FinMultiHop: A Cross-Document Multi-Hop Numerical Reasoning Benchmark for Financial Agents

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) expand their context windows to millions of tokens, evaluating reasoning requires moving beyond static, single-document extraction. Real-world financial analysis inherently relies on multi-hop, cross-document reasoning across time-series and cross companies. To address this problem, We introduce FinMultiHop, the benchmark designed to evaluate agents on cross-document numerical reasoning and logical bridging. FinMultiHop is constructed from a corpus of 150 S&P 500 annual reports (5 years) across 27 companies, totaling over 16 million tokens. Unlike previous benchmarks relying on manually curated snippets, FinMultiHop operates on raw SEC 10-K filings, allowing agents to autonomously navigate, extract, and synthesize information. We propose a pipeline that leverages FMP API for ground truth, using edgartools to perform numerical alignment with raw textual fillings. We systematically define 12 multi-hop reasoning paths, requiring models to perform retrieval, logical bridging, and high-precision computation. Our evaluation employed a funnel evaluation system, which can diagnose failure modes in terms of retrieval accuracy, bridging accuracy, and final computational accuracy. Experimental results show when facing highly similar filings, some traditional RAG methods fail to find the gold content. Some agent paradigms are also unable to find the correct content effectively through tool calls. FinMultiHop provides a financial benchmark that can test an agent's financial knowledge, tool usage, and financial calculation abilities.

1 Introduction

Reasoning over vast quantities of heterogeneous textual and tabular data is a hallmark of professional financial analysis, yet it remains a significant challenge for Large Language Models. Early benchmarks in automating financial numerical reasoning, such as FinQA (Chen et al., 2021b) and TAT-QA (Zhu et al., 2021), primarily focused on localized document snippets where relevant evidence is pre-identified. While recent efforts like DocFinQA (Reddy et al., 2024) have extended this task to the full-document level, they remain confined to intra-document reading comprehension. In practice, professional analysis is inherently cross-document, requiring agents to benchmark performance against industry peers and trace multi-year trends across hundreds of pages.

However, a fundamental gap persists between existing benchmarks and the reality of financial analysis: current evaluation frameworks ignore the complexity of retrieval in a highly redundant semantic space. Professional financial analysis involves searching within an extensive corpus where reports share an identical regulatory structure and overlapping financial terminology. In this environment, traditional retrieval models often fail to distinguish between nearly identical filings across different years or companies, falling victim to retrieval anchor drift. Current evaluation frameworks bypasses the core competencies required of a financial agent: autonomous evidence localization, handling semantic ambiguity across documents, and maintaining logical consistency through multi-stage, multi-hop computation.

The trajectory of AI benchmarks across high-stakes domains reflects a clear paradigm shift: moving from isolated, single-document extraction toward complex, cross-document logical synthesis. Domains such as software engineering, law, and tabular reasoning have all evolved to evaluate multi-file and multi-document capabilities to better mirror real-world analytical workflows. Despite this progression, financial reasoning evaluation has remained largely constrained to single-document.

To bridge this gap, we introduce FinMultiHop, the first financial QA benchmark designed to evaluate agents on autonomous, cross-document numerical reasoning. We curate a high-fidelity cor-

pus of 150 S&P 500 annual reports, comprising over 16 million tokens of raw financial text. FinMultiHop shifts the focus from passive reading to active exploration, requiring agents to navigate the retrieval space to synthesize evidence scattered across temporal and cross-sectional corporate entities. Furthermore, unlike prior benchmarks that treat reasoning as a black box, FinMultiHop introduces a diagnostic evaluation paradigm. By explicitly assessing the accuracy of intermediate logical bridges, we move beyond result-oriented grading to process-oriented diagnosis.

Our contributions are summarized as follows:

- **A Novel Agent-Centric Financial Retrieval Benchmark.** We propose FinMultiHop, the first benchmark dedicated to autonomous exploration in the financial domain. By providing a 16M-token corpus of raw filings, it challenges agents to perform cross-document retrieval in a highly redundant semantic environment, mirroring real-world analytical workflows.

- **Diagnostic Funnel Evaluation System.** We introduce a granular framework that isolates failures across the entire reasoning chain. Unlike prior datasets, we explicitly evaluate the accuracy of intermediate logical bridging. This allows for a systemic diagnosis of failure modes including retrieval drift and synthesis collapse, which is critical for iterative model improvement.

- **Quantitative Analysis of Semantic Landscape.** We visualize the unique difficulty of financial filings by plotting chunk level heatmaps. Our analysis reveals that financial filings occupy a compressed semantic space, significantly more redundant than general-domain benchmarks, analyzing why the current static retrieval paradigm fails and common errors in the agentic paradigm.

2 Related Work

Financial Question Answering. Prior studies (Oh et al., 2019; Lutz et al., 2018) in financial question answering initially focused on non-numerical reasoning and sentiment analysis. The domain shifted toward grounded numerical reasoning with the introduction of benchmarks such as FinQA (Chen et al., 2021b) and TAT-QA (Zhu et al., 2021). These datasets pioneered the field by requiring models to perform arithmetic over localized document snippets and pre-structured tables. However, they inherently rely on gold chunks, where all relevant evidence is manually pre-selected, effectively bypassing the challenge of evidence localization in a large-scale corpus.

Recently, DocFinQA (Reddy et al., 2024) significantly advanced the field by extending FinQA's (Chen et al., 2021b) context from short snippets to full SEC 10-K filings. While DocFinQA introduced the challenge of finding pertinent information within hundreds of pages, it remains primarily an intra-document reading comprehension task. The paper explored several methods, such as BM25 (Robertson et al., 2004) and SentenceBert (Reimers and Gurevych, 2019), for locating gold chunks. In contrast, professional financial analysis is inherently cross-document.

Long-Document and Multi-Hop Reasoning. Long-document QA has been studied extensively in the general domain through benchmarks such as LiteraryQA (Bonomo et al., 2025), Longgenbench (Liu et al., 2024), and LONGCODEU (Li et al., 2025). Recent advances in attention mechanisms and long-context window models allow for processing of contexts exceeding 1M+ tokens. Furthermore, general-purpose multi-hop reasoning datasets like HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) do not capture the unique constraints of the financial domain, such as strict accounting logic and hierarchical report structures. FinMultiHop adapts the multi-hop paradigm to an industrial financial setting, utilizing 12 distinct reasoning paths to test the limits of agents.

Benchmark Evolution. The trajectory of AI benchmarks across high-stakes domains reflects a clear paradigm shift: moving from isolated, single-document extraction toward complex, cross-document logical synthesis. In software engineering, evaluation has evolved from simple function-level tasks, such as HumanEval (Chen et al., 2021a), to cross-file codebase understanding exemplified by LoCoBench (Qiu et al., 2025). Similarly, legal reasoning has progressed from analyzing isolated contract clauses in LegalBench (Guha et al., 2023) to multi-document case discovery in Harvey's LAB (Harvey AI, 2025), while tabular reasoning has advanced from single-table QA to multi-table logical synthesis in ReasonTabQA (Pan et al., 2026). This evolution underscores a broader consensus in the research community: as models improve, evaluation must shift from static extraction to multi-document reasoning to better mirror

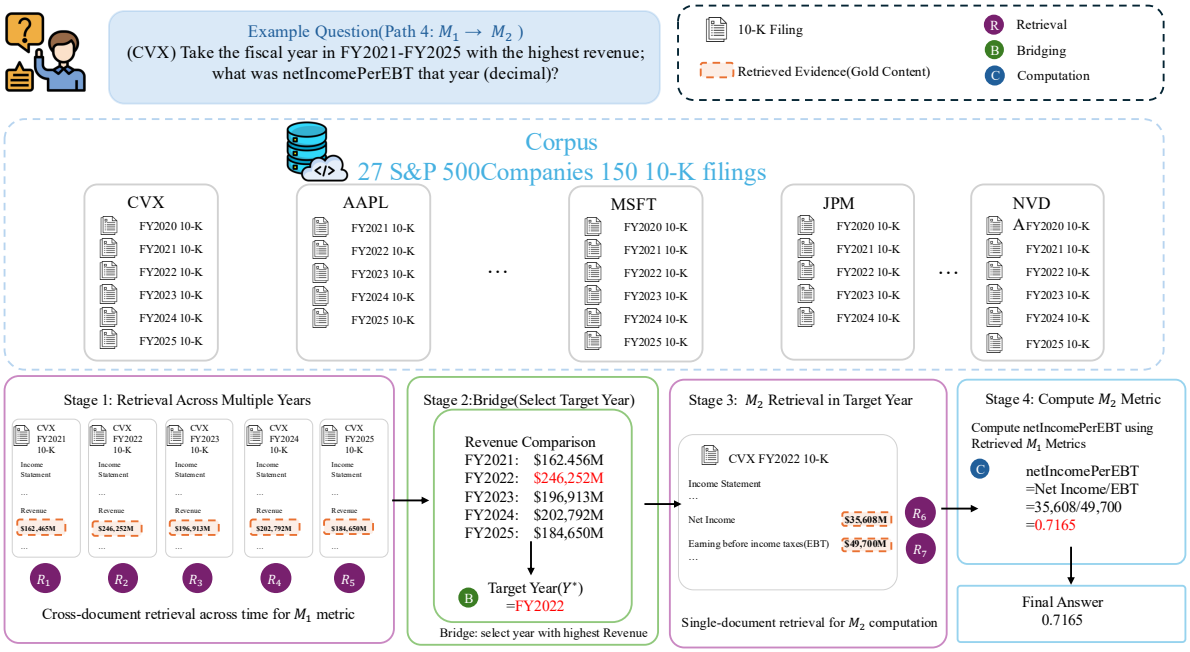


Figure 1: Example of cross-document multi-hop reasoning in FinMultiHop (Path 4: L1 → L2).

real-world analytical workflows.

Model	Correct	Accuracy
Gemini-3-Flash-Preview	25	2.18%
DeepSeek-V3.2-exp	23	2.01%
Grok-4.1-Fast	5	1.67%
GPT-5-mini	22	1.92%

Table 1: Closed-book, zero-shot performance on FinQA without providing context (N=1,147).

Dataset	Evidence Scope	Context Size	Accuracy
FinQA	Localized Snippet	1K Tokens	34.90%
DocFinQA	Full SEC Filing	100K Tokens	16.67%

Table 2: Retrieval-free performance of Gemini-3-Flash-Preview on legacy benchmarks.

3 Methodology

To show the drawbacks, we first conduct two preliminary studies on existing benchmarks: (1) *Parametric Memorization* in static datasets, and (2) *Attention Dilution* in passive reading tasks.

3.1 Analysis of Existing Paradigms

Obsolescence of Static Benchmarks. We test the integrity of existing static datasets FinQA by providing the query without document context. As shown in Table 1, LLMs achieve a non-trivial accuracy of ∼2%, indicating clear parametric memorization. This means we need a generative platform that continuously synthesizes new questions.

The Bottleneck of Passive Reading. We evaluate the "brute-force" context consumption of Gemini-3-Flash-Preview on full SEC 10-K filings. As shown in Table 2, accuracy collapses from 34.9% to 16.67%. This decay confirms that expansive context windows lead to severe attention dilution, rendering passive reading ineffective for long context financial analysis.

3.2 Benchmark Generation Pipeline

The construction of FinMultiHop relies on a fully automated, zero-human-annotation pipeline. We decouple Ground Truth Generation from the Retrieval Environment, ensuring that our benchmark provides both verifiable answers and perfectly aligned evidence anchors within raw documents.

1. Ground Truth Generation via FMP API. We utilize the Financial Modeling Prep (FMP) API to collect financial data for 27 representative S&P 500 companies spanning the fiscal years 2020–2025. We formalize this financial knowledge into an atomic fact quintuple $A = \langle T, Y, M, V, U \rangle$, representing the Ticker (T), Fiscal Year (Y), Financial Metric (M), numerical Value (V), and Unit (U). Within this framework, data points extracted directly from the three primary financial statements are labeled as M_1, while those derived through

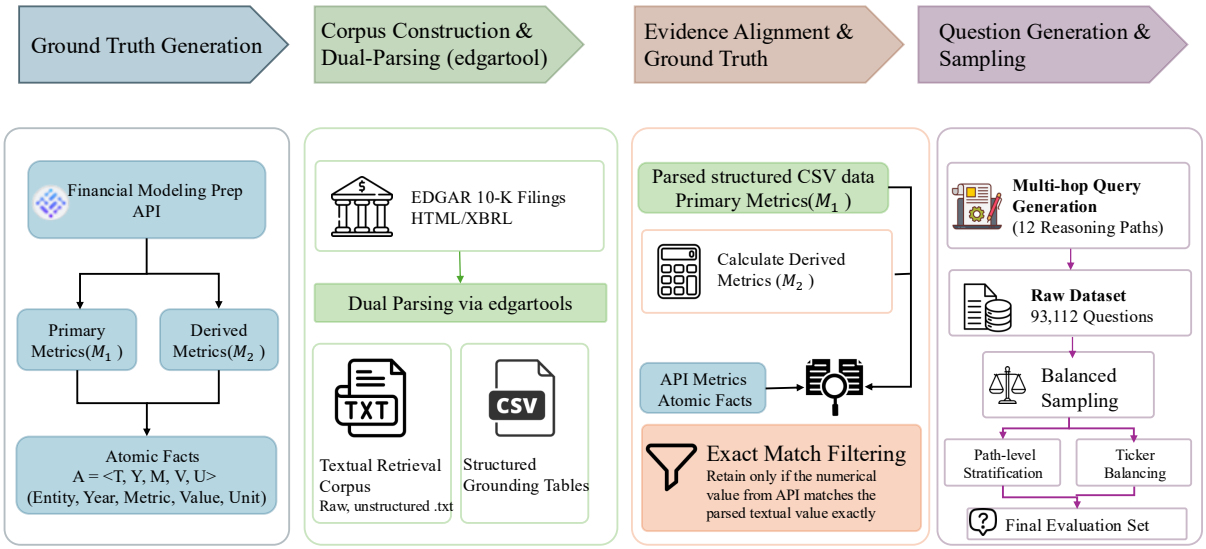


Figure 2: Pipeline of the FinMultiHop construction. The framework decouples ground truth generation from the retrieval environment, and enforces deterministic evidence alignment to produce multi-hop reasoning benchmarks.

calculations based on M_1 are denoted as M_2.

2. Corpus Construction and Grounded Alignment. We leverage the `edgartools` framework to perform a dual-parsing operation on the original HTML/XBRL filings, creating a dataset for both retrieval and verification:

- **Textual Retrieval Corpus.** We parse complete filings into raw, unstructured `.txt` documents. This serves as the retrieval space for RAG and agents, preserving the inherent noise, length, and complexity of original financial reports.
- **Deterministic Evidence Anchoring.** To ensure absolute precision, we implement a Strict Alignment Protocol. First, `edgartools` extracts core financial statements into structured CSVs, from which we programmatically derive M_2 metrics. These values, alongside M_1 data, are then rigorously cross-referenced against the FMP API records. A data point is designated as a Gold Content only if the numerical value from the API matches the parsed textual value exactly. This stringent filtering eliminates noise from accounting restatements and ensures that every metric is accurately grounded within the source text.

3.3 Task Taxonomy and Logical Paths

With the verified fact repository and aligned textual corpus, we generate multi-hop queries based on 12 predefined reasoning paths (Trivedi et al., 2022). As summarized in Table 3, these paths represent a combination of retrieval ($\mathcal{R}$), computation ($\mathcal{C}$), and logical bridging ($\mathcal{B}$). Bridging paths (e.g., Path 6 and 12) represent the highest complexity, where intermediate results determine the subsequent retrieval target, mirroring the iterative workflow of real human analysts.

3.4 Data Distribution and Balanced Sampling

Our automated pipeline generates an initial pool of 93,112 raw multi-hop reasoning pairs. As illustrated in Table 4, the distribution across the 12 reasoning paths exhibits a heavy imbalance. To construct a fair and diagnostic evaluation set, we implement a two-level stratified sampling strategy to derive a balanced subset of 360 high-quality samples ($n = 30$ per path).

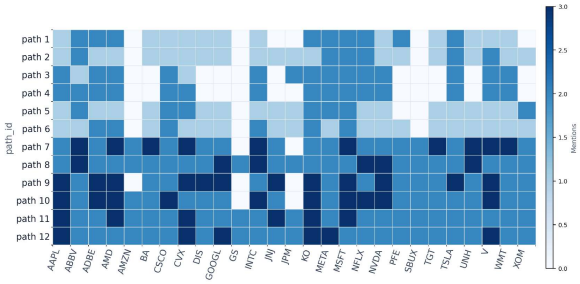


Figure 3: Entity and Path Diversity of the Evaluation Subset across 27 tickers and 12 reasoning paths.

First, we apply Path-Level Stratification to ensure that extreme-complexity tasks are represented equally alongside simpler patterns, preventing aggregate performance from being dominated by a single logic type. Second, we perform Intra-bucket Ticker Balancing using a Minimax Greedy algorithm. This algorithm iteratively selects instances

ID	Logical Description	Formalized Path	Hops
Path 1	M_1 Metric Growth Comparison	$\mathcal{R}(Y_1) + \mathcal{R}(Y_2) \rightarrow \mathcal{C}$	2
Path 2	M_2 Metric Growth Comparison	$[\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{Y_1,Y_2} \rightarrow \mathcal{C}$	3–4
Path 3	Range Bridge: $M_1 \rightarrow M_1$	$\arg\max_{Y_i \in \{Y_{1\ldots n}\}} \{\mathcal{R}(Y_i)\} \rightarrow Y^* \rightarrow \mathcal{R}(Y^*)$	3–6
Path 4	Range Bridge: $M_1 \rightarrow M_2$	$\arg\max_{Y_i \in \{Y_{1\ldots n}\}} \{\mathcal{R}(Y_i)\} \rightarrow Y^* \rightarrow [\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{Y^*}$	4–8
Path 5	Bridge: $M_2 \rightarrow M_1$	$[\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{Y_1,Y_2} \rightarrow \mathcal{B}(Y^*) \rightarrow \mathcal{R}(Y^*)$	4
Path 6	Bridge: $M_2 \rightarrow M_2$	$[\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{Y_1,Y_2} \rightarrow \mathcal{B}(Y^*) \rightarrow [\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{Y^*}$	5
Path 7	M_1 Metric Cross-Comparison	$\mathcal{R}(T_1) + \mathcal{R}(T_2) \rightarrow \mathcal{C}$	2
Path 8	M_2 Metric Cross-Comparison	$[\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{T_1,T_2} \rightarrow \mathcal{C}$	3–4
Path 9	Bridge: $M_1 \rightarrow M_1$	$\mathcal{R}(T_1) + \mathcal{R}(T_2) \rightarrow \mathcal{B}(T^*) \rightarrow \mathcal{R}(T^*)$	3
Path 10	Bridge: $M_1 \rightarrow M_2$	$\mathcal{R}(T_1) + \mathcal{R}(T_2) \rightarrow \mathcal{B}(T^*) \rightarrow [\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{T^*}$	4
Path 11	Bridge: $M_2 \rightarrow M_1$	$[\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{T_1,T_2} \rightarrow \mathcal{B}(T^*) \rightarrow \mathcal{R}(T^*)$	4
Path 12	Bridge: $M_2 \rightarrow M_2$	$[\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{T_1,T_2} \rightarrow \mathcal{B}(T^*) \rightarrow [\mathcal{R} + \mathcal{R} \rightarrow \mathcal{C}]_{T^*}$	5

Table 3: Taxonomy of Reasoning Paths in FinMultiHop, categorized by M_1 and M_2.

Path ID	Count	Percentage
Path 1–6 (Intra-company)	4476	4.81%
Path 7–12 (Inter-company)	88,636	95.19%
Total	**93,112**	**100.00%**

Table 4: Raw distribution of generated reasoning pairs.

that minimize the maximum frequency of any ticker within the subset, ensuring near-uniform entity coverage. As visualized in Figure 3, this strategy achieves exhaustive representation across all 27 S&P 500 constituents.

3.5 Task Complexity and Structural Span

We analyze the complexity of FinMultiHop along two critical dimensions: Evidence Intensity and Document Breadth. As illustrated in Figure 4, we measure the average number of atomic M_1 facts and unique 10-K filings required for reasoning.

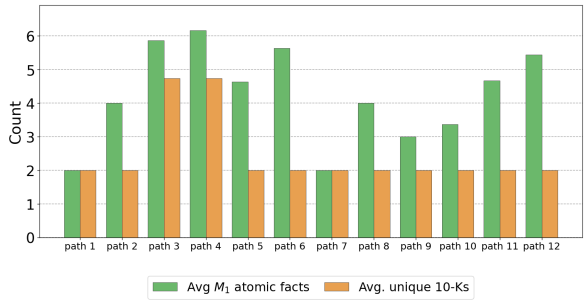


Figure 4: Structural complexity across reasoning paths.

Evidence Intensity and Density. The green bars in Figure 4 reveal a tiered complexity: while simple paths require only 2 anchors, complex multi-hop paths demand up to 6 distinct facts which are hard to find them all. This variance is further detailed in the heatmaps provided in Appendix D.

