

Article

ESG-SASB Label Stability: A Curated Benchmark and Reproducible Pipeline for Reusing Sentence-Level Sustainability Disclosure Labels

Yufei Li [†] , Tianhao Chen [†] , Wei Ke  and Patrick Pang ^{*} 

Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China; yufei.li@mpu.edu.mo (Y.L.); tianhao.chen@mpu.edu.mo (T.C.); wke@mpu.edu.mo (W.K.)

^{*} Correspondence: mail@patrickpang.net

[†] These authors contributed equally to this work.

Abstract

Annotated text datasets are increasingly reused as classifier targets, annotation candidates, and inputs to aggregate profiles, yet their labels often circulate without enough information about how they were produced. This article presents a reproducible benchmark and validation workflow for the public *SASB-Aligned ESG Sentences* corpus, a sentence-level sustainability disclosure dataset organized around standards-based categories such as those used in Sustainability Accounting Standards Board (SASB) analytics. Using the downloaded 6460-row version of the corpus, we construct fixed train/validation/test splits, map released child labels to parent categories, and evaluate label reuse through supervised classifiers, prompted GPT-4o classification, blind and candidate-visible Claude annotation, and Monte Carlo aggregation into ESG/Non-ESG category profiles. The reproducibility artifacts provide split metadata, label mappings, prompt templates, model predictions, LLM annotation outputs, profile sensitivity outputs, figure inputs, and scripts for reproducing the reported tables and figures. Results show that label reproduction is strongest at coarser label levels, blind annotation flags 40.3% of held-out sentences as ambiguous, candidate-visible annotation increases agreement while changing the task format, and aggregate profiles remain sensitive to label source. The benchmark supports transparent reuse of sentence-level ESG labels by reporting label source, annotation condition, prompt family, and aggregation level.

Keywords: annotated text datasets; label provenance; ESG disclosure analytics; SASB labels; reproducible benchmark; large language models; data reuse



Academic Editor: Olga Kurasova

Received: 14 May 2026

Revised: 29 June 2026

Accepted: 2 July 2026

Published: 3 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Public annotated text datasets are increasingly reused beyond their original release context. A released label can become a supervised learning target, a candidate shown to an annotator, or a count inside an aggregate dashboard. For such reuse, downstream users need evidence about both predictive reproducibility and the pathway by which each label was produced.

Sentence-level sustainability disclosure labels provide a useful setting for studying this reuse problem. Corporate sustainability reports are written as long narrative documents, while automated and AI-assisted disclosure analytics often reduce those reports to labeled sentences. Sustainability Accounting Standards Board (SASB)-aligned analytics are especially informative because SASB standards organize disclosure through industry-specific

issue categories and metrics [1,2]. A sentence-level SASB-aligned dataset therefore turns a context-rich standards interpretation into a short text fragment plus a reusable label.

The reuse challenge arises because standards-based interpretation is normally performed with report context. A report reader can use the industry classification, disclosure topic, section heading, neighboring sentences, and position of a statement within the wider narrative. A sentence-level dataset usually gives later users a text fragment and a released label. Once that label becomes a classifier target, an annotation candidate, or an input to an aggregate profile, it becomes a reusable data assumption that needs visible provenance.

The public *SASB-Aligned ESG Sentences* corpus provides the source corpus for this article because it pairs corporate sustainability report sentences with released SASB-aligned labels [3]. We use *category* to denote a node in the SASB-aligned taxonomy and *label* to denote the category assigned to a sentence. A *released label* is the dataset-supplied reference label attached to a sentence by the public corpus creators; it is used as a reference label for reproduction analysis, not as an official SASB judgment. A *label source* is the procedure that produces sentence labels, including released labels, model predictions, and large language model (LLM) annotation labels. An *annotation condition* is the information condition under which an annotator assigns a label. An *ESG/Non-ESG category profile* reports the distribution of sentence labels across SASB parent categories and the separate Non-ESG corpus category.

The label hierarchy also makes exact-match disagreement an incomplete description of reuse behavior. Fine-grained child labels map to broader SASB parent categories, while Non-ESG is retained as a separate corpus category. A child label change from Customer Privacy to Data Security remains within the Social Capital parent branch and is semantically adjacent in a way that differs from a move into Non-ESG or from Social Capital to Leadership & Governance. Table 1 summarizes the label scheme used in this study. The full 6460-sentence corpus defines the label universe and split frame; final comparisons use the stratified 969-sentence held-out test set so that classifier, annotation, and aggregation outputs are compared on the same sentences across binary, parent category, and child category views.

Table 1. ESG/Non-ESG category scheme used in this study. The five ESG categories follow the SASB parent category structure used by the corpus; Non-ESG is a separate corpus category included in profile analyses where relevant and excluded from parent-level classifier evaluation. Child labels are illustrative examples, and the corpus contains additional child labels.

Category	Role in the Label Scheme	Illustrative Child Labels
Environment	Environmental impacts and resource use	GHG Emissions; Energy Management; Water & Wastewater Management
Social Capital	Relationships with customers, communities, and society	Customer Privacy; Data Security; Product Quality & Safety
Human Capital	Workforce-related issues	Labor Practices; Employee Health & Safety; Employee Engagement, Diversity & Inclusion
Business Model & Innovation	Sustainability issues embedded in products, supply chains, and business models	Product Design & Lifecycle Management; Supply Chain Management; Business Model Resilience
Leadership & Governance	Governance, ethics, risk, and regulatory management	Business Ethics; Competitive Behavior; Critical Incident Risk Management
Non-ESG	Separate corpus category used alongside the SASB ESG categories in profile analyses	Sentences outside the SASB ESG child categories

This study asks how released sentence-level SASB-aligned labels behave when reused as benchmark targets, annotation inputs, and aggregate profile data. Three descriptive research questions organize the analysis:

- RQ1. How closely do automated label sources reproduce released SASB-aligned labels at binary, parent category, and child category levels?

- RQ2. How do retention, agreement, and self-reported ambiguity differ when a fixed LLM annotator labels the same held-out sentences under blind and candidate-visible annotation conditions?
- RQ3. How sensitive are aggregate ESG/Non-ESG category profiles to the sentence-level label source used for aggregation?

The article contributes a reusable benchmark and validation workflow for AI-assisted sustainability disclosure analytics. It shows how released sentence-level labels behave as classifier targets, objects of LLM annotation, and inputs to aggregate category profiles. It also provides reproducibility artifacts containing split metadata, label mappings, prompt templates, model outputs, annotation outputs, profile sensitivity outputs, figure inputs, and reproduction scripts. The practical output is a reporting norm for label provenance: downstream uses should document the pathway by which a label or profile output is produced, including the label source, annotation condition, prompt family, and aggregation level.

Section 2 situates the paper in standards-based sentence labeling, automated disclosure analytics, annotation reliability, and ESG profile aggregation. Section 3 describes the source corpus, reproducibility artifacts, and validation workflow. Section 4 reports results for the three research questions, while Sections 5 and 6 interpret their implications for data reuse.

2. Literature Review

2.1. Sustainability Reporting and SASB-Aligned Sentence Labels

Corporate sustainability reports are context-rich documents, while sentence-level datasets reduce them to short text units with category labels. This reduction is useful for scalable analysis, but it changes the information available for standards-based interpretation. Research on sustainability reporting shows that materiality and comparability remain interpretive even under shared standards [4–6]. For a reusable sentence label, the relevant question is whether the label remains interpretable when industry, topic, and report context are no longer visible.

SASB-aligned analytics sharpen this issue because SASB standards organize disclosure through industry-specific topics and metrics [1]. A sentence-level SASB-aligned label compresses a standards-based interpretation into one short text unit. Prior work on SASB standards and ESG information environments motivates treating that compression as a data reuse problem that extends beyond classifier performance [2].

2.2. Automated ESG Disclosure Analytics and Standards-Based Text Classification

Automated disclosure analytics promise to turn narrative sustainability reports into variables, labels, and profiles that can be compared at scale. Reviews of accounting and sustainability reporting text analysis describe automated methods as a way to convert corporate communication into measurable constructs [7,8]. In ESG-specific work, contextual sentence models have been used to recognize ESG concepts [9], NLP models have quantified corporate ESG communication [10], and LLM or retrieval pipelines have extracted structured ESG information from reports [11]. Other work extends the evidence base to news, benchmarks LLMs on sustainability knowledge, and reviews how LLMs can support environmental sustainability [12–14]. Related health domain work, including SDG-oriented resource allocation analysis, uses LLM-based aspect analysis and hierarchical aspect models to convert user-generated text into structured service and resource signals [15,16]. Across these systems, text functions as an intermediate input that is converted into labels or scores for later comparison.

Standards-based sentence labels make that conversion especially consequential. A released corpus that pairs one sentence with one SASB-aligned label invites later systems to

reproduce that label through supervised learning, prompting, or annotation [3,17]. The label therefore does more than describe the corpus; it defines the task that later systems inherit. Exact child label agreement is useful for reproduction analysis, but an exact mismatch can collapse different semantic situations: a sibling category substitution within the same parent branch, a cross-parent reassignment, and a move between ESG and Non-ESG are all counted as child-level errors. Text-as-data validation research argues that computational text outputs should be evaluated against the use case in which they are applied [18,19]. This paper applies that logic to sentence-level SASB-aligned labels by following the same released labels through classifier reproduction, LLM annotation, and profile construction.

Sentence representation research provides the broader NLP framing for this distinction. Sentence-BERT models sentence meaning through dense embeddings [20]; SimCSE shows how contrastive learning can improve sentence embeddings for semantic similarity tasks [21]; and SetCSE extends contrastive sentence representation toward set-like semantic operations and category relationships [22]. Work on text embedding validity reaches a compatible conclusion from measurement theory: text representations require validity checks when they are used as analytic inputs, not just when they are used as predictors [23]. For a reusable standards-based label, this literature motivates looking beyond exact label matching to the semantic proximity and hierarchy level at which disagreement occurs.

