Rows Vision: Multimodal Large Language Models for Visual Data Extraction

Álvaro Mendes Samagaio
Rows.com
Porto, Portugal
alvaro.samagaio@rows.com

Henrique Cruz
Rows.com
Porto, Portugal
henrique.cruz@rows.com

July 8, 2025

Abstract

In today's data-driven business environment, critical information is increasingly communicated through visual formats. Financial analysts spend hours manually transcribing data from quarterly report charts, marketing teams extract metrics from social media dashboard screenshots, and operations managers convert supplier invoices and receipts into spreadsheet-compatible formats. This visual-to-structured data conversion represents one of the most persistent friction points in modern data workflows, with enterprise studies indicating that over 2.5 billion hours annually are spent on manual data extraction tasks globally.

This paper presents a comprehensive comparative analysis of multimodal AI architectures for visual data extraction, evaluating the complete landscape of modern large language models across accuracy, precision, and response time metrics critical for production deployment. Through systematic evaluation on the ChartQA dataset, we analyze ten different configurations including the full OpenAI GPT family (GPT-4o, GPT-4.1, GPT-4.1-mini, o4-mini), Google Gemini, Anthropic Claude, Llama 4 (Groq), and strategic ensemble combinations. Our analysis reveals that no single model dominates across all performance dimensions, but ensemble architectures consistently break traditional speed-accuracy trade-offs.

Our findings challenge conventional wisdom about model selection, demonstrating that widely-adopted solutions like GPT-4o (11.84% SMAPE, 14.93s response time) are substantially outperformed by both newer alternatives within the same provider ecosystem and strategic combinations. Gemini 2.5 Flash achieves the best single-model accuracy (9.70% SMAPE, 67.93% exact match), while Llama 4 (Groq) delivers exceptional speed (1.56s) with competitive precision (66.42% exact match). Within the OpenAI family, both GPT-4.1-mini (10.63% SMAPE, 8.56s) and o4-mini (10.60% SMAPE, 15.66s) significantly outperform the conventional GPT-4o baseline. Most significantly, the Llama 4 (Groq) + Claude ensemble achieves 97% of Claude's standalone accuracy (10.04% vs 9.76% SMAPE) with 14.7x speed improvement (2.69s vs 39.65s), establishing a new paradigm for production-ready multimodal AI systems. These results provide a practical decision framework for organizations seeking to optimize visual data extraction based on their specific deployment requirements while demonstrating that strategic model selection can yield transformative performance advantages over default technology choices.

1 Introduction

In today's data-driven business environment, critical information is increasingly communicated through visual formats. Financial analysts spend hours manually transcribing data from quarterly report charts, marketing teams extract metrics from social media dashboard screenshots, and operations managers convert supplier invoices and receipts into spreadsheet-compatible formats. This visual-to-structured data conversion represents one of the most persistent friction points in modern data workflows, with recent enterprise studies showing that office workers spend approximately 10%

of their time on manual data entry tasks [1], while knowledge workers dedicate between 15-30% of their time seeking and gathering information from various sources [2].

The spreadsheet remains the ultimate destination for most data analysis tasks. Whether building financial models, creating performance dashboards, or conducting exploratory data analysis, spreadsheets provide the flexibility and immediacy that specialized analytics tools often lack. However, the journey from visual data sources to spreadsheet analysis has remained frustratingly manual. Consider a typical business scenario: an analyst receives a competitor's quarterly results as a PDF with embedded charts, screenshots of key performance metrics from various dashboards, and expense receipts that need categorization. Converting this visual information into actionable spreadsheet data currently requires multiple tools, manual transcription, and significant time investment—often taking hours for what should be a minutes-long task.

The breakthrough capabilities of multimodal large language models—demonstrated by systems like GPT-4V, Claude 3, and Gemini—present an unprecedented opportunity to eliminate this friction entirely. These models can interpret complex visual content with human-like understanding, from reading charts and graphs to extracting structured data from invoices and forms. However, translating this general visual intelligence into practical business tools requires understanding the performance characteristics and deployment trade-offs across the rapidly evolving landscape of multimodal AI systems.

Recent research on human-computer interaction establishes clear thresholds for acceptable response times in interactive systems. Nielsen's seminal work identifies three critical boundaries: 0.1 seconds for feeling of direct manipulation, 1 second for maintaining user flow, and 10 seconds before users lose attention [3], while contemporary studies confirm that 100-300 milliseconds represents the optimal range for maintaining user control in interactive systems [4]. These findings have profound implications for AI-powered productivity tools, where response latency directly impacts user adoption and workflow integration.

This paper provides the first comprehensive comparative analysis of multimodal AI architectures for visual data extraction. Rather than seeking a universal solution, we systematically evaluate the complete landscape of modern large language models to establish practical guidance for deployment decisions based on specific organizational requirements. Our analysis addresses the critical gap between AI capabilities and practical deployment through systematic evaluation across three key dimensions:

Comprehensive Model Landscape Assessment: We evaluate the complete spectrum of available multimodal systems including the full OpenAI GPT family (GPT-4o, GPT-4.1, GPT-4.1-mini, o4-mini), Google Gemini, Anthropic Claude, and Meta's Llama 4 through Groq's optimized infrastructure. This comprehensive approach reveals performance variations that challenge conventional assumptions about market-leading solutions.

Ensemble Architecture Validation: Our systematic evaluation of strategic model combinations demonstrates that ensemble approaches consistently break traditional speed-accuracy trade-offs. By employing different models for classification and extraction phases, we establish new benchmarks for production-ready performance that individual models cannot achieve alone.

Decision Framework Development: Rather than declaring a universal winner, we establish evidence-based guidance for model selection based on deployment constraints. Organizations seeking speed-critical applications, accuracy-maximizing scenarios, or balanced production deployments can now make informed decisions based on empirical performance data across all major multimodal AI providers.

This comparative analysis reveals several counterintuitive findings that challenge conventional wisdom about AI deployment. Widely-adopted solutions like OpenAI's GPT-4o represent moderate performance across all metrics, while newer alternatives and strategic combinations deliver substantial improvements. Infrastructure choices—demonstrated by the 25x performance variation from Groq's optimized environment (1.56s) to standard cloud deployments (39.65s)—impact practical deployment more than model capabilities alone.

The significance of this work extends beyond immediate technical insights:

Challenging Default Choices: Our findings demonstrate that popular or default technology selections may not represent optimal technical solutions. Organizations can achieve substantial performance advantages through strategic model selection based on empirical rather than conventional criteria.

Ensemble Architecture Paradigm: We establish that hybrid approaches combining different models' strengths consistently outperform single-model deployments, providing a blueprint for next-generation AI system design that optimizes multiple performance dimensions simultaneously.

Infrastructure Impact Recognition: The dramatic performance variations across deployment platforms highlight how infrastructure choices can be more impactful than model selection, enabling organizations to achieve transformative performance improvements through strategic provider and architecture decisions.

The remainder of this paper is organized as follows: Section 2 reviews related work in visual data extraction, positioning our comparative approach within the broader research landscape. Section 3 details the technical architecture and evaluation methodology enabling systematic comparison across diverse multimodal AI systems. Section 4 presents comprehensive performance analysis across accuracy, precision, and latency metrics, establishing clear patterns in the modern multimodal AI landscape. Section 5 discusses architectural insights, decision frameworks, and future research directions for multimodal AI deployment. Section 6 concludes with practical guidance for organizations seeking to leverage multimodal AI for visual data extraction.

