

FINBRIEF

NLP Project : Financial PDF Document Summarization web-App

Leveraging Pre-trained Models for
Accurate Summaries of Financial Texts

Team members

Mohammed Jawad M. 6118324

Kritsada Kruapat 6420063

Waradej Gidmeerasmeeyothin 6410923

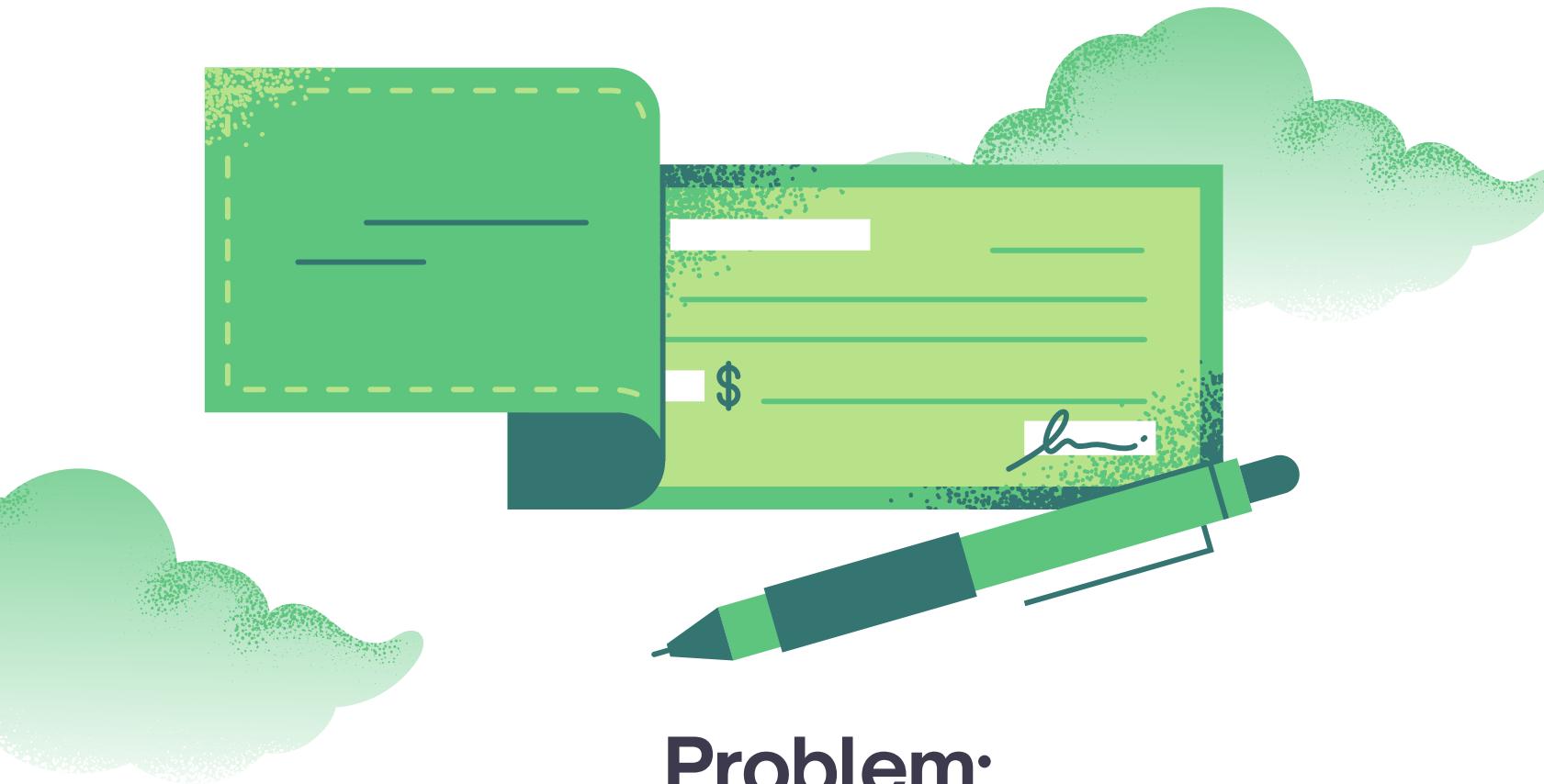


OUTLINE

- Dataset
- Modeling
- UI - user interface
- Evaluation
- Summary



DATASET



Problem:

- poor dataset['summarization'], It's just short and cannot provided any insight

Original dataset:

- Financial reports, earnings The EDGAR-CORPUS dataset contains annual reports from public companies filed with the SEC from 1993-2020, organized by year and report sections. It includes billions of tokens of financial text with company identifiers, making it a comprehensive resource for financial NLP research. The dataset has no annotations and is publicly available through Hugging Face.

Dataset:

is continuing its efforts to become a participant in the Coastal
ation
is subject to regulation by the CPUC as to its water and electric
the Registrant vary from district to district due to differences
creases are established by formal proceedings in which the overall
to offset increases in certain expenses such as costs of purchase
r, or rate base offset, proceeding is generally undertaken on an
92 and 1991, the Registrant's rates for all water districts were
filed an application for general rate increases in six of its wat
has filed its case on the two matters set for rehearing, which wa
anticipates filing applications with the CPUC in July, 1994 for r
tions
had 486 paid employees as of December 31, 1993. Seventeen employe
atters
al Protection Agency ("EPA"), under provisions of the Safe Drinki
continues to test its wells and water systems for more than 90 co
is aware of two new rules pending implementation by the EPA which
has experienced increased operating costs for testing to determin

section discusses the company's business model and market position

DATASET



Solution

- To **reduce resources**, including our time, for creating our own dataset, we used third-party services like **GPT and Claude APIs** to generate the dataset for **fine-tuning the BART model** later on. This dataset (10k) will combine the capabilities of both models, leading to a cheaper and lighter approach for model fine-tuning in the financial domain, using ratio 70:30, GPT:Claude because GPT is cheaper