Document Breadth. The orange bars highlight the cross-document requirement: every query necessitates access to at least 2 filings, with temporal paths (Paths 3, 4) spanning up to 5 disparate reports. This variance is further detailed in the heatmaps provided in Appendix D.

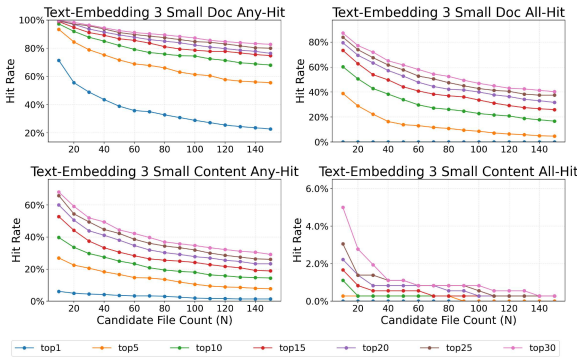


Figure 5: Scalability Stress: Text Embedding 3 Small

4 Experiments

Experimental Setup. For retrieval-based baselines, we process the documents into textual chunks of 1200 characters with a 200-character overlap.

Evaluation Metrics. We introduce a three-tier Funnel Evaluation System to diagnose agent performance across three dimensions: retrieval, intermediate logic, and numerical accuracy.

Multi-Granular Retrieval (RAG). We assess retrieval quality at document and content levels using **Any-Hit** (at least one correct match) and **All-Hit** (complete evidence capture) metrics. These measure the agent's ability to locate ground-truth anchors within the 16M-token corpus.

Reasoning Paradigm	Gemini-3-Flash-Preview		DeepSeek-V4-Flash		Grok-4.1-Fast		GPT-5-mini	
	Acc (360)	Bridge (240)	Acc (360)	Bridge (240)	Acc (360)	Bridge (240)	Acc (360)	Bridge (240)
Static RAG (Top-10)								
BM25	0.0	38.3	1.4	10.4	0.0	2.1	0.0	58.3
TFIDF	11.5	38.8	0.0	5.8	0.0	2.1	1.9	58.3
SentenceBERT	10.7	44.6	0.0	0.0	0.0	4.5	1.3	49.1
Text Embed 3 Small	8.4	53.8	1.1	33.3	2.0	22.1	3.1	55.4
Gemini Embed 2 Prev	17.0	**77.9**	3.8	**39.7**	9.8	**47.9**	9.1	**73.8**
Agentic Paradigm								
DeepRead (Round=8)	6.4	9.6	1.5	1.5	1.7	36.3	0.3	2.1
10k-Agent	**44.8**	75.4	**10.17**	38.3	**14.8**	**47.9**	**32.4**	70.4

Note: Bridging Accuracy is only calculated for the questions that require entities. Static RAG uses Top-10 Chunks.

Table 5: Comprehensive evaluation results.

Trajectory and Logic Accuracy. To measure reasoning path fidelity, we compute **Bridge Accuracy (B-Acc)**, which evaluates the agent's ability to correctly identify intermediate entities (T^* or Y^*). This disentangles planning errors from final synthesis collapse.

End-to-End Accuracy. Numerical Accuracy (Acc) measures the final answer correctness against ground-truth values.

Agentic Search Dynamics. To diagnose search stagnation, we measure: (1) **Sequential Query Consistency (QMCC)**: Quantifies retrieval loops via sequential query similarity; (2) **Retrieval Mean Novelty (RMN).** Measures information gain across retrieval steps; (3) **Cycle Death Rate (CDR).** Evaluates task completion efficiency under constraints.

Full mathematical definitions and explanations for all metrics are provided in Appendix C.

Scalability Stress Test. To investigate the performance boundaries of static retrieval, we conduct a scalability stress test by incrementally expanding the retrieval corpus from 10 to 150 filings. We evaluate several representative models: (i) Lexical (BM25 (Robertson and Zaragoza, 2009), TF-IDF (Aizawa, 2003)); (ii) Neural Dense, including open-source models (Sentence-BERT (Reimers and Gurevych, 2019)) and industry-leading models (`Text Embedding 3 Small` (Petukhova et al., 2025) from OpenAI, `Gemini Embedding 2 Preview` (Lee et al., 2025) from Google). The retrieval efficacy is assessed across varying Top-K values: *Doc-AnyHit*, *Doc-AllHit*, *Content-AnyHit*, and *Content-AllHit* (the *Sufficient Evidence Rate*).

Diagnostic Analysis: The Retrieval Wall. As visualized in Figure 5 and Appendix E, our experiments reveal a systematic collapse of evidence localization within FinMultiHop by using questions as query keywords. This "Retrieval Wall" signifies a fundamental bottleneck where performance degrades sharply as the corpus scales. We categorize this phenomenon into three critical failure modes:

The Doc-Content Divergence. While *Doc Any-Hit* (Top-Left panels) remains relatively high ($>80\%$ for BM25), the Content Any-Hit (Bottom-Left panels) collapses toward zero. This indicates thst retrievers can identify the correct corporate filing but are semantically blind to the specific financial line items within them.

The All-Hit Bottleneck and Evidence Vanishing. For multi-hop queries, capturing all required evidence is a prerequisite for success. However, Doc All-Hit (Top-Right) exhibits sharp exponential decay as the filing pool grows. More critically, the Content All-Hit (SER) is closed to 0% across all methods and scales (Bottom-Right). Since LLM never receives a complete evidence set, logical reasoning becomes physically impossible, regardless of the LLM's reasoning capacity.

The Fallacy of Top-K Expansion. Increasing retrieval depth ($K = 30$) provides marginal gains for document recall but fails to lift the Content-AllHit above the 0% floor, as shown in Table 6.

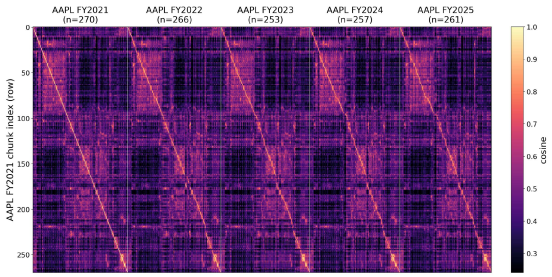
Cross-Document Retrieval Complexity. We visualize the semantic landscape of FinMultiHop and compare it with existing benchmarks across code, legal, and tabular reasoning. As shown in Figure 6 and Appendix I, these benchmarks exhibit fundamentally different semantic distributions.

Structural and Semantic Resonance. As shown in Figure 6 and Appendix E, FinMultiHop exhibits pervasive structural resonance, where regulatory mandates and standardized accounting terminology create a dense, semantically entangled landscape. In contrast, benchmarks in software en-

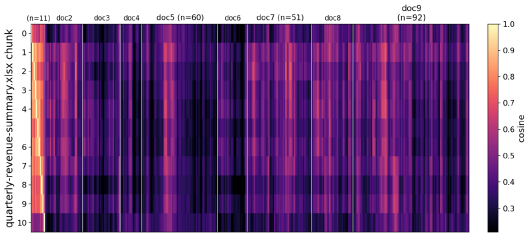
Model	Doc Any Hit	Doc All Hit	Content Any Hit	Content All Hit
Static RAG (Top-30 Chunks)				
BM25	74.7	43.6	0.8	0.0
TFIDF	48.1	13.3	0.0	0.0
SentenceBERT	52.2	8.6	6.9	0.2
OpenAI: Text Embedding 3 Small	82.8	40.3	29.1	0.3
Google: Gemini Embedding 2 Preview	**99.4**	**87.8**	62.8	3.3
10K-Agent (Top 4 Chunks/Doc)				
GPT-5-mini	98.6	86.1	**97.2**	40.8
Gemini-3-Flash-Preview	98.3	87.2	96.6	**44.3**
Grok4.1-Fast	98.1	87.2	96.1	37.9
DeepSeek-V4-Flash	98.3	86.7	94.6	29.1

Note: For 10K-Agent, the Docs for RAG are chosen by agent. The average number of documents is 2.86.

Table 6: Retrieval Hit Rate performance across 150 SEC filings with Top-30 chunks.



(a) FinMultiHop (Cross-Year Financial Filings)



(b) Harvey LAB (Legal Discovery Benchmark)

Figure 6: Visualizing the semantic landscape.

gineering (LoCoBench), legal discovery (Harvey LAB), and open-domain QA (LongBench) demonstrate sparse semantic alignment, where similarity is strictly localized to functional or thematic dependencies. This contrast explains the unique retrieval difficulty of FinMultiHop.

Diagnostic Analysis of Agentic Reasoning. To evaluate the agent's capacity for autonomous exploration, we anonymize the entire corpus by renaming all 150 filings into randomized identifiers $\{n_1, n_2, \ldots, n_{150}\}$. We analyze agent trajectories on Path 2 (multi-year comparison of M_2 metrics) to identify reasoning bottlenecks of DeepRead (Li et al., 2026), as summarized in Table 7.

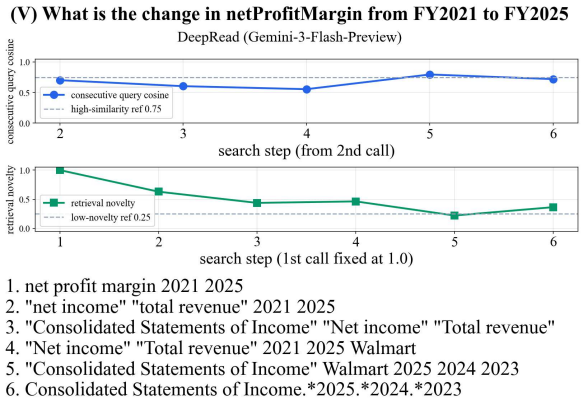


Figure 7: Sequential Query Similarity

Atomic Retrieval Bias. Agents exhibit a strong bias toward retrieving synthetic M_2 metrics as atomic entities (ARB 63.3%–100%), ignoring that these require decomposition into M_1 primary indicators. This indicates that the agent is unfamiliar with filings and does not know that these metrics will not appear in the filings. Avg. First M_1 Turn show the average first turn that the agent try to search M_1 to calculate the M_2.

Logical Blindness. Agents frequently fail to retrieve complete sets of M_1 variables (Coverage 40%–73.3%). This means Agent do not know how to calculate these metrics. High QMCC values (> 0.69) indicate that agents repeatedly issue similar queries rather than adjusting for formulaic requirements, suggesting an inability to map financial metrics to their underlying accounting formula.

Search Stagnation. We observe a strong correlation between search stagnation (low RMN) and resource exhaustion (high CDR). Agents often become trapped in semantic loops, repeatedly retrieving redundant chunks due to the highly overlapping semantic space of filings, as shown in Figure 7.

Backbone	ARB (%)	Avg. First M_1 Turn	Coverage (%)	QMCC	RMN	CDR (%)
Gemini-3-Flash-Preview	70.00	2.16	70.00	0.74	0.38	93.33
DeepSeek-V4-Flash	86.67	2.30	73.33	0.69	0.33	86.67
Grok-4.1-Fast	63.33	1.82	40.00	0.79	0.39	40.00
GPT-5-mini	100.0	3.17	56.67	0.80	0.29	96.67

Table 7: Agentic diagnostic metrics for Path 2 reasoning on DeepRead.

Result Analysis and Discussion. The evaluation results in Table 5 reveal a significant gap between the capabilities of frontier LLMs and the demands of cross-document financial analysis.

The Failure of Static Retrieval. Static Retrieval-Augmented Generation (RAG) methods struggle significantly with multi-hop financial queries, achieving only 0% to 17% End-to-End Accuracy. As detailed in Appendix E, even top-tier retrieval models (e.g., Gemini Embedding 2) fail at the fine-grained level. While these models can correctly identify the right financial document, they cannot accurately pinpoint the specific passage containing the numerical evidence. Within our highly redundant semantic space (GSR ≈ 0.87), text chunks from different years or companies look nearly identical to the retriever. Consequently, even when the retriever finds the correct document, it fails to separate relevant information from a large number of semantically similar but irrelevant chunks, making high-precision numerical grounding impossible.

The Stagnation of Agentic Paradigm. Existing long-context agents like DeepRead demonstrate significant limitations when applied to FinMultiHop. Despite the limitation of 8 rounds, the agent often fails to synthesize the final answer. As evidenced by the high QMCC and low RMN values in Table 7, these models frequently get stuck in "retrieval loops," repeatedly querying the same semantic clusters. They fail to aggregate the complete evidence set because their tool-use policy is tightly coupled with the retrieval process: the agent issues a query, receives a mixture of relevant and irrelevant chunks, and—failing to differentiate them—re-issues an identical query. Different backbones have different behavior patterns; for example, DeepSeek-V4-Flash is prone to this problem.

Decoupling as a Mitigation Strategy. Our 10K-Agent demonstrates that the key to resolving semantic drift is to decouple document selection from evidence extraction. DeepRead relies on a unified, high-frequency tool-use loop that retrieves from the entire corpus; while flexible, this approach exposes the agent to massive semantic noise, as similar financial metrics across different companies and years trigger frequent irrelevant matches.

In contrast, our 10K-Agent performs a two-stage process: first, it autonomously maps Tickers and Fiscal Years, then restricts evidence retrieval to these pre-selected filings using a single-pass BM25 iteration. This approach serves as an ablation study on retrieval complexity for agentic paradigm: despite using fewer tool calls and simpler tools than DeepRead, the 10K-Agent achieves significantly higher accuracy. The significant increase in Accuracy proves that for highly redundant financial corpora, robustness is achieved not through deeper tool-use rounds, but through focused retrieval within a limited scope for agentic paradigm.

5 Conclusion

We introduce FinMultiHop, a benchmark that shifts financial reasoning evaluation from static, single-document extraction to autonomous, cross-document exploration. Our empirical analysis reveals that traditional retrieval-augmented generation paradigms struggle with retrieval anchor drift, a failure mode driven by the high semantic redundancy inherent in financial filings. We demonstrate that decoupling document selection from evidence synthesis significantly improves reasoning robustness. Furthermore, our funnel evaluation system exposes a critical synthesis bottleneck: while frontier models excel at initial planning, they frequently fail to reliably aggregate multi-hop numerical evidence. FinMultiHop provides a rigorous diagnostic testbed to identify these failure modes, driving the transition from passive reading systems toward autonomous financial agents capable of reliable end to end analytical reasoning.

Limitations

Our current study has several limitations. First, regarding scope and generalization, our current dataset is limited to S&P 500 reports governed by SEC regulations; future work should expand

to multi-lingual disclosures. Second, in terms of agentic diversity, we do not analyze other agentic frameworks, such as multi-agent collaboration.

Ethics Statement

We confirm that we have fully complied with the ACL Ethics Policy in this study. All the environments are publicly available and have been extensively used in the research. We do not foresee any risks that may raised by this paper.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Tommaso Bonomo, Luca Gioffré, and Roberto Navigli. 2025. Literaryqa: Towards effective evaluation of long-document narrative qa. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34074–34095.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021b. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Harvey AI. 2025. Harvey LAB: Legal agent benchmark — an open-source benchmark for evaluating agents on real legal work. https://github.com/harveyai/harvey-labs. Accessed: 2026-05-23.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, and 1 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Jia Li, Xuyuan Guo, Lei Li, Kechi Zhang, Ge Li, Zhengwei Tao, Fang Liu, Chongyang Tao, Yuqi Zhu, and Zhi Jin. 2025. Benchmarking long-context language models on long code understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27309–27327.

Zhanli Li, Huiwen Tian, Lvzhou Luo, Yixuan Cao, and Ping Luo. 2026. Deepread: Document structure-aware reasoning to enhance agentic search. *arXiv preprint arXiv:2602.05014*.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024. Longgenbench: Long-context generation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 865–883.

Bernhard Lutz, Nicolas Pröllochs, and Dirk Neumann. 2018. Sentence-level sentiment analysis of financial news using distributed text representations and multi-instance learning. *arXiv preprint arXiv:1901.00400*.

Kyo-Joong Oh, Ho-Jin Choi, Sungtae Kwon, and Suhwan Park. 2019. Question understanding based on sentence embedding on dialog systems for banking service. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–3. IEEE.

Changzai Pan, Jie Zhang, Kaiwen Wei, Chenshuo Pan, Yu Zhao, Jingwang Huang, Jian Yang, Zhenhe Wu, Haoyang Zeng, Xiaoyan Gu, and 1 others. 2026. Reasontabqa: A comprehensive benchmark for table question answering from real world industrial scenarios. *arXiv preprint arXiv:2601.07280*.

Alina Petukhova, João P Matos-Carvalho, and Nuno Fachada. 2025. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6:100–108.

Jielin Qiu, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Jianguo Zhang, Haolin Chen, Shiyu Wang, Ming Zhu, Liangwei Yang, Juntao Tan, and 1 others. 2025. Locobench: A benchmark for long-context large language models in complex software engineering. *arXiv preprint arXiv:2509.09614*.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 3277–3287.

A Appendix: List of 38 Computable Secondary Metrics (M_2)

The FinMultiHop benchmark evaluates agents across a comprehensive set of 38 secondary financial metrics (M_2), programmatically derived from primary financial statements using deterministic formulas. Any metric with a zero denominator or missing primary data is excluded.

`grossProfitMargin`, `ebitMargin`, `ebitdaMargin`, `operatingProfitMargin`, `pretaxProfitMargin`, `netProfitMargin`, `contOpsProfitMargin`, `bottomLineProfitMargin`.

`assetTurnover`, `fixedAssetTurnover`, `inventoryTurnover`, `payablesTurnover`.

`debtToAssetsRatio`, `debtToEquityRatio`, `debtToCapitalRatio`, `longTermDebtToCap`, `financialLeverageRatio`.

`currentRatio`, `cashRatio`, `effectiveTaxRate`, `netIncomePerEBT`, `solvencyRatio`.

`operatingCashFlowSalesRatio`, `freeCashFlowOperatingCashFlowRatio`, `operatingCashFlowRatio`, `shortTermOperatingCashFlowCoverageRatio`, `operatingCashFlowCoverageRatio`, `capitalExpenditureCoverageRatio`, `dividendPaidAndCapexCoverageRatio`, `dividendPayoutRatio`.

`revenuePerShare`, `netIncomePerShare`, `operatingCashFlowPerShare`, `freeCashFlowPerShare`, `bookValuePerShare`, `shareholdersEquityPerShare`, `tangibleBookValuePerShare`, `dividendPerShare`.

B Code and Data Availability

To facilitate the reproducibility of our findings while maintaining the double-blind review process, the complete source code, datasets, and preprocessing scripts used in this study have been made anonymously available.

You can access the repository via the following link: https://anonymous.4open.science/r/FinMultiHop-D05C/

C Formal Definitions of Evaluation Metrics

This section provides the formal mathematical definitions for the metrics employed in our Funnel Evaluation System and the diagnostic analysis of agentic search dynamics.

C.1 Retrieval and Reasoning Metrics

Let $\mathcal{G}_i$ and $\mathcal{P}_i$ denote the ground-truth set and the retrieved evidence set for query i respectively, with N representing the total number of evaluation samples.

- **Any-Hit.** Measures the proportion of queries where at least one ground-truth chunk is successfully retrieved.

$$\text{AnyHit} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\mathcal{P}_i \cap \mathcal{G}_i \neq \emptyset) \quad (1)$$

- **All-Hit.** Measures the proportion of queries where the entire ground-truth set is captured.

$$\text{AllHit} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\mathcal{G}_i \subseteq \mathcal{P}_i) \quad (2)$$

- **Bridge Accuracy (B-Acc).** Measures the agent's ability to identify the intermediate entity (T^* or Y^*) in multi-hop reasoning. Let $\mathcal{Q}_{bridge}$ be the subset of queries requiring bridging:

$$B\text{-}Acc = \frac{1}{|\mathcal{Q}_{bridge}|} \sum_{q \in \mathcal{Q}_{bridge}} \mathbb{1}(\hat{e}_q = e_q^*) \quad (3)$$

- **Numerical Accuracy (Acc).** Measures the proportion of tasks where the predicted numerical value $V_{pred,i}$ exactly matches the ground-truth $V_{gold,i}$.

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(V_{pred,i} = V_{gold,i}) \quad (4)$$

C.2 Agentic Search Dynamics Metrics

To characterize the exploratory behavior of agents, we track the sequence of queries $\{q_1, \ldots, q_n\}$ and retrieved snippets $\{r_1, \ldots, r_m\}$ across the reasoning chain.

- **Query Mean Consecutive Cosine (QMCC).** Quantifies the stability of the agent's search intent.

$$QMCC = \frac{1}{N(n-1)} \sum_{j=1}^{N} \sum_{i=1}^{n-1} \cos(q_{j,i}, q_{j,i+1}) \quad (5)$$

where $q_{j,i}$ is the i-th query of the j-th task.

