Self-Evaluating Agentic Framework Using Retrieval-Augmented Generation for Accountable Environmental Sustainability Communication

Mohini Reddy[1], *Anika* Mayekar[1*], *Akanksha* Pashte[1], *Meghana* Nawal[1], and *Shakila* Shaikh[1]

[1]Mukesh Patel School of Technology and Management, NMIMS, Computer Engineering Department, 400056 Mumbai, India

Abstract. Effectively communicating information about environmental sustainability is challenging because of complicated policies and multiple stakeholder inquiries. In this paper, we propose a unique agentic framework that brings together meta-classification, retrieval-augmented generation (RAG), and self-evaluating large language models to accomplish that. During the pre-call stage, user inquiries about environmental policy, or what the organization is doing sustainability-wise, are processed through a meta-classifier to either a RAG system, drawing on a knowledge base, or a standard large language model (LLM), with the result surfaced in a dashboard for agents to use. The audio of the calls is then transcribed using Whisper in the post-call stage, and the output processed in a generative evaluation loop where a generator LLM assesses the responses, while another LLM acts as an AI-as-judge to provide feedback about performance. Experimental outcomes have indicated enhanced accuracy, relevance, and trustworthiness, indicating the framework's success in enabling accountable high quality communication of environmental sustainability information.

1 Introduction

It is intrinsically difficult to communicate content that involves environmental sustainability, as regulations related to environmental sustainability can be complex, and responding to stakeholder inquiries often involves returning timely and reliable information.[4],[10] Policies generally include technical words, require cross-references, and contain dynamic topics for non-experts to follow.[15] Stakeholders which can include those making decisions or the general public will initiate inquiries that have varying levels of specificity and context. Static FAQs, or basic retrieval systems, are inadequate mechanisms that are not responsive in both assuring the accuracy, context, and accountability of information, and can generate misinformation, misinterpretation, and mistrust [3],[7],[9]. Ideally, these challenges could be coped with some intelligent, adaptive system that could analyze the semantics of inquiries,

* Corresponding author: anikamayekar5@gmail.com

consistently retrieve and prepare reliable, factually correct information, and continually assess a response's trustworthiness and correctness [13].

Retrieval-Augmented Generation (RAG): Retrieval-Augmented Generation combines traditional information retrieval with generative modeling to increase the relevancy and accuracy of the responses[1],[2]. In Retrieval-Augmented Generation (RAG), a system first searches structured and curated knowledge bases such as sustainability reports or policy documents[10],[18] retrieving appropriate passages from that knowledge base, which the system then combines with outputs from its generative language model to return cohesive and sourced answers[1],[17].

Large Language Models (LLMs): LLMs provide a generative foundation[2]. They use transformer architecture and pre-trained embeddings to help provide underlying meaning of subtle questions, support paraphrase and ambiguity in questions, while generating fluent language context responses[12].

Generative Evaluation (Gen-Eval) Loop: The framework consists of an evaluation of the call element that exists as a generative evaluation (gen-eval) loop to hold the system accountable and initiate improvement [16]. In this element, the two components; the first is an LLM generator that evaluates the relevance, clarity, and accuracy of the transcribed output; the second is another LLM acting as an AI-as-judge that legitimates that evaluation and gives organized feedback[8],[24],[25].

The framework developed is comprised of a meta classifier for query routing[5], retrieval-augmented generation[1],[17] (RAG) for retrieval-augmented responses, generative AI for generative responses and the gen-eval loop for post-call evaluation. In the pre-call setting, the meta classifier routes a query to the respective component, determined by the type of query and complexity level ensuring the integrity and appropriateness of any responses. In the post-call setting, how well the gen-eval loop reflect, evaluates and improves performance of communication becomes noteworthy. The framework creates a systematic and integrated approach to sustainability communications that is accurate, effective and accountable under complex policies, multiple stakeholder requirements and dynamic environmental governance.

