FinCriticalED: A Visual Benchmark for Financial Fact-Level OCR Evaluation

Yueru He*[2], Xueqing Peng*[1], Yupeng Cao[5], Yan Wang[1], Lingfei Qian[1], Haohang Li[5], Yi Han[3],
Ruoyu Xiang[4], Mingquan Lin[6], Prayag Tiwari[10], Jimin Huang[1], Guojun Xiong[7], Sophia Ananiadou[8,9]

[1]The FinAI, [2]Columbia University, [3]Georgia Institute of Technology, [4]New York University,
[5]Stevens Institute of Technology, [6]University of Minnesota, [7]Harvard University, [8] University of Manchester,
[9]Archimedes, Athena Research Center, [10]Halmstad University
* These authors contributed equally
xueqing.peng2024@gmail.com

Abstract

*We introduce FinCriticalED (**Fin**ancial **Critical** **E**rror **D**etection), a visual benchmark for evaluating OCR and vision language models on financial documents at the fact level. Financial documents contain visually dense and table heavy layouts where numerical and temporal information is tightly coupled with structure. In high stakes settings, small OCR mistakes such as sign inversion or shifted dates can lead to materially different interpretations, while traditional OCR metrics like ROUGE and edit distance capture only surface level text similarity. FinCriticalED provides 500 image-HTML pairs with expert annotated financial facts covering over seven hundred numerical and temporal facts. It introduces three key contributions. First, it establishes the first fact level evaluation benchmark for financial document understanding, shifting evaluation from lexical overlap to domain critical factual correctness. Second, all annotations are created and verified by financial experts with strict quality control over signs, magnitudes, and temporal expressions. Third, we develop an LLM-as-Judge evaluation pipeline that performs structured fact extraction and contextual verification for visually complex financial documents. We benchmark OCR systems, open source vision language models, and proprietary models on FinCriticalED. Results show that although the strongest proprietary models achieve the highest factual accuracy, substantial errors remain in visually intricate numerical and temporal contexts. Through quantitative evaluation and expert case studies, FinCriticalED provides a rigorous foundation for advancing visual factual precision in financial and other precision critical domains.*

1. Introduction

Optical character recognition (OCR) has achieved near-saturated performance on general benchmarks [18], but financial documents remain a persistent failure case for both OCR systems and multimodal large language models [28, 29]. Financial reports, regulatory filings, and earnings disclosures are numerically dense and visually complex, containing an order of magnitude more digits, tables, and structured layouts than typical scanned documents [11]. In high stakes financial settings, even a small recognition error, such as misreading a parenthesis as a positive value, shifting a decimal, or altering a reporting date, can lead to

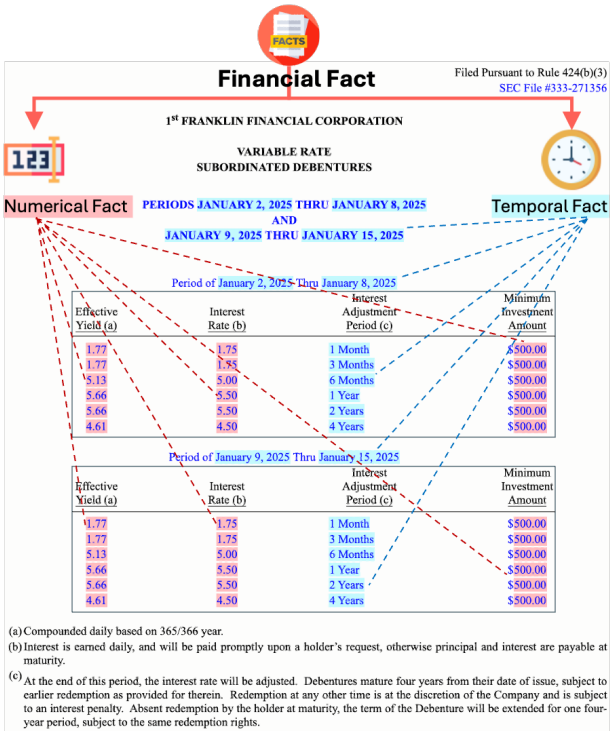

Figure 1. Illustration of financial fact extraction in **FinCriticalED**. Both Numerical Facts (e.g., yields, interest rates, minimum investment amounts) and Temporal Facts (e.g., adjustment periods) appear within the tabular sections of real SEC filings, requiring models to accurately identify and link fact values across table rows and columns.

materially different interpretations of financial statements. Such errors are rare but economically consequential, creating a precision regime that traditional OCR metrics cannot meaningfully capture.

Despite rapid progress in multimodal models, existing OCR and document understanding benchmarks evaluate only syntactic fidelity rather than factual reliability. General OCR datasets, including OCR-VQA [21] and ChartQA-X [9], assess surface-level extraction or QA accuracy through lexical metrics such as exact match, BLEU [26], ROUGE [14], or edit distance [12]. Large multi-task suites such as OCRBench [16], CC-OCR [36], and OmniDocBench [25] extend evaluation to layout analysis and bounding box detection, but they still measure transcription quality without verifying whether numerical or temporal facts remain correct. Financial-domain datasets [19, 22] instead focus on reasoning or long-context understanding and often rely on pre-extracted, error-free text, therefore bypassing OCR precision entirely.

To address this gap, we propose **Fin**ancial **C**ritical **E**rror **D**etection (FinCriticalED), a fact-centric evaluation framework for financial OCR that emphasizes factual reliability rather than textual similarity. We curate a dataset of full-page visually realistic and structurally diverse financial documents, including filings, statements, and tabular-heavy reports. Domain experts annotate financially important numerical and temporal facts directly in ground truth output, capturing subtle conventions such as parenthesized negatives and period-specific dates. This annotation process achieves high inter-annotator agreement, ensuring reliable supervision in a precision-sensitive domain. Based on the experts' annotations, we define a critical fact-level OCR task in which a model must reconstruct document structure from image and preserve all financial facts contained within it. For evaluation, we employ an LLM-as-Judge protocol that inspects factual correctness by comparing predicted output with expert-annotated evidence.

We comprehensively evaluate state-of-the-art OCR systems and multimodal vision language models on **FinCriticalED**, ranging from DeepSeek-OCR [34] to GPT-5 [24]. Our results show that conventional OCR metrics disproportionately reward non-critical character accuracy and fail to indicate whether financially meaningful facts are preserved. In contrast, fact-level evaluation reveals that even the strongest proprietary models still make errors on financially critical numerical and temporal entities, despite achieving high overall accuracy. Temporal entities exhibit moderately higher robustness than numerical fields, although factual inconsistencies remain across all model categories. These findings suggest that factual precision, rather than surface-level transcription, is the primary obstacle to achieving trustworthy financial document vision understanding.

Our contributions are threefold. (1) We introduce FinCriticalED, the first fact-level evaluation benchmark for financial OCR, focusing on the preservation of numerical and temporal financial facts. (2) We provide a domain-expert annotated dataset with high inter-annotator agreement, enabling precise and reliable evaluation. (3) We propose an LLM-as-Judge evaluation pipeline that quantifies factual correctness across numerical accuracy, temporal integrity, and evidence alignment. Together, these components establish a standardized foundation for evaluating factual reliability in financial OCR and support future research on high-precision, domain-critical document understanding.

2. Related Work

Benchmark	Financial Centric	OCR Included	Long Context	Fact Focus
OmniDocBench [25]	✗	✓	✓	✗
CC-OCR [36]	✗	✓	✓	✗
OCRBench / v2 [6]	✗	✓	✗	✗
OCR-VQA [21]	✗	✓	✗	✗
ChartQA-X [20]	✗	✗	✗	✗
CONTEXTUAL [31]	✗	✓	✓	✗
SEED-Bench-2-Plus [13]	✗	✓	✓	✗
MMMU[37] / MMMU-Pro[38]	✗	✗	✗	✗
FOX[15]	✗	✓	✓	✗
MultiFinBen [28]	✓	✓	✓	✗
FinMME [17]	✓	✗	✗	✗
MME-Finance [7]	✓	✗	✗	✗
FinCriticalED	✓	✓	✓	✓

Table 1. Comparison of multimodal OCR and financial benchmarks. ✓indicates inclusion and ✗exclusion of a property.

