VLSP 2025 challenge: Numerical Reasoning Question and Answer

Le Ngoc Toan, Ha My Linh, Pham Thi Duc, Ngo The Quyen, Nguyen Thi Minh Huyen
VNU University of Science, Hanoi Vietnam
{lengoctoan, hamylinh, phamthiduc ngoquyenbg, huyenntm}@hus.edu.vn
Correspondence: lengoctoan@hus.edu.vn

Abstract

The VLSP 2025 Shared Task on Numerical Reasoning Question Answering (NumQA) is the first initiative to address numerical reasoning in Vietnamese financial texts. To support this effort, we constructed ViNumQA, a large-scale benchmark dataset comprising over 4,000 manually validated question-program-answer triples. The dataset integrates two complementary sources: a human-verified Vietnamese translation of FinQA and newly constructed QA pairs derived from domestic corporate financial reports. Each instance requires systems to generate a transparent mathematical reasoning program and produce a final numerical answer, enabling explicit evaluation of both reasoning correctness and result accuracy. The shared task included two subtasks: (1) a constrained track focusing on efficient, reproducible modeling without external APIs, and (2) an unconstrained track allowing LLM-assisted training. The best-performing constrained model achieved the highest Program Accuracy ($PA = 76.6\%$), while an inference-only agent attained the highest Execution Accuracy ($EA = 84.0\%$) without fine-tuning. By releasing ViNumQA and evaluating multiple methods, this work provides a key resource for Vietnamese financial NLP and reveals the balance between interpretability and accuracy in numerical reasoning systems.

1 Introduction

The increasing availability of digital financial documents—such as annual reports, balance sheets, and corporate disclosures—has created an urgent demand for intelligent systems capable of understanding and reasoning over numerical information. Within this domain, Numerical Reasoning Question Answering (NumQA) plays a central role in enabling users to query complex data and extract quantitative insights directly from unstructured and semi-structured sources. Unlike conventional QA tasks, NumQA requires models not only to comprehend financial text and tables but also to perform mathematical operations and generate explicit reasoning programs that can be verified for correctness.

Significant progress has been achieved in English, driven by benchmark datasets like FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021), which integrate textual and tabular financial data and have spurred the development of program-based reasoning models. These works inspired a variety of neural semantic parsers and reasoning architectures, including models that generate symbolic programs (Mishra et al., 2022; Zhao et al., 2022) or employ chain-of-thought prompting for numerical inference (Wei et al., 2022). Beyond the financial domain, several datasets such as DROP (Dua et al., 2019), MathQA (Amini et al., 2019), and TabFact (Chen et al., 2020) have advanced numerical reasoning and fact verification over textual and tabular data. In parallel, recent research has explored multi-modal reasoning combining tables, charts, and text (Zhao et al., 2024), and instruction-tuned large language models have demonstrated strong zero-shot capabilities for numerical tasks (Wang et al., 2022; Shao et al., 2024).

Although most progress has centered on high-resource languages such as English, Vietnamese still lacks dedicated benchmarks for numerical reasoning. To address this, we present the VLSP 2025 Shared Task on Numerical Reasoning Question Answering for the financial domain. Its centerpiece, ViNumQA, is the first manually validated Vietnamese dataset, built from a verified FinQA translation and newly curated QA pairs from local financial reports (2020-2025), each paired with a gold-standard reasoning program and executable answer.

Through this shared task, we aim to (1) provide a rigorous benchmark for evaluating numerical reasoning in Vietnamese, (2) encourage model

transparency and interpretability through program-based reasoning, and (3) analyze the performance trade-offs between compact fine-tuned models and inference-only agentic workflows. By releasing ViNumQA and reporting the competition outcomes, this work establishes the first foundation for advancing Vietnamese financial NLP and contributes to the broader goal of multilingual numerical reasoning research.

This paper summarizes the VLSP 2025 Numerical Reasoning Shared Task organized as follows: Section 2 outlines the task and evaluation settings; Section 3 details the dataset construction and composition; Section 4 describes the participating systems and methodologies; and Section 5 reports the results and discusses directions for future work.

2 Shared Task Description

This section provides an overview of the Vietnamese Numerical Reasoning QA shared task, including the task objectives, subtasks, and evaluation metrics.

