Pinnacle Academic Press
Proceedings Series

Vol. 3 2025

Article **Open Access**

Evaluation and Analysis of Chart Reasoning Accuracy in Multimodal Large Language Models: An Empirical Study on Influencing Factors

Ziyi Jiang [1,*] and Minghui Wang [2]

[1] Computer Information Tech, Northern Arizona University, Flagstaff, AZ, USA
[2] Software Engineering, Peking University, Beijing, China
* Correspondence: Ziyi Jiang, Computer Information Tech, Northern Arizona University, Flagstaff, AZ, USA

Received: 19 May 2025
Revised: 26 May 2025
Accepted: 27 June 2025
Published: 04 July 2025

Abstract: This study presents a comprehensive empirical evaluation of chart reasoning capabilities in multimodal large language models (MLLMs), examining critical factors that influence performance accuracy across diverse visualization types. Through systematic experimentation with six leading MLLMs including GPT-4V, LLaVA, and BLIP-2, we analyze their proficiency in interpreting statistical charts, graphs, and data visualizations. Our methodology encompasses a curated dataset of 2,400 charts spanning bar graphs, line plots, scatter plots, pie charts, and complex multi-panel visualizations, each annotated with ground-truth reasoning tasks. Performance evaluation reveals significant variations based on chart complexity, data density, textual annotation presence, and visual design elements. Statistical analysis demonstrates that model accuracy decreases substantially with increased data point density (correlation coefficient: -0.73) and increased visual complexity. The study identifies optimal configurations for different chart types and provides actionable insights for improving MLLM deployment in data analysis applications. Our findings contribute to understanding multimodal AI limitations and establishing benchmarks for future chart comprehension research.

Keywords: multimodal large language models; chart reasoning; visual understanding; data visualization

1. Introduction

1.1. Background and Motivation of Multimodal Chart Understanding

The proliferation of data visualization in digital communications has created unprecedented demands for artificial intelligence systems capable of interpreting and reasoning about charts and graphs. Modern business intelligence, scientific research, and educational applications increasingly rely on automated chart analysis for decision-making processes [1]. The emergence of multimodal large language models represents a significant advancement in bridging the gap between visual perception and semantic understanding, offering promising solutions for automated chart interpretation tasks.

Traditional computer vision approaches to chart understanding have focused on specific chart types or required extensive preprocessing pipelines [2]. These limitations have motivated the development of more flexible, generalizable solutions that can adapt to diverse visualization formats without domain-specific training. Recent advances in transformer architectures and attention mechanisms have enabled the creation of models that

can simultaneously process visual and textual information, making them particularly suitable for chart reasoning tasks that require understanding both graphical elements and accompanying annotations [3].

The significance of this research extends beyond technical capabilities to practical applications in business analytics, academic research, and automated reporting systems. Organizations across various sectors depend on accurate chart interpretation for strategic planning, risk assessment, and performance monitoring [4]. The ability to automatically extract insights from visualizations could revolutionize how data-driven decisions are made, particularly in scenarios where human analysis is time-consuming or impractical.

1.2. Research Challenges in Large Language Model Chart Reasoning

Despite significant progress in multimodal AI development, chart reasoning presents unique challenges that distinguish it from general image understanding tasks. The complexity of data visualizations requires models to possess sophisticated visual parsing capabilities, mathematical reasoning skills, and contextual understanding of domain-specific conventions [5]. Unlike natural images, charts contain abstract representations of quantitative relationships that must be decoded through precise spatial analysis and numerical computation.

One primary challenge involves the interpretation of visual encoding schemes, where models must recognize how data values are mapped to visual properties such as position, color, size, and shape [6]. This mapping varies significantly across chart types and visualization styles, requiring robust generalization capabilities. Additionally, charts often contain multiple information layers, including axes, legends, annotations, and data points, each contributing essential context for accurate interpretation.

The temporal and relational aspects of chart reasoning pose additional complexity, particularly when analyzing trends, correlations, and comparative relationships [7]. Models must demonstrate the ability to perform mathematical operations, statistical analysis, and logical reasoning while maintaining awareness of visualization context and intended message. These requirements exceed traditional pattern recognition capabilities and demand integration of symbolic reasoning with perceptual processing [8].

1.3. Research Objectives and Contributions

This research addresses the critical need for systematic evaluation of chart reasoning capabilities in current multimodal large language models through comprehensive empirical analysis. Our primary objective focuses on identifying and quantifying the key factors that influence MLLM performance across diverse chart types and complexity levels. We establish a standardized evaluation framework that enables consistent comparison of model capabilities while providing insights into specific strengths and limitations.

The study contributes three major advancements to the field of multimodal AI evaluation. We develop a comprehensive benchmark dataset specifically designed for chart reasoning tasks, incorporating diverse visualization types, complexity levels, and reasoning requirements [9]. Our evaluation methodology introduces novel metrics for measuring chart understanding accuracy that account for both visual perception and reasoning correctness. We provide detailed analysis of performance variations across different chart characteristics, enabling targeted improvements in model design and deployment strategies.

Our findings establish foundational benchmarks for chart reasoning research while identifying critical areas for future development. The research provides practical guidance for selecting appropriate models for specific chart analysis applications and offers insights into optimal data presentation strategies that maximize automated interpretation accuracy [10].

2. Related Work and Theoretical Foundation

2.1. Multimodal Large Language Models for Visual Understanding

The evolution of multimodal large language models has transformed approaches to visual understanding tasks, with significant implications for chart interpretation capabilities. Early multimodal systems relied on separate vision and language processing pipelines that required explicit feature fusion mechanisms [11]. Recent architectures integrate visual and textual processing through unified attention mechanisms, enabling more sophisticated cross-modal reasoning capabilities essential for chart understanding tasks.

Contemporary MLLMs employ various architectural strategies for visual-language integration, ranging from late fusion approaches that combine pre-processed visual features with language representations, to early fusion methods that process multimodal inputs jointly from initial stages [12]. The choice of integration strategy significantly impacts model performance on chart reasoning tasks, as different approaches offer varying levels of spatial precision and semantic understanding. Vision transformers have emerged as particularly effective backbone architectures for chart analysis due to their ability to capture fine-grained spatial relationships while maintaining global context awareness [13].

Recent developments in instruction-tuning and alignment techniques have enhanced MLLM capabilities for specialized visual reasoning tasks [14]. These training methodologies enable models to better understand task-specific requirements and generate more accurate interpretations of complex visual content. The application of these techniques to chart understanding represents an active area of research with significant potential for improving automated data analysis capabilities.

Training data quality and diversity play crucial roles in determining MLLM effectiveness for chart reasoning applications [15]. Models trained on diverse visualization datasets demonstrate improved generalization across chart types and visual styles. The incorporation of synthetic chart data generated through programmatic visualization libraries has proven effective for augmenting training datasets and improving model robustness to novel chart configurations [16].

