From Words to Returns: Sentiment Analysis of Japanese 10-K Reports using Advanced Large Language Models

Okada et al. (2025), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.3349

Prensenter: 岡田克彦*
中筋萌*, 月岡靖智*, 山崎高広**

* 関西学院大学
**大阪産業大学

略語の整理

- 10K report: 有価証券報告書
- MD&A:Management’s Discussion and Analysis
  経営者による財政状態及び経営成績の検討及び分析
- NLP:Natural Language Processing 自然言語処理

背景: アノマリー研究の系譜

イベント・スタディ	Cross section / Time series
Post-earnings announcement drift (PEAD) • Ball & Brown (1968) • Bernard & Thomas (1989) • Fink (2021) • Jinushi (2023) News-related" drifts (media / macro announcements) • Tetlock (2007) • Engelberg & Parsons (2011) • Savor & Wilson (2013) • Peress (2014)	Profitability factors・Investment factors • Novy-Marx (2013) • Fama & French (2015) • Hou, Xue & Zhang (2015) Accruals anomaly • Sloan (1996) • Dechow, Richardson & Sloan (2011) • Green, Hand & Soliman (2011) Low-volatility / low-beta anomaly • Ang, Hodrick, Xing & Zhang (2006) • Frazzini & Pedersen (2014) Factor zoo • Harvey, Liu & Zhu (2016) • Hou, Xue & Zhang (2020)

背景:　これまでにないアノマリーの発見

Cohen, Malloy, and Nguyen (2020), *JF,* 『怠け者の市場*(Lazy Price)*』

- シグナルの源：有価証券報告書（10k）の言葉*year-over-year*の差異
- 言葉の違い（10kレポートの年度間の言葉の距離ーコサイン距離やJaccard距離ー）
  - *2*四半期後の営業利益
  - *2*四半期後の売上
  - *2*四半期後の純利益

  と関連していることを発見.
- 著者らのNLPによると86%の変化は悪い変化.

市場における価格形成で

- 従来から知られていたこと
  - 市場はすぐには情報を株価に反映させることはできない．*(PEAD*等）
  - 市場は公開情報の中でも，直接的な情報でない限りは，その咀嚼に時間がかかる（*ACCRUAL*アノマリー等）
  - ニュースのセンチメント等，定量化しにくい情報は株価に反映されるまでに時間がかかる．

- *Cohen et.al(2020)*で明らかにされたこと
  - 「怠け者市場」では，株価は長期間にわたって情報が反映されていない状態に放置されている．

本研究の位置づけ

- 本研究では，「怠け者の市場」が日本の10KのMD＆A部分含まれるセンチメント情報を，どう評価するかを明らかにする.

