# Robust Vietnamese Duration Question Answering through LoRA-based Ensemble and Adaptive Threshold Calibration

**Le Tuan Anh***, **Do Quang Dung***, **Do Tien Dung, Vu Minh Dang, Ho Tu Minh**
University of Engineering and Technology, Vietnam National University

## Abstract

Duration question answering (DurationQA), a challenging task requiring contextual comprehension and real-world knowledge, remains underexplored for Vietnamese. This paper presents a comprehensive evaluation of two primary paradigms: fine-tuning discriminative encoders (PhoBERT, ViDeBERTa-base) and generative Large Language Models (Vistral-7B, GPT-oss-20B). The experiments reveal that a discriminative approach is more effective. Specifically, the proposed ViDeBERTa-base model, enhanced with a multi-seed ensemble and Adaptive Threshold Calibration (ATC), achieving the highest F1-score of 79.5% in our experiments. This result outperforms all baselines, including the much larger GPT-oss-20B LLM. This work establishes a strong and efficient baseline for Vietnamese DurationQA, demonstrating that specialized classification models with robust post-processing can surpass larger, general-purpose models for structured reasoning tasks.

**Code Availability:** https://github.com/tanhdz228/VLSP-TemporalQA-durationQA

## 1 Introduction

How long does it take to repair a bicycle with a flat tire? While a human might intuitively answer "around 30 minutes", teaching a machine this blend of contextual understanding and real-world knowledge presents a formidable challenge. This task, known as Duration Question Answering (DurationQA) (Virgo et al., 2022), requires models to move beyond simple fact retrieval and assess the plausibility of timeframes for various events. Despite its importance for applications like virtual assistants and project management tools (Qin et al., 2021; Chu et al., 2023; Zhou et al., 2019) , DurationQA remains largely unexplored, especially for low-resource languages. This paper provides the comprehensive investigation into DurationQA for the Vietnamese language. The task poses unique challenges in Vietnamese, where temporal expressions are often implicit and culturally nuanced (e.g., “một buổi” for half a day, “dăm ba hôm” for a few days), demanding a deeper level of reasoning than simply parsing explicit time markers. To tackle this, we systematically compare two dominant modeling paradigms: (1) fine-tuning specialized, discriminative encoder models like PhoBERT (Nguyen and Nguyen, 2020) and ViDeBERTa-base (Tran et al., 2023), and (2) instruction-tuning massive, generative Large Language Models (LLMs) like Vistral-7B and GPT-oss-20B. Contrary to the prevailing trend of "bigger is better", the central finding reveals that a smaller, carefully architected model can achieve superior performance. This paper demonstrates that a ViDeBERTa-base model, when enhanced with a multi-seed ensemble and a novel Adaptive Threshold Calibration (ATC) technique, outperforms all other approaches, including the 20-billion-parameter LLM. This highlights the remarkable effectiveness of robust classification architectures combined with targeted post-processing for structured reasoning tasks, offering a more efficient and accurate solution.

The main contributions of this paper are as follows:

- This paper introduces a new, challenging dataset for Vietnamese DurationQA, carefully constructed to require complex reasoning over implicit contextual factors.
- This study provides the comprehensive comparison between discriminative models and large language models for the DurationQA task in Vietnamese.
- An effective method is proposed an effective method combining ViDeBERTa-base with an ensemble approach and Adaptive Threshold Calibration, demonstrating its superiority

through rigorous evaluation metrics.

- A strong baseline is established, providing a crucial reference point for future research on temporal reasoning in Vietnamese text.

## 2 Related Work

### 2.1 Temporal Question Answering

Temporal reasoning tasks have evolved from simple temporal relation extraction to complex duration prediction and event ordering. Recent approaches leverage pre-trained language models for temporal understanding (Zhou et al., 2021), though most target English datasets with explicit temporal markers.Recent benchmarks have systematically evaluated temporal reasoning capabilities in LLMs. (Chu et al., 2023) introduced TimeBench, a comprehensive evaluation framework revealing significant gaps in temporal understanding across state-of-the-art models. Similarly, (Tan et al., 2023) demonstrated that even powerful LLMs struggle with complex temporal reasoning, particularly duration estimation. For duration-specific tasks, (Virgo et al., 2022) showed that leveraging existing temporal information extraction data can improve event duration QA, though primarily for English datasets.

