SemViQA: A Semantic Question Answering System for Vietnamese Information Fact-Checking

Dien X. Tran[1*] **Nam V. Nguyen**[2*] **Thanh T. Tran**[1] **Anh T. Hoang**[1]
Tai V. Duong[1] ***Di T. Le***[1] ***Phuc-Lu Le***[3]
[1]*Industrial University of Ho Chi Minh City, Vietnam*
[2]*Applied AI Researcher*
[3]*University of Science, VNU-HCM, Vietnam*
Correspondence: *lplu@fit.hcmus.edu.vn*

Abstract

Recent advances in LLMs have accelerated both information generation and misinformation, especially in low-resource languages like Vietnamese, motivating robust fact-checking systems. Existing methods struggle with semantic ambiguity, homonyms, and complex linguistic structures, often trading accuracy for efficiency. We introduce SemViQA, a novel Vietnamese fact-checking framework integrating Semantic-based Evidence Retrieval (SER) and Two-step Verdict Classification (TVC). Our approach balances precision and speed, achieving state-of-the-art results with 78.97% strict accuracy on ISE-DSC01 and 80.82% on ViWikiFC, securing 1st place in the UIT Data Science Challenge. Additionally, SemViQA Faster improves inference speed 7× while maintaining competitive accuracy. SemViQA sets a new benchmark for Vietnamese fact verification, advancing the fight against misinformation. The source code is available at: https://github.com/DAVID-NGUYEN-S16/SemViQA.

1 Introduction

The rapid advancement of large language models (LLMs), such as OpenAI's ChatGPT, Google Gemini (Team et al., 2024), Llama3.1 (Touvron et al., 2023), Qwen2.5 (Qwen et al., 2025), DeepSeek V3, (DeepSeek-AI et al., 2024), Phi3.5 (Abdin et al., 2024) has significantly improved information retrieval and processing across various domains. However, a major challenge with these systems is their tendency to generate factually incorrect or hallucinated content seemingly plausible information that lacks factual grounding (Soleimani et al., 2020). This issue is particularly critical in domains requiring high accuracy, such as healthcare, law, and journalism, where misinformation can have serious consequences. Consequently, developing reliable fact-checking systems capable of retrieving and evaluating evidence from real-world sources has become an urgent need in Natural Language Processing (NLP).

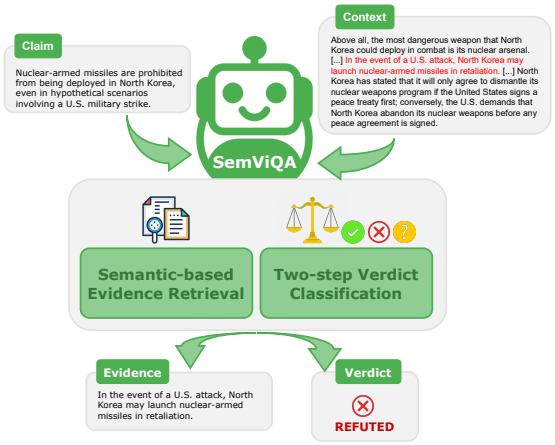

Figure 1: Overview of a Sample Information Fact-Checking Task

Although fact verification has been widely studied in high-resource languages like English, applying these methods to low-resource languages such as Vietnamese remains a significant challenge. Transformer-based models, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have demonstrated strong performance but their adaptation to Vietnamese is still limited. ViNSV (Tran et al., 2024b) employs BM25 and SBERT (Reimers and Gurevych, 2019a) for evidence retrieval but suffers from SBERT's 256-token input constraint, making it ineffective for complex, long-context claims. Graph-based reasoning methods (Zhong et al., 2020) offer promising semantic inference but are often computationally expensive. Traditional retrieval methods like TF-IDF and BM25, while efficient, rely heavily on exact keyword matching, limiting their ability to capture nuanced semantics. Recent large language model (LLM) approaches (Huo et al., 2023; Schimanski et al., 2024) show potential but typically

*Equal contribution.

require substantial computational resources, creating a trade-off between speed and accuracy.

To address these challenges, we propose **SemViQA**, a Vietnamese fact-checking framework that balances semantic accuracy and computational efficiency. As shown in Figure 1, SemViQA comprises three key components:

1. **Semantic-based Evidence Retrieval (SER):** Includes a preprocessing step that efficiently handles long-token contexts by splitting them into manageable subcontexts (e.g., 400 tokens). It combines fast TF-IDF retrieval with selective Question Answering Token Classification (QATC) to strike a balance between speed and semantic accuracy.

2. **Two-step Verdict Classification (TVC):** Employs a hierarchical classification strategy with both three-class and binary classification stages to enhance robustness and improve performance on challenging claim verification tasks.

SemViQA achieves 78.97% strict accuracy on ISE-DSC01 and 80.82% on ViWikiFC (Le et al., 2024), outperforming existing baselines (see Table 2). These results validate SemViQA's potential to enhance Vietnamese fact verification, supporting misinformation mitigation and improved transparency.

The rest of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the methodology, Section 4 reports experimental results, and Section 5 concludes the paper with future directions.

2 Related Works

Advances in Natural Language Processing (NLP) have driven rapid progress in fact verification and evidence extraction. Early BiLSTM-based models, such as the Neural Semantic Matching Network (NSMN) (Nie et al., 2018) augmented with WordNet features improved accuracy but struggled with complex sentence relations due to sequential processing limitations (Graves and Schmidhuber, 2005). Transformer models, notably BERT (Devlin et al., 2019), introduced bidirectional contextual encoding and achieved state-of-the-art results on FEVER (Soleimani et al., 2020; Zhou et al., 2019; Malon, 2018; Aly et al., 2021; Lin et al., 2024; Yuan and Vlachos, 2024; DeHaven and Scott, 2023). Recent work demonstrates that fine-tuned transformers specifically adapted for fact-checking tasks outperform larger language models such as GPT-4 (OpenAI et al., 2024) in claim detection and veracity prediction while being significantly more cost-effective (Setty, 2024). However, their input-length constraints (typically 512 tokens) hinder long-document fact-checking, which is particularly problematic for real-world datasets where contexts frequently exceed 4,000 tokens. Graph-based reasoning methods (Zhong et al., 2020; Thorne et al., 2018) and AdMIRaL's logic-driven retrieval (Aly and Vlachos, 2022) enhance multi-hop verification capabilities but incur considerable computational costs.

Evidence Retrieval Methods. TF-IDF remains the industry standard for document retrieval due to its speed, simplicity, interpretability, and ability to handle long contexts effectively (Reddy et al., 2018; Qaiser and Ali, 2018; Li, 2021; Azevedo et al., 2022). However, its reliance on surface-level keyword matching limits its capability to handle paraphrases, contextual nuances, and multi-hop reasoning, thereby reducing recall accuracy on complex queries. BM25 and SBERT (Reimers and Gurevych, 2019a) offer improvements but face similar limitations. Recent work emphasizes evidence retrieval quality as the dominant factor in fact verification performance (Zheng et al., 2024), with multi-stage reranking pipelines (Malviya and Katsigiannis, 2024) achieving 93.63% recall through the integration of dense retrieval models and list-aware rerankers.

LLM and RAG-based Approaches. Large language models, including GPT-4 (OpenAI et al., 2024), Gemini (Team et al., 2024), Llama 3 (Touvron et al., 2023), Qwen 2.5 (Qwen et al., 2024), and DeepSeek V3 (DeepSeek-AI et al., 2024), demonstrate impressive capabilities but struggle with factual hallucinations (Soleimani et al., 2020). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) mitigates this limitation by grounding generation in retrieved evidence. Recent RAG systems for fact-checking (Asai et al., 2024; Khaliq et al., 2024) convert claims into structured queries to retrieve information from knowledge bases. Advanced RAG architectures employ dynamic retrieval triggering mechanisms (Su et al., 2024; Jiang et al., 2023) based on generation uncertainty.