2.1. OCR Benchmarks

OCR Models. Recent advances in OCR have been driven by the emergence of MLLMs that replace traditional detection–recognition pipelines. DeepSeek-OCR [34], GOT-OCR 2.0[32, 33], and MinerU2.5 [23] exemplify this shift by integrating visual encoders and text decoders to perform holistic text understanding, layout interpretation, and structural reasoning. DeepSeek-OCR introduces long-context token compression for page-scale reasoning, while adopts a decoupled coarse-to-fine design for preserving high-resolution details. The PaddleOCR family [3, 4] extended OCR to 5 languages and tasks like table and formula recognition. Specialized frameworks have also emerged for domain-specific scenarios: OneChart [1] targets chart content extraction and Slow Perception [35] focuses on fine-grained visual-text fusion for structured reasoning. However, despite these modeling advances, their evaluation has remained on emphasizing general text recognition accuracy or BLEU-like scores, with no assessment of domain-specific reliability in high-stakes contexts such as finance.

Document Understanding Benchmarks. A wide range of benchmarks now evaluate document understanding, spanning general and specialized tasks. Early OCR benchmarks such as OCR-VQA [21] and ChartQA-X [9] focus on question-answering from image or charts. Recent benchmark suites such as OCRBench-v1 [16], OCRBench-v2 [6], and OmniDocBench [25] broaden coverage by including tasks like text localization, recognition, layout analysis, and table extraction, evaluated with character accuracy, edit distance, and mean average precision. CC-OCR [36] further includes multilingual setting and information extraction task. Recent comprehensive benchmarks MMMU [37] and MMMU-Pro [38] likewise span numerous OCR-related tasks and modalities. Beyond text reading, SEED-Bench-2-Plus [13] introduce over 2.3k vision-language Q&A pairs including structured diagrams and charts, and CONTEXTUAL [31] targets text-rich understanding with dense grounding and multi-hop reasoning across pages. More recently, FOX [15] enables fine-grained multi-page document tasks from region-level OCR/translation to cross-page VQA.

Despite the diversity and scale of these benchmarks, they predominantly measure textual accuracy or semantic alignment (e.g. Rouge scores, edit distances, localization precision) in reading and parsing tasks. Factual correctness for the recognized content remains unexamined.

2.2. OCR for Long Financial Documents

In the financial domain, some benchmarks have been proposed but with limited focus on OCR fidelity. For instance, FinMME [17] and MME-Finance [7] evaluate multimodal financial question answering using textual filings or cleaned documents, effectively sidestepping the OCR step and assuming perfect text input. MultiFinBen[28] as the most recent comprehensive financial benchmark also includes OCR task, but the scope limits to text extraction, failing to address the importance of critical facts in financial documents. Similarly, LongFin[19] introduces a model for long financial reports, but it still evaluates representation quality, not cross-page factual consistency or numeric correctness.

All of these benchmarks and evaluation stops at syntactic correctness. **There is a gap in Financial Fidelity Evaluation.** None explicitly addresses the asymmetric cost of factual errors: a single incorrect number, unit, or sign can render a financial report invalid. In addition, existing financial datasets assess reasoning over text; OCR datasets assess perception and layout; but no benchmark integrates both to measure factual reliability. We address these needs with FinCriticalED, a new benchmark for fact-level evaluation in financial documents. FinCriticalED combines full-page OCR extraction with financial context reasoning under factual constraints, directly assessing a model's ability to recognize and verify critical numeric and temporal facts in situ. This framework thus establishes the first standardized setting for measuring factual reliability in financial OCR, bridging the gap between general OCR performance and domain-specific understanding.

3. Methodology

To address this gap, we constructed the FinCriticalED dataset, highlighting key financial numerical and temporal expressions for evaluating models in high-stakes financial scenarios.

3.1. FinCriticalED Dataset Curation

Financial documents differ fundamentally from ordinary text-centric documents. They are dominated by numerical data and structured tables, where even minor OCR errors can lead to significant factual distortions. Existing OCR datasets focus primarily on general-domain text and use lexical metrics such as ROUGE-1, ROUGE-L, or edit distance, which fail to capture the factual accuracy required in financial contexts. To address this limitation, we curate the FinCriticalED dataset to highlight financial-critical facts that matter most for downstream financial analysis. Specifically, we annotate key numerical values and temporal expressions to enable fine-grained evaluation of factual consistency rather than lexical similarity in practical financial applications. The evaluation focuses exclusively on financially relevant text where factual correctness directly affects downstream interpretation.

3.1.1. Task Definition

The Financial Critical Error Detection task is formulated as a *structured OCR and factual verification* problem. Given an input image I from a financial document, the model generates an HTML-formatted output:

$$T = \text{OCR}(I), \tag{1}$$

which preserves both textual content and layout structure, including tables, headers, and footnotes.

To detect financial fact–specific errors, each ground-truth HTML is accompanied by fine-grained annotations of *financial facts*, denoted as:

$$F = \{f_1, f_2, \ldots, f_m\}, \tag{2}$$

where each fact f_i belongs to one of two categories:

$$F = F_n \cup F_t, \tag{3}$$

with F_n representing *numeric* facts and F_t representing *temporal* facts.

Numeric facts ($f_n \in F_n$) include quantities, percentages, and monetary values, where even subtle OCR errors, such as missing signs or misinterpreting parentheses (e.g.,

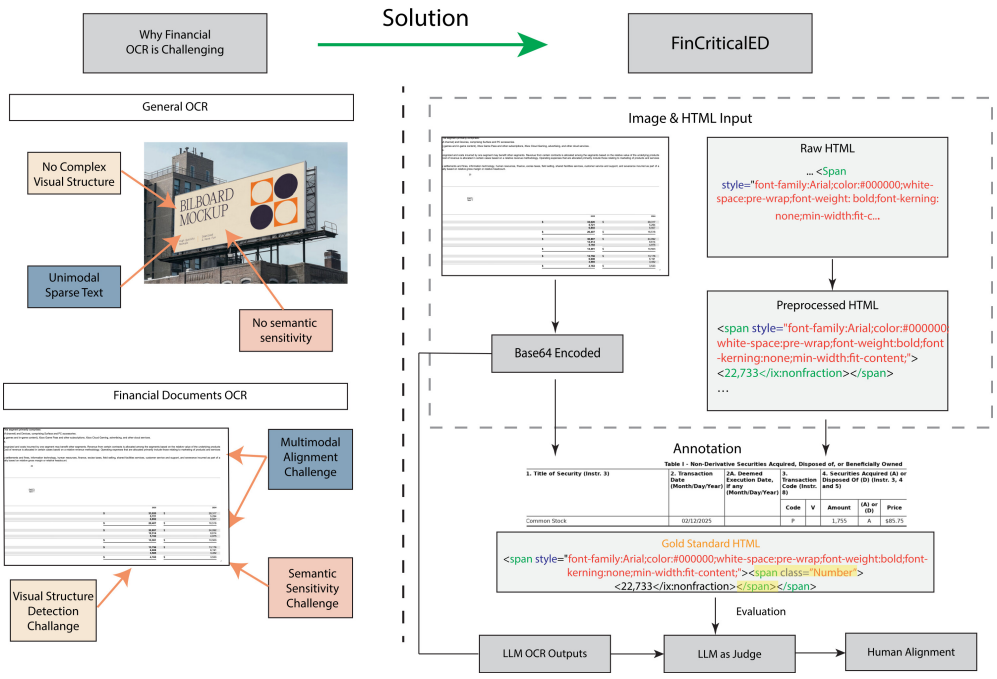

Figure 2. Challenges in financial OCR and the FinCriticalED solution pipeline. **Left**: Unlike general OCR with sparse, unimodal text and simple layouts, financial documents contain dense tables, hierarchical structures, and semantically sensitive numeric values that require multimodal alignment and layout-aware reasoning. **Right**: FinCriticalED addresses these challenges by combining page-level images, rendered and preprocessed HTML, expert-annotated financial entities, and an LLM-as-Judge evaluation framework to assess factual OCR reliability.

"(1,200)" as "1,200"), can invert the factual meaning. Temporal facts ($f_t \in F_t$) capture time-related expressions such as dates, periods, and durations that are essential for financial interpretation and timeline consistency.

3.1.2. Raw Data Collection

We collect financial documents from publicly available repositories, including the U.S. Securities and Exchange Commission (SEC) EDGAR database, corporate disclosure portals, and regulatory archives. The corpus covers five major document types: financial statements, supplemental reports, tax forms, securities transaction records, and financial legal documents.

To ensure diversity and representativeness, we perform stratified sampling across industries and fiscal years (2023–2025). Documents with complex layouts containing tables, numerical footnotes, and dated disclosures are prioritized to capture diverse financial structures. Each document is downloaded in its official structured HTML format and converted to PNG images through intermediate PDF rendering.

