Extracting airline emission KPIs from sustainability reports using large language models (LLMs)

Luis Martín-Domingo [a,b,*] , Jaime B. Fernandez [c], Marina Efthymiou [a], Muhammad Intizar Ali [c]

[a] *Business School, Dublin City University, Dublin 9, Ireland*
[b] *Faculty of Aviation, Ozyegin University, Istanbul, Turkey*
[c] *Insight Research Ireland Centre for Data Analytics, Dublin City University, Dublin 9, Ireland*

ARTICLE INFO

ABSTRACT

The extraction of environmental Key Performance Indicators (KPIs) from airline sustainability reports is essential for assessing environmental sustainability metrics and regulatory compliance within the European aviation sector. Manual extraction from extensive, unstructured documents is laborious and often inconsistent. This study systematically investigates the potential of advanced Large Language Models (LLMs) –specifically –GPT-4.0, o3-mini, and Deepseek R1- to automate the extraction of emissions-related KPIs from the 2023 sustainability reports of 16 publicly traded European airline groups. Utilizing the Perplexity platform, the research contrasts manual expert extraction with automated approaches, exploring various models, prompt strategies, and data formats. Results indicate that the accuracy of LLM extraction depends significantly on prompt specificity. Attempts to extract data from unstructured documents without guidance yielded low accuracy. However, incorporating explicit KPI terms into prompts increased accuracy from below 30% to above 70%. The format of the data source was also influential, with HTML formats producing superior extraction results compared to PDFs. Despite ongoing challenges in standardizing data and extracting precise KPI metrics, the findings demonstrate that LLMs can substantially streamline environmental, social and governance (ESG) data collection when prompt engineering and source standardization are prioritized. This study represents a novel, interdisciplinary approach by combining advances in large language models (LLMs) with expertise in environmental, social, and governance (ESG) analysis within the aviation sector, offering empirical benchmarking of LLM performance in real-world regulatory contexts. Recommendations for LLM integration into ESG analysis workflows are provided, and future research directions for advancing automation in sustainability reporting are discussed.

Introduction

Environmental sustainability has been a matter of growing global concern, capturing the attention of policymakers, scholars, corporations and the general public. Although numerous frameworks and methodologies exist for evaluating sustainability outcomes, the diversity in corporate reporting formats and levels of detail across companies has made consistent assessment and comparison of environmental sustainability metrics particularly difficult (Zou et al., 2025).

One solution to this challenge is to extract environmental Key Performance Indicators (KPIs) from annual and sustainability reports, also referred as Environment Social and Government (ESG) reports, which is a well-established approach for evaluating corporate environmental

performance and adherence to sustainability objectives (Caraveo Gomez Llanos et al., 2023; Zieba and Johansson, 2022). For publicly traded companies, annual and sustainable reports are mandatory in Europe under the Non-Financial Reporting Directive (NFRD) and the Corporate Sustainability Reporting Directive (CSRD), making them a reliable and standardized source of data (EU, 2014, 2024). However, manual extraction from these complex and often unstructured documents is a labour-intensive and time-consuming process that can lead to inconsistencies and errors (Ong et al., 2025). Halteh et al. (2024) demonstrated how automation through machine learning can enhance the extraction and analysis of complex financial patterns in the aviation industry.

In recent years, the rapid development of LLMs such as GPT-4.0 and

---

Deepseek R1 has revolutionized the field of Natural Language Processing (NLP), enabling more sophisticated and context-aware data extraction from unstructured text sources (Ong et al., 2025). These models, accessible through user-friendly platforms, offer the potential to automate and standardize the extraction of ESG-related KPIs from large volumes of airline sustainability disclosures. As recent European regulatory frameworks increase the complexity and volume of required disclosures (IATA, 2024a; Zieba and Johansson, 2022), leveraging LLMs could address the growing need for scalable, accurate, and consistent data extraction methods.

This paper presents a systematic evaluation of LLM-based data extraction workflows using real-world airline sustainability reports, comparing manual and automated extraction performance across multiple models and prompt strategies. By identifying the strengths and limitations of current LLMs in this context, the research contributes to the ongoing discourse about digital transformation in sustainability reporting and highlights practical considerations for integrating AI-driven tools in ESG data analysis. The interdisciplinary nature of this research bridges the fields of natural language processing, sustainability science, regulatory compliance, and aviation management, undertaking a rare contribution to both AI research and applied sustainability practices. This cross-sectoral integration lays the groundwork for future studies seeking to automate complex ESG data extraction tasks. Ultimately, this study provides a foundation for future research in improving the accuracy, consistency, and scalability of LLMs for environmental sustainability data extraction.

Literature review

As environmental challenges intensify, organisations are under pressure to monitor, improve and report their environmental performance. Several recent studies have addressed aviation-related emissions and sustainability metrics. For instance, Bruzzone et al. (2023) developed a performance indicator framework for integrated passenger–freight systems, demonstrating how transport models can be evaluated for environmental sustainability and energy efficiency. Ayesu (2023) analysed the link between the shipping sector and environmental emissions in African economies, emphasizing the importance of emissions data in transport policy. Dobruszkes and Efthymiou (2020) critiqued the framing of environmental indicators in aviation, particularly in the context of aircraft noise in Europe, and demonstrated that even established environmental metrics may embed policy assumptions. Sobieralski and Mumbower (2022) examined emissions trends linked to the rise in private aviation during COVID-19, while Calderon-Tellez and Herrera (2021) simulated the effects of pandemic-related travel restrictions on aviation emissions in Latin America. Wild et al. (2021) focused on the behavioural and regulatory outcomes of market-based instruments such as flight taxes and carbon offsetting.