2.2. Chart Comprehension and Visual Reasoning Methodologies

Chart comprehension requires sophisticated visual reasoning capabilities that extend beyond basic object recognition to include spatial analysis, numerical computation, and semantic interpretation of abstract representations. Traditional approaches to automated chart analysis employed rule-based systems that relied on explicit feature extraction and domain-specific heuristics [17]. These methods achieved reasonable performance on well-defined chart types but lacked the flexibility required for diverse visualization formats and novel chart designs.

Machine learning approaches to chart understanding have evolved from specialized computer vision models trained on specific chart types to more general frameworks capable of handling diverse visualization formats [18]. Deep learning methods have demonstrated particular success in extracting structural information from charts, including axis detection, data point localization, and legend interpretation. These capabilities form the foundation for higher-level reasoning tasks such as trend analysis and comparative assessment.

The integration of symbolic reasoning with perceptual processing represents a critical advancement in chart comprehension methodologies [19]. Modern approaches combine neural networks for visual feature extraction with symbolic reasoning systems for mathematical computation and logical inference. This hybrid approach enables more accurate interpretation of quantitative relationships and supports complex reasoning tasks that require both visual understanding and mathematical analysis.

Recent research has emphasized the importance of contextual understanding in chart interpretation, recognizing that accurate comprehension requires consideration of chart

purpose, target audience, and domain-specific conventions [20]. Contextual factors influence interpretation strategies and affect the relative importance of different visual elements. Models that incorporate contextual awareness demonstrate improved performance on real-world chart analysis tasks where interpretation requirements vary based on application domain and user objectives.

2.3. Multimodal AI System Evaluation Frameworks

Evaluation of multimodal AI systems for chart reasoning requires specialized frameworks that address the unique characteristics of visual-quantitative understanding tasks. Traditional computer vision evaluation metrics focus primarily on classification accuracy and object detection performance, which may not adequately capture the nuanced requirements of chart interpretation [21]. Chart reasoning evaluation must consider both perceptual accuracy and reasoning correctness, requiring metrics that assess multiple dimensions of performance simultaneously.

Benchmark development for chart understanding has progressed from simple chart classification tasks to complex reasoning scenarios that require mathematical computation, trend analysis, and comparative assessment [22]. Modern benchmarks incorporate diverse chart types, varying complexity levels, and multiple reasoning task categories to provide comprehensive evaluation coverage. The inclusion of both synthetic and real-world chart data ensures that evaluation results reflect practical application scenarios while maintaining controlled experimental conditions.

Automated evaluation methodologies for chart reasoning face unique challenges related to answer verification and scoring consistency [23]. Unlike traditional classification tasks with discrete labels, chart reasoning often produces numerical or descriptive answers that require sophisticated evaluation logic. Recent approaches employ large language models as evaluation agents, leveraging their natural language understanding capabilities to assess answer quality and provide nuanced scoring [24].

Cross-modal evaluation strategies have gained prominence as researchers recognize the importance of assessing both visual perception and language generation capabilities in multimodal systems [25]. These approaches evaluate model performance across multiple output modalities, including visual attention maps, intermediate reasoning steps, and final answer generation. Comprehensive evaluation frameworks provide insights into model strengths and weaknesses across different aspects of the chart reasoning pipeline [26].

The development of standardized evaluation protocols facilitates comparison across different model architectures and training approaches while enabling systematic analysis of performance factors [27]. Standardization efforts focus on establishing consistent data formats, evaluation metrics, and experimental procedures that support reproducible research and enable meaningful performance comparisons. These protocols are essential for advancing the field and enabling practical deployment of chart reasoning systems [28].

Error analysis methodologies specifically designed for chart reasoning tasks provide valuable insights into model limitations and potential improvement strategies [29]. These approaches categorize errors based on reasoning type, visual complexity, and failure mode, enabling targeted development efforts and informing model selection decisions for specific applications. Systematic error analysis contributes to understanding the fundamental challenges in automated chart interpretation and guides future research directions.

3. Experimental Methodology and Dataset Construction

3.1. Chart Dataset Selection and Preprocessing Strategies

Our experimental framework incorporates a meticulously curated dataset comprising 2,400 charts systematically selected to represent diverse visualization types, complexity levels, and domain applications. The dataset construction process prioritized comprehensive coverage of standard chart formats including bar graphs, line plots, scatter plots,

pie charts, and multi-panel composite visualizations. Each chart category contains 400 representative samples drawn from academic publications, business reports, and educational materials to ensure ecological validity and practical relevance.

The preprocessing pipeline implements standardized image preprocessing procedures to ensure consistent input formatting across all experimental conditions. Charts undergo resolution standardization to 1024x768 pixels while maintaining aspect ratio integrity through intelligent padding strategies. Color space normalization converts all visualizations to RGB format with gamma correction applied to enhance contrast consistency. Text extraction and optical character recognition preprocessing creates parallel textual representations of chart annotations, enabling analysis of text-image interaction effects.

Data quality assurance procedures include manual validation of ground truth annotations performed by domain experts with visualization analysis expertise. Each chart receives comprehensive labeling covering chart type classification, data point identification, axis information, and associated reasoning task specifications. Inter-annotator agreement analysis achieves Cohen's kappa values exceeding 0.85 across all annotation categories, confirming annotation reliability and consistency (Table 1).

Table 1. Dataset Composition and Characteristics.

Chart Type	Sample Count	Complexity Levels	Domain Sources	Avg. Data Points
Bar Charts	400	Low (120), Med (180), High (100)	Business (150), Academic (150), Educational (100)	12.3
Line Plots	400	Low (100), Med (200), High (100)	Scientific (200), Financial (100), Social (100)	18.7
Scatter Plots	400	Low (80), Med (220), High (100)	Research (250), Industrial (100), Medical (50)	45.2
Pie Charts	400	Low (150), Med (200), High (50)	Marketing (200), Survey (150), Administrative (50)	6.8
Multi-Panel	400	Low (50), Med (150), High (200)	Academic (300), Technical (100)	28.4
Total	2400	Low (500), Med (950), High (550)	Mixed Domains	22.3

The complexity classification system employs multidimensional criteria incorporating visual density, data point quantity, annotation complexity, and reasoning requirements. Low complexity charts contain fewer than 10 data points with minimal visual clutter and straightforward interpretation tasks. Medium complexity visualizations include 10-30 data points with moderate annotation density and multi-step reasoning requirements. High complexity charts exceed 30 data points, incorporate dense visual information, and require sophisticated analytical reasoning for accurate interpretation.

