BIS Innovation Hub

Project Gaia

Enabling climate risk analysis using generative AI

March 2024

BANCO DE ESPAÑA
Eurosistema

DEUTSCHE BUNDESBANK
EUROSYSTEM

EUROPEAN CENTRAL BANK

www.bis.org
email@bis.org

Follow us

Executive summary

Project Gaia – a collaboration between the BIS Innovation Hub Eurosystem Centre, the Bank of Spain,[1] the Deutsche Bundesbank and the European Central Bank – leverages generative artificial intelligence (AI) to facilitate the analysis of climate-related risks in the financial system.

Central banks, supervisory authorities and financial institutions need higher quality and more accessible data to model the financial risks posed by climate change. Today, due to the lack of global reporting standards, accessing relevant climate-related indicators takes significant effort. In financial institutions' corporate reports, climate-related data are buried among other financial and non-financial information and, in many cases, information pertaining to one company is split across multiple reports, and relevant information is contained in texts, tables, footnotes and figures. These challenges constrain the usability of climate-related information.

Project Gaia aims to help analysts search corporate climate-related disclosures and extract data quickly and efficiently using AI, particularly large language models (LLMs). Gaia Phase I has surveyed climate risk experts from central banks and supervisory authorities, designed a solution that addresses the requirements articulated by these experts and delivered a proof of concept (PoC) demonstrating the technical feasibility of the concept.

By automating information extraction, Gaia opens up the possibility of analysing climate-related indicators at a scale that was not previously feasible. Furthermore, Gaia offers harmonised metrics despite the heterogeneity of naming conventions and definitions across different jurisdictions. The combination of semantic search together with iterative and systematic LLM prompting enables Gaia to overcome differences in disclosure frameworks. This offers much needed transparency and comparability of climate-related information.

Project Gaia breaks new ground by integrating LLMs into an application and leveraging it for data extraction. This poses several technical challenges, including LLMs' long response times, randomness (non-repeatability) in their responses and hallucinations. This report explains a set of concrete design choices that allow the Gaia PoC to overcome these challenges.

Gaia demonstrates the power of creating AI-enabled intelligent tools to automate existing workflows. For example, macro analysis results presented in this report cover 20 key performance indicators (KPIs) for 187 financial institutions over five years and adding more institutions or KPIs is quick and easy. Due to its flexible design, the platform is relevant in a much broader context than climate-related data analysis. This paves the way for AI-enabled applications for central banks and the financial sector to address, for example, regulatory and supervisory use cases. Generative AI promises to change the way we work in the future and Project Gaia is one of the first comprehensive studies investigating how this can be done in practice.

1 *Building on previous work published by Moreno and Caminero (2020, 2022, 2023).*

List of abbreviations and acronyms

AI	Artificial intelligence
BIS	Bank for International Settlements
DC	Design choice
ESG	Environmental, social and governance
JSON	Java script object notation
KPI	Key performance indicator
LLM	Large language model
NGFS	Network for Greening the Financial System
NLP	Natural language processing
PoC	Proof of concept
SQL	Structured query language
TCFD	Task Force on Climate-related Financial Disclosures

Table of contents

1. The climate data challenge

The climate data challenge

Climate-related financial risks include physical and transition risks. Physical risks can result from extreme weather events (such as droughts, floods and wildfires) or from long-term shifts in weather patterns. Transition risks relate to the financial consequences of moving to a climate-friendly economy, for example due to changing firm valuations. These risks represent a new challenge for financial institutions and there is a growing need to assess each organisation's exposure to climate-related risks.

Climate change can also have an impact on overall financial stability. (Battiston et al (2021)). Central banks and supervisors increasingly need to perform climate risk analysis to assess the financial system's vulnerabilities to climate change. To perform these assessments, they need high-quality, comparable and accessible data.

The richest insights into a company's climate-related practices are contained in their environmental, social and governance (ESG) disclosures. Regulations, frameworks and standards form the backbone of ESG disclosures and play interconnected roles in guiding how organisations report their sustainability practices. They are also essential for climate data transparency as they establish guidelines for disclosing and addressing climate-related risks. They promote transparency, foster accountability and help mitigate risks associated with climate issues.

There are several challenges impeding the availability of information about an institution's objectives and actions in relation to climate risks:

- **Lack of standardisation** – ESG reporting lacks consistent global standards, leading to a proliferation of different frameworks and metrics. This makes it difficult to extract and compare ESG indicators across companies and industries.

- **Data quality and accuracy** – ensuring the accuracy and reliability of ESG data can be challenging, as it usually relies on self-reported information and sometimes on proxies, which may be subject to human error, inaccuracies, bias or lack of transparency.

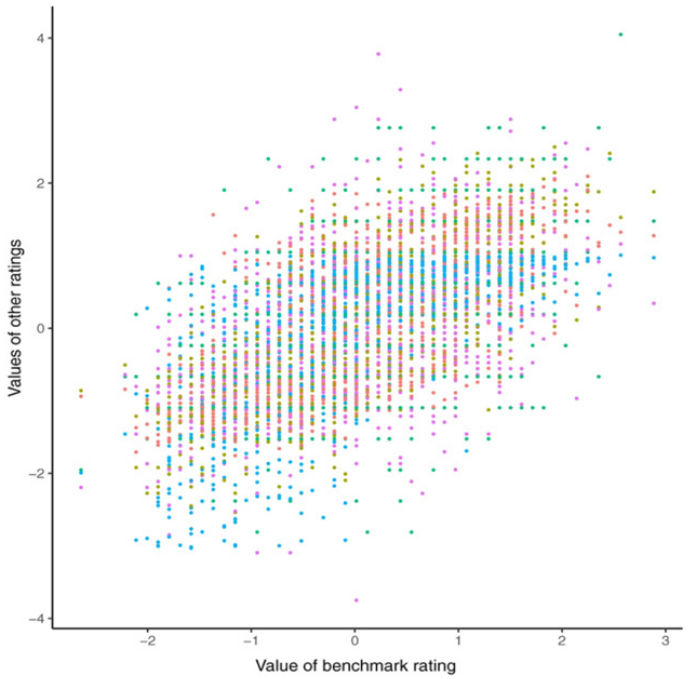
- **Scope and materiality** – determining which ESG factors are most material to a company's performance and which should be disclosed is an ongoing challenge, as these vary by industry and context.

- **Regulatory fragmentation** – different countries and regions have developed their own ESG reporting regulations, leading to a complex and fragmented landscape.

In addition, collecting climate-related indicators from corporate ESG reports often requires significant manual effort. Climate-related data are buried among other financial and non-financial information, and in many cases, information pertaining to one company is split between multiple reports. A further challenge is that relevant information may be contained in a combination of text fragments, tables, footnotes and figures – all of which may need to be analysed to extract the desired indicator.

These data challenges are well illustrated by the divergence of ESG ratings between different rating agencies (Graph 1) Berg et al (2022) show that such divergence is due to differing interpretations of the underlying data. The analysis questions the reliability of ESG data and highlights the need for analysts to obtain more granular information.

Divergence of ESG ratings - Graph 1



This graph illustrates ESG ratings for 924 firms. Each dot corresponds to one firm. The horizontal axis indicates the value of one rating provider. Rating values by the other five raters are plotted on the vertical axis in different colours. For each rater, the distribution of values has been normalised to zero mean and unit variance.
Source: Berg et al (2022).

An alternative to relying on ESG reports directly is to use commercial data providers. Various companies offer data sets on sustainability risks, usually relying on manually extracted information from ESG reports, enriched with estimated data based on proprietary algorithms. Commercial data currently fuel sustainability risk analysis conducted by financial market actors worldwide but there is certainly room for improvement. Data reliability remains an issue, since variance among vendors can be large. Accessibility is limited by licensing restrictions despite the data often being public at the source. Furthermore, proprietary algorithms make it difficult for supervisors and regulators to base reproducible and explainable decisions on proprietary data.

Consequently, many international standard-setting bodies highlight the need to close climate data gaps and render existing information usable. It is emphasised that the need for high-quality, comparable climate-related data continues to be a pressing issue (NGFS 2022).[2] Some also point to the difficulties of working with data from third-party climate data providers and the restrictions on usage rights associated with this. Ferreira et al (2021) have argued for strengthening the "climate information architecture" built on data availability, disclosure standards and classification approaches. Gaia aims to fill this gap by improving the accessibility and usability of a currently underused source of climate-related information.

2. The Network for Greening the Financial System (NGFS) is a group of central banks and supervisors willing, on a voluntary basis, to share best practices and contribute to the enhancement of environmental and climate risk management, and to mobilise mainstream finance to support the transition towards a sustainable economy. The NGFS has identified three building blocks for reliable and comparable climate-related data: (i) a rapid convergence towards a common and consistent set of global disclosure standards; (ii) efforts towards a minimally accepted global taxonomy/shared principles for sustainable finance; and (iii) the development of transparent metrics and methodological standards.

