VNU Journal of Science: Comp. Science & Com. Eng., Vol. 42, No. 3 (2026) 34–53

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

VLSP 2025 Shared Task: Vietnamese Temporal Question Answering (TemporalQA)

Pham Thi Duc*, Ha My Linh, Le Ngoc Toan, Nguyen Thi Minh Huyen

VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam

Received 06th December 2025;
Revised 12th March 2026; Accepted 19th August 2026

Abstract: In 2025, the eleventh edition of the Vietnamese Language and Speech Processing workshop (VLSP 2025) launched its first shared task on Vietnamese Temporal Question Answering (TemporalQA). This task aims to benchmark systems in their ability to comprehend and perform reasoning over temporal information expressed in Vietnamese. TemporalQA is composed of two subtasks. The first, Date Arithmetic (date-arith), evaluates models on questions involving temporal computation, such as adding or subtracting time intervals from a reference date, requiring precise interpretation of temporal expressions within a given context. The second, Duration Question Answering (durationQA), assesses systems on estimating the duration of events or activities based on contextual descriptions. By formalizing these two subtasks and releasing curated datasets, TemporalQA provides the first dedicated benchmark for temporal reasoning in Vietnamese, filling an important resource gap and supporting future research on temporal understanding and reasoning for Vietnamese NLP. The benchmark was hosted on the AIHUB platform, where system rankings were determined using a private test set following a public testing phase. Results show that the leading system for Subtask 1 reached an accuracy of 99%, while the best-performing system in Subtask 2 achieved an F1-score of 81.89% and an Exact Match score of 47.52%.

Keywords: Temporal Reasoning, Question Answering, TemporalQA, VLSP 2025, Vietnamese

1. Introduction

Temporal reasoning refers to a system's ability to understand and manipulate time-related information, including identifying when events occur, determining their durations, and inferring the temporal relationships between them. As a core component of natural language understanding, it enables models to accurately interpret temporal expressions such as dates, durations, and event sequences. Accurate temporal understanding is essential for a wide range of downstream NLP applications, including question answering [1], information extraction [2], event and narrative understanding [3], timeline construction [4], and temporal summarization [5].

*Corresponding author.
E-mail address: ptduc@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6505

Despite its importance, temporal reasoning remains challenging due to the complexity, ambiguity, and context-dependence of temporal expressions in natural language. In recent years, several datasets and benchmarks have been introduced to advance temporal reasoning in English, such as TimeQA [6], TORQUE [7], and McTACO [8]. These resources have facilitated significant progress by providing standardized evaluation protocols and encouraging the development of specialized models for temporal understanding.

To date, research on temporal reasoning for Vietnamese remains limited. Although a few studies have addressed temporal phenomena-for example, Annika Strötgen et al. [9] extended HeidelTime, a rule-based multilingual temporal tagger, to support Vietnamese, and Philippe Lambert et al. [10] conducted a detailed linguistic analysis of Vietnamese morphological expressions centered on hôm-these efforts remain narrow in scope. The lack of large-scale benchmarks and curated datasets constrains both the development and systematic evaluation of Vietnamese temporal reasoning systems, especially for tasks requiring fine-grained temporal inference.

To address this gap, the eleventh Vietnamese Language and Speech Processing Workshop (VLSP 2025) introduced the first shared task on Vietnamese Temporal Question Answering (TemporalQA[1]). The primary goal of this shared task is to promote research on temporal reasoning in Vietnamese by providing a curated dataset, well-defined subtasks, and a unified evaluation framework. TemporalQA consists of two subtasks:

i. Date Arithmetic (date-arith), which evaluates systems' ability to perform temporal computations based on date expressions; and

ii. Duration Question Answering (durationQA), which requires identifying and reasoning about the duration of events or actions given a surrounding context.

The competition was conducted on the AIHUB platform[2], where participating systems were first evaluated on a public test set and subsequently ranked according to their performance on a private test set. In addition to summarizing participating systems, we introduce task-specific baseline models, including a rule-based baseline for Subtask 1 and an encoder-based baseline for Subtask 2, together with prompting-based large language model baselines, to provide reference points for assessing dataset difficulty.

2. Shared Task Description

The Vietnamese Temporal Question Answering (TemporalQA) shared task at VLSP 2025 aims to advance research on processing temporal information in Vietnamese text. The task focuses on developing systems capable of interpreting and computing time-related information through two subtasks: *Date Arithmetic* and *Duration Question Answering*.

2.1. Subtask 1: Date Arithmetic

The *Date Arithmetic* problem represents one of the core challenges in temporal information processing. Unlike tasks that merely identify or extract temporal entities [9, 10], this setting requires systems to interpret complex temporal expressions and perform precise calendar-based computations. In other words, models must not only understand the semantics of phrases such as "2 months and 5 days" but also apply them correctly to a given reference date to produce an exact temporal outcome.

[1] https://vlsp.org.vn/vlsp2025/eval/tempqa

[2] https://aihub.ml/

For instance, given the question "Thời điểm 2 tháng và 5 ngày sau ngày 15 tháng 7 năm 1997 là khi nào? (*When was 2 months and 5 days after July 15, 1997?*)", a system must detect the reference point (*15 tháng 7 năm 1997*), recognize the temporal offset (*2 tháng và 5 ngày*), and carry out the corresponding addition to obtain the correct answer: *19 tháng 9 năm 1997 (September 19, 1997)*. This process involves multiple layers of reasoning: (i) parsing and representing multi-unit temporal expressions, (ii) performing accurate calendar arithmetic-taking into account month lengths, year transitions, and leap-year cases-and (iii) integrating linguistic understanding with quantitative logical operations.

In English, this capability has been systematically evaluated through datasets such as **TimeBench** [11] and the **Date Arithmetic** benchmark [12]. These resources focus on testing a model's ability to manipulate abstract temporal constructs independently of any supporting documents. Each instance typically provides a reference date and a temporal interval, for example: "What is the time 6 years and 4 months after Nov, 1185?", with the expected output "Mar, 1192". Data samples include three fields: the `question`, the `answer`-a normalized date after computation-and an empty `context` field, reflecting the context-free nature of the task. For Vietnamese, we follow the same data structure as TimeBench, for example:

- Question: *Thời gian 1 năm và 2 tháng trước tháng 6, 1297 là khi nào? (What is 1 year and 2 months before June 1297?*
  Expected Answer: *Tháng 4, 1296. (April, 1296)*

Another example involves addition:

- Question: *Ngày 15 tháng 5 năm 2000 sau 10 ngày là khi nào? (What is 10 days after May 15, 2000?)*
  Expected Answer: *Ngày 25 tháng 5 năm 2000. (May 25, 2000.)*

In contrast to context-dependent question answering, *Date Arithmetic* isolates and directly probes a model's ability to perform semantic temporal reasoning. By requiring systems to internalize time-unit relationships and to execute accurate temporal transformations (e.g., adding 5 months to August yields January of the following year), this task serves as a rigorous diagnostic for evaluating fine-grained temporal reasoning. Its relevance extends beyond benchmarking: robust performance in Date Arithmetic is foundational for real-world applications such as intelligent personal assistants, timeline analysis, event forecasting, and any system that must reconcile or compute temporal information with high reliability.