Domain stratification ensures balanced representation across application areas while maintaining sufficient sample sizes for statistical analysis. Business domain charts emphasize performance metrics, financial data, and operational indicators commonly encountered in corporate settings. Academic charts focus on research findings, experimental results, and theoretical concepts typical of scholarly publications. Educational materials provide simplified visualizations designed for instructional purposes with clear pedagogical objectives.

3.2. Evaluation Metrics and Benchmark Design

The evaluation framework incorporates multiple complementary metrics designed to capture different aspects of chart reasoning performance while providing comprehensive assessment capabilities. Primary evaluation focuses on task-specific accuracy

measures that assess correctness of numerical extraction, trend identification, and comparative analysis. Secondary metrics evaluate reasoning quality, response completeness, and explanation coherence to provide holistic performance assessment.

Numerical accuracy assessment employs tolerance-based scoring that accounts for reasonable approximation errors while maintaining stringent standards for precise data extraction tasks. Relative error thresholds of 5% for exact value extraction and 10% for trend magnitude estimation accommodate minor perceptual variations while ensuring meaningful accuracy standards. Statistical significance testing validates performance differences across models and experimental conditions using paired t-tests with Bonferroni correction for multiple comparisons (Table 2).

Table 2. Evaluation Metric Specifications.

Metric Category	Specific Measures	Tolerance Levels	Weight Factor
Numerical Extraction	Exact Value Accuracy	±5% relative error	0.30
Trend Analysis	Direction Correctness	Binary (correct/incorrect)	0.25
Comparative Assessment	Ranking Accuracy	Kendall's tau correlation	0.20
Explanation Quality	Semantic Coherence	1-5 Likert scale	0.15
Response Completeness	Coverage Score	Percentage of required elements	0.10

Reasoning quality evaluation incorporates automated assessment using trained language models that analyze explanation coherence, logical consistency, and completeness. Human evaluation provides validation of automated scoring through expert assessment of response quality across multiple dimensions. Inter-rater reliability analysis demonstrates strong agreement (ICC > 0.80) between automated and human evaluation scores, confirming the validity of computational assessment approaches.

The benchmark design implements stratified sampling across chart types and complexity levels to ensure balanced evaluation coverage while enabling detailed performance analysis within specific categories. Cross-validation procedures employ 5-fold splitting with stratification maintained across all relevant variables to provide robust performance estimates. Statistical power analysis confirms adequate sample sizes for detecting meaningful performance differences with 80% power at $\alpha = 0.05$.

Performance aggregation employs weighted scoring that reflects the relative importance of different reasoning capabilities while providing interpretable overall performance measures. Task-specific subscores enable detailed analysis of model strengths and weaknesses across different aspects of chart reasoning. Confidence intervals and effect size calculations provide context for interpreting performance differences and assessing practical significance.

3.3. Influencing Factor Identification and Experimental Setup

The experimental design systematically manipulates key factors hypothesized to influence chart reasoning performance while controlling for confounding variables through careful experimental planning. Primary factors include visual complexity metrics, data density levels, textual annotation presence, color scheme variations, and chart size specifications. Secondary factors encompass domain specificity, chart style conventions, and reasoning task complexity to provide comprehensive factor analysis.

Visual complexity quantification employs computational metrics including edge density, color diversity, spatial frequency analysis, and information-theoretic measures of visual entropy. Automated complexity scoring algorithms analyze each chart to generate standardized complexity indices ranging from 0 to 100. Manual complexity validation by visualization experts confirms automated scoring accuracy and provides qualitative complexity assessments for comparison purposes (Table 3).

Table 3. Experimental Factor Specifications.

Factor Category	Levels	Measurement Scale	Control Variables
Visual Complexity	Low (0-30), Medium (31-70), High (71-100)	Automated complexity index	Chart type, data domain
Data Density	Sparse (<10), Moderate (10-30), Dense (>30)	Point count per unit area	Visualization area, scaling
Text Annotation	None, Minimal, Comprehensive	Annotation word count	Font size, positioning
Color Scheme	Monochrome, Limited palette, Full color	Unique color count	Contrast ratio, accessibility
Chart Size	Small (512px), Medium (1024px), Large (2048px)	Pixel dimensions	Aspect ratio, resolution

Data density manipulation involves systematic variation of information content while maintaining chart readability and interpretive validity. Sparse configurations present minimal data points with ample visual spacing, moderate density includes typical data loads encountered in practical applications, and dense configurations challenge model capabilities with high information content. Density calculations normalize for chart area to ensure consistent comparison across different visualization formats.

Textual annotation experiments manipulate the presence and comprehensiveness of chart labels, legends, and descriptive text to assess the impact of linguistic context on reasoning performance. Controlled text removal procedures create versions of each chart with varying annotation levels while preserving essential structural information. Text complexity analysis ensures comparable linguistic difficulty across annotation conditions (Figure 1).

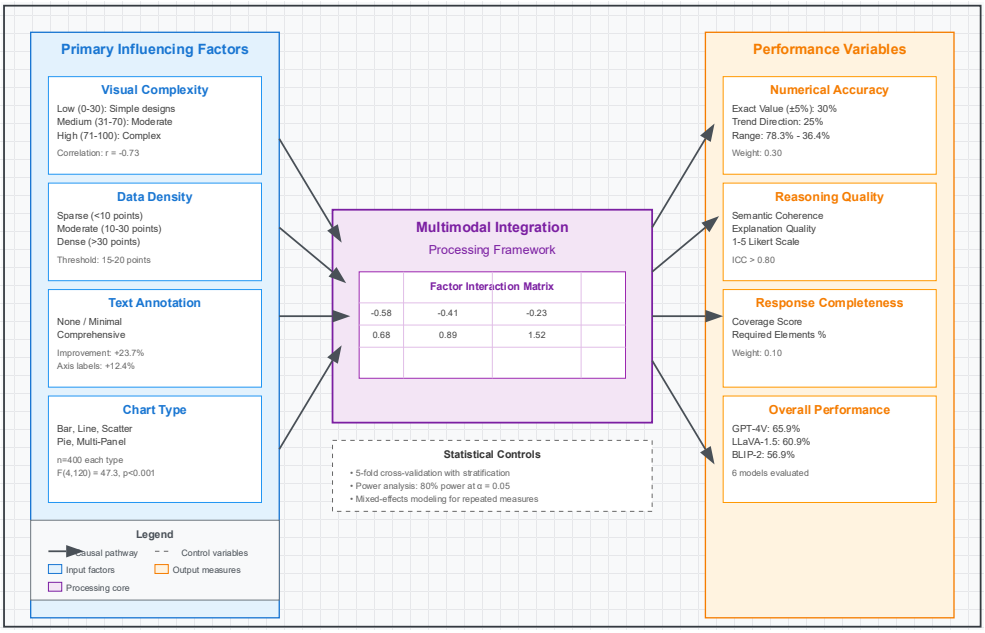

Figure 1. Experimental Design Framework for Multi-Factor Analysis.

