

Article

LLM-Assisted and Rule-Based Assessment of ESG Disclosure Quality and Its Association with External ESG Ratings: Exploratory Evidence from S&P 500 Energy Firms

Hae Sun Jung ¹, Shaopeng Che ²  and Haein Lee ^{3,*} 

¹ Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul 03063, Republic of Korea; jestiriel@g.skku.edu

² School of Humanities and Social Science, Chang'an University, Xi'an 710064, China; cheshao@chd.edu.cn

³ School of Interdisciplinary Studies, Dongguk University, Seoul 04620, Republic of Korea

* Correspondence: lhi00034@dongguk.edu; Tel.: +82-2-2260-8930

Abstract

Environmental, social, and governance (ESG) disclosure has become an important source of information for external stakeholders. As sustainability reporting has expanded, distinguishing disclosure quantity from disclosure quality has become increasingly important. This study examines how sustainability disclosure quality is associated with external ESG evaluation outcomes among Standard & Poor's (S&P) 500 Energy Sector firms. ESG-related claims were identified and classified from sustainability reports using large language model (LLM)-assisted structured content analysis. Based on the resulting corpus, three disclosure quality indicators were constructed: the Quantitative Evidence Ratio (QER), the Target Accountability Ratio (TAR), and the Reporting Infrastructure Score (RIS). These indicators were integrated into a composite Disclosure Quality Index (DQI) and examined in relation to S&P Global ESG Scores using Spearman's rank correlation. The results indicated limited positive associations for the individual indicators, whereas the composite DQI showed a more pronounced positive relationship. Disclosure-rating divergence patterns were also observed, indicating that relatively favorable disclosure quality positions do not consistently correspond to higher ESG Score rankings. Overall, the findings suggested that sustainability disclosure quality may be multidimensional and that its association with external ESG evaluation outcomes became more apparent when disclosure characteristics were considered in combination. However, because the analysis is restricted to a small sample of S&P 500 Energy Sector firms, the findings should be interpreted as exploratory sector-specific evidence with limited generalizability.

Keywords: ESG disclosure quality; sustainability reporting; ESG evaluation; LLM-assisted analysis



Academic Editors: Michalis Skordoulis, Rohail Hassan and Costinela Fortea

Received: 28 May 2026

Revised: 26 June 2026

Accepted: 4 July 2026

Published: 6 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

Environmental, social, and governance (ESG) disclosure has become an important communication mechanism through which firms explain their environmental-, social-, and governance-related activities to investors, regulators, rating agencies, and other stakeholders [1]. This issue is especially important in the Energy Sector, where greenhouse gas emissions, methane leakage, flaring, operational safety, water management, and governance risks are simultaneously emphasized [2]. In this context, sustainability reports

serve not only as communication documents, but also as important inputs for external ESG evaluation and investment decision-making [3].

At the same time, extensive ESG disclosure does not necessarily imply high-quality or highly accountable disclosure. Prior research has shown that firms may provide broad sustainability narratives without offering sufficiently measurable, verifiable, or accountability-oriented disclosure structures [4–6]. Existing ESG disclosure studies have therefore distinguished between more verifiable forms of disclosure and more symbolic or narrative-oriented forms of communication while also emphasizing that disclosure quality depends on features such as measurability, comparability, reliability, and informational usefulness rather than disclosure volume alone [6]. Related research on ESG ratings has further shown that external ESG evaluation outcomes often diverge across rating agencies because of differences in scope, measurement approaches, weighting structures, and information sources [7]. These findings suggest that sustainability disclosure quality and external ESG evaluation outcomes may be related, but they should not be treated as directly equivalent.

To address this gap, this study investigates how sustainability disclosure quality is associated with external ESG evaluation outcomes among S&P 500 Energy Sector firms. Rather than treating ESG scores as absolute representations of ESG performance, this study focuses on whether firms with more measurable, target-linked, and structurally supported sustainability disclosures exhibit stronger correspondence with external ESG evaluations. Specifically, this study adopts a two-stage analytical design. First, sustainability reports are processed through a Python-based text extraction and sentence-level preprocessing pipeline, followed by LLM-assisted ESG claim identification and E/S/G classification. Second, disclosure quality indicators are constructed through rule-based feature extraction and examined in relation to external ESG evaluation outcomes. Within this design, the role of the LLM is limited to structured content identification based on a predefined codebook, while the disclosure quality measures themselves are operationalized through observable textual features. This approach supports methodological transparency and interpretability while maintaining scalability across firm-level reports.

Based on the classified ESG disclosure corpus, this study constructs three disclosure quality indicators. The Quantitative Evidence Ratio (QER) measures the extent to which ESG disclosures contain measurable quantitative evidence such as numerical metrics, key performance indicator references, target years, and baseline information. The Target Accountability Ratio (TAR) captures whether ESG targets are linked to accountability-related elements including baseline references, progress indicators, and temporal commitments. The Reporting Infrastructure Score (RIS) evaluates whether sustainability reports contain structured reporting infrastructures such as framework disclosure, ESG performance tables, assurance practices, reporting boundaries, and materiality assessment procedures. These indicators are then integrated into a composite Disclosure Quality Index (DQI) in order to examine the broader structure of sustainability disclosure quality.

This study contributes to the literature in three ways. First, it extends prior ESG disclosure quality research by offering a structured operationalization of measurable, accountability-oriented, and infrastructure-related disclosure characteristics at scale. Second, it proposes a hybrid methodological framework that combines LLM-assisted ESG claim identification with rule-based construction of disclosure quality indicators. Third, it provides exploratory empirical evidence on how multidimensional disclosure quality is associated with external ESG evaluation outcomes and where disclosure–rating divergence patterns emerge across firms. Accordingly, this study addresses the following research questions:

- RQ1. How can ESG disclosure quality be operationalized through quantitative evidence, target accountability, and reporting infrastructure in firm-level sustainability reports?

- RQ2. To what extent are these disclosure quality indicators associated with external ESG evaluation outcomes among S&P 500 Energy Sector firms?
- RQ3. What disclosure–rating divergence patterns can be identified between sustainability disclosure quality and external ESG evaluations?

The principal findings indicate that the composite DQI showed stronger positive associations with the external ESG evaluation outcomes examined in this study, namely the S&P Global ESG Score and the S&P Global Corporate Sustainability Assessment (CSA) Score, than the individual disclosure quality indicators considered separately. The results also suggest that sustainability disclosure quality and external ESG evaluation outcomes were related but not fully aligned, with several firms showing notable disclosure–rating divergence patterns. These findings support the view that multidimensional disclosure quality may provide a useful analytical basis for examining sustainability report structure and its relationship with external ESG assessment outcomes.

The remainder of this paper is organized as follows: Section 2 reviews prior literature on ESG disclosure, disclosure quality, ESG rating divergence, and LLM-assisted content analysis. Section 3 presents the proposed analytical framework, including data collection procedures, LLM-assisted ESG claim identification, rule-based feature extraction, and the construction of disclosure quality indicators. Section 4 reports the empirical findings regarding disclosure quality indicators, ESG evaluation relationships, and disclosure–rating divergence patterns. Section 5 discusses the implications of the findings. Section 6 concludes this study.

2. Literature Review

2.1. ESG Disclosure Quality as Strategic Corporate Communication

ESG disclosure has increasingly been recognized not only as an information disclosure mechanism, but also as a strategic form of corporate communication directed toward investors, regulators, rating agencies, and broader stakeholder groups [1]. Prior social and environmental accounting research has argued that sustainability reporting plays an important role in shaping corporate legitimacy, stakeholder perception, and accountability structures [8–10]. From this perspective, sustainability reports function not merely as repositories of environmental and social information, but also as organizational narratives through which firms communicate how they manage ESG-related responsibilities and risks [11].

This communication function is particularly visible in environmentally sensitive industries such as the Energy Sector, where greenhouse gas emissions, methane leakage, operational safety, water management, and governance risks are closely scrutinized by external audiences [12,13]. In such contexts, sustainability reporting has evolved from relatively broad symbolic disclosure toward increasingly structured forms of ESG communication incorporating reporting frameworks, assurance statements, key performance indicator (KPI)-oriented metrics, governance disclosures, and target-related information. As a result, the informational value of ESG reporting cannot be adequately understood through disclosure quantity alone [14].

A central issue in prior ESG disclosure research concerns the distinction between disclosure quantity and disclosure quality. Clarkson et al. [6] distinguished between more verifiable forms of quantitative environmental disclosure and broader narrative-oriented disclosure practices. Hard disclosures generally include measurable indicators, numerical performance metrics, quantified targets, and externally verifiable reporting elements, whereas soft disclosures rely more heavily on symbolic language, broad commitments, or declarative sustainability rhetoric. Subsequent studies extended this distinction beyond environmental reporting and emphasized that disclosure quality depends not only on the

amount of information disclosed, but also on balance, comparability, precision, reliability, and informational usefulness [11,15].

