
MODELING THE EVOLUTION OF AI LANGUAGE

A Single-Firm Case Study

Marco Siliezar ^{1, 2, †}

1 Northwestern University

2 <https://meatloaf02.github.io>

† marco.siliezar@gmail.com

Abstract

With the popularity and fast evolution of artificial intelligence (AI) over the last several years, enterprise software firms have increasingly redistributed internal resources toward the development and integration of AI technologies (McKinsey & Company 2025). This includes R&D spending, product roadmaps, and strategic priorities. As noted in the McKinsey article, executives of these companies view AI as likely to drive “potentially higher margins and earnings” which translate to increases in shareholder value. This research paper attempts to answer the question, has this shareholder value been recognized?

This research aims to review the evolution of AI related language, product capabilities, market positioning, and risk disclosures from 2015 to present for Workday, Inc., a business-to-business enterprise software company. Information about this company will be extracted from various sources online including regulatory filings, official company press releases, and events. This data will be ingested and stored in a knowledge base. This will aim to contribute to a predictive model for predicting the next month return direction for this company’s stock as a proxy for shareholder value. The work completed so far: developing a structured plan, creating seed URLs from which this information will be extracted, drafting of a knowledge graph schema blueprint, and a ground-truth approach.

Table of Contents

Abstract.....	i
Introduction and Problem Statement	1
Literature Review	1
Textual Analysis of SEC Filings as Financial Signal	1
Corporate Disclosures and Knowledge Graphs	2
AI-Related Language in Disclosures and Risks.....	2
Research Gap and Contribution	2
Data	2
Methods.....	3
Results	6
Conclusions	6
References	6
Code Availability.....	8
Appendix	9

Introduction and Problem Statement

Core enterprise software-as-a-service enterprise platform companies sell to other businesses products that aim to increase productivity. Workday is a publicly traded company that builds these products for managing human resources, finances, payroll, and other core aspects of corporate and government operations. These aspects are essential to corporate operations thus there is an inherent efficiency maximization problem that is to be solved. Therefore, Workday is an interesting single-company case study to research whether its use of AI addresses this efficiency problem for its customers.

This research aims to compare the increase in AI-related language over time against the stock price. A predictive model will be powered by a knowledge graph containing information about this company. The predictive model will attempt to predict next month's stock direction. The model will not answer whether the stock should be bought or sold, rather it will output the probability the direction next month is up based on signals and measures derived from the knowledge base.

Likely users of this application will be analysts and investors conducting research who will be looking for support for a trading-style direction rule using this firm's AI strategy as input. The application will not be for users who want to answer the question, "should I buy or sell WDAY?"

Literature Review

Textual Analysis of SEC Filings as Financial Signal

Finance researchers have previously studied the link between financial tone of text in 10-K filings to financial performance and outcomes. By creating finance-domain specific word lists, Loughran and McDonald (2011) showed that these lists provide a better understanding of the impact of filing text on stock returns more than generic nonbusiness word lists. The text in filings is an important aspect but more so is its readability. When analyzing the relationship between annual report readability and company performance, Li (2008) found that firms with poor performance have more difficult to read annual reports. Thus, financial tone and readability are signals and are features worth comparing against embeddings.

Corporate Disclosures and Knowledge Graphs

Financial narrative researchers have previously explored how to process unstructured corporate disclosure documents using natural language processing techniques to store them in knowledge graphs to discover underlying networks. An approach by Cavar (2018) creates a pipeline to process 10-K filings, analyze their semantics using publicly available Python packages for NLP processing, creates a uniform data structure, and implements a knowledge graph for storing raw text as tuples. This demonstrates the feasibility of domain-specific knowledge graphs for storing and analyzing corporate disclosure data.

AI-Related Language in Disclosures and Risks

The rise of AI-related language has been quantified by industry analysis. In a report by Arize (2024), an AI observability and LLM evaluation company, quantified how often companies mention AI and how often this language is associated with risk. Their report shows that in 2024 the Fortune 500 companies mention AI 250.1% more than in 2022. These companies cited AI as a risk factor more than 473.5% more in 2024 than in 2022. This analysis supports the hypothesis about AI as an opportunity versus a risk.