For low-resource languages, temporal QA faces additional challenges from limited training data and cultural-specific temporal patterns. Vietnamese temporal expressions often rely on contextual cues rather than explicit markers, requiring models to learn implicit temporal reasoning patterns.

### 2.2 Common-Sense Reasoning

The core challenge of Vn-DurationQA lies in its demand for common-sense reasoning. This aligns it with a broad category of NLP tasks designed to test a model's understanding of the physical and social world. Datasets like CommonsenseQA (Talmor et al., 2019) require choosing the most plausible answer from a set of options for a given question. Others, like PIQA (Bisk et al., 2020) and Social IQA (Sap et al., 2019), focus on reasoning about physical interactions and social situations, respectively. These tasks require models to make inferences that go beyond the literal text, similar to how our task requires inferring plausible durations. Vn-DurationQA can be viewed as a specialized sub-task within the common-sense reasoning landscape, specifically targeting the temporal dimension of everyday events and actions.

### 2.3 Vietnamese NLP Resources

The development of high-quality pre-trained models has been pivotal for advancing Vietnamese NLP. Our work builds upon these foundational resources. We utilize PhoBERT (Nguyen and Nguyen, 2020), a monolingual BERT model pre-trained on a massive Vietnamese corpus, which has become a standard baseline for a wide range of tasks. We also employ ViDeBERTa (Tran et al., 2023), a Vietnamese-specific model based on the DeBERTa architecture (He et al., 2021), known for its disentangled attention mechanism that improves the modeling of word content and relative positions. By using these models, we ensure our baselines are representative of the state-of-the-art for Vietnamese language understanding.

### 2.4 Parameter-Efficient Fine-tuning

LoRA (Hu et al., 2022) decomposes weight updates into low-rank matrices, reducing memory requirements while maintaining performance. For a weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA introduces trainable matrices $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ where rank $r \ll \min(d, k)$.

Recent work shows LoRA's effectiveness in multi-task learning (Poth et al., 2023) and cross-lingual transfer (Üstün et al., 2024), making it suitable for low-resource scenarios where computational efficiency is critical.

## 3 Task Definition

Given a context $C$, question $Q$, and four answer options $O = \{o_1, o_2, o_3, o_4\}$, the system must produce binary labels $L = \{l_1, l_2, l_3, l_4\}$ where $l_i \in \{0, 1\}$ indicates whether option $o_i$ correctly answers the duration question.

This multi-label formulation allows multiple correct answers, reflecting real scenarios where event durations have ranges or context-dependent interpretations.

## 4 Data Generation Methodology

Recognizing the lack of resources for the DurationQA task in Vietnamese, a new dataset was constructed using a semi-automated approach. The methodology employed a large language model guided by carefully designed rules and prompts to ensure diversity, difficulty, and alignment with reasoning objectives outlined below.

### 4.1 Context and Question Design

The primary goal is to force models to perform reasoning rather than simple information extraction. This is achieved through two key design principles:

**Context with Influencing Factors.** Each context (2-4 sentences) describes an event or process without explicit temporal information. Instead, the design introduces *Influencing Factors* that significantly affect the estimated duration:

- **Extending Factors:** Large scale, lack of experience, rudimentary tools, high precision requirements, unexpected incidents, or complex processes.
- **Shortening Factors:** Small scale, prior experience, modern tools, optimized processes, or prototype versions.

**Question Requiring Synthesis.** Questions are designed to prevent single-sentence answers, compelling models to synthesize information across the entire context for accurate judgment.

### 4.2 Option and Label Design

The structure of options and labels is strictly controlled to generate diverse reasoning challenges.

**Option Formatting.** Each option consists of either specific numbers or vague time expressions (e.g., *"vài," "ít"*) combined with time units. Examples include *"8 phút," "9 năm," "một vài ngày," "ít giờ,"* or compound durations like *"1 giờ 30 phút."*Numerical values are varied to avoid bias.