Vietnamese Fact-Checking. Research on Vietnamese fact verification remains limited compared to high-resource languages. ViNSV (Tran et al.,

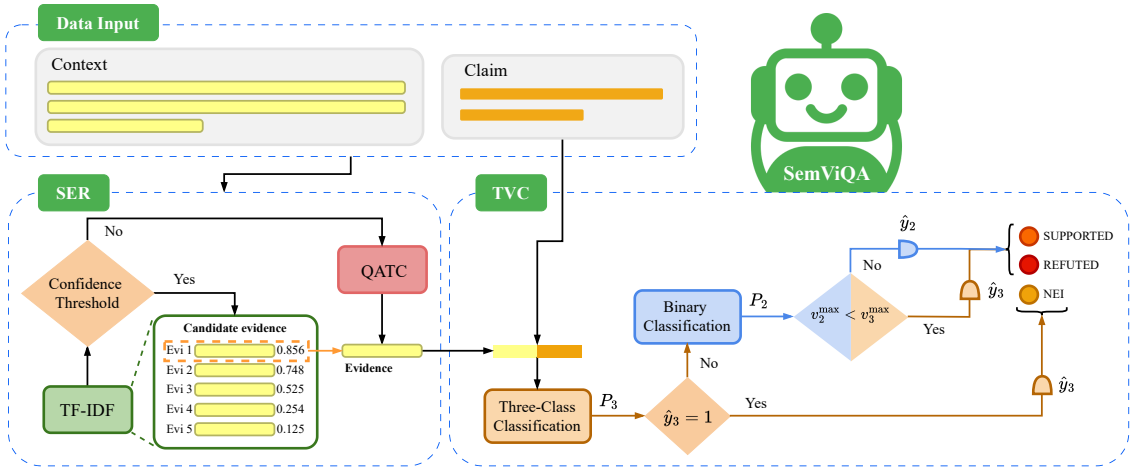

Figure 2: **SemViQA**: A Two-Stage Method for Semantic-based Evidence Retrieval (SER) and Two-step Verdict Classification (TVC), where P_2 and P_3 represent the probabilities of the two-class and three-class classifications, respectively, and $\hat{y}_2$ and $\hat{y}_3$ denote their corresponding predictions.

2024b) pairs BM25 with SBERT but falters on complex reasoning tasks due to static embeddings and input constraints. Recent datasets include ViWikiFC (Le et al., 2024), comprising over 20,000 Wikipedia-based claims; ISE-DSC01 from the UIT Challenge 2023; ViFactCheck (Hoa et al., 2025) for multi-domain benchmarking; and ViNumFCR (Luong et al., 2025) for numerical reasoning. Knowledge graph approaches (Duong et al., 2023) integrate Datalog reasoning with KG-BERT for structured knowledge representation.

Ensemble learning methods (Hannichenko et al., 2023; Wang et al., 2021; Liu et al., 2024; Ganaie et al., 2022) mitigate individual model weaknesses by aggregating diverse architectures and training signals, yielding robust performance gains. Building on these insights, **SemViQA** integrates fast TF-IDF retrieval, semantic reasoning via QATC, and hierarchical classification, delivering high accuracy, low latency, and practical scalability for Vietnamese fact verification.

3 SemViQA - Semantic Vietnamese Question Answering

We formulate Vietnamese fact verification as a multi-output classification task, where the input is a pair (C, X), with C being a claim and X its corresponding context paragraph or document. The objective is to (i) identify the most relevant evidence sentence(s) from X and (ii) predict the veracity label of the claim as one of three categories: Supported, Refuted, or Not Enough Information (NEI).

To address challenges such as long input sequences and semantic ambiguity, we propose **SemViQA**, a three-stage framework consisting of data preprocessing, semantic-based evidence retrieval, and two-step verdict classification. An overview of the architecture is shown in Figure 2, with detailed descriptions provided in the following subsections.

3.1 Evidence Extraction via Question Answering with Token Classification

3.1.1 Data Processing

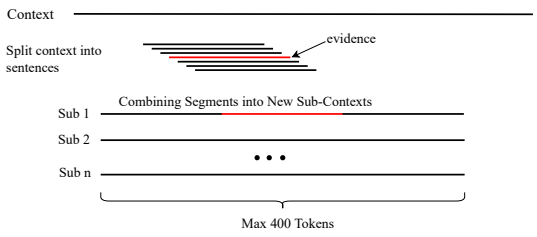

Figure 3: Long context processing solution.

To effectively support evidence retrieval and claim classification, we apply distinct preprocessing strategies to the context based on the specific requirements of each downstream task. A key challenge arises from the considerable length of many context passages, frequently exceeding the token limits of Vietnamese BERT-based models (see Appendix A for detailed token length analysis). A key challenge arises from the considerable length of many context passages, frequently exceeding the token limits of Vietnamese BERT-based models.

Figure 3 illustrates our approach for handling long input contexts. First, the context is segmented into individual sentences. Next, sentences are sequentially aggregated into subcontexts until reaching approximately 400 tokens. Each completed subcontext is then processed by the QATC model to identify potential evidence. The next subcontext begins from the subsequent sentence, and this process continues until all sentences are processed. However, processing subcontexts sequentially can be time-consuming. Therefore, we developed SemViQA Faster, which batches and processes subcontexts in parallel, significantly accelerating the retrieval process.

3.1.2 Question Answering with Token Classification (QATC)

Traditional Question Answering models typically predict the start and end positions of an answer span. In our framework, we enhance this approach by introducing a **token-level classification objective**, enabling the model to focus explicitly on tokens within evidence sentences in the context. This dual formulation provides improved supervision for evidence extraction. Drawing inspiration from rationale tagging (Ju et al., 2019), we treat token labeling as a binary classification task: tokens within evidence sentences receive a label of 1, and all other tokens receive a label of 0. In cases marked NEI (Not Enough Information), every token is labeled as 0. We employ a feed-forward classification layer on the token representations:

$$p_t = \sigma\left(W_2 \cdot \text{ReLU}(W_1 h_t)\right), \tag{1}$$

where h_t is the contextual representation of token t, W_1, W_2 are learnable weights in the neural network and $\sigma(\cdot)$ denotes the sigmoid function. The loss for this task is the Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{RT} = -\frac{1}{T}\sum_{t=1}^{T} \text{BCE}(y_t, p_t), \tag{2}$$

3.2 Semantic-based Evidence Retrieval (SER)

Accurate claim verification requires reliable evidence. To improve both efficiency and robustness, we adopt a two-stage evidence retrieval strategy combining TF-IDF with a QATC.

Stage 1: TF-IDF-based Retrieval. We segment the context X into smaller passages and pair each with the claim C. Preprocessing includes noise removal and tokenization using ViTokenizer[1]. TF-IDF is effective for simple claims particularly refuted ones but struggles with semantically complex cases due to its reliance on keyword overlap. To enrich short segments (i.e., those with fewer than 60% of C's tokens), we merge them with preceding segments to improve evidence completeness. Retrieved segments are then ranked, and a confidence threshold is applied to identify easy cases (handled by TF-IDF) and hard cases (passed to QATC).

