FIND: Toward Multimodal Financial Reasoning and Question Answering for Indic Languages

Sarmistha Das[1*], **Vaibhav Vishal**[1*], **Syed Ibrahim Ahmad**[1*]
Manish Gupta[2], **Sriparna Saha**[1]
[1]Indian Institute of Technology Patna, India [2]Microsoft, India
sarmistha1515@gmail.com, vvaibhav728@gmail.com, syediahmad0@gmail.com
gmanish@microsoft.com, sriparna@iitp.ac.in

Abstract

Financial decision-making in multilingual settings demands accurate numerical reasoning grounded in diverse modalities, yet existing benchmarks largely overlook this high-stakes, real-world challenge, especially for Indic languages. We introduce *FinVQA*, a benchmark for evaluating financial numerical and multimodal reasoning in multilingual Indic contexts. *FinVQA* spans English, Hindi, Bengali, Marathi, Gujarati, and Tamil, and comprises 18,900 samples across 14 financial domains. The dataset captures diverse reasoning paradigms under realistic constraints, and is structured across three difficulty levels (easy, moderate, hard) and four question formats: multiple choice, fill-in-the-blank, table matching, and true/false. To address these challenges, we propose *FIND*, a framework that combines supervised fine-tuning with constraint-aware decoding to promote faithful numerical reasoning, robust multimodal grounding, and structured decision-making. Together, *FinVQA* and *FIND* establish a rigorous evaluation and modeling paradigm for high-stakes multilingual multimodal financial reasoning[1].

1 Introduction

Recently, driven by advances in train-time and test-time scaling (Kaplan et al., 2020; OpenAI, 2024), large language models (LLMs) have demonstrated substantial improvements in long-horizon reasoning through structured inference and strategic deliberation (Yue et al., 2024; Das et al., 2024b). These reasoning enhanced models (Ghosh et al., 2025), commonly referred to as Large Reasoning Models (LRMs) (OpenAI et al., 2024; Jaech et al., 2024; OpenAI, 2025; Liu et al., 2024; Guo et al., 2025; Qwen Team, 2024; Kimi Team et al., 2025; Mallick and Kilpatrick, 2025), have achieved strong performance on complex multi-step tasks in domains such as, advisory (Das et al., 2025a,b) programming (Chen, 2021; Jain et al., 2024), mathematics (Lightman et al., 2023; Mao et al., 2024) and science (Lu et al., 2023; Wang et al., 2023; Yue et al., 2024).

However, real-world domain-specific numerical reasoning, particularly in finance, remains a significant challenge, as it requires precise mathematical computation, faithful application of domain knowledge, and robust reasoning over hybrid multimodal contexts such as tables, charts, and textual descriptions (Chen et al., 2023; Das et al., 2023, 2024a; Romera-Paredes et al., 2024; Wang and Zhao, 2024). Existing models often struggle with such hybrid reasoning, leading to errors in numerical accuracy and inconsistencies in multimodal grounding (Tang et al., 2025; Deng et al., 2025). Moreover, vision-language models (VLMs) tend to over-rely on textual cues while underutilizing visual financial signals, further limiting their effectiveness in realistic financial question answering scenarios (Vo et al., 2025).

These challenges are further amplified in multilingual settings such as India, where financial literacy initiatives and educational systems heavily rely on regional Indic languages. As the country undergoes rapid economic growth, enabling accurate and accessible financial reasoning across languages becomes increasingly critical. However, existing benchmarks largely focus on English and lack coverage of Indic languages, multimodal financial data, and controlled reasoning complexity.

To address this gap, we introduce FinVQA, a multilingual benchmark for financial visual question answering in Indic contexts. FinVQA comprises 18,900 samples spanning 14 financial domains, and is designed to capture diverse reasoning paradigms under realistic constraints. The dataset covers English, Hindi, Bengali, Marathi, Gujarati, and Tamil, and is organized into three difficulty levels (easy, moderate, hard) and four question formats: multiple choice, fill-in-the-blank, true/false,

[1*] These authors contributed equally.

Findings of the Association for Computational Linguistics: ACL 2026, pages 32750–32770

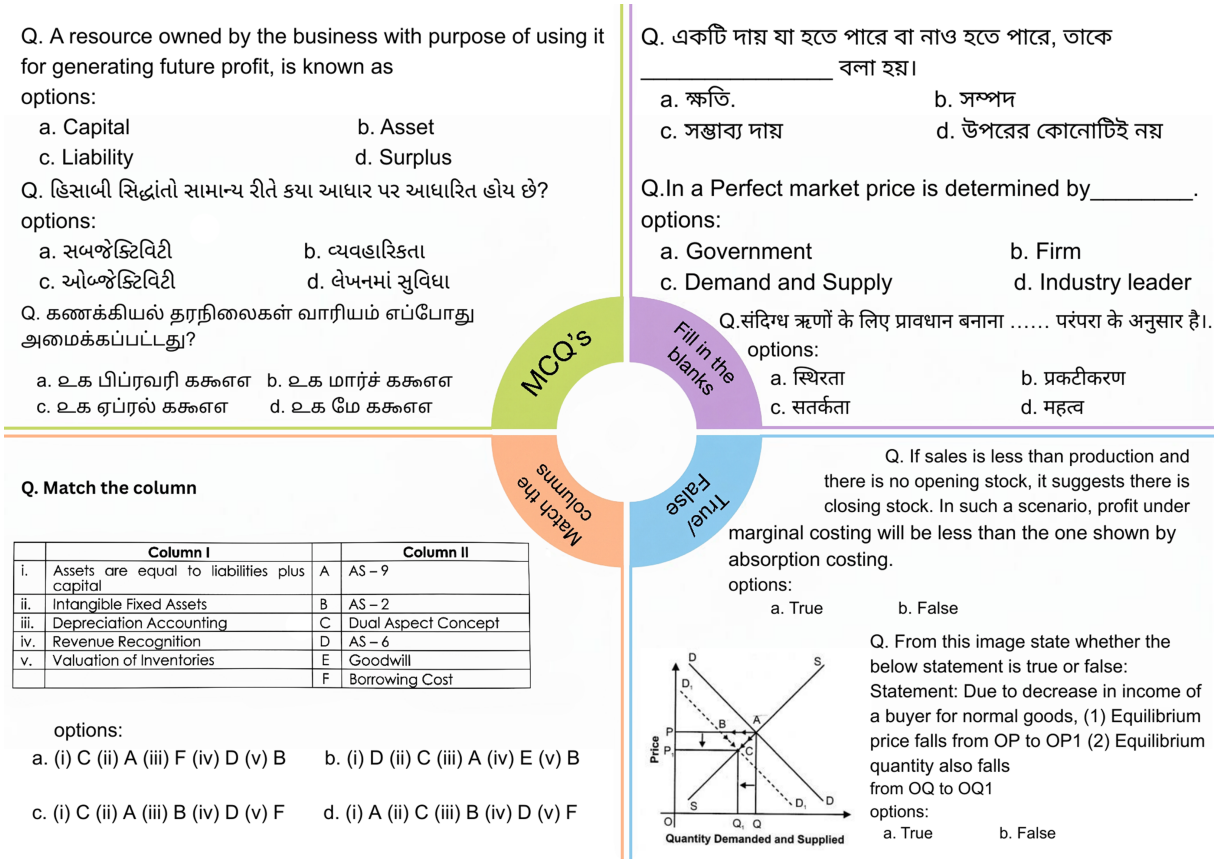


Figure 1: FinVQA Corpus Overview. Illustrative samples spanning four distinct question paradigms: multiple-choice (MCQ), fill-in-the-blanks, true/false, and table-matching.

and table matching. Figure 1 illustrates representative examples from the dataset.

Building on this benchmark, we propose *FIND*, a framework for reliable multimodal financial reasoning that combines supervised fine-tuning with constraint-aware decoding. FIND enforces structured outputs, improves numerical faithfulness, and promotes consistent multimodal grounding, enabling more trustworthy reasoning in high-stakes financial applications.

Our main contributions are as follows:

- We introduce *FinVQA*, a large-scale benchmark for evaluating multilingual financial numerical with supported reasonings, comprising 18,900 samples across 14 financial domains, organized into three difficulty levels and four reasoning formats.
- Considering the linguistic diversity in India[2], *FinVQA* spans English and five major Indic languages, Hindi, Bengali, Marathi, Gujarati, and Tamil. The dataset is designed to evaluate realistic visual-textual financial reasoning.
- We propose *FIND*, a constraint-aware framework that integrates supervised fine-tuning with structured decoding to improve numerical correctness, multimodal alignment, and reliability in financial reasoning tasks.

2 Background

Recent progress in large language models equipped with explicit deliberative and multi-step inference mechanisms, such as OpenAI o1 (OpenAI et al., 2024) and DeepSeek-R1 (Guo et al., 2025), has significantly improved multi-step reasoning performance across domains, including code (Chen, 2021; Jain et al., 2024), mathematics (Lightman et al., 2023; Mao et al., 2024), and science (Lu et al., 2023; Wang et al., 2023; Yue et al., 2024). However, their capabilities on domain-specific numerical reasoning, particularly in finance, remain insufficiently validated.

Existing financial tasks (Das et al., 2024c, 2025c) and their reasoning benchmarks, such as FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), ConvFinQA, and their programmatic exten-

[2]Most Widely Spoken Languages in India

Table 1: Comparison of multilingual financial question answering and related financial reasoning datasets

	Dataset Name	Count (approx.)	Type	Speciality	Modality	Language
Non-Indic	FinQA (Chen et al., 2021)	∼8k QA pairs	Numerical QA	Multi-step numerical reasoning over financial reports	Text + Tables	English
	TAT-QA (Zhu et al., 2021)	∼16.5k QA pairs	Table QA	Numerical reasoning over financial tables with text	Text + Tables	English
	ConvFinQA (Chen et al., 2022)	∼14.1k conv. turns	Conversational QA	Multi-turn numerical reasoning over financial data	Text + Tables	English
	MultiHiertt (Klimaszewski et al., 2025)	∼10k sentence pairs	NLI / reasoning	Hierarchical text-table entailment, includes finance	Text + Tables	English
	FinMME (Luo et al., 2025)	∼11k samples	Multimodal QA	Financial chart/table understanding and reasoning	Text + Charts + Tables	English
	FinDER (Choi et al., 2025)	∼5.7k queries	Retrieval QA	Retrieval-augmented financial QA	Text	English
	FinLLMs (Yuan et al., 2024) (core dataset)	∼15k synthetic QA	Numerical QA	Programmatic generation of financial reasoning problems	Text + Tables	English
	FinTruthQA (Xu et al., 2024)	∼6k QA pairs	Investor-firms interactions	Investor questions and Company responses	Text	Chinese
Indic	IndicFinNLP (Exaggerated Numeral Detection) (Ghosh et al., 2024)	∼6.5k financial statements	Binary classification	Sustainability Classification in financial texts	Text	Hindi, Bengali, Telugu
	Proposed ***FinVQA***	18,900	MCQ, Fill in the blanks, True & False, Column Matching	Financial Problem specific reasonings	Text + Images	English, Hindi, Bengali, Gujarati, Marathi, Tamil

sions CodeFinQA and CodeTAT-QA (Krumdick et al., 2024) largely focus on tabular extraction and shallow arithmetic, limiting their ability to distinguish advanced LRMs from standard LLMs. Other datasets, including FinCode (Krumdick et al., 2024) and FinanceMath (Zhao et al., 2024), suffer from limited complexity, ambiguous problem formulations, and relaxed evaluation protocols, hindering accurate assessment of true reasoning ability.

