VNU Journal of Science: Comp. Science & Com. Eng., Vol. 41, No. 2 (2025) 100–112

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

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

Received 06th December 2025;
Revised 10th December 2025; Accepted 22nd December 2025

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. We established robust baselines using the LLaMA model family and compared them against state-of-the-art proprietary LLMs (GPT-4o, GPT-5 mini). The results demonstrate that supervised fine-tuning is essential for adherence to reasoning schemas, as few-shot prompting strategies suffered from high invalid generation rates. 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 in both Program and Execution Accuracy. Meanwhile, an inference-only agent attained a highly competitive Execution Accuracy without any 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.

Keywords: Numerical Reasoning, Question Answering, viNumQA, VLSP 2025, Vietnamese

1. Introduction

Numerical Reasoning Question Answering (NumQA) has emerged as a critical task in financial Natural Language Processing (NLP), driven by the increasing availability of digital financial documents such as annual reports and corporate disclosures. A key aspect of NumQA

*Corresponding author.
E-mail address: lengoctoan@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6507

lies in its requirement for numerical reasoning, enabling users to query complex data and extract insights directly from unstructured and semi-structured sources. While conventional QA tasks primarily focus on text comprehension or span extraction, modern NumQA systems are required to not only comprehend financial text and tables but also to perform mathematical operations and generate explicit reasoning programs that can be verified for correctness. This work is presented as part of the VLSP 2025 Shared Task on Numerical Reasoning Question Answering, which introduces ViNumQA as the first initiative to benchmark these capabilities in Vietnamese financial texts and evaluates the balance between model transparency and execution accuracy.

Significant progress has been achieved in English, driven by benchmark datasets like FinQA [1] and TAT-QA [2], 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 [3, 4] or employ chain-of-thought prompting for numerical inference [5]. Beyond the financial domain, several datasets such as DROP [6], MathQA [7], and TabFact [8] 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 [9], and instruction-tuned large language models have demonstrated strong zero-shot capabilities for numerical tasks [10, 11].

Despite significant advancements in numerical reasoning for high-resource languages like English, there remains a critical absence of benchmarks dedicated to Vietnamese. To bridge this gap, we introduce ViNumQA, a dataset synthesizing verified FinQA translations and QA pairs extracted from domestic financial reports (2020-2025), complete with gold-standard reasoning programs and executable answers.

We aim to benchmark Vietnamese numerical reasoning, foster model transparency via program generation, and compare the efficiency of compact fine-tuned models versus inference-only agents. By releasing ViNumQA and detailing the task outcomes, this work serves as a pioneering resource for Vietnamese financial NLP and contributes significantly to multilingual 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 primary objective of this shared task is to advance and benchmark numerical reasoning systems specifically for the Vietnamese financial domain. Given a hybrid context of financial text and tabular data, participating systems must generate two outputs: an executable mathematical reasoning program and the final numerical answer derived from that program. This dual requirement ensures that models are evaluated not only on result accuracy but also on the transparency and validity of their logical process.

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

Subtask 1: Focuses on low-resource settings. Models must be self-contained, reproducible, and less or equal 13B parameters. No external APIs are allowed at any stage.

Subtask 2: Focuses on high performance. Participants may use any resources (including LLMs) for training. However, the final test phase requires the model to run offline, generating reasoning programs without calling external services.

2.2. Data Format

The ViNumQA dataset utilizes a JSON structure to represent financial reasoning problems. Each instance links a specific natural language question to a relevant context composed of unstructured text and structured tables.

The context is segmented into `pre_text` and `post_text` representing paragraphs immediately preceding and following the table and the table itself, which is formatted as a nested list of strings. The core question-answering task is encapsulated in a `qa` object containing the Vietnamese question.

To facilitate supervised learning, training examples include a `program` field (the executable reasoning path) and an `exe_ans` field (the correct numerical result). However, to ensure rigorous assessment, these solution fields are removed from the evaluation datasets, forcing models to independently generate the reasoning logic and calculation.

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: [... bên cạnh mức tăng trưởng doanh thu rất cao của ngành bất động sản, lợi nhuận ròng của ngành này cũng đã tăng đáng kể ở mức 23,8% (yoy, ttm), xấp xỉ với mức quân bình của tất cả các ngành (23,1%).]