In these examples, the system must correctly handle temporal unit composition (year, month, day), perform arithmetic over mixed units, and output a normalized temporal expression. Errors in parsing or normalization can propagate and cause incorrect results, making this subtask an evaluation of fine-grained temporal reasoning and normalization capabilities.

Each record in the dataset is provided in JSON format. For Date Arithmetic, each entry contains a question and its computed date:

```
{
  "id": "date_001",
  "question": "Thời gian 3 tháng sau
  ↪ tháng 5, 2001 là khi nào? (When is
  ↪ the 3 months after May 2001?)",
  "answer": "Tháng 8, 2001 (August,
  ↪ 2001)"
}
```

2.2. Subtask 2: Duration Question Answering

While Date Arithmetic primarily tests precise temporal computation, the second subtask, Duration Question Answering, evaluates the ability to reason about plausible durations of events in context. Each instance in the dataset consists of a short narrative context, a specific question about the duration of the described

event, and a set of candidate durations. The model's task is to classify each candidate as either "yes" (indicating a plausible duration) or "no" (indicating an implausible duration) based on information explicitly stated in the context, as well as commonsense and real-world temporal knowledge.

For example:

- Context: *Tôi đang sửa chữa chiếc xe đạp bị hỏng.* (I am repairing a broken bicycle.)
  Question: *Mất bao lâu để sửa xong chiếc xe đạp.* (How long does it take to repair the bike?)
  Candidates: [30 phút (30 minutes), 1 tháng (a month), 10 phút (10 minutes), 2 giờ (2 hours)]
  Labels: [yes, no, yes, yes]

In this case, the model must recognize that repairing a bicycle generally requires minutes to a few hours, and that durations of a month are unreasonable for such a task. This requires combining linguistic understanding of the context with commonsense knowledge about everyday activities.

Another example involves longer activities:

- Context: *Họ đang xây một trường đại học mới trong thành phố.* (They are building a new university in the city.)
  Question: *Thời gian thực hiện hành động trên là bao lâu?* (How long does it take to perform the above action?) Candidates: ["3 tháng (3 months)", "1 năm (1 year)", "4 năm (4 years)", "10 năm (10 years)"],
  Labels: ["no", "yes", "yes", "no"]

Here, the model must consider the scale and complexity of the described activity, inferring that bridge construction is a long-term endeavor that typically takes months to years, rather than days or a single week. In some cases, models need to compare or integrate information from multiple sentences or events to assess the plausibility of a duration

DurationQA poses significant challenges compared to more straightforward temporal reasoning tasks. It not only requires understanding the literal meaning of the described events, but also demands contextual reasoning, the ability to leverage commonsense temporal knowledge, and cross-event inference. Unlike Date Arithmetic, which focuses primarily on calculating specific dates from temporal expressions, DurationQA emphasizes plausibility judgment. Models must integrate semantic understanding, typical durations of real-world activities, and contextual cues to determine whether each candidate duration is reasonable.

Each record in the DurationQA dataset is represented in JSON format. A typical entry contains a context, a question, a list of candidate durations, and a corresponding list of binary labels:

```
{
  "id": "dur_001",
  "context": "Tôi đang đọc một cuốn tiểu
  ↪ thuyết dài. (I am reading a long
  ↪ novel.)",
  "question": "Thời gian thực hiện hành
  ↪ động trên là bao lâu? (How long
  ↪ does it take to perform the above
  ↪ action?)",
  "candidates": ["5 phút (5 minutes)", "2
  ↪ giờ (2 hours)", "1 tuần (a week)",
  ↪ "3 tháng (3 months)"],
  "labels": ["no", "yes", "yes", "no"]
}
```

This structured representation allows models to learn to discriminate between plausible and implausible durations, encouraging reasoning that goes beyond surface-level text understanding. By leveraging linguistic, real-world knowledge, and commonsense inference, DurationQA challenges models to mimic human-like temporal reasoning in Vietnamese narratives.

Together, these two subtasks provide a comprehensive benchmark for evaluating the temporal reasoning capabilities of Vietnamese

NLP systems, from precise date calculation to commonsense-informed duration inference.

2.3. Evaluation

To measure system performance on each subtask, we adopt task-specific metrics that reflect both computational accuracy and reasoning ability.

For the **Date Arithmetic** subtask, the primary metric is **Accuracy**, which measures the proportion of correctly computed temporal outputs:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}},$$

where N_{correct} is the number of correctly predicted answers, and N_{total} is the total number of questions.

For the **Duration Question Answering** subtask, evaluation relies on two complementary metrics: **Exact Match (EM)** and **F1-score**. EM measures the percentage of instances where the predicted sequence of binary labels exactly matches the gold labels:

$$\text{EM} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i),$$

where $\hat{y}_i$ and y_i denote the predicted and gold label sequences for instance i, and $\mathbf{1}(\cdot)$ is the indicator function.

The F1-score evaluates the harmonic mean of precision and recall across all predicted labels:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN},$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively. F1 can be computed over all candidate labels across instances to account for multi-label evaluation.

Together, these metrics assess both the correctness of temporal computations and the model's ability to reason about plausible durations in context, providing a comprehensive measure of performance in Vietnamese Temporal Question Answering.

3. Data Preparation

This section outlines the creation of two Vietnamese temporal question answering datasets: **Date-arith**, designed for arithmetic reasoning over dates, and **DurationQA**, focused on estimating event durations. These resources were developed to address the absence of publicly available datasets for temporal reasoning in Vietnamese. Building on the English **TimeBench** benchmark [11], we constructed the datasets through three major stages: translation into Vietnamese, automatic expansion of data, and manual validation. A group of nine native Vietnamese speakers was responsible for the manual validation stage.

3.1. Date-arith: Date Arithmetic QA

Original Dataset. Date-arith is derived from the TimeBench benchmark, which contains questions requiring precise temporal computations, for example, "What is the time 9 years and 1 month after Nov, 1543?". Each entry consists of a question-answer pair without any contextual passage.

Stage 1: Translation to Vietnamese

The first stage involved translating the English questions into natural Vietnamese while maintaining their arithmetic correctness. For example, an English instance:

```
{
  "question": "What is the time 9 year
  ↪  and 1 month after Nov, 1543",
  "answer": ["Dec, 1552"],
  "context": ""
}
```

was converted into Vietnamese with clear, natural phrasing while preserving the intended temporal calculation. This resulted in a set of 4,000 validated Vietnamese question-answer pairs, standardized in a format with `question` and `answer` fields to ensure consistency.