Moreover, as evidenced by existing benchmarks (Table 1), financial reasoning datasets are overwhelmingly English-centric, with minimal support for multilingual or Indic-language settings, and limited coverage of multimodal financial VQA. This gap is particularly critical in regions such as India, where financial education and decision-making often rely on regional languages. These limitations highlight the need for a multilingual, multimodal, and rigorously annotated financial reasoning benchmark that evaluates complex, domain-aware reasoning beyond surface-level numerical accuracy motivating the development of *FIND*[3].

3 FinVQA Dataset Curation

To promote societal impact through education-oriented artificial intelligence, we curate a high-quality collection of *financial and economic reasoning questions* from authoritative Indian educational sources. Our primary data source is textbooks published by the National Council of Educational Research and Training (NCERT), India's apex body for school education, covering core financial concepts in *Accountancy, Business Studies, and Economics* for Classes 9–12[4]. These materials provide structured coverage of topics such as financial accounting, financial management, and demand supply analysis. To

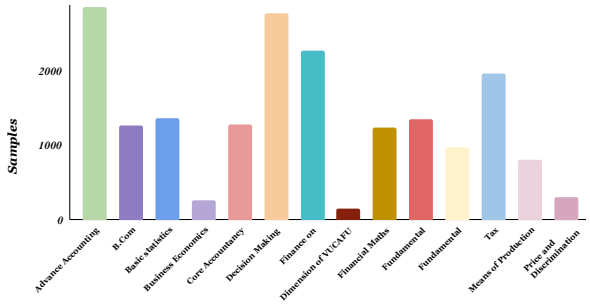


Figure 2: Domain-wise distribution of FinVQA.

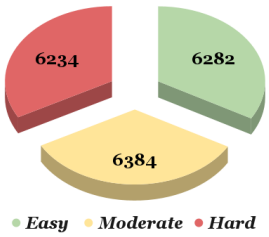


Figure 3: Difficulty-wise distribution of FinVQA.

complement this foundation, we additionally incorporate professionally oriented content from the ICMAI–CMA program, administered by the *Institute of Cost Accountants of India (ICMAI)*, a statutory body under the Ministry of Corporate Affairs[5]. From these curated sources, we initially

[3]Resources for *FIND* and *FinVQA*

[4]NCERT website

[5]ICMAI website

collect over 3,000 English-language samples spanning both *unimodal textual questions* and *multimodal visual question answering (VQA)* formats. Following rigorous manual scrutiny and quality control, we retain 2519 text-only samples and 631 image-based financial reasoning samples. This filtering process ensures unambiguous problem formulation, numerical correctness, and suitability for evaluating both conceptual understanding and quantitative reasoning in financial contexts.

3.1 Multilingual Data Curation

To construct the multilingual textual samples, we employed GPT-4o to translate the original English instances into five target languages: Hindi, Bengali, Gujarati, Marathi, and Tamil (see Appendix B for further details). Translation quality was quantitatively evaluated using cosine similarity over sentence embeddings generated by the *multilingual-MiniLM-L12-v2* (Reimers and Gurevych, 2019) Sentence Transformer model, thereby ensuring semantic alignment across languages. To further strengthen translation reliability, a back-translation procedure was additionally performed. Finally, for authoritative validation, we engaged five native speakers one per target language who manually reviewed each translated instance and resolved linguistic ambiguities, idiomatic inconsistencies, and contextual deviations, ensuring high linguistic accuracy and cultural fidelity of the multilingual corpus.

3.2 OCR-Based Text Replacement

Images in the dataset contain English text which also needs to be translated to target languages. We utilized PaddleOCR (Cui et al., 2025) that outputs structured information for each detected text string from these images, including the recognized text, its bounding-box coordinates, and an orientation flag indicating horizontal or vertical alignment. Leveraging this metadata, the original English text is replaced with its translated counterpart at the exact spatial location from which it was extracted. Figure 6 in the appendix represents one instance.

To maintain visual coherence, the original text region is first removed via background-matched inpainting, followed by rendering the translated text within the same bounding box, thereby preserving the original layout and spatial structure.

For rendering Indic-language translations, we employ the `NotoSans-{IndicLanguage}` font families to ensure consistent glyph coverage and script fidelity. The font size is adaptively determined based on the detected region geometry; specifically, for horizontally aligned text, the font size is set to 90% of the bounding-box height, ensuring optimal readability while respecting spatial constraints.

3.3 Back-Translation

We employ back-translation to enhance linguistic diversity and robustness in multilingual settings by translating instances to an intermediate language and back, preserving semantics. For *FIND*, back-translations are generated using GPT-4o, and semantic fidelity is verified via cosine similarity, with most instances achieving 85–90% similarity to the originals. Low-similarity cases are manually reviewed to ensure numerical consistency and constraint preservation.

3.4 Dataset Annotation and Quality Check

To construct *FinVQA*, we adopted a structured human annotation pipeline as represented in Figure 4a involving six annotators. Among them, four junior annotators were undergraduate-level students, while the remaining two senior annotators were domain experts with professional experience in finance. To establish consistent annotation standards, we initially used 3150 English data samples before translation, and senior experts first annotated 200 seed samples, comprising both image-based and text-based questions, which served as reference exemplars for training junior annotators. Detailed annotation guidelines are provided in the Appendix B.

During the final annotation phase, all junior annotators were required to strictly adhere to the predefined annotation rules. Upon completion, the senior experts conducted an *inter-annotator agreement* analysis, achieving a Fleiss' Kappa score (Gwet, 2014) of 0.78, indicating substantial agreement. The dataset was subsequently refined through expert-led validation based on the following criteria:

1. **Question Framing and Option Balance:** Each question was examined to ensure clear formulation and an equitable distribution of multiple-choice options (a, b, c, d). For numerical problems, structured and logically

consistent reasoning traces were verified and retained.

2. **Difficulty and Question-Type Distribution:** The dataset was balanced across *easy*, *moderate*, and *hard* difficulty levels, while maintaining diversity in question formats, including fill-in-the-blank, true/false, and table-matching tasks.

3. **Expert Correction and Multimodal Validation:** Any detected error in questions, answer options, numerical values, or reasoning steps were manually corrected by senior annotators. For multimodal samples, image quality and semantic alignment with the corresponding textual context were rigorously verified to ensure clarity and relevance.

Conclusively, the proposed *FinVQA* corpus comprises a total of 18,900 carefully curated instances, including 15,114 text-only and 3,786 image-based samples. The dataset spans 14 distinct financial domains and is systematically organized across three levels of question difficulty: easy (6282), moderate (6384), and hard (6234). The domain-wise distribution and difficulty-wise distribution are illustrated in Figures 2 and 3 respectively, while comprehensive dataset statistics and construction details are provided in Appendix A.

4 The *FIND* Methodology

4.1 Multimodal Financial QA Problem

We address multimodal financial reasoning using a multiple-choice question answering (MCQA) formulation. Each instance in the dataset $\mathcal{D}$ comprises a question q, a set of candidate answers $\mathcal{C} = \{c_1, \ldots, c_K\}$, optional visual input $\mathcal{I}$ (e.g., chart, table, receipt, or document), the correct answer $c^* \in \mathcal{C}$, and a ground-truth explanation represented as $\{q, \mathcal{C}, \mathcal{I}, c^*\} \sim \mathcal{D}$. The learning objective must encourage solutions that are both accurate and logically coherent with their explanations, reflecting the reasoning demands of financial tasks. Our proposed framework is illustrated in Figure 4b.

4.2 The *FIND* Learning Framework

To systematically study multimodal financial reasoning under varying degrees of supervision and output control, we design the experimental pipeline with three-variations, namely : (i) zero-shot inference, (ii) constrained decoding, and (iii) supervised fine-tuning (SFT). Each phase progressively increases task specialization and structural enforcement while operating over the same underlying MCQA formulation.

Let π_θ denote a vision-language model parameterized by θ, which defines a conditional distribution over output sequences given multimodal inputs. To conduct these tasks the subsequent three variant experimental pipelines are discussed below. The corresponding prompts are given in Appendix C.

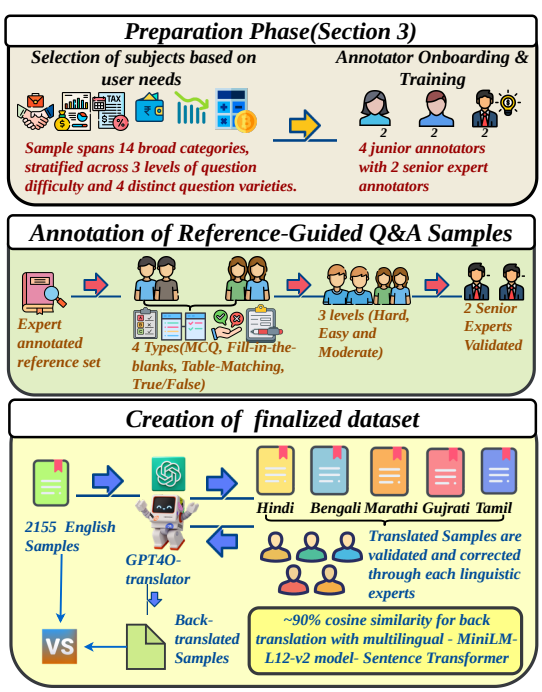
4.2.1 Variation I: Zero-Shot Inference

In the first phase, we evaluate the intrinsic reasoning capability of pretrained vision–language models without task-specific adaptation. Given an input instance $x = \{q, \mathcal{C}, I\}$, the model directly estimates the conditional probability $\pi_\theta(y \mid x) = \pi_\theta(\hat{e}, \hat{c} \mid q, \mathcal{C}, I)$, where $\hat{e}$ denotes the generated reasoning sequence and $\hat{c} \in \mathcal{C}$ is the predicted answer option. Inference is performed autoregressively by factorizing the output distribution as $\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid y_{<t}, q, \mathcal{C}, I)$, without explicit constraints on output structure or length. While this setup enables unbiased evaluation of generalization, it often results in formatting instability, verbosity, and language inconsistency, particularly in multilingual financial settings.