(Translation: [... besides the very high revenue growth of the real estate sector, the net profit of this sector also increased significantly at 23.8% (yoy, ttm), approximating the average of all sectors (23.1%).])

Table:

Ngành	Lợi nhuận	Doanh thu
VN-Index	23.13	22.29
Tiêu dùng không thiết yếu (Consumer Discretionary)	2.49	23.76
Tiêu dùng thiết yếu (Consumer Staples)	8.72	6.46
Năng lượng (Energy)	-7.02	25.98
Tài chính (Financials)	48.78	25.68
Y tế (Health Care)	-3.03	6.05
Công nghiệp (Industrials)	15.25	22.82
Công nghệ thông tin (Information Technology)	37.70	-13.75
Nguyên vật liệu (Materials)	2.42	25.70
Bất động sản (Real Estate)	23.83	52.04
Tiện ích (Utilities)	27.47	13.64

Post_text: [... sự khác biệt giữa mức tăng trưởng doanh thu và lợi nhuận của ngành bất động sản nêu trên có thể được giải thích bởi 2 nguyên nhân sau.]

(Translation: [... the difference between the revenue and profit growth of the aforementioned real estate sector can be explained by the following 2 reasons.])

Question: Tăng trưởng lợi nhuận của ngành Tài chính cao hơn bao nhiêu phần trăm so với mức tăng trưởng lợi nhuận trung bình của tất cả các ngành?

(Translation: By how many percentage points is the profit growth of the Financials sector higher than the average profit growth of all sectors?)

Answer: divide(48.78, 100), subtract(#0, 23.1%)

Figure 1. An illustrative example of a hybrid question answering instance. The input consists of a financial table and textual context. To answer the question, the model must extract the profit growth of the Financials sector from the table and compare it with the average growth mentioned in the text, generating the corresponding arithmetic program.

2.3. Evaluation Metrics

To evaluate model performance on the FinQA benchmark, the official evaluation protocol proposed by Chen et al. [1] 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.

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 details the construction of ViNumQA, the first large-scale benchmark for numerical reasoning in the Vietnamese financial domain. To ensure the dataset possesses both authentic local context and structural diversity, we employed a hybrid construction strategy: (1) creating a core dataset from native Vietnamese financial reports, and (2) augmenting this with a high-quality translation of the FinQA benchmark.

3.1. Datasets and Resource

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 [1], 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.

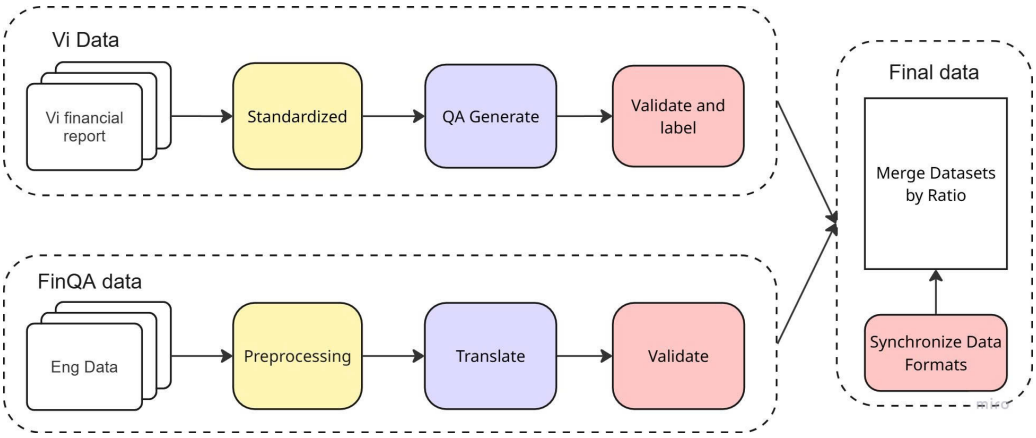

Figure 2. Overview of the data processing integrating Vi Data and FinQA-Vi datasets.