This distinction is especially important in ESG reporting contexts where firms may disclose extensive sustainability-related narratives without providing sufficiently measurable or accountable evidence structures. Prior research has therefore increasingly emphasized the role of quantitative indicators, target accountability, external assurance, and reporting transparency in evaluating ESG disclosure quality [16,17]. Related discussions of symbolic disclosure and selective emphasis have similarly suggested that extensive or favorable ESG narratives do not necessarily imply substantively informative disclosure [18–20]. In this sense, broad keyword frequency, positive tone, or general ESG rhetoric may overstate the informational quality of sustainability disclosure when measurable evidence and accountability structures remain limited.

Collectively, these studies suggest that ESG disclosure quality is better understood as a multidimensional construct than as a simple function of disclosure volume or thematic intensity. Building on this literature, the present study operationalizes disclosure quality through three observable and comparatively reproducible dimensions: quantitative evidence, accountability-oriented target disclosure, and reporting infrastructure. This framework does not treat ESG disclosure quality as a purely interpretive or symbolic construct. Instead, it focuses on whether firms provide measurable, traceable, and structurally supported disclosure characteristics that can be identified systematically across reports.

2.2. External ESG Evaluation and Rating Divergence

External ESG ratings are widely used by investors, regulators, and financial institutions to evaluate corporate sustainability performance [1]. At the same time, prior research has consistently shown that ESG evaluation outcomes often diverge across rating agencies because of differences in scope, measurement approaches, weighting structures, and methodological assumptions [21–23]. Berg et al. [21] showed that ESG rating divergence is driven by variation in measurement, scope, and weighting methodologies across providers. Christensen et al. [24] similarly argued that ESG ratings often capture heterogeneous and only partially overlapping dimensions of corporate sustainability performance rather than a single unified construct. These findings complicate the interpretation of ESG scores as direct or objective representations of ESG performance.

An important implication of this literature is that external ESG evaluations are not determined solely by sustainability report disclosures. Rating methodologies frequently incorporate additional information sources such as media analysis, controversy assessments, stakeholder information, peer-relative benchmarking, external datasets, and proprietary modeling approaches [21,24]. Accordingly, sustainability disclosure quality and external ESG evaluation outcomes may be related without showing direct one-to-one correspondence. A firm may provide relatively structured and accountable sustainability disclosure while receiving only moderate external ESG evaluations, and the reverse pattern may also occur.

This issue is particularly relevant for the present study because it does not treat external ESG scores as direct measures of disclosure quality. Instead, external ESG scores are used as benchmark outcomes against which the alignment and divergence of disclosure quality can be examined. From this perspective, the relationship between disclosure quality and external ESG evaluation should be interpreted as an empirical association between two related but non-identical constructs. This framing is consistent with prior ESG rating divergence research and provides the conceptual basis for examining both correlation patterns and disclosure–rating divergence across firms while recognizing that external

ESG assessments also reflect broader methodological and informational inputs beyond sustainability report content alone.

2.3. Methodological Positioning of LLM-Assisted Structured Content Analysis

Recent advances in large language models (LLMs) have increased interest in LLM-assisted text classification and annotation tasks. Gilardi et al. [25] showed that GPT-based annotation approaches can perform competitively relative to crowd-sourced annotation in several text classification settings. Other recent studies have similarly explored the use of LLMs for document classification, semantic labeling, and structured content analysis tasks [26,27]. These developments suggest that LLMs can support large-scale qualitative coding workflows that would otherwise be difficult to implement consistently across extensive document collections.

At the same time, methodological concerns remain regarding interpretive subjectivity, reproducibility, hallucination risk, and transparency in LLM-assisted research workflows [28,29]. These concerns are particularly important in ESG analysis, where direct model-based judgments about whether disclosures are good, credible, substantive, or responsible may introduce substantial ambiguity. In ESG disclosure research, LLMs may therefore be better positioned as tools for supporting the scalable implementation of predefined analytical frameworks, rather than as standalone evaluators of disclosure quality.

In the present study, the LLM is used for ESG claim identification and E/S/G classification based on a predefined codebook. Substantive disclosure quality indicators are then constructed through rule-based feature extraction grounded in observable textual characteristics, including numerical evidence, temporal references, target structures, KPI-related disclosures, assurance statements, and reporting framework references.

This hybrid design combines the scalability of LLM-assisted content identification with the transparency and reproducibility of rule-based operationalization. This study therefore contributes methodologically by offering a structured approach for operationalizing established disclosure quality concepts across a systematically collected set of sustainability reports, and by examining whether multidimensional disclosure quality characteristics are associated with external ESG evaluation outcomes.

3. Materials and Methods

This study employed a multi-stage ESG disclosure analysis framework combining LLM-assisted ESG claim identification, rule-based feature extraction, and firm-level disclosure quality measurement.

First, sustainability disclosure documents and external ESG benchmark data were collected from S&P 500 Energy Sector firms. Second, the collected documents were processed through a sentence-level text extraction and preprocessing pipeline. Third, ESG-related disclosure claims were identified and classified using an LLM-assisted content coding procedure. Subsequently, rule-based feature extraction methods were applied to construct disclosure quality indicators, including QER, TAR, RIS, and the composite Disclosure Quality Index (DQI). Finally, the relationships between disclosure quality indicators and external ESG evaluation outcomes were examined using correlation analysis and disclosure-rating divergence analysis. Figure 1 presents the overall research framework and analytical procedure employed in this study.

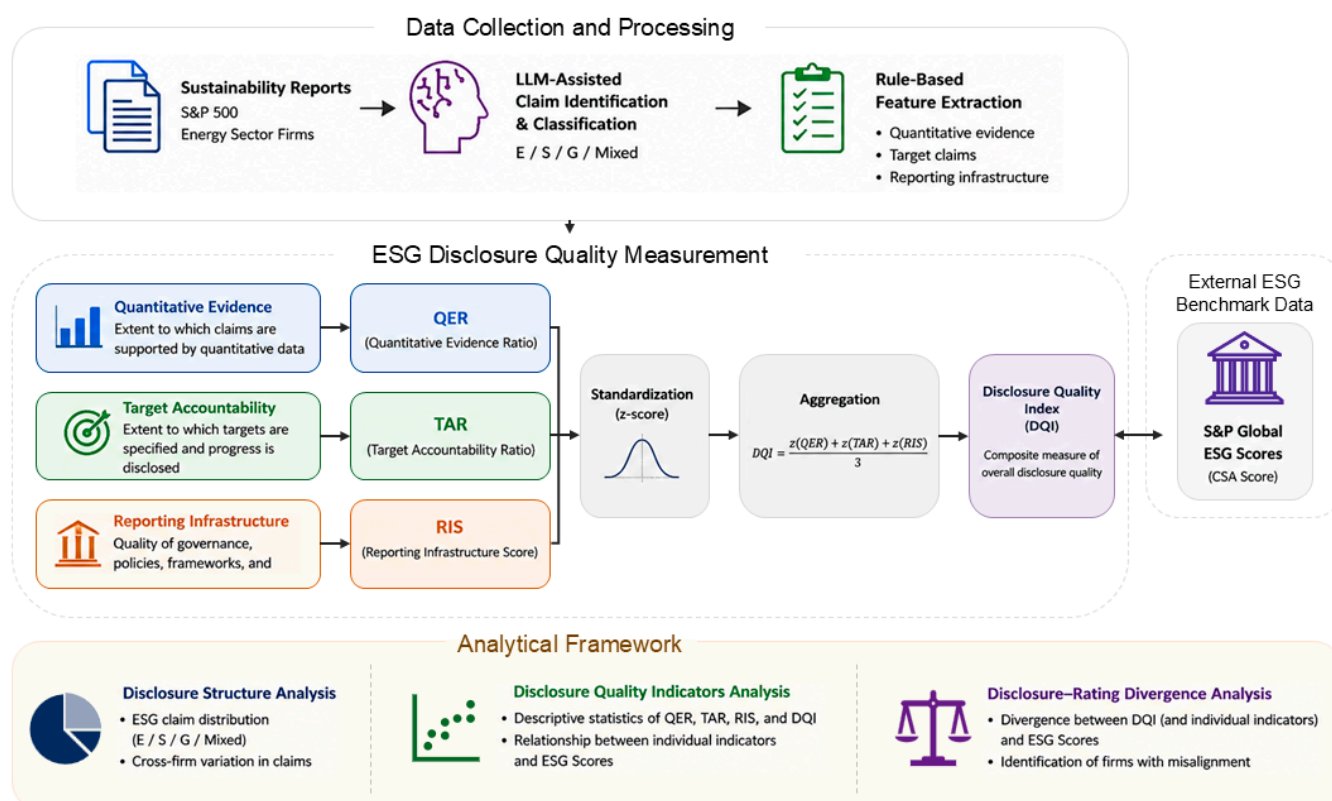


Figure 1. Overall research framework and analytical procedure.

