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

Anonymous ACL submission

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[1].

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 in the study 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.

Claim:
Các công dân trẻ tiêu biểu cũng tham gia vào giải chạy bộ "Bước chân xanh" nhằm hưởng ứng chiến dịch Giờ Trái đất năm 2023.
English: Exemplary young citizens also participate in the "Green Steps" running event to support the Earth Hour campaign in 2023.

Support ✔

Context:
TPO-Sáng 25/3, Thành Đoàn, Hội LHTN Việt Nam TPHCM , Hội Sinh viên Việt Nam TPHCM tổ chức Giải chạy bộ "Bước chân xanh" lần thứ 2. Giải chạy thu hút hơn 1.000 người tham gia hưởng ứng chiến dịch Giờ Trái đất năm 2023. Bên cạnh đông đảo đoàn viên, thanh niên, sinh viên, giải chạy bộ "Bước chân xanh" còn thu hút các gương công dân trẻ tiêu biểu TPHCM, các hoa hậu, á hậu, văn nghệ sĩ trẻ... cùng tham gia.
English: TPO-March 25th, the HCM Youth Union and the Vietnam National Union of Students in HCM City organized the 2nd "Green Steps" running event. The race attracted over 1,000 participants in response to the Earth Hour campaign in 2023. In addition to a large number of union members, youth, and students, the "Green Steps" running event also attracted notable young citizens of HCM City, beauty queens, runners-up, young artists, and others to participate.

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"

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;

[1]The supplementary material will be made publicly available upon acceptance

Schuster et al., 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 dataset tailored to Vietnamese news, which spans multiple domains. Our main contributions are described as follows:

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 on several state-of-the-art language models using the ViFactCheck dataset to evaluate their effectiveness in verifying information within the Vietnamese context. Our study encompasses the use of both pre-trained and large language models, specifically adapted to this linguistic framework, to explore 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 dataset (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 et al., 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, excel at understanding and analyzing the context within texts, making them exceptionally effective for tasks such as evidence retrieval and claim verification (Nie et al., 2019; Soleimani et al., 2020; Liu et al., 2020; Zhong 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 et al., 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 (Popat et al., 2018; Zhong et al., 2020; Baek et al., 2023), leading to more robust and accurate fact-checking solutions.

	Dataset	Labels	# Claims	Real-World	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
	ViFactCheck (Ours)	3	7,232	✔	Vietnamese	News	Multi

Table 1: Comparative overview of typical open-domain fact-checking datasets. Real-World indicates datasets comprising claims generated by humans about events that have actually occurred. The type of Reasoning Steps (#RS) column reflects the complexity involved in verifying the claims in each dataset.

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. A notable study by Duong et al. (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

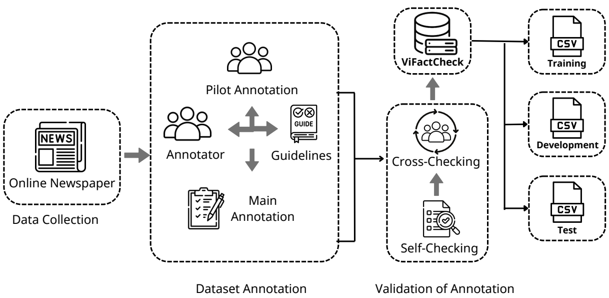


Figure 2: The ViFactCheck dataset contruction process.

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.

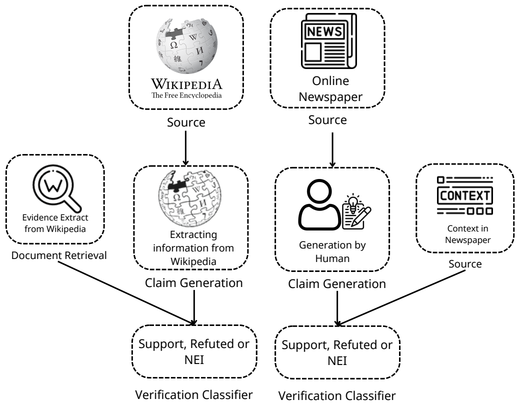
3.1 Data Collection

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, 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, article descriptions 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.



(a) Thorne et al. (2018). (b) Our proposed process.

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

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.

3.2.1 Pilot Annotation

The first stage of dataset annotation is the pilot annotation, which is used to familiarize the annotators with the claim generation and verification classifier process described above. 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 specific annotation 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.

3.2.2 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 research.

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 E). 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 in fact-checking research.

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 analysis

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 et al. (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 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).

For LLMs, we utilized the Unsloth framework with supervised fine-tuning using LoRA adaptation. The hyper-parameters 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 and parameters can be found in Appendices H and I.

4.2 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	Context	Evidence	Δ
Fine-tuning PLMs			
PhoBERT$_{base}$	68.55	77.76	↑9.21
PhoBERT$_{large}$	62.93	79.76	↑<u>16.83</u>
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$_{large}$	<u>75.42</u>	<u>88.02</u>	↑12.60
Fine-tuning LLMs			
Gemma	<u>**85.94**</u>	<u>**89.90**</u>	↑3.96
Mistral	70.13	88.63	↑18.50
Llama2	41.47	79.53	↑<u>**38.06**</u>
Llama3	79.65	88.67	↑9.02
Prompting LLMs			
Gemini	<u>76.26</u>	<u>74.88</u>	↓1.38
Gemma	45.05	39.47	↓5.58
Mistral	61.02	57.31	↓3.71
Llama2	63.54	51.64	↓<u>11.90</u>
Llama3	56.66	50.81	↓5.85

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, 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 Con-

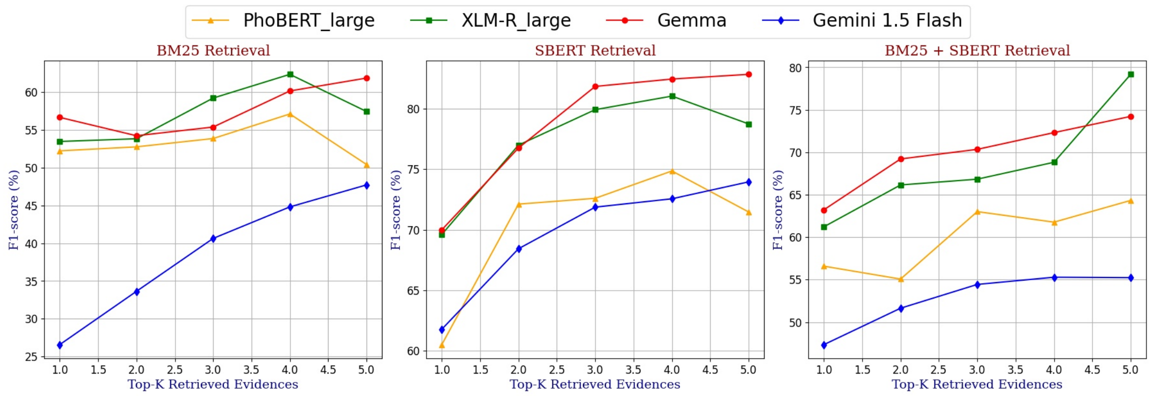


Figure 4: Comparative performance of various text retrieval models across different Top-K settings.

text 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.

