# Using Large Language Models for the Assessment of Sustainable Forest Investment Projects

Maria Letizia Jannibelli
University of Zurich
marialetizia.jannibelli@uzh.ch

Jiayu Luo
University of Zurich
jiayu.luo@uzh.ch

Kilian Sprenkamp
University of Zurich
kilian.sprenkamp@uzh.ch

Liudmila Zavolokina
University of Lausanne & University of Zurich
liudmila.zavolokina@unil.ch

## Abstract

*The integration of Large Language Models (LLMs) into the assessment processes of sustainable forest investment projects is a compelling prospect, given the limitations present in manual assessment. This paper examines how such an LLM-based assessment tool can be designed and whether such a tool can serve as a viable alternative to human experts in this task through the development and subsequent evaluation of a prototype. The analysis shows that the use of retrieval augmented generation to extract and summarize relevant information from project documents is promising but reveals challenges in the use of LLMs for more complex analysis and grading tasks. Design principles and possible steps for further development of the tool are proposed.*

**Keywords:** Sustainable Forest Investments, Large Language Models, Retrieval Augmented Generation

## 1. Introduction

The urgency of addressing contemporary global challenges, such as climate change and biodiversity loss, has increased the importance of measures to protect, sustainably manage and restore ecosystems, such as sustainable forest projects. These projects are crucial in sequestering carbon and preserving biodiversity. For instance, biodiversity loss can lead to the collapse of ecosystems that support agriculture and clean water, directly impacting human well-being (Cohen-Shacham et al., 2016).

While most funding for such projects comes from the public sector, the private sector also invests in sustainable forest projects, incentivized through carbon credits, public image benefits, and as a way to protect their supply chains (Deutz et al., 2020; Forum, 2021). These investments are often driven by environmental, social, and governance (ESG) considerations, emphasizing the need for reliable assessments to ensure that projects meet sustainability goals and provide long-term benefits. Hence, investors rely on professional assessments to ensure the long-term efficacy and sustainability of the projects they invest in (Raymond et al., 2017; Seddon et al., 2020). Domain experts rely on various sources, including project documents, certification reports, audits, and land titles, to assess projects across multiple areas. However, the large volume of lengthy documents makes the process time-consuming and limits the depth of analysis a single expert can perform.

In collaboration with Xilva, a company specializing in the assessment of sustainable forestry projects, this research explores how large language models (LLMs) can enhance the assessment process. LLMs, like BERT (Devlin et al., 2019) or GPT-4 (OpenAI et al., 2024), are transformer-based models capable of analyzing and interpreting text with high accuracy, making them ideal for tasks like extracting data from ESG reports and analyzing sustainability-related documents (Luccioni et al., 2020; Ni et al., 2023; Vaghefi et al., 2023; Zou et al., 2023). Despite their success in these areas, the application of LLMs to create a unified tool specifically for assessing sustainable forest investments has not yet been realized. Hence we ask the following research questions:

**RQ1** *How can we design an LLM-based tool to support the assessment of sustainable forest investment projects?*

**RQ2** *Can LLM-based assessment serve as a viable alternative to human experts in assessing sustainable forest investment projects?*

URI: https://hdl.handle.net/10125/109408
978-0-9981331-8-8

As a part of a larger design science research (DSR) project (Hevner et al., 2004; Peffers et al., 2007) in collaboration with Xilva, we develop GRAIDY, a prototype of such an LLM-based assessment tool. These research questions are, therefore, primarily addressed through the design of the prototype and the subsequent comparative analysis of an assessment performed by GRAIDY with the analysis performed by domain experts. Finally, we derive a set of six design principles (Gregor et al., 2020) for such a tool. These design principles can be used by researchers and practitioners aiming to employ LLMs for information retrieval in complex domain-specific assessments, like sustainable forest investments.

## 2. Related Work

While sustainable investments contribute to mitigating the effects of climate change, investors struggle to find investable projects due to a lack of standardized reporting, inconsistent data quality, and the complexity of analyzing diverse ESG metrics across different projects. Although the assessment of such projects is often outsourced to experts, they still face significant challenges due to the unstructured nature of documents like sustainability reports, environmental impact assessments, and legal compliance documents. These documents are characterized by complex, domain-specific language and diverse formats, such as tables, images, and PDFs, which complicate the analysis (Bronzini et al., 2024). This lack of standardized information makes it difficult for experts to efficiently and accurately assess sustainability investment projects in a timely manner.