3.1. Sample Selection and Benchmark Data Collection

The initial sample was constructed from S&P 500 firms classified under the Energy sector according to the Global Industry Classification Standard (GICS) framework developed by Morgan Stanley Capital International (MSCI) and S&P Dow Jones Indices [30]. The constituent list was cross-checked against publicly available S&P 500 information to ensure data integrity.

Based on the filtered firm list, the most recent sustainability disclosure document for each firm was manually collected from official corporate websites. The primary analysis materials consisted of standalone Sustainability Reports, ESG Reports, Corporate Sustainability Reports, or equivalent sustainability disclosure documents. Condensed highlights versions, summary reports, ongoing reports, draft documents, and legal review versions were excluded to ensure document consistency and comparability. Additionally, reports with severe text extraction or encoding problems that prevented stable sentence-level analysis were excluded from the final sample. Accordingly, among the GICS-classified S&P 500 Energy firms, only firms with comparable standalone sustainability disclosure documents suitable for stable sentence-level analysis were retained, resulting in a final sample of 17 Energy Sector firms.

Following the document collection process, external benchmark variables were additionally collected from the S&P Global ESG database. These variables included the S&P Global ESG Score, CSA Score, industry classification, Score Under Review status, and Last Updated information. The S&P Global ESG database was selected because it provides publicly accessible and relatively up-to-date ESG benchmark information across firms, thereby improving data consistency and comparability.

To improve temporal transparency, the report year and ESG benchmark update date were documented for all firms. While some companies publicly disclose precise report publication dates, such information was not consistently available across the entire sample.

Therefore, report year was used as the primary temporal reference for sustainability disclosures. In addition, historical S&P Global ESG Score values corresponding exactly to each report year were not publicly available to the researchers. As a result, perfect temporal synchronization between the sustainability reports and external benchmark data could not be achieved. The S&P Global ESG Score and CSA Score were therefore interpreted as temporally proximate external benchmark outcomes rather than perfectly synchronized contemporaneous measures. Detailed company information and benchmark variables are presented in Table 1.

Table 1. Final sample firms and ESG benchmark information.

Company (Ticker)	Report Year	CSA Score	ESG Score	Score Updated
APA Corporation (APA)	2024	35	43	July 2025
Baker Hughes Company (BKR)	2025	62	63	October 2025
ConocoPhillips (COP)	2024	53	54	February 2025
Devon Energy Corporation (DVN)	2025	41	45	January 2026
Diamondback Energy, Inc. (FANG)	2025	39	43	October 2025
EOG Resources, Inc. (EOG)	2024	34	40	December 2025
EQT Corporation (EQT)	2024	43	45	January 2026
Expand Energy Corporation (EXE)	2024	38	43	October 2025
Exxon Mobil Corporation (XOM)	2025	29	34	March 2026
Kinder Morgan, Inc. (KMI)	2024	45	47	August 2025
Marathon Petroleum Corporation (MPC)	2024	51	51	October 2025
Occidental Petroleum Corporation (OXY)	2025	48	50	February 2026
ONEOK, Inc. (OKE)	2025	56	57	December 2025
Phillips 66 (PSX)	2025	40	41	August 2025
Targa Resources Corp. (TRGP)	2024	34	39	December 2025
Valero Energy Corporation (VLO)	2024	37	41	April 2026
The Williams Companies, Inc. (WMB)	2024	68	68	January 2026

The S&P Global ESG Score and CSA Score were used as external benchmark outcomes rather than direct measures of ESG disclosure quality. The S&P Global ESG Score reflects a broader ESG evaluation process that incorporates company disclosures, Media and Stakeholder Analysis (MSA), modeling approaches, and company engagement through the CSA. The CSA is an annual industry-specific sustainability assessment that provides an important basis for the S&P Global ESG Score. Accordingly, these benchmark indicators reflect broader ESG evaluation processes than sustainability report disclosures alone.

3.2. Text Extraction and Preprocessing

The sustainability disclosure documents selected through the filtering procedures were processed using a Python version 3.12.13. Text was extracted from Portable Document Format (PDF) files using the PyMuPDF version 1.28.0, and the extracted text was subsequently segmented into sentence-level observations using the Natural Language Toolkit (NLTK) version 3.9.1 sentence tokenizer [31]. The resulting dataset retained metadata including company file name, PDF path, page number, sentence identifier, and sentence text.

Afterwards, several preprocessing procedures were applied to improve textual consistency and reduce noise generated during PDF extraction. First, text normalization procedures were conducted to remove soft hyphens, corrupted PDF characters, excessive whitespace, and unnecessary spacing artifacts introduced during PDF conversion. Second, short text fragments containing fewer than 20 characters were excluded to reduce fragmented or semantically incomplete observations. Third, boilerplate or non-substantive text patterns including table of contents entries, forward-looking statements, cautionary state-

ments, copyright notices, intentionally blank page markers, and generic navigation phrases were removed to improve analytical relevance and reduce repetitive disclosure artifacts.

3.3. LLM-Assisted ESG Content Coding

The preprocessed sentence-level dataset was analyzed using an LLM-assisted ESG content coding procedure. ESG claim identification was conducted using the GPT-5.2 model with a temperature setting of 0 to improve output consistency across firms and processing batches. Sentences were processed in batches, and a standardized coding prompt and codebook were applied across the entire dataset (see Appendix A).

3.3.1. ESG Claim Identification and ESG Dimension Classification

For each sentence, the LLM performed two primary classification tasks: (1) identifying whether the sentence constituted an ESG-related disclosure claim and (2) classifying ESG claim sentences into Environmental, Social, Governance, or Mixed categories. ESG claims were operationalized as disclosures related to environmental performance, workforce and community issues, governance structures, risk management, reporting practices, sustainability targets, corporate policies, or accountability mechanisms. The coding categories were broadly informed by prior ESG disclosure studies and commonly used ESG reporting frameworks [23].

Environmental (E) disclosures included topics such as greenhouse gas emissions, climate change, biodiversity, water management, energy transition, resource efficiency, pollution prevention, and environmental stewardship. Social (S) disclosures included workforce management, employee health and safety, diversity and inclusion, training and development, human rights, labor practices, community engagement, and social impact activities. Governance (G) disclosures included board oversight, ethics and compliance, risk management, cybersecurity, reporting governance, transparency, assurance practices, and regulatory compliance. Mixed disclosures were assigned when multiple ESG dimensions were simultaneously reflected within a single disclosure statement.

In addition, auxiliary claim-type labels were generated to support manual inspection and coding consistency assessment during the pilot validation stage. These auxiliary labels included categories such as emissions performance, climate targets, workforce and human capital, safety and health, governance and compliance, reporting framework and assurance, supply chain governance, and generic ESG statements.

To improve output stability and parsing reliability, the LLM responses were constrained to a structured output format. The ESG coding outputs were subsequently merged with the original sentence-level dataset using sentence identifiers, after which only sentences classified as ESG claims were retained.

3.3.2. Validation of LLM-Based Coding

To evaluate the reliability of the LLM-assisted coding procedure, a random sample of 200 sentences was independently reviewed by two authors. The validation sample included both ESG and non-ESG candidate sentences. Inter-coder reliability between the two reviewers was assessed using Cohen's Kappa and demonstrated substantial agreement ($\kappa = 0.83$) [32].

Disagreements were subsequently resolved through discussion and consensus, and the consensus labels were used as the reference standard for validation. Compared with the consensus labels, GPT-5.2 achieved an agreement rate of 90.7% for ESG claim identification. Furthermore, among sentences classified as ESG claims, agreement for ESG dimension classification (E, S, G, or Mixed) reached 93.68%. Among the validation samples, disagreements primarily occurred in the distinction between Governance and Mixed categories, whereas ESG claim identification generally exhibited high consistency.

These results suggest that the LLM-assisted coding procedure was broadly consistent with human evaluation and was sufficiently reliable for the purposes of the subsequent analysis. Representative classification examples are provided in Appendix B.

3.4. Rule-Based Feature Extraction and Measurement

Following the LLM-assisted ESG claim identification, the retained ESG claim sentences were processed using rule-based regular expression patterns (regex) to identify observable disclosure features. This step extracted explicit textual features required for disclosure quality measurement, including quantitative evidence, target-related disclosures, accountability-related target features, and reporting framework references. The regex patterns were iteratively refined through manual inspection during the pilot coding stage, and the full set of rules is summarized in Appendix C. This post-LLM rule-based design supports transparency, reproducibility, and auditability because feature extraction is based on predefined and inspectable patterns. However, it also involves a trade-off: regex-based extraction has limited semantic flexibility and may miss valid ESG disclosures that are expressed in nuanced, indirect, or unconventional language.

