ClimateAI: A multimodal Large Language Model to generate Project Design Documents that comply with the Verified Carbon Standard

Chuqi Wang, Shengtong Sun, Zhihang Feng,
Baijiang Wang, Yu Song, Yijia Sun

Abstract

This report introduces ClimateAI, a multimodal large language model system designed to generate Project Design Documents (PDDs) compliant with the Verified Carbon Standard (VCS). By leveraging a Retrieval-Augmented Generation (RAG) framework integrated with large language models (LLMs), ClimateAI addresses the complexities and inefficiencies inherent in PDD development. The system uses advanced data collection, preprocessing, and vectorization techniques to construct a robust knowledge base, enabling accurate and context-aware PDD generation. Evaluations highlight the system's ability to ensure compliance with VCS guidelines, clarity, and technical accuracy. This innovation represents a significant step in streamlining the carbon market's documentation processes, offering a scalable, efficient, and reliable solution for managing environmental projects.

A live demonstration of ClimateAI, showcasing its ability to generate PDDs aligned with VCS standards, is available for exploration. This interactive demo allows users to experience the system's functionality firsthand, including its capability to handle diverse environmental project scenarios and produce context-aware, technically accurate documentation. You can access the demonstration at: Climate AI Demo

1 Introduction

1.1 Overview

In the context of global initiatives aimed at combating climate change, the significance of carbon markets has grown increasingly prominent. These markets facilitate the trading of carbon credits, which are generated through projects designed to reduce greenhouse gas emissions. Effective management and documentation of such projects are essential to ensure their integrity and uphold the credibility of the credits produced. Project Description Documents (PDDs) serve as a comprehensive record detailing the planning, implementation, and anticipated impacts of these projects. The Verified Carbon Standard (VCS) acts as a benchmark that guarantees projects are conceived and executed in a manner that delivers genuine and verifiable environmental benefits.

1.2 Problem Statement

Despite the pivotal role that Project Design Documents (PDDs) play in carbon trading, the process of developing these documents is frequently characterized by complexity, time consumption, and a susceptibility to errors. Such challenges can result in delays during project validation, compliance issues, and potential skepticism from investors and regulators regarding the credibility of reported emission reductions. Furthermore, the dynamic nature of environmental projects coupled with evolving regulatory requirements necessitates regular updates to PDDs. This not only adds to the administrative burden but also heightens the risk of non-compliance with standards such as the Verified Carbon Standard (VCS).

1.3 Background Information

The cap-and-trade system serves as a vital regulatory mechanism employed globally to mitigate greenhouse gas emissions in a cost-effective manner. This market-oriented approach entails establishing a limit on the total volume of greenhouse gases that can be emitted by various entities, including corporations and entire sectors. Entities are allocated emissions allowances, which they may trade based on their capacity to meet or surpass their designated emissions targets. This framework not only provides financial incentives for reducing emissions but also fosters investments in sustainable technologies and practices.

The Project Design Documents (PDDs) play a pivotal role in the cap-and-trade system, particularly within carbon markets. They serve to validate

emission reductions by providing comprehensive accounts of carbon reduction projects, encompassing methodologies, anticipated outcomes, and monitoring strategies. This thorough documentation is essential for substantiating claims of emission reductions and constitutes the primary basis for the issuance of carbon credits.

2 Datasets

Our dataset is consist of 1824 project description documents that are registered from the Verified Carbon Standard (VCS) in Verra Registry.

2.1 Exploratory Data Analysis

2.1.1 EDA: Frequent words from the datasets

The EDA conducted on the datasets involved extracting frequent words, which provided significant insights into the prevailing themes and terminologies across various PDDs. This analysis has led to the categorization of keywords into seven distinct groups, each representing a critical aspect of the documentation required under the VCS and other related carbon management frameworks, as shown in Table 1

Table 1: Categorization of Keywords from EDA Analysis.

Category	Keywords
Project Management and Description	project, description, activity, activities, methodology, procedures
Technical and Measurement Methods	data, emissions, emission, baseline, calculation, parameter, factor, measurement, methodology, efficiency
Specific Measures and Activities	monitoring, electricity, fuel, biomass, methane, energy, waste, manure, cookstove, biogas
Quantitative Data and Statistics	number, total, annual, fraction, estimated, calculated
Policies and Regulations	vcs, ics, applicable, scenario, grid, carbon, crediting
Project Duration and Timing	year, period, years, time
General and Administrative Language	shall, used, use, using, required, available

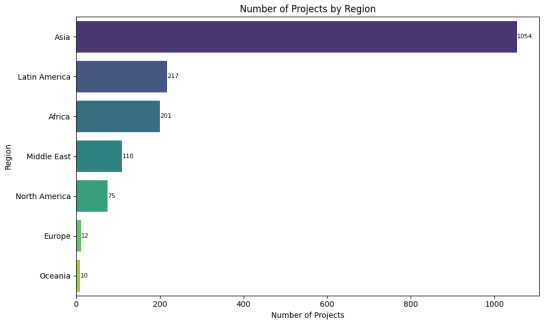
Analyzing the frequency of words in our datasets is crucial for effectively developing our AI chatbot. This process enables us to identify the primary topics discussed within the climate industry, such as renewable energy, energy efficiency, and waste management. A comprehensive understanding of these key areas allows us to customize the chatbot to address the specific needs of this sector. By recognizing significant terms, the chatbot can generate responses that are more relevant and beneficial to users, thereby enhancing their overall experience. Furthermore, this word analysis may reveal potential gaps in information coverage. If certain important terms are infrequently represented, it could indicate that our dataset is incomplete or fails to encompass all essential aspects of the climate industry. Consequently, we can focus on addressing these deficiencies to enhance the chatbot's intelligence and reliability, ultimately providing improved support for users seeking climate-related information.