- **Retrieval Mean Novelty (RMN).** Measures the information gain at each retrieval step by comparing the new snippet r_i to all previously retrieved content.

$$RMN = \frac{1}{m-1} \sum_{i=2}^{m} \left(1 - \max_{j<i} \cos(r_i, r_j)\right) \quad (6)$$

- **Cycle Death Rate.** This metric measures task completion efficiency under a reasoning budget $T_{\max}$. It is the proportion of tasks where the agent reaches the maximum rounds without producing a final answer. Let A_i be 1 if an answer is produced and 0 otherwise.

$$CDR = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(R_i = T_{\max} \wedge A_i = 0) \quad (7)$$

where R_i denotes the total iteration rounds for task i. A high CDR indicates that the agent is trapped in an inefficient loop.

C.3 Agentic Diagnostic Metrics

To diagnose the structural and logical reasoning failures of agents, we define three diagnostic metrics focused on financial knowledge and reasoning path fidelity.

- Atomic Retrieval Bias (ARB): This metric measures the proportion of tasks in which the agent incorrectly attempts to retrieve synthetic M_2 metrics (which are not explicitly present as raw text in filings) as atomic entities. Let N be the total number of tasks and Q_k the query sequence for task k.

$$ARB = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}(\exists q \in Q_k \text{ s.t. } q \in \mathcal{M}_2) \quad (8)$$

where $\mathcal{M}_2$ denotes the set of synthetic metrics. A high ARB indicates that the agent fails to recognize the hierarchical structure of financial statements.

- Avg. First M_1 Turn: This metric represents the average reasoning step t at which the agent initiates its first search for primary M_1 indicators necessary to compute the target M_2 metric.

$$T_{\text{first}} = \frac{1}{N}\sum_{k=1}^{N} \min\{t \mid q_{k,t} \cap \mathcal{M}_{1,k} \neq \emptyset\} \quad (9)$$

where $q_{k,t}$ is the query at step t for question q_k and $\mathcal{M}_{1,k}$ is the set of primary indicators for question q_k. This measures the agent's efficiency in transitioning from M_2 search to a logical decomposition strategy.

- M_1 Coverage: This metric measures the agent's ability to retrieve the complete set of primary M_1 indicators required to resolve the target financial formula.

$$\text{Coverage} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{I}\big(\text{Required}(M_{1,k}) \subseteq \text{Retrieved}(M_{1,k})\big) \quad (10)$$

where $\text{Required}(M_{1,k})$ represents all primary metrics needed to calculate the target M_2 in task k, and $\text{Retrieved}(M_{1,k})$ is the set of indicators successfully retrieved by the agent.

D Extended Complexity Analysis

In this section, we provide detailed granularity regarding the structural complexity of the FinMultiHop evaluation subset.

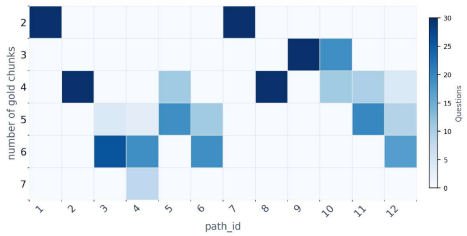


Figure 8: Granular distribution of evidence density: number of required chunks for each reasoning path.

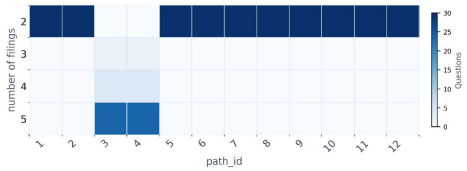


Figure 9: Granular distribution of document breadth: number of required SEC filings for each reasoning path.

E Appendix: Detailed Scalability Analysis

This section provides the full scalability stress test results for additional retrieval paradigms. These results demonstrate that the "Retrieval Wall" is a systematic failure across diverse lexical and dense retrieval architectures.

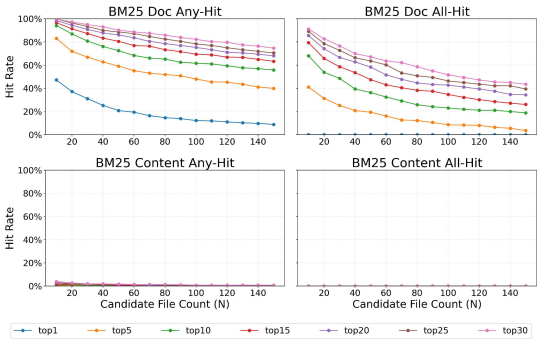


Figure 10: Scalability and stress test analysis for BM25 retrieval paradigm under varying context window loads.

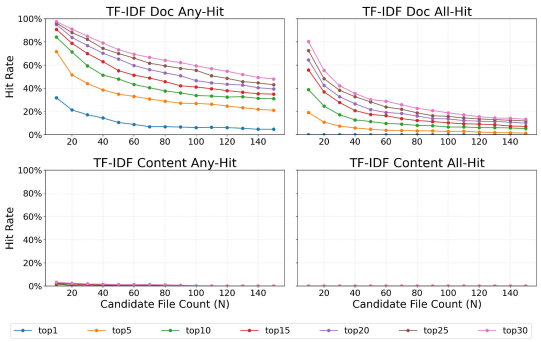


Figure 11: Scalability and stress test analysis for TF-IDF retrieval paradigm under varying context window loads.

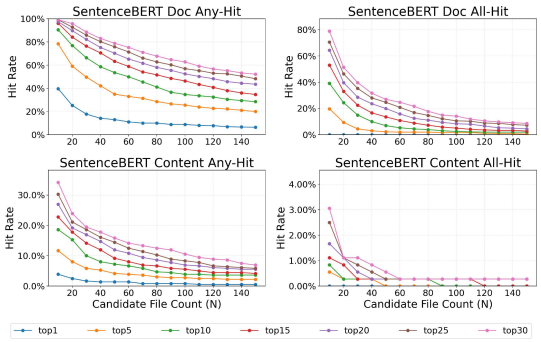


Figure 12: Scalability and stress test analysis for SentenceBert retrieval paradigm under varying context window loads.

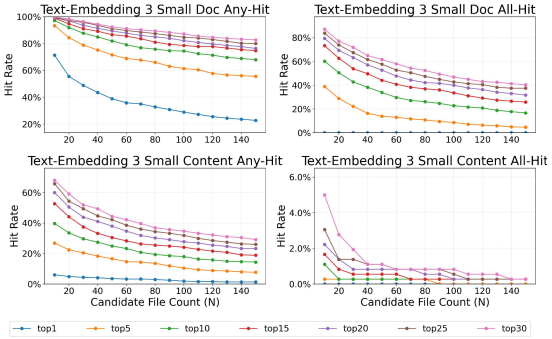


Figure 13: Scalability and stress test analysis for OpenAI (Text Embedding 3 Small) retrieval paradigm under varying context window loads.

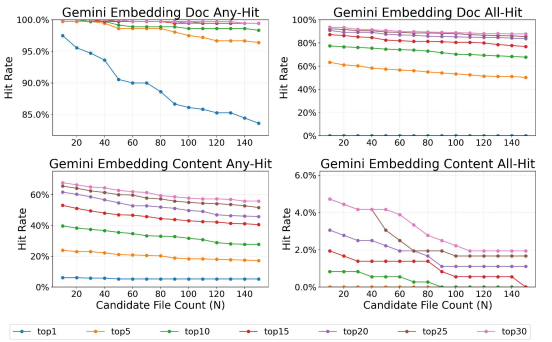


Figure 14: Scalability and stress test analysis for Google (Gemini Embedding 2 Preview) retrieval paradigm under varying context window loads.

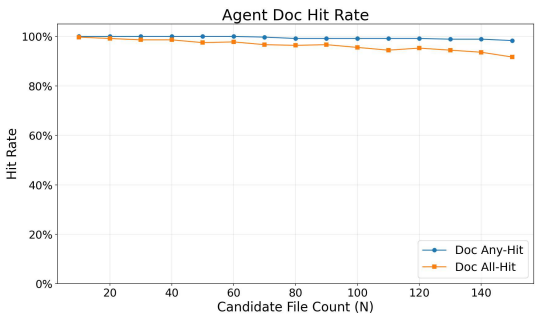


Figure 15: Scalability and stress test analysis for 10K-Agent under varying context window loads.

F Extended Semantic Landscape Analysis

We provide additional similarity heatmaps for other benchmarks to further demonstrate the unique semantic uniformity of financial corpora.

G Semantic Landscape and Structural Analysis

To understand why multi-document financial reasoning is uniquely difficult, we characterize the semantic landscape of FinMultiHop and contrast it with established benchmarks in software engineering (LoCoBench) and open-domain QA (LongBench).

G.1 Global Semantic Redundancy (GSR)

We define *Global Semantic Redundancy* (GSR) to quantify retrieval difficulty. Given a corpus $D = \{d_1, \ldots, d_n\}$, where each document is represented by the mean vector v_i of its internal chunks (normalized to unit length), we define GSR as the average off-diagonal cosine similarity:

$$GSR = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} v_i^\top v_j \qquad (11)$$

As reported in Table 8, FinMultiHop exhibits a GSR of ~ 0.874, significantly higher than the LoCoBench and LongBench. This increase in redundancy confirms that financial filings inhabit a highly compressed semantic space. While documents in general-domain or software engineering benchmarks are theme-distinct, providing natural retrieval barriers, financial filings are *structurally entangled*, forcing agents to perform high-precision discrimination within a dense cluster of semantically near-identical documents.

G.2 Visualizing Structural Resonance and Retrieval Ambiguity

The semantic density is visualized in Figure 6. Temporal stickiness (a) and cross-company structural blueprints (b) create a "field of needles," where raw semantic similarity fails to isolate relevant content.

G.3 Diagnostic Analysis: The Flat Retrieval Landscape

To examine how semantic entanglement affects performance, we analyze the *Query-Chunk Similarity Distribution* in Figure 17. The heatmap displays the top 5 retrieved chunks for representative queries. We observe that even for the same query, similarity scores across different chunks remain uniformly high (> 0.5), indicating a *flat semantic landscape* where the retriever fails to rank "Gold Content" significantly above noise. This explains the *Retrieval Anchor Drift*: when an agent attempts to retrieve a specific metric, it finds hundreds of semantically similar candidates that are structurally indistinguishable, forcing the agent to rely on metadata-aware filtration to identify the ground truth.

H 10K-Agent Framework

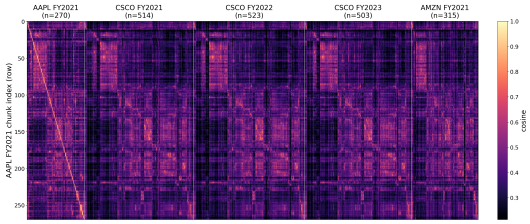
Phase 1: Autonomous Corpus Indexing. The agent emulates a human analyst's orientation by sacnning the first 50 lines of each anonymized file to map its Ticker and Fiscal Year. As shown in Table 9, all models achieved 100% accuracy.

Phase 2: Targeted Evidence Extraction and Reasoning. The agent selects limited documents(Max 8, Avergae 2.86) to retrieve. The agent invokes **BM25-based retrieval tool** to extract evidence from specific documents.We enforced equivalent environment constraints: (1) Chunk configuration is fixed at 1,200 characters with a 200-character overlap; (2) Retrieval depth is restricted to the Top-4 chunks; and (3) a Single-pass constraint.

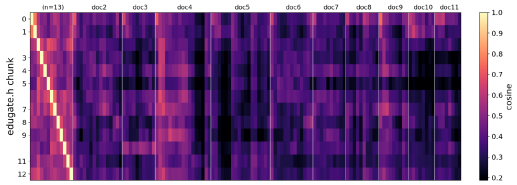
Benchmark	Domain	Mean GSR	AD/Q	MM/Q	Chunks/Q	Chunks/Doc
LongBench	General NLP	0.217	16.2	1/40	44.3 ± 21.0	2.7 ± 4.3
LoCoBench	Software	0.586	23.27	1/85	271.5 ± 231.4	11.7 ± 3.4
TQA-Bench	Multi-Table	0.576	3.3	2/6	174.6 ± 47.7	52.9 ± 46.8
Harvey LAB	Legal	0.714	7.61	2/60	289.4 ± 225.4	38.0 ± 35.3
FinMultiHop	**Filings**	**0.874**	**150**	**150/150**	$\mathbf{76433 \pm 0}$	$\mathbf{509.6 \pm 178.4}$

Note: AD/Q: Average Number of Documents Per Question; **MM/Q**: Min/Max Documents Per Question; **Chunks/Q** and **Chunks/Doc** report the mean and standard deviation ($\mu \pm \sigma$) of the number of chunks per question and per document, respectively.

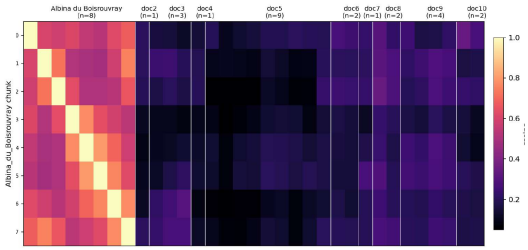
Table 8: Comparison of Cross Document Benchmark.



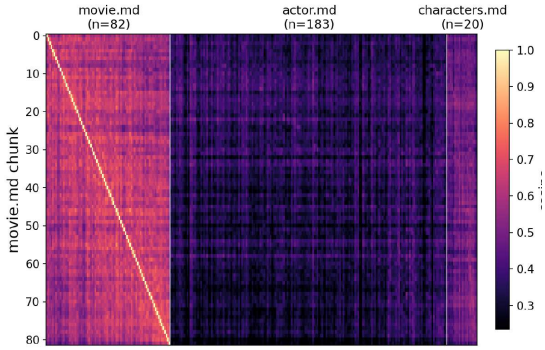
(a) FinMultiHop (Cross-Company)



(b) LoCoBench (Code)



(c) LongBench (Open-Domain QA)



(d) TQA-Bench (Multi-Table)

Figure 16: Extended similarity heatmaps for benchmark domains.

Model	Accuracy
DeepSeek-V4-Flash	100.0%
Gemini-3-Flash-Preview	100.0%
Grok-4.1-Fast	100.0%
GPT-5-mini	100.0%

Table 9: Phase 1 Indexing Accuracy (N=150).

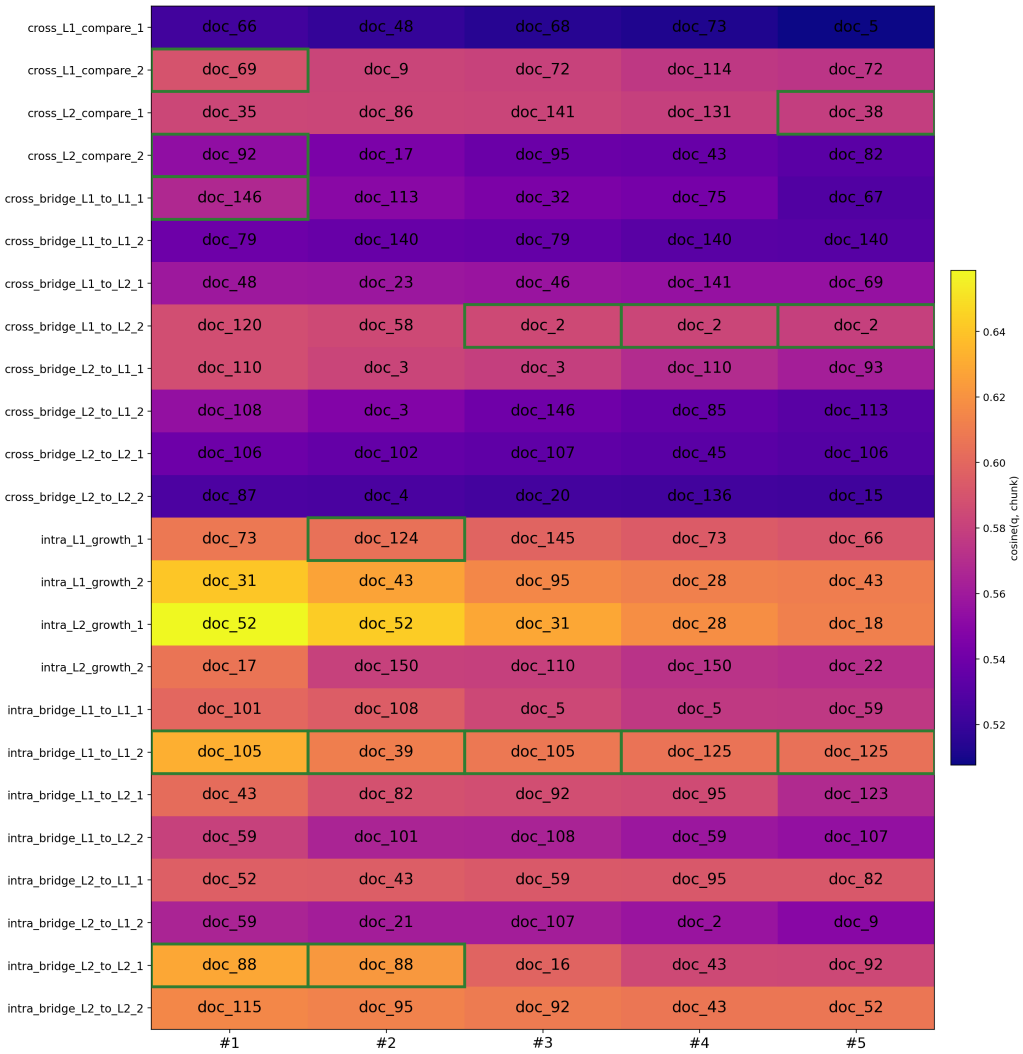


Figure 17: Query-Chunk similarity distribution for top 5 retrieved chunks. Green boxes denote Gold Document. The uniform color intensity illustrates the difficulty of isolating Gold Document from noise (BM25).

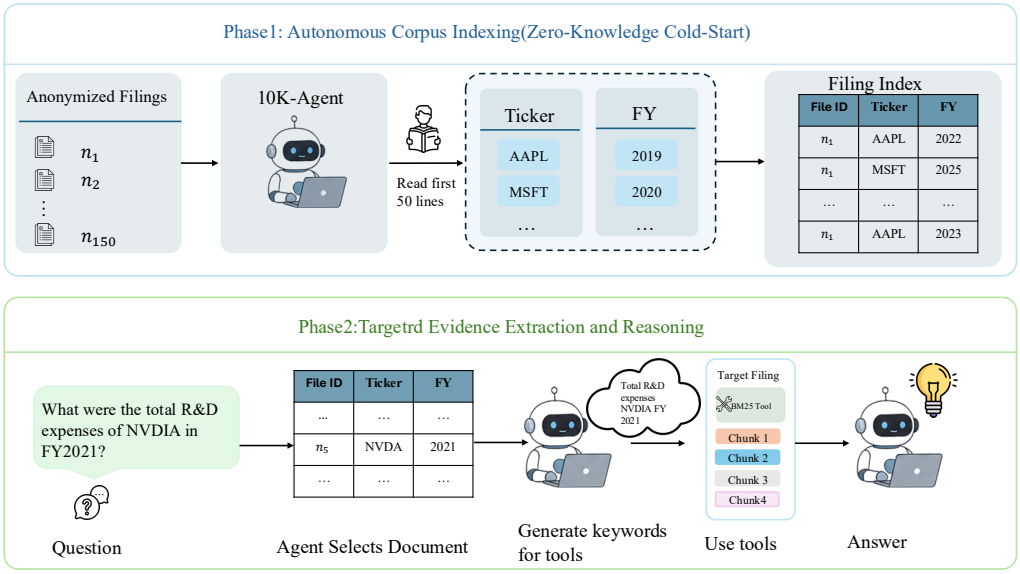


Figure 18: Agentic Paradigm

I Filing Heatmaps

This appendix presents the complete collection of temporal consistency heatmaps (2021 vs. 2021–2025) for 26 evaluated tickers.

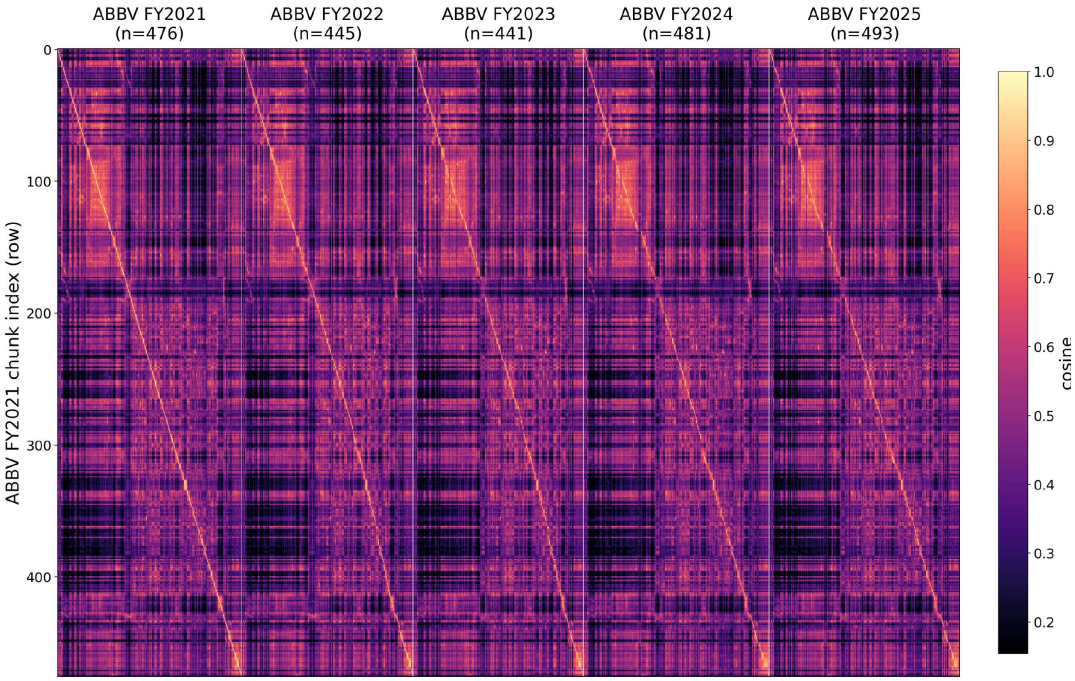


Figure 19: Temporal consistency heatmap for ABBV.

