SASB スタンダードに準拠したコンプライアンスチェックのための ESG 日本語データセット構築

解良智紀 [1]　関洋平 [2]　中尾悠利子 [3]　高村大也 [4]　欅惇志 [1]
[1] 一橋大学大学院　[2] 筑波大学　[3] 関西大学　[4] 産業技術総合研究所
dm250004@g.hit-u.ac.jp

概要

本研究では，企業の ESG (Environment·Social·Governance) 情報が ESG 情報開示フレームワークに準拠しているかを評価するためのデータセットを構築した．本データセットは，日本語のサステナビリティレポートを対象としている．また，人手での作業効率化を目的として，大規模言語モデル（LLM）を用いた事前アノテーションを行った．その結果，データセット構築に要する工数の削減に一定の効果が確認された一方，自動的な情報抽出における性能が低いことが明らかとなった．

1　はじめに

ESG 情報の透明性と比較可能性の向上のために，SASB スタンダード (Sustainability Accounting Standards Board Standards) や GRI スタンダード（Global Reporting Initiative Sustainability Reporting Standards）などの ESG 情報開示フレームワークが提案されている．これらのフレームワークでは，エネルギー消費量や労働災害件数など，開示すべき具体的な項目（指標）が定められている．

企業は当該項目に基づいて，ESG 情報を開示する．また，投資家は，企業のレポート内に指標に関した記載が存在するか否かをを評価する．このような評価は，コンプライアンスチェックと呼ばれ，企業の持続可能性に関する取り組みの妥当性を判断する上で，重要な役割を果たしている．

企業が公表する主要な ESG 情報源の一つに，サステナビリティレポートがある．企業のサステナビリティレポートは，一般的に数十ページから，多い企業では数百ページに及ぶ長文の文書であり，企業ごとに記述内容や表現形式が大きく異なる．さらに，ESG 情報は用語の定義や解釈が必ずしも統一されていない．このような特性により，コンプライアンスチェックには多大な時間的・人的コストが発生している．これらの問題を解決するために，自動コンプライアンスチェック手法が提案された [1, 2].

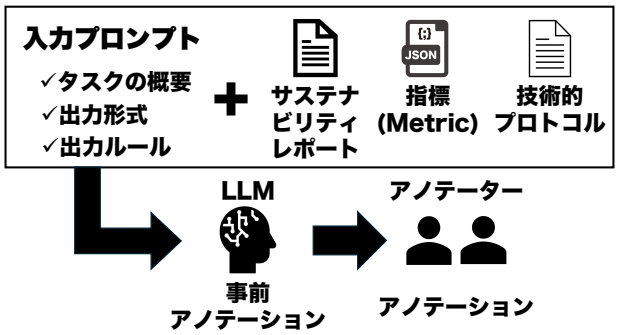

図 1: アノテーションの流れ

自動コンプライアンスチェックにおいては，ESG 情報開示フレームワークに定められた指標に関する記載が，レポート内に含まれているかの判定が求められる．また，レポート内の指標に関する記載を抽出し，レポートの準拠状況を評価する．このような評価を自動で行うためには，指標関連記述が明示されたデータセットが必要となる．しかしながら，自動コンプライアンスチェックのためのデータセットの整備は未だに進んでいない．この課題を踏まえ，NTCIR-19 のタスクの一つである “Multinational，Multilingual，MultiIndustry Regulatory Compliance Checking (RegCom)[1)]” では，言語・業種を横断した，ESG コンプライアンスチェックデータセットを新たに構築している．このデータセットには，英語・フランス語・中国語・タイ語・韓国語・日本語の 6 言語のサステナビリティレポートにおける，指標関連記述が収録されている．

本稿では，著者らが担当した日本語 ESG コンプライアンスチェックデータセットについて報告する．本データセットは，今まで研究で取り扱われてこな

1） https://sites.google.com/view/ntcir19regcom/regcom

かった，日本語のサステナビリティレポートを対象としている．データセット構築において，レポート内に存在する指標に関した記載や，その記載があったページなどをアノテーションした．

その際，アノテーションコストとアノテーション漏れの軽減のために，LLM による事前アノテーションを行った．その結果，アノテーション作業の効率化に一定程度寄与することが確認された．

2 背景

基本事項 サステナビリティレポートとは，ESG に関する取り組みの詳細を記載したレポートのことである．顧客や従業員，地域社会や NGO などの，幅広い読者を想定している [3]．