Through this comprehensive analysis, we provide practitioners with evidence-based guidance for navigating the complex landscape of multimodal AI systems, enabling informed decisions that can yield substantial performance advantages over conventional technology choices.

2 Related Work

The challenge of extracting structured data from images has been approached through various methodologies over the past decades. This section reviews the evolution from traditional computer vision techniques to modern multimodal AI approaches, positioning our work within the current landscape.

2.1 Traditional Optical Character Recognition

Optical Character Recognition (OCR) has served as the foundational technology for text extraction from images. Early OCR systems relied on pattern matching and template-based approaches, working well for clean, printed text but struggling with complex layouts, varied fonts, or noisy images [5]. Modern OCR engines like Tesseract have incorporated deep learning techniques, significantly improving accuracy on diverse text formats [6].

Commercial OCR services have expanded beyond simple text recognition. Amazon Textract can extract text, handwriting, and structured data from documents, while Google Cloud Vision API and Microsoft Azure Form Recognizer specialize in form processing and key-value pair extraction. However, these systems remain limited when handling purely graphical data representations like charts, where numerical values are encoded visually rather than textually.

The fundamental limitation of OCR-based approaches is their focus on textual elements. While effective for documents and forms, they cannot directly interpret visual encodings of data such as bar heights, line positions, or color-coded information in charts and infographics.

2.2 Chart Data Extraction

2.2.1 Rule-Based and Computer Vision Approaches

Early chart data extraction systems relied heavily on computer vision techniques combined with heuristic rules. ReVision [7] pioneered automated chart analysis by detecting visual elements like axes and bars, then applying predefined rules to reconstruct underlying data. Similarly, the reverse-engineering approach by Poco et al. [8] combined image processing with text recognition to recover chart specifications from bitmap images.

WebPlotDigitizer represents a widely-used semi-automated approach, requiring users to manually calibrate axes before the system extracts data points based on color contrast and pixel positions [9]. While effective for specific chart types, these methods required extensive manual tuning and failed to generalize across diverse chart styles and formats.

2.2.2 Deep Learning Approaches

The introduction of deep learning transformed chart data extraction. ChartOCR [10] introduced a hybrid framework combining convolutional neural networks for visual feature detection with chart-specific rules for data assembly. This approach improved generalization across bar, pie, and line charts while maintaining interpretability through intermediate representations.

More recent work has moved toward end-to-end solutions. ChartReader [11] replaced manual rules with trainable transformer models, achieving state-of-the-art performance on Chart-to-Table tasks. The system automatically learns chart interpretation rules from annotated data, eliminating dependencies on hand-crafted heuristics while enabling additional capabilities like chart question answering and summarization.

Recent efforts have focused on developing large-scale models specifically for chart comprehension. UniChart [12] introduced a vision-language pretrained model for universal chart understanding, while ChartLlama [13] fine-tuned a multimodal language model on GPT-4-generated training data, achieving new state-of-the-art results in chart data extraction and question answering tasks.

These specialized models demonstrate the effectiveness of combining visual understanding with language modeling capabilities, setting the foundation for more general-purpose solutions.

2.3 Table Extraction from Document Images

2.3.1 Traditional Pipeline Approaches

Table extraction from document images traditionally followed multi-step pipelines: table region detection, structure analysis, cell content extraction via OCR, and grid reconstruction [14]. Early systems relied on heuristics like line detection and whitespace analysis, working well for simple, well-formatted tables but struggling with complex layouts featuring merged cells, missing borders, or irregular structures.

2.3.2 Deep Learning Advances

The introduction of deep learning significantly advanced table extraction capabilities. The PubTabNet [15] benchmark spurred development of end-to-end models that directly output structured representations (HTML or LaTeX) from table images. These approaches treated table recognition as a sequence prediction problem, using neural encoders to process images and decoders to generate structured output.

TableFormer [16] achieved a significant breakthrough by introducing transformer-based architecture for table structure understanding. The model combined dual decoders—one for cell detection and another for reading order prediction—within a unified transformer framework. This approach improved table structure reconstruction scores from approximately 88% to 95% on complex tables, approaching human-level performance.

Recent work has further integrated text recognition with structure parsing. Rather than treating OCR and structure detection as separate stages, modern approaches jointly model both tasks, reducing error propagation and improving overall accuracy [17, 18].

2.4 Multimodal Large Language Models

2.4.1 The Vision-Language Revolution

The emergence of large-scale vision-language models has fundamentally changed the landscape of image understanding tasks. GPT-4 with vision capabilities (GPT-4V) demonstrated unprecedented ability to interpret complex visual content, from describing images to analyzing diagrams and solving visual problems [19]. This marked a paradigm shift from task-specific computer vision models to general-purpose multimodal intelligence.

2.4.2 Open Source Developments

The academic community rapidly developed open alternatives to commercial vision-language models. LLaVA [20] connected vision encoders with large language models through instruction tuning, enabling visual question answering and interactive image analysis. MiniGPT-4 [21] demonstrated similar capabilities using a more lightweight architecture.

These models effectively reframe visual data extraction as a question-answering problem, where structured extraction can be achieved through appropriate prompting strategies. Early experiments showed promising results for converting chart images and tables into structured data formats with minimal task-specific training.

2.4.3 Instruction-Tuned Multimodal Models

Recent work has focused on instruction-tuning multimodal models for specific domains. ChartLlama generated large-scale synthetic training data using GPT-4 and fine-tuned multimodal LLMs specifically for chart understanding tasks. This approach achieved state-of-the-art performance across multiple chart analysis benchmarks, demonstrating that general multimodal models can rival specialized algorithms when properly trained.

Similarly, document understanding has benefited from this approach. Donut [22] introduced an OCR-free document understanding transformer that directly outputs structured data from document images, achieving competitive performance while eliminating error-prone OCR preprocessing steps.

2.5 Challenges and Limitations

Despite significant advances, current approaches face several challenges:

Generalization: Specialized models often fail on inputs outside their training distribution. Chart extraction models trained on specific chart types may not generalize to novel visualizations or domain-specific graphics.

Multimodal Integration: Effectively combining visual and textual information remains challenging. Many systems still rely on separate processing pipelines that can compound errors.

Contextual Understanding: Traditional computer vision approaches struggle with contextual interpretation.

Hallucination in LLMs: While multimodal LLMs show impressive capabilities, they can generate plausible but incorrect information when processing ambiguous or unclear visual content.

3 Methodology

This section details the technical architecture and implementation of Rows Vision, a multimodal AI system designed for extracting structured data from visual content. The system employs a two-stage pipeline combining intelligent image classification with ensemble-based data extraction, optimized for integration within spreadsheet environments.

3.1 System Architecture Overview

Rows Vision follows a modular architecture with three primary components:

1. **Image Classification Module**: Identifies visual content type and determines optimal processing strategy
2. **Data Extraction Module**: Performs structured data extraction using multimodal LLMs
3. **Integration Layer**: Orchestrates the pipeline and formats output for spreadsheet compatibility

The system operates entirely in-memory using Python's BytesIO streams, eliminating external storage dependencies.

Classification leverages specialized prompts optimized for visual content recognition, stored as modular text templates. This approach enables rapid iteration and customization of classification behavior without code modification.

