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, Hanoi, Vietnam
{phamthiduc, hamylinh, lengoctoan_t65, huyenntm}@hus.edu.vn
Correspondence: phamthiduc@hus.edu.vn

Abstract

In 2025, the eleventh workshop on Vietnamese Language and Speech Processing (VLSP 2025[1]) organized the first shared task on Vietnamese Temporal Question Answering (TemporalQA). The primary goal of the TemporalQA challenge is to evaluate the performance of systems in understanding and reasoning over temporal information in Vietnamese. The task consists of two subtasks: Subtask 1 (Date Arithmetic – date-arith), focuses on handling questions related to date calculations, such as adding or subtracting time intervals from a given date, by understanding and manipulating time expressions to compute answers based on the provided context, and Subtask 2 (Duration Question Answering – durationQA), which requires systems to answer questions about the duration of events or actions based on a given context. The competition was conducted on the AIHUB platform[2] , where systems were ranked based on their performance in the private test phase, following a period of public testing. The top-performing team in Subtask 1 achieved an accuracy of 99%, while the best system in Subtask 2 obtained an F1-score of 81.89% and an Exact Match (EM) score of 47.52%.

Keywords: temporal reasoning, question answering, TemporalQA, VLSP 2025, Vietnamese

1 Introduction

Temporal reasoning is a fundamental yet challenging aspect of natural language understanding. It involves identifying, interpreting, and reasoning over temporal expressions such as dates, durations, and temporal relations between events in a text. Temporal information plays a crucial role in many downstream applications of natural language processing (NLP), including question answering (Saxena et al., 2021), information extraction (Cowie and Lehnert, 1996), event understanding (Leonard et al., 2014), timeline construction (Chambers et al., 2014), and temporal summarization (Aslam et al., 2013).

In recent years, several datasets and benchmarks have been developed to evaluate temporal reasoning capabilities in English, such as TimeQA (Chen et al., 2021), TORQUE (Ning et al., 2020), and McTACO (Zhou et al., 2019). These benchmarks have significantly advanced research in temporal understanding and reasoning. However, for Vietnamese, research on temporal reasoning remains limited, with no large-scale datasets or shared evaluation tasks available to date. This scarcity of resources poses challenges for the development and assessment of temporal reasoning systems for Vietnamese.

To address this gap, the eleventh Vietnamese Language and Speech Processing (VLSP 2025) organized the first shared task on Vietnamese Temporal Question Answering (TemporalQA[3]). The main goal of the TemporalQA shared task is to encourage the development of datasets, systems, and evaluation methodologies for temporal reasoning in Vietnamese. The task consists of two subtasks: (1) **Date Arithmetic (date-arith)** – performing computation over date expressions, and (2) **Duration Question Answering (durationQA)** – answering questions about the duration of events or actions based on a given context.

The competition was hosted on the AIHUB platform[4], where systems were first evaluated on a public test set and then ranked based on their performance on a private test set. The top-performing team achieved an accuracy of **99%** on Subtask 1 (Date Arithmetic), while the best system in Subtask 2 (DurationQA) obtained an **F1-score of** 81.89% and an **Exact Match (EM)**

[1] https://vlsp.org.vn/vlsp2025
[2] https://aihub.ml/
[3] https://vlsp.org.vn/vlsp2025/eval/tempqa
[4] https://aihub.ml/

score of 47.52%. These results demonstrate the promising potential of temporal reasoning research in Vietnamese and provide valuable baselines for future studies.

This paper provides a comprehensive overview of the TemporalQA shared task at VLSP 2025: Section 2 introduces the task, Section 3 presents the dataset construction and annotation process, Section 4 describes the methods and systems submitted by participants, and Section 5 discusses the results and outlines future directions.

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*.

In the first subtask, **Date Arithmetic**, systems are required to perform arithmetic operations over temporal expressions in Vietnamese text. Each question includes a reference time (e.g., a *date*, *month*, or *year*) and a temporal relation expressed as "before (trước)" or "after (sau)", which corresponds to an arithmetic operation of "subtract" or "add", respectively. The task aims to compute the resulting date or time point after applying this operation.

This subtask challenges systems to (1) accurately identify the reference time within the question, (2) interpret the operation and its associated temporal quantity, and (3) normalize the result into a canonical date format (e.g., *Tháng 4, 1296*).