Based on the extracted disclosure features, three ESG disclosure quality indicators were subsequently constructed: QER, TAR, and RIS. These indicators were designed to capture complementary dimensions of ESG disclosure quality, including quantitative evidence, accountability-oriented target disclosure, and reporting infrastructure.

3.4.1. Quantitative Evidence Ratio (QER)

QER measures the proportion of ESG claim sentences containing explicit quantitative disclosure evidence. This concept was developed based on prior ESG disclosure studies distinguishing between verifiable hard disclosure and more narrative-oriented soft disclosure [6,11]. The QER is calculated as follows:

$$QER_i = \frac{\text{Quantitative ESG Claims}_i}{\text{Total ESG claims}_i} \quad (1)$$

where *Quantitative ESG Claims_i* denotes the number of ESG claim sentences containing explicit quantitative evidence for firm *i*, and *Total ESG Claims_i* denotes the total number of ESG claim sentences for firm *i*. Quantitative evidence was identified using the rule-based feature extraction procedure described in Section 3.4.

3.4.2. Target Accountability Ratio (TAR)

TAR measures the proportion of target-related ESG claims containing accountability-related disclosure elements. This concept was informed by prior studies emphasizing accountability, traceability, and the distinction between symbolic sustainability communication and substantively supported disclosure practices [4,11]. The TAR is calculated as follows:

$$TAR_i = \frac{\text{Accountability-Linked Target Claims}_i}{\text{Total ESG Target Claims}_i} \quad (2)$$

where *Accountability-Linked Target Claims_i* denotes the number of ESG target-related claims containing accountability-related disclosure elements for firm *i*, and *Total ESG Target Claims_i* denotes the total number of ESG target-related claims for firm *i*.

ESG target-related claims were identified using rule-based target detection patterns, while accountability-related target disclosures were identified using combinations of temporal, baseline-related, and progress-oriented disclosure features.

3.4.3. Reporting Infrastructure Score (RIS)

RIS measures the extent to which a sustainability report contains formalized disclosure infrastructure and reporting governance mechanisms. RIS was constructed through report-level binary coding because reporting infrastructure is primarily reflected at the document level rather than within isolated sentence-level disclosures. The RIS framework was conceptually informed by prior sustainability reporting and assurance literature emphasizing reporting structure, external verification, and disclosure governance mechanisms [33,34].

RIS was operationalized using five disclosure infrastructure components: reporting framework disclosure, ESG data table disclosure, assurance or verification disclosure, reporting boundary disclosure, and materiality assessment disclosure. Each component was coded as either 0 or 1 based on predefined criteria. The RIS is calculated as follows:

$$RIS_i = \frac{Framework_i + Table_i + Assurance_i + Boundary_i + Materiality_i}{5} \quad (3)$$

where $Framework_i$, $Table_i$, $Assurance_i$, $Boundary_i$, and $Materiality_i$ denote the binary coding results for each disclosure infrastructure component for firm i . The operational coding criteria for each RIS component are summarized in Table 2.

Table 2. Operational criteria for RIS.

Component	Coding Criteria
Framework	Global Reporting Initiative (GRI), Sustainability Accounting Standards Board (SASB), Task Force on Climate-related Financial Disclosures (TCFD), International Sustainability Standards Board (ISSB), and industry-specific reporting frameworks
Table	Presence of structured ESG performance tables or metric summary sections
Assurance/verification	Presence of third-party assurance, limited assurance, independent verification, or issue-specific external verification
Boundary	Explicit disclosure of reporting boundary, operational scope, exclusions, or methodology
Materiality	Disclosure of materiality assessment or sustainability topic prioritization process

3.4.4. Disclosure Quality Index (DQI)

Subsequently, QER, TAR, and RIS were combined to construct a composite DQI. Building on prior ESG disclosure quality studies arguing that disclosure quality should not be reduced to a single textual or quantitative characteristic [6,11], this study conceptualizes ESG disclosure quality as a multidimensional construct composed of quantitative evidence, accountability-oriented target disclosure, and reporting infrastructure.

Consistent with this multidimensional conceptualization, the three indicators were treated as parallel components of disclosure quality rather than as hierarchical predictors. Equal weighting was applied as a transparent baseline aggregation strategy to preserve conceptual balance among the three dimensions and to avoid assigning unsupported priority to any single component in the absence of a prior theoretical or empirical basis for differential weighting. This weighting scheme should not be interpreted as assuming that QER, TAR, and RIS exert identical causal effects on disclosure quality or external ESG evaluations. Rather, it reflects the exploratory purpose of this study and provides a neutral composite measure for examining whether measurable evidence, target accountability, and

reporting infrastructure operate jointly as a broader disclosure quality profile. Accordingly, the DQI was constructed as the average of the standardized z-scores of QER, TAR, and RIS:

$$DQI_i = \frac{z(QER_i) + z(TAR_i) + z(RIS_i)}{3} \quad (4)$$

where QER_i , TAR_i and RIS_i represent the corresponding metric values for firm i , and $z(\cdot)$ denotes z-score standardization. QER, TAR, and RIS were measured on different scales and exhibited different value ranges across firms, each indicator was standardized prior to aggregation. This standardization allows the three components to be compared on a common scale and prevents differences in measurement range from disproportionately affecting the composite index. As a result of this z-score standardization, DQI values can take either positive or negative values. Negative DQI values indicate below-average disclosure quality relative to the sample mean, whereas positive values indicate above-average disclosure quality.

The DQI was used as a composite indicator for comparing overall ESG disclosure quality across firms, while the individual results for QER, TAR, and RIS were also analyzed separately to preserve interpretability across different disclosure quality dimensions.

3.5. Correlation and Divergence Analysis

The relationships between S&P Global ESG Scores and the proposed disclosure quality indicators (QER, TAR, RIS, and DQI) were examined using Spearman's rank correlation analysis. Because the sample size was limited and ESG Scores largely reflect relative within-industry evaluation, Spearman's correlation was considered more appropriate than Pearson's correlation for evaluating rank-based relationships under potentially non-normal conditions [35].

In addition to the correlation analysis, disclosure-rating divergence was analyzed by comparing firm-level DQI rankings with S&P Global ESG Score rankings:

$$Divergence_i = Rank(DQI_i) - Rank(ESG\ Score_i) \quad (5)$$

where $Rank(DQI_i)$ denotes the disclosure quality ranking of firm i based on DQI, and $Rank(ESG\ Score_i)$ denotes the corresponding ranking based on S&P Global ESG Score.

Positive divergence values indicate firms with relatively stronger disclosure quality relative to external ESG ratings, whereas negative divergence values indicate firms with comparatively stronger ESG ratings relative to disclosure quality. These divergence patterns were examined to evaluate whether sustainability disclosure quality and external ESG evaluation outcomes exhibited systematic alignment or relative discrepancy across firms.

4. Results

4.1. Overview of the Sustainability Disclosure Dataset

Table 3 summarizes the sustainability disclosure characteristics of the sampled Energy Sector firms. The statistics include total extracted sentence counts, ESG claim sentence counts, and the distribution of E, S, G, and Mixed disclosure categories identified through the LLM-assisted coding procedure.

Substantial variation was observed across firms in both report volume and ESG disclosure composition. Total extracted sentence counts ranged from 761 (Diamondback Energy) to 2712 (Williams Companies), while ESG claim sentence counts ranged from 635 to 2181. Firms such as Williams Companies, Baker Hughes, Kinder Morgan, and APA Corporation produced comparatively extensive sustainability disclosure reports, whereas Diamondback Energy and Phillips 66 relied on relatively more condensed reporting structures.

Table 3. Descriptive statistics of the sustainability disclosure dataset.

Ticker	Total Extracted Sentences	ESG Claim Sentences	E	S	G	Mixed
APA	2222	1797	538	570	283	406
BKR	2501	2108	700	561	350	497
COP	1424	1181	519	268	119	275
DVN	1752	1473	415	457	216	385
FANG	761	635	228	184	96	127
EOG	1242	994	438	205	137	214
EQT	2015	1624	538	468	225	393
EXE	1239	981	193	339	147	302
XOM	1486	1220	304	414	182	320
KMI	2227	1739	751	432	227	329
MPC	1128	986	275	309	147	255
OXY	1479	1214	439	315	150	310
OKE	1422	1099	279	330	224	266
PSX	824	643	194	203	105	141
TRGP	1050	881	225	267	155	234
VLO	1379	1063	327	390	126	220
WMB	2712	2181	583	679	383	536

Environmental disclosures generally constituted the largest category across firms, reflecting the Energy Sector’s strong emphasis on emissions management, climate targets, operational efficiency, methane reduction, and environmental reporting. In addition, social disclosures also represented a substantial proportion of ESG claim sentences in several firms, particularly in relation to workforce safety, employee development, contractor management, and community engagement. Governance disclosures were comparatively smaller in volume but consistently appeared across firms in areas such as compliance, ethics, reporting governance, cybersecurity, and assurance practices.