2.3. Annotation Reliability, Ambiguity, and Candidate-Visible LLM Annotation

Annotation research treats labels as products of a coding procedure, which means the procedure itself becomes part of what a label represents. Reliability measures were developed because coders can vary in how they apply a scheme [24], and observer error models treat annotator performance as an estimable component of a labeling process [25]. Work on crowd annotation further shows that disagreement can reveal more than errors; in semantic tasks it can expose ambiguity in the item, the category boundary, or the intended use of the label [26]. Soft label approaches make the same point from another direction by preserving annotator disagreement as information about the labeling task [27]. The label noise literature adds the downstream risk: once contested labels become training targets, the noise can shape model performance and evaluation [28].

LLMs make large-scale reannotation feasible, yet their outputs still inherit the information structure of the annotation task. In this study, Claude Sonnet 4.6 is used as a fixed LLM annotator across the held-out sentences, while GPT-4o is evaluated separately as a prompted classifier and candidate generator. Survey work on LLM-based annotation supports using LLMs as scalable annotation aids while treating assessment of their outputs as part of the workflow [29]. Human–AI qualitative analysis research has reached a compatible design principle: AI support should be located explicitly within the analytic process, because different stages invite different forms of assistance and accountability [30,31]. The LLM-as-a-judge literature adds a specific caution for candidate-visible annotation, since judgments can shift when prior outputs or visible candidates are presented to the evaluator [32]. For ESG disclosure analytics, blind annotation and candidate-visible annotation represent distinct information conditions.

2.4. From Sentence Labels to Aggregate Profiles

ESG disagreement research has shown that divergence can remain substantial after sustainability information has been selected, coded, and aggregated [33,34]. This literature mostly studies provider-level ratings or scores. A prior data question remains at the sentence-label stage: whether category assignments entering those systems remain stable under alternative annotation and modeling conditions.

Sentence-level disagreement matters for data reuse when it survives aggregation into the profiles that analysts inspect. ESG applications rarely stop at sentence classification; the final user often sees a dashboard, portfolio profile, sector comparison, or disclosure coverage summary. Each artifact aggregates many local labeling decisions into a small number of category proportions or counts. Work on financial narrative measurement shows that design choices in text measures can materially affect empirical signals [35]. We address the corresponding sentence-label problem by following the same released labels through classifier reproduction, LLM annotation, and aggregate profile construction.

3. Materials and Methods

The empirical design treats label reuse as a data processing and validation problem. First, a reproduction analysis assesses whether automated tools can learn or approximate the released label scheme. Second, two annotation conditions test whether LLM labeling changes when sentences are judged blindly or with visible candidate labels. Third, a profile sensitivity analysis examines whether sentence-level differences remain visible after aggregation. Applying all three modules to the same held-out test set allows local label disagreement and aggregate profile shifts to be compared on a common sentence pool.

The data flow separates classifier evaluation from candidate-visible annotation. In classifier reproduction, training rows fit supervised models, validation rows select the GPT-4o prompt family, and test rows provide final predictions and reference labels for metrics after those choices are fixed. Candidate-visible annotation then simulates a label comparison workflow in which Claude sees both the released label and the GPT-4o candidate for the same test sentence. Table 2 summarizes the practical question, comparison, and output for each module. Released labels serve as the comparison baseline for label reproduction and profile sensitivity.

Table 2. Overview of the empirical modules. Final comparisons use the same held-out test set; the classifier module also uses training and validation splits for fitting and prompt selection. κ denotes Cohen’s kappa, a chance-corrected agreement measure.

Module	Practical Question	What We Compare	Main Output
Classifier reproduction	Can automated tools reproduce the released label scheme?	TF-IDF, SBERT, and GPT-4o label sources	Macro F1, top-3 accuracy
LLM annotation	Which label does a fixed LLM annotator assign under blind and candidate-visible conditions?	Blind annotation and candidate-visible LLM annotation	Retention, κ , ambiguity
Profile sensitivity	Do label source choices change ESG/Non-ESG profiles?	Released labels and alternative label sources	Category share shifts

3.1. Reproducibility Artifacts and Reusable Records

The reproducibility artifacts make the reuse workflow inspectable without presenting the third-party source corpus as a newly authored dataset. The released files record the source corpus manifest, fixed split metadata, label hierarchy, prompt templates, model and annotation outputs, metric summaries, profile sensitivity outputs, figure inputs, and scripts used to regenerate the reported tables and figures. They are archived at Zenodo: <https://doi.org/10.5281/zenodo.20105936>. Table 3 summarizes the reusable records.

Table 3. Reproducibility artifacts for reusable analysis. The released files document the treatment, validation, and analysis of the source corpus while avoiding unnecessary redistribution of raw sentence text.

Record Group	Release Files	Purpose	Redistribution Note
Source manifest	source_corpus_manifest.json	Records source URL, access date, loaded row count, landing page count, and SHA-256 hash	Raw text is omitted; users reconstruct from Kaggle
Label hierarchy	label_hierarchy/	Provides the child-to-parent mapping used in all projections	Derived metadata
Split metadata	splits/split_indices.csv	Reconstructs train, validation, and test membership with sentence hashes	Text column removed; hashes support alignment
Prompt templates	prompts/	Documents GPT-4o classification and Claude annotation calls	Author-generated templates
Predictions and annotations	predictions/, annotations/	Supports reproduction of classifier, prompt, and annotation metrics	Public outputs retain labels and diagnostics, not raw sentence text
Metrics and profile outputs	metrics/, profile_sensitivity/	Supports reported results and profile sensitivity summaries	Derived analysis outputs
Figure inputs, code, and tests	manuscript_inputs/, code/, tests/	Recreates manuscript tables, figures, and validation checks	Reproducibility materials

The source corpus was downloaded from Kaggle on 4 May 2026. The Kaggle landing page describes 6454 labeled examples, while the locally loaded `companies.csv` file contains 6460 rows. The source manifest records this discrepancy and the local file hash. The Kaggle source dataset was listed under Apache-2.0 at the time of access. To avoid unnecessary redistribution of third-party report text, the artifacts omit the raw Text column from public split and prediction files. Users can download the source corpus, verify the hash, and align records through labels, split metadata, and sentence-level SHA-256 hashes.

3.2. Corpus, Label Hierarchy, and Split

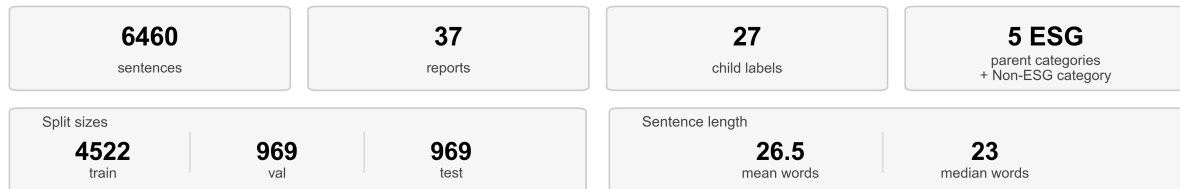
We study a sentence-level ESG resource with a different scope from a firm-level disclosure panel. The *SASB-Aligned ESG Sentences* dataset is distributed through Kaggle [3]. The downloaded file used in this analysis contains 6460 sentences extracted from 37 U.S.-based corporate sustainability reports together with one released child label. The Kaggle landing page describes 6454 labeled examples; in the downloaded version used here, `companies.csv` loads as 6460 rows, and all reported analyses use that analysis table. We derive parent categories by mapping each released child label to its corresponding SASB parent category according to the corpus hierarchy. Prior sentence-level ESG classification work provides the immediate empirical background for our classifier reproduction comparisons [9], while the public SASB-ESG classifier illustrates how this particular SASB-aligned dataset can be reused as a classification resource [17]. The corpus supports classifier training, annotation, and aggregate ESG category profiles at the sentence level.

The corpus has a long-tailed sentence classification structure. Figure 1 summarizes the main empirical constraints: Non-ESG sentences account for more than half of the corpus, the five ESG parent categories are smaller and unevenly distributed, and seven child labels have fewer than 50 examples. Customer Privacy is the rarest child label, with six examples. These counts matter because minority labels are statistically difficult for classifiers to learn [36] and potential disagreement is concentrated in the categories where the training signal is weakest. Keeping these rare labels preserves the released label scheme and makes the child-level task reflect the full corpus taxonomy.

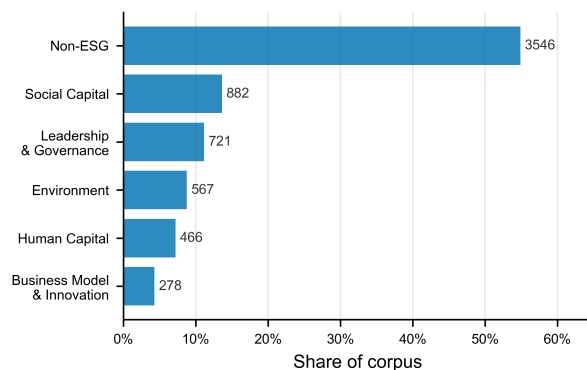
The sentence texts create a second source of difficulty because each observation is short and isolated from the surrounding report narrative. The downloaded corpus has a mean sentence length of 26.5 words and a median of 23 words, with substantial right-tail variation across parent categories. Table 4 gives paraphrased orientation examples. These examples illustrate why the same sentence can be useful for automated disclosure analytics

while still depending on how the label was assigned: some sentences name a sustainability topic directly, some are accounting or report fragments, and some depend on surrounding context to distinguish adjacent SASB issues.