3.1.3. Financial Fact Annotation

To construct reliable factual supervision for FinCriticalED, we annotate financial facts directly on the ground-truth HTML representations. Two fact types are considered: *numeric* and *temporal*. Numeric facts include numbers or signed quantities with possible decimal points, thousand separators, fractions, or percentages (e.g., "1,000,000", "2,345", "0.37", "1/3", "-2.3", "(10,234)", "25.63%", "Section 30(h)"). Temporal facts include time expressions such as calendar dates, periods, and durations (e.g., "March 24, 2025", "1 month", "Q2 2025").

A detailed annotation guideline (Appendix B) was developed and refined through multiple pilot rounds to ensure high consistency across annotators and document types. Following this protocol, four trained financial annotators with backgrounds in finance, accounting, and computer technology manually labeled all factual spans using the Label Studio platform (Figure 4, Appendix B), ensuring a reproducible and auditable workflow. Each annotation precisely delineates the text span corresponding to a financial fact without including auxiliary tokens, enabling accurate alignment between OCR predictions and ground-truth facts.

3.1.4. Quality Control

To ensure the reliability and consistency of the financial fact annotations, we conducted a comprehensive quality validation using both pairwise and multi-annotator agreement metrics. Specifically, we computed Cohen's κ for all annotator pairs and Fleiss' κ for overall inter-rater reliability across all four annotators.

Annotator	1	2	3	4
1	—	0.8789	0.8952	0.9159
2	0.8789	—	0.8511	0.8294
3	0.8952	0.8511	—	0.9334
4	0.9159	0.8294	0.9334	—
Overall (Fleiss' κ)		**0.8837**		

Table 2. Inter-annotator agreement measured by pairwise Cohen's (κ) and overall Fleiss' (κ). Higher values indicate stronger annotation consistency.

Category	Total Count	Average per Document	Type Ratio (%)
Dataset Size (pairs)	500	–	–
All Financial Facts (F)	739	2.44	100.0
Numeric Facts (F_n)	444	1.48	60.1
Temporal Facts (F_t)	295	0.96	39.9

Table 3. Statistics of the FinCriticalED dataset containing 500 document–image pairs. Each value represents the count or average across all samples.

As shown in Table 2, pairwise agreement scores range from 0.83 to 0.93, indicating strong cross-annotator consistency. The overall Fleiss' κ reaches 0.8837, confirming high annotation reliability and stable adherence to the annotation guidelines. These results demonstrate that the annotation process achieved robust factual agreement across annotators, ensuring the dataset's suitability for precise financial fact evaluation.

3.1.5. Dataset Statistics

The final FinCriticalED dataset contains 500 annotated document–image pairs, each with a corresponding ground-truth HTML file. Across all samples, we identify a total of 739 financial fact entities, consisting of 444 numeric and 295 temporal facts. On average, each document contains approximately 1.48 numeric and 0.96 temporal facts, reflecting the heterogeneous density of financial information across document types.

The dataset covers diverse financial document structures, including tabular statements, narrative disclosures, and time-dependent reporting sections. This balanced composition enables both *value-centric* (numeric) and *time-centric* (temporal) factual evaluations for comprehensive OCR benchmarking.

3.2. Financial Fact Evaluation

Traditional OCR evaluation relies on surface-level lexical metrics such as ROUGE-1, ROUGE-L, or edit distance. While these metrics capture string similarity, they fail to assess whether a model preserves the *financial meaning* of a document. Unlike lexical metrics that reward superficial string overlap, our evaluation directly measures whether the model preserves domain-critical factual content. This shift from text similarity to *fact-level correctness* is essential for financial applications, where even small OCR deviations can alter the underlying numerical or temporal meaning.

3.2.1. Overview

FinCriticalED operationalizes this perspective by evaluating whether a model correctly reproduces key financial facts, specifically the numerical and temporal entities that determine the validity of financial statements. Given a ground-truth HTML annotated with explicit `<Number>` and `<Date>` tags and a model-generated HTML without tags, the objective is to verify the correctness of all annotated financial facts.

$$F = F_n \cup F_t = \{f_1, f_2, \ldots, f_m\}, \tag{4}$$

where F_n and F_t denote the sets of *numerical* and *temporal* facts respectively. We report three metrics: (1) *Financial Fact Accuracy (FFA)*, (2) *Numerical FFA (N-FFA)*, and (3) *Temporal FFA (T-FFA)*.

3.2.2. Fact Matching Strategy

For each fact $f_i \in F$, the evaluation checks whether its textual value appears in a semantically equivalent local context in the model-generated HTML. Each fact is assigned a binary correctness indicator:

$$\delta_i = \begin{cases} 1, & \text{if } f_i \text{ is correctly recovered,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

For *numerical facts* ($f_i \in F_n$), the match requires exact preservation of signs, commas, decimal points, and parenthetical negative notation. For example, converting "(1,200)" to "1,200" is marked incorrect because it inverts the financial meaning. For *temporal facts* ($f_i \in F_t$), the date or duration must also be exactly reproduced.

3.2.3. Financial Fact Accuracy (FFA)

For each document, overall Financial Fact Accuracy (FFA) is computed as:

$$\alpha = \frac{\sum_i \delta_i}{|F|} \quad \text{(FFA)}. \tag{6}$$

Numerical and temporal accuracies are:

$$\alpha_n = \frac{\sum_{f_i \in F_n} \delta_i}{|F_n|} \quad \text{(N-FFA)}, \tag{7}$$

$$\alpha_t = \frac{\sum_{f_i \in F_t} \delta_i}{|F_t|} \quad \text{(T-FFA)}. \tag{8}$$

These metrics directly quantify a model's ability to preserve the financial facts necessary for correct downstream interpretation.

To obtain dataset-level performance, we aggregate correct and total fact counts across all samples. Let C_n and

T_n denote the aggregated correct and total numerical facts; Numerical FFA is:

$$\text{N-FFA} = 100 \times \frac{C_n}{T_n}. \tag{9}$$

Temporal FFA is computed analogously using aggregated temporal facts, and overall FFA is computed over all facts in F. This aggregation normalizes across documents with varying fact densities and yields robust, benchmark-level estimates focused solely on factual fidelity.

3.2.4. LLM-as-Judge Implementation

We implement the evaluation framework using an LLM-as-Judge design. A large language model (GPT-4o) serves as the evaluator responsible for extracting financial facts from the ground-truth HTML and verifying their presence in the model-generated HTML. The LLM processes both inputs under a structured evaluation prompt (provided in the Appendix C), enabling it to perform normalization, contextual matching, and fine-grained fact checking.

This LLM-based evaluation allows reasoning over financial fact correctness, capabilities that are difficult to capture with rule-based string matching. The evaluator outputs per-fact correctness decisions that are directly used to compute FFA, N-FFA, and T-FFA.

3.2.5. Financial Expert Analysis

We further assess the reliability of the LLM-as-Judge evaluator through an expert analysis. Financial-domain experts review sampled cases judged by the gpt-4o framework, focusing on errors with material impact such as sign inversion, changes in numerical magnitude, parenthetical negative notation, and temporal inconsistencies related to fiscal periods.

For each case, experts evaluate whether the LLM's decision is correct and whether its reasoning aligns with financial reporting conventions. This expert analysis provides human-grounded validation that the evaluation framework reflects domain-critical fact correctness.

3.3. Benchmark Settings and Baselines

To assess the difficulty and coverage of the proposed benchmark, we evaluate a diverse set of representative systems spanning three categories: SOTA OCR models, open-source multimodal vision-language models, and proprietary multimodal vision-language models models. All evaluations follow the same pipeline described in Section 4.2 to ensure comparability.

3.3.1. Experimental Setup

State-of-the-Art OCR Models This category comprises State-of-the-Art OCR systems that represent the highest performance tier in text detection and structured document reconstruction. Representative models include DeepSeek-OCR[34] and MinerU 2.5 [23]. These models are optimized for high-accuracy visual text extraction across complex document layouts and multilingual inputs. DeepSeek-OCR is a fine-tuned, large-scale OCR system trained for multi-page document comprehension and structure-preserving recognition, while MinerU 2.5 adopts a decoupled vision–language architecture that improves generalization across financial and administrative document types. Together, these models provide a strong OCR-oriented baseline, enabling direct comparison between pure visual-text extraction and multimodal reasoning approaches in financial document understanding.

Open-Source Vision–Language Models The open-source VLM group includes Qwen2.5-VL-72B-Instruct [30], Llama-4-Scout-17B-16E-Instruct [27], and other publicly available multimodal models. These models integrate both OCR and reasoning components, allowing them to interpret tabular, textual, and visual features jointly. They are evaluated to understand how open models handle long-context factual extraction in visually dense financial documents. Their open accessibility also enables reproducible benchmarking within FinCriticalED, forming the mid-tier baseline for cross-modal factual grounding.