2.1.2 EDA: distribution of projects by region

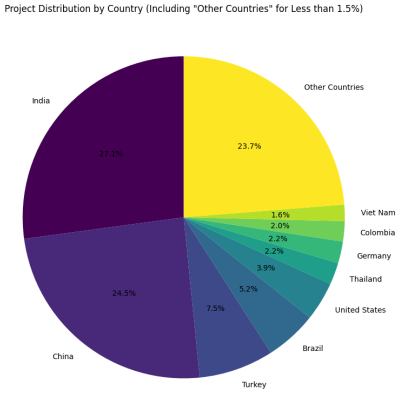
Figure 1 illustrates significant geographic disparities in climate-related initiatives. Asia leads with a substantial total of 1,054 projects, reflecting its extensive industrial base and the urgent need to address environmental impacts. This is followed by Latin America, which has initiated 217 projects, highlighting an increasing emphasis on sustainable practices and climate change mitigation. Africa and the Middle East also demonstrate considerable activity, contributing 201 and 110 projects respectively. This trend likely indicates a growing awareness of climate issues as well as the influence of international development aid and local environmental policies. In contrast, North America and Europe exhibit fewer initiatives, with only 75 and 12 projects respectively; this may be attributed to the maturity of existing environmental programs or differing regulatory frameworks and market dynamics. Oceania presents the fewest initiatives at just 10 projects, potentially suggesting smaller market sizes or alternative priorities within environmental policy. This figure is essential for understanding where global climate action is concentrated and can inform strategic decisions regarding future investments and policy efforts in climate project development as well as sustainability initiatives.

2.1.3 EDA: distribution of projects by type

Figure 2 underscores a predominant emphasis on energy industries, which encompass both renewable and non-renewable sources, with a total of 1,028 projects. This highlights the critical importance of transformations within the energy sector in addressing climate change. Additionally, agriculture, forestry, and land use are significant contributors as well, featuring 237 projects that reflect a commitment to sustainable practices aimed at enhancing carbon sequestration. Waste management follows closely with 228 projects, illustrating its potential role in emissions reduction through recycling initiatives and methane capture technologies. Projects focused on energy demand amount to 203, drawing attention to efforts directed towards improving energy efficiency and conservation measures. Other sectors such as mining and manufacturing remain active but are less prominently featured; this suggests potential areas for future climate action initiatives.



(a) Number of Projects by Region



(b) Project Distribution by Country

Figure 1: Overview of Project Distribution by Region and Country

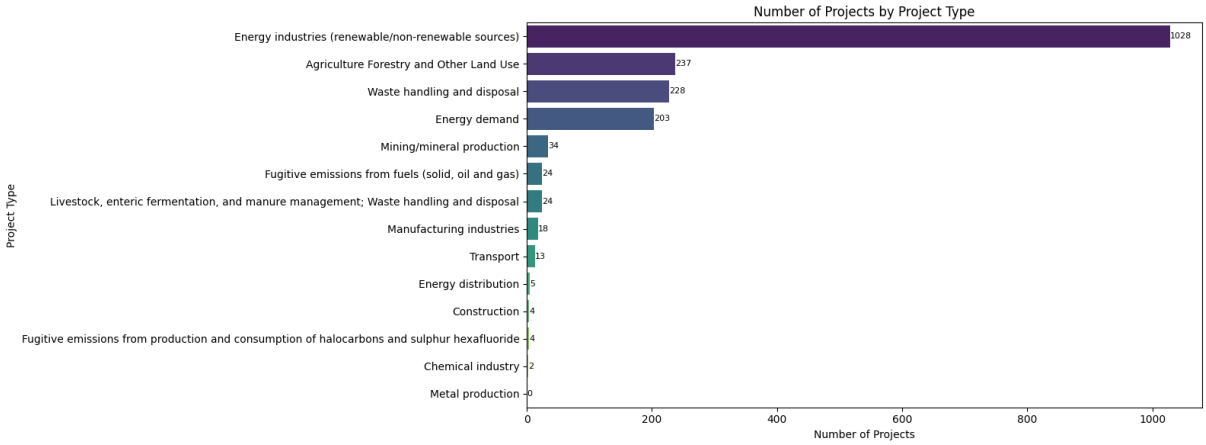


Figure 2: Number of Projects by Project Type

.

3 Methods

3.1 Overview

The Climate AI project aimed to develop an AI-driven system capable of autonomously generating Project Design Documents (PDDs) that adhere to the Verified Carbon Standard (VCS). This system employed a combination of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and Large Language Models (LLMs) to ensure both accuracy and compliance with VCS guidelines.

3.2 Data Collection and Preprocessing

We collected over 500 documents from the Verra website, including standards, templates, and certified Project Design Documents (PDDs). These documents are originally in PDF format, the data of the certified PDDs included texts and graphs within PDFs.

The PDFs underwent preprocessing to extract textual information compatible with our system. We utilized the indexing job of the RAG framework to clean and segment the data. Following this, vectorization was conducted on the segmented text contained within the JSON files after cleaning and segmentation had been completed. To improve the accuracy of the retrieval phase, we manually annotated critical sections of the knowledge base. This annotation process concentrated on key areas identified from Verra standards and PDD structures, including methodologies application, emission quantification, monitoring requirements, and risk assessments. We initiated this process by defining the scope of annotations, specifying essential elements such as methodology names, sources of emission factors, as well as details regarding monitoring parameters and their respective measurement methods.