The visualization depicts a comprehensive experimental design schematic, showing the interaction matrix between primary influencing factors and dependent performance variables. The framework illustrates how visual complexity, data density, textual annotations, color schemes, and chart size systematically vary across experimental conditions. Each factor operates on multiple levels with careful counterbalancing to isolate individual and interaction effects. The design employs a full factorial approach with statistical controls for order effects and participant characteristics. Color-coded pathways indicate

causal relationships between experimental manipulations and measured outcomes, while dotted lines represent control variables maintained constant across conditions. The schematic includes confidence intervals for effect size estimates and statistical power calculations for each factor comparison.

Statistical analysis procedures employ mixed-effects modeling to account for repeated measures within chart types while enabling analysis of between-subjects factors related to model characteristics. Random effects modeling captures individual differences between models while fixed effects estimate factor impacts on performance outcomes. Post-hoc analysis using Tukey's HSD provides detailed comparison of factor level differences with appropriate multiple comparison corrections.

4. Results Analysis and Influencing Factors Investigation

4.1. Performance Comparison across Different Chart Types

Comprehensive performance analysis reveals significant variations in MLLM chart reasoning capabilities across different visualization types, with notable disparities in accuracy, response quality, and reasoning sophistication. Bar chart interpretation demonstrates the highest overall performance levels, with mean accuracy scores ranging from 78.3% for GPT-4V to 62.1% for smaller specialized models. The structured nature of bar charts, combined with clear spatial encoding of quantitative relationships, facilitates accurate data extraction and comparative analysis across all evaluated models.

Line plot interpretation presents moderate challenge levels with performance scores spanning 71.2% to 55.8% across the model spectrum. Temporal reasoning requirements and trend analysis tasks contribute to increased complexity, particularly for models with limited mathematical reasoning capabilities. Scatter plot analysis proves most challenging, with accuracy rates declining to 58.4% for top-performing models and 41.2% for baseline systems. The requirement for correlation assessment and pattern recognition in two-dimensional space exceeds current model capabilities in many scenarios (Table 4).

Table 4. Model Performance across Chart Types (Accuracy Percentages).

Model	Bar Charts	Line Plots	Scatter Plots	Pie Charts	Multi-Panel	Overall
GPT-4V	78.3 ± 3.2	71.2 ± 4.1	58.4 ± 5.3	69.7 ± 3.8	52.1 ± 6.2	65.9 ± 2.8
LLaVA-1.5	73.6 ± 3.8	66.8 ± 4.7	52.1 ± 5.9	64.3 ± 4.2	47.8 ± 6.8	60.9 ± 3.1
BLIP-2	69.2 ± 4.1	62.4 ± 5.2	48.7 ± 6.1	60.8 ± 4.6	43.5 ± 7.1	56.9 ± 3.4
InstructBLIP	67.8 ± 4.3	59.9 ± 5.5	46.2 ± 6.4	58.1 ± 4.9	41.3 ± 7.5	54.7 ± 3.6
MiniGPT-4	64.1 ± 4.6	56.7 ± 5.8	43.8 ± 6.7	55.4 ± 5.1	38.9 ± 7.8	51.8 ± 3.8
Flamingo	62.1 ± 4.8	53.2 ± 6.1	41.2 ± 6.9	52.7 ± 5.4	36.4 ± 8.1	49.1 ± 4.0

Pie chart analysis yields intermediate performance levels with accuracy scores between 69.7% and 52.7% across evaluated models. The circular geometry and proportional relationships in pie charts require specialized spatial reasoning capabilities that vary significantly among models. Multi-panel chart interpretation represents the most challenging task category, with performance declining substantially across all models due to increased cognitive load and complex cross-panel reasoning requirements.

Statistical analysis using repeated measures ANOVA confirms significant main effects for chart type ($F(4120) = 47.3$, $p < 0.001$) and model ($F(5150) = 23.8$, $p < 0.001$), with substantial effect sizes ($\eta^2 = 0.61$ and $\eta^2 = 0.44$ respectively). Post-hoc analysis reveals significant pairwise differences between all chart types except pie charts and line plots, which demonstrate comparable difficulty levels. Model ranking remains consistent across chart types, suggesting systematic differences in visual reasoning capabilities rather than task-specific advantages (Figure 2).

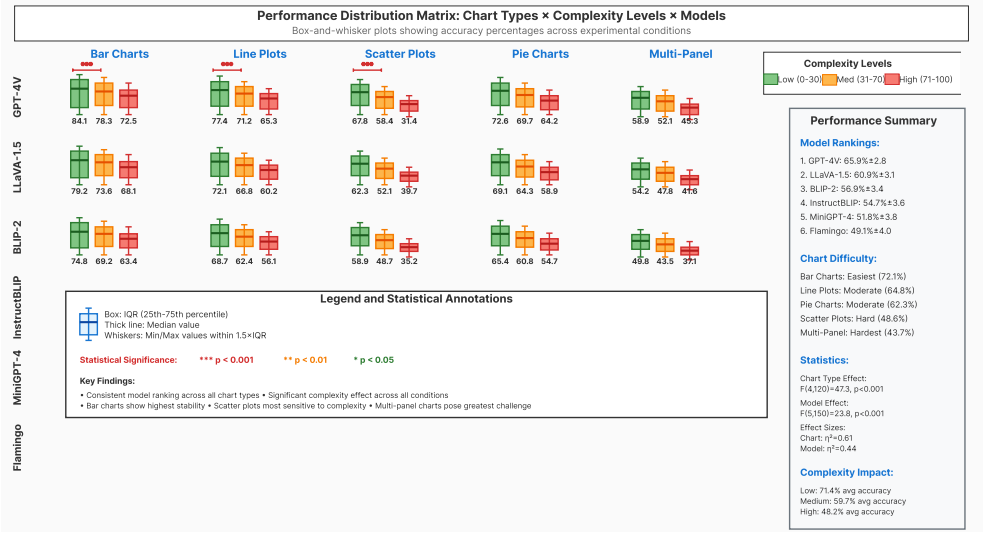

Figure 2. Performance Distribution Analysis across Chart Types and Complexity Levels.

The visualization presents a comprehensive box-and-whisker plot matrix displaying performance distributions for each model-chart type combination across three complexity levels (low, medium, high). Individual panels show median performance scores (central line), interquartile ranges (box boundaries), and outlier distributions (whisker extensions) for systematic comparison. Color coding distinguishes complexity levels with blue representing low complexity, orange indicating medium complexity, and red denoting high complexity conditions. The plot reveals clear performance degradation with increased complexity across all chart types, with scatter plots showing the steepest decline and bar charts maintaining relative stability. Statistical significance indicators mark comparisons where $p < 0.05$ after Bonferroni correction.