Stage 2: QATC-based Refinement. For complex cases, QATC is applied to segmented subcontexts rather than the full context due to the input length limitation of BERT models. The detailed processing approach is described in Section 3.1.1. At this time, we consider three scenarios: (1) If multiple subcontexts yield conflicting answers, we collect all predicted spans and re-rank them using TF-IDF. (2) If a single evidence span consistently appears, it is selected directly. (3) If no evidence is found, fallback to TF-IDF is used.

This hybrid approach balances speed and semantic accuracy, improving evidence selection for downstream verdict classification. Examples are provided in Appendix E.

3.3 Two-step Verdict Classification (TVC)

We adopt a two-stage classification framework to enhance claim verification robustness and mitigate label imbalance, especially the overrepresentation of *NEI*.

Stage 1: Three-Class Classification. Given a claim-evidence pair (C, E), a BERT-based model $f_{\text{3-class}}$ predicts a probability distribution over three labels: *Supported*, *Refuted*, and *Not Enough Information (NEI)*:

$$P_3 = f_{\text{3-class}}(C, E), \quad \hat{y}_3 = \arg\max_k P_3. \tag{3}$$

This step is optimized using Cross-Entropy Loss.

Stage 2: Binary Classification. If $\hat{y}_3 \neq$ NEI, we apply a refined binary classifier $f_{\text{2-class}}$ to distinguish between *Supported* and *Refuted*:

$$P_2 = f_{\text{2-class}}(C, E), \quad \hat{y}_2 = \arg\max_k P_2. \tag{4}$$

This model uses Focal Loss (Lin et al., 2018) to address class imbalance.

[1] https://github.com/trungtv/pyvi

Final Prediction Rule. The final verdict $\hat{y} \in \{1, 2, 3\}$ where $1 =$ NEI, $2 =$ Supported, $3 =$ Refuted is determined by comparing the confidence scores from both classifiers. Here, $\hat{y}$ represents the index of the predicted label, where each index corresponds to a specific class description.

$$\hat{y} = \begin{cases} \hat{y}_3, & \text{if } \hat{y}_3 = 1, \\ \hat{y}_3, & \text{if } v_3^{\max} > v_2^{\max}, \\ \hat{y}_2, & \text{otherwise}, \end{cases} \quad (5)$$

where $v_3^{\max} = \max(P_3)$, $v_2^{\max} = \max(P_2)$ represents the highest probability.

This hybrid strategy allows the three-class model to handle general cases, especially detecting NEI early, while the binary model specializes in distinguishing difficult SUP/REF cases.

3.4 SemViQA Pipeline System

We now describe the full SemViQA pipeline, as illustrated in Figure 2. First, we prepare input for TF-IDF by splitting the context paragraph X into sentences, then concatenating each sentence with the claim C. We calculate the matching score for each sentence and select the one with the highest probability. If this score exceeds the threshold t, we directly use this sentence as the evidence. If the score is below t, we proceed to prepare input for the QATC model. We segment X into subcontexts (as detailed in Section 3.1.1). Each subcontext is sequentially processed by the QATC model. If QATC identifies zero or multiple candidate evidence spans, we collect all predicted spans and re-rank them using TF-IDF. If QATC finds exactly one candidate evidence span, we confidently select it as the final evidence. Finally, we move to the two-step verdict classification (TVC) stage. We prepare input for TVC by concatenating the claim C with the final evidence. We first apply the three-class model. If it predicts NEI, the process ends. Otherwise, we use an ensemble method to combine the weights from both the three-class and binary models to make the final prediction. All coefficients used in our system are provided in the ablation study, as shown in Appendix C.

4 Experiments

4.1 Dataset

We evaluate our approach on two Vietnamese fact verification datasets: **ISE-DSC01** from the UIT Challenge 2023 and **ViWikiFC** (Le et al., 2024). ISE-DSC01 comprises real-world news-based claims with complex, multi-domain contexts, while ViWikiFC contains over 20,000 Wikipedia-based claims with annotated evidence for all three labels, including "Not Enough Information" (NEI).

Table 1 presents the detailed statistics of both datasets. ISE-DSC01 provides 37,967 training samples, 4,794 development samples, and 5,396 test samples, offering substantial data for model training and evaluation. ViWikiFC includes 16,738 training samples, 2,090 development samples, and 2,091 test samples. These statistics contextualize the scale and evaluation coverage of our methods across different data distributions and complexity levels.

	ISE-DSC01	**ViWikiFC**
Train	37,967	16,738
Dev	4,794	2,090
Test	5,396	2,091

Table 1: Dataset statistics for ISE-DSC01 and ViWikiFC.

4.2 Experimental Setup

We conducted extensive experiments on NVIDIA A100 GPUs, fine-tuning key hyperparameters while keeping consistent settings across runs. The final configuration, selected via rigorous validation, improved both accuracy and strict accuracy on ISE-DSC01 and ViWikiFC. Full details are provided in Appendix D. For fair evaluation, all methods were tested on a Kaggle instance with an NVIDIA T4 GPU.

The large language model was fine-tuned in a distributed A100 setup using a structured prompt-based reformulation. Raw data were converted into prompt format to align with LLM training objectives and maximize task-specific performance. Training setup, prompt design, and preprocessing are also detailed in Appendix D.

4.3 Main Results

The results in Table 2 demonstrate that SemViQA outperforms previous methods in Vietnamese fact-checking tasks. Specifically, our model achieves the highest Strict Accuracy, reaching 80.82% on ViWikiFC and 78.97% on ISE-DSC01, establishing a new benchmark for automated fact-checking systems in Vietnamese language.

Method		ViWikiFC				ISE-DSC01				Avg Strict Acc
ER	VC	Strict Acc	VC Acc	ER Acc	Time (s)	Strict Acc	VC Acc	ER Acc	Time (s)	
Traditional Baselines										
	InfoXLM$_{large}$	75.56	82.21	90.15	131	73.59	78.08	76.61	378	74.58
TF-IDF	XLM-R$_{large}$	76.47	82.78	90.15	134	75.61	80.50	78.58	366	76.04
	Ernie-M$_{large}$	75.56	81.83	90.15	144	78.19	81.69	80.65	403	76.88
	InfoXLM$_{large}$	70.44	79.01	83.50	130	72.09	77.37	75.04	320	71.27
BM25	XLM-R$_{large}$	70.97	78.91	83.50	132	73.94	79.37	76.95	333	72.46
	Ernie-M$_{large}$	70.21	78.29	83.50	141	76.58	80.76	79.02	381	73.40
	InfoXLM$_{large}$	74.99	81.59	89.72	195	71.20	76.59	74.15	915	73.10
SBert	XLM-R$_{large}$	75.80	82.35	89.72	194	72.85	78.78	75.89	835	74.33
	Ernie-M$_{large}$	75.13	81.44	89.72	203	75.46	79.89	77.91	920	75.30
QA-based Approaches										
	InfoXLM$_{large}$	77.28	81.97	92.49	3778	54.36	64.14	56.84	9798	65.82
ViMRC$_{large}$	XLM-R$_{large}$	78.29	82.83	92.49	3824	53.98	66.70	57.77	9809	66.14
	Ernie-M$_{large}$	77.38	81.92	92.49	3785	56.62	62.19	58.91	9833	67.00
	InfoXLM$_{large}$	78.14	82.07	93.45	4092	53.50	63.83	56.17	10057	65.82
InfoXLM$_{large}$	XLM-R$_{large}$	79.20	83.07	93.45	4096	53.32	66.70	57.25	10066	66.26
	Ernie-M$_{large}$	78.24	82.21	93.45	4102	56.34	62.36	58.69	10078	67.29
LLMs										
Qwen2.5-1.5-Instruct		51.03	65.18	78.96	7665	59.23	66.68	65.51	19780	55.13
Qwen2.5-3B-Instruct		44.38	62.31	71.35	12123	60.87	66.92	66.10	31284	52.63
	InfoXLM$_{large}$	66.14	76.47	78.96	7788	64.40	68.37	66.49	19970	65.27
Qwen2.5-1.5-Instruct	XLM-R$_{large}$	67.67	78.10	78.96	7789	64.66	69.63	66.72	19976	66.17
	Ernie-M$_{large}$	66.52	76.52	78.96	7794	65.70	68.37	67.33	20003	66.11
	InfoXLM$_{large}$	59.88	72.50	71.35	12246	65.72	69.66	67.51	31477	62.80
Qwen2.5-3B-Instruct	XLM-R$_{large}$	60.74	73.08	71.35	12246	66.12	70.44	67.83	31483	63.43
	Ernie-M$_{large}$	60.02	72.21	71.35	12251	67.48	70.77	68.75	31512	63.80
Ours: SER Faster + TVC										
ViMRC$_{large}$	Ernie-M$_{large}$	**79.44**	**82.93**	**94.60**	410	**78.32**	**81.91**	**80.26**	995	**78.88**
InfoXLM$_{large}$		**79.77**	**83.07**	**95.03**	487	**78.37**	**81.91**	**80.32**	925	**79.07**
Ours: Full SER + TVC										
	InfoXLM$_{large}$	80.25	83.84	94.69	2731	75.13	79.54	76.87	5191	77.69
ViMRC$_{large}$	XLM-R$_{large}$	80.34	83.64	94.69	2733	76.71	81.65	78.91	5219	78.53
	Ernie-M$_{large}$	79.53	82.97	94.69	2733	**78.97**	**82.54**	**80.91**	5225	79.25
	InfoXLM$_{large}$	80.68	**83.98**	**95.31**	3860	75.13	79.60	76.87	5175	77.91
InfoXLM$_{large}$	XLM-R$_{large}$	**80.82**	83.88	**95.31**	3843	76.74	81.71	78.95	5200	78.78
	Ernie-M$_{large}$	80.06	83.17	**95.31**	3891	**78.97**	82.49	**80.91**	5297	**79.52**

