Enabling Temporal Commonsense in Vietnamese LLMs – Date-Arith and DurationQA

Duc Dinh Chu*, **Thanh-Bac Nguyen Ba***, **Duy Dinh Le, Khanh Van Tran**
FPT Smart Cloud, FPT Corporation
{duccd4, bacnbt, duyld10, khanhtv26}@fpt.com

Abstract

This work investigates two subtasks in temporal reasoning: 1. Date Arithmetic (date-arith) and 2. Duration Question Answering (durationQA). For date-arith, we focus on fine-tuning large language models (LLMs) to directly extract and compute answers. For durationQA, the challenge lies in identifying both explicit and implicit duration expressions in text and reasoning with world knowledge to assess correctness. We explore multiple approaches, from naive supervised fine-tuning (SFT) to SFT augmented with reasoning-based synthetic data and GRPO. Our findings highlight the critical role of carefully constructed data and appropriate training strategies in enabling effective temporal reasoning. Our source code can be found here.

1 Introduction

Temporal reasoning is a fundamental component of natural language understanding, yet it remains a challenging task for current language models—especially in low-resource languages such as Vietnamese. Successfully answering temporal questions requires not only the identification and normalization of time expressions, but also symbolic reasoning and integration of commonsense or real-world knowledge.

The VLSP 2025 shared task introduces two subtasks designed to evaluate temporal reasoning capabilities in Vietnamese:

Subtask 1: Date Arithmetic. This subtask assesses a model's ability to perform abstract date calculations. Given a reference date, the model must compute the correct date that occurs a specified amount of time before or after it. For example, the question "Thời gian 1 năm và 2 tháng trước tháng 6, 1297 là khi nào?" ("What is the date 1 year and 2 months before June, 1297?") should yield the correct answer: "Tháng 4, 1296" ("April, 1296").

Subtask 2: Duration Question Answering. This subtask evaluates a model's understanding of event-related temporal knowledge, including typical durations, frequencies, sequences, and plausible timelines. Given a context and a temporal question, the model must classify each answer option as a plausible duration ("yes") or implausible ("no"). For instance, given the context "Tôi đang sửa chữa chiếc xe đạp bị hỏng." ("I am repairing a broken bicycle.") and the question "Mất thời gian bao lâu để sửa chữa chiếc xe đạp?" ("How long does it take to repair a bicycle?"), with candidate answers ["30 phút", "1 tháng", "10 phút", "2 giờ"], the correct labels are ["yes", "no", "yes", "yes"].

In this paper, we present our approach to both subtasks. For **Date Arithmetic**, we design systems capable of performing symbolic date computations, such as adding or subtracting temporal intervals. For **Duration Question Answering**, we develop methods to detect explicit and implicit temporal cues in text and assess candidate answers using both linguistic patterns and real-world knowledge. To tackle these challenges, we explore a range of techniques, including supervised fine-tuning (SFT), reasoning-aware data augmentation, and reinforcement-based prompt optimization. Our results offer insights into the effectiveness and limitations of current methods for temporal reasoning in Vietnamese.

2 Related Works

Date Arithmetic The Date Arithmetic task requires models to compute dates by adding or subtracting specified time intervals (years, months, days) from a given reference date, testing both temporal understanding and arithmetic reasoning. One approach (Tan et al., 2023) is to finetune T5 model to solve this problem. To deal with this problem,

*These authors contributed equally.

(Chu et al., 2024) use LLM with standard prompting (zero-shot and few-shot) and Chain-of-thought prompting (Wei et al., 2022).