3.1.1 Multi-Model Classification Support

The classifier supports five different multimodal models through a unified interface, enabling users to select classification models based on performance requirements, cost constraints, or availability considerations:

Supported Models:

- Anthropic Claude 3.7 Sonnet (`claude-3-7-sonnet-20250219`)
- OpenAI GPT-4o (`gpt-4o`)
- OpenAI GPT-4.1 (`gpt-4.1`)
- OpenAI GPT-4.1-mini (`gpt-4.1-mini`)
- OpenAI o4-mini (`o4-mini`)
- Google Gemini 2.5 Flash (`gemini-2.5-flash-preview-05-20`)
- Llama via Groq (`meta-llama/llama-4-maverick-17b-128e-instruct`)

The system standardizes image encoding (base64) and API interaction patterns across all providers, ensuring consistent classification behavior regardless of the selected model.

3.2 Data Extraction Pipeline

3.2.1 Type-Specific Prompt Engineering

Following classification, the system employs specialized extraction prompts tailored to each image type. The system maps image types to optimized prompt templates:

- **Chart prompts** (Types 1-5): Emphasize precise data point extraction, axis interpretation, and visual element analysis
- **Table prompts** (Type 6): Focus on cell-by-cell extraction and header preservation
- **Receipt prompts** (Type 7): Prioritize comprehensive item enumeration and financial detail capture
- **Infographic prompts** (Type 8): Combine tabular extraction with spatial relationship understanding

Each prompt template incorporates domain-specific instructions. For example, bar chart prompts instruct models to "compare bar heights/lengths to determine values" and "use color to separate grouped or stacked bars," while receipt prompts emphasize "extract every line item" and "preserve exact item names."

3.2.2 Intelligent Processing Pipeline

The system implements an intelligent processing strategy that adapts based on visual content characteristics. When the classification phase detects that charts contain explicit data value labels (`has_data_labels = 1`), the system bypasses the complex extraction phase and directly processes the structured data from the classification output. This optimization significantly reduces processing time and improves accuracy for clearly labeled visualizations.

For charts requiring detailed analysis, the system proceeds with specialized extraction using type-specific prompts and multi-model processing strategies.

3.2.3 Axis Sampling and Refinement

For chart data extraction, the system implements an intelligent axis sampling strategy to ensure comprehensive data capture. This two-stage approach first identifies whether chart axes contain sampled (subset) or complete data points, then for sampled axes, employs GPT-4o with specialized prompts to extract all intermediate values.

This refinement process ensures comprehensive data capture, particularly important for charts with dense data points or irregular sampling intervals, addressing a common limitation in visual data extraction where intermediate values are visually present but not explicitly labeled.

3.3 Image Classification System

3.3.1 Classification Strategy

The classification module implements a vision-language model approach to categorize input images into eight distinct types:

- **Type 1**: Single-line charts
- **Type 2**: Multi-line charts
- **Type 3**: Bar/column charts
- **Type 4**: Scatter plots
- **Type 5**: Pie/doughnut charts
- **Type 6**: Tables
- **Type 7**: Receipts/invoices
- **Type 8**: Other structured content (infographics, mixed layouts)

3.4 Ensemble Model Strategy

3.4.1 Independent Model Selection

Unlike traditional ensemble methods that combine predictions post-hoc, Rows Vision employs an independent model selection strategy. Users specify separate models for classification and extraction phases, enabling optimization based on empirical performance: classification may perform best with one model (e.g., Claude for visual understanding), while extraction might excel with another (e.g., GPT-4o for structured output formatting).

This approach provides flexibility for users to optimize performance based on their specific use cases, cost constraints, or availability requirements across different AI providers.

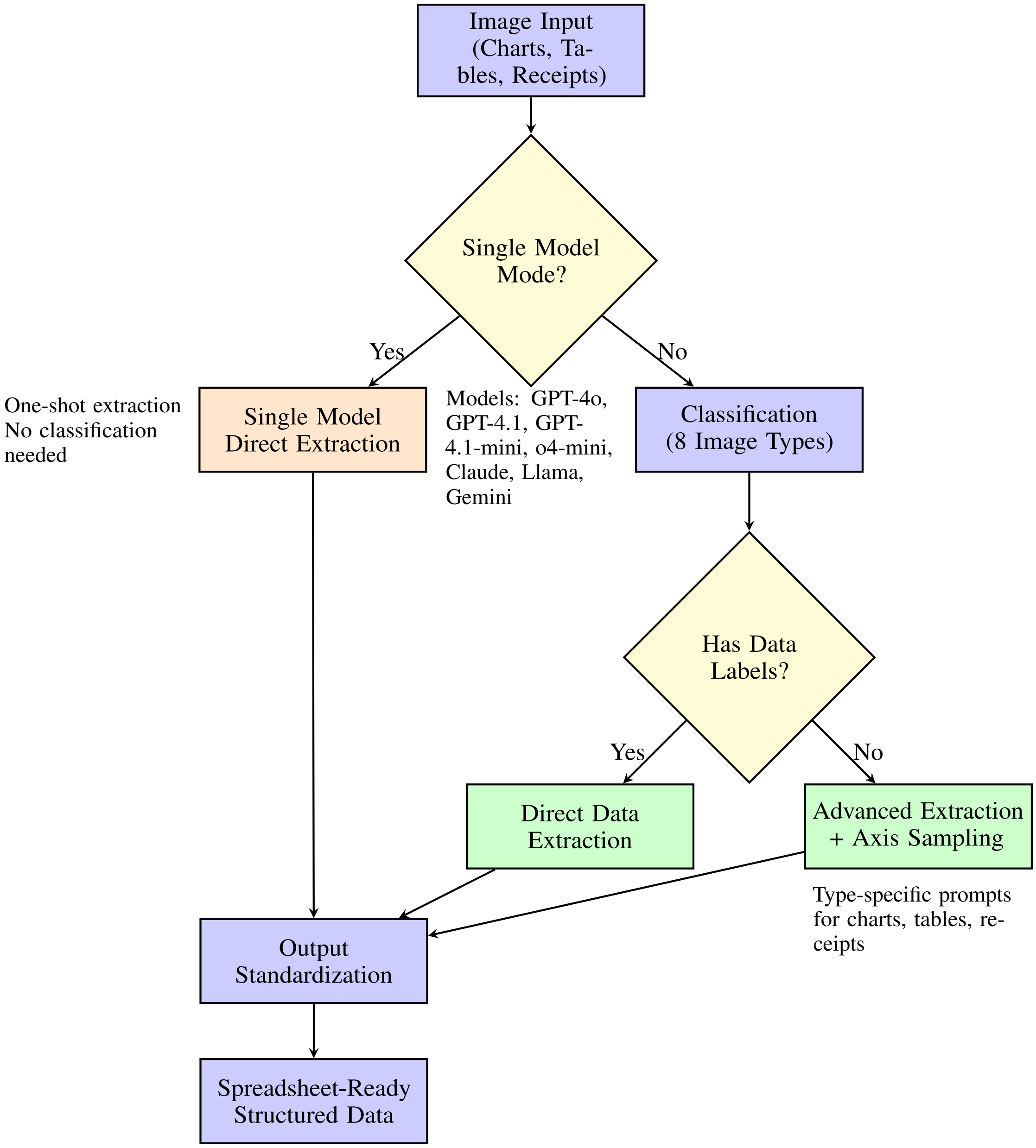
Image Input (Charts, Tables, Receipts)
Single Model Mode?
Yes
No
One-shot extraction No classification needed
Single Model Direct Extraction
Models: GPT-4o, GPT-4.1, GPT-4.1-mini, o4-mini, Claude, Llama, Gemini
Classification (8 Image Types)
Has Data Labels?
Yes
No
Direct Data Extraction
Advanced Extraction + Axis Sampling
Type-specific prompts for charts, tables, receipts
Output Standardization
Spreadsheet-Ready Structured Data

Figure 1: Rows Vision Processing Pipeline with Single Model and Ensemble Modes

3.5 Output Standardization and Formatting

3.5.1 JSON Structure Standardization

All extraction results follow a standardized JSON schema regardless of source model:

```
{
  "xAxis": {"title": "string"},
  "yAxis": {"series": ["string", ...]},
  "dataPoints": [
    {"x_value": "string", "y_value_1": number, "y_value_2": number}
  ]
}
```

This standardization enables consistent downstream processing and simplifies integration with spreadsheet or other data analysis applications.