Table 2: Performance comparison on the ViWikiFC test set and the ISE-DSC01 private-test dataset. The results highlight differences among models based on several criteria: Strict Accuracy (Strict Acc), Veracity Classification Accuracy (VC Acc), and Evidence Retrieval Accuracy (ER Acc). Time represents the total inference time required to generate the complete results. **Avg Strict Acc** is the average of Strict Accuracy across both datasets. Bold values indicate the best performance in each metric.

4.3.1 Performance Comparison

a) Handling Long Token Sequences in Fact-Checking A major limitation of conventional Question Answering (QA) models in fact verification is their inability to process long-context claims due to the 512-token input limit of transformer-based models such as ViMRC$_{large}$[2], InfoXLM$_{large}$ (Chi et al., 2021), XLM-R$_{large}$ (Conneau et al., 2020), and Ernie-M$_{large}$ (Ouyang et al., 2021). Real-world datasets like ISE-DSC01 often contain contexts exceeding 4800 tokens, severely degrading QA-based performance by limiting access to full evidence. To overcome this, SemViQA employs an efficient retrieval-based strategy (see Section 3.1.1) that handles long-token sequences effectively. On ISE-DSC01, SemViQA outperforms traditional QA models by fully leveraging extended contexts, confirming that the long-token constraint is a critical bottleneck. Conversely, on ViWikiFC, where contexts average around 512 tokens, QA models perform competitively. Yet, even in this setting, integrating our Semantic-based Evidence Retrieval (SER) yields a 1.86% improvement in evidence retrieval accuracy, demonstrating the versatility and efficiency of our approach. These findings emphasize that long-token limitations significantly hinder fact verification, and SemViQA successfully mitigates this issue while enhancing QA models across varying dataset conditions.

b) Performance and Inference Time Optimization SemViQA significantly reduces inference time while maintaining high accuracy, making it highly practical for real-world applications. Key highlights include:

- On ISE-DSC01, SemViQA averages 5200s per run, over **6 times faster** than large LLM-based models like Qwen2.5-3B-Instruct (Qwen et al., 2024), which require over 31,000s.
- Compared to ViMRC$_{large}$ (9800s), SemViQA

[2] https://huggingface.co/nguyenvulebinh/vi-mrc-large

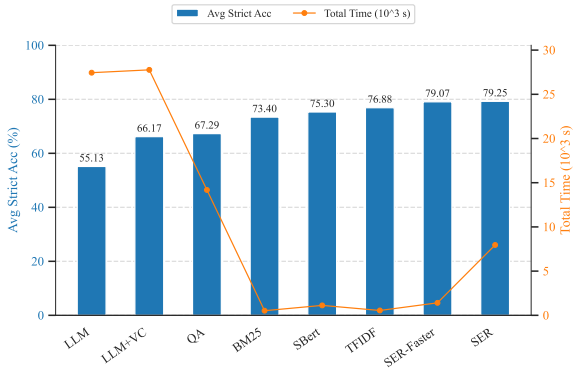

Figure 4: Comparison of methods in terms of peak performance and total inference time across datasets. Each retrieval approach is evaluated by its best score, while overall efficiency is reflected through cumulative inference time. See Table 2 for details.

halves inference time while achieving superior Strict Accuracy and Veracity Classification Accuracy.

- Although BM25 and SBERT (Reimers and Gurevych, 2019b) are faster, they struggle with complex, multi-step reasoning where SemViQA maintains a strong balance between speed and accuracy.

SemViQA Faster: We further introduce SemViQA Faster, which accelerates inference by up to **7**× through batch processing of subcontexts (see Section 3.1.1). As shown in Figure 4, this variant achieves inference speeds comparable to traditional methods while retaining competitive accuracy. The minor performance trade-off is acceptable given the substantial time savings, making SemViQA Faster ideal for scalable, real-world fact-checking systems.

4.3.2 Comparison with other results in the competition

Methods	Strict Acc	VC Acc	ER Acc
SemViQA	**78.97**	82.54	**80.91**
DS@UIT Dynasty[3]	78.05	**84.76**	80.13
URA_FNU[4]	77.87	83.71	79.96
ViNSV (Tran et al., 2024b)	76.33	81.67	78.11
(Tran et al., 2024a)	75.11	82.30	76.82

Table 3: Private leaderboard comparison of top systems in the ISE-DSC01 competition.

[3]https://github.com/minhquan6203/Verdict-Classification-for-Fact-Checking-at-DSC-2023

[4]https://github.com/virrosluo/URA_UIT_Data_Science_Challenge

The results presented in Table 3 indicate that our SemViQA approach outperforms other competing teams, achieving the highest Strict Accuracy and demonstrating exceptional effectiveness in information processing and verification. This achievement highlights SemViQA's capability to deliver significantly more accurate and reliable results.

5 Conclusion and Future Works

We introduced SemViQA, a Vietnamese fact-checking framework that integrates Semantic-based Evidence Retrieval (SER) and Two-step Verdict Classification (TVC) to enhance claim verification. Our approach outperforms existing methods, including LLMs, TF-IDF, BM25, SBERT, and QA-based models, particularly in handling long-token sequences and complex reasoning tasks. Extensive experiments demonstrated SemViQA's state-of-the-art performance on ISE-DSC01 and ViWikiFC. Additionally, the SemViQA Faster variant accelerates inference by up to 7×, improving its practicality for real-world applications. By addressing key challenges such as semantic ambiguity and multi-step reasoning, SemViQA lays the groundwork for advancing Vietnamese NLP, with potential applications in misinformation detection and low-resource language fact-checking.

