Harnessing the potential of AI for LCI data compilation from company reporting

Charaf Bejjit[1*], Diae Hennioui[1], Frédéric Lai[1], Andreas Kroiss[2], Aina Mas-Fons[1], Basile Guth[1], Stéphanie Muller[1], Daniel Monfort Climent[1], Stephan Lutter[2], Victor Maus[2], Antoine Beylot[1]

Affiliations

[1]BRGM - French Geological Survey, F-45060 Orléans, France
[2]WU Vienna University of Economics and Business, Institute for Ecological Economics, 1020 Vienna, Austria

SUPPLEMENTARY INFORMATION (SI 1)

1. Additional details on the screening module

1.1 Overview

The screening module is designed to evaluate, prior to any large-scale data extraction, whether corporate and technical reports contain **quantitative information suitable for life cycle inventory (LCI) construction**.

Processing industrial reports using multimodal large language models (LLMs) is computationally expensive and time-consuming. At the same time, a significant proportion of reports contain only:

- aggregated company-level indicators,
- narrative descriptions without numerical values,
- or financial metrics not suitable for environmental modeling.

The screening module therefore acts as a filtering and prioritization layer, ensuring that only reports containing relevant, quantitative, and sufficiently granular data are forwarded to the extraction pipeline.

The screening specifically targets reports containing:

- production metrics (e.g., ore mined, concentrate produced, refined metal),
- environmental metrics (e.g., energy consumption, water use, emissions),
- expressed with physical units,
- and associated with industrial sites or specific products,
- within copper (Cu) and nickel (Ni) value chains.

1.2 Structural reconstruction: implementation

Industrial reports exhibit highly heterogeneous layouts, including:

- multi-column text,
- embedded tables,
- mixed graphical and textual elements,
- and non-standard formatting.

To ensure reliable extraction of quantitative information, the screening module implements a systematic structural reconstruction strategy combining tabular and textual extraction. Unlike a conditional fallback approach, the reconstruction procedure is applied systematically to all documents. This conservative design choice prioritizes data recall over computational efficiency, reducing the risk of missing structured numerical data.

For each page of a PDF document, two complementary extraction processes are performed:

- **Table extraction:**

Tables are extracted using the pdfplumber library:

```
p.extract_tables()
```

Table detection relies on the geometric structure of the PDF, including:

  - alignment of text elements,
  - presence of horizontal and vertical vector lines,
  - consistent spacing patterns between rows and columns.

These structural cues are used to reconstruct tabular layouts and identify cell boundaries.

Detected tables are then converted into a structured textual representation (row–column format using tab separators). This preserves:

  - alignment between values and headers,
  - relationships between sites and associated metrics,
  - the integrity of tabular numerical data.

This approach enables the recovery of structured data even in complex multi-page or non-standard table layouts.

- **Text extraction:**

Continuous textual content is extracted using:

```
p.extract_text(x_tolerance=2)
```

This step captures:

- narrative descriptions of operations,
- contextual explanations of reported metrics,
- qualitative information supporting quantitative disclosures

- **Unified representation:**

Extracted tables and text are combined into a single representation:

```
--- PAGE X ---
[STRUCTURED TABLES FOUND]
<table content>
<text content>
```

This unified structure allows the model to process both structured and unstructured information simultaneously.

- **Limitations:**

Reports consisting exclusively of scanned images without an extractable text layer cannot be processed automatically and are flagged for manual evaluation.

1.3 LLM-based evaluation:

The structured JSON output constitutes the central artifact of the screening module. It encodes the information required to assess the usability of each report for LCI construction and serves as the interface between the screening and extraction stages.

The schema captures multiple complementary dimensions of the report content, including:

• industrial context identification, through detection and categorization of sites and assets,

• product and material coverage, including copper, nickel, and associated by-products,

• quantitative data availability, distinguishing between site-level, product-level, and financial metrics,

• presence of physical units, ensuring that numerical values are interpretable and usable,

• data structure characteristics, such as the presence of tables and separation by site or metal,

• traceability elements, including verbatim numerical excerpts and associated page references,

• industrial process information, describing key inputs, outputs, and residues when available.

This structured representation allows heterogeneous reports to be transformed into a consistent, machine-readable format while preserving key contextual information required for validation.

1.4 Screening decision logic

Each report is assigned a preliminary screening decision. The classification follows a rule-based logic prioritizing production and environmental data over financial indicators.

The decision process evaluates the following conditions:

- presence of quantitative data associated with physical units,
- relevance to copper or nickel value chains,
- existence of site-level or product-level information,
- degree of structural organization (e.g., tabular data separated by site or metal).