LLMs have shown a potential to aid in information retrieval for domain-specific unstructured text due to several technological advancements. LLMs are pretrained using self-supervised learning on vast amounts of unstructured text data (Devlin et al., 2019; OpenAI et al., 2024). This extensive pretraining allows them to acquire a broad understanding of language, including syntax, semantics, and general knowledge across various domains. Moreover, these models can be fine-tuned to various domains, e.g. sustainable investments, broadening the knowledge base of the LLM. Another way to enhance the knowledge base of LLMs is retrieval-augmented generation (RAG) (Lewis et al., 2020), which combines the model's internal knowledge with up-to-date external information sources to improve accuracy and relevance. RAG utilizes embeddings, which are vector representations of text that capture semantic meaning. These embeddings are stored in a vector database, which allows for efficient similarity searches. When a query is made, the system retrieves relevant information from the vector database based on the similarity of embeddings. The retrieved information is then combined with the LLM's response, grounding the output in specific, contextually relevant external data. This approach not only expands the scope of information accessible to the model but also ensures that the generated text is accurate and aligned with the most current information available. Furthermore, LLMs benefit from prompt-based learning, a technique that guides the model's responses using carefully crafted prompts, enhancing their ability to produce accurate and contextually appropriate outputs tailored to specific needs or queries (OpenAI et al., 2024). These advancements collectively make LLMs powerful tools for processing and analyzing unstructured text in domain-specific applications.

The potential of LLMs in analyzing ESG-related domain-specific text has been discussed by various scholars utilizing LLMs like BERT, RoBERTa, and GPT-4 (e.g. Angin et al., 2022; Luccioni et al., 2020; Ni et al., 2023). Domain-specific pretraining demonstrated significant improvement in the performance of models like BERT on various climate-related tasks (H. Lee et al., 2024; Luccioni et al., 2020; Pasch and Ehnes, 2022; Webersinke et al., 2022). However, it could also be shown that grounding responses in external sources through the use of RAG (Lewis et al., 2020) can be an efficient and cost-effective alternative to pretraining. A comparison of GPT-4's performance with varying dependencies on external reports indicated a synergistic effect when models utilize a combination of innate knowledge and external data to answer climate-related questions (Vaghefi et al., 2023).

However, research on further use of identified relevant text beyond simple statistical or sentiment analysis is scarce. Notable exceptions include ChatReport (Ni et al., 2023), a tool that automates the assessment of sustainability reports on their adherence to certain guidelines, with a chat function that allows users to further analyze a report through question answering. Similarly, Zou et al. (2023) present ESGReveal, a system that provides summarization of reports with a particular focus on the correct extraction of numerical data. Luccioni et al. (2020) developed ClimateQA, a model based on RoBERTa, which is further pretrained using financial and sustainability reports, that enables question answering based on provided documents.

One challenge in the analysis of sustainability reports and other relevant documents is that they are often in unstructured formats like PDFs and frequently contain data in tables or images. Converting these

documents into a format usable by LLMs, such as structured text or embeddings, is a significant challenge (Bronzini et al., 2024; Polignano et al., 2022; Zou et al., 2023). Inaccurate extraction of information from tables has been identified as a major source of error and a critical area for further research to enhance the performance of LLMs in this domain (Luccioni et al., 2020). Another limitation is the tendency of LLMs to hallucinate, sometimes providing incorrect information with high confidence, making it hard for users to identify these errors (Braunschweiler et al., 2023). Grounding responses in external sources through RAG can significantly reduce, but not completely eliminate, hallucinations (Luccioni et al., 2020; Vaghefi et al., 2023; Zou et al., 2023).

Regarding the use of LLMs for automated assessment of text, insights can be gained from literature in the domain of automated scoring of essays and other student-written texts, as well as from research on the use of LLMs to evaluate text generated by LLMs. Chiang and Lee (2023) provides a framework for automated scoring by detailing a sequence where LLMs receive comprehensive instructions on the rating task, including a definition and rating criteria of the attribute to be rated. Their research suggests that ratings given by ChatGPT correlate more closely with human judgment when it is not only prompted to provide a numeric score but also to explain its reasoning.

The existing research shows the capabilities of LLMs both for information retrieval from sustainability-related text and the assessment of text, but they have not been jointly applied in a unified application to perform complex assessments based on large amounts of unstructured text regarding forest investments.

## 3. Methodology

This paper is a part of a larger DSR project. DSR is a well-established methodology that guides the design of IT artifacts (e.g., systems, models, methods) that aim to solve real-world problems (Hevner et al., 2004). The DSR process after Peffers et al. (2007) encompasses six steps: (1) Problem Identification and Motivation, (2) Definition of Objectives of a Solution, (3) Design and Development, (4) Demonstration, (5) Evaluation, and finally, (6) Communication. The *(1) Problem Identification and Motivation* are described based on the problems that Xilva, a company specializing in sustainable forestry investment assessment, experiences in their work. *(2) Definition of Objectives of a Solution* were formulated and continuously adapted through regular meetings between the research team and the company's domain experts. This iterative approach during *(3) Design and Development* resulted in the creation of an IT artifact in the form of an LLM-based assessment tool called GRAIDY to analyze forest investment projects.

