景気ウォッチャー調査を用いた金融・経済ドメインのデータセットとタスク

鈴木　雅弘[†]　坂地　泰紀[††]

† 東京大学 大学院工学系研究科　〒113-8656 東京都文京区本郷 7-3-1
†† 北海道大学 大学院情報科学研究院　〒060-0814 北海道札幌市北区北 14 条西 9 丁目
E-mail: †msuzuki@g.ecc.u-tokyo.ac.jp, ††sakaji@ist.hokudai.ac.jp

あらまし　本研究では，景気ウォッチャー調査を用いて，センチメント分析を含む 3 つの金融・経済ドメインの文分類タスクに対応する大規模データセットを構築する．景気ウォッチャー調査とは，内閣府が毎月公開し，日本の経済状況を迅速に把握するための重要なデータソースである．毎月公開される調査結果を自動で統合・公開するためのフレームワークを構築することで，いつでも最新のタスクデータセットを利用できるようになる．
キーワード　データセット，日本語，文分類，金融テキストマイニング

EWS: the Economic Watcher Survey Datasets and Tasks for the Financial and Economic Domain

Masahiro SUZUKI[†] and Hiroki SAKAJI[††]

† School of Engineering, The University of Tokyo　7-3-1 Hongo, Bunkyo, Tokyo, 113–8656 Japan
†† Faculty of Information Science and Technology, Hokkaido University　Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
E-mail: †msuzuki@g.ecc.u-tokyo.ac.jp, ††sakaji@ist.hokudai.ac.jp

Abstract　We construct a large dataset corresponding to three financial and economic domain text classification tasks, including sentiment analysis, using the Economy Watchers Survey. The Economy Watchers Survey is a crucial data source released monthly by the Cabinet Office to swiftly grasp the economic situation in Japan. We ensure that the latest task datasets are always available by building a framework to automatically integrate and release the monthly survey results.
Key words　Dataset, Japanese, Sentence Classification, Financial Text Mining

1. はじめに

ChatGPT や GPT-4 をはじめとする大規模言語モデル (LLMs) が，自然言語処理のタスクで高い性能を発揮している．LLM の活用は一般的なドメインにとどまらず，医療や法律，金融など専門的なドメインにも広がりを見せている [1], [2].

従来のファインチューニングモデルに比べ，LLM は異なる性質を持っている．LLM は zero-/few-shot でタスクを解いたり，指示に従って質問に答えることができるなど，従来と異なるモデルの挙動を示す [3], [4]．また従来モデルと異なり，複数の言語や専門ドメインについて特別に学習しなくとも，複数の言語やドメインを 1 つのモデルによって高い性能でタスクを解くことが可能となっている．金融ドメインにおいても，LLM が少ないサンプル数で BERT などの従来のファインチューニングを必要とするモデルと同等の性能を発揮するなど，高い性能を示している [5].

従来と異なる性質を持つ LLM を評価するために，いくつかのタスクやフレームワークが提案されているが，専門ドメインや英語以外の言語については未だ不十分である．例えば，英語の汎用ドメインにおいては language model evaluation harness [6] や LLM Leaderboard(注1) など，複数のタスクを同時に検証できるフレームワークが存在する．また英語の専門ドメインにおいても法律ドメインのベンチマークである LegalBench [7] や金融ドメインのタスクを集めた FinBen [8] などの整備が進んでいる．英語以外の言語における専門ドメインについても構築は進めら

† 責任著者

(注1)：https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

れているものの [9], [10]，数は多くなく発展途上であるといえる．特に，英語以外の専門ドメインで LLM を評価するためのデータセットは，その専門性から，データセットのサンプル数が少ない事が多い．例えば，金融分野の日本語 LLM のベンチマークでは，5 つのタスクのうち 500 以上のサンプル数を持つタスクは 1 つのみである [9]．英語以外の専門領域では，評価タスクが十分なサンプル数を持たないのが現状である．