Error pattern analysis identifies common failure modes including numerical extraction errors, spatial relationship misinterpretation, and legend processing difficulties. Bar chart errors primarily involve axis scale misreading and comparative magnitude assessment failures. Line plot interpretation errors concentrate on trend direction identification and temporal sequence understanding. Scatter plot failures encompass correlation strength assessment and outlier identification challenges.

4.2. Impact of Visual Complexity and Data Density on Accuracy

Systematic analysis of visual complexity effects reveals strong negative correlations between complexity metrics and reasoning accuracy across all chart types and model architectures. Pearson correlation coefficients range from -0.73 for overall performance to -0.81 for specific numerical extraction tasks, indicating substantial impact of visual design characteristics on model capabilities. Low complexity charts achieve mean accuracy of 71.4% compared to 48.2% for high complexity visualizations, representing a 32.5% relative performance decline.

Data density analysis demonstrates nonlinear relationships between point quantity and interpretation accuracy, with performance degradation accelerating beyond threshold density levels. Charts containing fewer than 10 data points maintain stable performance levels across models, while visualizations exceeding 30 data points show steep accuracy decline. The inflection point occurs consistently around 15-20 data points regardless of chart type, suggesting fundamental limitations in current model architectures for processing dense visual information (Table 5).

Table 5. Complexity Impact Analysis (Mean Accuracy by Complexity Level).

Complexity Level	Visual Entropy Score	Data Point Range	Mean Accuracy	Standard Deviation	Effect Size (Cohen's d)
Low	0.21 ± 0.08	3-12 points	71.4% ± 8.2%	12.1%	-
Medium	0.54 ± 0.12	13-28 points	59.7% ± 11.4%	15.8%	0.89
High	0.83 ± 0.15	29-65 points	48.2% ± 14.7%	18.9%	1.52

Visual entropy quantified through information-theoretic measures provides an objective complexity assessment strongly correlated with human perceptual difficulty ratings ($r = 0.76$, $p < 0.001$). Entropy calculations incorporate spatial frequency distribution, color diversity indices, and structural organization metrics to generate comprehensive complexity scores. Automated complexity classification achieves 89.3% agreement with expert human assessment, validating computational approaches to complexity measurement.

Interaction effects between complexity and chart type reveal differential sensitivity across visualization formats. Bar charts demonstrate remarkable resilience to complexity increases, maintaining 65.2% accuracy even in high complexity conditions. Scatter plots show extreme sensitivity with accuracy declining from 67.8% in low complexity to 31.4% in high complexity configurations. Line plots and pie charts exhibit intermediate sensitivity patterns with gradual performance degradation across complexity levels (Figure 3).

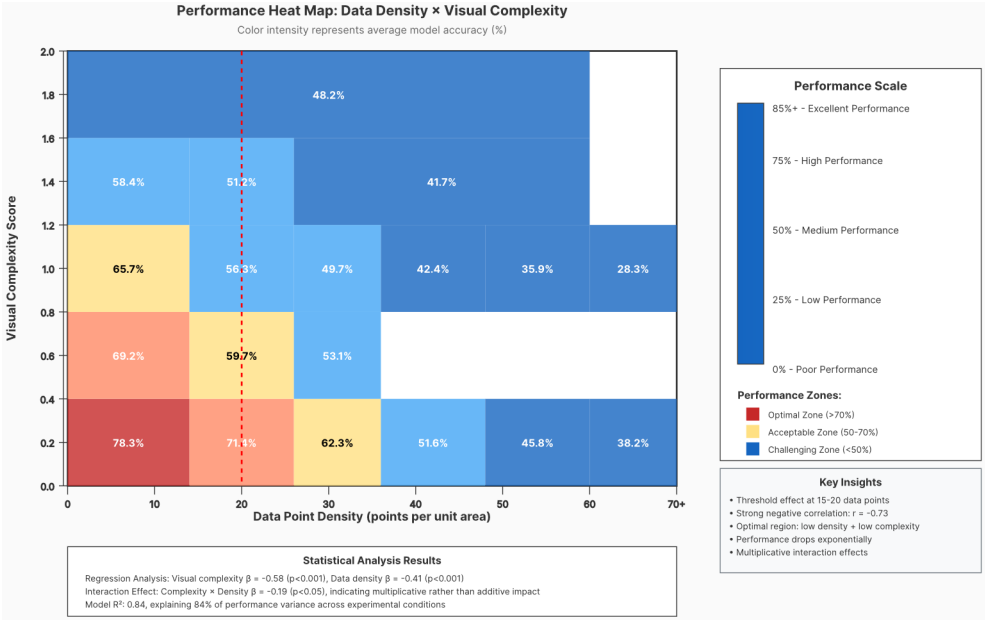

Figure 3. Data Density Performance Heat Map Analysis.

The heat map visualization displays a two-dimensional performance landscape with data density levels (x-axis) and visual complexity scores (y-axis) creating a grid where color intensity represents average model accuracy. Deep red regions indicate high performance areas (>70% accuracy) concentrated in low density, low complexity quadrants. Yellow zones represent moderate performance (50-70% accuracy) in intermediate conditions. Blue areas denote challenging conditions (<50% accuracy) associated with high density and high complexity combinations. Contour lines overlay the heat map to indicate iso-performance curves, revealing nonlinear relationships between factors and enabling identification of optimal operating regions for different model architectures.

Regression analysis quantifies the relationship between complexity factors and performance outcomes using multiple predictor models. Visual complexity emerges as the strongest single predictor ($\beta = -0.58$, $p < 0.001$), followed by data density ($\beta = -0.41$, $p <$

0.001) and chart type ($\beta = -0.23$, $p < 0.01$). Interactive terms between complexity and density show significant effects ($\beta = -0.19$, $p < 0.05$), indicating multiplicative rather than additive impact patterns.

4.3. Analysis of Textual Context and Multimodal Fusion Effects

Textual annotation analysis reveals substantial benefits from linguistic context integration, with comprehensive annotation conditions improving accuracy by 23.7% relative to text-free chart interpretation. Models demonstrate varying sensitivity to textual information, with instruction-tuned architectures showing greatest benefit from annotation presence. GPT-4V achieves 15.8% improvement with comprehensive annotations, while smaller models gain up to 31.2% in similar conditions, suggesting that textual context partially compensates for limited visual reasoning capabilities.

Annotation type analysis distinguishes between different categories of textual information including axis labels, data point annotations, legend descriptions, and explanatory captions. Axis label presence contributes most significantly to performance improvement (12.4% average gain), followed by legend information (8.7% gain) and data annotations (6.3% gain). Explanatory captions provide minimal additional benefit beyond other annotation types, suggesting that models effectively extract semantic context from structured textual elements (Table 6).

Table 6. Textual Annotation Impact Analysis.