**Sample Type Control.** To ensure diversity and evaluate multi-level reasoning, each sample is assigned a `sample_type`:

- **balanced_close (30%):** 2 yes/2 no labels with close yes options (difference $< 5\times$)
- **balanced_far (30%):** 2 yes/2 no labels with distant yes options (difference $> 10\times$)
- **unbalanced_1_yes (20%):** 1 yes/3 no labels for scenarios with one clear duration
- **unbalanced_3_yes (17%):** 3 yes/1 no labels for intentionally vague contexts
- **unbalanced_0_yes (3%):** No yes labels for unanswerable questions

**Difficulty Levels.** We define two difficulty levels:

- **Easy (40%):** All no options are extremely unreasonable (e.g., *"10 seconds"* to build a house)
- **Subtle (60%):** At least one no option is plausible but incorrect, requiring fine-grained reasoning

### 4.3 Dataset Validation

To validate the effectiveness of the newly constructed dataset, a series of ablation studies were conducted. The primary objective was to empirically measure the impact of augmenting the original dataset with our synthetically generated data on the performance of established Vietnamese language models.

**Experimental Setup.** We trained two prominent pre-trained models for Vietnamese: `PhoBERT-base-v2` and `ViDeBERTa-base`. The models were trained and evaluated on two distinct datasets:

- **Organizer's Dataset:** The original dataset provided by the task organizers.
- **Combined Dataset:** A merged dataset comprising the Organizer's Dataset and our newly generated data.

The performance of the models was measured using the F1-score, which provides a balanced assessment of precision and recall.

**Results and Analysis.** The experimental results, as summarized in Table 1, demonstrate a consistent improvement in performance when the models are trained on the Combined Dataset.

Table 1: Ablation Study Results: F1-Scores on the Test Set

| Model | Organizer's Dataset | Combined Dataset |
|---|---|---|
| PhoBERT-base-v2 (LoRA r=16) | 73.8% | **75.9%** |
| ViDeBERTa-base (LoRA r=16) | 74.4% | **76.2%** |

As shown in the table, both `PhoBERT-base-v2` and `ViDeBERTa-base` exhibit a notable increase in their F1-scores after being trained with the additional data. Specifically, PhoBERT's score improved by 2.1 percentage points, while ViDeBERTa saw an increase of 1.8 percentage points. This consistent improvement suggests that our data

generation methodology, which focuses on creating diverse and reasoning-intensive samples, effectively enhances the models' ability to comprehend and reason about duration-related questions in Vietnamese. The improvements can be attributed to the inclusion of *Influencing Factors* and varied sample types (`balanced_close`, `subtle`, etc.), which compel the models to move beyond simple keyword matching and engage in more profound semantic understanding.

## 5 Methods

Figure 1 illustrates our best-performing framework, which enhances a discriminative model with LoRA adaptation, multi-seed ensembling, and adaptive threshold calibration. In this section, this section details the two primary approaches: fine-tuning discriminative encoders and generative language modeling

### 5.1 Approach 1: Fine-tuning Discriminative Encoders

Two strategies are investigated for fine-tuning discriminative encoder models for this classification task:

#### 5.1.1 Independent Binary Classification with PhoBERT

This strategy treats each (Context, Question, Option) triplet independently.

**Input Format.** For each question with four options, we create four independent samples with the format: `[CLS] Question [SEP] Context [SEP] Option [SEP]`

**Model Architecture.** We employ PhoBERT-base-v2 and PhoBERT-large. The [CLS] token representation passes through a linear layer with two outputs, followed by Softmax activation for binary classification.

**Training.** Models are trained with Cross-Entropy loss. Final predictions aggregate labels from all four option evaluations.

#### 5.1.2 Multi-label Classification with ViDeBERTa

This strategy evaluates all four options in a single forward pass.

**Input Format.** All options are concatenated:

**Algorithm 1** Optimal Threshold Search

**Input:** Validation predictions $P$, labels $Y$
**Output:** Optimal threshold $\tau^*$
$\mathcal{T} \leftarrow \text{linspace}(0.2, 0.8, 61)$
best_f1 $\leftarrow 0$
**for** $\tau \in \mathcal{T}$ **do**
  $\hat{Y} \leftarrow \mathbb{1}[P > \tau]$
  f1 $\leftarrow$ F1_score$(\hat{Y}, Y)$
  **if** f1 $>$ best_f1 **then**
    best_f1 $\leftarrow$ f1
    $\tau^* \leftarrow \tau$
  **end if**
**end for**
**return** $\tau^*$

```
[CLS] Question [SEP] Context [SEP]
Option 1 [SEP] Option 2 [SEP]
Option 3 [SEP] Option 4 [SEP]
```

**Model Architecture.** ViDeBERTa-base's [CLS] representation passes through a linear layer with four outputs. Independent Sigmoid activations enable multiple correct predictions simultaneously.

