Autonomous Summarization of Corporate Social Responsibility Reports using Agentic AI Workflows

Vedant Desai[1], Anvay Khuperkar[1], Svara Dadhe[1], and Dr. Ruchi Sharma[2]

[1] Department of Information Technology, Mukesh Patel School of Technology Management and Engineering, NMIMS, Mumbai, India {vedant.desai49, anvay.khuperkar22, anant.kekre57, svara.dadhe81, eeshaan.garkhedkar40}@nmims.in
[2] ruchi.k6508@gmail.com

Abstract. CSR reporting is becoming a major concern with regard to sustainable investment and corporate transparency. But as the present is, that is not the case.CSR reports are massive, very heterogeneous, and time-consuming to statistically respond to manually. This process is also dissimilar, resource-intensive. To beat these odds and this project presents the CSR. Summarizer, autonomous, containerized AI architecture. that is provided to fully automate the summarization of CSR reports [6, 7, 9]. The system works with PDF-based.CSR reports extract clean textual data, and uses AIbased logic to produce structured summaries. The centre of the solution will be developed around a locally hosted Large. A workflow engine, which is provided in Docker, and Language Model (LLM). The self-hosted architecture is essential since it can guarantee high portability, scalability as well as, above all, holds all personal information in secrecy. By the system operates locally, reducing the concerns. linked to the privacy of data, lock-in of vendors and the recurrence. expenses related to AI services delivered by the cloud. The workflow is an automated pipeline which begins with file-system, new PDF report monitoring, text extraction is done using a Python based engine and the clean text is then fed into.
the local LLM.

1 Introduction

Corporate Social responsibility (CSR) reports have come into the limelight of Corporate Sustainability in the past few years as the global interest in the issue of sustainability and corporate accountability is considered a key factor by investors, regulators, and consumers in determining the corporate responsibility of the corporation [12,14]. Such reports provide an insight into long-term value-creating policies and risk management policies of a company. Nevertheless, these types of documents have huge challenges that tend to impair their practical utility. The reports on CSR are voluminous, disordered, and inconsistent across companies in their structure and content, which makes the manual review inherently an impractical and subjective process [1,13]. Not only is such a manual procedure highly susceptible to human error and bias, but also large amounts of domain knowledge are needed, which is also a major obstacle to scalable analysis. With the aim of dealing with these inefficiencies, the CSR Summarizer was created to help exploit the strength of artificial intelligence to automate the full process of data interpretation. The essence of this project is to develop a system that has the capability of handling the raw and unstructured CSR documents, and then extracting the most significant insight items intelligently and then compiling them in a structure form of a summary [2,16]. In so doing, the system will strive to remove the acts of inconsistency and subjectivity of manual review and hence offer quicker, more valid and operational CSR intelligence to all the involved parties.

2 Literature Review

Artificial intelligence (AI) and natural language processing (NLP) have found their prominent place in finance and corporate reporting [2,24,17]; though, in the current literature, more emphasis is laid on their potential and demand in the field of Corporate Social Responsibility (CSR) analytics. The recent international focus on sustainability, corporate responsibility, and Environmental, Social, and Governance (ESG) standards and criteria has made the manual examination of CSR disclosures prohibitively resource-intensive and uneven [14].

2.1 Deep Learning and Improved CSR Interpretation

Scholarly studies still indicate that deep-learning models prove highly efficient at decoding unstructured and generally complex text of voluminous CSR disclosures [1,11]. These models are sophisticated to the level of being able to do more than just a simple keyword search, enabling them to:

- Determine complex subject matters and stories concerning sustainability initiatives, allowing a more profound awareness of business activities.

- Calculate emotion and mood expressed in different sections of a report, which enables evaluators to differentiate between aspirational and substantive actions [10].

- Derive significant measures and Key Performance Indicators (KPIs), which are necessary factors that determine quantitatively the social and environmental footprint of a company.

- Increase productivity and fiscal performance: AI-based CSR processes have been observed to be better and smoother than traditional approaches, consequently enhancing operational impact, ESG business operations, and investor confidence [12, 23].