2.1 Task Overview

The goal of this shared task is to develop and evaluate systems for numerical reasoning question answering in the Vietnamese financial domain. Given a context consisting of textual paragraphs and a table extracted from a financial report, along with a natural language question, participating systems are required to perform two main actions:

1. **Generate a Reasoning Program:** Systems must parse the question and provided context to create an executable, step-by-step mathematical program. This program serves as a transparent and verifiable reasoning path to the final answer.

2. **Provide a Final Answer:** Systems must execute the generated program to compute the final numerical answer.

This dual-output formulation assesses not only answer accuracy but also the model's logical and interpretable reasoning process. By requiring transparent reasoning programs, the task promotes the development of trustworthy and accountable AI systems for financial analysis.

The competition consists of two distinct subtasks with different resource constraints:

Subtask 1: Constrained-Resource Numerical Reasoning This subtask focused on developing efficient and reproducible models under strict resource limitations. The primary goal was to encourage innovation in low-resource settings. The key constraints were:

- Models were constrained to contain at most 13 billion parameters.
- The use of external Large Language Models (LLMs) or any API services during both training and inference was strictly prohibited.
- All system components were required to be self-contained and reproducible.

Subtask 2: Unconstrained Training with LLM-Supported Reasoning This subtask lifted the restrictions on model size and training resources, allowing participants to leverage state-of-the-art tools to build high-performance systems. The shared task imposed several constraints: there was no restriction on model size, and participants could use LLMs or external APIs during training for data augmentation or synthetic data generation. However, during the final testing phase, systems were required to run independently without external API or LLM access, while still producing transparent reasoning programs for each answer.

2.2 Data Format

The entire dataset is provided in JSON format. Each data instance corresponds to a single question-answering problem and is structured as a JSON object with the following key fields:

- pre_text: A list of strings, representing the textual paragraphs that appear before the table in the original document.
- post_text: A list of strings, representing the textual paragraphs that appear after the table.
- table: A two-dimensional list of strings representing the financial data table, with the first inner list being the header row.
- id: A unique string identifier for the data sample.
- qa: A JSON object containing the question-answering pair, which includes:
  - question: The question in Vietnamese.

- program: The gold-standard reasoning program (e.g., divide(914, 391)). This field is provided in the training data for model learning.
- exe_ans: The final numerical answer obtained by executing the gold program. This is also provided for training.

An illustrative example of a data instance is shown in Figure 1. It demonstrates the relationship between textual and tabular inputs and the corresponding reasoning process expressed as a program.

pre_text: ["mục đích phát triển dự án bất động sản của doanh nghiệp.", "hệ số nợ trên vốn chủ sở hữu (d/e) của công ty cổ phần phát triển đô thị từ liêm là 0.1 lần và thường xuyên ở mức thấp, tránh rủi ro thanh khoản cho doanh nghiệp."]
("The purpose of the company's real estate development project is to support its strategic growth. The debt-to-equity ratio (D/E) of Tu Liem Urban Development Joint Stock Company is 0.1 and consistently remains low, helping the company avoid liquidity risks.")

table:

Metric (VND billion)	2022	2023
Net revenue	391	914
Gross profit	163	513
...	...	...

post_text: ["."]

qa:
question: *Doanh thu thuần năm 2023 gấp bao nhiêu lần doanh thu thuần năm 2022?*
(How many times is the net revenue in 2023 compared to the net revenue in 2022?)
program: divide(914, 391)
exe_ans: 2.338

Figure 1: An example instance from the ViNumQA dataset.

For the evaluation sets (both public and private), each instance includes the full context (pre_text, post_text, and table) along with the question. The program and exe_ans fields are withheld, as these are the target outputs that participating systems must predict during evaluation.

2.3 Evaluation Metrics

To evaluate model performance on the FinQA benchmark, the official evaluation protocol proposed by Chen et al. (Chen et al., 2021) is adopted. This protocol employs two main metrics: Execution Accuracy (EA), which measures the correctness of the final numerical result, and Program Accuracy (PA), which evaluates the mathematical equivalence between the generated reasoning program and the gold-standard program.

2.3.1 Execution Accuracy (EA)

Execution Accuracy measures the percentage of questions for which the model's generated program, upon execution, produces the correct final answer. The evaluation process involves taking the sequence of operations predicted by the model, computing the final numerical result, and comparing this result directly against the gold answer in the dataset. While EA provides a straightforward measure of task completion, it can potentially overestimate a model's true reasoning ability, as an incorrect program may coincidentally yield the correct answer.

