Uncovering Information: Can AI Tell Us Where to Look?

Anna Costello, Bradford Levy, Valeri Nikolaev (2024)

Present by Li Ziming

Motivation

- Disclosure redundancy
  - Stakeholders need quickly and easily to find **relevant information**.
  - But information buried in sea of unstructured disclosure released by firm.
- Information content and associated market reactions
  - Prior studies mostly focus on quantitative information (Ball and Brown, 1968).
  - Narrative information highly **multidimensional** and challenging to find surprise.
  - What is expected values for textual data is not well-defined.

Motivation

- Measuring new information in narrative disclosures
  - **Properties of the text**(tone, proportion of words belong to reference dictionary)
  - Rely on subjective classification of words and unable to assess surprise.
  - **Similarity between documents** (bag of words)
  - Difficult to compare different disclosure type from period to period.
  - **Shannon information**
  - Consistent with “surprise” used in prior literature
  - Same modeling objective as LLMs (standard causal language modeling).

Research Questions

- (1) How to measure surprise in narrative disclosures?
- (2) Where new information is most prevalent?
- (3) When surprise is released?
- (4) Can explain market reaction?

Contribution

- Contribute to literature using output of LLMs to predict returns.
  - Prior literature: LLM transform a source document into another representation (Lopez-Lira and Tang, 2023; Kim et al., 2023; Chen et al., 2022).
  - Extend: preserve source and overlay measure of information eliminating **output hallucination** and **lookahead bias** in predictability.
- Contribute to literature on costly information processing.
  - Prior literature: measure informativeness as similarity of text across annual and quarterly reports (Cohen et al., 2020).
  - Extend: develop a method for **forming priors** and **identifying new information**.

Information theory

- Shannon information
  - Information of a message depends on degree to which content is surprising.
  - The message is more informative, if a highly unlikely event occurs.
  - Information content of an event E (self-information): $I(E) = -\log p(E)$
- Definition of information in a given textual disclosure
  - Each disclosure considered as a sequence of random tokens: $\widetilde{D} = \{\widetilde{\tau_1}, \widetilde{\tau_2}, \dots, \widetilde{\tau_n}\}$
  - A set of publicly available information set: $\Omega_t = \{D_j | j \in J\}$
  - $I_t(D) = -\log p_{\Omega_t}(\tau_1, \dots, \tau_n) = -\sum_{i=1}^{n} \log p_{\Omega_t}(\tau_n | \tau_{n-1}, \dots, \tau_1)$

Sample Construction

Criteria	N-firms
Valid link between Compustat and CRSP as of 2006-12-31	6,312
Financial statement criteria: $atq > 0$, $saleq > 0$, $csohq > 0$, $prccq > 0$, ibq.notnull()	5,643
Stock criteria: $shrcd <= 11$, $primexch \in \{A, N, Q\}$	4,548
Non-financial firms	3,558
Random sample	500
Decile 0: (\$0, \$39]	50
Decile 1: (\$39, \$85]	50
Decile 2: (\$85, \$152]	50
Decile 3: (\$152, \$263]	50
Decile 4: (\$263, \$426]	50
Decile 5: (\$426, \$713]	50
Decile 6: (\$713, \$1,194]	50
Decile 7: (\$1,194, \$2,212]	50
Decile 8: (\$2,212, \$5,928]	50
Decile 9: (\$5,928, >\$5,928)	50