The use of AI optimizes data management and helps predict sustainability trends, which in turn improves the overall efficiency, accuracy, and transparency of CSR reporting. [13]

2.2 Uncritical Moving to Sovereign AI

Although the early sophisticated analysis tools were numerous, they were very primitive. I have been designed on cloud-based artificial intelligence systems provided by big techs [8] whose benefits are frequently disputed changes Riddled by great disadvantages, especially in the area of corporate reporting. Among the major concerns associated with these platforms are:

- **Data Protection Uncertainty:** * Sensitive company information and strategic knowledge is often stored as part of the corporate social responsibility (CSR) reports, have to be backed-up to third party servers. This raises important questions that relate to data sovereignty, compliance, and confidentiality, in particular, in highly controlled sectors [7].
- **Vendor Lock-in and Cost Volatility:** Dependence on a single cloud provider creates a vendor lock-in situation, making migration to another service provider difficult and costly due to proprietary formats and non- standard APIs. Additionally, pricing models are often unpredictable—costs can escalate rapidly with increased usage—resulting in long-term financial uncertainty and lack of budget reliability [15].

2.3 Local, Open-source AI Architecture SMV

The recent academic research has found the effectiveness of local implementation of AI application, using open-source tools to tackle the previously mentioned limitations. The latter strategy is typified by a sovereign and locally based AI pipeline to analyse Corporate Social Responsibility [7].

- **Self-Hosted LLM:** Delivering large language models like Llama3 via Olla- max systems ensure organisations have increased scalability and control of data and infrastructure [3]. This will make sure that the sensitive infor- mation is not stored on the external network and therefore help to eliminate the issue of privacy and enable ease of adherence to data-protection rules.
- **Open-Source Workflow Automation:** Open-source workflow automa- tion systems like n8n [4,21] also provide the building blocks to a full-fledged end-to-end processing system through the integration of these LLMs. The platforms chain the operation of the tasks, without involving human labor between the stages of document ingestion, data mining, and AI logic into a

single, automated pipeline. As a result, organisations are able to build document parsing and summarisation pipelines in state of art without external data leaks, which maximises the security as well as scalability.

The CSR Summarizer perfectly responds to this need in the industry, providing an innovative, LLM-based, fully containerised automated workflow that leverages on the locally available and open-source technologies.

3 Activities Proposed in the Working Project

The CSR Summarizer is simply a pipeline which is easily run with a few modules to keep the operation lean and mean. The logic of the system is coordinated by a workflow automation engine and links together all the services in a way that it can accept data in at the ingestion to the final output. We have created our project in two stages, a prototype of the project on the lightning-fast cloud (to check the essence logic), and a containerized local infrastructure that can be expanded to production levels.

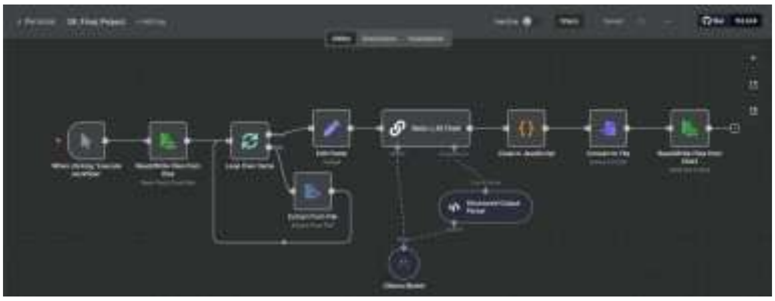


Fig. 1. Automated CSR Report Processing Workflow

3.1 Workflow: CSR Reports Summarizer Workflow

It is a workflow that is meant to receive CSR reports in the form of PDF files, process the PDF files, and extract the data with the help of an LLM and then format the outcomes into a CSV file.The complete workflow consists of the following stages:

- **Manual Workflow Trigger:** The workflow is started with a manual trigger, which is a manual process performed when the user starts the workflow.
- **File Ingestion:** Read/Write Files node of Disk node reads a collection of that have been locally stored.