4.3 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 (Robert-

	Overall				Single-evidence				Multiple-evidence			
	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.

son et al., 2009), and their hybrid configurations under various conditions, with a focus on how well these models understand the semantic complexities of 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 4.2.

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.

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 types 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.

4.4 Human Performance

Table 5 presents an evaluation of fine-tuned models in fact-checking tasks, 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$_{base}$	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.

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.

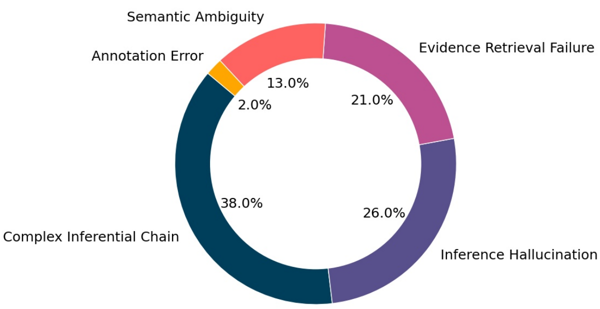


Figure 5: Distributions of errors.

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.

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 overall reasoning and inference capabilities (Kim et al., 2023; Pan et al., 2023).

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).

Limitations and Ethics consideration

The ViFactCheck dataset and methods present a significant advancement in Vietnamese fact-checking; however, certain limitations must be acknowledged. One notable limitation pertains to potential bias introduced during data labeling by human annotators. These biases, whether conscious or unconscious, can impact the fairness and generalizability of fact-checking models trained on the dataset. Addressing this limitation requires the implementation of transparent guidelines and rigorous quality control measures to minimize bias and ensure consistency in the annotations.

During the construction of the ViFactCheck dataset, we prioritized ethical principles to protect the privacy of individuals. Informed consent was obtained from the data contributors and data privacy regulations were strictly adhered to. We established clear annotation guidelines and performed regular quality control checks to minimize potential biases. The dataset was anonymized to protect the confidentiality of the sources and individuals mentioned in the claims. We commit to using the ViFactCheck dataset solely for research purposes, ensuring its reliability and ethical integrity.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Jinheon Baek, Alham Fikri Aji, Jens Lehmann, and Sung Ju Hwang. 2023. Direct fact retrieval from knowledge graphs without entity linking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10038–10055, Toronto, Canada. Association for Computational Linguistics.

The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 2020. Improving sequence tagging for Vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 13–20, Hanoi, Vietnam. Association for Computational Linguistics.

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.

Huong T Duong, Van H Ho, and Phuc Do. 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.

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

Ashim Gupta and Vivek Srikumar. 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)*, pages 675–682, Online. Association for Computational Linguistics.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 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*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023. Zero-shot faithful factual error correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5660–5676, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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. *Preprint*, arXiv:1907.11692.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6859–6866.

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, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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.

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.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

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.

Yihe Wang, Yitong Li, Yasheng Wang, Fei Mi, Pingyi Zhou, Xin Wang, Jin Liu, Xin Jiang, and Qun Liu. 2022. Pan more gold from the sand: Refining open-domain dialogue training with noisy self-retrieval generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 636–647, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shan Wu, Chunlei Xin, Hongyu Lin, Xianpei Han, Cao Liu, Jiansong Chen, Fan Yang, Guanglu Wan, and Le Sun. 2023. Ambiguous learning from retrieval: Towards zero-shot semantic parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14081–14094, Toronto, Canada. Association for Computational Linguistics.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

A Task Definition

Fact-checking can be defined as the task of automatically assessing the veracity of a given claim based on available evidence. This task typically involves two main steps: **1) Evidence Retrieval (Fact Extraction):** This step aims to find relevant evidence from a given corpus to support or refute the claim. **2) Claim Verification (Fact Verification):** This step determines the truthfulness of the claim based on the retrieved evidence.

The proposed system is designed to assign labels to the claims, categorizing them as follows:

- **Support:** The claim is confirmed to be correct according to the available evidence.
- **Refute:** The claim is determined to be inaccurate compared to the available evidence.
- **Not Enough Info (NEI):** The claim is not sufficiently supported by the evidence within the corresponding news article, making it impossible to definitively verify or refute.

The goal of fact-checking systems is to assist human fact-checkers in their efforts to combat the spread of disinformation and false news. These systems provide automated tools that assess the credibility of claims in various sources, including news articles, social media, and political speeches.

B Data Collection Source

Website	Organization	URL
Bao Chinh Phu	Government of Vietnam	https://baochinhphu.vn
VnExpress	MOST Vietnam	https://vnexpress.net
Dan Tri	MOLISA Vietnam	https://dantri.com.vn
Nguoi Lao Dong	HCM City Committee	https://nld.com.vn
Tuoi Tre	HCM Communist Youth Union	https://tuoitre.vn
Tin Tuc	Vietnam News Agency	https://baotintuc.vn
Phap Luat HCM	HCM City People's Committee	https://plo.vn
Thanh Nien	Vietnam Youth Union	https://thanhnien.vn

Table 6: Details of the sources and organizations of the online news sites in the ViFactCheck dataset.

C Human Recruitment

Annotation Recruitment In our study, we recruited seven university students as annotators, all native Vietnamese speakers aged 20 to 22, representing a diverse range of academic disciplines. These disciplines included social sciences, natural sciences, and Vietnamese studies. Their selection was based on exceptional linguistic skills, demonstrated by high scores in Vietnamese literature exams, and a deep familiarity with various media platforms, ensuring annotations that accurately reflect current linguistic trends.