SASB スタンダードとは，非営利団体であるサステナビリティ会計基準審議会によって制定された，国際的な ESG 情報開示基準である．11 セクター・77 業種に対して，それぞれ異なる企業の財務パフォーマンスに影響を与える可能性が高いサステナビリティ課題を特定し，開示すべき情報を定めている．SASB スタンダートは「産業の説明」，「開示トピック」，「指標」，「技術的プロトコル」，「活動指標」の 5 つの要素から構成される．表 3 は商社業界における SASB スタンダードの例である．

LLM を用いた ESG コンプライアンスチェック LLM を用いた ESG 文書のコンプライアンスチェックに関する既存研究は，主として ESG フレームワークや基準への準拠性を判定するタスクと [1, 4, 5] ，文書中から関連情報を抽出するタスクの二つに大別される [2, 6, 7, 4]．

前者においては，文書中の記述が特定の ESG フレームワークや基準に適合しているかを Yes /No の二値で判定する分類タスクが主に扱われている．この種の研究では，各要件に対する記述の有無や適合度を評価することが目的とされる．後者においては，文書中から ESG フレームワークや基準の各要件に対応する記述を特定し，抽出するタスクが中心となる．抽出対象は，温室効果ガス排出量などの数値情報や，方針・取り組み内容を示すテキスト情報など多岐にわたる．

ESG コンプライアンスチェックデータセット ESG コンプライアンスチェック評価に特化したデータセットは依然として数が限られているものの，近年，その整備を試みる研究が報告されている．

Birti et al. [5] は，企業が ESG 側面への貢献を強化するために実施すべき活動を体系化した「ESG タクソノミー」への準拠状況を 2 値分類するためのデータセットを開発した．このデータセットは，イタリアの主要運輸企業 4 社が公表した非財務情報開示文書のテキストを基に構築されている．また，土橋ら [8] は，有価証券報告書から ESG 情報開示基準の 1 つである GRI スタンダードに関連する文章を抽出し，ESG 関連文データセットを構築した．

本研究では，従来は扱われてこなかった日本語のサステナビリティレポートを対象としている．

表 1: 対象企業

	自動車	ガス	商社
大規模	ホンダ	東京ガス	伊藤忠商事
中規模	日産	東邦ガス	丸紅
小規模	マツダ	北海道ガス	住友商事

3 データセット構築

本節では，国際的な ESG 情報開示基準である SASB スタンダードにおける，日本語コンプライアンスチェックデータセットについて報告する．

対象データ データセット構築にあたり，「自動車」，「商社」，「ガス」の 3 業種から時価総額を参照し[2)]，大規模・中規模・小規模の各規模に該当する企業 3 社を選出した．表 1 に，対象企業を示す．

アノテーション 対象とするサステナビリティレポートについて，図 1 の流れに従って SASB スタンダードの指標関連記述にアノテーションを行った．具体的には，各業種ごとのスタンダードにおける「指標」に関連すると判断される，Value（数値またはテキスト）を抽出した．

抽出の際には，Value が数値またはテキストかを定める「カテゴリー」を参照した．数値を抽出する際には，定義された「測定単位」を用いて再計算した．例を挙げると，Wh から GJ への変換や，m^3 から MBtu への変換が挙げられる．テキストを抽出する際には，2 文以下の場合は原文をそのまま抽出し，3 文以上の場合は LLM を用いて要約した．

指標の関連性評価にあたっては，SASB スタンダードの「技術的プロトコル」に示される (1) 開示トピックの定義，(2) 測定指標の算定方法・単位，(3) 対象とする事象の範囲を参照した．これらの記述と指標の定義・算定方法を照合し，概念的に整合するものを採用した．指標に該当する記述がレポートか

2) 業種の選定基準については付録 A にて述べる．

表 2: データセットのアノテーション件数の内訳

(a) データ形式別件数

	数値	テキスト	N/A
GPT	63	18	86
Gemini	62	25	63
最終結果	134	104	54

(b) 業種別件数

	自動車	ガス	商社
GPT	72	47	48
Gemini	54	48	48
最終結果	137	70	85

(c) 企業規模別件数

	大規模	中規模	小規模
GPT	49	70	48
Gemini	42	59	49
最終結果	93	111	88

ら見つからなかった場合は，N/A と記載した．

アノテーションに先立ち，LLM による事前アノテーションを行った．これにより，アノテーション作業の負荷とアノテーション漏れの軽減を図った．LLM に対して，PDF 形式のサステナビリティレポート，各業界の「指標」をまとめた JSON ファイル，および各指標に対応する「技術的プロトコル」を記載したテキストファイルを入力した．これらの入力データを基に，SASB スタンダードの指標と関連するページの特定，ならびに特定ページにおける Value の抽出・要約を実施した[3)]．