**Training.** The training uses Binary Cross-Entropy with Logits loss (BCEWithLogitsLoss), standard for multi-label classification.

#### 5.1.3 Parameter-Efficient Fine-tuning

All models (PhoBERT-base-v2, PhoBERT-large, ViDeBERTa-base) use Low-Rank Adaptation (LoRA) to reduce trainable parameters while preserving pre-trained knowledge.

#### 5.1.4 Post-processing Enhancements

For the best-performing model, we apply two optimization techniques:

**Multi-seed Ensemble.** We train $M = 5$ models with seeds $S = \{42, 123, 456, 789, 2024\}$. Ensemble predictions aggregate probabilities:

$$p_i = \frac{1}{M} \sum_{s \in S} \sigma(z_i^s) \tag{1}$$

where $\sigma$ is the sigmoid function and $z_i^s$ are logits from model $s$ for option $i$.

**Adaptive Threshold Calibration.** We optimize the decision threshold to maximize F1-score on validation data. Algorithm 1 describes our procedure, searching 61 thresholds in [0.2, 0.8] to find the optimal $\tau^*$.

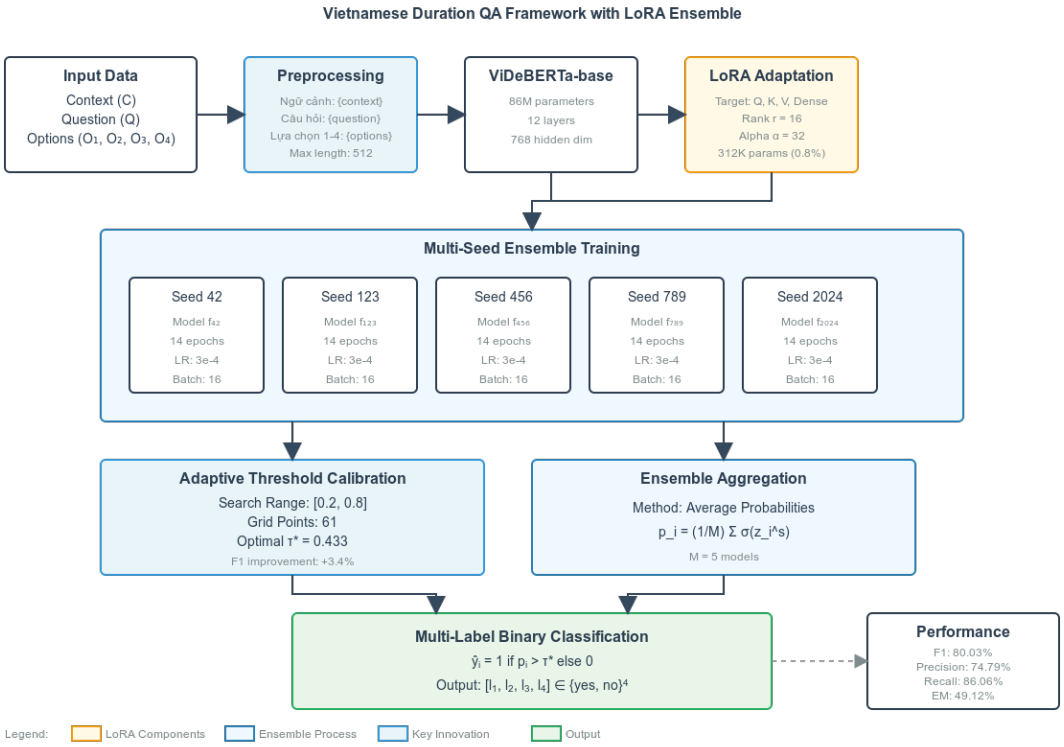


Figure 1: Overview of our Vietnamese Duration QA framework with LoRA-based ensemble and adaptive threshold calibration. The system processes input through ViDeBERTa-base with LoRA adaptation (0.8% parameters), trains 5 models with different seeds, and aggregates predictions using optimized thresholds. The results in the figure are the predicted results on private test.