Proprietary Multimodal Models The final category includes proprietary large-scale models such as GPT-4o [10] and GPT-5 [24], which represent the current frontier of multimodal reasoning. These models can directly process images and text through the OpenAI API, providing end-to-end financial document interpretation without external OCR preprocessing. Their performance indicates the upper bound of multimodal factual comprehension under instruction-tuned supervision, highlighting the remaining gaps between closed and open systems in OCR accuracy and financial fact fidelity.

3.3.2. Implementation Details

All OpenAI- and TogetherAI-hosted models, including Qwen2.5-VL, Llama-4-Scout, and Gemma-3, are accessed via their official APIs with temperature set to zero for deterministic outputs. DeepSeek-OCR is self-hosted using four NVIDIA GPUs, while MinerU 2.5 is accessed through its public API endpoint.

All models share a unified input–output format aligned with the FinCriticalED annotation schema. The evaluation harness automatically normalizes numeric expressions and date formats prior to comparison to ensure factual consistency across architectures. This rigorous implementation guarantees fair comparison across open and proprietary models, ensuring that performance differences reflect factual accuracy rather than deployment variance.

Model	Size	General (%)				Fact-Level (%)			
		R1	RL	E↓	Rank	N-FFA	T-FFA	FFA	Rank
OCR Models									
MinerU2.5 [23]	1.2B	47.58	47.33	63.39	7	68.84	83.97	74.55	7
DeepSeek-OCR [34]	6B	55.46	55.05	57.09	6	77.09	77.82	77.15	6
Open-source VL Models									
Gemma-3n-E4B-it [8]	4B	60.08	59.08	52.59	3	78.07	81.38	78.07	5
Llama-4-Scout-17B [27]	17B	59.50	58.65	53.38	4	81.95	88.97	84.67	3
Qwen2.5-VL-72B [30]	72B	**60.51**	**59.96**	52.23	2	75.63	82.77	78.24	4
Proprietary VL Models									
GPT-4o [10]	-	60.41	59.13	**49.76**	**1**	83.99	91.25	86.95	2
GPT-5 [24]	-	57.37	56.65	50.47	5	**91.46**	**95.69**	**93.61**	**1**

Table 4. Comparison of OCR and multimodal models on FinCriticalED across general OCR metrics and fact-level evaluations. General Rank is calculated using $(\mathrm{R1} + \mathrm{RL} + 1 - \mathrm{E})/3$. Fact Rank is determined by overall FFA. (R1 = ROUGE-1; RL = ROUGE-L; E = Edit Distance↓; N-FFA = numerical fact accuracy; T-FFA = temporal fact accuracy; FFA = overall financial fact accuracy).

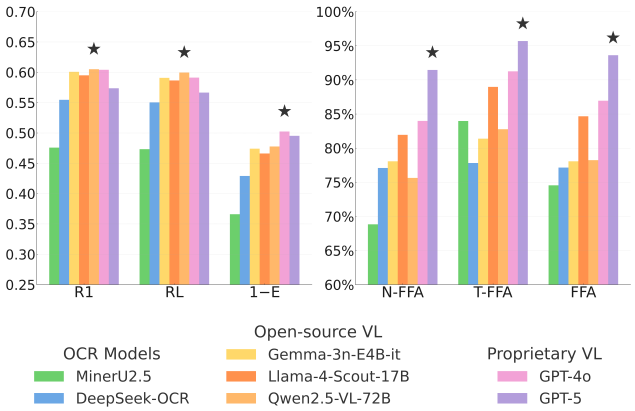

Figure 3. Comparison of OCR and multimodal models on general OCR metrics and financial fact-level accuracy. Best results per metric are highlighted with a star (⋆). (R1 = ROUGE-1; RL = ROUGE-L; E = Edit Distance↓; N-FFA = numerical fact accuracy; T-FFA = temporal fact accuracy; FFA = overall financial fact accuracy).

4. Results and Analysis

4.1. Main Results

4.1.1. RQ1: Fact-level vs. Text-level Performance

Research Question: How do models differ in their financial OCR performance on traditional text-level metrics (ROUGE, Edit Distance) compared with fact-level financial accuracy (FFA)?

Table 4 compares LLM output HTML with the gold standard HTML from original financial documents. Across these models, we observe a consistent gap between traditional text-level metrics and fact-level accuracy. This difference should not be interpreted as a direct performance delta but rather the distinct evaluation focuses: surface-level textual similarity vs. semantic-level factual correctness.

Traditional textual metrics (ROUGE-1, ROUGE-L, and Edit Distance) capture string-level overlap between predicted and reference HTML, emphasizing syntactic reconstruction instead of factual fidelity. OCR-oriented models like DeepSeek-OCR and MinerU 2.5 achieve moderate text-level scores (55.46% ROUGE-1 and 57.09% Edit Distance for DeepSeek-OCR; 47.58% and 63.39% for MinerU 2.5), indicating that these models can partially reconstruct the document structure. However, the textual overlap mainly reflects consistency in token sequences such as tag boundaries or CSS tokens, rather than accuracy in the underlying financial facts. Minor differences in table alignment, tag nesting, or style normalization often lead to large variations in ROUGE and Edit Distance, even when the numerical and temporal content remains correct.

However, the Fact-Level Accuracy (FFA) metrics reveal a different pattern that better aligns with semantic correctness. Models such as GPT-5 maintain high factual accuracy (91.46% and 95.69% for numerical and temporal entities, respectively) despite moderate textual similarity scores (57.37% ROUGE-1). Qwen2.5-VL-72B also demonstrates balanced factual precision (75.63% N-FFA and 82.77% T-FFA) with only 60.51% ROUGE-1, suggesting that even limited textual resemblance can preserve essential financial content. Conversely, MinerU 2.5, which records the highest Edit Distance value (63.39%), performs the worst in factual accuracy (68.84% N-FFA and 83.97% T-FFA). This inconsistency confirms that high token-level similarity does not guarantee accurate semantic reproduction.

Key takeaway: Fact-level evaluation focuses on the accurate recognition of numerical and temporal entities, which are more critical to financial understanding, indicating that models such as Qwen2.5-VL-72B and GPT-5 achieve high factual accuracy even when their textual overlap is limited.

4.1.2. RQ2. Numerical vs. Temporal Fact OCR

Research Question: How do LLMs differ in their handling of numerical versus temporal entities within financial OCR?

Fact-level results in Table 4 reveal a consistent discrepancy between numerical and temporal accuracy across models. While numerical fields such as revenues, costs, and percentages vary in notation, sign, and unit representation, temporal expressions follow more standardized formats, including fiscal quarter labels and reporting dates. This structural regularity contributes to systematically higher T-FFA across all models.

For example, MinerU 2.5 records 68.84% N-FFA and 83.97% T-FFA, showing a clear 15% difference. DeepSeek-OCR achieves similar accuracies on both dimensions (77.09% and 77.82%), suggesting better normalization for numeric data than most open-source models. Proprietary systems further amplify this gap: GPT-4o reaches 83.99% N-FFA and 91.25% T-FFA, while GPT-5 attains 91.46% and 95.69%, respectively. These results confirm that tem-

poral recognition is a more deterministic subtask, primarily pattern-driven, whereas numerical interpretation requires semantic reasoning over contextual scales, signs, and currency conventions.

Key takeaway: All models display a consistent temporal advantage, typically three to five percentage points higher in T-FFA than in N-FFA. This finding highlights a key bottleneck in financial OCR: despite strong multimodal reasoning, current systems still underperform on numerical fields where factual precision depends on context-sensitive normalization rather than pattern matching. Bridging this gap will require integrating symbolic reasoning or structured financial priors into future multimodal LLM architectures.

4.1.3. RQ3. Proprietary vs. Open-Source Model Reliability

Research Question: To what extent do proprietary large multimodal models outperform open-source VLMs in factual and temporal accuracy?

Proprietary models exhibit the highest factual fidelity across all evaluation dimensions. GPT-4o achieves a numerical FFA of 83.99%, temporal accuracy of 91.25%, and an overall factual score of 86.95%. GPT-5 further improves performance, reaching 91.46% in numeric accuracy, 95.69% in temporal accuracy, and an overall factual score of 93.61%, setting a new benchmark for factual correctness in multimodal financial understanding.

The superior factual fidelity of proprietary models reflects not only larger model capacity but also more comprehensive multimodal alignment and layout-aware supervision during pretraining. In contrast, open-source systems, despite rapid progress, still rely on limited document-style corpora, which constrain their ability to capture hierarchical relationships between text, numbers, and layout. Nevertheless, open-source models such as Qwen2.5-VL-72B and Llama-4-Scout-17B, although trailing by 8–10 points, demonstrate comparable consistency and steady improvement across both text-level and fact-level metrics.