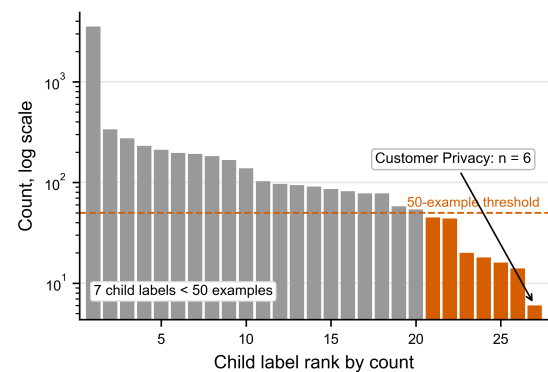
A. Corpus overview



B. Released ESG/Non-ESG category distribution



C. Child label long tail



D. Sentence length distributions by released parent category label

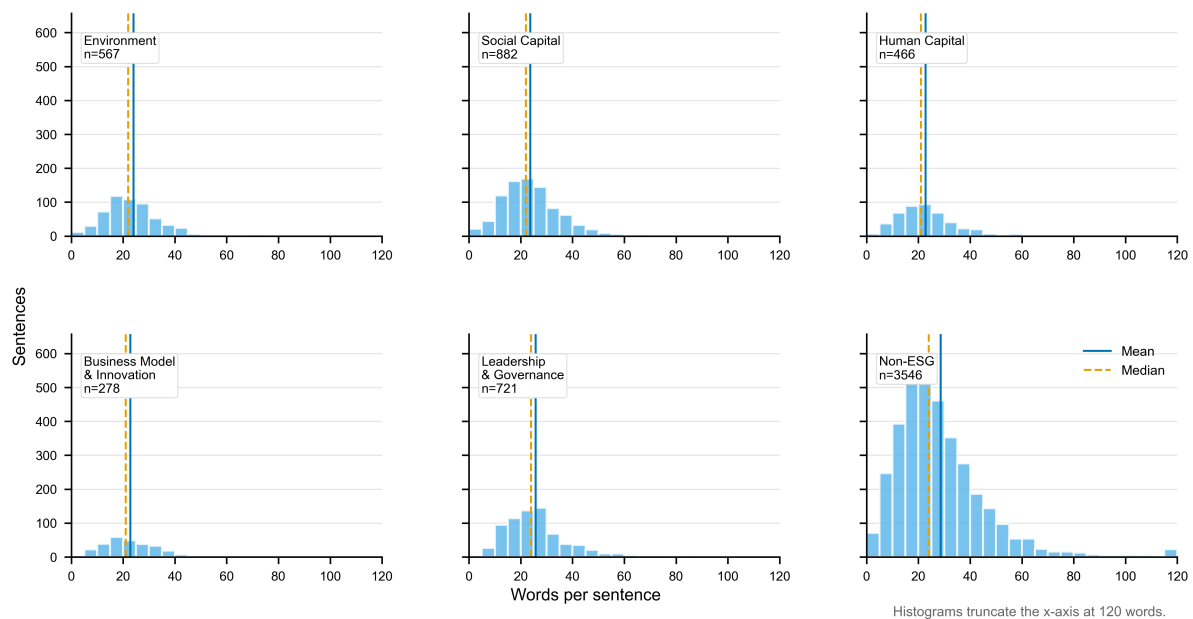


Figure 1. Corpus structure and sentence-level classification constraints. Panel (A) summarizes corpus size, report coverage, label hierarchy, split sizes, and sentence length. Panels (B–D) show full-corpus parent category imbalance, the child label long tail, and parent-specific sentence length histograms. The child label count panel uses a log scale because child label counts are highly uneven, and the sentence length histograms truncate the x-axis at 120 words for readability. Table A1 compares the full corpus with the held-out test set used for final reuse analyses.

Table 4. Paraphrased sentence label examples used for corpus orientation. The public artifacts omit the full raw Text column; these short paraphrases illustrate label use without redistributing source sentences.

Paraphrased Sentence Description	Released Child Label	Parent Category Label	Comment
A sentence describes projected annual carbon-dioxide-equivalent avoidance from environmental projects.	GHG Emissions	Environment	Direct topical signal
A sentence describes a third-party valuation review related to goodwill impairment.	Non-ESG	Non-ESG	Accounting/report fragment
A sentence states that customer data are protected according to contractual commitments.	Customer Privacy	Social Capital	Potentially adjacent SASB issue
A sentence reports global employee completion of cybersecurity training.	Data Security	Social Capital	Context-dependent issue
A sentence describes procedures for responding to non-permitted wastewater discharge.	Water & Wastewater Management	Environment	Direct environmental disclosure

The corpus uses a fixed two-level hierarchy. Each sentence first receives one child label, and the child label maps deterministically to one parent category. We work with three projections of that same hierarchy: a binary ESG versus Non-ESG task, a 27-class child label task, and a parent-level classification task defined on the ESG subset. Parent-level classifier evaluation therefore excludes Non-ESG and is computed only among ESG-labeled sentences. Profile aggregation later retains Non-ESG as a separate corpus category because dashboard-like summaries may show the ESG/Non-ESG balance. This design follows the broader logic of hierarchical classification, where model behavior and evaluation can change materially across levels of the label tree [37,38].

The hierarchy lets us distinguish two substantively different forms of disagreement: some departures from the released label preserve the broad ESG dimension while changing the child label; others move a sentence across parent categories or into Non-ESG. This separation matters because local disagreement within a parent branch differs from large interpretive shifts across branches. A change from Customer Privacy to Data Security differs from a change from either label to Non-ESG, and both differ from a change from Social Capital to Leadership & Governance. Standard flat metrics score all three outcomes identically if they deviate from the released child label, so our design supplements task metrics with retention and label change rates that track whether disagreements remain within a parent branch or cross the broader hierarchy.

The corpus is split into training, validation, and test sets using stratified random sampling on the child-label distribution with a fixed random seed of 42. The split proportions are 70%/15%/15%, yielding 4522 training sentences, 969 validation sentences, and 969 test sentences. The classifier pipelines use all three splits for distinct roles: training sentences fit the supervised classifiers, validation sentences select the GPT-4o prompt family, and test sentences provide the final classifier metrics and GPT-4o candidates used in candidate-visible annotation. All reported final comparisons use the same held-out test set of 969 sentences, so annotation condition agreement, classifier-based reproduction, and ESG category profile sensitivity are evaluated on a common sample pool. Stratification preserves child label composition by design; Appendix A additionally checks parent category shares, ESG/Non-ESG shares, and sentence length summaries in the held-out test set.

3.3. Classifier-Based Reproduction of Released Labels

The classifier reproduction module compares three automated label sources that a downstream user might use to reproduce or extend the released labels: lexical supervised classifiers, sentence embedding classifiers, and prompted GPT-4o classification. In this module, reproduction refers to agreement with released labels on held-out sentences after

model fitting or prompt selection. Macro F1 is used as the main reproduction measure because it gives each category equal weight, preventing frequent categories from dominating the score. Table 5 summarizes how each source converts a sentence into a label and how it enters the later profile analysis. The supervised models are trained on the training split and evaluated on the same held-out test set used for annotation. GPT-4o is zero-shot evaluated under prompt families selected on the validation split.

Table 5. Automated label sources in the classifier reproduction module. The table gives each source's pipeline role; implementation details follow in the text.

Label Source	Sentence Representation	Split Use	Role in This Study
TF-IDF classifiers	scikit-learn lowercasing and tokenization, word unigram–bigram TF-IDF features	Train for fitting; test for final evaluation	Lexical supervised baseline; LinearSVC predictions enter profile aggregation
SBERT classifier	all-mpnet-base-v2 tokenizer and frozen 768-dimensional sentence embeddings	Train for fitting; test for final evaluation	Embedding supervised baseline and profile aggregation label source
GPT-4o classifier	Prompted classification using the sentence, label resources, and JSON output schema	Validation for prompt family selection; test for final evaluation	Prompted LLM label source and candidate label shown in candidate-visible annotation

Two supervised model families provide non-LLM baselines and serve as label sources in the aggregation analysis. Both start from the raw sentence string and convert it into model input through the preprocessing attached to the feature extractor. The lexical family uses the default scikit-learn `TfidfVectorizer` settings for lowercasing and tokenization, builds word unigram–bigram features, and represents each sentence as a TF-IDF vector [39]. Logistic regression and LinearSVC heads are trained on these features [40]. The embedding family tokenizes sentences with the pretrained Sentence-Transformers tokenizer associated with all-mpnet-base-v2, then uses the frozen encoder to produce 768-dimensional sentence embeddings [20]. A logistic regression head is trained on the resulting embeddings. Table 6 lists the implementation settings for the supervised and GPT-4o classifiers.

All deterministic preprocessing, model fitting, metric calculation, profile aggregation, and figure generation are implemented in the accompanying Python 3.9.13 scripts. The repository includes an environment file specifying Python dependencies, together with scripts and tests for reproducing preprocessing, model fitting, metric calculation, profile aggregation, and figure generation. The Supplementary Materials provide the archived reproducibility artifacts, including split metadata, prompt templates, model predictions, LLM annotation outputs, profile sensitivity outputs, figure inputs, and scripts. The supervised classifiers, split construction, Monte Carlo partitioning, and figure inputs use random seed 42 where randomization is involved. LLM calls are not rerun during ordinary reproduction; cached GPT-4o and Claude outputs are included with the prompt families, model identifiers, temperature settings, and token budget used in the study.

GPT-4o is evaluated as a zero-shot classifier under three prompt families that vary the information supplied alongside the sentence. Here, zero-shot means that no labeled examples, task-specific fine-tuning, or retrieval from surrounding report context is supplied at inference time. The *Minimal* family supplies the 27 child label names, the *Definitions* family adds a one-sentence definition for each SASB issue category, and the *Hierarchy* family supplies both the parent–child mapping and the child label definitions. The latter two prompts are therefore context-enriched relative to the minimal label list prompt while still preserving the isolated sentence reuse setting studied in this paper. Differences across prompt families are treated as prompt sensitivity evidence. Appendix B reproduces the full prompt text.

Table 6. Implementation settings for classifier reproduction. All supervised classifiers use raw sentence text and a fixed random seed of 42.