その後，著者 2 人が独立にアノテーションを行った．事前アノテーションにより抽出されたページおよび Value が正しく抽出されているかを確認し，誤りが認められた場合には修正を行った．さらに，レポートの目次を参照して，アノテーションの漏れがないかを確認し，漏れがあった場合には追記した．アノテーションが完了した後に，著者ら（うち 1 人はサステナビリティ経営学の専門家）による検証を 1 件ずつ実施した．アノテーションが難しいケースに対しては，注釈を加えた．

データセットのアノテーション件数の内訳を表 2 に掲載する．また商社の SASB スタンダードとアノーテーションの例を表 3 と表 4 に示す．

4 事前アノテーション結果の分析

本節では，LLM による事前アノテーション結果を分析・考察する．

4.1 実験設定

業種や企業規模による，サステナビリティレポートの性質の違いがモデルの抽出性能に影響するという仮説の下，「データ形式」，「業種」，「企業規模」の三つの観点から性能評価を行った．実験には Gemini 3.0 Pro[4)] および ChatGPT-5 Thinking[5)] を使用して，ページおよび Value を抽出した．

数値抽出では，ページ番号および指標に対応する数値の双方が正しく抽出された場合を正解と定義した．テキスト抽出については，正解文が要約されている場合もあり，正解の判定が必ずしも自明ではない．そのため，著者が人手で確認をし，正誤の判定をした[6)]．抽出対象のページ番号が誤っている場合は，数値抽出・テキスト抽出のいずれにおいても不正解として扱った．また，モデルがページおよび値を N/A と出力した場合には，正解データも N/A であるときに限り正解とした．これらの評価基準に基づき，数値抽出およびテキスト抽出の性能について Precision，Recall，F1-score (F1) を算出した．

表 5 に示すデータ形式別の結果では，両モデルともテキストよりも数値の方が F1 が高かった．これは，数値の方が定まったフォーマットで記載されることが多く，抽出がより容易であるからだと考えられる．また，N/A の判定においては，いずれのモデルにおいても最も高い性能を示した．

表 6 の業種別評価では，GPT-5 Thinking は商社領域で最も高い F1 を示し，Gemini 3 Pro はガス業界で最大値を示した．全体として，どちらのモデルも業種間で性能にばらつきがみられた．これは，業種ごとにサステナビリティに関する記載内容や表現が異なることが要因となっていると考えられる．それらの違いがモデルの抽出難易度に影響を及ぼした可能性を示唆している．モデル間に見られる性能差の要因については，本研究の範囲では明らかにできておらず，今後の課題とする．

表 7 の企業規模別評価では，GPT-5 Thinking は小規模企業，Gemini 3 Pro は大規模において企業比較的良好な結果を示した．業種と同様に，両モデルとも企業規模の違いによる性能にばらつきが見られた．企業規模によってサステナビリティ情報の詳細度，ならびに文書構造が異なり，それらの違いがモ

3) 事前アノテーションで使用したプロンプトは付録 B にまとめた．
4) https://deepmind.google/models/gemini/pro/
5) https://openai.com/ja-JP/index/introducing-gpt-5/
6) 抽出文と正解文の類似度に基づく BERTScore [9] による自動判定を試みたが，要約文の事実の正確性を担保できなかったため，本項では人手による評価を採用した．

表 3: SASB スタンダード（商社）

トピック	指標	カテゴリー	測定単位
小売及び流通における エネルギー管理	(1) エネルギー総消費量 (2) 電力系統からの電気の 割合及び (3) 再生可能エネルギーの割合	定量	ギガジュール（GJ）、 パーセンテージ（%）