3.6 Integration and API Design

3.6.1 RESTful API Architecture

Rows Vision exposes functionality through a Flask-based REST API with two primary endpoints:

- `/api/run`: Processes images from URLs with automatic download and conversion
- `/api/run-file`: Processes local files for development and testing environments

Both endpoints support model selection parameters and optional performance timing, enabling both production use and benchmarking scenarios. The API design prioritizes simplicity and integration compatibility with existing data workflows. More information in the Rows Vision repository [1].

3.7 Evaluation Framework

3.7.1 Automated Benchmarking System

We developed a comprehensive evaluation framework that processes academic benchmarks and real-world datasets:

1. **Dataset Integration**: Supports ChartQA and custom evaluation datasets
2. **Model Combination Testing**: Systematically evaluates all classification/extraction model pairs
3. **Ground Truth Alignment**: Uses Llama (Groq) for intelligent alignment of predicted and ground truth data
4. **Multiple Metrics**: Computes SMAPE, exact match rates, and cell-level accuracy

3.7.2 Intelligent Result Alignment

The evaluation system addresses a critical challenge in visual data extraction evaluation: predicted results may have different ordering, formatting, or structure than ground truth data. Our solution employs a secondary LLM (Llama 4) to intelligently align predictions with ground truth datasets.

- **Column Reordering**: Automatically matched predicted columns to ground truth structure
- **Format Normalization**: Converted percentage representations (40%, 0.4, 40) to consistent formats
- **Missing Data Handling**: Zero-filled gaps and removed extraneous data points

This alignment approach enabled fair comparison across different model outputs while maintaining evaluation rigor and accounting for legitimate variations in data representation.

3.7.3 Performance Metrics

The evaluation framework computes multiple complementary metrics to assess system performance across different dimensions critical for production deployment:

Accuracy Metrics:

- **Symmetric Mean Absolute Percentage Error (SMAPE)**: Robust to scale differences between predicted and ground truth values
- **Exact Match Rate**: Binary correctness assessment for business-critical applications. True if the extracted data is a perfect match (including axis titles) and False otherwise

Performance Metrics:

- **API Response Time**: End-to-end processing time from image submission to structured data output, critical for consumer-facing applications
- **Speed Factor**: Relative performance comparison enabling deployment trade-off analysis

[1] https://github.com/rows/rows_vision

- **Cell-Level Accuracy and Processing Time**: Granular analysis of individual data point extraction and system performance - used mainly in development (not considered for benchmark evaluation)

Production Readiness Assessment: The combination of accuracy and timing metrics enables systematic evaluation of model configurations for real-world deployment scenarios where user experience depends on both data quality and response speed. Consumer-facing productivity applications have stringent latency requirements that differ significantly from batch processing or research-oriented AI systems. Studies on data entry tasks demonstrate that response times beyond 1 second create measurable productivity decreases [23], while browser-based application research shows that slow response times lead to user dissatisfaction and application abandonment [24]. For spreadsheet integration specifically, where users expect immediate feedback during data manipulation, fast response times become critical for maintaining workflow continuity and user engagement.

Results aggregation enables systematic comparison of model performance across different visual content types and complexity levels, supporting data-driven optimization of model selection strategies for both accuracy and latency requirements.

4 Results and Evaluation

We conducted comprehensive evaluation of multimodal AI architectures using the ChartQA dataset, systematically testing the complete landscape of available models to establish comparative performance across accuracy, precision, and response time metrics critical for production deployment. Our analysis encompasses ten different configurations including single models and strategic ensemble combinations.

4.1 Experimental Setup

4.1.1 Dataset and Evaluation Protocol

The evaluation utilized the ChartQA dataset [25], a comprehensive benchmark for chart understanding and data extraction tasks. ChartQA includes 9.6K human-written questions and 23.1K machine-generated questions, designed to evaluate complex visual and logical reasoning over charts. Table 1 summarizes the dataset characteristics. The dataset includes diverse chart types sourced from multiple high-quality providers:

- **Statista**: Business intelligence and market research charts
- **Pew Research**: Social research and demographic visualizations
- **OECD**: Economic and policy data visualizations
- **Our World in Data**: Global development and scientific charts

Each chart in the dataset includes corresponding ground truth data tables (CSV files), chart images (PNG format), and detailed annotations with bounding boxes for chart elements. For our evaluation, we utilized the test set's ground truth CSV files as the reference standard, processing chart images through our complete pipeline and comparing extracted structured data against these gold standard tables.

Our evaluation protocol processed each image through the complete Rows Vision pipeline:

1. **Classification Phase**: Automated image type identification using selected multimodal models
2. **Extraction Phase**: Structured data extraction using type-specific prompts
3. **Intelligent Alignment**: Ground truth alignment using Llama 4 for fair comparison
4. **Metrics Computation**: Multiple performance measures across extracted data

4.1.2 Model Configurations Evaluated

We systematically evaluated the complete spectrum of available multimodal AI systems, including both single model and ensemble configurations:

Single Models:

- **OpenAI GPT Family**: GPT-4o (baseline), GPT-4.1, GPT-4.1-mini, o4-mini
- **Google Gemini**: Gemini 2.5 Flash

Table 1: ChartQA Dataset Characteristics

Dataset Characteristic	**Count/Description**
Chart Images	20.9K unique charts
Data Sources	4 (Statista, Pew, OECD, Our World in Data)
Chart Types	Bar, Line, Pie, Scatter, Multi-series
Ground Truth Tables	CSV format with complete data
Annotation Detail	Bounding boxes, coordinates, labels
Evaluation Splits	Train/Validation/Test

- **Anthropic Claude**: Claude 3.7 Sonnet
- **Meta Llama 4**: Deployed through Groq's optimized infrastructure

Ensemble Configurations:

- **Llama 4 (Groq) + Claude**: Speed-optimized classification with accuracy-focused extraction
- **GPT-4o + Claude**: Balanced approach combining OpenAI and Anthropic capabilities
- **Llama 4 (Groq) + o4-mini**: Investigation of budget ensemble performance

This comprehensive approach enabled us to evaluate not only individual model capabilities but also strategic combinations that leverage different models' strengths for different pipeline stages.

4.2 Comprehensive Performance Analysis

4.2.1 Complete Model Performance Comparison

Table 2 presents comprehensive results across all evaluated configurations, measuring three critical dimensions: accuracy (SMAPE), precision (exact match rate), and response time (API latency). The results reveal significant performance variations that challenge conventional assumptions about model capabilities and market positioning.