より多くの評価タスクが渇望される中で，日本語の金融・経済ドメインにおいては，景気ウォッチャー調査という重要なリソースがある．景気ウォッチャー調査は，内閣府が 2000 年 1 月から毎月公開している，2,050 人に対して行う調査で，景気動向についての評価 (センチメント) やコメント，評価の理由やコメントが関連する分野のラベルなど，様々な情報が含まれている．景気ウォッチャー調査は，調査結果から得られる Diffusion Index (DI) などの指数が景気動向の把握のために役立てられている [11]〜[13] 一方で，景気ウォッチャー調査が含むテキスト情報やセンチメント以外の情報はほとんど用いられず，日本語の金融分野における自然言語処理においての活用例も多くない．この背景としては，景気ウォッチャー調査の収集が容易でないことや，煩雑な形式によって公開されているために，利用しやすいデータセットの形式に変換することが困難である事が挙げられる．

本研究では，日本語の金融・経済ドメインのリソースである景気ウォッチャー調査から，現状と先行きの景気についてのセンチメントやテキストの説明などが含まれる 2 つの大規模なデータセットを構築する．構築したデータセットを用い，3 クラスと 12 クラスのカテゴリ分類タスクと 5 クラスのセンチメント分類による，3 つの大規模なタスクを構築する．統合的に利用できるデータセットのフレームワークについても構築し，評価タスク以外の領域においても活用可能となっている．本研究の貢献は次のとおりである．

- 景気ウォッチャー調査を用い，現在と将来の景気動向についてのコメントなどが含まれる，それぞれ 30 万以上のサンプルを持つ日本語のデータセットを構築した
- 構築したデータセットを用い，それぞれ約 65 万サンプルと約 30 万サンプルからなる 2 つの文分類タスクと，約 65 万サンプルからなる 1 つのセンチメント分析のタスクによる，3 つの大規模な日本語の評価タスクを構築した
- 構築したタスクを用い，日本語のファインチューニングを行う言語モデルや ChatGPT，GPT-4o による評価を行った

タスク用に整形されたデータセットやタスクは Hugging Face Hub[(注2)][(注3)] にて，またデータセットを構築するための実装や抽出したデータは GitHub[(注4)] にて公開している．

2. 関 連 研 究

2.1 金融ドメインにおける言語タスク

金融ドメインにおける言語モデルの評価のために，様々な評価タスクが提案されている．Shah らは，6 つのタスクを用いて金融ドメインの言語モデルを評価した [14]．彼らは，Financial Phrase Bank (FPB) [15]，FiQA [16] に含まれるセンチメント分析と QA，ヘッドライン分類 [17]，NER [18]，FinSBD-3[(注5)] を用いた．FPB と FiQA に含まれるセンチメント分析では，ブログや発表，ニュースのヘッドラインをポジティブ，ネガティブ，中立の 3 つのラベルに分類する，Shah らは，Federal Open Market Committee (FOMC) の声明や議事録から，タカ派とハト派の分類を行うタスクを提案した [19]．さらに構築したデータセットを学習したモデルから，新しい経済インデックスを構築した．

金融ドメインに特化した言語モデルの評価を行うために，Suzuki らは 3 つのタスクを用いた [20]．1 つは chABSA データセット[(注6)] によるセンチメント分析，2 つ目は決算短信のテキストがどのセクションに含まれるかについてのドキュメント分類タスク，3 つ目は経済新聞の記事のテキストに因果関係が含まれるかを判定する因果文判定タスクである．しかしながら，これらの 3 つのタスクのうち公開されているのは 1 つめの chABSA のみである．さらに，鈴木らは，上記の 3 つのタスクに加え，決算短信に含まれるテキストについての因果文判定を行った [21]．彼らが追加した決算短信の因果文判定のデータセットも，非公開である．日本語金融ドメインにおける大規模言語モデルの評価タスクとしては，平野によるものがある [9]．彼は，上記の chABSA に加え公認会計士試験における監査に関するタスク [22]，証券分析における基礎知識タスク，ファイナンシャルプランナー試験の選択肢問題のタスク，証券外務員試験の模擬試験タスクによる，5 つのタスクを大規模言語モデルが対話形式で回答するフレームワークを構築した．