Based on these criteria, reports are classified into three categories:

- **GO:** reports containing explicit site- or product-level production and/or environmental data with sufficient granularity for automated extraction.
- **MAYBE:** reports containing partial information (e.g., limited granularity, text-based site data, or predominantly financial indicators) requiring manual verification prior to extraction.
- **NO-GO:** reports containing only aggregated company-level indicators or narrative descriptions without usable quantitative data for LCI construction.

1.5 Output files and integration into the pipeline

For each processed report, the screening module generates three complementary outputs:

- **Structured JSON file**

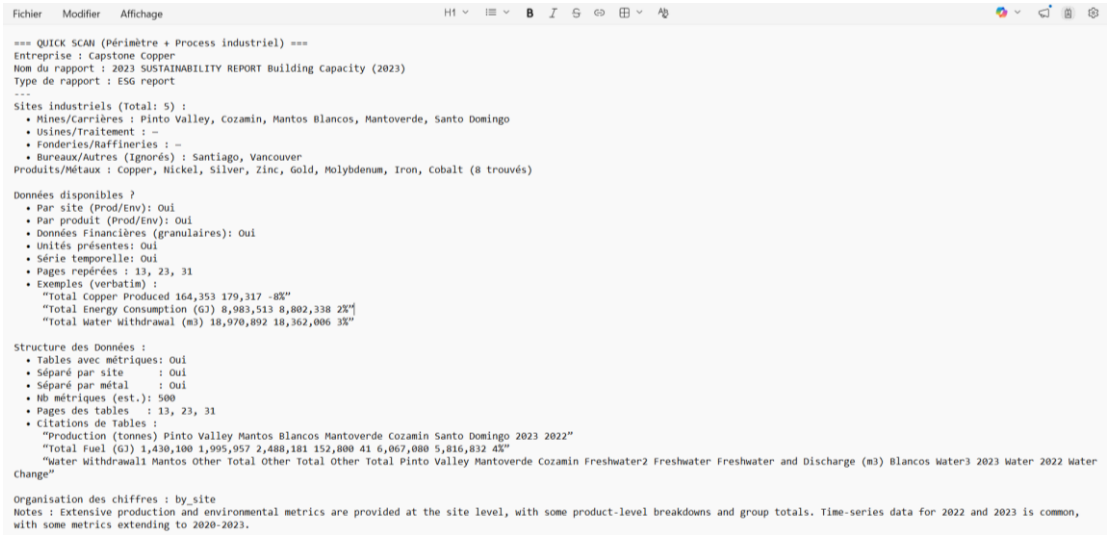
A machine-readable representation containing all normalized screening variables and the preliminary decision. This file ensures reproducibility and serves as the primary input for downstream extraction modules.

```
{"company_or_group":"Capstone Copper","report_name":"2023 SUSTAINABILITY REPORT Building
Capacity","year":2023,"report_type":"ESG report","sites_categorized":{"Mine_Quarry":["Pinto
Valley","Cozamin","Mantos Blancos","Mantoverde","Santo Domingo"],"Processing_Plant":
[],"Smelter_Refinery":[],"Office_Other":["Santiago","Vancouver"]},"products":["Copper
concentrate","Copper cathode","Silver","Zinc","Gold","Molybdenum","Iron
ore","Cobalt"],"metals_or_materials_covered":
["Copper","Nickel","Silver","Zinc","Gold","Molybdenum","Iron","Cobalt"],"site_product_mapping":
[],"data_existence":
{"has_site_level_data":true,"has_product_level_data":true,"has_financial_metrics":true,"has_units_prese
nt":true,"has_time_series":true,"examples_verbatim":["Total Copper Produced 164,353 179,317 -8%","Total
Energy Consumption (GJ) 8,983,513 8,802,338 2%","Total Water Withdrawal (m3) 18,970,892 18,362,006
3%"],"pages":[13,23,31]},"data_organization":{"primary_axes":"by_site","notes":"Extensive production
and environmental metrics are provided at the site level, with some product-level breakdowns and group
totals. Time-series data for 2022 and 2023 is common, with some metrics extending to 2020-
2023."},"data_structure":
{"has_tables_with_metrics":true,"metrics_separated_by_site":true,"metrics_separated_by_metal":true,"tot
al_metric_count_estimate":500,"table_verbatims":["Production (tonnes) Pinto Valley Mantos Blancos
Mantoverde Cozamin Santo Domingo 2023 2022","Total Fuel (GJ) 1,430,100 1,995,957 2,488,181 152,800 41
6,067,080 5,816,832 4%","Water Withdrawal1 Mantos Other Total Other Total Other Total Pinto Valley
Mantoverde Cozamin Freshwater2 Freshwater Freshwater and Discharge (m3) Blancos Water3 2023 Water 2022
Water Change"],"table_pages":["13","23","31"]},"industrial_process":{"flow_summary":"Capstone Copper
operates open-pit and underground mines. Processing involves milling, flotation recovery, solvent
extraction and electrowinning (SX/EW), and heap leaching. Products include copper concentrate and
copper cathode. Tailings management includes dry-stack tailings. Desalinated water is used in some
operations.","key_inputs":["Ore","Sulphuric acid","Electricity","Fuel (diesel, gasoline, propane,
liquefied petroleum gas)","Reagents","Water (freshwater, desalinated seawater, third-party water,
reclaimed process water)"],"key_outputs":["Copper concentrate","Copper
cathode","Gold","Silver","Zinc","Molybdenum","Iron ore","Cobalt"],"residues_waste":["Tailings","Waste
rock","Sludge","Hazardous waste","Non-hazardous waste","Brine"],"process_verbatims":["Pinto Valley is a
copper-molybdenum open-pit mine and one of only two operating mines located in the historic Globe-Miami
mining district of Arizona, one of the oldest and most productive mining districts in the US. Pinto
Valley is currently the second-largest private employer in the district. Pinto Valley has a current
life of mine plan that extends through 2039 but is being assessed for possible extension. Type of Mine
and Open pit with milling and flotation recovery; solvent Production Process extraction and
electrowinning (SX/EW) plant Product(s) Copper concentrate and copper cathode","Cozamin is a copper-
silver underground mine with a surface milling facility and is located near the city of Zacatecas in
the mineral-rich state of Zacatecas, Mexico. The mine currently has a life of mine plan that extends
through 2030. However, in an effort to extend its mine life, brownfield exploration continues. Type of
Mine and Underground mine with surface milling facility Production Process Product(s) Copper
concentrate","Mantoverde is an open-pit, oxide heap leach copper mine in the Atacama region of Chile. A
significant expansion – the Mantoverde Development Project – is underway to support mining and mineral-
processing of sulphide ore. Construction of the copper-gold project was completed in 2023, and
commissioning commenced in December. Proximity to our project at Santo Domingo presents possibilities
for district integration. Type of Mine and Open pit processing oxide ore; development project to
Production Process process sulphide ore Product(s) Copper cathode; copper concentrate with significant
gold by-product ramping up in 2024"],"process_pages":
[7,8]},"decision_tag":"go","decision_reason":"Complex report (multiple sites/metals), but data is
clearly separated by Site AND Metal (tables/structure). (Financial metrics also present)."}
```