Research Gap and Contribution

Prior work demonstrates that language in 10-K filings is informative, provides signals, and representing their information in a knowledge graph, provides relationships that would otherwise be nonobvious. However, there is limited work integrating longitudinal NLP, knowledge graph modeling, and firm-specific context within a single framework. This research contributes by modeling Workday, Inc. as a longitudinal single-firm case study, constructing a knowledge graph that captures the evolution of AI-related language from 2015 to present. This approach enables exploratory signals relevant to investment decision-making.

Data

To build a strong dataset for the predictive model, data from 2015 to 2026 was obtained about this single company from its official SEC filings, official corporate webpages, earnings prepared remarks, and public news coverage. Table 2 in the Appendix summarizes the data. Most of the data originates from SEC filings. All data was obtained from public and free sources. Unfortunately, richer data available is only accessible via paid subscriptions to services that provide such data.

However, the public data contains adequate context about the activity of this company including acquisitions, product capabilities, risk topics mentioned by leadership, and announcements. For example, the earnings prepared remarks dated November 25, 2025 obtained through web scraping provided the CEO's announcement about a recent acquisition of an AI platform company: "Last week, we announced the intent to acquire Pipedream, a low-code integration platform for AI agents..." (Workday, Inc., 2025). This document, as an example, contributes to high-signal product detail which can be fed into a knowledge graph that becomes the generator of features whose provenance can be traced back to this specific document.

The company announcing an acquisition is one perspective and it is important to also obtain the perspective of such announcement from analysts and the media. Most analyst reports are behind paywalls, but this may be mitigated by focused web crawling of public websites that post analyst highlights in a future iteration. Media perspective can be obtained through future focused web crawling and API endpoints. Further research will explore the GDELT database for extracting events not covered through existing sources. Really Simple Syndication (RSS) feeds will be used for incremental updates to the database. Additional further research will explore obtaining earnings call transcripts as these will provide deeper insight from unscripted Q&A with analysts.

The MVP scope of this research is to ingest 300 – 500 documents to balance comprehensive coverage with project feasibility. Ideally, these documents are distributed across the analysis period (2015 to present) with the priority of documents sourced from SEC filings. The success criteria in as far as document ingestion is that MVP is achieved when at least one document is ingested for every quarter in the analysis period, all SEC filings are included, and there are at least 30 AI-relevant documents per year at least starting in 2020. From an exploratory analysis of GDELT reveals that this may be viable.

Methods

To arrive at utilizing the predictive model for predicting the next month's return direction of this company's stock, a five-stage pipeline was created. This pipeline aims to provide the predictive model with an aggregation of quarterly signals along with entities extracted from a knowledge graph created in a prior step. The benefits of creating a knowledge graph at this stage is for identifying, exploring relationships (nodes/edges) that otherwise wouldn't be obvious in a relational database, and easy query the ingested data. The data for knowledge graph ingestion is processed through its own pipeline to extract text, dates, split

sentences, and store the processed documents in a JSON format. To summarize the architecture, the first stage ingests data from the mentioned sources and stores the raw documents in a PostgreSQL database. Table 4 in the Appendix provides a summary of the pipeline stages.



Figure 1. Architecture Overview

The ingest and process steps are as follows. Prior to ingesting data, seed URLs were hand curated. The goal of this approach is to obtain high coverage of company information and minimize noise. In this stage, sitemaps were also pulled with the purpose of enabling the web crawler to obtain documents deeper in these websites in a focused manner. The seed URLs were from a mix of content sources as seen in Table 3 in the Appendix. The focused web crawler takes the seed URL as an allowed and focused website list to crawl and fetches documents while maintaining compliance to robots.txt and implements rate limiting for polite crawling. Additionally, the data was normalized and deduped since duplicate data is surely to exist as information is sourced from various web locations. This web information extraction implementation uses the Beautiful Soup (Richardson, 2025) Python package to extract data from HTML files ingested from the internet, for example, for SEC filing documents. The ingested data is stored in a PostgreSQL table using the SQL Alchemy package (SQLAlchemy, 2026) to handle document metadata storage for hand-off to the knowledge graph.