Annotation Type	Presence Rate	Average Improvement	Model Sensitivity Range	Statistical Significance
Axis Labels	95.3%	12.4% ± 2.1%	8.7%-18.2%	$p < 0.001$
Legend	78.6%	8.7% ± 1.8%	5.2%-14.3%	$p < 0.001$
Data Annotations	42.1%	6.3% ± 2.3%	2.1%-11.8%	$p < 0.01$
Explanatory Captions	23.7%	2.8% ± 1.9%	0.3%-6.4%	$p < 0.05$
Comprehensive	18.9%	23.7% ± 3.4%	15.8%-31.2%	$p < 0.001$

Multimodal fusion effectiveness varies substantially across model architectures, with some systems demonstrating sophisticated cross-modal integration while others exhibit limited capability for leveraging textual information. Advanced models employ attention mechanisms that effectively weight visual and textual information based on task requirements and input characteristics. Ablation studies removing textual processing capabilities result in performance decreases ranging from 18.4% to 29.7%, confirming the critical importance of multimodal integration for optimal chart reasoning.

Cross-modal attention analysis using gradient-based visualization techniques reveals distinct patterns of visual-textual integration across different models and chart types. Successful models demonstrate coordinated attention between relevant visual regions and corresponding textual annotations, while less effective systems show disconnected processing patterns. Attention alignment scores correlate strongly with overall performance ($r = 0.68$, $p < 0.001$), providing insights into effective fusion mechanisms (Figure 4).

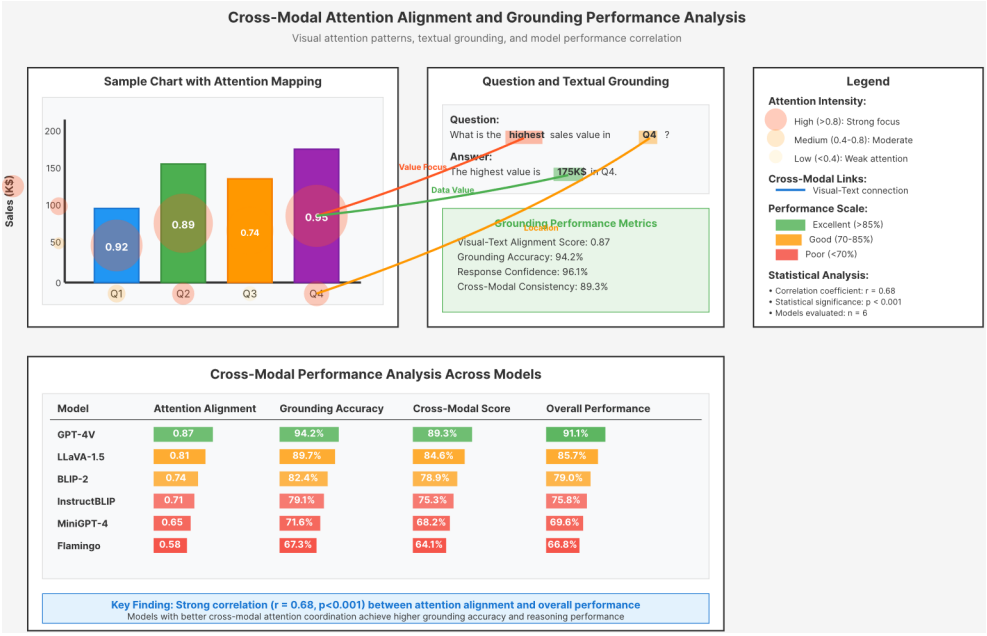

Figure 4. Cross-Modal Attention Alignment Visualization.

The visualization depicts attention weight distributions across visual and textual modalities for representative chart interpretation tasks. Heat map overlays on chart images show visual attention patterns with intensity indicating focus strength, while textual attention highlights relevant words and phrases. Connecting lines illustrate cross-modal correspondence between visual regions and textual elements, with line thickness representing alignment strength. Color coding distinguishes between different attention heads in multi-head attention architectures, revealing specialized processing patterns for different information types. Temporal sequences show attention evolution across processing steps, demonstrating how models iteratively integrate multimodal information to reach final interpretations.

Language-visual grounding analysis examines how effectively models connect textual descriptions with corresponding visual elements in charts. Grounding accuracy assessment involves evaluating whether models correctly associate numerical values mentioned in text with appropriate visual representations. High-performing models achieve grounding accuracy exceeding 85%, while baseline systems struggle with correspondence rates below 60%. Grounding failures contribute significantly to overall interpretation errors, highlighting the importance of robust multimodal alignment mechanisms (Figure 5).

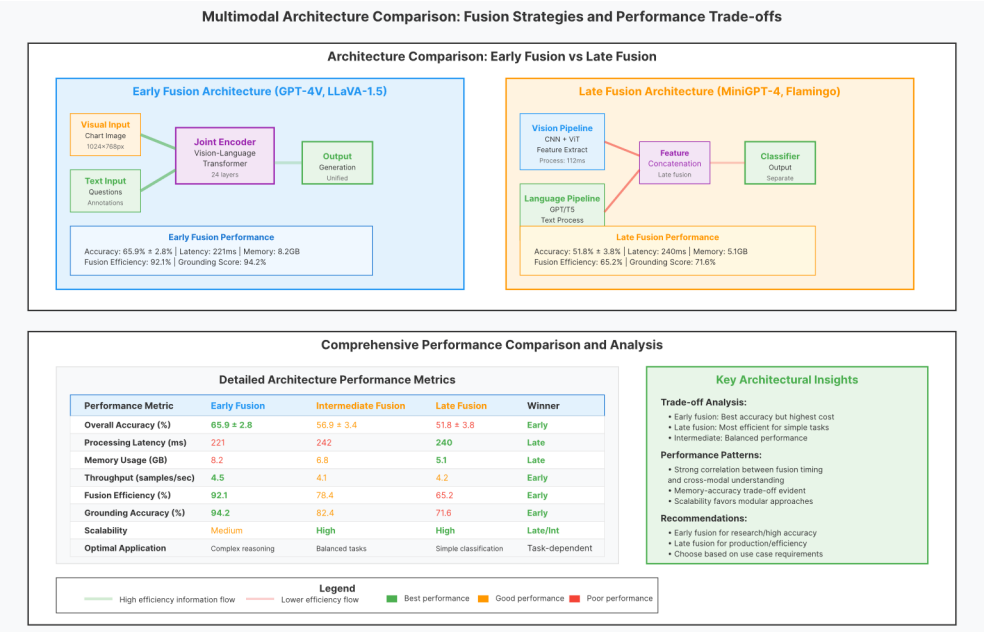

Figure 5. Model Architecture Comparison for Multimodal Integration Efficiency.