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. *Construction of Native Vietnamese Dataset (Vi Data)*

The core component of our benchmark, the Vi Data subset, consists of authentic QA pairs derived from Vietnamese corporate financial reports published between 2020 and 2025. The source documents were collected from major financial service providers, including MBS[1], SSI[2], and MASVN[3]. The construction pipeline comprised three rigorous stages:

- Standardization: We extracted textual paragraphs and data tables from the original PDF reports. Special attention was paid to layout analysis to ensure tables were aligned correctly with the source documents. Furthermore, numerical values were normalized to address inconsistent formatting (e.g., varying decimal separators and thousands delimiters) common in Vietnamese documents.
- Automated QA Generation: To scale the dataset efficiently, we utilized advanced Large Language Models (LLMs), specifically Gemini Pro and OpenAI GPT-4o. These models were prompted to analyze the extracted financial contexts and generate Question-Answer (QA) pairs accompanied by executable reasoning programs
- Human Validation and Labeling: To ensure data quality, human annotators meticulously reviewed all LLM-generated instances. Annotators checked for clarity, factual correctness against the source report, and the logical validity of the reasoning program. Ambiguous or factually incorrect samples were discarded to establish a high-quality gold standard.

This process resulted in 2,069 verified samples, representing the first dedicated corpus for Vietnamese financial numerical reasoning.

[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/

3.3. Data Augmentation via FinQA Translation (FinQA-Vi)

To enhance the dataset's diversity and incorporate complex reasoning structures proven in English benchmarks, we constructed the FinQA-Vi subset by adapting the established FinQA dataset. This phase involved two key steps:

- Preprocessing: The original English data underwent cleaning and normalization, including the correction of Unicode inconsistencies and sentence segmentation, to facilitate accurate translation.
- Translation and Verification: All textual components (questions, context passages, and table headers) were translated into Vietnamese using Gemini Pro. Subsequently, human annotators manually validated the translations to ensure linguistic fluency and the preservation of precise financial terminology. Crucially, the numerical values and mathematical logic were verified to remain consistent with the original reasoning programs.

This augmentation phase contributed 2,005 samples to the final dataset.

3.4. 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**	

Table 3 presents a comparative statistical analysis between the native Vi Data and the translated FinQA subset. A closer examination reveals significant structural and distributional differences that enhance the robustness of the Vietnamese benchmark.

Regarding the input context, the Vi Data exhibits considerably higher structural complexity. As evidenced by the table

Table 3. Counts of general text/table attributes and the distribution of arithmetic operations reported in Vi Data and FinQA data

Category	Vi Data	FinQA
General Statistics (Average)		
Word count	1,594.46	3,047.40
Sentence count	14.42	24.30
Table Rows	10.05	6.36
Table Cols	6.68	3.86
Operations (Count and Percentage)		
add	551 (17.47%)	1,952 (15.33%)
subtract	1,034 (32.78%)	3,676 (28.87%)
divide	897 (28.44%)	5,901 (46.35%)
multiply	161 (5.10%)	759 (5.96%)
table_max	218 (6.91%)	66 (0.52%)
table_min	119 (3.77%)	36 (0.28%)
table_average	118 (3.74%)	129 (1.01%)
table_sum	56 (1.78%)	50 (0.39%)
greater	0 (0.00%)	154 (1.21%)
exp	0 (0.00%)	9 (0.07%)

dimensions, the Vietnamese financial reports feature substantially larger tables, averaging 10.05 rows and 6.68 columns, compared to just 6.36 rows and 3.86 columns in FinQA. This expanded tabular size presents a richer and more challenging context for evidence retrieval models. In terms of logical operations, the Vi Data construction process intentionally streamlined the operator set by excluding comparison-based operations such as greater and exp (0.00%). More importantly, the dataset addresses the class imbalance observed in the original FinQA by achieving a more balanced distribution for complex aggregation operations. While FinQA is heavily dominated by simple arithmetic (divide, subtract), Vi Data significantly boosts the representation of low-frequency operations. Specifically, operations like table_max (6.91% vs. 0.52%), table_min (3.77% vs. 0.28%), and table_average (3.74% vs. 1.01%) appear with much greater frequency. This shift ensures that models are evaluated on their ability to perform complex reasoning over table ranges rather than merely retrieving single cells for arithmetic calculation.

4. Numerical Reasoning Question and Answer Methods

4.1. Baseline Implementation

To establish a robust performance benchmark for the shared task, we implemented a baseline study utilizing the LLaMA model family. We selected three specific variants to evaluate reasoning capabilities across different parameter scales: LLaMA 3.2 1B and LLaMA 3.2 3B, and LLaMA 3 8B[12].