For example:

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

Another example involves addition:

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

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.

The second subtask, **Duration Question Answering**, focuses on estimating the duration of events or actions mentioned in Vietnamese contexts. Each instance includes a short passage, a question about the event's duration, and a list of candidate durations. The system must classify each candidate as "yes" (plausible duration) or "no" (implausible), depending on contextual clues and commonsense knowledge.

For example:

- Context: *Tôi đang sửa chữa chiếc xe đạp bị hỏng.*
  Translation: "I am repairing a broken bicycle."
  Candidates: [30 phút, 1 tháng, 10 phút, 2 giờ]
  Correct labels: [yes, no, yes, yes]

This example tests the system's ability to infer realistic durations based on world knowledge (e.g., repairing a bicycle usually takes minutes or hours, not months).

Another example involves longer activities:

- Context: *Chúng tôi xây dựng một cây cầu bắc qua sông.*
  Translation: "We are building a bridge across the river."
  Candidates: [3 ngày, 2 năm, 5 tháng, 1 tuần]
  Correct labels: [no, yes, yes, no]

This subtask is challenging because it requires not only understanding the literal meaning of the event but also leveraging commonsense and real-world temporal knowledge. Unlike the first subtask, DurationQA emphasizes contextual reasoning, temporal plausibility, and cross-event inference rather than arithmetic computation.

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?",
  "answer": "Tháng 8, 2001"
}
```

For DurationQA, each entry includes a context, a question, candidate durations, and binary labels:

```
{
  "id": "dur_001",
  "context": "Tôi đang đọc một cuốn tiểu thuyết
  ↪ dài.",
  "question": "Thời gian thực hiện hành động trên
  ↪ là bao lâu?",
  "candidates": ["5 phút", "2 giờ", "1 tuần", "3
  ↪ tháng"],
  "labels": ["no", "yes", "yes", "no"]
}
```

The data for both subtasks are provided in JSON format, where each entry represents a single instance. For the Date Arithmetic subtask, each record includes a question and its corresponding computed date. For the DurationQA subtask, each record includes a context, a question, a list of candidate durations, and binary labels.

System performance is assessed using task-specific evaluation metrics.

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} denotes 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 metrics: **Exact Match (EM)** and **F1-score**.

The **Exact Match (EM)** score 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 represent the predicted and gold label sequences for instance i, and $\mathbf{1}(\cdot)$ is the indicator function.

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

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

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

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively.

These metrics collectively assess both the precision of computation and the robustness of reasoning in temporal question answering.

3 Data Preparation

This section describes the construction of two Vietnamese temporal question answering datasets: **Date-ArithQA** for temporal arithmetic reasoning and **DurationQA** for temporal duration understanding. Both datasets were built to evaluate the temporal reasoning ability of language models in Vietnamese, as no public resource currently exists for this purpose. Building on the English TimeBench benchmark (Chu et al., 2023), the dataset was prepared in three phases: translation into Vietnamese, generation of new data, and manual verification of grammar and temporal expressions.

3.1 Date-ArithQA: Date Arithmetic QA

Data Source. The Date-ArithQA dataset is adapted from the English TimeBench benchmark (Chu et al., 2023), which focuses on temporal arithmetic questions such as "What is the time 9 years and 1 month after Nov, 1543?". Each sample consists of a pair $\{\text{question}, \text{answer}\}$ without context.

Phase 1: Translation from English to Vietnamese. We first translated the English dataset into Vietnamese while preserving semantic intent and arithmetic correctness. The process produced 4,000 Vietnamese question–answer pairs.

Phase 2: Automatic Data Generation. To enhance linguistic diversity, we automatically generated 5,000 additional samples in Vietnamese. Each question was created by randomly sampling:

- a starting month and year (1000–2000),
- a duration (0–10 years, 0–12 months),
- a temporal direction ("before" or "after").

The target time was computed by normalizing months to integers, applying the shift, and converting back to "Month, Year" format. We designed five question templates to capture various Vietnamese expressions, such as:

- "Ngày tháng nào sẽ là 9 năm 2 tháng trước (sau) tháng 3, 1992?"
- "Giả sử bạn đang ở tháng 7, 1894, nếu trôi qua 8 năm 8 tháng thì là khi nào?"

Phase 3: Verification. All translation and generated data were manually checked for grammatical accuracy and correctness of the computed answers by a team of nine contributors. The final dataset contains 9,000 validated QA pairs, stored in .jsonl format.

Table 1: Statistics of the Date-ArithQA dataset.

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

3.2 DurationQA-Vi: Duration Question Answering

Data Source. We also construct a Vietnamese version of the DurationQA subset from TimeBench, which evaluates temporal reasoning over event durations. Each English instance includes context, question, options, and labels.