Figure 1: Example of structured JSON output generated by the screening module

- **Diagnostic text file (TXT)**

A human-readable summary presenting identified sites, products, and key quantitative evidence, along with verbatim excerpts. This file supports rapid expert validation without requiring full document review.

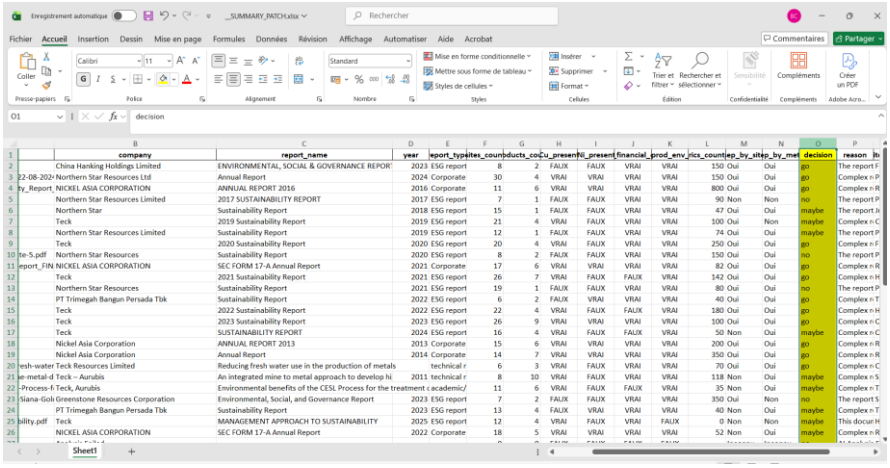


```
=== QUICK SCAN (Périmètre + Process industriel) ===
Entreprise : Capstone Copper
Nom du rapport : 2023 SUSTAINABILITY REPORT Building Capacity (2023)
Type de rapport : ESG report
---
Sites industriels (Total: 5) :
  • Mines/Carrières : Pinto Valley, Cozamin, Mantos Blancos, Mantoverde, Santo Domingo
  • Usines/Traitement : –
  • Fonderies/Raffineries : –
  • Bureaux/Autres (Ignorés) : Santiago, Vancouver
Produits/Métaux : Copper, Nickel, Silver, Zinc, Gold, Molybdenum, Iron, Cobalt (8 trouvés)

Données disponibles ?
  • Par site (Prod/Env): Oui
  • Par produit (Prod/Env): Oui
  • Données Financières (granulaires): Oui
  • Unités présentes: Oui
  • Série temporelle: Oui
  • Pages repérées : 13, 23, 31
  • Exemples (verbatim) :
     "Total Copper Produced 164,353 179,317 -8%"
     "Total Energy Consumption (GJ) 8,983,513 8,802,338 2%"
     "Total Water Withdrawal (m3) 18,970,892 18,362,006 3%"

Structure des Données :
  • Tables avec métriques: Oui
  • Séparé par site      : Oui
  • Séparé par métal     : Oui
  • Nb métriques (est.): 500
  • Pages des tables   : 13, 23, 31
  • Citations de Tables :
     "Production (tonnes) Pinto Valley Mantos Blancos Mantoverde Cozamin Santo Domingo 2023 2022"
     "Total Fuel (GJ) 1,430,100 1,995,957 2,488,181 152,800 41 6,067,080 5,816,832 4%"
     "Water Withdrawal1 Mantos Other Total Other Total Other Total Pinto Valley Mantoverde Cozamin Freshwater2 Freshwater Freshwater and Discharge (m3) Blancos Water3 2023 Water 2022 Water
Change"

Organisation des chiffres : by_site
Notes : Extensive production and environmental metrics are provided at the site level, with some product-level breakdowns and group totals. Time-series data for 2022 and 2023 is common,
with some metrics extending to 2020-2023.
```