Mixed-category disclosures were also frequently observed. Firms such as Williams Companies, Baker Hughes, APA Corporation, and EQT Corporation exhibited relatively large numbers of Mixed disclosures, indicating that ESG-related discussions were often interconnected across Environmental, Social, and Governance dimensions rather than presented as strictly isolated categories.

Overall, the descriptive statistics indicate substantial heterogeneity in sustainability disclosure structure, reporting emphasis, and ESG communication strategies across the sampled Energy Sector firms.

4.2. ESG Disclosure Quality Indicators Analysis

Table 4 presents the firm-level ESG disclosure quality indicators derived from the proposed analytical framework. In addition to the three component indicators, it also reports the numbers of quantitative evidence disclosures, target-related ESG claims, and accountability-linked target claims identified through the rule-based feature extraction procedure. The inclusion of DQI alongside QER, TAR, and RIS allows for comparison of both the multidimensional structure of disclosure quality and each firm’s overall relative position within the sample.

Substantial variation was observed across firms in all disclosure quality indicators. QER values ranged from 0.194 for ExxonMobil to 0.378 for Kinder Morgan, indicating variation in the extent to which firms supported ESG-related statements with measurable quantitative evidence. Kinder Morgan, EQT Corporation, Valero Energy, and Occidental Petroleum showed relatively high QER values, whereas ExxonMobil, Expand Energy,

and EOG Resources showed lower QER values within the sample, despite maintaining substantial report volume in some cases.

Table 4. Firm-level ESG disclosure quality indicators.

Ticker	Quantitative Evidence	Target Claims	Accountable Targets	QER	TAR	RIS	DQI
APA	528	343	58	0.293	0.169	0.4	−0.803
BKR	629	444	91	0.298	0.205	1.0	0.697
COP	363	241	52	0.307	0.216	0.8	0.492
DVN	365	304	52	0.247	0.171	1.0	0.013
FANG	163	124	24	0.257	0.194	0.8	−0.071
EOG	244	181	28	0.245	0.155	0.8	−0.527
EQT	555	305	56	0.341	0.184	0.8	0.426
EXE	198	139	18	0.202	0.129	1.0	−0.705
XOM	237	184	19	0.194	0.103	1.0	−1.009
KMI	659	293	53	0.378	0.181	0.8	0.659
MPC	241	194	35	0.244	0.180	1.0	0.078
OXY	390	230	58	0.321	0.252	1.0	1.308
OKE	276	210	30	0.251	0.143	1.0	−0.226
PSX	186	95	15	0.289	0.158	0.6	−0.562
TRGP	221	159	25	0.251	0.157	1.0	−0.092
VLO	350	171	37	0.329	0.216	0.6	0.273
WMB	615	383	57	0.282	0.149	1.0	0.050

Variation in TAR was also observed across firms. Occidental Petroleum recorded the highest TAR value (0.252), followed by ConocoPhillips (0.216) and Valero Energy (0.216). These firms more frequently linked ESG targets to baseline references, target years, progress information, or quantitative evidence. In contrast, ExxonMobil (0.103), Expand Energy (0.129), and ONEOK (0.143) showed lower TAR values, indicating more limited accountability linkage in target-related disclosure.

RIS values suggested that many firms disclosed relatively extensive reporting infrastructure elements. Baker Hughes, Devon Energy, Expand Energy, ExxonMobil, Marathon Petroleum, Occidental Petroleum, ONEOK, Targa Resources, and Williams Companies all received RIS values of 1.0, indicating the presence of reporting framework disclosure, ESG data tables, assurance or verification practices, reporting boundary descriptions, and materiality assessment procedures. In contrast, APA Corporation and Phillips 66 showed lower RIS values within the sample.

The composite DQI provides an integrated view of these three dimensions. Because DQI is based on the average of the standardized QER, TAR, and RIS values, positive DQI values indicate above-average overall disclosure quality relative to the sample mean, whereas negative DQI values indicate below-average overall disclosure quality relative to the sample mean. Occidental Petroleum recorded the highest DQI (1.308), followed by Baker Hughes (0.697), Kinder Morgan (0.659), and ConocoPhillips (0.492). These firms tended to show comparatively favorable positions across multiple disclosure quality dimensions within the sample. At the other end of the distribution, ExxonMobil (−1.009), APA Corporation (−0.803), and Expand Energy (−0.705) showed lower overall disclosure quality positions within the sample.

The DQI results also show that higher values in one disclosure dimension did not necessarily correspond to uniformly higher values across all dimensions. Expand Energy and ExxonMobil, for example, received RIS values of 1.0 but showed relatively low QER and TAR values. Conversely, firms such as Valero Energy and EQT Corporation showed relatively high quantitative and accountability-oriented disclosure characteristics despite

not receiving the highest RIS values. These descriptive patterns are consistent with the interpretation that ESG disclosure quality may be multidimensional and that the composite DQI is more appropriately read together with the individual indicators than as a substitute for them.

Overall, the results show cross-firm variation in the measurability, accountability orientation, reporting infrastructure, and overall disclosure quality of sustainability reports across the sampled Energy Sector firms.

4.3. Relationship Between Disclosure Quality Indicators and External ESG Metrics

Table 5 presents the Spearman's rank correlation results between the proposed ESG disclosure quality indicators and external ESG evaluation metrics.

Table 5. Spearman's correlations between disclosure quality indicators and ESG evaluations.

Benchmark	Indicator	Spearman's ρ	p -Value
CSA Score	QER	0.345	0.1745
	TAR	0.319	0.2120
	RIS	0.313	0.2216
	DQI	0.645	0.0052 *
ESG Score	QER	0.311	0.2237
	TAR	0.300	0.2425
	RIS	0.330	0.1958
	DQI	0.613	0.0089 *

Note. * $p < 0.05$. Bold values indicate statistically significant correlations.

The correlation results showed different association patterns between the individual disclosure quality indicators and the composite disclosure quality index. QER, TAR, and RIS each showed modest positive associations with both ESG Score and CSA Score, although none of the individual correlations reached statistical significance.

For ESG Score, RIS showed the highest correlation among the individual indicators ($\rho = 0.330$), followed by QER ($\rho = 0.311$) and TAR ($\rho = 0.300$). For CSA Score, QER ($\rho = 0.345$), TAR ($\rho = 0.319$), and RIS ($\rho = 0.313$) also showed modest positive relationships. These results suggest that firms with higher values in reporting infrastructure, quantitative evidence, and accountability-oriented target disclosure also tended to receive higher external ESG evaluations, although these individual relationships were not statistically significant in the present sample.

By contrast, the composite DQI showed stronger positive correlations with both ESG Score ($\rho = 0.613$, $p = 0.0089$) and CSA Score ($\rho = 0.645$, $p = 0.0052$). This pattern suggests that external ESG evaluations were more closely associated with the combined structure of quantitative evidence, accountability-oriented target disclosure, and reporting infrastructure than with any single disclosure dimension considered separately.

Given the limited sample size, additional robustness checks were conducted to assess the stability of the main correlation results. First, a leave-one-out procedure was applied to examine whether the observed associations between the composite DQI and the two external ESG evaluation metrics were driven by any single firm [36]. In this procedure, one firm was sequentially removed from the sample, and Spearman's correlation between DQI and each external ESG metric was recalculated for the remaining firms. This test evaluates the sensitivity of the main results to firm-level exclusion and is particularly useful for identifying whether a statistically significant association is disproportionately influenced by a single observation.

Second, Kendall's tau was estimated as an additional rank-based check [37]. While Spearman's correlation evaluates the monotonic association between ranked variables,

Kendall's tau provides a complementary measure based on the concordance and discordance of ranked pairs. Therefore, Kendall's tau analysis was used to examine whether the positive association between DQI and external ESG evaluation metrics remained consistent under an alternative nonparametric rank-correlation measure.

Table 6 reports the results of these robustness checks. The leave-one-out results showed that the positive associations between DQI and both the CSA Score and the ESG Score were retained across all firm-exclusion specifications. For the CSA Score, the leave-one-out Spearman's correlations ranged from 0.574 to 0.746, with p -values ranging from 0.0009 to 0.0201. For the ESG Score, the leave-one-out Spearman's correlations ranged from 0.535 to 0.711, with p -values ranging from 0.0020 to 0.0328. These results indicate that the main DQI associations were not driven by a single influential firm.

Table 6. Robustness checks for the relationship between DQI and external ESG evaluation metrics.

