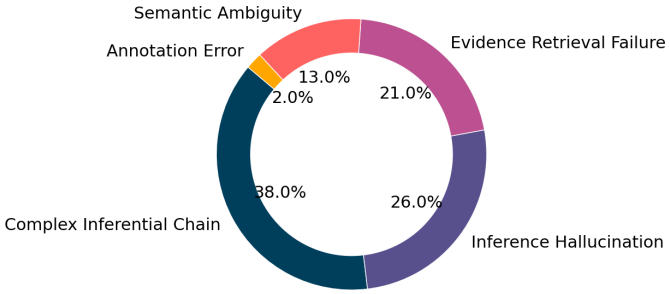

Figure 5: Distributions of errors.

The analysis revealed significant challenges in handling Semantic Ambiguity and Complex Inferential Chains, both of which are pivotal for refining NLP technologies. Semantic Ambiguity issues particularly highlight the necessity for context-aware processing (Baek et al. 2023; Wang et al. 2022; Wu et al. 2023). By integrating transformer-based models, the ability of the Gemma model to interpret complex linguistic contexts could be substantially improved, enhancing its accuracy in environments where nuance is critical.

	Single-evidence				Multiple-evidence				Overall			
	Avg. F1	**Support**	**Refute**	**NEI**	**Avg. F1**	**Support**	**Refute**	**NEI**	**Avg. F1**	**Support**	**Refute**	**NEI**
Fine-tuning PLMs												
PhoBERT$_{base}$	69.79	71.04	65.53	72.80	64.92	66.89	67.10	60.77	68.47	69.75	69.75	69.75
PhoBERT$_{large}$	75.01	76.27	70.81	77.95	62.72	64.44	67.12	56.58	71.47	72.64	69.59	72.18
ViBERT	56.42	61.38	46.31	61.59	58.11	63.64	56.23	54.46	57.16	62.08	49.57	59.81
mBERT	71.92	70.45	67.17	78.13	61.72	62.19	60.57	62.41	68.97	67.80	65.03	74.08
XLM-R$_{base}$	71.52	74.96	64.75	74.86	67.46	67.35	69.49	65.56	70.48	72.57	66.33	72.54
XLM-R$_{large}$	80.06	81.38	78.00	80.80	74.97	78.96	77.82	68.14	78.75	80.60	77.94	77.72
Fine-tuning LLMs												
Gemma	**83.99**	84.77	82.33	**84.87**	**79.52**	81.71	**81.96**	**74.88**	**82.85**	83.75	**82.21**	**82.59**
Mistral	83.62	85.26	**83.01**	82.60	77.35	**82.99**	78.11	70.94	81.89	84.52	81.41	79.75
Llama2	38.99	38.77	33.62	44.59	38.05	45.02	33.09	36.04	38.81	40.73	33.45	42.27
Llama3	83.45	**85.88**	80.64	83.82	75.02	82.01	77.51	65.55	81.18	**84.59**	79.65	79.29
Prompting LLMs												
Gemini	73.96	80.85	71.58	69.46	75.33	80.55	72.70	72.75	69.96	81.46	69.29	59.13
Gemma	49.53	53.33	54.16	41.09	52.08	55.67	56.19	46.55	49.76	61.26	52.31	35.71
Mistral	51.54	68.79	53.38	32.45	53.97	68.20	53.01	40.69	49.99	68.06	50.14	31.78
Llama2	36.12	65.43	11.95	30.99	43.58	64.08	30.83	35.83	33.16	67.14	19.68	22.66
Llama3	48.65	61.16	50.41	34.37	52.31	55.31	59.56	42.06	43.71	48.21	46.25	36.67

Table 4: Performance comparison of language models across Single and Multiple evidence scenarios.

Moreover, the frequent errors associated with Complex Inferential Chains expose the limitations of the model in synthesizing and reasoning across diverse informational inputs. The adoption of memory networks and knowledge graphs could markedly improve its capacity to process and link extended data sequences, thereby enhancing its reasoning and inference capabilities (Kim et al. 2023; Pan et al. 2023).