```
{
"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: Machine Translation. Automatic translation was applied to all text fields (context, question, and options), yielding 687 Vietnamese samples.

Phase 2: Data Generation with LLM. To enrich linguistic diversity, we used GPT-4o mini (Hurst et al., 2024), based on GPT-4 (Achiam et al., 2023) to generate 806 additional Vietnamese samples following the same schema. Each sample was manually reviewed to ensure consistency between options and labels, as well as natural Vietnamese phrasing.

Phase 3: Verification. A team of nine native Vietnamese annotators independently verified all 2,873 samples to ensure linguistic fluency and semantic accuracy, with any disagreements resolved through group discussion.

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

Table 2: Statistics of the DurationQA-Vi dataset.

Summary. Table 3 presents an overview of both datasets. Together, they form the first comprehensive Vietnamese temporal QA resource supporting two complementary reasoning types: temporal arithmetic and temporal duration inference.

Table 3: Summary of Vietnamese Temporal QA datasets.

Dataset	#Samples	Reasoning Type
Date-ArithQA	9,000	Temporal arithmetic
DurationQA-Vi	2,873	Temporal duration
Total	**11,873**	

3.3 Dataset Split and Evaluation Setup

For the **Date-ArithQA** subtask, the dataset was divided into three parts: *training*, *public test*, and *private test*. The training set consists of 3,000 questions generated from five Vietnamese question templates, while the public and private test sets contain 500 and 1,000 questions, respectively. Each question template represents a distinct linguistic structure expressing temporal arithmetic reasoning. Table 4 summarizes the distribution across templates and splits.

Table 4: Statistics of the Date-ArithQA dataset across templates and splits.

Question Template	Train	Public	Private
Bạn có thể dự đoán...	624	100	191
Giả sử bạn đang ở...	610	97	207
Ngày tháng nào sẽ là...	595	112	196
Thời gian...	604	94	194
Hãy tính thời điểm...	567	97	212
Total	**3,000**	**500**	**1,000**

For the **DurationQA-Vi** subtask, a similar data splitting strategy was applied. The training set contains 1,490 samples, while the public and private test sets consist of 400 and 625 samples, respectively. Each instance includes a short context, a question about the event duration, and multiple candidate durations annotated with binary labels. Table 5 summarizes the dataset distribution.

Table 5: Dataset split for the DurationQA-Vi subtask.

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

All public and private test sets were hosted on the AIHUB platform[5], where participants submitted model predictions 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 (Dettmers et al., 2023) and DoRA (Liu et al., 2024) were commonly used to adapt large Vietnamese or multilingual models including Qwen3 (Yang et al., 2025), Vistral-7B (Van Nguyen et al., 2023), and Gemma (Team et al., 2024a). Almost all systems integrated a canonicalization module that verified and standardized the predicted date to maintain consistency with the gold format.

4.1.2 Participants Approaches

Team UIT-NTTT adopted a retrieval-augmented prompting framework that leveraged **LLaMA3.1–8B** and **LLaMA3.1–70B** (Touvron et al., 2023) models. They first generated a large synthetic corpus of Vietnamese temporal arithmetic examples, each accompanied by detailed reasoning steps. These examples were embedded using **multilingual-e5-large** (Wang et al., 2022) and indexed in a **Qdrant** (Zhou et al., 2019) vector store. During inference, the system retrieved semantically similar examples and incorporated them into the prompt as demonstrations. The final prompt template guided the model through identifying the base date, computing the offset, and producing a normalized result in structured format. This approach emphasized in-context reasoning with explicit retrieval rather than parameter optimization, demonstrating the utility of few-shot adaptation for arithmetic reasoning in Vietnamese.