Sample Construction

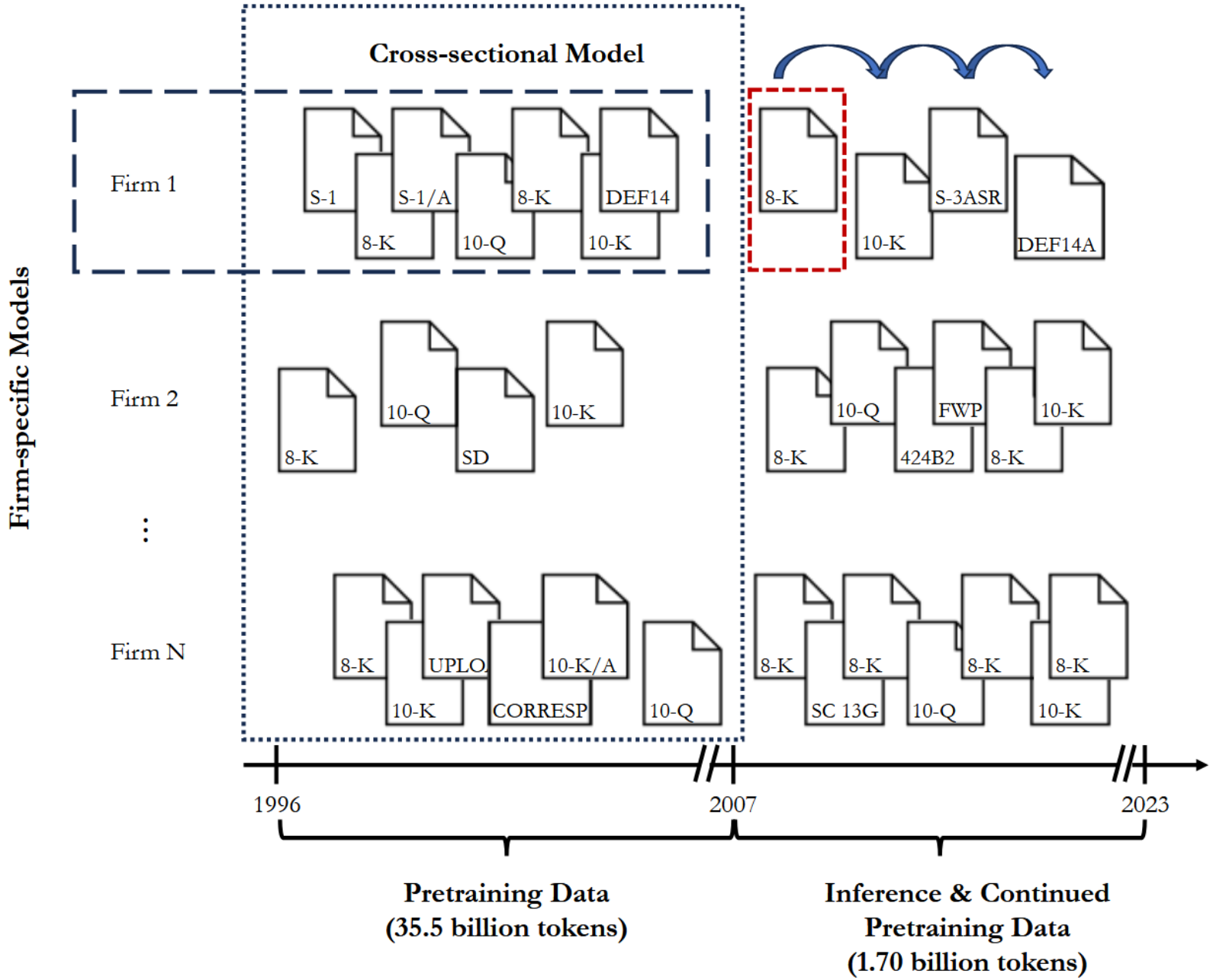
- Extracting narrative disclosure
  - BeanCounter corpus (Wang and Levy, 2024)
  - Plain text versions of all filings accepted by SEC EDGAR system from 1996 to 2023
  - Annual and quarterly reports (10-K, 10-Q), current reports (8-K), proxy statements (DEF 14A), registrations of material M&A information (S-4)
- Tokenization
  - BPE: Byte-Pair Encoding (Sennrich et al., 2016)
  - Train a tokenizer from scratch using **in-sample** disclosures within corpus

Model Training and Inference

- Estimating investors' prior beliefs
  - $\theta = argmin_\theta \sum_{j=1}^{m} \sum_{i=1}^{n_j} -\log \widehat{p_{\Omega_t}}\left(\tau_{j,i} \middle| \tau_{j,i-1}, \dots, \tau_{j,i-k}; \theta\right)$
  - $\widehat{p_{\Omega_t}}\left(\tau_{j,i} \middle| \cdot; \theta\right)$ is a GPT-style model parameterized by θ, k is size of context window.
- Model architecture selection
  - Train models following Pythia scaling suite (Biderman et al., 2023)
  - Decoder-only Transformer (GPT style)
  - Baseline model size: 410M parameters