2.2 景気ウォッチャー調査の活用

景気ウォッチャー調査は，含まれるテキストとセンチメントを用いてモデルを学習し，経済動向をナウキャスティングする指標を構築するためのデータセットとして用いられる事が多い [11], [13], [23]．毎月新しいデータが追加され，中央省庁によって発表されることから信頼度も高いため，日本の中央銀行である日本銀行のリサーチにも用いられるなど [12]，非常に注目度の高いデータセットであると言える．その一方で，データの収集や構築が煩雑で，データセットのフィルタリングも研究によって異なることから，統一的なデータセットの構築が再現性の担保と既存データの活用の推進の観点から必要とされている．

3. 景気ウォッチャー調査

景気ウォッチャー調査[(注7)] は，2000 年 1 月より毎月，内閣府

(注2)：https://huggingface.co/datasets/retarfi/economy-watchers-survey
(注3)：https://huggingface.co/datasets/retarfi/economy-watchers-survey-task-finsbd-3
(注4)：https://github.com/retarfi/economy-watchers-survey
(注5)：https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021/shared-evaluation
(注6)：https://github.com/chakki-works/chABSA-dataset
(注7)：https://www5.cao.go.jp/keizai3/watcher-e/index-e.html

表 1　EWS データセットのサンプル例

現状 / 先行き	追加説明及び具体的状況の説明 / 景気の先行きに対する判断理由	地域	業種・職種	関連分野	景気の現状 / 先行き判断	判断の理由
現状	Ｗｅｂ媒体、紙媒体共に、広告売上が前年を上回っている。	近畿	広告代理店（営業担当）	企業動向	◎	受注量や販売量の動き
現状	同業者も当店と同様に今月は暇だと言っている。景気はやや悪い。	東北	一般レストラン（経営者）	家計動向	▲	競争相手の様子
現状	新型コロナウイルスの影響で休業する企業が増加傾向であり、景気は悪い。	東北	学校 [専門学校]	雇用	×	周辺企業の様子
先行き	夏季は団体需要も増えるので期待したい。	北海道	都市型ホテル（スタッフ）	家計動向	○	-
先行き	今後も過度な水準の円安が続くとみられることから、景気は変わらない。	北海道	衣料品専門店（経営者）	家計動向	□	-

が行っている調査である．景気に敏感な 2,050 人を対象に行い，毎回約 1,800 人の有効回答が得られている．景気ウォッチャー調査は，地域の景気に関連の深い動きを観察できる立場にある人々の協力を得て，地域ごとの景気動向を的確かつ迅速に把握し，景気動向判断の基礎資料とすることを目的としている．調査対象は，地域の経済活動に密接に関わる業種・職種の人々で，具体的には，小売業，サービス業，製造業，建設業，不動産業などの現場で働く人々が対象となっている．調査はアンケート形式で行われ，現状の景気感や今後の見通しについての質問がなされる．回答は主観的な評価に基づいて行われ，数値化される．景気ウォッチャー調査の結果は，政府の経済政策の立案や企業の経営戦略の策定，金融機関の投資判断など，さまざまな場面で利用されている．

4. EWS データセットの構築

我々は，2000 年 1 月から 2024 年 5 月までに行われた景気ウォッチャー調査を用いて統合的なデータセット (EWS データセット) をする．景気ウォッチャー調査は大きく現状と先行きの 2 つのファイル形式に分かれる．先行きの調査では，景気の先行きを 5 つのラベル (良い方から順に◎，○，□，▲，×) から判断する．またその理由がテキストによって記載される．さらに，景気動向の評価が家計動向，企業動向，雇用のどの分野に関連しているかのラベルが付与されている．現状の調査では，上記の 3 つに加え，判断するにあたり着目した点 (判断の理由) を，「来客数の動き」や「お客様の様子」などから選択する．また現状と先行きの調査に共通して，それぞれのサンプルがインタビューされた地域と，インタビューを受けた人の業種や職種の項目も存在する．