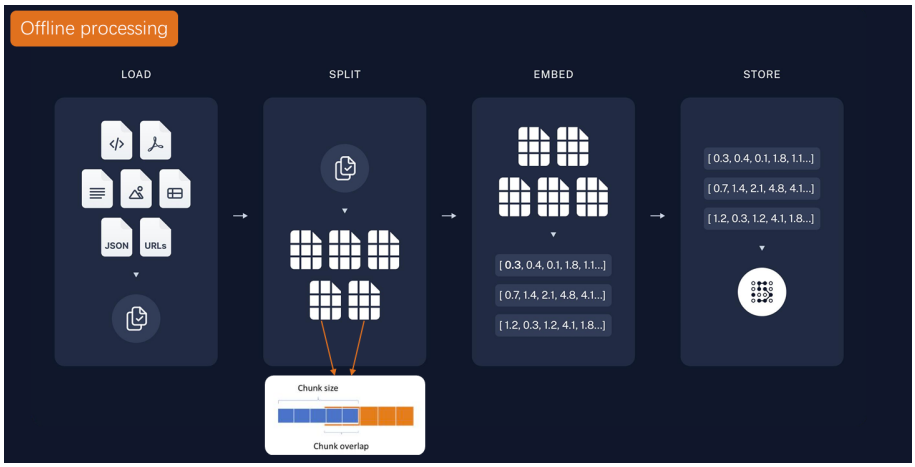


Figure 3: Workflow of Offline Processing in Retrieval-Augmented Generation (RAG)

3.3 Vectorization and Knowledge Base Construction

To facilitate semantic retrieval, we employed a pre-trained BERT model to vectorize the segmented content. In this phase, we transformed the text into a collection of N-dimensional floating-point numbers, commonly referred to as text vectors or embeddings, which serve as the foundation of our knowledge base. The distance between these vectors can be computed; consequently, their proximity corresponds to the degree of semantic similarity among the embeddings. The resulting embeddings were subsequently stored in a FAISS vector database, thereby enabling efficient searches based on user queries. In addition to this automated processing, we manually annotated key sections of the knowledge base to highlight critical content such as monitoring guidelines and emission reduction methodologies, thus enhancing the accuracy of the retrieval process. As shown in Figure 3

3.4 RAG Model Integration

At the core of our system lies the RAG framework, as illustrated in Figure 4. This framework seamlessly integrates semantic retrieval with contextual generation. User queries are meticulously processed through the retrieval module, which identifies the most relevant content from the knowledge base based on semantic similarity. The retrieved information is then integrated into the large language model as context, ensuring that the generated text is both accurate and adheres strictly to Verra standards.

To enhance the efficiency of Project Design Document (PDD) creation, we have developed a structured workflow that delineates the process into five distinct sections: Project Details, Safeguards, Application of Methodology, Quantification of Estimated Greenhouse Gas Emission Reductions and Removals, and Monitoring. Each section is generated independently utilizing the retrieved content, after which the outputs are integrated into a comprehensive and compliant PDD. This modular approach ensures adherence to Verra standards while providing precision and flexibility in the final document.

The decision to prioritize the RAG framework is motivated by its remarkable capability to integrate precise content retrieval with adaptable text generation. By utilizing a search engine alongside an external knowledge base, RAG accomplishes two essential functions: it incorporates external knowledge into the model and facilitates In-Context Learning, thereby significantly enhancing the contextual scope. This dual-process approach, which encompasses indexing as well as Retrieval and Generation, guarantees that the produced text is not only coherent and structured but also firmly anchored in the retrieved content.

For instance, when a user queries the system regarding the optimization of carbon emissions for a landfill gas-to-energy project, the retrieval module identifies the most pertinent content from the indexed knowledge base. This retrieved information is subsequently integrated into the input context of the large language model (LLM), thereby enabling the model to generate customized responses that align seamlessly with Verra standards. This iterative process allows users to refine their inputs based on generated outputs, ensuring that the final document adheres to specific requirements and regulations. By grounding generated responses in

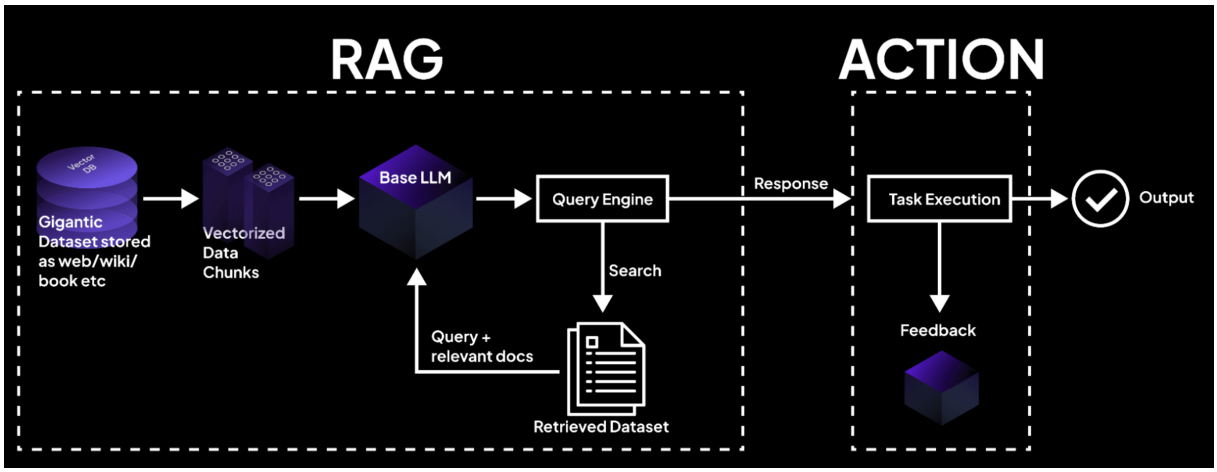


Figure 4: Framework of Retrieval-Augmented Generation

retrieved knowledge, we effectively mitigate the limitations associated with purely generative models, thus delivering more reliable and accurate results that conform to Verra standards.

