Retrieval-Guided Fine-tuning for Vietnamese Event Duration Question Answering

Duong Nguyen Dao
FPT Smart Cloud Company
Hanoi, Vietnam
nguyenduongyht@gmail.com

Thanh Xuan Nguyen
Hanoi University of Science and Technology
Hanoi, Vietnam
xuanthanh2201.work@gmail.com

Abstract

In the VLSP 2025 Temporal QA Challenge[1], our team participates in Sub-Task 2 (DurationQA) and presents a retrieval-guided fine-tuning system. Given a Vietnamese context, a question about event duration and multiple candidate duration options, the task requires predicting a yes/no label for each option. Our method combines (i) QLoRA fine-tuning techniques with large language models; (ii) a retrieval module that sources relevant examples from the training set to guide the reasoning process; (iii) an ensemble that combines model outputs through majority voting to produce the final prediction. The approach is transparent, data-efficient, and challenge-compliant.[2]

1 Introduction

Temporal reasoning about event durations requires sophisticated understanding of both linguistic expressions and real-world constraints. In Vietnamese, duration expressions exhibit remarkable diversity: the same temporal quantity may appear as "1 giờ 30 phút," "1.5 giờ," "1 giờ rưỡi," "90 phút," or approximate descriptions like "tầm 10–15 phút," "dưới 2 giờ". This linguistic variability, combined with the need for commonsense reasoning when contexts are underspecified, makes Vietnamese event duration question answering particularly challenging.

The VLSP 2025 Sub-Task 2 (DurationQA) formalizes this challenge as a multi-label binary classification problem. Given a Vietnamese context, a duration-related question, and multiple candidate duration options, systems must predict yes/no labels for each option. Evaluation employs F1-score, balancing precision (proportion of correct yes predictions) and recall (proportion of ground-truth yes labels successfully identified), together with Exact Match (EM), which measures the proportion of samples where the predicted option set exactly matches the ground-truth answers.

Recent advances in temporal reasoning have produced datasets like McTACO (Zhou et al., 2019) and UDST-DurationQA (Virgo et al., 2022), which share structural similarities with VLSP 2025 Sub-Task 2. However, existing approaches have primarily relied on either commercial large language models with multi-stage pipelines (Chu et al., 2024) or BERT-based fine-tuning strategies (Virgo et al., 2022). These methods have not fully explored the potential of open-source large language models operating under resource constraints—a central requirement of this challenge.

We hypothesize that training examples contain valuable patterns for Vietnamese duration reasoning: linguistic realizations, answer templates, and event types exhibit systematic recurrence across the dataset. Building on this insight, we develop a retrieval-guided reasoning system that leverages analogous labeled examples to enhance inference-time decision making.

Our contributions are threefold: (1) We demonstrate effective QLoRA-based fine-tuning strategies for Vietnamese temporal reasoning using resource-constrained models, (2) We introduce a retrieval mechanism that dynamically incorporates relevant training examples as few-shot guidance during inference, and (3) We show that ensemble methods combining multiple model configurations substantially improve robustness and performance on this challenging task.

2 Related Work

Temporal reasoning and benchmarks The field of temporal reasoning has evolved significantly with benchmarks like TimeBench (Chu et al., 2024), which demonstrates the brittleness of purely

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

[2] Our code is available at https://github.com/duongkstn/durationqa-vlsp-solution

parametric LLM reasoning on temporal tasks. Follow-up studies (Tan et al., 2023) have reinforced these findings, highlighting the need for structured approaches to temporal understanding. VLSP 2025 extends these concepts to Vietnamese while introducing a multi-label decision framework that requires simultaneous evaluation of multiple duration candidates.

Event duration question answering The McTACO dataset (Zhou et al., 2019) established foundational work in temporal commonsense reasoning through multiple-choice questions about event durations. Building on this foundation, Virgo et al. (2022) advanced the field by leveraging temporal information extraction resources and formulating duration understanding as a classification task, leading to successful BERT-based fine-tuning approaches. Their work demonstrates the effectiveness of supervised learning for duration reasoning, though it focuses primarily on English-language contexts.