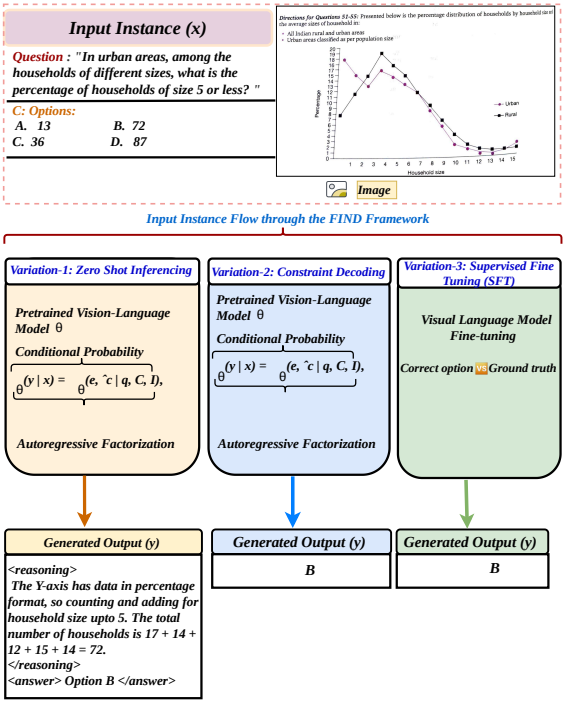
4.2.2 Variation II: Constrained Decoding for Structured Prediction

To mitigate the common failure modes such as overly verbose responses, format violations, code-switching across languages, and the generation of repetitive or irrelevant content, observed in zero-shot inference we introduce constrained decoding at inference time implemented using vLLM (Kwon et al., 2023). Technically, this is realized by restricting the model's output space to valid multiple-choice answer option tokens (e.g., A/B/C/D) while preserving the internal reasoning process.

Let $\mathcal{V}$ denote the full tokenizer vocabulary and $\mathcal{V}_{\text{valid}} \subset \mathcal{V}$ the subset corresponding to valid answer options. During decoding, the token-level distribution is constrained as $\tilde{\pi}_\theta(y_t \mid \cdot) = \pi_\theta(y_t \mid \cdot)$ if $y_t \in \mathcal{V}_{\text{valid}}$, and 0 otherwise, implemented by masking invalid logits with $-\infty$ prior to softmax. The decoding objective is thus $\hat{c} = \arg\max_{c_k \in \mathcal{C}} \sum_{t=1}^{|c_k|} \log \pi_\theta(c_{k,t} \mid c_{k,<t}, x)$, where $|c_k|$ denotes the number of subword tokens in option c_k. To account for tokenizer variability, we apply language-aware constraints: single-token decoding for English, Hindi, and Marathi, and up to



(a) End-to-End Dataset Annotation and Construction Pipeline.



(b) High-level architecture of the proposed *FIND* framework for multimodal financial QA.

Figure 4: Overview of the *FinVQA* dataset construction pipeline and the proposed *FIND* framework.

three tokens for Bengali, Gujarati, and Tamil. This strategy enforces strict format compliance, suppresses spurious generations, and ensures stable, comparable predictions across languages without modifying model parameters.

4.2.3 Variation III: SFT of Vision Language Models

In this variant, we apply SFT to align the vision–language model with the financial **MCQA** task under answer-only supervision. Given a labeled dataset $\mathcal{D} = \{(x_i, c_i^*)\}_{i=1}^{N}$, where $c_i^* \in \mathcal{C}_i$ denotes the correct option for input x_i, the model is trained to generate the target answer text corresponding to the correct choice.

Since answer options may consist of multiple tokens, the model learns to predict the sequence of tokens forming the correct answer conditioned on the input. The training objective maximizes the likelihood of generating this correct answer sequence given the input.

Parameter-Efficient Attention Adaptation. To limit overfitting and training cost, we fine-tune only the query, key, and value projections in self- and cross-attention layers. For each layer, projections are computed as $Q = XW_Q$, $K = XW_K$, $V = XW_V$, and optimization is restricted to $\theta_{\text{train}} = \{W_Q^{(\ell)}, W_K^{(\ell)}, W_V^{(\ell)}\}_{\ell=1}^{L}$, while all remaining parameters $\theta_{\text{frozen}} = \theta \setminus \theta_{\text{train}}$ are held fixed. Attention is computed as $\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$. To further enhance parameter efficiency, we utilise Low-Rank Adaptation (LoRA) (Hu et al., 2022) with rank $r = 8$, scaling factor $\alpha = 32$, and dropout rate of 0.1.

This design enables selective recalibration of multimodal interactions, particularly between visual representations $\mathcal{I}$ and textual queries, while mitigating catastrophic forgetting. Although reasoning tokens are not verbalised or explicitly supervised, latent reasoning patterns emerge through attention reweighting induced by answer-level supervision. To prevent verbosity and format violations, the model is trained with SFT and deployed with the constrained decoding strategy from variation II, ensuring high predictive accuracy and strict structural validity.

The dataset was split into training and testing sets in the ratio 70:30. For all variations the metrics are reported using the test set. We train models for 3 epochs, with the learning rate for training fixed at 1e-5, and the optimiser used was the 8-bit variant of AdamW. All of the experiments are conducted on a machine with one Nvidia A100 80GB GPU.

5 Experiments and Results

This section presents the experimental setup, research questions, and the principal conclusions drawn from the empirical evaluation. Our study is guided by the following research questions:

- **RQ1:** What is the overall performance of the evaluated models?
- **RQ2:** Which modeling technique exerts the greatest influence on performance?
- **RQ3:** What language-specific performance trends and insights emerge?
- **RQ4:** How do the experimental results justify the chosen model framework design?

Models and Evaluation Protocol. We evaluate a diverse set of state-of-the-art VLMs spanning multiple parameter scales, including compact (4B), medium (7B–12B), and large-scale (>27B) architectures. Specifically, our experiments consider models from the *Qwen2.5-VL* (Bai et al., 2025), *Qwen3-VL* (Yang et al., 2025) and *Gemma-3* (Gemma Team, 2025) families. Each model is assessed under zero-shot, constrained-decoding, and SFT settings. Model performance is quantitatively evaluated using accuracy as the primary metric.

5.1 Analytical Discussion

This section synthesises the empirical findings and distils key insights in relation to the research questions.

5.1.1 Answer to RQ1 (Overall Performance)

Across both text-only and multimodal settings, model performance exhibits a clear hierarchy governed by model scale and training regime. In the zero-shot text setting (Table 2), smaller models ($\leq$4B) show weak multilingual generalization, with accuracies often below 30 for Marathi, Gujarati, and Tamil (e.g., Qwen2.5-VL-3B: 14.3 in Marathi, 21.2 in Tamil; Qwen3-VL-4B: 17.5 in Tamil). In contrast, large-scale models perform substantially better; for instance, Qwen3-VL-32B-Instruct achieves 70.9 (Hindi), 66.1 (Bengali), and 60.6 (Marathi) in zero-shot inference. Following supervised fine-tuning, text-only performance improves considerably, even with Qwen3-VL-32B-Instruct reaching balanced accuracies across Indic languages such as 68.9 (Hindi), 68.5 (Bengali), 69.4 (Marathi), and 68.0 (Gujarati) while smaller models (3B–7B) remain below 40 in low-resource languages. In multimodal evaluations, constrained decoding and SFT (Table 3) progressively enhance performance over zero-shot baselines, particularly for Indic languages. Overall, English consistently attains the highest accuracy, and large-scale VLMs outperform smaller counterparts across all modalities and evaluation regimes, confirming strong capacity-dependent cross-lingual generalization.

5.1.2 Answer to RQ2 (Greatest Influence)

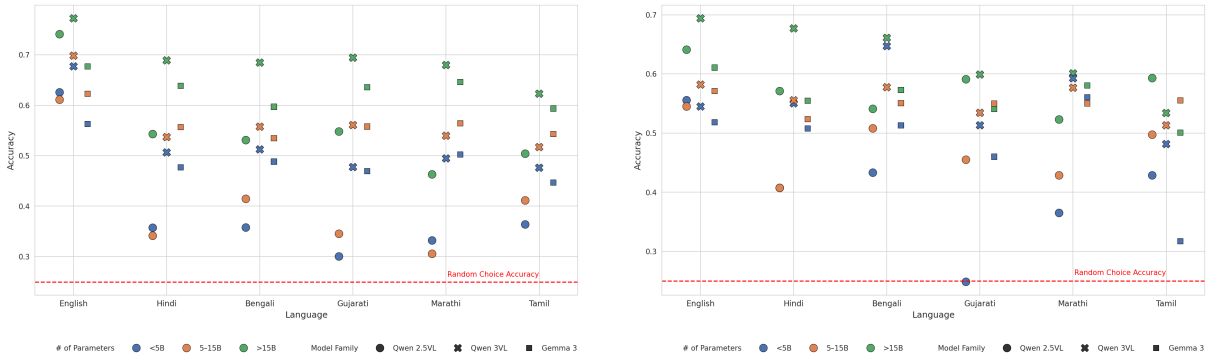
The results show that SFT exerts the strongest influence on performance, substantially outweighing gains from constrained decoding alone. While constrained decoding yields modest and sometimes inconsistent improvements, SFT produces pronounced cross-lingual gains. For example, Qwen3-VL-4B improves Marathi accuracy from 26.2 (zero-shot) to 47.8 (+21.6 points) after SFT, with similar trends in Gujarati and Tamil.

In the multimodal setting, Qwen3-VL-8B increases Marathi accuracy from 27.5 (zero-shot) to 52.9 with constrained decoding, and further stabilizes around 53.4 after SFT. Likewise, Qwen3-VL-32B improves Bengali accuracy from 49.1 (zero-shot) to 65.2 (constrained decoding) and reaches 66.1 post-SFT. Overall, while constrained decoding aids structural validity, SFT delivers the most consistent and robust multilingual improvements, with additional epochs yielding diminishing but stabilizing gains.

5.1.3 Answer to RQ3 (Linguistic trends)

Language-wise analysis reveals stable and interpretable trends across modalities and training regimes. Hindi and Bengali consistently achieve the highest accuracies, often exceeding 60 in large models, reflecting higher resource availability and closer alignment with English. Marathi and Gujarati exhibit the largest relative gains from constrained decoding and SFT, frequently improving by 15-30 absolute points compared to zero-shot multimodal inference (e.g., Qwen3-VL-4B Marathi: 19.0 $\rightarrow$ 50.8 $\rightarrow$ 51.3), highlighting their sensitivity to alignment-based training.

Tamil remains the most challenging language, with lower absolute accuracies (below 65 even after SFT), likely due to greater typological distance and script complexity; nevertheless, it shows consistent post-SFT improvements. Overall, SFT substantially reduces cross-lingual performance variance, indicating enhanced language-agnostic rea-



(a) Text-only dataset

(b) Text + image dataset

Figure 5: Inference results of the SFT models.