Stage 2: Automatic Data Generation

To enhance linguistic and temporal variety, we generated an additional 5,000 QA pairs automatically. The generation process included:

1. Random sampling of temporal parameters: Start dates were randomly selected, with months between 1 and 12 and years between 1000 and 2000. Temporal shifts were defined as a number of years (0-10) and months (0-12), applied either "before" (*trước*) or "after" (*sau*). The target date was computed by converting months to integers, applying the temporal offset, and converting back into the "Tháng, Năm" format.

2. Template-based question generation: Five Vietnamese templates were designed to capture diverse ways of asking temporal questions:

   Direct question: Ngày tháng nào sẽ là 9 năm 2 tháng trước (sau) tháng 3, 1992? (*What date would be 9 years and 2 months before (or after) March 1992?*)

   Using "khi nào": Thời gian 3 năm 6 tháng trước (sau) tháng 1, 1783 là khi nào? (*When was 3 years and 6 months before (after) January 1783?*)

   Hypothetical: Giả sử bạn đang ở tháng 7, 1894, nếu trôi qua (lùi lại) 8 năm 8 tháng thì là thời điểm nào? (*Assume you are in July 1894; if you go back 8 years and 8 months, what point in time would that be?*)

   Calculation request: Hãy tính thời điểm 8 năm và 8 tháng trước (sau) tháng 7, 1894. (*Calculate the point in time that is 8 years and 8 months before (or after) July 1894*)

   Prediction inquiry: Bạn có thể cho biết thời điểm trước (sau) 8 năm 8 tháng từ tháng 7 năm 1894 là khi nào không? (*Could you tell me what the date is that lies 8 years and 8 months before (or after) July 1894?*)

Each template was instantiated 1,000 times with randomly chosen start dates and temporal offsets, yielding 5,000 automatically generated pairs. Answers were normalized to the format `["Tháng X, Năm Y"]` with consistent capitalization, and all entries were saved in `.jsonl` format.

Stage 3: Manual Validation

All translated and generated QA pairs were manually reviewed to ensure: (i) grammatical correctness, (ii) semantic accuracy, and (iii) correctness of the computed target date.

This step guarantees that the dataset is reliable for training and evaluating models on temporal reasoning tasks.

The final Date-arith dataset contains 9,000 validated instances, split into translated and automatically generated subsets (Table 1).

Table 1. Summary of the Date-arith dataset

Subset	#Samples	#Templates	Source
Translated	4,000	1	TimeBench (EN)
Generated	5,000	5	Automatic synthesis
Total	**9,000**	5	

Data Splits. For the VLSP 2025 shared task, the dataset was partitioned into training, public test, and private test sets, ensuring balanced representation of templates across splits (Table 2).

This dataset provides a robust benchmark for assessing the ability of Vietnamese language models to perform precise temporal arithmetic

Table 2. Train/public/private splits of Date-arith for VLSP 2025

Template	Train	Public	Private
Direct question	624	100	191
Hypothetical	610	97	207
Using "khi nào"	595	112	196
Calculation request	604	94	194
Prediction inquiry	567	97	212
Total	**3,000**	**500**	**1,000**

and establishes a foundation for future research in Vietnamese temporal question answering.

3.2. DurationQA: Duration Question Answering

Data Source. We also developed a Vietnamese version of the DurationQA subset from TimeBench, which evaluates temporal reasoning over event durations. Each English instance contains the fields `context`, `question`, `options`, and `labels`. For example:

```
{
"context": "Drove all the way over from
↪  the highway ... closed at 7.",
"question": "How long did it take to
↪  close?",
"options": ["8 minutes", "almost
↪  instantly", "18 hours", "4 days"],
"labels": ["no", "yes", "no", "no"]
}
```

Phase 1: Translation to Vietnamese.

All textual fields (`context`, `question`, `options`) were automatically translated into Vietnamese, producing an initial set of 687 samples.

Phase 2: Data Augmentation with LLM.

To increase linguistic diversity, we used GPT-4o mini [3], a lightweight model in the GPT-4o family [13], to generate an additional 2,186 Vietnamese samples following the same schema. Each generated sample was manually reviewed to ensure consistency between `options` and `labels` and to guarantee natural Vietnamese phrasing.

Phase 3: Verification.

The manual validation team independently reviewed all 2,873 samples to confirm linguistic fluency and semantic correctness. Any disagreements were resolved through group discussion.

Table 3. Statistics of the DurationQA dataset

Type	Number of Samples	Source
Translated	687	TimeBench (EN)
Generated	2,186	LLM synthesis
Total	**2,873**	

Data Splits. The DurationQA dataset follows a similar splitting strategy as Date-arith. The training set contains 1,490 samples, while the public and private test sets consist of 400 and 625 samples, respectively. Each instance contains a context, a question about event duration, and multiple candidate durations with binary labels. Table 4 summarizes the dataset distribution.

Table 4. Dataset split for the DurationQA subtask for VLSP 2025

Split	Train	Public	Private
Number of samples	1,490	400	625

Table 5 reports the option-level label distribution for DurationQA across the training, public test, and private test splits. The results show a highly balanced distribution between plausible (yes) and implausible (no) durations across all splits, indicating that the dataset does not suffer from label imbalance despite the substantial use of LLM-generated instances. Despite the heavy use of LLM-generated data, the balanced label distribution across all splits and the moderate performance of encoder-based baselines suggest that the dataset does not exhibit trivial artifacts exploitable by surface cues.

[3] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Table 5. Option-level label distribution for DurationQA

Split	Total Options	Yes (%)	No (%)
Training	5,960	2,957 (49.61)	3,003 (50.39)
Public Test	1,600	811 (50.69)	789 (49.31)
Private Test	2,500	1,241 (49.64)	1,259 (50.36)

The public and private test sets were hosted on AIHUB[4] for leaderboard evaluation. Final system rankings were determined based on the hidden private test results.

4. Method

This section provides an overview of the technical methodologies proposed by participating teams in the VLSP 2025 Temporal Question Answering (TemporalQA) shared task, covering its two subtasks: *Date Arithmetic* (Subtask 1) and *Duration Question Answering* (Subtask 2). While both subtasks involve temporal understanding, they differ fundamentally in the types of reasoning required and the strategies employed by the participants.

4.1. Subtask 1: Date Arithmetic

4.1.1. Overview of Methodologies

The first subtask evaluates the ability of models to compute new dates from natural language expressions that describe temporal offsets. Most systems followed a unified architecture comprising temporal parsing, reasoning, and canonical decoding. While some teams relied on prompting techniques to elicit structured reasoning without parameter updates, others fine-tuned language models using synthetic rule-based datasets. A few systems combined neural inference with symbolic solvers to ensure logical accuracy and consistent date normalization. Parameter-efficient tuning techniques such as QLoRA [14] and DoRA [15] were commonly used to adapt large Vietnamese or multilingual models including Qwen3 [16], Vistral-7B [17], and Gemma [18]. Almost all systems integrated a canonicalization module that verified and standardized the predicted date to maintain consistency with the gold format. To provide a comprehensive performance reference for the shared task, a suite of baseline systems-including both proprietary and open-source LLMs-was also evaluated using zero-shot and few-shot prompting strategies. These baselines help contextualize the results of participating teams and highlight the relative difficulty of the task settings.