Benchmark	Leave-One-Out ρ Range	Leave-One-Out p -Value Range	Kendall's τ	Kendall's τ p -Value
CSA Score	0.574–0.746	0.0009–0.0201	0.494	0.0057
ESG Score	0.535–0.711	0.0020–0.0328	0.442	0.0146

The Kendall's tau results also supported the robustness of the main findings. DQI showed positive and statistically significant associations with the CSA Score ($\tau = 0.494$, $p = 0.0057$) and the ESG Score ($\tau = 0.442$, $p = 0.0146$). Although the Kendall's tau coefficients were smaller than Spearman's correlations, which is expected because the two measures are scaled differently, the results remained directionally consistent and statistically significant. Taken together, these robustness checks suggest that the observed DQI–external ESG evaluation relationships are stable under both firm-level exclusion and an alternative rank-based correlation measure.

Because the composite DQI was constructed using equal weights, an additional sensitivity analysis was conducted to examine whether the main correlation results were dependent on this aggregation choice. The DQI was recalculated under alternative weighting specifications that emphasized QER, TAR, and RIS separately, while retaining z -score standardization for all components. Additional balanced alternatives were also examined to reflect greater emphasis on textual evidence or reporting infrastructure. As shown in Table 7, the recalculated composite indices remained positively and significantly associated with both the CSA Score and the ESG Score across all alternative specifications. This result suggests that the main finding was not an artifact of the equal-weighting assumption, although the equal-weighted DQI should still be interpreted as an exploratory composite indicator rather than as an empirically optimized weighting model.

Table 7. Sensitivity analysis under alternative DQI weighting specifications.

Specification	CSA ρ	CSA p -Value	ESG ρ	ESG p -Value
Equal weighting	0.645	0.0052	0.613	0.0089
QER-emphasized	0.630	0.0067	0.597	0.0115
TAR-emphasized	0.586	0.0134	0.558	0.0198
RIS-emphasized	0.722	0.0011	0.711	0.0014
Textual evidence-emphasized	0.571	0.0166	0.530	0.0286
Infrastructure-emphasized	0.671	0.0032	0.643	0.0053

Note. Equal weighting assigns 0.333 to QER, TAR, and RIS. QER-, TAR-, and RIS-emphasized specifications assign 0.500 to the focal component and 0.250 to each remaining component. The textual evidence-emphasized specification assigns weights of 0.400, 0.400, and 0.200 to QER, TAR, and RIS, respectively. The infrastructure-emphasized specification assigns weights of 0.300, 0.300, and 0.400 to QER, TAR, and RIS, respectively. All composite indices were calculated using standardized component values. Spearman's ρ is reported.

To further examine disclosure–rating divergence patterns, firm-level DQI rankings were compared with ESG Score rankings for all sampled firms. Table 8 reports the full ranking comparison.

Table 8. Full disclosure–rating divergence results across sampled firms.

Ticker	DQI Rank	ESG Score Rank	Divergence
OXY	1	7	+6
BKR	2	2	0
KMI	3	6	+3
COP	4	5	+1
EQT	5	10	+5
VLO	6	13	+7
MPC	7	9	+2
WMB	8	1	−7
DVN	9	8	−1
FANG	10	12	+2
TRGP	11	15	+4
OKE	12	4	−8
EOG	13	14	+1
PSX	14	13	−1
EXE	15	12	−3
APA	16	11	−5
XOM	17	17	0

The full divergence results suggest that sustainability disclosure quality and external ESG evaluation outcomes were related but not identical within the sample. Several firms with higher DQI ranks did not receive proportionally high ESG Score ranks, whereas some firms with stronger ESG Score ranks showed more moderate DQI positions. For example, OXY, EQT, and VLO ranked more favorably on DQI than on ESG Score, whereas WMB and OKE ranked more favorably on ESG Score than on DQI.

The divergence patterns reported in Table 6 also help contextualize the correlation results in Table 5. Although DQI showed a positive and statistically significant correlation with ESG Score, the firm-level rank differences indicate that the correspondence was not uniform across all firms. In this sense, disclosure quality and external ESG evaluation appeared directionally related at the aggregate level, while still showing firm-level differences.

Document structure and reporting format may also be relevant to some of these differences. Firms relying more heavily on tables, appendix-oriented reporting structures, condensed sustainability summaries, or visually organized disclosures may show lower sentence-level accountability measurements despite containing substantial ESG-related information at the document level. This suggests that sentence-level ESG disclosure analysis may partially underrepresent information embedded primarily in tables, visual layouts, or condensed reporting formats.

Overall, the results suggest a positive but incomplete correspondence between disclosure quality indicators derived from sustainability reports and external ESG evaluation outcomes. While firms with higher disclosure quality positions often also showed higher ESG and CSA Scores, the divergence analysis suggests that external ESG evaluations may reflect broader assessment processes beyond sustainability report disclosure structure alone.

5. Discussion

This study examined whether the sustainability disclosure quality of S&P 500 Energy firms is aligned with external ESG evaluation outcomes. Rather than treating sustainability reports as simple collections of ESG-related keywords or sentiment expressions, this study

focused on whether firms provided measurable, accountability-oriented, and structurally supported disclosures. To operationalize this perspective, this study constructed three disclosure quality indicators and subsequently integrated them into a composite DQI.

A central finding of this study is that the individual disclosure quality indicators showed only modest positive relationships with ESG Score and CSA Score, whereas the composite DQI showed stronger positive associations with both external benchmark metrics. This result suggests that external ESG evaluations may be more closely aligned with the combined structure of disclosure quality than with any single disclosure characteristic considered separately. Quantitative evidence, accountability-oriented target disclosure, and reporting infrastructure therefore appear to operate less as isolated signals and more as complementary dimensions of a broader disclosure quality profile.

This interpretation is broadly consistent with prior ESG disclosure research distinguishing between hard and soft disclosure and emphasizing that disclosure quality should not be reduced to disclosure volume alone [6,11,15]. Previous studies suggested that more credible ESG communication depends on measurable, comparable, and reliable information rather than on broad declarative sustainability language [16,17]. The present findings extend that perspective by indicating that disclosure quality may be more meaningfully captured when these characteristics are considered together. In this sense, numerical disclosure alone may be less informative when target accountability remains weak, while formal reporting infrastructure may also be less informative when disclosures lack traceable targets, baselines, or progress-related information.

The findings also contribute to the interpretation of sustainability reporting as a form of corporate communication. Prior studies have described ESG disclosure not only as an information mechanism but also as a strategic communication process through which firms shape legitimacy, accountability, and stakeholder perception [8–10]. The present results are consistent with that view, but they also indicate that more extensive disclosure should not automatically be interpreted as higher-quality disclosure. Firms with substantial report volume did not always show stronger QER, TAR, or DQI values. This suggests that disclosure extensiveness and disclosure quality should be treated as related but distinct concepts. Within this framework, sustainability reporting appears more informative when firms move beyond broad ESG narratives and provide measurable evidence, accountable target structures, and formalized reporting support.

The divergence analysis provides a further basis for interpreting the relationship between disclosure quality and external ESG evaluations. Although DQI showed positive associations with the ESG Score and the CSA Score, the firm-level ranking comparison showed that alignment was incomplete. Some firms ranked more favorably on DQI than on ESG Score, whereas others showed the opposite pattern. This divergence is consistent with prior ESG rating research showing that external ESG evaluations are shaped by differences in scope, measurement, weighting, and information sources [21–24]. At the same time, the observed divergence also suggests that disclosure quality and external ESG evaluation outcomes should be interpreted as related but distinct constructs. Whereas DQI captures the measurability, target linkage, and reporting structure of sustainability disclosure, external ESG evaluation outcomes may also reflect broader assessments of ESG performance, risk exposure, sector-relative positioning, and information sources beyond the sustainability report itself.

This interpretation is illustrated by firms such as OXY, which ranked first in DQI but did not rank first in ESG Score. In this case, the high DQI rank suggests that the firm's sustainability report provided relatively measurable, target-linked, and structurally supported disclosure. However, this does not necessarily imply that the firm would receive the highest external ESG ranking within the sector. In the Energy Sector, external ESG

evaluations may be influenced by factors such as carbon-intensive operations, emissions exposure, fossil-fuel asset dependence, transition risk, operational controversies, regulatory issues, or peer-relative decarbonization assessments. Accordingly, a firm may provide relatively high-quality sustainability disclosure while still receiving a more moderate ESG evaluation if broader performance- or risk-related factors are assessed less favorably. Conversely, firms with stronger external ESG rankings may not always show the highest sentence-level DQI if relevant ESG information is embedded in other reporting formats or if external evaluations place greater emphasis on performance outcomes, controversies, or proprietary assessment inputs.

These divergence patterns therefore provide a useful caution for interpreting the relationship between disclosure quality and ESG ratings. The positive correlation results suggest that firms with stronger multidimensional disclosure quality tended to receive more favorable external evaluations at the aggregate level. However, the firm-level differences indicate that high disclosure quality may be better interpreted as a reporting-quality signal rather than as direct evidence of higher ESG performance or lower ESG risk. In this sense, DQI may be more appropriately understood as an indicator of the structure and accountability of ESG communication. This distinction helps explain why disclosure quality and external ESG evaluation outcomes can move together at the aggregate level while still diverging for specific firms.