To ensure the reliability and accuracy of our annotation process, we engaged two linguistic experts with a solid background in Vietnamese grammar and syntax. These experts, also native speakers, were tasked with developing the guidelines and continuously monitoring the annotation process. Their expertise is grounded in extensive academic achievements and a critical ability to evaluate news content across various media platforms, adding a significant layer of scientific rigor and depth to our data creation methodology.

Human Evaluation Recruitment To evaluate human performance in the fact-checking process, we engaged three native Vietnamese-speaking students who had no prior exposure to the task of fact-checking. They were tasked with annotating a representative subset consisting of 200 samples. Comprehensive instructions were provided to ensure their understanding of the task, including clarifications on the significance of each label and additional information to assist them in determining the appropriate labels for each sample. The final label for each claim was determined through a majority consensus among the assessors.

D Data Annotation Tool and Guideline

Data annotation tool During the annotation phases of the dataset, we utilized Label Studio, an open-source platform that provides an intuitive interface and supports various labeling tasks across different types of data. Our annotation interface is shown in Figure 6.

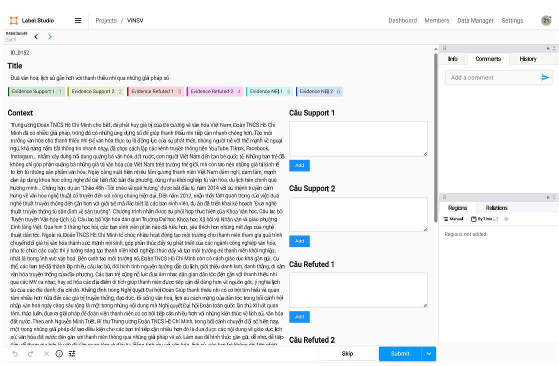


Figure 6: Label Studio UI for our annotation task.

Data annotation guideline Comprehensive guidelines[2] were provided to the annotators to ensure a cohesive and systematic approach:

(1) The annotation process required the generation of six claim pairs for each article in the dataset, resulting in two pairs for each designated label: **Support, Refute, and NEI (Not Enough Information). (2)**. For the **Support** and **Refute** labels, annotations were grounded in the intrinsic information and contextual evidence derived directly from the corresponding news articles. The **NEI** label required a more nuanced approach, involving the addition of external information and context, which could either align with or deviate from the truth. **(3)** The generated claims must adhere to certain rules: paraphrasing sentences from the article, inferring claims by combining multiple pieces of information, and meticulously avoiding spelling and abbreviation errors that could compromise the quality of the dataset. **(4)** To enrich the dataset with diverse perspectives and challenges, annotators were encouraged to leverage their broad vocabulary and skilled sentence-writing techniques, thereby introducing valuable nuances into the annotations.

E Data Examples

The ViFactCheck dataset includes various examples of written claims, as illustrated in Table 7. To create a challenging and realistic context, annotators were tasked with generating claims based on multiple pieces of evidence, which are highlighted within the textual context provided. This methodological approach not only enhances the complexity and challenge of the annotation task but also contributes significantly to the reliability and practical value of the dataset for fact-checking tasks in the Vietnamese language. By ensuring that claims are grounded in verifiable pieces of evidence, the dataset fosters a robust environment for training and evaluating language models specifically tailored to the nuances of fact verification.

F Human-Generated Rules

In the ViFactCheck datasets, annotators were encouraged to leverage their broad vocabulary and skilled sentence-writing techniques, thus introducing valuable nuances into the annotations. The basic rules for the use of the generation by annotators are summarized in Table 8.

[2] The detailed annotation guideline will be provided upon acceptance.

Context	TPO - Tổng Công ty Cảng Hàng không Việt Nam (ACV) vừa chính thức gia hạn thời gian mời thầu thêm 1 tháng, kéo dài thời gian thực hiện gói thầu thi công nhà ga sân bay Long Thành từ 33 tháng lên 39 tháng. Như vậy, "siêu sân bay" Long Thành sẽ chỉ có thể đưa vào khai thác từ năm 2026 thay vì mục tiêu năm 2025 như trước đó. Tin từ ACV cho hay, đơn vị chính thức điều chỉnh kế hoạch và hồ sơ mời thầu gói thầu thi công xây dựng và lắp đặt thiết bị nhà ga hành khách sân bay Long Thành giai đoạn 1 (do ACV làm chủ đầu tư). Cụ thể, thời gian mời thầu được gia hạn thêm 1 tháng, kéo dài tới sáng ngày 28/4, thay vì tới ngày 28/3 như trước đó. ... Gói thầu thi công nhà ga hành khách sân bay Long Thành trị giá hơn 35 nghìn tỷ đồng do ACV làm chủ đầu tư. Đây là gói thầu lớn nhất dự án sân bay Long Thành... (**English:** TPO - Vietnam Airport Corporation (ACV) has officially extended the bidding period by an additional month, prolonging the implementation time for the construction contract of the Long Thanh Airport passenger terminal from 33 to 39 months. Consequently, the "mega airport" Long Thanh will only be operational by 2026 instead of the previous target of 2025. According to ACV, the organization has formally adjusted the plan and tender documents for the construction and installation of the passenger terminal at Long Thanh Airport Phase 1 (with ACV as the main investor). Specifically, the bidding period has been extended by one month, now ending on the morning of April 28, instead of the previous deadline of March 28. ... The construction contract for Long Thanh Airport's passenger terminal, valued at over 35 trillion VND is being managed by ACV. This is the largest contract within the Long Thanh Airport project.)
Support	Việc nhà thầu thi công xây dựng và lắp đặt thiết bị nhà ga hành khách sân bay Long Thành giai đoạn 1 bị điều chỉnh, thời gian bị kéo dài tới sáng ngày 28/4 thay vì tới ngày 28/3 như dự kiến. **English:** *The construction and installation contract for the Long Thanh Airport Phase 1 passenger terminal has been adjusted, with the timeline extended to the morning of April 28 instead of the originally anticipated March 28.*
Refute	Tổng Công ty Cảng Hàng không Việt Nam (ACV) vừa gia hạn thời gian mời thầu thêm thời gian 2 tháng, tức "siêu sân bay" Long Thành sẽ chỉ có thể đưa vào sử dụng từ năm 2026 thay vì năm 2025 như dự kiến ban đầu. **English:** *Vietnam Airport Corporation (ACV) has recently extended the bidding period by an additional 2 months, meaning that the "mega airport" Long Thanh will only be operational by 2026 instead of the originally planned year 2025.*
NEI	Gói thầu lớn nhất dự án sân bay Long Thành là gói thầu thi công nhà ga hành khách với trị giá hơn 35 nghìn tỷ đồng, được tài trợ bởi công ty Hàn Quốc. **English:** *The largest contract within the Long Thanh Airport project is the construction of the passenger terminal, valued at over 35 trillion VND, and it is sponsored by a South Korean company.*

Table 7: Typical samples from the **ViFactCheck** dataset with three labels **Support**, **Refute**, and **NEI**. The highlighted words is the evidence of the claim.