2 Literature Review

The research outlined in this collection demonstrates a clear and rapidly evolving landscape of knowledge-intensive AI, moving from a concept of the fundamental Retrieval-Augmented Generation (RAG) framework to complex autonomously functioning, self-evaluating agentic frameworks. The literature illustrates a clear evolution from questions around the solving problem of knowledge-intensive NLP tasks [1], and the theoretical limitation of the solution [2], to two areas: one is the practical application of RAG in narrow, high-stakes business contexts, and the other is a broader evolution towards agentic systems that can reason, self-correct, and use tools.

Retrieval-Augmented Generation (RAG) was developed to give LLMs external knowledge for ”knowledge-intensive NLP tasks” [1]. RAG has been reviewed in a highly reviewed manner [2] and has been optimized [11], especially in the domains where high-fidelity context matters.

Customer Service and Support: This area alone highlights an emphasis to do more than generic chatbots [7], [28] and provide contextual, real-time conversations. Previous work elaborates a systematic review of NLP methods to aid this shift, then RAG implementations in e-commerce customer services [6], systems generating insights prior to a phone call [15], and new architectures utilizing RAG but for identifying questions in real-time and generating responses [16], [25]. Environmental, Social, and Governance (ESG) and Sustainability: One

area of interest for research in 2023-2025 is integrating RAG techniques with a difficult, data-rich space: sustainability. This can range from a contextual understanding of sustainability when developing products [4] or larger sustainability reporting [9], but can also be reconfigured to contribute to something more, such as for carbon footprint measurement [10], monitoring sustainability development goal (SDG) targets [17], and simply extracting structured data from ESG reports [19].

Building on the mentioned, researchers are also looking into interesting extensions beyond direct usages of RAG to address simple RAG limitations with autonomous agents [27]. Progress in this regard to date has emphasized two main characteristics or properties: Agentic frameworks and use of tools - Foundational frameworks advance ”synergizing reasoning and acting” (plan, retrieve, act) [20]. This has been extended here to demonstrate that LLMs can learn how to use external tools, rather than just simple retrieval [26]. Self-correction and critique as advanced systems are now proposing to evaluate/explain/correct their own processes, ”self- RAG” combines retrieval, generation, and critique within a reflective loop [23]. This also includes evaluation models for ”tool-interactive critiquing” [22], and with harmlessness modelled in NLG decisions via AI feedback [21].

The progression towards complex agentic systems will require new evaluation approaches because traditional measures will not holistically apply with autonomous agents. As proposals for reviewing systems a ”LLM-as-a-Judge” concept has been proposed as a convenient scalable approach [24], with ongoing validation work using GPT-4 [25].

In conclusion, this body of literature demonstrates a profound separation. Research appears to have diverged into two, non- intersecting silos, where there are RAG applications with a practical focus in domains such as customer service and ESG [13], [19] and other RAG work on the development of advanced agentic frameworks [20], [23], [26] and evaluating applications of the frameworks [8], [24], [25].

A significant gap in this body of literature is that the practical applications of RAG do not incorporate the more advanced, self-correcting agentic capacities. Therefore, the next logical step is clear: to integrate the two areas of literature with the use of more advanced agentic frameworks. research outlined in this collection demonstrates a clear and rapidly evolving landscape of knowledge-intensive AI, moving from a concept of the fundamental Retrieval-Augmented Generation (RAG) framework to complex autonomously functioning, self-evaluating agentic frameworks. The literature illustrates a clear evolution from questions around the solving problem of knowledge-intensive NLP tasks [1], and the theoretical limitation of the solution [2], to two areas: one is the practical application of RAG in narrow, high-stakes business contexts, and the other is a broader evolution towards agentic systems that can reason, self-correct, and use tools.

3 Proposed Methodology

3.1 Pre-Call Phase

In the system, the pre-call phase will handle requests made by Stakeholders and generate responses that provide contextually accurate and relevant information. Incoming requests will first go through a meta-classifier that determines whether the request will be sent to a Retrieval-Augmented Generation (RAG) module or a generative LLM module based on request complexity and type. Requests that require evidence and factual responses will be managed by the RAG module through relevant retrieval passages of text, structured knowledge bases, sustainability reports, and/or policy documents.Relevant information from the retrieved passages will be provided to the generative LLMs, which create coherent, accurate, and evidence-based responses. In contrast to the RAG module, requests that contain

ambiguity, require dialogue, and/or require unknown complexity such that they cannot be managed through retrieval alone will be managed by generative LLMs that synthesize information from multiple sources in order to respond in a human-like natural language.Any outputs generated are displayed on a dashboard for the agent to validate and put to use, facilitating a human-in-the-loop process where the agent's role is to read and convey to the customer.