Opferkuch et al. (2021) argue that there is a lack of standardisation in corporate sustainability reporting, which makes it difficult for companies to extract, organize, and disclose meaningful information. As a result, companies are often left to independently select definitions, assessment methods, and performance indicators, leading to highly variable, which are difficult to compare reports. This inconsistency undermines the credibility and transparency of sustainability claims, making it harder for external stakeholders to assess companies' true environmental sustainability metrics.

In the context of air transport, the increasing expectation for ESG's to be integrated into the strategies of European airlines and regulatory frameworks has made emissions-related KPIs central to both compliance and strategic decision-making, with transparent and reliable emissions data now critical for investor confidence and regulatory adherence (EU, 2023; Lufthansa Group, 2023). Evolving regulatory frameworks such as the Non-Financial Reporting Directive (NFRD) and the Corporate Sustainability Reporting Directive (CSRD) have introduced stricter requirements for sustainability reporting, reinforcing the need for robust,

comparable, and transparent emissions metrics (EU, 2014, 2024). Despite these advances, the academic literature consistently highlights a lack of standardization in the manner in which airlines report their environmental sustainability metrics —a challenge identified over two decades ago and still present today (Caraveo Gomez Llanos et al., 2023; Hooper and Greenall, 2005; Zieba and Johansson, 2022). Establishing a clear set of greenhouse gas (GHG) KPIs and closely monitoring them through periodic reporting remains a critical foundation for achieving net zero GHG emissions by 2050, yet ongoing efforts are needed to harmonize reporting standards and methodologies across the sector.

ESG and sustainability reporting by airlines are useful sources of information for researchers, and content analysis is a frequently employed method to extract pertinent information (Zieba and Johansson, 2022). Thus far, content analysis of airline sustainability reporting has predominantly been conducted manually (Coles et al., 2014; Cowper-Smith and de Grosbois, 2011; Ringham and Miles, 2018). More recently, Li et al. (2023) used a quantitative analysis tool to assist in data coding. However, these manual and semi-automated approaches are increasingly challenged by the growing complexity, volume, and unstructured nature of sustainability metrics, especially as regulatory requirements increase in number and complexity under frameworks like the CSRD and NFRD (IATA, 2024a; Zieba and Johansson, 2022). Manual content analysis is resource-intensive, time-consuming, and susceptible to inconsistencies and subjective bias, particularly when analysts must interpret nuanced or ambiguous language across diverse reporting formats and languages. In this context, Natural Language Processing (NLP) offers a robust alternative for extracting and analyzing, for example, data from airline sustainability and ESG reports. Zou et al. (2025) have used an LLM to evaluate the sustainability performance of 166 companies and found that GPT-4 achieved higher accuracy in identifying and quantifying ESG claims, arguing that it is a sufficient and effective tool for extracting data from unstructured ESG reports.

NLP (Natural Language Processing), despite its challenging complexity, has several useful applications such as Machine Translation, Information Extraction and Summarization (Khurana et al., 2023) and is deployed in different fields, including business (Bahja, 2021), medicine (Locke et al., 2021) and construction (Ding et al., 2022). In recent years, several methodologies have been developed to guide the execution of such tasks (Koroteev, 2021). Traditionally, text extraction methods have been employed to retrieve data from PDF documents. In cases where documents contain images, Optical Character Recognition (OCR) techniques have been used to generate useable text from images. However, with the recent development of new technologies in terms of both hardware (GPUs) and software (Deep Learning, Long Short-Term Memory networks–LSTMs, Attention mechanism, and Transformers), LLMs (Large Language Models) have emerged as powerful tools for performing Natural Language Processing (NLP) tasks with high accuracy.

From 2018 onward, researchers have focused on building increasingly larger models. In 2019 researchers from Google introduced BERT, the two-directional, 340-million parameter model (the third largest model of its kind) that could determine context, which enabled it to adapt to a greater variety of tasks. By pre-training BERT on a wide variety of unstructured data via self-supervised learning, the model was able to understand the relationships between words. Almost immediately, BERT became the preeminent tool for natural language processing tasks. In fact, it was BERT that was behind every English-based query administered via Google Search. After BERT, GPT-1 *continued* the evolution of LLMs by carrying out simple tasks such as answering questions. When GPT-2 came out in 2019, the model had grown significantly, expanding more than ten times the size of GPT-1. GPT-2 could now produce human-like text and perform certain tasks automatically. With the introduction of GPT-3 in 2020, the public was able to access this innovative technology. In 2022 GPT-3 introduced problem-solving as an even further functionality. GPT-3.5 broadened the system's capabilities, becoming more streamlined and less costly. Recently, in 2023 and 2024,

GPT-4 and GPT-4o, feature significant enhancements, such as the ability to use computer vision to interpret visual data. These models accept both text and images as input (Raiaan et al., 2024; Toloka, 2023; Wang et al., 2024).

For researchers, the great benefit of all this development is that these models have been made available for local execution or as web-based tools like Perplexity and ChatGPT. These tools provide accessible and user-friendly interfaces with which users can execute queries (prompts), select their preferred LLM model, fine tune these models by proving their own database of knowledge as files (pdfs, websites, docs, excel) and even to give instructions for how the model should behave. Further, these tools also have the flexibility to permit that the model be told that its answers should not include newly generated information but be limited to knowledge provided in the database.