Team HUET designed a multi-phase fine-tuning pipeline for **Qwen3** (Yang et al., 2025) and **Gemma** (Team et al., 2024a) models that combined synthetic data generation, iterative supervised fine-tuning, and reasoning enhancement. A rule-based generator was used to produce a diverse collection of date arithmetic samples covering a range of operations (addition and subtraction) and time units (days, weeks, months, years). Each sample was automatically validated by a large teacher model (**Qwen3–235B**) to guarantee correctness before being included in training. Fine-tuning was conducted in several iterations, where model outputs were compared with solver-computed dates, and mispredicted cases were corrected and reintroduced into the dataset. In later stages, the team incorporated a "thinking-enabled" training mode that encouraged the model to generate intermediate reasoning traces along with the final date, effectively bridging symbolic and natural reasoning.

Team 777 proposed a bilingual hybrid architecture that separates language understanding from symbolic computation. Vietnamese inputs were first normalized and converted into English templates using **Qwen2.5–1.5B** (Team et al., 2024b). The structured English representation, which explicitly encoded the base date and temporal offset, was then processed by an English-trained reasoning model (**Flan-T5-base**) (Chung et al., 2024) to perform the date calculation. The output was translated back into Vietnamese and verified through a canonicalization and validation step using Python's `datetime` module. This architecture allowed the system to exploit

[5] https://aihub.ml/

strong English temporal reasoning models while preserving compatibility with Vietnamese inputs, achieving robust cross-lingual generalization.

Another team (**AI5**) employed a neural–symbolic hybrid system based on **Vistral-7B-iSMART** (Van Nguyen et al., 2023) fine-tuned using **QDoRA** (Liu et al., 2024). This training configuration combined low-rank adaptation (LoRA) (Hu et al., 2022) with directional regularization (DoRA) (Liu et al., 2024) for efficient yet stable optimization. To enrich the training data, additional Vietnamese examples were generated through retrieval-augmented generation and validated by a symbolic solver. At inference time, the model produced reasoning traces which were passed through a canonical decoder that enforced format standardization and logical consistency. The design emphasized a balance between flexible neural reasoning and deterministic computation, improving interpretability and reliability.

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 (Hu et al., 2022) 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 Participants Approaches

Team Engineers introduced a retrieval-guided fine-tuning framework centered on **Qwen2.5–7B** (Team et al., 2024b) and **Qwen3–8B** (Yang et al., 2025). For each question, semantically similar examples were retrieved using multilingual embeddings and appended to the model input to provide contextual analogues during reasoning. Fine-tuning was conducted using QLoRA (Dettmers et al., 2023) under 4-bit quantization for efficiency. During inference, multiple fine-tuned checkpoints were ensembled through a voting mechanism to increase stability. The design effectively combined retrieval-augmented prompting with efficient model adaptation.

Team Softmind_AIO developed the *Dual-Prompt Ensemble (DP-Ens)* method using **Qwen3–4B-Thinking** (Yang et al., 2025). Their system used two complementary prompting modes: the first prompted the model to produce a chain-of-thought explanation describing its reasoning process, while the second refined this explanation into a final binary label. During training, reasoning rationales generated by a large teacher model (Gemini 2.0 Flash (Comanici et al., 2025)) were distilled into the smaller student model to guide it toward interpretable inference. At inference time, outputs from both prompting paths were aggregated through a log-probability ensemble, ensuring that the final decision maintained both consistency and transparency.

Team UIT_BlackCoffee employed a task-specific expert prompting approach using **Qwen3–24B** (Yang et al., 2025) fine-tuned with LoRA (Hu et al., 2022) adapters. The model was instructed to act as a "temporal reasoning specialist," promoting disciplined and logically structured reasoning. Training data were reformatted into an instruction–response style that encouraged the model to output its reasoning process and decision in a standardized JSON structure. This design emphasized semantic clarity and consistency across different event types and duration categories.

Team HUET adopted an instruction-based fine-tuning strategy using the **Gemma-3-12B-it** model (Team et al., 2024a). Their system was trained on a merged corpus combining the translated **McTACO** dataset (Zhou et al., 2019) with the official VLSP-provided data, resulting in over fifteen thousand verified samples. Each instance was reformulated into a natural instruction–response format to leverage Gemma's strong instruction-following capabilities. The model was fine-tuned under the supervised fine-tuning (SFT) paradigm with reasoning-oriented examples that encouraged the generation of concise justifications for each

decision. This design achieved high recall and stable generalization, although subtle translation ambiguities and vague quantifiers occasionally led to borderline classifications, revealing the need for improved linguistic normalization in future iterations.

