DP-Ens DurationQA: Dual-Prompt Fine-Tuning with Log-Probability Ensemble for Duration Question Answering

Van Thai Le[1*] **Dong Duy Nguyen**[1*]
[1]Ton Duc Thang University
vanlethai12042002@gmail.com, duydong.tdtu@gmail.com

Ngoc Bui Lam Quang[2†]
[2]AI Viet Nam
ngoc.bui150019@vnuk.edu.vn

Abstract

This paper presents DP-Ens DurationQA, a method for Duration Question Answering (DurationQA), a multi-label temporal QA task where multiple candidate durations may be correct. DurationQA is challenging because it requires multi-step reasoning and consistent prediction across multiple candidate labels. To address this, the proposed approach fine-tunes a single LLM under two complementary prompting strategies: Chain-of-Thought (CoT) to generate intermediate reasoning steps, and Refinement to predict labels conditioned on these reasoning traces. At inference, outputs from both prompts are combined via a numerically stable log-probability ensemble, producing reliable binary labels for each candidate option. Experiments on the VLSP 2025 DurationQA benchmark show that the proposed ensemble approach outperforms single-prompt and non-ensemble baselines, achieving state-of-the-art F1 while maintaining efficient computation with a quantized 4-bit model. The results demonstrate that combining reasoning and reasoning-conditioned label prediction can enhance multi-label temporal QA performance without increasing model parameters.

1 Introduction

Duration Question Answering (DurationQA) is a subtask of temporal question answering that requires predicting the duration of events mentioned in a textual context. Unlike standard QA, multiple candidate durations can be correct simultaneously, making it a multi-label prediction problem and increasing the complexity of reasoning.

Temporal QA has attracted growing attention due to applications in timeline construction, event understanding, and knowledge extraction (Tan et al., 2023; Virgo et al., 2022).

Existing approaches include rule-based reasoning (Harabagiu and Bejan, 2005), neural sequence models (Dhingra et al., 2022; Wang et al., 2023a), and statistical models (Berberich et al., 2010). However, these methods often struggle to reason accurately across multiple candidates while maintaining label consistency (Ton et al., 2025; Zhao et al., 2025).

Recently, Large Language Models (LLMs) have shown strong capabilities in generating intermediate reasoning steps (Wei et al., 2023a,b), opening a promising direction for multi-label temporal QA.

Predicting consistent labels in DurationQA presents several challenges.

First, candidate durations may be conflicting (Tan et al., 2024) or overlapping (Virgo et al., 2022), which requires careful reasoning over context.

Second, many questions demand multi-step reasoning to infer the correct durations. Single-step prediction models often fail in such cases. For example, GPT-4 (OpenAI et al., 2024) may miss implicit contextual clues or overlook some valid answers in multi-answer settings. This limitation leads to lower accuracy compared to tasks with explicit, single-step temporal information (Tan et al., 2024).

Third, maintaining consistency across multiple labels remains difficult, particularly when using generative models that produce free-form outputs (Ton et al., 2025; Zhao et al., 2025).

These challenges highlight the need for methods that integrate explicit reasoning with structured label prediction.

To address these challenges, this study introduces DP-Ens DurationQA, a framework that employs a single LLM with two complementary prompting strategies.

[*]Equal contribution (first author)
[†]Corresponding author

Chain-of-Thought (CoT) prompting (Wei et al., 2023a) elicits intermediate reasoning steps, encouraging the model to capture temporal dependencies between events.

Refinement prompting then predicts binary labels for each candidate duration, conditioned on the reasoning traces. This step bridges free-form reasoning with structured outputs (Yun et al., 2025; Shen et al., 2025).

During inference, outputs from both prompts are integrated through a numerically stable log-probability ensemble, which preserves relative confidence scores and stabilizes predictions.

This design enables a quantized 4-bit LLM (Wang et al., 2025) to handle both reasoning generation and label prediction without expanding model size. It achieves a balance between exploratory reasoning and reliable decision-making.