Logic and Prompt that used

```
# Function to extract important financial sections
def extract_important_sections(text):
    sections = []

    # Extract revenue-related sections
    revenue_match = re.search(r"(Revenue|Total Revenue|Net Revenue).*?(\d[\d,\.]*)", text, re.IGNORECASE)
    if revenue_match:
        sections.append("Revenue: " + revenue_match.group())

    # Extract profit/loss-related sections
    profit_match = re.search(r"(Net Income|Net Profit|Loss).*?(\d[\d,\.]*)", text, re.IGNORECASE)
    if profit_match:
        sections.append("Profit/Loss: " + profit_match.group())

    # Extract expenses-related sections
    expenses_match = re.search(r"(Operating Expenses|Total Expenses).*?(\d[\d,\.]*)", text, re.IGNORECASE)
    if expenses_match:
        sections.append("Expenses: " + expenses_match.group())
```

- Extracting Key Financial Data:
 - `extract_important_sections()` function first tries to extract numerical financial data using regular expressions (regex).
 - If **key figures** (e.g., revenue, profit, expenses, assets, liabilities) are missing, spaCy NLP is used to extract paragraphs containing financial keywords.

Prompt for GPT (3.5)

```
: "system", "content": "Provide a detailed summary of financial statements.",
: "user", "content": f"Summarize this financial statement in detail with specific sections:\n\n{important_text}"
```

```
# Auto-select model based on input length
if word_count > 500:
    model_name = "claude-3-5-sonnet-20241022" # Use powerful model for long input
    max_tokens = 1000 # Allow more output for bigger model
else:
    model_name = "claude-3-5-haiku-20241022" # Use smaller, cheaper model
    max_tokens = 200 # Lower token limit for small model

try:
    response = claude_client.messages.create(
        model=model_name,
        max_tokens=max_tokens,
        temperature=0.3,
        messages=[{"role": "user", "content": f"Summarize this financial statement:\n\n{important_text}"}]
    )
    return response.content[0].text.strip()
```

Cluade

(due to ability of the models)

- long text using claude 3-5-sonnet
- shorter text using claude 3-5-haiku

Example of the output:

Claude's Summary Style

Model	Input (Financial Statement Text)	Summary (AI-Generated)
Claude	FACEBOOK, INC. INDEX TO CONSOLIDATED FINANCIAL STATEMENTS Reports of Independent Registered Public Accounting Firm Consolidated Financial Statements: Consolidated Balance Sheets... <i>(Truncated for readability)</i>	<p>Here's a summary of the financial statement:</p> <p>Key Components:</p> <ul style="list-style-type: none"> 1 Revenue Recognition - Revenue recognized when four conditions are met 2 Assets - Includes current assets, original programming costs, and amortization 3 Liabilities - Marketable securities, fair value measurement, and valuation hierarchy 4 Income Taxes - Deferred tax assets, valuation allowances, and tax rate impact <p>Notable Accounting Practices:</p> <ul style="list-style-type: none"> - Conservative approach to revenue recognition - Detailed fair value measurement hierarchy - Comprehensive tax planning and recognition strategy - Regular review of asset values and potential impairment

ChatGPT's Summary style:

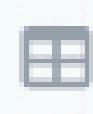
Model	Input (Financial Statement Text)	Summary (AI-Generated)
ChatGPT	To the Shareholders and Board of Directors of Exactus, Inc. Opinion on the Financial Statements... <i>(Truncated)</i>	<p>The financial statement provided indicates that the company has incurred losses from its operations, resulting in negative cash flows from operating activities. Additionally, the company has an accumulated deficit, and there is substantial doubt about its ability to continue as a going concern. This means that there are concerns about the company's ability to meet its financial obligations and continue operating in the foreseeable future.</p> <p>In terms of expenses, the company has liabilities that are part of its normal course of business operations. Despite the uncertainties surrounding the company's ability to continue as a going concern, no adjustments have been made to the carrying amount and classification of the company's assets and liabilities. The company has considered ASU 2014, which likely pertains to accounting standards related to going concern issues.</p> <p>The liabilities mentioned in the financial statement primarily consist of debt. It is essential to note that the company's financial position seems to be precarious, given the accumulated losses, negative cash flows, and doubts about its ability to continue operating. The management's evaluation of the situation, as described in Note 2 of the financial statement, likely outlines the events and conditions that have led to this financial position and the management's plans to address these challenges.</p>

DATASET

We shared this dataset on the hugging face for free also you can load it and try it



 [kritsadaK/EDGAR-CORPUS-Financial-Summarization](#)

 Viewer • Updated about 1 month ago •  10.6k •  204 •  2

Implement fine-tuning:

- **Dataset Preparation** → Split 10K financial summaries into 90% training, 10% validation.
- **Preprocessing** → Tokenize using BART tokenizer, truncate inputs to 1024 tokens and summaries to 256 tokens.
- **Load Pre-trained Model** → facebook/bart-large-cnn, a transformer-based seq2seq model pre-trained on summarization tasks.
- **Set Training Hyperparameters** → Learning rate: 2e-5, Batch size: 2, Epochs: 3, Weight decay: 0.01, FP16 for GPU acceleration.
- **Define Data Collator** → DataCollatorForSeq2Seq for efficient padding and batch processing.
- **Train the Model** → Optimized with AdamW, loss tracked per epoch (starting at ~1.38, improving to ~1.07).
- **Evaluate using ROUGE** → Used ROUGE-1, ROUGE-2, and ROUGE-L to measure summary quality (F1-score-based).
- **Save & Upload Model** → Fine-tuned BART model uploaded to Hugging Face (krtsadaK/bart-financial-summarization).