Acknowledgements

We would like to express our sincere gratitude to all those who contributed to the successful completion of this research. We are particularly grateful to Dr. Hung Bui, Head of the Data Science Department, as well as other professors and reviewers, for taking the time to review this paper and for their suggestions. We thank the organizers of the UIT Data Science Challenge 2023 for providing the ISE-DSC01 dataset. This work was supported by the Industrial University of Ho Chi Minh City and the University of Science, VNU-HCM. We also thank the anonymous reviewers for their constructive feedback.

Limitations

While SemViQA demonstrates strong performance in Vietnamese fact verification, several limitations remain. First, our reliance on TF-IDF for initial evidence retrieval, while efficient, limits the model's ability to capture deep semantic relationships and retrieve implicit evidence. To mitigate this, we employ a threshold-based mechanism to identify hard samples and process them with a more advanced retrieval model. However, this approach relies on manually defined thresholds, which may not generalize well across different datasets, underscoring the need for adaptive and data-driven retrieval strategies in future work. Second, our Two-step Verdict Classification (TVC) framework improves claim verification accuracy but requires multiple classification stages, increasing inference time compared to single-step approaches. This additional computational cost is particularly significant in three-class classification tasks, where optimizing model efficiency without compromising accuracy remains a key challenge. Future work should focus on refining retrieval mechanisms and classification strategies to enhance efficiency and robustness, ensuring broader applicability of SemViQA in real-world fact verification scenarios.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6123–6135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, volume 2024, pages 9112–9141.

Pedro Azevedo, Gil Rocha, Diego Esteves, and Henrique Lopes Cardoso. 2022. Towards better evidence extraction methods for fact-checking systems. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, page 277–284, New York, NY, USA. Association for Computing Machinery.

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

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.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Mitchell DeHaven and Stephen Scott. 2023. BEVERS: A general, simple, and performant framework for automatic fact verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65, Dubrovnik, Croatia. 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 Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(10).

Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks : the official journal of the International Neural Network Society*, 18:602–10.

Tetyana Hannichenko, Peter Bidyuk, Irina Kalinina, and Oleksandr Zhebko. 2023. Classification system based on ensemble methods for solving machine learning tasks.

Tran Thai Hoa, Tran Quang Duy, Khanh Quoc Tran, and Kiet Van Nguyen. 2025. Vifactcheck: a new benchmark dataset and methods for multi-domain news fact-checking in vietnamese. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 11–20. ACM.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *Preprint*, arXiv:1909.10772.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Kun Li. 2021. Haha at fakedes 2021: A fake news detection method based on tf-idf and ensemble machine learning. In *IberLEF@ SEPLN*, pages 630–638.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

Ying-Jia Lin, Chun-Yi Lin, Chia-Jen Yeh, Yi-Ting Li, Yun-Yu Hu, Chih-Hao Hsu, Mei-Feng Lee, and Hung-Yu Kao. 2024. Cfever: A chinese fact extraction and verification dataset. *Preprint*, arXiv:2402.13025.

Yichen Liu, Abhijit Dasgupta, and Qiwei He. 2024. Music genre classification: Ensemble learning with subcomponents-level attention. *Preprint*, arXiv:2412.15602.

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.

Nhi Ngoc Phuong Luong, Anh Thi Lan Le, Tin Van Huynh, Kiet Van Nguyen, and Ngan Nguyen. 2025. ViNumFCR: A novel Vietnamese benchmark for numerical reasoning fact checking on social media news. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 134–147, Hanoi, Vietnam. Association for Computational Linguistics.

Christopher Malon. 2018. Team papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113, Brussels, Belgium. Association for Computational Linguistics.

Shrikant Malviya and Stamos Katsigiannis. 2024. Evidence retrieval for fact verification using multi-stage reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7295–7308, Miami, Florida, USA. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2018. Combining fact extraction and verification with neural semantic matching networks. *Preprint*, arXiv:1811.07039.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. 2018. Defactonlp: Fact verification using entity recognition, tfidf vector comparison and decomposable attention. *arXiv preprint arXiv:1809.00509*.

Nils Reimers and Iryna Gurevych. 2019a. 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.

Nils Reimers and Iryna Gurevych. 2019b. 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.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931, Bangkok, Thailand. Association for Computational Linguistics.

Vinay Setty. 2024. Surprising efficacy of fine-tuned transformers for fact-checking over larger language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2842–2846, New York, NY, USA. Association for Computing Machinery.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, page 359–366, Berlin, Heidelberg. Springer-Verlag.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Bao Tran, T. N. Khanh, Khang Nguyen Tuong, Thien Dang, Quang Nguyen, Nguyen T. Thinh, and Vo T. Hung. 2024a. *BERT-Based Model for Vietnamese Fact Verification Dataset*, page 219–231. Springer Nature Switzerland.

Quang-Duy Tran, Thai-Hoa Tran, and Khanh Quoc Tran. 2024b. Advancing vietnamese fact extraction and verification through multi-stage text ranking. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–7.

Guangtao Wang, Qinbao Song, and Xiaoyan Zhu. 2021. Ensemble learning based classification algorithm recommendation. *Preprint*, arXiv:2101.05993.

Moy Yuan and Andreas Vlachos. 2024. Zero-shot fact-checking with semantic triples and knowledge graphs. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 105–115, Bangkok, Thailand. Association for Computational Linguistics.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9274–9281, Bangkok, Thailand. 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.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

A Data Analysis on Context Lengths

Figure 5 illustrates the distribution of token lengths across input contexts in the ViWikiFC and ISE-DSC01 datasets. As shown, many samples significantly exceed the 512-token input limitation of standard Transformer models. The ISE-DSC01 dataset, in particular, contains several contexts with over 4,000 tokens. This analysis highlights the necessity for effective context segmentation strategies to ensure full coverage of relevant evidence while maintaining compatibility with model constraints.

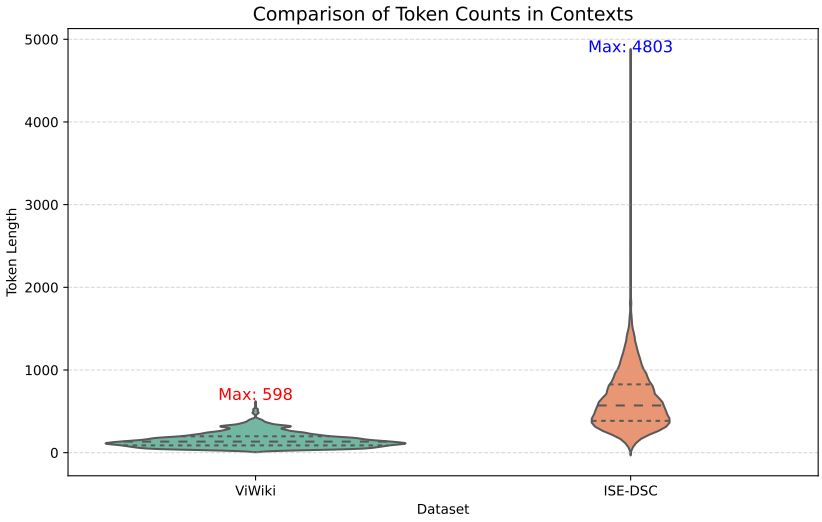

Figure 5: Graph representing the lengths of contexts.

Figure 3 illustrates our strategy for handling long input contexts. First, the context is segmented into individual sentences. Next, sentences are sequentially aggregated into subcontexts until reaching approximately 400 tokens. Each completed subcontext is then passed to the QATC model to identify potential evidence. The next subcontext begins from the subsequent sentence, continuing until all sentences have been processed. However, processing subcontexts sequentially can be time-consuming. Therefore, we developed SER Faster, which batches subcontexts and processes them in parallel, significantly accelerating the retrieval process.