- **PDF Processing Loop:** Loop Over Items node loops over every file that is ingested. In the case of each item, the Extract from File node is used to extract the raw text content of the PDF document.[5,22].

- **Data Preparation:** The extracted text is then staged by the Edit Fields node (rename/set) after the loop, probably with formatting or renaming data fields to make them ready as input to the LLM.

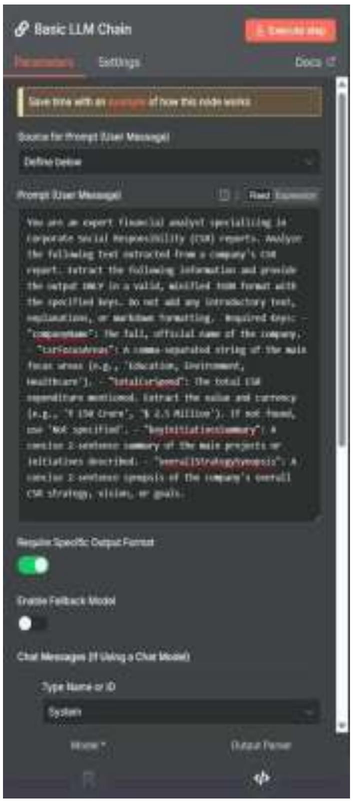
- **LLM Processing:** The processed data is transmitted to a Basic LLM Chain. A local Ollama Model (e.g., LLaMA 3) is used to power this chain and analyzes the text.[3,19,22]

- **Structured Output Generation:** Structured Output Parser is a part of the LLM chain. This drives the output of the model into a structured format, which is predefined (such as JSON), and is consistent..

- **Transformation of Output:** The output of the LLM is a structured output that is fed to a Code node. At this stage, post-processing of the output is done to clean and reorganize the results.

- **Format Conversion:** The Convert to File node will convert the resulting JSON or text output into a CSV format, to store the output in the CSV file.

- **File Output and Storage:** The last node in the pipeline is the Read/Write Files from Disk node which writes the new CSV file to disk.



Source for Prompt (User Message)

Define below

Prompt (User Message)

```
You are an expert financial analyst specializing in
Corporate Social Responsibility (CSR) reports. Analyze
the following text extracted from a company's CSR
report. Extract the following information and provide
the output ONLY in a valid, minified JSON format with
the specified keys. Do not add any introductory text,
explanations, or markdown formatting.  Required Keys: -
"companyName": The full, official name of the company.
- "csrFocusAreas": A comma-separated string of the main
focus areas (e.g., 'Education, Environment,
Healthcare'). - "totalCsrSpend": The total CSR
expenditure mentioned. Extract the value and currency
(e.g., '₹ 150 Crore', '$ 2.5 Million'). If not found,
use 'Not specified'. - "keyInitiativesSummary": A
concise 2-sentence summary of the main projects or
initiatives described. - "overallStrategySynopsis": A
concise 2-sentence synopsis of the company's overall
CSR strategy, vision, or goals.
```

Require Specific Output Format

●

Enable Fallback Model

○

Chat Messages (if Using a Chat Model)

Type Name or ID

System

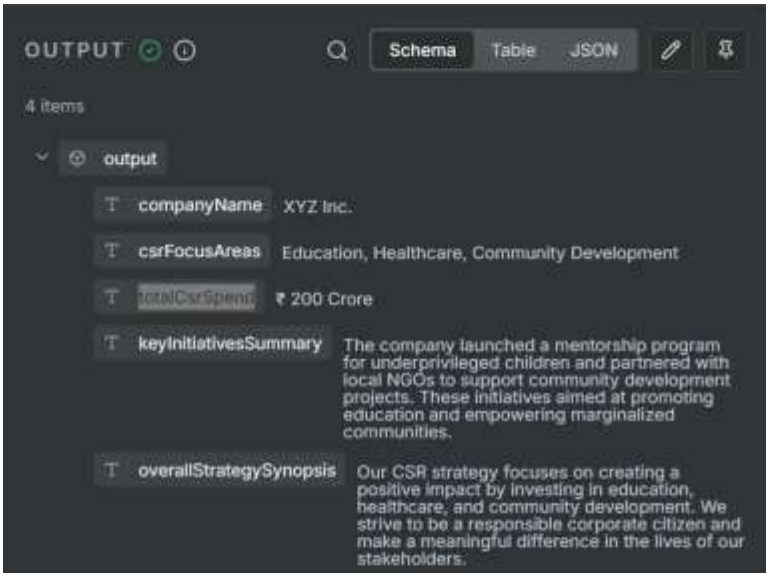
Fig. 2. Sample prompt used for the LLM.