Retrieval-augmented generation Retrieval-augmented generation (RAG) has emerged as a powerful paradigm for improving factual accuracy by incorporating relevant contexts at inference time. This approach draws conceptual inspiration from classical case-based reasoning (CBR) (Aamodt and Plaza, 1994), which emphasizes learning from analogous past instances. Our work operationalizes these concepts within an reasoning framework, treating the training dataset as a dynamic knowledge repository, where labeled cases provide supervisory signals that guide fine-tuned models during inference.

Large language models (LLMs) The development of large language models has transformed natural language processing. The research community has seen the emergence of powerful open-source models such as LLaMA 2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), FLAN-T5-Large (Wei et al., 2022), Gemma (Gemma Team, 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), alongside proprietary, API-based systems such as GPT-4 (Achiam et al., 2024) and Gemini (Gemini Team et al., 2023). This ecosystem has accelerated innovation, with open models supporting reproducible research and closed models driving real-world applications.

Large language model fine-tuning The development of parameter-efficient fine-tuning methods has democratized access to large language model adaptation. QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) extends the LoRA framework (Hu et al., 2022) to enable memory-efficient fine-tuning through quantization techniques. This approach has proven particularly valuable for resource-constrained environments while maintaining competitive performance across diverse tasks.

3 Methodology

3.1 Task Definition

Each example is a tuple (C, Q, O) where C is a Vietnamese context, Q stands for question and $O = \{o_i\}_{i=1}^m$ are candidate duration options. The system outputs answer $A = \{y_i\}_{i=1}^m$, $y_i \in \{\text{yes}, \text{no}\}$. The evaluation metrics comprise Exact Match, Precision, Recall, and F1-score, which are defined as follows.

- **Exact Match (EM)** checks whether the predicted label sequence exactly matches the ground truth.
- **Precision** refers to the proportion of instances predicted as "yes" that are indeed correct, reflecting the reliability of positive predictions.
- **Recall** refers to the proportion of actual "yes" instances that are correctly identified by the system, reflecting the model's ability to retrieve all relevant positive cases.
- **F1-score** is the harmonic mean of Precision and Recall, balancing correctness and completeness.

3.2 QLoRA Fine-Tuning

QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) leverages the LoRA framework (Hu et al., 2022) to enable memory-efficient fine-tuning of LLMs. This technique addresses the high computational demands of adapting very large models by combining low-rank adaptation with quantization. Specifically, QLoRA stores the base model in 4-bit precision, drastically reducing GPU memory usage, while training LoRA adapters in higher precision to preserve performance.

The key idea is to retain the expressive power of LoRA while lowering memory requirements through quantization. During training, the pre-trained model weights are quantized and frozen, ensuring no gradient updates are applied to them.

Instead, trainable low-rank matrices are introduced alongside the frozen projections, similar to LoRA, and optimized in higher precision. This design makes it possible to fine-tune larger models on a single modern GPU.

3.3 Supervised Fine-tuning

Among current state-of-the-art LLMs which support multilingual capabilities, the Qwen family (developed by Alibaba) has emerged as a strong choice for multilingual applications, with notable support for Vietnamese. Qwen models are trained on large-scale multilingual corpora and provide strong performance across a wide range of natural language understanding and generation tasks.

To address this challenge, we apply supervised fine-tuning to Qwen2.5 (Yang et al., 2025b) and Qwen3 (Yang et al., 2025a) using both the provided training set and the UDST-DurationQA dataset (Virgo et al., 2022). Additionally, due to limited computational resources, 7B and 8B (billion parameters) models have been employed to facilitate training.

3.4 Retrieval-Guided Prompting

To effectively leverage the training data for few-shot learning, we implement a retrieval mechanism that identifies relevant examples to guide the model's reasoning process. We design task-specific instruction prompts that satisfy the following requirements:

- **Task alignment:** the prompt must clearly reflect the target task.
- **Explicit output format:** the expected answer format should be well-defined.
- **Role-based system instruction:** the model is assigned a role to condition its behavior.
- **Retrieval integration:** relevant examples are dynamically incorporated to provide contextual guidance.

Different output formats can be employed depending on the downstream requirement, such as JSON format or list format. In our framework, prompts are formulated as follows:

$$\mathrm{PROMPT} = \{\mathrm{ROLE}\}\{\mathrm{T}\}\{\mathrm{C}\}\{\mathrm{Q}\}\{\mathrm{O}\}\{\mathrm{F}\} \tag{1}$$

where $ROLE$ is instruction prompt, T is target task and F is output format.