B Strict Accuracy in Fact-Checking

Strict Accuracy: This metric is a stringent measure that requires both the verdict and the evidence to be predicted correctly compared to the ground truth sample.

Verdict (v and v'): refers to the verdict of the sample and the predicted verdict (supported, refuted, nei).

Evidence (e and e'): refers to the evidence of the sample and the predicted evidence.

$$StrAcc = f(v, v').f(e, e') \tag{6}$$

Where:

$$f(v, v') = \begin{cases} 1 & \text{if } v = v' \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$f(e, e') = \begin{cases} 1 & \text{if } e = e' \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Strict accuracy is the average of all StrAcc values.

Configuration		ViWikiFC			ISE-DSC01		
ER Model	**VC Model**	Strict Acc	VC Acc	ER Acc	Strict Acc	VC Acc	ER Acc
SemViQA (Full)							
	InfoXLM$_{large}$	80.68	**83.98**	**95.31**	75.13	79.60	76.87
InfoXLM$_{large}$	XLM-R$_{large}$	**80.82**	83.88	**95.31**	76.74	81.71	78.95
	Ernie-M$_{large}$	80.06	83.17	**95.31**	**78.97**	**82.49**	**80.91**
w/o Binary Classification (One-step VC)							
	InfoXLM$_{large}$	79.63	82.88	95.31	73.87	78.35	76.89
InfoXLM$_{large}$	XLM-R$_{large}$	80.73	83.69	95.31	75.96	80.80	78.97
	Ernie-M$_{large}$	79.91	83.07	95.31	78.47	81.89	80.93
w/o QATC (TF-IDF-based ER)							
	InfoXLM$_{large}$	76.57	83.26	90.15	74.89	79.36	76.61
TF-IDF	XLM-R$_{large}$	76.47	82.93	90.15	76.39	81.41	78.58
	Ernie-M$_{large}$	75.75	81.97	90.15	78.71	82.28	80.65

Table 4: Ablation results of SemViQA.

C Ablation Study

C.1 Ablation Experiments

Table 4 presents the ablation results to evaluate the contribution of each component in the SemViQA framework. When employing the full model with both the QATC-based evidence retrieval and the two-step verdict classification (TVC), SemViQA achieves the best performance across both datasets. Notably, using Ernie-M$_{large}$ yields the highest strict accuracy (78.97%) and evidence retrieval accuracy (80.91%) on the ISE-DSC01 dataset. Removing the binary classification stage (i.e., using only one-step VC) leads to a noticeable performance drop, especially on ISE-DSC01, indicating that the binary classifier enhances the distinction between *Supported* and *Refuted* labels. Furthermore, replacing the QATC module with a TF-IDF based retriever results in a significant decline (about 5%) in evidence retrieval accuracy, which subsequently affects the overall performance. These findings highlight the critical role of both QATC and the two-step classification scheme in improving SemViQA's effectiveness on fact verification in Vietnamese.

C.2 Analysis of Confidence Threshold in SemViQA

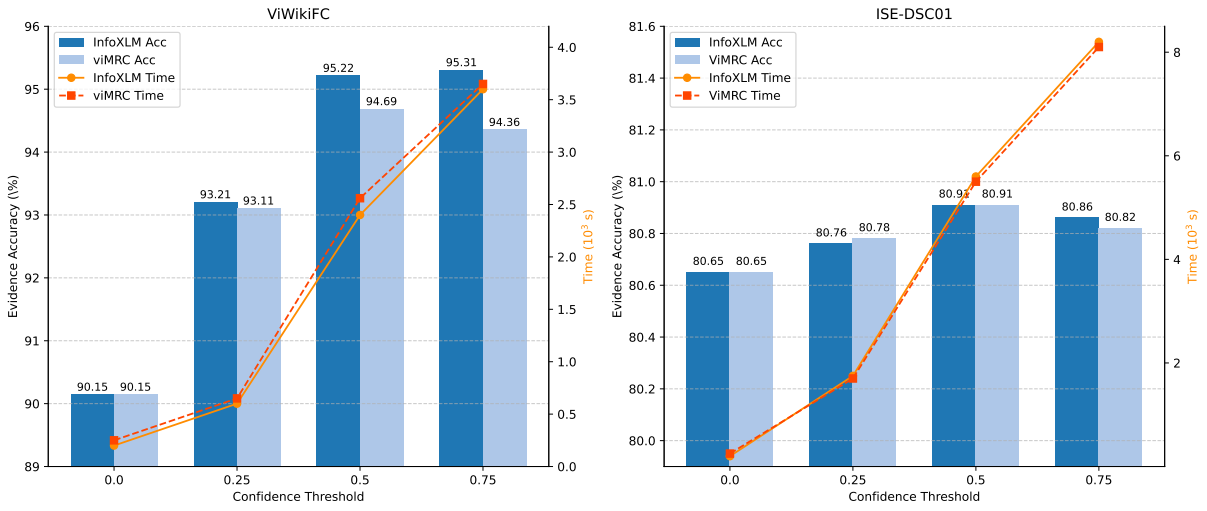

Figure 6: Impact of confidence threshold on evidence retrieval accuracy in SemViQA.

The confidence threshold plays a crucial role in balancing accuracy and inference time in SemViQA's evidence retrieval process. Analysis from Figure 6 indicates that as the threshold increases from 0.0 to 0.5, evidence retrieval accuracy improves significantly, particularly on ViWikiFC (95%) and ISE-DSC01 (80.8%). However, beyond 0.5, accuracy gains plateau, while inference time decreases sharply due to the

system filtering out low-confidence evidence more aggressively. Setting an optimal threshold in the range of 0.4 - 0.5 achieves a trade-off between efficiency and accuracy, ensuring that SemViQA operates swiftly while maintaining precise evidence retrieval.

C.3 Effectiveness of QATC over Traditional QA-based Evidence Retrieval

SemViQA		ViWikiFC			ISE-DSC01		
SER	TVC	Strict Acc	VC Acc	ER Acc	Strict Acc	VC Acc	ER Acc
QATC-based ER							
InfoXLM$_{\text{large}}$	InfoXLM$_{\text{large}}$	80.68	**83.98**	**95.31**	75.13	79.60	76.87
	XLM-R$_{\text{large}}$	**80.82**	83.88	**95.31**	76.74	81.71	78.95
	Ernie-M$_{\text{large}}$	80.06	83.17	**95.31**	**78.97**	**82.49**	**80.91**
ViMRC$_{large}$	InfoXLM$_{large}$	80.25	83.84	94.69	75.13	79.54	76.87
	XLM-R$_{large}$	80.34	83.64	94.69	76.71	81.65	78.91
	Ernie-M$_{large}$	79.53	82.97	94.69	**78.97**	**82.54**	**80.91**
QA-based ER							
InfoXLM$_{\text{large}}$	InfoXLM$_{\text{large}}$	79.96	83.50	94.45	74.02	78.95	75.83
	XLM-R$_{\text{large}}$	80.11	83.60	94.45	75.61	80.95	77.91
	Ernie-M$_{\text{large}}$	79.24	82.74	94.45	77.82	81.76	79.82
ViMRC$_{large}$	InfoXLM$_{large}$	79.77	83.84	94.26	74.05	78.93	75.87
	XLM-R$_{\text{large}}$	79.87	83.79	94.26	75.65	80.93	77.95
	Ernie-M$_{\text{large}}$	79.01	82.78	94.26	77.84	81.73	79.86

Table 5: Comparison of QATC-based vs. QA-based evidence retrieval in SemViQA. QATC consistently improves all metrics.