2. Project scope and methodology

Project scope and methodology

Problem statement

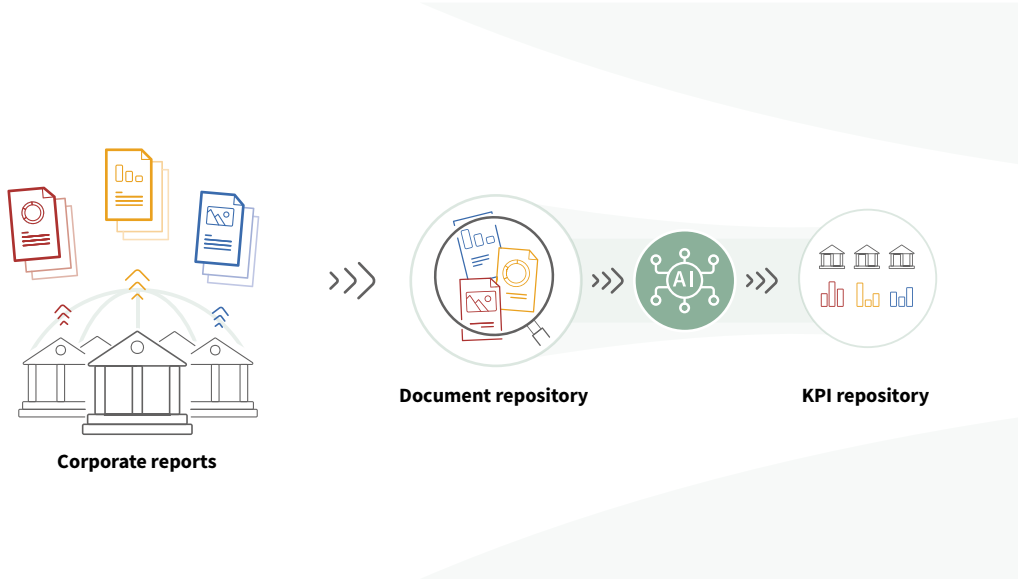
In order for the financial sector to conduct effective climate risk analysis, investors, central banks, supervisors and other relevant stakeholders need access to relevant climate-related indicators. Today, accessing such indicators requires significant manual effort. Additionally, the complex landscape of regulations, frameworks and standards has led to varying reporting practices and data, which limit the usability and comparability of available information.

These challenges underscore the necessity for a comprehensive solution to ensure relevant, reliable and comparable climate-related financial data.

Gaia value proposition

The vision of Project Gaia is to create an open web-based tool that helps analysts and supervisors search corporate climate-related disclosures and extract data, thereby reducing the manual effort involved in climate assessments. The tool works by extracting structured information from unstructured PDF documents, combining all the information elements, such as text, tables and figures. It is based on cutting-edge technologies, notably AI, more specifically large language models (LLMs), to extract relevant information and present it in an easy-to-use form (Graph 2).

Gaia value proposition - Graph 2



Gaia adds value by extracting and structuring data for climate risk analysts.
Source: Project Gaia.

This approach does not require the harmonisation of guidelines or regulatory requirements but takes advantage of detailed public information available today. Gaia promises to overcome the challenges of differences between disclosure frameworks and provides an independent and reliable source of previously underused climate risk-related data. Project Gaia Phase I has developed a functioning PoC of such a tool and demonstrated the technical feasibility of the concept.

This opens up the possibility of analysing climate-related key performance indicators (KPIs) at a scale that was previously unimaginable. With the traditional manual approach of collecting KPIs for analysis, each additional KPI and each additional institution requires dedicated manual work. The analyst either needs to search for the information in public corporate reports or contact the institution for information. With the Gaia approach, once the platform is available, searching for new KPIs or adding new institutions comes at near-zero costs and with very little delay.

Project Gaia is an experiment carried out for purely research purposes (ie: an applied research project), in the public interest and on a not-for-profit basis.

The value proposition of Project Gaia can be summarised as follows:

- **Enhanced accessibility to climate-related disclosures** – Project Gaia envisions an open web-based tool that makes it easier for financial supervisors and macro analysts to access and search corporate climate-related disclosures.

- **Efficient data extraction** – the project leverages AI techniques, particularly LLMs, to extract climate-related data from corporate reports quickly and efficiently. This automation reduces the time and effort required for climate-related analysis.

- **Harmonised climate metrics** – Project Gaia relies on each KPI's definition rather than its name when searching corporate reports, which allows it to offer harmonised metrics despite the heterogeneity of naming and definitions across different jurisdictions and companies.

- **Transparently generated data** – Project Gaia provides transparency and traceability by providing a justification and direct view into the sources for each KPI extracted.

- **Scalability and reliability** – the PoC implementation is constructed with enterprise-grade components and designed for high scalability to meet the changing needs of analysts and the growing demand for climate-related data.

- **Flexibility** – the flexible design ensures that the Gaia platform can be easily configured to extract other types of KPIs besides the ones used in the project and is hence potentially applicable to a much broader context than climate-related data analysis.

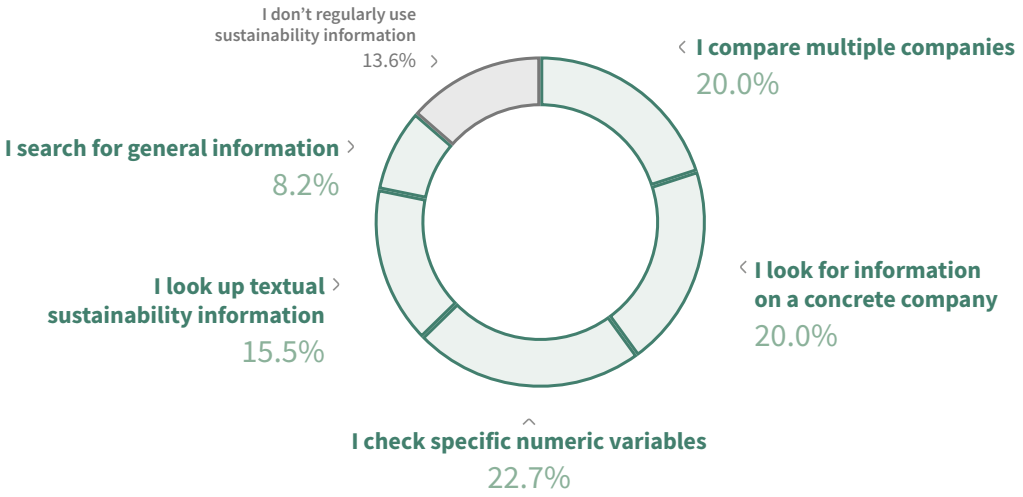
User survey

Project Gaia adopted a user-centric approach putting the needs, preferences and experiences of users at the forefront of the project's development. This principle was implemented by engaging users and incorporating feedback at various stages of the project.

The project started by researching and defining in detail the problem to be addressed. This phase was paramount for gaining a deep understanding of the challenge and forming a solid foundation for developing an innovative and user-centred solution. For this purpose, a forum called the Green Finance AI Working Group was established. The working group encompasses climate-related data users from 18 different central banks and supervisory and regulatory institutions. They work in a wide range of business areas.

Through the engagement of the Green Finance AI Working Group, Project Gaia was able to distil user stories to capture and communicate requirements and expectations. A survey was used to check and refine user stories from a diverse central bank and supervisor audience working with climate-related data.[3] Out of all respondents, 71% currently use sustainability reports. This includes 66% of respondents who primarily rely on sustainability reports in their work to investigate specific institutions, as well as 20% who compare multiple institutions.

Current usage of sustainability reports by financial supervisor experts - Graph 3



Project Gaia survey among 45 financial supervisor experts from 17 countries.
Source: Project Gaia.

3. The survey yielded 45 responses from international experts. Responses came from the following countries: Austria, Belgium, Croatia, France, Germany, Greece, Hong Kong SAR, Ireland, Lithuania, Luxembourg, the Netherlands, North Macedonia, the Philippines, Spain, Türkiye, the United States and one supranational organisation.

According to the survey results, the most valuable variables from sustainability reports include CO_2 equivalent emissions (highlighted by 18.6% of respondents), decarbonisation targets (15.8%), emission intensities (14.5%), energy intensities (11.8%) and energy consumption (11.3%). Furthermore, the survey confirmed that 22.7% of respondents use sustainability reports to check specific numeric values, while 20% use them for comparisons or information about specific companies (Graph 3).

Input from the Green Finance AI Working Group also informed several concrete design decisions in the Gaia PoC platform. For example, the working group emphasised the need to be able to trace an extracted KPI to the source, which became a key feature of the PoC.

Based on the learnings obtained from the working group, the project identified two personas as initial hypothetical users of the solution. The supervisor needs detailed information about one institution. They filter for this organisation and access all possible KPIs, including potential relations between the KPIs or comparisons between different years. In contrast, the macro analyst is interested in the overall development of climate risk and in climate-related actions based on established KPIs.

