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.

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*, *required SEC filings*, *tax forms*, *securities transaction records*, and *financial legal documents*. To ensure annotation quality, each sampled filing must contain at least one type of critical fields.

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, filing year) is recorded and stored in structured JSON format. To ensure compatibility with OCR models, we further normalize text alignment, table borders, and visual regions. Empty or blank pages are automatically filtered out.

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 the five critical value types. 20% of the documents are independently labeled by two annotators to ensure inter-annotator agreement, with conflicts resolved by consensus. An internal agreement table reports consistent reliability across both entity types (see Table 4).

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.

Final Statistics and Accessibility. The final dataset contains 859 fully annotated financial documents, totaling 9,481 annotated entities.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 2.

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 FinCriticalED. Annotators followed a standardized set of rules to ensure consistency across financial documents.

B.1.1 Entity List.

- Number
- Temporal
- Monetary Unit
- Reporting Entity
- Financial Concept

B.1.2 General Rules.

- Identify all entities in the HTML file that belong to one of the five 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.
- Highlight **all valid entities** in each task without omission.

B.1.3 Guidelines on entity type definition and label instructions.

- **Number**
  - The numeric value should be **financially critical**, including percentages, monetary amounts, and share amounts.
  - Annotate only the number itself, or the number with attached signs, which may include decimal points, commas for thousands, and negative markers.
  - Do not include non-financially-related numbers, such as serial numbers or policy section numbers.
  - Examples: *1,000,000*, *2,345*, *0.37*, *10*, *1/3*, *-2.3*, *(10,234)*, *25.63%*. Do not annotate examples such as *Section 30(h)*.
- **Temporal (Duration and Dates)**
  - Only include specific dates and time periods. Do not include time frequency expressions such as *monthly* or *weekly*.
  - If the annotation target is a date range, annotate only the **date** itself and remove any context words between the two dates.
  - Examples: *March 24, 2025*, *1 month*, and *2 years* should be annotated.
  - For a span such as *PERIODS JANUARY 2, 2025 THRU JANUARY 8, 2025*, only annotate *JANUARY 2, 2025* and *JANUARY 8, 2025*; exclude *PERIODS* and *THRU*.
- **Monetary Unit**
  - Monetary-related units should be included.
  - Examples: *$*, *USD*, *thousands*, *millions*, *share*.
- **Reporting Entity**
  - If the entity, such as a company name, contains the full name followed by the abbreviation, annotate them together.
  - Example: *1st Franklin Financial Corporation ("1FFC")*.
  - Additional examples: *1st FRANKLIN FINANCIAL CORPORATION*, *1832 Asset Management L.P.*
  - Include the whole entity, e.g., *American Century Investment Management, Inc.*
- **Financial Concepts**

- Only label financial concepts if they appear inside a table or form. There is no need to label table titles or footnotes.
- Label the whole line item if it contains a financial concept.
- Example: *Maximum Sales Charge (Load) Imposed on Purchases (as a percentage of offering price).*
- Label a term as **Financial Concept** if it belongs to one of the following categories:
  1. **Revenue and Sales**
     Concepts describing inflows from core business activities.
     Examples: *Revenue*, *Net revenue*, *Gross revenue*, *Sales*, *Net sales*, *Service revenue*, *Total Revenue*.
  2. **Income, Profit, and Earnings**
     Concepts describing profitability or earnings outcomes.
     Examples: *Income*, *Net income*, *Operating income*, *Gross profit*, *Profit*, *Earnings*, *Earnings per share (EPS)*.
  3. **Costs, Expenses, and Losses**
     Concepts describing outflows, reductions, or negative performance.
     Examples: *Cost*, *Cost of revenue*, *Cost of sales*, *Operating expenses*, *Expense*, *Loss*, *XXX Fee*, *XXX Expenses*.
  4. **Taxes**
     Concepts related to taxation and tax-related outcomes.
     Examples: *Income tax*, *Tax expense*, *Deferred tax*, *Tax liability*, *Effective tax rate*.
  5. **Margins and Ratios**
     Common, high-level ratios directly tied to revenue or profit.
     Examples: *Gross margin*, *Operating margin*, *Profit margin*, *Net Income/Loss*.
  6. **Financial Obligations and Commitments**
     Concepts describing required or expected payments.
     Examples: *Lease obligation*, *Purchase obligation*, *Debt*, *Interest expense*, *Interest income*.
- **Explicit Exclusions (Do NOT label)**
  1. **Operational or Functional Activities**: *Research and Development*, *Marketing*, *Sales and distribution*, *General and administrative*.
  2. **Accounting or Reporting Metadata**: *Accounting policy*, *Notes to the financial statements*, *Segment information*, *Management discussion and analysis (MD&A)*.
  3. **Broad Business or Strategy Terms**: *Growth strategy*, *Market expansion*, *Customer acquisition*, *Product roadmap*.
  4. **Non-Financial Metrics or Qualitative Descriptions**: *Headcount*, *Employee engagement*, *ESG initiatives*.