To compare the learning capabilities of QATC with traditional QA models, we construct two versions of the SemViQA pipeline that are identical in structure, differing only in the evidence retrieval model. As shown in Table 5, using QATC consistently outperforms QA-based retrieval across both datasets. Specifically, on ViWikiFC, QATC achieves up to 80.82% Strict Accuracy and 95.31% ER Accuracy, while QA-based models peak at 80.11% and 94.45%, respectively. The improvement is even more evident on ISE-DSC01, where QATC reaches 78.97% Strict Accuracy and 80.91% ER Accuracy. These results confirm that QATC is more effective at learning to identify relevant evidence, leading to superior performance across the entire fact-checking pipeline.

C.4 Analysis of Confusion Matrix in Verdict Classification

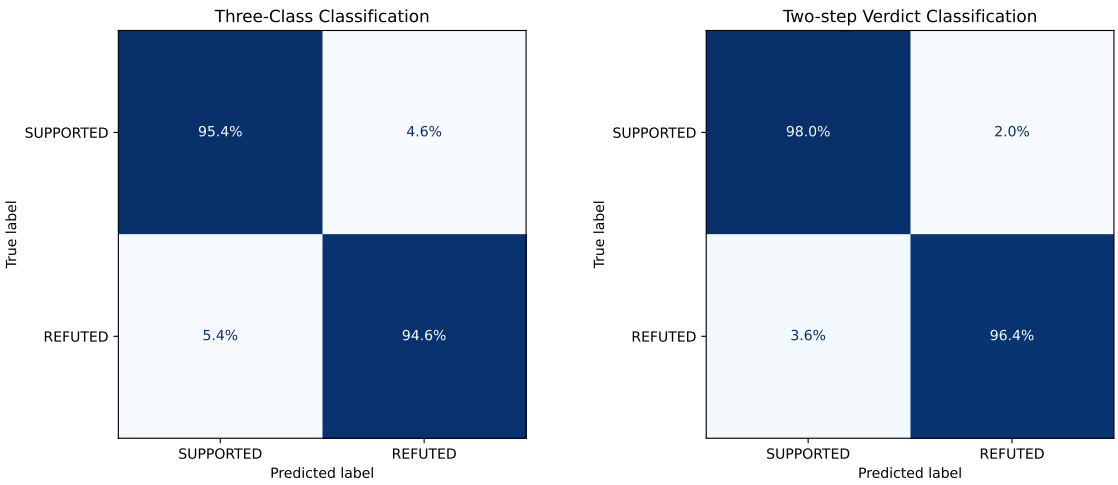

Figure 7: Confusion matrix of the Three-Class and Two-Step Verdict Classification (TVC) on the ISE-DSC01 dataset.

To evaluate the effectiveness of our Two-step Verdict Classification (TVC) strategy, we analyze the confusion matrices in Figure 7, which compare the standard three-class classification with our proposed

two-step approach, focusing solely on **Supported** and **Refuted** claims. The three-class classifier achieves strong performance, with accuracy of 95.4% for Supported and 94.6% for Refuted; however, it still exhibits notable confusion between the two classes, likely due to their semantic proximity and shared evidence patterns. In contrast, the two-step approach by isolating NEI cases early and applying a dedicated binary classifier further improves accuracy to 98.0% for Supported and 96.4% for Refuted. These results support our hypothesis that decomposing the verdict classification task into semantically coherent subtasks enhances the model's precision in detecting factual consistency.

D Hyperparameter and LLM Training Configuration

In this section, we present the detailed hyperparameter settings and training configurations for both our SemViQA models and the Large Language Model (LLM) fine-tuning process. Table 6 consolidates all hyperparameters used across different models, including Binary Classification (BC), Three-Class Classification (TC), Question Answering with Token Classification (QATC), and LLM fine-tuning.

Hyperparameter	**BC**	**TC**	**QATC**	**LLM**
Epochs	20	20	20	1
RT Loss	-	-	✓	-
Cross-Entropy Loss	-	✓	✓	-
Focal Loss	✓	-	-	-
Learning Rate	$1e^{-5}$	$1e^{-5}$	$2e^{-6}$	$5e^{-5}$
Batch Size	104	104	36	2
Gradient Accumulation	1	1	2	1
Optimizer (AdamW)	✓	✓	✓	✓
Max Token Length	256	256	512	4096
GPUs	A100	A100	A100	A100
Zero	-	-	-	Zero3
LR Schedule	Linear	Linear	Cyclic	Cosine
Mixed Precision	-	-	-	bf16

Table 6: Consolidated hyperparameter and training configuration for SemViQA models and LLM fine-tuning.

We fine-tune a Large Language Model (LLM) using a restructured version of the original datasets, ViWikiFC and ISE-DSC01, as detailed in Figure 7. These datasets have been carefully adapted for training to improve performance and ensure compatibility with our model. For training, we utilize the official Qwen LLM implementation from the QwenLM repository[5]. Our training setup follows the full configuration outlined in Table 6, ensuring optimal efficiency and alignment with best practices.

[5]https://github.com/QwenLM/Qwen

Question: You are tasked with verifying the correctness of the following statement.
- We provide you with a claim and a context. Please classify the claim into one of three labels: "Supported", "REFUTED", or "NEI" (Not Enough Info).
- Your answer should include the classification label and the most relevant evidence sentence from the context.
- Remember, the evidence must be a full sentence, not part of a sentence or less than one sentence.

Given a claim and context as follows:
Context: *The actress revealed her secrets to maintaining a youthful appearance as follows: Eating three balanced meals a day. For dinner, Ivy Chen usually eats early to ensure her body has enough time to digest food, metabolize energy, and avoid putting pressure on the stomach and other organs. A recent study published in *Frontiers in Nutrition* suggests that eating dinner earlier can lead to a longer lifespan, with the ideal time being 7 PM. If this is not possible, experts recommend having the last meal of the day 2-3 hours before bedtime. Drinking ginger tea: To keep her body warm, promote blood circulation, and enhance circulation, Ivy Chen drinks ginger tea daily. Her ginger tea is typically made with ground ginger, black tea, turmeric powder, and brown sugar. This drink is a natural remedy that not only boosts the immune system and reduces inflammation but also fights oxidation, supports weight loss, improves skin health, and helps maintain a youthful look. Regular exercise: Ivy Chen is a fitness enthusiast who loves physical activities and exercises daily, even during pregnancy. The Taiwanese actress shared that if she is not busy with work, she runs for at least 30 minutes every day. Even when traveling abroad, she maintains her running habit. A recent study published in *Progress in Cardiovascular Disease* found that regular runners live three years longer than non-runners. Running significantly helps with weight loss, maintaining a balanced physique, toning muscles, relaxing the mind, and benefiting heart health. Besides running, Ivy Chen also swims, practices yoga, and hikes to maintain physical fitness and endurance. Skincare: Regarding her skincare routine, the actress emphasized the importance of hydration. The Taiwanese beauty revealed that she always carries a facial mist to ensure her skin stays hydrated while outdoors.*

Claim: *Even when traveling abroad, Ivy Chen maintains her running habit.*

Answer: This claim is classified as **Supported**. The evidence is: *Even when traveling abroad, she maintains her running habit.*

Table 7: Example of a fact-checking task prompt used for LLM training. Note: Some parts of the Context and Claim were originally in Vietnamese. In this paper, we have translated them into English for better readability. Sentences highlighted in blue indicate the evidence.