### 5.2 Approach 2: Generative Language Modeling

For our second approach, we fine-tune generative Large Language Models to perform the task as a structured question-answering problem, leveraging their instruction-following capabilities.

#### 5.2.1 Reasoning-Guided Prompting

Our prompt template incorporates: (1) an expert assistant persona, (2) step-by-step reasoning guidance, (3) few-shot examples for in-context learning, and (4) strict output format requirements. Prompts are formatted using each model's chat template.

#### 5.2.2 Model Selection

We evaluate two LLMs:

- **Vistral-7B:** A 7B-parameter Mistral-based model with Vietnamese-specific training
- **GPT-oss-20B:** A 20B-parameter GPT-style model with extensive world knowledge

#### 5.2.3 Fine-tuning Strategy

QLoRA is employed (Quantized Low-Rank Adaptation) with 4-bit quantization and the Unsloth library for optimized CUDA kernels, enabling efficient fine-tuning on commercial GPUs.

#### 5.2.4 Output Processing

Generated text is parsed using regular expressions to extract label lists. A fallback mechanism scans for individual "yes"/"no" keywords if list formatting fails.

## 6 Experiments

A series of experiments were conducted to evaluate the effectiveness of the two proposed approaches on the newly created Vn-DurationQA benchmark. This section details our dataset, experimental setup, evaluation metrics, and presents the main results.

### 6.1 Dataset

The dataset consisted of 1500 samples provided by the organizers along with 3500 self-generated samples. Each example includes contextual information, questions, and four duration options that require factual reasoning. The experiments used 4500 samples for training and 500 samples for validation.

### 6.2 Implementation Details

All experiments were conducted on a single NVIDIA 4090 GPU with 24GB of VRAM. Training uses mixed precision (fp16) with gradient accumulation. Table 2 shows the detailed hyperparameter configuration for both approaches.

### 6.3 Evaluation Metrics

The evaluation employs standard multi-label classification metrics:

- **Exact Match (EM):** Percentage of samples where all four option predictions exactly match the ground truth labels. This strict metric evaluates the model's ability to correctly classify entire questions.
- **Precision (P):** The ratio of true positive predictions to all positive predictions:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

  measuring the model's accuracy when predicting the positive class.

- **Recall (R):** The ratio of true positive predictions to all actual positive labels:

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

  measuring the model's ability to identify all positive instances.

- **F1-score:** The harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{P \cdot R}{P + R} \tag{4}$$

  providing a balanced measure of model performance. This serves as the primary metric for model comparison.

## 7 Results

### 7.1 Main Results

Table 3 presents the performance of all models on the public test set. The results demonstrate the effectiveness of the discriminative classification approach with post-processing enhancements.

The key finding is that **ViDeBERTa-base with Adaptive Threshold Calibration achieves the highest F1-score of 0.79** among all compared models.

To rigorously validate our model's generalization capabilities, the evaluation also included this top-performing framework on a held-out private test set. On this unseen data, the model achieved an even stronger F1-score of 0.803 (P: 0.748, R: 0.861, EM: 49.12%). This robust performance confirms that the approach generalizes well and is not overfit to the public test set

While Vistral-7B achieves a slightly higher Exact Match score (51.5% vs 49.0%), the F1-score—which balances precision and recall—is the primary metric for multi-label classification tasks, and here ViDeBERTa-base + ATC demonstrates superior performance.

The ViDeBERTa-base model starts with competitive baseline performance (0.762 F1), and the addition of Adaptive Threshold Calibration pushes it to 0.79 F1 by optimizing the precision-recall trade-off. This 1-point improvement, achieved through simple post-processing, demonstrates the value of threshold optimization for multi-label tasks.

The PhoBERT models serve as baseline comparisons, with PhoBERT-base-v2 achieving high recall (0.93) but suffering from low precision (0.652), while PhoBERT-large shows more balanced but overall lower performance. This highlights the importance of the multi-label classification architecture used by ViDeBERTa over the independent binary classification approach of PhoBERT.