Considering the potential confusion that has accompanied the rapid advance and sheer novelty of these models, this study presents a real use case for emission-related KPI information retrieval from airlines annual and sustainability reports. The Perplexity platform—a versatile tool that integrates multiple LLMs, has been selected as the tool best suited to this task due to its user-friendly interface and access to the latest LLM models, such as GPT-4.0 (2023) and Deepseek R1 (January 21, 2025). These models offer advanced capabilities, making them the most up-to-date options available in Perplexity. Comparisons between these models can be found in sources like DocsBot (2025) and Elecrow (2025). For evaluation purposes we also selected a light version of the GPT called o3-mini, released on January 31, 2025, which, similar to Deepseek R1, is a reasoning model. The rapid advance of the technology has created a climate of uncertainty which this study hopes to ameliorate by providing a real practicable usage.

Methodology

Four key steps were involved in the methodology: first, selecting major European airline groups based on regional and regulatory criteria; second, collecting their 2023 annual and sustainability reports from investor relations websites; third, two experts manually extracting emission-related KPIs through systematic review of report sections; and fourth, using three large language models (LLMs) on the Perplexity platform to automatically extract the same KPIs via both unguided and guided prompts. The results from manual and automated extractions were then systematically compared to assess accuracy and consistency in KPI reporting across the sample.

This research includes 16 European airline groups operating from European Economic Area (EEA), the UK, and Switzerland. This selection represents adherence to common regulatory frameworks for sustainability and includes major European airline groups (EU, 2014). The researchers created a comprehensive list beginning with 121 European IATA members, which transport 83 % of global air passenger traffic (IATA, 2024b), and narrowed down to 88 airlines within the specified European region. Each airline was researched at its investor relation web pages, which also helped to establish the relationships between large publicly traded European airline groups and their subsidiary airlines. The Pitchbook[1] database was used to verify ownership when necessary, and airlines providing cargo services alone were excluded from the analysis. To ensure comprehensive coverage of non-IATA members, the researchers also examined the top 40 airlines by number of flights from Eurocontrol, identifying additional non-IATA publicly traded European airlines (e.g., Ryanair and Wizzair). This process resulted in a final sample of 31 airlines belonging to 16 airline groups, as detailed in Appendix 1. The non-financial reports (part of annual reports) of the 16

airline groups were downloaded from airline group investor relations websites for analysis. The year covered for analysis was 2023. However, some airlines do not report data from January to December. For example, the reporting period for easyJet was October 2022 – September 2023 and for Ryanair April 2023 – March 2024. In all cases, 2023 included the largest number of reported months for all airlines analyzed.

The main data format used to extract airline emission related KPIs were PDF documents as this was the common format provided by most of the European airline groups analysed. However, one exception was the Icelander group which provided their annual and sustainability report only in HTML format (Icelandair, 2024). The first approach was to convert the annual Icelandair report to PDF so as to ensure the same data format in all the analysed airline groups. The platform used for data extraction also allows processing of files in different formats, in addition to URL links. Thus, for Icelandair, the data extraction was carried out using both forms of data (i.e. PDF and HTML).

Both financial and non-financial data was provided in the same report by all the analysed airlines, although different names were adopted: "Annual Report" (Aegean, Croatian, Finnair, Lufthansa, Norse, Norwegian, Ryanair, and TUI); "Annual Reports and Accounts" (easyJet, Jet2, IAG, and Wizzair); "Annual and Sustainability Report" (Air Baltic, Icelandair, and SAS); and "Universal Registration Document" (Air France). Some airline groups publish a specific sustainability report, in addition to the annual report including: "Sustainability Report" (Aegean, and Ryanair); "ESG Factsheet" (easyJet); "Consolidated Statement of Non-Financial Information" (IAG); and "Sustainability Factsheet" (Lufthansa). For these, both annual reports and sustainability reports were downloaded for analysis. The analysed PDF reports ranged from 4 pages (easyJet ESG Factsheet) to 488 pages (AF-KLM Universal Registration document). A full list of page number can be found in Appendix 1.

Emission-related KPIs were manually extracted by two experts in sustainable aviation. Their approach was to review the section related to emissions of each airline's group report, which were titled as follows: Sustainability (Aegean, easyJet, Finnair, IAG, Jet2, TUI, and Wizzair); Environment (Air France-KLM, Lufthansa, Ryanair, and SAS); Climate (Air Baltic and Icelandair); and Environmental Responsibility (Norse and Norwegian). They then extracted the full metric name, unit, and values of each KPI and stored them in one table. When a discrepancy appeared between the experts, a joint revision of the report was conducted until an agreement was reached. A table of published KPIs in 2023 was created for each airline group.

The three models used between 17 and 28 March 2025 were ChatGPT 4.0, Chat GPT o3-mini (aka o3-mini), and Deepseek R1. Using three LLMs allows comparison of their performance in extracting structured insights from European airline reports. First, the LLM models were tasked with directly extracting relevant data from the annual and sustainability reports. This step allows the ability of each LLM to interpret unstructured text, tables, and charts within the documents, to be evaluated, providing insights into their capacity to handle diverse formats and complex content. Second, a predefined list of environmental KPIs—manually curated from prior analyses—was supplied to the models to refine their focus during extraction from airline annual and sustainability reports. This guided approach tests the ability of the models to locate specific data points efficiently when prompted with explicit instructions, enhancing precision in extracting targeted information. Both approaches, first and second, were compared for performance.