10K_MD&Aの解析方法

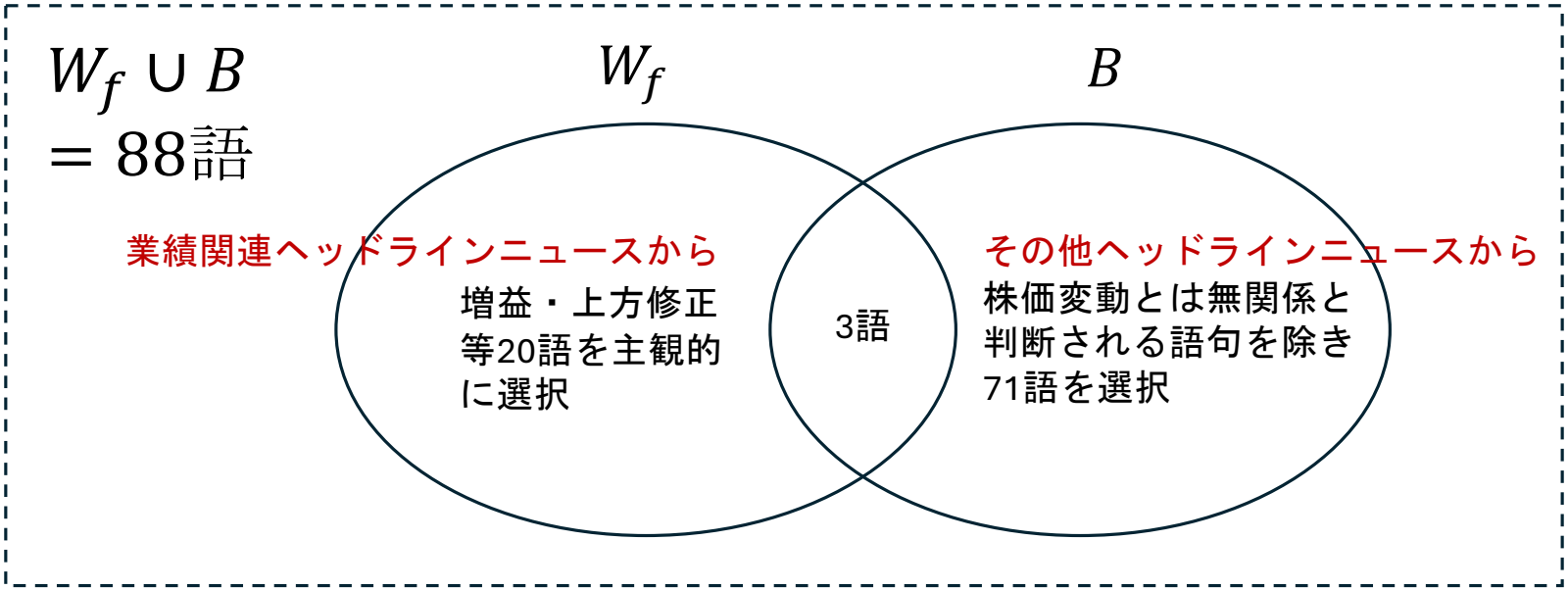
手法の系譜

- 辞書による方法
  - Loughran and McDonald (2011), JF. /　東京大学和泉研究室金融極性辞書
  - Cao et al. (2023), RFS, が問題点を指摘．経営者は公開辞書ディスクロージャーの文言を調整
  - Lin, Nakano and Takahashi (2025), WP, ヘッドラインニュースの文言より金融専門独自辞書作成．88単語，極性をマーケットリターンに基づいて付与.

- LLM アーキテクチャを用いる方法
  - FinBERT Huang et, al (2023)
  - FtBERT Goyenko et. Al (2023)
  - OPS　　Kirtac and Gemano (2024)

- Off the shelf LLMを使う

Lin Nakano and Takahashi (2025)



→ 各語のpolarity scoreを作成．その語の出現後の株価変化率の銘柄平均
14:57分までのニュース：当日の引値ー翌日引値
14:57分以降のニュース:翌日寄付値ー翌日引値

Huang et al.(2023)　FinBERT

Google $BERT_{base}$

- 一般的な語彙の使用（BaseVocab）
  30,522 tokens
- WikipediaやBookCorpusからPre-Trainingされる

Fin BERT

- 金融専門語彙の使用（FinVocab）
  30,873 tokens
- 10K, 10Q，アナリストレポートから
  Pre-Trainingされる
  - 60,490 10-K and 142,622 10-Q filings from Russell 3000 firms (1994-2019).
  - 476,633 reports for S&P 500 firms (2003-2012)
  - 136,578 transcripts from public firms (2004-2019).
  - Total 4.9 billion tokens

最初からpre-trainするために
BERT__baseモデルを使用

FinBERTが辞書よりも優れている点

- 同社の純損失は昨年に比べて大幅に縮小し、回復の兆しを示唆している。
  - 極性辞書：判定：ネガティブ　理由：損失という単語をネガティブと判定し、文全体を否定的に評価する。文脈の「縮小」や「回復」を理解できない。
  - FinBERT:判定：ポジティブ　理由：「損失が縮小」「回復の兆し」という改善の文脈を捉え、ポジティブと正しく分類。
- 株価は当初のボラティリティにもかかわらず上昇し、今期最高値で引けた。
  - 極性辞書：判定：ネガティブ　理由：「ボラティリティ」という単語をネガティブと判断し、全体のポジティブな「上昇」，「最高値」をうまく総合できない。
  - FinBERT:判定：ポジティブ　理由：株価が最高値で引けたことを重要と判断し、ポジティブと分類。

Kirtac and Germano (2023)

Meta OPS

- 一般的な語彙の使用（BaseVocab）
  30,522 tokens
- WikipediaやBookCorpusからPre-Trainingされる.
- BERT_{small}との違いは，ネットワークのサイズだけ.

Goyenko et al.(2023) Ft_BERT

Google $BERT_{Large}$

- 一般的な語彙の使用（BaseVocab）30,522 tokens
- WikipediaやBookCorpusからPre-Trainingされる.
- BERT_{small}との違いは，ネットワークのサイズだけ.

fine-tune

FtBERT

- BERT_{Large}を使用しfine tuning
- 目的変数：次の四半期業績サプライズ(SUE)
- MD&Aのテキストを入力し，次期SUE scoreをlabelとして入力.