This study also has methodological implications for AI-assisted ESG analysis. Prior ESG text studies have used sentiment analysis, keyword frequency approaches, or topic modeling to identify themes and communication patterns [1,23]. These approaches remain useful for detecting visibility, tone, or thematic emphasis, but they are less directly suited to evaluating whether disclosures are measurable, accountable, or structurally supported. The present study addressed this issue by limiting the role of the LLM to ESG claim identification and ESG dimension classification based on a predefined codebook. Disclosure quality indicators were then constructed through rule-based feature extraction grounded in observable textual characteristics. This hybrid design preserves scalability while also improving transparency in how the quality measures are operationalized.

Several limitations should also be acknowledged. First, the analysis focused only on S&P 500 Energy Sector firms, which limits generalizability across industries with different reporting structures and ESG evaluation logics. Second, the final sample remained relatively small because this study prioritized document comparability and standalone sustainability reports suitable for stable sentence-level analysis. Third, the analysis relied primarily on sentence-level textual extraction from PDF documents and may not have fully captured ESG information embedded in tables, appendices, graphics, metric dashboards, or visually structured sections. This limitation may underrepresent firms that condense sustainability metrics into tabular or visual formats, even when such formats provide structured and decision-useful ESG disclosure. Fourth, the publication years of the corporate sustainability reports and the update dates of the S&P Global ESG benchmark data were not perfectly synchronized across all firms. Because historical S&P Global ESG Scores corresponding exactly to each report year were not publicly, some degree of temporal mismatch may remain between the disclosure measures and external ESG evaluation outcomes. In addition, the external ESG metrics used in this study reflect broader assessment processes than sustainability report content alone. For this reason, the observed relationships should not be interpreted as direct equivalence between disclosure quality and ESG performance.

These limitations also indicate several directions for future research. Subsequent studies could examine whether similar relationships appear in other industries or in larger cross-sector samples. Applying the framework to larger cross-sector datasets would also allow future research to examine whether the relative importance of quantitative evidence,

target accountability, and reporting infrastructure varies across industries with different regulatory pressures, reporting practices, and ESG evaluation logics. Future work could also extend the framework through table-aware extraction, layout-sensitive parsing, and multimodal document analysis in order to better capture visually structured ESG information. When historical ESG benchmark data aligned with corporate sustainability report years become available, future studies could further examine whether closer temporal matching between report publication periods and ESG benchmark update dates changes the observed correlation and divergence patterns. Additional research may also compare whether different ESG rating providers show similar relationships with multidimensional disclosure quality or whether these relationships vary depending on provider methodology.

Overall, the findings suggest that LLM-assisted and rule-based structured content analysis provides a transparent and scalable approach for evaluating ESG disclosure quality. Although disclosure quality did not fully explain external ESG evaluation outcomes, the proposed framework showed that measurable, accountability-oriented, and structurally supported disclosure characteristics were associated with external ESG assessments and helped identify where disclosure quality and rating outcomes converged or diverged across firms.

6. Conclusions

This study examined the relationship between sustainability disclosure quality and external ESG evaluation outcomes among S&P 500 Energy Sector firms. Using LLM-assisted ESG claim identification and rule-based feature extraction, this study developed a multidimensional framework for evaluating ESG disclosure quality based on quantitative evidence, target accountability, and reporting infrastructure characteristics.

The findings indicated that the composite DQI had stronger positive associations with the ESG Score and the CSA Score than the individual disclosure quality indicators considered separately. This result suggests that external ESG evaluations may be associated with the overall structure of disclosure quality to a greater extent than with isolated disclosure characteristics alone. At the same time, the observed disclosure–rating divergence patterns indicate that sustainability disclosure quality and external ESG evaluation outcomes should not be treated as interchangeable constructs.

This study has practical implications for both corporate reporting and ESG analysis. For firms, the results suggest that ESG communication quality may be related not only to disclosure volume but also to the structure of disclosure itself. In practical terms, this includes providing clearer quantitative evidence, linking targets to baselines and target years, reporting progress more explicitly, and strengthening formal reporting infrastructure through framework disclosure, assurance, boundary definition, and materiality-related explanation. For researchers and analysts, the proposed framework provides a transparent and scalable basis for comparing disclosure quality across firms and for identifying where disclosure quality and external ESG evaluations converge or diverge.

More broadly, this study suggests that LLM-assisted and rule-based structured content analysis may provide a relatively transparent and scalable approach for ESG disclosure assessment. However, because the analysis was restricted to a small sample of S&P 500 Energy Sector firms with comparable standalone sustainability reports, the findings should be interpreted as exploratory sector-specific evidence with limited generalizability. Future research may extend this framework through table-aware extraction, multimodal document analysis, temporally aligned ESG benchmark datasets, and larger cross-sector samples that allow the framework to be tested across industries with different regulatory pressures, reporting practices, and ESG evaluation logics.

Author Contributions: Conceptualization, H.S.J., S.C. and H.L.; methodology, H.S.J.; software, H.S.J.; validation, H.S.J. and H.L.; formal analysis, H.S.J.; investigation, H.S.J.; resources, H.L.; data curation, H.S.J.; writing—original draft preparation, H.S.J. and S.C.; writing—review and editing, H.L.; visualization, H.S.J.; supervision, H.L.; project administration, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Dongguk University Research Fund under Grant S-2025-G0001-00139. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25481681).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

Appendix A ESG Coding Prompt and Classification Framework

To improve methodological transparency and reproducibility, the representative ESG coding prompt used in the LLM-assisted content analysis procedure is provided below.

Table A1. Representative ESG claim identification prompt.

You are an ESG disclosure content-analysis coder for an academic study. Your task is to classify each sentence. Definitions: - ESG claim: A sentence that describes a company's environmental, social, or governance-related performance, target, policy, risk, program, activity, governance structure, reporting practice, stakeholder action, or accountability mechanism. - Do NOT classify generic business descriptions as ESG claims unless they explicitly relate to E, S, or G. - Do NOT classify generic marketing slogans as ESG claims unless they describe a specific ESG-related policy, target, action, performance, or risk. - Do NOT infer beyond the sentence. Code only what is explicitly stated. Dimension definitions: - E: climate, GHG emissions, Scope 1/2/3, methane, flaring, energy use, water, waste, biodiversity, spills, air quality, environmental incidents, environmental targets. - S: safety, workforce, employees, training, human rights, community investment, diversity, inclusion, employee health, contractors, Indigenous communities, stakeholder engagement. - G: board oversight, ethics, compliance, risk management, cybersecurity, lobbying, executive compensation, supply chain governance, reporting governance, assurance, reporting framework. - Mixed: more than one of E/S/G is clearly present. - None: not an ESG claim. Claim type options: - emissions_performance - climate_target - water_waste_biodiversity - safety_health - workforce_human_capital - community_human_rights - governance_ethics_compliance - risk_cybersecurity - reporting_framework_assurance - supply_chain - generic_esg_statement - non_esg Return JSON only. Do not include markdown. Do not include explanation outside JSON.
--

Appendix B Representative ESG Claim Classification Examples

To improve methodological transparency and illustrate the ESG coding procedure, representative examples of ESG claim classifications generated through the LLM-assisted content analysis framework are provided below.

Table A2. Representative ESG claim classification examples.

Sentence	Dimension
Annual training on our Anti-Corruption and Bribery guidelines is required for every Targa employee.	G
As a member of the ONE Future Coalition, ONEOK is committed to helping reduce methane emissions across the natural gas value chain and is participating in working toward a sector-specific methane emissions intensity target for our natural gas transmission and storage operations.	E
In 2024, 88.5% of EQT employees participated in the survey and we received a 79% workplace experience score.	S
Oxy has zero tolerance for the disregard or circumvention of Oxy policies or engaging in unethical dealings in connection with Oxy's businesses.	G
ONEOK voluntarily enrolled in Candidate Conservation Agreements with Assurances (CCAAs) for the lesser prairie chicken and monarch butterfly.	E
Nearly 100% of our salaried employees participate in our performance process each year.	S
This function is overseen by the vice president of Internal Audit, Risk Management and Compliance, who verifies that procedures are in place for the corporatwide identification and management of both operational and nonoperational risks and provides oversight of ongoing, companywide monitoring and risk management.	G
We do not operate within national parks or federally designated wetlands, unless permitted, and are committed to the conservation of wildlife and species protection.	E
We adhere to the Payment Card Industry Data Security Standard, which provides requirements that protect payment card account data of our customers.	G
All employees were invited to participate in the series, and the resources accompanying these events provide employees with simple ways to focus on and improve their mental well-being.	S
Through transparent and inclusive engagement, we work closely with stakeholders to understand and address the social and environmental impacts of our business.	Mixed
Workplace Culture Policies All Expand Energy decisions regarding recruiting, hiring, training, evaluation, assignment, advancement and termination of employment are made without unlawful discrimination.	S

Appendix C Rule-Based Feature Extraction Patterns

To improve methodological transparency and reproducibility, representative regular expression patterns used for rule-based ESG feature extraction are provided below.