The main contributions of this work are:

- Introducing a dual fine-tuning strategy with CoT and Refinement prompts for multi-label DurationQA, effectively combining intermediate reasoning and consistent label prediction.
- Proposing a numerically stable log-probability ensemble to merge CoT and Refinement predictions, preserving relative confidence scores and stabilizing final outputs.
- Empirical validation on the VLSP 2025 DurationQA benchmark, showing that the ensemble improves F1 and overall accuracy over single-prompt or non-ensemble baselines.

2 Related Work

2.1 Temporal QA and temporal commonsense

Research on temporal reasoning in NLP spans event ordering, implicit events, conversational temporal phenomena, and time-sensitive facts. TORQUE targets temporal ordering questions over news passages (Ning et al., 2020), while TRACIE evaluates reasoning over implicit events (Zhou et al., 2021). In dialogue settings, TIMEDIAL frames temporal commonsense as a multiple-choice cloze task and reveals sizable model–human gaps (Qin et al., 2021). More comprehensively, TIMEBENCH proposes a hierarchical benchmark covering diverse temporal phenomena and documents persistent performance gaps between SOTA LLMs and humans (Chu et al., 2024). Complementing these, TIMEQA focuses on questions grounded in time-evolving facts, probing both temporal understanding and reasoning (Chen et al., 2021).

2.2 Duration reasoning and multi-label formats

MC-TACO formalizes five temporal commonsense categories—including *duration*—and adopts a multi-label setup in which multiple options may be plausible for a single question (Zhou et al., 2019). This design aligns closely with DURATIONQA in VLSP 2025 (VLSP Organizing Committee, 2025), which labels each option as yes/no for plausibility (VLSP Organizing Committee, 2025; Chu et al., 2024).

2.3 Reasoning with rationales

Reasoning with intermediate rationales includes prompting-based and learning-based approaches. Chain-of-Thought (CoT) improves complex reasoning when models are sufficiently large (Wei et al., 2023a); Self-Consistency samples diverse traces and marginalizes to a stable answer (Wang et al., 2023c); Least-to-Most decomposes problems into ordered sub-problems (Zhou et al., 2023). Beyond prompting, methods like STAR generate, filter, and fine-tune on rationales yielding correct answers.(Zelikman et al., 2022); DISTILLING STEP-BY-STEP shows that small models trained with rationale supervision can outperform few-shot LLMs using less data (Hsieh et al., 2023); SCOTT distills self-consistent CoT from a large teacher to a smaller student, improving faithfulness and downstream accuracy (Wang et al., 2023b). In multi-label duration settings like DURATIONQA, option-wise rationales can regularize the mapping from context–question to plausibility judgments, complementing label-only supervision (Zhou et al., 2019; VLSP Organizing Committee, 2025).

2.4 Positioning of this work

Following the VLSP setup and its links to TIMEBENCH, only the training split is augmented with Gemini-generated free-form CoT rationales (Team et al., 2025), while the test split remains reasoning-free to avoid information leakage (Chu et al., 2024; VLSP Organizing Committee, 2025). At inference, base predictions are combined

Thank AI Vietnam for financial support.

with rationale-conditioned predictions via a log-probability ensemble. This design aligns with findings that aggregating multiple reasoning views (e.g., Self-Consistency) better captures option-level plausibility in multi-label QA (Wang et al., 2023c; Zelikman et al., 2022; Hsieh et al., 2023; Wang et al., 2023b).

3 Methodology

3.1 Task Formulation

Sub-Task 2: Duration Question Answering (DurationQA) from the VLSP 2025 Challenge on Temporal QA (VLSP Organizing Committee, 2025) is considered. Given a context c and a question q about the duration of an event, each candidate option $o_i \in \mathcal{O}$ is assigned a binary label $y_i \in \{\text{yes}, \text{no}\}$ indicating whether it is a plausible answer.