B.2 Annotator Demography

FinCriticalED was annotated by a four-person team with complementary expertise in finance, economics, auditing, and computer technology. The team combines professional experience in financial analysis and FinTech with graduate-level training in business analytics, financial mathematics, computer science, and computer technology.

One annotator is a principal analyst at a major U.S. financial institution with academic training in business analytics, statistics, and economics, as well as experience in LLMs, financial data analysis, and multilingual reasoning. Another annotator has training in financial mathematics and computer science, together with more than seven years of experience in strategic finance and consulting in the FinTech industry. The remaining two annotators are graduate students in computer technology with backgrounds in auditing, financial analysis, data processing, annotation workflows, and LLM evaluation and adaptation for domain-specific tasks.

Overall, the team reflects a balance between financial domain knowledge and technical expertise, which supports accurate, consistent, and contextually grounded annotation across the dataset.

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 critical fields are highlighted and later enclosed within specialized `<span>` tags (e.g., `<span class="Number">...</span>`, `<span class="Time">...</span>`).

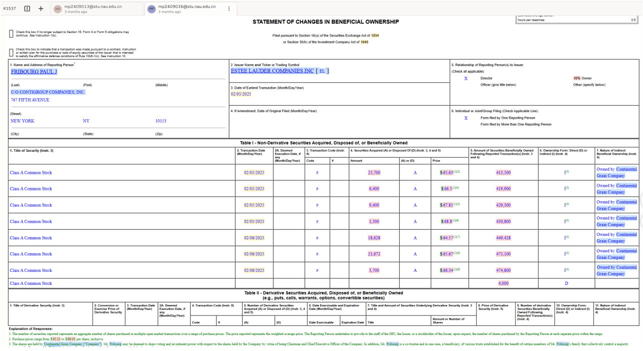
20% of the data is annotated independently by at least two annotators to ensure inter-annotator agreement for quality measurement. Upon completion, annotations are exported into JSON format, preserving the hierarchical document structure and linking each annotated entity to its source metadata. This 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 exported JSON schema.

Second, an **inter-annotator consistency check** evaluates pairwise agreement between annotators using both token-level and



#1537 mp2409013@stu.nau.edu.cn 3 months ago mp2409036@stu.nau.edu.cn 3 months ago

hours per response: 0.5

STATEMENT OF CHANGES IN BENEFICIAL OWNERSHIP

□ Check this box if no longer subject to Section 16. Form 4 or Form 5 obligations may continue. See Instruction 1(b).

□ Check this box to indicate that a transaction was made pursuant to a contract, instruction or written plan for the purchase or sale of equity securities of the issuer that is intended to satisfy the affirmative defense conditions of Rule 10b5-1(c). See Instruction 10.