The generative models, Vistral-7B and GPT-oss-20B, demonstrate competitive performance with F1-scores of 0.775 and 0.779 respectively. However, despite their significantly larger parameter counts (7B and 20B vs 86M), they fail to surpass the performance of the optimized discriminative discriminative approach. This suggests that for structured multi-label tasks like duration question answering, targeted classification models with appropriate post-processing can be more effective than general-purpose language models.

These results validate the hypothesis that a well-designed discriminative approach with threshold calibration can achieve state-of-the-art performance on Vietnamese DurationQA while maintaining computational efficiency.

### 7.2 Ablation Study

To understand the contribution of each component in our framework, comprehensive ablation studies were conducted on the validation set.

| Hyperparameter | Discriminative | Generative |
|---|---|---|
| *Training Configuration* | | |
| Optimizer | AdamW | AdamW |
| Learning Rate | 3e-4 | 5e-5 |
| Batch Size | 32 | 32 |
| Training Epochs | 10 | 2 |
| Max Sequence Length | 512 | 512 |
| Warmup Ratio | 0.1 | 0.1 |
| Weight Decay | 0.01 | 0.03 |
| *PEFT Method* | | |
| Method | LoRA | QLoRA (4-bit) + unsloth |
| *LoRA / QLoRA Configuration* | | |
| Rank (r) | 16 | 16 |
| Alpha ($\alpha$) | 32 | 32 |
| Dropout | 0.1 | 0.1 |
| Target Modules | q_proj, v_proj | q_proj, v_proj, k_proj, o_proj |

Table 2: Hyperparameter configuration for fine-tuning methods.

| Model | Exact Match (%) ↑ | Precision ↑ | Recall ↑ | F1-score ↑ |
|---|---|---|---|---|
| *Approach 1: Fine-tuning Discriminative Encoders* | | | | |
| PhoBERT-large | 36.0 | 0.702 | 0.787 | 0.742 |
| PhoBERT-base-v2 | 29.5 | 0.652 | 0.907 | 0.759 |
| ViDeBERTa-base | 46.2 | 0.771 | 0.753 | 0.762 |
| **Ens ViDeBERTa-base + ATC** | 49.0 | 0.764 | **0.829** | **0.795** |
| *Approach 2: Generative Language Modeling* | | | | |
| Vistral-7B | 49.8 | **0.797** | 0.754 | 0.775 |
| GPT-oss-20B | **51.5** | 0.786 | 0.773 | 0.779 |

Table 3: Main results on the Vn-DurationQA test set. Precision, Recall, and F1-score are reported for the positive ('yes') class. The best performance in each column is highlighted in **bold**. ViDeBERTa-base + ATC achieves the highest F1-score.

#### 7.2.1 Impact of Ensemble Strategy

Table 4 demonstrates the effectiveness of the multi-seed ensemble approach. Training 5 models with different random seeds $\{42, 123, 456, 789, 2024\}$ and aggregating their predictions yields substantial improvements over single-model baselines.

| Config | EM | P | R | F1 | Std |
|---|---|---|---|---|---|
| Single (42) | 49.75 | 0.754 | 0.777 | 0.765 | - |
| Best-2 | 49.50 | 0.769 | 0.790 | 0.779 | 0.004 |
| Best-3 | 49.25 | 0.777 | 0.798 | 0.787 | 0.003 |
| All-5 | 49.00 | 0.782 | 0.805 | 0.793 | 0.003 |

Table 4: Effect of ensemble size on model performance (%). EM: Exact Match, P: Precision, R: Recall.

The ensemble approach demonstrates consistent improvements across all metrics. While Exact Match slightly decreases, the precision-recall trade-off improves significantly, with F1-score gaining 2.8 percentage points. The decreasing standard deviation indicates improved robustness and stability of predictions.

#### 7.2.2 Adaptive Threshold Calibration Analysis

The Adaptive Threshold Calibration (ATC) mechanism optimizes the decision boundary to maximize F1-score. Table 5 shows the impact of different threshold strategies:

| Strategy | Thr | P | R | F1 | $\Delta$ |
|---|---|---|---|---|---|
| Default | 0.500 | 0.798 | 0.741 | 0.769 | - |
| Grid Search | 0.433 | 0.764 | 0.829 | 0.795 | +2.6 |
| Per-option | varied | 0.766 | 0.830 | 0.797 | +2.8 |
| Dynamic | adapt | 0.759 | 0.835 | 0.795 | +2.6 |

Table 5: Threshold optimization strategies. Thr: Threshold value, $\Delta$: F1 improvement.