Component	Settings
TF-IDF vectorizer	word unigrams and bigrams; sublinear term frequency scaling; maximum vocabulary of 50,000 features
TF-IDF classifier heads	logistic regression with <code>lbfgs</code> , $C = 1.0$, balanced class weights, and 1000-iteration limit; LinearSVC with $C = 1.0$, balanced class weights, and 2000-iteration limit
SBERT encoder	<code>all-mpnet-base-v2</code> ; frozen encoder; 768-dimensional sentence embeddings
SBERT classifier head	logistic regression with $C = 1.0$, balanced class weights, and 1000-iteration limit
GPT-4o classifier calls	zero-shot prompted classification; temperature 0; maximum token budget of 900 tokens; OpenRouter API
GPT-4o output schema	<code>top1_label</code> , <code>top3_labels</code> , <code>confidence</code> , and <code>rationale</code>

Classifier reproduction is summarized with the metrics in Table 7. Prompt family selection is based on the validation set child Macro F1 against the released labels. The prompt family with the highest validation child Macro F1 is carried forward as the validation-selected GPT-4o prompt for test set evaluation and candidate-visible annotation. This step produces the GPT-4o candidate label shown to Claude in the candidate-visible annotation condition. Binary scores evaluate ESG versus Non-ESG, child scores evaluate the 27 released child labels, and parent-level reproduction scores are computed on ESG-labeled sentences only.

Table 7. Evaluation metrics for classifier-based label reproduction. These metrics measure agreement with released labels; they do not adjudicate substantive SASB correctness.

Metric	Definition	Use in This Study
Macro F1	Macro F1 averages per-category F1 scores, giving rare and common categories equal weight [41]	Main closeness-to-released-labels measure for binary, parent, and child projections. A value closer to 1 indicates closer reproduction of the released labels
Validation child Macro F1	Macro F1 on the validation split for the 27 released child labels	Selects the GPT-4o prompt family carried forward to the test set
Top-3 accuracy	Share of test sentences where the released child label appears among GPT-4o's three returned labels	Reports whether GPT-4o retrieves the released label as a ranked candidate
Prompt family flip rate	Share of test sentences where two GPT-4o prompt families assign different child or parent labels	Summarizes prompt sensitivity across Minimal, Definitions, and Hierarchy prompts

3.4. LLM Annotation Under Blind and Candidate-Visible Conditions

The annotation module uses Claude Sonnet 4.6, accessed through the OpenRouter API as `anthropic/claude-sonnet-4.6`, as a single LLM annotator operating under two information conditions on the held-out test set. The two LLMs have separate roles in the design: GPT-4o, accessed as `openai/gpt-4o`, serves as the prompted classifier and candidate label generator, and Claude Sonnet 4.6 serves as the fixed LLM annotator. In blind annotation, the LLM annotator receives only the sentence and the author-prepared short definitions for the SASB child labels used in the prompts. These definitions are fixed before inference and reused across all sentences. The system message asks for the best-fitting child label and parent category, a confidence value, and an ambiguity flag for sentences with ambiguous child assignment. The JSON output contains `child_label`, `parent_label`, `confidence`, `ambiguity_flag`, and `note`. All annotation calls use temperature 0 and a maximum token budget of 900 tokens; Appendix B gives the full prompt text.

Candidate-visible annotation uses the same held-out sentences and the same LLM annotator to simulate a workflow in which a labeler sees the released dataset label beside a model-generated candidate. The GPT-4o candidate is produced by the validation-selected prompt described in Section 3.3. The prompt instructs the model to choose among four verdicts: released label, model candidate, both visible labels, or another label. The JSON output contains `verdict` (one of *original*, *model*, *both*, *other*), `recommended_child`, `recommended_parent`, `confidence`, `ambiguity_flag`, and `note`; the *original* verdict value

denotes the released dataset label. Metric computation and profile aggregation use the `recommended_child` field as the candidate-visible annotation label, including cases where `verdict` is *both*. The `verdict` is retained as a diagnostic field. The ambiguity rate is the share of test sentences for which the LLM annotator sets the ambiguity flag under a given prompt; this field records the LLM annotator's uncertainty signal and is not a human-validated ambiguity measure. Child retention is the share of test sentences for which the annotation label equals the released child label; parent retention is the corresponding share after mapping child labels to parent categories. Within-parent label changes alter the child label while retaining the released parent category, whereas cross-parent label changes alter the parent category assignment or move between ESG and Non-ESG. For reported parent-level metrics, LLM annotator parent labels are derived from the selected child label using the corpus hierarchy, while model-supplied parent strings are retained only as raw diagnostic fields. Annotation condition agreement is summarized with child retention, parent retention, within-parent label change rate, cross-parent label change rate, ambiguity rate, Cohen's κ (a chance-corrected agreement statistic), and blind-candidate-visible disagreement.

3.5. Aggregation into ESG/Non-ESG Category Profiles

Profile sensitivity concerns whether local label source differences remain visible after aggregation. In ordinary use, sentence labels may be aggregated within a company report, a portfolio, a sector dashboard, or another disclosure summary unit. The public corpus used here lacks firm identifiers for constructing actual firm-level profiles. We therefore conduct a Monte Carlo profile sensitivity analysis on artificial aggregation units. In each repetition, the 969 held-out sentences are randomly partitioned into eight approximately equal batches, each containing about 121 sentences. Repeating this procedure 500 times yields 4000 artificial aggregation units ($8 \text{ batches} \times 500 \text{ repetitions}$). For each unit and each label source, we compute the share of sentences assigned to each SASB parent category and to Non-ESG. This procedure varies the composition of artificial sentence pools and estimates how much the visible category profile changes when released labels are replaced by model predictions or LLM annotation labels. The resulting batches serve as stress test aggregation units, not as companies, industries, or portfolios.

The six label sources are the released labels, TF-IDF (LinearSVC) predictions, SBERT predictions, GPT-4o predictions from the validation-selected prompt, blind annotation labels, and candidate-visible annotation labels. Supervised model predictions are generated by applying the fitted models from Section 3.3 to the test set; annotation-based labels come from the outputs of Section 3.4; GPT-4o predictions come from the same validation-selected prompt family used in the classifier module. For aggregation unit b , label source s , and category k , let $p_{b,s}(k)$ be the share of sentences assigned to category k . We summarize profile sensitivity using the maximum absolute category share difference, $L_{\infty,b}(s) = \max_k |p_{b,s}(k) - p_{b,\text{released}}(k)|$, and the signed category share difference for each category. This maximum is the largest single-category share shift a profile user would see when an alternative label source replaces released labels. Child-level L_1 summaries are retained in Appendix A to show how much parent aggregation compresses fine-grained disagreement.

4. Results

4.1. Reuse Validation Across Label Granularities

Reuse validation shows that released labels are most stable after the label space is coarsened. Figure 2 reports agreement with dataset-supplied SASB-aligned labels across hierarchy levels. Macro F1 against released labels summarizes how closely each automated source matches those labels while giving rare and frequent categories equal weight.

A higher Macro F1 indicates closer agreement with the released labels. Binary is the coarsest task because it only separates ESG from Non-ESG. Parent-level classification is intermediate because it uses broad SASB parent categories for ESG-labeled sentences. Child-level classification is the most fine-grained task because it distinguishes the 27 released child labels.

Panel A compares classifier families and highlights the label sources carried into the aggregation analysis, with GPT-4o represented by the validation-selected *definitions* prompt. TF-IDF + LinearSVC reproduces the released child labels most closely with Macro F1 = 0.5707, followed by SBERT at 0.4612 and GPT-4o with *definitions* at 0.3604. Differences are smaller at coarser label levels: binary Macro F1 is high for the supervised models, and parent-level Macro F1 lies in a narrow band around 0.72 for TF-IDF and SBERT. These scores quantify agreement with the released label scheme under each hierarchy projection.

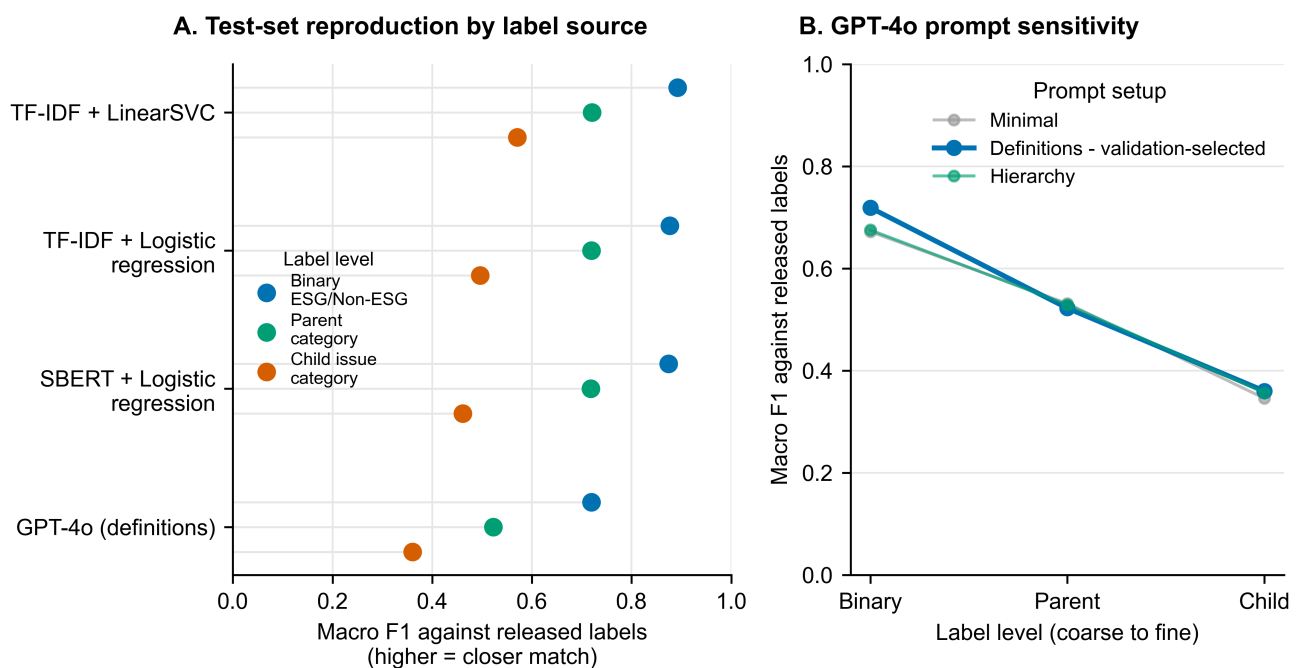


Figure 2. Test set reproduction of released dataset labels by automated label sources. Macro F1 is computed against released labels. Binary refers to ESG versus Non-ESG, parent refers to SASB parent categories among ESG-labeled sentences, and child refers to the 27 released child labels. Panel (A) compares TF-IDF and SBERT supervised classifiers with GPT-4o using the validation-selected *Definitions* prompt. Panel (B) shows GPT-4o prompt family sensitivity across Minimal, Definitions, and Hierarchy prompts.