3.5 User Interaction and Deployment

To facilitate user experience, we developed a web-based interface utilizing Streamlit for the frontend and FastAPI for backend processing. This interface offers users clearly defined input fields and prompts that guide them through the process of generating Project Design Documents (PDDs). Users have the ability to iteratively refine their inputs based on initial outputs, thereby enabling a customized and interactive experience. The system is deployed across both local servers and cloud platforms. This hybrid deployment strategy ensures that sufficient computational resources are available for model operations while maintaining scalability and accessibility for users. Regular updates to the knowledge base and model parameters are seamlessly integrated into the cloud environment, allowing the system to remain aligned with evolving Verra standards.

4 Tradeoffs Analysis

In the preliminary phases of the project, we explored two primary methodologies to tackle the issue at hand: fine-tuning a large language model (LLM) and employing a Retrieval-Augmented Generation (RAG) framework in conjunction with an LLM. Following thorough evaluation and experimentation, we concluded that the RAG+LLM approach emerged as the most appropriate solution for our specific requirements.

4.1 Fine-Tuning a Large Language Model

Fine-tuning a large language model (LLM) involves customizing the model by training it on our specific dataset of Project Design Documents (PDDs). While this approach offers the potential advantage of generating highly domain-specific responses without necessitating retrieval, it also presents significant challenges:

Data Complexity: Extracting and structuring critical information from PDDs to create a high-quality fine-tuning dataset has proven to be a complex and resource-intensive endeavor. Numerous PDDs contained non-standard formats, embedded tables, and technical jargon that posed significant challenges for effective parsing.

Technical Limitations: Training or fine-tuning a large language model locally necessitated significant computational resources that exceeded the capabilities of our existing infrastructure. Additionally, the cost and complexity associated with establishing such a training environment further diminished the feasibility of this approach.

4.2 RAG+LLM Approach

In contrast, the RAG+LLM approach offered several clear advantages that aligned well with our project goals:

Flexibility and Scalability: By separating retrieval and generation, the RAG framework allowed us to focus on extracting and organizing domain-specific knowledge in a lightweight, scalable manner without modifying the underlying language model.

Accurate Contextualization: Using RAG, the system could dynamically retrieve the most relevant content from the knowledge base for each query. This ensured that the generated responses were grounded in domain-specific knowledge, reducing the likelihood of hallucinations often seen in purely generative models.

Efficiency: Unlike fine-tuning, the RAG approach did not require computationally expensive training processes. The system leveraged pre-trained LLMs, integrating them with a knowledge base of PDD content to deliver accurate, context-aware outputs without additional model retraining.

4.3 Model Selection

We evaluated several large language models, including Llama 3.2 (Team, 2024b), GPT-4o-mini (Team, 2024c), and Gemini 1.5 Pro (Team, 2024a), to identify the most suitable option for our system. Gemini 1.5 Pro outperformed the others in terms of accuracy and generation speed, making it the ideal choice. The system's outputs were validated against real-world PDD examples, ensuring alignment with Verra standards and high-quality results. Ultimately, the RAG+LLM framework proved to be the best solution for our problem due to its ability to:

- Leverage existing pre-trained LLMs, eliminating the need for resource-intensive fine-tuning.
- Dynamically integrate structured domain-specific knowledge through retrieval, ensuring compliance with Verra standards and producing accurate, detailed PDDs.
- Scale efficiently, enabling us to handle updates to the knowledge base (e.g., new PDDs or standards) without retraining the model. By adopting the RAG+LLM approach, we not only addressed the challenges posed by fine-tuning but also created a system that is both adaptable and reliable for generating PDDs aligned with Verra standards.