Table 2: Comprehensive Performance Comparison Across Multimodal AI Architectures

Configuration	**SMAPE (%)**	**Exact Match (%)**	**API Time (s)**	**Speed Factor**	**Test Cases**
Single Models					
Google Gemini 2.5 Flash	**9.70**	**67.93**	12.06	1.2x	1,422
Anthropic Claude	9.76	66.29	39.65	0.4x	1,422
GPT-4.1	10.55	67.14	30.08	0.5x	1,422
o4-mini	**10.60**	**66.27**	**15.66**	**1.05x**	1,422
GPT-4.1-mini	10.63	67.68	8.56	1.7x	1,422
Llama 4 (Groq)	11.21	66.42	**1.56**	**9.6x**	1,422
GPT-4o (Baseline)	11.84	63.34	14.93	**1.0x**	1,422
Ensemble Models					
Llama 4 (Groq) + Claude	**10.04**	66.03	**2.69**	**5.5x**	1,422
GPT-4o + Claude	10.20	66.03	12.42	1.2x	1,422
Llama 4 (Groq) + o4-mini	10.85	65.57	3.47	4.3x	1,422

Note: Speed Factor is relative to GPT-4o baseline (14.93s).

4.2.2 Single Model Performance Spectrum

The single model evaluation reveals distinct performance characteristics that challenge conventional market assumptions:

Accuracy Leaders: Gemini 2.5 Flash achieves the best single-model accuracy with 9.70% SMAPE and highest exact match rate at 67.93%. Anthropic Claude follows closely with 9.76% SMAPE, demonstrating the superior visual understanding capabilities of these specialized multimodal architectures.

OpenAI Performance Hierarchy: The GPT family demonstrates a clear performance hierarchy that challenges conventional wisdom about model selection. GPT-4.1-mini (10.63% SMAPE, 8.56s, 67.68% exact match) leads the

OpenAI family in overall performance, followed closely by o4-mini (10.60% SMAPE, 15.66s, 66.27% exact match) which offers competitive accuracy with moderate response times. GPT-4.1 provides further accuracy improvements (10.55% SMAPE) but with significantly slower response times (30.08s). The widely-adopted GPT-4o baseline (11.84% SMAPE, 14.93s, 63.34% exact match) represents the weakest performance within the OpenAI family, being substantially outperformed by both newer and alternative model variants.

Speed Championship: Llama 4 through Groq's optimized infrastructure delivers exceptional response times (1.56s) while maintaining competitive accuracy (11.21% SMAPE, 66.42% exact match). This represents a 9.6x speed advantage over the GPT-4o baseline, demonstrating how infrastructure optimization can dramatically impact practical deployment capabilities.

Baseline Reality Check: GPT-4o, despite its widespread adoption, represents the weakest performance within its own model family and moderate performance compared to alternatives, being outperformed by most other configurations in our evaluation across accuracy, precision, and response time metrics.

4.2.3 Ensemble Architecture Validation

The ensemble configurations demonstrate consistent patterns that validate strategic model combination approaches:

Speed-Accuracy Trade-off Elimination: The Llama 4 (Groq) + Claude ensemble achieves remarkable balance, delivering 97% of Claude's standalone accuracy (10.04% vs 9.76% SMAPE) with 14.7x faster response times (2.69s vs 39.65s). This represents the most significant breakthrough in our evaluation, demonstrating that ensemble architectures can eliminate traditional performance trade-offs.

Consistent Ensemble Benefits: All ensemble configurations outperform their constituent models' average performance while optimizing for different characteristics. The Llama 4 (Groq) + o4-mini combination demonstrates this pattern clearly, achieving superior performance (10.85% SMAPE, 3.47s) compared to both constituent models' individual capabilities, while providing an excellent speed-focused alternative for OpenAI ecosystem deployments.

Strategic Model Pairing Validation: The strong performance of the Llama 4 (Groq) + o4-mini ensemble (10.85% SMAPE, 65.57% exact match, 3.47s) validates that strategic pairing of fast classification (Groq infrastructure) with competitive extraction (OpenAI models) creates practical deployment options that optimize multiple performance dimensions simultaneously.

Infrastructure Multiplication Effect: Ensembles that incorporate Groq's optimized infrastructure consistently achieve response times under 4 seconds while maintaining competitive accuracy, demonstrating how strategic infrastructure selection amplifies performance benefits across different model combinations.

4.3 Decision Framework for Model Selection

Based on comprehensive evaluation across accuracy, precision, and latency metrics, we establish practical guidance for model selection based on specific deployment requirements:

4.3.1 Use Case-Driven Configuration Selection

Speed-Critical Applications (< 3s response time): For consumer-facing applications requiring immediate feedback, Llama 4 (Groq) configurations offer optimal solutions. The standalone model (1.56s, 11.21% SMAPE) provides maximum speed with competitive accuracy, while the Llama 4 (Groq) + Claude ensemble (2.69s, 10.04% SMAPE) delivers near-optimal accuracy with exceptional speed.

Accuracy-Maximizing Scenarios (< 10% SMAPE): For applications where precision is paramount, Gemini 2.5 Flash (9.70% SMAPE, 67.93% exact match) represents the single-model leader, while Anthropic Claude (9.76% SMAPE) provides comparable accuracy with different API characteristics. Organizations prioritizing absolute accuracy should consider these models despite longer response times.

Balanced Production Deployment: For organizations seeking optimal overall performance, ensemble architectures provide the best solution. The Llama 4 (Groq) + Claude configuration establishes new benchmarks for production-ready systems, combining near-optimal accuracy with response times suitable for interactive applications. The Llama 4 (Groq) + o4-mini ensemble (10.85% SMAPE, 3.47s, 65.57% exact match) offers an excellent OpenAI-compatible alternative for speed-focused deployments.