Finally, to assess whether GRAIDY is useful to solve the problem and to answer our RQs, we conducted and describe the results of an early *(4) Demonstration* and *(5) Evaluation* of the GRAIDY prototype. A forest investment project suitable for demonstration was selected by the domain experts. The chosen project was from a hardwood company that plans to take the majority of the forests in South America under their management out of the planned logging cycle and into management for ecosystem services. A total of 23 documents about this project were provided, including project description documents by the hardwood company themselves, certification documents, and audit reports. These documents were given as input to GRAIDY which then performed an assessment of the project.

During the *(5) Evaluation*, a comparative analysis involving domain experts was conducted to evaluate the information retrieval, grading accuracy, and overall usefulness of GRAIDY to the expert assessors. For this, an expert with extensive experience in forest ecology provided a manual assessment of the same project using Xilva's standard methodologies. This included referencing external validation tools in addition to the project documents. It is important to note that GRAIDY does not yet integrate these external validation sources. This deviation was recognized as a limitation but deemed necessary to preserve the authenticity of the expert's assessment process. The comparative analysis was done in two strands:

1. **Qualitative Analysis**: The evaluation of GRAIDY involved multiple stakeholders. Initially, four expert interviews (*E1-E4*) were conducted with experts from the partner company. During these interviews, the outputs generated by GRAIDY were compared with those from a manual assessment. The interview focused on each subarea, assessing the quality of textual responses and grades provided by GRAIDY and how well they aligned with the manual assessment. Furthermore, broader feedback on GRAIDY's overall performance was collected. Additionally, GRAIDY was showcased in a workshop (*W1*) attended by the company's customers and investors.

2. **Quantitative Analysis**: Grades from 0 (insufficient information for grading) to 5

(fully fulfills the criteria of this subarea) were assigned to each of the twelve subareas. Spearman's correlation coefficient was calculated to determine the relationship between the grades given by GRAIDY and the expert. Additionally, the textual responses were analyzed for incorrect citing or hallucinations by cross-referencing the information against the stated sources.

This evaluation, which combines quantitative and qualitative analysis, allowed for a comprehensive examination of the prototype's abilities and whether it achieved the design goals set. The final step of the DSR process, the *(6) Communication* of the research happens with the paper at hand.

# 4. Results

## 4.1. Problem Identification and Solution Objectives

Through internal discussions with the partner company Xilva, in the form of workshops and regular sprint planning meetings, the following main problems in their current assessment process could be identified, and corresponding design goals (DG) were formulated.

**Problem 1** *Relevant information being missed.* Due to the large amount of documents and time limitations present, some documents or parts of documents are not read in detail but skimmed or not considered at all. This leads to information being missed that would have been relevant for the accurate assessment of a subarea. Based on this, we defined *DG1 Gather relevant information from provided documents.* The objective of this DG was to facilitate reading through all documents to find relevant information for each subarea, both saving time and reducing instances of information being missed by being able to consider all documents with equal depth.

**Problem 2** *Difficulty in adequately reporting findings of assessment.* The need to rigorously document the findings of the assessment was identified as a time-consuming and challenging aspect of the assessment process. Therefore, we defined *DG2 Provide summaries of relevant information and findings.* The objective of this DG is to facilitate the documentation of the assessment by providing textual responses for each step of the assessment process.

**Problem 3** *Large amount of time needed to perform an assessment.* The aforementioned large amount of time needed to read through all relevant documents, document the findings, and assign grades presents a problem to the partner company as it leads to them not being able to meet the demand for their assessment services. Based on this problem, we defined *DG3 Perform a complete assessment in a timely manner.* The objective of this DG was to provide the expert assessors with a preliminary assessment, which combines the steps of information retrieval, grading, and documentation of findings, saving time on these steps and leaving the experts free to focus on areas that warrant further examination based on the findings of this preliminary assessment.

## 4.2. Design and Development

Based on the design goals, GRAIDY was developed to quickly find relevant information in a set of documents relating to the project and to provide a preliminary assessment based on this information. The first working prototype of GRAIDY being studied in this paper is limited to only 1 of the 6 core domains in Xilva's assessment methodology with the idea being to later expand it to all 6 areas. The domain selected for this initial demonstration is the ESG safeguards, which is comprised of 12 subareas and assesses the project on the proper analysis of ESG risks and if it has put in place the appropriate safeguards and mitigations against them.