Duration Question Answering The Duration Question Answering lies at the intersection of temporal commonsense reasoning and multiple-choice QA, requiring models to judge the plausibility of candidate durations for events described in context. McTACO (Zhou et al., 2019) was introduced to evaluate diverse temporal phenomena—including event duration, frequency, and order—in a multiple-choice QA format. It contains 13000 human-authored questions covering five temporal dimensions (duration, ordering, typical time, frequency, and stationarity); the duration subtask specifically tests whether models can select plausible durations from foil options. Despite fine-tuning on transformer models (e.g., RoBERTa (Liu et al., 2019)), performance on duration questions lags behind human accuracy by over 20 percentage points, highlighting challenges in abstract duration reasoning. (Zhou et al., 2020) present a transformer-based temporal common sense language model called TACOLM, trained on temporal data that are extracted using patterns from a large corpus. It uses temporal data from 3 temporal commonsense dimensions: duration, frequency, and typical time. It outperforms BERT (Devlin et al., 2019) in various temporal tasks, including McTACO. Recently, with the development of LLMs, (Virgo et al., 2022) proposes methods using these models to deal with this task.

3 Methodology

3.1 Subtask 1: Date Arithmetic

In this subtask, we conducted experiments with Qwen3 models (Yang et al., 2025). In addition to the dataset released by the task organizers, we developed a supplementary dataset to further enrich the training material. Specifically, to synthesize approximately 40,000 samples, we adopted a rule-based methodology. For each sample, we randomly selected a date (month and year), an arithmetic operation (addition or subtraction), and a time interval (in months or years). We applied the chosen operation to ensure that the corresponding sample was assigned a correct label. To guarantee that all prompt instances complied with the format mandated by the evaluation framework, we employed a few-shot prompting strategy to regularize and standardize the synthesized dataset.

Hyperparameter	**Gemma (SFT)**	**Qwen (SFT)**
Precision	BF16	BF16
Cutoff length	2048	2048
Fine-tuning type	Full	Full
Attention	FlashAttention-2	FlashAttention-2
Gradient accumulation	1	1
Batch size / device	4	16
Learning rate	5.0e-5	5.0e-5
Scheduler	cosine	cosine
Warmup steps	25	25
Max grad norm	1.0	1.0
Epochs	5	5
Optimizer	AdamW	AdamW
DeepSpeed config	ZeRO-3	ZeRO-3
Template	gemma3	qwen3
Enable thinking	false	true

Table 1: Hyperparameters used for supervised fine-tuning (SFT) of Gemma and Qwen models for subtask 1.

To double-check the synthesized dataset, we employed the powerful Qwen3-235B model to produce outputs in the following JSON format: { "start_month": . . . , "start_year": . . . , "operator": . . . , // “add” or “subtract” "interval_month": . . . , "interval_year": . . . }.

We then used a rule-based mechanism to compare the model’s output against the gold labels for consistency and correctness.

Our training pipeline followed an iterative refinement strategy. We first designed an initial prompt and leveraged the Qwen3-235B model to improve the prompt design and generate reasoning traces. These reasoning-augmented outputs were then used to fine-tune the smaller-scale models. The fine-tuned models were subsequently applied for inference to identify erroneous predictions. Based on the observed errors, we refined the prompts and repeated the process iteratively until stable performance was achieved.

3.2 Subtask 2: Duration Question Answering

Using the same model selection strategy as in Subtask 1, we identified two promising candidates for this subtask: Qwen3 and Gemma-3 (Team et al., 2025).

In addition to the dataset provided by the task organizers, we also incorporated the McTACO dataset. Specifically, we translated and merged the organizer’s dataset with McTACO, then used Qwen3-235B to evaluate and determine the translation quality. This process yielded a combined corpus of 15,618 samples. From this corpus, we generated reasoning data using Qwen3, and subsequently performed verification, resulting in 15,120

Hyperparameter	SFT	GRPO
Attention	FlashAttention-2	FlashAttention-2
Batch size / device	64	16
Learning rate	5.0e-5	1.0e-6
Epochs	3	5
Optimizer	AdamW	AdamW
DeepSpeed config	ZeRO-3	ZeRO-3

Table 2: Hyperparameters used for supervised fine-tuning (SFT) and GRPO in subtask 2.

high-quality samples, which we used for supervised fine-tuning (SFT).