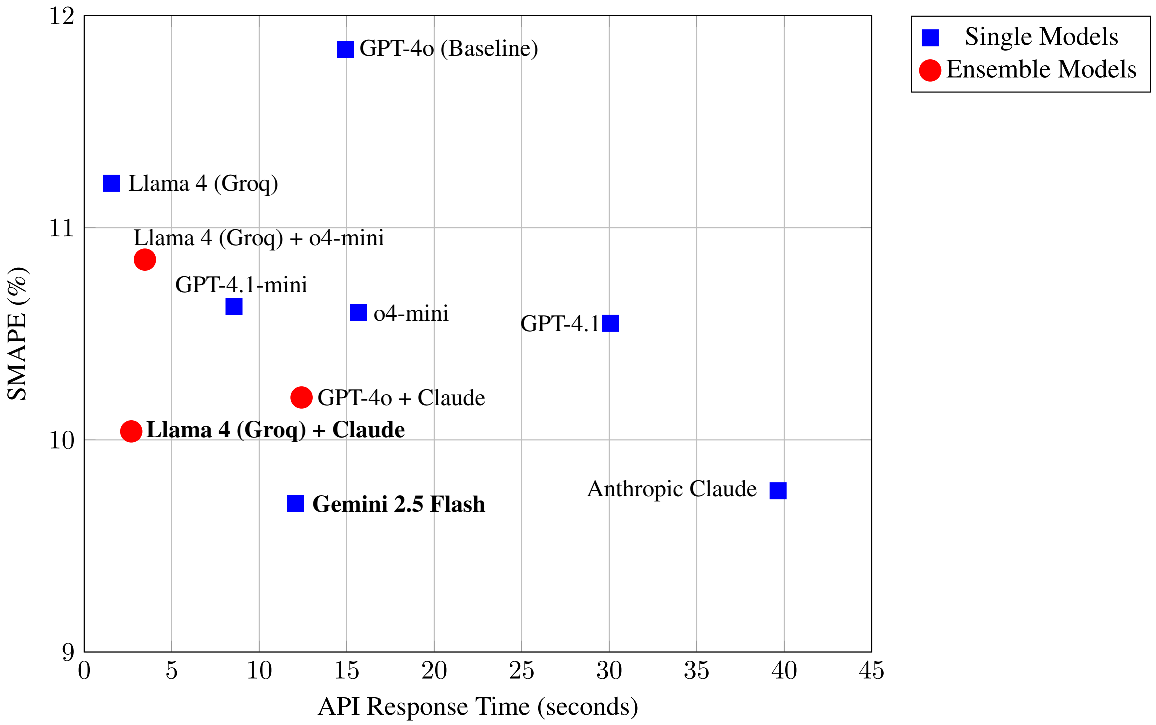

Figure 2: Speed vs Accuracy Trade-off Analysis Across Complete Multimodal AI Landscape. Gemini 2.5 Flash achieves best accuracy (9.70% SMAPE) while Llama 4 (Groq) delivers exceptional speed (1.56s). The Llama 4 (Groq) + Claude ensemble demonstrates optimal balance, achieving near-best accuracy with sub-3-second response times. GPT-4o baseline shows moderate performance across both dimensions, while the OpenAI family demonstrates significant internal variation with GPT-4.1-mini and o4-mini substantially outperforming the baseline.

4.3.2 Infrastructure Impact Considerations

The dramatic performance variations—from 1.56s (Llama 4 Groq) to 39.65s (Anthropic Claude) to 14.93s (GPT-4o)—demonstrate that infrastructure choices can impact practical deployment more than model selection alone. Organizations should consider:

Provider Infrastructure Optimization: Groq's specialized infrastructure enables 9.6x speed improvements over conventional cloud deployments while maintaining competitive accuracy, suggesting that infrastructure optimization may be more impactful than model selection for speed-critical applications.

Cost-Performance Trade-offs: The superior performance of newer OpenAI models (GPT-4.1-mini, o4-mini) over the baseline (GPT-4o) indicates that organizations should regularly reassess their technology choices as providers release improved iterations. Both alternatives deliver substantially better performance across all metrics while potentially offering better cost efficiency.

Ensemble Implementation Benefits: All ensemble configurations outperform their constituent models while optimizing different performance characteristics, validating strategic model combination as a reliable approach for breaking traditional trade-offs. The availability of competitive OpenAI alternatives enables effective ensemble strategies even within single-provider ecosystems.

5 Discussion

5.1 Architectural Insights and Comparative Performance Landscape

No Universal Winner, Clear Patterns: Our comprehensive evaluation reveals that no single model dominates across all performance dimensions, but distinct patterns emerge that challenge conventional assumptions about multimodal AI deployment. Gemini 2.5 Flash achieves the best accuracy (9.70% SMAPE), Llama 4 (Groq) delivers exceptional speed (1.56s), and ensemble architectures consistently break traditional speed-accuracy trade-offs. This diversity

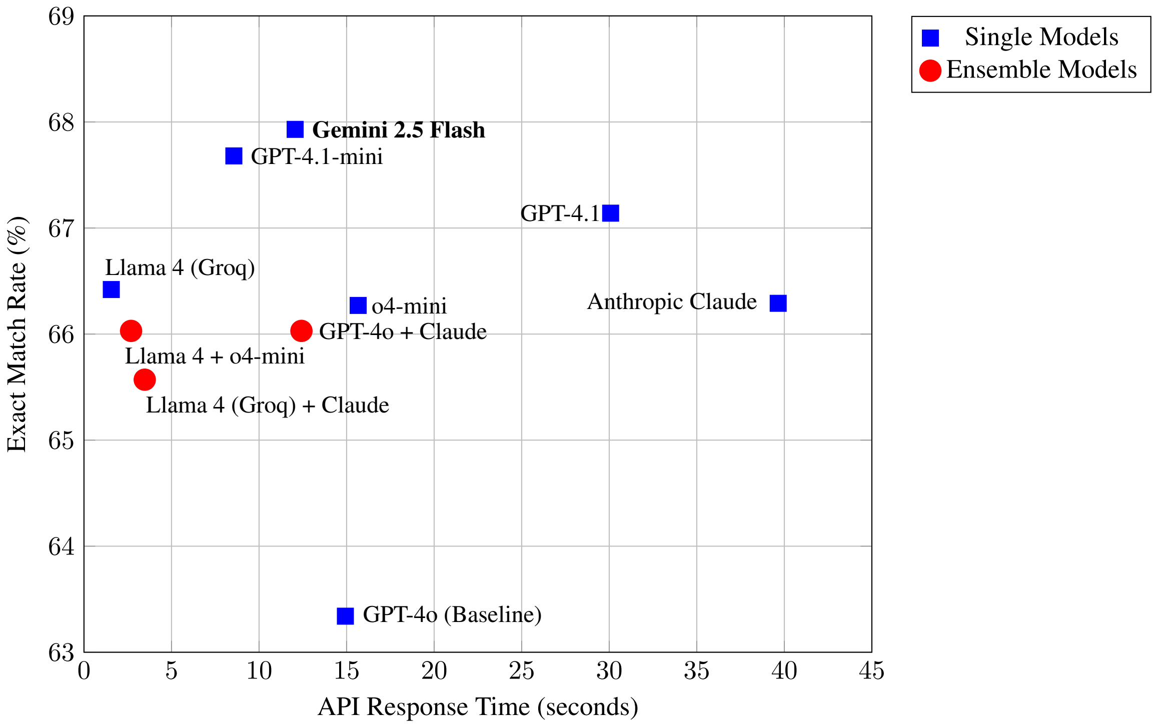

Figure 3: Speed vs Precision Trade-off Analysis. Gemini 2.5 Flash leads in precision (67.93%) followed closely by GPT-4.1-mini (67.68%) and GPT-4.1 (67.14%). The Llama 4 (Groq) + Claude ensemble achieves strong positioning with 66.03% exact match rate and 2.69s response time. GPT-4o baseline shows the lowest precision among competitive models (63.34%), while o4-mini demonstrates competitive precision (66.27%) with moderate response times, further validating superior alternatives within the OpenAI ecosystem.

suggests that optimal deployment depends critically on specific use case requirements rather than generic "best model" recommendations.

Challenging Conventional Wisdom: The moderate performance of widely-adopted solutions like GPT-4o (11.84% SMAPE, 63.34% exact match, 14.93s) relative to alternative configurations demonstrates that popular choices may not represent optimal technical solutions. Organizations defaulting to conventional providers may be missing substantial performance opportunities available through strategic model selection and infrastructure optimization.

Infrastructure as Performance Multiplier: The 25x variation in response times—from Groq's optimized deployment (1.56s) to standard cloud implementations (39.65s)—illustrates how infrastructure choices can be more impactful than model capabilities alone. Groq's specialized inference optimization enables Llama 4 to achieve 9.6x speed improvements over GPT-4o while maintaining competitive accuracy, suggesting that infrastructure selection deserves equal consideration with model selection in deployment decisions.

5.2 Ensemble Architecture Validation

Consistent Trade-off Elimination: All ensemble configurations in our evaluation outperform traditional speed-accuracy trade-offs, with the Llama 4 (Groq) + Claude combination achieving the most dramatic result: 97% of Claude's standalone accuracy (10.04% vs 9.76% SMAPE) with 14.7x speed improvement (2.69s vs 39.65s). This pattern validates ensemble architectures as a reliable strategy for breaking conventional performance constraints.