Evaluation uses the official metrics: Exact Match (EM), Precision, Recall, and F1. EM counts predictions as correct only if the entire label sequence matches the ground truth. Precision is the proportion of correctly predicted "yes" answers among all "yes" predictions, Recall is the proportion of correctly predicted "yes" answers among all actual "yes" labels, and F1 is their harmonic mean.

All metrics are computed at the option level and averaged over all questions.

Example.

"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"]
(English translation: "Context": "I am repairing the broken bicycle." "Question": "How long does it take to repair the bicycle?" "Options": ["30 minutes", "1 month", "10 minutes", "2 hours"] "Labels": ["yes", "no", "yes", "yes"])

3.2 Model Overview

The DurationQA task is cast into a multi-label sequence prediction framework, where the system receives the full set of candidate options and outputs a corresponding sequence of binary labels. Unlike traditional multiple-choice QA, this setting allows multiple options to be labeled yes.

To improve reasoning ability, the model is fine-tuned under two complementary prompting strategies: (i) a *Chain-of-Thought (CoT)* (Wei et al., 2023a) setting, where the model produces intermediate reasoning before generating the labels, and (ii) a *Refinement* (Yun et al., 2025; Shen et al., 2025) setting, where the model is provided with reasoning traces in the input and is trained to map them directly into the final label sequence. This dual setup bridges free-form reasoning and consistent prediction.

During inference, predictions from CoT and Refinement are combined using a log-probability ensemble. A subsequent softmax normalization produces a binary (yes/no) distribution for each candidate option. This approach balances the exploratory reasoning of CoT with the stability of Refinement and mitigates the effect of extreme probabilities, yielding more reliable final labels.

An overview of the proposed framework is illustrated in Figure 1.

3.3 Fine-tuning Strategy

For each training example, two complementary input formats are constructed to fully leverage the reasoning capabilities of the model.

The first format, **CoT input**, prompts the model to generate intermediate reasoning steps (Chain-of-Thought) before producing the final label.

The second format, **Refinement input**, provides the model with reasoning traces derived from the CoT input and requires it to predict the final label only, without generating additional reasoning.

Both types of inputs are used jointly to fine-tune the model in a single training process. Supervised learning with cross-entropy loss over the target label tokens is employed.

By combining these two complementary approaches within a single model, the system can simultaneously handle free-form reasoning and reasoning-conditioned classification, reducing inconsistencies that may arise when using CoT alone and achieving more stable and accurate predictions across all candidate options, without increasing the number of trainable parameters.

3.4 Inference and Ensemble

At inference time, the finetuned model is queried with both CoT and Refinement prompts. For each candidate option, we extract the log-probability of the token actually generated by the model within <labels> </labels>. If the generated token is yes, we take p_{yes} directly from the model logits and

Thank AI Vietnam for financial support.