This narrowing performance gap suggests that open-source VLMs are rapidly approaching proprietary models in factual accuracy, largely due to enhanced instruction-tuning and layout-aware pretraining.

Key takeaway: While proprietary systems such as GPT-5 remain superior in both numerical and temporal accuracy, open-source models (notably Qwen2.5-VL-72B) show competitive factual reliability, signaling that transparent, community-driven multimodal architectures can soon match closed-source performance in financial OCR.

4.2. LLM-as-Judge Reliability

To validate the LLM-as-Judge evaluation scheme, we conduct a case-level alignment study comparing its decisions with human expert assessments. Two representative examples are selected: one where both human evaluators and the LLM-as-Judge agree on **high** factual accuracy, and another where both concur on **low** OCR fidelity.

In the high-accuracy case (Figure 5, Appendix), all monetary and date fields match exactly after normalization. The LLM-as-Judge correctly identifies every numeric and percentage value, mirroring human judgment.

As shown in Figure 6, the errors include not only missing durations and hallucinated headings, but also cases where correct numeric entities are matched to incorrect rows. Further case studies are conducted to confirm that the LLM-as-Judge flags the same errors with same explanation from human experts, indicating that it can detect semantically significant OCR failures. More details of the human alignment assessments can be found in Appendix C.2.

Key takeaway: LLM-as-Judge shows strong alignment with human evaluation, demonstrating (1) robust multimodal reasoning over image and HTML inputs, (2) reliable comparison of structured and unstructured content, and (3) accurate, instruction-following scoring. These properties make it a scalable and dependable framework for financial OCR assessment.

When paired with quantifiable metrics such as Fact-Level Accuracy (FFA), LLM-as-Judge provides an effective foundation for automated and interpretable evaluation in precision-critical financial OCR.

5. Conclusion

We present FinCriticalED, a fact-level visual benchmark for evaluating financial OCR reliability. Unlike traditional OCR benchmarks that emphasize character or word recognition, FinCriticalED focuses on visually grounded financial facts. The benchmark includes an expert-annotated dataset covering critical numerical and temporal facts, together with an LLM-as-Judge protocol for structured fact-level assessment.

Our experiments yield three key findings. First, conventional text similarity metrics do not capture fact-level correctness and often mask financially meaningful errors. Second, models consistently achieve slightly higher accuracy on temporal facts than on numerical ones, reflecting the greater structural regularity of temporal expressions compared with the context-dependent variability of numeric values. Third, although proprietary vision language systems achieve the highest factual accuracy, open source models are improving rapidly, indicating that precise modeling of layout and structure is more decisive than model scale.

In summary, FinCriticalED reframes financial OCR as a visual reasoning task that requires alignment across layout, structure, and factual meaning. The benchmark supports progress toward high-precision, layout-aware financial document understanding and provides a foundation for future research on visually grounded factual reliability in precision-critical domains.

References

[1] Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Onechart: Purify the chart structural extraction via one auxiliary token. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 147–155, 2024. 2

[2] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. 3

[3] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*, 2025. 2

[4] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 2

[5] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. 3

[6] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2025. 2, 3

[7] Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 12867–12874, New York, NY, USA, 2025. Association for Computing Machinery. 2, 3

[8] Google. google/gemma-3n-E4B-it: Gemma 3N Instruction-Tuned Model, 2025. 7

[9] Shamanthak Hegde, Pooyan Fazli, and Hasti Seifi. Chartqa-x: Generating explanations for visual chart reasoning. *arXiv preprint arXiv:2504.13275*, 2025. 2, 3

[10] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6, 7

[11] Yichao Jin, Yushuo Wang, Qishuai Zhong, Kent Chiu Jin-Chun, Kenneth Zhu Ke, and Donald MacDonald. Multi-stage field extraction of financial documents with ocr and compact vision-language models. *arXiv preprint arXiv:2510.23066*, 2025. 1

[12] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965. 2

[13] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 2, 3

[14] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2

[15] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024. 2, 3

[16] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. 2, 3

[17] Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. *arXiv preprint arXiv:2505.24714*, 2025. 2, 3

[18] Souvik Mandal, Nayancy Gupta, Ashish Talewar, Paras Ahuja, Prathamesh Juvatkar, and Gourinath Banda. Idpleaderboard: A unified leaderboard for intelligent document processing tasks, 2025. 1

[19] Ahmed Masry and Amir Hajian. Longfin: A multimodal document understanding model for long financial domain documents. *arXiv preprint arXiv:2401.15050*, 2024. 2, 3

[20] Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*, 2020. 2

[21] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952, 2019. 2, 3

[22] Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. Dolfin – document-level financial test set for machine translation. *arXiv preprint arXiv:2502.03053*, 2025. 2

[23] Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, Zirui Tang, Boyu Niu, Ziyang Miao, Hejun Dong, Siyi Qian, Junyuan Zhang, Jingzhou Chen, Fangdong Wang, Xiaomeng Zhao, Liqun Wei, Wei Li, Shasha Wang, Ruiliang Xu, Yuanyuan Cao, Lu Chen, Qianqian Wu, Huaiyu Gu, Lindong Lu, Keming Wang, Dechen Lin, Guanlin Shen, Xuanhe Zhou, Linfeng Zhang, Yuhang Zang, Xiaoyi Dong, Jiaqi Wang, Bo Zhang, Lei Bai, Pei Chu, Weijia Li, Jiang Wu, Lijun Wu, Zhenxiang Li, Guangyu Wang, Zhongying Tu, Chao Xu, Kai Chen, Yu Qiao, Bowen Zhou, Dahua Lin, Wentao Zhang, and Conghui He. Mineru2.5: A decoupled vision-language model for

efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2023. 2, 6, 7

[24] OpenAI. Gpt-5 system card, 2025. 2, 6, 7

[25] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *arXiv preprint arXiv:2412.07626*, 2024. 2, 3

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318, USA, 2002. Association for Computational Linguistics. 2

[27] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *arXiv preprint arXiv:2204.05149*, 2022. 6, 7

[28] Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, Shanshan Yang, Yilun Zhao, Zhiwei Liu, Peng Lu, Jerry Huang, Suyuchen Wang, Triantafillos Papadopoulos, Polydoros Giannouris, Efstathia Soufleri, Nuo Chen, Zhiyang Deng, Heming Fu, Yijia Zhao, Mingquan Lin, Meikang Qiu, Kaleb E Smith, Arman Cohan, Xiao-Yang Liu, Jimin Huang, Guojun Xiong, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Sophia Ananiadou, and Qianqian Xie. Multifinben: Benchmarking large language models for multilingual and multimodal financial application. *arXiv preprint arXiv:2506.14028*, 2025. 1, 2, 3

[29] Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. Exploring ocr capabilities of gpt-4v(ision) : A quantitative and in-depth evaluation, 2023. 1

[30] Qwen Team. Qwen2.5-vl, 2025. 6, 7

[31] Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. *arXiv preprint arXiv:2401.13311*, 2024. 2, 3

[32] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023. 2

[33] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. 2

[34] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025. 2, 6, 7

[35] Haoran Wei, Youyang Yin, Yumeng Li, Jia Wang, Liang Zhao, Jianjian Sun, Zheng Ge, Xiangyu Zhang, and Daxin Jiang. Slow perception: Let's perceive geometric figures step-by-step. *arXiv preprint arXiv:2412.20631*, 2025. 2

[36] Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*, 2024. 2, 3

[37] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2024. 2, 3

[38] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria, 2025. Association for Computational Linguistics. 2, 3

FinCriticalED: A Visual Benchmark for Financial Fact-Level OCR Evaluation

Supplementary Material

A. Dataset Construction Details

We provide the complete pipeline for data collection, preprocessing, and annotation. This includes document source distribution, sampling criteria, and anonymization protocols. Each document is stripped of personally identifiable information and re-rendered into page-level PDFs for OCR benchmarking.

A.1. Document Sources and Sampling

Documents are collected from publicly available financial repositories, including the U.S. Securities and Exchange Commission (SEC) EDGAR database, stock exchange disclosure portals, and open-access corporate filings. We include five primary document types: *financial statements*, *supplementary financial reporting*, *tax forms*, *securities transaction records*, and *financial legal disclosures*.

To ensure annotation quality, each sampled filing must contain at least one section with numeric tables, textual footnotes, and dated references. The final dataset comprises 500 curated documents evenly distributed across the five document types.