Model Training and Inference

- Cross-sectional model
  - Pre-train **a single model** on narrative disclosures of all firms from 1996 to 2007
- Firm and time-specific models
  - Initialize cross-sectional model and continue pretrain using a single firm's disclosures prior to 2007 to yield a **firm-specific model** for each firm.
  - Iteratively apply each firm-specific model to firm's new disclosures in time order from 2007 through 2023 to measure information content and then update firm-specific model by continued pretraining on the new disclosure.


Cross-sectional Model
Firm-specific Models
Firm 1
S-1
S-1/A
8-K
DEF14
8-K
10-Q
10-K
8-K
10-K
S-3ASR
DEF14A
Firm 2
8-K
10-Q
SD
10-K
8-K
10-Q
424B2
FWP
8-K
10-K
Firm N
8-K
10-K
UPLO
CORRESP
10-K/A
10-Q
8-K
SC 13G
8-K
10-Q
8-K
10-K
8-K
1996
2007
2023
Pretraining Data
(35.5 billion tokens)
Inference & Continued
Pretraining Data
(1.70 billion tokens)

Measure of Information

- Information of the i-th token
  - Given disclosure D produced by firm f at time t, apply firm and time-specific prior
  - $I(\tau_i) = -\log\widehat{\Omega_{f,t}}\left(\tau_i \middle| \tau_{i-1}, \dots, \tau_{i-k}; \theta_{f,t}\right)$
- Information of a giving filing
  - Mean of information of all tokens
  - Weight given to each token is equal regardless of where it is located
- High information
  - Label token τ_i as “high information” if $I(\tau_i)$ is in the top quartile across all filings.

Exhibit 99.1

Apple Reports First Quarter Results

Revenue Exceeds $7 Billion; Record Profit of $1 Billion

CUPERTINO, California—January 17, 2007—Apple® today announced financial results for its fiscal 2007 first quarter ended December 30, 2006. The Company posted record revenue of $7.1 billion and record net quarterly profit of $1.0 billion, or $1.14 per diluted share. These results compare to revenue of $5.7 billion and net quarterly profit of $565 million, or $.65 per diluted share, in the year-ago quarter. Gross margin was 31.2 percent, up from 27.2 percent in the year-ago quarter. International sales accounted for 42 percent of the quarter's revenue.

Apple shipped 1,606,000 Macintosh® computers and 21,066,000 iPods during the quarter, representing 28 percent growth in Macs and 50 percent growth in iPods over the year-ago quarter.

"We are incredibly pleased to report record quarterly revenue of over $7 billion and record earnings of $1 billion," said Steve Jobs, Apple's CEO. "We've just kicked off what is going to be a very strong new product year for Apple by launching Apple TV and the revolutionary iPhone."

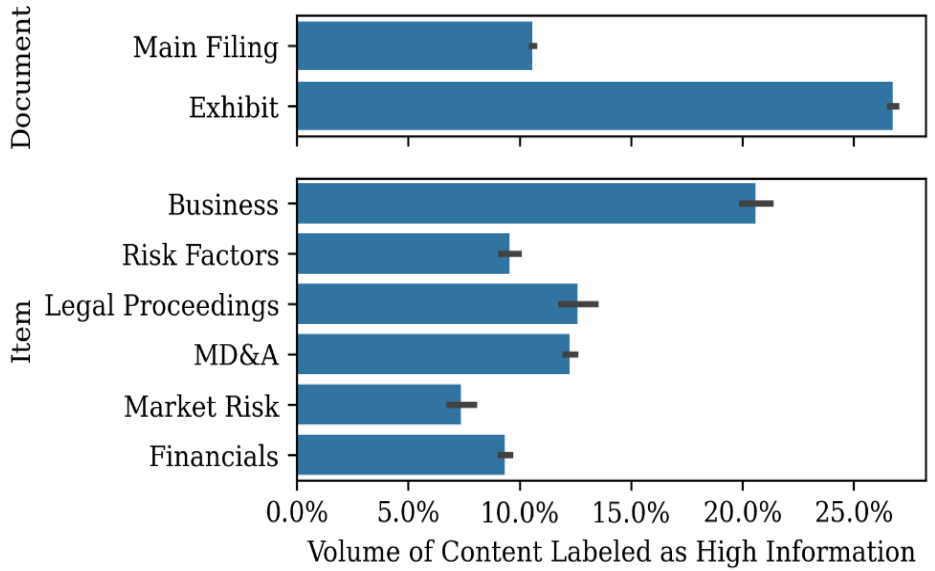
"We generated over $1.75 billion in cash during the quarter to end with $11.9 billion," said Peter Oppenheimer, Apple's CFO. "Looking ahead to the second fiscal quarter of 2007, we expect revenue of $4.8 to $4.9 billion and earnings per diluted share of $.54 to $.56."