Table 2: Model-wise Accuracy for text only part of dataset across Indic languages using zero shot inference, constrained decoding, and SFT models inferenced with constrained decoding

Model Name	Zero shot inference						Constrained decoding						Supervised finetuning					
	en	hi	bn	mr	gu	ta	en	hi	bn	mr	gu	ta	en	hi	bn	mr	gu	ta
Qwen2.5-VL-3B-Instruct	57.5	26.7	27.3	14.3	25.4	21.2	63.1	36.9	20.9	29.4	32.4	20.2	62.6	35.7	35.8	30.0	33.2	36.4
Qwen3-VL-4B-Instruct	69.6	51.2	48.5	26.2	39.6	17.5	67.2	49.6	40.8	14.3	49.1	46.7	67.7	50.7	51.3	47.8	49.5	47.6
Gemma-3-4b-it	61.5	50.3	52.2	50.4	41.3	45.2	56.3	46.3	49.1	48.4	49.6	40.9	56.3	47.8	48.9	47.0	50.3	44.7
Qwen2.5-VL-7B-Instruct	55.6	36.6	37.4	20.5	31.3	27.6	61.2	39.8	34.2	18.3	34.1	40.1	61.1	34.1	41.5	34.5	30.6	41.1
Qwen3-VL-8B-Instruct	74.1	58.5	58.4	49.5	56.5	46.4	69.0	53.4	45.0	28.0	53.6	52.1	69.8	53.7	55.8	56.1	54.0	51.7
Gemma-3-12b-it	71.7	62.7	66.6	63.4	64.3	57.8	62.3	58.2	58.5	57.3	59.3	55.3	62.3	55.7	53.5	55.8	56.5	54.4
Qwen2.5-VL-32B-Instruct	73.8	40.6	27.9	22.6	34.4	33.9	70.5	48.9	35.1	35.4	45.0	50.4	74.1	54.3	53.1	54.8	46.3	50.4
Qwen3-VL-32B-Instruct	77.8	70.9	66.1	60.6	69.4	57.3	76.9	68.9	52.1	33.6	67.2	61.2	77.2	68.9	68.5	69.4	68.0	62.3
Gemma-3-27b-it	73.7	70.4	67.4	69.7	68.9	65.6	69.2	64.0	61.3	63.4	64.3	61.2	67.7	63.9	59.7	63.6	64.7	59.4

soning rather than superficial transfer.

5.1.4 Answer to RQ4 (Framework Justification)

The empirical results strongly validate the proposed framework, which combines large-scale instruction-tuned VLMs with supervised alignment and constrained decoding. Large models such as Qwen3-VL-32B and Gemma-3-27B consistently achieve the highest and most stable performance across text-only and multimodal settings (Figure 5), often yielding 15-30 point gains over zero-shot baselines in low-resource Indic languages after SFT.

Smaller models become competitive only after supervised alignment, underscoring the importance of capacity-aware architectures. In contrast, weakly aligned models (e.g., LLaVA variants) fail under constrained decoding, exhibiting near-zero accuracies across Indic languages. Additionally, the diminishing yet stabilizing gains across successive SFT epochs confirm that early alignment combined with controlled inference rather than reliance on zero-shot generalization is critical for robust multilingual multimodal reasoning, thereby justifying the framework's design choices.

5.2 Human Evaluation

We conduct a comprehensive human evaluation of the zero-shot inference results adhering to a well-defined evaluation protocol in appendix D. The assessment is grounded in a structured rubric designed to capture critical dimensions of financial question answering quality, including Financial Domain Understanding, Problem Interpretation, Mathematical Correctness, Formula Application, Reasoning Consistency, and Formatting Compliance, enabling fine-grained assessment of reasoning and usability.

The results (Table 5 in the appendix) show a strong scale–performance relationship. While Formatting Compliance remains relatively stable across models (≈0.7–0.8), higher-order competencies particularly Financial Domain Understanding and Problem Interpretation are strongly capacity-dependent. Smaller models (e.g., Qwen2.5-VL-3B, Gemma-3-4B) score as low as 0.1-0.2, whereas large models (Gemma-3-27B, Qwen3-VL-32B) approach near-ceiling performance (0.9-1.0). The Qwen3-VL family consistently outperforms comparable baselines, with Qwen3-VL-32B achieving the highest aggregate score (0.967). Overall, syntactic compliance is achievable at low compute, but high-fidelity financial reasoning re-

Table 3: Model-wise Accuracy for multimodal version of dataset across Indic languages using zero shot inference, constrained decoding, and SFT models inferenced with constrained decoding

Model Name	Zero shot inference						Constrained decoding						Supervised finetuning					
	en	hi	bn	mr	gu	ta	en	hi	bn	mr	gu	ta	en	hi	bn	mr	gu	ta
Qwen2.5-VL-3B-Instruct	52.4	33.9	31.0	13.2	25.9	22.2	54.5	41.8	46.5	24.3	35.4	40.2	55.6	40.7	43.3	24.9	36.5	42.9
Qwen3-VL-4B-Instruct	61.9	46.0	47.1	19.0	37.0	12.2	55.6	55.0	62.0	50.8	60.8	46.0	54.5	55.0	64.7	51.3	59.3	48.1
Gemma-3-4b-it	37.6	44.4	42.8	41.8	43.4	30.2	53.4	52.9	54.5	44.4	54.0	29.6	51.9	50.8	51.3	46.0	56.1	31.7
Qwen2.5-VL-7B-Instruct	61.4	46.6	49.2	26.5	42.9	45.5	53.4	41.3	52.4	47.6	47.6	49.2	54.5	40.7	50.8	45.5	42.9	49.7
Qwen3-VL-8B-Instruct	64.6	49.2	47.6	27.5	44.4	41.8	58.2	55.0	57.8	52.9	56.6	51.3	58.2	55.6	57.8	53.4	57.7	51.3
Gemma-3-4b-it	66.7	55.0	60.4	58.7	60.3	58.2	59.3	52.4	53.5	52.4	54.5	54.5	57.1	52.4	55.1	55.0	55.0	55.6
Qwen2.5-VL-32B-Instruct	67.2	36.0	22.5	9.0	32.8	30.1	58.7	55.6	54.5	54.5	48.1	55.0	64.1	57.1	54.1	59.1	52.3	59.3
Qwen3-VL-32B-Instruct	68.3	51.8	49.1	31.3	49.1	45.9	62.4	63.0	65.2	58.7	61.9	53.4	69.4	67.7	66.1	59.9	60.1	53.4
Gemma-3-4b-it	69.3	62.4	59.9	55.6	61.9	63.5	59.3	52.9	58.8	52.9	59.8	50.8	61.1	55.5	57.3	54.1	58.1	50.1

quires large-scale models.

5.3 Qualitative Analysis

A qualitative analysis of the models across three parameter scales reveals a distinct progression in multimodal and mathematical reasoning capabilities. As showcased using a representative sample in Table 6 in the appendix, at the 3-4B parameter scale the models exhibited foundational failures in visual grounding, inaccurately extracting data points from the chart and demonstrating a severe logical disconnect by selecting a final option that contradicted its own flawed reasoning trace. Scaling to the 7-12B parameter range yielded significant improvements in visual parsing and logical architecture, enabling the model to formulate a valid and efficient problem-solving strategy; however, its execution remained brittle, as evidenced by minor arithmetic inaccuracies during the calculation phase.

In contrast, the 27-32B model demonstrated highly robust and flawless execution across the entire visual-reasoning pipeline. It accurately extracted all relevant data, executed a precise multi-step calculation with the correct unit conversions, and logically arrived at the correct answer, ultimately highlighting that reliable, end-to-end performance on complex chart interpretation and mathematical reasoning tasks currently necessitates larger parameter scales.

5.4 Error Analysis

Analysis of the failure cases reveals clear patterns across model families and scales. We show an example using multi-lingual responses in Table 7. Smaller models (3-4B) frequently show weak visual grounding, evidenced by inaccurate extraction of data points, misunderstanding of core task objectives, and severe logical disconnects between intermediate reasoning traces and final answer selections. As capacity increases (7-12B), these foundational parsing issues largely diminish, with mid-sized models exhibiting improved vision-language alignment and the ability to formulate valid, multi-step problem-solving strategies. However, performance bottlenecks persist at this intermediate scale, predominantly manifesting as brittle mathematical execution where models falter on minor arithmetic inaccuracies despite sound logical frameworks.

At the larger 27-32B scale, models successfully bridge these gaps, demonstrating robust data extraction and flawless end-to-end mathematical execution. Nonetheless, the persistent arithmetic fragility of intermediate models indicates that while structural visual understanding emerges at lower parameter counts, reliable execution of coupled visual-mathematical tasks remains highly dependent on increased model scale.

6 Conclusion and future work

This paper introduces *FinVQA*, a multimodal Financial VQA benchmark covering English and five Indic languages such as Hindi, Bengali, Marathi, Gujarati, and Punjabi. It evaluates multilingual financial numerical problems along with their corresponding reasoning across 14 domains, structured into three difficulty levels (easy, moderate, hard) with diverse task formats. By integrating supervised fine-tuning with constraint-aware decoding, *FIND* enables rigorous and reliable assessment of numerical faithfulness, multimodal alignment, and structured decision-making in high-stakes financial settings.

As a future direction, we aim to investigate reward-based methods in reinforcement learning environments and undertake a more comprehensive analysis of multilingual financial reasoning, with a particular focus on examining how performance varies across languages and identifying the factors that influence reasoning quality in each linguistic setting.

7 Limitations

Despite demonstrating encouraging performance, the models exhibit several notable limitations. We observe frequent failures in verbal output quality, including repetitive generations, overly verbose responses, and inconsistent or unstable reasoning. In multiple cases, the reasoning presented by the model does not align with the final answer selected, indicating a disconnect between intermediate inference and answer selection. Additionally, inconsistent adherence to the required output format often results in multiple or ambiguous answer predictions, further reducing reliability. We employ two complementary techniques: constrained decoding and supervised fine-tuning which mitigate a subset of these issues. However, several challenges remain unresolved:

- **Multi-step reasoning consistency:** Difficulty in maintaining consistent and correct reasoning across multiple inference steps.
- **Visual information extraction and utilization:** Inability to accurately extract and effectively use information from the visual modality.
- **Financial domain understanding:** Limited robustness in financial knowledge and conceptual understanding, especially in complex or multimodal scenarios.
- **Sample Coverage:** The current benchmark includes 18,900 multilingual instances generated from approximately 3,150 unique source questions translated across multiple languages. This construction supports systematic cross-lingual comparison by preserving semantic alignment across languages; however, it also implies that the diversity of underlying source instances is narrower than the total sample count alone might indicate. We view this as a practical first step toward multilingual financial benchmarking, and in future work we plan to expand the resource with more natively authored and semantically distinct questions to strengthen linguistic diversity, cultural nuance, and benchmark coverage.
- **Closed-Source Baselines**: Our evaluation does not yet include a wider set of closed-source baselines or more extensive Chain-of-Thought analysis, mainly due to limited computational and API resources. As the study was conducted without dedicated external funding, large-scale experimentation with high-cost proprietary models was not feasible. We will clarify this constraint in the camera-ready version and discuss the resulting evaluation gap accordingly.