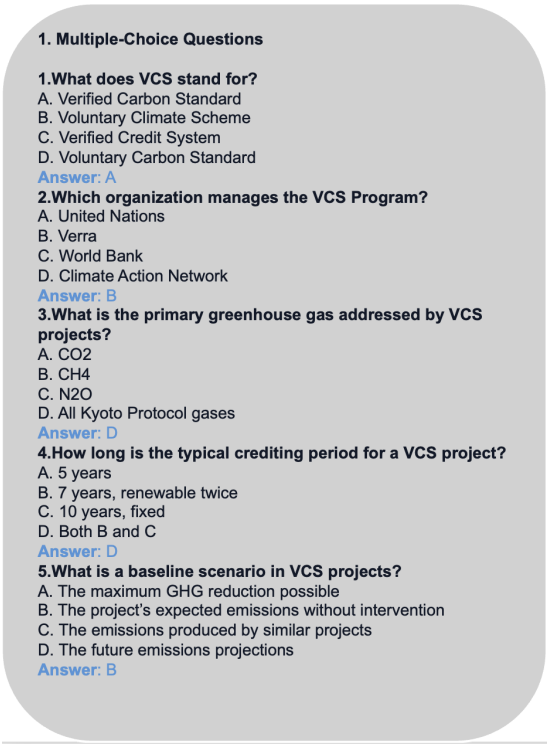
5 Evaluation and Validation

5.1 Model Choice Reasoning

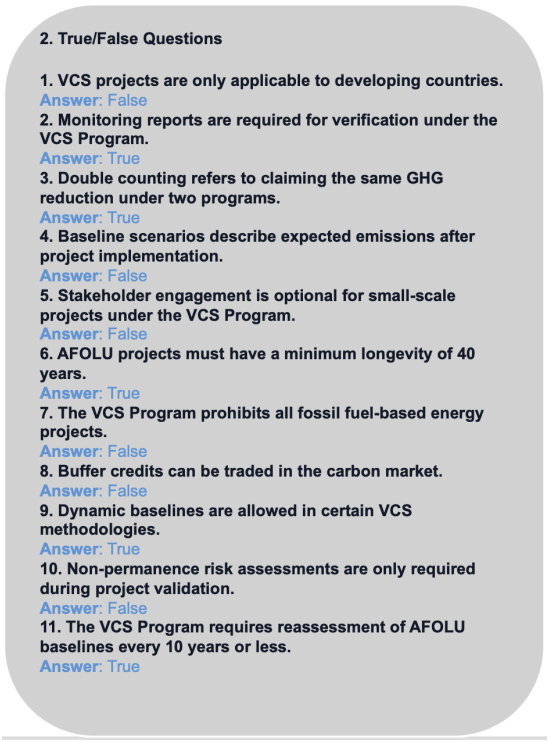
In our Climate AI project, we selected Llama 3.2-vision-11B, GPT-4o Mini, and Gemini-1.5 Pro to ensure a balance of efficiency, adaptability, and precision in generating Project Design Documents (PDDs) compliant with Verified Carbon Standard (VCS) guidelines. Llama 3.2-vision-11B excels in efficiency and scalability, offering faster inference times and cost-effective deployment. Its support for domain-specific fine-tuning allows us to tailor it to VCS requirements, ensuring high-quality outputs while maintaining budget-conscious resource usage. GPT-4o Mini provides advanced language understanding and adaptability, handling the complex nature of PDDs with its ability to process both structured and unstructured data, such as formulas and legal terminology. Gemini-1.5 Pro's exceptional multimodal processing capabilities enable effective management of textual descriptions, graphs, and charts. Its contextual accuracy ensures precise, compliance-specific outputs across both general and specialized queries, making it indispensable for our diverse project needs. Seamlessly integrating with the retrieval-augmented generation (RAG) framework, it minimizes hallucination and stays aligned with evolving compliance standards. The combination of these models enables our chatbot to process and analyze complex datasets, generate accurate and compliant PDDs. Through comparison of these models, we will provide reliable responses, creating a robust and cost-effective solution for climate-related AI applications.

5.2 Evaluation Criteria

The evaluation of Project Design Documents (PDDs) was based on four criteria: Completeness and Relevance (30%) examined whether the PDD included essential sections like the executive summary, baseline methodology, and monitoring plan, and provided detailed, relevant content supported by quantitative data. Full marks were awarded for well-developed sections, with deductions for missing or incomplete areas. Alignment with VCS Guidelines (30%) evaluated adherence to the VCS framework, including template compliance, methodology application, and regulatory alignment. Full marks required complete alignment, with deductions for placeholders or insufficient justifications. Clarity and Professional Presentation (20%) assessed organization, language, and overall professionalism. Full marks were given for clear, error-free documents. Technical and Methodological Accuracy (20%) reviewed the rigor of methodology application, emissions quantification, and monitoring plans. Full marks required detailed and accurate calculations. Final



(a) Multiple-Choice Questions Sample



(b) True/False Questions Sample

Figure 5: Overview of Climate AI benchmark sample

scores were calculated by summing the criteria, providing a fair evaluation of each PDD's strengths and weaknesses.

5.3 Model Evaluations

5.3.1 Completeness and Relevance

In this part of evaluation, GPT-4o Mini and Gemini 1.5 Pro with RAG performed similarly, both scoring 26/30, with nearly identical observations about the document's structured approach and VCS guideline alignment. They both noted the PDD's strengths in project basics, description, and sustainable contributions, while critiquing the lack of quantified GHG emissions reductions and comprehensive monitoring plans. Llama 3.2-vision-11B came in third place with a score of 25/30. Gemini 1.5 Pro without the RAG model scored the lowest at 24/30, offering a less nuanced evaluation that, while recognizing the professional organization and key VCS requirements, provided a more generic analysis and noted additional gaps in risk assessment and stakeholder engagement sections. Across all models, a consistent theme emerged: the PDD presents a solid foundational structure that requires significant technical refinement to achieve full VCS compliance.

5.3.2 Alignment with VCS Guidelines

In the Alignment with VCS Guidelines evaluation, Llama 3.2-vision-11B and Gemini 1.5 Pro without RAG performed similarly, both scoring 22/30, demonstrating near-identical observations about the document's VCS guideline adherence. Both models recognized the PDD's strengths in maintaining the standard template structure and referencing the "High-Efficiency Fossil Fuel Fired Boilers" methodology, while critically highlighting significant gaps in technical documentation. Their analyses consistently pointed out the lack of detailed methodology justification, incomplete baseline assessments, and vague monitoring plans that omitted crucial metrics, roles, and data collection tools. The GPT-4o Mini and Gemini 1.5 Pro with RAG both scored 23/30, offering a slightly more nuanced evaluation that acknowledged the basic VCS template alignment while still exhibiting deficiencies in critical sections like investment analysis, barrier assessments, and stakeholder engagement. Across all models, a consistent theme emerged: the PDD presents a superficial structural compliance that requires substantial technical refinement to achieve meaningful VCS guideline alignment, revealing the current limitations of AI-generated technical documentation in handling complex, evidence-driven

scientific reporting.