Filed pursuant to Section 16(a) of the Securities Exchange Act of 1934
or Section 30(h) of the Investment Company Act of 1940

1. Name and Address of Reporting Person*
FRIBOURG PAUL J
(Last) (First) (Middle)
C/O CONTIGROUP COMPANIES, INC.
767 FIFTH AVENUE
(Street)
NEW YORK NY 10153
(City) (State) (Zip)

2. Issuer Name and Ticker or Trading Symbol
ESTEE LAUDER COMPANIES INC [EL]

3. Date of Earliest Transaction (Month/Day/Year)
02/05/2025

4. If Amendment, Date of Original Filed (Month/Day/Year)

5. Relationship of Reporting Person(s) to Issuer
(Check all applicable)
X Director 10% Owner
Officer (give title below) Other (specify below)

6. Individual or Joint/Group Filing (Check Applicable Line)
X Form filed by One Reporting Person
Form filed by More than One Reporting Person

Table I - Non-Derivative Securities Acquired, Disposed of, or Beneficially Owned

1. Title of Security (Instr. 3)	2. Transaction Date (Month/Day/Year)	2A. Deemed Execution Date, if any (Month/Day/Year)	3. Transaction Code (Instr. 8)		4. Securities Acquired (A) or Disposed Of (D) (Instr. 3, 4 and 5)			5. Amount of Securities Beneficially Owned Following Reported Transaction(s) (Instr. 3 and 4)	6. Ownership Form: Direct (D) or Indirect (I) (Instr. 4)	7. Nature of Indirect Beneficial Ownership (Instr. 4)
			Code	V	Amount	(A) or (D)	Price			
Class A Common Stock	02/05/2025		P		25,700	A	$ 65.63(1)(2)	413,500	I(3)	Owned by Continental Grain Company
Class A Common Stock	02/05/2025		P		6,400	A	$ 66.5(1)(4)	419,900	I(3)	Owned by Continental Grain Company
Class A Common Stock	02/05/2025		P		9,400	A	$ 67.81(1)(5)	429,300	I(3)	Owned by Continental Grain Company
Class A Common Stock	02/05/2025		P		1,500	A	$ 68.8(1)(6)	430,800	I(3)	Owned by Continental Grain Company
Class A Common Stock	02/06/2025		P		18,628	A	$ 64.57(1)(7)	449,428	I(3)	Owned by Continental Grain Company
Class A Common Stock	02/06/2025		P		21,672	A	$ 65.67(1)(8)	471,100	I(3)	Owned by Continental Grain Company
Class A Common Stock	02/06/2025		P		3,700	A	$ 66.34(1)(9)	474,800	I(3)	Owned by Continental Grain Company
Class A Common Stock								4,000	D	

Table II - Derivative Securities Acquired, Disposed of, or Beneficially Owned
(e.g., puts, calls, warrants, options, convertible securities)

1. Title of Derivative Security (Instr. 3)	2. Conversion or Exercise Price of Derivative Security	3. Transaction Date (Month/Day/Year)	3A. Deemed Execution Date, if any (Month/Day/Year)	4. Transaction Code (Instr. 8)		5. Number of Derivative Securities Acquired (A) or Disposed of (D) (Instr. 3, 4 and 5)		6. Date Exercisable and Expiration Date (Month/Day/Year)		7. Title and Amount of Securities Underlying Derivative Security (Instr. 3 and 4)		8. Price of Derivative Security (Instr. 5)	9. Number of derivative Securities Beneficially Owned Following Reported Transaction(s) (Instr. 4)	10. Ownership Form: Direct (D) or Indirect (I) (Instr. 4)	11. Nature of Indirect Beneficial Ownership (Instr. 4)
				Code	V	(A)	(D)	Date Exercisable	Expiration Date	Title	Amount or Number of Shares				

Explanation of Responses:
1. The number of securities reported represents an aggregate number of shares purchased in multiple open market transactions over a range of purchase prices. The price reported represents the weighted average price. The Reporting Person undertakes to provide to the staff of the SEC, the Issuer, or a stockholder of the Issuer, upon request, the number of shares purchased by the Reporting Person at each separate price within the range.
2. Purchase prices range from $ 65.11 to $ 66.03 per share, inclusive.
3. The shares are held by Continental Grain Company ("Company"). Mr. Fribourg may be deemed to share voting and investment power with respect to the shares held by the Company by virtue of being Chairman and Chief Executive Officer of the Company. In addition, Mr. Fribourg is a co-trustee and in one case, a beneficiary, of various trusts established for the benefit of certain members of Mr. Fribourg's family that collectively control a majority

Figure 2: Annotation interface used in FinCriticalED. Annotators highlight entities directly within HTML while referencing rendered page images for layout validation.