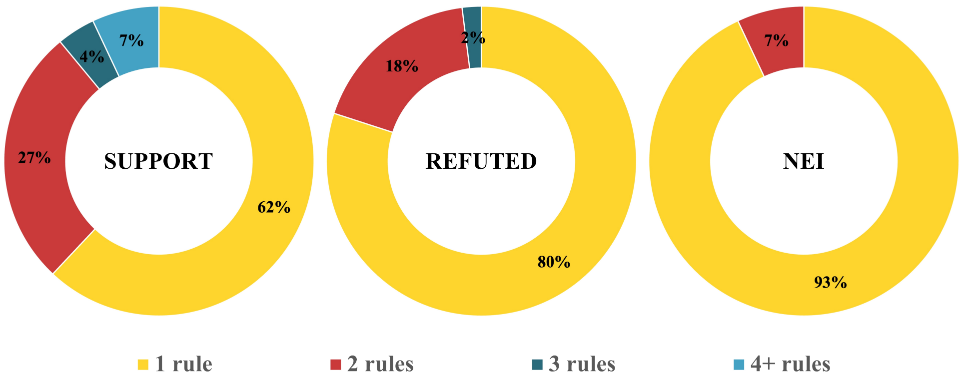


Figure 7: The ratio of combining different rules to create claims in ViFactCheck.

	Rules	Ratio (%)
Support	Restructuring the evidences	73.68
	Eliminating or adding words	44.21
	Substituting numbers, time, or mathematical inferences	7.34
	Altering the word order in a sentence	8.42
Refute	Employing Negation	8.16
	Replacing Words with Antonyms	17.35
	Misrepresenting quantity	22.45
	Misrepresenting Temporal Logic	16.37
	Misinterpreting Entity Relationships	5.11
	Misjudging Event Dynamics	47.96
NEI	Inferring sentences with unspecified information	90.20
	Utilizing external knowledge	10.78

Table 8: Approaches and rules for generating claims by humans in the ViFactCheck dataset. Note that a claim could involve multiple rules

Annotators are required to follow guidelines to create diverse and challenging data. The distribution of data-generating rule usage for claims related to Support, Refute, and Not Enough Information (NEI) is shown in Figure 7. To understand how annotators behave in creating ViFactCheck, we analyzed the number of rules used to generate claims. We randomly selected 100 context-claim pairs for Support, Refute, and NEI categories.

The primary trend in this dataset reveals an obvious bias towards using 1-2 rules, reflecting a standardized annotation process. However, some annotators deviated from this trend, opting for four or more rules, demonstrating an awareness of the complexity and diversity of data. This underscores the importance of judiciously combining rules for reliable and accurate annotation.

The use of multiple rules presents challenges for language model development, introducing complexity into inference and decision-making processes dependent on rule combinations. However, it also offers an opportunity to improve more adaptable language models, ensuring greater accuracy in making inferences.

G Additional Dataset Analysis

Dataset basic statistic. The ViFactCheck dataset contains 7,232 samples divided into three subsets: training, development, and test with a ratio of 7:1:2. The basic statistics of the three subsets are shown in Table 9. We observed that the average length of a context in the dataset is approximately 700 words, with the longest context extending to 3,602 words. Such richness in context proves highly beneficial for models with large parameter sets, such as Gemma, as they can effectively capture the maximum features of the data. On average, each claim sentence contains about 36 words, with the longest reaching 165 words.

		Training	Development	Test
Context	Total samples	1035	496	758
	Avg length	693.2	670.2	690.5
	Max length	3602	2534	3602
	Min length	71	71	71
	Total vocab size	25,382	16,522	21,263
claim	Total samples	5062	723	1447
	Avg length	35.9	35.6	35.8
	Max length	165	145	135
	Min length	7	10	7
	Total vocab size	12,189	4,555	6,711

Table 9: Basic statistic of ViFactCheck dataset. The size and length of the vocab are computed at word level.

Topic Distribution Analysis The ViFactCheck dataset covers 12 popular topics frequently found in Vietnamese newspapers, which are often subjected to misinformation. These topics, summarized in Figure 8, include “Headlines”, “World”, “Education”, and “Economics”, among others. “Headlines”, covering updates on social issues and events, appears most frequently, demonstrating a significant presence in the dataset. The other notable topics, “World", “Education”, and “Economics”, contribute 12.4%, 12.9%, and 10.9% to the dataset, respectively. In contrast, “National Security” accounts for the lowest percentage at 2.0%. This lower representation is attributed to the relatively few articles on this topic in real life. Despite its smaller volume, due to the critical need for accuracy in information pertaining to national security, a concerted effort was made to include articles related to this topic.

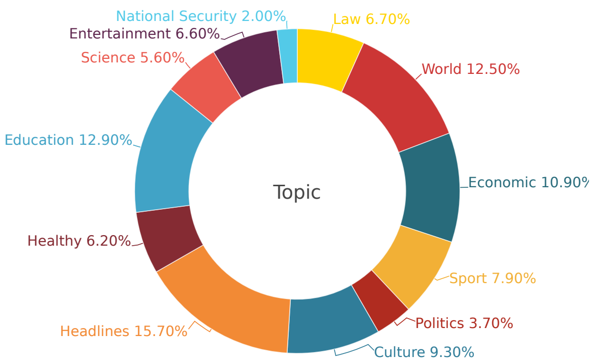


Figure 8: Topic distribution on ViFactCheck dataset.

Evidence Distribution Analysis Figure 9 from the ViFactCheck dataset shows the distribution of samples with varying numbers of evidence per claim. Single evidence refers to using only one piece of information to verify a claim, while multi-evidence involves integrating findings from multiple sources, necessitating advanced analytical skills to synthesize and validate information.

The distribution reveals a predominant reliance on single pieces of evidence, where claims are supported by one source, reflecting simpler verification tasks. Multi-evidence scenarios, where claims are substantiated by two or more sources, demonstrate a steep decline to 1,765 and 293 samples for two and three evidences, respectively. This indicates the increasing complexity and computational demand of integrating diverse evidences. Notably, the rise to 130 samples for claims with more than five evidences suggests some scenarios necessitate extensive, complex reasoning, highlighting the capability of the dataset to train models for robust, multifaceted fact verification.