類似研究の実験結果

	Model	Market	Text	Key variable	Test period	Method	Alpha/ Excess return
Lin et al. (2025)	Dictionary	Japan	Japanese financial news platform	Daily aggregate setiment score for each stock	2016-24	Long only	Dictionary:27.35% SR 1.06 ChatGPT3.5 14.57% SR 0.62 TOPIX 10.68% SR0.61
Goyenko et al (2023)	FtBERT	US	MD&A and Risk Factor sections of all U.S. 10-K and 10-Q	Expected normalized rank in the cross-section (0-1)	2003-21	Quintile Long / Short Monthly Rebalance	6.70% VW
	FinBERT			The ratio of the number of FinBERT-negative sentences	2003-21	Quintile Long / Short Monthly Rebalance	3.00% VW
Kirtac and Gemano (2024)	FinBERT	US	Refinitiv's global news feed	Probability that the news article belonged to the positive class	2021-23	Quintile Long / Short Daily rebalance	165% SR=2.07
	OPS			Probability score predicting a positive future stock return	2021-23	Quintile Long / Short Daily rebalance	355% SR=3.05
Our experiment	Dictionary	Japan	MD&A of Japanese 10K	2 kinds of sentiment score	2014-23	Quintile Long / Short Yearly Rebalance	
	DeBERTaV2			Probability that the news article belonged to the positive class	2014-23	Quintile Long / Short Yearly Rebalance	
	ChatGPT 4o			Sentiment score (0-1)	2014-23	Quintile Long / Short Yearly Rebalance	
	Claude haiku			Sentiment score (0-1)	2014-23	Quintile Long / Short Yearly Rebalance	
	Gemini 1.5			Sentiment score (0-1)	2014-23	Quintile Long / Short Yearly Rebalance	

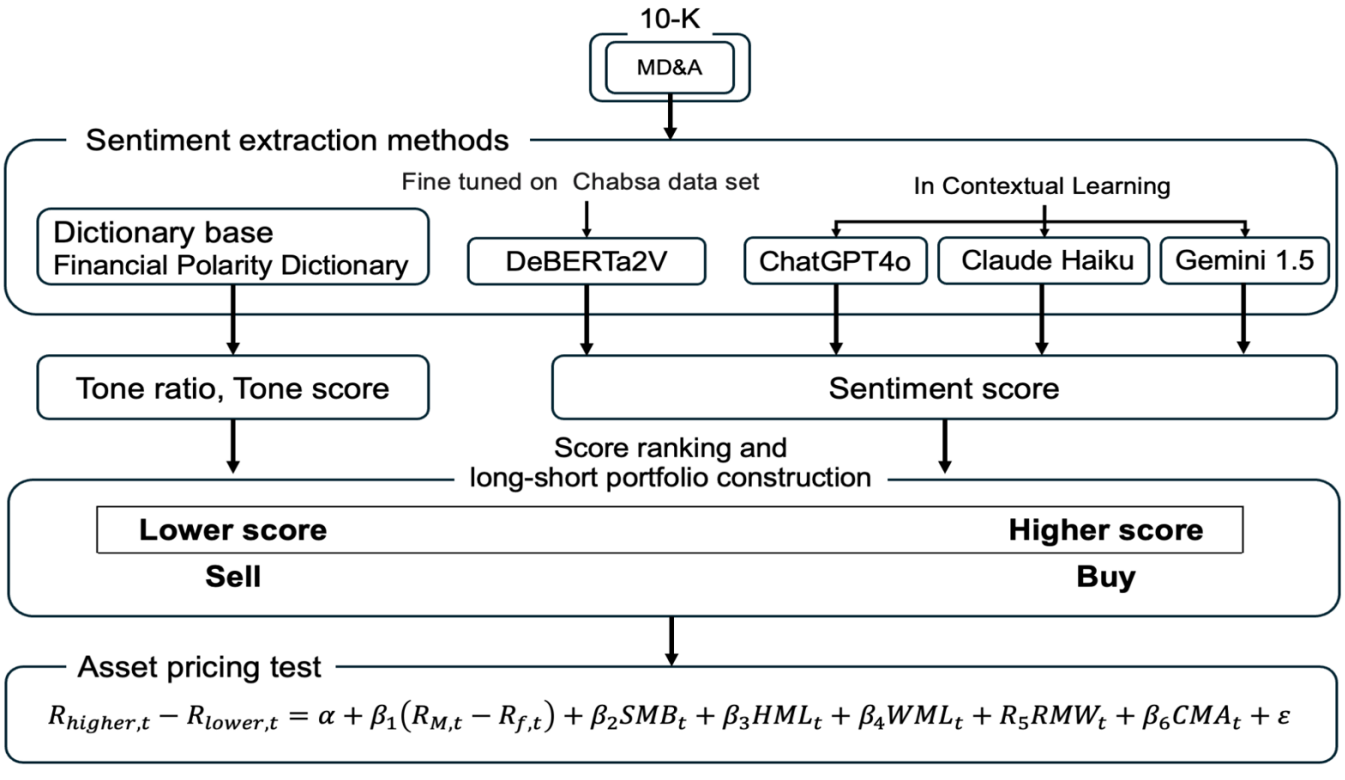
本研究のリサーチデザイン

センチメント・スコアの算出

クロスセクションにランキング

上位20%と下位20%でL/Sポートフォリオを組成

L/Sポートフォリオをファクターモデルで検証



3つのセンチメント分析アプローチ

- 辞書ベース (Tone Ratio, Tone Score): 伝統的“bag-of-words”アプローチ.
  - 金融極性辞書によるポジティブ，ネガティブの判定
- トランスフォーマーモデルのファインチューニング (DeBERTaV2):
  日本の有価証券報告書データでファインチューニングされたモデル
- 汎用LLMs:（GPT-4o, Claude haiku, Gemini 1.5)
  プロンプトによるIncontextual learning

データ: 有価証券報告書

- **Source:** 経営者による財政状態および経営成績の検討と分析(MD&A)
- **Sample:** 11,135 firm-years, from 2014 to 2023.
- 東京証券取引所第1部とプライム．3月決算企業のみ.