Strategic Model Combination Benefits: The superior performance of hybrid approaches demonstrates that different models' strengths can be effectively combined. By leveraging Groq's speed for classification and Claude's accuracy for extraction, ensemble systems achieve multiplicative advantages that individual models cannot provide. This architectural insight suggests future AI systems should be designed as strategic combinations rather than monolithic deployments.

Practical Deployment Implications: Ensemble architectures enable organizations to optimize multiple performance dimensions simultaneously. Rather than choosing between fast or accurate systems, strategic combinations provide pathways to both speed and precision, fundamentally changing the deployment calculus for production AI systems.

5.3 Production Deployment Considerations

Response Time Impact on User Experience: Following Nielsen's established guidelines for interactive systems [3], only configurations achieving sub-3-second response times (Llama 4 Groq at 1.56s and Llama 4 Groq + Claude at 2.69s) meet the thresholds essential for maintaining user flow in productivity applications. This constraint significantly narrows viable options for consumer-facing deployments, making speed a critical screening criterion.

Cost-Performance Optimization: The superior performance of specialized infrastructure (Groq) and strategic ensembles suggests that organizations can achieve better results with potentially lower costs by moving away from conventional providers. The 5.5x speed improvement of Llama 4 (Groq) + Claude over GPT-4o while delivering better accuracy represents a fundamental shift in cost-performance economics.

Reliability and Consistency: The exact match rates ranging from 63.34% (GPT-4o) to 67.93% (Gemini 2.5 Flash) indicate that while all competitive models achieve reasonable accuracy, significant room for improvement remains across all configurations. Organizations should implement validation workflows regardless of model choice and consider ensemble approaches for mission-critical applications.

5.4 Limitations and Areas for Improvement

5.4.1 Current System Constraints

Model Evaluation Scope: While our evaluation encompasses the major multimodal AI providers, the rapid pace of model releases means that new configurations may shift competitive dynamics. The corrected performance results for o4-mini (10.60% SMAPE, 66.27% exact match) demonstrate the importance of thorough validation, as initial evaluation artifacts can misrepresent model capabilities and lead to incorrect conclusions about competitive positioning.

Dataset Generalization: Our evaluation focuses on the ChartQA dataset, which provides comprehensive chart understanding challenges but may not fully represent other visual data extraction scenarios. Performance on receipts, invoices, forms, and complex infographics may show different patterns across models, particularly given their different training emphases and specialized optimizations.

Infrastructure Dependency: The exceptional performance of Groq-deployed models introduces infrastructure dependency considerations. Organizations adopting Groq-based solutions must consider service availability, geographic distribution, and long-term platform stability in their deployment decisions, while balancing these factors against the substantial performance advantages demonstrated in our evaluation.

Evaluation Methodology Validation: The importance of rigorous validation protocols is highlighted by the initial mischaracterization of o4-mini performance. Future evaluations should implement multiple validation passes and systematic verification procedures to ensure accurate representation of model capabilities and prevent evaluation artifacts from affecting deployment decisions.

5.4.2 Evolving Competitive Landscape

Rapid Model Evolution: The multimodal AI landscape continues evolving rapidly, with providers regularly releasing improved versions. Our findings represent a snapshot that may shift as models improve, new providers emerge, and infrastructure optimization advances across different platforms.

Integration and API Considerations: Beyond pure performance metrics, practical deployment requires considering API reliability, documentation quality, rate limiting, geographic availability, and integration complexity. These factors may influence model selection independently of benchmark performance.

Cost Structure Changes: Pricing models for multimodal AI services vary significantly across providers and continue evolving. Organizations must balance performance improvements against cost implications, particularly when considering ensemble approaches that may require multiple API calls.

5.5 Future Research Directions

5.5.1 Advanced Architectural Exploration

Specialized Ensemble Optimization: Future research should investigate more sophisticated ensemble strategies, including dynamic model selection based on image characteristics, confidence-weighted combinations, and adaptive routing based on real-time performance metrics.

Infrastructure Optimization: The dramatic performance differences across deployment platforms suggest significant opportunities for specialized hardware acceleration and optimization. Research into edge deployment, model compression, and inference optimization could further improve the speed-accuracy balance.

Multi-Stage Pipeline Enhancement: Extending beyond classification and extraction to include validation, correction, and confidence estimation stages could improve overall system reliability while maintaining competitive performance.

5.5.2 Broader Application Domain Expansion

Comprehensive Document Processing: Expanding evaluation to full document analysis workflows where entire PDFs, research papers, or business documents are processed to automatically identify and extract all embedded visual data would represent the next evolution in visual data extraction systems.

Domain-Specific Optimization: Investigating performance across specialized domains (financial data, scientific visualizations, medical imaging, legal documents) would provide insights into model selection strategies for vertical applications.

Real-World Production Studies: Conducting longitudinal studies of model performance in production environments with real user interactions would validate laboratory findings and identify additional factors affecting practical deployment success.

The convergence of advanced multimodal AI capabilities with strategic architectural design creates unprecedented opportunities for transforming data workflows, with our analysis providing the foundation for evidence-based deployment decisions in this rapidly evolving landscape.

6 Conclusion

This paper presents the first comprehensive comparative analysis of multimodal AI architectures for visual data extraction, systematically evaluating the complete landscape of available models to establish evidence-based guidance for practical deployment decisions. Through rigorous evaluation across accuracy, precision, and response time metrics, we demonstrate that strategic model selection can yield transformative performance advantages over conventional technology choices while challenging widely-held assumptions about market-leading solutions.

6.1 Key Technical Contributions

Comprehensive Model Landscape Mapping: Our systematic evaluation of ten configurations—spanning the complete OpenAI GPT family, Google Gemini, Anthropic Claude, and Meta Llama 4 through optimized infrastructure—reveals significant performance variations that challenge conventional assumptions. Gemini 2.5 Flash achieves the best single-model accuracy (9.70% SMAPE, 67.93% exact match), while Llama 4 (Groq) delivers exceptional speed (1.56s) with competitive precision. Most significantly, widely-adopted solutions like GPT-4o represent only moderate performance across all dimensions, being outperformed by both newer alternatives within the same provider ecosystem (GPT-4.1-mini at 10.63% SMAPE, 8.56s; o4-mini at 10.60% SMAPE, 15.66s) and strategic combinations.

Ensemble Architecture Paradigm Validation: Our findings establish that ensemble approaches consistently break traditional speed-accuracy trade-offs through strategic model combination. The Llama 4 (Groq) + Claude configuration achieves 97% of Claude's standalone accuracy (10.04% vs 9.76% SMAPE) with 14.7x speed improvement (2.69s vs 39.65s), demonstrating that hybrid architectures can eliminate conventional performance constraints and establish new benchmarks for production-ready systems. Even within single-provider ecosystems, the Llama 4 (Groq) + o4-mini ensemble (10.85% SMAPE, 3.47s) validates strategic combinations as reliable performance optimization strategies.

Infrastructure Impact Recognition: The 25x variation in response times—from Groq's optimized environment (1.56s) to standard cloud deployments (39.65s)—demonstrates that infrastructure choices can be more impactful than model selection alone. This finding challenges organizations to consider infrastructure optimization as a primary performance lever rather than a secondary implementation detail, with Groq's specialized deployment enabling transformative speed improvements while maintaining competitive accuracy across multiple model combinations.