To ensure a rigorous and fair comparison across these varying model scales, we standardized the experimental setup for both training and inference. During the Supervised Fine-Tuning (SFT) phase, we employed Low-Rank Adaptation (LoRA[13]) with an alpha value of 32 to optimize parameter efficiency. The training process was conducted over 2 epochs with a learning rate of 2×10^{-4} and a total batch size of 4, incorporating 5 warm-up steps to stabilize the optimization trajectory. Additionally, the maximum sequence length was fixed at 4096 tokens to accommodate the extensive context often required in financial documents. For the inference phase, we prioritized deterministic outputs to evaluate reasoning consistency; therefore, we set strict sampling parameters with both *temperature* and *min_p* at 0.1, while capping the generation at 128 new tokens.

The selection of LLaMA 3.2 (1B and 3B) was driven by the objectives of Subtask 1, aiming to test the efficacy of lightweight, edge-optimized models in constrained resource environments. Conversely, the LLaMA 3 8B model served as a standard generalist baseline to assess how a mid-sized open-weight model performs on Vietnamese financial reasoning without the specialized reinforcement learning pipelines used by the top-tier teams. These models were subjected to supervised fine-tuning (SFT) on the ViNumQA training set to provide a reference

point for evaluating the specialized architectures submitted to the competition.

To complement the fine-tuned open-source baselines, we also evaluated the intrinsic numerical reasoning capabilities of state-of-the-art proprietary Large Language Models (LLMs) via in-context learning. We selected GPT-4o and GPT-4o mini[14], to represent the frontier of commercial model performance. Unlike the fine-tuned models, these systems were assessed using a few-shot prompting strategy, where a single representative example of a (Question, Context, Program) triplet was provided within the prompt context. This experimental setup was designed to gauge the models' immediate ability to synthesize financial reasoning programs without the computational overhead of parameter updates, serving as a high-level benchmark for the "Unconstrained" subtask.

4.2. Participating Systems

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 [15] model family. Key trends included multi-stage training pipelines combining supervised fine-tuning with reinforcement learning, extensive data augmentation, and advanced inference-time strategies.

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) [15] 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 [16] stage employed a carefully designed reward function that prioritized program accuracy, balancing it with execution correctness and conciseness, which proved critical for their success.

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) [17] 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) [13] 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.

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 [18] was detrimental, as it often removed essential information. The UIT_BlackCoffee team also used a two-stage SFT [19] 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.

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

Table 4. Performance comparison of LLMs and competitive systems

Model	Invalid	Eval.	Table Only	Text Only	Table & Text	Total
Baseline Models						
Unsloth LLaMA 1B	3	EA (%)	71.82	29.41	34.38	60.76
		PA (%)	68.23	29.41	32.81	57.95
Unsloth LLaMA 3B	1	EA (%)	80.22	57.35	46.88	72.64
		PA (%)	77.75	50.00	43.75	69.42
Unsloth LLaMA 8B	1	EA (%)	82.69	51.47	46.88	73.64
		PA (%)	79.95	48.53	43.75	70.82
LLMs						
GPT-4o mini	76	EA (%)	69.16	42.19	46.94	52.92
		PA (%)	65.91	35.94	44.9	49.9
GPT-4.1 mini	75	EA (%)	59.24	43.4	63.64	49.09
		PA (%)	57.64	37.74	58.18	46.88
GPT-5 mini	42	EA (%)	48.81	43.94	37.74	42.86
		PA (%)	45.83	34.85	33.96	39.24
Participating Systems						
HUSTUET	0	EA (%)	86.85	82.35	57.81	82.49
		PA (%)	84.11	67.65	53.12	77.87
Vietnam Finance	12	EA (%)	68.54	46.27	38.71	60.16
		PA (%)	63.76	38.81	30.65	54.73
Innovation-LLM	3	EA (%)	82.92	70.59	65.08	78.47
		PA (%)	76.86	60.29	58.73	71.83
UIT_BlackCoffee	38	EA (%)	78.29	60.29	42.19	65.19
		PA (%)	74.92	52.94	39.06	61.57

The experimental results presented in Table 4 highlight distinct performance trends across supervised baselines, few-shot proprietary LLMs, and specialized participating systems.