8 Ethical Considerations

FIND targets financial numerical and multimodal reasoning, a high-stakes domain where incorrect or misleading outputs can have real economic consequences. A primary data-level ethical concern is the risk of misuse, as models evaluated on *FIND* may be perceived as reliable financial advisors despite the benchmark not being designed for prescriptive or personalized financial decision-making. Accordingly, *FIND* is explicitly intended for evaluation and research only and contains no real user data or individualized financial scenarios.

From a linguistic perspective, although *FIND* covers multiple Indic languages, uneven model performance across languages may reinforce existing disparities in financial literacy if not carefully interpreted. To mitigate this risk, the dataset maintains balanced difficulty levels and domain coverage across languages, and emphasizes language-wise reporting rather than aggregate claims.

At the model level, *FIND* exposes a critical ethical issue: fluency without correctness. Large reasoning models often generate confident explanations even when numerical reasoning or formula application is incorrect, posing a risk of user over-trust in financial contexts. Additionally, vision–language models frequently under-utilize visual financial evidence, relying instead on textual priors, which can lead to systematically flawed multimodal reasoning.

Finally, while constraint-aware decoding improves format compliance, it may also mask underlying reasoning errors, creating an illusion of reliability. Overall, *FIND* highlights that model scale and structured outputs alone do not guarantee trustworthy financial reasoning, underscoring the necessity of human oversight, transparent evaluation, and cautious interpretation in any downstream use.

9 Acknowledgement

All authors sincerely acknowledge the invaluable contributions of the dataset annotators Paavne, Jheel, Yajant, and Alekhya who were instrumen-

tal in the successful completion of the annotation process through their dedicated, proactive, and sustained efforts.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Mark Chen. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. 2025. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 638–646.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *Preprint*, arXiv:2510.14528.

Sarmistha Das, Ujjwal Chowdhury, Nandakumar S Lijin, Atulya Deep, Sriparna Saha, and Alka Maurya. 2024a. Investigate how market behaves: toward an explanatory multitasking based analytical model for financial investments. *IEEE Access*, 12:30928–30940.

Sarmistha Das, Tuhinangshu Gangopadhyay, Atulya Deep, Sriparna Saha, and Alka Maurya. 2023. "find the table": A contrastive learning-based approach with faster rcnn for establishing tabular entity relationships. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Sarmistha Das, RE Zera Marveen Lyngkhoi, Sriparna Saha, and Alka Maurya. 2024b. Wealth guide: a sophisticated language model solution for financial trading decisions. In *Proceedings of the eighth financial technology and natural language processing and the 1st agent AI for scenario planning*, pages 133–140.

Sarmistha Das, RE Zera Marveen Lyngkhoi, Sriparna Saha, and Alka Maurya. 2025a. Unlocking financial insights: An advanced multimodal summarization with multimodal output framework for financial advisory videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11976–11985.

Sarmistha Das, Priya Mathur, Ishani Sharma, Sriparna Saha, Kitsuchart Pasupa, and Alka Maurya. 2025b. Fin-ally: Pioneering the development of an advanced, commonsense-embedded conversational ai for money matters. *arXiv preprint arXiv:2509.24342*.

Sarmistha Das, Basha Mujavarsheik, R E Zera Lyngkhoi, Sriparna Saha, and Alka Maurya. 2025c. Deciphering the complaint aspects: towards an aspect-based complaint identification model with video complaint dataset in finance. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 7195–7204.

Sarmistha Das, Apoorva Singh, Sriparna Saha, and Alka Maurya. 2024c. Negative review or complaint? exploring interpretability in financial complaints. *IEEE Transactions on Computational Social Systems*, 11(3):3606–3615.

Shuangyan Deng, Haizhou Peng, Jiachen Xu, Chunhou Liu, Ciprian Doru Giurcuaneanu, and Jiamou Liu. 2025. Understanding financial reasoning in ai: A multimodal benchmark and error learning approach. *arXiv preprint arXiv:2506.06282*.

Team Gemma Team. 2025. Gemma 3.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*.

Sohom Ghosh, Arnab Maji, Aswartha Narayana, and Sudip Kumar Naskar. 2024. Indicfinnlp: Financial natural language processing for indian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9010–9018.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in

llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kimi Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Mateusz Klimaszewski, Pinzhen Chen, Liane Guillou, Ioannis Papaioannou, Barry Haddow, and Alexandra Birch. 2025. Avenibench: Accessible and versatile evaluation of finance intelligence. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 111–117.

Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. Bizbench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8309–8332.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, and 1 others. 2025. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. *arXiv preprint arXiv:2505.24714*.

Shrestha Basu Mallick and Logan Kilpatrick. 2025. Gemini 2.0: Flash, flash-lite and pro.

Yujun Mao, Yoon Kim, and Yilun Zhou. 2024. Champ: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities. *arXiv preprint arXiv:2401.06961*.

A Jaech OpenAI, Adam Kalai, Adam Lerer, Adam Richardson, A El-Kishky, A Low, A Helyar, A Madry, A Beutel, A Carney, and 1 others. 2024. Openai o1 system card, 2024. *URL https://arxiv. org/abs/2412.16720*.

Brian Zhang OpenAI. 2025. Openai o3-mini system card. *https://cdn. openai. com/o3-mini-system-card-feb10. pdf*.

Team OpenAI. 2024. Learning to reason with llms. *OpenAI Blog*.

Qwen Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Zichen Tang, Jiacheng Liu, Zhongjun Yang, Rongjin Li, Zihua Rong, Haoyang He, Zhuodi Hao, Xinyang Hu, Kun Ji, Ziyan Ma, and 1 others. 2025. Finmmr: make financial numerical reasoning more multimodal, comprehensive, and challenging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3245–3257.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025. Vision language models are biased. *arXiv preprint arXiv:2505.23941*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926.

Ziyue Xu, Peilin Zhou, Xinyu Shi, Jiageng Wu, Yikang Jiang, Dading Chong, Bin Ke, and Jie Yang. 2024. Fintruthqa: A benchmark dataset for evaluating the quality of financial information disclosure. *arXiv preprint arXiv:2406.12009*.

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.

Ziqiang Yuan, Kaiyuan Wang, Shoutai Zhu, Ye Yuan, Jingya Zhou, Yanlin Zhu, and Wenqi Wei. 2024. Finllms: A framework for financial reasoning dataset generation with large language models. *IEEE Transactions on Big Data*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Dataset Details

A.1 Domain Coverage and Definition

The dataset is systematically organized across **14 financial, economic, and decision-centric domains**, each targeting distinct reasoning competencies relevant to real-world financial understanding. Below, we define each domain along with representative attributes illustrating its scope.

- **Advance Accounting (2,856 samples):** Focuses on higher-order accounting treatments requiring multi-step reasoning. *Attributes:* partnership admission and retirement, goodwill valuation, amalgamation, revaluation accounts, company final accounts.
- **Core Accountancy (1,290 samples):** Covers procedural accounting fundamentals emphasizing accuracy and rule-based reasoning. *Attributes:* journal entries, ledger posting, trial balance preparation, error rectification.
- **Fundamental Accountancy (1,356 samples):** Introduces basic accounting principles and conceptual foundations. *Attributes:* matching concept, accrual accounting, depreciation, asset–liability classification.
- **Finance on Accountancy (2,274 samples):** Integrates financial analysis with accounting data for decision-making. *Attributes:* ratio analysis, capital structure decisions, breakeven analysis, financial statement interpretation.
- **Financial Mathematics (1,236 samples):** Targets numerical and quantitative reasoning in financial contexts. *Attributes:* simple and compound interest, annuities, time value of money, percentage change, ratio computation.
- **Basic Statistics (1,374 samples):** Covers descriptive and introductory inferential statistics applied to financial data. *Attributes:* mean, median, mode, probability basics, data interpretation from tables and charts.
- **Business Economics (270 samples):** Examines firm-level economic reasoning and managerial decision contexts. *Attributes:* cost functions, revenue analysis, production decisions, profit maximization.

- **Fundamental Economics (978 samples):** Introduces core microeconomic concepts and market mechanisms. *Attributes:* demand and supply analysis, elasticity, consumer behavior, equilibrium pricing.
- **Price and Discrimination (300 samples):** Focuses on pricing strategies and market regulation. *Attributes:* price floors and ceilings, degrees of price discrimination, monopolistic pricing.
- **Means of Production (804 samples):** Addresses the role of production factors in economic output. *Attributes:* land, labor, capital, entrepreneurship, factor remuneration.
- **Decision Making (2,784 samples):** Captures applied reasoning under uncertainty and constraints. *Attributes:* logical reasoning, risk assessment, optimization, scenario-based choices.
- **Taxation (1,962 samples):** Covers compliance-oriented financial reasoning and tax computation. *Attributes:* income tax calculation, GST, exemptions, deductions, tax slabs.
- **B.Com (1,266 samples):** Represents integrated undergraduate-level commerce problems combining multiple disciplines. *Attributes:* mixed accounting–finance problems, numerical aptitude, applied economics.
- **Dimensions of VUCAFU (150 samples):** Targets modern strategic reasoning under dynamic economic conditions. *Attributes:* volatility, uncertainty, complexity, ambiguity, fragility, unpredictability.

A.2 Difficulty Stratification

Each domain is uniformly stratified across three levels of cognitive complexity: **Easy (6,282 samples)** targeting conceptual recall and single-step reasoning, **Moderate (6,384 samples)** emphasizing applied multi-step computation, and **Hard (6,234 samples)** requiring complex, multi-hop financial reasoning and decision analysis.

B Text–Image Annotation Guidelines

B.1 Textual Annotation Principles

To ensure semantic accuracy, consistency, and reasoning fidelity in textual annotations, annotators must adhere to the following principles:

1. **Semantic Faithfulness:** Text annotations must accurately reflect the intended meaning of the input question or content, without introducing unsupported assumptions or hallucinated information.
2. **Conceptual Correctness:** Ensure that all domain-specific concepts (e.g., financial, economic, statistical) are used correctly and consistently, adhering to standard definitions and conventions.
3. **Clarity and Precision:** Textual explanations should be concise, unambiguous, and logically structured, avoiding unnecessary verbosity while preserving essential reasoning steps.
4. **Reasoning Transparency:** When explanations are required, intermediate reasoning must align coherently with the final answer, ensuring traceability between assumptions, computations, and conclusions.
5. **Context Appropriateness:** Incorporate only the contextual information necessary to interpret or solve the task. Irrelevant or speculative details should be excluded.
6. **Consistency and Formatting:** Maintain a uniform annotation style across all samples, including consistent terminology, notation, and response structure.
7. **Bias and Neutrality Check:** Avoid subjective, cultural, or demographic bias in textual descriptions and reasoning. All annotations should remain neutral and task-focused.

