Scaling Species Diversity Analysis in Carbon Credit Projects with Large-Context LLMs

Jessica Walkenhorst[1], Colin McCormick[1,2]
[1]Carbon Direct, New York, USA [2]Georgetown University, Washington DC, USA
Correspondence: jwalkenhorst@carbon-direct.com

Abstract

Reforestation and revegetation projects can help mitigate climate change because plant growth removes CO_2 from the air. However, the use of non-native species and monocultures in these projects may negatively affect biodiversity. Here, we describe a data pipeline to extract information about species that are planted or managed in over 1,000 afforestation/reforestation/revegetation and improved forest management projects, based on detailed project documentation. The pipeline leverages a large-context LLM and results in a macro-averaged recall of 79% and a macro-averaged precision of 89% across all projects and species.

1 Introduction

Reforestation and revegetation projects can help mitigate climate change because plant growth removes CO_2 from the air. The voluntary carbon market (VCM) includes carbon credits from both "afforestation/reforestation/revegetation" (ARR) and "improved forest management" (IFM) projects. The major VCM registries have issued more than 300 million credits (tons of CO_2 equivalent, tCO_2e) to date for ARR and IFM projects ($\sim$14% of the total volume) (Haya et al., 2025).

However, the use of non-native species and monocultures in these projects may negatively affect biodiversity (Cunningham et al., 2015), (Andres et al., 2022), (Moyano et al., 2024). ARR and IFM projects may plant or manage one or more native species, use a mixture of native and non-native species, or use entirely non-native species due to faster growth rates that reduce the cost per tCO_2e mitigated (Busch et al., 2024).

Unfortunately, comprehensive metrics to track planted and managed ("p/m") species in ARR and IFM projects are not readily available. Manual examination of project documents is difficult because there are more than 1,000 ARR and IFM projects in major VCM registries (Haya et al., 2025). A single project's documentation may have tens to hundreds of pages across multiple documents with no common format. Species may be named in the text by botanical (Latin) or common names, and/or be misspelled. A species may be mentioned to indicate it will be planted, it will not be planted, it will be reduced/suppressed, or without clear implications. Given these complexities, advanced natural-language-processing methods are needed.

Here we describe a data pipeline that uses large-context large language models (LLM) to extract information about p/m species in ARR and IFM projects from project documentation. We apply the pipeline to > 1,000 ARR and IFM projects and compare our results to expert human annotation of a subsample. Our pipeline performs well, although validation against expert-annotated "ground truth" data is challenging. Optimizing across two different LLMs, our pipeline results in a macro-averaged recall of 79% and a macro-averaged precision of 89%. We present an analysis of our system's performance using the better-performing model, an error analysis, and a comparison between the two LLMs.

2 Background

There are two main approaches for LLM-based information extraction from long documents. Retrieval-augmented generation (RAG) uses vector similarity between an input prompt and a document database to identify relevant documents, then sends the result to an LLM for response generation. The emergence of large-context (LC) LLMs has led to an alternative approach in which an LC LLM is directly prompted with tasks, with the entire document appended as context. The relative strengths of these two approaches continue to be debated (Xu et al., 2024)(Li et al., 2024). LLMs have been used in biology and ecology for information

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 188–193
July 31, 2025 ©2025 Association for Computational Linguistics

extraction, including the use of GPT-4 to extract information about pests from scientific abstracts (Scheepens et al., 2024); the use of GPT-3.5, GPT-4, and LLaMA-2-70B to extract species distribution data from news articles and research papers (Castro et al., 2024); and the use of *text-bison-001* to extract information about plant pathogens from scientific reports (Gougherty and Clipp, 2024). While curated test datasets are needed for evaluating LLM performance, human annotation is known to produce errors in domains ranging from medicine (Sylolypavan et al., 2023) to online search (Peters et al., 2023); careful annotation guidelines and procedures can partially mitigate this problem.