5.3.3 Clarity and Professional Presentation

The clarity and presentation of the PDDs varied across models, reflecting their strengths and areas for improvement. Llama 3.2-Vision-11B scored 16/20, demonstrating a logical structure and professional tone but was hindered by placeholders and insufficient depth in areas like additionality and leakage assessment. GPT-4o Mini earned 17/20, with a well-written, professionally presented document featuring concise language and clearly defined sections; however, the lack of specific GHG emissions reductions and detailed explanations for additionality and leakage impacted its overall clarity. Gemini-1.5 Pro without RAG performed better, scoring 18/20 due to its professional tone, well-structured sections, and logical flow, though placeholder sections and missing quantitative details weakened its overall impact. Similarly, Gemini-1.5 Pro with RAG also achieved 18/20 for clarity and presentation, benefiting from professional writing and logical structure, but it was slightly undermined by the absence of quantified GHG reductions and comprehensive justifications for additionality and leakage.

5.3.4 Technical and Methodological Accuracy

The technical and methodological accuracy of the PDDs highlighted varying levels of rigor across the models. Llama 3.2-Vision-11B scored 14/20 due to significant gaps in methodology application, including missing equations, monitoring parameters, and risk mitigation strategies. GPT-4o Mini performed slightly better, earning 15/20, but was still limited by the absence of critical technical details such as equations, parameters, and comprehensive risk mitigation measures. Gemini-1.5 Pro without RAG also scored 14/20, reflecting shortcomings in baseline and project emissions data, incomplete monitoring methodologies, and a lack of risk mitigation strategies. Similarly, Gemini-1.5 Pro with RAG achieved 15/20, with its methodological rigor affected by missing equations, emission factors, and robust risk mitigation measures. These gaps collectively impacted the robustness and technical precision of the PDDs.

5.4 Evaluation Results

The four models—Gemini-1.5 Pro with RAG, Gemini-1.5 Pro without RAG, GPT-4o Mini, and Llama 3.2-Vision-11B—were evaluated for their ability to generate VCS-compliant Project Design Documents (PDDs) based on completeness, alignment with VCS guidelines, clarity, and technical accuracy. Based on these criteria, Gemini-1.5 Pro with RAG emerged as the best-performing model, scoring 82/100 in total, followed by GPT-4o Mini with 81/100 in total, Gemini-1.5 Pro without RAG with 78/100 in total, and Llama 3.2-Vision-11B, which scored 77/100 in total. Gemini-1.5 Pro with RAG excelled due to its comprehensive and professional output, leveraging Retrieval-Augmented Generation (RAG) to enhance depth and accuracy in sections like additionality, methodology application, and monitoring plans. While it still had minor gaps, such as missing specific equations, its overall adherence to VCS guidelines and polished presentation made it the top-performing model. Gemini-1.5 Pro without RAG also delivered a strong and structured PDD, but the absence of RAG's additional depth led to slightly less robust technical accuracy, particularly in areas like baseline emissions and risk mitigation. GPT-4o Mini produced a coherent and well-structured document. However, placeholders for key data, missing quantified GHG reductions, and insufficient detail in additionality and leakage assessments lowered its overall performance. It showed promise but lacked the technical depth to surpass the Gemini-1.5 models. Llama 3.2-Vision-11B, performed the weakest. Its PDDs were incomplete and lacked critical sections like sustainable development contributions and detailed monitoring methodologies. Significant gaps in content, coherence, and technical rigor made it the least suitable model for generating high-quality PDDs. In conclusion, Gemini-1.5 Pro with RAG delivered the most comprehensive and VCS-compliant PDDs, followed closely by GPT-4o Mini and Gemini-1.5 Pro without RAG. Llama 3.2-Vision-11B fell short of meeting the necessary standards, highlighting the advantages of RAG-enhanced approaches for producing reliable and high-quality PDDs.

5.5 Validation Through Benchmarking Tests

Table 2: Performance comparison across different LLMs on Multiple Choice and True/False questions.

Model Type	Model Name	Multiple Choice Accuracy	True/False Accuracy	Overall Accuracy
General	Llama 3.2-Vision-11B	12/15 = 80%	10/15 = 66.7%	22/30 = 73.3%
	GPT-4o-mini	**14/15 = 93.3%**	11/15 = 73.3%	25/30 = 83.3%
Optimized	Gemini-1.5-Pro	13/15 = 86.7%	**15/15 = 100%**	**28/30 = 93.3%**

The validation is done through benchmark test(Figure 5), which contains multiple questions and T/F questions. The table compares the perfor-

mance of three AI models—Llama 3.2-Vision-11B, GPT-4o-mini, and Gemini-1.5-Pro—in answering Multiple Choice and True/False questions, as well as their overall accuracy through our designed benchmark tests. Gemini-1.5-Pro performed best, with 86.7% accuracy in MCQs and a perfect 100% in True/False questions, achieving the highest overall accuracy of 93.3%. GPT-4o-mini excelled in MCQs with 93.3% accuracy but had 73.3% in True/False, resulting in 83.3% overall. Llama 3.2-Vision-11B showed consistent but lower performance, with 80% in MCQs, 66.7% in True/False, and 73.3% overall. In summary, Gemini-1.5-Pro excelled overall, while GPT-4o-mini led in MCQs but fell short in True/False questions. Llama 3.2-Vision-11B trailed behind in all categories. For our chatbot demo results, please see our appendix.

6 Conclusion

ClimateAI demonstrates the potential of combining Retrieval-Augmented Generation frameworks with large language models to address the challenges of creating VCS-compliant Project Design Documents. By automating the generation of PDDs, the system reduces the administrative burden, minimizes errors, and enhances adherence to regulatory standards. The integration of domain-specific knowledge through advanced retrieval and vectorization ensures accurate, structured, and professional outputs.