4 Results and Analysis

The n8n workflow (described in Section 3) was executed successfully, which automated the process of extracting Corporate Social Responsibility (CSR) data of a batch of PDF reports. The findings indicate the usefulness of the pipeline in converting unstructured text to a structured, analysis-ready format.

4.1 Intermediate Structured Output.

The predefined data points were correctly recognized and extracted by the Basic LLM Chain, which is driven by the Ollama model and is controlled by Struc- tured Output Parser, of each PDF document that is processed in the Loop Over Items node.Figure 4.1 demonstrates the n8n result of one file that was processed (XYZ Inc.). The key-value pairs defined by the parser are stored in the output object in the form of companyName, csrFocusAreas, totalCsrSpend, keyInitiativesSummary, and overallStrategySynopsis. This middle-level finding proves that the LLM chain read the raw PDF text and interpreted the context correctly and organized it into the preferred schema.



output

companyName XYZ Inc.

csrFocusAreas Education, Healthcare, Community Development

totalCsrSpend ₹ 200 Crore

keyInitiativesSummary The company launched a mentorship program for underprivileged children and partnered with local NGOs to support community development projects. These initiatives aimed at promoting education and empowering marginalized communities.

overallStrategySynopsis Our CSR strategy focuses on creating a positive impact by investing in education, healthcare, and community development. We strive to be a responsible corporate citizen and make a meaningful difference in the lives of our stakeholders.

Fig. 3. Structured output of n8n for one processed document.

`	**CSR Focus Areas**	**Total CSR Spend**	**Key Initia- tives Sum- mary**	**Overall Strategy Synopsis**
ABC Corporation	Education, Environment, Healthcare	150 Crore	Our flagship program aims to provide education to underprivileged children. We also support local communities through environmental initiatives.	Our CSR strategy focuses on empowering marginalized sections of society and promoting sustainable development.
ABC Corporation	Education, Environment, Healthcare, Community Development	100 Crore	The company's key initiatives include a literacy program for underprivileged children and a community-based environmental project.	Our CSR strategy aims to create positive social impact through sustainable practices, stakeholder engagement, and strategic partnerships.
ABC Inc.	Education, Community Development, Environment	500 Crore	Implemented education programs for underprivileged children and supported community-development initiatives. Also, launched a sustainable environment project to reduce carbon footprint.	Our CSR strategy focuses on empowering local communities through education and sustainable development while promoting environmental stewardship.
XYZ Corporation Ltd.	Education, Healthcare, Community Development	200 Crore	The company launched a new mentorship program for underprivileged children and expanded its community	Our CSR strategy focuses on empowering local communities, promoting education and healthcare, and fostering a

Table 1. The resulting CSV file with aggregated results of all reports processed.\

4.2 Final Aggregated Output

The main objective of the workflow was to summarize the extracted information of all documents processed into one and consolidated file. This was done by the last nodes in the pipeline Convert to File and Read/Write Files from Disk, which combined the individual JSON output of each loop iteration and transformed it into a single CSV file. The final aggregated output is illustrated in Table 1. A row in the table represents a single PDF report that was worked with by the workflow. The columns are directly associated with the data fields that are extracted by the LLM. The successful creation of such CSV file confirms the end-to-end functionality of the pipeline. The workflow was used to batch-process several unstructured documents efficiently, and provide a structured dataset [18], which could be further analyzed, reported or ingested into a database [16].