Furthermore, we experimented with the GRPO method (DeepSeek-AI et al., 2025), which allows the model to perform self-reasoning before generating final answers, thereby encouraging more structured and accurate temporal reasoning. We used F1-score as the reward function for the training process. We employ the Unsloth framework for training, using the hyperparameter settings detailed in Table 2.

For both subtasks, the computing resource is 8xH100 GPUs.

4 Analysis

4.1 Subtask 1

For subtask 1, we report accuracy, defined as the proportion of system predictions that exactly match the gold-standard answers.

$$EM = \frac{1}{N}\sum_{i=1}^{N} 1(y_i^* = y_i)$$

where:

- N is the number of samples (predictions),
- y_i^* is the predicted output for sample i,
- y_i is the ground truth (reference) output for sample i,
- $1(\cdot)$ is the indicator function, equal to 1 if the argument is true, and 0 otherwise.

Our preliminary experiments yielded the following results across foundation models in Table 1. These results indicate that Qwen3 achieved the best performance on the evaluation dataset. During the iterative training process, we encountered several challenges:

Iteration 1: The model produced systematic errors by performing unintended unit conversions.
Example:

Question: Hãy tính thời điểm 9 tháng sau tháng 8, 1040.

Expected answer: Tháng 5, 1041

Wrong reasoning path:

1. *Phân tích câu hỏi*:
- *Thời điểm gốc*: tháng 8, năm 1040
- *Khoảng thời gian*: 9 tháng (tương đương 1 năm)
- *Hướng*: "sau" (cộng)
2. *Thực hiện phép toán*:
- *Cộng năm (trước)*: năm mới = 1040 + 1 = 1041
- *Cộng tháng (sau)*: tháng tạm = 8 + 9 = 17
- *Điều chỉnh tháng* (vì 17 > 12):
- năm mới = 1041 + 1 = 1042
- tháng mới = 17 - 12 = 5
3. *Trình bày kết quả*:
- *Tháng 5, năm 1042*
- *Kiểm tra*:
- Tháng 5 ∈ [1, 12]
- Năm 1042 > 0
4. *Kết luận*: *Tháng 5, 1042*

In other cases, it simultaneously converted months into years while also adding them separately, leading to inconsistent answers.

Example:

Question: Giả sử bạn đang ở tháng 1, 1421, thời gian sau 12 tháng, thì là thời điểm nào?

Expected answer: Tháng 1, 1422

Wrong reasoning path:

1. **Phân tích câu hỏi**:
- **Thời điểm gốc**: tháng 1, năm 1421
- **Khoảng thời gian**: 12 tháng (tương đương 1 năm)
- **Hướng**: "sau" (cộng)
2. **Thực hiện phép toán**:
- **Cộng năm**: năm mới = 1421 + 1 = 1422
- **Cộng tháng**: tháng tạm = 1 + 12 = 13
- **Điều chỉnh (vì 13 > 12)**:
- năm mới = 1422 + 1 = 1423
- tháng mới = 13 - 12 = 1
3. **Trình bày kết quả**:
- **Tháng 1, 1423**
- **Kiểm tra**: tháng 1 ∈ [1,12], năm 1423 > 0
4. **Kết luận**: "Tháng 1, 1423"

Iteration 2: The model often misidentified temporal spans, particularly in cases where the duration was expressed solely in months or solely in years.
Example:

Question: Ngày tháng nào sẽ là 4 năm trước tháng 10, 1124?

Expected answer: Tháng 10, 1120

Iteration	Result
1	0.26
2	0.96
3	0.98

Table 3: Qwen3's improvement on each iteration on public test in subtask 1.

Setting	Result
Qwen3 + SFT no thinking	0.98
Qwen3 + SFT thinking	0.99

Table 4: Final results on private test subtask 1.