	Number of firms	Average length of MD&A (Number of characters)	Min	25%	Median	75%	Max
2014	1,012	6,159	1,910	4,409	5,339	6,654	43,996
2015	1,052	6,136	2,230	4,384	5,365	6,764	43,943
2016	1,069	6,095	1,956	4,414	5,412	6,821	40,281
2017	1,091	5,829	2,094	4,290	5,149	6,491	32,411
2018	1,165	6,294	2,291	4,720	5,751	7,106	29,553
2019	1,190	6,510	2,233	4,896	5,960	7,289	29,542
2020	1,188	7,501	2,966	5,633	6,834	8,536	29,622
2021	1,191	7,345	2,855	5,537	6,639	8,233	29,576
2022	1,176	7,278	2,663	5,506	6,630	8,251	29,614
2023	1,001	7,210	2,797	5,497	6,607	8,117	29,558
Total	11,135	6,636	1,910	4,899	5,052	7,575	43,996

センチメントスコアの分布

	Mean	Min	25%	Median	75%	Max
Tone ratio	−0.144	−0.393	−0.184	−0.142	−0.102	0.083
Tone score	−3.265	−194.445	−11.188	−1.872	6.108	92.435
DeBERTaV2	0.689	0.004	0.182	0.998	0.999	0.999
GPT-4	0.575	0.100	0.400	0.600	0.700	0.900
Claude	0.616	0.000	0.500	0.600	0.700	0.800
Gemini	0.517	0.100	0.400	0.600	0.700	0.900

*Tone ratioは各ドキュメントのポジティブ・ネガティブ比率．理論的最小は-1, 最大は1.
**Tone scoreはポジティブ・ネガティブ単語の総和. 最小最大は単語数に依存.
***DeBERTaV2は，各ドキュメントが「ポジティブ」，「中立」，「ネガティブ」である各確率の加重平均値. 理論的最小は0, 最大は1

センチメント抽出方法間のランク相関

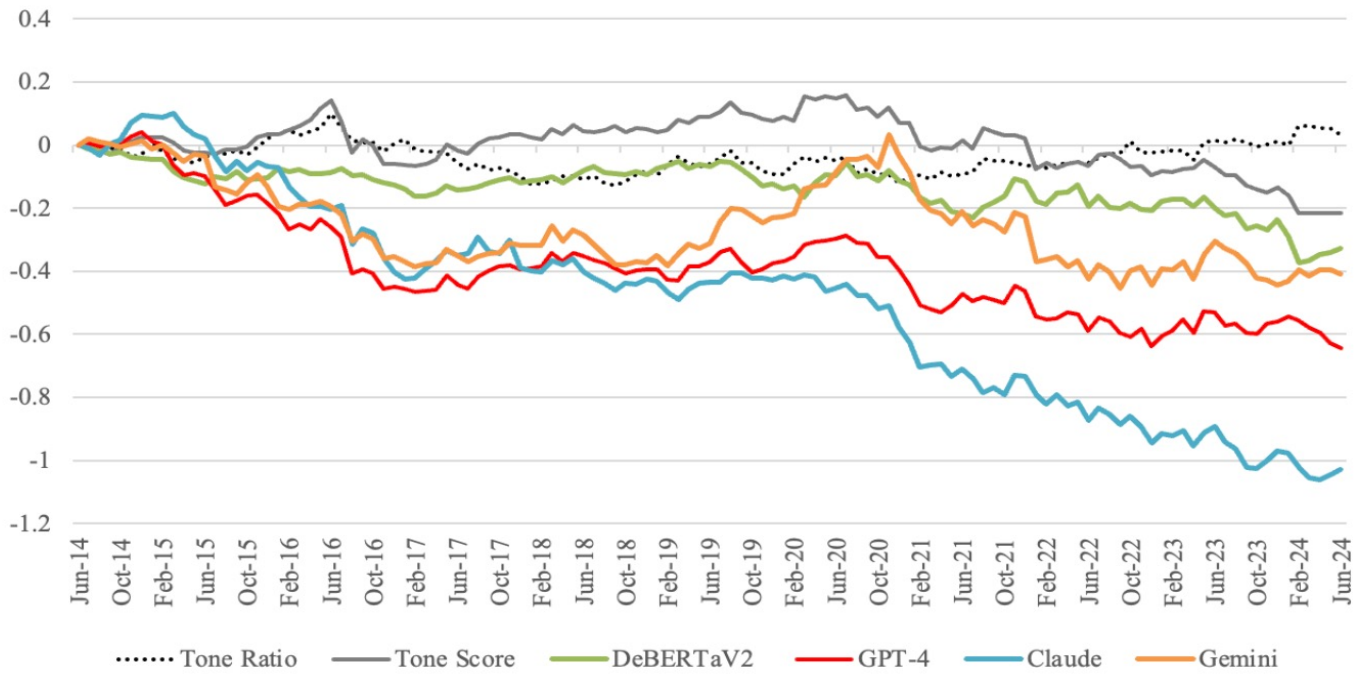
Table : Rank Correlation Matrix of MD&A Sentiment Scores