In the Perplexity platform the following variables were adjusted: i) The LLM used (i.e. ChatGPT 4.0, o3-mini, and Deepseek R1); ii) The sources were limited to those uploaded by the researchers (i.e. Airlines annual and sustainability reports and manually predefined list of environmental KPIs); and iii) the instructions given for all queries were "Always respond in a formal tone and prioritize data-driven insights. Extract only data from uploaded sources". The prompts used are

---

[1] The Pitchbook database was used for reliability for those airlines whose investor relations websites couldn't confirm whether they were publicly traded companies. The link to access the online database is https://pitchbook.com/profiles/company/129567–61.

included in Table 1.

Perplexity allowed for the creation of dedicated workspaces for each airline that prevented mixing data between airlines. In each space, the annual and sustainability reports for each airline group were uploaded. Each space was configured by unselecting the given options (i.e. Web, Academic, and social) so that the LLM only extracted data from the sources provided. After running each prompt, the extracted table of KPIs produced by each LLM and prompt was copied into an Excel document for manual comparison with a master table containing KPI metrics, units, and values.

Extracted KPIs using LLM models were only counted as correct when the metric name, unit, and value of the KPI were the same as those of the mandatory KPIs manually extracted by the two experts. There were situations in which the KPI was not expressed using the same units of the metric, but if the combination of unit and value was the same as that in the reference table, it was considered a correct extraction. For example, in the data extraction from the IAG group report, the KPI "Emissions net reduction due to SAF" was as indicated in the below Table 2.

From this example, we can observe that the KPI name is not transcribed exactly from the report but shows minor differences, as is the case with LLM o3-mini writing out the full name of the acronym SAF. The original KPI value and unit were 157.1 $ktCO_2$, where kilo tonnes refers to thousands of tonnes. Chat GPT 4.0 and o3-mini refer to thousand tonnes. However, Deepseek R1 extracts the KPI in tonnes of CO2 providing the actual number of tonnes (157,100).

The methodological framework exemplifies interdisciplinary research by merging expertise from AI (prompt engineering, model selection), sustainability analytics, and regulatory studies. By systematically comparing manual expert processes with automated LLM workflows, the study fuses technical, domain, and policy perspectives in ESG analysis in an innovative manner.

Results and discussion

This section first presents an evaluation of the accuracy and reliability of large language models (LLMs) in extracting emissions-related Key Performance Indicators (KPIs) from European airline sustainability reports. Subsequently, the analysis covers multiple data sources, extraction strategies, and model architectures, providing a comprehensive overview of the factors influencing automated KPI extraction performance.

The data extraction accuracy, measured as the percentage of KPIs matching the manual extraction, of each of the three LLMs used (Chat GPT 4.0, o3-mini, and Deepseek R1) is presented in Fig. 1.

The comparative analysis of LLM performance in extracting emissions-related KPIs from European airline reports shown in Fig. 1 show substantial differences based on both the prompt strategy and the model used. When tasked with unguided extraction—using only the uploaded reports and the prompt to "extract all environmental KPIs related to emissions"—all three models performed poorly (28 %), with overall accuracy rates of 29 % for GPT 4.0, 32 % for o3-mini, and just 22 % for Deepseek R1. This low performance is consistent across most airline groups, and particularly Norse, Ryanair, and AF-KLM, which could be an indication of more complex or less standardized reporting formats. This also indicates that LLMs struggle to independently identify and extract relevant emission-related KPIs from unstructured documents without explicit guidance.

Introducing a guided extraction approach by providing the models with a master list of KPIs in Excel format resulted in modest improvements (34 % vs. 28 %). o3-mini and Deepseek R1 both reached an average accuracy of 35 %, while that of GPT 4.0 increased slightly to 31 %. The results for some airlines, like easyJet and Finnair, saw marked improvements, with o3-mini achieving 100 % accuracy for Finnair and Deepseek R1 reaching 85 %. However, persistent challenges remained for an accurate assessment for several airline groups, as many models still failed to match the manual extraction benchmark, highlighting limitations in cross-referencing and matching KPI definitions even when structured references are available.

The most dramatic increase in extraction accuracy was observed (66 %) when explicit KPI metric names were included directly in the prompt. Under this highly guided scenario, GPT 4.0 achieved 71 % accuracy, o3-mini 65 %, and Deepseek R1 63 %. Results for several airlines, including Finnair, IAG, Lufthansa, Norse, SAS, and Wizzair, reached near-perfect or perfect extraction rates across all models, demonstrating that prompt specificity is the dominant factor in LLM data extraction performance. While GPT 4.0 led in optimal conditions, the differences between models diminished as prompt guidance increased, underscoring that the quality and clarity of instructions are more critical than the choice of model for reliable automated KPI extraction from complex sustainability reports.

Fig. 2 below shows the results of data extraction for the Icelandair Group when using PDFs as a data source, compared to those obtained using the annual and sustainability report in HTML format provided on its website.