3 Methodology

Dataset Creation

We identified all ARR and IFM carbon credit projects listed on three major VCM registries (Verra, CAR and ACR) resulting in a total of 339 ARR and 750 IFM projects. We automatically downloaded all existing project documents for these projects and selected all PDFs for further processing. The resulting dataset contains 4196 PDFs with a total of 148,778 pages. Projects in our dataset contain up to 72 PDFs each, with an average of 10 documents per project. PDFs contain up to 870 pages. The maximum number of document pages in a project is 2502. Once downloaded, we converted PDFs to plain text using LangChain's PyPDFLoader and concatenated the text, resulting in one single, large document per project.

Test Set Annotation

We randomly selected 53 ARR and 21 IFM projects for validation. These were distributed among 3 internal subject matter experts (SMEs), who annotated each one with a list of p/m species and an indication of where the information was found in the documentation. SMEs used keyword search and visual scanning to find species information. On average, the annotators spent 15 to 20 minutes per project. Due to resource constraints, only a single SME annotated each document. In a second step, we automatically extracted p/m species information from the documents using each of our LLMs (see below) and manually validated the extracted output. The final list of annotations combines the SME annotations with corrections/additions from the manually validated outputs of the LLMs.

Extracting Species Information

To extract species information from project descriptions, we worked with *gemini-1.5-flash-002* (written *gemini-1.5* in the following) and *gemini-2.0-flash-001* (*gemini-2.0* in the following) through the VertexAI platform. We chose to combine multiple questions into a single prompt to minimize costs. The prompt was as follows:

The context below describes a nature-based carbon credit project. Based on the context given, answer the following questions:

** Which native plant species will be planted or managed (if any)? Only list the native plants that will be planted or actively managed, do not list other native plants.*

** Which non-native or invasive plant species will be planted or managed (if any)? Only list the non-native or invasive plants that will be planted or actively managed, do not list other non-native or invasive plants.*

** Will native plant species be planted and/or managed (true/false)?*

** Will non-native or invasive plant species be planted and/or managed (true/false)?*

For each of the answers, provide an explanation based on the context. Think in steps. If the information is not in the text, simply say "I don't know".

We instructed the model to generate structured output by providing VertexAI with a json structure. If the generated response was not in valid json format, we retried the query once, and skipped the project if that was also invalid. We also skipped projects where the extracted text exceeded the LLM's (large) context window.

Post-Processing

LLM responses were cleaned by replacing "[I don't know]" and "[species]" with empty lists. The prompt asked the LLM to distinguish between native and non-native plant species. Since in this paper, we focus on analysing the complete list of all p/m species, we discarded the answers to the final two questions and aggregated the native and non-native species lists to one final list, which we de-duplicated, first automatically, then manually. Manual de-duplication consisted of unifying species that were mentioned with both their botanical and common names, as well as de-duplicating species that were clearly the same with minor spelling variations.

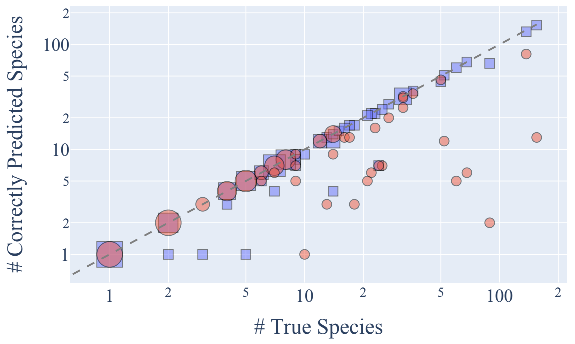

Figure 1: Number of correctly identified vs true number of species for *gemini-1.5-flash* (red circles) and human expert annotators (blue squares) for all test set projects.