4.6 Human Performance

Table 5 presents an evaluation of fine-tuned models, offering crucial insights into their varied performances in the Support, Refute, and NEI compared to human performance. Models such as Gemma and Llama3 demonstrate strong capabilities in the Support and NEI categories, indicating their robustness in handling both direct and ambiguous information. However, their performance declines in the Refute category, highlighting a critical gap in the ability of AI to effectively process and analyze contradictory information.

Model	**F1 score**	**Support**	**Refute**	**NEI**
Fine-tuning PLMs				
PhoBERT$_{base}$	71.29	75.19	63.89	74.80
PhoBERT$_{large}$	73.08	79.70	62.30	77.24
ViBERT	55.66	68.70	48.28	50.00
mBERT	66.94	71.79	61.84	67.18
XLM-R$_{base}$	66.33	71.64	64.97	62.39
XLM-R$_{large}$	74.95	76.47	73.02	75.36
Fine-tuning LLMs				
Gemma	83.95	91.73	77.52	82.61
Mistral	66.61	77.46	62.69	59.68
Llama2	46.10	50.45	40.94	46.91
Llama3	84.24	91.97	77.05	83.69
Human Evaluating				
Human	**84.93**	**81.25**	**80.95**	**82.38**

Table 5: Evaluation results of human performance compared to the models on the test set of 200 samples. Models that outperform human evaluators are marked in gray.

This pattern is not isolated but is evident across various models, suggesting that current AI architectures and training paradigms may lack the sophisticated reasoning required to handle complex linguistic challenges that humans manage more adeptly. The comparative underperformance of AI in the Refute category underscores the need for integrating deeper contextual understanding and advanced reasoning mechanisms into AI systems to better mimic human cognitive abilities in processing contradictions and complex arguments.

5 Conclusion & Future Work

The development of the ViFactCheck dataset marks a transformative advancement in fact-checking for Vietnamese. This dataset comprises 7,232 entries across 12 topics, providing a substantial resource to assess various SOTA baseline models. Our work demonstrates the potential of using advanced language models, fine-tuned on this dataset, to achieve high levels of accuracy, as evidenced by a macro F1 score of 89.90%. This validates the efficacy of our dataset and methodologies in a real-world context, setting a new benchmark for fact-checking performance in low-resource languages. The challenges identified through our in-depth analysis, such as semantic ambiguity and evidence retrieval failures, not only underscore the complexity of fact-checking in such environments but also pave the way for targeted improvements.

Future research will focus on addressing the identified challenges to further enhance model performance. Efforts will include refining semantic understanding and evidence retrieval capabilities to handle ambiguous and complex datasets more effectively (Wang et al. 2022; Wu et al. 2023). In addition, we plan to develop methods to mitigate inference hallucinations and improve reasoning across complex inferential chains (Kim et al. 2023; Pan et al. 2023). Expanding the dataset to incorporate a wider range of misinformation types and correcting labeling errors will also be crucial (Gupta and Srikumar 2021; Augenstein et al. 2019).

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund of the VNUHCM University of Information Technology.

References

Aly, R.; Guo, Z.; Schlichtkrull, M.; Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Cocarascu, O.; and Mittal, A. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Augenstein, I.; Lioma, C.; Wang, D.; Chaves Lima, L.; Hansen, C.; Hansen, C.; and Simonsen, J. G. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4685–4697. Hong Kong, China: Association for Computational Linguistics.

Baek, J.; Aji, A. F.; Lehmann, J.; and Hwang, S. J. 2023. Direct Fact Retrieval from Knowledge Graphs without Entity Linking. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10038–10055. Toronto, Canada: Association for Computational Linguistics.

Bui, T. V.; Tran, T. O.; and Le-Hong, P. 2020. Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models. In Nguyen, M. L.; Luong, M. C.; and Song, S., eds., *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, 13–20. Hanoi, Vietnam: Association for Computational Linguistics.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Duong, H. T.; Do, P.; et al. 2022. Vietnamese Fact Checking based on the Knowledge Graph and Deep Learning. In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 530–535. IEEE.

Duong, H. T.; Ho, V. H.; and Do, P. 2023. Fact-checking Vietnamese Information Using Knowledge Graph, Datalog, and KG-BERT. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(10): 1–23.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Gupta, A.; and Srikumar, V. 2021. X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 675–682. Online: Association for Computational Linguistics.