A.2. Preprocessing and Rendering

All source documents are converted into canonical HTML using a rule-based extraction pipeline. Each HTML file is then paired with its corresponding page-level image deriving from intermediate PDFs, rendered at 300 dpi using Chromium headless mode to preserve layout fidelity. During rendering, embedded objects such as charts, tables, and formulas are retained as rasterized elements. For each document, metadata (document type, source URL, filing year) is recorded and stored in structured JSON format.

To ensure compatibility with OCR models, we further process each document into a page-level dataset, normalizing text alignment, table borders, and visual regions. Empty or blank pages are automatically filtered out.

A.3. Annotation and Quality Control

Annotations are performed directly on HTML text, using the corresponding page image as visual reference for validation. Annotators highlight entities within HTML and embed specialized `<span>` tags encoding entity type (*numeric* or *temporal*), normalized value, and sign conventions. Each document is independently labeled by two annotators, with conflicts resolved by consensus. An internal agreement table reports consistent reliability across both entity types (see Table 2).

All annotations undergo rigorous validation, including cross-checks for missing tags, invalid numeric formats, or inconsistent temporal normalization. Post-annotation, results are reviewed by a senior annotator for consistency and schema compliance.

A.4. Final Statistics and Accessibility

The final dataset contains 500 fully annotated financial documents, totaling 739 annotated entities (444 numeric and 295 temporal).

Each page is accompanied by its source metadata, HTML text, and annotated span-level markup. All files are distributed under a research-only license via a secure hosting repository.

B. Annotation Guidelines

Annotators were instructed to prioritize visually salient, economically meaningful facts such as totals, subtotals, important dates and durations appearing in tables or textual statements. An excerpt from the annotation manual and UI interface is shown in Figure 4.

FinCriticalED annotation procedure revolves around entity labeling for financial documents. The goal is to identify financially critical entities, focusing on numbers and time.

B.1. Annotation Rules

This section outlines the detailed annotation protocol used in *FinOCRBench*. Annotators followed a standardized set of rules to ensure consistency across financial documents, with two primary entity categories and unified markup conventions applied to rendered HTML text.

B.1.1. Entity List

- Number
- Date and Duration

B.1.2. General Rules

- Identify all entities in the HTML file that belong to one of the two categories listed above.
- Use the **rendered HTML** as the primary source of truth; use the corresponding page image only for visual assistance when the layout or OCR text appears ambiguous.
- The numeric value must be **financially relevant**, excluding items such as phone numbers, postal codes, address numbers, or document item numbers.
- For dates, annotate only the date expression itself and remove any connective context between two dates. For example, in *PERIODS JANUARY 2, 2025 THRU JANUARY*

8, 2025, only annotate *JANUARY 2, 2025* and *JANUARY 8, 2025*, removing *PERIODS* and *THRU*.
- Highlight **all valid entities** in each task without omission.

B.1.3. Specific Rules

Examples:
- **Numeric values:** annotate the number only, including its sign and any relevant punctuation. This may include decimal points, commas for thousands, or percentage symbols. Examples: "1,000,000", "2.345", "0.37", "10", "1/3", "-2.3", "(10,234)", "25.63%", "Section 39(h)".
- **Date and Duration:** annotate explicit temporal expressions such as calendar dates or duration phrases. Examples: "March 24, 2025", "1 month", "2 years".

B.2. Annotator Demography

The construction of the FinCriticalED dataset relies on the deep domain expertise of a team of highly qualified annotators with strong backgrounds in finance, economics, and computer technology. Their interdisciplinary training and multilingual fluency ensure accurate and contextually grounded annotations across both financial statements and cross-lingual news articles.

One annotator, currently working as a senior analyst at a major U.S. financial institution, holds a master's degree in Business Analytics from a leading Ivy League university and a bachelor's degree in Statistics and Economics. Their background includes research on large language models (LLMs), financial data analysis, and economics, enabling precise annotation of complex financial information. Fluent in Chinese and experienced in multilingual reasoning, she brings a high level of rigor to aligning financial data with multilingual news content.

Another annotator earned their master's degree in Financial Mathematics from a prominent U.S. institution. They are currently pursuing a second master's degree in Computer Science with a focus on machine learning at a major public research university in the United States. With over seven years of professional experience in strategic finance and consulting within the FinTech industry, they contribute practical expertise in corporate finance, along with a strong foundation in computational methods.

The third annotator is a graduate student in Computer Technology with a solid foundation in auditing, financial analysis, and data processing. Having participated in multiple financial data annotation projects, this annotator possesses extensive familiarity with annotation workflows, schema design, and quality-control standards. A prior internship at a technology company provided experience in data preprocessing and model support, reinforcing the annotator's capability to ensure precision and consistency across large-scale auditing and financial data tasks.

The fourth annotator is also a graduate student specializing in Computer Technology, with research focused on the integration of large language models (LLMs) into auditing and assurance applications. Her experience includes LLM evaluation and fine-tuning for domain adaptation, bringing a research-oriented perspective to the annotation process. Her technical background complements the team's financial expertise, contributing to the dataset's balance between linguistic fidelity and computational reproducibility.

Together, the annotation team represents a balanced combination of domain-specific expertise and technical proficiency. Two annotators bring professional experience in financial analytics and quantitative modeling, while two others contribute strong computational and auditing backgrounds with hands-on experience in LLM evaluation and data engineering. This interdisciplinary composition ensures both semantic precision in financial reasoning and methodological consistency in data labeling, forming a robust foundation for the dataset's reliability and reproducibility.

B.3. Annotation Process

The annotation workflow of FinCriticalED follows a structured, multi-stage process designed to ensure both factual precision and consistency across annotators. Each annotator operates within a controlled web-based interface that displays paired `HTML` text and its corresponding rendered page image. This dual-view setup enables annotators to reference the visual layout when verifying line breaks, superscripts, or numerical formatting inconsistencies.

The annotation process begins with a **pre-screening phase**, where annotators review extracted HTML text for structural integrity and identify potential OCR or layout errors. Once validated, they proceed to the **entity marking phase**, where relevant segments—numerical values, signs, dates, and durations—are highlighted and later enclosed within specialized `<span>` tags (e.g., `<span class="Number">...</span>`, `<span class="Time">...</span>`).

Each page is annotated independently by two annotators, ensuring redundancy for quality measurement. Upon completion, annotations are exported into JSON format, preserving the hierarchical document structure and linking each annotated entity to its source metadata. This structure enables downstream comparison against model-generated predictions for entity-level evaluation.

B.4. Validation Guideline

To ensure reliability and reproducibility, FinCriticalED employs a multi-layered validation protocol combining automated checks, cross-annotator comparison, and expert review.

First, an **automated integrity validation** step scans all annotated files to detect malformed or nested `<span>` tags, missing entity attributes, or misaligned indices within the

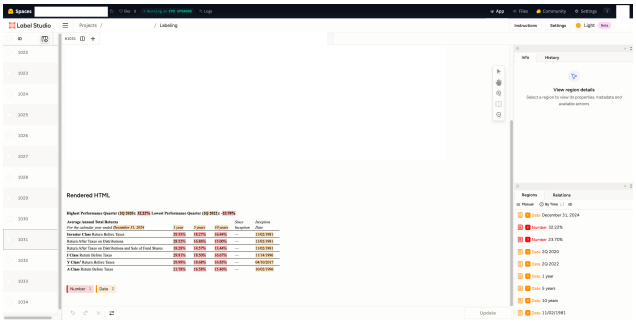

Figure 4. Annotation interface used in FinCriticalED. Annotators highlight entities directly within HTML while referencing rendered page images for layout validation.

exported JSON schema.

Second, an **inter-annotator consistency check** evaluates pairwise agreement between annotators using both token-level and entity-level overlap metrics. Entities with disagreement scores below a fixed confidence threshold are automatically queued for adjudication. Discrepancies are resolved through a consensus meeting facilitated by a lead annotator, who applies majority agreement rules and reviews edge cases (e.g., ambiguous table headers or multi-line numeric ranges).

Third, a **domain validation review** is conducted after every 100 annotated pages. In this phase, a senior financial expert randomly samples 5% of the annotations to assess factual correctness, sign consistency, and adherence to normalization conventions. Feedback from this phase is incorporated into a continuously refined annotation guideline that codifies new patterns or exceptions encountered during labeling.

Finally, the validated annotations are version-controlled and re-exported to ensure traceability across dataset releases. The combination of automated validation, human consensus, and domain-level oversight establishes a high-confidence ground truth benchmark for evaluating factual accuracy and numeric fidelity.

