MIXPEEK

Financial Document Intelligence

Capability Benchmark Report

Version	1.0
Date	December 2025
Test Corpus	SEC 10-K Filings (AAPL, MSFT, NVDA)
Total Chunks	3,847

Executive Summary

This report presents benchmark results for Mixpeek's financial document retrieval system. We measure three core capabilities: table extraction accuracy, calculation precision, and retrieval quality. All tests are conducted on real SEC 10-K filings with manually verified ground truth.

94.2%	96.3%	94%	<200ms
Table Extraction Accuracy	Calculation Precision	Retrieval P@3	Query Latency (P50)

What We Test: Specific capabilities for financial document processing—not general intelligence. For general Q&A;, use GPT-4 or Claude. For structured extraction from SEC filings, use Mixpeek.

Benchmark 1: Table Extraction Accuracy

Financial tables in SEC filings have complex structures: multi-level headers, merged cells, footnote references, and inconsistent formatting. We measure cell-level extraction accuracy using TableFormer against standard alternatives.

Methodology

- **Test Set:** 50 tables from AAPL, MSFT, NVDA 10-K filings (FY2023-2024)
- **Table Types:** Income statements, balance sheets, segment breakdowns, quarterly summaries
- **Validation:** Manual cell-by-cell verification against source PDFs
- **Metric:** (Correctly extracted cells / Total cells) × 100%

Results

Extraction Method	Cell Accuracy	Header Detection	Merged Cells	Footnotes
Mixpeek (TableFormer)	94.2%	98.1%	91.3%	89.7%
Google Document AI	86.1%	91.2%	73.8%	71.2%
AWS Textract	82.4%	88.5%	68.2%	65.8%
Tesseract OCR	62.3%	71.5%	38.2%	32.1%

What Makes Financial Tables Hard

- **Multi-level headers:** "2024" spanning Q1-Q4 columns
- **Merged cells:** "Total Revenue" spanning multiple rows
- **Footnote markers:** "Revenue (1)" requiring reference resolution
- **Number formatting:** $(1,234) vs -1234 vs (1,234)

Example: AAPL Quarterly Revenue Table

Source: Apple Inc. 10-K FY2024, Page 31 — Products and Services Performance

	Q1 2024	Q2 2024	Q3 2024	Q4 2024	FY 2024
iPhone	$69.7B	$45.9B	$39.3B	$46.2B	$201.2B
Mac	$7.8B	$7.5B	$7.0B	$7.7B	$30.0B
iPad	$7.0B	$5.6B	$7.2B	$6.9B	$26.7B
Services	$23.1B	$23.9B	$24.2B	$25.0B	$96.2B

Our Output: Structured JSON with cell values, data types, row/column indices, and bounding box coordinates for each cell. Every extracted value links back to its exact location in the source PDF.

Benchmark 2: Numerical Calculation Precision

Financial analysis requires exact calculations—not estimates. Our system extracts numbers from documents, generates Python code to perform calculations, executes in a sandbox, and returns verified results with full audit trail.

Methodology

- **Test Set:** 80 calculation queries requiring document context
- **Categories:** YoY growth, margins, ratios, multi-step analysis
- **Validation:** Ground truth calculated from source documents
- **Metric:** Exact match (0% tolerance for financial figures)

Results by Calculation Type

Calculation Type	Test Cases	Accuracy	Example
Simple Arithmetic	20	100%	Revenue - COGS = Gross Profit
YoY/QoQ Growth	20	100%	(Rev_2024 - Rev_2023) / Rev_2023
Margin Calculations	20	95%	Operating Income / Revenue
Multi-Step Analysis	20	90%	Segment growth vs company avg
Overall	80	96.3%	—

How It Works: Code-Verified Calculations

Unlike LLM-based estimation, our system generates and executes verifiable Python code:

```
Query: "What was Apple's YoY iPhone revenue growth in FY2024?"

Step 1: Retrieve relevant chunks from indexed 10-K
Step 2: Extract values: iPhone_2024 = $201.2B, iPhone_2023 = $200.6B
Step 3: Generate code: growth = ((201.2 - 200.6) / 200.6) * 100
Step 4: Execute in sandbox → Result: 0.30%
Step 5: Return answer + sources + calculation trace
```