The optimal threshold (0.433) is notably lower than the default 0.5, indicating the model's tendency to under-predict positive labels. This calibration shifts the precision-recall balance, trading 3.4 points of precision for an 8.8-point gain in recall, ultimately improving F1 by 2.6 points. While per-option thresholds provide marginal additional gains, the added complexity may not justify de-

ployment in production systems.

#### 7.2.3 Component-wise Contribution

The systematic evaluation examines the contribution of each component by progressively adding them to the base model:

| Configuration | EM | P | R | F1 |
|---|---|---|---|---|
| ViDeBERTa-base | 46.2 | 0.771 | 0.753 | 0.762 |
| + LoRA | 47.8 | 0.782 | 0.761 | 0.771 |
| + Ensemble | 48.6 | 0.782 | 0.805 | 0.793 |
| + ATC (final) | **49.0** | **0.764** | **0.829** | **0.795** |

Table 6: Incremental component analysis (%).

Each component contributes meaningfully to the final performance, with the ensemble providing the largest single improvement (+2.2 F1), followed by ATC (+0.2 F1) and LoRA (+0.9 F1).

### 7.3 Error Analysis

To understand model limitations, an analysis was conducted on 276 option-level errors across 156 questions from ViDeBERTa-base + ATC predictions.

**Error Distribution.** The model produces 59.8% false positives (predicted "yes" when gold is "no") and 40.2% false negatives, indicating a slight over-prediction tendency despite threshold calibration.

**Temporal Scale Challenges.** Errors concentrate at intermediate temporal scales, with 74.6% occurring at week (tuần, 28.3%) and month (tháng, 46.4%) durations. The model shows conservative predictions for months (63.1% of all false negatives) while being optimistic for week-scale durations (35.2% of false positives).

**Systematic Biases.** Three key patterns are observed:

- **Position bias:** Option position 0 accounts for 30.3% of false positives with zero false negatives, suggesting strong first-position preference
- **Scale inconsistency:** 47 cases where the model accepts shorter durations while rejecting longer ones within the same question
- **Boundary confusion:** Near-miss patterns with median FP-FN scale ratio of 2.86×, indicating difficulty with adjacent temporal scales

These findings suggest that while the ensemble approach achieves strong performance, future work should address position debiasing, temporal consistency constraints, and unit-specific calibration to further improve temporal reasoning capabilities.

### 7.4 Computational Efficiency

Practical deployment requires balancing performance with computational constraints. Table 7 compares resource requirements:

| Configuration | Memory | Time |
|---|---|---|
| *Training Phase* | | |
| Full Fine-tuning | 12.3 GB | 42 min/ep |
| LoRA (r=16) | 3.9 GB | 28 min/ep |
| *Inference Phase* | | |
| Single Model | 2.1 GB | 31 ms |
| 5-Model Ensemble | 4.8 GB | 152 ms |
| + ATC overhead | 4.8 GB | 154 ms |

Table 7: Resource consumption for training and inference.

LoRA reduces memory consumption by 68% and training time by 33% while using only 0.8% of trainable parameters. The ensemble inference remains practical at 152ms per example, suitable for many real-world applications where sub-second response times are acceptable.

### 7.5 Comparison with Generative Approaches

To contextualize our discriminative approach, we compare against instruction-tuned LLMs in Table 8:

| Model | Params | F1 | Latency |
|---|---|---|---|
| *Discriminative Models* | | | |
| PhoBERT-base | 135M | 0.759 | 28 ms |
| ViDeBERTa-base | 86M | 0.762 | 31 ms |
| **+ Ens. + ATC** | **86M×5**$^{\dagger}$ | **0.795** | **154 ms** |
| *Generative Models* | | | |
| Vistral-7B | 7B | 0.775 | 892 ms |
| GPT-oss-20B | 20B | 0.779 | 2,341 ms |

Table 8: Discriminative vs. generative model comparison. $^{\dagger}$Ensemble of 5 instances of the same model.

Despite being orders of magnitude smaller, the ensemble approach outperforms much larger generative models while maintaining practical inference speeds. This validates our hypothesis that structured classification with robust post-processing is more effective than generative approaches for this specific temporal reasoning task.