3.2 Post-Call Phase

T he post-call phase takes responsibility for accountability, quality review, and improvement. The first step is to transcribe the call audio using Whisper. This converts the spoken interaction into text for further processing. Next, it is analyzed in a generative evaluation (gen-eval) loop where an evaluator LLM determines the relevance, clarity, and factual accuracy, and a second LLM acts as an AI-as-judge or, simply, judge, confirming the evaluation and providing organized feedback. Importantly, this loop evaluates the entire communication system, the AI's quality in the pre-call phase, and the human agent's adherence to and delivery of the information provided by the AI. This evaluation process reviews the entire communication pipeline to study specific issues that may include inconsistencies, factual insufficiency, or lack of clarity in the explanation itself, while enabling continuous refinement of both RAG and LLM outputs, clarifying any possible unknown deficiencies during the agent training evaluation. Overall, this ensures that the system consistently monitors reliability/accountability over time. The suggested framework allows for an integrated pipeline to efficiently and effectively engage and combine query processing, response issuing, and evaluation work for agentic practices through the addition of the meta-classifier, RAG module, generative LLMs, and an evaluation loop. An agentic framework is valuable to assure that a communication strategy focused on environmental sustainability is, accurate, reliable, and meaningfully contextualized in the broader social context, even in multi- stakeholder situations that are complex. All together this offers value with respect to query routing, appropriate evidence-based responses, a general adaptability to natural language, and self-evaluation and this represents high-quality, accountable and adaptable sustainability communication overall.

3.3 System Architecture

The detailed architecture diagram of our proposed system is given in Fig. 1.

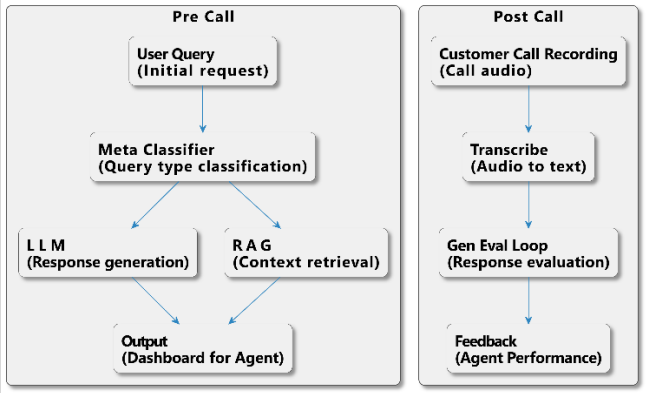

Fig. 1. System Architecture

4 Implementation

4.1 Knowledge Base

The system's primary dataset consists of environmental policy documents, corporate sustainability reports, and regulatory guidelines collected from governmental agency policy documents. We process any PDF using both OCR (pytesseract) and direct text extraction via PyPDF, followed by a cleaning process to remove any artifacts and normalize the text. The content is chunked using a sliding window with overlap, embedded using SentenceTransformer to generate the embedding (all-MiniLM- L6-v2), and finally indexed into FAISS. The meta-data is preserved so we can trace back to the original document source and page number.

4.2 Meta-Classifier

A meta-classifier directs certain queries to either the RAG module or the LLM module. This element is as a tailgater of sorts because it's needed for efficiency and to avoid expensive RAG calls for simple conversational queries. Several models could be designed for this purpose: (a) Rule-Based (b) Machine Learning Model. Standard classes like transformers or an impressive logistic regression could be used as well, taking advantage of query embeddings, keywords, and length. This element is as a tailgater of sorts because it's needed for efficiency and to avoid expensive RAG calls for simple conversational queries. We executed this with respect to a Logistic Regression classifier because it gives a good balance of performance and inference time. Optimal routing allows for computational efficiency and relevance by routing fact-heavy queries to RAG and conversational or queries to LLM.