表 4: アノテーションの例

ページ	指標	Value	測定単位
56	エネルギー総消費量	16,992,000	GJ

表 5: データ形式ごとの性能指標

モデル	データ形式	Recall	Precision	F1
GPT-5 Thinking	数値	0.38	0.31	0.34
	テキスト	0.01	0.01	0.01
	N/A	**0.68**	**0.51**	**0.58**
Gemini 3 Pro	数値	0.34	0.35	0.34
	テキスト	0.06	0.31	0.10
	N/A	**0.75**	**0.71**	**0.73**

表 6: 業種ごとの性能指標

モデル	業種	Recall	Precision	F1
GPT-5 Thinking	自動車	0.19	0.35	0.25
	ガス	0.09	0.33	0.12
	商社	**0.22**	**0.40**	**0.29**
Gemini 3 Pro	自動車	0.11	0.27	0.16
	ガス	**0.29**	**0.42**	**0.34**
	商社	0.20	0.37	0.26

表 7: 企業規模ごとの性能指標

モデル	企業規模	Recall	Precision	F1
GPT-5 Thinking	大規模	0.18	0.33	0.23
	中規模	0.13	0.18	0.15
	小規模	**0.28**	**0.35**	**0.31**
Gemini 3 Pro	大規模	**0.20**	**0.39**	**0.27**
	中規模	0.14	0.25	0.18
	小規模	**0.20**	0.31	0.24

デルの抽出性能に影響を及ぼした可能性を示唆している．業種別要因と同様に，企業規模別のモデル間性能差の要因については今後の検討が必要である．

4.2 LLM を用いた事前アノテーションの有用性および限界

本研究で使用した ChatGPT-5 Thinking および Gemini 3 Pro には，共通した有用性が確認された．LLM はレポート内における指標関連記述の有無の判定においては，比較的高い精度を示した．特に Gemini 3 Pro は N/A 判定において F1 が 0.73 と高い値を示し，該当情報が文書内に存在しない場合を適切に識別できていた．このことは，LLM が文章内容の有無を把握する能力については，一定の信頼性を有していることを示唆している．

アノテーション箇所の特定および見落としの低減にも有用であると考えられる．事前アノテーションにより，アノテーション範囲が絞り込まれ，注視すべき点が明確になった．さらに，事前アノテーションによりアノテーション対象が明確になるため，後続のアノテーション作業における見落としの低減にも寄与した．なお，仮に事前アノテーションに誤りが含まれていた場合でも，作業対象の絞り込みという点では有用であった．これらの効果により，人手のみで行うアノテーションと比較して，LLM による事前アノテーションは，工数削減において一定の効果があった．

一方，LLM における事前アノテーションには一定の限界も見られた．第一に，ページ内にテキスト抽出の該当項目が複数存在する場合に，そのうちの一つのみを抽出する傾向が見られた．また，抽出された数値自体は正しいものの，ページ番号を誤って出力したケースも確認された．このような事例は，ChatGPT-5 Thinking では 11 件，Gemini 3 Pro では 4 件見受けられた．このことから，数値そのものの認識能力とは別に，出典位置を正確に対応付ける処理における，モデルごとの違いが影響している可能性が考えられる．

5 おわりに

本稿では，サステナビリティレポートを対象とした，日本語コンプライアンスチェックデータセット構築の概要と設計方針を示した．また，LLM による事前アノテーションを試行した．アノテーション範囲の事前絞り込みと候補抽出の自動化により，全文を網羅的に確認する場合と比較し，人手作業の負荷が軽減された．また，指標関連記述の有無判定において LLM が一定の性能を示したことから，アノテーション対象の見落とし抑制にも寄与した．

謝辞

本研究の一部は，JSPS 科研費（基盤研究 (B) (課題番号: 23H03686, 25K03178)，基盤研究 (C) (課題番号: 24K15066)，令和７年度次世代人工知能技術等研究開発拠点形成事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」，株式会社デンソーアイティーラボラトリとの共同研究の支援による．ここに記して謝意を表す．

参考文献

[1] Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, Zongxiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. Esgreveal: An llm-based approach for extracting structured data from esg reports. **Journal of Cleaner Production**, Vol. 489, p. 144572, January 2025.

[2] 濱田祐馬, 石野亜耶, 中尾悠利子. 大規模言語モデルを活用した ESG 評価. 第 32 回金融情報学研究会, Vol. 2023, No. FIN-032, pp. 45–52, 2024.

[3] T. M. Rusu, A. Odagiu, H. Pop, and L. Paulette. Sustainability performance reporting. **Sustainability**, Vol. 16, No. 19, p. 8538, 2024.