4.1.2. Baseline Systems

We consider two complementary types of baselines for the Date Arithmetic subtask: prompting-based large language models, which reflect the capabilities of general-purpose neural reasoning systems, and a deterministic rule-based baseline, which serves as a diagnostic upper bound under ideal parsing assumptions.

Prompting-based LLM Baselines. To provide non-specialized reference points for the Date Arithmetic subtask, we implemented prompting-based baselines using instruction-tuned large language models. These baselines assess how well generic LLMs can perform temporal arithmetic in Vietnamese without task-specific training.

Models. We evaluated a compact yet competitive LLM widely used in downstream applications, namely GPT-4o-mini. The model was queried via API under identical experimental settings, using deterministic decoding (temperature = 0) to ensure reproducibility.

Prompting Strategies. Two prompting regimes were employed:

- **Zero-shot Chain-of-Thought (CoT):** The model receives the question and is instructed to reason step by step before giving the final date.

[4] https://aihub.ml/

Zero-shot CoT Prompting

```
prompt_template = """
Question: {question}
Final answer format: "Tháng X, YYYY"
Think step by step.
Before giving the final answer, add the phrase: "Therefore, the answer is".
"""
```

- **Few-shot Chain-of-Thought (CoT):** In this setting, the model is shown two manually crafted exemplars demonstrating typical temporal offset computations (addition and subtraction of years and months). These examples serve as guiding patterns, enabling the model to generalize reasoning steps to unseen questions.

Few-shot CoT Prompting

```
prompt_template = """
Question: 4 years and 1 month after April 2000 is when?
Think step by step.
First, adding 4 years to 2000 gives 2004. Next, adding 1 month to April gives May. Therefore, the answer is Tháng 5, 2004.

Question: 3 years and 4 months before June 1840 is when?
Think step by step.
First, subtracting 3 years from 1840 gives 1837. Next, subtracting 4 months from June gives February. Therefore, the answer is Tháng 2, 1837.

Question: {question}
Final answer format: "Tháng X, YYYY"
Think step by step.
Before giving the final answer, add the phrase: "Therefore, the answer is".
"""
```

Rule-based Baseline. To validate the linguistic realizations and internal consistency of the Date-arith dataset, we implement a deterministic rule-based baseline as a diagnostic reference rather than a competitive system.

The system parses Vietnamese temporal expressions using handcrafted regular-expression patterns to identify the reference date, temporal direction (trước/sau), and offset units (years, months, days). Calendar arithmetic is then performed symbolically using a deterministic solver, followed by normalization into the canonical output format required by the task.

Importantly, this baseline does not rely on any learned parameters or external language models. Its primary purpose is to verify that the translated and rule-generated Vietnamese questions do not introduce unintended linguistic ambiguity that could hinder correct temporal interpretation.

4.1.3. Participants Approaches

Team UIT-NTTT adopted a retrieval-augmented prompting framework built on the LLaMA3.1-8B and LLaMA3.1-70B [19] models to address Vietnamese date arithmetic reasoning. The team first constructed a synthetic corpus by prompting LLaMA3.1-70B with zero-shot chain-of-thought instructions, generating examples that contained both the final answer and detailed reasoning traces. Only samples that were predicted correctly were retained to ensure high-quality supervision. These validated instances were then embedded using the multilingual-e5-large model [20] and indexed in a **Qdrant** vector database [8].

At inference time, the system performed semantic search to retrieve the top-k most relevant synthetic examples for each input question. These retrieved demonstrations were incorporated into a structured few-shot prompt that guided the model through the steps of identifying the base date, extracting the temporal operation, computing the interval, and generating a normalized output. The prompting template

explicitly encouraged consistent chain-of-thought reasoning and enforced a standardized JSON output format.

This design emphasizes contextual retrieval and prompt engineering rather than parameter optimization, enabling strong performance without fine-tuning. The approach shows that retrieval-augmented few-shot prompting is highly effective for Vietnamese temporal arithmetic, particularly in scenarios where training data are limited and linguistic variations require flexible reasoning.

Team HUET developed a multi-stage fine-tuning pipeline for the Qwen3 and Gemma models to perform symbolic date arithmetic in Vietnamese. To enrich training data beyond the official dataset, the team generated approximately 40,000 synthetic samples using a rule-based generator. Each synthetic instance was constructed by randomly selecting a base date, choosing an arithmetic operation (addition or subtraction), and sampling a time interval in months or years. The resulting target date was computed deterministically to ensure label correctness.

To further guarantee data quality, all synthesized samples were verified using the large Qwen3-235B model, which produced normalized JSON outputs describing the base date, operator, and interval. A rule-based comparator was then applied to check consistency between the model output and the gold label.

The team followed an iterative supervised fine-tuning strategy: an initial prompt was refined using Qwen3-235B's reasoning traces, after which smaller Qwen3/Gemma models were fine-tuned and evaluated. Error cases were analyzed and fed back into the prompt design, progressively improving stability and accuracy. In the final configuration, Qwen3 fine-tuned with "thinking-enabled" mode achieved the highest performance, demonstrating the benefit of incorporating intermediate reasoning traces during training.

Team 777 developed a bilingual hybrid architecture that explicitly separates Vietnamese language understanding from symbolic temporal reasoning. The system employs Qwen2.5-1.5B-Instruct as a cross-lingual adapter to normalize Vietnamese questions and convert them into structured English templates encoding the base date and the temporal operation.

The resulting English representation is then processed by an English-trained reasoning model, Flan-T5-base, which had been fine-tuned on a 400k-sample temporal-reasoning corpus. Flan-T5 performs the date calculation and produces the final result in English. This output is subsequently translated back into Vietnamese using the Qwen adapter, ensuring that the answer conforms to the required canonical format.

By decoupling linguistic processing from symbolic computation, the architecture leverages strong English temporal-reasoning capabilities while maintaining seamless compatibility with Vietnamese inputs. This design yields robust cross-lingual generalization and competitive performance without requiring Vietnamese-specific fine-tuning.

Another participating team (**AI5**) developed a neural-symbolic hybrid system built upon the Vistral-7B-iSMART model [17], fine-tuned using the QDoRA framework [15]. QDoRA integrates low-rank adaptation (LoRA) [21] with directional regularization (DoRA) to achieve parameter-efficient training while maintaining stability during optimization.

To enhance data coverage, the team generated additional Vietnamese training examples through a retrieval-augmented generation pipeline, followed by verification with a symbolic solver to ensure logical correctness. During inference, the model produced step-wise reasoning traces, which were subsequently normalized by a canonical decoder enforcing consistent output formatting and basic logical constraints.

Overall, this hybrid architecture was designed to balance flexible neural reasoning with

deterministic computation. The combination improved interpretability, reduced formatting errors, and increased reliability in symbolic temporal reasoning tasks.