B.5. Annotation Agreement Metric

To assess the reliability and consistency of human annotation in FinCriticalED, we employ both **Cohen's Kappa** [2] for pairwise annotator agreement and **Fleiss' Kappa** [5] for overall multi-annotator reliability. These chance-corrected metrics provide a robust measure of agreement that accounts for the likelihood of random coincidence in categorical labeling.

Cohen's Kappa (κ) quantifies the level of agreement between two annotators and is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{10}$$

where p_o represents the observed agreement (the proportion of items on which both annotators agree) and p_e denotes the expected agreement based on random chance. A value of $\kappa = 1$ indicates perfect agreement, while $\kappa = 0$ reflects agreement equivalent to chance.

For more than two annotators, we adopt Fleiss' Kappa, which generalizes the formulation of inter-rater agreement to multiple raters. It is computed as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{11}$$

where $\bar{P}$ denotes the mean proportion of observed agreement across all annotation units, and $\bar{P}_e$ indicates the expected probability of agreement under a random labeling assumption. Similar to Cohen's Kappa, higher values signify stronger consensus among annotators.

In FinCriticalED, we calculate Cohen's Kappa for each annotator pair to evaluate pairwise consistency and Fleiss' Kappa for the full group to measure overall reliability across both entity categories (*numeric* and *temporal*). These statistics confirm a high level of agreement among the annotators, underscoring the robustness of the dataset's annotation process.

C. LLM-as-Judge Setup And Output

LLM Judge (GPT-4o) Prompt

```
JUDGE_PROMPT_TEMPLATE = """ #Instruction: You
are an expert HTML layout inspector for
financial documents.
You will be given:
    1. Ground truth HTML that contains special
    entity tags such as <Number>...</Number>
    and <Date>...</Date>.
    2. Model-generated HTML that was produced
    from the same image but does not contain
    those tags.
Your goal is to judge how correct the
generated HTML is compared with the
ground truth HTML.
Follow all steps carefully and output only
one JSON object as the final result.
# Step 1: Normalize the ground truth
structure
The ground truth HTML contains entity
tags that are not part of the visual HTML
structure. Create a \structure-only" version
of the ground truth by removing the following
tags but keeping their inner text:
* <Number> and </Number>
* <Date> and </Date>
Keep everything else | tag order, nesting,
table structure, fonts, and text | unchanged.
Call this result GT_STRUCT.
# Step 2: Check structural fidelity
Compare GT_STRUCT with the model-generated
HTML. Judge:
* Whether the generated HTML is syntactically
complete (properly closed tags, valid structure).
* Whether the main blocks, headings, and
paragraphs appear in the same order.
* Whether tables, rows, and columns are
preserved and aligned.
* Whether obvious style containers (<div>,
<font>, etc.) are retained so layout can be
reconstructed.
```

```
Assign a structure_score in (0, 1, 2):
* 2 - Structure complete and strongly
aligned with GT_STRUCT.
* 1 - Mostly aligned but missing or
misplaced parts.
* 0 - Incomplete or clearly inconsistent.
Provide a short textual explanation in
structure_explanation.
# Step 3: Extract and match entities
From the original ground truth HTML,
extract all GT entities explicitly enclosed
by the special tags:
* <Number> ... </Number> → entity type =
"Number"
* <Date> ... </Date> → entity type =
"Date"
Do not infer entities by meaning; only
extract those wrapped by these tags.
Each entity record must include:
* type - "Number" or "Date" (derived
from tag name)
* value - exact inner text between the tags
* context_hint - short fragment of surrounding
text that helps locate it
Example:
[
  {"type": "Number", "value": "50,000",
  "context_hint": "sales charge discounts"},
  {"type": "Date", "value": "December
  31, 2024", "context_hint": "fiscal year ended"}
]
Then, in the generated HTML (which lacks tags),
locate each entity's value within a similar
textual or positional context. Count a
match as correct only if the same text
appears in the correct or very similar
paragraph, sentence, or table cell.
Compute and report:
* total_entities = count of GT entities
* total_entities_with_Number_type = count
of GT Number entities
* total_entities_with_Date_type = count
of GT Date entities
* correct_entities = the number of entities
that is correctly found
* correct_entities_with_Number_type = the
number of entities with Number type that
is correctly found
* correct_entities_with_Date_type = the
number of entities with Date type that
is correctly found
* entity_accuracy = correct_entities
/ total_entities
    (entity_accuracy should be a
          ]
    }
    floating-point number between 0 and
    1, rounded to two decimal places.)
* entity_score = entity_accuracy × 5
    (entity_score should be a floating-
    point number between 0 and 5, rounded
    to two decimal places.)
If an entity is partially matched (e.g.,
missing comma or currency symbol), mark
it incorrect and explain why in
entity_diagnostics.
# Step 4: Overall judgment
Give an overall_score (0{10).
Consider both structure and entity accuracy:
* Formula: overall_score = structure_score × 2
+ entity_score (clipped to 0{10)
* If the structure is severely broken, lower
overall_score even if entities match.
* Add a short justification
(overall_explanation).
```

```
# Step 5: Output format
Output exactly one valid JSON object:
{
  "structure_score": 0,
  "structure_explanation": "",
  "total_entities": 0,
  "total_entities_with_Number_type": 0,
  "total_entities_with_Date_type": 0,
  "correct_entities": 0,
  "correct_entities_with_Number_type": 0,
  "correct_entities_with_Date_type": 0,
  "entity_accuracy": 0.0,
  "entity_score": 0.0,
  "overall_score": 0.0,
  "overall_explanation": ""
}
Rules:
* Output only valid JSON (no extra text).
* Numbers must be numeric, not strings.
* List every entity in order in
entity_diagnostics.
"""
```

C.1. LLM Judge output sample

LLM Judge Output Sample (case of high FFA)


```
{"structure_score": 2,
  "structure_explanation": "The model-generated
  HTML is syntactically complete and the
  structure closely aligns with the GT_STRUCT.
  All main blocks, headings, and tables are
  presented in the same order as the GT_STRUCT
  with proper nesting and alignment preserved.",
  "total_entities": 19,
  "total_entities_with_Number_type": 13,
  "total_entities_with_Date_type": 6,
  "correct_entities": 19,
  "correct_entities_with_Number_type": 13,
  "correct_entities_with_Date_type": 6,
  "entity_accuracy": 1.0,
  "entity_score": 5.0,
  "overall_score": 9.0,
  "overall_explanation": "Both the structure and
  content of the model-generated HTML match the
  ground truth very well, leading to a high
  overall score.}
```


LLM Judge Output Sample (case of low FFA)


```
{"structure_score": 1,
  "structure_explanation": "The generated HTML
  table structure captures the rows and columns but
  misaligns the headers and has an extra data row.
  The main blocks are in order but with variability
  in class and entity rows.",
  "total_entities": 28,
  "total_entities_with_Number_type": 28,
  "total_entities_with_Date_type": 0,
  "correct_entities": 17,
  "correct_entities_with_Number_type": 17,
  "correct_entities_with_Date_type": 0,
  "entity_accuracy": 0.61,
  "entity_score": 3.05,
  "overall_score": 5.05,
  "overall_explanation": "The structure had some
  discrepancies, mainly due to an extra row and
  misaligned headers. Entities were mostly matched
   with some incorrect matches, contributing to a
   medium overall score."}
```

C.2. Case Study: Human Alignment with LLM-as-Judge

We propose an LLM-as-Judge evaluation scheme that assesses financial OCR outputs by comparing model-generated HTML against gold-standard critical numerical and temporal facts.

To evaluate the robustness of the LLM-as-Judge procedure, we incorporate an expert-in-the-loop review process and conduct multiple human–LLM alignment assessments. We draw 2% of sample from the dataset and let the human expert independently compare each financial document image with the corresponding model-generated HTML and manually count all incorrectly transcribed numeric and temporal entities. In addition, they provide qualitative comments on *structural correctness* and *overall factual fidelity*, mirroring the evaluation categories produced by the LLM-as-Judge. The following case studies illustrate how human judgments align with the LLM-as-Judge across both high-accuracy and low-accuracy scenarios.

On Figures 5 and 6 show two representative cases demonstrating strong agreement between human experts and the LLM-as-Judge. Each case includes the raw page image provided to the model and the corresponding HTML reconstruction produced by the model.

High-accuracy case (FFA = 100%). On Figure 5, both the human expert and the LLM-as-Judge classify the OCR result as fully accurate. The LLM-as-Judge recognizes all numerical and temporal entities exactly after normalization and reports no mismatches.

Human expert review: We have our financial expert independently review the financial documents image and corresponding LLM output, who notes that all monetary values, percentages, and dates in the model output match the ground-truth document.