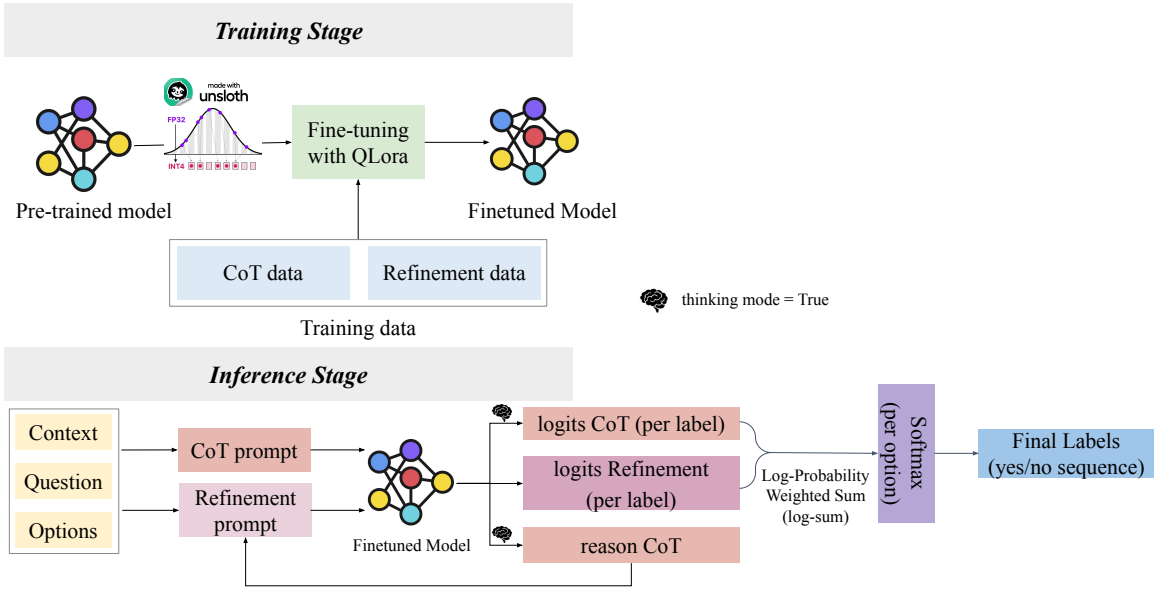

Figure 1: Overview of our framework for DurationQA, consisting of a training stage (fine-tuning a 4-bit Qwen model with CoT and Refinement prompts) and an inference stage (combining predictions via log-probability ensemble).

set $p_{\text{no}} = 1 - p_{\text{yes}}$; if the generated token is no, we take p_{no} from the logits and set $p_{\text{yes}} = 1 - p_{\text{no}}$. This approach reduces memory usage and computation, as only one logit per option needs to be retrieved.

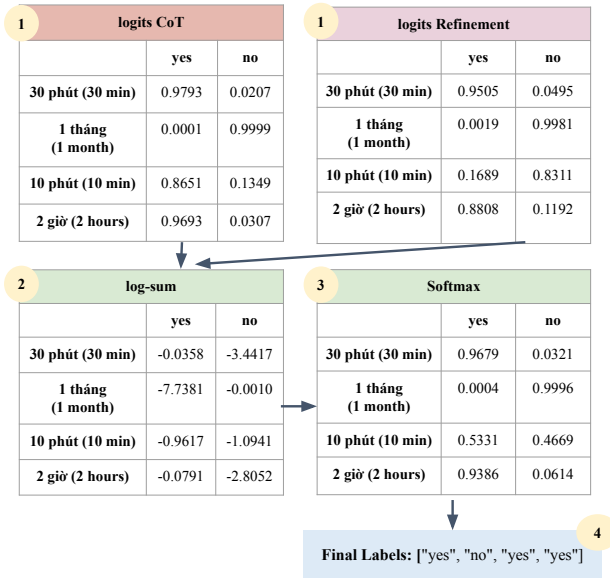

Figure 2: Logits from CoT and Refinement are combined with a weight $w = 0.5$, passed through a binary softmax, and converted into the final yes/no label sequence for all options.

The ensemble logits are computed as a weighted sum in log space:

$$\ell_{\text{yes}} = (1 - w) \cdot \log p_{\text{CoT}}(\text{yes}) + w \cdot \log p_{\text{Ref}}(\text{yes}),$$

$$\ell_{\text{no}} = (1 - w) \cdot \log p_{\text{CoT}}(\text{no}) + w \cdot \log p_{\text{Ref}}(\text{no}),$$

where $w \in [0, 1]$ is the weight assigned to the Refinement model. For each option, a binary softmax is applied to $(\ell_{\text{yes}}, \ell_{\text{no}})$ to obtain normalized probabilities, and the final prediction is a binary label sequence assigned to all candidate options.

Since each label token is generated from the same model under different prompts, the raw probabilities are not directly comparable. The CoT prompt produces a token as part of a reasoning chain, while the Refinement prompt generates it conditioned on the refined rationale. Converting probabilities to log space allows a numerically stable weighted combination, preserving the relative confidence of each prompt while preventing one from dominating due to scale differences.

Figure 2 illustrates this process, showing how CoT and Refinement logits are combined to produce the final label sequence.