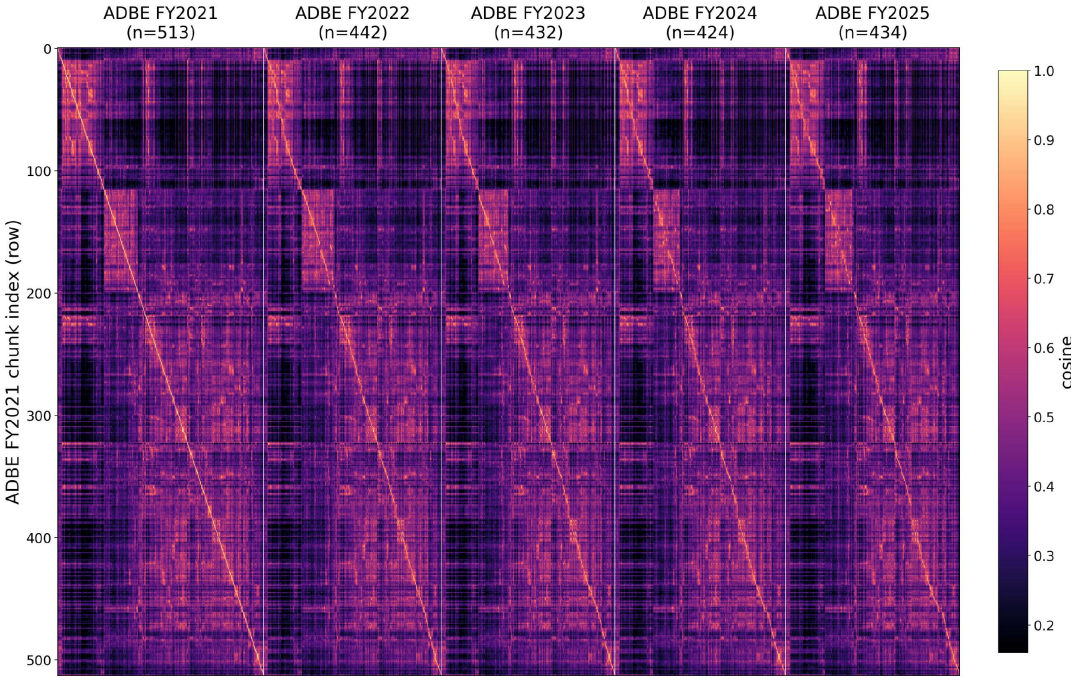


Figure 20: Temporal consistency heatmap for ADBE.

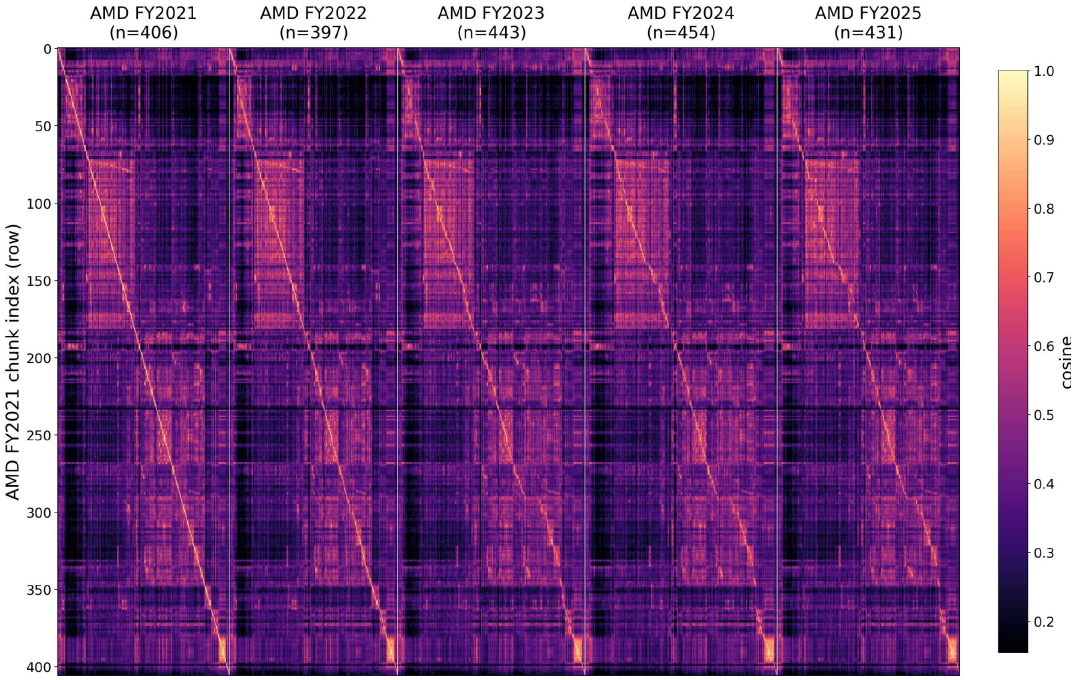


Figure 21: Temporal consistency heatmap for AMD.

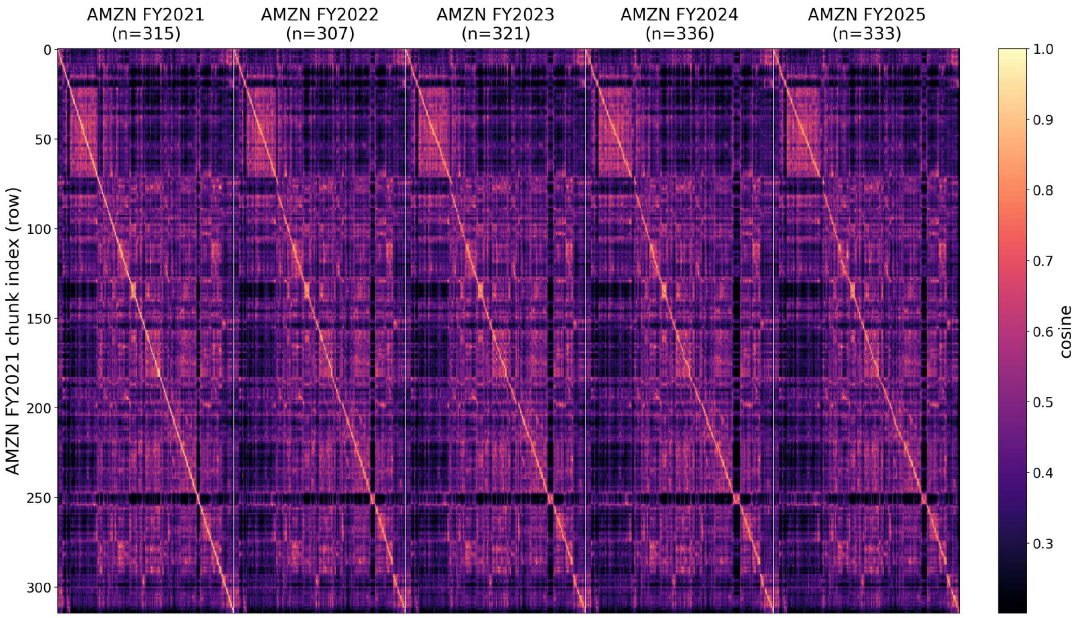


Figure 22: Temporal consistency heatmap for AMZN.

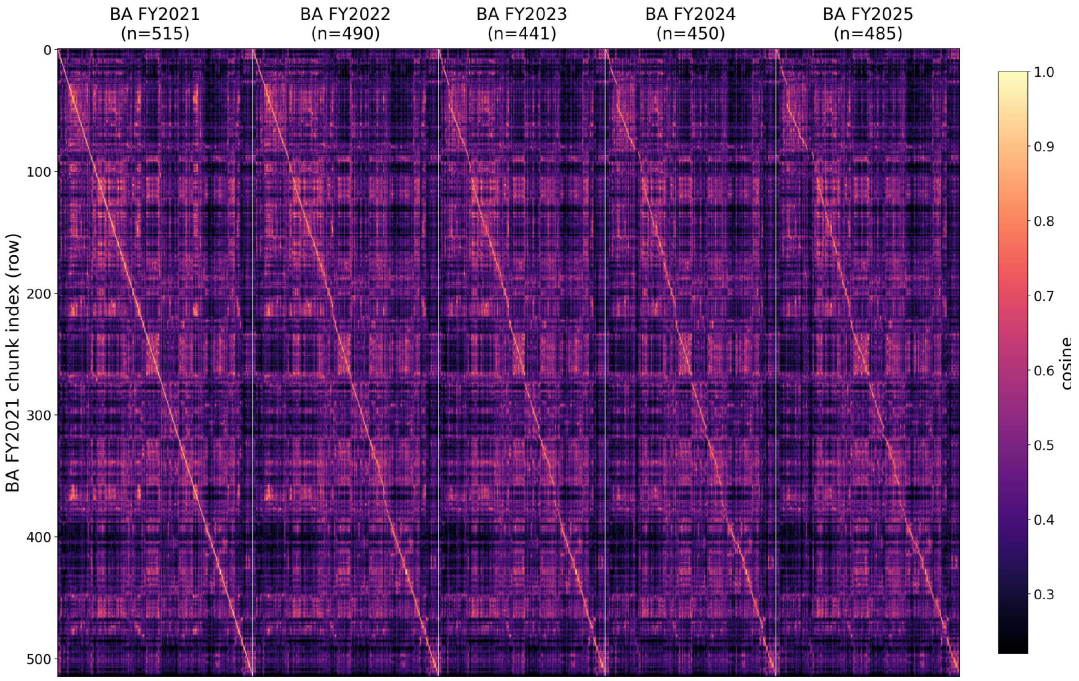


Figure 23: Temporal consistency heatmap for BA.

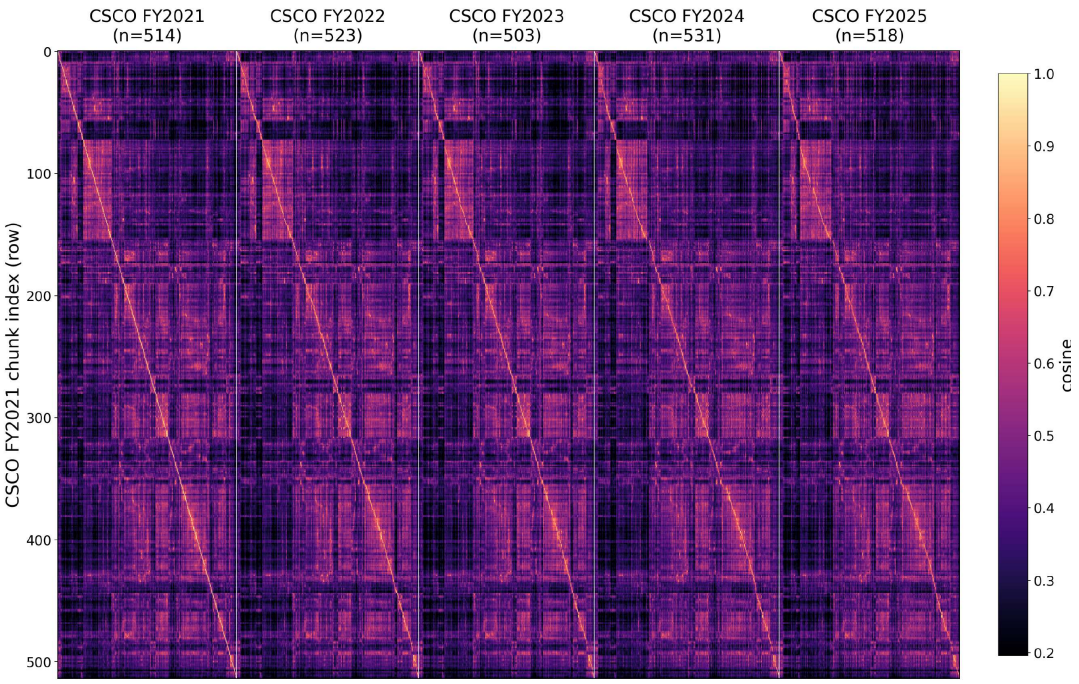


Figure 24: Temporal consistency heatmap for CSCO.

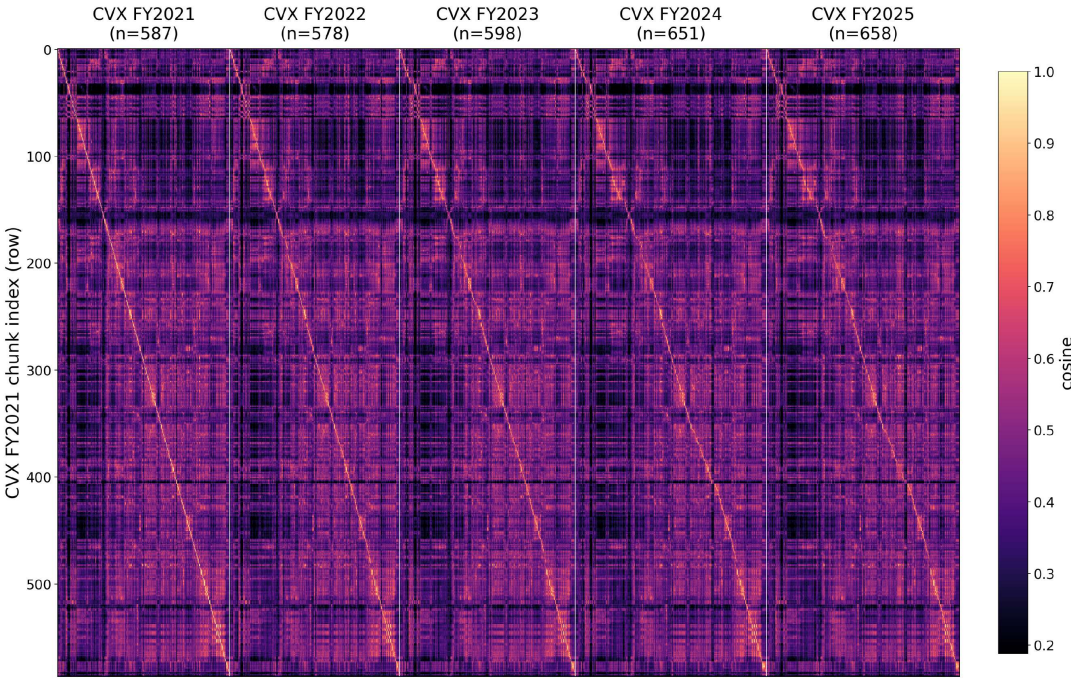


Figure 25: Temporal consistency heatmap for CVX.

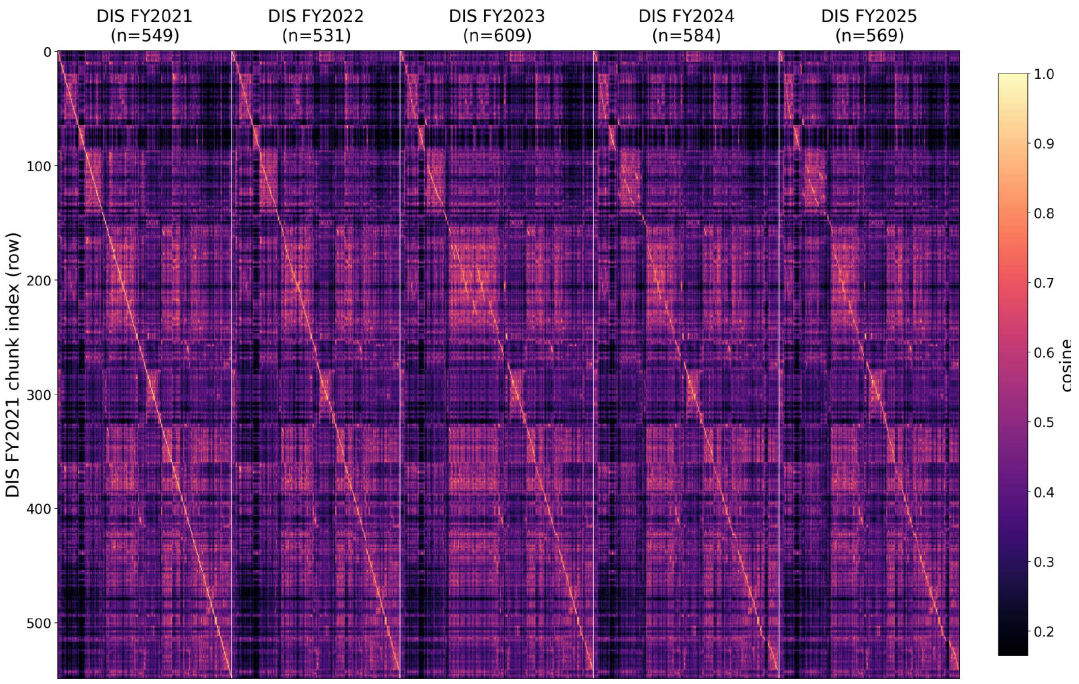


Figure 26: Temporal consistency heatmap for DIS.

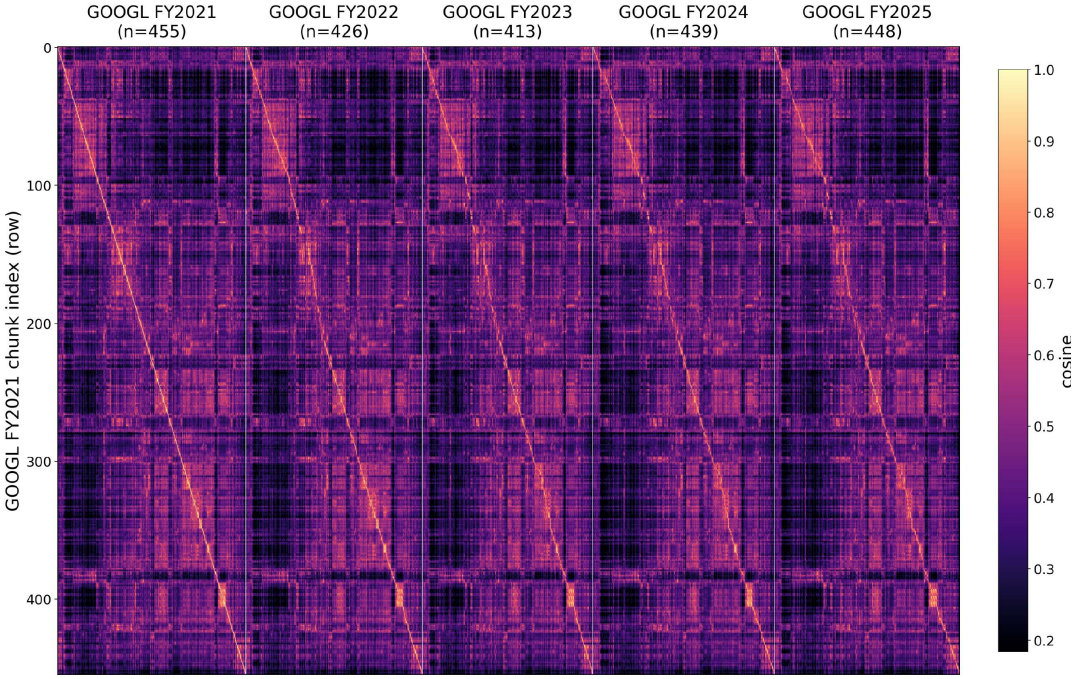


Figure 27: Temporal consistency heatmap for GOOGL.