4.2. Subtask 2: Duration Question Answering

4.2.1. Overview of Methodologies

The second subtask focuses on assessing the ability of models to judge whether a proposed duration is contextually appropriate for an event. This requires commonsense reasoning rather than arithmetic precision. Most teams adopted large language models fine-tuned with low-rank adapters [21] and trained under task-specific prompting schemes. Approaches can generally be grouped into two categories:

1. *Generative reasoning models*, which produce intermediate explanations before making a binary decision
2. *Discriminative classifiers*, which directly predict the plausibility label based on contextual embeddings.

Enhancements such as retrieval-based prompting, dual-prompt fine-tuning, rationale distillation, and adaptive threshold calibration were employed to strengthen interpretability and prediction robustness.

4.2.2. Baseline Systems

To characterize dataset difficulty and provide reference points for performance comparison, we consider two complementary types of baselines for Subtask 2: prompting-based large language models, which reflect general-purpose commonsense reasoning ability, and a lightweight encoder-based classifier, which serves as a purely neural lower bound.

Prompting-based LLM Baselines. We developed prompting-based baselines using instruction-tuned large language models to assess how well general-purpose LLMs can perform commonsense duration reasoning in Vietnamese without task-specific fine-tuning. These baselines encompass two standard reasoning-oriented prompting strategies, namely zero-shot CoT and few-shot CoT.

In the zero-shot CoT setting, models were instructed to verbalize intermediate reasoning steps before producing the final answer.

Zero-shot CoT Prompting

```
prompt_template = """
Answer the following question, select all possible correct options.
Each question has at least one correct option.
Let's think step by step.
Before the final answer, add: "Therefore, the answer is".
Context: context
Question: question
Options: options Answer:
"""
```

For the few-shot CoT baseline, we constructed a small set of manually curated exemplars covering representative duration reasoning patterns, such as temporal aggregation, subtraction, and unit conversion.

Few-shot CoT Prompting

```
prompt_template = """
Answer the following question, select all possible correct options. Each question has at least one correct option.
Context: actually i have an project on it so please give me as much as you have information about migratory birds in punjab
Question: How long did it take for them to have information about migratory birds in punjab?
Options: A. several months B. 12 weeks
```

C. a few minutes D. almost instantly
Answer: This is a conversation scenario. Providing information happens in real-time and requires very little time. Therefore, the answer is C. a few minutes, D. almost instantly.
Context: Hope she stops laying eggs because she will get really skinny !
Question: How long did it take for her to lay eggs?
Options: A. 1 week B. 22 hours C. 2 years D. 4 years
Answer: Based on commonsense knowledge, birds typically lay eggs within a range of hours to several days. Therefore, the answer is A. 1 week, B. 22 hours.
Let's think step by step.
Before the final answer, add: "Therefore, the answer is".
Context: context
Question: question
Options: options
Answer:
"""

We evaluated two contemporary instruction-tuned LLMs GPT-4o-mini and Gemini-2.0-flash using identical prompting templates to ensure controlled comparisons. All experiments were executed deterministically with temperature set to 0 to minimize output variance.

Encoder-based Baseline (PhoBERT). For Subtask 2, we implement a simple encoder-based baseline using PhoBERT, a pre-trained Vietnamese Transformer model. The task is formulated as an option-level binary classification problem, where each candidate duration is evaluated independently for plausibility.

Given a context c, a question q, and a candidate duration option o, the input sequence is constructed as:

$$[\text{CLS}]\, c\, [\text{SEP}]\, q\, [\text{SEP}]\, o\, [\text{SEP}]$$

The concatenated sequence is encoded by PhoBERT, and the contextualized representation of the `[CLS]` token is passed to a linear classification layer to predict whether the duration is plausible (*yes*) or implausible (*no*).

The model is fine-tuned using binary cross-entropy loss with a batch size of 16 for three epochs. To prevent overfitting, we select the best checkpoint based on the F1 score on the public development set. This baseline intentionally avoids external knowledge sources, rule-based heuristics, or symbolic reasoning, and thus serves as a lower bound for evaluating the inherent difficulty of commonsense duration reasoning in Vietnamese.

Since each option is classified independently, the model does not enforce consistency constraints across options belonging to the same question, which partially explains the observed gap between F1 and exact match (EM) scores.

4.2.3. Participants Approaches

Team **Engineers** proposed a retrieval-guided fine-tuning framework built on Qwen2.5-7B [22] and Qwen3-8B [16]. For each query, semantically similar examples were retrieved using multilingual embeddings and appended to the model input, providing contextual analogs that support more grounded reasoning. The models were fine-tuned using QLoRA [14] under 4-bit quantization to reduce memory footprint while preserving training effectiveness.

At inference time, multiple fine-tuned checkpoints were ensembled through a voting strategy to enhance robustness and mitigate variance across runs. This design integrates retrieval-augmented prompting with parameter-efficient adaptation, yielding a stable and context-aware reasoning pipeline.

Team **Softmind_AIO** proposed the Dual-Prompt Ensemble (DP-Ens) approach built upon Qwen3-4B-Thinking [16]. The system employed two complementary prompting modes: a Chain-of-Thought (CoT) prompt that elicited intermediate reasoning traces, and a Refinement prompt that converted these traces into a final sequence of binary labels.

To enhance the model's reasoning quality, free-form rationales generated by a large teacher model (Gemini 2.0 Flash [23]) were distilled into the student model during fine-tuning, guiding it toward more interpretable and structured inference.

At inference time, predictions from both prompting paths were combined using a log-probability ensemble, enabling the final decision to reflect both the exploratory reasoning of CoT and the stability of refinement-based prediction. This dual-prompt design improved consistency, transparency, and overall robustness in multi-label temporal reasoning.

Team **UIT_BlackCoffee** adopts a task-oriented expert-prompting strategy built on top of the Qwen3-24B model [16], further adapted using LoRA lightweight fine-tuning [21]. To enhance temporal understanding, the model is explicitly instructed to operate as a "temporal reasoning specialist", encouraging systematic, constraint-aware reasoning rather than surface-level pattern matching.

To support this behavior, all training instances are reformulated into an instruction-response format. Each sample embeds the context, question, and duration options within a unified prompt, while the target output is expressed in a normalized JSON structure. This representation standardizes how the model articulates its reasoning steps and final decisions, yielding more interpretable outputs. This design choice aims to improve semantic clarity, reduce label ambiguity, and maintain consistency across different event types and duration categories within the DurationQA task.

Team HUET approached DurationQA using an instruction-based fine-tuning pipeline built on Gemma-3-12B-it. The training data combined the translated McTACO corpus with the official VLSP samples, and all instances were automatically checked using Qwen3-235B to ensure correctness, yielding over fifteen thousand verified examples. Each question was reformulated into an instruction-response format in which the model selected the correct option and provided a brief justification. The model was then fine-tuned under supervised fine-tuning to encourage stable, concise reasoning.