4 Experiments and Results

4.1 Dataset

The official VLSP 2025 – Temporal QA, Sub-task 2: Duration Question Answering (durationQA) dataset in Vietnamese.(VLSP Organizing Committee, 2025) Each instance contains a *context*, a *question* about an event's duration, a list

Thank AI Vietnam for financial support.

of candidate *options*, and corresponding binary *labels* ("yes"/"no"). The dataset was constructed by translating English material from *TimeBench* and extending it via GPT-based generation to mirror the original structures and semantics.(VLSP Organizing Committee, 2025)

For our experiments, the released training set is split into **85% training** and **15% testing**:

Split	Train	Test
# Instances	1,250	221

Chain-of-Thought augmentation. To enrich the supervision signal, we additionally generate natural language rationales for each candidate option using *Gemini 2.0 Flash* (Team et al., 2025). Given the full input (context, question, options, and labels), the model produces a step-by-step justification of why each option is or is not a plausible duration.
For example:

Example Instance

Context: Cô ấy đang chuẩn bị tài liệu cho buổi họp.
(She is preparing documents for a meeting)
Question: Mất bao lâu để chuẩn bị tài liệu cho buổi họp?
(How long does it take to prepare the documents for the meeting?)
Options: [15 phút (15 minutes), 30 phút (30 minutes), 2 tháng (2 months), 3 tháng (3 months)]
Labels: [yes, yes, no, no]
Reasoning (generated):
– 15 phút (15 minutes): Có thể đủ nếu tài liệu đã sẵn sàng, chỉ cần in ấn hoặc chỉnh sửa nhỏ (Reasonable if the documents are already prepared, requiring only minor editing or printing.)
– 30 phút (30 minutes): Cũng hợp lý cho chuẩn bị đơn giản, như rà soát và in ấn tài liệu. (Also plausible for simple preparation such as reviewing and printing.)
– 2 tháng (2 months): Quá dài, chỉ phù hợp cho các dự án lớn cần nghiên cứu chuyên sâu. (Too long; only realistic for large projects requiring extensive research.)
– 3 tháng (3 months): Tương tự như 2 tháng, không phù hợp với ngữ cảnh thông thường. (Similar to 2 months; unrealistic for a regular meeting.)

These rationales serve as auxiliary training signals in our method, while the evaluation strictly follows the official labels.

4.2 Experimental Setup

We experiment with four standard prompting paradigms: *Zero-shot*, *Zero-shot CoT*, *Few-shot*, and *Few-shot CoT*, applied consistently to both Gemini-2.5-Pro (Team et al., 2025) and Qwen3-4B-Thinking-2507 (Yang et al., 2025). For fine-tuning, we use the quantized version unsloth/Qwen3-4B-Thinking-2507-unsloth-bnb-4bit to reduce memory consumption, and train with different supervision signals, including labels only, chain-of-thought (CoT), refinement prompts, and their combinations. In particular, the best-performing setting integrates both CoT and refinement, and ensembles their predicted label probabilities.

Training employs trl's SFTTrainer on a single Kaggle P100 GPU for two epochs, batch size 2 with gradient accumulation 4, learning rate 2×10^{-5}, AdamW-8bit optimizer (weight decay 0.01), linear scheduler with warmup 0.03, mixed precision fp16, and fixed seed 3407.

Models are evaluated on Precision, Recall, F1, Exact Match, as well as inference and total training time.

4.3 Experimental Results

Table 1 summarizes the performance of different models and training strategies on DurationQA. Among zero-shot and few-shot settings, Gemini-2.5-Pro (Team et al., 2025) achieves strong precision. However, its chain-of-thought (CoT) variants show reduced recall and F1, suggesting that naive CoT prompting does not always improve performance. For Qwen3-4B-Thinking-2507, zero-shot and few-shot variants underperform compared to fine-tuned models, highlighting the benefits of task-specific supervision.