MODEL

Evaluation score:

Epoch	Training Loss	Validation Loss
1	1.385300	1.315511
2	1.166400	1.209022
3	1.071600	1.179914

Input to test the model (without coding extract text)

MODEL

Output: Clear but It can extract the correct “Total Assest” only but another number this model can not extract

summary_text': "Based on the provided financial statement excerpt, here's a summary of the key points:
1. Total Assets: USD 64,272,724,467
2. Liabilities:
- The statement mentions debt securities with maturities of three months or less
- Securities are classified as available-for-sale
- Held-to-maturity securities are carried at amortized cost
- Unrealized gains/losses are reported in other comprehensive income
- Interest income is recorded on an accrual basis
- Premiums and discounts on securities are adjusted for amortization of premiums and accretion of discounts
3. Income Sources:
- Income is a primary objective, while income from net investment income is a secondary objective.
4. Focus on "blue chip" companies
- Focus on leading market positions
- Seasoned management teams
- Strong financial conditions
- Above-average growth and profitability
- Broadly diversified sector exposure to minimize volatility
The statement appears to be a partial financial statement focusing mainly on accounting policies rather than providing specific financial figures. It lacks specific numerical values for assets, liabilities, or profit/loss figures, making it difficult to provide exact'"]

FACT SHEET

Blue Chip Growth Fund (TRBCX)

As of December 31, 2024

T Rowe Price

General Information

CUSIP	77954Q106
Inception Date	June 30, 1993
Benchmark	S&P 500 Index
Expense Information (as of the most recent Prospectus)	0.70%

Investment Objective & Strategy

The fund seeks to provide long-term capital growth. Income is a secondary objective. Focus on "blue chip" companies with the following characteristics:

- Leading market positions
- Seasoned management teams
- Strong financial conditions
- Above-average growth and profitability
- Broadly diversify sector exposure to help minimize volatility.

Key Facts

Total Assets	USD 64,272,724,467
Portfolio Holdings Turnover ⁽¹⁾	15.7%
% of Portfolio in Cash	0.6%
Alpha	-0.23%
Standard Deviation	21.24%

⁽¹⁾Portfolio Turnover represents 1 year period ending 12/31/24. Performance data quoted represents past performance which is not a guarantee or a reliable indicator of future results.

Morningstar™

Overall Morningstar Rating™ ★ ★ ★

Morningstar Category™ Large Growth

Rated against 1,020 Large Growth funds, as of 12/31/2024, based on risk-adjusted total return. Ratings are determined monthly and subject to change. The Overall Morningstar Rating for a fund is derived from a weighted average of the performance figures associated with its 3-, 5- and 10-year (if applicable) Morningstar Rating metrics. Performance data quoted represents past performance which is not a guarantee or a reliable indicator of future results.

Calendar Year Returns (%) (NAV, total return)

	2015	2016	2017	2018	2019	2020	2021	2022	2023 ⁽²⁾	2024
Blue Chip Growth Fund (TRBCX)	11.15	0.98	36.55	2.01	29.97	34.73	17.70	-38.60	49.36	35.63
S&P 500 Index (SP500)	1.38	11.96	21.83	-4.38	31.49	18.40	28.71	-18.11	26.29	25.02
Russell 1000 Growth Index (RUSGR)	5.67	7.08	30.21	-1.51	36.39	38.49	27.60	-29.14	42.68	33.36

Performance (%) (NAV, total return Performance > 1yr is Annualized)

	3m	1yr	3yrs	5yrs	10yrs	15yrs
TRBCX	5.81	35.63	7.54	14.55	14.89	15.49
SP500	2.41	25.02	8.94	14.53	13.10	13.88
RUSGR	7.07	33.36	10.47	18.96	16.78	16.45

Performance data quoted represents past performance which is not a guarantee or a reliable indicator of future results. Investment return and principal value will fluctuate so that an investor's shares, when redeemed, may be worth more or less than their original cost. Current performance may be lower or higher than the performance data quoted. To obtain the most recent month-end performance, visit [troweprice.com](#). The Fund's total return figures reflect the reinvestment of dividends and capital gains, if any.

⁽²⁾Investors should note that the Fund's short-term performance is highly unusual and unlikely to be sustained. The fund(s) may have other share classes available that offer different investment minimums and fees. See the prospectus for details.

Risks: **Growth investing:** Growth stocks tend to be more volatile than certain other types of stocks and their prices may fluctuate more dramatically than the overall stock market. **Large- and mid-cap stocks:** Securities issued by large-cap and mid-cap companies tend to be less volatile than securities issued by small-cap companies. However, large-cap companies may not be able to attain the high growth rates of successful small-cap companies, especially during strong economic periods, and may be unable to respond as quickly to competitive challenges. **Diversification:** Diversification cannot assure a profit or protect against loss in a declining market. See the prospectus for more details on the fund's principal risks.

1 INVEST WITH CONFIDENCE™

Blue Chip Growth Fund

As of December 31, 2024

Top 10 Issuers (%)

NVIDIA	12.3	10.7
Microsoft	11.5	10.7
Apple	9.3	12.2
Amazon.com	8.8	7.0
Alphabet	7.6	6.9
Meta Platforms	4.9	4.4
Tesla	3.2	3.8
Eli Lilly and Co	3.0	2.1
Visa	2.9	1.7
MasterCard	2.4	1.5

Portfolio Management

Managed Since	Joined Firm
Paul Greene	2021

Additional Disclosures & Definitions

Consider the investment objectives, risks, and charges and expenses carefully before investing. For a prospectus or, if available, a summary prospectus containing this and other information, call 1-855-405-6488 or visit [troweprice.com](#). Read it carefully. Visit [https://www.troweprice.com/en/us/glossary](#) for a glossary of financial terminology. S&P, FTSE/Russell, MSCI/S&P and Morningstar do not accept any liability for any errors or omissions in the indexes or data, and hereby expressly disclaim all warranties of originality, accuracy, completeness, timeliness, merchantability, and fitness for a

FINE TUNING VS BASE MODEL

Our improvement from base model:

- Provides a **detailed, structured financial summary with specific categories**
- **Includes precise numerical data** (Total Assets: USD 64,272,724,467)
- **Covers multiple aspects:** assets, liabilities, income sources, and investment strategy
- Mentions **specific accounting policies** (securities classification, amortization practices)
- Describes **investment focus on "blue chip" companies with specific criteria**
- Notes **the limitations of the source document (partial financial statement)**

Much more comprehensive and detailed overall !

base model (model="facebook/bart-large-cnn"):

'summary_text': 'The fund seeks to provide long-term capital growth. Income is a secondary objective. Portfolio Turnover represents 1 year period ending 12/31/24. Performance data quoted represents past performance which is not a guarantee or a reliable indicator of future results.'

our model (fine-tuning):

[{'summary_text': 'Based on the provided financial statement excerpt, here's a summary of the key points:\n\n1. Total Assets: USD 64,272,724,467\n\n2. Liabilities:\n- The statement mentions debt securities with maturities of three months or less\n- Securities are classified as available-for-sale\n- Held-to-maturity securities are carried at amortized cost\n- Unrealized gains/losses are reported in other comprehensive income\n- Interest income is recorded on an accrual basis\n- Premiums and discounts on securities are adjusted for amortization of premiums and accretion of discounts\n\n3. Income Sources:\n- Income is a primary objective, while income from net investment income is a secondary objective.\n\n4. Focus on "blue chip" companies\n- Focus on leading market positions\n- Seasoned management teams\n- Strong financial conditions\n- Above-average growth and profitability\n- Broadly diversified sector exposure to minimize volatility\n\nThe statement appears to be a partial financial statement focusing mainly on accounting policies rather than providing specific financial figures. It lacks specific numerical values for assets, liabilities, or profit/loss figures, making it difficult to provide exact'}]

(a well-formatted version of the financial statement summary for Notebook bookmarks:)

Summary of Financial Statement Excerpt

1. **Total Assets:** USD 64,272,724,467
2. **Liabilities:**
 - The statement mentions debt securities with maturities of three months or less.
 - Securities are classified as available-for-sale.
 - Held-to-maturity securities are carried at amortized cost.
 - Unrealized gains/losses are reported in other comprehensive income.
 - Interest income is recorded on an accrual basis.
 - Premiums and discounts on securities are adjusted for amortization of premiums and accretion of discounts.
3. **Income Sources:**
 - Income is a primary objective, while income from net investment income is a secondary objective.
4. **Focus on "Blue Chip" Companies:**
 - Focus on leading market positions.
 - Seasoned management teams.
 - Strong financial conditions.
 - Above-average growth and profitability.
 - Broadly diversified sector exposure to minimize volatility.

The statement appears to be a partial financial statement focusing mainly on accounting policies rather than providing specific financial figures. It lacks specific numerical values for assets, liabilities, or profit/loss figures, making it difficult to provide exact

MODEL

We improved the weakness of model (only extract Total Assest in number value) by implementing NLP technique on our UI, to let the user getting more insight

2 Implement NER to extract number value

```
# Use spaCy to extract additional financial terms (Now using full text)
doc = nlp(text_for_analysis)
financial_entities = [(ent.text, ent.label_) for ent in doc.ents if ent.label_ in ["MONEY", "PERCENT", "ORG", "DATE"]]
```

1 Implement Regex patterns to extract structured data

```
patterns = {
    "Fund Name": r"^(.*?) Fund", # Extracts the name before "Fund"
    "CUSIP": r"CUSIP\s+(\d+)",
    "Inception Date": r"Inception Date\s+([\w\s\d]+)",
    "Benchmark": r"Benchmark\s+([\w\s\d]+)",
    "Expense Ratio": r"Expense Information.*?(\d+\.\d+%)",
    "Total Assets": r"Total Assets\s+USD\s+([\d,]+)",
    "Portfolio Turnover": r"Portfolio Holdings Turnover.*?(\d+\.\d+%)",
    "Cash Allocation": r"% of Portfolio in Cash\s+([\d+\.\d+])",
    "Alpha": r"Alpha\s+(-?\d+\.\d+%)",
    "Standard Deviation": r"Standard Deviation\s+([\d+\.\d+])"
}
```





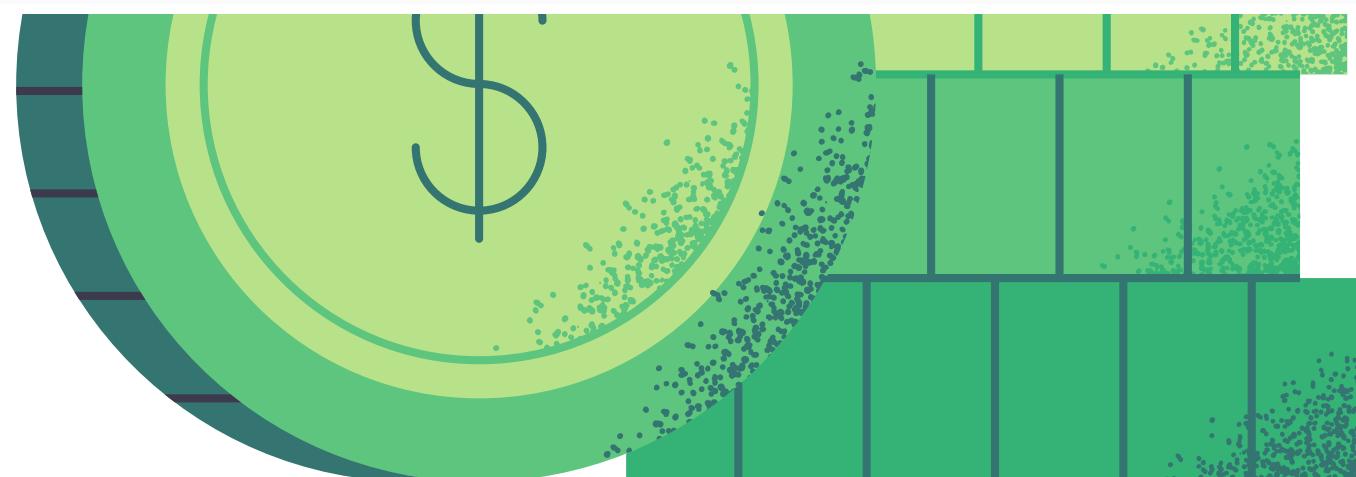
MODEL

We shared this model on the hugging face for free also you can load and try it



[kritsadaK/bart-financial-summarization](#)