We can observe that the data extraction from the PDFs was very poor in this case; however, the data extraction from the corresponding website was much more effective, especially when the list of KPI names was included directly in the prompt. The LLM o3-mini achieved a 75 % success rate, Chat GPT-4.0 67 %, and Deepseek R1 50 %. These results for Icelandair suggest that using a data source in HTML format should lead to better data extraction results. However, the same type of comparison could not be made with other airlines, as no others provided the full environmental information used in the analysis in HTML format.

Alternative data sources to PDF were provided by three airline groups (easyJet, Finnair, and IAG). These airline groups provided their annual and sustainability reports in iXBRL (inline eXtensive Business Reporting Language) format, an open standard that enables a single document to provide both human-readable and structured, machine-readable data (XBRL, 2025). Unfortunately, data extraction could not be tested on this format as it was not supported by the LLM platforms used for this research.

European publicly traded companies (including airlines) are now required to publish financial information in the European Single Electronic Format (ESEF) with XBLR tagging (ESMA, 2025). However, sustainability data reporting, following ESEF with XBLR tagging is only expected to be mandatory in 2027 (ESMA, 2024). Thus, close monitoring of the implementation of ESMA and its adoption by European

Table 1
Prompts used to extract KPIs from European airline's annual and sustainability reports (authors).

ID	Prompt	KPI names extracted manually	Airline annual and sustainability report
1	*"Extract all the environmental KPIs related to emissions of airline-name 2023. Organize it on a table with three columns: KPIs name, KPIs units, KPI value"*	No	PDF Uploaded
2	*"Using the list of KPIs placed in the airline-name-Master-table related to emissions. Extract airline-name KPIs for 2023 from the attached reports (KPI name, KPI unit, KPI value). Organize them on a table with three columns: KPI name, KPI unit, KPI value*	Yes on Excel file uploaded	PDF Uploaded
3	"Extract all the following KPIs of *airline-name* related to emissions in 2023. Organize it on a table with three columns: KPIs name, KPIs units, KPI value. $KPIname_1$, $KPIname_2$, $KPIname_3$, …,$KPIname_n$"	Yes Included in prompt	PDF Uploaded

Table 2

Data extraction comparison of emissions net reduction from the use of Sustainable Aviation Fuel (SAF) (Authors).

KPI	Chat GPT 4.0	o3-mini		Deepseek R1	Manual by Experts
Name	CO$_2$ Saved from SAF Use	CO$_2$ saved from Sustainable Aviation Fuel (SAF) use		CO$_2$ saved from SAF use:	Net reduction (SAF uplift):
Unit	Thousand tonnes CO$_2$	Thousand tonnes (KT)		Tonnes CO$_2$	ktCO2
Value	157.1		157.1	157,100	157.1

Fig. 1. KPI data extraction using different LLMs and methods (authors).

Fig. 2. KPI data extraction comparison for Icelandair using different data source formats, LLMs and methods (authors).

publicly listed airlines will be important steps towards improving the effectiveness of sustainable KPI data extraction. LLMs are already starting to be used with iXBRL formats (Aavang et al., 2025) and present very promising new avenues for automated data extraction.

During the LLM data extraction, there were different situations that led to incorrect data extraction: 1) Correct KPI name and unit, but incorrect value; 2) Correct name and value, but incorrect units; and 3) Both the unit and the value were incorrect. The following tables show examples of each category.

The first two categories of errors (incorrect unit and value extraction of the KPIs) can be observed in the Air Baltic report data extraction of the metric Sustainable Aviation Fuel (SAF) as shown in Table 3 below.

In the Chat GPT 4.0 data extraction, the units were incorrect (litres instead of kg), and in the Deepseek R1 data extraction, the value was

incorrect (3,105 instead of 71,372). Thus, the o3-mini data extraction was the only one that was fully correct. The example shown below in Table 4 is from SAS data extraction for the metric flight operations emissions.

In the Chat GPT 4.0 extraction, both the unit and the value do not match those from the manual extraction. However, the combination of unit and value resulted in the same value as our reference (i.e. 3,091,000 tonnes of CO2e) and therefore is considered a correct extraction. In the o3-mini data extraction, the value was the same as in the previous case; however, the units were not aligned to work out the same value (i.e. 3,091 tonnes of CO2e) In the Deepseek R1 extraction, the same units and value were extracted (i.e. 3,091,000 tonnes of CO2e). The example in Table 5 below shows data extraction for Aegean for the metric of Fuel

Table 3

Data extraction comparison (using **prompt ID 3** of Table 1) from **Air Baltic** report on Sustainable Aviation Fuel (SAF) (Authors).

KPI	Chat GPT 4.0	o3-mini	Deepseek R1	Manual by Experts
Name	SAF	SAF (Sustainable Aviation Fuel)	SAF	SAF
Unit	~~Litres~~	Kilograms (kg)	Kg	Kg
Value	71,372	71,372	~~3,105~~	71,372

Table 4

Data extraction comparison (using **prompt ID 3** of Table 1) from **SAS** report on total flight operations CO$_2$equivalent (e) emissions (Authors).

KPI	Chat GPT 4.0	o3-mini	Deepseek R1	Manual by Experts
Name	CO2e total Flight operations	CO2e total Flight operations	CO2e total Flight operations	total Flight operations
Unit	1,000 tonnes	Metric Tonnes	Tonnes	tonnes CO2e
Value	3,091	~~3,091~~	3,091,000	3,091,000

Table 5
Data extraction comparison (using **prompt ID 3** of Table 1) from **Aegean** report on Fuel Efficiency per passenger kilometre (Authors).