Fine-tuning with label-only supervision yields substantial gains across all metrics. CoT-based fine-tuning further enhances the model's reasoning capabilities. Using refinement at inference alone, however, can produce unstable outputs, as missing or extra labels reduce overall scores.

Jointly fine-tuning with both CoT and refinement prompts leads to more consistent predictions. The proposed log-probability ensemble, which combines outputs from both branches, achieves the best F1 and recall. Combining CoT, refinement, and ensembling yields more accurate, reliable DurationQA. The table also shows that these improvements incur only modest additional inference cost, while training time remains manageable.

4.4 Ablation on Ensemble Weight

To study the impact of the ensemble weight on model performance, the interpolation parameter w is varied between the CoT and Refinement branches from 0 (only CoT) to 1 (only Refinement). For each weight, we compute Precision, Recall, F1, and Exact Match on the DurationQA validation set, with results shown in Fig. 3.

The ablation shows that combining predictions from both CoT and Refinement consistently improves F1 compared to using either branch alone. In particular, equal weighting ($w = 0.5$) achieves the highest F1 and Recall while

Thank AI Vietnam for financial support.

Model	Setting	Prec.	Rec.	F1	EM	Inf. Time	Train Time
Gemini-2.5-Pro	Zero-shot	**87.86**	69.25	77.45	57.47	11.19	–
	Zero-shot CoT	79.05	37.81	51.16	18.55	24.98	–
	Few-shot	84.59	68.79	75.88	56.56	11.52	–
	Few-shot CoT	83.81	60.14	70.03	50.68	11.36	–
Qwen3-4B-Thinking-2507	Zero-shot	72.89	47.15	57.26	15.84	25.38	–
	Zero-shot CoT	65.83	24.75	48.89	13.51	26.47	–
	Few-shot	68.42	8.88	15.73	0.45	27.52	–
	Few-shot CoT	51.52	3.87	7.20	0.90	30.09	–
	Finetune (labels only)	84.23	85.19	84.71	74.21	14.12	1h 15m
	Finetune (CoT only, no refine @ inf.)	81.98	82.92	82.45	68.33	24.13	1h 39m
	Finetune (CoT only, refine @ inf.)	48.79	39.74	42.68	28.05	31.71	1h 39m
	Finetune (CoT+Refine, no refine @ inf.)	84.84	85.42	85.13	**75.11**	24.02	5h 03m
	Finetune (CoT+Refine, refine @ inf.)	85.06	84.28	84.67	73.76	26.45	5h 03m
	Ensemble (CoT+Refine, ours)	85.94	**86.33**	**86.14**	74.66	26.45	5h 03m

Table 1: Experimental results on DurationQA. Precision, Recall, F1, and Exact Match (EM) are reported, together with inference time per sample and total training time. All Qwen fine-tuned variants are trained from the quantized checkpoint unsloth/Qwen3-4B-Thinking-2507-unsloth-bnb-4bit. *Finetune (labels only)* uses supervision on labels only. *Finetune (CoT only)* is trained with chain-of-thought prompts, with inference either without refinement ("no refine @ inf.") or with refinement ("refine @ inf."). *Finetune (CoT+Refine)* jointly trains with CoT and refinement prompts, and predictions are taken either from CoT alone ("no refine @ inf.") or from refinement alone ("refine @ inf."). *Ensemble (ours)* combines both branches by interpolating their log-probabilities with equal weights ($w = 0.5$), followed by softmax normalization.

maintaining competitive Precision and Exact Match. This confirms that both reasoning paths provide complementary information, and that combining log-probabilities is sufficient to exploit them effectively.

Moreover, Exact Match remains stable around $w = 0.5$, while Precision, Recall, and F1 vary slightly, indicating that the ensemble is relatively robust to small changes in the weight.