A diagram illustrating the components and process of GRAIDY can be found in Figure 1. The first step of the process is **1. Document Upload**, for which we designed a simple user interface that enables the uploading of PDF documents. To fulfill DG1 of gathering relevant information from the documents, RAG was chosen as the first design feature (DF1) and served as the core principle of the tool. To prepare the uploaded PDF documents for retrieval, the text is parsed to extract plain text and tables. Given their complex structure, tables are reformatted to enhance readability and consistency, involving cell merging, text orientation correction, and standardization. The plain text is cleaned by removing headers, footers, and extra spaces. Finally, the text is segmented into chunks of 500 tokens, with a 100-token overlap, to ensure complete coverage and facilitate efficient search and retrieval. An embedding model, specifically `text-embedding-ada-002`, which is based on the GPT model family, is used to encode each chunk into a vector representation of length 1536, stored in a vector database. Then, for the next step of the process (see **2. RAG** in Figure 1), the user can enter a prompt, with a default one already provided, for each of the 12 subareas to specify what information should be retrieved. This design feature (DF2), allowing the user to specify exactly what information should be retrieved, again aims to fulfill DG1. The user can then initiate the assessment process with a button. This starts the RAG process by first vectorizing the given prompts with

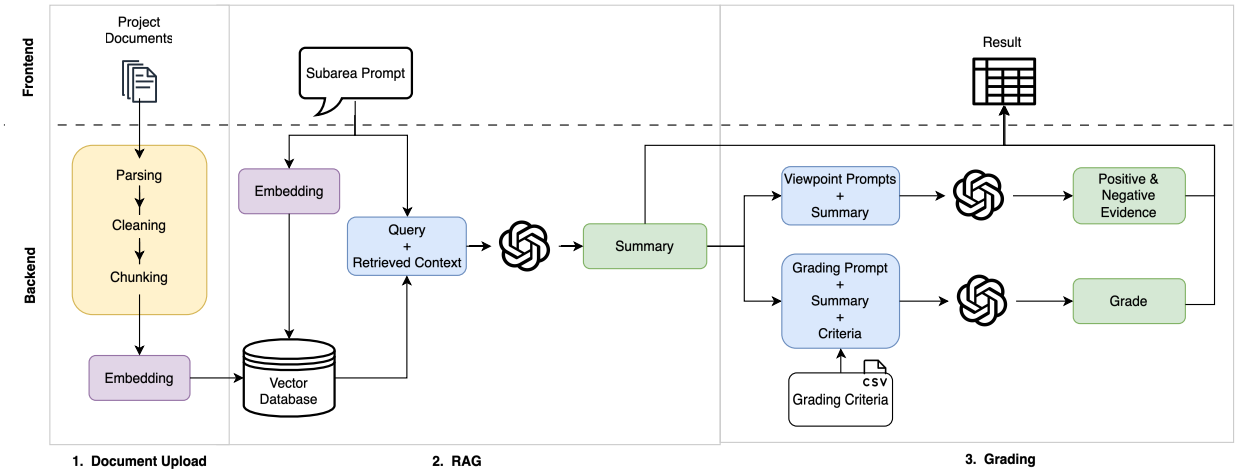


**Figure 1. Prototype process diagram**

the same embedding model used on the documents, enabling the calculation of a similarity score between the query vector and the vectors already in the database. The chunks with the greatest similarity are returned. In the subsequent generation step, the LLM is prompted with a combination of the original query and the retrieved information. The result of this is a summary of the retrieved information (DF3). Along with this summary, the sources of the information are returned in the form of a document name and page number. This sourcing feature (DF4) was implemented to facilitate the verification of information and, therefore, the identification of hallucinations. It also aims to facilitate any further in-depth analysis by the domain experts by providing them with information on what documents and pages are relevant for each subarea. The summary given for each subarea then serves as the basis for the grading process (see **3. Grading** in Figure 1). This process encompasses two design features: Firstly, DF5 of providing positive and negative evidence concerning the subarea, aligning with the company's structuring of analysis results and fulfilling the DG2 of documenting the assessment results. Secondly, DF6 provides a numerical grade. This is done by again querying the LLM with a prompt consisting of the summary, a grading prompt, and grading criteria specific to each subarea. The grade, summary, sources, positive evidence, and negative evidence of each subarea are then combined and returned to the frontend, where they are presented to the user as a table and fulfill DG3 of providing a complete assessment.

## 4.3. Demonstration and Evaluation

An assessment of the selected forestry project was carried out using GRAIDY, with GPT-4 as the base LLM model. The 23 documents available were uploaded to the GRAIDY system, resulting in 5046 text chunks for the vector database. The default information retrieval prompts were used. The assessment of one subarea ("Natural Hazards") by both GRAIDY and the expert is shown in Table 1 and serves as an example.

**Qualitative Analysis** The following summarizes the main findings from the expert interviews and workshop and illustrates them through a few examples.