Why Code Execution Matters

- **Auditability:** Every calculation has a verifiable code trace
- **Precision:** Exact numbers, not "approximately 18%"
- **Reproducibility:** Same inputs always produce same outputs
- **Source Attribution:** Every number links to source document + page + cell

Benchmark 3: Retrieval Precision

Retrieval quality determines whether the right information reaches the answer generation step. We measure Precision@K: the percentage of queries where the correct answer appears in the top K results.

Methodology

- **Test Set:** 200 queries with manually labeled ground truth chunks
- **Query Types:** Factual lookup, numerical, comparative, detail extraction
- **Corpus:** 3,847 chunks from 3 company 10-K filings
- **Metric:** Precision@K = (Queries with correct chunk in top K) / Total

Results: Hybrid Search vs Alternatives

Search Method	P@1	P@3	P@5	P@10
Mixpeek Hybrid (Vector + BM25 + RRF)	76%	94%	98%	100%
Single Vector (text-embedding-3-large)	62%	81%	89%	96%
BM25 Keyword Only	54%	73%	84%	91%
Dense Retrieval (BERT)	58%	78%	87%	94%

Why Hybrid Search Wins

Single-method search has blind spots. Vector search misses exact keyword matches; BM25 misses semantic similarity. Our hybrid approach combines both with Reciprocal Rank Fusion (RRF):

- **Vector Search:** Finds semantically similar content ("revenue" ≈ "net sales")
- **BM25 Keyword:** Finds exact matches ("Q4 2024" must match exactly)
- **RRF Fusion:** Combines rankings, boosting results that appear in both
- **FinBERT Embeddings:** Financial-domain vectors understand "EBITDA" context

Results by Query Type

Query Type	Example	P@3	Notes
Factual Lookup	"What is AAPL's fiscal year end?"	98%	High keyword overlap
Numerical	"What was Q3 2024 services revenue?"	96%	Requires table context
Comparative	"Compare iPhone vs Mac margins"	88%	Needs multiple chunks
Detail Extraction	"List all risk factors"	92%	Long-form retrieval

Unique Capabilities

Beyond benchmark scores, our system provides capabilities that general-purpose tools cannot:

1. Bounding Box Coordinates

Every extracted element includes normalized PDF coordinates. Your application can highlight the exact source location—not just "page 31" but the specific cell at coordinates (0.125, 0.456).

```
{
  "answer": "$201.2B",
  "source": {
    "file": "AAPL_10K_2024.pdf",
    "page": 31,
    "bbox": {"x": 0.125, "y": 0.456, "w": 0.08, "h": 0.02}
  },
  "xbrl_tag": "aapl:IPhoneMember"
}
```

2. XBRL-Native Extraction

SEC filings include machine-readable XBRL data. We parse it directly—no OCR errors, standardized taxonomy, built-in validation via calculation linkbases. Every XBRL fact links to its visual location in the PDF.

3. Multi-Vector Search

Each chunk is embedded into 7 specialized vectors: title, summary, full_text, propositions, contextual, visual, and financial (FinBERT). Different query types activate different vectors for optimal retrieval.

Limitations & Scope

- **Scope:** Optimized for financial documents (10-K, 10-Q, earnings). Not a general-purpose assistant.
- **Corpus:** Benchmarks run on 3 company filings. Results may vary on other document types.
- **Multi-step reasoning:** 90% accuracy on complex queries—some edge cases require human review.
- **Scanned PDFs:** OCR quality affects extraction accuracy on scanned/image-based documents.

Recommendation: Use Mixpeek for structured extraction and retrieval from financial documents. Use general-purpose LLMs for creative tasks, coding, and broad Q&A.;

—

For questions about this benchmark or to reproduce results on your documents:
mxp.co/finance | info@mixpeek.com