4.1　データ収集

2024 年 6 月 17 日現在，景気ウォッチャー調査のデータの月ごとのページは期間ごとに，2000 年 1 月から 2009 年 12 月について(注8)，2010 年 1 月から 2018 年 12 月について(注9)，2019 年 1 月以降について(注10)，の 3 つの異なる web ページにて公開されている．それぞれの年・月に対応するページには，現状と先行きの景気判断理由集を，それぞれ watcher4.csv と watcher5.csv を含むリンクから取得することができる．なお，2000 年 1 月から 2009 年 12 月についてのデータの取得においてはファイル名や URL に含まれる年・月と調査を実施した年月が一致するものの，それ以降のデータの取得においては，ファイル名や URL に含まれる年・月は調査が公表された年月 (調査を実施した翌月) となっているため，注意が必要である．本研究では，データに含まれる全ての年月の表記で調査が実施された年月を用いる．これらの実装は，景気ウォッチャー調査を取得できる同様のリポジトリ(注11)を参考に行う．

現状についての景気判断理由集の CSV ファイル) は，分野，景気の現状判断，業種・職種，判断の理由，追加説明及び具体的状況の説明の 5 列を持つ．分野の列は，判断理由が関連する分野 (家計動向関連，企業動向関連，雇用関連) と調査がなされた地域の 2 つの情報を含むため，これらをそれぞれ地域と関連の 2 列に分ける．景気の現状判断は良い方から順に◎，○，□，▲，×の 5 つのラベルのどれかが付与される．CSV ファイルには景気判断のラベルを含まない行も含まれるため，それらの行は削除する．

先行きについての景気判断理由集の CSV ファイルは，分野，景気の先行き判断，業種・職種，景気の先行きに対する判断理由の 4 列を持つ．現状についての景気判断理由集の CSV ファイルの処理と同様に，分野の列を地域と関連の 2 列に分け，景気判断のラベルを含まない行を削除する．

4.2　フィルタリング

現状の調査に含まれる追加説明及び具体的状況の説明と，先行きの調査に含まれる景気の先行きに対する判断理由のテキストについて，次の処理を行う．これらのデータには，「－」と「＊」のみ記載のあるサンプルがあり，これらはそれぞれ「回答が存在しない」ことと「主だった回答等が存在しない」ことを示す．これらのサンプルは自然言語処理の分析には用いづらいことから，これらの文字列のみが記載されたサンプルは削除する．また，それ以外のサンプルは全て itemize を示す「・」とい

(注8)：https://www5.cao.go.jp/keizai3/kako_csv/kako2_watcher.html
(注9)：https://www5.cao.go.jp/keizai3/kako_watcher.html
(注10)：https://www5.cao.go.jp/keizai3/watcher_index.html
(注11)：https://github.com/MitsuruFujiwara/KeikiWatcherScraping

表 2　EWS データセットの構成

Data Type	Train	Dev	Test
現状	279,652	14,747	15,493
先行き	296,143	17,240	18,115

表 3　関連分野の分類タスクのラベル分布

ラベル	Train	Dev	Test
家計動向	387,858	21,741	22,863
企業動向	131,803	6,947	7,319
雇用	64,296	3,299	3,426
合計	583,957	31,987	33,608

う文字から始まることから，先頭の「・」という文字も削除する．表 1 に構築したデータセットのサンプル例を示す．

4.3 分　　割

現状と先行きのそれぞれの調査から構築したデータセットは，全データのうち最新 15,000 件以上がテストセットに，テストと dev セットの合計が 30,000 件以上になり，また現状と先行きの両方のデータセットが同じ期間によって区切られるように分割する．表 2 に構築したデータセットのサンプル数を示す．