During inference with retrieval guidance, we use a prompt that integrates few-shot relevant examples:

$$\begin{aligned}\mathrm{PROMPT} = \{\mathrm{ROLE}\}\{\mathrm{T}\}\{\mathrm{EXAMPLES}\} \\ \{\mathrm{C}\}\{\mathrm{Q}\}\{\mathrm{O}\}\{\mathrm{F}\}\end{aligned} \tag{2}$$

where

$$\begin{aligned}\mathrm{EXAMPLES} = \{\mathrm{C}_1\}\{\mathrm{Q}_1\}\{\mathrm{O}_1\}\{\mathrm{A}_1\}\ldots \\ \ldots\{\mathrm{C}_\mathrm{k}\}\{\mathrm{Q}_\mathrm{k}\}\{\mathrm{O}_\mathrm{k}\}\{\mathrm{A}_\mathrm{k}\}\end{aligned} \tag{3}$$

with each set of context, question, options and answer from the K most relevant examples. All prompt formats are clearly declared in Appendix A.1 and A.2.

3.5 Ensemble Strategy

To enhance robustness and reduce variance across individual models, we employ an ensemble strategy at the prediction stage. Specifically, multiple fine-tuned instances of Qwen2.5-7B (Yang et al., 2025b) and Qwen3-8B (Yang et al., 2025a) are independently trained with identical data and prompt settings. At inference time, each model produces a binary decision (yes/no) for every candidate option.

The final predictions are obtained through majority voting across the ensemble. For each option, if the majority of models predict yes, the system outputs yes; otherwise, it outputs no. This simple yet effective aggregation reduces the risk of overfitting to individual model biases and improves overall performance.

4 Datasets

We use two resources: VLSP 2025 DurationQA and UDST-DurationQA (Virgo et al., 2022). To illustrate the data format, a sample from the DurationQA dataset is shown in Table 1, which highlights the context, question, candidate options, and corresponding labels.

Table 2 reports their statistics. The VLSP dataset has a relatively consistent structure, with an average of 4 options per sample, whereas UDST-DurationQA contains more candidate answers per question, averaging 7.0. Moreover, UDST-DurationQA is much larger in size compared to VLSP. The proportion of "yes" answers also differs between the two datasets (49% in VLSP vs. 40% in UDST-DurationQA), showing that they follow different data distributions.

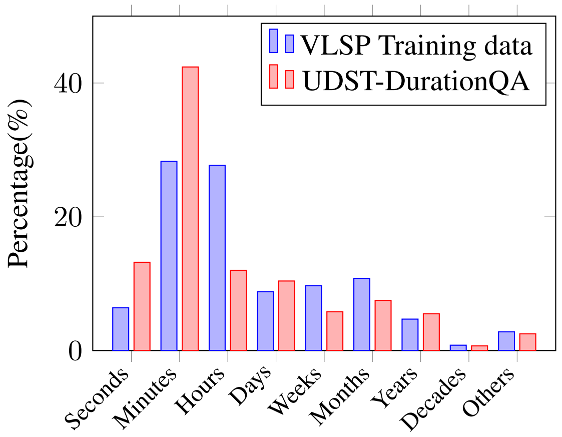

Figure 1: Duration distribution of "yes" answers in VLSP 2025 DurationQA training data and UDST-DurationQA

In detail, Figure 1 illustrates the percentage distribution of "yes" answers across different time units in the VLSP 2025 DurationQA training data and UDST-DurationQA. We observe clear differences in the distributions between the two datasets. While minutes dominate in both datasets, the VLSP data shows a more balanced distribution with a substantially higher proportion of hours compared to UDST-DurationQA. For longer time units such as days, weeks, months, years and decades, the two datasets exhibit relatively similar proportions. In addition to standard time units, the VLSP data also contains expressions categorized as *Others*, such as *"gần như ngay lập tức"* or *"Nó xảy ra ngay lập tức"* which correspond to *"almost instantly"* or *"it takes instantly"* in the UDST-DurationQA. Similarly, expressions like *"mãi mãi"* are aligned with *"it takes forever."*.