This approach resulted in strong recall and consistent generalization, with most remaining errors arising from ambiguous translations or inherently vague temporal expressions.

Team **AI5** team pursued a discriminative approach based on ViDeBERTa-base [24] enhanced with LoRA [21] adapters. To increase diversity, they constructed an expanded dataset containing more than three thousand additional examples, including complex linguistic patterns such as nested temporal phrases and duration modifiers. The model was fine-tuned for binary classification using contrastive hard negatives [25] to sharpen boundary discrimination between plausible and implausible cases. An adaptive threshold calibration module was introduced during inference to dynamically adjust the decision boundary, and multiple random-seed runs were ensembled to mitigate stochastic effects. This setup highlighted the effectiveness of smaller discriminative models when paired with careful calibration and augmentation

5. Results and Discussion

5.1. Overall Results

Table 6 summarizes the systems submitted by participating teams for both subtasks of the VLSP 2025 TemporalQA shared task. All runs were evaluated on the hidden test set using exact-match accuracy. Overall, the results confirm

that large language models (LLMs) can perform robust temporal reasoning in Vietnamese when coupled with retrieval augmentation, structured prompting, or parameter-efficient fine-tuning.

Across both subtasks, retrieval-augmented and fine-tuned models consistently outperformed pure prompting baselines, highlighting the benefit of targeted adaptation even for large pretrained LLMs. Symbolic verification improved precision and stability in Subtask 1, while dual-prompt and rationale-driven models enhanced interpretability in Subtask 2.

5.2. Subtask 1 - Date Arithmetic

Table 7 presents the baseline and official leaderboard results for Subtask 1 (Date Arithmetic), which evaluates systems' ability to interpret Vietnamese temporal expressions and convert them into normalized calendar dates.

Overall, participating systems achieved consistently strong performance on both the public and private test sets. This suggests that date arithmetic with deterministic ground truth is a tractable problem when temporal expressions are correctly interpreted, while still requiring careful handling of linguistic variation and normalization.

In addition to prompting-based baselines and participating systems, we report results from a deterministic rule-based baseline introduced by the task organizers as a diagnostic reference. Following the shared task regulations, participating teams were required to avoid purely rule-based solutions; accordingly, this system was not part of the official competition ranking. Despite its simplicity, the rule-based baseline achieves 99.60% accuracy on the public test set and 99.00% on the private test set. These results confirm that the Vietnamese linguistic realizations in the Date-arith dataset are largely unambiguous and internally consistent, allowing symbolic temporal parsing and calendar arithmetic to be performed reliably. Importantly, this baseline is not intended as a competitive system, but rather as a validation of dataset quality and temporal expression clarity.

Table 7. Official results for Subtask 1 (Date Arithmetic)

Model	**Public test (%)**	**Private Test (%)**
GPT (Zero)	85.00	87.50
GPT (Few)	83.00	78.80
Rule-based	99.60	99.00
UIT-NTTT	98.00	99.00
HUET	98.00	99.00
777	98.00	99.00
AI5	98.00	99.00

The retrieval-augmented prompting approach developed by **UIT-NTTT** exhibited highly consistent performance by leveraging dynamically selected in-context examples without any parameter updates. **HUET**'s multi-stage fine-tuning pipeline achieved similarly strong results through solver-guided validation and iterative synthetic data expansion. The **777** team demonstrated that bilingual hybrid pipelines can effectively transfer English-centric deterministic reasoning to Vietnamese through a canonicalization layer. Meanwhile, **AI5**'s neural-symbolic architecture combined QDoRA fine-tuning with solver-assisted post-correction to match the highest accuracy while maintaining interpretability.

The GPT-based baseline demonstrated moderate performance, with zero-shot prompting outperforming the few-shot setting. These results indicate that naive prompting remains insufficient for reliably handling Vietnamese temporal expressions, thereby underscoring the advantages of the more structured neural–symbolic approaches adopted by participating teams.

Overall, systems that integrated neural inference with deterministic symbolic correction delivered the most robust and generalizable performance, substantially surpassing the prompting-based baselines.

Table 6. Overview of participating systems and methodological orientations across both subtasks

Team	**Subtask 1 (Date-Arith)**	**Subtask 2 (Duration)**
UIT-NTTT	Retrieval-Augmented Prompting	–
HUET	Iterative Fine-tuning with Synthetic Data	Instruction-based SFT (Gemma-3-12B-it)
777	Bilingual Hybrid (Qwen2.5 + Flan-T5)	–
AI5	Neural-Symbolic Hybrid (Vistral-7B, QDoRA)	Discriminative Model + Calibration
The Engineers	–	Retrieval-Guided QLoRA Fine-tuning
Softmind_AIO	–	Dual-Prompt Ensemble (Qwen3-4B-Thinking)
UIT_BlackCoffee	–	LoRA Fine-tuning + Expert Prompting

5.3. Subtask 2 - Duration Question Answering

This subtask evaluates whether systems can assess the plausibility of event durations-an ability grounded primarily in commonsense temporal reasoning rather than explicit symbolic computation. Table 8 presents the results of three types of baselines for Subtask 2, including a neural encoder baseline (PhoBERT), prompting-based large language model baselines (GPT and Gemini), and the official leaderboard scores.

Table 8. Official private test results for Subtask 2 (Duration QA)

Model	**F_1 (%)**	**EM (%)**
phoBert	63.93	57.84
GPT (Zero)	92.90	92.32
GPT (Few)	93.27	92.64
Gemini (Zero)	90.24	89.92
The Engineers	**81.89**	47.52
UIT_BlackCoffee	80.13	42.72
AI5	80.03	**49.12**
HUET	79.97	40.32
Softmind_AIO	79.06	34.08

The PhoBERT baseline serves as a purely neural lower bound that does not leverage external commonsense knowledge or prompting, and its substantially lower performance highlights the challenge of duration plausibility assessment.

Among the prompting-based baselines, few-shot chain-of-thought prompting consistently outperformed zero-shot prompting for GPT, confirming that structured demonstrations improve duration reasoning. For Gemini-2.0-Flash, only the zero-shot setting was evaluated, and its performance was slightly lower than GPT's zero-shot baseline. These results suggest that while prompting-based approaches can capture coarse duration plausibility, their scores are not directly comparable to the leaderboard systems due to differences in evaluation protocols,[5] highlighting the inherent difficulty of duration estimation without task-specific supervision.

Table 8 also summarizes the performance of participating teams, whose systems-trained with supervised, retrieval-augmented, or hybrid methods-demonstrate strong capability in modeling Vietnamese commonsense duration reasoning.