Future work aims to create a larger and more comprehensive benchmark to evaluate model performance more accurately and rigorously. Additionally, the workflow will be improved to enhance the intelligence of the artificial intelligence agents, enabling more sophisticated and context-aware document generation. Expanding the dataset collection from verified sources such as guidelines and standards will further strengthen the system's knowledge base, ensuring ClimateAI remains a robust and reliable tool for advancing sustainable practices in the carbon market.

Limitations

While our ClimateAI system demonstrates significant advancements in automating the generation of Project Design Documents (PDDs) that adhere to the Verified Carbon Standard (VCS), it is not without its limitations. The current system produces text-based PDDs, which necessitate manual formatting or the use of external tools to convert them into fully compliant PDF documents. This requirement adds an additional layer of effort for users seeking ready-to-submit formats. Furthermore, the system encounters challenges when generating lengthy or highly detailed PDDs due to inherent constraints within the underlying language models and the Retrieval-Augmented Generation (RAG) framework, thereby affecting its scalability for complex projects that demand extensive documentation. Moreover, there is currently no established benchmark or standard dataset available to rigorously evaluate the model's performance in generating PDDs. This absence limits our ability to compare our system with others in this domain effectively. Lastly, assessing the quality and compliance of generated PDDs poses a challenge since validation presently relies on manual inspection—a process that is subjective and resource-intensive. Therefore, automated evaluation mechanisms are essential to ensure consistent quality and reliability of outputs.

Acknowledgments

We would like to express our sincere gratitude to the California Climate Exchange Company (CCEX) for their invaluable collaboration and support throughout this capstone project. Their expertise and resources have been instrumental in shaping the direction and success of our work. We are also deeply thankful to the professors and teaching assistants who provided their guidance and insights at every stage of the project. Additionally, we extend our appreciation to the Master of Data Science students from the University of California, Irvine, whose hard work, dedication, and teamwork were critical to achieving our objectives. This project would not have been possible without the collective contributions of everyone involved.

References

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Llama Team. 2024b. Llama 3.2: Advanced ai model. Accessed: 2024-12-11.

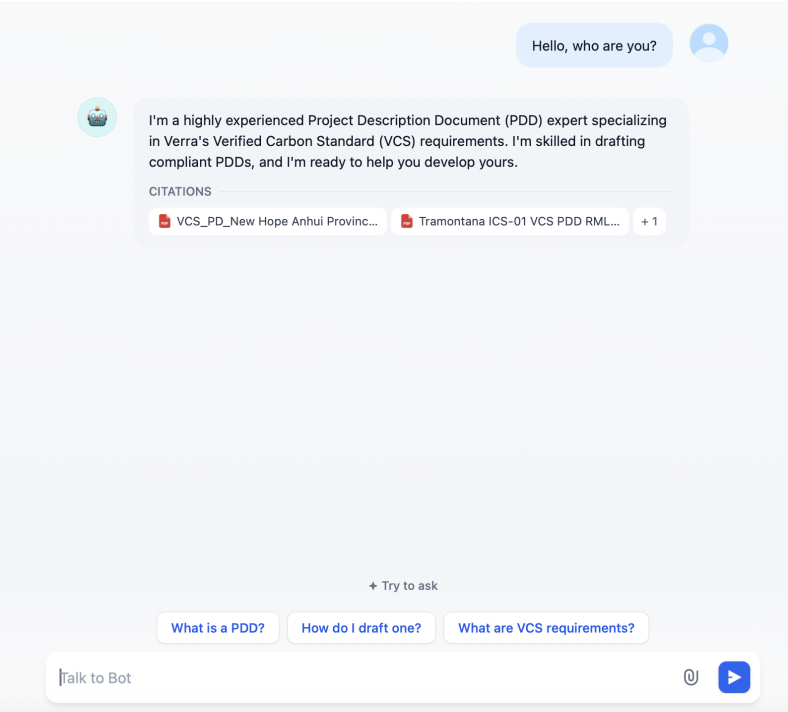
OpenAI Team. 2024c. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-12-11.

A Team Member Contribution Statements

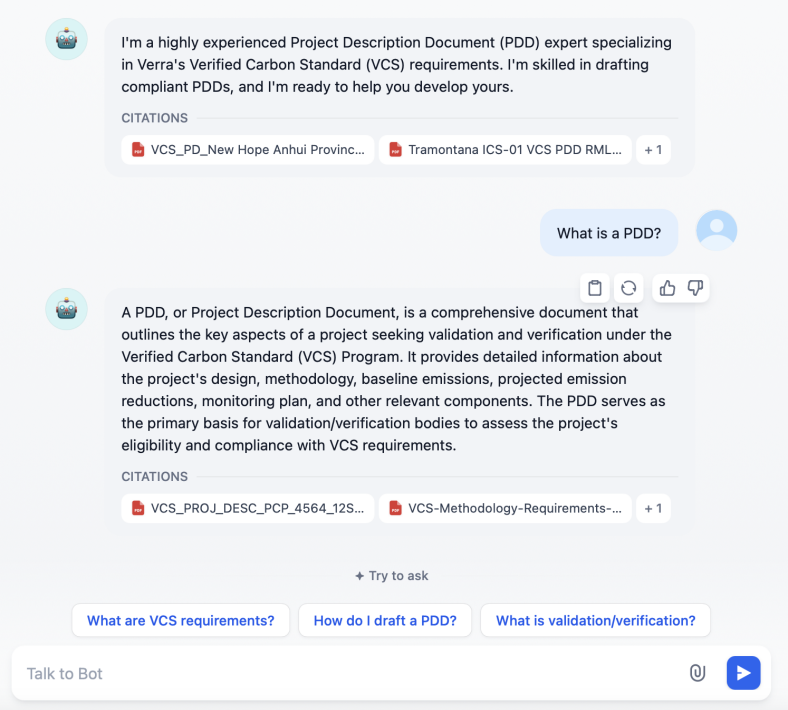
Yu Song, Yijia Sun, and Baijiang Wang: Focused on data-related tasks, including crawling data from verified websites, preprocessing and annotating the collected data, and creating a comprehensive PowerPoint presentation to summarize the project's findings.