B.2 Image Annotation and Validation Rules

The following rules govern the annotation and validation of visual inputs to ensure alignment, quality, and reliability in multimodal settings:

1. **Rule 1: Text–Image Semantic Alignment** The image must provide visual evidence that is directly relevant to the associated text or question. Visual content should support, not contradict, the intended reasoning task.
2. **Rule 2: Visual Evidence Sufficiency** Annotate only those images where the necessary information is visually discernible. Im-

ages that are ambiguous, misleading, or insufficient to support reasoning should be excluded.

3. **Rule 3: No Embedded or Overlay Text**
   Images must not contain embedded text, labels, watermarks, or annotations, as these can introduce unintended shortcuts or textual bias.

4. **Rule 4: Visual Quality and Readability**
   Images should be clear, properly cropped, and well-illuminated. Low-resolution, blurred, or cluttered visuals that hinder interpretation are not permitted.

5. **Rule 5: Numerical and Structural Integrity**
   For charts, tables, or financial visuals, ensure that axes, symbols, and numerical values are visually intact and correctly rendered, without distortion or truncation.

6. **Rule 6: Human and Object Realism**
   Any depicted humans or objects must appear natural and undistorted. Images exhibiting anatomical inconsistencies or AI-induced artifacts should be rejected.

These guidelines ensure that textual and visual annotations remain semantically aligned, logically coherent, and visually reliable, thereby enabling robust multimodal reasoning, evaluation, and model generalisation.

C Prompts

Prompt used for Zero-shot inference variation

System Role: You are a helpful assistant.
Instruction: You will be given a multiple-choice question (MCQ). Solve the question and choose the correct option from the given choices (A, B, C, or D). Explain your reasoning step by step before providing your final answer.
Format your output as:
<reasoning> reasoning_text </reasoning>
<answer> Option [A] or [B] or [C] or [D] </answer>
The question and the options are given in {langauge} language. Provide the answer and reasoning in the same language - {langauge}, and the tags in English as shown above. The question and the possible answers are as follows:
Question: {question} Options: {options}

Prompt used for Constrained decoding variation

System Role: You are a helpful assistant.
Instruction: You will be given a multiple-choice question (MCQ). Solve the question and choose the correct option from the given choices (A, B, C, or D). Explain your reasoning step by step before providing your final answer.
The output should be a single letter only corresponding to the correct option: A or B or C or D
Do not output anything else.
The question and the options are given in {langauge} language. Provide the answer option in the same language - {langauge}. The question and the possible answers are as follows:
Question: {question} Options: {options}

D Human Evaluation Details

Human evaluation was conducted exclusively under the zero-shot setting (Variation-1). This restriction is intentional: in the zero-shot regime, models generate complete outputs comprising reasoning traces, option selection, and final answer justification, enabling a faithful assessment of inter-

Table 4: Multilingual Sample Instance (Sample ID: 3) Across Six Languages

Language	Details
English	**Sub-Domain:** Core Accountancy **Question:** If average inventory is 1,25,000 and closing inventory is 10,000 less than opening inventory, then the value of closing inventory will be. **Options:** [A] 1,35,000 [B] 1,15,000 [C] 1,30,000 [D] 1,20,000 **Answer:** Option [D] **Reasoning:** Average Inventory = (Opening + Closing)/2. Let Opening = X, Closing = X − 10,000. 1,25,000 = (2X − 10,000)/2 ⇒ 2,50,000 = 2X − 10,000 ⇒ 2,60,000 = 2X ⇒ X = 1,30,000. Closing = 1,20,000.
Hindi	**Sub-Domain:** कोर अकाउंटेंसी **Question:** यदि औसत इन्वेंटरी ₹१,२५,००० है और समापन इन्वेंटरी, उद्घाटन इन्वेंटरी से ₹१०,००० कम है, तो समापन इन्वेंटरी का मूल्य क्या होगा? **Options:** [क] ₹१,३५,००० [ख] ₹१,१५,००० [ग] ₹१,३०,००० [घ] ₹१,२०,००० **Answer:** विकल्प [घ] **Reasoning:** औसत इन्वेंटरी = (उद्घाटन इन्वेंटरी + समापन इन्वेंटरी)/२। मान लें उद्घाटन इन्वेंटरी = X, तब समापन इन्वेंटरी = X − ₹१०,०००। ₹१,२५,००० = (२X − ₹१०,०००)/२ ⇒ ₹२,५०,००० = २X − ₹१०,००० ⇒ ₹२,६०,००० = २X ⇒ X = ₹१,३०,०००। समापन इन्वेंटरी = ₹१,२०,०००।
Bengali	**Sub-Domain:** কোর অ্যাকাউন্টেন্সি **Question:** যদি গড় ইনভেন্টরি ₹১,২৫,০০০ হয় এবং সমাপনী ইনভেন্টরি উদ্বোধনী ইনভেন্টরির থেকে ₹১০,০০০ কম হয়, তবে সমাপনী ইনভেন্টরির মূল্য কত হবে? **Options:** [ক] ₹১,৩৫,০০০ [খ] ₹১,১৫,০০০ [গ] ₹১,৩০,০০০ [ঘ] ₹১,২০,০০০ **Answer:** বিকল্প [ঘ] **Reasoning:** গড় ইনভেন্টরি = (উদ্বোধনী ইনভেন্টরি + সমাপনী ইনভেন্টরি)/২। ধরা যাক উদ্বোধনী ইনভেন্টরি = X, তবে সমাপনী ইনভেন্টরি = X − ₹১০,০০০। ₹১,২৫,০০০ = (২X − ₹১০,০০০)/২ ⇒ ₹২,৫০,০০০ = ২X − ₹১০,০০০ ⇒ ₹২,৬০,০০০ = ২X ⇒ X = ₹১,৩০,০০০। সমাপনী ইনভেন্টরি = ₹১,২০,০০০।
Marathi	**Sub-Domain:** कोअर अकाउंटन्सी **Question:** जर सरासरी साठा ₹१,२५,००० असेल आणि समापन साठा उद्घाटन साठ्यापेक्षा ₹१०,००० ने कमी असेल, तर समापन साठ्याची किंमत किती असेल? **Options:** [क] ₹१,३५,००० [ख] ₹१,१५,००० [ग] ₹१,३०,००० [घ] ₹१,२०,००० **Answer:** पर्याय [घ] **Reasoning:** सरासरी साठा = (उद्घाटन साठा + समापन साठा)/२। गृहित धरा उद्घाटन साठा = X, तर समापन साठा = X − ₹१०,०००। ₹१,२५,००० = (२X − ₹१०,०००)/२ ⇒ ₹२,५०,००० = २X − ₹१०,००० ⇒ ₹२,६०,००० = २X ⇒ X = ₹१,३०,०००। समापन साठा = ₹१,२०,०००।
Gujarati	**Sub-Domain:** કોર એકાઉન્ટન્સી **Question:** જો સરેરાશ ઇન્વેન્ટરી ₹१,२५,००० છે અને ક્લોઝિંગ ઇન્વેન્ટરી ઓપનિંગ ઇન્વેન્ટરી કરતાં ₹१०,००० ઓછી છે, તો ક્લોઝિંગ ઇન્વેન્ટરીનું મૂલ્ય કેટલું હશે? **Options:** [ક] ₹१,३५,००० [ખ] ₹१,१५,००० [ગ] ₹१,३०,००० [ઘ] ₹१,२०,००० **Answer:** વિકલ્પ [ઘ] **Reasoning:** સરેરાશ ઇન્વેન્ટરી = (ઓપનિંગ ઇન્વેન્ટરી + ક્લોઝિંગ ઇન્વેન્ટરી)/૨। માનીએ કે ઓપનિંગ ઇન્વેન્ટરી = X, તો ક્લોઝિંગ ઇન્વેન્ટરી = X − ₹१०,०००। ₹१,२५,००० = (२X − ₹१०,०००)/२ ⇒ ₹२,५०,००० = २X − ₹१०,००० ⇒ ₹२,६०,००० = २X ⇒ X = ₹१,३०,०००। ક્લોઝિંગ ઇન્વેન્ટરી = ₹१,२०,०००।
Tamil	**Sub-Domain:** கோர் அக்கௌண்டன்சி **Question:** சராசரி சரக்குக,உடு,ooo ஆகவும், நிறைவுச் சரக்குதொடக்கச் சரக்கை விட ₹க०,००० கறவைக இருந்தால், நிறைவுச் சரக்கின் மதிப்பாவ்வளவு? **Options:** [க] ₹க,ஙடு,००० [உ] ₹க,கடு,००० [ங] ₹க,ங०,००० [ச] ₹க,உ०,००० **Answer:** விரல்பம் [ச] **Reasoning:** சராசரி சரக்கு= (தொடக்கச் சரக்கு+ நிறைவுச் சரக்கு)/உ. தொடக்கச் சரக்கு= X எனக் கொள்ளுங்கள். அப்பொதுநிறைவுச் சரக்கு X − ₹க०,०००. ₹க,உடு,००० = (உX − ₹க०,०००)/உ ⇒ ₹உ,டு०,००० = உX − ₹க०,००० ⇒ ₹உ,சு०,००० = உX ⇒ X = ₹க,ங०,०००. நிறைவுச் சரக்கு ₹க,உ०,०००.