KPI	Chat GPT 4.0	o3-mini	Deepseek R1	Manual by Experts
Name	Fuel Efficiency – Revenue Passenger Kilometers (RPK)	Fuel Consumption – Passenger Flights	Fuel Efficiency – Revenue Passenger Kilo-meters	Fuel Efficiency – Revenue Passenger Kilo-meters (RPK)
Unit	kg/100 RPK	~~Tons~~	~~RPK/kg fuel~~	kg/100 RPK
Value	2.47	~~Not explicitly stated~~	~~40.654~~	2.47

efficiency as measured by Revenue per passenger kilometre.

In the above case, only the data extraction by Chat GPT 4.0 was correct (i.e. 2.47 kg/100 RPK). In the other two cases (o3-mini and Deepseek R1), both the unit and the value were incorrect extractions.

These findings are not only technically significant, but they also demonstrate the practical benefits of interdisciplinary dialogue between AI, ESG analytics, and regulatory policy. The results offer insights for computer scientists, sustainability officers, and regulatory agencies alike, showing the tangible impact of interdisciplinary collaborations in overcoming sector-wide reporting challenges.

Conclusions

This manuscript makes a novel contribution by demonstrating the power of interdisciplinary scholarship, uniting artificial intelligence, regulatory policy, and transport environmental sustainability in a single analytical workflow. Environmental KPIs and, in general, ESG claims are critical for investors to make informed decisions and for companies like airlines to demonstrate their commitment to sustainability practices and address their climate change contributions. This study demonstrates that Large Language Models (LLMs) such as GPT-4.0, o3-mini, and Deepseek R1 can significantly streamline the extraction of environmental KPIs from European airline sustainability reports, particularly when guided by explicit and well-structured prompts. The results show that extraction accuracy is highly dependent on the specificity of instructions provided to the LLMs and the format of the source data. When KPI metric names were included directly in prompts, extraction accuracy increased substantially, with GPT-4.0 achieving up to 71 % accuracy. Conversely, unguided extraction from unstructured PDF documents yielded poor results across all models.

Furthermore, the study highlights the importance of the data source format: HTML and structured data formats (such as iXBRL) offer superior extraction results compared to PDFs, although current LLM platforms may not fully support all structured formats. The findings underscore the value of prompt engineering and data standardization in maximizing the effectiveness of LLM-assisted ESG data extraction.

This paper contributes to the literature by empirically demonstrating that Large Language Models (LLMs) can automate ESG KPI extraction from sustainability reports, with accuracy highly dependent on prompt specificity and data structure. It bridges a gap by providing systematic benchmarking across multiple LLMs and prompts, thereby advancing knowledge on AI applications in sustainability reporting. Practically, it offers a roadmap for companies to integrate LLMs into ESG workflows, showing how prompt design and source standardization can greatly improve extraction efficiency. A major wider implication of the paper is that it can help accelerate the digital transformation of sustainability reporting and facilitates regulatory moves towards machine-readable ESG claims.

The findings of this study have several practical implications: For airlines, LLM-powered automation can reduce the labor and time required for ESG report generation, thereby improving the consistency, reproducibility, and transparency of reported emissions data. This is particularly important for small companies under resources constrains. Regulators stand to benefit through enhanced oversight as LLM-enabled workflows can facilitate real-time monitoring and verification of compliance and highlight data inconsistencies, provided that the push towards machine-readable and structured reporting formats is maintained. Finally, investors, third-party data providers, and civil society organizations will gain from more reliable, comparable, accessible and timely ESG disclosures, which support due diligence, benchmarking, and advocacy efforts. However, all stakeholders must remain aware of ongoing challenges in data standardization, accuracy, and the limitation of current models.

This study is limited by its focus on a sample of 16 European airline groups and its reliance on commercial LLM interfaces, which may restrict the adaptability and scalability of the extraction process.

Additionally, the analysis primarily targeted emissions-related KPIs and faced challenges with unstructured PDF formats and the inability to process certain structured data formats like iXBRL, potentially limiting the generalizability and completeness of the findings. In addition to standard KPI extraction demonstrated in this study, the evolving capabilities of LLMs open promising avenues for their expanded use in the transport sustainability context. These include automated Q&A interfaces for ESG data, multi-report summarization, regulatory compliance validation, harmonization of disparate metric formats, and even predictive scenario modelling. However, widespread adoption faces notable challenges. Extraction accuracy remains highly dependent on document format and prompt clarity, and current LLMs often struggle with complex, non-standard data presentation. LLMs often fail to link extracted KPIs to specific document locations or sources, undermining auditability and trust and complicating compliance audits. Fully automated KPI extraction without human review risks serious misinterpretation in regulatory or operational settings. Using commercial LLM APIs means data might be made public beyond the jurisdiction of the responsible organization, potentially violating GDPR or other regulations and reports processed by cloud-hosted LLMs, risk exposing sensitive information to external systems.

Future research could broaden the scope to include a wider range of ESG indicators, different transport modes or additional sectors, and the use of multilingual and further fine-tuned LLMs. Further exploration of LLM integration with structured data extraction tools, as well as cost-effectiveness and operational scalability assessments, would help advance automated ESG data extraction and support evolving regulatory requirements. By fostering collaboration across disciplinary boundaries, AI, sustainability, and regulatory compliance, this research sets a precedent for ongoing digital transformation and standardisation within ESG reporting for airlines and stakeholders beyond.