Efficacy of Supervised Fine-Tuning (SFT): The fine-tuned LLaMA baseline models demonstrate the critical importance of task-specific adaptation. The LLaMA 3 8B model achieved a Total Execution Accuracy (EA) of 73.64%, which is highly competitive, outperforming the participating system UIT_BlackCoffee (65.19%) and significantly surpassing all proprietary LLMs. The scaling trend is evident, with the 8B model outperforming the 1B variant by nearly 13 percentage points in Total EA. Furthermore, the extremely low "Invalid" counts for the LLaMA models confirm that supervised fine-tuning effectively constrains the models to generate the required executable program format.

Limitations of Few-Shot Prompting in Proprietary LLMs: In sharp contrast, the proprietary LLMs (GPT-4o mini, GPT-5 mini, etc.) struggled significantly under the few-shot setting.

- High Failure Rates: The "Invalid" column reveals a major bottleneck for these models. GPT-4o mini failed to produce a valid response in 76 instances, this indicates a consistent issue with instruction adherence in these models.
- Reasoning Gap: Despite their general knowledge, these models achieved Total EAs ranging only between 42.86% and 52.92%. This suggests that without fine-tuning, even powerful models struggle to adhere to the strict program syntax or perform the specific multi-step financial reasoning required by the ViNumQA benchmark.

Superiority of Specialized Architectures: The winning team, HUSTUET, demonstrated the upper bound of current performance with a Total EA of 82.49% and a remarkably perfect record of 0 Invalid answers. This validates their knowledge distillation approach, where a larger "teacher" model guides the reasoning process. Notably, the Innovation-LLM agent (Total EA 78.47%) outperformed the strong LLaMA 8B baseline, proving that their inference-only decomposition strategy is effective. However,

the LLaMA 8B baseline notably outperformed the UIT_BlackCoffee system, suggesting that a simple, well-tuned strong baseline can beat more complex architectures if the latter are not optimized correctly.

5.1. Analysis by Question Type

A granular analysis of the results by question type, as detailed in Table 4, reveals specific strengths and weaknesses across the different model classes. The data confirms a distinct performance hierarchy where models excel on structured data but face significant hurdles with unstructured and hybrid contexts.

Structured Data Reasoning (Table Only): This category yielded the highest performance across fine-tuned systems. Notably, the LLaMA 3 8B baseline achieved an Execution Accuracy (EA) of 82.69%, performing on par with the top-tier participating systems (Innovation-LLM at 82.92% and HUSTUET at 86.85%). This suggests that for structured financial data, a standard supervised fine-tuning (SFT) approach on a mid-sized model is sufficient to capture the reasoning patterns. In sharp contrast, proprietary LLMs utilizing few-shot prompting struggled significantly, with GPT-4o mini and GPT-4.1 mini achieving only 69.16% and 49.09% respectively. This disparity highlights that without task-specific tuning to enforce schema constraints, even powerful generalist models fail to reliably parse financial tables into executable programs.

Unstructured Text Reasoning (Text Only): Performance divergences became pronounced in the Text Only category. While the specialized HUSTUET system maintained a high EA of 82.35%, the LLaMA 8B baseline dropped significantly to 51.47%. Interestingly, LLaMA 3B (57.35%) sightly outperformed the LLaMA 8B (51.47%) in this specific category. For instance, the LLaMA 3 8B baseline drops precipitously from 82.69% EA on structured tables to 51.47% EA on unstructured text. This confirms that extracting numerical facts and logic from prose is considerably more difficult for these models than processing structured tabular schemas.

Hybrid Reasoning (Table & Text): The Table & Text category remains the most challenging, requiring the synthesis of disparate evidence sources. Here, the limitations of standard SFT became evident: the LLaMA 8B baseline achieved only 46.88% EA. The proprietary LLMs fared even worse, averaging below 43%. Crucially, the Innovation-LLM agent maintained a superior performance of 65.08%, significantly outperforming both the robust LLaMA baseline and the HUSTUET system (57.81%).This result also shows a fairly similar performance to gpt-4.1 mini (63.64%) and strongly validates the hypothesis that for complex, multi-modal financial reasoning, a decomposition-based agentic workflow is superior to monolithic fine-tuning.

5.2. 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). the Innovation-LLM system exhibits the largest gap (6.64%) - between EA and PA, suggesting that inference-only agents are highly effective at deriving correct answers through flexible decomposition but are less constrained to the specific programmatic schema of the dataset. 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.