GPT-4o prompt family variation is concentrated at the child level. Panel B shows that all three prompts decline sharply from binary to child-level classification. On validation, *Definitions* is closest to the released child labels and is therefore selected for later use. On test, *Definitions* remains closest at the child level, but the margins over the other prompt families are small. Pairwise prompt flips are also concentrated at the child level, reaching roughly one quarter of test sentences in some prompt comparisons. Exact prompt family metrics, flip rates, and top-3 accuracy values are provided in Appendix A.

The child task remains difficult across model families, and prompt family effects are concentrated at the level where the label space is most granular. GPT-4o often places the released label among its top candidates, yet the top-1 child Macro F1 remains well below binary and parent scores. In this corpus, the lexical baseline reproduces released label patterns more closely than the pretrained language models, consistent with disclosure

measurement research showing that domain-specific lexical signals can remain competitive when the task is to reproduce a particular coding scheme [35].

Additional held-out accuracy, macro precision, and macro recall diagnostics are reported in Appendix A, Table A5; the revised reproducibility artifacts provide the corresponding per-class precision, recall, F1, support, and confusion matrix CSV files.

4.2. Annotation Conditions Change Retention and Ambiguity Signals

The annotation comparison shows different retention and ambiguity signals across the two information conditions. Blind annotation retains 54.8% of the released child labels and 59.9% of the derived parent category labels, while candidate-visible annotation raises those rates to 64.8% and 70.0%. Chance-corrected agreement follows the same pattern, with child-level κ rising from 0.395 to 0.504 and parent-level κ rising from 0.432 to 0.554. Figure 3 separates label retention against released labels, chance-corrected agreement, ambiguity flags, and candidate-visible annotation decisions so that each panel is read on its own scale, while the full annotation condition agreement table is available in Appendix A.

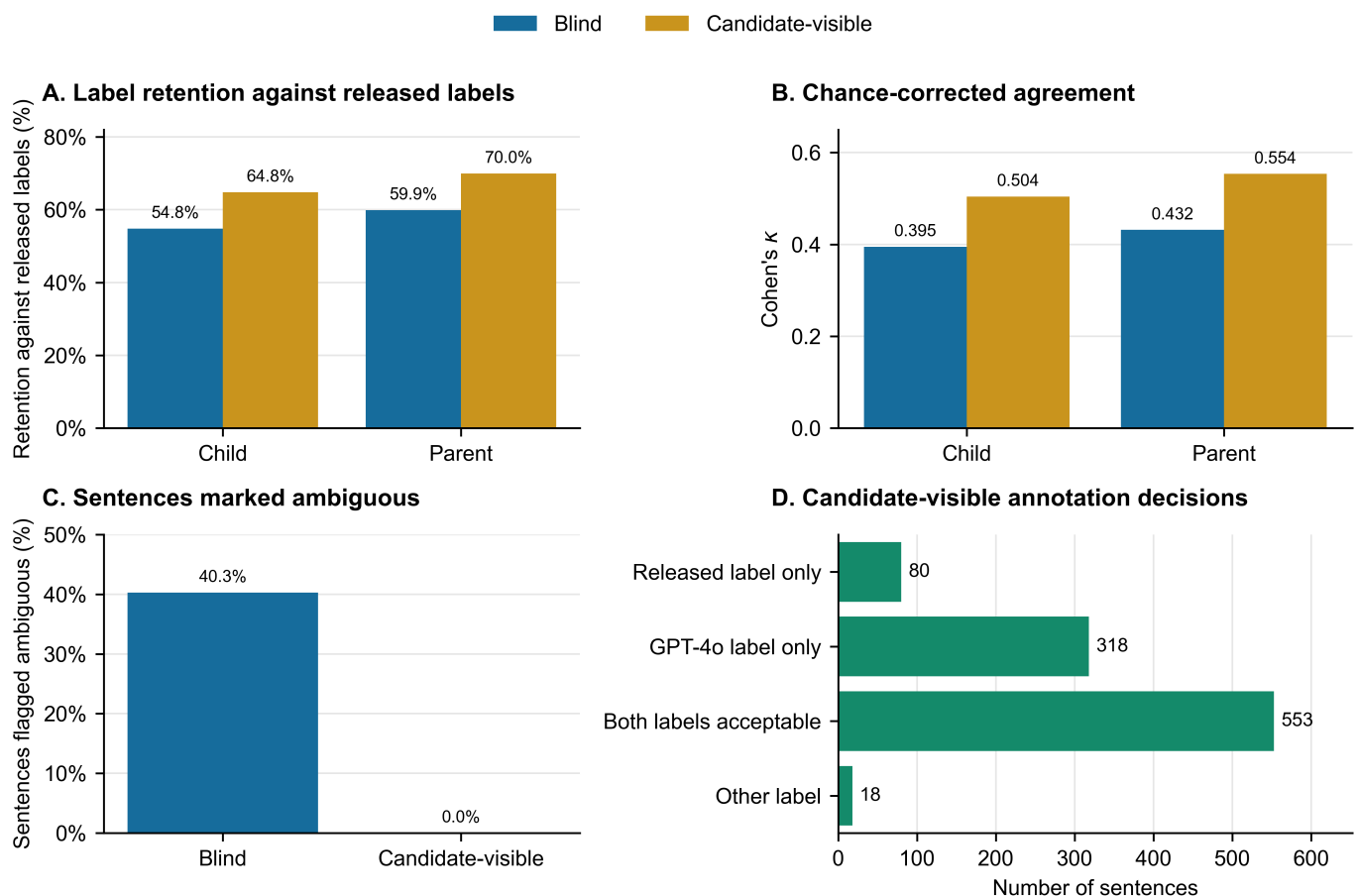


Figure 3. Annotation outcomes under blind and candidate-visible conditions. In blind annotation, the LLM annotator sees only the sentence and SASB label definitions; in candidate-visible annotation, the LLM annotator also sees the released dataset label and the GPT-4o candidate label. Panels (A,B) report agreement with the released labels at child and parent levels using label retention and Cohen's κ . Panel (C) reports the share of sentences flagged as ambiguous. Panel (D) reports candidate-visible annotation decision counts: released label only, GPT-4o label only, both labels acceptable, or another label.

Blind annotation flags ambiguity in 40.3% of the test sentences, while candidate-visible annotation records zero ambiguity flags. Blind and candidate-visible annotation assign different child labels to 19.1% of the same sentences. Within the candidate-visible condition, 553 cases judge both visible labels acceptable, 318 choose the GPT-4o label only, 80 choose the released label only, and 18 move to another label. Because the two prompts define different information conditions, Section 5.2 interprets the assisted format separately from the descriptive results.

4.3. Aggregate Profiles Are Sensitive to Label Source

The aggregation analysis shows that ESG/Non-ESG profiles remain sensitive to sentence-level label source. Figure 4 separates the largest single-category share shift from which categories gain or lose share. The Monte Carlo batches provide artificial aggregation units for a held-out sentence pool stress test. Across these batches, TF-IDF has the smallest mean largest single-category share shift from the released label profile, candidate-visible annotation remains relatively close, GPT-4o and blind annotation move farther away, and SBERT produces the largest single-category shifts. The full parent-level distance table is provided in Appendix A.