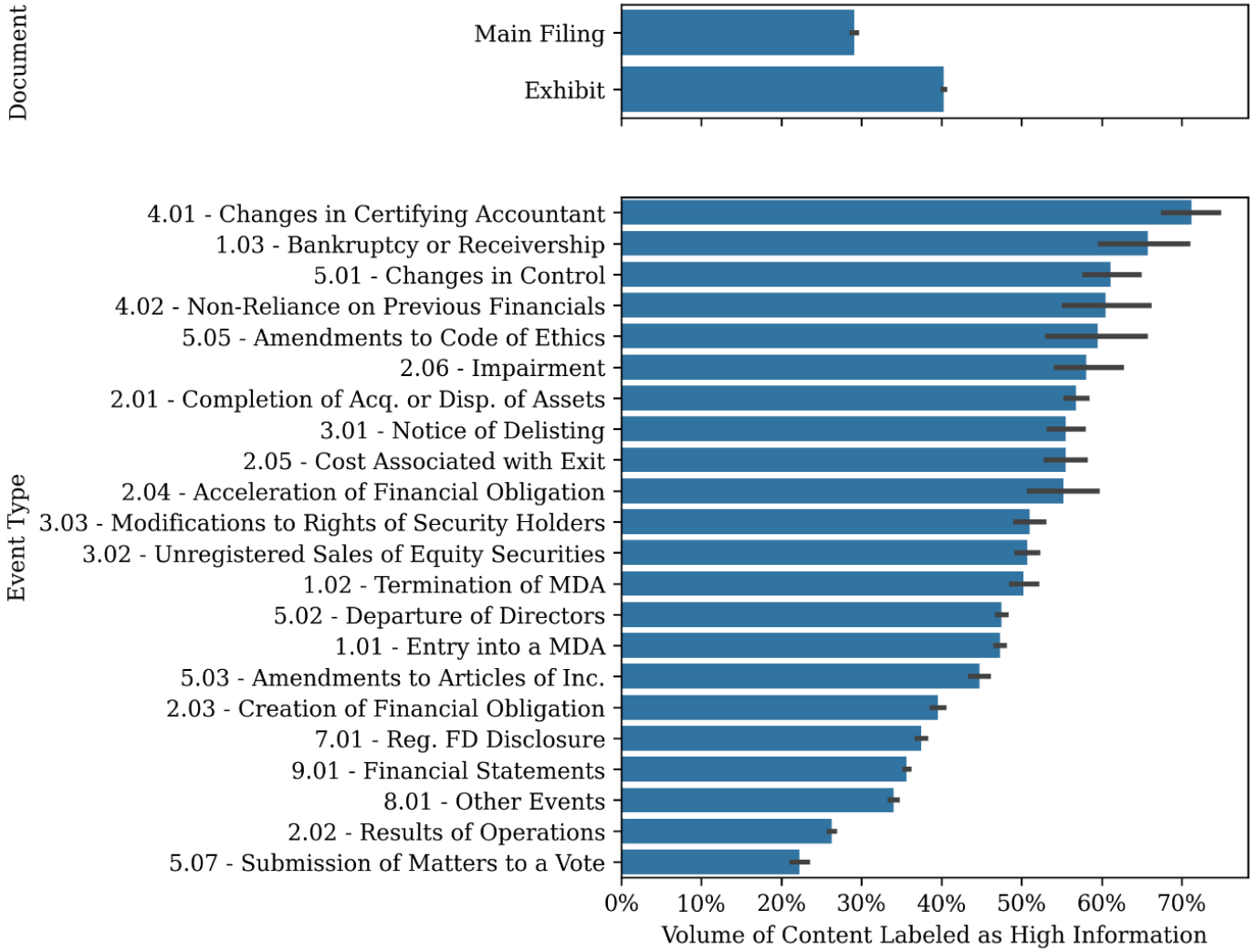
Apple will provide live streaming of its Q1 2007 financial results conference call utilizing QuickTime®, Apple's standards-based technology for live and on-demand audio and video streaming. The live webcast will begin at 2:00 p.m. PST on Wednesday, January 17, 2007 at www.apple.com/quicktime/qtv/earningsq107/ and will also be available for replay. The QuickTime player is available free for Macintosh and Windows users at www.apple.com/quicktime.