5 Conclusion

In the CSR Summarizer project, it was accomplished to create a containerized and autonomous AI platform [6], which would help to deal with such accumulating issues as the time-consuming manual analysis of large and non-uniform Corporate Social Responsibility (CSR) reports. The architecture has a philosophy based on Sovereign AI, using a locally hosted Large Language Model (LLM)- an Ollama-provided Llama 3 model [19]- and an open-source workflow automation engine n8n [21]. Such a self-hosted strategy is essential in securing high levels of data confidentiality and sovereignty [7], avoiding the risk of data privacy, and reducing the challenges of vendor lock-in and unpredictable cost variability in the case of cloud-based AI services The automated workflow reads and extracts clean text via automated scanning of PDF reports into a workflow designed to facilitate the complex AI reasoning through the application of the LLM guided by a particular prompt as a kind of an expert financial analyst. A Structured Output Parser makes the resulting analysis consistent and produced in a structured format (JSON), which will be converted into one, aggregate CSV file and analyzed along with other business intelligence tools or databases [16]. The pipeline effectiveness is evidenced by its successful running and production validation that leads to a great deal of analysis-ready data that is structured out of unstructured CSR documents (Table 1). The CSR Summarizer, therefore, provides a privacy-conscious, useful, high-scaling technology of modern CSR analytics to allow the stakeholders to make more valid, quicker, and consistent decisions based on the data and reduce the anomalies and biases of manual analysis [10]. The architecture offers a solid blueprint of complete autonomous CSR intelligence of any industry.

6 Future Scope

Although the existing CSR Summarizer is a powerful, privacy-oriented data-extraction tool, future development can broaden its services to include a batch-processing service instead of a full-fledged, analytical service. It can be developed in the following directions:

- **Full Autonomous Ingestion:** Develop the manual trigger into an event driven system. The n8n workflow can be restructured to watch the cloud- based sources (such as a particular Google Drive folder, an S3 bucket, or an email inbox) and automatically initiate the processing pipeline when a new CSR report is uploaded.
- **Interactive RAG Pipeline:** The extracted text fragments can be embed- ded with the help of the local Ollama model and stored in a vector database (e.g., Qdrant, Chroma) instead of simply being outputted to a CSV. This would make the system a Retrieval-Augmented Generation (RAG) system where users can query the system using natural language questions (Com- pare the key initiatives of Company A and Company B or What was the total CSR spend of all companies in the healthcare sector?) and the system would synthesize-grounded answers.
- **High-order Analysis:** The task that the LLM could be used to do more than summarizing may be extended to encompass more analysis. This in- cludes:
  - **Sentiment Analysis:** Determining the mood and the feeling of various parts of the report.
  - **Framework Alignment:** Building prompts to rate or mark the adherence to the report to the standard sustainability frameworks (e.g., GRI, SASB).
  - **Longitudinal Analysis:** Generating reports of the same company across years to automatically define and summarize the trends in spending, ar- eas of focus, and strategic objectives.
- **Live Dashboard Integration:** The structured data of the workflow may not be terminated as a fixed CSV file, but instead may be ingested directly into a database (such as PostgreSQL or MongoDB) and displayed in an interactive real-time BI dashboard (such as Grafana or a custom web application), offering stakeholders a real-time interactive view of CSR data.
- **Moving to a Multi-Agent System:** Develop the linear workflow to a real multi-agent system. This might include separate ReportMonitoring agents, DataExtraction agents, ComparativeAnalysis agents, and StakeholderAlerting agents so that more complex, dynamic and parallel processing of corporate data may be done.

References

1. A. Patel et al., ”AI-Driven Corporate Report Analysis: Automating Sustainability Insights.” Springer, 2022.
2. S. Lee and R. Kumar, ”Corporate Intelligence Systems Using NLP Techniques,” *IEEE Access*, 2023.
3. Ollama Inc., ”Llama3 Model Architecture and Deployment Guide,” 2024.
4. n8n Documentation, ”Self-Hosted Workflow Automation Platform,” 2024.
5. PyMuPDF and pdfplumber, ”Python Libraries for Document Text Extraction,” 2024.