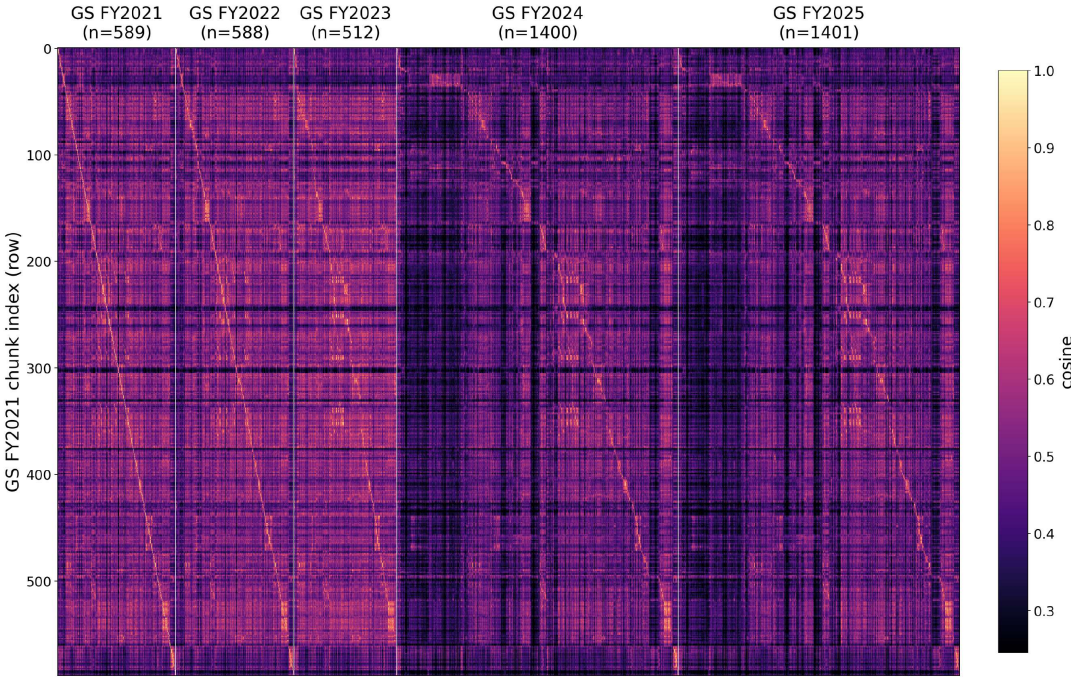


Figure 28: Temporal consistency heatmap for GS.

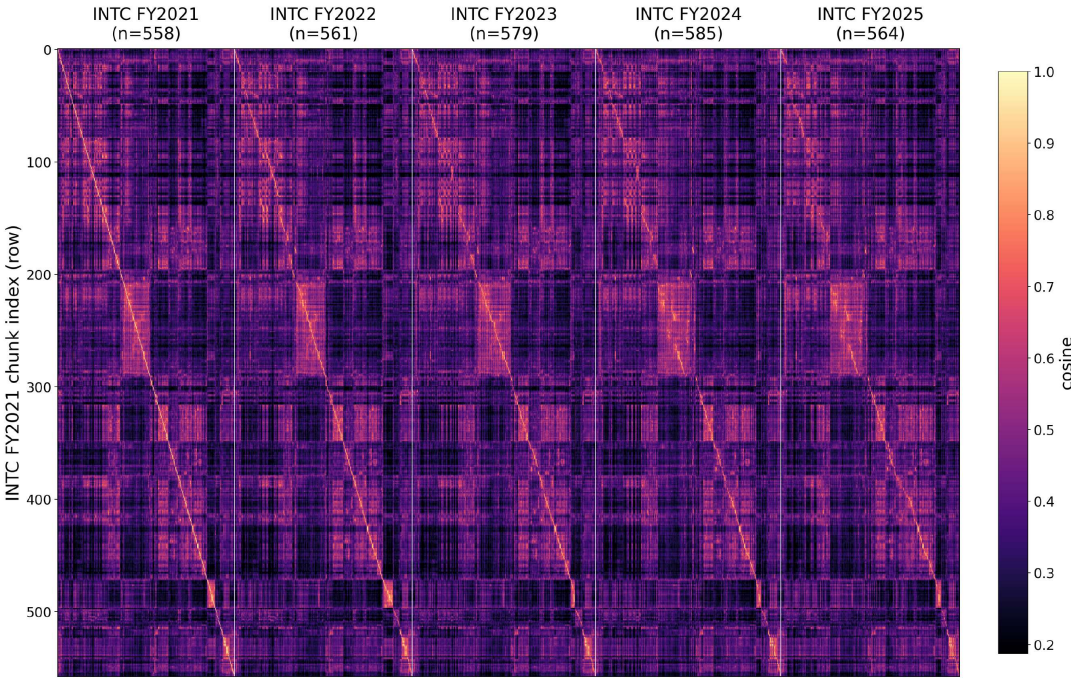


Figure 29: Temporal consistency heatmap for INTC.

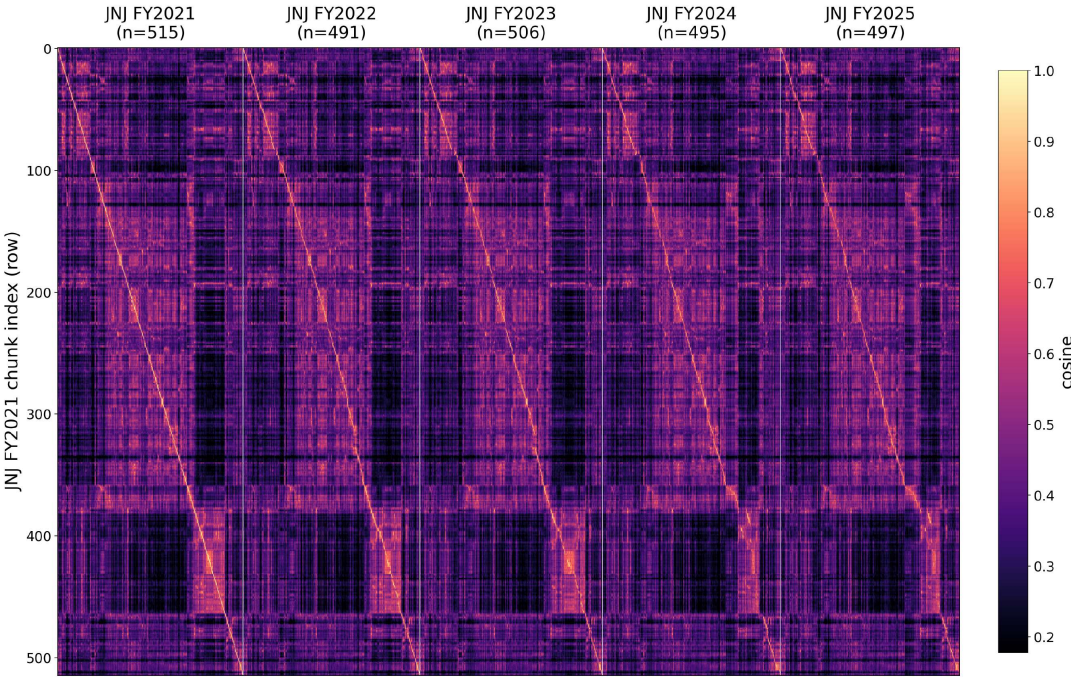


Figure 30: Temporal consistency heatmap for JNJ.

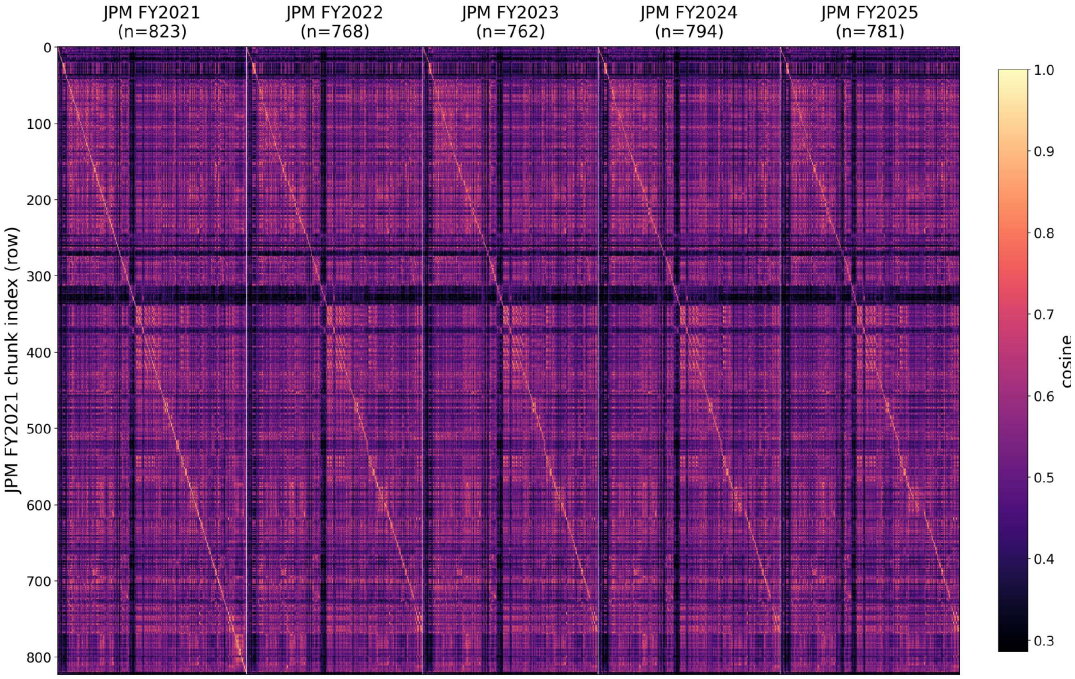


Figure 31: Temporal consistency heatmap for JPM.

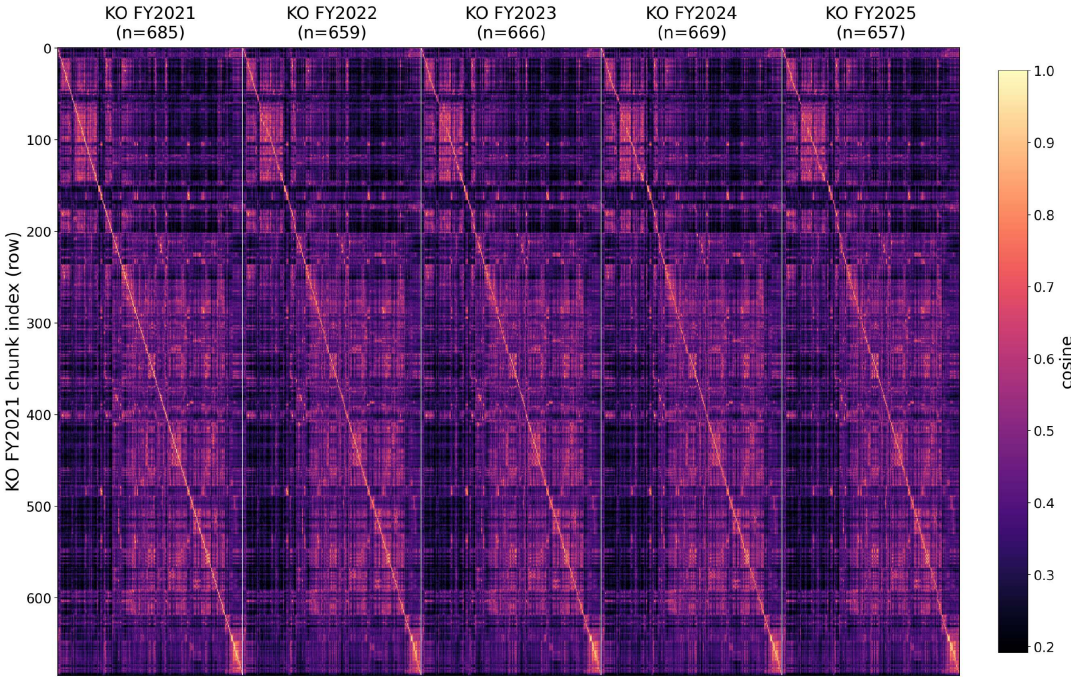


Figure 32: Temporal consistency heatmap for KO.

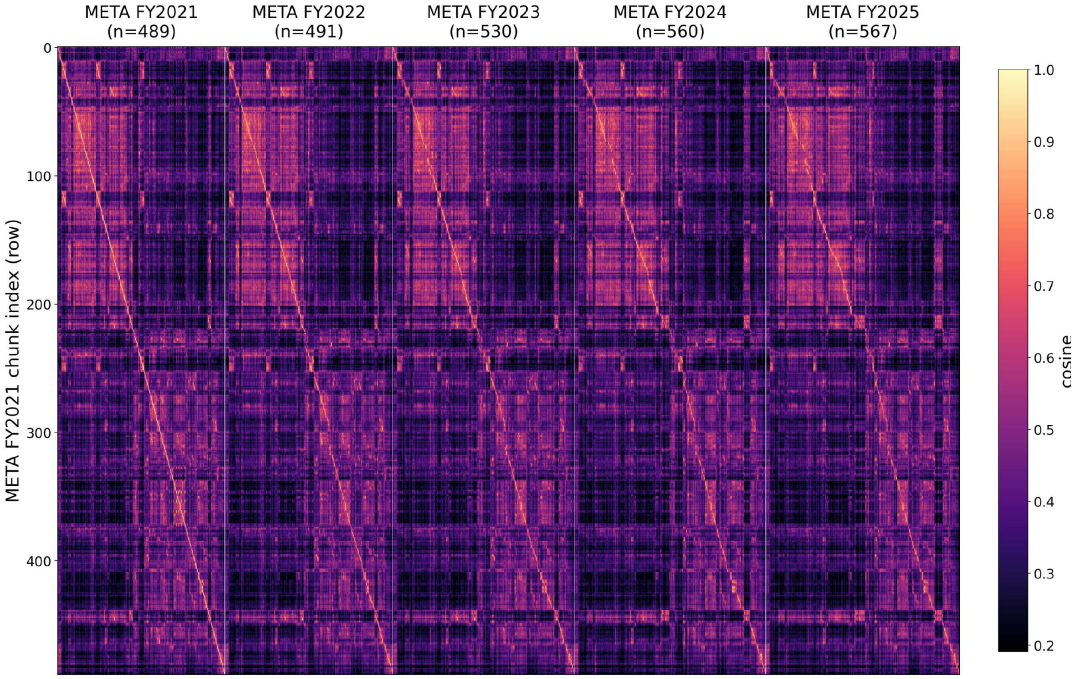


Figure 33: Temporal consistency heatmap for META.

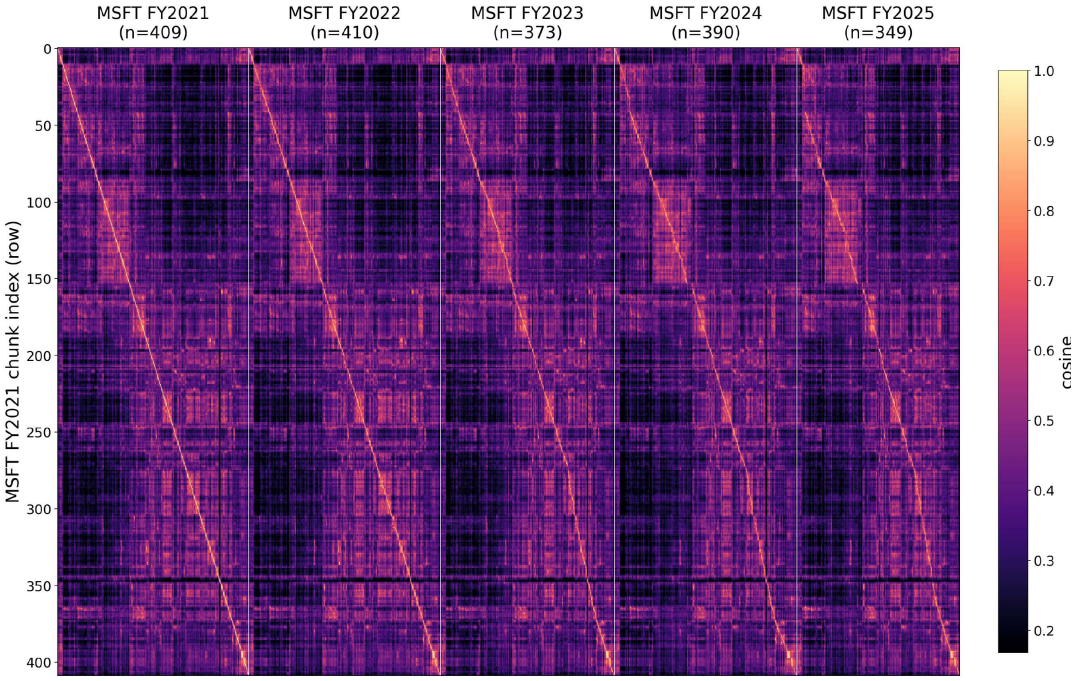


Figure 34: Temporal consistency heatmap for MSFT.

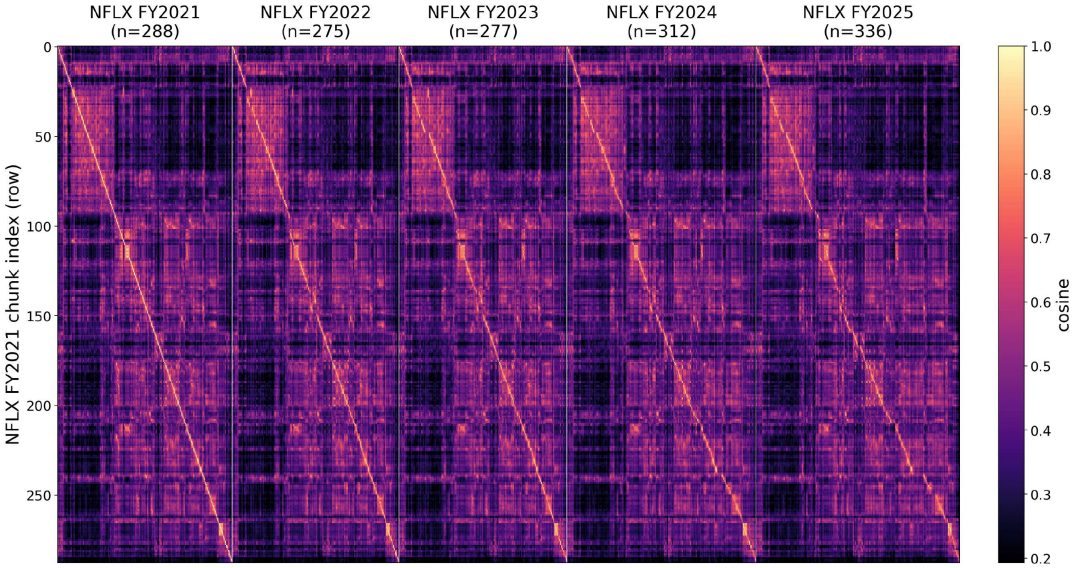


Figure 35: Temporal consistency heatmap for NFLX.

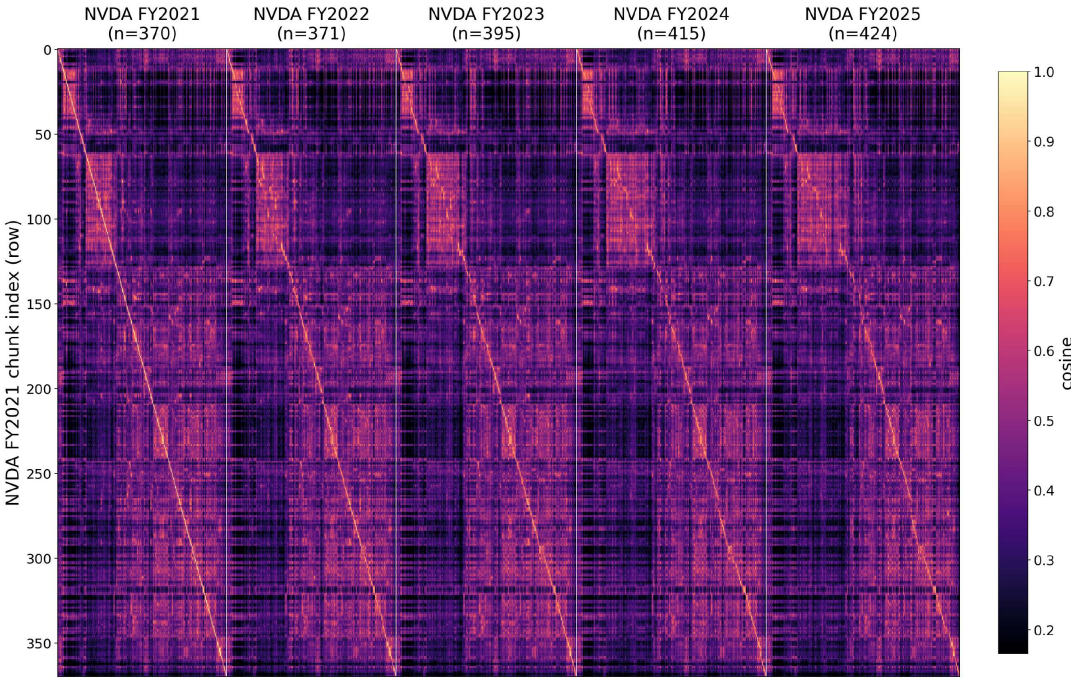


Figure 36: Temporal consistency heatmap for NVDA.