VLSP 2025 DurationQA Sample
Context: Tôi đang sửa chữa chiếc xe đạp bị hỏng. **Question:** *Mất thời gian bao lâu để sửa chữa chiếc xe đạp?* **Options:** 30 phút | 1 tháng | 10 phút | 2 giờ **Labels:** yes | no | yes | yes
UDST-DurationQA Sample
Context: Remember that what you do to yourself affects me and everybody else , remember that what you do to me and anyone else shapes your destiny . **Question:** *How long does it take for what you do to me and anyone else to shape your destiny?* **Options:** 1 year | several years | for years | a few hours | for weeks | 12 hours | 6 days **Labels:** yes | yes | yes | no | no | no | no

Table 1: Sample examples from the VLSP 2025 and UDST-DurationQA datasets

Dataset	**# Samples**	**Avg. #Options**	**% "yes" labels**
VLSP training data	1490	4.0	49
VLSP public test	400	4.0	-
VLSP private test	983	4.0	-
UDST-DurationQA	5512	7.0	40

Table 2: Dataset statistics

5 Experiments

5.1 Implementation Details

We fine-tune three models using the Hugging Face's **TRL library**[3]: one based on Qwen2.5 and two based on Qwen3. The Qwen2.5 model is trained with a JSON output format, while the other two are trained with a List output format. For training data, the provided dataset is used to fine-tune Qwen2.5 and one Qwen3 model, whereas the remaining Qwen3 model is trained with the UDST-DurationQA dataset as additional training data.

To fine-tune with QLoRA, specific configurations are applied, including a LoRA rank of 8, LoRA alpha of 16. LoRA target modules are q_proj, k_proj, v_proj, o_proj, up_proj, $down\_proj$, $gate\_proj$. Other training hyperparameters are presented in Table 3.

For our retrieval-guided approach (the few-shot setting), we implement the retriever as follows: we encode each training example by concatenating the context (C) and the question (Q), and then obtaining its vector representation using the multilingual-e5-large model (Wang et al., 2024). At inference time, a cosine similarity search is performed to find the most relevant examples for the current input.

[3] https://huggingface.co/docs/trl/index

Hyperparameter	**Value**
batch_size	8
max_seq_length	512
learning_rate	2e-05
epochs	3 or 5
precision	fp16
optimizer	AdamW

Table 3: Training Hyperparameters

5.2 Inference Process

During inference, each Qwen3 model generates predictions under two configurations: a zero-shot setting (without in-context examples) and a retrieval-guided setting (with few-shot prompting). We use the default generation configurations of each model, with temperature set to 0.7 for Qwen2.5 and 0.8 for Qwen3.

In the few-shot setting, the context and question are concatenated and encoded for cosine similarity retrieval over the training set. We conduct experiments with different values of K ($K \in \{1, 2, 5\}$) on the public test set to assess their influence on model performance. It is noted that with $K = 1$ and $K = 2$, we set the $\mathrm{max\_seq\_length}$ of input prompts to 512 tokens, while for $K = 5$ we increase it to 1024 tokens to accommodate the longer prompts resulting from retrieving more relevant samples. Based on the results from the public test set, we determine the most suitable value of K for experiments on the private test set. To illustrate retrieval guidance, we present qualitative examples of retrieved cases in Appendix A.3.

6 Results

6.1 Public Test Results

Table 4 summarizes the experimental results obtained during the public test phase, comparing different model configurations and K values. We evaluate three main settings: fine-tuned Qwen2.5^{1}, Qwen3 under a zero-shot setting^{2}, and Qwen3 under a retrieval-guided setting^{3,4,5} with varying values of K retrieval samples ($K \in \{1, 2, 5\}$). Among these, Qwen2.5 with JSON-formatted outputs achieves the lowest performance, with an EM of 0.3850 and an F1-score of 0.7157. Qwen3 in the zero-shot setting significantly improves performance, reaching an EM of 0.4425 and an F1-score of 0.7786. Incorporating retrieval guidance further enhances results, with different K values showing varying performance: $K = 1$ yields an EM of 0.4100 and F1-score of 0.7770, $K = 2$ achieves the highest F1-score of 0.7935 (EM of 0.4050), and $K = 5$ slightly lowers F1 to 0.7900 while significantly reducing EM to 0.3675.