Team AI5 team pursued a discriminative approach based on **ViDeBERTa-base** (He et al., 2020) enhanced with LoRA (Hu et al., 2022) 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 (Gao et al., 2021) 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 reports the official results for the Date-Arith subtask, which required computing normalized calendar dates from Vietnamese temporal expressions. All participating systems achieved high accuracy on both public and private test sets, confirming that symbolic computation is well captured by retrieval or fine-tuning strategies.

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

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

Team	**Public Test (%)**	**Private Test (%)**
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 system from **UIT-NTTT** demonstrated strong reasoning consistency using contextual in-context examples rather than parameter updates. **HUET**'s iterative fine-tuning pipeline achieved similarly high performance by integrating solver verification and synthetic data refinement across multiple rounds of supervised learning. The **777** team's bilingual hybrid design proved that cross-lingual symbolic reasoning can effectively transfer from English models to Vietnamese inputs, aided by a canonicalization layer. Finally, **AI5**'s neural–symbolic hybrid with QDoRA fine-tuning achieved comparable accuracy while maintaining interpretability through solver-based validation. Across all approaches, systems that combined neural reasoning with deterministic symbolic correction achieved the most stable and generalizable results.

5.3 Subtask 2 – Duration Question Answering

The Duration subtask evaluated whether models could correctly judge the plausibility of event durations, a task emphasizing commonsense reasoning rather than direct computation. Performance across teams is shown in Table 8.

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

Team	F1	P	R	EM
The Engineers	81.89	76.45	88.15	47.52
UIT_BlackCoffee	80.13	73.06	88.72	42.72
AI5	80.03	74.79	86.06	49.12
HUET	79.97	70.71	92.02	40.32
Softmind_AIO	79.06	70.28	90.33	34.08

The **Engineers** team achieved the best overall F1 performance through retrieval-guided QLoRA fine-tuning, leveraging contextual examples retrieved by semantic similarity. **UIT_BlackCoffee** obtained competitive accuracy with a LoRA-based expert prompting strategy on Qwen3–24B, which enhanced reasoning discipline and structural consistency. **AI5**'s discriminative ViDeBERTa-based model, aided by adaptive threshold calibration, proved that compact encoders can remain competitive with careful augmentation and calibration. **HUET** fine-tuned Gemma-3-12B-it on a combined corpus of the translated McTACO dataset and the official VLSP data, achieving strong recall and balanced reasoning through instruction-based supervised fine-tuning. The **Softmind_AIO** system, using the Dual-Prompt Ensemble (DP-Ens), combined reasoning and decision prompts to deliver interpretable outputs and balanced recall–precision performance. Collectively, these results reveal that retrieval-based context enrichment and multi-stage prompting significantly improve commonsense duration 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. Despite near-perfect accuracy, residual errors mainly arose from three sources: (1) *Boundary ambiguity*, where models misinterpreted the inclusion or exclusion of the starting date (e.g., "after three days" vs. "in three days"). (2) *Month overflow errors*, in which models failed to handle cases that crossed month or year boundaries, such as "two months after December 25." (3) *Format normalization*, particularly inconsistencies in output canonicalization (e.g., "2025-3-5" instead of "2025-03-05"). Systems that incorporated symbolic solvers or Python-based validation (e.g., **HUET**, **777**, **AI5**) were generally able to detect and correct such cases automatically. In contrast, purely prompt-based models (e.g., **UIT-NTTT**) occasionally produced logically correct but syntactically inconsistent answers, highlighting the need for tighter integration between reasoning and canonicalization.