Output Validation

We manually validated the output by comparing the list of species produced by the SME annotators and the LLM. The validation was performed manually since many species were mentioned in the documents using both botanical and common names; both the SMEs and the LLM would sometimes choose one and sometimes the other. Manual validation allowed us to accept a predicted species as correct regardless of whether the botanical or common name was used. Species were also deemed correct if they were captured with minor misspellings. These variations were typically the consequence of a species being mentioned multiple times with different spelling in the document. In some cases the LLM output a higher-level taxonomic grouping (e.g. the family or genus) rather than a species, for example *conifers* or *oaks* instead of *red cedar* or *white oak*. This was not counted as correct.

4 Results and Discussion

We successfully extracted data for 1006 out of the 1089 projects, with the remaining failing either because the documents were too long for the LLM's context window or the LLM repeatedly failed to create a valid json response. Given the prompt above, *gemini-1.5* performed better on our test set than *gemini-2.0*. In the following we discuss the results obtained with *gemini-1.5* in detail, followed by a short comparison with the results obtained by *gemini-2.0* and a qualitative discussion of the differences.

Recall

In total, the 74 test set documents contained references to 1241 p/m species. Of these, the human SME annotators found 1147 and the LLM found 628, leading to a micro-averaged recall of 92% for the human experts and 51% for the LLM. The much lower micro-averaged recall of the LLM is mainly caused by the LLM's failure to correctly detect the majority of p/m species for a small number of projects with a large number of species. Fig. 1 shows the number of correctly identified species by both the LLM and the human expert annotators as a function of the true number of p/m species in the project documentation for all test set projects. The LLM performs very well for projects with relatively few species, finding all p/m species for all test set projects with up to six p/m species. However, the LLM's recall drops as the number of p/m species in a project increases. In contrast, human expert annotators tend to miss relatively few p/m species and do so independently of the true number of species in a project.

This pattern can be understood as follows. P/M species are detailed in project documents in multiple ways, but are mostly listed in large tables. SME annotators almost always found these, with occasional entries missing or cases where nearly identical tables exist in the project documents and only the species in one table are annotated. The LLM often did not capture all the species mentioned in these tables, missing more for tables with large numbers of listed species. Species can also be mentioned in the main body of the text. In some cases, this is the only place in the documentation where species names occur (there are no tables), and this may include only a small number of species in total. SME annotators missed these species more frequently than the species which are listed in tables, while the LLM was able to identify them. Finally, in other cases, species are mentioned in graphs and figures, which were often not parsed correctly using our current data pipeline, and therefore were not found by the LLM.

Recall can also be understood at the project level, or macro-averaged recall, for which each project is given the same weighting regardless of its number of p/m species. The macro-averaged recall is 79% for the LLM and 88% for the SME annotators. The median recall is 100% for both LLM and SMEs. The LLM found all p/m species for 62% of the projects, and the SMEs found all p/m species for 68% of the projects. For the remaining projects, the recall is uniformly distributed. Note that the given prompt works better for ARR than IFM projects, reaching a macro-averaged recall of 87% for ARR and 58% for IFM projects.

Precision

Since our test set is a combination of SME-extracted species and LLM-found/human-verified species, human recall for our test set is less than 100%. As for precision, since each document was annotated by a single SME, our test-set creation methodology does not allow us to identify when human annotations are incorrect, resulting in a SME-expert precision of 100%. In contrast, when assessing the precision of the LLM, there are two separate sources of incorrect predictions. The first is hallucination, when the LLM outputs species that do not occur in the project documents. The second is misinterpretation, when the LLM outputs species that occur in the project documents, but in a context that makes it clear they are not planted or managed as part of project activities. For example, in one project the LLM output eucalyptus as a p/m species, despite it only being mentioned in the introductory text as a plant that is often used in reflectivity measurements for monitoring purposes. In another project, the LLM output species that were mentioned in the project documents as having previously been present in the project area, but were later destroyed by fire.

The micro-averaged precision of the LLM across all projects and species was 87%, and the macro-averaged precision was 89%. Of all incorrect predictions, 17% were due to hallucinations, and the remaining 83% were due to misinterpretation, i.e., the LLM output a species present in the text but only in a context different from being planted or managed.