One manual intermediate step was taken to define seed data for entities. The goal was to simplify the information extraction and to accurately populate the knowledge graph without noise. Seed data for entities included company products, capabilities and risk topics. Of course, a seed company was defined. However, the knowledge graph is abstracted enough and isn't one-company specific. It can be applied to any similar company. Table 1 in the Appendix provides examples.

The knowledge base was constructed using a curated approach. The curated method (Nickel et al., 2015) was chosen to leverage the author's expert knowledge of this company (and its subsector) to create an accurate representation. With this approach, a simple but

effective set of triples was created to construct the knowledge graph. A visual representation of the knowledge graph schema is in Figure 2.

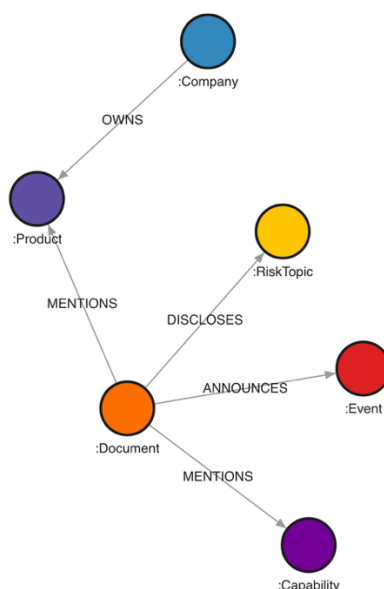


Figure 2. Knowledge Graph Schema (Memgraph)

The knowledge base schema includes entities and relationships that are Resource Description Framework (RDF) compatible. As the goal for this KG is to deliver buy/sell signal features which originate from information in documents created at a point in time, the database is designed to store documents which contain excerpts about mentions and claims. These mentions will become the evidence for the signal. Events will also become important. Events include filings, earnings calls, acquisition announcements, among other activity. These events will be time anchors. If analyst reports become obtainable such that they can be stored in the database, they are also documents and measures. Additionally, because Cypher (Neo4j, 2026) is used as the language to query the knowledge database, properties were declared for relationships to leverage this Cypher capability, particularly for maintaining timestamps. The complete knowledge graph schema is summarized in the Appendix: the entities in Table 3 and the relationships and their properties in Table 4.

An important ontological choice was the definition of an event for the purposes of the knowledge graph. A concrete definition was created to minimize scope and not introduce additional complexity from such techniques as named entity recognition. An event is either an earnings call, product launch, acquisition, partnership, leadership change, conference, or regulatory filing.

Results

Conclusions

The knowledge base that will be built for this research is designed to power a predictive model for the analyst and investor with the ability to answer questions such as: “What AI capabilities does Workday claim, where, and how often?”, “Which products are emphasized over time?”, “What risk topics are emerging?”. Additionally, it will power buy and sell signal features from media measures, impacts of events, and shifts in the language of filings over time. This depends on the quality and quantity of data which can be gathered from public sources. At the same time, the timespan of 2015 to present should provide sufficient data for the model. Lastly, this application will not claim certainty for a buy or sell position of a stock in the company, and it will not claim it has solved stock prediction. The success of this application in production for its users paves the way to implement future applications to additional B2B SaaS companies in the information technology sector.