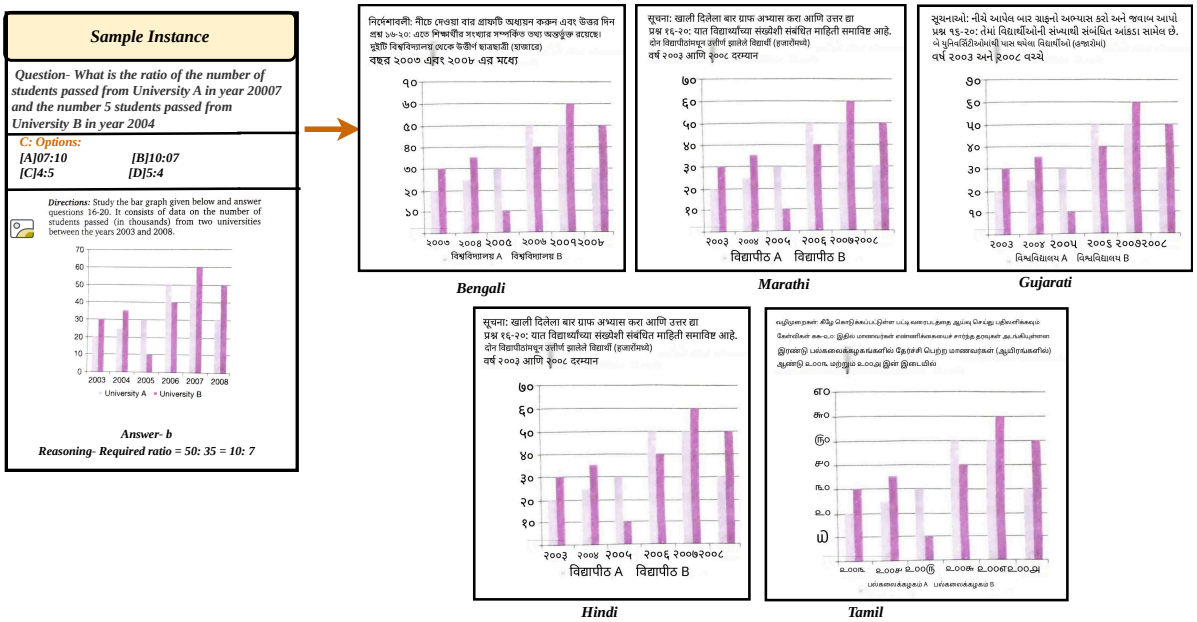


Figure 6: OCR-Based Text Conversion.

mediate reasoning quality. In contrast, under supervised fine-tuning and constraint decoding settings, the generation process becomes restricted, with models predominantly producing option letters without explicit reasoning traces. This limits the ability to evaluate logical coherence, and stepwise correctness. Consequently, the proposed human evaluation framework is applied solely to zero-shot outputs. To ensure balanced multilingual evaluation, we sample 100 instances per language, and all scores reported in Table 5 are derived from this controlled human annotation protocol. The evaluation follows a structured, rubric-driven methodology designed to capture multiple dimensions of financial reasoning beyond surface-level correctness. Specifically, the rubric evaluates six core criteria:

- Financial Domain Understanding, assessing whether the model correctly captures domain-specific financial concepts.
- Problem Interpretation and Assumption Validity, measuring the correctness of question comprehension and the plausibility of implicit assumptions.
- Mathematical Correctness, verifying numerical accuracy and computational validity.
- Formula Selection and Application, evaluating whether appropriate financial or mathematical formulations are correctly identified and applied.
- Reasoning Consistency and Answer Selection, assessing internal logical coherence and alignment between reasoning and final answer.
- Formatting and Output Compliance, ensuring adherence to the prescribed response structure.

This evaluation strategy enables a fine-grained, reasoning-centric analysis, moving beyond exact-match metrics to assess the faithfulness, interpretability, and practical usability of model outputs in financial QA. The results in Table 5 reveal a consistent scaling trend across all model families, where performance improves with increasing parameter size. For instance, the average score for Qwen2.5VL increases from 0.46 (3B) to 0.70 (7B) and 0.94 (32B); Gemma 3 improves from 0.60 (4B) to 0.80 (12B) and 0.96 (27B); and Qwen3VL rises from 0.66 (4B) to 0.85 (8B) and 0.98 (32B). This progression indicates that larger models not only achieve higher answer accuracy but also exhibit substantially improved reasoning fidelity, formula grounding, and structural compliance.

A metric-wise analysis further highlights that Formatting and Output Compliance consistently achieves near-perfect scores for medium and large models, often reaching 1.00, indicating strong adherence to output constraints. Similarly, Formula Selection and Application and Reasoning Consistency demonstrate significant gains with scale, with several large models attaining perfect scores, reflecting robust capability in structured financial reasoning. In contrast, Financial Domain Understanding remains comparatively challenging for smaller models (e.g., 0.13 for Qwen2.5VL-3B, 0.24 for Gemma 3-4B, and 0.32 for Qwen 3VL-4B), suggesting that such models often lack deep semantic grounding despite occasionally producing correct outputs.

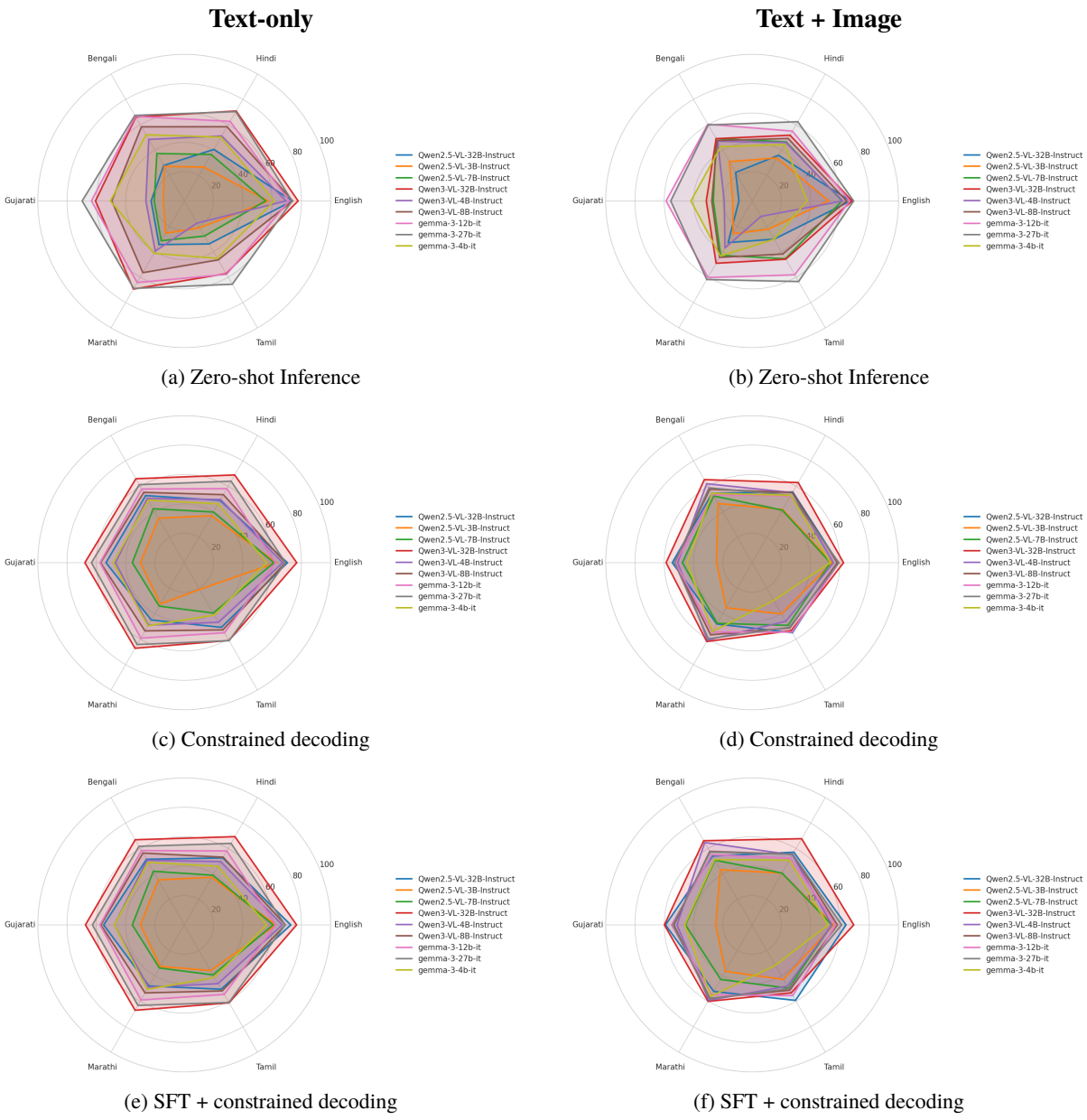
Overall, this human evaluation framework provides a rigorous, multi-dimensional assessment of financial QA systems, explicitly capturing reasoning quality, numerical faithfulness, and usability. By leveraging a multilingual, balanced sample (100 instances per language) and a carefully designed rubric, the analysis offers a comprehensive diagnostic perspective, complementing automatic metrics and enabling a more reliable evaluation of model behavior in high-stakes financial reasoning tasks.

E LLM usage

We used large language models (LLMs) to assist with code development and minor editing of the final manuscript.

Table 5: Human evaluation scores across model families and parameter scales.

Metrics	Qwen2.5VL			Gemma 3			Qwen3VL		
	3B	7B	32B	4B	12B	27B	4B	8B	32B
Financial Domain Understanding	0.13	0.33	0.82	0.24	0.43	0.91	0.32	0.53	0.92
Problem Interpretation & Assumption Validity	0.34	0.63	0.83	0.54	0.73	0.84	0.64	0.74	0.93
Mathematical Correctness	0.52	0.74	1.00	0.63	0.84	1.00	0.72	0.94	1.00
Formula Selection & Application	0.63	0.74	1.00	0.73	0.83	1.00	0.73	0.83	1.00
Reasoning Consistency and Answer Selection	0.44	0.83	1.00	0.63	1.00	1.00	0.73	1.00	1.00
Formatting & Output Compliance	0.74	1.00	1.00	0.83	1.00	1.00	0.83	1.00	1.00
Average Score	0.46	0.70	0.94	0.60	0.80	0.96	0.66	0.85	0.98