Our LLM-assisted annotation procedure (as described above) impacts the LLM precision analysis. Analysing the LLM's precision taking into account purely human expert annotations (without corrections identified by the LLM-assisted procedure) gives a macro-averaged precision of 78%, 11% lower than the true macro-averaged precision of 89%. The values for the micro-averaged precision are 80% for the manually-corrected data in comparison to 87% without the correction. This highlights the usefulness of LLM-assisted annotation procedures.

gemini-2.0-flash vs. *gemini-1.5-flash*

LLMs are being developed quickly, typically delivering better performance each iteration. Having evaluated our setup in detail using *gemini-1.5*, we also tested it with a newer Gemini model, *gemini-2.0*. Contrary to our initial expectation, we find that *gemini-2.0* performs worse against our test-set using the original prompt, with the overall macro-averaged recall dropping to 60% and the macro-averaged precision dropping slightly to 88%. However, splitting the analysis by project type reveals a more faceted picture. Replacing *gemini-1.5* with *gemini-2.0* leaves the macro-averaged recall for ARR projects roughly unchanged at 87%, but decreases the macro-averaged recall for IFM projects from 58% to 18%. The model frequently outputs that species are mentioned but no species are explicitly stated to be planted or managed, which is true for many documents. Thus *gemini-2.0* behaves as a more literal reviewer than our SMEs, who will infer that a mentioned species is planted or managed from the overall context of being mentioned in the project documents. How to prompt *gemini-2.0* to be more permissive for these species whilst also not extracting species mentioned in other contexts will be the focus of further research. Separate prompts for ARR and IFM projects will be a key step.

5 Summary and Conclusions

In this work, we developed a dataset of over 1,000 ARR and IFM projects listed on three major VCM registries (Verra, CAR, ACR). We used a combination of manual annotations and LLM-derived corrections/additions to create a test set of planted/managed species in 74 projects. Next we developed a data pipeline to extract species information from project documentation documents by prompting a large-context LLM with questions regarding the species that would be planted and/or managed as part of project activities, with the full text of the project documentation appended as context. The LLM achieved a macro-averaged recall of 79% and a macro-averaged precision of 89% whilst human annotators achieved a macro-averaged recall of 88%. Notably, human annotators tended to miss a small number of planted/managed species per project, while the LLM missed more species if more species were mentioned in the text. Our results demonstrate the possibility of using large-context LLMs to extract species diversity information from lengthy project description documents.

6 Future Work

In future work we will address several areas. First, we will explore prompt-engineering to get the LLM to consistently choose a species' botanical name over its common name and explore the use of a second, smaller LLM for automatic de-duplication. Second, we will further explore prompt-engineering techniques to better distinguish when species are actively planted or managed vs. simply mentioned in passing. Third, we will analyze which errors are caused by PDF parsing errors. Fourth, we will split not only our analysis but also our prompt engineering by project type (IFM vs ARR projects). Fifth, we will extend our approach to look into the contrast between native and non-native species use. Sixth, we will examine the performance of the LLM on non-English documents. Finally, we will explore the use of RAG instead of large-context LLMs in order to improve scalability.

Limitations

Labeling Accuracy

Annotating complex datasets is challenging and the annotations created in this work might not yet be completely correct. In particular, in IFM projects, it is sometimes not possible (not even for human annotators) to correctly label species as being under active management or merely present in a project area. Additionally, because of our use of a single SME annotator per project, we were unable to inter-compare manual annotations, a process used to increase the reliability of a labelling process. Species that are part of a project might still be missing, making the reported recall appear higher than it truly is. In particular, albeit using two different LLMs, we still used the same LLMs during annotation and testing process, making this scenario more likely.

PDF Parsing

In the current work, errors in recall are analysed on a pipeline level, without distinguishing whether the species information was present in the parsed text or not (we only know that it was present in the PDF). Distinguishing errors in recall into errors caused by parsing issues vs errors caused by the LLM would give further insights into the maximum possible performance of the LLM pipeline.