- N_{correct} be the number of questions with correct final answers,
- N be the total number of questions.

Then EA is defined as:

$$EA = \frac{N_{\text{correct}}}{N} \times 100\%. \quad (1)$$

2.3.2 Program Accuracy (PA)

Program Accuracy is a more rigorous metric designed to evaluate the logical correctness of the generated reasoning steps. It measures the percentage of instances where the predicted program is mathematically equivalent to the gold program. This is determined through a symbolic evaluation process:

1. All numerical arguments and table references within both the predicted and gold programs are replaced with abstract symbols (e.g., a1, a2).
2. These symbolic programs are then converted into formal mathematical expressions.
3. The expressions are subsequently simplified to a canonical form using a symbolic math library. This ensures that mathematically equivalent operations (e.g., $a + b$ and $b + a$) are treated as identical.
4. A prediction is only considered accurate if its simplified symbolic expression exactly matches that of the gold program.

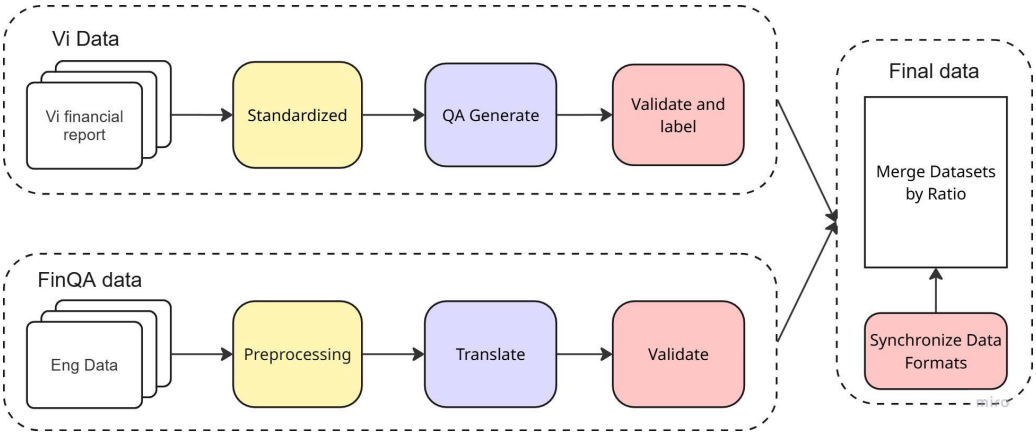

Figure 2: Overview of the data processing integrating Vi Data and FinQA-Vi datasets.

PA serves as a direct assessment of the model's ability to learn the correct reasoning procedure. However, it may be overly strict and produce false negatives if a question can be solved by multiple, distinct, yet equally valid programs.

- N_{equiv} be the number of questions with mathematically equivalent programs (i.e., the predicted program is algebraically / semantically equivalent to the gold-standard program),
- N be the total number of questions.

Then PA is defined as:

$$PA = \frac{N_{\text{equiv}}}{N} \times 100\%. \quad (2)$$

For example, the following two programs are mathematically equivalent.
ADD(A1, A2), ADD(A3, A4), SUBTRACT(#0, #1)
ADD(A4, A3), ADD(A1, A2), SUBTRACT(#1, #0)

3 Data Preparation

This section describes the construction and composition of the dataset in the NumQA shared task. The final corpus, named **ViNumQA**, integrates two complementary sources: (1) a Vietnamese-translated and validated version of the **FinQA-Vi** dataset, and (2) a newly curated Vietnamese financial dataset (**Vi Data**). Figure 2 illustrates the complete data processing pipeline, from raw reports to the finalized dataset.

3.1 Datasets and Resources

Participants are provided with a comprehensive set of resources for training and evaluation.

Training Data The official training corpus combines two main sources:

- A Vietnamese-translated version of the FinQA dataset (Chen et al., 2021), carefully preprocessed and manually verified for translation fidelity.
- A newly curated collection of financial data extracted from publicly available Vietnamese corporate reports (2020-2025).

Participants are encouraged to utilize additional publicly available or appropriately licensed Vietnamese financial datasets to improve model robustness.

Evaluation Data The evaluation corpus is divided into two parts:

- **Public Test Set:** Released for model validation and hyperparameter tuning.
- **Private Test Set:** Reserved for final leaderboard ranking; this portion remains confidential until the competition concludes.