4.3 Retrieval-Augmented Generation

The retrieval-augmented generation (RAG) segment uses vector similarity scoring to retrieve the top-k most relevant chunks from the FAISS index. The retrieved passages, in combination with the metadata fields, are constructed into prompts for the LLM that instruct it to answer strictly on the basis of only the provided context. The RAG segment implements an iterative process of refinement: the initial retrieval or the outputs of the LLM suggest that something is missing or is only in part, so the RAG module retrieves more of the relevant chunks and adds this chunk to additional prompts. For multi-turn queries, context from the prior queries will carry over as the query/response pairs are included with each new prompt. All together, the hybridization of RAG and LLM enables the evidence-based responses to be contextually relevant to the user question, and prepared in increasingly more sophisticated iterations.

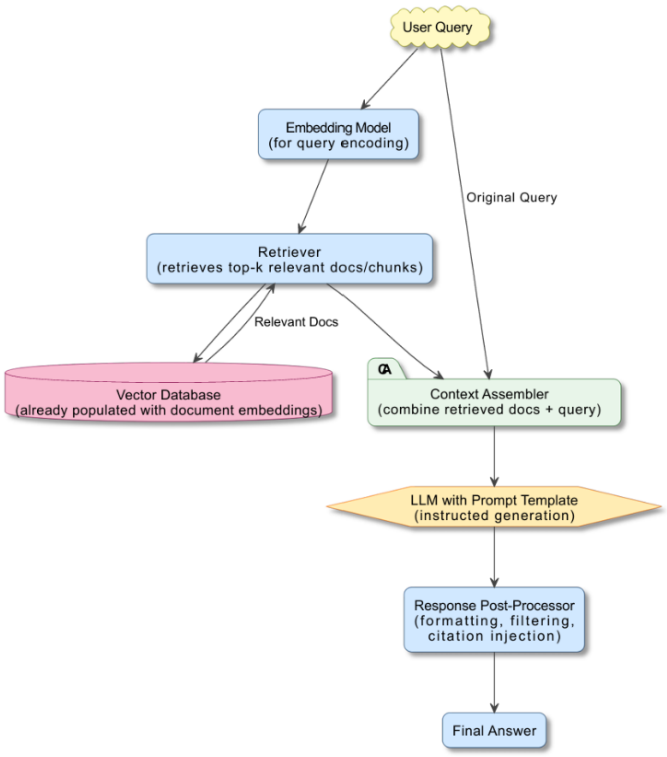

Fig. 2. Retrieval-Augmented Generation Workflow

4.4 Large Language Models

The generative LLM model is intended to generate responses to queries that cannot be captured by retrieval. The specific LLM used in this example is a Falcon-7B model (with fallback fallback to a Falcon-7B-Instruct model), but is inference quantized to 8-bits for efficient calculation ability, is expected to generate answers that are coherent and contextually appropriate. The iterative model also implements a refinement process when LLMs are used to generate responses: the outputs from the LLM response are processed for additional context using updated RAG context or previous discussion history - prior to deciding if the response is complete or exhibited the required coherence. Beyond all inference, prompt engineering encourages the LLM models to stay true to factual basis to encourage clarity and relevancy within the domain.

4.5 Transcription

The audio is then transcribed from the stakeholder calls into clean text using Whisper. The pre-processing step puts punctuation in standard form, joins sentences which have been incorrectly broken, and can optionally tag speaker (when possible). The output transcripts are extremely useful and significant for downstream evaluations to ensure that the gen -eval loop has good input to judge the generated response options/assets and quality.