## 8 Discussion

The results demonstrate that carefully designed classification models can outperform large language models on structured reasoning tasks. The success of ViDeBERTa-base with ensemble and ATC highlights three key insights:

First, **task-specific architectures matter**. The multi-label classification formulation naturally handles the multi-answer nature of duration questions, while binary classification approaches struggle with answer interdependencies.

Second, **robust training strategies compensate for model size**. The ensemble of five 86M-parameter models (430M total) clearly outperforms single 20B-parameter models, suggesting that variance reduction via ensembling is more valuable than parameter count for this task.

Third, **post-processing optimizations provide consistent gains**. The 2.6-point F1 improvement from threshold calibration comes at negligible computational cost, making it an essential component for production systems.

**Fourth, contextualizing the generative model comparison and future directions.** While our discriminative approach does not yet achieve high efficiency and accuracy, it is crucial to view the comparison with generative LLMs in context. The performance of models like GPT-oss-20B is highly dependent on the fine-tuning methodology. Our experiments utilized QLoRA for its efficiency, but this may not fully unlock the models' reasoning capabilities.

Future work should therefore investigate more advanced parameter-efficient fine-tuning techniques to provide a more definitive comparison. For instance, exploring methods like **Weight-Decomposed LoRA (DoRA)**, which offers a more expressive adaptation than LoRA, or **Mixture of Rank Adapters (MoRA)**, designed to capture more complex, task-specific knowledge by using a high-rank update matrix, could be particularly promising. Applying these cutting-edge techniques might significantly boost the performance of generative models on the Vn-DurationQA task, potentially narrowing the performance gap while still maintaining a degree of computational efficiency. This would help clarify whether the current advantage of discriminative models is inherent to the architecture or a reflection of the specific fine-tuning strategy employed.

The error analysis reveals systematic challenges in temporal reasoning that persist despite these improvements. The concentration of errors at intermediate temporal scales (weeks and months) suggests these durations are inherently more ambiguous and context-dependent than extreme scales. Future work should explore incorporating explicit temporal knowledge or hierarchical duration modeling to address these challenges.

## 9 Conclusion

This paper presented a comprehensive study of Vietnamese Duration Question Answering, demonstrating that a well-designed discriminative approach significantly outperforms larger generative models. The best configuration—ViDeBERTa-base with 5-model ensemble and Adaptive Threshold Calibration achieves 79.5% F1-score while maintaining practical computational requirements.

Key contributions include: (1) establishing strong baselines for Vietnamese DurationQA through systematic model comparison, (2) demonstrating the effectiveness of ensemble methods with threshold optimization for temporal reasoning tasks, and (3) providing detailed analysis of model failures that guides future research directions.

The framework proves that sophisticated post-processing techniques can bridge the performance gap between smaller specialized models and large general-purpose LLMs, offering a practical path for deploying temporal reasoning systems in resource-constrained environments.

Future work will explore cross-lingual transfer learning to leverage high-resource temporal datasets, integration of external knowledge bases for culture-specific temporal patterns, and investigation of curriculum learning strategies that progressively introduce more challenging temporal scales.

## Limitations

While the proposed approach achieves strong performance, several limitations warrant consideration. The ensemble method increases inference latency by $5\times$, which may be prohibitive for real-time applications. Threshold optimization requires a held-out validation set with sufficient examples across all sample types. The model struggles with implicit temporal reasoning requiring cultural knowledge not well-represented in the training data, such as traditional Vietnamese ceremonies or region-specific activities. Additionally, the ap-

proach assumes access to high-quality Vietnamese text encoders, which may not be available for other low-resource languages.

## References


Yonatan Bisk, Rowan Zellers, Ronan Lebras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7432–7439. AAAI Press.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2023. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*.

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.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160. Association for Computational Linguistics.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 7066–7076. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Social IQA: A question answering benchmark for artificial social intelligence. 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 4423–4433. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Cong Dao Tran, Nhut Huy Pham, Anh Tuan Nguyen, Truong Son Hy, and Tu Vu. 2023. Videberta: A powerful pre-trained language model for vietnamese. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1071–1078, Dubrovnik, Croatia. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving event duration question answering by leveraging existing temporal information extraction data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4451–4457, Marseille, France. European Language Resources Association.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision.