ATLAS: A Reasoning-Guided Attribution Framework for Mathematical Chart Analysis

Anonymous ACL submission

Abstract

The human-like capability of Multimodal Large Language Models (MLLMs) like GPT-4o to process both text and images enables them to help humans with quantitative analysis of charts. However, these models are known to hallucinate, more so on vision language tasks; our initial study on a sample from the ChartQA dataset (Masry et al., 2022) indicates that GPT-4o provides accurate answers only 58% of the time for questions on chart images. In this paper, we introduce attribution for chart-based mathematical questions, where bounding boxes identify the key regions that justify answers, building on recent work in factual verification for text-based question answering. Taking inspiration from Chain-of-Thought (CoT)-like prompting strategies, we hypothesize that understanding step-by-step reasoning can help in improving attribution accuracy in chart-based mathematical question-answering. We propose a semi-automatic approach to obtain a benchmarking dataset comprising 7,819 diverse samples with charts, questions, reasoning steps, and attribution annotations. We introduce a method using the open-source Internlm-XComposer2 model with Partial Low-Rank Adaptation, treating vision and language tokens equally to generate high-quality attributions through detailed reasoning steps. Our experimental results show that our approach enhances attribution quality by $\sim$15%, advancing the development of interpretable and trustworthy chart-based AI systems.

1 Introduction

Data visualizations like bar charts and line charts are among the most straightforward tools for representing and analyzing data, helping people make informed decisions. Analyzing these charts often requires performing mathematical calculations and applying formulas to extract insights or answers (Kim et al., 2020). Studies by Satpute et al. (2024); Srivastava et al. (2024); Ahn et al. (2024); Gupta et al. (2024) evaluated Large Language Models (LLMs) (Brown et al., 2020; Jiang et al., 2024; Touvron et al., 2023; Achiam et al., 2023) and Multi-modal Large Language Models (MLLMs) (OpenAI, 2023; Team et al., 2023; Su et al., 2023; Chen et al., 2023) for their ability to answer mathematical questions or provide reasoning using various datasets across tasks like solving geometrical problems combining diagram and text interpretation (Seo et al., 2015) and mathematical word problems (Wang et al., 2017), etc. While LLMs and MLLMs have demonstrated impressive performance in mathematical question-answering tasks, establishing trust in their generated answers through attribution mechanisms is important. This is particularly crucial for mathematical questions involving charts, where numerical accuracy and proper interpretation of visual elements directly impact decision-making in real-world settings.

Prior work on attribution has primarily focused on general text-based question-answering and visual question-answering tasks (Yue et al., 2023; Phukan et al., 2024a,b; Bohnet et al., 2022; Qi et al., 2024). However, directly applying these approaches to mathematical chart question answering presents significant limitations. For instance, in fig 1, when applied to complex mathematical questions involving charts, existing attribution methods often fail to correctly identify the relevant chart regions that contribute to the final answer. To the best of our knowledge, attributing the generated answers to the charts for such complex math questions has been unexplored. In this paper, we address the task of attributing generated answers to specific regions in charts for complex mathematical questions. We focus specifically on line and bar charts, covering a range of mathematical operations including aggregations, comparisons, and trend analysis. Drawing inspiration from Chain-of-Thought (CoT) prompting strategies (Wei et al., 2024; Zhang et al., 2022), we hypothesize that incorporating step-by-

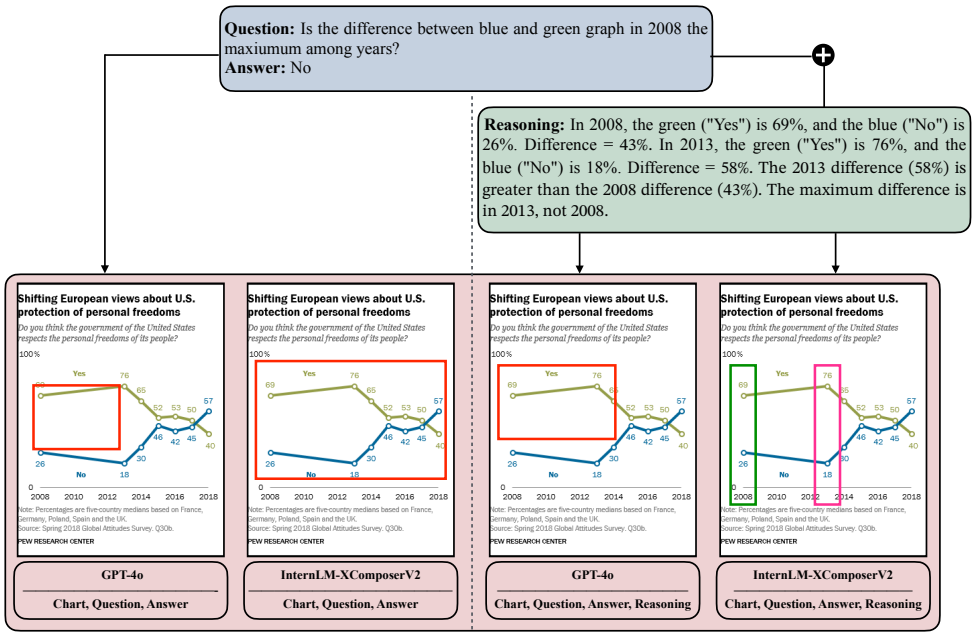


Figure 1: Comparison of attribution methods between GPT-4o and InternLM-XComposer2 on a chart reasoning task. Left: Both models receive only the chart, question, and answer as input. Right: Models additionally receive reasoning steps, leading to more precise attributions. The example shows how incorporating reasoning steps helps InternLM-XComposer2 correctly attribute the relevant data points for comparing differences between lines in 2008 and 2013, while GPT-4o struggles with accurate attribution even with reasoning provided.

step reasoning in the attribution process can improve performance by mimicking human mathematical problem-solving approaches.

We make four main contributions in our work **ATLAS**: **A** Reasoning-Guided A**T**tribution Framework for Mathematica**L** Ch**A**rt Analysi**S**: **(1)** We introduce the **task of attribution for mathematical question answering in charts**, addressing a critical gap in current visual mathematical question answering, as outlined in figure 1. **(2)** We present a systematic **data curation strategy** that combines MLLM-generated reasoning and attribution annotations with human corrections. This results in a high-quality dataset derived from ChartQA (Masry et al., 2022), comprising annotated examples spanning line and bar chart types and mathematical operations. **(3)** We **propose an automatic attribution and reasoning method** that utilizes InternLM-XComposer2 (Dong et al., 2024) model to generate attribution for chart QA (Masry et al., 2022) dataset by taking the reasoning for the answers as input. Our approach utilizes an InternLM-XComposer2 model that uses Partial Low-Rank Adaptation (PLoRA) on to generate reasoning steps from chart-question-answer triples, then uses these to produce attribution bounding boxes. **(4)** Through extensive empirical evaluation, we demonstrate that high-quality **reasoning steps significantly improve attribution accuracy**. Our results show an average of 15% improvement from baseline approaches through our proposed methodology, highlighting current limitations and areas for future enhancement in reasoning generation.

2 Related Work

Recent work has increasingly focused on attribution mechanisms to improve the trustworthiness of AI systems' outputs. For text-based systems, Bohnet et al. (2022) survey attribution methods in open-domain generative systems, highlighting challenges like ambiguous knowledge sources and biases. In the context of question answering, Phukan et al. (2024b) leverage LLMs' hidden state representations to attribute parts of generated answers to source documents, while Qi et al. (2024) propose MIRAGE, a model internals-based approach for faithful answer attribution in retrieval-augmented generation. For multimodal systems, Phukan et al. (2024a) extend the logit lens technique to detect and ground visual hallucinations using contextual token embeddings from middle layers of MLLMs, improving bounding box precision and spatial understanding. Chart-based question answering has emerged as a crucial task for visual data interpre-

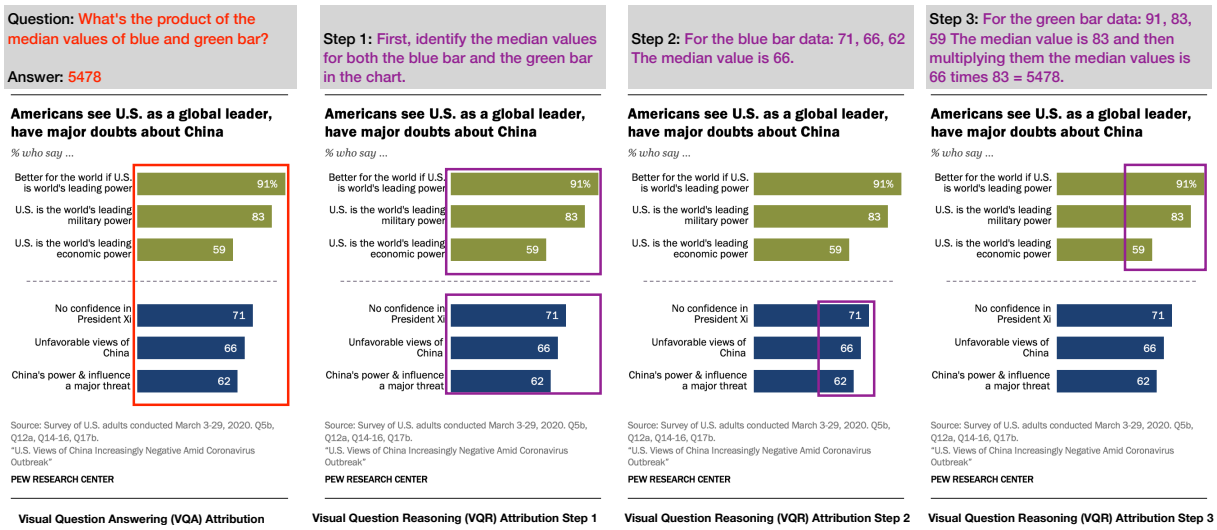


Figure 2: ATLAS's attribution process for mathematical reasoning on charts. Left to right: (1) VQA attribution highlights all relevant data bars needed for computing the final answer, (2) Step 1 attribution identifies the specific bars needed to find median values, (3) Step 2 attribution focuses on the blue bars for calculating their median (66), and (4) Step 3 attribution shows the green bars used for the final multiplication step ($66 \times 83 = 5478$). The progressive attribution demonstrates how our framework traces both the final answer and intermediate reasoning steps to specific chart regions.

tation. The ChartQA dataset (Masry et al., 2022) provides a comprehensive benchmark with 9.6K human-written and 23.1K generated questions for visual and logical reasoning over various chart types. Recent advances like Chart Llama (Han et al., 2023) have demonstrated superior performance in tasks like ChartQA, Chart-to-text, and chart extraction. Supporting technologies such as ChartOCR (Luo et al., 2021) combine deep learning and rule-based methods to effectively extract chart segments, facilitating better chart understanding.

Mathematical reasoning has become increasingly important in AI systems, particularly for chart interpretation where statistical understanding is crucial. Imani et al. (2023) proposed MathPrompter, generating multiple solution paths using zero-shot CoT prompting to improve arithmetic problem-solving. While CoT prompting shows promise with large models, Ranaldi and Freitas (2024) addressed its limitations in smaller models through instruction-tuning. As surveyed by Lu et al. (2023), mathematical reasoning serves as a crucial testbed for evaluating AI systems' capabilities, with implications for chart-based mathematical analysis. The need for trustworthy chart-based mathematical reasoning has highlighted the importance of attribution in this domain. While models like InternLM-XComposer2 (Dong et al., 2024) excel in multimodal understanding through techniques like Partial LoRA for image token processing, existing attribution methods face challenges with mathematical chart questions. Current approaches, while promising for general visual attribution, underperform when dealing with complex mathematical operations on charts, creating a critical gap in trustworthy chart-based mathematical reasoning systems.

3 Attribution Definition

Chart attribution aims to identify regions of a chart that support generated answers, similar to the Grounded Visual Question Answering (VQA) approach proposed by (Phukan et al., 2024a). For mathematical chart question answering, where complex reasoning steps are essential to arrive at answers, we propose a two-level attribution framework that provides transparency not only for final answers but also for intermediate reasoning steps.

3.1 Answer-Level Attribution

Basic chart attribution involves visually linking chart elements to answers using bounding boxes, highlighting the specific data points that support the answer. In the leftmost chart of Figure 2, while the bounding boxes highlight all datapoints contributing to the answer "5478", the reasoning behind this calculation remains unclear without additional

context. This demonstrates why incorporating reasoning steps becomes crucial for questions involving mathematical operations, where the path to the answer is as important as the answer itself.

3.2 Reasoning-Level Attribution

When solving mathematical questions with charts, the path to the answer often involves multiple reasoning steps. Our framework attributes each reasoning step to relevant chart regions, creating a traceable connection between the reasoning process and visual elements. As shown in Figure 2, the 2nd, 3rd and the 4th chart represents each of the reasoning steps. This granular-level attribution approach enhances trust in the system by making both final answers and the reasoning process transparent and verifiable against the source chart.

Additional examples for both answer level and reasoning level attribution is present in Appendix section A.

4 Dataset Curation

Currently, no datasets exist that provide reasoning steps for chart question answering or attribution annotations for mathematical chart QA. To address this gap, we first examine existing model capabilities before developing a semi-automatic annotation strategy.

4.1 Reasoning Capabilities of MLLMs for Charts

Based on performance evaluations in the Polymath benchmark (Gupta et al., 2024), we select Claude 3.5 Sonnet and GPT-4o as our primary models for analysis. We also include GPT-4v for its vision capabilities. To assess the performance on reasoning generation, we randomly select 100 examples from the ChartQA dataset (Masry et al., 2022), each containing a chart, question, and answer triple. Charts and questions are passed to these models as input and they are prompted to generate answers and reasoning. These answers and reasoning are annotated by human annotators on whether they are correct or not, and the results are presented in Table 1.

Model	Answer is Correct (Human Annotated)	Reasoning is Correct (Human Annotated)
Gpt-4o	58%	49%
Gpt-4v	64%	45%
Claude-3.5-sonnet	96%	75%

Table 1: Benchmarking performance based on Visual Question Answering and Visual Question Reasoning.

As shown in Table 1, while Claude 3.5 Sonnet demonstrates strong answer generation (96% accuracy), its reasoning capabilities show significant room for improvement (75% accuracy). Other models perform notably worse, with reasoning accuracies below 50%.

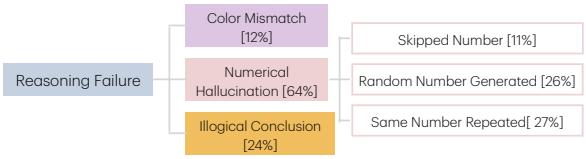


Figure 3: Taxonomy of Failure Cases that represents the categories of reasoning failure.

A more detailed error analysis, including word clouds of failure patterns in appendix section B figure 12 and specific examples, is provided in the appendix section B figure 13. Notably, providing correct answers alongside questions reduced reasoning failures from 51% to 25%, suggesting the potential for improved performance through better model guidance.

4.2 Task Setup

Our goal is to obtain three types of annotations for chart-based mathematical attribution: (1) reasoning steps for given chart-question-answer triples, (2) answer attribution, and (3) reasoning step attribution. Rather than annotating from scratch, we developed a semi-automatic approach leveraging Claude 3.5 Sonnet's capabilities to generate initial annotations for human correction.

We recruited two qualified annotators through the Upwork platform[1] after an initial screening of three candidates using 100 sample data points. The entire annotation process went for 120 hours and each annotator was paid 15 USD hourly.

For attribution annotation, we employed the VGG Image Annotator platform[2], which provides an intuitive interface for drawing bounding boxes and mapping them to textual reasoning steps. Screenshots from the annotation interface, more details on initial screening and examples of such annotations are provided in the appendix section A fig 6 and 10.

Stage 1: Reasoning Validation and Correction. In Stage 1, annotators perform reasoning validation through three key steps: (1) correction

[1] https://www.upwork.com
[2] https://annotate.officialstatistics.org/

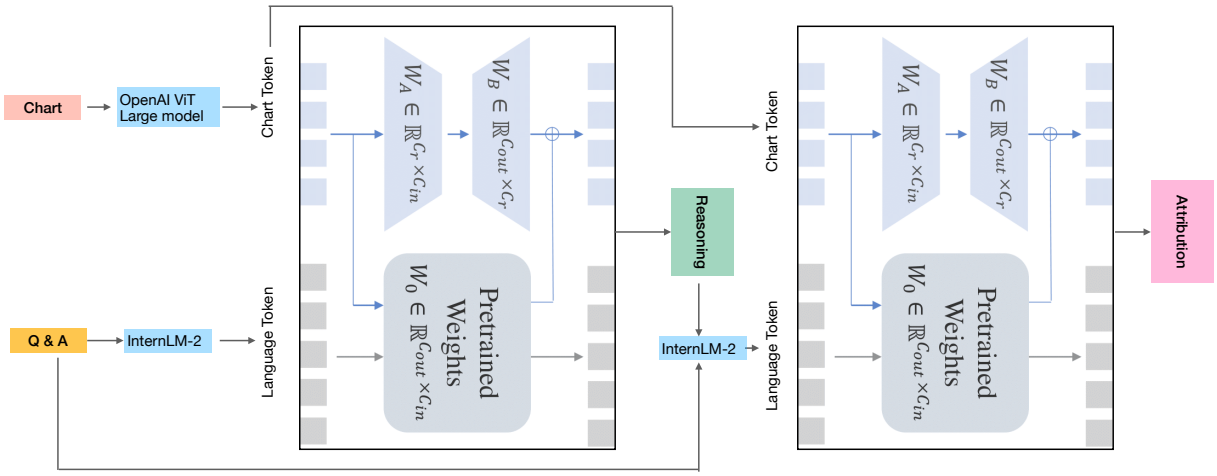


Figure 4: Overview of our proposed ATLAS framework for reasoning step generation. The architecture leverages InternLM-XComposer2 with Partial-LoRA for visual token adaptation. Given a chart and Q&A pair, the model processes visual tokens through CLIP ViT-Large and applies Partial-LoRA for chart-specific feature adaptation, while textual inputs are processed by InternLM-2. The final output provides reasoning steps through hidden state analysis.

by reviewing Claude-3.5-sonnet generated reasoning for chart-question-answer triples, (2) providing a binary correctness assessment (Yes/No) for each triple, and (3) categorizing errors in incorrect reasoning (such as color mismatches or illogical conclusions) while supplying corrected reasoning when the original is found to be inaccurate.

Stage 2: Answer Attribution. For each chart-question-answer triple, annotators draw bounding boxes using the VGG image annotator indicating chart regions supporting the answer. Figure 6 demonstrates this process, showing both input and resulting annotations.

Stage 3: Reasoning Attribution. Using the validated reasoning from Stage 1, annotators are instructed to provide bounding boxes using the VGG image annotator for each reasoning step. As shown in Figure 10, each statement (e.g., "orange line represents unfavorable") is linked to relevant chart regions.

4.3 Data Annotation & Analysis

To ensure annotation quality, we conducted initial screening to select mathematically proficient annotators, measured inter-annotator agreement using Kappa score (Cohen, 1960), and had authors manually verify a sample of annotations. This semi-automatic approach significantly reduced annotation effort while maintaining high quality through human validation and correction. More details on the InterAnnotator Agreement consisting of Kappa score and Intersection Over Union (IOU) score (Rezatofighi et al., 2019) calculation formula are present in the Appendix section B.1.

Chart Type	Stage 1 [Kappa Score]	Stage 2 [IOU Score]	Stage 3 [IOU Score]
Line	0.825	0.524	0.561
Bar	0.920	0.579	0.647

Table 2: Annotation scores across three stages for line and bar charts. Stage 1 shows high inter-annotator agreement (Kappa score > 0.8) for reasoning validation. Stage 2 demonstrates moderate agreement in answer-based attribution (IOU scores 0.5), while Stage 3 shows agreement (0.56 < IOU scores < 0.64) when incorporating reasoning-based attribution. This progression suggests that reasoning steps help annotators more consistently identify relevant chart regions. Also, this indicates the complexity of annotations for Stage 2 and Stage 3.

4.4 Data Analysis

After the annotation process, the key statistics about the data are summarized in the table 3. Table 3 shows the breakdown of the dataset by chart type. There are a total of 1000 charts, consisting of 500 line charts and 500 bar charts. For each chart, there are 2 QA pairs, resulting in a total of 2000 QA pairs. Additionally, the annotators identified a total of 3599 reasoning steps across all the charts (stage 1). The table also shows the number of image regions that were attributed to the QA-based annotations (stage 2) and the reasoning-based annotations (stage 3). For line charts, there are 1465 QA-based and 2691 reasoning-based attributed regions, while

for bar charts, there are 2627 QA-based and 4437 reasoning-based attributed regions.

Chart Type	No. of Charts	No. of QA pairs	No. of reasoning steps	No. of attributed image regions (QA-based)	No. of attributed image regions (reasoning-based)
Line	500	1000	1773	1465	2691
Bar	500	1000	1826	2627	4437
Total	**1000**	**2000**	**3599**	**4092**	**7128**

Table 3: Summary of the ATLAS Dataset. Our dataset contains an equal distribution of line and bar charts (500 each), with 2 QA pairs per chart. The table shows the progression from basic QA pairs through reasoning steps to attributed regions, with reasoning-based attribution requiring significantly more regions (7128) compared to QA-based attribution (4092).

5 ATLAS: Proposed Method

Our proposed method, ATLAS, addresses the challenge of attributing mathematical reasoning in charts through a two-stage pipeline (figure 4). Given a chart-question-answer triple, we first generate step-by-step reasoning using InternLM-XComposer2 model (figure 5), then leverage these reasoning steps along with the chart, question, and answer to produce attribution bounding boxes for both the final answer and intermediate reasoning steps. For reasoning generation, we utilize InternLM-XComposer2[3] output. The model architecture incorporates a vision encoder (CLIP ViT-Large (Radford et al., 2021)) that processes charts into a 35×35 grid (1225 visual tokens) and maps them to a shared 4096-dimensional embedding space with text from InternLM-2 (Cai et al., 2024). Using (Dong et al., 2024), we employ Partial LoRA (Hu et al., 2021) (PLoRA), which applies additional trainable parameters specifically to visual tokens while preserving the base 7B-parameter language model's capabilities.

We utilize (Phukan et al., 2024a)'s findings on attribution, and extract hidden states from layer 16, which empirically provides optimal semantic representations for our task. The attribution mechanism employs a GPU-accelerated sliding window approach, efficiently processing window configurations from 3×3 to 35×35 patches through normalized patch embedding averaging and cosine similarity metrics between textual descriptions and visual regions.

[3] https://github.com/InternLM/InternLM-XComposer

6 Experiments

We conduct experiments on the curated dataset presented in section 4.4. We experiment on this dataset for two tasks i.e. (i) Attribution based on Visual Question Answering (VQA) and (ii) Attribution based on Visual Question Reasoning (VQR).

6.1 Baselines

We evaluate ATLAS against three state-of-the-art MLLMs: GPT-4o, GPT-4v, and Claude 3.5 Sonnet. For each baseline, we test both zero-shot and few-shot prompting strategies for two tasks: attribution based on answer attribution and reasoning attribution.

Answer Attribution (VQA): Models must identify relevant chart regions using bounding boxes that support their answers to specific questions.

Reasoning Attribution (VQR): Models must attribute their mathematical reasoning steps using bounding boxes to specific chart elements. Each reasoning step has different granular attribution as described in VQR step 1, 2 and 3 of fig 2.

We collect attribution results through API calls to GPT-4o[4], GPT-4v[5], and Claude 3.5 Sonnet[6]. Since these models cannot directly output chart images with bounding boxes, we design prompts to obtain coordinates of the bounding boxes in the format of X1, Y1, X2, and Y2.

Zero Shot For answer attribution, we used the following zero shot prompt:

Zero-Shot Prompt

System Prompt: You are a helpful assistant that responds in markdown. Help me with my math question.
Input Format:

- Chart: [chart_image], Question: [question_text], Answer: [answer_text]

User Prompt: Given this chart and the question-answer pair: question = "question", answer = "answer"; ONLY generate bounding box coordinates in X1, Y1, X2, Y2 format - A list of tuples, each containing (x1, y1, x2, y2) representing the bounding box coordinates without additional text which represents which part of the chart corresponds to the answer.

[4] GPT-4o: "GPT-4o", "2023-05-15"
[5] GPT-4v: "gpt-4-vision-preview", "2023-07-01-preview"
[6] Claude 3.5 Sonnet: "claude-3-5-sonnet-20240620-v1:0"

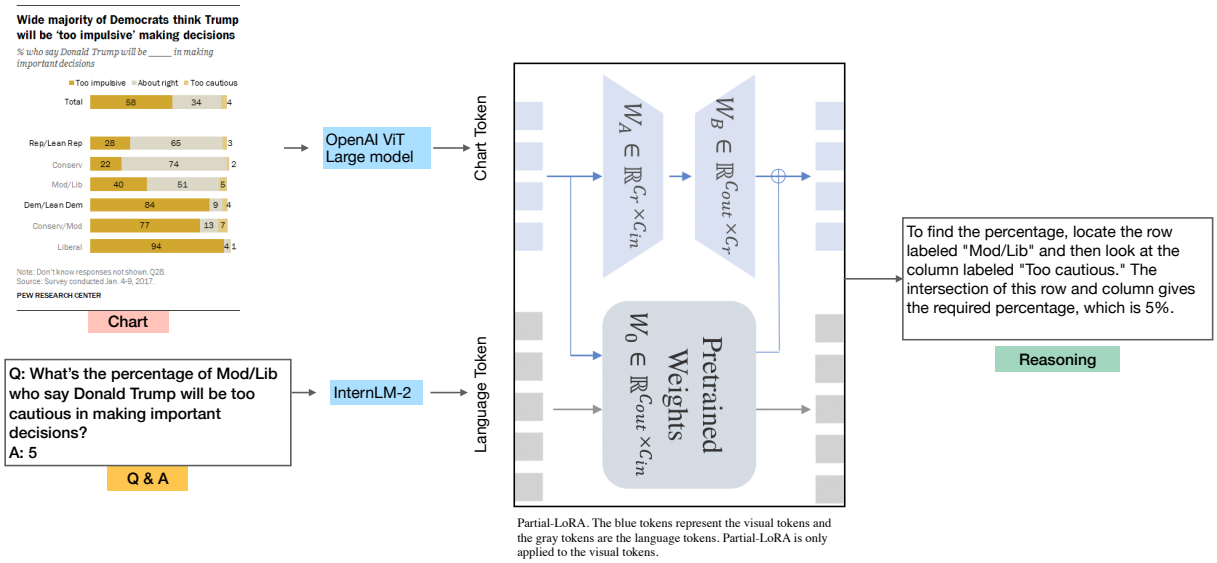


Figure 5: Overview of our proposed ATLAS framework. The architecture leverages InternLM-XComposer2 with Partial-LoRA for visual token adaptation. Given a chart and Q&A pair, the model processes visual tokens through CLIP ViT-Large and applies Partial-LoRA for chart-specific feature adaptation, while textual inputs are processed by InternLM-2. The final output provides both answer and reasoning-based attributions through hidden state analysis.

Few Shot Using Few-shot prompting for answer attribution, an example is passed in the prompt, and the prompt is described below:

Few-Shot Prompt

System Prompt: You are a helpful assistant that responds in markdown. Help me with my math question.
Example 1:

Chart: [bar_chart_image] Question: "What was the highest value in 2020?"
Answer: "85 units"
Bounding Box: (120, 45, 140, 230)

User Prompt: Given this chart and the question-answer pair: question = "question", answer = "answer" and examples; ONLY generate bounding box coordinates in X1, Y1, X2, Y2 format - A list of tuples, each containing (x1, y1, x2, y2) representing the bounding box coordinates without additional text which represents which part of the chart corresponds to the answer.

Appendix section C contains additional details.

For **Reasoning Attribution (VQR):** Few-shot and zero-shot prompting are conducted and the details of the prompt are present in Appendix section C figure 16 and 17 for zero-shot and few-shot respectively. The input charts are encoded before being passed as prompts and are decoded using base64 encoding/decoding after collecting the bounding box coordinates for further analysis.

6.2 ATLAS-Automatic Reasoning Step Generation

For automatic reasoning step generation, we leverage the Partial LoRA framework from (Dong et al., 2024) due to its effectiveness in preserving language capabilities while adapting to visual inputs. Given a chart-question-answer triple (C, Q, A), our goal is to generate reasoning steps R that explain the answer derivation while maintaining alignment with visual elements.

Following the existing Partial LoRA architecture, we process inputs $x = [x_v, x_t]$, where x_v represents visual tokens from the chart processed through CLIP ViT-Large, and x_t represents the concatenated question-answer tokens. The output features are computed as follows:

$$\hat{x} = [\hat{x}_v, \hat{x}_t] \quad (1)$$

where $\hat{x}_t$ follows the standard language model path, and $\hat{x}_v$ incorporates visual adaptation through the Partial LoRA matrices.

Reasoning generator is used to maximize:

$$P(R|C, Q, A) = \prod_{i=1}^{n} P(r_i|r_{<i}, C, Q, A) \quad (2)$$

where r_i represents the i-th token in the reasoning sequence.

This approach enables our model to generate step-by-step reasoning by utilizing chart-specific visual features while maintaining strong language capabilities, producing coherent explanations that explicitly reference chart elements, and describing the mathematical operations needed to arrive at the answer. The generated reasoning provides a transparent explanation of the answer derivation process, which is then used to guide our attribution mechanism for identifying relevant chart regions.

6.3 Metrics

We employ IOU score (Rezatofighi et al., 2019) as our primary evaluation metric:

$$IOU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \quad (3)$$

where B_p represents the predicted bounding box and B_{gt} represents the ground truth box. Additionally, we measure the cosine similarity between textual descriptions and visual regions to evaluate semantic alignment.

7 Results and Discussion

Table 4 presents the IOU scores across different models and prompting strategies. Our ATLAS framework shows substantial improvements across all metrics compared to baseline models.

Using the ATLAS method, we achieve improvements of 504% ($0.026 \rightarrow 0.157$) and 446% ($0.028 \rightarrow 0.153$) for VQA tasks (line and bar charts, respectively). With automated reasoning generated using fig 5, we report a 230% ($0.122 \rightarrow 0.037$) and 110% ($0.039 \rightarrow 0.082$) improvement for VQR tasks. These improvements further increase when using human-validated reasoning, particularly for VQR tasks, reaching 268% ($0.037 \rightarrow 0.136$) for line charts and 405% ($0.197 \rightarrow 0.039$) for bar charts, while maintaining the same VQA performance since in the VQA task, only question and answer is given as input. This demonstrates the value of both our automated approach and the potential for further improvements with human validation.

We note that ATLAS with automated reasoning is not very far from that with human reasoning, indicating that our reasoning generation using InternLM is fairly good, close to human-provided or corrected reasoning steps. However, there is still a very long way to go for attribution for charts, given that even the best-performing, human-based reasoning variant of ATLAS leads to $0.15 - 0.2$ IOU scores. One possible future direction could be to come up with better attribution bounding box prediction systems. A promising direction would be fine-tuning an MLLM to generate these attribution bounding boxes.

Model	VQA IOU		VQR IOU	
	Line	Bar	Line	Bar
GPT-4o (zero-shot)	0.026	0.028	0.025	0.021
GPT-4o (few-shot)	0.020	0.022	0.022	0.019
GPT-4v (zero-shot)	0.016	0.019	0.021	0.023
GPT-4v (few-shot)	0.014	0.017	0.022	0.024
Claude 3.5 (zero-shot)	0.024	0.025	0.032	0.035
Claude 3.5 (few-shot)	0.025	0.021	0.037	0.039
ATLAS (Automated Reasoning)	**0.157**	**0.153**	**0.122**	**0.082**
ATLAS (Human Reasoning)	**0.157**	**0.153**	**0.136**	**0.197**

Table 4: Attribution performance comparison across different models and settings. Performance is measured using IOU scores for both VQA and VQR tasks on line and bar charts. Baseline models (GPT-4o, GPT-4v, Claude 3.5) show limited performance in both zero and few-shot settings (<0.04 IOU). Our ATLAS framework demonstrates substantial improvements, achieving IOU scores >0.15 for VQA tasks and up to 0.197 for VQR tasks when using human-validated reasoning, highlighting the benefits of incorporating reasoning steps in the attribution process.

8 Conclusion

In this paper, we presented a novel framework for chart attribution that combines visual and mathematical reasoning capabilities. Our primary contributions include (i) the formalization of chart attribution for mathematical question answering and reasoning tasks, (ii) a systematic data curation strategy that combines MLLM-generated reasoning with human corrections for reliable attribution annotation, and (iii) a framework for using InternLM-XComposer2 model that utilizes automatic reasoning steps to improve attribution accuracy. While our approach demonstrates significant improvements over baselines, opportunities remain for enhancing reasoning generation, extending support for complex chart types, and integrating with downstream applications. Our framework provides a foundation for building more trustworthy and interpretable AI systems for mathematical reasoning tasks, paving the way for chart-based systems that can better explain their decision-making processes.

9 Limitations

While our framework demonstrates promising results for chart attribution through automated reasoning, several important limitations and areas for discussion emerge from our study:

Restrictive Prompting: To produce bounding boxes, we use a highly restrictive prompting approach. We instruct the model to generate bounding boxes as a list of coordinate tuples in the format (X1, Y1, X2, Y2). However, research works on restrictive prompting (Tam et al., 2024) has found that using overly restrictive prompts can lead to notable decreases in model performance compared to less constrained prompting techniques. Therefore

Attribution Task is Challenging for Humans: As shown in Table 2, the attribution task proved difficult for human annotators. In stages 2 and 3 of the annotation process, the agreement percentages ranged from just 52% to 64%. These relatively low levels of agreement underscore the inherent challenge of the attribution task, even for human raters with domain expertise.

Reasoning Quality Dependencies: Our attribution system's performance depends on the quality of generated reasoning steps. While fine-tuning InternLM-XComposer2 may improve reasoning generation, complex mathematical operations, and multi-step calculations still present challenges, potentially affecting attribution accuracy. We discuss failure cases for reasoning in fig 12.

Chart Type Constraints: The current implementation focuses primarily on line and bar charts, limiting its applicability to other visualization types. Complex charts with multiple axes, overlapping elements, or nested visualizations may pose additional challenges for both reasoning generation and attribution.

Computational Requirements: The sliding window mechanism used for attribution, while effective, requires significant computational resources, especially for high-resolution charts or when processing multiple reasoning steps. This may impact the system's practicality in real-time applications.

Human Validation Process: While our data curation strategy employs human validation to ensure quality, the subjectivity in reasoning annotation and attribution marking can introduce inconsistencies. The inter-annotator agreement scores suggest room for improvement in standardizing the validation process.

Model Architecture Limitations: The current approach relies on layer 16 hidden states of InternLM-XComposer2, which may not capture all relevant features for attribution. Alternative architectural choices or multi-layer approaches could potentially yield better results.

These limitations point to several promising directions for future research, including more robust reasoning generation mechanisms, efficient attribution algorithms, and improved validation methodologies.

10 Ethics Statement

We acknowledge several ethical considerations in our development of chart attribution systems. First, we prioritized transparency by openly documenting our methodology, model limitations, and potential biases in both reasoning generation and attribution accuracy. All training data was properly sourced from public datasets with appropriate licensing, and our human annotation process followed fair labor practices, including equitable compensation ($15/hour) and clear guidelines. While our system aims to improve accessibility and understanding of quantitative information through transparent reasoning steps, we recognize potential risks of misuse, such as automated generation of misleading chart interpretations. We recommend deploying this technology with appropriate human oversight in high-stakes scenarios and maintaining regular audits for systematic biases. Our goal is to advance chart interpretation capabilities while implementing safeguards that protect against potential misuse and ensure the technology serves its intended purpose of making quantitative information more accessible and understandable to diverse user groups.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev,

Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: Large language model as a unified interface for vision-language multitask learning. *arXiv preprint arXiv:2310.09478*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. 2024. Polymath: A challenging multi-modal mathematical reasoning benchmark.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

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

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Arxiv 2401.04088*.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1917–1925.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2023. GPT-4V(ision) system card.

Anirudh Phukan, Divyansh, Harshit Kumar Morj, Vaishnavi, Apoorv Saxena, and Koustava Goswami. 2024a. Beyond logit lens: Contextual embeddings for robust hallucination detection grounding in vlms.

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024b. Peering into the mind of language models: An approach for attribution in contextual question

answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11481–11495, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. *arXiv preprint arXiv:2406.13663*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. Can llms master math? investigating large language models on math stack exchange. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2316–2320.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476.

Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, and Dan Roth. 2024. Evaluating llms' mathematical reasoning in financial document question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3853–3878.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 845–854.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

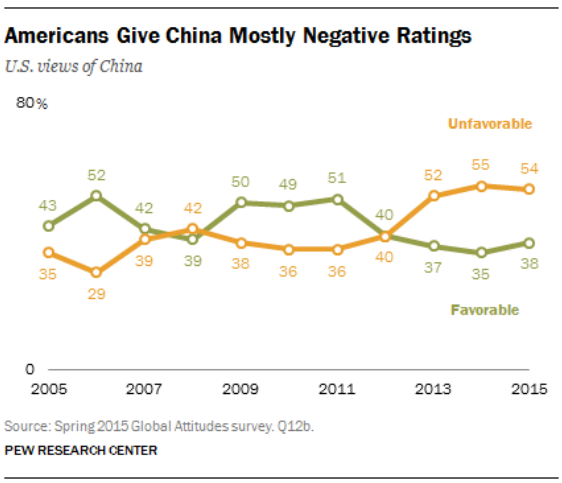
Appendix

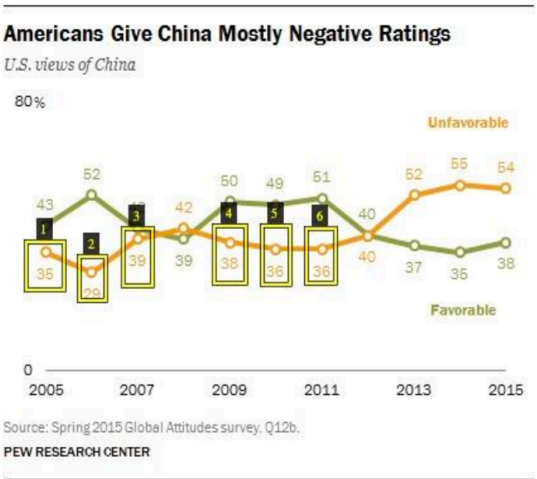
This section provides additional examples to assist in the understanding and interpretation of the research work presented.

- Section A: Attribution Examples
- Section B: Dataset Curation
- Section C: Experiments

A Attribution Definition & Examples

Examples of attribution based on visual question answering is present in fig 6 and 7.





ATTRIBUTION BASED ON VISUAL QUESTION ANSWERING
Question : How many value is below 40 in unfavorable graph? **Answer** : 6
Explantion: The chart on the right side shows Human-annotated Attribution based on VQA. Each datapoint contributing to the answer is represented through a distinct bounding box, with a total of six boxes displayed on the visualization.

Figure 6: This figure shows attribution based on question answering. Here the bounding boxes clearly identify six data points on the "Unfavorable" line that fall below 40%, directly supporting the answer to the question "How many values are below 40 in the Unfavorable graph?

For Visual Question Reasoning, examples are present in fig 8, 9 and fig 10.

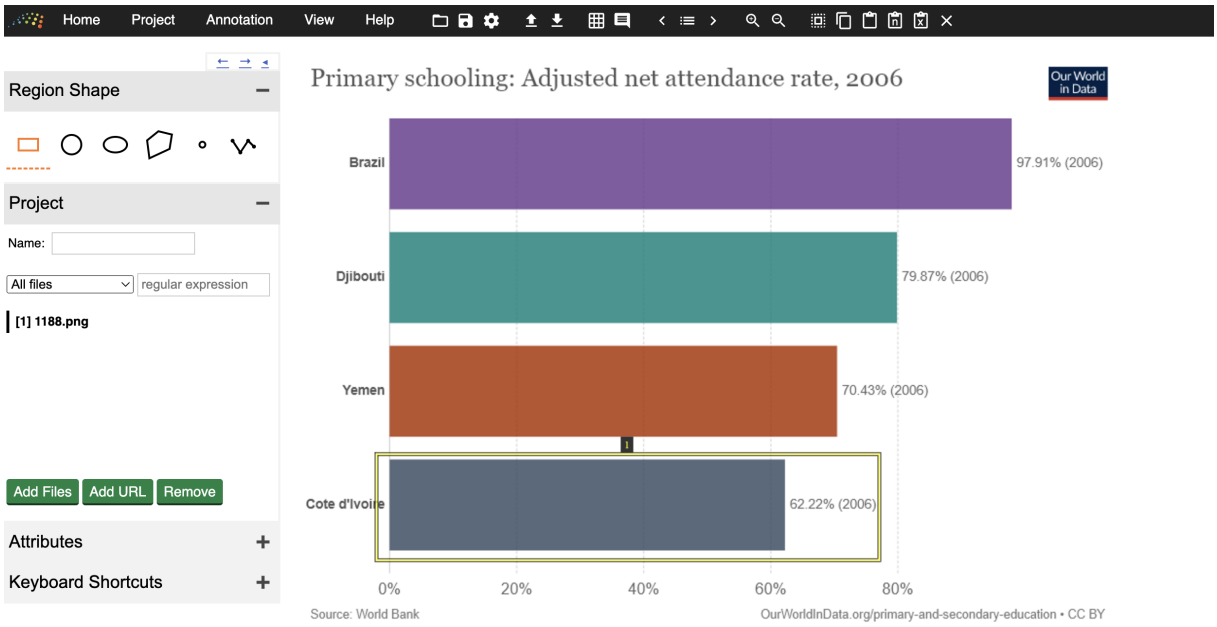


Figure 7: This figure shows attribution annotation platform based on question answering. Annotators are provided question = "What's the value of smallest bar?" and answer="62.22%". The annotators draw bounding box based on these as represented in the figure.

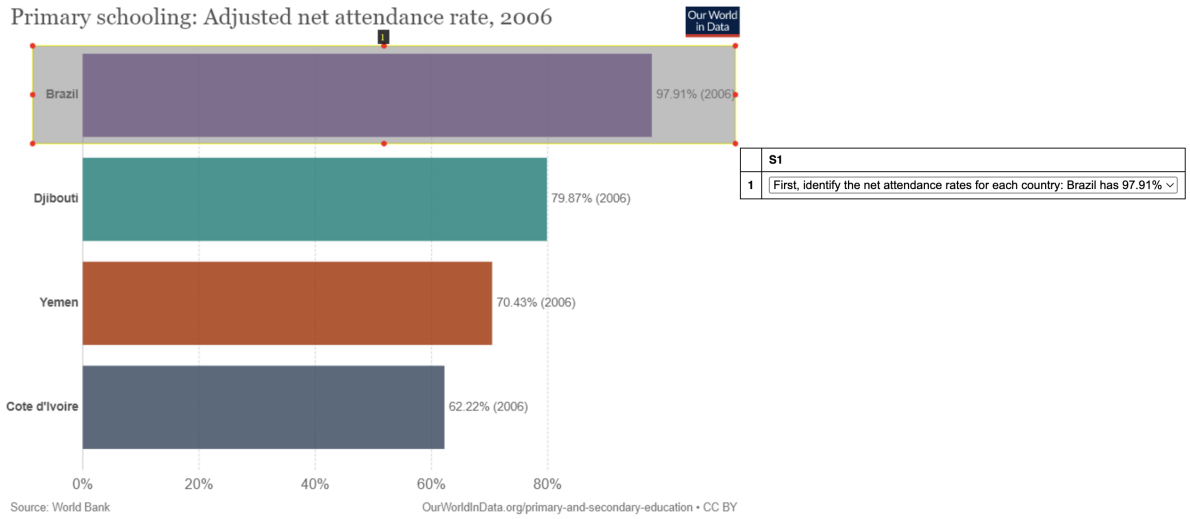


Figure 8: This figure shows how sentence-level reasoning is attributed in our dataset. Annotators are provided with a chart, question = What's the average of Yemen and Brazil?, answer= 84.17, and reasoning = "First, identify the net attendance rates for each country: Brazil has 97.91%. Yemen has 70.43%. Next, sum these values: 97.91 + 70.43 = 168.34. Then, divide by the number of countries to find the average: 168.34 / 2 = 84.17". The first reasoning statement "First, identify the net attendance rates for each country: Brazil has 97.91%." is directly linked to corresponding chart elements, ensuring each step of the mathematical reasoning process is grounded in the chart's components.

B Data Sources and Compilation

B.1 Data Annotation

To ensure annotation quality, we conducted initial screening to select mathematically proficient annotators, measured inter-annotator agreement using Kappa score (Cohen, 1960), and had authors manually verify a sample of annotations. This semi-automatic approach significantly reduced annotation effort while

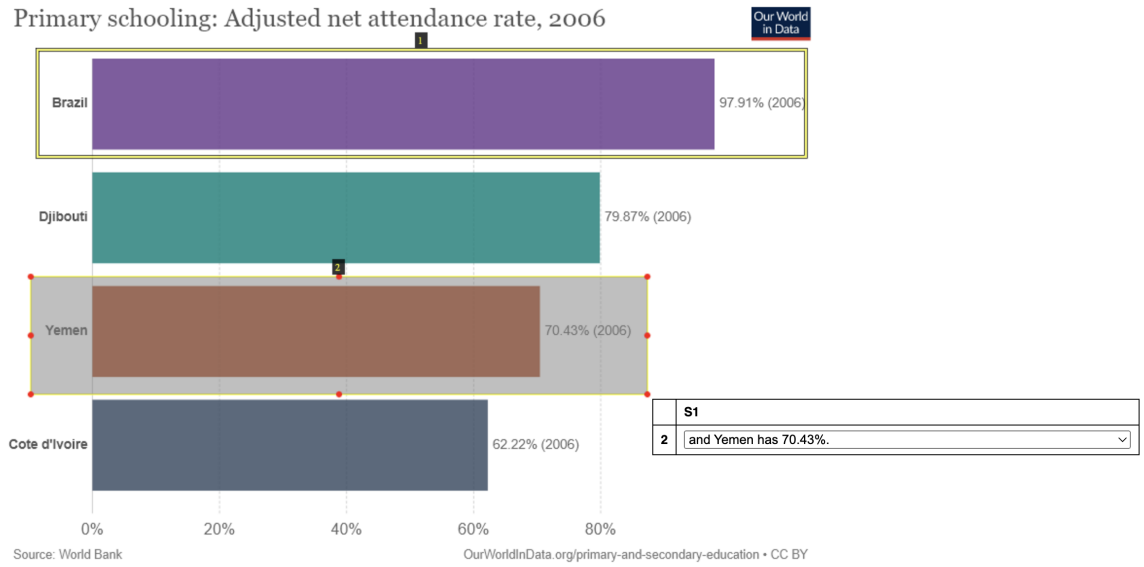
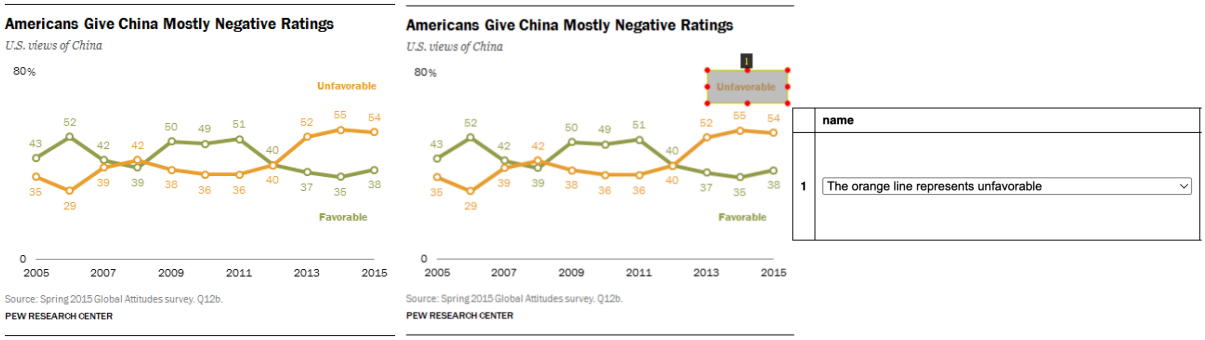


Figure 9: This figure shows how sentence-level reasoning is attributed in our dataset. Annotators are provided with a chart, question = What's the average of Yemen and Brazil?, answer= 84.17, and reasoning = "First, identify the net attendance rates for each country: Brazil has 97.91%. Yemen has 70.43%. Next, sum these values: 97.91 + 70.43 = 168.34. Then, divide by the number of countries to find the average: 168.34 / 2 = 84.17". The second reasoning statement "Yemen has 70.43%" is directly linked to corresponding chart elements, ensuring each step of the mathematical reasoning process is grounded in the chart's components.



ATTRIBUTION BASED ON VISUAL QUESTION REASONING
Question: What is the value of unfavorable in the year 2015? **Answer** : 54
Reasoning generated by Claude 3.5 Sonnet followed by correct annotation by human: The orange line represents unfavorable. For value that corresponds to the year 2015 in orange line is 54.
Explanation: The chart on the right side shows human annotation based on reasoning steps. Given each reasoning step, we attribute it on a sentence level. For example, the first sentence i.e. orange line represents unfavorable is attributed as displayed in the chart.

Figure 10: This figure shows how sentence-level reasoning is attributed in our dataset. The first reasoning statement "The orange line represents unfavorable" is directly linked to corresponding chart elements, ensuring each step of the mathematical reasoning process is grounded in the chart's components.

maintaining high quality through human validation and correction.

The agreement score for stage 1 is based on whether the reasoning is correct and is represented by Kappa score (Cohen, 1960). Kappa score is defined as a measure of inter-rater agreement for categorical items, taking into account the agreement occurring by chance. The Kappa score is defined mathematically as:

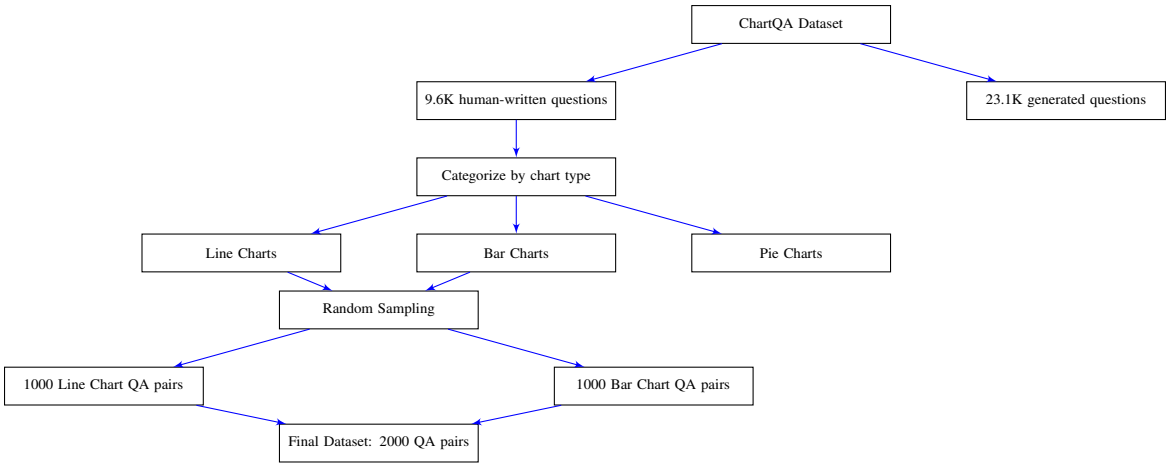


Figure 11: Data Compilation Process Flowchart. From the ChartQA dataset containing both human-written (9.6K) and generated (23.1K) questions, we focus on the human-written subset for quality assurance. We categorize these by chart type (line, bar, and pie charts), then use random sampling to create a balanced final dataset of 2000 QA pairs, comprising 1000 pairs each for line and bar charts.

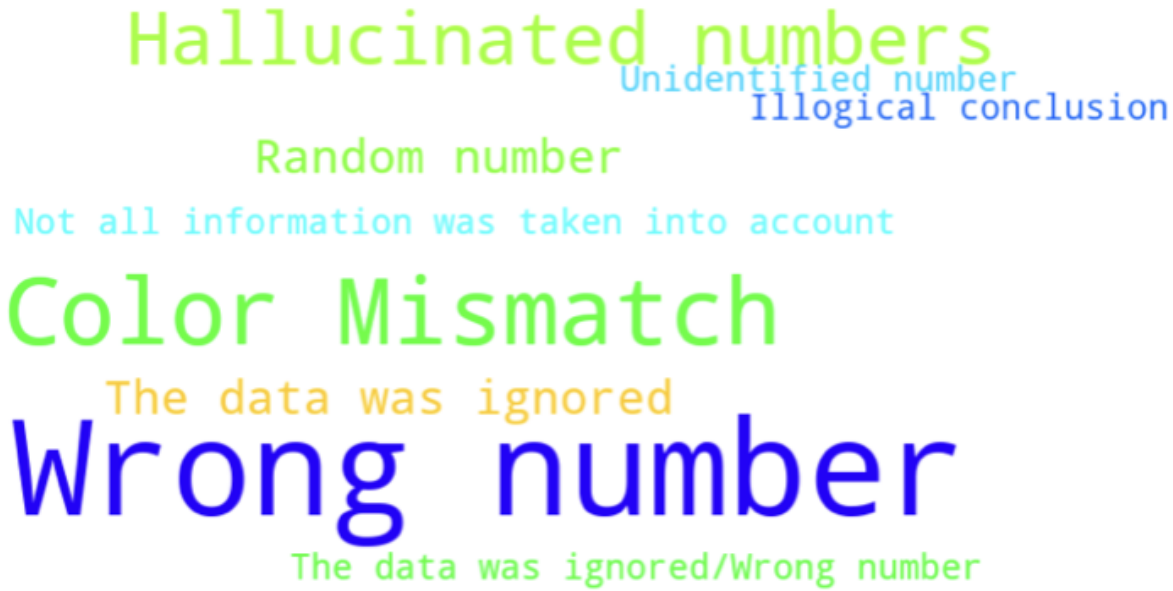


Figure 12: Word cloud representing the annotated labels for reasoning failure. These annotated labels are Hallucinated numbers, Illogical conclusions, Color mismatch, data points ignored, etc.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

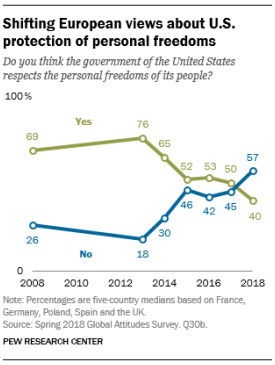
where:

p_o is the observed agreement between the two annotators p_e is the expected agreement by chance

The observed agreement p_o is calculated as:

$$p_o = \frac{a + d}{a + b + c + d}$$

where: a is the number of cases where both annotators agreed on "yes" b is the number of cases where the first annotator said "yes" and the second said "no" c is the number of cases where the first annotator said "no" and the second said "yes" d is the number of cases where both annotators agreed on "no"

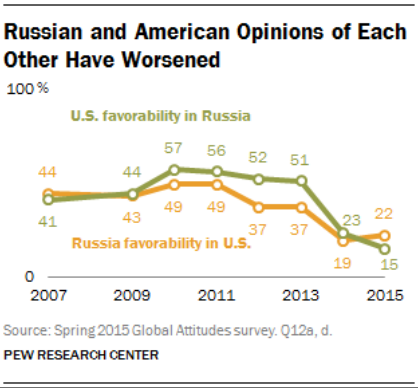


Reasoning Failure Category: Illogical Conclusion

Question: Is the difference between blue and green graph in 2008 the maxiumum among years?

Reasoning generated from GPT-4o: In 2008, the difference between the blue line (No) and the green line (Yes) is 43% (69% - 26%). Examining other years, the differences are: 58% in 2009, 48% in 2017, and 57% in 2018. The difference in 2008 is not the maximum among these years, with the maximum difference occurring in 2009.

Explanation: The years and numbers were incorrectly identified. An Illogical conclusion was drawn.



Reasoning Failure Category: Color Mismatch

Question: Is the sum of highest and lowest value of orange graph greater than 60?

Reasoning generated from GPT-4o: The highest value of the orange graph (Russia favorability in U.S.) is 49 in 2010, and the lowest value is 15 in 2015. Adding these two values gives 49 + 15, which equals 64. Since 64 is greater than 60, the answer is 'Yes'.

Explanation: Reasoning generated from GPT-4o for the orange graph contains an inaccuracy - it identified the lowest point as 15 in 2015 which is lowest for green graph; therefore its a color mismatch.

Figure 13: Reasoning failure case examples

The expected agreement p_e is calculated as:

$$p_e = \frac{(a+b)(a+c) + (c+d)(b+d)}{(a+b+c+d)^2}$$

For stage 2 and stage 3, the Intersection over Union (IOU) score (Rezatofighi et al., 2019) was calculated. IOU score is a measure of the overlap between the bounding box drawn by annotator 1 and bounding box drawn by annotator 2, defined as the ratio of the area of intersection to the area of union. The IOU score is mathematically defined as:

$$\text{IOU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$

where the area of intersection is the overlapping area between the predicted bounding box and the ground truth bounding box, and the area of union is the total area covered by both bounding boxes.

Let's denote the predicted bounding box as B_p and the ground truth bounding box as B_g. Then, the IOU score can be calculated as:

$$\text{IOU} = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}$$

Where $|B_p \cap B_g|$ is the area of the intersection and $|B_p \cup B_g|$ is the area of the union.

C Experiments

This section contains prompts and additional implementation details.

C.1 Computing Infrastructure Details

Our implementation uses PyTorch 2.0 and all experiments were conducted on 4 NVIDIA A100 GPUs with 80GB of memory each. The experiments were run on Amazon Elastic Compute Cloud (Amazon EC2) instances equipped with A100 Tensor Core GPUs and 400 Gbps networking capabilities. The complete experimental pipeline took approximately 100 hours.

C.2 Prompting Strategies for Attribution

We experimented with zero-shot and few-shot prompting strategies for both VQA-based and VQR-based attribution.

VQA based Attribution For VQA based Attribution, we used both zero shot and few shot prompting and the prompt is described in figure 14 and 15.

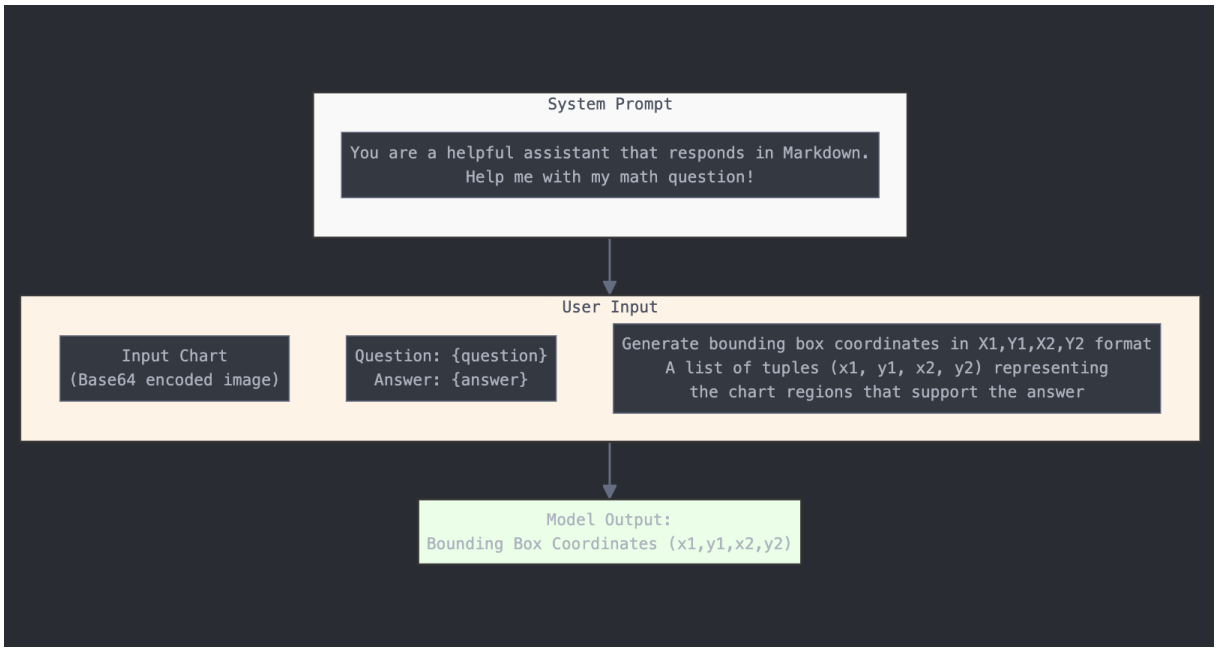


Figure 14: Zero Shot Prompting for Attribution based on VQA task.

Attribution based on VQR

For VQR based Attribution, we used both zero shot and few shot prompting and the prompt is described in figure 16 and 17.

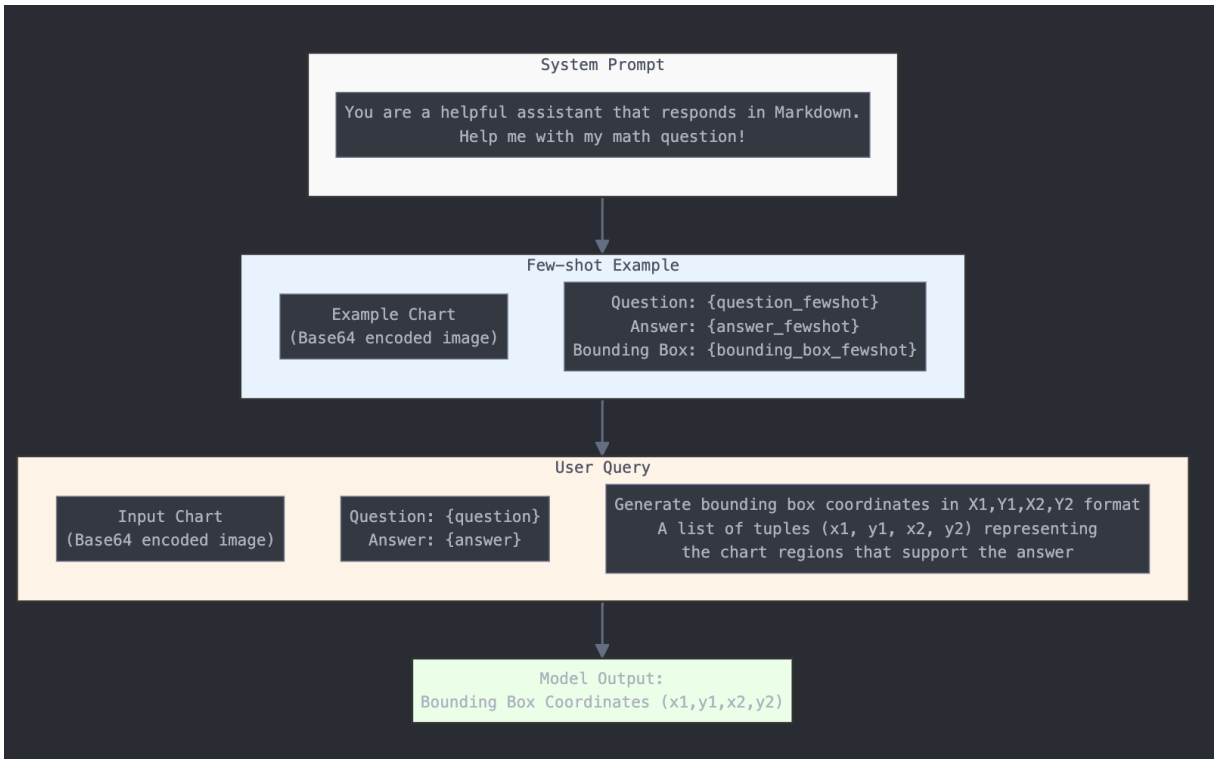


Figure 15: Few Shot Prompting for Attribution based on VQA task.

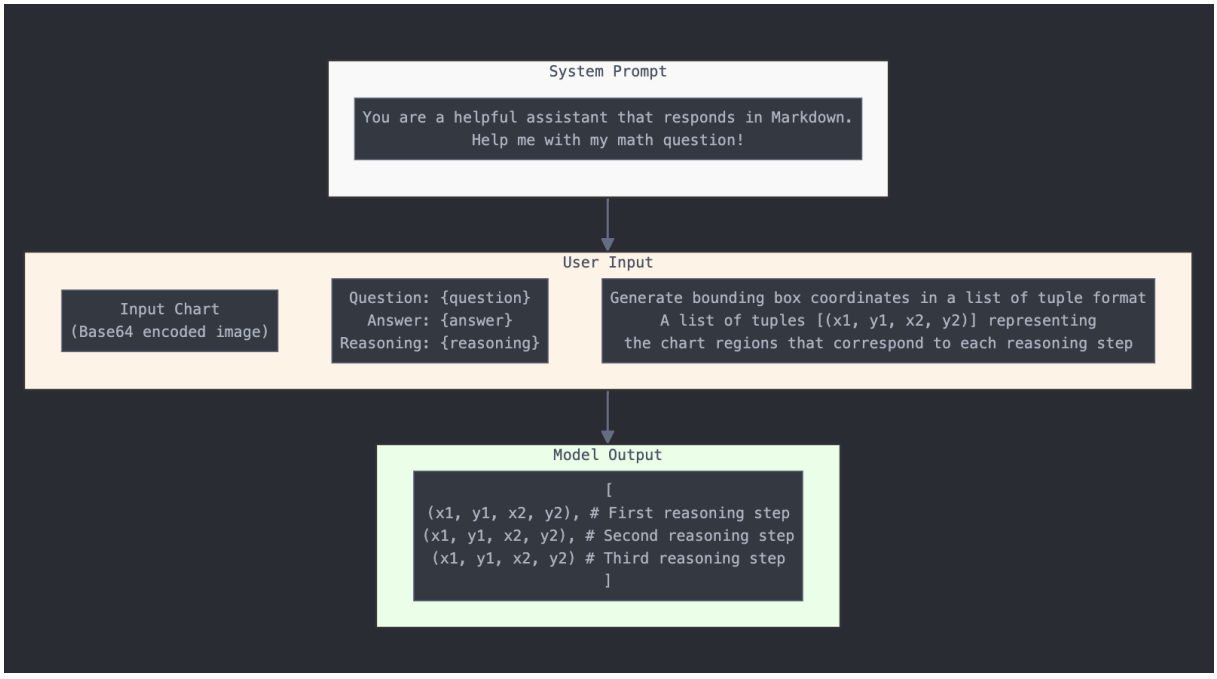


Figure 16: Zero Shot Prompting for Attribution based on VQR task.

C.3 Implementation Details - ATLAS

The proposed pipeline architecture for chart understanding consists of four integral stages that work in concert to process and analyze chart images with corresponding textual inputs.

In the first stage, Input Processing, the system handles three primary inputs: the chart image, which serves as the visual input for analysis; the question and answer prompt, which guides the analysis direction. These inputs undergo Base64 encoding for the image and are formatted into a specialized text prompt

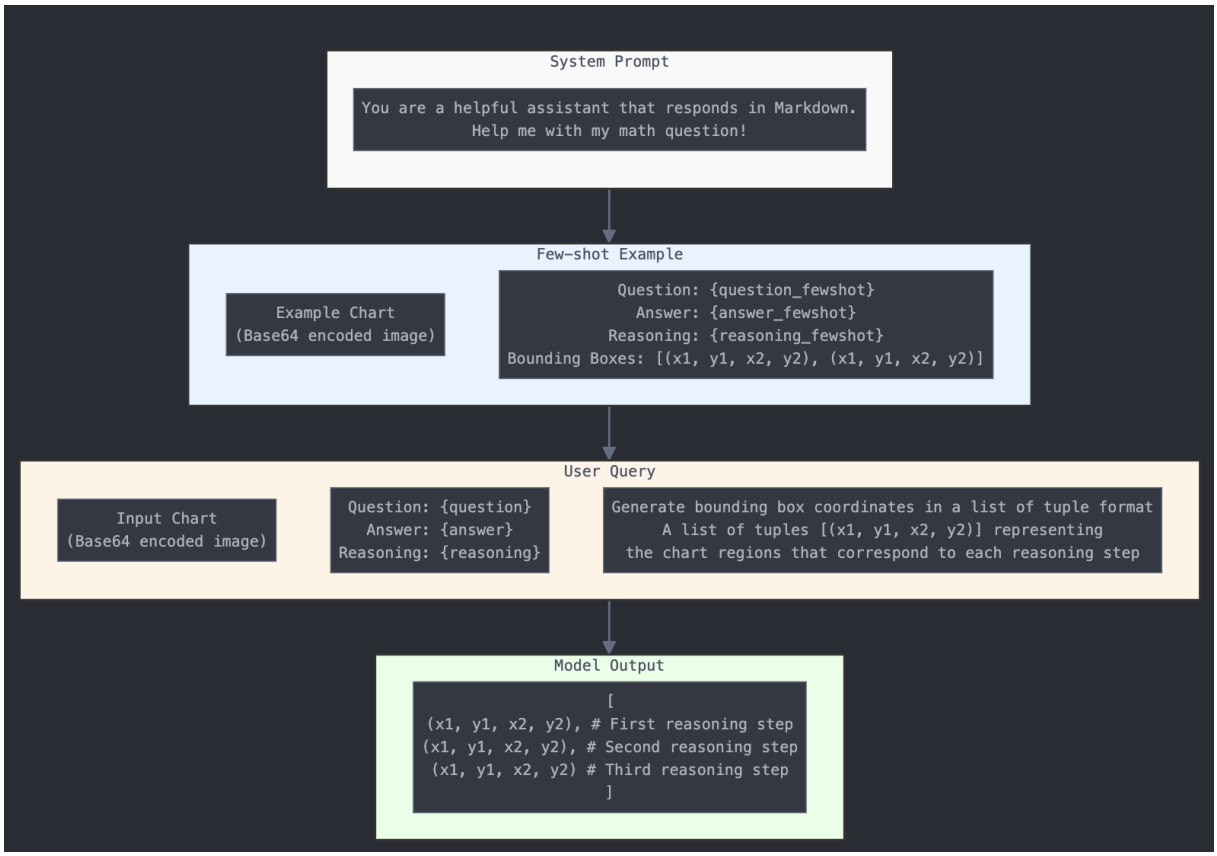


Figure 17: Few Shot Prompting for Attribution based on VQR task.

structure, resulting in encoded inputs suitable for model processing.