	TONE Ratio	TONE Score	DeBERTaV2	GPT-4	Gemini	Claude
TONE Ratio	1.0000					
TONE Score	0.4087	1.0000				
DeBERTaV2	0.2128	0.2896	1.0000			
GPT-4	0.2635	0.5298	0.3256	1.0000		
Gemini	0.2731	0.5168	0.3312	0.7544	1.0000	
Claude	0.2814	0.5027	0.3220	0.6638	0.6626	1.0000

結果：累積リターン

GPT-4とClaudeによる分類が，リターンの予測可能性を示した.

Figure 1: Cumulative Returns of Long-Short Portfolios Based on Sentiment Scores



Asset Pricing Test: Risk-Adjusted Alphas
コントラリアンシグナルとして機能

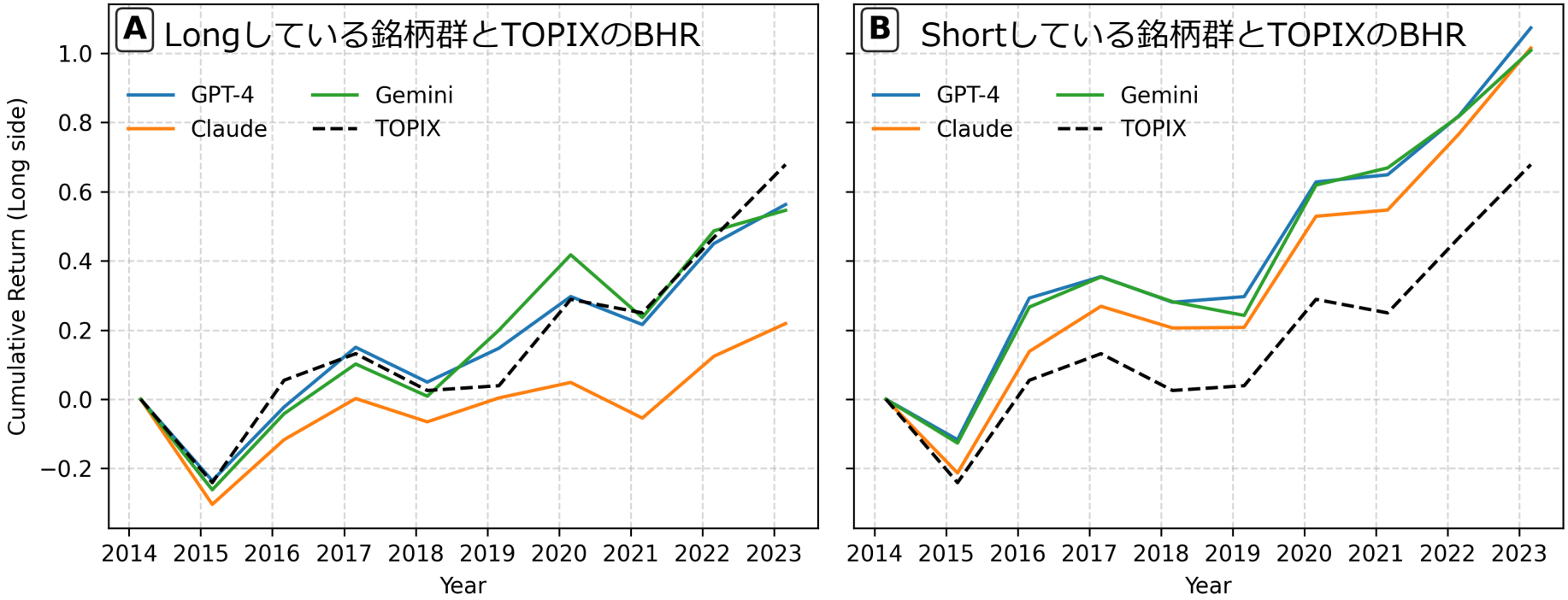
Model	FF3 Alpha (annualized)	FFC4 Alpha (annualized)	FF5 Alpha (annualized)
Tone Ratio	2.40%	1.20%	2.40%
Tone Score	0.00%	-1.20%	-1.20%
DeBERTaV2	0.00%	-1.20%	-1.20%
GPT-4	**-5.95%** (t=-2.01)	**-6.49%** (t=-2.21)	**-6.29%** (t=-2.18)
Claude	**-9.15%** (t=-2.66)	**-9.90%** (t=-2.92)	**-9.46%** (t=-2.76)
Gemini	-4.80%	-6.00%	-4.80%