*Information Retrieval:* Regarding GRAIDY's ability to retrieve all information that is relevant to a subarea, the experts agreed that the performance varied quite drastically from subarea to subarea. As noted by *E3*, this is likely due to information returned by the RAG system, stating: *"It might be related to the context given to the LLM. The LLM's response is very general, while the manual assessment is much more specific."*. However, in the subarea "Natural Hazards" (see Table 1) *E1* and *E4* noted that a piece of information was found by the prototype that was missed by the human assessment; that the audit team and local government agreed with the assessment of the risks as insignificant. Moreover, a workshop participant *W1* underlined the importance of structuring the information retrieval prompt based on the different subareas analyzed, stating: *"Through iterating and refining the prompt we obtained more relevant results for the given subarea."*

However, in many cases, there was information in the provided documents that would have been relevant for the subarea but was not found by GRAIDY. One stark example of this was in the subarea "Labour & Working Conditions," where the human assessment had found repeated reports of poor labor conditions in the provided audit reports. GRAIDY, however, missed this information, only referencing the documents written by the project owners themselves on this topic. The audit reports where this crucial information could be found were written in Spanish and structured in a

**Table 1. Assessment of subarea "Natural Hazards"**

| | GRAIDY | Expert |
|---|---|---|
| **Grade** | 5 | 3 |
| **Summary** | The only natural hazard identified in the project area is extreme weather, specifically tornados. Other risks, such as fire, pests and disease, and geological or other natural risks, do not occur in the project area. The project proponent has demonstrated that these risks are insignificant, a finding that was confirmed by the audit team and the regional government. | |
| **Positive Evidence** | The project area has been thoroughly assessed for natural hazards, and only one significant risk, extreme weather in the form of tornados, has been identified. This indicates a relatively low-risk profile for the project in terms of natural hazards. The project proponent's findings have been confirmed by both an audit team and the regional government, adding credibility to the assessment. | The project demonstrates awareness of the possibility of increased climatic anomalies in temperature and rainfall patterns and identifies riverine communities as particularly sensitive to floods. The Iquitos district has a very low incidence of forest fires both in terms of frequency and number of active days. |
| **Negative Evidence** | While the assessment seems comprehensive, it does not provide any mitigation strategies or contingency plans for the identified risk of tornados. Additionally, the assessment does not consider the potential impact of climate change, which could introduce new risks or exacerbate existing ones. | Think Hazard: river and urban flood (high risk), extreme heat, and wildfire (medium). The most sensitive areas to river flood are the riverine communities that are in the project area but out of the concession. However, there is no evidence that the project has conducted a risk analysis, e.g., flood zone mapping, or has a mitigation plan to minimize damage in the face of these environmental hazards. |

table format, which are both possible reasons for this information not being retrieved. One case where no reason was apparent why the information was missed was in the subarea "Political instability & corruption." Here, no information was retrieved by GRAIDY, and subsequently, no grading was performed. However, the following sentences can be found in one of the documents:

*"[The project company] adheres to the strictest local laws, reinforced by internal policies to prevent corruption and unethical behavior, evidenced in the Code of Conduct and Anti-Bribery Policy documents."*

This passage was not in table format or any other formatting known to cause issues, and it explicitly included the word "corruption," which was included in the prompt for retrieving information on this subarea.

*Quality of textual responses:* The summaries provided by GRAIDY were generally of satisfactory quality to the experts. One exception to this was in the "Anthropogenic Deforestation" subarea, where it found information missed in the manual assessment but summarized some technical details about monitoring strategies in a way the experts (*E1, E2, E3, E4*) agreed on being unclear:

*"Anthropogenic disturbances, such as intense felling, can lead to significant reductions in carbon stocks. These disturbances can be detected in areas where the cumulative $\Delta C$ in any grid of 3x3 pixels is greater than -10 tC."*

While this information is relevant and correct, it is unclear, for example, what "$\Delta$C" refers to in this context and what data or map is being used. *E2* noted that they would have preferred a direct quote from the document itself instead of a summary so that no information is lost.

*Importance of External Verification: E1* repeatedly stressed the importance of cross-checking the project documents with external sources for accurate assessment of several subareas, such as "Natural Hazards." They regard it as impossible to correctly assess a project for adequate handling of natural hazards without first assessing the hazards present in the area from impartial external sources. The lack of such cross-checking often resulted in the positive evidence found by GRAIDY to be similar to that of the expert, both using the project documents as the source. However, the negative evidence differed greatly, with GRAIDY often providing only little negative evidence based solely on the project documents, while the expert had found more significant negative evidence through external sources. This was further confirmed by *E4*, stating: *"In the prototype analysis, the language is too promotional, using positive terms unsuitable for the neutral tone of due diligence reports."*

*Evaluation of Document Credibility:* Related to the above, *E1* also noted the importance of treating documents differently depending on their type and source, such as in the previous example in the subarea "Labour & Working Conditions." A description of good working conditions in a document by the project owners themselves should be questioned, whereas a mention of poor working conditions in an audit report should be taken at face value and trumps the former. GRAIDY's

responses did show some challenging claims made, but not consistently. In some cases, when the project documents mention, for example, a strategy for dealing with a certain issue, the negative evidence given by GRAIDY highlights a lack of information or clarity regarding the implementation of the strategy, which is the desired behavior. However, in other cases where a similar lack of information or clarity is present, this is not stated.