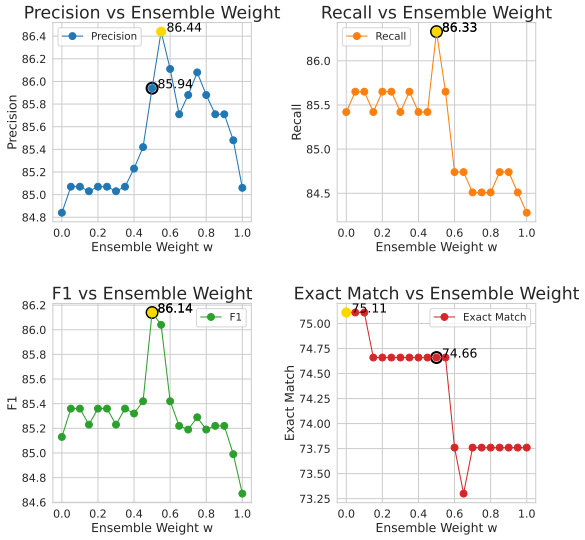

Figure 3: Performance metrics vs ensemble weight for Qwen3-4B-Thinking-2507. Highlighted best values (gold) and our method (w=0.5, black circle).

4.5 Full-dataset Training and Official Evaluation

While the above experiments use an 85%/15% split of the released training data to analyze model variants, the best-performing configuration is additionally fine-tuned ($w = 0.5$) on the **entire released training set (100%)** to prepare for the official VLSP 2025 evaluation.

As **Softmind_AIO**, predictions are then submitted to both the public and private test sets provided by the organizers. Table 2 reports the scores on the official leaderboards.

We achieve strong performance across Precision, Recall, F1, and Exact Match on the public test set, and maintain competitive results on the private test set, demonstrating that the approach generalizes beyond the held-out split used during development.

Split	Prec.	Rec.	F1	EM
Public Test	75.59%	90.14%	82.23%	50.00%
Private Test	70.28%	90.33%	79.06%	34.08%

Table 2: Performance of our best model (Softmind_AIO) trained on the full dataset, evaluated on the official VLSP 2025 public and private test sets.

4.6 Error Analysis on the Public Test Set

On the Public Test set, recall is relatively high while precision is lower (Table 2). This indicates that the model often predicts more candidate durations as correct than the ground truth, yielding false positives, and occasionally misses ground-truth

Thank AI Vietnam for financial support.

correct options, leading to false negatives. These behaviors explain the precision–recall trade-off: the model favors coverage (higher recall) at the cost of stricter selectivity (lower precision).

QID	Case
5	**Context:** Một nhóm phóng viên đang chuẩn bị cho một cuộc phỏng vấn độc quyền với một nhân vật nổi tiếng. Họ phải thu thập thông tin, lên kế hoạch và thực hiện nhiều công đoạn để đảm bảo cuộc phỏng vấn diễn ra thành công. (A group of journalists was preparing for an exclusive interview with a famous figure. They had to collect information, plan, and carry out several tasks to ensure the interview's success.) **Question:** Mất bao lâu để nhóm phóng viên chuẩn bị cho cuộc phỏng vấn độc quyền? (How long did it take the journalists to prepare for the exclusive interview?) **Options:** 3 ngày (3 days), 1 tuần (1 week), 12 giờ (12 hours), 5 ngày (5 days) **Ground truth/ Prediction** [yes, no, no, yes] / [yes, yes, no, yes]
45	**Context:** Trong một thành phố lớn, một nhóm tình nguyện viên đang làm việc để tổ chức buổi lễ trao quà cho những trẻ em nghèo. Họ đã chuẩn bị rất nhiều món quà và cần sắp xếp mọi thứ chu đáo. (In a large city, a group of volunteers was organizing a gift-giving event for underprivileged children. They prepared many gifts and needed to arrange everything carefully.) **Question:** Mất bao lâu để tổ chức buổi lễ trao quà? (How long did it take the volunteers to organize the event?) **Options:** 2 tuần (2 weeks), 4 tuần (4 weeks), 1 tháng (1 month), 6 tháng (6 months) **Ground truth/ Prediction** [yes, yes, no, no] / [yes, yes, yes, no]

Table 3: Representative cases on the Public Test set.