3.2 Dataset Construction

The ViNumQA dataset was developed through a multi-phase pipeline to ensure both linguistic quality and reasoning consistency. Each phase targeted a specific objective—from translation and normalization to question generation and schema harmonization.

Phase 1: FinQA Translation and Validation To broaden linguistic diversity and preserve financial reasoning structure, we created a Vietnamese version of FinQA. The process involved two key steps:

- **Preprocessing.** The original English data were cleaned and normalized, including fixing Unicode inconsistencies and ensuring proper sentence segmentation for reliable translation.

- **Translation and Verification.** All text fields (question, context, and table) were translated into Vietnamese using Gemini Pro. Human annotators then manually validated the translations to ensure linguistic fluency, correctness of financial terminology, and preservation of numerical precision.

This step produced the Vietnamese FinQA subset (FinQA-Vi), ensuring faithful preservation of numerical reasoning structures.

Phase 2: Vi Data Construction The Vi Data subset was derived from Vietnamese financial reports published between 2020 and 2025. The financial reports were collected from publicly available sources, including major Vietnamese financial service providers such as MBS[1], SSI[2], and MASVN[3]. The construction process followed a three-stage pipeline:

- **Standardization.** Textual and tabular content were extracted from PDF financial reports. Tables were checked for layout accuracy and alignment with their source documents. Numerical values were normalized to ensure consistent numeric formatting (e.g., decimal separators and thousands delimiters).

- **QA Generation.** Large language models (Gemini Pro (Team et al., 2023) and OpenAI 4o (Hurst et al., 2024)) were used to automatically generate question-answer (QA) pairs grounded in the extracted reports. Each QA instance includes both a natural-language question and its corresponding numerical reasoning program.

- **Validation and Labeling.** Human annotators reviewed all LLM-generated QA pairs to ensure clarity, factual correctness, and contextual consistency. Ambiguous or invalid samples were discarded, and the validated data were finalized for model training.

[1] https://mbs.com.vn/bao-cao-phan-tich-nganh/

[2] https://www.ssi.com.vn/khach-hang-ca-nhan/bao-cao-chien-luoc

[3] https://www.masvn.com/

The Vi Data subset thus provides authentic Vietnamese financial QA instances with verified reasoning traces.

Phase 3: Merging and Harmonization After preparation, the two subsets - FinQA-Vi and Vi Data - were merged into a unified schema. This process ensured consistency in field names, reasoning program representation, and JSON formatting across all instances. The resulting corpus, **ViNumQA**, serves as the official benchmark for Vietnamese numerical reasoning in the financial domain.

3.3 Data Statistics

The ViNumQA dataset comprises a total of **4,074** question-program-answer triplets, divided into **2,993** for training, **584** for validation, and **497** for testing. Following FinQA's categorization, questions are grouped based on the source of supporting evidence:

Table Only - Evidence entirely contained within the structured table.

Text Only - Evidence derived exclusively from unstructured text passages.

Table & Text — Evidence requiring integration of both table and text information.

Table 1: Statistics of the ViNumQA dataset across training, validation, and test splits, categorized by question type and data source.

	Train		Valid		Test	
Type	**Vi**	**Trans.**	**Vi**	**Trans.**	**Vi**	**Trans.**
Table Only	1,087	1,126	234	207	211	154
Table & Text	204	204	37	34	34	30
Text Only	183	189	37	35	42	26
Total	**1,474**	**1,519**	**308**	**276**	**287**	**210**

Table 1 presents a detailed breakdown across dataset splits and question types. The *Table Only* category dominates across all splits, underscoring the importance of reasoning over structured financial data. In contrast, *Table & Text* questions, while less frequent, pose greater challenges by requiring multi-source reasoning.

The dataset maintains a balanced distribution between original Vietnamese samples (Vi) and translated ones (Trans.), totaling 2,069 and 2,005 instances, respectively. This near 1:1 ratio helps mitigate potential source bias and promotes stronger generalization.

Table 2 summarizes the overall dataset composition. Together, the Vi Data and FinQA-Vi subsets constitute the first large-scale Vietnamese corpus for numerical reasoning in the financial domain, designed to benchmark multi-source and multilingual reasoning capabilities.

Table 2: Summary of ViNumQA dataset composition.