IFM vs ARR Projects

This work treats ARR and IFM projects similarly. The prompt is generalized, intended to work reasonably well for both types of projects. However, whilst these project types are similar, they are not the same. In particular, in IFM projects, it is sometimes not possible (not even for human annotators) to correctly label species as being under active management or merely present in a project area. A distinction between ARR and IFM projects in future prompts will be helpful. As we demonstrated in the present paper, this will become increasingly important with the development of more powerful LLMs which are capable of understanding ever more subtle nuances of human language.

Scalability

The presented approach used large-context LLMs to extract species information from project document descriptions. This approach works well for most projects, but already reaches its limits for some. Additionally, registries do not delete documents, making texts longer over time. Alternative architectures like RAG could help alleviate this issue.

Single-Purpose vs Multi-Purpose Prompts

Due to financial constraints, we tried to limit the number of times we queried the LLM. In particular, we combined multiple questions into a single prompt, where several, individual queries might have achieved better performance. This is a limitation of our set-up not of the LLM's capability.

Manual De-Duplication

LLM outputs were manually de-duplicated, unifying botanical with common names as well as correcting spelling errors. In particular, validation was done manually. This approach does not scale and makes the current process not suitable for techniques like automatic prompt-optimization. An automatic validation setup including prompting the LLM to always list species with their botanical names will be implemented in the future. This could be supplemented by using dictionaries mapping between common and botanical names.

Acknowledgments

The authors would like to thank Sarah Federman and Tristan Goodbody for their support regarding biodiversity concerns in ARR and IFM projects; Jared Stapp for acquiring the raw data; Mila Stieglitz-Courtney, Simon Sharp and Julia Yang for annotating the test data as well as manually validating the outputs; and Liam Kilroy and Daniel Garcia for additional manual validation. The authors would also like to acknowledge the use of ARR and IFM project data from ACR, Verra, and the Climate Action Reserve (CAR).

References

Samantha E. Andres, Rachel J. Standish, and et al. Paige E. Lieurance. 2022. Defining biodiverse reforestation: Why it matters for climate change mitigation and biodiversity. *Plants People Planet*, 5:27–38.

J. Busch, J.J. Bukoski, and S.C. et al. Cook-Patton. 2024. Cost-effectiveness of natural forest regeneration and plantations for climate mitigation. *Nat. Clim. Chang.*, 14:996–1002.

Andry Castro, João Pinto, and et al Luís Reino. 2024. Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics*, 82:102742.

S.C. Cunningham, R. Mac Nally, and et al. P.J. Baker. 2015. Balancing the environmental benefits of reforestation in agricultural regions. *Per. in Plant Ecol., Evo. and Syst.*, 17:301–317.

Andrew V. Gougherty and Hannah L. Clipp. 2024. Testing the reliability of an ai-based large language model to extract ecological information from the scientific literature. *npj biodiversity*, 3:13.

Barbara K. Haya, Aline Abayo, Xinyun Rong, Tyler G. Bernard, Ivy S. So, and Micah Elias. 2025. Voluntary registry offsets database v2024-12-year-end, berkeley carbon trading project, university of california, berkeley.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. In *TProceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*.

Jaime Moyano, Romina D. Dimarco, and et al. Juan Paritsis. 2024. Unintended consequences of planting native and non-native trees in treeless ecosystems to mitigate climate change. *J. Ecologya*, 112:2480–2491.

Heinrich Peters, Alireza Hashemi, and James Rae. 2023. Generalizable error modeling for human data annotation: Evidence from an industry-scale search data annotation program. *Journal of Data and Information Quality*, 16:1–15.

Daan Scheepens, Joseph Millard, Maxwell Farrell, and Tim Newbold. 2024. Large language models help facilitate the automated synthesis of information on potential pest controllers. *Methods in Ecology and Evolution*, 15:1261–1273.

Aneeta Sylolypavan, Derek Sleepman, Honghan Wu, and Malcolm Sim. 2023. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, 6.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.