[4] Steven Katz, Yu Gu, and Lanxin Jiang. Information extraction from ESG reports using NLP: A ChatGPT comparison. **SSRN**, 2024.

[5] Mattia Birti, Andrea Maurino, and Francesco Osborne. Optimizing large language models for ESG activity detection in financial texts. In **ICAI F '25: Proceedings of the 6th ACM International Conference on AI in Finance**, pp. 856–863, 2025.

[6] 児玉実優, 酒井浩之, 永並健吾, 高野海斗, 中川慧. 統合報告書からの ESG 関連情報の自動抽出. 2022 年度人工知能学会全国大会（第 36 回）, 京都国際会館 + オンライン, June 2022.

[7] Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. Glitter or gold? deriving structured insights from sustainability reports via large language models. **EPJ Data Science**, Vol. 13, No. 41, 2024.

[8] 土橋諒太, 中田和秀. BERT を用いた有価証券報告書からの ESG 関連文抽出. 人工知能学会第二種研究会資料, Vol. 2021, No. FIN-026, pp. 09–, 2021.

[9] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. 2020.

[10] Chung-Chi Chen, Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Juyeon Kang, Hanwool Lee, Min-Yuh Day, and Hiroya Takamura. SemEval-2025 task 6: Multinational, multilingual, multi-industry promise verification. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors, **Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)**, pp. 2461–2471, Vienna, Austria, July 2025. Association for Computational Linguistics.

[11] Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Hanwool Lee, Juyeon Kang, Min-Yuh Day, and Chung-Chi Chen. ML-promise: A multilingual dataset for corporate promise verification. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 20364–20377, Suzhou, China, November 2025. Association for Computational Linguistics.

A　業種の選定基準

業種の選定は，先行研究である SemEval-2025 Task 6: PromiseEval [10] において構築された ML-Promise [11] 中の業種に基づき選定しているが，対象企業については，SASB スタンダードの一貫性を考慮して一部を入れ替えている．PromiseEval は，企業の ESG に関する将来志向の約束を自動的に検証することを目的として提案された，SemEval-2025 の共有タスクである．本タスクでは，約束の識別，行動証拠の有無，約束と証拠の対応関係の明確性，および検証時期の推定という 4 つの下位タスクが定義されている．ML-Promise は，この PromiseEval の公式データセットとして構築されたものである．

B　事前アノテーションプロンプト

図 2 に LLM による事前アノテーションにて用いたプロンプトの例を挙げる．

＜タスクの概要＞
あなたは ESG の専門家です。以下のタスクを解いてください。第一に、Json ファイルに記載されている指標 (Metric) とそれに紐づいたカテゴリーを抽出してください。その抽出した各指標と関連がありそうなページを特定してください。ページを特定する際は、PDF に記載されているページを元に判断するようにしてください。 その後、特定したページの中から指標に関する数値もしくは文章を抜き出してください。 各指標に関する説明が載ったテキストファイルをアップロードしているので、参考にしてください。 以下のルールに出力して従ってください。
＜出力フォーマット＞
CSV(3 列：「Metric」と「Page」と「Value」列) の Shift-Jis 形式で出力する。
＜出力ルール＞
1. 共通ルール
1-1. 出力は CSV 形式（Shift-JIS）とする
1-2. 列は「Metric」「Page」「Value」の 3 列とする
1-3. 1 行につき、Page と Value は 1 つのみ含める
1-4. 同一指標に複数の Page または Value がある場合は、行を分けて出力する
2. 条件別ルール（指標カテゴリーによる分岐）
2-1. 指標のカテゴリーが「定量」の場合
・指標に対応する数値が記載されているページを特定する
・PDF に記載されているページ番号を「Page」に出力する
・数値のみを「Value」に出力し、文字列は含めない
・単位が SASB スタンダードと異なる場合は、SASB に準拠する単位に換算する
2-2. 指標のカテゴリーが「説明および分析」の場合
・指標に関する説明または分析が記載されているページを特定する
・該当する PDF ページ番号を「Page」に出力する
・該当箇所の文章を「Value」に出力する
・文章が 3 文以上の場合は要約する
2-3. 指標に該当するページが見つからなかった場合
・指標に該当するページが見つからなかった場合に適用する
・「Page」および「Value」に「N/A」を出力する

図 2: プロンプトの例