When asked for more general feedback on GRAIDY, *E1* stated that they would prefer for the information retrieval and grading aspect to be separate instead of one complete process and to be able to ask for more detail or clarification on answers given. All experts (*E1, E2, E3, E4*) noted the information retrieval and summarization features to be of use to their assessment process but found the grading component to not be usable in its current form.

**Quantitative Analysis**: Spearman's correlation coefficient analysis showed a very weak and non-significant negative correlation ($r = -0.10$, $p = 0.76$) of the grades given by GRAIDY compared to the experts. The negative correlation suggests an inverse relationship, where one set of grades tends to increase as the other decreases. However, the high p-value indicates this correlation is not statistically significant and likely due to random chance, suggesting no meaningful relationship between the grades. The current grading system may not be effective, possibly due to a lack of context or insufficient training examples. In the future, the grading could be done through a dedicated trained classifier instead of prompt-based learning, as suggested by *E3*. Additionally, a machine-generated grading could lead to more objectivity, as human grading tends to be subjective, as noted by *E4: "We agree on the content, but how you put it as a three or a two or four tends to be a little bit subjective."* Regarding hallucinations, no instances were found. All information stated came from somewhere in the documents provided. The citing quality was inconsistent, with correct citations in 6 of the 12 subareas. For several subareas, only page numbers were provided without filenames, making these references mostly unusable as they could correspond to any of the 23 documents provided. Still, citing has the potential to enhance the quality of the assessment, creating hyperlinks directly to the assessed document, as noted by *E4: "We often create due diligence reports under time pressure, which leads to cutting corners like citing without specific pages. A prototype that pinpoints the exact page is therefore highly useful."*

## 5. Discussion

Through the DSR methodology, we created GRAIDY, the first prototype of an LLM-based tool to assess sustainable forest investment projects. From the design and subsequent evaluation of this prototype, we derive the following early DPs (Gregor et al., 2020) for such a tool (Table 2) to answer *RQ1 How can we design an LLM-based tool to support the assessment of sustainable forest investment projects?*

Our first design goal was *DG1 Gather relevant information from provided documents*. Prior research shows LLMs can reliably retrieve sustainability-related information (Braunschweiler et al., 2023; Ni et al., 2023; Zou et al., 2023). This investigation found GRAIDY to perform inconsistently in this regard. It showed promise, often retrieving similar information to the expert assessor and, in some cases, finding data missed during manual assessment. Additionally, while LLMs' tendency to hallucinate was a reoccurring issue discussed in the literature, interestingly, no instance of hallucination was found in this evaluation, showing the effectiveness of RAG in reducing or even eliminating hallucination. Therefore we propose the DF1 of our prototype as *Design Principle 1: Use of RAG* as an effective way to find relevant information in documents.

However, there were many cases where clearly relevant information was not retrieved. One cause of this was found to most likely lie in the difficulty of LLMs in dealing with information contained in a table format, a finding shared by Luccioni et al. (2020). In some cases, the cause seemed to lie in the LLM having too little knowledge of what the desired information is, an issue that could be mitigated by using more comprehensive information retrieval prompts that more clearly outline the scope of each subarea. DF2 addressed this by allowing users to specify retrieval prompts. The use of just the default prompts in this paper's evaluation constitutes a limitation, but we regard this functionality as a further *Design Principle 2: Easily Adaptable Information Retrieval Prompts*.

From DF3 and DF5 of providing textual summaries of the retrieved information and positive and negative evidence, respectively, we define the combined *Design Principle 3: Textual Responses*. While the content varied in quality, the textual responses were generally regarded as a useful feature by domain experts, and we recommend their inclusion in the design to fulfill our second design goal *DG2 Provide summaries of relevant information and findings*. Similarly, while the implementation of DF4 of providing sources for all information stated was not yet of satisfactory quality, the functionality was regarded as helpful in principle by

**Table 2. Relationship between Design Goals, Features, and Principles**

| Design Goals | Design Features | Design Principles |
|---|---|---|
| DG1: Gather relevant information from provided documents | DF1: Use RAG as the core principle for gathering relevant information | DP1: Use of RAG for effective information retrieval. |
| | DF2: Allow the user to specify what information should be retrieved. | DP2: Easily adaptable information retrieval prompts |
| DG2: Provide summaries of relevant information and findings | DF3: Provide summaries of the retrieved information. | DP3: Textual responses including summaries and evidence. |
| | DF5: Provide positive and negative evidence | |
| DG3: Perform a complete assessment in a timely manner | DF6: Assign numerical grades. | DP4: Citing for verification and in-depth analysis. |
| | DF4: Provide sources for all information stated. | DP5: Document labeling for differentiated handling. |
| | | DP6: Collaboration through chat function. |

the experts, and we therefore nevertheless define *Design Principle 4: Citing*.