Research data for this article

Research data set used in this article can be found in the following citation:

Martín-Domingo, L. (2025). European Public Listed Airlines Annual and Sustainability Reports 2023 [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14876202.

CRediT authorship contribution statement

Luis Martín-Domingo: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jaime B. Fernandez:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Marina Efthymiou:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition. **Muhammad Intizar Ali:** Writing – review & editing, Supervision, Software, Methodology, Conceptualization.

Funding

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

A1: List of European publicly traded airline groups, reporting period analysed and reports used for the analysis (authors based on airlines investor relations data).

ID	Airline Group	Period	pp. – Combined report	pp. – Sustainability report
1	Aegean Group	Jan-Dec23	258 pp. – Ann. report	139 pp. – Sustainability report
2	Air Baltic Corporation	Jan-Dec23	182 pp. – Ann. and sustainability	–
3	Air France-KLM Group	Jan-Dec23	488 pp. – Universal registration doc.	–
4	Croatian Airlines	Jan-Dec23	135 pp. – Ann. report	–
5	easyJet plc	Oct-22-Sep23	204 pp. – Ann. reports and accounts	4 pp. – ESG Factsheet
6	Finnair Group	Jan-Dec23	160 pp. – Ann. report	–
7	IAG Group	Jan-Dec23	316 pp. – Ann. reports and accounts	113 pp. – Non-financial info
8	Icelandair Group	Jan-Dec23	13 pp. – Ann. and sustainability	–
9	Jet2 plc	Jan-Dec23	89 pp. – Ann. reports and accounts	–
10	Lufthansa Group	Jan-Dec23	329 pp. – Ann. report	32 pp.- Sustainability Factsheet
11	Norse Atlantic ASA	Jan-Dec23	60 pp. – Ann. report	–
12	Norwegian Group	Jan-Dec23	153 pp. – Ann. report	–
13	Ryanair Group	Apr23-Mar24	244 pp. – Ann. report	72 pp. – Sustainability report
14	SAS Group	Nov22-Oct23	158 pp. – Ann. and sustainability	–
15	Tui Group	Nov22-Oct23	297 pp. – Ann. report	–
16	Wizzair Holdings plc	Apr23-Mar24	246 pp. – Ann. reports and accounts	–

pp.: Number of pages of the report; Ann. (Annual).

References

Aavang, R., Rizzi, G., Bøggild, R., Iolov, A., Zhang, M., & Bjerva, J. (2025). *HiFi-KPI: A Dataset for Hierarchical KPI Extraction from Earnings Filings* (arXiv:2502.15411). arXiv. https://doi.org/10.48550/arXiv.2502.15411.

Ayesu, E.K., 2023. Does shipping cause environmental emissions? evidence from african countries. Transp. Res. Interdisci. Perspect. 21, 100873. https://doi.org/10.1016/j.trip.2023.100873.

Bahja, M. (2021). Natural language processing applications in business. In *E-Business: Higher Education and Intelligence Applications*. IntechOpen. https://doi.org/10.5772/intechopen.92203.

Bruzzone, F., Cavallaro, F., Nocera, S., 2023. Environmental and energy performance of integrated passenger–freight transport. Transp. Res. Interdiscip. Perspect. 22, 100958. https://doi.org/10.1016/j.trip.2023.100958.

Calderon-Tellez, J.A., Herrera, M.M., 2021. Appraising the impact of air transport on the environment: lessons from the COVID-19 pandemic. Transp. Res. Interdiscip. Perspect. 10, 100351. https://doi.org/10.1016/j.trip.2021.100351.

Caraveo Gomez Llanos, A.F., Vijaya, A., Wicaksono, H., 2023. Rating ESG key performance indicators in the airline industry. Environ. Dev. Sustain. https://doi.org/10.1007/s10668-023-03775-z.

Coles, T., Fenclova, E., Dinan, C., 2014. Corporate social responsibility reporting among European low-fares airlines: challenges for the examination and development of sustainable mobilities. J. Sustain. Tour. 22 (1), 69–88. https://doi.org/10.1080/09669582.2013.790391.

Cowper-Smith, A., de Grosbois, D., 2011. The adoption of corporate social responsibility practices in the airline industry. J. Sustain. Tour. 19 (1), 59–77. https://doi.org/10.1080/09669582.2010.498918.

Ding, Y., Ma, J., Luo, X., 2022. Applications of natural language processing in construction. Autom. Constr. 136, 104169. https://doi.org/10.1016/j.autcon.2022.104169.

Dobruszkes, F., Efthymiou, M., 2020. When environmental indicators are not neutral: assessing aircraft noise assessment in Europe. J. Air Transp. Manag. 88, 101861. https://doi.org/10.1016/j.jairtraman.2020.101861.

DocsBot, 2025. GPT-4 vs DeepSeek-R1—detailed performance & feature comparison. DocsBot AI. https://docsbot.ai/models/compare/gpt-4/deepseek-r1.

Elecrow. (2025). *DeepSeek R1 vs. OpenAI GPT-4: A Comparative Analysis*. https://www.elecrow.com/blog/deepseek-r1-vs-chatgpt-4-a-comparative-analysis.html.

ESMA. (2024). *ESMA consults on proposals to digitalise sustainability and financial disclosures*. https://www.esma.europa.eu/press-news/esma-news/esma-consults-proposals-digitalise-sustainability-and-financial-disclosures.