6. P. Sharma and R. K. Singh, “SustAI-SCM: Intelligent Supply Chain Process Automation with Agentic AI for Sustainability and Cost Efficiency,” *Sustainability (MDPI)*, vol. 17, no. 5, pp. 2501–2515, Jan. 2025.
7. A. Davies, ”AI Agents and Agentic Systems: Redefining Human Contribution, Autonomy, Industry Structures, and Governance,” *Cronfa - Swansea University Repository*, Nov. 2024.
8. Google Cloud, “Real-world gen AI use cases from the world’s leading organiza- tions (Focus on Data Agents),” *Google Cloud Blog*, Mar. 2025. [Online]. Available: https://cloud.google.com/blog/... (Hypothetical Link).
9. IBM Institute for Business Value, *Orchestrating Agentic AI for Intelligent Business Operations*, IBM Corp., White Paper, 2025.
10. A. B. Chen and Z. M. Li, “A Detailed Study on LLM Biases Concerning Corporate Social Responsibility and Green Supply Chains,” *arXiv e-prints*, abs/2501.01234, Feb. 2025.
11. S. R. Patel, “Large Language Models for Social Research: Potentials and Chal- lenges,” *YouTube/MZES Methods Bites*, Video Presentation, 2025.
12. M. C. Liu and S. Q. Han, ”Research on the impact of artificial intelligence on corporate sustainability performance and its mechanisms: an empirical analysis based on text analysis,” *Res. Gate Proc.*, pp. 101–115, 2025.
13. J. Smith, “Harnessing Big Data and AI to Revolutionize Sustainability Accounting and Integrated Corporate Financial Reporting,” *Res. Gate Proc.*, pp. 50–65, 2025.
14. R. Kumar, “Artificial Intelligence as a Tool to Evaluate Corporate Sustainability Reporting,” in *Advancements in Financial Technology*, Springer Nature Switzerland, 2025, pp. 300–320.
15. E. V. Johnson, “A Meta-Analysis of the Economic, Social, Legal, and Cultural Impacts of Widespread Adoption of Large Language Models such as ChatGPT,” *OxJournal of Econ. Stud.*, vol. 45, no. 2, pp. 112–130, Apr. 2024.
16. K. L. Wong, ”Machine Learning and AI in Business Intelligence: Trends and Opportunities,” *Int. J. of Comput.*, vol. 18, no. 4, pp. 450–465, 2023.
17. S. G. Ross, ”The State of the Art of Natural Language Processing—A Systematic Automated Review of NLP Literature Using NLP Techniques,” *Comp. Ling. NLP Series*, MIT Press Direct, 2024.
18. Z. H. Khan and T. Ali, “Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective,” *PMC - PubMed Central*, vol. 45, no. 3, pp. 101–120, Mar. 2021.
19. Meta AI, *Llama 3 Model Card/Release Notes*, Meta AI Technical Report, 2024.
20. Ollama Inc., *Ollama Documentation and Deployment Guides*, Ollama Technical Manual, 2024. [Online]. Available: https://ollama.com/docs (Hypothetical Link).
21. n8n, *n8n Documentation: Advanced Workflow Chaining and Custom Code Integration*, n8n Inc., Technical Documentation, 2024.
22. PyMuPDF Developers and pdfplumber Contributors, *PyMuPDF/pdfplumber Of- ficial Documentation and GitHub Repositories*, PyPI GitHub Repositories, 2024.
23. M. F. Zaki and H. Q. Noor, “Enhancing Communication Networks in the New Era with Artificial Intelligence: Techniques, Applications, and Future Directions,” *Computers (MDPI)*, vol. 14, no. 1, pp. 105–120, Feb. 2025.
24. M. O. Al-Qadi, “AI-Driven Financial Transparency and Corporate Governance: Enhancing Accounting Practices with Evidence from Jordan,” *J. Risk Financial Manag. (MDPI)*, vol. 18, no. 3, pp. 150–165, Mar. 2025.
25. A. K. Gupta et al., “Development and Validation of a Machine Learning Model That Predicts Short Inpatient Stays Among Urgent Admissions,” *Emerg. Care Med.*, vol. 20, no. 1, pp. 50–62, Jan. 2025.