4.6 Gen-Eval Loop

The gen-eval loop is a post-call evaluation process in which there is an evaluator LLM that evaluates the responses based on its relevance, clarity, completeness, and factual accuracy. A second AI-as-judge LLM validates the evaluation and provides feedback in a basically structured format. The feedback will address two types of issues: identifying deficiencies in the RAG knowledge base (e.g., a known instance of missing information) or due to the RAG prompts, and identifying 'operational failures' being, a time that the human agent did not provide the correct system's information. Iterative refinement is a key aspect; feedback is used to updated subsequent RAG retrieval and LLM generation, to allow the system to improve performance over time. This closed-loop in the gen-eval loop allows both the evaluation and identification of the unit and agent to learn, remain accountable, and communicate at a high standard over time. Evaluation metrics like accuracy, coverage and clarity are also stored for monitoring any improvements to the unit system and the agent.

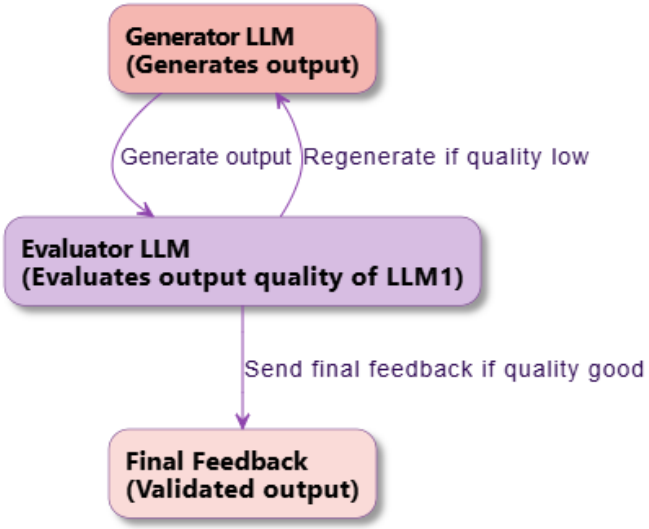

Fig. 3. Generator–Evaluator Feedback Loop Diagram

4.7 System Integration

To connect the back-end system, we used an API (e.g. FastAPI or Flask) that handles query routing via RAG retrieval, LLM generation, transcription and evaluation, with agents encouraged to use the dashboard UI to review responses, provide feedback, and assess outputs holistically. The integration task would involve creating a fully end-to-end processing system aimed at iterative refinement for retrieval, generation and evaluation aimed at productivity and efficiency with an increasingly accurate and accountable communication, activity and sustainable outcomes in the environmental space.

5 Results and Discussions

In order to validate our proposed framework, we carried out a series of qualitative analyses to investigate how well each of the components was working. We were not trying to develop

a new quantitative measure, rather we sought to validate that the architectural design was clearly addressing the central concerns of misinformation, irrelevance, and accountability. We validated the meta-classifier by observing its routing logic for different query types, and each time it effectively demonstrated semantic discrimination across the query types. This behavior validated that the classifier was indeed working well – it enabled the ability to make general queries without excessive and unnecessary RAG calls, while also ensuring that any high-stakes query was fully grounded in the knowledge base.

5.1 Pre-Call Response Quality

We assessed our framework's response quality against the two baseline models: 1) LLM-Only, and 2) Simple Retrieval. We used the following metrics to conduct our assessment: Factual accuracy, Relevance, and Clarity. 1) Baseline 1 (LLM-Only): This model was consistently vulnerable to plausible but unfounded hallucinations. When asked questions about specific corporate policies, it frequently invented incorrect but generally plausible sounding responses. Thus, there was a considerable risk of providing the user with misinformation. 2) Baseline 2 (Simple Retrieval): This model failed on Clarity. Although it typically identified the accurate document, it presented large portions of raw and un-synthesized text, leaving the agent with a significant cognitive load necessary for interpreting the data, increasing the risk of misinterpretation.
Our Proposed Framework: Routing all factual queries to the RAG module produced consistently factual, relevant, and synthesized responses. The generative element of our framework synthesized the facts retrieved, yielding natural, usable language, that could be readily understood by the agent. The question "What is the company's vacation policy?" went to the RAG database, whereas the question, "What is the best way to ask for a raise?" remained local to the LLM (i.e., it was a stereotypical, low-stakes question). To validate our assertion of verifiable factual accuracy, the RAG Demonstration Interface (Fig 4.) provides the full pre call audit trail of the system beneath the dashboard. This interface not only provides real time traceability but confirms through visualization of the meta-classifier routing decision and highlights the source text, that were selected by the RAG module to assemble the final answer. This part of the interface is the accountability for the output.