4.4 自動アップデート

景気ウォッチャー調査は毎月新しいデータが公開されるため，一度データセットを構築しても，定期的にアップデートがなされなければ，最新のデータで分析を行うことはできない．本研究では上記のデータ取得・フィルタリング・分割に加えデータのアップロードの処理を GitHub Action によって半自動化することで，毎月継続的にアップデートを行う．

具体的には，定期的にクローリングを行い，データのダウンロードがあった場合には (差分が生じた場合には)GitHub 上で自動で Pull Request を立てる．これらの更新内容に誤りが無いかを人手で確認し，マージする．マージがなされると，GitHub のタグ付けとリリース，Hugging Face Hub へのデータセットのアップロードが自動で行われる．データの更新のみを人手で確認するため，更新の労力を削減しつつ，予期していないデータが自動でマージされることも防ぐことができる．

5. タスクの構築

構築した EWS データセットをもとに，関連分野と判断理由についての 2 つの文分類タスクと 1 つのセンチメント分析を含む 3 つのタスクを構築する．

5.1 関連分野の分類タスク

EWS データセットは，現状と先行きの両方の調査に共通して，景気判断の理由のテキストに対し，その理由が家計動向，企業動向，雇用の 3 つの分野のうちどれに関連したものであるかがラベルづけされている．関連分野の分類タスクでは，景気判断の理由のテキストから関連する分野を 3 つのラベルから選ぶ．現状と先行きについてのデータセットから抽出したテキストと関連分野のデータを用いてタスクデータセットを構築する．表 3 に関連分野の分類タスクの統計情報を示す．

表 4　センチメント分析のラベル分布

ラベル	Train	Dev	Test
◎	11,103	1,233	1,112
○	120,556	8,863	8,849
□	260,640	12,759	15,846
▲	137,395	7,094	6,470
×	54,263	2,038	1,331
合計	583,957	31,987	33,608

表 5　判断の理由の分類タスクのラベル分布

ラベル	Train	Dev	Test
来客数の動き	58,730	3,620	3,663
販売量の動き	57,794	2,949	3,210
お客様の様子	41,617	2,293	2,349
受注量や販売量の動き	31,693	1,627	1,790
単価の動き	18,217	783	971
取引先の様子	18,217	818	814
求人数の動き	16,917	746	719
競争相手の様子	6,745	139	193
受注価格や販売価格の動き	5,420	332	305
周辺企業の様子	4,323	254	287
求職者数の動き	3,058	157	192
その他	16,921	1,029	1,000
Overall	279,652	14,747	15,493

5.2 センチメント分析

EWS データセットでは，現状と先行きの両方の調査に共通して，それぞれ現状と先行きの景気の見通しが 5 つのラベル (良い方から順に◎，○，□，▲，×) によって付与されている．センチメント分析は，景気判断の理由のテキストから景気の見通しの 5 つのラベルを選択するタスクである．表 4 にセンチメント分析の統計情報を示す．「どちらとも言えない」を示す□のラベルが多く，◎ (良い) と× (悪い) のラベルが特に少ないため，不均衡データではあることがわかる．

5.3 判断の理由の分類

EWS データセットでは，現状についての調査において，判断するにあたり着目した点 (判断の理由) の列がある．判断の理由についての分類タスクでは，景気判断の理由のテキストから，着目した点を分類するタスクとして構築する．景気ウォッチャー調査の調査票(注12)によると，当該項目は「その他」を含め 15 項目の中から選択する．しかし実際のデータには，表記揺れなどにより 15 項目以外の記述も存在する．また 15 項目の中に含まれていても，データにほとんど含まれないラベルも存在する．本研究では，train セットに 1% 以上含まれるラベルのみを用い，1% 未満のラベルは全て「それ以外」のラベルに統合する．表 5 に理由分類タスクの統計情報を示す．ラベルのフィルタリングの結果，「その他」を含め 12 種類のラベルをタスクでは用いる．