Applying majority voting to aggregate predictions from Qwen2.5-7B, Qwen3-8B, and Qwen3-8B with retrieval guidance consistently improves performance across across the range of K values considered. Among them, the ensemble using $K = 2$ achieves the highest F1-score of 0.8076.

Based on these observations, we select $K = 2$ retrieval samples for the subsequent evaluation on the private test set, as it provides the best balance between performance and computational cost while ensuring sufficient context for understanding similar temporal examples without introducing too many retrieval samples.

Model	EM	F1-score
Qwen2.5-7B^{1}	0.3850	0.7157
Qwen3-8B^{2}	0.4425	0.7786
Qwen3-8B w/ retrieval *K*=*1*^{3}	0.4100	0.7770
Qwen3-8B w/ retrieval *K*=*2*^{4}	0.4050	0.7935
Qwen3-8B w/ retrieval *K*=*5*^{5}	0.3675	0.7900
Ensemble 3 models^{1,2,3}	0.4775	**0.8025**
Ensemble 3 models^{1,2,4}	0.4650	**0.8076**
Ensemble 3 models^{1,2,5}	0.4625	**0.8073**

Table 4: Public test results

6.2 Private Test Results

Table 5 reports the submission results on the private test set. It is noted that in this table, we use $K = 2$ determined from the previous test phase. Compared with the public test phase (Table 4), the performance trend remains consistent. Qwen2.5-7B^{1} attains the lowest F1-score of 0.7737, while Qwen3-8B^{2} performs better at 0.7848. Incorporating retrieval guidance^{3} further improves Qwen3-8B to 0.8006. Augmentation with UDST-DurationQA (with or without retrieval guidance)^{4,5} yields mixed results, achieving 0.7911 and 0.7925, respectively. When combining models, the ensemble of three models (Qwen2.5-7B, Qwen3-8B, and Qwen3-8B with retrieval guidance) yields a strong score of 0.8143, which is the best overall performance. By extending the ensemble to all five models, we reach an F1-score of 0.8137.

Precision and recall exhibit a trade-off across different models. Retrieval guidance generally improves recall — for instance, Qwen3-8B with retrieval achieves the highest recall of 0.8896. However, in terms of precision, the highest value (0.7680) is attained when retrieval is combined with data augmentation. While the ensemble of three or five models does not achieve the highest precision or recall individually, it demonstrates the challenge of applying majority voting in ensembles to simultaneously improve both metrics.

The ensemble of three models improves Exact Match to 0.4576, demonstrating the benefit of combining multiple models. However, the ensemble of

Model	EM	Precision	Recall	F1-score
Qwen2.5-7B[1]	0.4352	0.7575	0.7905	0.7737
Qwen3-8B[2]	0.4304	0.7317	0.8461	0.7848
Qwen3-8B w/ retrieval[3]	0.4032	0.7278	0.8896	0.8006
Qwen3-8B w/ aug.[4]	0.4496	0.7459	0.8421	0.7911
Qwen3-8B w/ aug. + retrieval[5]	0.5120	0.7680	0.8187	0.7925
Ensemble 3 models[1,2,3]	0.4576	0.7514	0.8888	**0.8143**
Ensemble 5 models[1,2,3,4,5]	0.4720	0.7543	0.8832	**0.8137**

Table 5: Private test results

five models does not consistently outperform the best single model, as its EM score (0.4720) remains below the highest EM observed (0.5120). This difference in trend compared to F1-score arises because Exact Match requires perfect agreement for the entire answer sequence, making it more sensitive to specific model outputs, whereas F1-score aggregates correctness over individual labels and is therefore less affected by occasional mismatches.

7 Discussion

Model comparison Qwen3 consistently outperforms Qwen2.5. We hypothesize that this improvement can be attributed to the stronger capacity of Qwen3 as well as the efficiency of using the List output format compared with JSON. Under the JSON output format, the model must reproduce many tokens that already appear in the input prompt, rather than focusing on the decisive tokens (e.g., yes/no).

Effectiveness of retrieval guidance The retrieval-guided prompting strategy, which incorporates dynamically selected few-shot examples, provides a consistent boost in F1-score. The additional context helps the model better align with task-specific reasoning patterns found in the training data.