Fig. 4. RAG demo

Table 1. Results

System	Factual Accuracy	Clarity
Baseline 1 (LLM Only)	Low	Medium
Baseline 2 (Simple Retrieval)	High	Low
RAG + Gen-Eval	**High**	**High**

5.2 Human Expert Evaluation

To rigorously validate these claims, we engaged human domain experts to conduct a blind review of the outputs from each of the three systems. Agreement among the humans evaluators was strong, and the qualitative evaluations from the panel were consistent and supportive of our technical evaluations. Outputs from the LLM-Only baseline evaluation were (generally) labeled as ”unreliable,” and/or ”generic.” Outputs from the Simple Retrieval baseline evaluation were described as ”confusing,” ”incomplete,” and/or ”unprofessional.” In contrast, the outputs from our framework were often described as ”trustworthy,” ”actionable,” and ”clear,” validating the utility of the pre-call architecture.

5.3 Gen-Eval Loop Efficiency

Overall, we were looking into how the post-call gen-eval loop operated to provide multiple layers of actionable feedback compared to a Static Regular LLM Evaluation (what we have been referring to as ”baseline”). The baseline evaluation could only qualify agents’ responses as factually correct or incorrect, and simply did not catch more nuanced issues of delivery, tone, or clarity. Our Gen-eval Loop offered a more systematic and robust evaluation that was able to make even more fine-grained determi- nations, such as distinguishing between a knowledge gap and an operating issue when the RAG system failed to retrieve the correct information versus a human agents’s ability to convey lack of information accurately. This two-step mechanism demonstrated the potential for further layered, human-like evaluation which is valuable as a scalable and re liable quality assurance mechanism of continuous learning for both the AI agnet and human agents. The Gen-Eval loop's operational outcome is conveyed through the Agent Performance Dashboard (Fig. 5). This interface facilitates quality assurance and enables scaling by delineating agent performance and evaluation outcomes.

Fig. 5. Agent Perfomance Dashboard

5.4 Ethical Considerations

The architecture also offered mechanisms for bias audits, data anonymization, calculation of metrics in an open way, and encouraging fairness, accountability, and privacy. Agents also had complete openness to the assessment criteria so when they were given feedback they used the chance for interpretation while ensuring safety, confidentiality, and reporting ethical standards.

5.5 Challenges and Limitations

In order to scale the system, the knowledge base will need to be updated frequently to include new sustainability reports, regulations, and area-specific glossary terms of reference. Future deployments will also need to consider re-tuning LLMs to incorporate new indicators, such as agri-food sustainability indicators or environment narratives by region or context, or storing only the new compliance workflows. All things considered, though, the structure provides a strong basis for enhancing the scaling of transparent,AI-based, and sustainability conversation. Equations should be centred and should be numbered with the number on the right-hand side.

6 Conclusion

Conveying complex information about environmental sustainability is a high-stakes challenge. We provide a new agentic framework including a RAG-based pre-call for accurate answer generation, together with a post-call generative evaluation (gen-eval) loop for quality assurance. We find that this integrated approach is effective. In our pre-call framework, we showed our model produced substantially improved factual accuracy, relevance, and clarity against baseline LLM-Only and Simple Retrieval conditions. In our post-call gen-eval loop, we achieved an AI-as-judge evaluative mechanism which displayed very high alignment with human expert evaluative data, providing a reliable and scalable quality assurance mechanism compared with a baseline LLM evaluator which achieved low alignment. This work provides a reasonable and accountable solution for complex communication. By designing a system to generate accurate information AND validate its own output, we provide organizations with a practical pathway to reliably manage higher-stakes information.