The **Engineers** team achieved the highest F1 score using a retrieval-augmented QLoRA fine-tuning framework, in which semantically relevant exemplars provided contextual grounding for duration plausibility. **UIT_BlackCoffee** obtained competitive results by combining expert-designed prompting with LoRA fine-tuning on Qwen3-24B, producing controlled and structurally consistent reasoning chains. **AI5** showed that compact encoder-based architectures remain effective: their ViDeBERTa-based discriminative model, enhanced through targeted

[5]Baseline scores are reported on internally evaluated private sets, whereas leaderboard scores are computed on the official blind test set.

augmentation and adaptive threshold tuning, delivered the strongest EM score. **HUET** fine-tuned Gemma-3-12B-it on a merged corpus combining the translated McTACO dataset with VLSP-provided data, achieving balanced performance through instruction-guided supervision. The **Softmind_AIO** team adopted a Dual-Prompt Ensemble (DP-Ens) strategy that integrates reasoning-oriented and decision-oriented prompts, offering interpretable outputs and a well-balanced precision–recall tradeoff.

Rather than relying on surface-level diversity metrics, we characterize dataset difficulty through baseline performance. The substantial gap between the encoder-based baseline and top systems, as well as the persistent EM–F1 discrepancy, indicates that DurationQA requires non-trivial commonsense reasoning beyond lexical matching.

Overall, these results indicate that retrieval-enhanced prompting, multi-stage training pipelines, and solver-assisted verification play central roles in improving duration-based commonsense reasoning in Vietnamese.

5.4. Error Analysis

To gain deeper insights into system behavior, we conducted a qualitative and quantitative error analysis for both subtasks. Although overall accuracies were high, several recurring error categories were identified, revealing common limitations in temporal reasoning and data generalization.

Subtask 1 - Date Arithmetic. Although most systems achieved near-perfect performance on the date arithmetic task, a close inspection of the remaining errors reveals several systematic patterns. These errors primarily stem from challenges in handling month-year boundaries, interpreting temporal direction, and performing consistent arithmetic operations. Representative cases are shown below for illustration.

- Month Overflow Errors. Models incorrectly handled transitions across year boundaries when adding or subtracting months. *Example:* "4 tháng sau tháng 3, 1000 (*4 months after March, 1000*)" → *Gold:* Tháng 7, 1000 (*July, 1000*); *Predicted:* Tháng 1, 1001 (*January, 1001*).

- Temporal Direction Misinterpretation. Some systems confused the directional operators "trước" (before) and "sau" (after), leading to inversed date computations. *Example:* "Trước 8 năm 1 tháng từ tháng 2, 1185 (*8 years and 1 month before February 1185*)" → *Gold:* Tháng 1, 1177 (*January 1177*); *Predicted:* Tháng 3, 1193 (*March 1193*).

- Month Offset Miscalculation. Errors involving small month offsets, often keeping the year correct but shifting the month incorrectly. *Example:* "1 năm trước từ tháng 9, 1620 (*1 year before September 1620*)" → *Gold:* Tháng 9, 1619 (*September 1619*); *Predicted:* Tháng 5, 1619 (*May 1619*).

- Arithmetic Miscalculation. Larger numerical mistakes resulting in implausible year outcomes, indicating failures in multi-step reasoning. *Example:* "3 năm 7 tháng trước tháng 4, 1141 (*1 year before September 1620*)" → *Gold:* Tháng 9, 1137 (*September 1137*); *Predicted:* Tháng 3, 1266 (*March 1266*).

Overall, systems integrating symbolic modules or calendar-tool verification tended to avoid such mistakes, whereas purely prompt-based models were more susceptible to directional confusion and normalization inconsistencies. These findings highlight the importance of coupling neural reasoning with structured temporal validation for robust performance in date arithmetic tasks.

Subtask 2 - Duration Question Answering. For the DurationQA subtask, error patterns were more diverse due to the inherently open-ended nature of temporal duration inference. Two frequent categories emerged across systems:

- Commonsense Gaps. Models occasionally failed to apply real-world knowledge about plausible event durations, resulting in systematically overestimated or underestimated answers. *For example:* "Mất bao lâu để sơ cứu người bị thương? (*How long does it take to give first aid to an injured person?*)" *Gold:* 30 phút (*30 minutes*) *Predict:* 1 giờ (*a hour*)

- Temporal Scope Misinterpretation. Some systems misread the intended scope of the question, especially when the event could be interpreted at multiple granularity levels (e.g., production vs. announcement). *For example:* "Mất bao lâu để ban nhạc phát hành album mới? (*How long does it take for a band to release a new album?*)" *Gold:* 3 tuần (*3 weeks*) *Predict:* all options (2-5 tháng (*2-5 months*), 3-4 tuần (*3-4 weeks*)).

- Event-Duration Mismatch. Systems sometimes assign implausible durations to certain activity types, despite the presence of strong commonsense priors.

Overall, generative reasoning models such as Softmind_AIO and UIT_BlackCoffee tended to produce more coherent explanations but occasionally over-generated justifications. In contrast, discriminative approaches (e.g., AI5) showed sharper decision boundaries but were more sensitive to phrasing variations and out-of-distribution expressions.

Cross-Task Observations. Across both subtasks, systems that incorporated retrieval tended to reduce factual mistakes but sometimes introduced unintended biases by surfacing loosely related examples that diverted the reasoning process. Models trained with rationale supervision also showed occasional mismatches between their generated explanations and the final predicted answers, revealing instability in reasoning consistency. Overall, these patterns indicate that Vietnamese temporal reasoning remains highly sensitive to lexical nuances and contextual variations. Future work would benefit from more selective retrieval mechanisms, stronger handling of negation and modifiers, and tighter integration of explicit temporal logic to improve model robustness and interpretability.

5.5. Comparative Discussion

Across both subtasks, several overarching trends can be observed. Retrieval-based augmentation and structured prompting emerged as critical components, supplying models with the semantic cues needed for stable temporal inference. Parameter-efficient fine-tuning techniques-such as LoRA, QLoRA, and DoRA-proved effective for adapting large language models under constrained computational resources.

For Subtask 1, incorporating symbolic checking and canonical decoding yielded near-perfect performance, underscoring the complementary strengths of deterministic temporal arithmetic and neural reasoning. Conversely, Subtask 2 demanded broader commonsense understanding and linguistic flexibility; in this setting, ensemble prompting and rationale-guided training were particularly beneficial, improving both robustness and interpretability.

6. Conclusion

The VLSP 2025 Vietnamese TemporalQA Shared Task showcased the potential of NLP techniques in tackling temporal reasoning for Vietnamese. The task introduced two complementary subtasks: *Date Arithmetic* for

temporal computation and *DurationQA* for reasoning about event durations. To support this challenge, we constructed and released two high-quality Vietnamese datasets, **Date-arith** and **DurationQA**, covering both arithmetic and commonsense temporal reasoning.

The shared task attracted several participants who employed diverse approaches, ranging from retrieval-augmented prompting and hybrid neural-symbolic systems to fine-tuned generative and discriminative LLMs. Results indicated that large pretrained models, when combined with structured prompting, retrieval, or parameter-efficient adaptation, achieved strong performance in temporal reasoning for Vietnamese. Symbolic verification proved particularly useful for arithmetic tasks, while dual-prompt and rationale-driven strategies improved interpretability for duration reasoning.