(a) Zero-shot Inference

(b) Zero-shot Inference

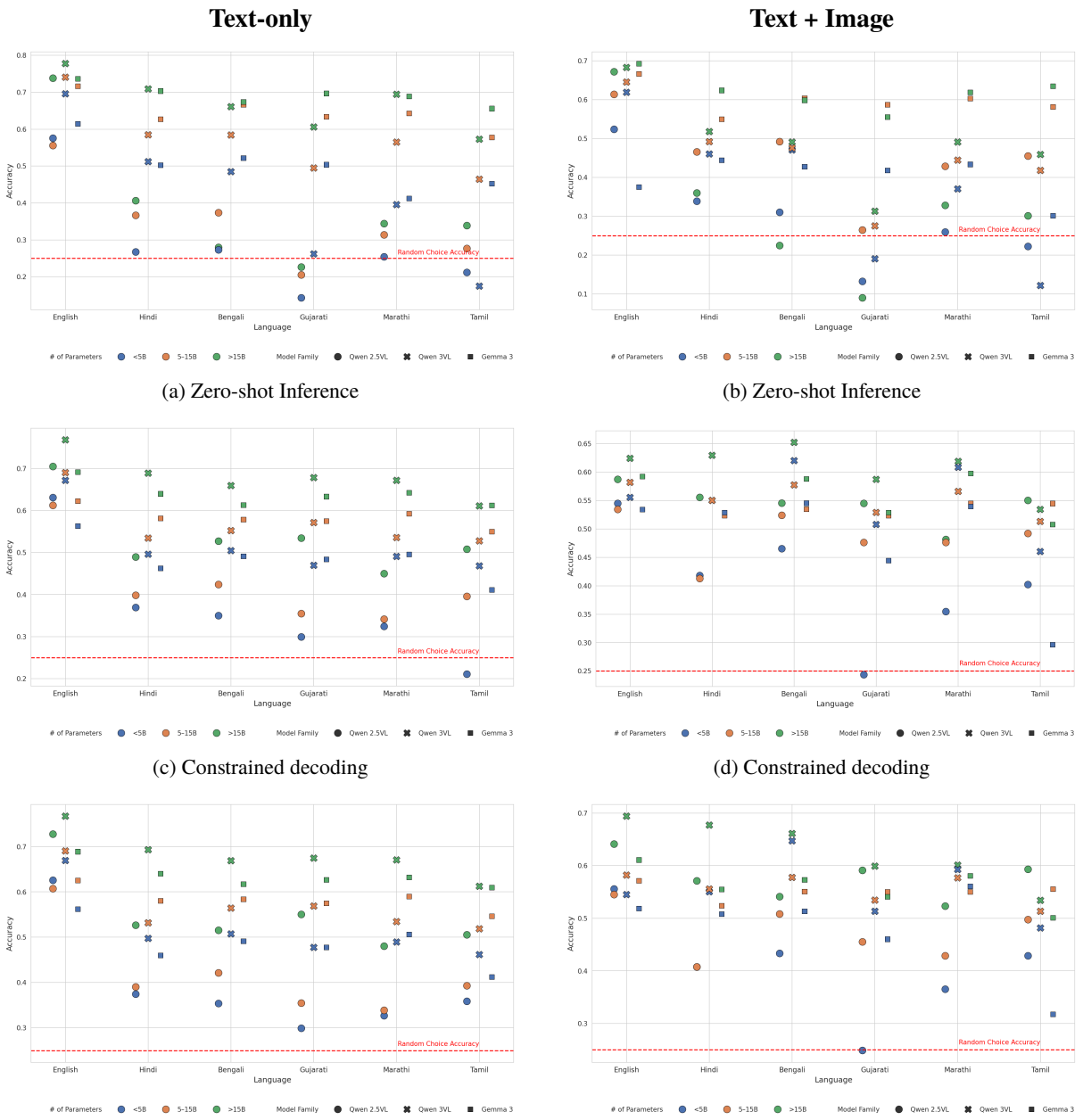
(c) Constrained decoding

(d) Constrained decoding

(e) SFT + constrained decoding

(f) SFT + constrained decoding

Figure 7: Comparison of model performance across inference and training strategies. Rows correspond to different generation strategy, while columns contrast text-only and text+image inputs.



(a) Zero-shot Inference

(b) Zero-shot Inference

(c) Constrained decoding

(d) Constrained decoding

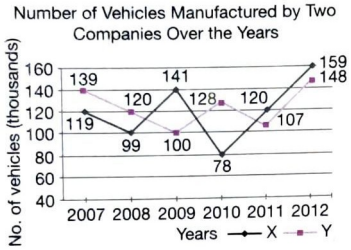
(e) SFT + constrained decoding

(f) SFT + constrained decoding

Figure 8: Comparison of model performance across inference and training strategies. Rows correspond to different generation strategy, while columns contrast text-only and text+image inputs.

Table 6: Responses generated by various models for an English sample from FinVQA

Directions: Study the following graph carefully and answer Questions 61-65. The graph consists of data about the number of vehicles manufactured by two companies over the years.



Question- What is the difference between the total production of the two companies in the given years?
Options- A. 19,000 B. 22,000 C. 26,000 D. 28,000

Answer- C

Reasoning:
Total production of Company X from 2007 to 2012 in thousands = 119 + 99 + 141 + 78 + 120 + 159
= 716.
Total production of Company Y from 2007 to 2012 in thousands = 139 + 120 + 100 + 128 + 107 + 148
= 742.
Difference = 742 - 716 thousands = 26,000; so the correct option is C.

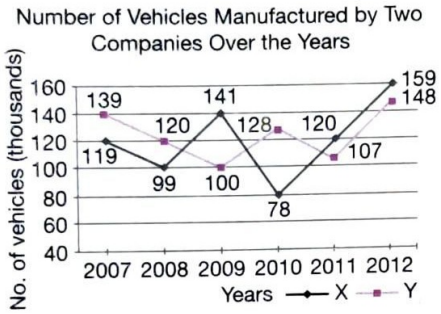
Response by Qwen 2.5 VL 3B -
<reasoning>
First, I need to extract the data points for both companies from the graph and find their total sums.
Company X values: 119, 99, 141, 78, 120, 159.
Sum of X = 119 + 99 + 141 + 78 + 120 + 159 = 716.
Company Y values: 139, 120, 100, 128, 107, 148.
Sum of Y = 139 + 120 + 110 + 128 + 107 + 148 = 752.
Next, find the difference between the two totals: 752 - 716 = 36.
The y-axis is labeled ”No. of vehicles (thousands)”, so the actual difference is 34 * 1000 = 34,000. This is closest to the option D.
</reasoning>
<answer>D </answer>

Response by Qwen 3 VL 7B -
<reasoning>
To find the total difference between the two companies, an alternative to summing all values first is to calculate the net difference year by year (Company Y - Company X) and then sum those differences:
2007: 139 - 119 = +20
2008: 120 - 99 = +21
2009: 100 - 141 = -41
2010: 128 - 78 = +50
2011: 107 - 120 = -11
2012: 148 - 159 = -11
Now, sum these annual differences: 20 + 21 - 41 + 50 -11 - 11 = 26.
According to the vertical axis, the numbers are in thousands. Therefore, a difference of 26 corresponds to 26,000 vehicles.
</reasoning>
<answer>Option C </answer>

Response by Gemma 3 27B -
<reasoning>
Based on the provided line graph, we must determine the total production for Company X and Company Y over the six-year period from 2007 to 2012, calculate the absolute difference between these totals, and adjust for the units. Note the y-axis explicitly states ”No. of vehicles (thousands)”.
Step 1: Calculate the total production for Company X (represented by the black line with diamond markers).
2007: 119
2008: 99
2009: 141
2010: 78
2011: 120
2012: 159
Total for Company X = 119 + 99 + 141 + 78 + 120 + 159 = 716 (in thousands).
Step 2: Calculate the total production for Company Y (represented by the purple line with square markers).
2007: 139
2008: 120
2009: 100
2010: 128
2011: 107
2012: 148
Total for Company Y = 139 + 120 + 100 + 128 + 107 + 148 = 742 (in thousands).
Step 3: Find the difference between the two total productions.
Difference = Total Y - Total X = 742 - 716 = 26 (in thousands).
Step 4: Convert to the final numerical value. 26 thousands is equal to 26,000. Checking the options provided, 26,000 corresponds to Option C.
</reasoning>
<answer>C </answer>

Table 7: Multilingual Error Analysis for a graph-based financial reasoning sample

Directions: Study the following graph carefully and answer Questions 61-65. The graph consists of data about the number of vehicles manufactured by two companies over the years.



Question- What is the difference between the total production of the two companies in the given years?
Options- A. 19,000 B. 22,000 C. 26,000 D. 28,000
Answer- C
Reasoning- Total production of Company X from 2007 to 2012 in thousands = 119 + 99 + 141 + 78 + 120 + 159 = 716. Total production of Company Y from 2007 to 2012 in thousands = 139 + 120 + 100 + 128 + 107 + 148 = 742. Difference = 742 - 716 thousands = 26,000; so the correct option is C.

Bengali

<reasoning>
আমাদের কোম্পানি X এবং কোম্পানি Y-এর মোট উৎপাদনের পার্থক্য হিসাব করতে হবে। প্রথম ধাপে, কোম্পানি X এর মোট উৎপাদন যোগ করি: 119 + 99 + 141 + 78 + 120 + 159 = 716। দ্বিতীয় ধাপে, কোম্পানি Y এর মোট উৎপাদন যোগ করি: 139 + 120 + 100 + 128 + 107 + 148 = 754। এখন, এই দুটির মধ্যে পার্থক্য বের করি: 754 - 726 = 28। Y-অক্ষ নির্দেশ করে যে মানগুলি হাজারে রয়েছে। তাই মোট পার্থক্য 28,000।
</reasoning>
<answer>D </answer>

Hindi

<reasoning>
ग्राफ़ के अनुसार, हमें कंपनी X और कंपनी Y का कुल उत्पादन निकालना होगा। कंपनी X का कुल उत्पादन = 119 + 99 + 141 + 78 + 120 + 159 = 716. कंपनी Y का कुल उत्पादन = 139 + 120 + 100 + 128 + 107 + 148 = 742. अब दोनों का अंतर निकालते हैं: 742 - 716 = 16. चूँकि संख्याएँ हजारों में हैं, इसलिए यह 16,000 होगा। यह सीधे तौर पर किसी विकल्प से मेल नहीं खाता, लेकिन सबसे करीब विकल्प A(19,000) है।
</reasoning>
<answer>A </answer>

Marathi

<reasoning>
दोन कंपन्यांच्या उत्पादनातील एकूण फरक शोधण्यासाठी मला डेटा जोडावा लागेल. ग्राफ वाचताना, मी मूल्ये जवळच्या संख्येपर्यंत पूर्ण करतो: कंपनी X ची अंदाजे बेरीज = 120 + 100 + 140 + 80 + 120 + 160 = 720. कंपनी Y ची अंदाजे बेरीज = 140 + 120 + 100 + 130 + 110 + 150 = 750. दोन्हीमधील फरक = 750 - 720 = 30. ग्राफमधील मूल्ये हजारात असल्याने, हा फरक 30,000 असायला हवा. दिलेल्या पर्यायांमध्ये 28,000 हा सर्वात जवळचा पर्याय आहे.
</reasoning>
<answer>D </answer>

Gujarati

<reasoning>
કંપની X અને કંપની Y ના ઉત્પાદનનો તફાવત શોધવાનો છે. હું માત્ર શરૂઆતના અને છેલ્લા વર્ષનો તફાવત જોઈશ. 2007 માં તફાવત: 139 - 119 = 20. 2012 માં તફાવત: 159 - 148 = 11. જો આપણે આ સરેરાશ તફાવતો ઉમેરીએ, તો તે આશરે 19 થાય છે. વાય-અક્ષ મુજબ આ મૂલ્યો હજારોમાં છે, તેથી જવાબ 19,000 છે. આ વિકલ્પ A સાથે મેળ ખાય છે.
</reasoning>
<answer>A </answer>

Tamil

<reasoning>
இரண்டுநிறுவனங்களின் மொத்த உற்பத்தியின் வித்தியாசத்தை நம் கணக்கிட வேண்டும். நிறுவனம் X-இன் தரவுகள்: 119, 99, 141, 78, 120, 159. இதன் மொத்த மதிப்பு 716. நிறுவனம் Y-இன் தரவுகள்: 139, 120, 100, 128, 107, 148. இதன் மொத்த மதிப்பு 742. இப்போதுவித்தியாசத்தை கண வேண்டும்: 742 - 716 = 22. வரைபடத்தில் மதிப்புகள் ஆயிரங்களில் ($thousands$) கொடுக்கப்பட்டள்ளன. எனவே, சரியான வித்தியாசம் 22,000 ஆகும்.
</reasoning>
<answer>B </answer>