ESMA. (2025). *2024 ESEF XBRL files and ESEF conformance suite*. https://www.esma.europa.eu/press-news/esma-news/esma-publishes-2024-esef-xbrl-files-and-esef-conformance-suite.

EU. (2014). *Directive (EU) 2014/95 of the European Parliament and of the Council of 22 October 2014 amending Directive 2013/34/EU as regards disclosure of non-financial and diversity information by certain large undertakings and groups Text with EEA relevance*. http://data.europa.eu/eli/dir/2014/95/oj/eng.

EU. (2023). *Sustainable finance: Council agrees negotiating mandate on ESG ratings*. https://www.consilium.europa.eu/en/press/press-releases/2023/12/20/sustainable-finance-council-agrees-negotiating-mandate-on-esg-ratings/.

EU. (2024). *Regulation—EU - 2024/3005 -on the transparency and integrity of Environmental, Social and Governance (ESG) rating activities, and amending Regulations (EU) 2019/2088 and (EU) 2023/2859*. https://eur-lex.europa.eu/eli/reg/2024/3005/oj/eng.

Halteh, K., AlKhoury, R., Adel Ziadat, S., Gepp, A., Kumar, K., 2024. Using machine learning techniques to assess the financial impact of the COVID-19 pandemic on the global aviation industry. Transp. Res. Interdiscip. Perspect. 24, 101043. https://doi.org/10.1016/j.trip.2024.101043.

Hooper, P.D., Greenall, A., 2005. Exploring the potential for environmental performance benchmarking in the airline sector. BIJ 12 (2), 151–165. https://doi.org/10.1108/14635770510593095.

IATA. (2024a). *Beginners Guide to Airline Sustainability Reporting*. https://www.iata.org/contentassets/77ec9a8c8a864daaa00bdb7f5de02902/beginners-guide-to-airline-sustainability-reporting-april2024.pdf.

IATA. (2024b). *IATA Members*. https://www.iata.org/en/about/members/.

Icelandair. (2024). *Icelandair Annual and Sustainability Report 2023*. http://annualreport2023.icelandairgroup.is/.

Khurana, D., Koli, A., Khatter, K., Singh, S., 2023. Natural language processing: State of the art, current trends and challenges. Multimed. Tools Appl. 82 (3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4.

Koroteev, M. V. (2021). *BERT: A Review of Applications in Natural Language Processing and Understanding* (arXiv:2103.11943). arXiv. https://doi.org/10.48550/arXiv.2103.11943.

Li, W., Cui, J., Gao, J., Xiong, J., 2023. Corporate social responsibility in China's airline industry: a longitudinal content analysis of related reports. J. Air Transp. Manag. 111, 102420. https://doi.org/10.1016/j.jairtraman.2023.102420.

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., Kitchen, G.B., 2021. Natural language processing in medicine: a review. Tren. Anaesth. Crit. Care 38, 4–9. https://doi.org/10.1016/j.tacc.2021.02.007.

Lufthansa Group. (2023). *Lufthansa Group Annual Report 2022*. https://www.lufthansagroup.com/en/themes/annual-report-2022.html.

Ong, K., Mao, R., Satapathy, R., Filho, R.S., Cambria, E., Sulaeman, J., Mengaldo, G., 2025. Explainable natural language processing for corporate sustainability analysis. Inf. Fusion 115, 102726. https://doi.org/10.1016/j.inffus.2024.102726.

Opferkuch, K., Caeiro, S., Salomone, R., Ramos, T.B., 2021. Circular economy in corporate sustainability reporting: a review of organisational approaches. Bus. Strateg. Environ. 30 (8), 4015–4036. https://doi.org/10.1002/bse.2854.

Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S., 2024. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access 12, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742.

Ringham, K., Miles, S., 2018. The boundary of corporate social responsibility reporting: the case of the airline industry. J. Sustain. Tour. 26 (7), 1043–1062. https://doi.org/10.1080/09669582.2017.1423317.

Sobieralski, J.B., Mumbower, S., 2022. Jet-setting during COVID-19: environmental implications of the pandemic induced private aviation boom. Transp. Res. Interdiscip. Perspect. 13, 100575. https://doi.org/10.1016/j.trip.2022.100575.

Toloka. (2023). *The history, timeline, and future of LLMs*. https://toloka.ai/blog/history-of-llms/.

Wang, Z., Chu, Z., Doan, T.V., Ni, S., Yang, M., Zhang, W., 2024. History, development, and principles of large language models: an introductory survey. AI Ethics. https://doi.org/10.1007/s43681-024-00583-7.

Wild, P., Mathys, F., Wang, J., 2021. Impact of political and market-based measures on aviation emissions and passenger behaviors (a Swiss case study). Transp. Res. Interdiscip. Perspect. 10, 100405. https://doi.org/10.1016/j.trip.2021.100405.

XBRL. (2025). *What is iXBRL?* https://www.xbrl.org/the-standard/what/ixbrl/.

Zieba, M., Johansson, E., 2022. Sustainability reporting in the airline industry: current literature and future research avenues. Transp. Res. Part D: Transp. Environ. 102, 103133. https://doi.org/10.1016/j.trd.2021.103133.

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., Zhou, W., 2025. ESGReveal: an LLM-based approach for extracting structured data from ESG reports. J. Clean. Prod. 489, 144572. https://doi.org/10.1016/j.jclepro.2024.144572.