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 samples. In this phase, a senior financial expert randomly samples 5% of the annotations to assess factual correctness. 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** [6] 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:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{6}$$

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:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{7}$$

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 five entity categories. Statistics in table 4 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-as-Judge prompt

```
JUDGE_PROMPT_TEMPLATE = """
#Instruction: You are an expert OCR results inspector for financial
documents. You will be judging the quality of model generated HTML
results based on image inputs.
You will be given:
1. Ground truth HTML that contains special entity tags that label
information that are financially critical, including 5 types: Number,
Temporal, Monetary Unit, Reporting Entity, and Financial Concepts,
each entity are wrapped with special spans:
*Number: <number>. . . </number>
*Temporal: <temporal>. . . </temporal>
*Monetary Unit: <monetaryunit>. . . </monetaryunit>
*Reporting Entity: <reportingentity>. . . </reportingentity>
*Financial Concepts: <financialconcepts>. . . </financialconcepts>

    2. Model-generated HTML that was produced from the same image
    but does not contain those tags.

Your goal is to judge the quality of the financially critical
information by comparing it with the ground truth HTML.

Follow all steps carefully and return one JSON object as the final result.

# Step 1: Extract and match entities
    From the ground truth HTML, extract all entities
    explicitly enclosed by the special tags:
* <number>. . . </number>  ⟶ entity type = "Number"
* <temporal>. . . </temporal> ⟶ entity type = "Temporal"
* <monetaryunit>. . . </monetaryunit> ⟶ entity type =
"Monetary Unit"
* <reportingentity>. . . </reportingentity> ⟶ entity type =
"Reporting Entity"
* <financialconcepts>. . . </financialconcepts> ⟶ entity type =
"Financial Concepts"

Do not infer entities by meaning; only extract those wrapped
by these tags. Each entity record must include:
* type - "Number" ,”Temporal”,”Monetary Unit”,
”Reporting Entity”,”Financial Concepts”
* 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": "Temporal", "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_Temporal_type = count of GT Temporal entities
* total_entities_with_Monetary_Unit_type = count of
GT Monetary Unit entities
* total_entities_with_Reporting_Entity_type = count of
GT Reporting Entity entities
* total_entities_with_Financial_Concepts_type = count of
```

```
GT Financial Concepts 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_Temporal_type = the number of
entities with Temporal type that is correctly found
* correct_entities_with_Monetary_Unit_type = the number of
entities with Monetary Unit type that is correctly found
* correct_entities_with_Reporting_Entity_type = the number
of entities with Reporting Entity type that is correctly found
* correct_entities_with_Financial_Concepts_type = the number of
entities with Financial Concepts type that is correctly found
* entity_accuracy = correct_entities / total_entities * 100%
            (entity_accuracy should be a percentage number
            between 0% and 100%, rounded to two decimal places.)
    If an entity is partially matched (e.g., missing comma
    or currency symbol), mark it incorrect.

# Step 2: Overall judgment
    Add a 1-2 sentence short justification (overall_explanation)
    on the LLM outputs.

# Step 3: Output format
Output exactly one valid JSON object:
        {
"total_entities": 0,
"total_entities_with_Number_type": 0,
"total_entities_with_Temporal_type": 0,
“total_entities_with_Monetary_Unit_type”:0,
“total_entities_with_Reporting_Entity_type”:0,
“total_entities_with_Financial_Concepts_type”:0,
"correct_entities": 0,
"correct_entities_with_Number_type": 0,
"correct_entities_with_Temporal_type": 0,
"correct_entities_with_Monetary_Unit_type": 0,
"correct_entities_with_Reporting_Entity_type": 0,
"correct_entities_with_Financial_Concepts_type": 0,
"entity_accuracy": 0.00%,
"overall_explanation": “”
        }

Rules:
* Output only valid JSON (no extra text).
* Numbers must be numeric, not strings.
"""
```