Data augmentation Augmenting the training data with the additional English dataset UDST-DurationQA shows mixed results, suggesting that such cross-lingual augmentation may not always be beneficial in this setting. This effect is largely explained by differences in training data composition. UDST-DurationQA contains far more examples than the VLSP 2025 training data, and its questions typically include more candidates with a lower proportion of "yes" answers. In addition, differences in how the two datasets were constructed may introduce subtle stylistic mismatches, further contributing to the inconsistency.

Ensemble methods Ensemble methods tend to stabilize predictions and yield overall stronger results, as they combine the strengths of different models and reduce the variance associated with individual predictions.

Error analysis For a detailed analysis of typical failure cases, please refer to Appendix A.4.

8 Conclusion

Reasoning about event duration in Vietnamese is a challenging problem due to diverse linguistic realizations and the need for commonsense inference. In this work, we introduce a retrieval-guided fine-tuning framework that integrates fine-tuned LLMs, a dynamic retrieval mechanism over the training data, and ensemble strategies to address Sub-Task 2 of the VLSP 2025 Temporal QA Challenge. Through extensive experiments, we show that our approach consistently improves over strong baselines, with ensembles providing the most stable and competitive results. The retrieval-guided approach demonstrates the effectiveness of leveraging similar training examples to guide model reasoning. Our findings highlight the promise of using retrieved, in-distribution examples to guide LLM reasoning for specialized tasks like event duration question answering, while also revealing the open challenges that remain in cross-lingual augmentation, motivating future research on more robust retrieval strategies and improved multilingual adaptation.

Acknowledgments

We sincerely thank the VLSP 2025 organizers for preparing the dataset and setting up the evaluation system, which made it possible for us to carry out this work and share the results with the community.

References

Agnar Aamodt and Enric Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 261 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

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, Kang Guan, Jiaqi Wang, Damai Dai, Kai Dong, Liyue Zhang, Yuli Gan, Fangjun Li, Hongxuan Zhang, and 21 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, and 231 others. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Daohan Ge, Dayiheng Liu, Dejian Yang, Fei Huang, Guanting Dong, Haoran Wei, Haowei Zhang, Huan Lin, Jian Yang, and 36 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

A Appendix

A.1 Prompt (JSON Output Format) for Qwen2.5-7B

Prompt

Bạn là Option Judge. Đánh nhãn cho từng phương án dựa trên:
- Ngữ cảnh/câu hỏi sau (tiếng Việt).
- Bản tóm tắt (distilled evidence) dưới đây.
- KHÔNG quy đổi đơn vị thời lượng sang số học; chỉ so sánh theo bậc đơn vị (giây < phút < giờ < ngày < tuần < tháng < năm) và theo ngôn ngữ tự nhiên.

Ngữ cảnh: {context}
Câu hỏi: {question}

Distilled evidence:
""""""
Các phương án: {opts_join}

Hãy trả về MỘT JSON ARRAY, mỗi phần tử tương ứng một phương án, theo schema:
{
"option": string,
"label": "yes" | "no",
}

Chỉ xuất JSON array, không giải thích thêm.

A.2 Prompt (List Output Format) for Qwen3-8B

Prompt

BẠN LÀ CHUYÊN GIA ĐÁNH GIÁ THỜI GIAN cho câu hỏi tiếng Việt.
NHIỆM VỤ: Đánh giá từng phương án thời gian có HỢP LÝ trong thực tế hay không (yes | no) cho hoạt động được mô tả.

VÍ DỤ THAM KHẢO:
—
{examples}

—
Ngữ cảnh: {context}
Câu hỏi: {question}

Các phương án: {opts_join}

Hãy trả về MỘT ARRAY, mỗi phần tử tương ứng một phương án theo thứ tự, theo schema:
["yes", "no", ...]
Chỉ xuất ARRAY, không giải thích thêm.