The Core Puzzle

最も情報効率性が高いユニバースでも頑健

Model	Universe	FF3 Alpha (annualized)	t-Stat	FFC4 Alpha (annualized)	t-Stat	FF5 Alpha (annualized)	t-Stat
GPT-4	Full Sample	**-5.95%**	**(-2.01)**	**-6.49%**	**(-2.21)**	**-6.29%**	**(-2.18)**
	TOPIX 500	-3.96%	(insig.)	-3.96%	(insig.)	-3.96%	(insig.)
	TOPIX 100	**-6.92%**	**(-2.11)**	**-7.45%**	**(-2.26)**	**-7.20%**	**(-2.20)**
Claude	Full Sample	**-9.15%**	**(-2.66)**	**-9.90%**	**(-2.92)**	**-9.46%**	**(-2.76)**
	TOPIX 500	**-9.60%**	**(-2.90)**	**-10.30%**	**(-3.15)**	**-9.60%**	**(-2.78)**
	TOPIX 100	**-9.83%**	**(-2.79)**	**-10.62%**	**(-3.08)**	**-10.32%**	**(-2.99)**

Long legとShort legの比較



Cohen et.al(2020)との比較

論文	Cohen, Malloy & Nguyen (2020) — “Lazy Prices” (U.S.)	Okada, Nakasuji, Tsukioka & Yamasaki (2025) — “From words to returns” (Japan)
市場・期間	U.S. listed firms, 10-K & 10-Q filings, 1995–2014.	東京証券取引所第1部とプライム市場 FY 2014–2023; 11,135 firm-years, >70M words of Japanese 10-Ks.
解析対象テキスト	Full 10-K / 10-Qのすべて	有価証券報告書MD&Aのナラティブ
取引シグナル	類似度のみ / コサイン距離等(cosine, Jaccard, edit distance, etc.) 自然言語処理なし“semantics”なし.	LLMs (GPT-4, Claude, Gemini) と伝統的なBOW辞書& DeBERTaV2.
悪いサインは何か	10Kに変化があること	Very positive LLM sentiment であること.
ポートフォリオ組成方法	Monthly L/S: long Non-changers (Q5), short Changers (Q1), equal- and value-weighted, 3-month holding, rolling.	Annual L/S on sentiment quintiles: long top 20% sentiment, short bottom 20%, value-weighted, held July t → June t+1, for each sentiment measure.
アルファ	L/S (Non-changers – Changers): 18–45 bps/month EW alpha; up to 58 bps/month VW alpha.	L/S (High – Low sentiment) for GPT-4 & Claude: significantly negative alphas. GPT-4 ≈ −5.95%/yr, Claude ≈ −9.15%/yr under FF3; still significant under FFC4 & FF5.
どちらのLegが強い	言葉に最も大きな差異のある銘柄群20%が最悪のパフォーマンス.	センチメントの高い方はINDEXとほぼ差異なし．低い方が大幅にアウトパフォームする.
ファクター係数	Alphas robust to FF3 + momentum + liquidity; not captured by known U.S. factors.	GPT-4 & Claude alphas remain significantly negative after FF3, FFC4, FF5 in Japan. Magnitude exceeds local SMB & HML premia (~1.3–1.4%/yr).
価格修正タイミング	発表日に有意な変化なし．6ヶ月かけて価格に反映	GPT-4 & Claude のセンチメントは1年間をかけて価格に反映
Mechanism story	10Kの変化の86%が負の情報によって構成されている．負の情報は発表時点では価格に未反映.投資家の不注意，怠慢.	LLMsによって悲観的な書き方と判断された銘柄群は，過小評価されている可能性を示唆．楽観的な記述の企業群はミスプライスが少ない.