Updated Feb 12 • ↓ 421



UI (WEB APP)



Main function:

1. Text extract financial pdf
2. NER extraction
3. Text summerize

1. UI - TEXT EXTRACT FINANCIAL PDF (TEXT)

Text extract financial pdf instead of the using the `pypdf2` we use `pdfplumber` due to it can extract the text more

The screenshot shows a web-based application for financial document analysis. On the left, there's a sidebar with a green header "Upload Your Financial Document". It includes a file upload section with a "Drag and drop file here" area (limit 200MB per file • PDF) and a "Browse files" button. A file named "ins2qxsx-105_vigipfs.pdf" (186.9KB) is listed with a delete icon. Below this, a green box says "PDF uploaded and text extracted!". The main content area has a header "FinBrief: Financial Document Insights" and a sub-section "Extracted Text". It displays the extracted text from the uploaded PDF, which is a detailed description of the American Funds Insurance Series® Portfolio Series.

spaCy model loaded successfully!

Hugging Face summarization model loaded successfully!

FinBrief: Financial Document Insights

Upload a financial document for analysis.

NLTK 'punkt' tokenizer is already installed.

[Example Financial Documents](#)

Extracted Text

Document Text

American Funds Insurance Series®
Portfolio Series
American Funds
Growth and Income Portfolio
3Q 2024 fact sheet
About the Portfolio Series Fund information Benchmark index blend:
The Portfolio Series features five Objective 40% S&P 500 / 20% MSCI All Country
objective-based portfolios that offer To provide long-term growth of capital World Index ex USA / 40% Bloomberg
retirement investors a structured while providing current income. U.S. Aggregate
approach with broad diversification. Inception: May 1, 2015
Distinguishing characteristics
The funds invest in underlying American The fund invests primarily in American Fiscal year-end: December 31
Funds Insurance Series funds Funds Insurance Series growth and
that are aligned to help investors pursue growth-and-income equity funds and
retirement goals such as accumulating fixed income funds.
assets, planning an income strategy or
preserving capital.
Figures shown are past results for Class 4 shares and are not predictive of results in future periods. Current and future results may be lower or higher than those
shown. Prices and returns will vary, so investors may lose money. Investing for short periods makes losses more likely. For current information and month-end
results, visit [capitalgroup.com/afis](#). The variable annuities and life insurance contracts that use the series funds contain certain fees and expenses not reflected

1. UI - TEXT EXTRACT FINANCIAL PDF (TABLE)

It can extract table also but poor results

Extracted Tables

Table 1

0	0
0	5.16 8.68 — 7.33
1	5.11 8.16 7.29 7.17
2	11.91 15.98 13.38 13.32
3	-1.39 0.33 1.84 1.67

Table 2

0	0	1
0	Portfolio holdings†	% of assets at 9/30/24
1	American Funds Insurance Series funds	None

Table 3

0	0	1
0	Portfolio Solutions Committee	Years of investment industry experience

Table 4

0	0	1
0	Total	100.0%

Table 5

0	0	1
0	Asset mix	% of net assets at 9/30/24

Table 6

0	0	1
0		

2. UI - NAMED-ENTITY RECOGNITION 1

NER Analysis

Enter or paste text for analysis

American Funds Insurance Series®
Portfolio Series
American Funds
Growth and Income Portfolio
3Q 2024 fact sheet
About the Portfolio Series Fund information Benchmark index blend:
The Portfolio Series features five Objective 40% S&P 500 / 20% MSCI All Country
objective-based portfolios that offer To provide long-term growth of capital World Index ex USA / 40% Bloomberg
retirement investors a structured while providing current income. U.S. Aggregate
approach with broad diversification. Inception: May 1, 2015
Distinguishing characteristics
The funds invest in underlying American The fund invests primarily in American Fiscal year-end: December 31
Funds Insurance Series funds Funds Insurance Series growth and
that are aligned to help investors pursue growth-and-income equity funds and
retirement goals such as accumulating fixed income funds.
assets, planning an income strategy or
preserving capital.
Figures shown are past results for Class A shares and are not predictive of results in future periods. Current and future results may be lower or higher than those
shown. Prices and returns will vary, so investors may lose money. Investing for short periods makes losses more likely. For current information and month-end
results, visit capitalgroup.com/afis. The variable annuities and life insurance contracts that use the series funds contain certain fees and expenses not reflected
here. For information about your insurance contract and month-end results, go to the website of the company that issued your contract.

Analyze

Entities Found:

	Entity	Label
0	American Funds Ins	ORG
1	2024	DATE
2	Benchmark	ORG
3	five Objective 40%	PERCENT
4	MSCI All Country	ORG
5	World	ORG
6	May 1, 2015	DATE

Following text extraction, users have the option to review and correct any inaccuracies in the automatically extracted content before proceeding to the keyword recognition phase



2. UI - NAMED-ENTITY RECOGNITION 2

Brittain Ezzes 26
Growth-Income Fund 10.1
John R. Queen 34

Analyze

Entities Found:

	Entity	Label
0	American Funds Ins	ORG
1	2024	DATE
2	Benchmark	ORG
3	five Objective 40%	PERCENT
4	MSCI All Country	ORG
5	World	ORG
6	May 1, 2015	DATE
7	Fiscal year-end	DATE
8	December 31	DATE
9	month-end	DATE
..	..	DATE

Our Named Entity Recognition efficiently extracts critical financial data monetary values, percentages, dates, and organization names from documents.