The second stage, MLLM Processing, leverages the InternLM-XComposer2 model's multimodal capabilities to process the encoded inputs. This stage extracts Layer 16 Hidden States, which contain rich semantic information from both modalities. The image features are processed as a 35×35 patch grid, while the text features are encoded into 4096-dimensional vectors, enabling comprehensive semantic representation of both visual and textual content. This dual-stream processing ensures that both modalities contribute effectively to the final analysis.

The third stage implements a Sliding Window Attribution mechanism, which is crucial for identifying relevant regions within the chart. This process begins with window generation, where variable-sized windows are created over the image feature space. The system then computes cosine similarity between the text and image features for each window, enabling the identification of regions most pertinent to the textual input. This stage culminates in the selection of the best region, outputting coordinates (i, j, h, w) that specify both the location and dimensions of the most relevant area within the chart.

The final stage focuses on Visualization, transforming the mathematical outputs into interpretable visual representations. This involves coordinate mapping, where the model's internal coordinate space is transformed into image pixel space, followed by bounding box generation that creates visible overlays highlighting the relevant regions identified by the model. This visualization stage is crucial for making the model's decisions interpretable and useful for end users.

The entire pipeline demonstrates flexibility in handling both reasoning-based and answer-based attribution scenarios through the same architectural framework. This unified approach allows for consistent processing while accommodating different types of chart analysis tasks, from simple identification to complex reasoning about chart elements. The system maintains a consistent flow of information through each stage, ensuring that the final output effectively bridges the gap between the visual elements of the chart and the textual understanding required for comprehensive chart analysis.

Figure 18 and 19 represents VQA based and VQR based attribution details respectively.

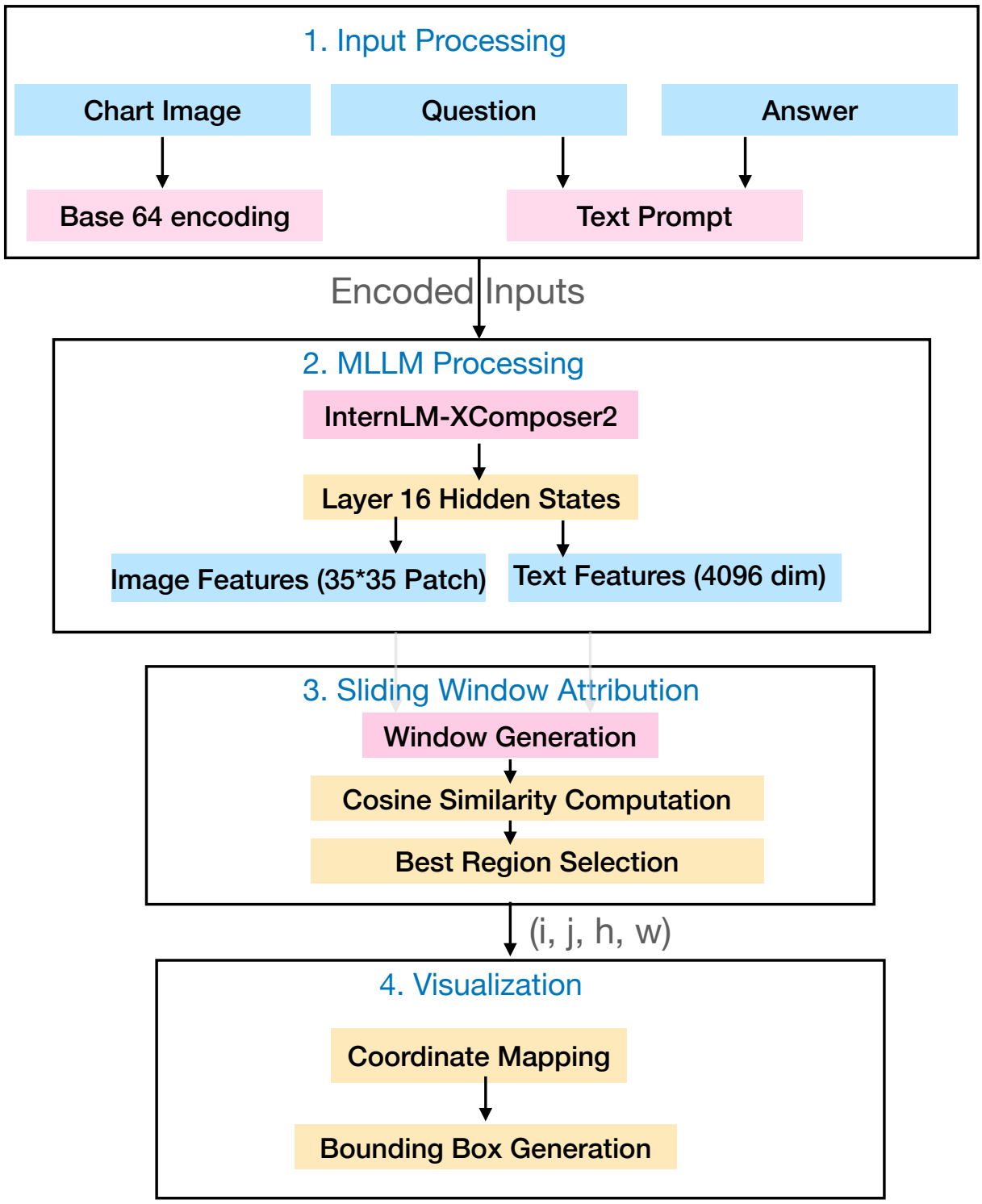


Figure 18: The pipeline architecture for chart understanding with InternLM-XComposer2 illustrates a four-stage process that bridges visual and textual modalities in chart analysis. The system progresses through Input Processing (encoding of chart images and text), MLLM Processing (multimodal feature extraction), Sliding Window Attribution (region identification), and Visualization (interpretable output generation), enabling comprehensive chart understanding through a unified architectural framework. This architecture supports answer-based attribution.

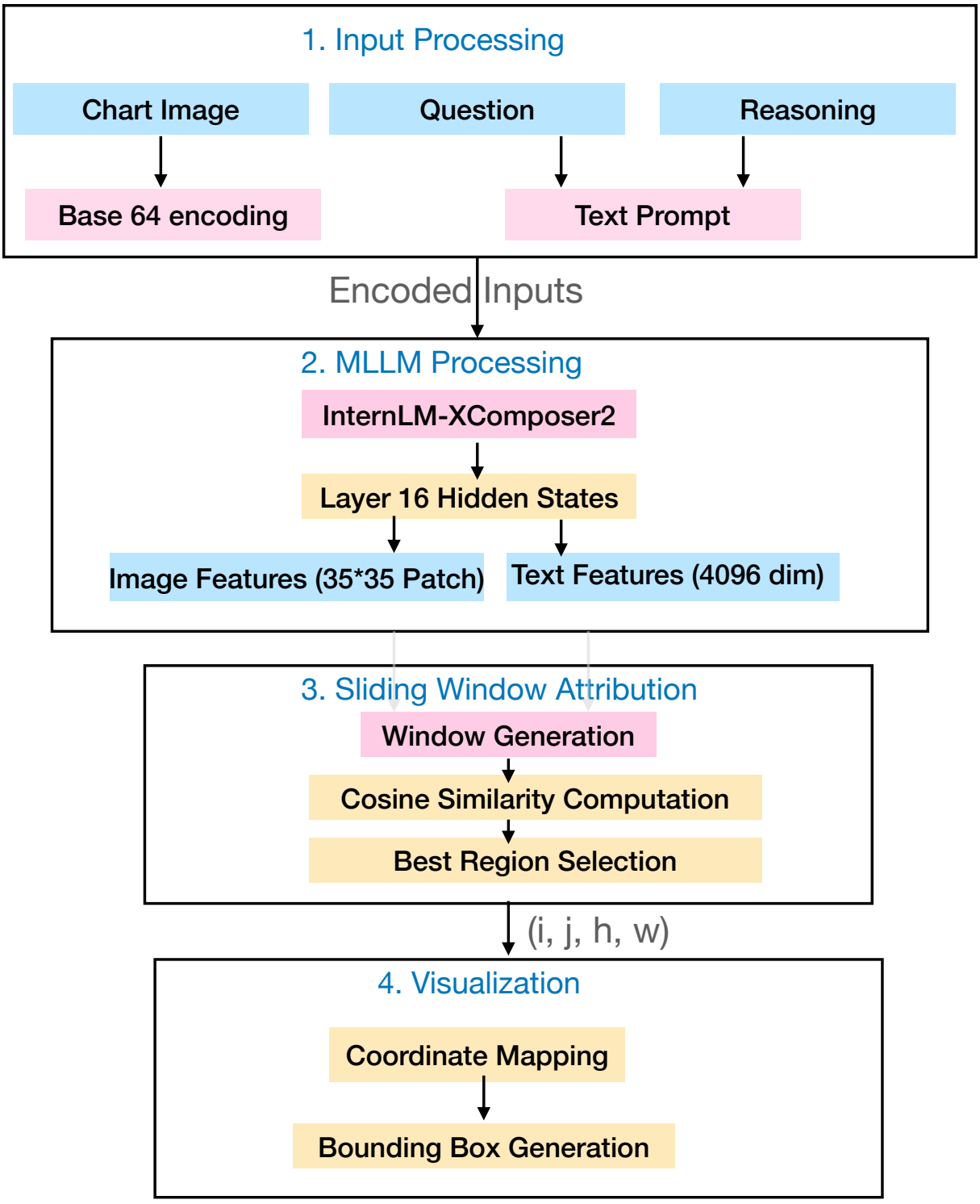


Figure 19: The pipeline architecture for chart understanding with InternLM-XComposer2 illustrates a four-stage process that bridges visual and textual modalities in chart analysis. The system progresses through Input Processing (encoding of chart images and text), MLLM Processing (multimodal feature extraction), Sliding Window Attribution (region identification), and Visualization (interpretable output generation), enabling comprehensive chart understanding through a unified architectural framework. This architecture supports reasoning based attribution.