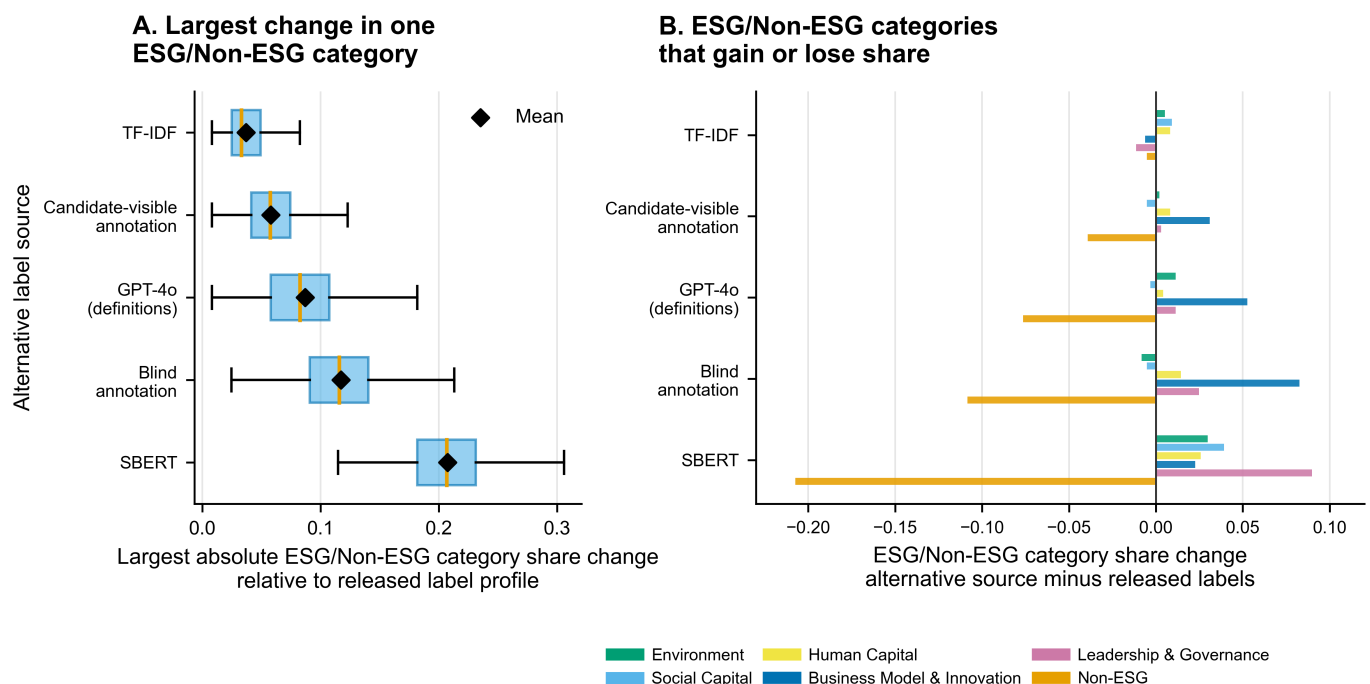


Figure 4. Sensitivity of aggregate ESG/Non-ESG category profiles to the choice of label source. A profile is the percentage distribution of sentences across Environment, Social Capital, Human Capital, Business Model & Innovation, Leadership & Governance, and the separate Non-ESG corpus category. Because firm identifiers are unavailable, the held-out test sentences are repeatedly partitioned into artificial Monte Carlo aggregation units and summarized under each label source. Panel (A) shows, for each alternative label source, the distribution of the largest absolute change in any one ESG/Non-ESG category share relative to the released label profile, with diamonds marking means; 0.05 corresponds to five percentage points. Panel (B) shows the average direction of change by category, calculated as the alternative source minus released labels. Positive values mean that the alternative source assigns a larger share of sentences to that category, while negative values mean a smaller share. Released labels provide the profile baseline.

The category-level shifts show why distance magnitude matters for downstream interpretation. Figure 4B shows which category shares increase or decrease under each source.

SBERT has the largest profile change, driven by a Non-ESG share that is 20.7 percentage points lower relative to the released label profile (-0.207) and a Leadership & Governance share that is 9.0 percentage points higher ($+0.090$). Blind annotation also lowers the Non-ESG share by 10.8 percentage points (-0.108) and raises Business Model & Innovation by 8.3 percentage points ($+0.083$), while candidate-visible annotation produces a smaller version of the same pattern, with Non-ESG being 3.9 percentage points lower (-0.039) and Business Model & Innovation 3.1 percentage points higher ($+0.031$). TF-IDF stays closest to the released label profile because it was trained directly on those labels; annotation-based sources involve independent judgment that can change released assignments and shift the distribution.

The child-level L_1 distances in Table A7 exceed the corresponding parent-level L_1 distances in Table A6 for every alternative source, showing that parent aggregation compresses total fine-grained reallocation. The L_∞ comparison is narrower because the largest single-category shift can be similar at child and parent levels. The remaining parent-level differences are still visible enough to affect a dashboard or disclosure coverage summary built from the same sentences.

5. Discussion

5.1. Reusable Annotated Text Labels Need Stability Evidence

The results frame sentence-level SASB-aligned labels as reusable but condition-sensitive data records. The consistent gap between child- and parent-level agreement suggests that fine-grained SASB categories require more contextual information than isolated sentences provide. Parent categories are coarser and more durable, yet they remain condition-sensitive. Distinctions such as Customer Privacy versus Data Security often depend on section headers, neighboring sentences, and industry framing, which disappear when a disclosure is reduced to one sentence. Parent categories survive that reduction more often because they encode the coarser semantic features that short text fragments still carry. The result helps explain why a sentence-level SASB-aligned dataset label can be useful as data infrastructure while still requiring provenance when it is reused.

The classifier results add a complementary signal about reusability. TF-IDF reproduces the released label scheme most closely in this corpus, which indicates that lexical patterns can carry much of the dataset's released labeling logic. The drop from binary and parent categories to child labels shows where that logic becomes harder to extend to held-out sentences. It also shows why all child-level mismatches should not be read with the same substantive severity: a sibling category change within Social Capital is closer to a semantic boundary disagreement than a cross-parent move or an ESG/Non-ESG reversal.

This interpretation places the study in the field of text-as-data validation as an analysis of label reuse. It is also consistent with semantic representation research, where sentence embeddings and contrastive objectives have been evaluated by whether they preserve the semantic relations needed by a downstream task [20–23]. In the present benchmark, the operational version of that question is whether an alternative label source preserves the released label exactly, stays within a semantically adjacent parent branch, or changes the broader ESG/Non-ESG interpretation. The results also fit with sustainability reporting research showing that materiality judgments retain interpretive margins even when standards provide a common vocabulary [5,6]. Missing context remains one possible cause of the observed disagreement, alongside annotator error, LLM annotator miscalibration, and redundancy in the label set. A document window condition would be needed to separate those explanations.

5.2. Annotation Condition Should Travel with the Label

Candidate-visible annotation is best treated as an assisted annotation condition, not as independent validation. Because the released label and GPT-4o candidate are visible, the LLM annotator resolves a bounded comparison and then supplies a single recommended label for downstream metrics. The zero ambiguity flags and the high share of *both labels acceptable* verdicts describe the decision format as much as the underlying sentences. This interpretation is consistent with the anchoring or candidate visibility effects discussed in the LLM-as-a-judge research [32] and with human–AI collaboration research that treats AI support as an explicit stage of the analytic workflow [30,31].

These patterns point to label provenance as a necessary part of dataset documentation. Dataset documentation work argues for recording provenance, intended use, scope, and collection conditions [42]; model cards extend the same logic to model outputs and evaluation settings [43]. Disagreement research adds a label-level reason for such documentation, because preserving annotator uncertainty can yield richer representations than forcing consensus on a single label [26]. In this setting, concrete documentation fields include the annotation condition under which each label is assigned, whether the LLM annotator sees candidate labels, prompt sensitivity estimates for LLM-generated labels, and a list of child categories that prove fragile under reannotation.

5.3. Aggregate Profiles Should Cite Their Label Source

The aggregation exercise shows that label source choices remain visible after sentence labels are summarized into ESG/Non-ESG profiles. TF-IDF stays closest to the released label profile, while SBERT and blind annotation produce the largest profile shifts. The category share changes clarify the direction of those distances: alternative sources generally reduce the Non-ESG share and reallocate sentences into ESG parent categories, most visibly for SBERT and blind annotation. The result concerns one sentence corpus and one profile construction exercise, but it is still consequential for disclosure analysis because changing the sentence-level label source changes the category proportions that an analyst would see.

Profile comparisons should therefore cite the source of labels, the annotation or prompt condition where applicable, and the aggregation level used to summarize disclosure emphasis across documents or systems. This is a data record requirement separate from claims about provider-level ESG ratings; the same sentence pool can yield different visible profiles before any weighting scheme, rating architecture, or score construction is introduced.

5.4. Usage Notes and Limitations

The benchmark should be reused with its scope constraints visible. This study validates reuse behavior of released sentence-level labels; it does not adjudicate whether each released label is the substantively correct SASB interpretation. The empirical scope is one public sentence corpus, one released label hierarchy, and a held-out sentence pool analysis. The Monte Carlo batches are artificial aggregation units for profile sensitivity analysis; firm-, industry-, and portfolio-level ESG measurement would require records that group sentences by reporting entity. The corpus also represents disclosures as isolated sentences, so the study evaluates label reuse after context removal.

Finance-specific language models such as BloombergGPT, a 50B-class model pre-trained on financial and general domain corpora, provide another natural extension of this benchmark [44]. Future work can combine such domain-pretrained models with document window retrieval, neighboring paragraphs, section headings, and industry metadata to test whether richer financial context reduces the child label instability observed here.

The annotation and model results should be read as workflow-specific evidence. Claude Sonnet 4.6 provides a fixed LLM annotator for the two annotation conditions,

while expert SASB coders or human adjudication would provide a different evidential standard. Ambiguity flags are LLM-reported fields, and proprietary LLM outputs can vary by model version, provider route, and access date even with temperature set to zero. The reported analysis therefore uses cached outputs from the stated model identifiers. Classifier metrics quantify reproduction of released labels. A broader validation program could combine document windows, multiple annotator families, human adjudication, and richer uncertainty models, following the text-as-data principle that computational text outputs should be evaluated against their intended use case [19,23].

6. Conclusions

This article presented a reproducible benchmark and validation workflow for reusing sentence-level SASB-aligned ESG labels. In a public corpus of 6460 SASB-labeled sentences, label reproduction is strongest at coarser levels, blind LLM annotation flags substantial ambiguity in isolated sentences, candidate-visible annotation records assisted resolution, and alternative label sources reshape the ESG/Non-ESG category profiles produced after aggregation. The evidence supports reuse under documented provenance: released labels, supervised classifier predictions, prompted LLM predictions, and LLM annotation labels are usable analytic inputs when their source and information condition are visible.

For downstream users, label provenance belongs beside the output. When sentence labels are reused for classifier training, annotation, or aggregate ESG/Non-ESG category profiles, downstream reports should identify the label source, annotation condition, prompt family, and aggregation level. The reproducibility artifacts turn this reporting norm into reusable records by providing split metadata, label mappings, prompt templates, cached model outputs, annotation outputs, profile sensitivity outputs, figure inputs, and reproduction scripts. This design lets users interpret AI-assisted sustainability disclosure analytics as documented, context-sensitive summaries.