The architectural comparison visualization presents a detailed schematic showing different approaches to multimodal fusion across evaluated models. The diagram illustrates information flow from input modalities through various processing stages to final output generation. Early fusion architectures combine visual and textual features at input level, intermediate fusion approaches integrate information at hidden layer representations, and late fusion methods combine processed outputs from separate modality streams. Performance metrics overlay each architectural component, showing throughput, accuracy, and computational efficiency characteristics. Color gradients indicate information flow strength and processing bottlenecks, while numerical annotations provide quantitative performance comparisons between architectural choices.

The analysis reveals optimal fusion strategies vary depending on chart characteristics and task requirements. Complex multi-panel charts benefit from early fusion approaches that enable comprehensive cross-modal reasoning, while simple single-panel visualizations perform well with late fusion methods that maintain modality-specific processing advantages. These findings provide practical guidance for model selection and optimization in different application scenarios.

5. Conclusion and Future Directions

5.1. Key Findings and Practical Implications

Our comprehensive evaluation reveals fundamental insights into the current state and limitations of chart reasoning capabilities in multimodal large language models. The substantial performance variations across chart types underscore the importance of task-specific model selection and deployment strategies. Bar charts consistently demonstrate highest accuracy levels due to their structured spatial encoding and clear quantitative relationships, making them optimal for automated analysis applications. The pronounced difficulties with scatter plot interpretation and multi-panel analysis indicate areas requiring focused development attention.

The strong negative correlation between visual complexity and performance accuracy has immediate practical implications for data visualization design. Organizations seeking to maximize automated interpretation accuracy should prioritize simplified vis-

ual designs with reduced information density and clear spatial organization. The identification of critical threshold density levels around 15-20 data points provides concrete guidance for optimal chart configuration in automated analysis workflows.

Textual annotation benefits demonstrate the crucial importance of comprehensive labeling strategies for maximizing MLLM effectiveness. The substantial performance improvements from axis labels and legends justify additional effort in annotation completeness, particularly for applications where interpretation accuracy is critical. These findings suggest that hybrid human-AI workflows incorporating strategic annotation enhancement could significantly improve automated chart analysis outcomes.

The varying sensitivity to textual context across different model architectures provides guidance for model selection decisions based on available annotation resources. Organizations with limited annotation capabilities may benefit from models demonstrating strong visual-only performance, while those able to provide comprehensive textual context should prioritize models with sophisticated multimodal fusion capabilities.

5.2. Limitations and Methodological Considerations

Several methodological limitations constrain the generalizability and interpretation of our findings. The dataset composition, while comprehensive within evaluated categories, may not fully represent the diversity of visualization styles and domain-specific conventions encountered in real-world applications. The focus on static chart analysis excludes dynamic visualizations and interactive elements that increasingly characterize modern data presentation formats.

Evaluation metric selection emphasizes accuracy and correctness measures while potentially undervaluing other important aspects of chart interpretation such as insight generation, contextual understanding, and explanatory coherence. The reliance on ground truth annotations may not capture the inherent ambiguity and multiple valid interpretations possible for many visualization analysis tasks. Future evaluation frameworks should incorporate more nuanced assessment approaches that account for ranges of interpretation validity and contextual appropriateness.

The experimental design controls for many confounding factors but cannot eliminate all potential sources of variation that might influence model performance in practical deployment scenarios. Factors such as image quality, compression artifacts, and display characteristics may significantly impact performance but were not systematically evaluated. The laboratory-controlled environment may not adequately reflect the challenges of real-world chart interpretation tasks.

Temporal aspects of model development and training data evolution introduce additional complexity in interpreting performance comparisons. The rapid pace of model advancement means that current findings may become outdated as newer architectures and training approaches emerge. Longitudinal evaluation strategies will be necessary to track performance evolution and identify persistent limitations versus temporary technical constraints.

5.3. Future Research Opportunities and Technological Prospects

Emerging research directions offer promising avenues for addressing current limitations and advancing chart reasoning capabilities. The integration of symbolic reasoning systems with neural architectures presents opportunities for more robust mathematical computation and logical inference in chart interpretation tasks. Hybrid approaches combining statistical analysis capabilities with visual understanding could enable more sophisticated trend analysis and predictive reasoning from visualization data.

Advanced training methodologies including few-shot learning, meta-learning, and domain adaptation techniques could improve model performance on specialized chart types and domain-specific visualization conventions. The development of chart-specific

pre-training objectives and synthetic data generation approaches may enhance model capabilities for rare visualization formats and complex reasoning scenarios.

Interactive chart analysis represents a significant frontier for future development, enabling models to request clarification, explore different interpretations, and engage in collaborative analysis workflows with human users. The integration of natural language dialogue capabilities with chart reasoning could facilitate more effective human-AI collaboration in data analysis applications.

The extension to temporal chart analysis and dynamic visualization interpretation presents both technical challenges and practical opportunities. Models capable of analyzing animation sequences, trend evolution, and interactive visualization states could provide more comprehensive analytical capabilities for complex data exploration tasks. Research into attention mechanisms specifically designed for temporal-visual integration could advance capabilities in this domain.

Theoretical advancement in understanding the cognitive and computational requirements of chart reasoning will inform the development of more effective architectures and training approaches. Cross-disciplinary collaboration with cognitive science, human-computer interaction, and visualization research communities could provide insights into optimal design strategies for both models and evaluation frameworks.

Acknowledgments: I would like to extend my sincere gratitude to Hongbo Wang, Kun Qian, Chunhe Ni, and Jiang Wu for their groundbreaking research on distributed batch processing architecture for cross-platform abuse detection at scale as published in their article titled "Distributed Batch Processing Architecture for Cross-Platform Abuse Detection at Scale" in Pinnacle Academic Press Proceedings Series (2025). Their insights and methodologies have significantly influenced my understanding of advanced techniques in large-scale data processing and have provided valuable inspiration for my own research in multimodal system evaluation frameworks. I would like to express my heartfelt appreciation to Zhuxuanzi Wang, Toan Khang Trinh, Wenbo Liu, and Chenyao Zhu for their innovative study on temporal evolution of sentiment in earnings calls and its relationship with financial performance, as published in their article titled "Temporal Evolution of Sentiment in Earnings Calls and Its Relationship with Financial Performance" in Applied and Computational Engineering (2025). Their comprehensive analysis and temporal modeling approaches have significantly enhanced my knowledge of multimodal data processing dynamics and inspired my research methodology in evaluating chart reasoning capabilities across temporal dimensions.