Table A3. Representative regex patterns for ESG feature extraction.

<pre> number_pattern = r"(\d+(\.\d+)?) (\\$?\d{1,3})(,\d{3})+)" percent_pattern = r"\d+(\.\d+)?\s?%" year_pattern = r"\b20\d{2}\b" unit_pattern = r"(mt metric tons? tonnes? tons? co2e co2 ch4 mwh gwh kwh barrels? bbl bcf mmbcf mmbtu boe mboe gallons? acres? hours? employees? workers? incidents? rate trir ltir million billion thousand mw hp dollars? \\$)" scope_pattern = r"scope\s*(1 2 3 1\s*and\s*2 1\s*\+\s*2 1/2 1\s*&\s*2)" baseline_pattern = r"(from\s+20\d{2}\s+(baseline base\s+year levels?) compared\s+to\s+(our\s+)?20\d{2} versus\s+(our\s+)?20\d{2} since\s+20\d{2} 20\d{2}\s+(baseline base\s+year))" target_year_pattern = r"(by\s+20\d{2} through\s+20\d{2} between\s+20\d{2}\s+and\s+20\d{2} 20\d{2}\s+target target\s+year\s+20\d{2})" progress_pattern = r"(reduced reduction decreased declined lowered cut increased improved achieved exceeded outperformed maintained reached completed implemented expanded on\s+track progress attained)" target_pattern = r"(target goal aim ambition commitment pledge objective net\s+zero zero\s+routine\s+flaring reduce reduction maintain achieve)" kpi_pattern = r"(scope\s*(1 2 3 1\s*and\s*2 1\s*\+\s*2 1/2 1\s*&\s*2) ghg greenhouse\s+gas methane flaring carbon\s+intensity emissions?\s+intensity trir ltir sif spill spills water\s+withdrawal freshwater waste biodiversity code\s+of\s+conduct training\s+hours volunteer\s+hours)" framework_pattern = r"(gri sasb tcfd issb ipieca xpc ogmp cdp sdg un\s+global\s+compact limited\s+assurance independent\s+assurance third-party\s+assurance performance\s+data data\s+table materiality reporting\s+boundary)" </pre>
--

References

1. Lee, H.; Kim, J.H.; Jung, H.S. From corporate earnings calls to social impact: Exploring ESG signals in S&P 500 ESG index companies through transformer-based models. *J. Clean. Prod.* **2025**, *501*, 145320.
2. Okeke, A. Towards sustainability in the global oil and gas industry: Identifying where the emphasis lies. *Environ. Sustain. Indic.* **2021**, *12*, 100145. [\[CrossRef\]](#)
3. Bronzini, M.; Nicolini, C.; Lepri, B.; Passerini, A.; Staiano, J. Glitter or gold? Deriving structured insights from sustainability reports via large language models. *EPJ Data Sci.* **2024**, *13*, 41. [\[CrossRef\]](#)
4. Boiral, O. Sustainability reports as simulacra? A counter-account of A and A+ GRI reports. *Account. Audit. Account. J.* **2013**, *26*, 1036–1071. [\[CrossRef\]](#)
5. Cho, C.H.; Laine, M.; Roberts, R.W.; Rodrigue, M. Organized hypocrisy, organizational façades, and sustainability reporting. *Account. Organ. Soc.* **2015**, *40*, 78–94. [\[CrossRef\]](#)
6. Clarkson, P.M.; Li, Y.; Richardson, G.D.; Vasvari, F.P. Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. *Account. Organ. Soc.* **2008**, *33*, 303–327. [\[CrossRef\]](#)
7. Lee, H.; Lee, S.H.; Park, H.; Kim, J.H.; Jung, H.S. ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models. *Heliyon* **2024**, *10*, e26404. [\[PubMed\]](#)
8. Gray, R.; Kouhy, R.; Lavers, S. Corporate social and environmental reporting: A review of the literature and a longitudinal study of UK disclosure. *Account. Audit. Account. J.* **1995**, *8*, 47–77.
9. Deegan, C. Introduction: The legitimising effect of social and environmental disclosures—A theoretical foundation. *Account. Audit. Account. J.* **2002**, *15*, 282–311.
10. Suchman, M.C. Managing legitimacy: Strategic and institutional approaches. *Acad. Manag. Rev.* **1995**, *20*, 571–610. [\[CrossRef\]](#)
11. Michelon, G.; Pilonato, S.; Ricceri, F. CSR reporting practices and the quality of disclosure: An empirical analysis. *Crit. Perspect. Account.* **2015**, *33*, 59–78. [\[CrossRef\]](#)
12. Cho, C.H.; Michelon, G.; Patten, D.M.; Roberts, R.W. CSR disclosure: The more things change...? *Account. Audit. Account. J.* **2015**, *28*, 14–35. [\[CrossRef\]](#)

13. Hahn, R.; Kühnen, M. Determinants of sustainability reporting: A review of results, trends, theory, and opportunities in an expanding field of research. *J. Clean. Prod.* **2013**, *59*, 5–21. [CrossRef]
14. Christensen, H.B.; Hail, L.; Leuz, C. Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Rev. Account. Stud.* **2021**, *26*, 1176–1248. [CrossRef]
15. Abraham, S.; Shrivies, P.J. Improving the relevance of risk factor disclosure in corporate annual reports. *Br. Account. Rev.* **2014**, *46*, 91–107. [CrossRef]
16. Hummel, K.; Schlick, C. The relationship between sustainability performance and sustainability disclosure—Reconciling voluntary disclosure theory and legitimacy theory. *J. Account. Public Policy* **2016**, *35*, 455–476.
17. Beretta, S.; Bozzolan, S. Quality versus quantity: The case of forward-looking disclosure. *J. Account. Audit. Financ.* **2008**, *23*, 333–376. [CrossRef]
18. Cho, C.H.; Roberts, R.W.; Patten, D.M. The language of US corporate environmental disclosure. *Account. Organ. Soc.* **2010**, *35*, 431–443. [CrossRef]
19. Marquis, C.; Toffel, M.W.; Zhou, Y. Scrutiny, norms, and selective disclosure: A global study of greenwashing. *Organ. Sci.* **2016**, *27*, 483–504. [CrossRef]
20. Lyon, T.P.; Montgomery, A.W. The means and end of greenwash. *Organ. Environ.* **2015**, *28*, 223–249. [CrossRef]
21. Berg, F.; Kölbel, J.F.; Rigobon, R. Aggregate confusion: The divergence of ESG ratings. *Rev. Financ.* **2022**, *26*, 1315–1344. [CrossRef]
22. 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.
23. Lee, H.; Kim, J.H.; Jung, H.S. ESG-KIBERT: A new paradigm in ESG evaluation using NLP and industry-specific customization. *Decis. Support Syst.* **2025**, *193*, 114440.
24. Christensen, D.M.; Serafeim, G.; Sikochi, A. Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *Account. Rev.* **2022**, *97*, 147–175.
25. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [PubMed]
26. Törnberg, P. How to use LLMs for text analysis. *arXiv* **2023**, arXiv:2307.13106.
27. Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Chen, H.; Liu, Z.; Dou, Z.; Wen, J.-R. Large language models for information retrieval: A survey. *ACM Trans. Inf. Syst.* **2025**, *44*, 1–54. [CrossRef]
28. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada, 3–10 March 2021.
29. Van Dis, E.A.; Bollen, J.; Zuidema, W.; Van Rooij, R.; Bockting, C.L. ChatGPT: Five priorities for research. *Nature* **2023**, *614*, 224–226. [CrossRef] [PubMed]
30. MSCI and S&P Dow Jones Indices. Global Industry Classification Standard (GICS) Methodology. Available online: <https://www.msci.com/gics> (accessed on 28 May 2026).
31. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.
32. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
33. Simnett, R.; Vanstraelen, A.; Chua, W.F. Assurance on sustainability reports: An international comparison. *Account. Rev.* **2009**, *84*, 937–967. [CrossRef]
34. Perego, P.; Kolk, A. Multinationals' accountability on sustainability: The evolution of third-party assurance of sustainability reports. *J. Bus. Ethics.* **2012**, *110*, 173–190. [CrossRef]
35. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1987**, *100*, 441–471. [CrossRef] [PubMed]
36. Tukey, J.W. Bias and confidence in not quite large samples. *Ann. Math. Stat.* **1958**, *29*, 614.
37. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. [CrossRef]

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.