Overall, the TemporalQA shared task established a benchmark for Vietnamese temporal reasoning, provided high-quality datasets and baselines, and paved the way for future research in multilingual temporal question answering and temporal commonsense reasoning, thereby contributing meaningfully to the Vietnamese NLP community.

Limitations and Future Directions

Although the VLSP 2025 TemporalQA shared task provides an important step toward advancing temporal reasoning in Vietnamese, several limitations remain. Due to the inherently subjective nature of commonsense duration plausibility, annotation in DurationQA focuses on consistency checking rather than cross-annotation. Although ambiguous cases were filtered during the verification phase, we do not report inter-annotator agreement, which may limit the reliability of judgments in borderline cases.

The existing datasets-constructed in part through translation and synthetic generation-still fall short of representing the full breadth, variability, and naturalness of temporal expressions found in real-world Vietnamese texts. In addition, evaluation metrics such as Accuracy and F_1 primarily measure surface-level correctness rather than the depth or transparency of a model's reasoning process. The predominance of monolingual system designs also means that the benefits of multilingual transfer have yet to be fully examined. To address these issues, future research should prioritize expanding naturally sourced data, developing more comprehensive evaluation schemes, and exploring unified neural-symbolic approaches to temporal inference.

Acknowledgments

We sincerely thank the VLSP 2025 Organizing Committee and the TemporalQA shared task organizers for providing the datasets, baselines, and evaluation framework that enabled this work. We also appreciate the participating teams for their valuable contributions and collaborative spirit, as well as the annotation and verification teams for constructing reliable Vietnamese temporal question-answering resources.

This research was funded by the VNU University of Science, grant number TN.25.19.

References

[1] A. Saxena, S. Chakrabarti, P. Talukdar, Question Answering over Temporal Knowledge Graphs, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 6663–6676.
URL https://doi.org/10.18653/v1/2021.acl-long.520

[2] J. Cowie, W. Lehnert, Information Extraction, Communications of the ACM, Vol. 39, No. 1, 1996, pp. 80–91.
URL https://doi.org/10.1145/234173.234209

[3] M. Leonard, S. Westra, A. Phatak, M. Lambert, B. van den Hurk, K. McInnes, J. Risbey, S.

Schuster, D. Jakob, M. Stafford-Smith, A Compound Event Framework for Understanding Extreme Impacts, Wiley Interdisciplinary Reviews: Climate Change, Vol. 5, No. 1, 2014, pp. 113–128.
URL https://doi.org/10.1002/wcc.252

[4] N. Chambers, T. Cassidy, B. McDowell, S. Bethard, Dense Event Ordering with a Multi-Pass Architecture, Transactions of the Association for Computational Linguistics, Vol. 2, 2014, pp. 273–284.
URL https://doi.org/10.1162/tacl_a_00182

[5] J. A. Aslam, M. Ekstrand-Abueg, V. Pavlu, F. Diaz, T. Sakai, TREC 2013 Temporal Summarization, in: Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013), NIST Special Publication 500-302, 2013.
URL https://doi.org/10.6028/NIST.SP.500-302.tempsumm-overview

[6] W. Chen, X. Wang, W. Y. Wang, A Dataset for Answering Time-sensitive Questions, arXiv preprint arXiv:2108.06314 (2021).
URL https://doi.org/10.48550/arXiv.2108.06314

[7] Q. Ning, H. Wu, R. Han, N. Peng, M. Gardner, D. Roth, TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1158–1172.
URL https://doi.org/10.18653/v1/2020.emnlp-main.88

[8] B. Zhou, D. Khashabi, Q. Ning, D. Roth, "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), 2019, pp. 3363–3369.
URL https://doi.org/10.18653/v1/D19-1332

[9] J. Str¨otgen, M. Gertz, HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions, in: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010, pp. 321–324.
URL https://dl.acm.org/doi/10.5555/1859664.1859735

[10] P. Lambert, S. R. Schwer, N. Boffo, A New Model of Time Expressions Detection and Annotation in Vietnamese: The Hôm Case, in: 2012 International Conference on Asian Language Processing, IEEE, 2012, pp. 181–184.
URL https://doi.org/10.1109/IALP.2012.15

[11] Z. Chu, J. Chen, Q. Chen, W. Yu, H. Wang, M. Liu, B. Qin, TimeBench: A Comprehensive Evaluation of Temporal Reasoning Abilities in Large Language Models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 1204–1228.
URL https://doi.org/10.18653/v1/2024.acl-long.66

[12] Q. Tan, H. T. Ng, L. Bing, Towards Benchmarking and Improving the Temporal Reasoning Capability of Large Language Models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 14820–14835.
URL https://doi.org/10.18653/v1/2023.acl-long.828

[13] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., GPT-4o System Card, arXiv preprint arXiv:2410.21276, 2024.
URL https://doi.org/10.48550/arXiv.2410.21276

[14] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 10088–10115.
URL https://doi.org/10.52202/075280-0441

[15] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, M.-H. Chen, DoRA: Weight-Decomposed Low-Rank Adaptation, arXiv preprint arXiv:2402.09353, 2024.
URL https://doi.org/10.48550/arXiv.2402.09353

[16] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 Technical Report, arXiv preprint arXiv:2505.09388 (2025).
URL https://doi.org/10.48550/arXiv.2505.09388

[17] C. Van Nguyen, T. Nguyen, Q. Nguyen, H. Nguyen, B. Plüster, N. Pham, H. Nguyen, P. Schramowski, T. Nguyen, Vistral-7B-Chat: Towards a State-of-the-Art Large Language Model for Vietnamese, 2023.
URL https://huggingface.co/Viet-Mistral/Vistral-7B-Chat

[18] Gemma Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open Models Based on Gemini Research and Technology, arXiv preprint arXiv:2403.08295, 2024.
URL https://doi.org/10.48550/arXiv.2403.08295

[19] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783, 2024.
URL https://doi.org/10.48550/arXiv.2407.21783

[20] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by Weakly-Supervised Contrastive Pre-training, arXiv preprint arXiv:2212.03533, 2022.
URL https://doi.org/10.48550/arXiv.2212.03533

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv preprint arXiv:2106.09685, 2021.

URL https://doi.org/10.48550/arXiv.2106.09685

[22] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, et al., Qwen2.5 Technical Report, arXiv preprint arXiv:2412.15115, 2025.
URL https://doi.org/10.48550/arXiv.2412.15115

[23] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al., Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, arXiv preprint arXiv:2507.06261, 2025.
URL https://doi.org/10.48550/arXiv.2507.06261

[24] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-Enhanced BERT with Disentangled Attention, arXiv preprint arXiv:2006.03654, 2020.
URL https://doi.org/10.48550/arXiv.2006.03654

[25] T. Gao, X. Yao, D. Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6894–6910.
URL https://doi.org/10.18653/v1/2021.emnlp-main.552