Figure 2: Example of diagnostic text file output generated by the screening module

- **Consolidated Excel file**

An aggregated overview of all processed reports, including metadata, detected metrics, structural indicators, and screening decisions. This file enables batch-level analysis, comparison, and workflow monitoring.



	company	report_name	year	report_type	ites_coun	ducts_co	u_present	il_present	financial_	prod_env_	rics_coun	ep_by_sit	ep_by_met	decision	reason
2	China Hanking Holdings Limited	ENVIRONMENTAL, SOCIAL & GOVERNANCE REPORT	2023	ESG report	8	2	FAUX	FAUX	VRAI	VRAI	150	Oui	Oui	go	The report F
3	22-08-202 Northern Star Resources Ltd	Annual Report	2024	Corporate	30	4	VRAI	VRAI	VRAI	VRAI	150	Oui	Oui	go	Complex n P
4	ty_Report_ NICKEL ASIA CORPORATION	ANNUAL REPORT 2016	2016	Corporate	11	6	VRAI	VRAI	VRAI	VRAI	800	Oui	Oui	go	Complex n R
5	Northern Star Resources Limited	2017 SUSTAINABILITY REPORT	2017	ESG report	7	1	FAUX	FAUX	VRAI	VRAI	90	Non	Non	no	The report P
6	Northern Star	Sustainability Report	2018	ESG report	15	1	FAUX	FAUX	VRAI	VRAI	47	Oui	Oui	maybe	The report N
7	Teck	2019 Sustainability Report	2019	ESG report	21	4	VRAI	FAUX	VRAI	VRAI	100	Oui	Non	maybe	Complex n C
8	Northern Star Resources Limited	Sustainability Report	2019	ESG report	12	1	FAUX	FAUX	VRAI	VRAI	74	Oui	Oui	maybe	The report P
9	Teck	2020 Sustainability Report	2020	ESG report	20	4	VRAI	FAUX	VRAI	VRAI	250	Oui	Oui	go	Complex n F
10	te-5.pdf Northern Star Resources	Sustainability Report	2020	ESG report	8	2	FAUX	FAUX	VRAI	VRAI	150	Oui	Oui	no	The report P
11	eport_FIN NICKEL ASIA CORPORATION	SEC FORM 17-A Annual Report	2021	Corporate	17	6	VRAI	VRAI	VRAI	VRAI	82	Oui	Oui	go	Complex n R
12	Teck	2021 Sustainability Report	2021	ESG report	26	7	VRAI	FAUX	FAUX	VRAI	142	Oui	Oui	go	Complex n H
13	Northern Star Resources	Sustainability Report	2021	ESG report	19	1	FAUX	FAUX	VRAI	VRAI	80	Oui	Oui	no	The report P
14	PT Trimegah Bangun Persada Tbk	Sustainability Report	2022	ESG report	6	2	FAUX	VRAI	VRAI	VRAI	40	Oui	Oui	go	Complex n T
15	Teck	2022 Sustainability Report	2022	ESG report	22	4	VRAI	FAUX	FAUX	VRAI	180	Oui	Oui	go	Complex n H
16	Teck	2023 Sustainability Report	2023	ESG report	26	9	VRAI	VRAI	VRAI	VRAI	100	Oui	Oui	go	Complex n C
17	Teck	SUSTAINABILITY REPORT	2024	ESG report	16	4	VRAI	FAUX	FAUX	VRAI	50	Non	Oui	maybe	Complex n C
18	Nickel Asia Corporation	ANNUAL REPORT 2013	2013	Corporate	15	6	VRAI	VRAI	VRAI	VRAI	200	Oui	Oui	go	Complex n R
19	Nickel Asia Corporation	Annual Report	2014	Corporate	14	7	VRAI	VRAI	VRAI	VRAI	350	Oui	Oui	go	Complex n R
20	resh-water Teck Resources Limited	Reducing fresh water use in the production of metals		technical r	6	3	VRAI	FAUX	VRAI	VRAI	70	Oui	Oui	go	Complex n C
21	e-metal-d Teck – Aurubis	An integrated mine to metal approach to develop hi	2011	technical r	8	10	VRAI	FAUX	VRAI	VRAI	118	Non	Oui	maybe	Complex n S
22	-Process-f Teck, Aurubis	Environmental benefits of the CESL Process for the treatment		academic/	11	6	VRAI	FAUX	FAUX	VRAI	35	Non	Oui	maybe	Complex n T
23	Siana-Gok Greenstone Resources Corporation	Environmental, Social, and Governance Report	2023	ESG report	7	2	FAUX	FAUX	VRAI	VRAI	350	Oui	Non	no	The report S
24	PT Trimegah Bangun Persada Tbk	Sustainability Report	2023	ESG report	13	4	FAUX	VRAI	VRAI	VRAI	40	Non	Oui	maybe	Complex n T
25	bility.pdf Teck	MANAGEMENT APPROACH TO SUSTAINABILITY	2025	ESG report	12	4	VRAI	FAUX	VRAI	FAUX	0	Non	Non	maybe	This docur H
26	NICKEL ASIA CORPORATION	SEC FORM 17-A Annual Report	2022	Corporate	18	5	VRAI	VRAI	VRAI	VRAI	52	Non	Oui	maybe	Complex n R