Subtask 2 – Duration Question Answering. Error patterns in this subtask were more diverse and cognitively complex. The most common issues included: (1) *Event–duration mismatch*, where models overestimated or underestimated plausible durations for certain event types (e.g., predicting "three hours" as plausible for "building a house"). (2) *Negation and modifier confusion*, particularly with sentences containing contrastive or comparative cues such as "no longer than" or "at least." (3) *Commonsense inconsistency*, where models relied on surface co-occurrence rather than real-world temporal knowledge (e.g., assuming "graduating" and "one week" co-occur frequently). (4) *Linguistic variability*, especially with nested or colloquial expressions like "chưa đầy hai tuần" ("less than two weeks") or "kéo dài ngót nghét một năm" ("almost a year"). Generative reasoning models such as **Softmind_AIO** and **UIT_BlackCoffee** tended to make fewer syntactic mistakes but occasionally overgenerated justifications, while discriminative models (e.g., **AI5**) exhibited sharper decision boundaries but struggled with out-of-distribution phrasing.

Cross-Task Observations. Across both subtasks, retrieval-based approaches reduced factual errors but occasionally introduced spurious contextual bias—retrieving irrelevant examples that misled reasoning chains. Inconsistencies between reasoning explanations and final predictions were also observed in models trained with rationale supervision. These findings suggest that

Vietnamese temporal reasoning remains sensitive to linguistic nuance and contextual diversity, and that future systems should integrate dynamic retrieval filtering, improved negation handling, and explicit temporal logic modules to enhance robustness and interpretability.

5.5 Comparative Discussion

Across both subtasks, several clear trends emerge. Retrieval augmentation and structured prompting were key enablers of temporal reasoning, providing models with explicit semantic context for inference. Parameter-efficient fine-tuning methods such as LoRA, QLoRA, and DoRA enabled effective adaptation of large LLMs under limited computational budgets. In Subtask 1, symbolic verification and canonical decoding led to near-perfect results, demonstrating that deterministic reasoning and neural inference can complement each other. In contrast, Subtask 2 required greater semantic flexibility—where ensemble prompting and rationale distillation improved interpretability and robustness.

Overall, the 2025 VLSP TemporalQA shared task shows that hybrid neural–symbolic and retrieval-augmented paradigms are particularly promising for advancing Vietnamese temporal reasoning, bridging the gap between symbolic computation and commonsense understanding.

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-ArithQA** and **DurationQA-Vi**, 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

While the VLSP 2025 TemporalQA shared task establishes a strong foundation for temporal reasoning in Vietnamese, several aspects warrant further improvement. The current datasets—partly translated and synthetically generated—do not yet capture the full linguistic diversity and spontaneity of temporal expressions in real-world Vietnamese. Moreover, evaluation metrics such as Accuracy and F_1 mainly assess correctness rather than reasoning depth or interpretability. Finally, most participating systems were developed under monolingual settings, leaving the potential of multilingual transfer underexplored. Future work should therefore focus on expanding naturally occurring data, designing richer evaluation protocols, and integrating symbolic and neural reasoning in a unified framework.

Acknowledgments

We would like to express our sincere gratitude to the VLSP 2025 Organizing Committee and the TemporalQA shared task organizers for providing the datasets, baseline systems, and evaluation framework that made this work possible. We also thank all participating teams for their valuable contributions, insightful discussions, and collaborative spirit throughout the competition. Special thanks go to the annotation and data verification teams for their meticulous efforts in constructing reliable temporal question-answering resources in Vietnamese.

This research was supported in part by the VLSP community and affiliated institutions contributing computational resources and organizational support. We also thank all participating teams for their valuable contributions and the reviewers for their constructive feedback, which helped improve this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Javed A Aslam, Matthew Ekstrand-Abueg, Virgil Pavlu, Fernando Diaz, and Tetsuya Sakai. 2013. Trec 2013 temporal summarization. In *TREC*.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Michael Leonard, Seth Westra, Aloke Phatak, Martin Lambert, Bart van den Hurk, Kathleen McInnes, James Risbey, Sandra Schuster, Doerte Jakob, and Mark Stafford-Smith. 2014. A compound event framework for understanding extreme impacts. *Wiley Interdisciplinary Reviews: Climate Change*, 5(1):113–128.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. Torque: A reading comprehension dataset of temporal ordering questions. *arXiv preprint arXiv:2005.00242*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Qwen Team and 1 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2:3.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chien Van Nguyen, Thuat Nguyen, Quan Nguyen, Huy Nguyen, Björn Plüster, Nam Pham, Huu Nguyen, Patrick Schramowski, and Thien Nguyen. 2023. Vistral-7b-chat-towards a state-of-the-art large language model for vietnamese.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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. *arXiv preprint arXiv:1909.03065*.