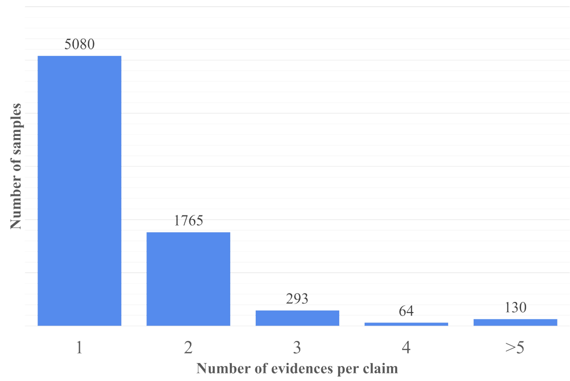


Figure 9: The distribution of single and multiple evidences samples in the ViFactCheck dataset.

H Details about the Baselines

H.1 Pre-trained Language Models

Based on the significant performance of transformer-based in prior fact-checking tasks (Thorne et al., 2018; Hu et al., 2022; Nørregaard and Derczynski, 2021), we employ pre-trained language models, specifically BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) architectures, for the fact-checking task. This study includes four models, comprising two multilingual models and two monolingual models (the details of each model are shown in Table 10).

mBERT (Devlin et al., 2019) is a transformer-based model trained on an extensive corpus of 104 languages, including Vietnamese. Its linguistic versatility makes mBERT invaluable for fact-checking tasks. As a multilingual model, mBERT enables comprehensive analysis and serves as an excellent tool for ensuring the credibility of data within the Vietnamese fact-checking framework.

Cross-lingual Language Model - RoBERTa (XLM-R) (Conneau et al., 2020) is a transformer-based model trained on 100 languages. This vast linguistic scope means XLM-R can understand and compare information across different languages, an advantage for fact-checking that offers a broader context beyond the Vietnamese language. The ability of XLM-R to process information from multilingual sources or across language barriers is especially valuable when dealing with content that transcends linguistic boundaries.

PhoBERT (Nguyen and Tuan Nguyen, 2020), leveraging the powerful Transformer architecture of RoBERTa (Liu et al., 2019), exhibits a profound understanding of the nuances and context of the Vietnamese language. This linguistic precision is highly beneficial for the Vietnamese fact-checking dataset, as it can discern subtle language nuances that general models might overlook. With its focus on Vietnamese, PhoBERT delivers exceptional efficiency and accuracy when applied to a corpus of the same language, facilitating high-quality fact-checking within the Vietnamese context.

ViBERT, based on the BERT architecture and specifically designed for Vietnamese, was introduced by Bui et al. (2020). Unlike mBERT, which is trained on a multi-language corpus, ViBERT is pre-trained on a substantial corpus of 10GB of uncompressed Vietnamese text, focusing solely on Vietnamese to achieve optimal performance.

By investigating the effectiveness of these BERT variants in Vietnamese fact-checking, we aim to enhance the field's ability to combat disinformation. The diversity of these models in terms of monolingual understanding, linguistic precision, and cross-lingual capabilities promises to make a significant contribution to the fact-checking landscape, advancing a more credible and precise information ecosystem.

H.2 Fine-Tuning 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 low-resource configurations. Specifically, we utilize the open-source models Llama2 7B, Llama3 8B, Gemma 7B, and Mistral 7B.

LLaMA (Touvron et al., 2023), or Large Language Model Meta AI, represents a significant leap in the development of foundational models for natural language inference (NLI) tasks. Introduced by Meta AI, LLaMA is designed to create a more accessible and efficient framework for researchers and developers. Available in various sizes including 7B, 13B, 33B, 65B, and 70B parameters, LLaMA caters to different computational needs and research objectives. It is trained on a diverse dataset comprising 1.4 trillion tokens from 20 languages, enabling it to perform a wide range of NLP tasks with high accuracy and efficiency. Innovations in its architecture, such as the SwiGLU activation function and rotary positional embeddings (Shazeer, 2020), contribute to its superior performance on NLP benchmarks.

Mistral (Jiang et al., 2023), developed by Mistral AI, stands out for its innovative approach to structured content generation and instruction-based modeling. Designed to generate high-quality, structured content similar to the functionalities offered by OpenAI models, Mistral achieves enhanced efficiency and lower resource requirements. The Mistral-7B model utilizes mechanisms like Grouped-query Attention (GQA) and Sliding Window Attention (SWA) to achieve faster inference times and handle longer text sequences. Its ability to parse and extract information using a JSON Schema makes it particularly suited for tasks requiring structured output.

Gemma (Team et al., 2024), developed by Google DeepMind, leverages technology from the Gemini model to offer state-of-the-art, open models. It includes a 7B parameter model for GPU/TPU use and a 2B parameter model for CPU and on-device applications, both trained on up to 6 billion tokens. These models excel in language understanding, reasoning, and safety benchmarks, outperforming similarly sized open models in 11 out of 18 tasks. Key enhancements in the Gemma models include multi-query attention, RoPE embeddings, GeGLU activations, and RMSNorm for stable training. The models are rigorously evaluated through automated and human benchmarks to ensure robustness and reliability, with a strong emphasis on responsible AI practices.

The LLMs were fine-tuned using the LoRA through the Unsloth library. Detailed configuration specifics and prompting procedures are described in Section 4.1 and Appendices J, respectively.

H.3 In-Context Learning Models

In addition to fine-tuning large language models (LLMs), we rigorously assessed the in-context learning capabilities of these models through zero-shot evaluations, where models are tasked with generating accurate responses without prior specific training on examples.