Figure 3: Example of the summary xlsx file output generated by the screening module

The end-to-end screening workflow combines table and text extraction with schema-constrained LLM evaluation to generate structured outputs suitable for downstream processing, as shown in Figure 2.

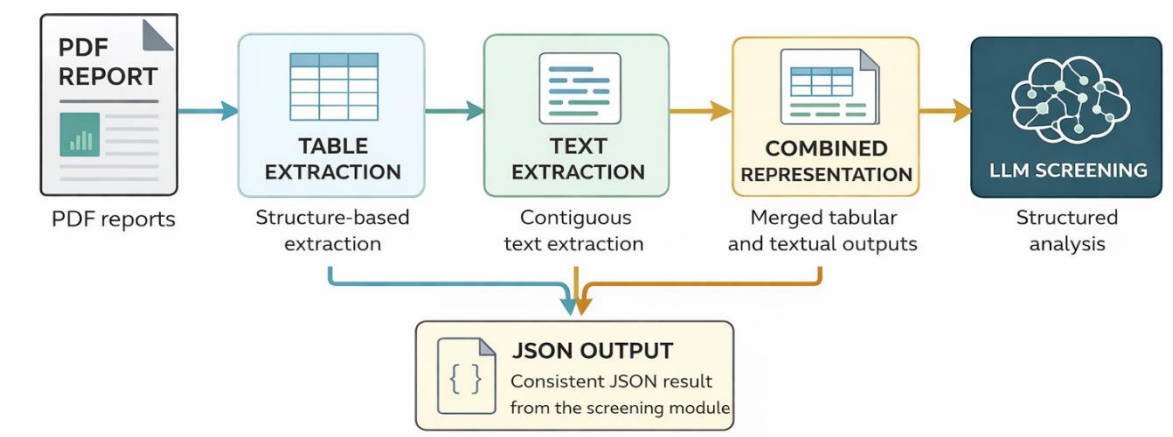


Figure 4: Data processing workflow for report screening

1.6 Human verification

The screening workflow concludes with a mandatory human validation step. Reports classified as GO or MAYBE are reviewed by analysts, who assign a final decision (GO or NO-GO) and document the main justification when applicable.

Reports initially classified as NO-GO are generally retained unless specific evidence justifies reassessment. Each report is evaluated independently, and cross-year consolidation is handled in a subsequent stage of the workflow.

2. Additional details on the extraction module

2.1 Overview

The extraction module is applied to reports classified as GO during the screening stage. Its purpose is to retrieve quantitative information from sustainability and ESG reports in a structured format suitable for subsequent data processing and LCI-oriented analysis.