Experimental results from the competing teams and the LLaMA 3 8B model indicate that the EA approach tends to slightly outperforms PA. An analysis of the LLaMA model's results reveals that this performance gap primarily stems from two prevalent error types in the model's program generation. First, the model frequently generates redundant, lengthy sequences of the four basic arithmetic operations instead of utilizing more efficient table-based operations. Second, there is a formatting inconsistency in arguments compared to the ground truth. For instance, the model may output *divide(6.2, 10.0)* when the gold standard requires *divide(6.2, 10)*.

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. Conclusion

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

[1] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. R. Routledge, et al., Finqa: A Dataset of Numerical Reasoning over Financial Data, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3697–3711.
URL https://aclanthology.org/2021.emnlp-main.300/

[2] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, T.-S. Chua, TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance, arXiv preprint arXiv:2105.07624 (2021).
URL https://arxiv.org/abs/2105.07624

[3] S. Mishra, A. Mitra, N. Varshney, B. Sachdeva, P. Clark, C. Baral, A. Kalyan, NumGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks, arXiv preprint arXiv:2204.05660 (2022).
URL https://arxiv.org/abs/2204.05660

[4] Y. Zhao, Y. Li, C. Li, R. Zhang, MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data, arXiv preprint arXiv:2206.01347 (2022).
URL https://arxiv.org/abs/2206.01347

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Advances in neural information processing systems, Vol. 35, 2022, pp. 24824–24837.
URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[6] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, M. Gardner, 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: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 2368–2378.
URL https://aclanthology.org/N19-1246/

[7] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, H. Hajishirzi, MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers), 2019, pp. 2357–2367.
URL https://aclanthology.org/N19-1245/

[8] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, W. Y. Wang, Tabfact: A Large-scale Dataset for Table-based Fact Verification, arXiv preprint arXiv:1909.02164 (2019).
URL https://arxiv.org/abs/1909.02164

[9] B. Zhao, T. Cheng, Y. Zhang, Y. Cheng, R. Feng, X. Zhang, Ct2c-qa: Multimodal Question Answering over Chinese Text, Table and Chart, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 3897–3906.
URL https://dl.acm.org/doi/abs/10.1145/3664647.3681053

[10] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency Improves Chain of Thought Reasoning in Language Models, arXiv preprint arXiv:2203.11171 (2022).
URL https://arxiv.org/abs/2203.11171

[11] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al., Deepseekmath: Pushing the Limits of Mathematical Reasoning in Open Language Models, arXiv preprint arXiv:2402.03300 (2024).
URL https://arxiv.org/abs/2402.03300

[12] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The Llama 3 Herd of Models, arXiv e-prints 2024, pp. arXiv–2407.
URL https://ui.adsabs.harvard.edu/abs/2024arXiv240721783G/abstract

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank Adaptation of Large Language Models., ICLR, Vol. 1, No. 2, 2022, pp. 3.
URL https://arxiv.org/pdf/2106.09685v1/1000

[14] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., Gpt-4o System Card, arXiv preprint arXiv:2410.21276 (2024).
URL https://arxiv.org/abs/2410.21276

[15] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 Technical Report, arXiv preprint arXiv:2505.09388 (2025).
URL https://arxiv.org/abs/2505.09388

[16] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al., Deepseekmath: Pushing the Limits of Mathematical Reasoning in Open Language Models, arXiv preprint

arXiv:2402.03300 (2024).
URL https://arxiv.org/abs/2402.03300

[17] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing Reasoning Capability in Llms via Reinforcement Learning, arXiv preprint arXiv:2501.12948 (2025).
URL https://arxiv.org/abs/2501.12948

[18] S. Robertson, H. Zaragoza, et al., The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends® in Information Retrieval, Vol. 3, No. 4, 2009, pp. 333–389.
URL https://www.nowpublishers.com/article/Details/INR-019

[19] G. Dong, H. Yuan, K. Lu, C. Li, M. Xue, D. Liu, W. Wang, Z. Yuan, C. Zhou, J. Zhou, How Abilities in Large Language Models are Affected by Supervised Fine-tuning Data Composition, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 177–198.
URL https://aclanthology.org/2024.acl-long.12/