LLM Judge output sample

```
{   "total_entities": 33,
    "total_entities_with_Number_type": 16,
    "total_entities_with_Temporal_type": 7,
    "total_entities_with_Monetary_Unit_type": 7,
    "total_entities_with_Reporting_Entity_type": 3,
    "total_entities_with_Financial_Concepts_type": 0,
    "correct_entities": 20,
    "correct_entities_with_Number_type": 10,
    "correct_entities_with_Temporal_type": 5,
    "correct_entities_with_Monetary_Unit_type": 5,
    "correct_entities_with_Reporting_Entity_type": 0,
    "correct_entities_with_Financial_Concepts_type": 0,
    "entity_accuracy": 60.61,
    "overall_explanation": "The model-generated output correctly
    captured a majority of Numbers, Temporals, and Monetary Units
    , but failed to match any Reporting Entities. The general
    format and layout aided in partial entity recognition."}
```

C.1 LLM-as-Judge with Human Alignment

Need to add new alignment case here after finishing LLM-as-Judge

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 3 and 4 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 3, 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.

```
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.

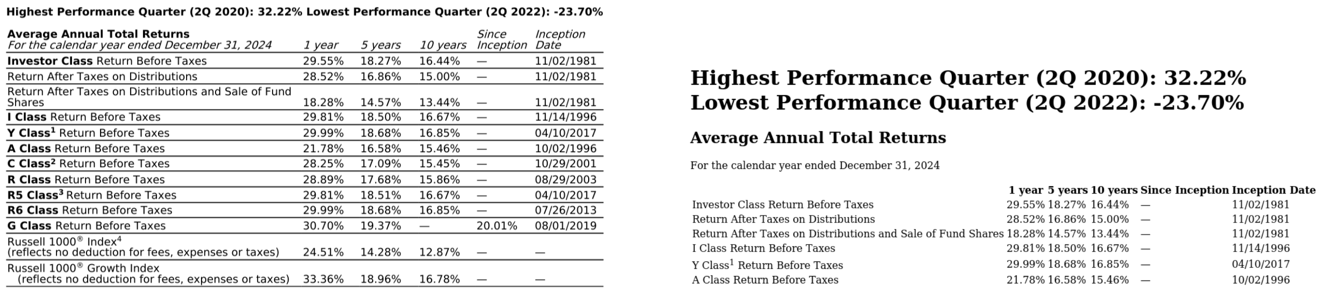
Low-accuracy case (FFA = 61%). On Figure 4, both human experts and the LLM-as-Judge classify the output as factually unreliable. The LLM-as-Judge identifies omissions of quantity fields, missing date and duration entities, and hallucinated headers that do not appear in the source image.

Human expert explanation: Experts likewise report multiple critical errors:

```
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.



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%

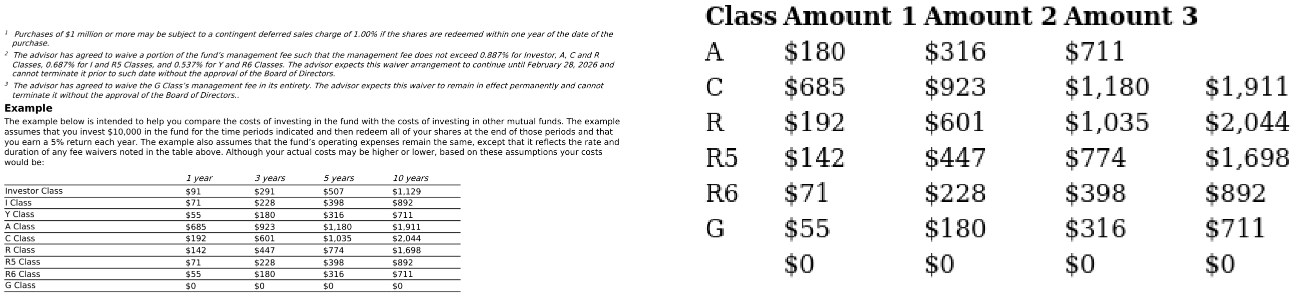
Average Annual Total Returns

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 3: 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 4: Human alignment with LLM-As-Judge paradigm in low FFA case (FFA=61%)