Chuqi Wang, Shengtong Sun, and Zhihang Feng: Concentrated on building the technical infrastructure. This included developing the Retrieval-Augmented Generation (RAG) framework, constructing the knowledge base, and deploying three distinct large language models. Their contributions also involved evaluating model performance through the creation of benchmarks, conducting comparative experiments to select the optimal model, and designing a user-friendly chatbot interface for interaction with the system.

B Climate AI Demo

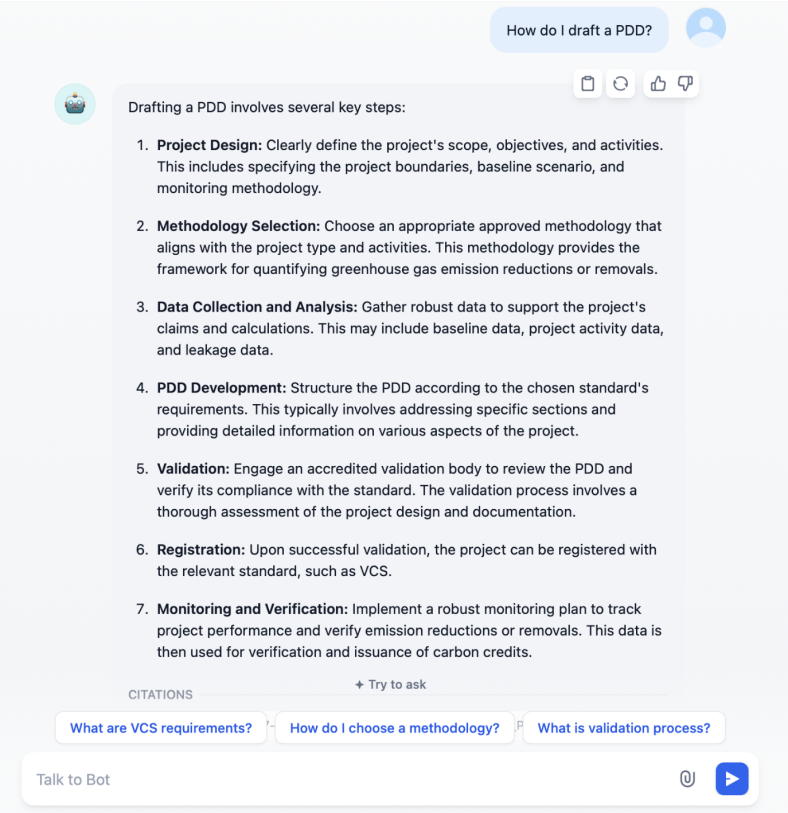


(a) Demo 1

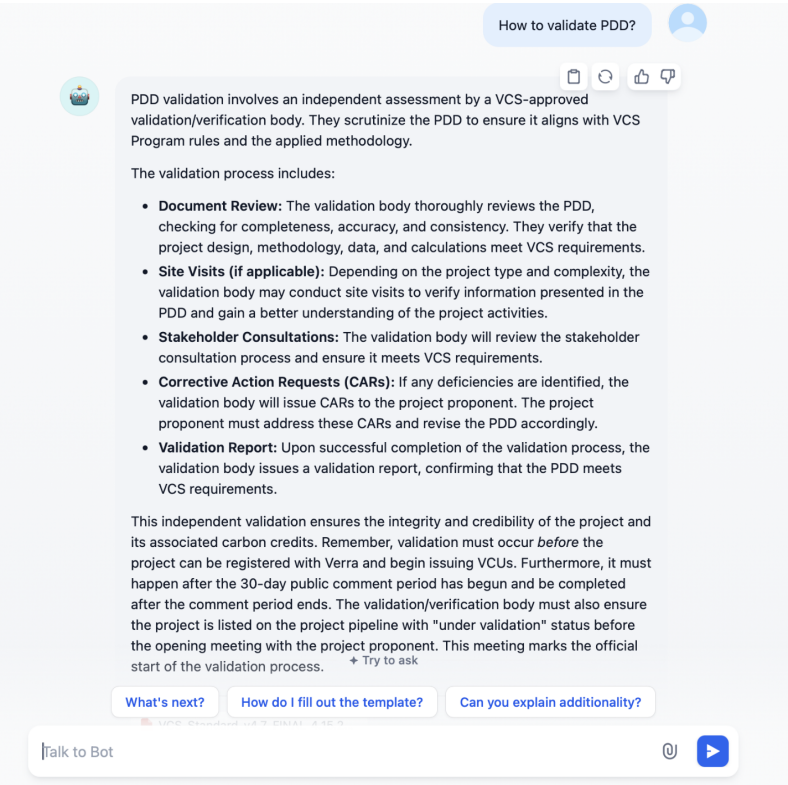


(b) Demo 2

Figure 6: Climate AI Demo



(a) Demo 3



(b) Demo 4

Figure 7: Climate AI Demo