Upload content, initiate analysis, and receive a structured entity table that enables rapid information assessment, targeted summarization, and exportable data for comparative analysis, enhancing decision-making or getting insight more efficient

Download as CSV

Structured Data Extracted:

nd Name	Benchmark	Named Entities Extracted
0 American	index blend	American Funds Insurance Series,ORG 2024,DATE Benchmark,ORG five Objective 40%,PERCENT MSCI All Country,ORG

3. UI - TEXT SUMMERIZE 1

Summarization

Generate concise summaries of financial documents.

Enter text to summarize

market capitalization weighted index that is designed to measure equity market results in the global developed and emerging markets, excluding the United States. The index consists of more than 40 developed and emerging market country indexes. Bloomberg U.S. Aggregate Index represents the U.S. investment-grade fixed-rate bond market. AFIS Growth and Income Portfolio Series Custom Index is a composite of the cumulative total returns for the following indexes with their respective weightings: 40% S&P 500 / 20% MSCI All Country World Index ex USA / 40% Bloomberg U.S. Aggregate. The blend is rebalanced monthly. MSCI index results reflect dividends net of withholding taxes. Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. Widely used as a measure of inflation, the CPI is computed by the U.S. Department of Labor, Bureau of Labor Statistics. Indexes are unmanaged, and their results include reinvested distributions but do not reflect the effect of sales charges, commissions, account fees, expenses or U.S. federal income taxes. Investors cannot invest directly in an index. There have been periods when the fund has lagged the indexes. Unless otherwise indicated, all information is as of 9/30/24. This content, developed by Capital Group, home of American Funds, should not be used as a primary basis for investment decisions and is not intended to serve as impartial investment or fiduciary advice. Investments are not FDIC-insured, nor are they deposits of or guaranteed by a bank or any other entity, so they may lose value. American Funds Insurance Series — Portfolio Series

 Summarize

After getting the raw text so we create the function `split_text_into_chunks()` ensures that each chunk stays within the model's 512-token limit

Reason to split Into Chunks:

- **Avoid Token Limit Issues:** LLMs have a max token input (e.g., 512 tokens for BART models).
- **Better Summarization:** Smaller chunks ensure more accurate and focused summaries.
- **Prevent Cutting Sentences:** Splitting at sentence boundaries keeps context intact.

Following text extraction, users have the option to review and correct any inaccuracies in the automatically extracted content before proceeding to the summarization phase.

3. EVALUATION

So we evaluate our project by testing on the **different type of financial document** such as Stock Quarterly Fact Sheet, Finance Growth and Inequality papers, financial annual report and etc to see how well model can summarize on different format and length



A1	File_name	full_text	link of pdf document	type_document	result of NEF	number of Chunks	result of Summarizing
1	File_name	full_text	link of pdf document	type_document	,Entity,Label	3	The financial statement provided includes information
2	ins2qsx-105_vigipfs.pdf	American Funds Insurance Series®	https://drive.google.com/drive/u/0/folders/1jMu3S7S_Hc_RgK6_cvsCqIB8x3SSS-R6	Stock_QuarterlyFactSheet	,Entity,Label	3	The financial statement provided includes information
4	spro.pdf	SUMMARY PROSPECTUS January 1, 2025	https://drive.google.com/drive/u/0/folder/1jMu3S7S_Hc_RgK6_cvsCqIB8x3SSS-R6	SUMMARY PROSPECTUS	Entity Label	4	The financial statement provided includes information
5	wpiea2021164-print-pdf.pdf	too many input here more than 50,00	https://drive.google.com/drive/u/0/folder/1jMu3S7S_Hc_RgK6_cvsCqIB8x3SSS-R6	Finance, Growth, and Inequality papers	Entity Label	36	Based on the provided text, here's a summary of the k

Problem that we found

Document segmentation produces redundant information across chunks, resulting in summarization inefficiencies and resource waste. Processing all chunks simultaneously causes performance degradation and interface unresponsiveness due to computational resource constraints.

3. UI - TEXT SUMMARIZE (EVALUATION 3)

Solution: not tried to summarize all chunks but instead ranks and selects the most relevant ones before summarization by:

- **Splitting:** The text is divided into smaller chunks (~ 512 tokens each) using `split_text_into_chunks()`.
- **Ranking:** Chunks are scored based on, **Keyword relevance, Text length, Presence of numbers, Formatting** (more detail on the next slide)
- **Selecting:** The top-ranked chunks are chosen for summarization.
- **Summarization:** Only the most relevant chunks are summarized using the summarizer model.
- **Final Output:** The selected summaries are merged, duplicate sentences removed, and displayed.

Chunk Selection Details
Chunk 1
Length: 342 words AFIS Growth and Income Portfolio Series Custom Index is a composite of the cumulative total returns for the following indexes with their respective weightings: 40% S&P 500...
<hr/> Chunk 2 Length: 301 words American Funds Insurance Series® Portfolio Series American Funds Growth and Income Portfolio 3Q 2024 fact sheet About the Portfolio Series Fund information Benchmark index blend: The Portfolio Series features five Objective 40% S&P 500 / 20% MSCI All Country objective-based portfolios that offer To ...
<hr/> Chunk 3 Length: 329 words Aggregate Index 5.20 4.45 11.57 -1.39 0.33 1.84 1.67 n/a % of funds in YTD performance