The prototype showed no ability to assign scores similar to those of an expert assessor. While some discrepancy has to be accounted for due to GRAIDY not having access to external sources, the scoring was found to not just deviate from the expert's scoring but also, in some cases, appear to contradict the evidence found by the prototype itself. Chiang and Lee (2023) and G.-G. Lee et al. (2024) significantly improved performance in scoring tasks using different prompting strategies. Based on their findings, refining the grading prompt to include more detailed instructions on the rating task and prompting the LLM to think step by step and explain its reasoning behind the numerical score might improve results. However, based on the results of this paper, we cannot recommend DF6 for assigning numerical scores as a design principle.

The discussion with Xilva's domain experts revealed two factors that are currently outside of GRAIDY's capability that are crucial for accurate assessment: the use of external sources to assess risk accurately and the differentiated handling of documents depending on their type and source. From this, we derive a design principle not yet implemented in the prototype but that we regard as a potential way to overcome these limitations: *Design Principle 5: Document Labeling*. Users could be asked to label each document they upload as being from the project team itself, a certification report, an audit report, etc. This would enable the differing treatment of documents based on their source as the LLM could then be prompted to treat these documents accordingly (e.g. "Use information in documents with label x to verify claims made in documents with label y."). With such a labeling system in place, external sources could then be introduced as well with their own label and specific instructions on how they should be used. The external sources are of different formats but are most commonly websites. Therefore, functionality could be added to extract text from more types of sources, such as websites. Since websites are in HTML (Hypertext Markup Language), which is a semi-structured format, their parsing is considered to be easier than PDFs. However, the domain experts noted many of the sources they use are maps, which cannot yet be meaningfully processed by a language model but would instead potentially require complex geospatial AI systems.

For some subareas in the grading methodology, such as "Natural hazards" and "Climate resilience to potential scenarios," the experts explained their process to start with assessing the level of risk through external sources. The information in the project documents is then assessed in the context of the identified level of risk, with a mitigation strategy not having to be as comprehensive when risk is judged to be low and vice versa. This risk assessment through external sources is complex and potentially impossible to achieve with an LLM-based system. We, therefore, recommend potentially leaving this step of the process to the expert assessor and providing a risk score or similar metric as additional input to the LLM for the relevant subareas. This leads us to propose again a design principle not yet implemented in GRAIDY: *Design Principle 6: Collaboration through Chat Function*. The idea of this principle is that a chat functionality, similar to Ni et al. (2023) ChatReport, would enable more input from the expert assessor into the tool's assessment. We propose this additionally as a solution for the cases where clarification or more information might be needed for answers provided by the tool.

The evaluation of the prototype further allows us to answer ***RQ2*** *Can LLM-based assessment serve as a viable alternative to human experts in assessing sustainable forest investment projects?* We found through our evaluation that LLM-based assessment is not currently a viable alternative to human assessment,

but we regard the information retrieval capabilities of GRAIDY to be useful to expert assessors and see further potential in the utility of LLM-based tools through the implementation of the above-defined design principles, which overall put the focus on designing a tool to assist human experts in their assessment process instead of attempting to complete the assessment in its entirety without human input.

## 6. Conclusion

Assessing sustainable forest investments for long-term efficacy and sustainability is an important but time-consuming task. This paper investigated using LLM-based assessment tools as an aid or potential alternative to human experts and proposed early design principles for such tools.

Retrieval-augmented generation has proven effective in mitigating hallucinations and is promising for finding and summarizing information from documents. Despite inconsistent quality, this paper concludes that LLM-based tools are well-suited for information retrieval. However, the prototype in its current state cannot meaningfully grade a project. Existing literature shows promise, and the prototype's performance could improve with the implementation of this paper's suggestions, though key challenges remain in using LLMs for this task. Regarding whether LLMs can replace human experts in assessing sustainable forest investment projects, this paper supports the conclusion that assessment should not be conducted by LLMs alone. Instead, LLM-based systems can assist domain experts by extracting relevant information from large document collections. For this process of assistance, we proposed design principles as our core contribution.

The prototype evaluated in this paper was intended to assist, not replace, domain experts, but its current implementation has limited utility. We suggest focusing future research on improving information retrieval and implementing "chat with your documents" functionality to enhance collaboration with AI and better support domain experts.

## Acknowledgements

We thank Xilva and their experts for their valuable support throughout the project. We also thank the University of Zurich and the Digital Society Initiative for (partially) financing the study through Liudmila Zavolokina's DIZH postdoc fellowship. Lastly, we thank Innosuisse for funding this research under the Innocheque Grant *70192.1 INNO-SBM* enabling our collaboration.