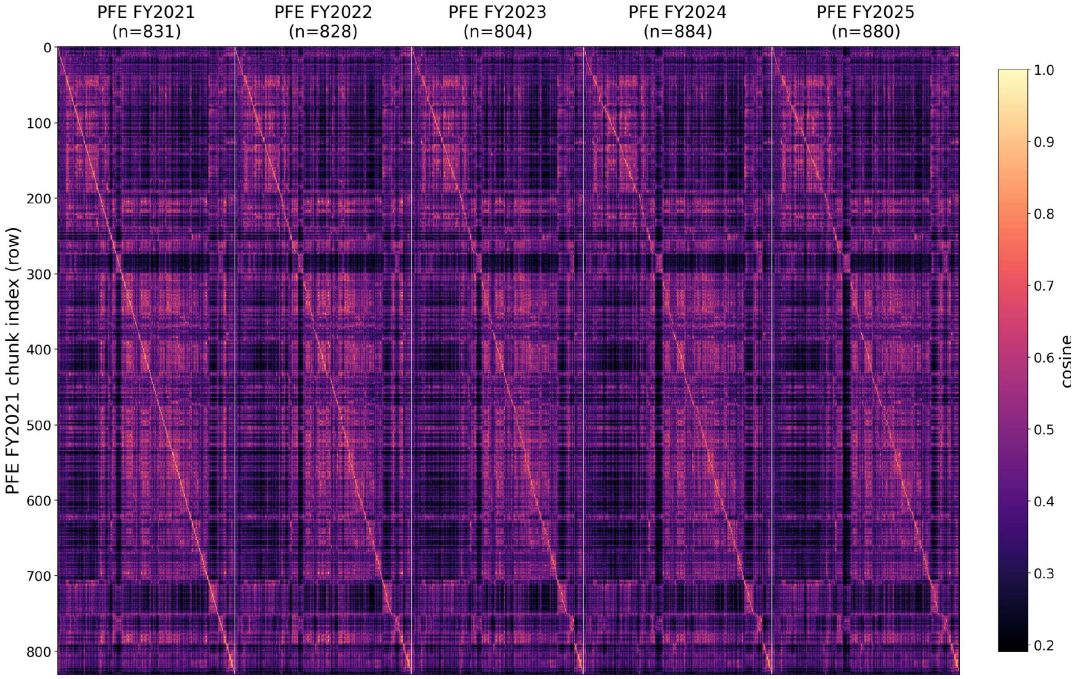


Figure 37: Temporal consistency heatmap for PFE.

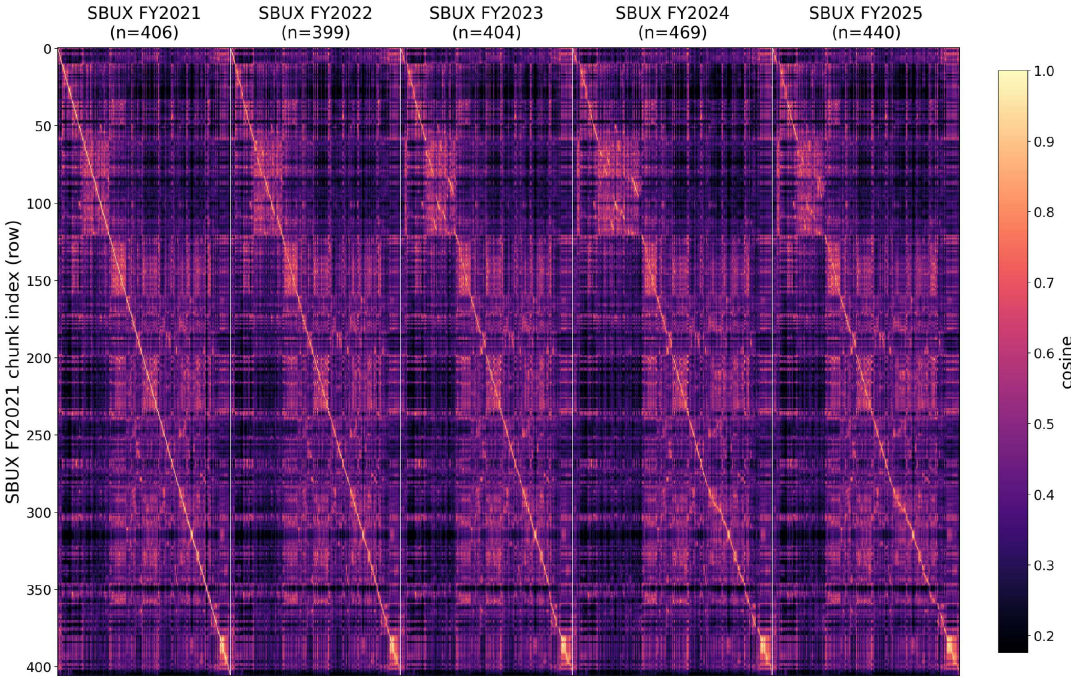


Figure 38: Temporal consistency heatmap for SBUX.

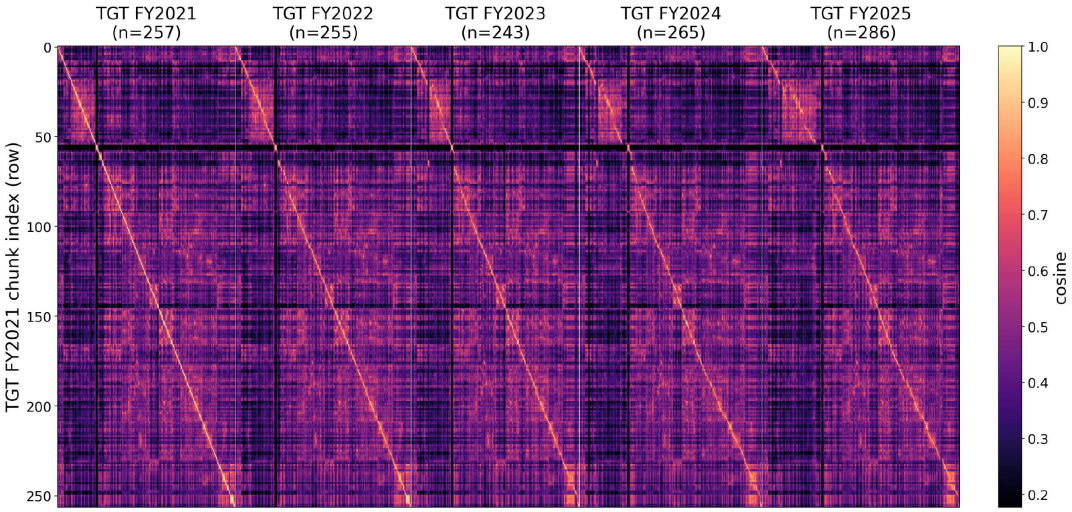


Figure 39: Temporal consistency heatmap for TGT.

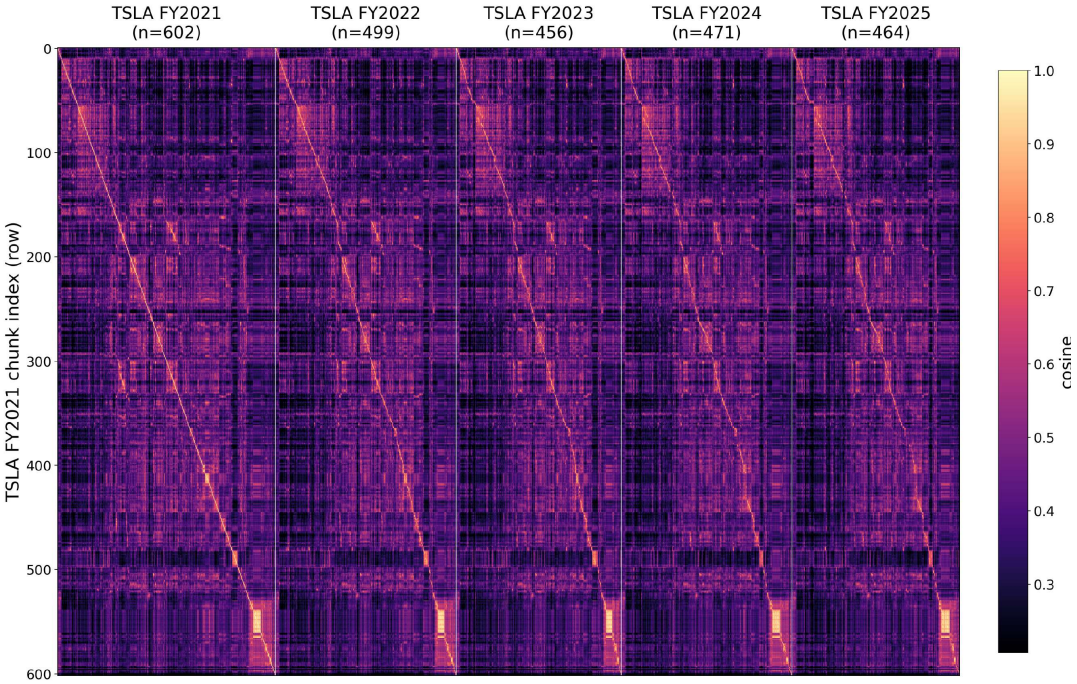


Figure 40: Temporal consistency heatmap for TSLA.

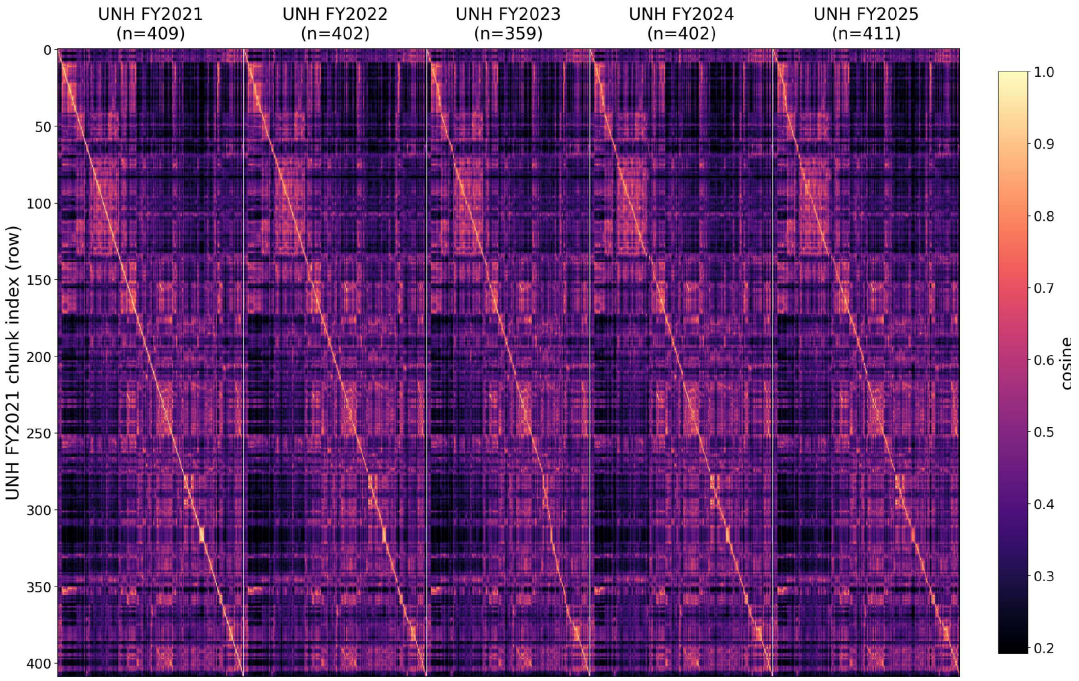


Figure 41: Temporal consistency heatmap for UNH.

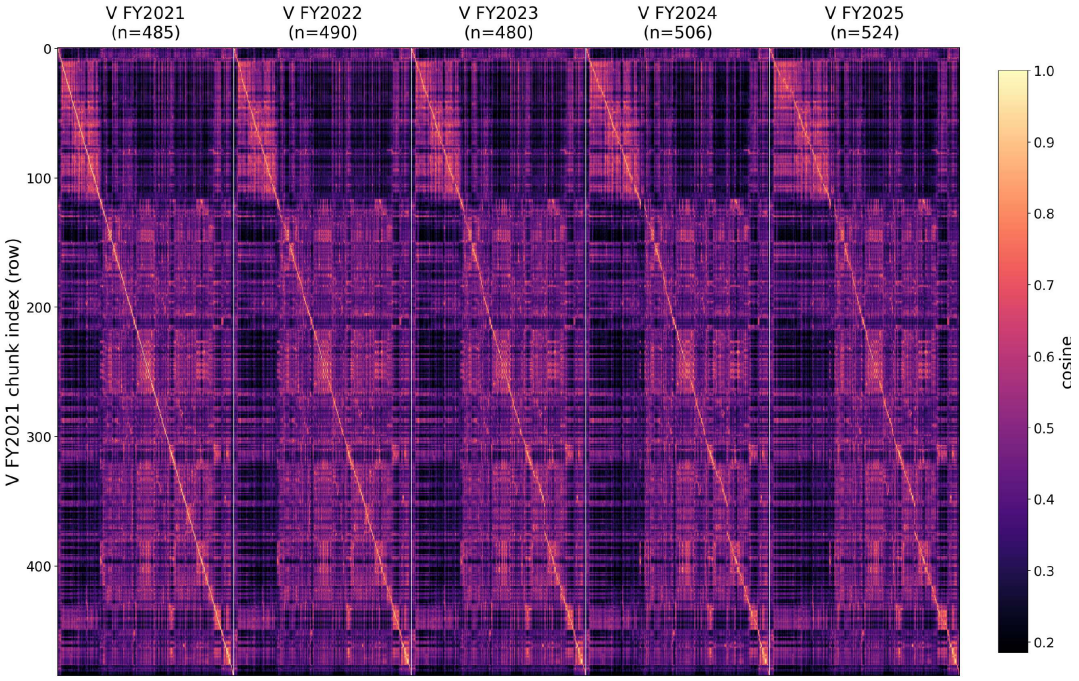


Figure 42: Temporal consistency heatmap for V.

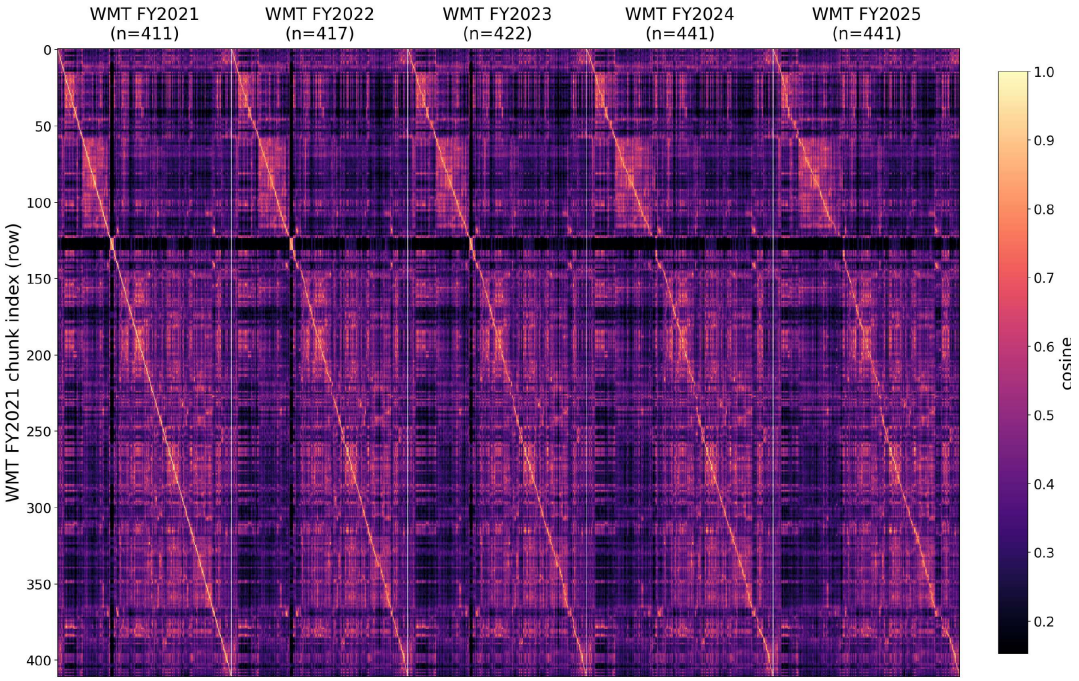


Figure 43: Temporal consistency heatmap for WMT.

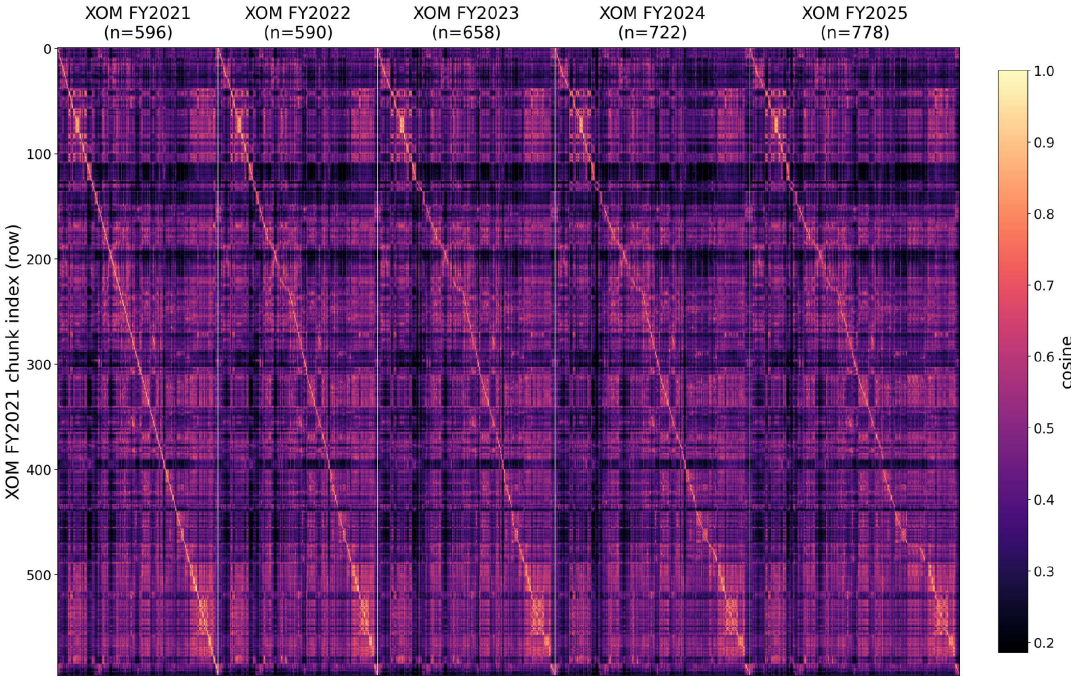


Figure 44: Temporal consistency heatmap for XOM.

J Cross-Company Comparison Heatmaps

This section presents the complete collection of cross-company comparison heatmaps for the evaluated 25 tickers, analyzed via row-wise similarities against self plus the nearest four matching cases.

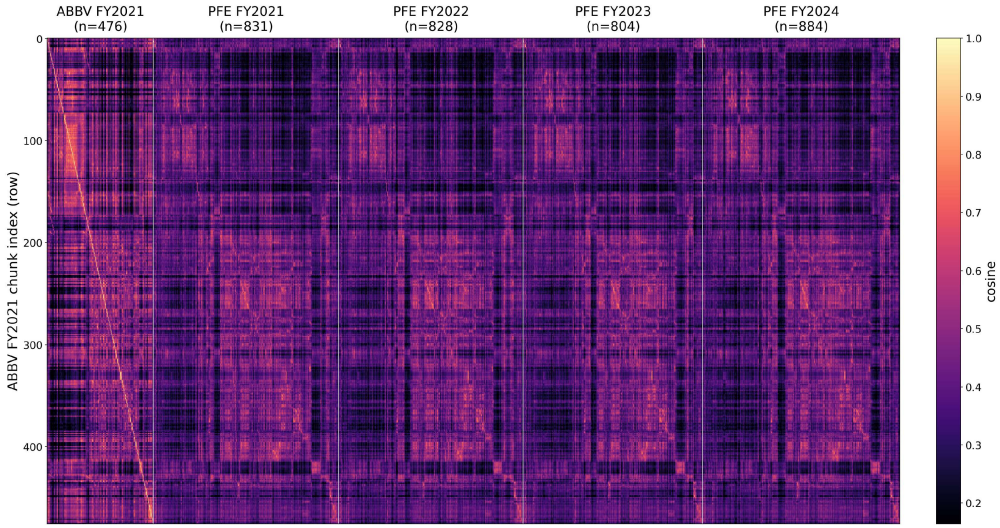


Figure 45: Cross-company comparison heatmap for ABBV.