(注12)：https://www5.cao.go.jp/keizai3/watcher/chousahyo.pdf

表 6　評価実験のファインチューニングで用いるハイパーパラメータ

Hyper-parameter	Values
Warmup Ratio	0.1
Learning Rates	{1e-5, 2e-5, 5e-5, 1e-4}
Batch Size	32, 64
Maximum Epochs	5
Maximum Seq Length	282 (FinBERT), 280 (DeBERTaV2)

表 7　評価実験の結果

	関連分野	センチメント	判断の理由
ChatGPT (0)	59.1	28.4	34.0
ChatGPT (5)	61.1	37.5	26.3
GPT-4o (0)	58.1	39.5	41.7
GPT-4o (5)	63.1	38.3	39.9
FinBERT	81.2	52.4	**55.1**
DeBERTaV2	**81.3**	**53.8**	55.0

6. タスクによる評価実験

構築したタスクを用い，いくつかの言語モデルに対して評価を行う．

6.1 実験設定

代表的な LLM として ChatGPT と GPT-4o モデルをテストする．ChatGPT と GPT-4o はそれぞれ，`gpt-3.5-turbo-0125` と `gpt-4o-2024-05-13` を用いる．さらに，日本語の金融ドメイン特化モデルである FinBERT [20] とより新しいアーキテクチャの汎用モデルである DeBERTaV2 [21] によるファインチューニングの結果との比較を行う．FinBERT と DeBERTaV2 のファインチューニングのパラメータは表 6 に示すものを用いる．評価指標はすべてのタスクにおいて macro F1 を用いる．ChatGPT と GPT-4o で用いるプロンプトは以下の通りである．

関連分野

与えられたテキストがどの項目に関連しているかを分類してください。答えは「家計動向」、「企業動向」、「雇用」のいずれかでなければなりません。

テキスト: 同業者も当店と同様に今月は暇だと言っている。景気はやや悪い。
答え:

センチメント分析

与えられたテキストのセンチメントをラベル付けしてください。答えは「◎」、「○」、「□」、「▲」、「×」のいずれかでなければなりません。なお、◎は良い、○はやや良い、□はどちらとも言えない、▲はやや悪い、×は悪い、を表します。

テキスト: 同業者も当店と同様に今月は暇だと言っている。景気はやや悪い。
答え:

判断の理由

与えられたテキストがどの点に着目しているかを分類してください。答えは「来客数の動き」、「販売量の動き」、「お客様の様子」、「受注量や販売量の動き」、「単価の動き」、「取引先の様子」、「求人数の動き」、「競争相手の様子」、「受注価格や販売価格の動き」、「周辺企業の様子」、「求職者数の動き」、「それ以外」のいずれかでなければなりません。

テキスト: 同業者も当店と同様に今月は暇だと言っている。景気はやや悪い。
答え:

6.2 結果と考察

表 7 に実験の結果を示す．すべてのタスクにおいてファインチューニングされた BERT または DeBERTaV2 モデルの性能が最も高かった．ChatGPT と GPT-4o はファインチューニングしたモデルに大きく劣後する結果となった．サンプル数が多いタスクにおいては，ChatGPT や GPT-4o への zero-shot や few-shot の適用よりも，ファインチューニングしたモデルの方が性能が高かった．

[5] におけるセンチメント分析の検証とは異なり，few-shot の効果が見られないケースが存在した．特に判断の理由のタスクは ChatGPT と GPT-4o の両方のモデルで few-shot の性能が低下した．ラベルの数が多いため，少ないサンプルでは適切に例を理解できなかった可能性がある．