## References

Angin, M., Taşdemir, B., Yılmaz, C. A., Demiralp, G., Atay, M., Angin, P., & Dikmener, G. (2022). A RoBERTa Approach for Automated Processing of Sustainability Reports [Number: 23]. *Sustainability*, *14*(23), 16139. https://doi.org/10.3390/su142316139

Braunschweiler, N., Doddipatla, R., Keizer, S., & Stoyanchev, S. (2023, September). Evaluating Large Language Models for Document-grounded Response Generation in Information-Seeking Dialogues [arXiv:2309.11838 [cs]]. https://doi.org/10.48550/arXiv.2309.11838

Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., & Staiano, J. (2024, January). Glitter or Gold? Deriving Structured Insights from Sustainability Reports via Large Language Models [arXiv:2310.05628 [cs]]. https://doi.org/https://doi.org/10.48550/arXiv.2310.05628

Chiang, C.-H., & Lee, H.-y. (2023, May). Can Large Language Models Be an Alternative to Human Evaluations? [arXiv:2305.01937 [cs]]. https://doi.org/10.48550/arXiv.2305.01937

Cohen-Shacham, E., Walters, G., Janzen, C., & Maginnis, S. (Eds.). (2016, August). *Nature-based solutions to address global societal challenges*. IUCN International Union for Conservation of Nature. https://doi.org/10.2305/IUCN.CH.2016.13.en

Deutz, A., Heal, G., Niu, R., Swanson, E., Townshend, T., Li, Z., Delmar, A., Meghji, A., Sethi, S., & Puente, J. (2020, September). *FINANCING NATURE: Closing the Global Biodiversity Financing Gap*. https://doi.org/10.13140/RG.2.2.26226.32968

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv:1810.04805 [cs]]. https://doi.org/10.48550/arXiv.1810.04805

Forum, W. E. (2021). Investing in forests – The business case. Retrieved June 12, 2024, from https://www3.weforum.org/docs/WEF_Investing_in_Forests_2021.pdf

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: The anatomy of a design principle. *Journal of the Association for Information Systems*, *21*(6), 2.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems

Research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*. https://doi.org/10.1016/j.caeai.2024.100213

Lee, H., Lee, S., Park, H., Kim, J., & Jung, H. (2024). ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models. *Heliyon*, *10*(4). https://doi.org/10.1016/j.heliyon.2024.e26404

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. Retrieved April 4, 2024, from https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Luccioni, A., Baylor, E., & Duchene, N. (2020, November). Analyzing Sustainability Reports Using Natural Language Processing [arXiv:2011.08073 [cs]]. https://doi.org/10.48550/arXiv.2011.08073

Ni, J., Bingler, J., Colesanti-Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Stammbach, D., Vaghefi, S. A., Wang, Q., Webersinke, N., Wekhof, T., Yu, T., & Leippold, M. (2023, October). CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-based Tools [arXiv:2307.15770 [cs]]. https://doi.org/10.48550/arXiv.2307.15770

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024, March). GPT-4 Technical Report [arXiv:2303.08774 [cs]]. https://doi.org/10.48550/arXiv.2303.08774

Pasch, S., & Ehnes, D. (2022). NLP for Responsible Finance: Fine-Tuning Transformer-Based Models for ESG. *2022 IEEE International Conference on Big Data (Big Data)*, 3532–3536. https://doi.org/10.1109/BigData55660.2022.10020755

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Polignano, M., Bellantuono, N., Lagrasta, F. P., Caputo, S., Pontrandolfo, P., & Semeraro, G. (2022, June). An NLP Approach for the Analysis of Global Reporting Initiative Indexes from Corporate Sustainability Reports. In M. Wan & C.-R. Huang (Eds.), *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference* (pp. 1–8). European Language Resources Association. Retrieved April 4, 2024, from https://aclanthology.org/2022.csrnlp-1.1

Raymond, C. M., Frantzeskaki, N., Kabisch, N., Berry, P., Breil, M., Nita, M. R., Geneletti, D., & Calfapietra, C. (2017). A framework for assessing and implementing the co-benefits of nature-based solutions in urban areas. *Environmental Science & Policy*, *77*, 15–24. https://doi.org/10.1016/j.envsci.2017.07.008

Seddon, N., Chausson, A., Berry, P., Girardin, C. A. J., Smith, A., & Turner, B. (2020). Understanding the value and limits of nature-based solutions to climate change and other global challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1794), 20190120. https://doi.org/10.1098/rstb.2019.0120

Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., Allen, S., Colesanti-Senni, C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Wang, Q., Webersinke, N., Huggel, C., & Leippold, M. (2023). ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, *4*(1), 1–13. https://doi.org/10.1038/s43247-023-01084-x

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022, December). ClimateBert: A Pretrained Language Model for Climate-Related Text [arXiv:2110.12010 [cs]]. https://doi.org/10.48550/arXiv.2110.12010

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2023, December). ESGReveal: An LLM-based approach for extracting structured data from ESG reports [arXiv:2312.17264 [cs]]. https://doi.org/10.48550/arXiv.2312.17264