Location of Information

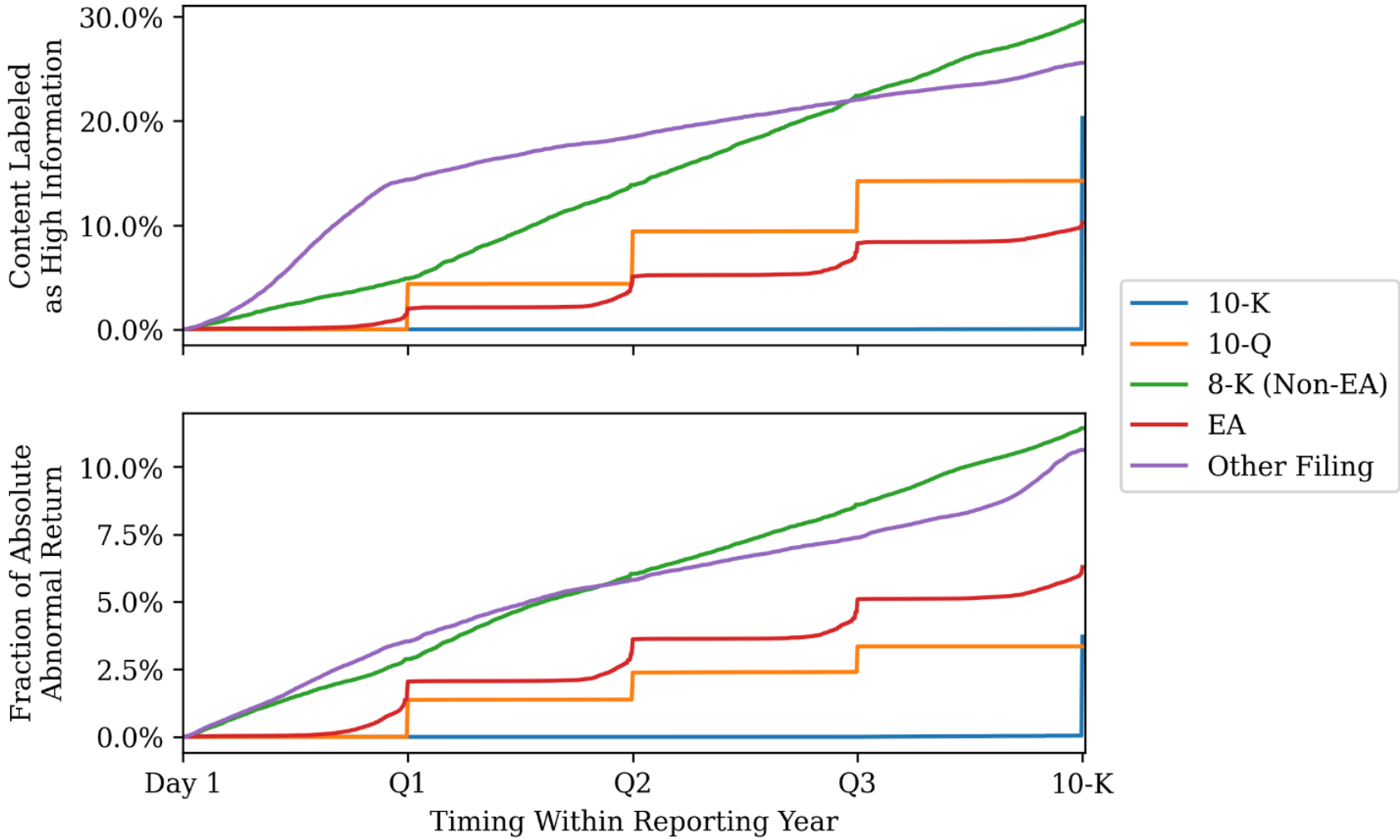
	Annual Reports (Forms 10-K)		Quarterly Reports (Forms 10-Q)		Current Reports (Forms 8-K)		Other Reports	
	Main	Exhibits	Main	Exhibits	Main	Exhibits	Main	Exhibits
Mean	1.244	1.430***	1.046	0.770***	1.533	2.854***	1.898	1.143***
Percentile								
5-th	0.632	0.086	0.424	0.000	0.067	0.241	0.001	0.001
25-th	0.906	0.361	0.703	0.022	0.354	1.451	0.348	0.004
50-th	1.153	0.905	0.967	0.121	1.126	2.553	1.162	0.180
75-th	1.472	1.957	1.300	0.821	2.433	3.980	2.983	1.190
95-th	2.147	4.779	1.933	4.127	4.258	6.416	5.976	5.954