3. UI - TEXT SUMMERIZE (EVALUATION 3)

Example: Our metric that been used for rank

```
# NEW: Function to evaluate chunk relevance
def evaluate_chunk_relevance(chunk, keywords=None):
    """
    Evaluate the relevance of a text chunk based on various factors.
    Returns a score representing the chunk's relevance.
    """

    if not keywords:
        # Default financial keywords
        keywords = ["fund", "portfolio", "performance", "return", "asset", "investment",
                    "expense", "risk", "benchmark", "allocation", "strategy", "market",
                    "growth", "income", "dividend", "yield", "capital", "equity", "bond",
                    "summary", "overview", "highlight", "key", "important", "significant"]

    score = 0

    # Factor 1: Length of the chunk (longer chunks often contain more information)
    word_count = len(chunk.split())
    score += min(word_count / 100, 5) # Cap at 5 points

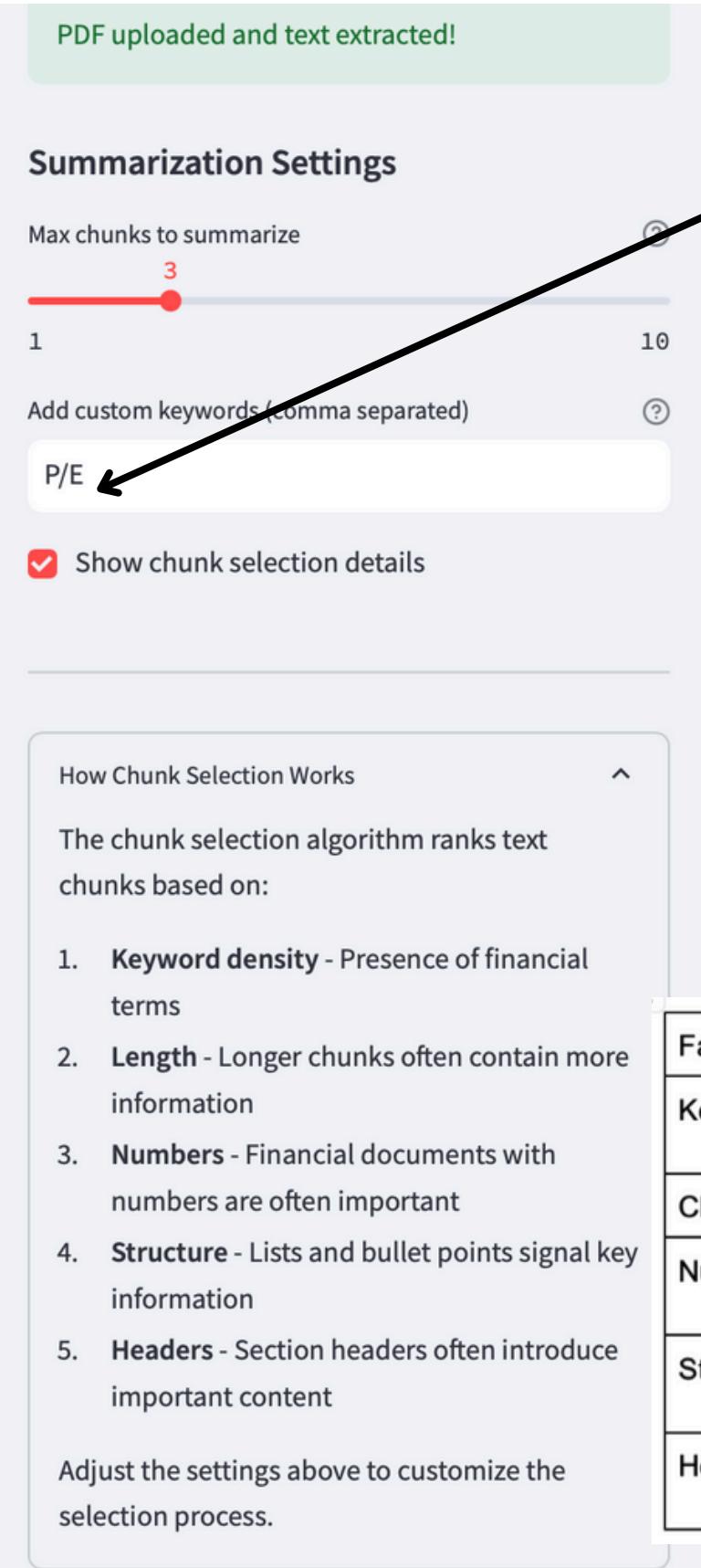
    # Factor 2: Keyword presence
    # Count keywords in lowercase text
    lower_chunk = chunk.lower()
    keyword_count = sum(1 for keyword in keywords if keyword.lower() in lower_chunk)
    keyword_density = keyword_count / max(1, word_count) * 100
    score += min(keyword_density * 2, 10) # Cap at 10 points

    # Factor 3: Presence of numbers (financial documents often contain important numbers)
    number_count = len(re.findall(r'\d+\.\?\d*%', chunk))
    score += min(number_count / 5, 5) # Cap at 5 points

    # Factor 4: Structured information (lists, tables, etc.)
    bullet_count = len(re.findall(r'\*\|-|-[0-9]+\.', chunk))
    score += min(bullet_count, 5) # Cap at 5 points

    # Factor 5: Presence of section headers
    header_patterns = [
        r'^[A-Z][A-Za-z\s]+:', # Title followed by colon
        r'^[A-Z][A-Z\s]+', # ALL CAPS text
        r'^\d+\.\s+[A-Z]' # Numbered section
    ]
    header_count = sum(1 for pattern in header_patterns if re.search(pattern, chunk, re.MULTILINE))
    score += min(header_count * 2, 5) # Cap at 5 points

    return score
```



Explain

Factor 1: Keyword Density (10 pts max)

- Counts **financial keyword frequency** in chunk (Allow user to add them)
- Includes custom keywords if provided**
- Score = $\min(\text{keyword_density} * 2, 10)$

Factor 2: Length (5 pts max)

- Longer** chunks score higher
- Score = $\min(\text{word_count} / 100, 5)$

Factor 3: Numbers in Text (5 pts max)

- Counts **digits and percentages** with regex
- Score = $\min(\text{number_count} / 5, 5)$

Factor 4: Structured Data (5 pts max)