Beyond the models detailed in Appendix H.2, we employed Gemini 1.5 Flash (Reid et al., 2024), a recent addition to the LLMs developed by Google AI. Introduced in May 2024, Gemini 1.5 Flash, part of the broader Gemini family, excels in handling multimodal tasks. Notable for its high-speed, large-

Model	#Layer	#Head	#Params	#Vocab	#MSL	Domain data	Language support
PhoBERT$_{base}$ (Nguyen and Tuan Nguyen, 2020)	12	12	135M	64K	256	ViWiki + ViNews	Vietnamese
PhoBERT$_{large}$ (Nguyen and Tuan Nguyen, 2020)	24	16	370M	64K	256	ViWiki + ViNews	Vietnamese
ViBERT (Bui et al., 2020)	12	12	-	30K	256	Vietnamese News	Vietnamese
mBERT (Devlin et al., 2019)	12	12	110M	30K	512	Wikipedia + BookCorpus	Multilingual
XLM-R$_{base}$ (Conneau et al., 2020)	12	12	270M	250K	512	CommonCrawl	100+ languages
XLM-R$_{large}$ (Conneau et al., 2020)	24	16	550M	250K	512	CommonCrawl	100+ languages
Gemini (Reid et al., 2024)	-	-	-	-	1,048,576	Mixed large datasets	Multilingual
Gemma (Team et al., 2024)	28	16	7B	256K	8,192	Mixed large datasets	Multilingual
Mistral (Jiang et al., 2023)	32	32	7B	32K	32,768	Mixed large datasets	Multilingual
Llama2 (Touvron et al., 2023)	32	32	7B	32K	4,096	Mixture of datasets	Primarily English
Llama3 (Touvron et al., 2023)	32	32	8B	128K	8,192	Mixture of datasets	Primarily English

Table 10: Detailed specifications of our baseline models. Abbreviations used are: #Layers (Number of Hidden Layers), #Heads (Number of Attention Heads), #Params (Total Number of Parameters), #Vocab (Vocabulary Size), and #MSL (Maximum Sequence Length).

scale information processing capabilities, Gemini 1.5 Flash is particularly suitable for real-time applications and environments requiring frequent updates. Despite its focus on efficiency, this model maintains robust reasoning capabilities across multiple modalities, including text, image, and audio, and supports an extensive context window of up to one million tokens. This feature is crucial for tasks that require a deep comprehension of prior information.

Furthermore, we conducted experiments on the ViFactCheck dataset using the prompt described in Appendix J, following a zero-shot approach. These experiments aimed to evaluate the ability of models to integrate and reason with various types of information without preliminary fine-tuning, showcasing its potential in real-world applications where training data may be sparse or unavailable.

I Number of Parameters

To establish the main baseline models, we utilized several state-of-the-art methods, including a pretrained and large language model, to support the Vietnamese Fact-Checking task. The details of each model are shown in Table 10.

J Prompts for Vietnames Fact-Checking

In this section, we outline the templates for the prompting methods used for fine-tuning and zero-shot evaluations with LLMs in the fact-checking task. The prompt structure is designed to test the ability of models to assess the veracity of a claim based on the given context or evidence.

Fine-Tune Instruction Prompting

You will be presented with a long context, followed by a claim. Your task is to fact-check the claim based on the provided context. You must categorize the claim into one of three categories:
- **Support:** Choose this if the claim is true and fully supported by the context.
- **Refute:** Choose this if the claim is false and contradicted by the context.
- **Not Enough Information:** Choose this if the claim contains content that is not covered by the context, making it impossible to determine its accuracy.
Context:
Claim:
Response:

Zeroshot Prompting

Return only the label in the format: Label: Support(0), Label: Refute(1), or Label: Not Enough Information(2).
Instructions:
1. Fact check the claim based on the provided evidence.
2. Use the following labels:
- **Support:** The claim is true and supported by the evidence.
- **Refute:** The claim is false and contradicted by the evidence.
- **Not Enough Information:** The claim contains content that is not covered by the evidence, making it impossible to determine its accuracy.
Example:
Label: Support

K Definition and Examples of Error Analysis

We introduce the error definition as follows and illustrate some error cases for Vietnamese fact-checking tasks in Figure 5:

- **Semantic Ambiguity:** Issues arising from context with ambiguous, verbose, or complex data, leading to interpretive difficulties (as shown in Figure 10).

- **Evidence Retrieval Failure:** Failures due to the inability of model to accurately and fully extract essential evidence from data sources (as shown in Figure 11).
- **Inference Hallucination:** Incorrect classifications produced by the model, despite the correct extraction and availability of relevant evidence (as shown in Figure 12).
- **Complex Inferential Chain:** Errors resulting from the necessity to synthesize insights across multiple sources or evidences through sequential reasoning (as shown in Figure 13).
- **Labeling Error:** Issues stemming from inaccuracies or inconsistencies introduced during the manual data labeling process.

Semantic Ambiguity

Context: ...Bên cạnh đó, Sở cũng yêu cầu các cơ sở giáo dục phối hợp với lực lượng chức năng tại địa phương đảm bảo trật tự, an toàn cho học sinh, sinh viên tại khu vực cổng trường.
(In addition, the Department also requested educational institutions to collaborate with local authorities to ensure order and safety for students in the school gate area.)

Claim: Sở cũng yêu cầu các cơ sở giáo dục phối hợp với lực lượng chức năng tại địa phương đảm bảo trật tự, an toàn cho học sinh, sinh viên tại những khu vực đông đúc gần trường.
(The department also requested educational institutions to collaborate with local authorities to ensure order and safety for students in crowded areas near schools.)

Gold labels: NEI **Gemma Prediction:** Refute

Figure 10: Examples of Semantic Ambiguity.

Evidence Retrieval Failure

Context: … Cuộc đua xe đạp Cúp Truyền hình TP.HCM 2023 với slogan "Non sông liền một dải - Niềm tin chiến thắng" quy tụ tất cả đội đua mạnh trên cả nước vốn quen thuộc với làng xe đạp chuyên nghiệp gồm: TP.HCM - Vinama, Tập đoàn Lộc Trời, Dược Domesco Đồng Tháp, Quân khu 7, Hà Nội, Kenda Đồng Nai, Bình Dương…
(…The 2023 Ho Chi Minh City Television Cup cycling race with the slogan "United Nation, Victory Belief" brings together all the strong teams across the country who are familiar with the professional cycling scene, including: Ho Chi Minh City - Vinama, Loc Troi Group, Domesco Dong Thap Pharmaceutical, Military Zone 7, Hanoi, Kenda Dong Nai, Binh Duong...)

Claim: Cuộc đua xe đạp quy tụ 100 đội đua xe chuyên nghiệp trong nước tham gia.
(The cycling race brought together 100 professional cycling teams from across the country.)