論文	Cohen, Malloy & Nguyen (2020) — “Lazy Prices” (U.S.)	Okada, Nakasuji, Tsukioka & Yamasaki (2025) — “From words to returns” (Japan)
市場・期間	U.S. listed firms, 10-K & 10-Q filings, 1995–2014.	東京証券取引所第1部とプライム市場 FY 2014–2023; 11,135 firm-years, >70M words of Japanese 10-Ks.
解析対象テキスト	Full 10-K / 10-Qのすべて	有価証券報告書MD&Aのナラティブ
取引シグナル	類似度のみ / コサイン距離等(cosine, Jaccard, edit distance, etc.) 自然言語処理なし“semantics”なし.	LLMs (GPT-4, Claude, Gemini) と伝統的なBOW辞書& DeBERTaV2.
悪いサインは何か	10Kに変化があること	Very positive LLM sentiment であること.
ポートフォリオ組成方法	Monthly L/S: long Non-changers (Q5), short Changers (Q1), equal- and value-weighted, 3-month holding, rolling.	Annual L/S on sentiment quintiles: long top 20% sentiment, short bottom 20%, value-weighted, held July t → June t+1, for each sentiment measure.
アルファ	L/S (Non-changers – Changers): 18–45 bps/month EW alpha; up to 58 bps/month VW alpha.	L/S (High – Low sentiment) for GPT-4 & Claude: significantly negative alphas. GPT-4 ≈ –5.95%/yr, Claude ≈ –9.15%/yr under FF3; still significant under FFC4 & FF5.
どちらのLegが強い	言葉に最も大きな差異のある銘柄群20%が最悪のパフォーマンス.	センチメントの高い方はINDEXとほぼ差異なし．低い方が大幅にアウトパフォームする.
ファクター係数	Alphas robust to FF3 + momentum + liquidity; not captured by known U.S. factors.	GPT-4 & Claude alphas remain significantly negative after FF3, FFC4, FF5 in Japan. Magnitude exceeds local SMB & HML premia (~1.3–1.4%/yr).
価格修正タイミング	発表日に有意な変化なし．6ヶ月かけて価格に反映	GPT-4 & Claude のセンチメントは1年間をかけて価格に反映
Mechanism story	10Kの変化の86%が負の情報によって構成されている．負の情報は発表時点では価格に未反映.投資家の不注意，怠慢.	LLMsによって悲観的な書き方と判断された銘柄群は，過小評価されている可能性を示唆．楽観的な記述の企業群はミスプライスが少ない.

Case study

Fiscal year	Company name	1-year BHR after 10K release	TOPIX BHR (Matched)	GPT-4 Sentiment rank	Claude sentiment rank
Panel A					
Best performers in GPT4 base negative sentiment					
2016	Yamashin filter corp.	4.169	0.296	1,058	896
2023	C&F logistics corp.	3.544	0.221	781	760
2014	Kubotek corp.	2.330	0.323	905	1,098
Worst performers in GPT4 base positive sentiment					
2021	IR Japan holdings	−0.859	−0.017	44	1
2021	SRE holdings corp.	−0.708	−0.038	44	1
2021	RareJob Inc.	−0.698	−0.054	44	38
Panel B					
Best performers in Claude base negative sentiment					
2016	Yamashin filter corp.	4.169	0.296	896	1,058
2023	C&F logistics corp.	3.544	0.221	760	781
2016	m-up holdings Inc.	2.592	0.300	896	728
Worst performers in Claude base positive sentiment					
2021	IR Japan holdings	−0.859	−0.017	1	44
2021	SRE holdings corp.	−0.708	−0.038	1	44
2015	MinebeaMitsumi Inc.	−0.675	−0.246	1	1

汎用LLM分析に関する批判について
Are the LLMs Just Cheating?

- リターンの予測可能性について: 学習データ中には，既にその後の株価パフォーマンスが含まれているのではないか？
  - 汎用LLMの学習データの中に，将来株価のパフォーマンスが含まれていれば，将来リターンを覗いてセンチメントを測定しているのではないか？
  - 負のαと整合しない

- 実証結果は，行動ファイナンス的解釈と整合的
  - 汎用LLMは，MD&Aに含まれる経営者の楽観/悲観度合いを検知.
  - 経営者の悲観は，株式投資家に過剰な負の期待形成を促すことが示唆された.
  - その結果，株価は過小評価され，Lazy Marketで時間をかけて修正される.

結論

1. BoWによるナラティブからのセンチメントの検出能力は限定的.
2. ファインチューニングにしたトランスフォーマーモデル(DeBERTaV2)を用いることで，検出能力は一定程度向上する.
3. 汎用LLMの検出能力がBoWやファインチューニングモデルよりも高かった.
   - とりわけ，GPT_4.0とClaudeの検出能力が，今回の実験では最も高かった.
4. MD&Aに悲観的な記述が多い銘柄は過小評価される可能性を示唆
5. 情報効率性が最も高いと考えられる対象銘柄群でもリターンの予測可能性は減じない.
   - 怠け者の市場においては，ACCRUALSがそうであったように，有価証券報告書のテキスト情報を瞬時に価格に反映するまで，10年以上時間がかかるかもしれない.

参考文献

Ang, Hodrick, Xing & Zhang (2006)
"The Cross-Section of Volatility and Expected Returns"
Journal of Finance 61(1)

Bernard & Thomas (1989)
"Post–Earnings-Announcement Drift: Delayed Price Response or Risk Premium?"
Journal of Accounting Research 27

Ball & Brown (1968)
"An Empirical Evaluation of Accounting Income Numbers"
Journal of Accounting Research.

Dechow, Richardson & Sloan (2011)
"The Persistence, Predictive Power, and Pricing of Accruals and Cash Flows"
Journal of Accounting and Economics

Engelberg & Parsons (2011)
"The Causal Impact of Media in Financial Markets"
Journal of Finance 66(1).