Location of Information

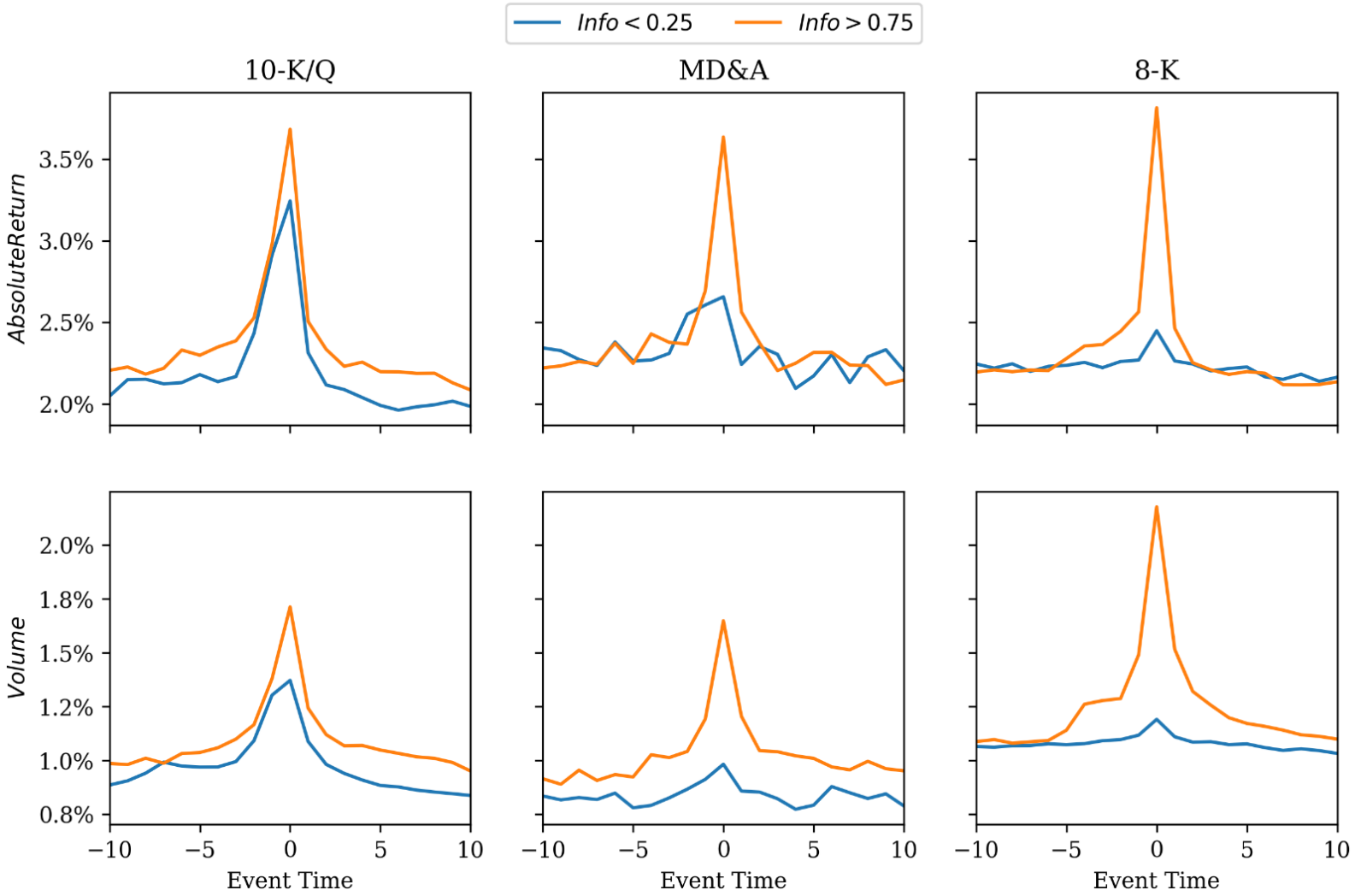


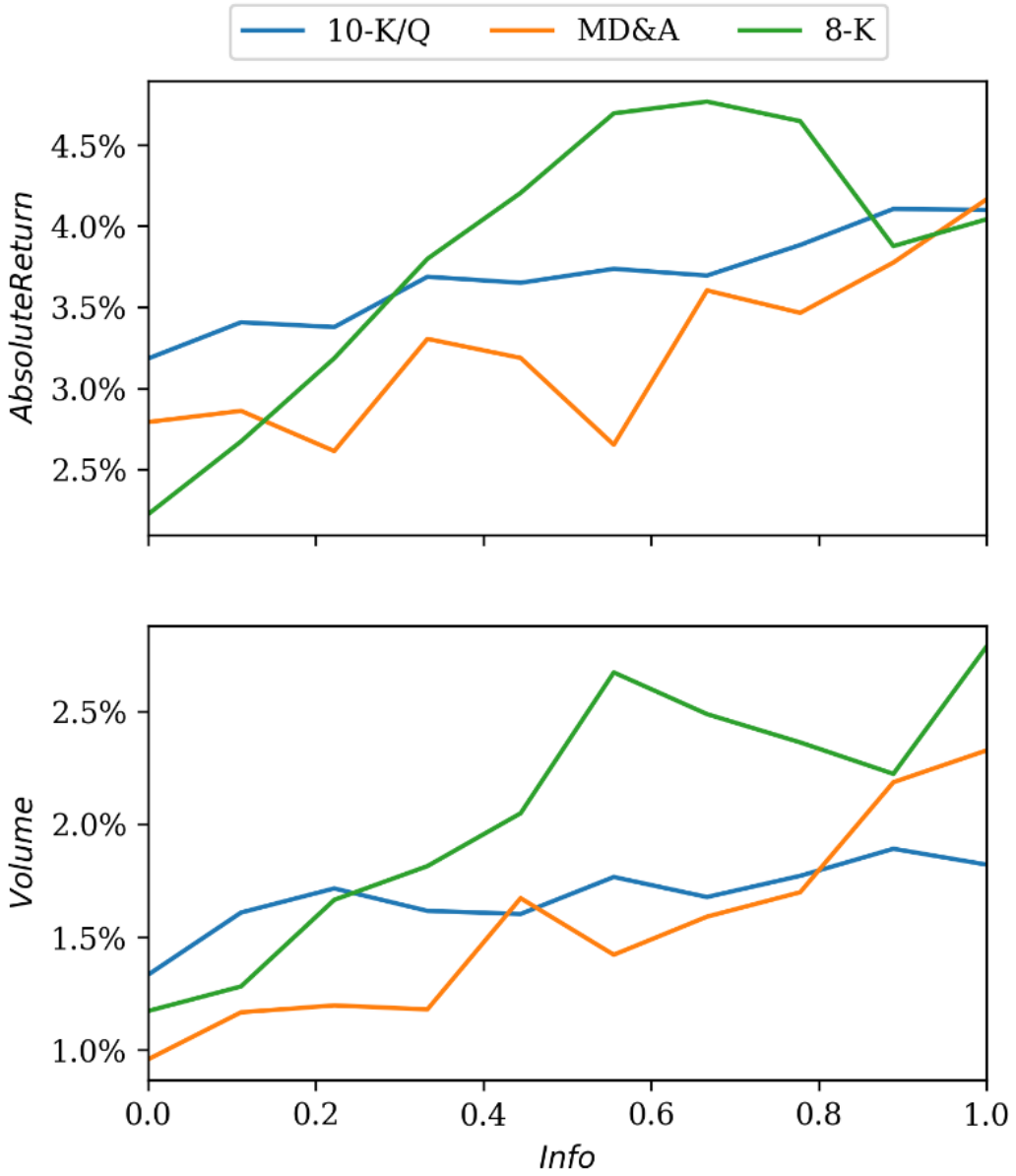


Timing of Information Release



Market Reaction




10-K/Q
MD&A
8-K
AbsoluteReturn
4.5%
4.0%
3.5%
3.0%
2.5%
Volume
2.5%
2.0%
1.5%
1.0%
0.0
0.2
0.4
0.6
0.8
1.0
Info

- Daily event study of market reaction
  - $y_{f,t} = \beta_1 Event[0] + \beta_2 Event[0] \times Info_{f,j} + \gamma X_{f,t} + \delta_t + \lambda_f + \epsilon_{f,t}$

Panel A. Annual and Quarterly Reports (Forms 10-K/Q)

	AbsoluteReturn			*Volume*		
	(1)	(2)	(3)	(4)	(5)	(6)
Event[0]	1.14***	0.30	0.71***	0.60***	-0.40	0.15
	(0.08)	(0.33)	(0.11)	(0.06)	(0.31)	(0.12)
Event[0] × *Length*		0.01**			0.01***	
		(0.00)			(0.00)	
Event[0] × *Info*			0.86***			0.88***
			(0.18)			(0.28)