Supplementary Materials: The reproducibility artifacts are archived at Zenodo: <https://doi.org/10.5281/zenodo.20105936>.

Author Contributions: Conceptualization and methodology, Y.L., T.C. and P.P.; software and data analysis, Y.L. and T.C.; writing—original draft, Y.L. and T.C.; writing—review and editing, all authors; supervision, P.P. and W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Institutional Review Board Statement: The institutional submission code for the Faculty of Applied Sciences (Faculdade de Ciências Aplicadas, FCA), Macao Polytechnic University, is fca.1337.a020.9.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source corpus analyzed in this study is the public *SASB-Aligned ESG Sentences* dataset [3], available at <https://www.kaggle.com/datasets/edwardjunprung/sasb-aligned-esg-sentences> (accessed on 4 May 2026). The Kaggle landing page describes 6454 labeled examples; the analysis uses the downloaded `companies.csv` file with 6460 rows after loading. Reproducibility artifacts are archived at Zenodo: <https://doi.org/10.5281/zenodo.20105936>. Raw sentence text from the third-party source corpus is not redistributed; users must obtain it from the source dataset and can reconstruct the analysis table using the provided manifest, split metadata, and sentence hashes. Code is released under the MIT License. Derived metadata and outputs are released under CC BY 4.0.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Supplementary Results

Appendix A.1. Corpus Split Check

Table A1. Full-corpus and held-out test set comparison. Sentence length is counted in words. Percentage point differences are absolute differences between the full corpus and the held-out test set.

Quantity	Full Corpus	Held-Out Test Set	Check
Sentences	6460	969	Final reuse comparisons use the held-out test set
Child labels represented	27	27	All released child labels remain represented
Smallest child label count	6	1	Rare released labels are retained in the task
Non-ESG share	54.89%	54.90%	0.01 percentage points
Largest ESG/Non-ESG category share difference	—	—	0.11 percentage points
Largest child label share difference	—	—	0.07 percentage points
Mean sentence length	26.53	26.27	0.26 words
Median sentence length	23	23	No difference
Interquartile range of sentence length	16–32	16–33	Similar middle range

Appendix A.2. Classifier and Annotation Results

Table A2. Annotation condition agreement with the released labels on the held-out test set. Retention and ambiguity are reported as rates; κ denotes Cohen’s chance-corrected agreement. Parent metrics are derived from LLM annotator child labels through the fixed corpus hierarchy.

Mode	N	Retention Rate		Cohen’s κ		Ambiguity Rate
		Child	Parent	Child	Parent	
Blind	969	0.5480	0.5986	0.3950	0.4317	0.4025
Candidate-visible	969	0.6481	0.6997	0.5042	0.5535	0.0000

Table A3. GPT-4o prompt family comparison. The validation child Macro F1 column reports the selection criterion; all remaining metrics are computed on the held-out test set.

Prompt	Validation Child Macro F1	Test Binary Macro F1	Test Parent Macro F1	Test Child Macro F1	Test Child Top-3 Acc.
Minimal	0.3790	0.6721	0.5317	0.3461	0.6347
Definitions	0.3823	0.7192	0.5225	0.3604	0.7203
Hierarchy	0.3769	0.6755	0.5292	0.3569	0.5841

Table A4. Classifier family comparison on the held-out test set. GPT-4o is reported with the validation-selected *Definitions* prompt, chosen on validation by child Macro F1 against the released child labels.

Method	Binary Macro F1	Parent Macro F1	Child Macro F1	Child Top-3 Acc.
TF-IDF + LinearSVC	0.8919	0.7204	0.5707	—
TF-IDF + Logistic regression	0.8764	0.7194	0.4959	—
SBERT + Logistic regression	0.8744	0.7182	0.4612	—
GPT-4o (definitions)	0.7192	0.5225	0.3604	0.7203

Appendix A.3. Profile Distance Results

For label source s in aggregation unit b , the category profile is the vector of category shares $p_{b,s}(k)$. The profile distance tables report two summaries relative to the released label baseline: $L_1 = \sum_k |p_{b,s}(k) - p_{b,\text{released}}(k)|$, the total profile reallocation across categories, and $L_\infty = \max_k |p_{b,s}(k) - p_{b,\text{released}}(k)|$, the largest single parent category share shift or child label share shift. The main text uses the parent-level L_∞ distribution because it is the most direct description of the largest category-level change a profile user would see.

Table A5. Additional held-out test set classifier diagnostics requested during review. Accuracy, macro precision, macro recall, and Macro F1 are computed against the released labels. The parent task is evaluated on ESG-labeled test sentences only; binary and child tasks use all 969 held-out test sentences. Full per-class precision, recall, F1, support, and confusion matrix CSV files are provided in the revised reproducibility artifacts.

Method	Task	N	Accuracy	Macro Precision	Macro Recall	Macro F1
TF-IDF + LinearSVC	Binary	969	0.893	0.891	0.893	0.892
TF-IDF + LinearSVC	Parent	437	0.739	0.731	0.713	0.720
TF-IDF + LinearSVC	Child	969	0.771	0.605	0.567	0.571
SBERT + LogReg	Binary	969	0.875	0.874	0.876	0.874
SBERT + LogReg	Parent	437	0.730	0.712	0.727	0.718
SBERT + LogReg	Child	969	0.612	0.444	0.593	0.461
GPT-4o (definitions)	Binary	969	0.719	0.723	0.724	0.719
GPT-4o (definitions)	Parent	437	0.563	0.590	0.484	0.522
GPT-4o (definitions)	Child	969	0.554	0.349	0.448	0.360

Table A6. Parent-level label distribution distances relative to the released labels across 4000 artificial aggregation units (8 batches $\times$ 500 Monte Carlo repetitions).

Source	Mean Parent L_1	P90 Parent L_1	Mean Parent L_∞	95% Interval for L_∞
TF-IDF	0.1051	0.1488	0.0369	[0.0165, 0.0738]
Candidate-visible annotation	0.1537	0.2149	0.0580	[0.0246, 0.1157]
GPT-4o (definitions)	0.2160	0.2975	0.0869	[0.0331, 0.1653]
Blind annotation	0.2859	0.3802	0.1171	[0.0574, 0.1967]
SBERT	0.4225	0.5124	0.2074	[0.1393, 0.2810]

Table A7. Child-level label distribution distances relative to the released labels across 4000 artificial aggregation units (8 batches $\times$ 500 Monte Carlo repetitions).

Source	Mean Child L_1	P90 Child L_1	Mean Child L_∞	95% Interval for L_∞
TF-IDF	0.2057	0.2645	0.0329	[0.0165, 0.0579]
Candidate-visible annotation	0.3516	0.4298	0.0593	[0.0248, 0.1157]
GPT-4o (definitions)	0.4334	0.5289	0.0846	[0.0331, 0.1653]
Blind annotation	0.5022	0.6116	0.1111	[0.0492, 0.1967]
SBERT	0.5354	0.6281	0.2074	[0.1393, 0.2810]

Appendix B. Full Prompt Text

This appendix reports the exact prompt templates used for LLM classification (Section 3.3) and label annotation (Section 3.4) in a structured format. Each call combines a short system instruction with a user message that inserts the sentence text and the relevant label resources; {text}, {label_list}, {definitions}, {hierarchy}, and {model_candidate} denote runtime fields substituted by the implementation, while {original_label} is the implementation field name for the released label shown in candidate-visible annotation. In the prompt text, *original label* refers to the released dataset label supplied by the corpus creators. All calls use temperature 0 and a maximum token budget of 900 tokens.

Table A8. Prompt overview for classification and annotation. The table records the decision task, the information exposed to the model, and the required JSON fields for each prompt family.

Prompt Family	Task	Information Supplied	Required JSON Fields
Minimal	GPT-4o child classification	Sentence text and the 27 child label names	top1_label, top3_labels, confidence, rationale
Definitions	GPT-4o child classification	Sentence text and child label definitions	top1_label, top3_labels, confidence, rationale
Hierarchy	GPT-4o hierarchy-aware classification	Sentence text, parent–child mapping, and child label definitions	parent_label, top1_label, top3_labels, confidence, rationale

Table A8. Cont.

Prompt Family	Task	Information Supplied	Required JSON Fields
Blind annotation	Claude LLM annotation	Sentence text and child label definitions	child_label, parent_label, confidence, ambiguity_flag, note
Candidate-visible annotation	Claude candidate-visible annotation	Sentence text, released label, model candidate, and child label definitions	verdict, recommended_child, recommended_parent, confidence, ambiguity_flag, note

Appendix B.1. Classification Prompts (GPT-4o)

The three classification families share the same sentence-level decision target, but they differ in how much task structure is exposed to GPT-4o. The templates below reproduce the exact wording while separating the system instruction from the user-supplied fields.

Minimal template.

The minimal template exposes only the sentence and the list of child label names, so it tests whether label names alone are sufficient for a useful top-1 and top-3 ranking.

System. “You classify corporate disclosure sentences into one of the provided SASB child labels. Use only the supplied label set and return JSON.”

User template. “Sentence: “{text}”. Child labels: {label_list}. Return valid JSON with keys top1_label, top3_labels, confidence, rationale.”

Definitions template.

The definitions template preserves the same output schema, and it adds the child label definitions so that the model can anchor each decision to a short issue description plus the label name.

System. “You are an ESG disclosure analyst classifying sentences with reference to SASB child category definitions.”

User template. “Sentence: “{text}”. Use the following SASB child labels and definitions: {definitions}. Return valid JSON with keys top1_label, top3_labels, confidence, rationale.”

Hierarchy template.

The hierarchy template adds the parent–child mapping alongside the child-level definitions, so the prompt can reason from the broad ESG branch to the final child assignment.

System. “You classify ESG disclosure sentences using the SASB hierarchy. Think first about broad parent meaning and then choose a child label.”