Fink (2021)
"A Review of the Post-Earnings-Announcement Drift"
International Review of Financial Analysis 76.

Frazzini & Pedersen (2014)
"Betting Against Beta"
Journal of Financial Economics 111(1).

Green, Hand & Soliman (2011)
"Going, Going, Gone? The Apparent Demise of the Accruals Anomaly"
Management Science 57(5).

Harvey, Liu & Zhu (2016) *"…and the Cross-Section of Expected Returns" Review of Financial Studies*, Volume 29, Issue 1

Hou, Xue & Zhang (2020)
"Replicating Anomalies"
Review of Financial Studies 33(5)

Hou, Xue & Zhang (2015)
"Digesting Anomalies: An Investment Approach"
Review of Financial Studies 28(3).

Jinushi (2023) *"Post-Earnings Announcement Drift and Ownership Structure in the Modern Japanese Stock Market"*, *The Japanese Accounting Review*.

Peress (2014)
"The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes"
Journal of Finance 69(5).

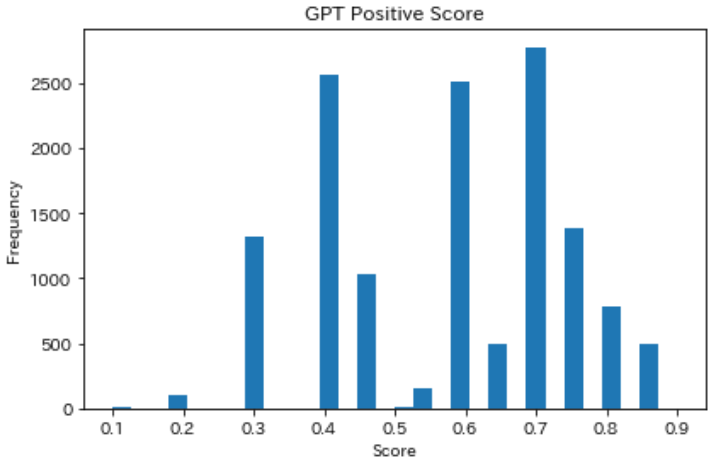
Novy-Marx (2013)
"The Other Side of Value: The Gross Profitability Premium"
Journal of Financial Economics 108(1)

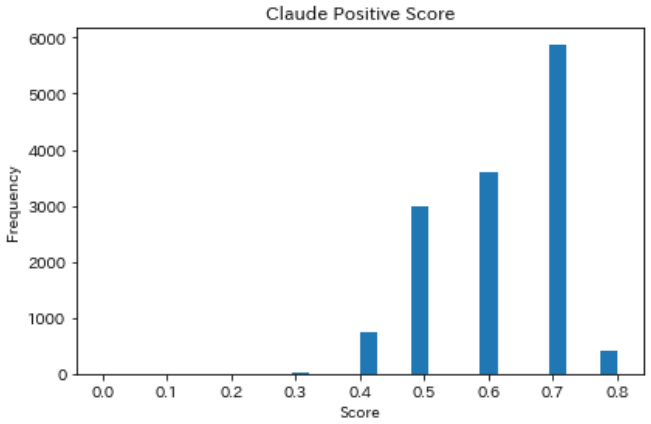
Sloan (1996)
"Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?"
The Accounting Review 71(3).

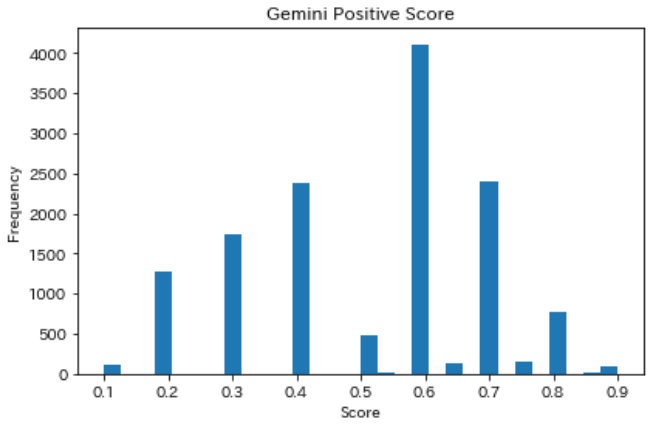
Savor & Wilson (2013)
"How Much Do Investors Care About Macroeconomic Risk? Evidence from Scheduled Economic Announcements"
Journal of Financial and Quantitative Analysis 48(2).

Tetlock (2007)
"Giving Content to Investor Sentiment: The Role of Media in the Stock Market"
Journal of Finance 62(3).

Appendix


GPT Positive Score
Frequency
Score
0
500
1000
1500
2000
2500
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


Claude Positive Score
Frequency
Score
0
1000
2000
3000
4000
5000
6000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Gemini Positive Score
Frequency
Score
0
500
1000
1500
2000
2500
3000
3500
4000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9