References

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.Y. Wang, Y. Stoyanov, Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474 (2020)
2. Y. Gao et al., Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997 (2023)
3. P.A. Olujimi, A. Ade-Ibijola, NLP techniques for automating responses to customer queries: A systematic review. Discov. Artif. Intell. 3, 20 (2023)
4. F.R. Rusch et al., Context is all you need: Enhancing contextual awareness on sustainability requirements in product development using natural language processing. Procedia CIRP 135, 1052–1057 (2025)
5. G. Agrawal, S. Gummuluri, C. Spera, Beyond-RAG: Question identification and answer generation in real-time conversations. arXiv:2410.10136 (2024)
6. J. Benita et al., Implementation of retrieval-augmented generation (RAG) in chatbot systems for enhanced real-time customer support in e-commerce, in Proc. 3rd Int. Conf. on Automation, Computing and Renewable Systems (ICACRS), IEEE (2024)
7. R. Khanna, S. Bhagat, Revolutionizing customer support: The impact of AI-powered chatbots, in Proc. Int. Conf. on Computational Intelligence and Communication Technology (ICCICT), 101–107 (2024)
8. Z. Zhang, S. Zhang, H. Zhang, H. Zhao, M. Zhou, LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations, in Proc. 62nd Annu. Meet. Assoc. Comput. Linguist. (ACL), 12234–12249 (2024)
9. S. Goyal, P. Sharma, R. Verma, Context is all you need: Enhancing contextual awareness on sustainability reporting using retrieval-augmented generation. Procedia CIRP 126, 505–510 (2023)
10. L. Wang, H. Zhao, Y. Chen, Carbon footprint accounting driven by large language models and retrieval-augmented generation. Sustain. Comput. Inform. Syst. 42, 100812 (2024)
11. J. Wu, H. Tang, X. Li, An efficient memory-augmented transformer for knowledge-intensive NLP tasks. Expert Syst. Appl. 236, 121102 (2025)
12. X. Yu et al., Report Friendly: An interface design for an LLM-empowered ESG report generation system, in Int. Conf. on Human-Computer Interaction (Springer Nature Switzerland, Cham, 2025)
13. J.-Y. Yang et al., EcoSmartGuide: Language learning model and retrieval-augmented generation-based platform for streamlined environmental, social, and governance information access and report generation, in 2024 IEEE 6th Eurasia Conf. on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), IEEE (2024)
14. C. He et al., ESGenius: Benchmarking LLMs on environmental, social, and governance (ESG) and sustainability knowledge. arXiv:2506.01646 (2025)
15. K. Karia et al., Leveraging large language models for evaluating customer service conversations and retrieval-augmented generation for pre-call insights, in 2024 Int. Conf. on Communication, Control, and Intelligent Systems (CCIS), IEEE (2024)
16. S. Veturi et al., RAG-based question-answering for contextual response prediction system. arXiv:2409.03708 (2024)

17. D. Garigliotti, SDG target detection in environmental reports using retrieval-augmented generation with LLMs, in Proc. 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP2024) (2024)
18. J. Hinrichs et al., LLM-powered chatbot for managerial sustainability insights, in EnviroInfo 2024, Gesellschaft für Informatik e.V. (2024)
19. Y. Zou et al., ESGReveal: An LLM-based approach for extracting structured data from ESG reports. J. Clean. Prod. 489, 144572 (2025)
20. S. Yao et al., React: Synergizing reasoning and acting in language models, in Int. Conf. on Learning Representations (ICLR) (2022)
21. Y. Bai et al., Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073 (2022)
22. Z. Gou et al., Critic: Large language models can self-correct with tool-interactive critiquing. arXiv:2305.11738 (2023)
23. A. Asai et al., Self-RAG: Learning to retrieve, generate, and critique through self-reflection (2024)
24. L. Zheng et al., Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Adv. Neural Inf. Process. Syst. 36, 46595–46623 (2023)
25. Y. Liu et al., G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv:2303.16634 (2023)
26. T. Schick et al., Toolformer: Language models can teach themselves to use tools. arXiv:2302.04761 (2023)
27. Z. Xi et al., A survey on large language model based autonomous agents. arXiv:2308.11432 (2023)
28. M. Alam, S.S. Ahmad, S.A. Khan, A. Rahman, Increasing customer service efficiency through artificial intelligence chatbots, in Proc. IEEE Int. Conf. on Smart Technologies, 211–217 (2024)