D Model Failure Analysis

D.1 Error Analysis Setup

We perform error analysis by comparing annotated ground-truth content with model-predicted OCR outputs after both are rendered in HTML format. Specifically, the ground-truth document and the corresponding model output are loaded into Label Studio, where annotators inspect them side by side and identify financially meaningful discrepancies. During review, annotators examine whether a model preserves both the local transcription and the financial meaning of each critical field under its rendered context.

We further organize the reviewed examples by document complexity and modality. Document complexity is divided into **low**, **medium**, and **high** complexity:

- **Low complexity**: [placeholder definition]
- **Medium complexity**: [placeholder definition]
- **High complexity**: [placeholder definition]

We also categorize examples by modality:

- **Table-only**: [placeholder definition]
- **Text-only**: [placeholder definition]
- **Mixed modality**: [placeholder definition]

Using this setup, we identify representative error cases and map them to the broader qualitative taxonomy introduced in Table 8.

D.2 Representative Model Error Cases

To complement the qualitative taxonomy in Table 8, we present a set of representative model failure cases. This section highlights a few financially consequential examples that illustrate the main recurring patterns observed across models, including header/period misalignment, hallucinated continuation, fixed lexical corruption, and format-sensitive transcription failures.

Case 1: Header/period misalignment. **Model:** [Model name]
Document type: [e.g., financial statement table]
Error pattern: The model correctly transcribes the value but attaches it to the wrong reporting period or column header, resulting in a semantically incorrect extraction despite plausible local OCR output. This corresponds to the *header/period misalignment* category in Table 8. Include fig as needed

Case 2: Hallucinated continuation. **Model:** [Model name]
Document type: [e.g., form or dense paragraph/table hybrid]
Error pattern: Triggered by a similar sentence prefix or repeated layout pattern, the model continues a previously seen phrase instead of transcribing the correct local content. This is a representative instance of *hallucinated continuation* in Table 8. Include fig as needed

Case 3: Fixed lexical or format-sensitive corruption. **Model:** [Model name]
Document type: [e.g., OCR-heavy filing page]
Error pattern: The model produces a repeated lexical corruption or

Table 8: Qualitative taxonomy of observed OCR failure patterns. OCR-specialized systems tend to show fixed and repetitive recognition errors, while general multimodal LLMs exhibit more variable and generative failure modes.

Error Type	Description	Common Models	Example
Header/period misalignment	Correct value is transcribed but attached to the wrong period, column, or local header context.	[General LLMs / placeholder]	[e.g., 2023 value aligned to 2024 column]
Hallucinated continuation	Similar sentence prefix causes the model to copy the continuation of a previously seen sentence instead of the correct local text.	[Llama-family / placeholder]	[placeholder example]
Fixed lexical corruption	The same token or word is repeatedly mistranscribed across pages.	[DeepSeek-OCR, MinerU / placeholder]	[e.g., transferor → transfessor]
Format-sensitive spacing	OCR errors consistently occur around dashes, punctuation, or spacing-sensitive formats.	[OCR-specialized models / placeholder]	[placeholder example]
Unstable free-form corruption	Errors vary substantially across samples, even under similar layouts or textual contexts.	[General LLMs / placeholder]	[placeholder example]

fails on spacing-/punctuation-sensitive formatting, such as symbols, dashes, decimals, or unit expressions. This aligns with the *fixed lexical corruption* or *format-sensitive spacing* categories in Table 8. Include fig as needed

Summary. These examples show that financially material OCR errors often arise not only from character-level recognition failures, but also from incorrect attachment of values to context, generative continuation artifacts, and brittle handling of formatting-sensitive expressions. Together with Table 8, these cases provide a concise qualitative view of how OCR failures manifest in FinCriticalED.

E 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 that unify OCR, layout understanding, and financial analysis.

F 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.

G 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 data ethics policies.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009