A.3 Qualitative Retrieval Examples

qid	Public test sample	Retrieved sample	Similarity type
#22	**Context:** A group of artists began organizing an art exhibition in a public space. They worked tirelessly to prepare for the event, finalizing the artworks and promoting the exhibition. **Question:** *How long does it take to organize an art exhibition in a public space?*	**Context:** A group of young people decided to organize an art event in the city park. They planned performances, street art, and food stalls. Everyone was excited and contributed to making the event a success. **Question:** *How long does it take to organize an art event in the city park?*	Topical relatedness and Syntactic similarity
#150	**Context:** A local radio station is striving to develop a new program aimed at raising awareness of environmental issues. They organized many workshops and discussions to find the best approach. **Question:** *How long does it take to develop a new program to raise awareness of environmental issues?*	**Context:** A research group is seeking solutions to the problem of air pollution in the city. They held many meetings and discussions to find ways to improve the situation. **Question:** *How long does it take for a research group to find a solution to the problem of air pollution?*	Topical relatedness, Syntactic and Semantic similarity
#10	**Context:** A theater has just premiered a new play that has attracted a large audience. The play offers an emotional experience for viewers and has received much praise from critics. **Question:** *How long does it take for a play to attract a large audience?*	**Context:** A newly opened store has attracted many customers. They held a grand opening with promotions and games. Customers were excited and enjoyed the new products on the shelves. Many returned to shop again after attending the event. **Question:** *How much time is needed to attract a large number of customers?*	Semantic similarity

Table 6: Examples of retrieved cases in VLSP public test. Original Vietnamese samples are translated into English for readability.

We provide three representative cases in VLSP public test in table 6. The first example illustrates both topical relatedness (art-related events) and syntactic similarity, as the questions share the same structure ("How long does it take"). The second example extends beyond syntactic overlap and exhibits semantic similarity, since both questions inquire about the time required to achieve outcomes in the environmental domain. The third example, in contrast, has no topical or syntactic similarity — the contexts belong to different domains and the questions use different structures. Their connection lies in expressing a similar meaning, specifically asking about the time needed to attract a large number of people. These types of similarity - topical, syntactic, and semantic - contribute to improving model performance by providing relevant context that helps the model produce more accurate results.

A.4 Error Analysis

qid	Example	Error analysis
#94	**Context:** A group of friends decided to organize an outdoor party over the weekend. They planned everything from food, music to entertainment activities to have a joyful day together. **Question:** *How long does it take to organize a perfect outdoor party?* **Options:** 3 weeks | 1 month | 6 weeks | 2 weeks **Labels:** no | no | no | yes **Prediction:** yes | no | no | no	Label ambiguity or vague question meaning; both "2 weeks" and "3 weeks" are reasonable answers, making the correct label unclear.
#65	**Context:** At a press conference, a reporter asked about the environmental crisis. A representative from an environmental protection organization emphasized the urgency of taking action to save the planet. **Question:** *How long does it take to implement an effective environmental protection campaign ?* **Options:** 3 months | 1 month | 5 years | 6 weeks **Labels:** yes | no | no | yes **Prediction:** yes | yes | no | no	Unable to differentiate between "weeks" and "months", treating them as distinct categories without considering their approximate equivalence.
#128	**Context:** A group of reporters is investigating a major scandal involving a large corporation. They must collect information and interview many people to clarify the truth. **Question:** *How long does it take for the reporters to collect enough information and interview everyone ?* **Options:** 2 weeks | 5 months | 6 weeks | 3 weeks **Labels:** yes | no | no | yes **Prediction:** yes | no | yes | no	Require deep domain knowledge in criminal investigation, the model struggles to estimate realistic time for completing the task.
#30	**Context:** A group of journalists is investigating a major scandal involving a multinational company. They have collected extensive documents and evidence to support their article. **Question:** *How long does it take for the group of journalists to complete the article about the scandal ?* **Options:** 2 weeks | 5 days | 1 month | 3 weeks **Labels:** no | no | yes | yes **Prediction:** yes | yes | no | no	Both involve specialized domain knowledge and ambiguous labeling.
#37	**Context:** During a performing arts event, a dance group delivered impressive performances that captivated the audience. They conveyed emotions through every dance step and melody, creating a vibrant and energetic atmosphere. **Question:** *How long does it take to create these artistic performances ?* **Options:** 2 months | 5 months | 3 hours | 4 months **Labels:** yes | no | no | no **Prediction:** yes | yes | no | yes	Struggle to distinguish between similar temporal durations within the same unit (e.g., months), fail to consistently recognize differences between "2 months", "5 months", and "4 months" in context.

Table 7: Investigation of errors in failure cases. Original Vietnamese samples are translated into English.