In line with the main text, the extraction module is designed to preserve not only numerical values, but also the contextual elements required for interpretation, including:

- the metric category,
- the physical unit,
- the reporting year when available,
- the local textual context,
- the site or asset name when explicitly identifiable,
- and the source modality (text, table, or graph).

The module targets nine predefined categories of information:

- **Production:** production volumes, ore mined or processed, concentrates, refined products, contained metal, throughput, and associated physical quantities.

• **Economics:** segment-level financial indicators explicitly tied to a commodity or product (e.g. realized price, C1 cost, AISC, unit costs by product).

• **Energy:** electricity, fuels, natural gas, steam, heat, renewable energy, and energy intensities.

• **Water:** water withdrawal, consumption, discharge, and water quality indicators.

• **Geology:** reserves, resources, grades, and classification of mineral inventories.

• **Waste:** tailings, effluents, waste streams, and related quantities.

• **Emissions:** greenhouse gas emissions and air pollutants (e.g. NOx, SOx, particulate matter).

• **Land:** disturbed, occupied, rehabilitated, or conserved land areas.

• **Other:** industrial or process inputs reported with a physical unit, such as reagents, chemicals, or other material inputs.

By contrast, the module explicitly excludes categories that are not relevant for this study, including workforce, health and safety indicators, training data, and generic corporate financial reporting not tied to a specific product.

2.2 Page-level multimodal processing

The extraction is performed at the page level. Each page of a PDF is processed independently to preserve local structure and avoid mixing information across sections.

For each page, two complementary inputs are generated:

```
img_bytes = render_page_png_bytes(pdf_path, page_num, dpi)
page_text = page.extract_text(x_tolerance=2)
```

Pages are rendered as high-resolution images (typically 320 DPI, increased to 400 DPI for graph-heavy pages), while the textual layer is extracted using *pdfplumber.* This dual representation ensures that both spatial layout (tables, charts) and textual content are captured.

2.3 Multimodal prompt and constrained extraction

The extraction relies on a constrained multimodal prompt applied to each page. The model receives:

- the rendered page image,
- the extracted page text,
- task-specific instructions, including a table extraction constraint.

```
resp = model.generate_content(
    [
        TABLE_HINT,
        PROMPT,
        f"PAGE_TEXT:\n{txt[:4000]}",
        Part.from_data(mime_type="image/png", data=img_bytes)
    ],
    generation_config=cfg
)
```

The prompt enforces strict rules:

- extraction limited to predefined categories,
- exact transcription of units,
- no inference or aggregation,
- full extraction of tables (all rows and numeric values),
- separation of values by panel or section,
- assignment of a source type (text, table, graph).

Each metric is returned as a structured JSON object containing:

- category
- metric_name
- value
- unit
- year
- source_type
- context
- site (operation)

The output is constrained to a single JSON array to ensure consistency and machine readability.

2.4 Deterministic configuration

To ensure reproducibility, the model operates under deterministic settings:

- temperature = 0.0
- low top_p
- JSON-only response format

This configuration prevents generative variability and ensures stable outputs across runs.

2.5 Robust JSON parsing

Model outputs may occasionally contain formatting inconsistencies. A dedicated parsing function is used to recover valid JSON structures:

```
data = parse_json_robuste(resp.text)
```

These routine handles:

- malformed arrays,
- missing separators,
- embedded JSON fragments,
- minor syntax errors.

This step ensures that extraction results remain usable even under imperfect model responses.

2.6 Rule-based filtering of extracted metrics

All extracted items are passed through a filtering function to retain only relevant LCI-compatible data:

```
if keep_item(it):
    all_rows.append(it)
```

This filtering stage removes:

- workforce and safety indicators,
- generic financial data not linked to a product,
- entries without meaningful values or units,
- irrelevant contextual noise.

For economic metrics, only values explicitly associated with a commodity (e.g. copper, nickel) and expressed in product-specific units are retained.

2.7 Context normalization and enrichment

To ensure consistency, contextual information is normalized after extraction.

For text-based metrics, missing context is reconstructed using local text matching. For table and graph entries, the context field is set to “N/A”.

This step ensures that all extracted metrics remain traceable to their original location in the document.

2.8 Page classification and adaptive resolution

Pages are dynamically classified based on their content (e.g. presence of charts, high numerical density). This classification determines the rendering resolution:

- **standard pages → 320 DPI**
- **graph-heavy pages → 400 DPI**

This improves extraction quality for complex visual elements without increasing computational cost unnecessarily.

2.9 Caching mechanism

To reduce redundant computations, a page-level caching system is implemented:

```
ck = cache_key(pdf_name, page_num, PROMPT, img_bytes)
data = load_cache(ck)
```

If a page has already been processed, the cached result is reused instead of re-calling the model.

The cache is reset between documents to avoid cross-report contamination.