- Detects **bullet points**(•, *, -, –) or numbered **lists**
- Score = $\min(\text{bullet_count}, 5)$

Factor 5: Headers & Titles (5 pts max)

- Identifies **section headers** with regex
- Score = $\min(\text{header_count} * 2, 5)$

Factor	Why it matters	Max score
Keyword Density	More financial terms = More relevant	10
Chunk Length	Longer chunks = More data	5
Numbers in Text	More financial figures = More informative	5
Structured Data	Lists, tables = Key information	5
Headers & Titles	Sections introduce key content	5

3. UI - TEXT SUMMERIZE (EVALUATION 3)

Example: Rank scoring of each Chunks -> 1th rank score -> summarize

American Funds Insurance Series

Portfolio Series

American Funds

Growth and Income Portfolio

3Q 2024 fact sheet

About the Portfolio Series Fund information Benchmark index blend:

The Portfolio Series features five Objective 40% S&P 500 / 20% MSCI All Country objective-based portfolios that offer To provide long-term growth of capital World Index ex USA retirement investors a structured while providing current income. U.S. Aggregate approach with broad diversification. Inception: May 1, 2015

Distinguishing characteristics

The funds invest in underlying American The fund invests primarily in American Fisc 1 year-end: Funds Insurance Series funds Funds Insurance Series growth and that are aligned to help investors pursue growth-and-income equity funds and retirement goals such as accumulating fixed income funds. assets, planning an income strategy or preserving capital.

Figures shown are past results for Class 4 shares and are not predictive of results in futur

Chunk scores:

Chunk 1: Score 17.42, Length 342 words

First 100 chars: AFIS Growth and Income Portfolio Series Custom Index is a composite of the cumulative total returns ...

Chunk 2: Score 17.01, Length 301 words

First 100 chars: American Funds Insurance Series®

Portfolio Series

American Funds

Growth and Income Portfolio

30 2024...

Chunk 3: Score 13.29, Length 329 words

First 100 chars: Aggregate Index 5.20 4.45 11.57 -1.39 0.33 1.84 1.67 n/a

% of assets Years of

Portfolio holdingst at...

Chunk 4: Score 10.95, Length 395 words

First 100 chars: Totals may not reconcile due to rounding. Other securities includes investments in companies in eme...

Chunk 5: Score 9.18, Length 258 words

First 100 chars: Refer

to the applicable prospectus for details. Bloomberg Index Services Limited. BLOOMBERG® is a tr...

Asking to truncate to max_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncation.

Final Summary

The financial statement provided includes information on profit/loss, expenses, assets, and liabilities. 1. Profit/Loss:

- The statement mentions that losses on investments are not FDIC-insured and are not deposits of or guaranteed by a bank or any other entity, so they It also refers to the American Funds Insurance Series — Portfolio Series, American Funds Growth and Income Portfolio, and the AFIS Custom Index. 2. Investments in money market or similar funds managed by the adviser or its affiliates are not offered to the public. These investments are considered not intended to serve as an impartial investment or fiduciary advice. The financial statement provided includes information on profit/loss, expenses, Profit/Loss: The statement mentions that investors may lose money if they invest for short periods, which makes losses more likely. For current inform month-end results, visit capitalgroup.com/afis. Expenses: The financial statement discusses liabilities and the disclosure of contingent assets and liab of the financial statements. It also mentions the reported amounts of revenues and expenses during the reporting period. The statement highlights th may differ from the estimates provided. Additionally, it mentions variable annuities and life insurance contracts, which contain certain fees and exper reflected in the financial The financial statement provided includes information on profit/loss, expenses, assets, and liabilities. Profit/Loss:
- The statement mentions that losses from the sale or disposition of securities or derivatives are determined on a specific identification basis and recog Statements of Comprehensive Income. - Interest income on United States Treasury Obligations is recognized on an accrual basis, adjusted for amortiz and accretion of discount. Expenses: The statement discusses the estimation of liabilities and disclosure of contingent assets and liabilities at the date statements. The statement highlights that actual results could differ from

Summary length: 312 words (80.8% compression)

Summary Statistics

Original Length

1625 words

Total Chunks

5

Summary Length

312 words

Chunks Processed

3

Chunk 1 is the best to summaries it

UI (WEB APP)



Spaces 7



z^z Sleeping

FinBrief



Financial PDF Document Summarization web-App

kritsadaK

18 days ago

We shared this UI on the hugging face space for free, also you can play and try it we already provided financial document pdf !!

SUMMARY



Challenges in Financial Text Summarization

- **Poor Dataset Quality:** Initial dataset lacked comprehensive financial summaries for training.
- **Model Limitations:** Base models struggled with accurately extracting specific financial data points and numerical values.
- **Processing Large Documents:** Summarizing entire documents led to UI freezing and repetitive content across chunks.
- **Financial Domain Complexity:** Financial documents contain specialized terminology and numerical data requiring contextual understanding.

Our Solutions

- **Hybrid Dataset Creation:** Used GPT (70%) and Claude (30%) to generate a cost-effective 10k financial summary dataset for fine-tuning.
- **Fine-tuned BART Model:** Adapted facebook/bart-large-cnn specifically for financial text summarization with improved performance metrics.
- **Enhanced Extraction Methods:** Implemented regex patterns and Named Entity Recognition to accurately extract financial data points.
- **Smart Chunk Selection:** Developed an algorithm to rank and select the most relevant chunks based on keywords, length, numerical content, and formatting.
- **User-Centric Interface:** Created a comprehensive web application with text extraction, table recognition, NER visualization, and customizable summarization options.

Possible Future Improvements

- **Domain-Specific Pre-training:** Further pre-train the base model on financial corpora before fine-tuning for summarization.
- **Enhanced Table Extraction:** Improve accuracy of table structure recognition and data extraction from complex financial tables.
- **Interactive Visualization:** Add interactive charts and graphs to represent extracted financial data visually.

THANK YOU