3. Platform design

Platform design

The evolution of Project Gaia is testimony to recent rapid development in the field of AI. At the time of project planning, LLMs had limited availability. The breakthrough in LLMs opened up new possibilities and a redesign of the PoC was therefore required. As much as changing the engine in a car from combustion to electric implies major internal design changes, changing from "classical" natural language processing (NLP) tools to large language models implies new paradigms throughout the application. To give an example: classical NLP tools require task-specific model training whereas LLMs are universally pretrained and can solve many tasks if they are given the task-specific information or "context".

Project Gaia breaks new ground by integrating LLMs into an application intended for the financial industry and leveraging it for text analysis at scale. For example, LLMs' long response times, randomness (non-repeatability) in their responses and hallucinations pose a real challenge in designing an LLM-based application. By addressing these and other challenges, Project Gaia has delivered insights that are applicable beyond climate data analysis, for example in regulatory and supervisory use cases. This section, intended for the expert reader, summarises these insights.

Design principles

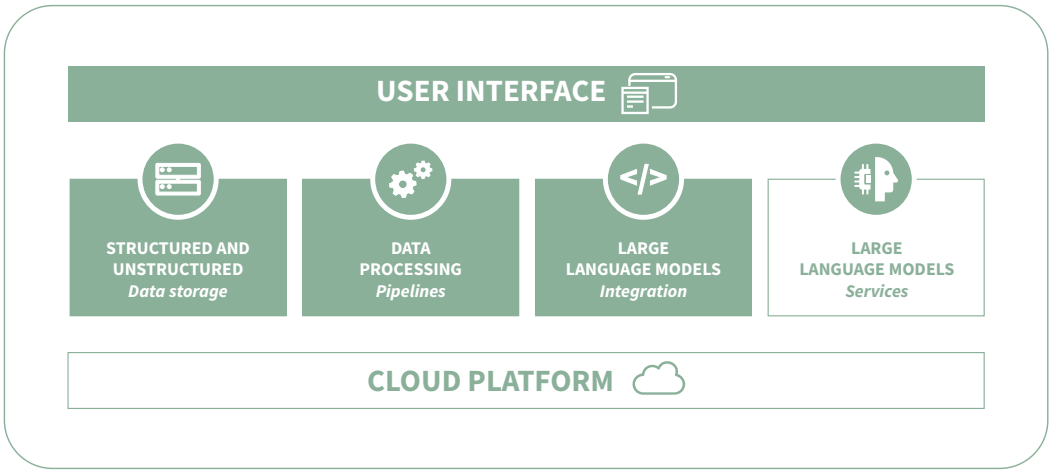
The Gaia PoC implementation is constructed with enterprise-grade components and designed for high scalability – facilitating rapid transition to a production environment in an eventual later stage. To achieve this objective, the project combined the following design principles:

- **Infrastructure as code** – all code is stored in version-controlled repositories, simplifying the deployment process and making it easily replicable at different locations.
- **Standard, enterprise-grade components** – by employing universally recognised, scalable and enterprise-ready components, the system ensures robust performance and adaptability. Wherever possible, it relies on standard cloud services to reduce the maintenance footprint.
- **Common design patterns** – the platform leans on established design patterns and widely used programming languages to promote code reusability and maintainability.
- **Loosely coupled architecture** – the codebase is entirely built around the concept of loosely coupling components. All components are set up as independent (containerised) microservices. This ensures Gaia can add, interchange, extend or delete components where needed. This is also true for the LLM, which can be exchanged for another model as technology progresses or the need for on-premise processing arises.
- **Cloud first –** reliance on cloud services reduces the infrastructure and maintenance burden.
- **Flexibility –** KPIs are not hard coded into the platform, but provided as configuration parameters along with their definition to be used in the extraction.

High-level architecture

Graph 4 illustrates the Gaia architecture, which consists of structured and unstructured data storage, data processing pipelines, LLM integration and a user interface, all on top of a cloud platform. The solution connects to an external LLM service via an application programming interface (API).

High-level architecture - Graph 4



High-level architecture showing the major building blocks of Gaia. Green items are part of Gaia that have either been custom built or configured. White items are external building blocks used by Gaia.
Source: Project Gaia.

Hosted on the Microsoft Azure cloud, the architecture aims to minimise the maintenance footprint and streamline the deployment process, enabling high scalability and easy customisation in the future. Most prominently, Gaia uses Azure Cognitive Services to parse the PDF documents, using the Form Recognizer and Azure OpenAI Service to access LLM. For processing, Gaia uses the Azure Kubernetes Service (AKS), which hosts Gaia's custom-built services, such as the user interface (front end) and the LLM-based processing pipeline. Kubernetes allows Gaia to dynamically add and remove service instances as the workload changes, making the platform very scalable. The scalability is further underpinned by connecting the services via a publish-subscribe messaging system (Kafka) instead of a more conventional API. This design choice also helps the application handle the long response times of LLMs.

Data processing and storage

Data processing is performed by four data processing pipelines:

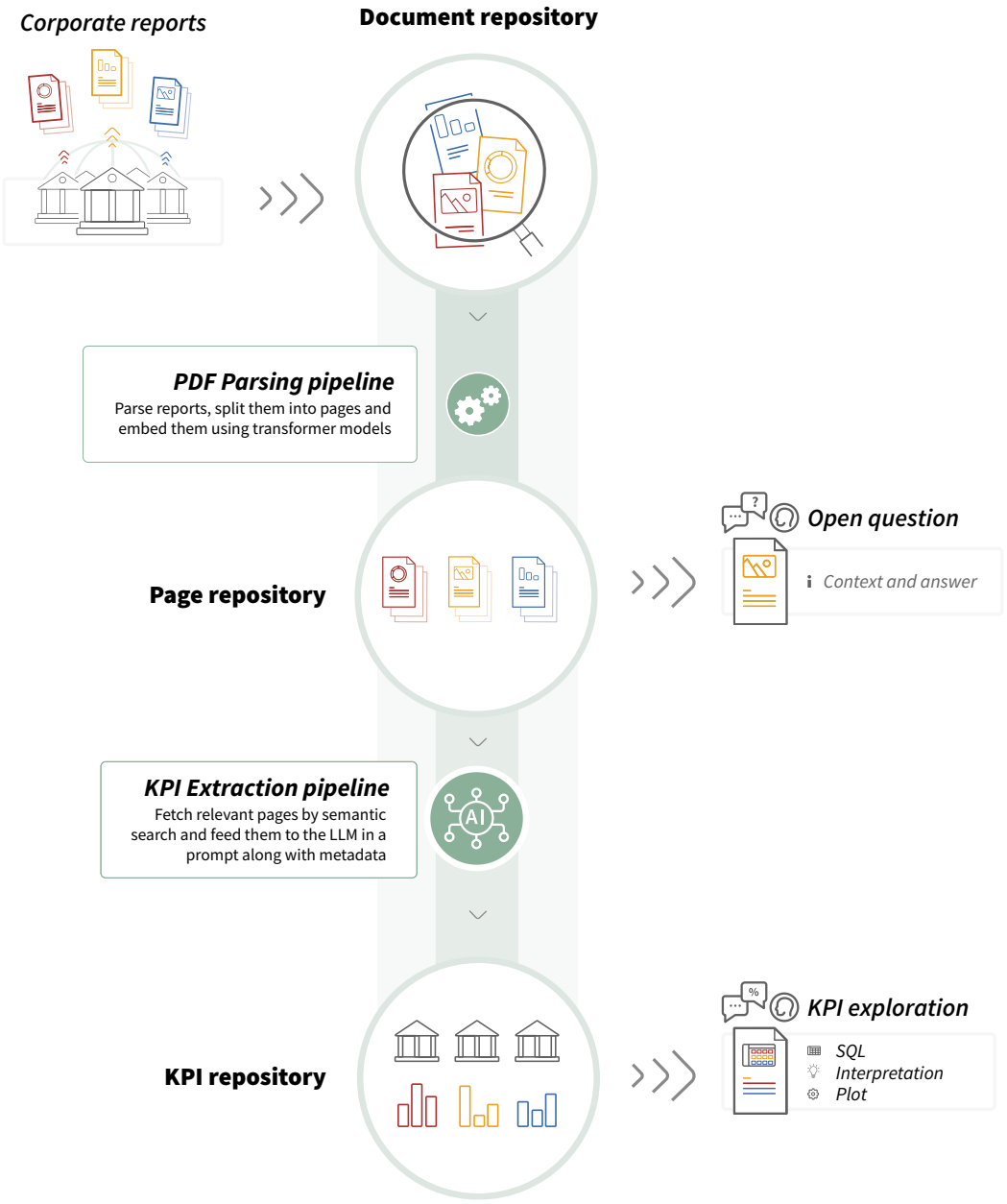
1. The PDF parsing pipeline.
2. The KPI extractor pipeline.
3. The open query pipeline.
4. The structured data exploration pipeline.