2.10 Company identification

The extraction workflow includes a dedicated step to identify the reporting company based on the first page of the document:

```
dominant_name = get_company_name_from_front_page(pdf_path, model, cfg)
```

This step uses a simplified prompt to retrieve the full company name from the title or logo area. If this step fails, a fallback mechanism retrieves the company name from an external summary file.

2.11 Company and site affiliation

Extracted data are linked to standardized company and site identifiers using an external mapping file. Fuzzy matching is applied to associate extracted names with reference entries:

```
process.extractOne(name, mapping_choices, scorer=fuzz.token_set_ratio)
```

This process assigns:

- company identifiers,
- site identifiers,
- standardized company names,
- associated commodities.

This step ensures consistency across reports and enables aggregation at the site and company levels.

2.12 Anomaly detection

A lightweight validation layer flags potentially inconsistent entries, including:

- out-of-range percentage values,
- mismatches between metric type and page context,
- environmental metrics appearing in unrelated sections,
- ambiguous values extracted from graphical elements.

Flagged entries are exported separately for manual inspection.

2.13 Output files

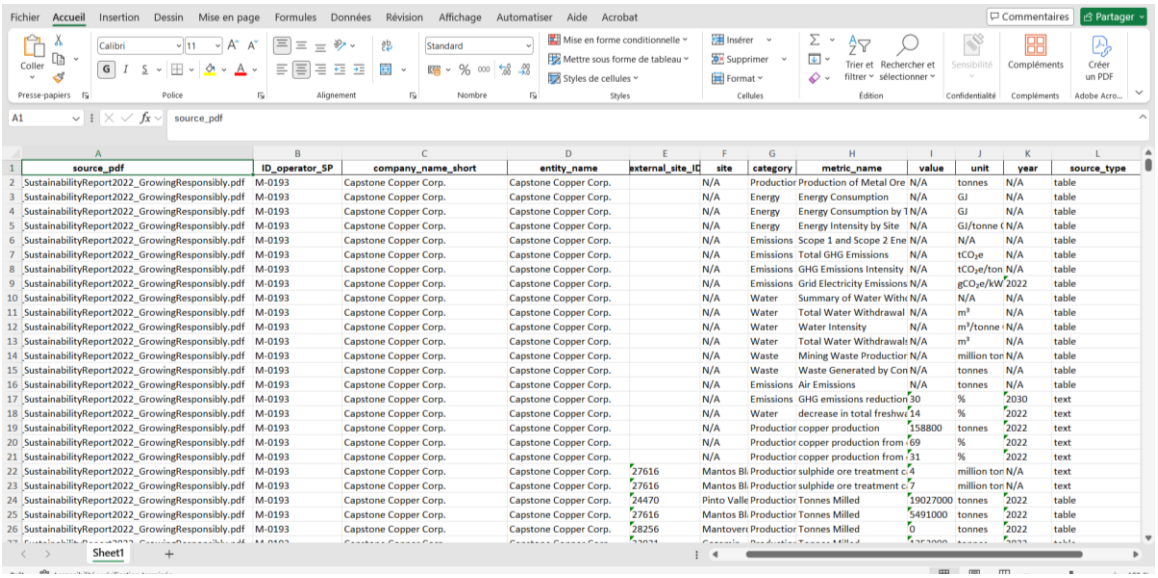
For each processed report, the extraction module generates two main output files:

- an Excel file containing all extracted and filtered metrics,
- a flags file listing entries identified as potentially inconsistent during post-processing.

The Excel output follows a standardized tabular structure directly derived from the extraction pipeline. Each row corresponds to a single extracted metric and includes the following fields:

- source_pdf: name of the processed report,
- ID_operator_SP: standardized company identifier obtained through mapping,
- company_name_short: normalized company name,
- entity_name: original or inferred company name associated with the report,
- external_site_ID: standardized site identifier when available,
- site: site or asset name extracted from the document,
- category: assigned metric category (e.g. Production, Energy, Water, Emissions),
- metric_name: name of the reported indicator,
- value: numerical value extracted from the document,
- unit: physical unit exactly as reported,
- year: reporting year when explicitly available,
- source_type: origin of the data (text, table, or graph),
- source_page: page number within the report,
- context: local textual context or snippet associated with the metric.