User template. “Sentence: “{text}”. Parent-to-child hierarchy: {hierarchy}. Child label definitions: {definitions}. Return valid JSON with keys parent_label, top1_label, top3_labels, confidence, rationale.”

Appendix B.2. Annotation Prompts (Claude Sonnet 4.6)

The annotation prompts keep Claude Sonnet 4.6 fixed as the LLM annotator and change only the information visible at decision time. Blind annotation exposes the sentence and the author-prepared child label definitions; candidate-visible annotation adds the released label and the validation-selected GPT-4o candidate, which turns the task into a bounded comparison across visible alternatives. The exact system prompts use the role phrase “ESG reviewer”, while the analysis treats the task as annotation because the output is a sentence-level category label. The prompt text uses the phrase “SASB child label definitions” for these author-prepared definitions.

Blind annotation template.

The blind annotation template requests an independent child label decision together with a parent label, confidence value, ambiguity flag, and short note.

System. “You are an ESG reviewer. Read the sentence and assign the best SASB child label and parent label. If the sentence is too ambiguous for a unique child label, mark `ambiguity_flag` as `ambiguous`.”

User template. “Sentence: “`{text}`”. SASB child label definitions: `{definitions}`. Return valid JSON with keys `child_label`, `parent_label`, `confidence`, `ambiguity_flag`, `note`.”

Candidate-visible template.

The candidate-visible template preserves the same sentence and author-prepared definition fields, then adds the released label, represented in the implementation by `original_label`, and the model candidate before asking for a verdict among *original*, *model*, *both*, and *other*. In the main analysis, `recommended_child` supplies the candidate-visible annotation label for retention, kappa, and profile aggregation, including cases where the verdict is *both*.

System. “You are an ESG reviewer comparing candidate labels for a sentence. Decide whether the original label, the model candidate, both, or neither is best supported.”

User template. “Sentence: “`{text}`”. Original label: `{original_label}`. Model candidate: `{model_candidate}`. SASB child label definitions: `{definitions}`. Return valid JSON with keys `verdict`, `recommended_child`, `recommended_parent`, `confidence`, `ambiguity_flag`, `note`. Allowed verdict values: `original`, `model`, `both`, `other`.”

References

1. IFRS Foundation. Understanding the SASB Standards, 2026. Guidance Page Describing the SASB Standards as Industry-Based Guidance Used Within the ISSB Framework. Available online: <https://www.ifrs.org/issued-standards/sasb-standards/understanding-sasb-standards/> (accessed on 4 May 2026).
2. Cahan, S.F.; Chen, L.; Wei, Y. Do sustainability standards improve the information environment? Evidence from the influence of SASB standards on disagreement among ESG rating agencies. *Meditari Account. Res.* **2025**, *ahead-of-print*. [CrossRef]
3. Junprung, E. SASB-Aligned ESG Sentences. Kaggle Datasets, 2023. License: Apache 2.0. Available online: <https://www.kaggle.com/datasets/edwardjunprung/sasb-aligned-esg-sentences> (accessed on 4 May 2026).
4. Korca, B.; Costa, E.; Bouten, L. Disentangling the concept of comparability in sustainability reporting. *Sustain. Account. Manag. Policy J.* **2023**, *14*, 815–851. [CrossRef]
5. Jørgensen, S.; Mjøs, A.; Pedersen, L.J.T. Sustainability reporting and approaches to materiality: Tensions and potential resolutions. *Sustain. Account. Manag. Policy J.* **2022**, *13*, 341–361. [CrossRef]
6. León, R.; Salesa, A. Is sustainability reporting disclosing what is relevant? Assessing materiality accuracy in the Spanish telecommunication industry. *Environ. Dev. Sustain.* **2024**, *26*, 21433–21460. [CrossRef]
7. Bochkay, K.; Brown, S.V.; Leone, A.J.; Tucker, J.W. Textual Analysis in Accounting: What’s Next? *Contemp. Account. Res.* **2023**, *40*, 765–805. [CrossRef]
8. Velte, P. Automated text analyses of sustainability & integrated reporting. A literature review of empirical-quantitative research. *J. Glob. Responsib.* **2023**, *14*, 530–566. [CrossRef]
9. Linhares Pontes, E.; Ben Jannet, M.; Moreno, J.G.; Doucet, A. Using Contextual Sentence Analysis Models to Recognize ESG Concepts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, Abu Dhabi, United Arab Emirates (Hybrid); Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 218–223. [CrossRef]
10. Schimanski, T.; Reding, A.; Reding, N.; Bingler, J.; Kraus, M.; Leippold, M. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Financ. Res. Lett.* **2024**, *61*, 104979. [CrossRef]
11. Zou, Y.; Shi, M.; Chen, Z.; Deng, Z.; Lei, Z.; Zeng, Z.; Yang, S.; Tong, H.; Xiao, L.; Zhou, W. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *J. Clean. Prod.* **2025**, *489*, 144572. [CrossRef]

12. Tseng, Y.M.; Chen, C.C.; Huang, H.H.; Chen, H.H. DynamicESG: A Dataset for Dynamically Unearthing ESG Ratings from News Articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, United Kingdom*; ACM: New York, NY, USA, 2023; pp. 5412–5416. [\[CrossRef\]](#)
13. He, C.; Zhou, X.; Wu, Y.; Yu, X.; Zhang, Y.; Zhang, L.; Wang, D.; Lyu, S.; Xu, H.; Wang, X.; et al. ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 14612–14653. [\[CrossRef\]](#)
14. Su, X.; Liu, T.; Pang, P.; Luo, Y.T.; Wong, D. How Can Large Language Models Drive Environmental Sustainability? A Systematic Scoping Review. *Sustainability* **2026**, *18*, 4327. [\[CrossRef\]](#)
15. Li, J.; Yang, Y.; Mao, C.; Pang, P.C.I.; Zhu, Q.; Xu, D.; Wang, Y. Revealing Patient Dissatisfaction with Health Care Resource Allocation in Multiple Dimensions Using Large Language Models and the International Classification of Diseases 11th Revision: Aspect-Based Sentiment Analysis. *J. Med. Internet Res.* **2025**, *27*, e66344. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Li, J.; Guo, J.; Pang, P.; Oliveira, H.G.; Ng, B.K.; Tan, T. Multi-task Specialized Expert Model for Hierarchical Aspect-based Sentiment Analysis in Consumer Healthcare. *Expert Syst. Appl.* **2026**, *331*, 133419. [\[CrossRef\]](#)
17. Junprung, E. SASB-ESG Category Classifier. Hugging Face Model Card, 2023. Available online: <https://huggingface.co/ejunprung/SASB-ESG-Category-Classifier> (accessed on 4 May 2026).
18. Adcock, R.; Collier, D. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *Am. Political Sci. Rev.* **2001**, *95*, 529–546. [\[CrossRef\]](#)
19. Birkenmaier, L.; Lechner, C.M.; Wagner, C. The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. *Commun. Methods Meas.* **2024**, *18*, 249–277. [\[CrossRef\]](#)
20. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3980–3990. [\[CrossRef\]](#)
21. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 6894–6910. [\[CrossRef\]](#)
22. Liu, K. SetCSE: Set Operations using Contrastive Learning of Sentence Embeddings. In *Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024*.
23. Fang, Q.; Nguyen, D.; Oberski, D.L. Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Sci.* **2022**, *11*, 39. [\[CrossRef\]](#)
24. Hayes, A.F.; Krippendorff, K. Answering the Call for a Standard Reliability Measure for Coding Data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [\[CrossRef\]](#)
25. Dawid, A.P.; Skene, A.M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Appl. Stat.* **1979**, *28*, 20. [\[CrossRef\]](#)
26. Aroyo, L.; Welty, C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Mag.* **2015**, *36*, 15–24. [\[CrossRef\]](#)
27. Fornaciari, T.; Uma, A.; Paun, S.; Plank, B.; Hovy, D.; Poesio, M. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2591–2597. [\[CrossRef\]](#)
28. Frenay, B.; Verleysen, M. Classification in the Presence of Label Noise: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 845–869. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; Liu, H. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 930–957. [\[CrossRef\]](#)
30. Feuston, J.L.; Brubaker, J.R. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 1–25. [\[CrossRef\]](#)
31. Eschrich, J.; Serman, S. A Framework For Discussing LLMs as Tools for Qualitative Analysis. Version Number: 1. *arXiv* **2024**, arXiv:2407.11198. [\[CrossRef\]](#)
32. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Version Number: 4. *arXiv* **2023**, arXiv:2306.05685. [\[CrossRef\]](#)
33. Berg, F.; Kölbel, J.F.; Rigobon, R. Aggregate Confusion: The Divergence of ESG Ratings. *Rev. Financ.* **2022**, *26*, 1315–1344. [\[CrossRef\]](#)
34. Chatterji, A.K.; Durand, R.; Levine, D.I.; Touboul, S. Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strateg. Manag. J.* **2016**, *37*, 1597–1614. [\[CrossRef\]](#)

35. Henry, E.; Leone, A.J. Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. *Account. Rev.* **2016**, *91*, 153–178. [[CrossRef](#)]
36. He, H.; Garcia, E. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [[CrossRef](#)]
37. Silla, C.N.; Freitas, A.A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* **2011**, *22*, 31–72. [[CrossRef](#)]
38. Kosmopoulos, A.; Partalas, I.; Gaussier, E.; Paliouras, G.; Androutsopoulos, I. Evaluation measures for hierarchical classification: A unified view and novel approaches. *Data Min. Knowl. Discov.* **2015**, *29*, 820–865. [[CrossRef](#)]
39. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [[CrossRef](#)]
40. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
41. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [[CrossRef](#)]
42. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Iii, H.D.; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92. [[CrossRef](#)]
43. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2019; pp. 220–229. [[CrossRef](#)]
44. Wu, S.; Irsoy, O.; Lu, S.; Dabrovolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; Mann, G. BloombergGPT: A Large Language Model for Finance. *arXiv* **2023**, arXiv:2303.17564.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.