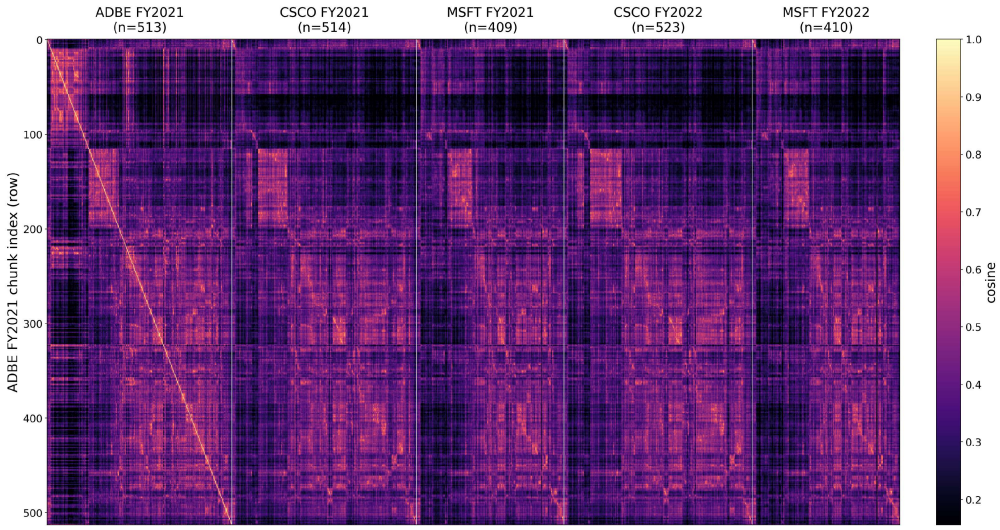


Figure 46: Cross-company comparison heatmap for ADBE.

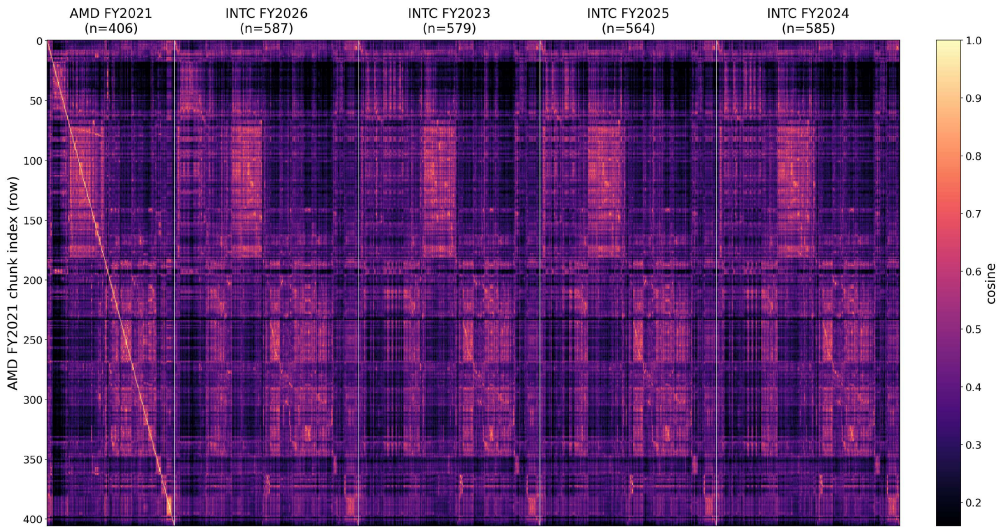


Figure 47: Cross-company comparison heatmap for AMD.

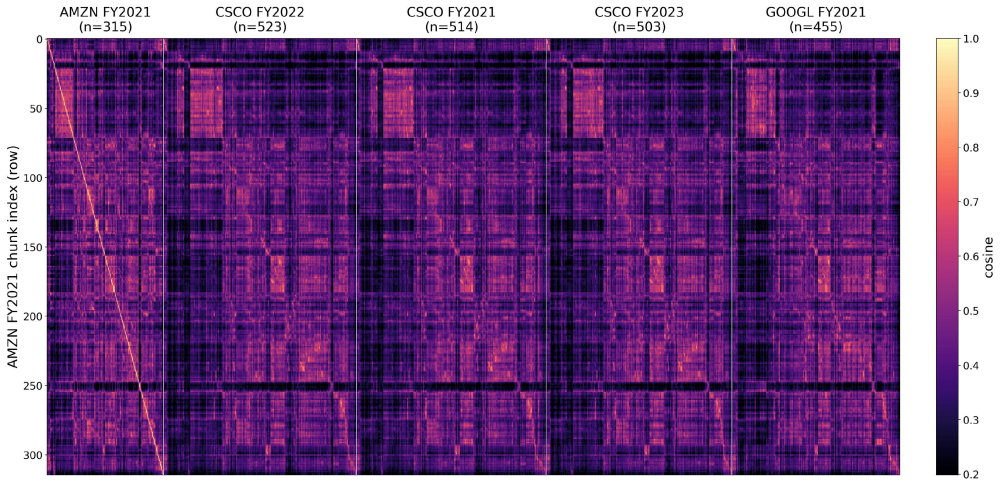


Figure 48: Cross-company comparison heatmap for AMZN.

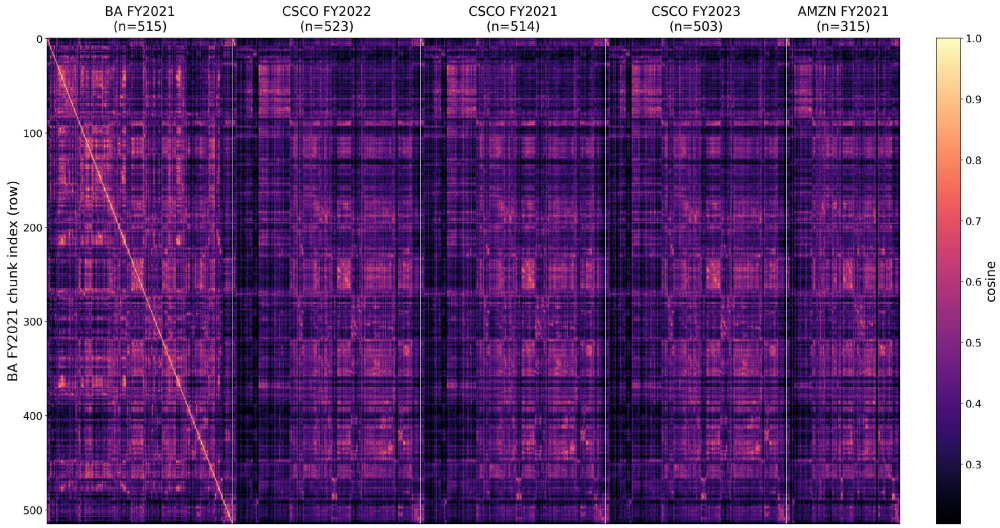


Figure 49: Cross-company comparison heatmap for BA.

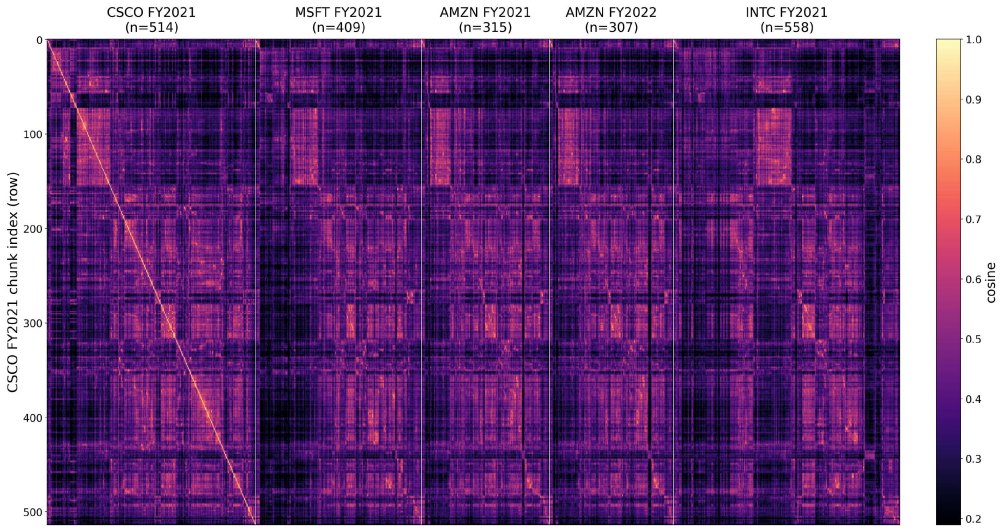


Figure 50: Cross-company comparison heatmap for CSCO.

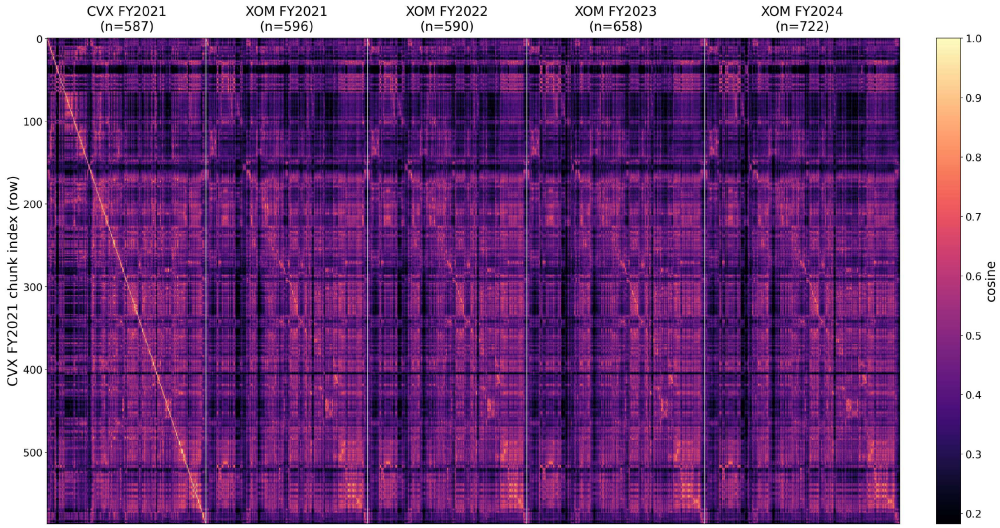


Figure 51: Cross-company comparison heatmap for CVX.

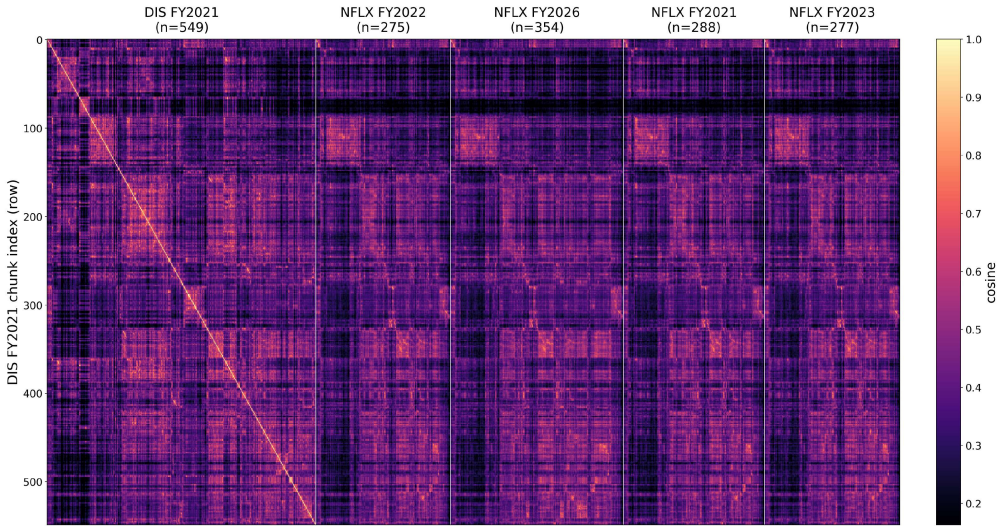


Figure 52: Cross-company comparison heatmap for DIS.

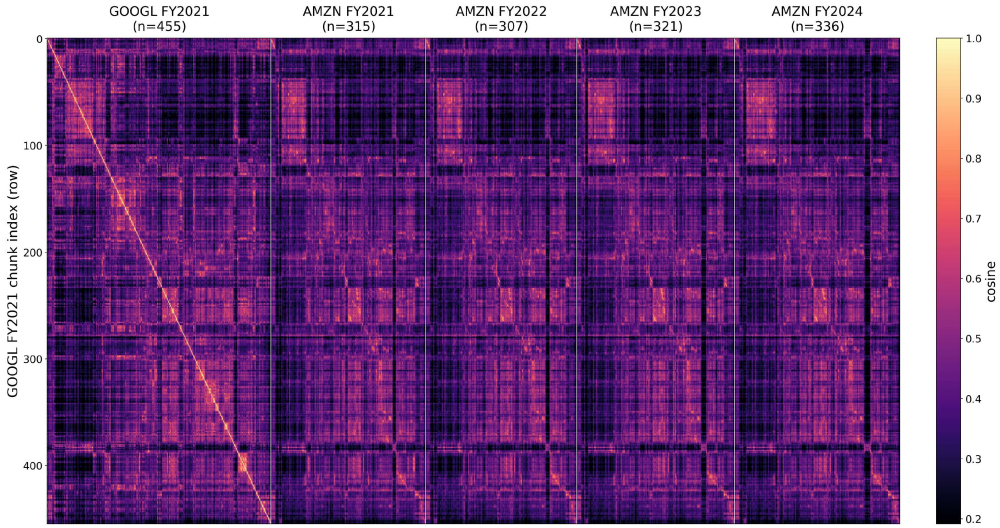


Figure 53: Cross-company comparison heatmap for GOOGL.

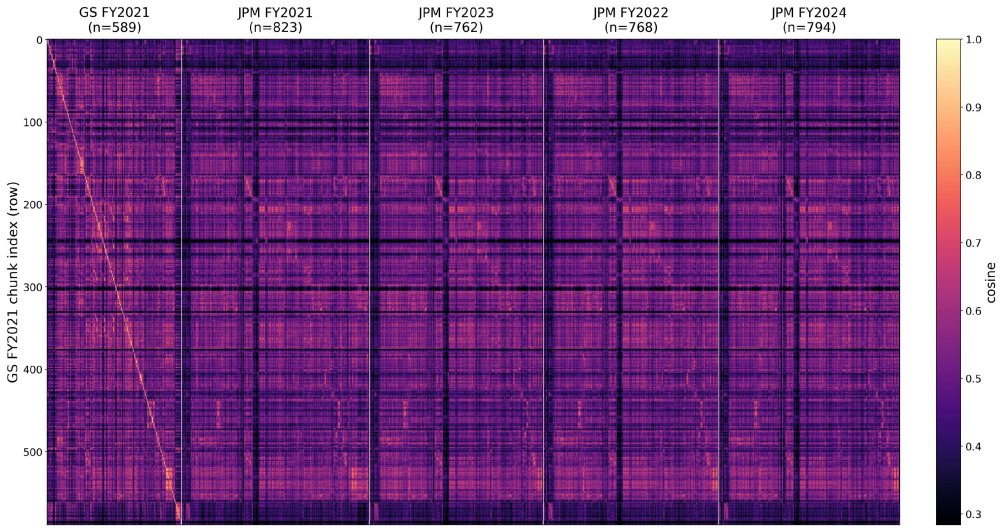


Figure 54: Cross-company comparison heatmap for GS.

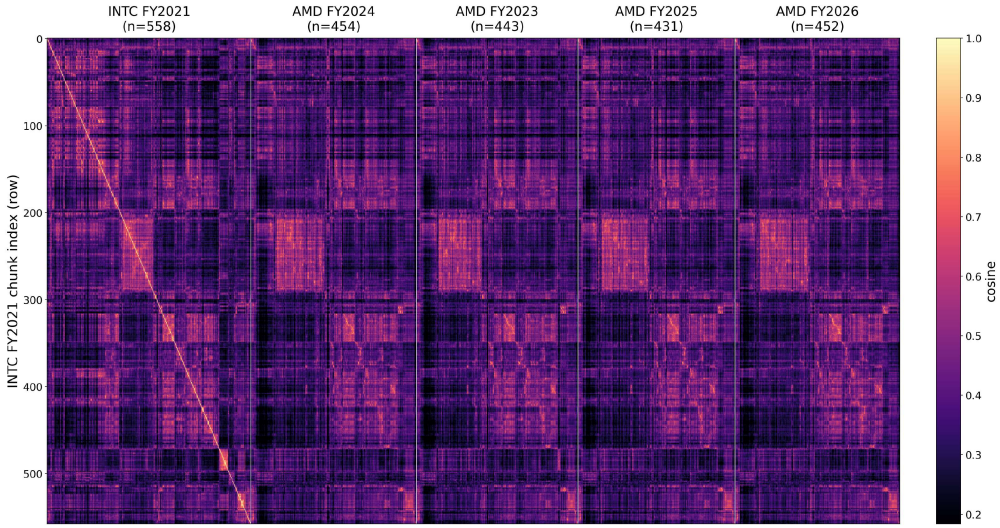


Figure 55: Cross-company comparison heatmap for INTC.

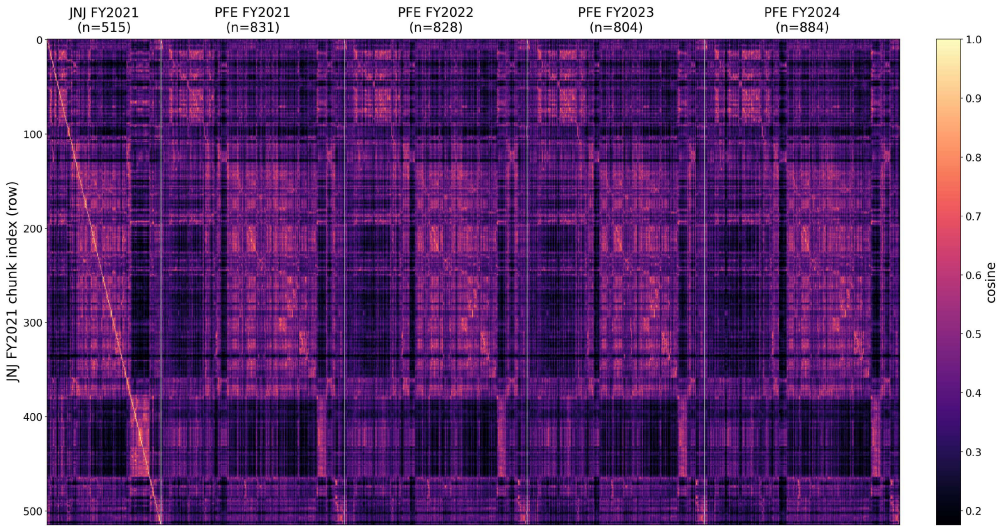


Figure 56: Cross-company comparison heatmap for JNJ.

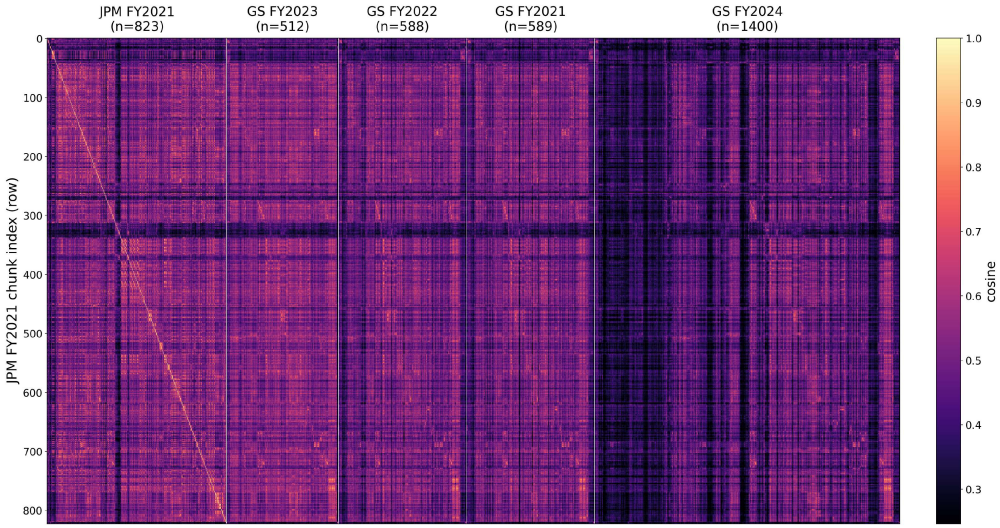


Figure 57: Cross-company comparison heatmap for JPM.

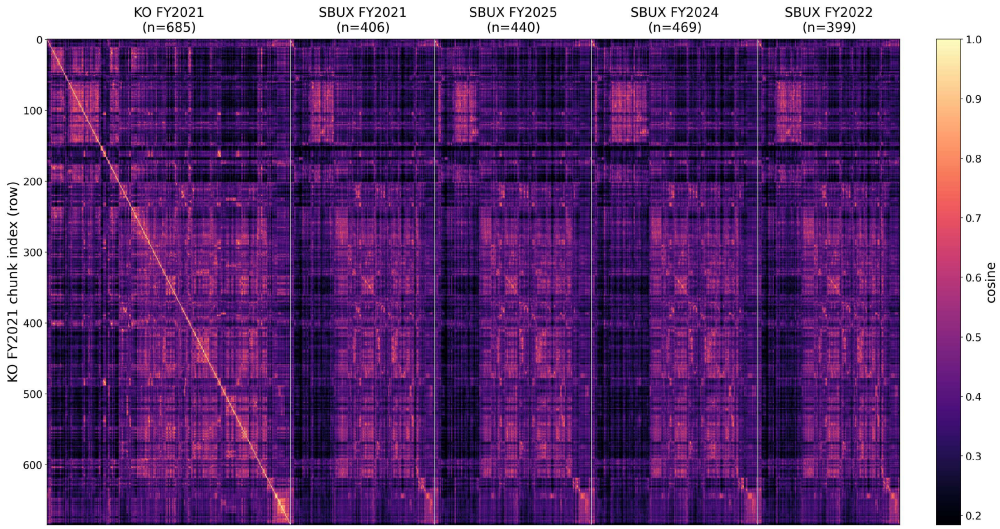


Figure 58: Cross-company comparison heatmap for KO.