The pipelines and the data storage components are illustrated in Graph 5.

Raw input files – a library of corporate reports – reside in the document repository, which is implemented in Azure Data Lake.

The PDF parsing pipeline takes its input from the document repository in the form of PDF documents. The documents are processed using the Azure AI Document Intelligence into a machine-processing friendly JSON format and stored alongside the original documents. Next, the documents are split into individual pages and uploaded into the page repository, an OpenSearch document database. For each page, a vector embedding is calculated using sentence transformer models and this is stored in the metadata of the respective page record.

Data processing pipelines - Graph 5



Gaia consist of a set of automated processing pipelines, transferring newly submitted reports to the page repository and extracting KPI into the KPI repository. From this pipeline, specific use cases are implemented serving Gaia's customers.
Source: Project Gaia.

The core of the Gaia PoC is the KPI extraction pipeline, which uses LLMs to extract a given KPI for a set of reports available in the page repository. It starts with a semantic search yielding the most likely pages to contain information on a given KPI. Next, those pages are integrated into a prompt that is submitted to the LLM. The LLM responds with the KPI's value, unit, source and other relevant metadata which are stored in a relational database called the KPI repository.

Two additional pipelines facilitate interaction for the non-technical user by providing AI-based workflows for explorative analysis. Both workflows can process a request in human language and aggregate and process the KPI data to address the needs of specific personas. The open question pipeline involves contextualising information for LLM-driven query answering. This contextualisation aids in formulating queries that are then processed by the LLM to provide answers to the user's queries. This capability is particularly useful for the supervisor persona, which poses broad ad hoc textual questions related to an organisation's operations or metrics.

In contrast, the macro data exploration query leverages structured data within the KPI repository and uses LLMs to automatically generate a structured query language (SQL) query based on the user's textual request. The SQL query is run on the KPI repository, and the resulting output, represented as a table, is fed back into the LLM for interpretation and amendments, which are subsequently returned to the user. This workflow is useful for the macro analyst persona who is interested in the overall climate risk landscape based on established KPIs.

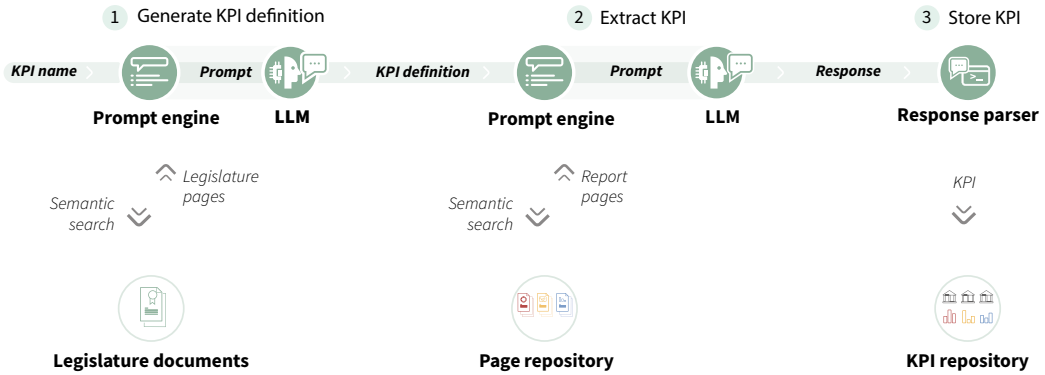
All pipelines are implemented as containerised microservices. The data pipeline consists of sequential steps; each step is executed by a dedicated microservice, communicating through a scalable streaming message service.

LLM integration

LLMs offer powerful capabilities in handling unstructured data, including an ability to summarise vast amounts of textual information[4]. It may seem straightforward to simply query the LLM for the required information, such as a KPI that one wants to extract. But on a closer look, however, this view is revealed to be oversimplified. The LLMs do not, and cannot, have all the information about the reports from their training and thus need to be provided with proper context from the reports themselves. The amount of text LLMs can process at once is limited to a few thousand words, which is exceeded by most corporate disclosures. In addition, too much information "confuses" current LLMs. These limitations mandate a more sophisticated solution than "just asking LLM", including the preselection of relevant information submitted to the LLM, as well as prompt engineering.

4. A recent study by Moreno and Caminero (2023) has shown the potential of LLMs in this context.

KPI extraction pipeline - Graph 6



Steps starting from the KPI definition through finding relevant pages in the report to compiling the prompt and feeding the result into the KPI repository.

Source: Project Gaia.

LLMs are a new and evolving technology, and there is only a limited set of best practice examples available. A major contribution of Project Gaia is the set of design choices that were made for efficient and reliable LLM integration (Graph 6 and Table 1). These design choices are a result of a thorough optimisation process, which has entailed over 50,000 LLM queries to date. In what follows, the key steps of KPI extraction using LLMs are described, along with the design choices related to each step.

Design choices for efficient LLM integration - Table 1

Nr	Design choice (DC)
DC1	Defining the KPI in the context of green finance
DC2	Tranche size of the document splitting
DC3	Embedding model for the semantic search
DC4	Semantic search similarity metric
DC5	Number of hits from the semantic search
DC6	Prompt engineering
DC7	LLM self-assessment
DC8	LLM model and version
DC9	LLM role
DC10	Creativity (temperature) parameter
DC11	Parsing the output

KPI context (DC1)

KPIs such as "gross direct GHG emissions (Scope 1)" or "total energy consumption" have a clear definition in the field of green finance but cannot be expected to be understood within the general training scope of the LLMs. The first step is to create a succinct definition of the KPI within the context of the field of green finance (DC1). To this end, the OpenSearch database also contains an index of reporting standards and legislature documents. The LLM is instructed to "Write a concise five to seven sentence definition for [KPI]" based on relevant pages from the standards and legislature documents, which are passed as context. The generated KPI definition becomes a cornerstone in the semantic search (DC4) and is also reused in the subsequent LLM prompt (DC6).

Page pre-selection (DC2 – DC5)

For a successful extraction of the KPI, the relevant information (eg text passage or table) needs to be passed into the limited context of the LLM. As an entire document is typically too large for the context, each report is broken down into smaller units, Gaia uses pages (DC2). Each page is stored together with its semantic vector encoding. Gaia uses the ADA 2 vector embedding model from the Azure Open AI Services (DC3). The KPI definition is also encoded into a vector and a similarity search yields the pages which are most likely to contain the desired information. To measure the similarity between vectors, Gaia uses an L2 norm (DC4). This semantic search is more resource intensive than a traditional keyword search, but it produces superior results in cases where specific keywords are not present and it works across languages. Gaia uses the top five pages (DC5) based on vector similarity to be passed on to the LLM's context.[5]

Prompt engineering (DC6 – DC7)

Graph 7 shows the LLM prompt (DC6), which combines several building blocks of instructions:

```
1.  Role
2.  What to do exactly
3.  The format in general (ie JSON)
4.  How to shape the specific answer within the format
5.  The content from the pages containing the desired information
6.  What NOT to do and how to behave in fringe cases
```

The prompt instructs the LLM to find a numeric value for the desired KPI, and report its unit, the document and the page it was found in, giving a short quotation to highlight in the text. Finally, it should comment on how the information was found: eg in a table or inferred from multiple numbers and lastly give self-assessed certainty about the provided information (DC7).

5. The design pattern of injecting information into the context of the LLM based on a semantic search has since been established as retrieval augmented generation (RAG).

Gaia LLM prompt template - Graph 7

```
System: You are a helpful climate and green finance risk analyst.

Human: Based only on the excerpts from the corporate reports provided below, we would
like to extract the KPI information for the year {year}. Specifically, we are looking
for the following:
- The numerical value and unit of "{KPI}".
- A short comment explaining how this value was obtained.
- The sources used, including the page number and a short quotation from that page.
- An indicator of certainty that ranges from 100 (absolutely certain) to 0 (cannot be
determined based on available information).

 If the information for the KPI cannot be found or determined from the provided
 documents, please generate a JSON object with the appropriate fields set to 'null' and
 include a comment stating that the information was not available.

Please note that the output will be a JSON object, structured according to a predefined
schema:
{format_instructions}

### Definition of'{KPI}':
{KPI_definition}

### Report Excerpts:
{reference_docs}

If no relevant information is found in the provided excerpts, the output should clearly
reflect this with a 'null' value or an appropriate indication of the absence of data.
The 'certainty' field should reflect the level of certainty of the information
provided, including a value of '0' if the KPI could not be determined based on
available information.
```

Variables "{...}" are filled with the necessary information for a given KPI. Colours correspond to listing in the text: black – general instructions, red – definition of the output format, orange – definition of the information to be supplied in the response, green – pages from the report.
Source: Project Gaia.