Retrieved Evidence: Chặng đua đồng đội tính giờ sẽ diễn ra tại Quảng Ngãi. Nét đặc biệt của giải đua năm nay là ngoài lộ trình lên Tây Nguyên và vòng xuống các tỉnh miền Đông Nam Bộ còn là chuyển tải nhiều thông điệp như quảng bá du lịch, giới thiệu văn hóa vùng miền, bảo vệ môi trường, góp phần cổ vũ, động viên phong trào đạp xe đạp trong cả nước và các hoạt động thiện nguyện, lan tỏa công tác đền ơn đáp nghĩa.
(The team time trial stage will take place in Quang Ngai. A special feature of this year's race is that in addition to the route up to the Central Highlands and down to the provinces of the Southeast, it also conveys many messages such as promoting tourism, introducing regional culture, protecting the environment, contributing to encouraging and motivating the cycling movement nationwide, and charitable activities, spreading gratitude and repayment.)

Gold labels: NEI **Gemma Prediction:** Refute

Figure 11: Examples of Evidence Retrieval Failure.

Inference Hallucination

Context: (Chinhphu.vn) – Việt Nam và Trung Quốc sẽ tiếp tục hợp tác chặt chẽ cùng nhau thúc đẩy du lịch hai nước phục hồi và phát triển lành mạnh….
((Chinhphu.vn) - Vietnam and China will continue to work closely together to promote the recovery and healthy development of tourism in both countries.)

Claim: Việt Nam và Trung Quốc sẽ tiếp tục hợp tác chặt chẽ trong lĩnh vực nông nghiệp.
(Vietnam and China will continue to strengthen cooperation in agriculture.)

Retrieved Evidence: (Chinhphu.vn) – Việt Nam và Trung Quốc sẽ tiếp tục hợp tác chặt chẽ cùng nhau thúc đẩy du lịch hai nước phục hồi và phát triển lành mạnh….
((Chinhphu.vn) - Vietnam and China will continue to work closely together to promote the recovery and healthy development of tourism in both countries.)

Gold labels: Refute **Gemma Prediction:** NEI

Figure 12: Examples of Inference Hallucination.

Complex Inferential Chain

Context: ...Rạng sáng 6/2, trận động đất độ lớn 7,8 có tâm chấn tại Thổ Nhĩ Kỳ đã gây thiệt hạilớn tại nước này và nước láng giềng Syria. Tính đến 16h ngày 12/2 (giờ Việt Nam), trận động đất này đã cướp đi sinh mạng của hơn 29.000 người tại cả hai nước, trong đó có 24.617 người tại Thổ Nhĩ Kỳ và hơn 4.500 người tại Syria, trong khi có hàng chục nghìn người bị thương.
(Early morning on February 6th, a 7.8 magnitude earthquake centered in Turkey struck, causing widespread devastation in both Turkey and neighboring Syria. As of 4 PM on February 12th (Vietnam time), the earthquake has claimed the lives of over 29,000 people in both countries, including 24,617 in Turkey and over 4,500 in Syria, while tens of thousands have been injured)

Claim: Trận động đất độ lớn 7,8 có tâm chấn tại Thổ Nhĩ Kỳ đã gây thiệt hại lớn tại nước này và hai nước láng giềng là Bulgaria và Syria, có hàng chục nghìn người bị thương, cướp đi sinh mạng của hơn 29.000 người.
(The 7.8 magnitude earthquake centered in Turkey has caused widespread devastation in the country and its neighboring nations, Bulgaria and Syria. The disaster has left tens of thousands injured and tragically claimed the lives of over 29,000 people.)

Gold labels: NEI **Gemma Prediction:** Refute

Figure 13: Examples of Complex Inferential Chain.

L Additional Qualitative Analysis

To obtain insights into the performance of language models, we conducted an in-depth analysis considering various factors such as the length of the context, the topic of discussion, the volume of training data, and the duration of model training.

Effects of Context Length We initiated our investigation by analyzing the test results with respect to the length of the context (see Figure 14). Notably, PhoBERT$_{large}$ and XLM-R$_{large}$ perform well when analyzing shorter texts (0-100 words). However, their performance declines as text length increases, particularly in the 400-500 and 500-600 word ranges, suggesting that longer texts may pose challenges for these models. In contrast, Gemma and Gemini exhibit more consistent performance across different text lengths, showing only minor fluctuations. This stability suggests their potential suitability for tasks involving a wide range of text lengths, where maintaining accuracy is crucial. The consistent performance of Gemma and

Gemini across various text lengths is particularly advantageous for fact-checking, which often involves analyzing claims of different lengths, from short social media posts to longer articles.

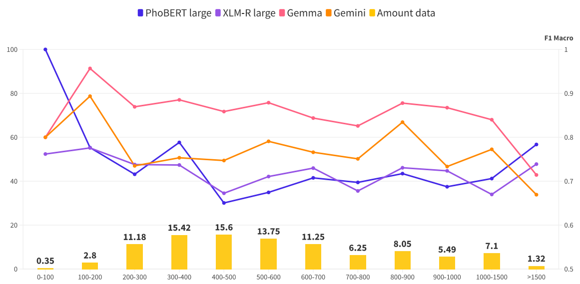


Figure 14: The effect of the length context on test set.

Effects of Topic Further analysis focused on the impact of topics on model performance, as illustrated in Figure 15. Gemma consistently outperforms other models across most topics, excelling particularly in the “Science”, “National Security”, and "Culture" categories. Gemini generally performs the second best, closely following Gemma in most areas but showing a slight dip in “National security” and “Entertainment”. While not as strong overall, PhoBERT$_{large}$ and XLM-R$_{large}$ have their strengths: PhoBERT$_{large}$ performs notably well in “Politics”, benefiting from being pre-trained on a large Vietnamese dataset that provides an advantage in this domain due to the specific vocabulary required. Conversely, XLM-R$_{large}$ shows a relative peak in the “World” category, leveraging its multilingual training data to gain an advantage over monolingual models like PhoBERT$_{large}$.

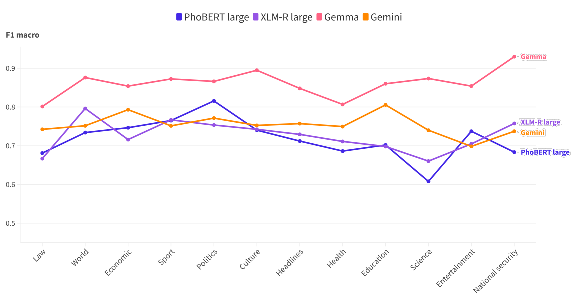


Figure 15: The effect of the topic on the test set.