日本語の金融ドメインのセンチメントタスクとしては，chABSA データセットについて検証した研究がある [9]．その ChatGPT の性能は accuracy が約 90 であった．それに対し，本研究のセンチメントタスクの ChatGPT の性能は 5-shot の macro F1 で 37.5(accuracy では 47.6) と ChatGPT のスコアが大きく下がっている．このことから，本研究で構築した景気ウォッチャー調査のセンチメントタスクはより難しいタスクであるといえる．

ファインチューニングされたモデルを比較すると，汎用モデルの DeBERTaV2 は金融コーパスで事前学習された FinBERT より性能が高いか同等の性能を示した．[21] でも指摘されているように，モデルの性能の向上やインターネットで公開されている幅広い文書で事前学習されるようになったことで，ドメインに特化したモデルの優位性が薄くなっていると考えられる．

7. まとめ

本研究では，景気ウォッチャー調査から構築された，テキストやセンチメントを含む大規模なオープンソースデータセットを構築した．またデータセットに存在する関連分野，センチメント，景気判断の理由のラベルを用い，3 つの文分類タスクを提案した．我々の構築したデータセットはそれぞれが約 30 万

件のサンプル，タスクのうち 1 つが約 30 万件，2 つが約 64 万件のサンプルを含み，非常に大規模なものである．いつでも最新の調査結果を利用できるようにするために，毎月公開される調査結果を自動で統合できるフレームワークを構築した．さらに，構築したタスクを用い，ファインチューニングされた日本語モデルと ChatGPT，GPT-4 による評価も行った．本研究は日本語の金融ドメインにおける評価タスクのみならず，経済動向を把握するための指数の構築などのために幅広く活用されることが期待される．

謝　　辞

本研究は JST さきがけ JPMJPR2267 の助成を受けたものです．

文　　献

[1] T. Jayakumar, F. Farooqui, and L. Farooqui, “Large Language Models are legal but they are not: Making the case for a powerful LegalLLM,” Proceedings of the Natural Legal Language Processing Workshop 2023, pp.223–229, Association for Computational Linguistics, Singapore, Dec. 2023. https://aclanthology.org/2023.nllp-1.22

[2] K. Singhal, S. Azizi, T. Tu, S.S. Mahdavi, J. Wei, H.W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., “Large language models encode clinical knowledge,” Nature, vol.620, no.7972, pp.172–180, 2023.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol.33, pp.1877–1901, 2020.

[4] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” Advances in neural information processing systems, vol.35, pp.22199–22213, 2022.

[5] X. Li, S. Chan, X. Zhu, Y. Pei, Z. Ma, X. Liu, and S. Shah, “Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks,” Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp.408–422, Association for Computational Linguistics, Singapore, Dec. 2023. https://aclanthology.org/2023.emnlp-industry.39

[6] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac’h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, “A framework for few-shot language model evaluation,” Dec. 2023. https://zenodo.org/records/10256836

[7] N. Guha, J. Nyarko, D.E. Ho, C. Ré, A. Chilton, A. Narayana, A. Chohlas-Wood, A. Peters, B. Waldon, D.N. Rockmore, D. Zambrano, D. Talisman, E. Hoque, F. Surani, F. Fagan, G. Sarfaty, G.M. Dickinson, H. Porat, J. Hegland, J. Wu, J. Nudell, J. Niklaus, J. Nay, J.H. Choi, K. Tobia, M. Hagan, M. Ma, M. Livermore, N. Rasumov-Rahe, N. Holzenberger, N. Kolt, P. Henderson, S. Rehaag, S. Goel, S. Gao, S. Williams, S. Gandhi, T. Zur, V. Iyer, and Z. Li, “LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models,” 2023. https://arxiv.org/abs/2308.11462