LLM model and parameters (DC9 – DC10)

GPT4 (DC8) receives the prompt and is instructed to act as a "helpful climate analyst" (DC9), setting the parameter controlling the creativity (ie the temperature) to zero (DC10) to get the highest degree of reproducibility.

Parsing the output (DC11)

The response is parsed and deconstructed into individual fields and uploaded into the relational database. Should the LLM fail to produce a valid JSON object, the query is retried once (DC11) and, in case there is a second successive failure, it is logged for manual resolution.

4. Testing & Evaluation

Testing & Evaluation

The PoC was evaluated using a test set of publicly available corporate reports from 187 financial institutions from across the world of systemic importance in terms of size, interconnectedness, complexity (including cross-border activity) and financial institution infrastructure. Financial institutions that are part of a group were represented at their highest level of accounting consolidation. The geographical provenance of institutions in the sample is 35% European, 35% Asia-Pacific, 15% North American and 16% from the rest of the world. In total, the test set included 2,328 documents (ESG, Pillar 3, financial statements, annual reports and other relevant public reports), covering a period of five years, from 2018 to 2022. The number of possible company-year combinations amount to 862 (note that there are 73 cases with no available report for a given company and year) and it gives the number of possible values for each type of KPI that can be potentially extracted from the documents. Currently, 20 KPI types have been extracted by the Gaia platform (list in Graph 10) and, due to the flexible design, new KPIs can be added easily by uploading additional reference documents (DC1).

Reliability of results

The lack of systematic, granular and reliable data poses a challenge when assessing the quality and accuracy of Gaia data extraction. As a statistical approach to benchmarking, one can consider the analyses of the Task Force on Climate-related Financial Disclosures (TCFD, 2023). The TCFD investigated the alignment of corporate reports with its recommendations for different sectors over a three-year period. It found that, within a sample of 235 global banks, from 2020 to 2022 between 40 and 58% of companies per year reported in line with the TCFD recommendations on the disclosure of Scope 1, Scope 2 and Scope 3. Project Gaia's data indicate that from 2020 to 2022, on average, 47% of the financial institutions in the data set reported all three scopes of GHG emission KPIs. This is in line with the TCFD's findings and provides a first indication of reliability.

For more granular benchmarking, Project Gaia relied on three commercial data sources, comparing them manually to the output obtained using the Gaia platform. The focus was on Scope 1 emissions as this is one of the most widely available KPIs. One challenge with this benchmarking exercise is that the commercial data sources do not cover Scope 1 emissions for all institutions and, in a significant portion of the sample, there are inconsistencies between the values reported by different commercial data sources. This can be due to the heterogeneity of their information sources, which include public corporate reports, as well as other sources. Despite this challenging environment, values extracted by Gaia match[6] at least one of the commercial data sources in 74% of cases where external data are available. Given the complexity of the ESG reporting landscape, this can be considered a very high initial rate of alignment and is promising for future applications of the approach.

A third level of benchmarking consisted of manual cross-checking against human data extraction in a random set of 163 examples. In Graph 8, the column on the left corresponds to 104 cases in which Gaia was able to extract a value from the report. Out of these cases, Gaia results were accurate in 79.8% of the tests, while showing a divergence rate of 18.3%. In the cases with divergent results, one typical reason was overly complicated table structures that the LLM was not able to interpret. In some cases, a mismatch was caused by the corporate report deviating from required KPI definitions. For example, some institutions reported "total emissions" defined as Scope 1 and Scope 2, leaving Scope 3

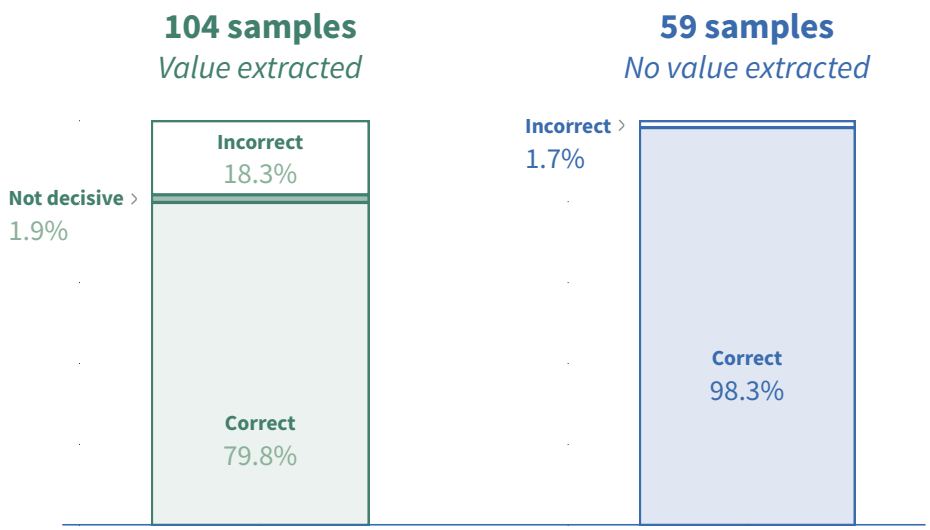
6. A match is defined as being within 2% after aligning the order of magnitude (correcting for diverging metric prefixes: eg kg/t or GWh/kWh).

emissions undisclosed. A third type of error was caused by values that were missing for a specific year. In such cases, Gaia incorrectly reported the value for the last available year. In most of these erroneous cases, the LLM response included hints to the mismatch, which makes it possible to eliminate the problems in a post-processing phase. In addition, the project has outlined possible LLM iteration techniques that can identify erroneous responses. Together, these methods promise to further reduce the rate of incorrect results. This falls within the potential scope of the next phase of Project Gaia.

In two cases (1.9%) the manual comparison was not decisive because it was not possible for the human tester to determine the correct value based on the corporate report, which illustrates the complexity of the task.

In the 59 cases in which Gaia was not able to extract a value (Graph 8, right-hand column) the accuracy was very high: in 98.3% of these cases, the human tester confirmed that the information is not present in the documents. This level of accuracy shows the effectiveness of Gaia's proper prompt engineering to mitigate the LLM's tendency to provide vague or false responses in cases of missing information. Some of the inconsistencies are the result of KPIs that are contained in infographics but not in text format in reports. Depending on the way in which infographics are inserted in a document, they may not currently be extracted by Gaia.

Manual cross-check of Gaia results - Graph 8



The graph presents the results of a manual cross-checking procedure that tested for potential type I and type II errors. It displays the proportions of results characterised as correct, incorrect and not decisive. The "correct" category represents cases in which extracted values align with those obtained from a manual cross-check. The "incorrect" category reflects a mismatch between the KPI definitions and values extracted. The "not decisive" category includes cases in which it was unclear from the manual review whether the information provided was accurate enough.
Source: Project Gaia.

Language independence

While most of the documents in the document repository are in English, a small number of Spanish and German language reports are also included. Tests have shown that Gaia is able to obtain correct values from the non-English documents, which underlines another important advantage of LLM-based text analysis: it can be made, to a large degree, language independent.

It is important to note that relying on LLMs does not, in itself, make the solution capable of handling multiple languages. Some other design choices must also be made with language independence in mind, in particular, the search for the relevant context inside the documents needs to be language independent. The semantic searches based on Open AI embeddings used in Gaia fulfil this criterion, but a simple keyword search would fail.

For a deeper analysis of language independence, two questions were submitted to Gaia's open query pipeline together with the relevant non-English reports as context: "Can you give me an overview of the climate transition plans?" and "What were the GHG emissions for scope 1, 2 and 3?"

All three scopes of emissions could be extracted correctly even though the Spanish translation "alcance" was used throughout the document, as illustrated in Graph 9. The same is true for the transition plans, which were also extracted correctly and translated into a reply in English.

Language independence of the LLM-based extraction approach - Graph 9

Emisiones totales de Gases de Efecto Invernadero (GEI) de la organización (2017-2019)

Categoría	2017	2018	2019	
	tCO_2e	tCO_2e	tCO_2e	%
Combustión en fuentes fijas	489	147	63	0,1
Combustión en fuentes móviles	755	817	770	1,1
Emisiones fugitivas	449	826	885	1,2
Subtotal alcance 1	1.693	1.790	1.718	2,4
Consumo eléctrico	15.382	15.910	14.643	20,4
Subtotal alcance 2	15.382	15.910	14.643	20,4
Transporte de clientes	41.861	39.450	35.284	49,3
Transporte de colaboradores	8.262	8.023	7.996	11,2
Servicio de centro de datos externo	3.425	4.000	4.304	6,0
Uso de banca electrónica por clientes	2.315	3.109	3.653	5,1
Viajes de negocios	1.356	1.662	1.714	2,4
Otros	3.507	2.419	2.330	3,2
Subtotal alcance 3	60.726	58.665	55.282	77,2
Emisiones totales de GEI	77.801	76.364	71.642	100,0

```
[
  {
    "company":"Banco [redacted]",
    "year":"2019",
    "KPI\/question":"What were the GHG emissions for scope 1, 2 and 3?",
    "short response":"In 2019, Banco [redacted]'s GHG emissions were 1,718 tCO2e for Scope 1, 14,643 tCO2e for Scope 2, and 55,282 tCO2e for Scope 3.",
    "background":"The GHG emissions for Banco [redacted] in 2019 were found in the 2019_Banco[redacted]_ESG.pdf document on page 129. The table titled 'Emisiones totales de Gases de Efecto Invernadero (GEI) de la organizacion (2017-2019)' provides the emissions for Scope 1, 2, and 3. The emissions are reported in metric tonnes of CO2 equivalent (tCO2e).",
    "certainty":"100.0",
    "GPT reference":"[{'document_name': '2019_Banco[redacted]_ESG.pdf', 'page_number': '129', 'quote': 'Emisiones totales de Gases de Efecto Invernadero (GEI) de la organizacion (2017-2019)'}]"
  }
]
```

Gaia is capable of extracting KPIs from non-English documents because the underlying LLM is language agnostic to most major languages.