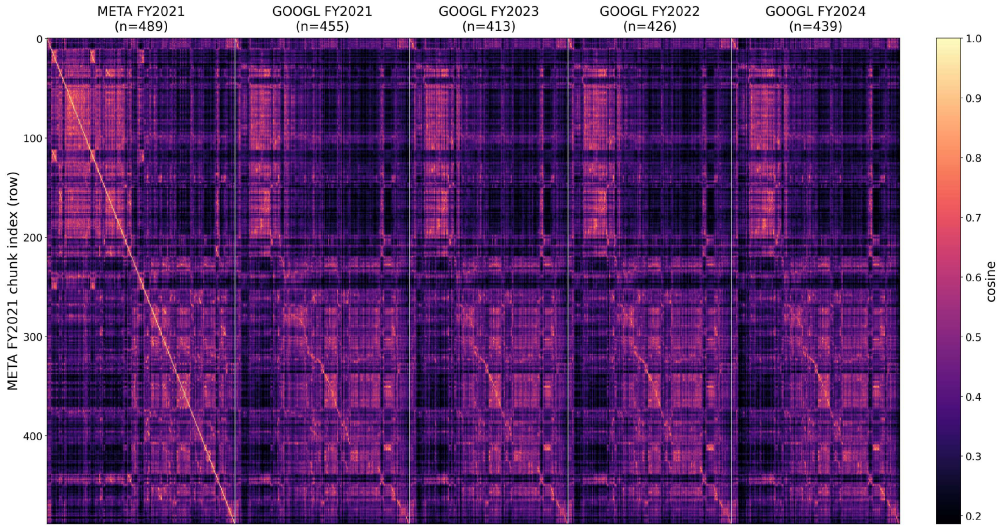


Figure 59: Cross-company comparison heatmap for META.

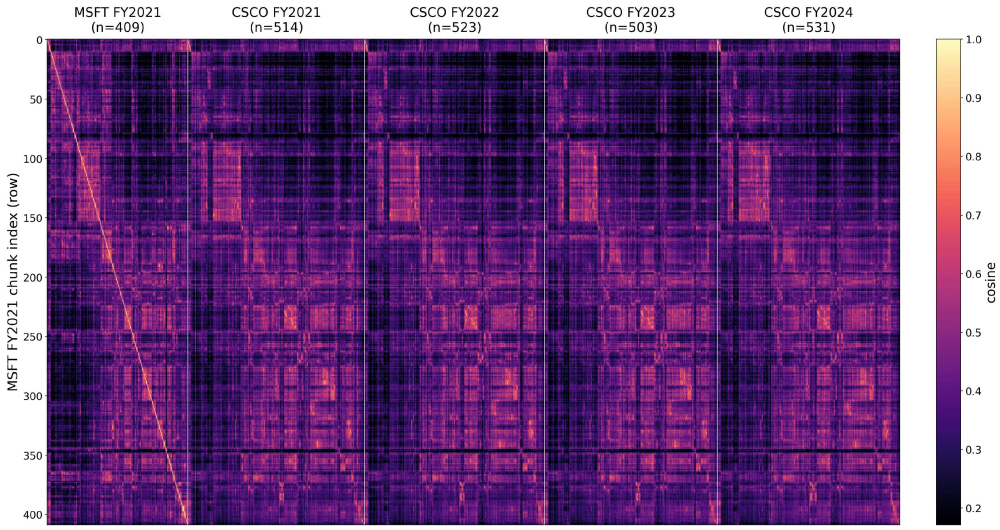


Figure 60: Cross-company comparison heatmap for MSFT.

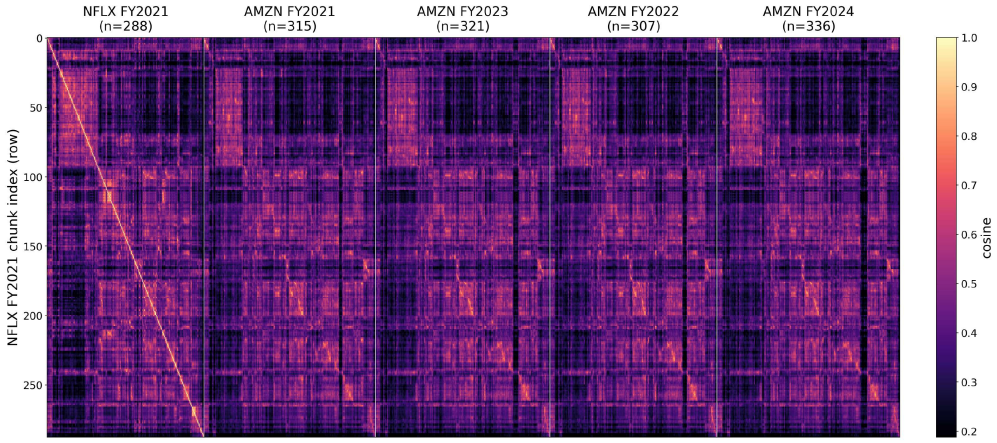


Figure 61: Cross-company comparison heatmap for NFLX.

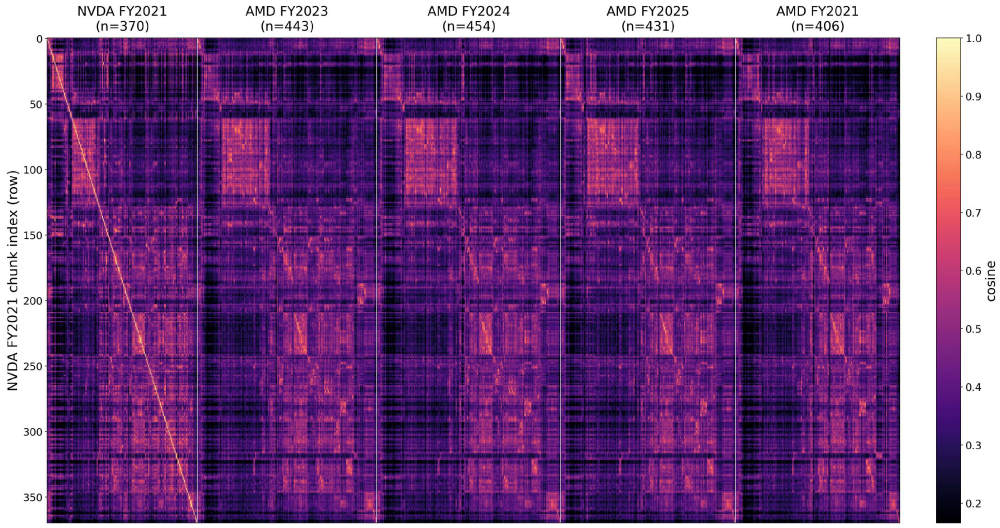


Figure 62: Cross-company comparison heatmap for NVDA.

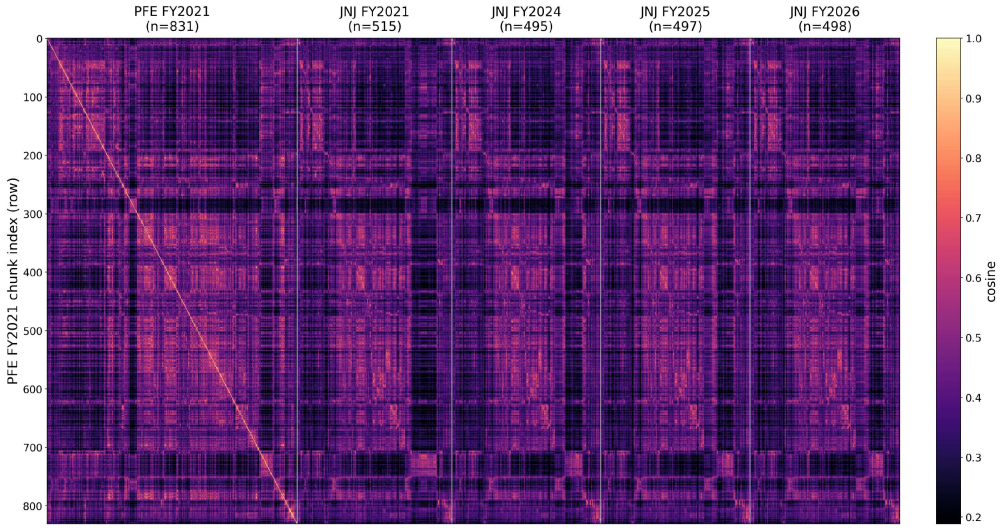


Figure 63: Cross-company comparison heatmap for PFE.

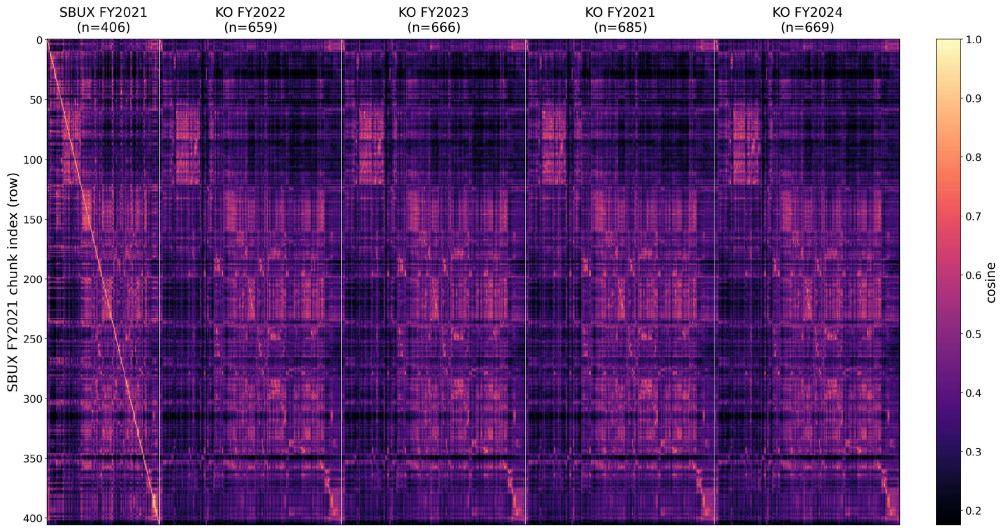


Figure 64: Cross-company comparison heatmap for SBUX.

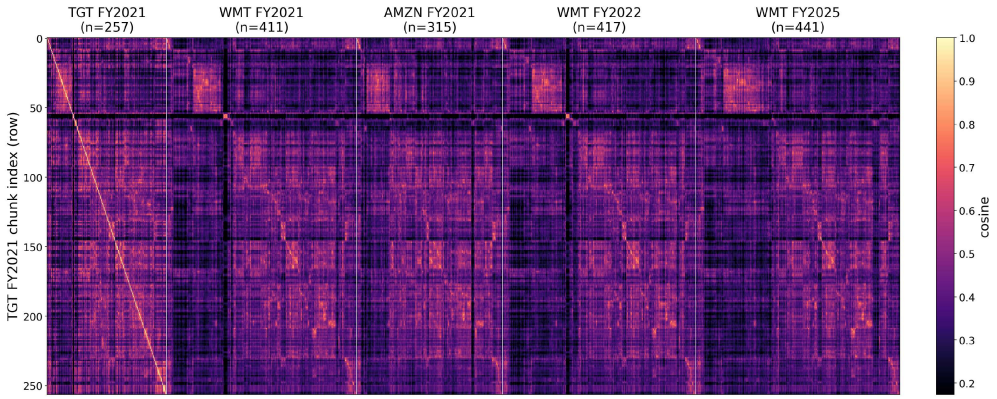


Figure 65: Cross-company comparison heatmap for TGT.

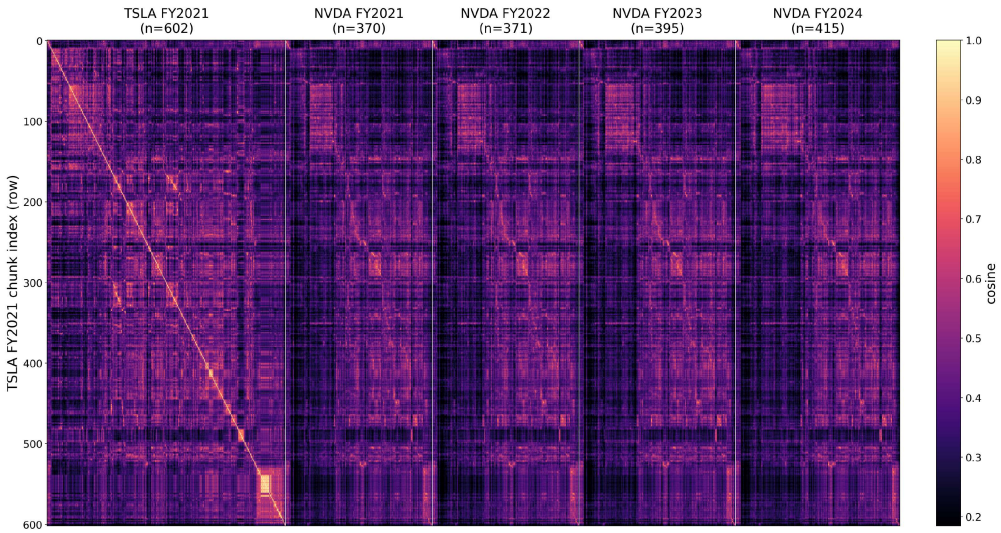


Figure 66: Cross-company comparison heatmap for TSLA.

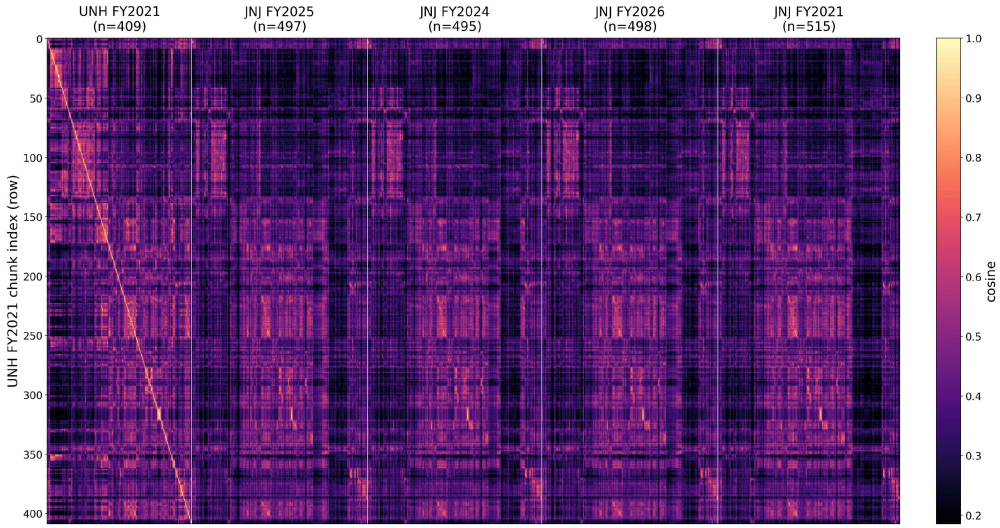


Figure 67: Cross-company comparison heatmap for UNH.

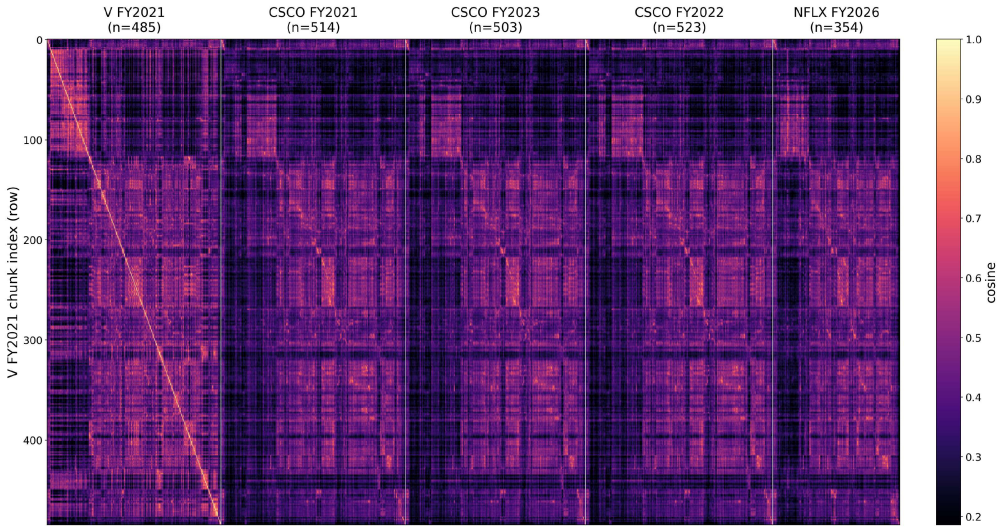


Figure 68: Cross-company comparison heatmap for V.

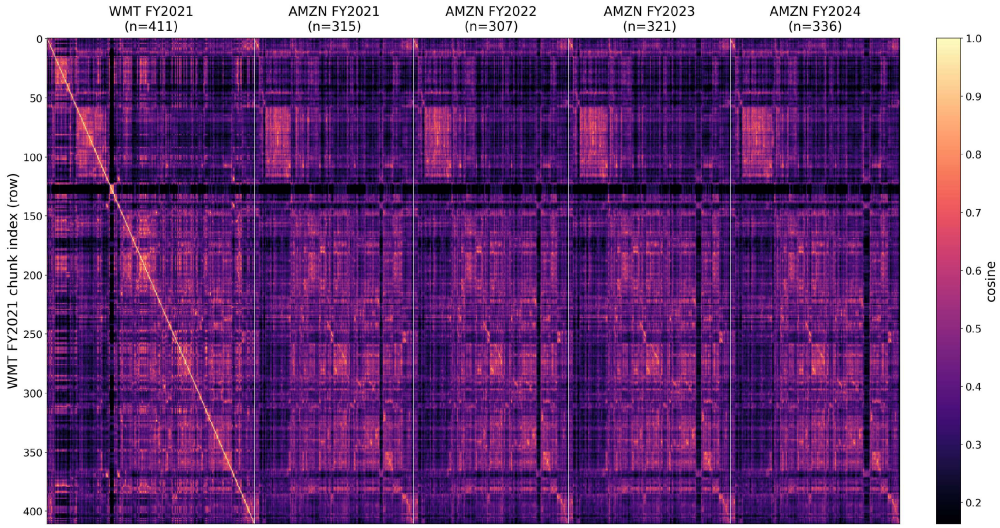


Figure 69: Cross-company comparison heatmap for WMT.

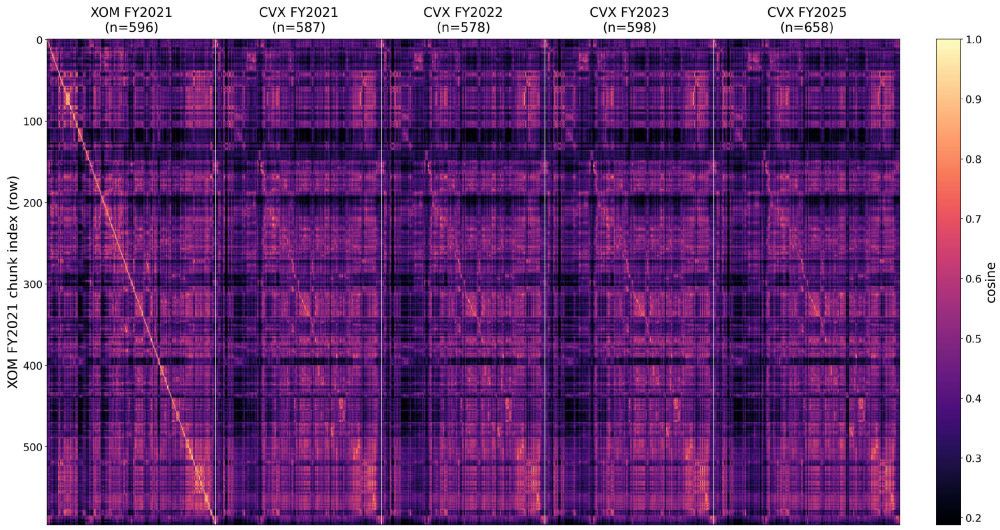


Figure 70: Cross-company comparison heatmap for XOM.