[8] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, Y. Xu, H. Kang, Z. Kuang, C. Yuan, K. Yang, Z. Luo, T. Zhang, Z. Liu, G. Xiong, Z. Deng, Y. Jiang, Z. Yao, H. Li, Y. Yu, G. Hu, J. Huang, X.-Y. Liu, A. Lopez-Lira, B. Wang, Y. Lai, H. Wang, M. Peng, S. Ananiadou, and J. Huang, “FinBen: A Holistic Financial Benchmark for Large Language Models,” 2024. https://arxiv.org/abs/2402.12659

[9] M. Hirano, “Construction of a Japanese Financial Benchmark for Large Language Models,” Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing @ LREC-COLING 2024, pp.1–9, ELRA and ICCL, Torino, Italia, May 2024. https://aclanthology.org/2024.finnlp-1.1

[10] Y. Dai, D. Feng, J. Huang, H. Jia, Q. Xie, Y. Zhang, W. Han, W. Tian, and H. Wang, “LAiW: A Chinese Legal Large Language Models Benchmark,” 2024. https://arxiv.org/abs/2310.05620

[11] K. Goshima, H. Ishijima, M. Shintani, and H. Yamamoto, “Forecasting Japanese inflation with a news-based leading indicator of economic activities,” Studies in Nonlinear Dynamics & Econometrics, vol.25, no.4, pp.111–133, 2021.

[12] J. Nakajima, H. Yamagata, T. Okuda, S. Katsuki, and T. Shinohara, “Extracting firms’ short-term inflation expectations from the economy watchers survey using text analysis,” Technical report, Bank of Japan, 2021.

[13] K. Seki, Y. Ikuta, and Y. Matsubayashi, “News-based business sentiment and its properties as an economic index,” Information Processing & Management, vol.59, no.2, p.102795, 2022. https://www.sciencedirect.com/science/article/pii/S0306457321002739

[14] R. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang, “When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain,” Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp.2322–2335, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, Dec. 2022. https://aclanthology.org/2022.emnlp-main.148

[15] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, “Good debt or bad debt: Detecting semantic orientations in economic texts,” Journal of the Association for Information Science and Technology, vol.65, no.4, pp.782–796, 2014.

[16] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur, “WWW’18 Open Challenge: Financial Opinion Mining and Question Answering,” Companion Proceedings of the The Web Conference 2018, p.1941–1942, 2018. https://doi.org/10.1145/3184558.3192301

[17] A. Sinha and T. Khandait, “Impact of news on the commodity market: Dataset and results,” Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2, pp.589–601, 2021. https://doi.org/10.1007/978-3-030-73103-8_41

[18] J.C.S. Alvarado, K. Verspoor, and T. Baldwin, “Domain adaption of named entity recognition to support credit risk assessment,” Proceedings of the Australasian Language Technology Association Workshop 2015, pp.84–90, 2015. https://aclanthology.org/U15-1010

[19] A. Shah, S. Paturi, and S. Chava, “Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis,” Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.6664–6679, Association for Computational Linguistics, Toronto, Canada, July 2023. https://aclanthology.org/2023.acl-long.368

[20] M. Suzuki, H. Sakaji, M. Hirano, and K. Izumi, “Constructing and analyzing domain-specific language model for financial text mining,” Information Processing & Management, vol.60, no.2, p.103194, 2023.

[21] 鈴木雅弘，坂地泰紀，平野正徳，和泉　潔，“FinDeBERTaV2: 単語分割フリーな金融事前学習言語モデル，” 人工知能学会論文誌，vol.39，no.4，pp.FIN23–G_1–14，2024.

[22] 増田　樹，中川　慧，星野崇宏，“ChatGPT は公認会計士試験を突破できるか？: 短答式試験監査論への挑戦，” 第 31 回 人工知能学会 金融情報学研究会（SIG-FIN），pp.81–88，2023.

[23] D. Bragoli, “Now-casting the japanese economy,” International Journal of Forecasting, vol.33, no.2, pp.390–402, 2017. https://www.sciencedirect.com/science/article/pii/S0169207016301297