Source: Project Gaia.

AI risks and challenges

Hallucinations

At one stage of experimentation, the manual cross-checking of results revealed a notable anomaly: a significant portion of extracted KPI values consistently showed the same figure, namely 1,500,000 metric tonnes of CO_2 equivalent (t CO_2eq).

Upon closer inspection of the sources provided by the LLM, two report names – "Corporate_Report_2020" and "Annual_Report_2020" – repeatedly appeared. Further analysis of the results exposed a surprising revelation: no reports with the aforementioned names actually existed. The LLM had generated fictional reports, complete with quotations such as, "In 2020, our gross direct GHG emissions (Scope 1) amounted to 1,500,000 metric tons of CO_2 equivalent (t CO_2eq). This graph was calculated based on our non-renewable fuel consumption and using consistent emission factors". All of this was from imaginary sources that convincingly mimicked actual reported sentences.

This peculiar behaviour was not a one-time occurrence. The LLM consistently followed the same pattern in multiple experiments, always returning a KPI value of 1,500,000 t CO_2eq and referencing the same fictional reports. This recurring pattern of hallucination was mitigated by a number of design choices, in particular (DC8) choosing the latest LLM version, GPT4, which adopts a more conservative approach instead of providing potentially incorrect values as compared with its predecessor; (DC6) prompt engineering, instructing the LLM only to refer to information given within the provided context and explicitly stating what to respond if no information is found; and (DC10) setting the temperature parameter to zero, which controls the creativity of LLM. With these design choices, hallucinations were significantly reduced and they do not seem to impact results in the final version of the PoC.

Overconfidence

LLM overconfidence occurs when the model asserts incorrect information from the data set, potentially due to incomplete data, without indicating the limitations of its result. This was illustrated, for example, by the open question pipeline, particularly in the case of short responses when posed with a question like, "Does company X have any partnerships with environmental organisations?". Relying solely on the information contained in the source pages of corporate reports, the LLM often concluded that the company had no partnerships with environmental organisations. This conclusion was drawn with a high level of certainty, reaching 100% in one instance and 70% in some others. However, these conclusions were based solely on excerpts provided from corporate reports and it cannot be asserted with a high degree of certainty that the company had absolutely no partnerships with environmental organisations. This underscores the risk of misinterpreting output from an LLM-based application. A human user forming conclusions based on LLM responses must always be cognisant that, while the response may be formulated as a generic statement, in reality it is based on a specific set of input data and its validity is limited to the context of that data.

5. Macro analysis use cases

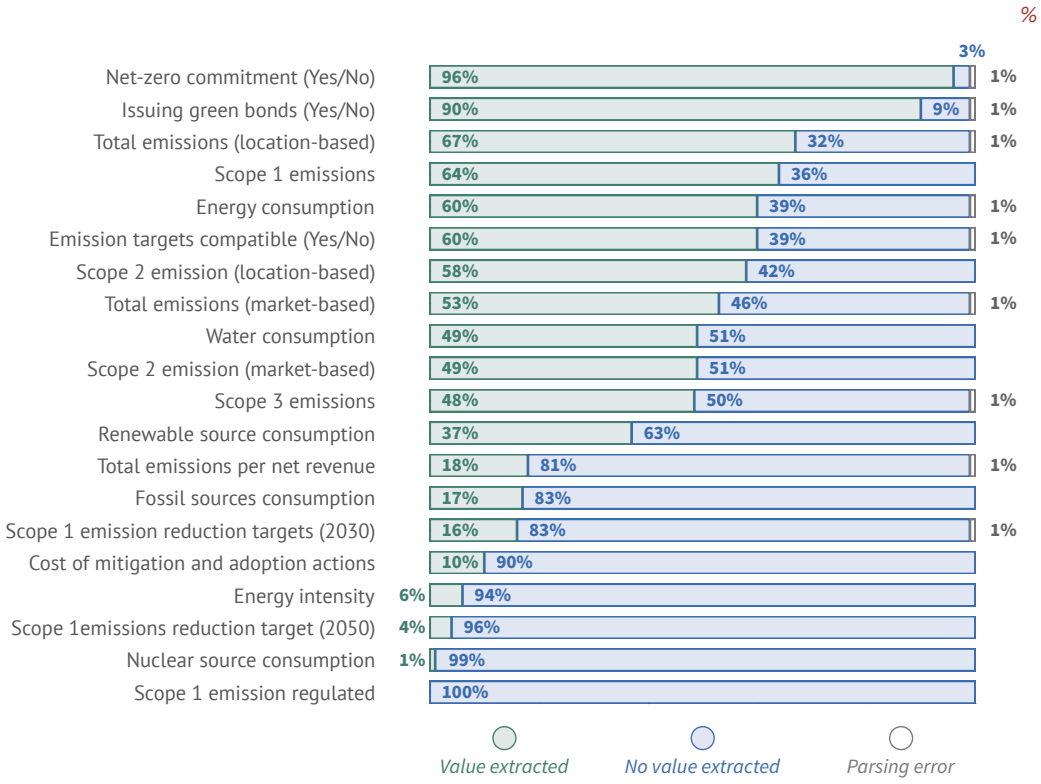
Macro analysis use cases

This section presents examples of green finance macro analysis use cases that are enabled by Gaia data. The examples build on the same test data set that was used for functional testing, as presented in the previous section: a total of 2,328 publicly available documents (ESG, Pillar 3, financial statements, annual reports and other relevant reports) of systemically important financial institutions, covering a period of five years from 2018 to 2022.

KPI coverage

Graph 10 illustrates the total rate of climate-related KPI coverage, for all 187 financial institutions across five years, in the test set. A notable finding from the graph is the prevalence of variables related to greenhouse gas (GHG) emissions. The most frequently reported metrics include total GHG emissions, along with the sub-component emissions for Scope 1, Scope 2 and Scope 3. Scope 1 emissions are direct emissions from owned or controlled sources, such as company facilities and vehicles. Scope 2 emissions refer to indirect emissions from the generation of purchased electricity, steam, heating and cooling consumed by the reporting financial institution. Further, Scope 3 emissions encompass all other indirect emissions arising within a company's value chain.

Results of the KPI extraction for common numeric and binary indicators - Graph 10



The graph depicts the percentage of extracted values for all available years of sampled companies. If no value is extracted, it is usually due to a lack of relevant information in the reports. However, the tool may extract additional supplementary comments.

Source: Project Gaia.

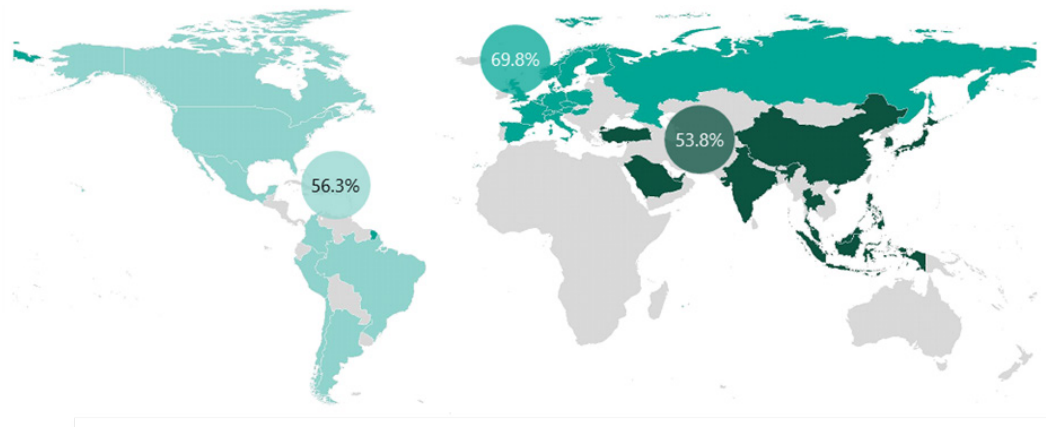
The results indicate that Scope 3 emissions are less commonly reported by institutions (48% of potential data points), probably due to the higher complexity involved in their assessment. In terms of reporting approaches, Scope 2 emissions are more commonly reported using a location-based method rather than a market-based one. The location-based approach considers the average emissions intensity of grids where energy consumption occurs, while the market-based approach accounts for the emissions from the specific types of energy that a company purchases.