Controls	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y
Firm FEs	Y	Y	Y	Y	Y	Y
N-obs	376,921	376,921	376,921	376,921	376,921	376,921
R^2	6.5%	6.7%	6.9%	9.0%	9.0%	9.1%

Panel B. Current Reports (Form 8-K)

	AbsoluteReturn			*Volume*		
	(1)	(2)	(3)	(4)	(5)	(6)
Event[0]	1.39***	1.00***	0.37***	0.99***	0.55***	0.29
	(0.06)	(0.06)	(0.06)	(0.13)	(0.08)	(0.20)
Event[0] ×*Length*		0.09***			0.10***	
		(0.01)			(0.03)	
Event[0] ×*Info*			2.04***			1.41***
			(0.13)			(0.33)
Controls	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y
Firm-Yr-Qtr FEs	Y	Y	Y	Y	Y	Y
N-obs	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697
R^2	0.30%	0.42%	1.2%	0.11%	0.13%	0.15%

Panel C. Management's Discussion and Analysis (Extracted from Forms 10-K/Q)

	AbsoluteReturn			*Volume*		
	(1)	(2)	(3)	(4)	(5)	(6)
Event[0]	0.81***	0.47***	0.13	0.44***	0.23***	-0.06
	(0.09)	(0.13)	(0.14)	(0.07)	(0.09)	(0.10)
Event[0] ×*Length*		0.08***			0.05**	
		(0.03)			(0.02)	
Event[0] ×*Info*			1.37***			1.00***
			(0.28)			(0.25)
Controls	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y
Firm FEs	Y	Y	Y	Y	Y	Y
N-obs	348,071	348,071	348,071	348,071	348,071	348,071
R^2	8.1%	8.4%	9.3%	2.3%	2.4%	2.9%

New ideas

- Combine measurement of multimodal data
  - Can LLMs measure new information in data and tables?
  - Directly input pdf type filings?
- Has information in filings been released through other channels?
  - Insider information, industry information, macro information
  - Conference Call, news, social media