Subset	#Samples	Source
Vi Data	2,069	Financial reports (VN)
FinQA-Vi	2,005	Translated from FinQA (EN)
Total	**4,074**	

4 Numerical Reasoning Question and Answer Methods

The submitted systems showcased a variety of sophisticated techniques, primarily centered around the fine-tuning of open-source Large Language Models (LLMs), with a notable preference for the Qwen (Yang et al., 2025) model family. Key trends included multi-stage training pipelines combining supervised fine-tuning with reinforcement learning, extensive data augmentation, and advanced inference-time strategies.

Team HUSTUET The winning **HUSTUET** team opted for a multilingual approach, directly incorporating the English FinQA data into their training set without translation to preserve semantic integrity and avoid translation errors. They also expanded their dataset by utilizing an alternative correct program provided in the FinQA program_re field, enhancing programmatic diversity. The **HUSTUET** team's first-place approach was distinguished by its use of knowledge distillation. They used a powerful 235B-parameter teacher model (Qwen3-235B-Thinking) (Yang et al., 2025) to generate high-quality, structured reasoning traces for the entire training set. These rich traces were then used to fine-tune a much smaller Qwen3-8B student model. Their subsequent GRPO (Zhihong Shao, 2024) stage employed a carefully designed reward function that prioritized program accuracy, balancing it with execution correctness and conciseness, which proved critical for their success.

Team Vietnam Finance The **Vietnam Finance** team translated the FinQA corpus into Vietnamese using OpenAI's GPT-o3, then employed a reasoning-specialized model (DeepSeek-R1-Distill-Qwen-7B) (DeepSeek-AI, 2025) to generate new chain-of-thought traces and programs. These generated examples were rigorously filtered for correctness, with a human-in-the-loop process involving financial analysts to ensure quality. The **Vietnam Finance** team applied parameter-efficient fine-tuning (LoRA) (Hu et al., 2021) for their SFT stage before moving to GRPO, where the reward was based on execution correctness and program parsability. The **Vietnam Finance** team significantly boosted their final accuracy by applying a majority-voting decoding strategy (self-consistency). At inference, they generated 10 candidate programs for each question and selected the most frequently occurring one as the final answer, which effectively reduced stochastic errors and improved robustness.

Team UIT_BlackCoffee For data preprocessing, the **UIT_BlackCoffee** team found that converting financial tables into Markdown format was more effective for model consumption than using the original list-based or JSON formats. Their experiments also showed that simplistic context filtering with BM25 (Robertson and Zaragoza, 2009) was detrimental, as it often removed essential information. The **UIT_BlackCoffee** team also used a two-stage SFT (Dong et al., 2023) and GRPO pipeline, but with a simpler reward function focused primarily on execution correctness. They also demonstrated the effectiveness of using a quantized version of the Qwen3 model (8.7B parameters) for greater efficiency in both training and inference.

Team Innovation-LLM In a stark contrast to the training-heavy methods, the **Innovation-LLM** team developed a pure inference-only AI agent that required no fine-tuning. Their system broke the problem down into a four-step pipeline: (1) Question Decomposition into subqueries, (2) Grounded Data Extraction to answer each subquery, (3) Multi-Path Program Generation using n-sampling (n=15) to create multiple candidate reasoning plans, and (4) Optimal Program Selection via majority voting over the generated program structures. This approach excelled at finding functionally correct solutions, achieving the highest Execution Accuracy in the competition's second subtask and a top-three rank in the first.

5 Result and Discussion

The competition revealed a clear contrast between two main approaches: teams that performed deep fine-tuning of Large Language Models (LLMs)

to optimize for the task (such as HUSTUET and UIT_BlackCoffee), and a team that adopted an inference-only approach without any fine-tuning (Innovation-LLM). The detailed results of these systems are presented in Table 3.

Table 3: Final results of two subtasks.

Subtask	Team	EA (%)	PA (%)
Subtask 1			
	HUSTUET	79.88	**76.63**
	Vietnam Finance	**81.95**	75.00
	Innovation-LLM	79.14	69.82
	UIT_BlackCoffee	74.26	69.67
Subtask 2			
	HUSTUET[1]	79.88	76.63
	Innovation-LLM	**84.00**	**74.07**
	UIT_BlackCoffee[1]	74.26	69.67

5.1 Analysis of Subtask 1: Constrained - Resource Environment

In this subtask, models were constrained to contain at most 13 billion parameters, and were not allowed to use external APIs, focusing on efficiency and reproducibility.