The data also indicate that 60% of company-year combinations reported on total energy consumption and 49% reported on total water consumption, highlighting these as key areas of focus in environmental sustainability reporting. However, a disparity is observed in the reporting of long-term environmental commitments versus specific targets. In 2022, while 82% of the sampled companies claimed to have a net zero commitment, only 30% had set a concrete target for absolute GHG reduction by 2030 and a mere 3% had done so for 2050. This suggests a potential gap between stated ambitions and detailed planning or goal setting. Moreover, intensity measures, which provide insights into emissions relative to a company's economic output, are less frequently reported. Only 18% of potential data points reported total GHG emissions per net revenue and just 6% reported on energy intensity. Other KPIs such as the share of Scope 1 emissions covered by emission trading schemes (Scope 1 regulated emissions) were found even less frequently, which is hardly surprising as they are less relevant for financial institutions than for other industries.

Regional variation of KPI adoption

By making climate-related KPIs available at scale, Gaia unlocks broad and granular macro analysis that was previously not possible. One example is regional comparisons. Specifically, Graph 11 illustrates how financial institutions in different regions reported their green bond issuance in 2018. The results indicate some notable variations across regions: 70% of European banks were reporting green bond issuance, while lower proportions were found in the other regions covered by the data set (the Americas and Asia-Pacific). Similar analysis can be performed for other KPIs, or, for example, at the country level.

Reported green bond issuance by region in 2018 - Graph 11



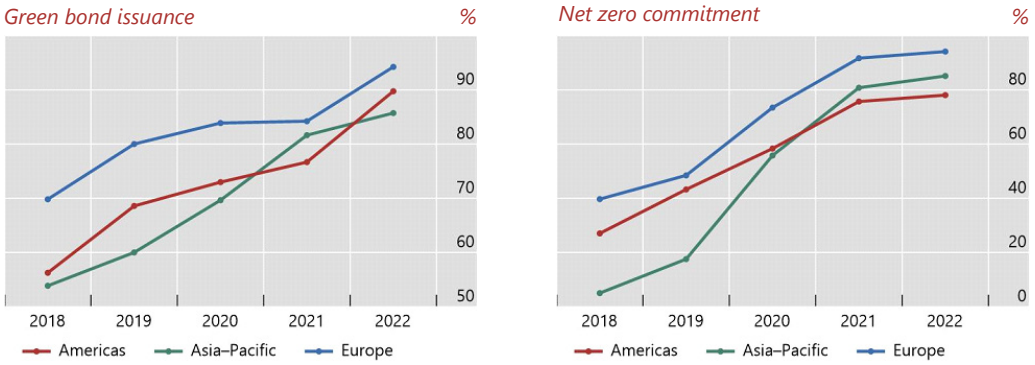
The graph depicts institution domiciles on a regional basis. Values represent the ratio of institutions issuing green bonds in 2018 based on disclosed reports relative to each region's total number of institutions in the data set. The use of this map does not constitute, and should not be construed as constituting, an expression of a position by the BIS regarding the legal status or sovereignty of any territory or its authorities, the delimitation of international frontiers and boundaries, and/or the name and designation of any territory, city or area.
Source: Project Gaia.

Evolution of KPI adoption

Gaia data show a steady increase in climate-related KPI adoption reported in all regions, throughout the time range available in the data set. At the beginning of the period, fewer than 40% of financial institutions had declared a voluntary commitment to net zero policies, with the largest share in Europe. Over time, the share of banks adopting the policy goals has increased across regions. The adoption rate among Asia-Pacific banks, in particular, increased significantly in 2020 and is now at a similar level to that seen in the Americas, based on the Gaia data set (Graph 12).

Conversely, the reported adoption of green bond issuance was already much higher in 2018, across regions. One explanation may be that the high demand for green bonds motivated additional market participants to step into the market, effectively decreasing their own costs of funding. Since then, there has been a robust increase in the share of participating banks. Gaia data show that Asia-Pacific banks have increased their emphasis on green bond issuance the most, with the share of participating banks reaching levels similar to that seen in Europe by the end of the time period investigated. Banks in the Americas, on the other hand, experienced more stable growth. By 2022, participation seems to have broadly equalised across regions.

Adoption trends of selected KPIs by region - Graph 12



Graphs depict the increase over time of institutions in various regions reporting voluntary net zero targets and issuance of green bonds, in relation to the total number of institutions present in these regions for which data are available.

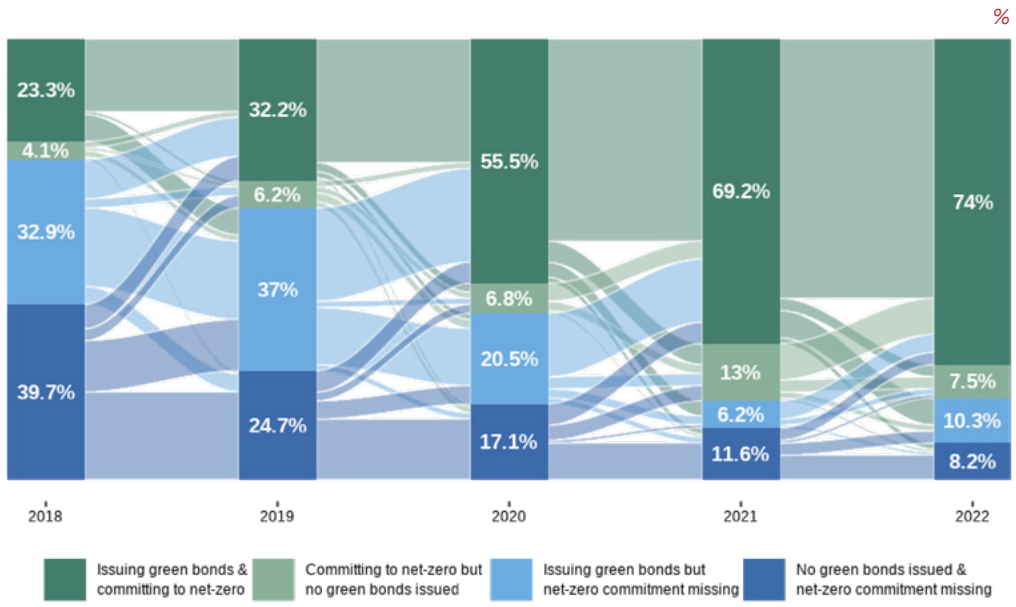
The test data set is incomplete for 2022. At the time of generating the data set, only 36% of annual or sustainability reports were available for 2022 compared with 2021. Despite covering 89% of banks in 2022, some reports dedicated to specific topics may be missing, resulting in lower KPI coverage.
Source: Project Gaia.

Journey of individual institutions

Behind the global trends illustrated above lie decisions by financial institutions that have gradually increased their level of commitment to climate transition plans. Gaia data can also be used to analyse the alignment of institutions.

Graph 13 illustrates the journey of financial institutions using green bond issuance and voluntary net zero commitment as examples. The graph shows that in 2018 more than half of banks globally (56.2%) issued green bonds while only 27.4% publicly made a commitment to net zero. Only 23.3% of banks reported both in their reports. Over time, bond issuance became more popular and banks gradually increased communications about their net zero commitments. In 2022, 84.3% of banks issued green bonds, while fewer than 20% did not commit to net zero. There was also a significantly lower proportion of banks issuing green bonds without communicating their commitment publicly (10.3%). At the same time, there was still a portion of banks (8.2%) that did not disclose such information publicly.

Voluntary commitment to net zero targets and green bond issuance over time - Graph 13



Evolution of institutions in their commitment to net zero targets and the issuance of green bonds over time. Note that the data presented only include the institutions for which KPIs are available across all periods. Different colour groups symbolise various combinations of engagement with the two KPIs, with the numbers indicating the relative size of each group in successive years. The flows represent changing attitudes towards the two KPIs found in their reports.

Source: Project Gaia.

6. Learnings and next steps

Learnings and next steps

A game changer for green finance

Project Gaia has proven the feasibility of climate-related financial risk analysis using AI and LLM. By tapping into public information readily available today, Gaia provides an efficient and reliable source of climate risk-related data.

With the traditional manual approach of collecting KPIs for analysis, each additional KPI and each additional institution requires dedicated manual work. The analyst either needs to search for the information in public corporate reports or contact the institution for information. With Gaia, adding new KPIs or new institutions is quick and easy. This makes it possible to extract and analyse a multitude of KPIs from a large number of institutions, opening up the possibility of climate risk analysis at a scale that was previously unimaginable.