Wrong reasoning path:
1. **Phân tích câu hỏi**:
- Thời điểm gốc: tháng 10, năm 1124
- Khoảng thời gian: 4 năm + 1124
- Hướng: "trước" (trừ)
2. **Thực hiện phép toán**:
- **Trừ năm (thực hiện trước)**: ‘năm mới = 1124 - 4 = 1120‘
- **Trừ tháng (thực hiện sau)**: ‘tháng tạm = 10 - 11 = -1‘
- **Điều chỉnh (vì tháng tạm < 1)**:
- năm mới = 1120 - 1 = 1119
- tháng mới = -1 + 12 = 11
3. **Trình bày kết quả**:
- Tháng 11, năm 1119
- Kiểm tra: tháng 11 ∈ [1,12], năm 1119 > 0
4. **Kết luận**: "Tháng 11, 1119"

To address these issues, we incorporated erroneous cases directly into the original prompt. We then used the Qwen3-235B model to evaluate these error examples and automatically refine the initial prompt accordingly. Details regarding the prompt optimization process can be found in the accompanying source code.

4.2 Subtask 2

Evaluation metrics used in subtask 2 include:

- **Precision** measures the proportion of correctly predicted "yes" answers among all "yes" predictions made by the system.
- **Recall** measures the proportion of correctly predicted "yes" answers among all actual "yes" answers in the ground truth.
- **F1-score** is the harmonic mean of Precision and Recall, providing a balanced measure of overall performance.

$$\textbf{F1-score:} \quad = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

We experimented with synthetic data generation; however, the process introduced substantial ambiguity, which limited its effectiveness. Consequently, we decided to rely primarily on translated data from existing resources rather than synthetic expansions. In addition, our experiments revealed that fine-tuning Gemma-3 with supervised fine-tuning (SFT) was more effective than the pipeline of fine-tuning Qwen3 with SFT followed by GRPO, highlighting model-specific differences in training efficiency. Our results are shown in table 5. The results above may be attributed to the fact that Gemma3 (12B) is an instruct-style model whose strength lies in general reasoning tasks, while Qwen3 (8B) is a thinking-style model that excels in higher-order reasoning tasks such as mathematics and code generation. Consequently, Gemma3 is better suited to tasks involving extended interaction or sustained reasoning (i.e. "task duration").

5 Discussion

For subtask 1, the problem was relatively simple, as it involved only months and years, unlike the TimeBench dataset (Chu et al., 2024).

For subtask 2, data augmentation contributed positively to performance; however, the reasoning data generated was not sufficiently robust. As a result, the improvement did not meet our initial expectations, suggesting the need for more effective strategies in generating high-quality reasoning traces.

6 Conclusion

In this work, we investigated two subtasks of temporal reasoning in Vietnamese: Date Arithmetic (date-arith) and Duration Question Answering (durationQA). For date-arith, we demonstrated that large language models can be fine-tuned to directly extract and compute temporal answers, while for durationQA the primary challenge lies in handling both explicit and implicit duration expressions and reasoning with external world knowledge.

Through a series of experiments, we compared multiple training strategies, ranging from naïve supervised fine-tuning to approaches that incorporate reasoning-augmented synthetic data and GRPO.

Model	Public			Private		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Qwen3-8B + SFT	0.7240	0.8508	0.7823	0.7089	0.8558	0.7755
Qwen3-8B + SFT + GRPO	0.7240	0.8508	0.7823	0.7134	0.8807	0.7883
Qwen3-30B + SFT	0.7474	0.8718	0.8048	0.7233	0.8824	0.7949
Gemma-3-12b-it + SFT	0.7148	0.9088	0.8002	0.7071	0.9202	0.7997

Table 5: Precision, Recall, and F1 results for each model in subtask 2.

Our results underscore the importance of carefully designed data and iterative refinement of training pipelines for achieving robust performance in temporal reasoning tasks.

For future work, we plan to extend the Date Arithmetic (date-arith) task to cover a broader set of temporal units beyond days and months, which were the primary focus in subtask 1. For subtask 2 (durationQA), we aim to address questions with more complex and diverse contexts, thereby pushing the limits of temporal reasoning in natural language.

Acknowledgments

We would like to thank FPT Smart Cloud for providing the computational resources that enabled the experiments.

References

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. 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)*, pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. 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)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference* *on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.