Team Vietnam Finance This team achieved the **highest Execution Accuracy (EA) at 81.95%**. This indicates their model was the most effective at producing the correct final numerical answer. However, their **Program Accuracy (PA) at 75.00%** was lower than team HUSTUET's, suggesting that while their results were correct, the reasoning process (the program) generated by their model did not always align perfectly with the gold-standard logic in the dataset.

Team HUSTUET This team achieved the **highest Program Accuracy (PA) at 76.63%**, demonstrating their model's superior ability to accurately reproduce the logical reasoning chain.This aligns perfectly with their methodology described in the report: using a massive 235B-parameter "teacher" model to generate high-quality reasoning traces, which were then used to fine-tune a much smaller "student" model. Prioritizing PA in their reward function was also a critical factor in this success. Their model can be considered the most comprehensive and reliable.

[1]Teams that do not resubmit their technical report will not be ranked, even though they still appear on the leaderboard.

Team	Eval.	Table Only	Text Only	Table & Text	Total
HUSTUET	EA (%)	86.85	82.35	57.81	79.88
	PA (%)	84.11	67.65	53.12	76.63
Vietnam Finance	EA (%)	68.54	46.27	38.71	81.95
	PA (%)	62.19	38.24	29.69	75.00
Innovation-LLM	EA (%)	82.92	70.59	65.08	79.14
	PA (%)	76.44	60.29	57.81	69.82
UIT_BlackCoffee	EA (%)	78.29	60.29	42.19	74.26
	PA (%)	67.12	52.94	30.06	69.67

Table 4: EA and PA results across different input settings.

Team Innovation-LLM Their performance is commendable for a no-fine-tuning approach. Their EA of 79.14% was highly competitive. The lower PA of 69.82% suggests that their method of generating multiple candidate programs and using majority voting could find a functionally correct path to the answer, but not necessarily the canonical one from the dataset.

Team UIT_BlackCoffee The team's results show the effectiveness of using a quantized version of the Qwen3 model for greater efficiency. Their preprocessing technique of converting financial tables into Markdown format was also noted as being more effective for model consumption than list-based or JSON.

5.2 Analysis of Subtask 2: Unconstrained Training Environment

This subtask allowed the use of large models and external APIs during the training phase, but systems had to operate independently during the final inference stage.

Team Innovation-LLM The team was ranked first because they were the only team to submit a technical report for this subtask. From a technical standpoint, they achieved an **outstandingly high EA of 84.00%**, the highest across the entire competition. This demonstrates that their inference-based agent approach, which involves question decomposition and multi-path program generation, is extremely effective when unconstrained. Their PA also saw a significant improvement to 74.07%.

5.3 Analysis by Question Type

A granular analysis of the results by question type reveals the specific strengths and weaknesses of the competing approaches and confirms that performance is highly dependent on the nature of the evidence source. The data clearly shows a performance hierarchy, with models excelling on struc-

tured data but struggling significantly as the reasoning becomes more complex and multi-modal. The detailed results of these systems are presented in table 4.

Table Only questions, where all necessary evidence is contained within the structured table, yielded the highest performance across all teams. This was the expected outcome, as reasoning over well-defined rows and columns is the most straightforward task. Team **HUSTUET** achieved the top scores in this category with an Execution Accuracy (EA) of 86.85% and a Program Accuracy (PA) of 84.11%. Their knowledge distillation method, which used a powerful teacher model to generate high-quality reasoning traces for a smaller student model, proved exceptionally effective for these structured problems where the reasoning paths are often unambiguous.

Performance saw a marked decline in the **Text Only** category, which requires deriving answers exclusively from unstructured text passages. While HUSTUET maintained a leading EA of 82.35%, the Program Accuracy (PA) for all teams dropped significantly more than their EA scores. This divergence suggests that while models could often extract the correct numerical entities from the text and perform the right calculation to get the right answer, they struggled to consistently formulate the correct, verifiable reasoning program. This highlights the inherent challenge of semantic parsing in unstructured financial text compared to structured tables.

The most significant performance drop occurred with **Table & Text** questions, which require integrating information from both sources and represent the most complex reasoning challenge. In this category, the fine-tuned models struggled, while the inference-only agent from team **Innovation-LLM** achieved the highest scores (EA 65.08%, PA 57.81%). This result is particularly insightful; Innovation-LLM's methodology, which involves decomposing a question into subqueries, extracting grounded data for each, and then generating multiple program paths, appears to be more robust for these complex, multi-hop reasoning tasks. While deep fine-tuning helps models master patterns seen in training, the agentic, step-by-step approach demonstrated a superior ability to handle novel problems requiring the synthesis of disparate data sources. This suggests that for the most challenging financial analysis tasks, flexible, decomposition-based reasoning may be a more promising direction than monolithic fine-tuning.