As Table 3 shows, QID 5 is a typical false positive: the model predicts an extra option ("1 week") that is not marked as correct in the ground truth. This prediction is still plausible, as 7 days is an ambiguous duration — neither very short nor very long — and fits the context. QID 45 illustrates another false positive: the model predicts an extra option ("1 month"), which is close to 4 weeks (28 days) and thus also somewhat ambiguous, making it plausible even though it is not selected in the ground truth.

These cases highlight the challenge of reaching consensus on predictions for ambiguous durations.

Overall, this explains the precision–recall trade-off: recall is relatively high due to inclusive predictions, while precision suffers from extra incorrect predictions, resulting in higher F1 than EM and emphasizing the need for better calibration.

5 Conclusion

This paper introduced a dual-prompt fine-tuning approach for DurationQA, combining Chain-of-Thought reasoning with a Refinement prompt. Predictions are combined through a log-probability ensemble for more stable multi-label outputs. Experiments on the DurationQA benchmark show consistent gains over single-prompt and non-ensemble baselines, achieving higher F1 scores without increasing model size.

Error analysis revealed frequent false positives—predicted durations that are plausible but not selected in the ground truth—and difficulties in reaching consensus on ambiguous durations, leading to missed correct options. Future work will address these issues by improving calibration to better balance precision and recall, extending to timeline extraction where multiple spans interact, and testing robustness in multilingual and domain-specific settings.

References

Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. 2010. A language modeling approach for temporal information needs. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, ECIR'2010, page 13–25, Berlin, Heidelberg. Springer-Verlag.

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

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. *Preprint*, arXiv:2311.17667.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Sanda Harabagiu and Cosmin Adrian Bejan. 2005. Question answering based on temporal inference. In *Proceedings of the AAAI-2005 Workshop on Inference for Textual Question Answering*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *Preprint*, arXiv:2305.02301.

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

OpenAI, 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, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Thank AI Vietnam for financial support.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. *Preprint*, arXiv:2106.04571.

Yuanzhe Shen, Zisu Huang, Zhengkang Guo, Yide Liu, Guanxu Chen, Ruicheng Yin, Xiaoqing Zheng, and Xuanjing Huang. 2025. Intentionreasoner: Facilitating adaptive llm safeguards through intent reasoning and selective query refinement. *Preprint*, arXiv:2508.20151.

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

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2024. Towards robust temporal reasoning of large language models via a multi-hop qa dataset and pseudo-instruction tuning. *Preprint*, arXiv:2311.09821.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. 2025. Understanding chain-of-thought in LLMs through information theory. In *Forty-second International Conference on Machine Learning*.

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.

VLSP Organizing Committee. 2025. Vlsp 2025 evaluation campaign – temporal question answering (tempqa), sub-task 2: Duration question answering. https://vlsp.org.vn/vlsp2025/eval/tempqa.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023a. Bitimebert: Extending pre-trained language representations with bi-temporal information. SIGIR '23, page 812–821, New York, NY, USA. Association for Computing Machinery.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. Scott: Self-consistent chain-of-thought distillation. *Preprint*, arXiv:2305.01879.

Weilan Wang, Yu Mao, Dongdong Tang, Hongchao Du, Nan Guan, and Chun Jason Xue. 2025. When compression meets model compression: Memory-efficient double compression for large language models. *Preprint*, arXiv:2502.15443.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023b. MenatQA: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447, Singapore. Association for Computational Linguistics.

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.

Taewon Yun, Jihwan Oh, Hyangsuk Min, Yuho Lee, Jihwan Bang, Jason Cai, and Hwanjun Song. 2025. Refeed: Multi-dimensional summarization refinement with reflective reasoning on feedback. *Preprint*, arXiv:2503.21332.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. 2025. Is chain-of-thought reasoning of llms a mirage? a data distribution lens. *arXiv preprint arXiv:2508.01191*.

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. *Preprint*, arXiv:1909.03065.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. *Preprint*, arXiv:2010.12753.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Thank AI Vietnam for financial support.