Hu, X.; Guo, Z.; Wu, G.; Liu, A.; Wen, L.; and Yu, P. 2022. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3362–3376. Seattle, United States: Association for Computational Linguistics.

Huang, K.-H.; Chan, H. P.; and Ji, H. 2023. Zero-shot Faithful Factual Error Correction. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5660–5676. Toronto, Canada: Association for Computational Linguistics.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Khouja, J. 2020. Stance Prediction and Claim Verification: An Arabic Perspective. In Christodoulopoulos, C.; Thorne, J.; Vlachos, A.; Cocarascu, O.; and Mittal, A., eds., *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, 8–17. Online: Association for Computational Linguistics.

Kim, J.; Park, S.; Kwon, Y.; Jo, Y.; Thorne, J.; and Choi, E. 2023. FactKG: Fact Verification via Reasoning on Knowledge Graphs. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16190–16206. Toronto, Canada: Association for Computational Linguistics.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Le, H. T.; To, L. T.; Nguyen, M. T.; and Van Nguyen, K. 2024. ViWikiFC: Fact-Checking for Vietnamese Wikipedia-Based Textual Knowledge Source. *arXiv preprint arXiv:2405.07615*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7342–7351. Online: Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. Florence, Italy: Association for Computational Linguistics.

McHugh, M. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3): 276–282.

Nguyen, D. Q.; and Tuan Nguyen, A. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1037–1042. Association for Computational Linguistics.

Nie, Y.; Chen, H.; and Bansal, M. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6859–6866.

Nørregaard, J.; and Derczynski, L. 2021. DanFEVER: claim verification dataset for Danish. In Dobnik, S.; and Øvrelid, L., eds., *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 422–428. Reykjavik, Iceland: Linköping University Electronic Press, Sweden.

Pan, L.; Wu, X.; Lu, X.; Luu, A. T.; Wang, W. Y.; Kan, M.-Y.; and Nakov, P. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6981–7004. Toronto, Canada: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Schuster, T.; Fisch, A.; and Barzilay, R. 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*, 624–643. Online: Association for Computational Linguistics.

Soleimani, A.; Monz, C.; and Worring, M. 2020. BERT for Evidence Retrieval and Claim Verification. In Jose, J. M.; Yilmaz, E.; Magalhães, J.; Castells, P.; Ferro, N.; Silva, M. J.; and Martins, F., eds., *Advances in Information Retrieval*, 359–366. Cham: Springer International Publishing. ISBN 978-3-030-45442-5.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. New Orleans, Louisiana: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vlachos, A.; and Riedel, S. 2014. Fact Checking: Task definition and dataset construction. In Danescu-Niculescu-Mizil, C.; Eisenstein, J.; McKeown, K.; and Smith, N. A., eds., *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22. Baltimore, MD, USA: Association for Computational Linguistics.

Vu, T.; Nguyen, D. Q.; Nguyen, D. Q.; Dras, M.; and Johnson, M. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In Liu, Y.; Paek, T.; and Patwardhan, M., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 56–60. New Orleans, Louisiana: Association for Computational Linguistics.

Wang, W. Y. 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)*, 422–426. Vancouver, Canada: Association for Computational Linguistics.

Wang, Y.; Li, Y.; Wang, Y.; Mi, F.; Zhou, P.; Wang, X.; Liu, J.; Jiang, X.; and Liu, Q. 2022. Pan More Gold from the Sand: Refining Open-domain Dialogue Training with Noisy Self-Retrieval Generation. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 636–647. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Wu, S.; Xin, C.; Lin, H.; Han, X.; Liu, C.; Chen, J.; Yang, F.; Wan, G.; and Sun, L. 2023. Ambiguous Learning from Retrieval: Towards Zero-shot Semantic Parsing. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14081–14094. Toronto, Canada: Association for Computational Linguistics.

Zhong, W.; Xu, J.; Tang, D.; Xu, Z.; Duan, N.; Zhou, M.; Wang, J.; and Yin, J. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6170–6180. Association for Computational Linguistics.