5.4 Discussion

The results of the VLSP 2025 Numerical Reasoning Shared Task provide several noteworthy insights into model behavior and design trade-offs. In particular, the comparison between participating systems highlights distinct strengths in reasoning accuracy, generalization, and efficiency, offering valuable perspectives for future research in interpretable financial NLP.

- **The EA vs. PA Trade-off:** The results reveal an interesting trade-off between "getting the right answer" (EA) and "reasoning in the right way" (PA). Innovation-LLM was the champion of EA, while HUSTUET was the champion of PA. In a domain like finance, a model with high PA, such as HUSTUET's, might be preferred for its transparency, reliability, and auditability.
- **Innovation-LLM's Breakthrough Approach:** Developing a no-fine-tuning agent that still achieves top-tier results marks a promising direction, especially regarding generalization capabilities and reducing training costs.
- **HUSTUET's Confirmed Quality:** The technique of knowledge distillation from a massive model to a smaller one proved to be exceptionally effective, producing a system that is both powerful in its logic (high PA) and compact enough to meet the strict requirements of Subtask 1.

Although the proposed approach achieves promising results, several limitations remain. The ViNumQA dataset partly relies on translated and LLM-generated content, which may introduce semantic shifts and bias. Its domain coverage is confined to corporate reports from 2020-2025, limiting generalization to other financial contexts. Moreover, the evaluation metrics may not fully capture the systems' true reasoning capabilities, suggesting room for improvement in both data and assessment design.

6 Conclusions

The VLSP 2025 Shared Task on Numerical Reasoning has successfully established the first benchmark and publicly available dataset, ViNumQA,

for the Vietnamese financial domain. This work addresses a critical resource gap for non-English languages, providing a rigorously validated foundation to foster research in Vietnamese financial NLP.

The competition results revealed a significant dichotomy between two dominant strategies. On one hand, deep fine-tuning methods, particularly the knowledge distillation approach, demonstrated superior performance in generating correct reasoning logic, achieving the highest Program Accuracy (PA). On the other hand, a novel, inference-only agentic workflow achieved the highest Execution Accuracy (EA) without any task-specific training, highlighting a promising direction for generalization and cost reduction. This underscores a key trade-off in financial AI between systems that are demonstrably reliable and auditable (high PA) and those that are effective at producing the correct final answer (high EA).

By providing this foundational benchmark and analyzing the competing methodologies, this initiative paves the way for the development of more sophisticated and tailored models for the unique challenges of the Vietnamese financial landscape.

Acknowledgments

We would like to thank all participating teams for their active contributions, innovative ideas, and collaborative spirit throughout the VLSP 2025 Numerical Reasoning shared task. We are deeply grateful to the annotation and data verification teams for their careful and dedicated work in constructing and validating the Vietnamese numerical reasoning resources. This work was made possible through the support of the VLSP 2025 Organizing Committee and affiliated institutions, which provided computational resources and organizational assistance. We also appreciate the valuable feedback from the reviewers, which helped improve the quality and clarity of this report.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of NAACL-HLT*, pages 2357–2367.

Wenhu Chen, Hongmin Chen, Jianshu Chen, Yunkai Zhang, Shiqi Yu, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *Proceedings of ICLR*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. *Proceedings of EMNLP 2021*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in Large Language Models are affected by Supervised Fine-tuning Data Composition. *arXiv preprint arXiv:2310.05492*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2368–2378.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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*.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. *arXiv preprint arXiv:2204.05660*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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. *arXiv preprint arXiv:2505.09388*.

Bowen Zhao, Tianhao Cheng, Yuejie Zhang, Ying Cheng, Rui Feng, and Xiaobo Zhang. 2024. Ct2c-qa: Multimodal question answering over chinese text, table and chart. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3897–3906.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*.

Qihao Zhu Runxin Xu Junxiao Song Mingchuan Zhang Y.K. Li Y. Wu Daya Guo Zhihong Shao, Peiyi Wang. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Fengbin Zhu, Wenqiang Lei, Youfang Huang, Chao Wang, Shuo Zhang, Jian Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3105–3115. Association for Computational Linguistics.