References

- Arize AI, Inc. 2024. “The Rise of Generative AI in SEC Filings”. <https://arize.com/wp-content/uploads/2024/07/The-Rise-of-Generative-AI-In-SEC-Filings-Arize-AI-Report-2024.pdf>
- Bayer, Michael. 2026. “SQLAlchemy”. Accessed Feb 2, 2026. <https://www.sqlalchemy.org/>
- Cavar, Damir and Matthew Josefy. 2018. “Mapping Deep NLP to Knowledge Graphs: An Enhanced Approach to Analyzing Corporate Filings with Regulators.” *Proceedings of The first financial narrative processing workshop (FNP 2018)*.
- Kejriwal, Mayank, Craig A. Knoblock, and Pedro Szekely. 2021. *Knowledge Graphs: Fundamentals, Techniques, and Applications*. Cambridge, MA: MIT Press.
- Li, Feng. 2008. “Annual report readability, current earnings, and earnings persistence”. *Journal of Accounting & Economics* 45, no. 2 – 3 (August 2008): 221 – 247. <https://doi.org/10.1016/j.jacceco.2008.02.003>

Loughran, Tim and Bill McDonald. 2011. “When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks”. *The Journal of Finance* 66, no. 1 (February 2011): 35 – 65.
<https://doi.org/10.1111/j.1540-6261.2010.01625>.

McKinsey & Company. 2025. “The AI-centric imperative: Navigating the next software frontier. October 16, 2025. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-ai-centric-imperative-navigating-the-next-software-frontier>

Neo4J. 2026. “Cypher Manual”, version 25. <https://neo4j.com/docs/cypher-manual/current/introduction/>

Nickel, Maximillian, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. “A Review of Relational Machine Learning for Knowledge Graphs.” *Proceedings of the IEEE*, 104(1): 11-33.
<https://arxiv.org/abs/1503.00759>

Richardson, Leonard. 2015. “Beautiful Soup”. Accessed Feb 2, 2026.
<https://www.crummy.com/software/BeautifulSoup/>

Workday. 2025. “Third Quarter Fiscal 2026 Prepared Remarks”. November 25, 2025.
https://investor.workday.com/files/doc_financials/2026/q3/workday-fiscal-2026-third-quarter-prepared-remarks.pdf

Code Availability

The essential code to reproduce this research are available at <https://github.com/meatloaf02/KG.git>. The repository includes a comprehensive README file.

Appendix

URL	Content	Document Count
https://blog.workday.com/	Corporate blog posts	7
https://investor.workday.com/	Investor Relations official website	8
https://techcrunch.com/tag/workday/	Public news coverage	1
https://www.prnewswire.com/	Press releases	2
https://sec.gov	10-k, 10-Q, 8-K filings, proxy statements	784

Table 1. Data Manifest

Entity	Seed Data Examples
Company	Workday
Product	Workday-HCM, Workday-Financials, Workday-Payroll
Capability	AI, ML, Predictive Analytics, Automation, LLM
Risk Topic	Cybersecurity-Risk, Data-Breach, Regulatory-Compliance, AI-Ethics

Table 2. Seed Examples

Entity	Field	Notes
Document		
	Content Hash	
	URL Hash	
	Doc Type	10-K, 10-Q, Press Release, etc.
	Source Type	SEC Filing, Investor Relations, etc.
	Publish Date	
	Published At	Converted to native Memgraph Date
	Source URL	
	Created At	
	Updated At	
Company		
	ID	
	Name	
	Ticker	
	SEC CIK	
	Created At	
Product		
	ID	
	Name	
	Description	
	First Seen	
	Created At	

	Updated At	
Capability		
	ID	
	Name	
	Category	AI, ML, Analytics, Automation, etc.
	First Seen	
	Created At	
	Updated At	
Risk Topic		
	ID	
	Name	
	Category	
	First Seen	
	Created At	
	Updated At	
Event	ID	
	Name	
	Event Type	Earnings Call, Product Launch, etc.
	Event Date	
	Description	
	Created At	
	Updated At	

Table 3. KG Entities

Relationship	Property	Notes
Mentions; Discloses; Announces		
	Evidence Text	Text Span
	Sentence ID	Sentence Identifier
	End Char	End Offset
	Confidence	0 - 1
	Extracted At	Date Time
Has_Capability		
	First Seen	
	Created At	Timestamp
Owns		
	Created At	Timestamp

Table 4. KG Relationships