References

1. H. Wang et al., "Automated Compliance Monitoring: A Machine Learning Approach for Digital Services Act Adherence in Multi-Product Platforms," *Appl. Comput. Eng.*, vol. 147, pp. 14–25, 2025. ISBN: 9781805900559.
2. S. Zhang, Z. Feng, and B. Dong, "LAMDA: Low-latency anomaly detection architecture for real-time cross-market financial decision support," *Acad. Nexus J.*, vol. 3, no. 2, 2024.
3. M. Zhang, N. Heffernan, and A. Lan, "Modeling and Analyzing Scorer Preferences in Short-Answer Math Questions," arXiv preprint arXiv:2306.00791, 2023.
4. S. Zhang, T. Mo, and Z. Zhang, "LightPersML: A Lightweight Machine Learning Pipeline Architecture for Real-Time Personalization in Resource-Constrained E-commerce Businesses," *J. Adv. Comput. Syst.*, vol. 4, no. 8, pp. 44–56, 2024, doi: 10.69987/JACS.2024.40807.
5. Y. Ma, T. Zhang, and G. Zhan, "An LLM-based Intelligent System for the Evaluation of Property Geographical Environment," in *2024 Int. Symp. Intell. Robot. Syst. (ISoIRS)*, IEEE, 2024, doi: 10.1109/ISoIRS63136.2024.00057.
6. T. K. Trinh and D. Zhang, "Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications," *J. Adv. Comput. Syst.*, vol. 4, no. 2, pp. 36–49, 2024, doi: 10.69987/JACS.2024.40204.
7. P. Liu et al., "Deep flow collaborative network for online visual tracking," in *ICASSP 2020 - IEEE Int. Conf. Acoust., Speech Signal Process.*, IEEE, 2020, doi: 10.1109/ICASSP40776.2020.9054590.
8. M. Li, W. Liu, and C. Chen, "Adaptive Financial Literacy Enhancement through Cloud-Based AI Content Delivery: Effectiveness and Engagement Metrics," *Ann. Appl. Sci.*, vol. 5, no. 1, 2024.
9. Y. Zhao et al., "Unit operation combination and flow distribution scheme of water pump station system based on Genetic Algorithm," *Appl. Sci.*, vol. 13, no. 21, p. 11869, 2023, doi: 10.3390/app132111869.
10. Z. Wang et al., "Scientific formula retrieval via tree embeddings," in *2021 IEEE Int. Conf. Big Data (Big Data)*, IEEE, 2021, doi: 10.1109/BigData52589.2021.9671942.

11. M. Sun, Z. Feng, and P. Li, "Real-Time AI-Driven Attribution Modeling for Dynamic Budget Allocation in US E-Commerce: A Small Appliance Sector Analysis," *J. Adv. Comput. Syst.*, vol. 3, no. 9, pp. 39–53, 2023, doi: 10.69987/JACS.2023.30904.
12. J. Fan, T. K. Trinh, and H. Zhang, "Deep Learning-Based Transfer Pricing Anomaly Detection and Risk Alert System for Pharmaceutical Companies: A Data Security-Oriented Approach," *J. Adv. Comput. Syst.*, vol. 4, no. 2, pp. 1–14, 2024, doi: 10.69987/JACS.2024.40201.
13. R. Chand et al., "Survey on Visual Speech Recognition using Deep Learning Techniques," in *2023 Int. Conf. Commun. Syst., Comput. IT Appl. (CSCITA)*, IEEE, 2023, doi: 10.1109/CSCITA55725.2023.10104811.
14. Q. Liu et al., "Multimodal recommender systems: A survey," *ACM Comput. Surv.*, vol. 57, no. 2, pp. 1–17, 2024, doi: 10.1145/3695461.
15. C. Ju et al., "AI-Driven Vulnerability Assessment and Early Warning Mechanism for Semiconductor Supply Chain Resilience," *Ann. Appl. Sci.*, vol. 5, no. 1, 2024.
16. Z. Wang et al., "Temporal Evolution of Sentiment in Earnings Calls and Its Relationship with Financial Performance," *Appl. Comput. Eng.*, vol. 141, pp. 195–206, 2025. ISBN: 9781835589977.
17. M. Smalenberger et al., "Automatic Short Answer Grading in College Mathematics Using In-Context Meta-learning: An Evaluation of the Transferability of Findings," in *Int. Conf. Artif. Intell. Educ.*, Cham: Springer Nature Switzerland, 2024, doi: 10.1007/978-3-031-64315-6_38.
18. M. Zhang et al., "Interpretable math word problem solution generation via step-by-step planning," arXiv preprint arXiv:2306.00784, 2023.
19. Z. Wang, X. Wang, and H. Wang, "Temporal graph neural networks for money laundering detection in cross-border transactions," *Acad. Nexus J.*, vol. 3, no. 2, 2024.
20. G. Rao, Z. Wang, and J. Liang, "Reinforcement Learning for Pattern Recognition in Cross-Border Financial Transaction Anomalies: A Behavioral Economics Approach to AML," *Appl. Comput. Eng.*, vol. 142, pp. 116–127, 2025. ISBN: 9781835589991.
21. J. Chen and Z. Lv, "Graph Neural Networks for Critical Path Prediction and Optimization in High-Performance ASIC Design: A ML-Driven Physical Implementation Approach," in *Global Conf. Adv. Sci. Technol.*, vol. 1, no. 1, 2025.
22. S. Chen et al., "CARES: Comprehensive Evaluation of Safety and Adversarial Robustness in Medical LLMs," arXiv preprint arXiv:2505.11413, 2025.
23. J. Liang et al., "Anomaly Detection in Tax Filing Documents Using Natural Language Processing Techniques," *Appl. Comput. Eng.*, vol. 144, pp. 80–89, 2025. ISBN: 9781805900214.
24. S. Zhang, C. Zhu, and J. Xin, "CloudScale: A Lightweight AI Framework for Predictive Supply Chain Risk Management in Small and Medium Manufacturing Enterprises," *Spectrum Res.*, vol. 4, no. 2, 2024.
25. M. Zhang et al., "Automatic short math answer grading via in-context meta-learning," arXiv preprint arXiv:2205.15219, 2022.
26. C. Ni et al., "Contrastive Time-Series Visualization Techniques for Enhancing AI Model Interpretability in Financial Risk Assessment," 2025, doi: 10.20944/preprints202504.1984.v1.
27. G. Rao et al., "Jump prediction in systemically important financial institutions' CDS prices," *Spectrum Res.*, vol. 4, no. 2, 2024.
28. A. Kang, J. Xin, and X. Ma, "Anomalous cross-border capital flow patterns and their implications for national economic security: An empirical analysis," *J. Adv. Comput. Syst.*, vol. 4, no. 5, pp. 42–54, 2024, doi: 10.69987/JACS.2024.40504.
29. Y. Chen, C. Ni, and H. Wang, "AdaptiveGenBackend A Scalable Architecture for Low-Latency Generative AI Video Processing in Content Creation Platforms," *Ann. Appl. Sci.*, vol. 5, no. 1, 2024.