Human Expert Review (case of high FFA)

```
Wrong numeric entities: 0;
Wrong date entities: 0;
Structural: There are minor formatting differences
in HTML like font styles,table lining, and spacing;
Overall:
- The model preserves decimal precision
for reported amounts;
- Percentage signs are correctly captured;
- Fiscal quarter dates are fully captured.
```

Their detailed assessment matches the LLM-as-Judge reasoning, confirming perfect factual alignment.

Low-accuracy case (FFA = 61%). On Figure 6, both human experts and the LLM-as-Judge classify the output as factually unreliable. The LLM-as-Judge identifies omissions of quantity fields, missing date and duration entities, and hallucinated headers that do not appear in the source image.

Human expert explanation: Experts likewise report multiple critical errors:

Human Expert Review (case of low FFA)

```
Wrong numeric entities: 11;
Wrong date entities: 0;
Structural: There are column misalignment, and a
hallucinated heading for the table;
Overall:
- One numeric fields on the rightmost columns
of the tables are missing entirely;
- The model omitted the transaction date and
plan duration;
- certain monetary values are repeated or
misplaced across rows; and
- the model adds a chart title that does
not exist in the original document.
```

These observations correspond precisely to the error types surfaced by the LLM-as-Judge (omissions, positional mismatches, and hallucinations). Experts conclude that the factual reliability of the reconstruction is insufficient for financial analysis, aligning with the LLM-as-Judge's evaluation.

Summary The human explanations and the LLM-as-Judge output converge on the same factual judgments and error categories. This consistency provides empirical evidence that the LLM-as-Judge operates with human-like sensitivity to semantically meaningful financial errors, supporting its use as a scalable and interpretable evaluation tool for financial OCR.

D. Limitations

While FinCriticalED advances fact-level evaluation for financial OCR, several limitations remain.

First, the FinCriticalED dataset primarily focuses on U.S. financial documents with relatively standardized visual conventions; broader coverage of international filings, multilingual layouts, and handwritten annotations may reveal additional challenges not captured here.

Second, our annotation pipeline relies on rendered HTML as the structural reference, which, although stable and reproducible, may differ from native PDF or scanned-document artifacts encountered in real-world OCR deployments.

Third, FinCriticalED evaluates factual correctness at the entity level but does not yet consider cross-page linking, hierarchical financial relationships, or large-context numerical grounding across multi-document collections.

These limitations suggest promising directions for future work, including expanding to multilingual financial datasets, incorporating human–machine hybrid evaluation for edge cases, and exploring richer reasoning benchmarks

Highest Performance Quarter (2Q 2020): 32.22% Lowest Performance Quarter (2Q 2022): -23.70%

Average Annual Total Returns *For the calendar year ended December 31, 2024*	*1 year*	*5 years*	*10 years*	*Since Inception*	*Inception Date*
Investor Class Return Before Taxes	29.55%	18.27%	16.44%	—	11/02/1981
Return After Taxes on Distributions	28.52%	16.86%	15.00%	—	11/02/1981
Return After Taxes on Distributions and Sale of Fund Shares	18.28%	14.57%	13.44%	—	11/02/1981
I Class Return Before Taxes	29.81%	18.50%	16.67%	—	11/14/1996
Y Class[1] Return Before Taxes	29.99%	18.68%	16.85%	—	04/10/2017
A Class Return Before Taxes	21.78%	16.58%	15.46%	—	10/02/1996
C Class[2] Return Before Taxes	28.25%	17.09%	15.45%	—	10/29/2001
R Class Return Before Taxes	28.89%	17.68%	15.86%	—	08/29/2003
R5 Class[3] Return Before Taxes	29.81%	18.51%	16.67%	—	04/10/2017
R6 Class Return Before Taxes	29.99%	18.68%	16.85%	—	07/26/2013
G Class Return Before Taxes	30.70%	19.37%	—	20.01%	08/01/2019
Russell 1000® Index[4] (reflects no deduction for fees, expenses or taxes)	24.51%	14.28%	12.87%	—	—
Russell 1000® Growth Index (reflects no deduction for fees, expenses or taxes)	33.36%	18.96%	16.78%	—	—

(a) Part of a quarterly portfolio management results report

Highest Performance Quarter (2Q 2020): 32.22%
Lowest Performance Quarter (2Q 2022): -23.70%

Average Annual Total Returns

For the calendar year ended December 31, 2024

	1 year	5 years	10 years	Since Inception	Inception Date
Investor Class Return Before Taxes	29.55%	18.27%	16.44%	—	11/02/1981
Return After Taxes on Distributions	28.52%	16.86%	15.00%	—	11/02/1981
Return After Taxes on Distributions and Sale of Fund Shares	18.28%	14.57%	13.44%	—	11/02/1981
I Class Return Before Taxes	29.81%	18.50%	16.67%	—	11/14/1996
Y Class[1] Return Before Taxes	29.99%	18.68%	16.85%	—	04/10/2017
A Class Return Before Taxes	21.78%	16.58%	15.46%	—	10/02/1996

(b) Corresponding LLM output in HTML

Figure 5. Human alignment with LLM-As-Judge paradigm in high FFA case (FFA=100%)

[1] *Purchases of $1 million or more may be subject to a contingent deferred sales charge of 1.00% if the shares are redeemed within one year of the date of the purchase.*

[2] *The advisor has agreed to waive a portion of the fund's management fee such that the management fee does not exceed 0.887% for Investor, A, C and R Classes, 0.687% for I and R5 Classes, and 0.537% for Y and R6 Classes. The advisor expects this waiver arrangement to continue until February 28, 2026 and cannot terminate it prior to such date without the approval of the Board of Directors.*

[3] *The advisor has agreed to waive the G Class's management fee in its entirety. The advisor expects this waiver to remain in effect permanently and cannot terminate it without the approval of the Board of Directors..*

Example

The example below is intended to help you compare the costs of investing in the fund with the costs of investing in other mutual funds. The example assumes that you invest $10,000 in the fund for the time periods indicated and then redeem all of your shares at the end of those periods and that you earn a 5% return each year. The example also assumes that the fund's operating expenses remain the same, except that it reflects the rate and duration of any fee waivers noted in the table above. Although your actual costs may be higher or lower, based on these assumptions your costs would be:

	1 year	*3 years*	*5 years*	*10 years*
Investor Class	$91	$291	$507	$1,129
I Class	$71	$228	$398	$892
Y Class	$55	$180	$316	$711
A Class	$685	$923	$1,180	$1,911
C Class	$192	$601	$1,035	$2,044
R Class	$142	$447	$774	$1,698
R5 Class	$71	$228	$398	$892
R6 Class	$55	$180	$316	$711
G Class	$0	$0	$0	$0

(a) Part of financial statement investment plan explained

Class	Amount 1	Amount 2	Amount 3	
A	$180	$316	$711	
C	$685	$923	$1,180	$1,911
R	$192	$601	$1,035	$2,044
R5	$142	$447	$774	$1,698
R6	$71	$228	$398	$892
G	$55	$180	$316	$711
	$0	$0	$0	$0

(b) Corresponding LLM output in HTML

Figure 6. Human alignment with LLM-As-Judge paradigm in low FFA case (FFA=61%)

that unify OCR, layout understanding, and financial analysis.

E. Potential Risks and Misuse

FinCriticalED is designed to advance research on high-precision financial OCR, but several potential risks should be acknowledged.

First, although the benchmark focuses on publicly available financial documents, improved OCR techniques may enable more effective extraction of sensitive information from documents that were not intended for automated analysis. Responsible deployment requires ensuring that downstream systems respect privacy, regulatory requirements, and data-handling policies.

Second, the LLM-as-Judge evaluation framework could be misused as an authoritative decision-making tool rather than as an assessment mechanism. While effective for measuring fact-level OCR fidelity, it is not intended to replace professional financial auditing, compliance checks, or legally binding document verification.

Third, as with any dataset, FinCriticalED may embed domain-specific biases stemming from the geographic, regulatory, or formatting characteristics of the source documents. Models trained or tuned exclusively on this dataset may inadvertently overfit to these conventions and perform poorly on documents with different cultural, linguistic, or structural characteristics.

Finally, highly accurate OCR models may be used to automate large-scale extraction of financial data for questionable purposes, such as unauthorized scraping, adversarial market strategies, or amplification of misleading financial narratives. Researchers and practitioners should consider both the positive and negative downstream impacts when deploying systems built on top of FinCriticalED.

F. Ethical Considerations and Licensing

All documents originate from publicly available financial filings distributed under open-access or research licenses. No proprietary or confidential information was included. The dataset is intended solely for research on document understanding and factual accuracy; commercial deployment requires separate compliance verification. The benchmark complies with ACM and CVPR data ethics policies.