We present the complete training progress of the LLM models and QATC in Figure 9 and Figure 8, respectively. Figure 9 illustrates the training dynamics of Qwen 1.5B and Qwen 3B, supporting the results presented in Table 2. Notably, the Qwen 1.5B model demonstrates more stable training dynamics compared to the Qwen 3B model during the initial stage. Meanwhile, Figure 8 showcases the completion of QATC training, depicting the loss curves of $\text{ViMRC}_{\text{large}}$ and $\text{InfoXLM}_{\text{large}}$. These results highlight the convergence behavior of QATC training across different architectures, further supporting the robustness of our approach.

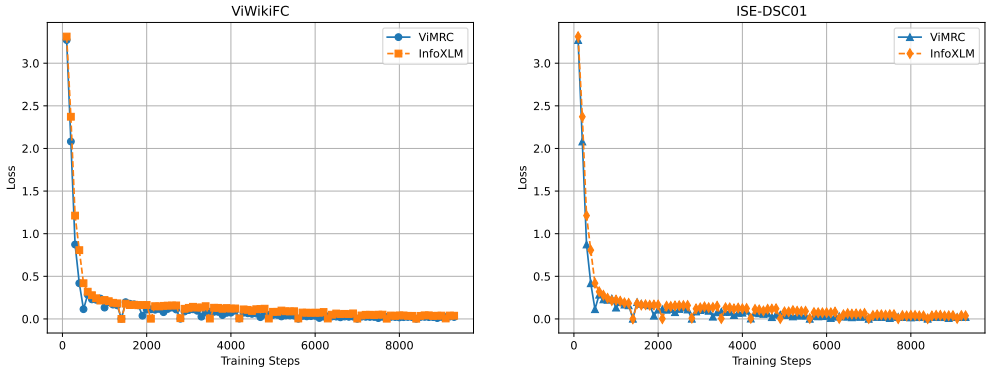

Figure 8: Training progress of the ViMRC$_{\text{large}}$ and InfoXLM$_{\text{large}}$ models.

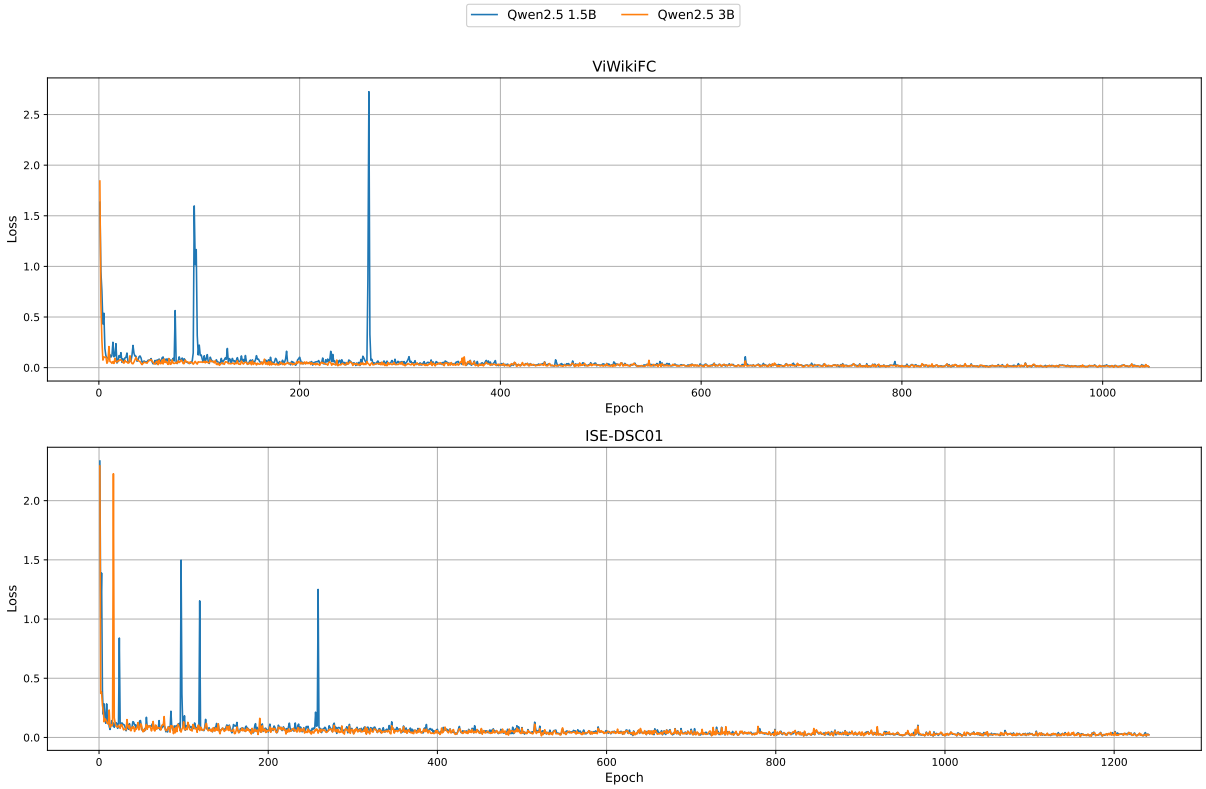

Figure 9: Training progress of the Qwen 1.5B and Qwen 3B models.

E Comparison of TF-IDF and QATC in Fact-Checking: Examples of Incorrect vs. Correct Evidence Selection

Claim	Evidence	TF-IDF	QATC
Du lịch Triều Tiên là điều mà chỉ có một số người được đi đến. (Traveling to North Korea is something only a few people can do.)	Theo nguyên tắc, bất kỳ ai cũng được phép du lịch tới Triều Tiên, và những ai có thể hoàn thành quá trình làm thủ tục thì đều không bị Triều Tiên từ chối cho nhập cảnh. (In principle, anyone is allowed to travel to North Korea, and those who complete the process are not denied entry.)	Khách du lịch không được đi thăm thú bên ngoài vùng đã được cho phép trước mà không được hướng dẫn viên người Triều Tiên cho phép nhằm tránh các điệp viên nằm vùng. (Tourists are not allowed to visit areas outside of the designated zones without a North Korean guide to prevent undercover spies.)	Theo nguyên tắc, bất kỳ ai cũng được phép du lịch tới Triều Tiên, và những ai có thể hoàn thành quá trình làm thủ tục thì đều không bị Triều Tiên từ chối cho nhập cảnh. (In principle, anyone is allowed to travel to North Korea, and those who complete the process are not denied entry.)
Nó có độ nóng chảy ở mức gần 30 độ C. (It has a melting point of about 30°C.)	Nó là một kim loại kiềm mềm, màu bạc, và với điểm nóng chảy là 28 °C (83 °F) khiến cho nó trở thành một trong các kim loại ở dạng lỏng tại hay gần nhiệt độ phòng. (It is a soft, silvery alkali metal with a melting point of 28°C (83°F), making it one of the metals that is liquid at or near room temperature.)	Nó là nguyên tố có độ âm điện thấp thứ hai sau franci, và chỉ có một đồng vị bền là caesi-133. (It is the second least electronegative element after francium, and has only one stable isotope, cesium-133.)	Nó là một kim loại kiềm mềm, màu bạc, và với điểm nóng chảy là 28 °C (83 °F) khiến cho nó trở thành một trong các kim loại ở dạng lỏng tại hay gần nhiệt độ phòng. (It is a soft, silvery alkali metal with a melting point of 28°C (83°F), making it one of the metals that is liquid at or near room temperature.)

Table 8: Comparison of TF-IDF and QATC in Fact-Checking: TF-IDF selects irrelevant evidence (Incorrect), while QATC selects accurate evidence (Correct).