Furthermore, Gaia offers harmonised metrics despite the heterogeneity of naming and definitions across different jurisdictions and companies. Since Gaia relies on each KPI's definition rather than its name when searching corporate reports, it does not require harmonisation of guidelines or regulatory requirements. This is crucial in instances where slightly different wording and definitions prevail for similar concepts (eg Scope 3 emissions are sometimes referenced as financed emissions or indirect emissions). Gaia promises to overcome potential differences in official disclosure frameworks, and offers much needed transparency and comparability of climate-related information.

Applicability beyond climate data analysis

The Gaia PoC is breaking new ground by embedding LLMs into an information technology application for the financial industry and its significance goes beyond climate data analysis. The KPIs extracted by Gaia are not hard coded into the platform. Instead, based on the name of the KPI, Gaia relies on an LLM to create a short definition of the KPI using some reference text, such as ESG standards and regulatory documents. This definition is then used by a further round of LLM prompts to search for answers in corporate reports.

This flexible approach ensures that the Gaia platform can be easily configured to extract new types of KPIs. Adding a new KPI consists of providing Gaia with the name of the KPI plus adding relevant reference documents that contain the definition of the KPI.

As a result of this approach, the Gaia platform is applicable to a much broader context than climate-related financial risk analysis. In fact, Gaia offers a generic solution to extract desired KPIs contained in a predefined set of PDF reports.

Within the financial sector, AI-based KPI extraction from large bodies of textual documents can be a game changer, for example, in regulatory and supervisory use cases. Prospectuses for financial instruments is another area in which information is dispersed in the form of unstructured texts that can be leveraged. By proving the feasibility of KPI extraction using LLM, Project Gaia is paving the way for a broad range of AI-enabled use cases within central banking, the financial sector and beyond.

Creating value in a fast-changing field

As is the case in other projects close to the edge of technological development, some of the findings may be used in a new context and some design choices may become obsolete after the project. But the learnings from hands-on usage and integration into real-life processes will help shape expectations and pave the way forward in the field.

Gaia shows that using LLM for data extraction is an emerging technology with great potential and the project is an early example of demonstrating its feasibility and exploring its capabilities at scale.

Gaia solves some of the existing limitations in current LLMs by adding specific context from a preselected set of data, and fine-tuning the prompts towards domain-specific and reproducible results. Furthermore, Gaia adds scalability to the data extraction process with prompts engineered for a low error rate and automated concurrent processes.

Gaia caters to domain-specific users, rendering their workflow as efficient as possible, by aggregating data into simple overviews. The source data are extracted in bulk and results are aggregated into overview tables. The resulting aggregated data sets are stored centrally and can be evaluated and checked for quality.

Next steps

The current phase of Project Gaia has proven the feasibility of extracting data from corporate climate-related disclosures using AI and LLMs. The PoC was created with enterprise-grade components, in an architecture designed for high scalability, to meet the needs of a future open tool. One possible continuation is to make the solution publicly available as an open web-based service for climate-related financial risk analysts and support the growing demand for climate-related data.

Another natural next step is to expand into use cases beyond green finance. Project Gaia is focused on climate-related financial KPIs, but the platform was designed in a flexible way, making it capable of extracting other types of KPIs as well. By offering a generic solution for extracting desired KPIs contained in a predefined set of PDF reports, the Gaia approach naturally lends itself to a wide variety of use cases in central banking and in the financial sector.

LLMs as a technology are rapidly evolving, increasing performance and expanding with new capabilities, such as internet search or image recognition. For example, online search greatly extends an LLM's knowledge base, enabling responses to current events and potentially covering information that was not part of the model's training. Future phases of Gaia will need to investigate and adopt these new developments to continue to harness the full power of cutting-edge AI.

Governance

Deployment of AI-enabled tools involves several non-technical challenges, which will need to be addressed in a potential future practical application of the Gaia technology. There are a growing number of initiatives addressing regulatory and policy issues to mitigate AI-related risks. Ethical and legal considerations must be taken into account and proper safeguards put in place to ensure privacy, security and accountability. Users of AI systems must be able to understand them and be comfortable using them. The environmental impact of large AI models needs to be monitored and minimised. To ensure that these and other considerations are properly accounted for, a future real-life application of the Gaia technology will need to be surrounded by proper governance structures and processes.

7. Conclusions

Conclusions

Transparency is a crucial requirement for the financial sector to handle climate-related risks. Although financial institutions increasingly disclose climate-related information in their corporate reports, these data are not easily accessible today due to the heterogeneity of regulations, frameworks and standards, among other challenges.

Project Gaia uses AI and LLMs to extract climate-related KPIs from corporate reports, which opens up the possibility of analysing climate-related financial risks at a scale that was previously unimaginable. The traditional manual approach of collecting KPIs for analysis requires dedicated effort to add each additional KPI and each additional institution. However, once the platform is available, adding new KPIs or new institutions comes at near-zero costs and with very little delay when using Gaia.

This report has presented the project's key findings including benchmarking results indicating the reliability of the approach, concrete examples of insights gained using the tool, and some design choices that ensure efficient and reliable KPI extraction from large volumes of text using LLMs.

Project Gaia breaks new ground by integrating LLM into an application and leveraging it for data extraction. As is the case in other projects close to the edge of technological development, some of the findings may be used in a new context and some design choices may become obsolete after the project. But the learnings from hands-on usage and integration into real-life processes will help shape expectations and pave the way forward in the field. The use of LLM for data extraction is an emerging technology with great potential and Project Gaia is an early example of demonstrating its feasibility and exploring its capabilities at scale.

The significance of Gaia goes beyond climate-related KPIs or, indeed, the financial industry. The Gaia PoC demonstrates the power of creating AI-enabled intelligent tools to automate existing workflows. This approach has the potential to change the way we work in the financial industry and beyond. Project Gaia is one of the first comprehensive studies investigating how this can be done in practice.

Project participants and acknowledgements

BIS Innovation Hub Eursosystem Center

Timothy Aerts, Adviser
Raphael Auer, Eurosystem Centre Head
Rudolf Biczok, Developer
Tiphanie Chabrol, Adviser
David Köpfer, Adviser and Technical Lead
María Molero, Adviser and Project Lead
Ivana Rajcic, Adviser
Maximilian Schrader, Data Analyst
Josef Švéda, Adviser
Andras Valko, Adviser
Violeta Vuletic, Data Analyst

Bank of Spain

Iván Balsategui, Manager
Teresa Caminero, Economist
Ana Fernández, Manager
Sergio Gorjón, Senior Manager
José Manuel Marqués, Director
Ángel Iván Moreno, Senior Data Scientist

Deutsche Bundesbank

Julia Biesen, Senior Manager
Hendrik Christian Doll, Economist
Manuel Fangmann, Economist
Lisa Reichenbach, Technology Expert

European Central Bank

Christoph Schaper, Head of Division
Sjoerd Van der Vaart, Strategy & Innovation Expert

Acknowledgements

The authors are grateful to Cecilia Skingsley, Francesca Hopwood and Teresa Lin for reviewing this report and providing extremely valuable feedback; to Simona Lambrinoc for the helpful discussions and legal insights; to Darko Micic and Bernard van den Boom for their crucial inputs on the technology side; to the Green Finance AI WG members and the members of the Expert Network on Data of the NGFS: Nathalie Rouille, Léa Grisey, Elena Triebskorn, Ong Li Ming, Maurice Fehr, Leandro D'Aurizio and Sunjin Park.

References

Battiston, S, Y Dafermos and, I Monasterolo (2021):
"Climate risks and financial stability" *Journal of Financial Stability,* vol 54, no 100867.

Berg, F, J Kölbel and R Rigobon (2022):
"Aggregate confusion: the divergence of ESG ratings",
Review of Finance, vol 26, no 6, pp 1315–44.

Ferreira, C, D Rozumek, R Singh and F, Suntheim (2021):
"Strengthening the climate information architecture",
IMF Staff Climate Note 2021/003.

Moreno, Á I and T Caminero (2020):
"Application of text mining to the analysis of climate-related disclosures",
International Review of Financial Analysis, vol 83.

Moreno, Á I and T Caminero (2022):
"Analysis of ESG disclosures in Pillar 3 reports: a text mining approach",
Bank of Spain Occasional Papers, no 2204.

Moreno, Á I and T Caminero (2023):
"Assessing the data challenges of climate-related disclosures in European banks.
A text mining study", *Bank of Spain Working Papers,* no 2326.

Network for Greening the Financial System (NGFS) (2022):
"*Final report on bridging data gaps*", July.
NGFS Technical Document.

Task Force on Climate-related Financial Disclosures (TCFD) (2023):
"*2023 Status Report*", October.

Bank for International Settlement (BIS)

978-92-9259-716-0 (Online)