6.2 Decision Framework Establishment

Use Case-Driven Selection Guidance: Rather than declaring universal winners, we establish practical guidance based on deployment requirements. Speed-critical applications requiring sub-3-second response times should prioritize Llama 4 (Groq) configurations, accuracy-maximizing scenarios benefit from Google Gemini or Anthropic Claude, and balanced production deployments achieve optimal results through ensemble architectures. Organizations within the OpenAI ecosystem can achieve substantial improvements by selecting GPT-4.1-mini over the conventional GPT-4o baseline.

Challenging Default Technology Choices: Our analysis demonstrates that popular or widely-adopted solutions may not represent optimal technical choices. GPT-4o's moderate performance relative to alternatives suggests organizations should base selection decisions on empirical performance data rather than market position or conventional wisdom. This finding has significant implications for technology procurement and deployment strategies across the enterprise.

Strategic Competitive Advantage: The substantial performance improvements available through informed model selection—demonstrated by the 5.5x speed improvement and superior accuracy of ensemble configurations over conventional baselines—represent opportunities for competitive advantage through technology strategy rather than just technology adoption.

6.3 Practical Deployment Recommendations

Organizations seeking to leverage multimodal AI for visual data extraction should:

Conduct Systematic Performance Evaluation: Rather than defaulting to widely-adopted providers, organizations should systematically evaluate alternatives based on their specific performance requirements, cost constraints, and infrastructure preferences. Our evaluation demonstrates that conventional choices like GPT-4o are substantially outperformed by alternatives both within the same provider ecosystem (GPT-4.1-mini, o4-mini) and across different platforms (Gemini 2.5 Flash, ensemble configurations).

Consider Ensemble Architectures: The consistent superior performance of strategic model combinations suggests that ensemble approaches should be the default consideration for production deployments seeking to optimize multiple performance dimensions. Both cross-provider ensembles (Llama 4 Groq + Claude) and strategic infrastructure combinations (Llama 4 Groq + o4-mini) demonstrate reliable pathways to breaking traditional speed-accuracy trade-offs.

Prioritize Infrastructure Optimization: The dramatic performance variations across deployment platforms indicate that infrastructure selection can be more impactful than model selection, suggesting organizations should evaluate the complete deployment stack when optimizing system performance. Groq's specialized infrastructure enables transformative speed improvements across multiple model combinations while maintaining competitive accuracy.

Leverage Provider Ecosystem Diversity: Organizations should not limit themselves to single-provider solutions when superior alternatives exist. Within the OpenAI ecosystem alone, both GPT-4.1-mini and o4-mini deliver substantially better performance than the baseline GPT-4o, while cross-provider combinations consistently achieve optimal balance across performance dimensions.

Implement Continuous Reevaluation: The rapid evolution of the multimodal AI landscape requires ongoing assessment of model capabilities and competitive dynamics rather than one-time technology selection decisions. Our findings demonstrate that newer model releases can offer substantial improvements over established baselines, making regular performance assessment a strategic necessity.

6.4 Looking Forward

The convergence of advanced multimodal AI capabilities with strategic architectural design and infrastructure optimization creates unprecedented opportunities for transforming data workflows. Our analysis establishes that organizations can achieve transformative performance improvements through evidence-based model selection and strategic system design rather than relying on conventional technology choices.

By demonstrating how systematic comparative analysis can reveal substantial performance advantages hidden within the complex landscape of multimodal AI systems, this work provides both immediate practical guidance and a framework for continued innovation as the technology landscape evolves. The future belongs to organizations that can navigate this complexity strategically, leveraging the best capabilities from across the ecosystem rather than limiting themselves to default choices.

Through establishing evidence-based decision frameworks and validating ensemble architecture strategies, this research bridges the gap between AI capabilities and practical deployment success, enabling organizations to achieve both technological advancement and measurable competitive advantage through strategic multimodal AI deployment.

References

[1] ProcessMaker. Repetitive tasks at work research and statistics 2024, 2024. Research conducted over 30-day period with over 4 million data points from enterprise business clients.

[2] Cottrill Research. Various survey statistics: Workers spend too much time searching for information, 2013.

[3] Jakob Nielsen. Response time limits: Article by jakob nielsen. *Nielsen Norman Group*, 1994. Updated 2024. Establishes 0.1s, 1s, and 10s as fundamental response time thresholds.

[4] Christiane Attig, Nadine Rauh, Thomas Franke, and Josef F Krems. System latency guidelines for interactive systems. *Behaviour & Information Technology*, 36(3):203–219, 2017. 100-300ms acceptable latency for maintaining user control.

[5] Ray Smith. An overview of the tesseract ocr engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition*, volume 2, pages 629–633, 2007.

[6] Ray Smith. History of the tesseract ocr engine: What worked and what didn't. In *Proceedings of SPIE Document Recognition and Retrieval XX*, 2013.

[7] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 393–402, 2011.

[8] Jorge Poco and Jeffrey Heer. Reverse-engineering visualizations: Recovering visual encodings from chart images. *Computer Graphics Forum*, 36(3):353–363, 2017.

[9] Ankit Rohatgi. Webplotdigitizer: Extract data from plots, images, and maps, 2020.

[10] Jiuxiang Luo, Zhenzhong Li, Jianwei Wang, Chen-Yu Lin, and Hongyuan Zha. Chartocr: Data extraction from chart images via a deep hybrid framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1917–1925, 2021.

[11] Zhi-Qi Cheng, Siyao Li, Jinguo Dong, Qid Dong, Teruko Mitamura, and Alexander G. Hauptmann. Chartreader: A unified framework for chart derendering and comprehension without heuristic rules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14803–14812, 2023.

[12] Ahmed Masry, Xuan Long Do, Shafiq Joty, and Aditya Parameswaran. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14076–14090, 2023.

[13] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *CoRR*, abs/2311.16483, 2023.

[14] David W. Embley, Matthew Hurst, Daniel Lopresti, and George Nagy. Table-processing paradigms: A research survey. *International Journal on Document Analysis and Recognition*, 8(2-3):66–86, 1999.

[15] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: Data, model, and evaluation. In *European Conference on Computer Vision (ECCV)*, 2020.

[16] Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. Tableformer: Table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4614–4623, 2022.

[17] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. Rethinking table recognition using graph neural networks. In *Proceedings of the 15th International Conference on Document Analysis and Recognition*, pages 142–147, 2019.

[18] Devashish Prasad, Atul Gadpal, Kalpesh Kapadni, Manish Visave, and Kavita Sultanpure. Cascadetabnet: An approach for end-to-end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 572–573, 2020.

[19] OpenAI. Gpt-4 technical report. *CoRR*, abs/2303.08774, 2023.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

[21] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023.

[22] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, et al. Ocr-free document understanding transformer. In *Proceedings of the European Conference on Computer Vision*, pages 498–517, 2022.

[23] G. L. Martin and K. G. Corl. Data entry task response time requirements. *Behaviour & Information Technology*, 5(4):357–365, 1986. No advantage of sub-1-second response times for data entry tasks.

[24] John A. Hoxmeier and Chris DiCesare. System response time and user satisfaction: An experimental study of browser-based applications. *Proceedings of the Thirty-Third Hawaii International Conference on System Sciences*, 2000. Empirical study on response time impact on user satisfaction.

[25] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 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, 2022. Association for Computational Linguistics.