Interestingly, except for Gemma, the remaining models seem to struggle with the “Science”, “Law”, and “Health” categories, indicating a potential area for improvement in Vietnamese fact-checking models. These categories require high accuracy and specialized vocabulary, which may explain the suboptimal performance of the other models. Additionally, there is a noticeable performance gap between Gemma and the other models in several topics, suggesting that the architecture or training data of Gemma might be better suited for fact-checking Vietnamese across diverse topics.

Effects of Training Data Size To investigate the effect of training data size on model performance, we conducted experiments with various data subsets, including those containing 1000, 2000, 3000, 4000, and 5062 data points. Figure 16 visually represents the evaluation performance across these subsets. Note that all models demonstrated improved performance as the dataset size increased. Given that Gemini is an API-based model and cannot be trained on custom datasets, it was excluded from this analysis.

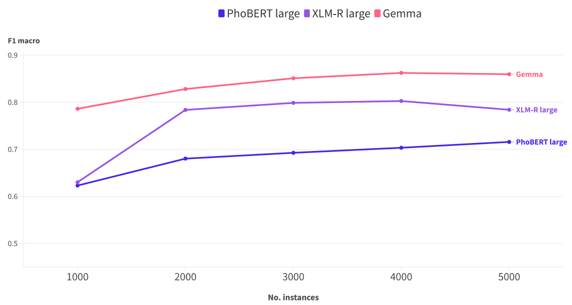


Figure 16: The impact of training data size on test set.

Our comprehensive analysis highlights the multifaceted factors influencing model performance. Gemma consistently outperforms both PhoBERT$_{large}$ and XLM-R$_{large}$ across all training sizes. While all models exhibit improved performance with increased data, F1-score of Gemma starts higher and increases at a steeper rate, especially up to 2,000 instances. Beyond this point, the rate of improvement for all models slows, indicating diminishing returns from additional training data. This consistent superiority demonstrates effectiveness of Gemma regardless of the available training data amount.

Moreover, our findings show that increasing the size of the training data improves the performance of Vietnamese models such as PhoBERT$_{large}$, highlighting the need for a robust and diverse training dataset to achieve optimal fact-checking results.

Analysis of Training Time Efficiency Finally, Figure 17 illustrates the training times of various models per epoch, measured in hours. The Mistral model has the longest training time at 1.1 hours

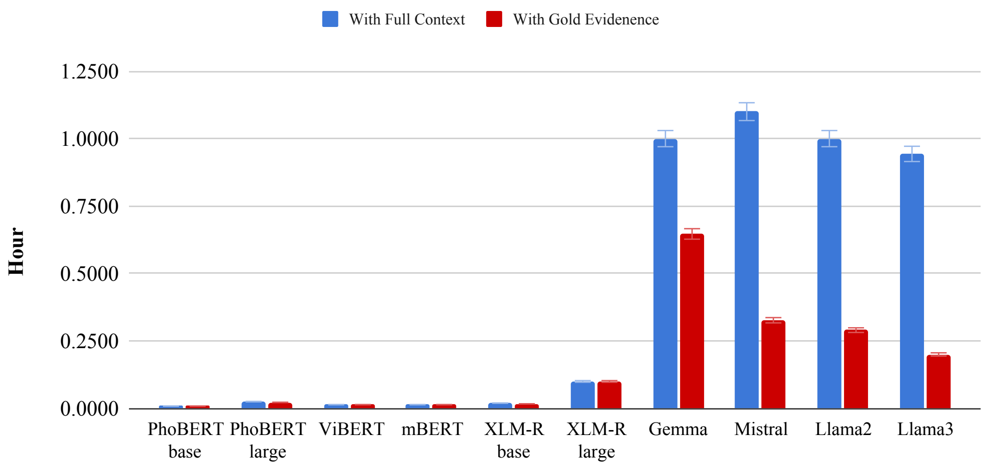


Figure 17: The comparison of training times per epoch for various baseline models.

for processing Full Context (FC), indicative of its complexity and computational demands. Gemma and Llama2 each require approximately 1.0 hour, while Llama3 requires significant time as well, at 0.94 hours. These durations illustrate the intricate computations these models undertake for handling detailed and extensive contexts.

In contrast, the XLM-R$_{large}$ model, though still demanding, is more time-efficient at only 0.1 hours, likely due to its optimized large-scale architecture. The PhoBERT$_{large}$ and XLM-R$_{base}$ models show moderate training times, striking a balance between computational efficiency and performance capabilities.

Models such as mBERT, ViBERT, and PhoBERT$_{base}$ demonstrate shorter training times, ranging from 0.0083 to 0.0139 hours. These reduced durations suggest higher operational efficiency but may also indicate a lower capacity for managing complex tasks requiring extensive contextual data.

When trained with Gold Evidence, which comprises shorter and more directly relevant sentences, Gemma still requires the most time at 0.6467 hours, although this is significantly less than with Full Context. Mistral, Llama2, and Llama3 also exhibit reduced training times at 0.32, 0.29, and 0.20 hours, respectively. This indicates that models can achieve greater efficiency when provided with concise and pertinent training data.

This analysis underscores the trade-offs between training time, model complexity, and performance, highlighting the substantial computational demands placed on advanced models to achieve high performance in Vietnamese fact-checking tasks. The reduced training times with Gold Evidence further emphasize the potential efficiency gains from using relevant training inputs.

In conclusion, our analysis elucidates the multifaceted effects of dataset characteristics and training time on the performance of language models. Larger and more diverse datasets generally improve model accuracy, particularly in specialized applications like fact-checking. However, the efficiency of model training also plays a critical role, as faster training can lead to quicker deployment and adaptation in dynamic environments. The results underscore the importance of optimizing both the data input and model architecture to achieve the best balance between performance and efficiency, which is crucial to develop robust AI systems capable of handling the intricacies of language-based tasks.

M Scientific Artifacts

The licenses for all the models and software used in this paper are listed in parentheses: Beautiful Soup 4 (MIT License), Selenium (Apache License 2.0), Fleiss Kappa (BSD License), mBERT (Apache License 2.0), ViBERT (Apache License 2.0), PhoBERT (MIT License), XLM-R (Apache License 2.0), VnCoreNLP (Apache License 2.0), Unsloth (Apache License 2.0), LoRa (Apache License 2.0), F1-score (BSD License), BM25 (MIT License), SBERT (Apache License 2.0), Gemma (Apache License 2.0), Mistral (Apache License 2.0), Llama3 (Apache License 2.0), Llama2 (Apache License 2.0) Gemini 1.5 Flash (Proprietary License), Label Studio (Apache License 2.0)