This structured output preserves both numerical information and contextual attributes, enabling traceability, validation, and direct integration into downstream analytical workflows:



source_pdf	ID_operator_SP	company_name_short	entity_name	external_site_ID	site	category	metric_name	value	unit	year	source_type
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Production	Production of Metal Ore	N/A	tonnes	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Energy	Energy Consumption	N/A	GJ	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Energy	Energy Consumption by T	N/A	GJ	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Energy	Energy Intensity by Site	N/A	GJ/tonne (	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Emissions	Scope 1 and Scope 2 Ene	N/A	N/A	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Emissions	Total GHG Emissions	N/A	tCO_2e	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Emissions	GHG Emissions Intensity	N/A	tCO_2e/ton	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Emissions	Grid Electricity Emissions	N/A	gCO_2e/kW	2022	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Water	Summary of Water Withd	N/A	N/A	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Water	Total Water Withdrawal	N/A	m^3	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Water	Water Intensity	N/A	m^3/tonne (	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Water	Total Water Withdrawals	N/A	m^3	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Waste	Mining Waste Production	N/A	million ton	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Waste	Waste Generated by Con	N/A	tonnes	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Emissions	Air Emissions	N/A	tonnes	N/A	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Emissions	GHG emissions reduction	30	%	2030	text
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Water	decrease in total freshw	14	%	2022	text
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Productior	copper production	158800	tonnes	2022	text
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Productior	copper production from	69	%	2022	text
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.		N/A	Productior	copper production from	31	%	2022	text
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.	27616	Mantos Bl	Productior	sulphide ore treatment c	4	million ton	N/A	text
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.	27616	Mantos Bl	Productior	sulphide ore treatment c	7	million ton	N/A	text
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.	24470	Pinto Valle	Productior	Tonnes Milled	19027000	tonnes	2022	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.	27616	Mantos Bl	Productior	Tonnes Milled	5491000	tonnes	2022	table
SustainabilityReport2022_GrowingResponsibly.pdf	M-0193	Capstone Copper Corp.	Capstone Copper Corp.	28256	Mantoverd	Productior	Tonnes Milled	0	tonnes	2022	table

Figure 5: Example of structured Excel output generated by the extraction module

The extraction module transforms PDF reports into structured datasets through multimodal analysis and rule-based post-processing:

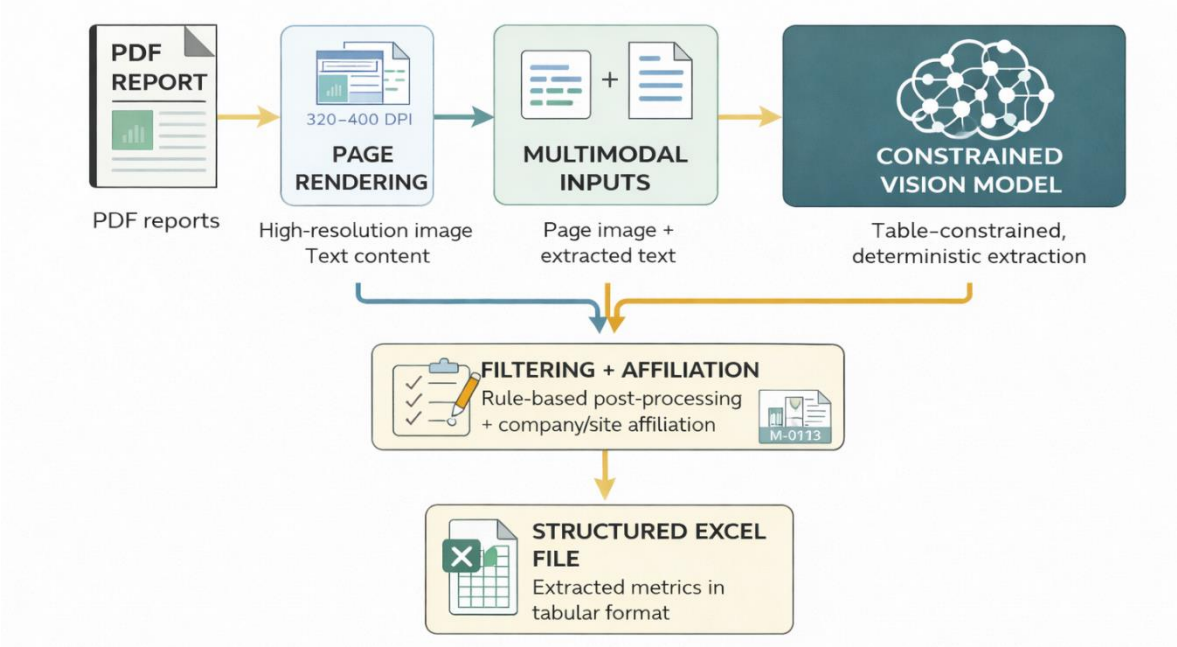


Figure 6 : Data processing workflow for data extraction from reports

2.14 Limitations

The extraction module remains subject to several limitations. First, graphical elements without explicit numerical labels may reduce extraction accuracy, particularly when values must be inferred from visual representations. Second, site attribution depends on explicit mention within the local context, which may not always be consistently provided across reports. Third, document quality, including formatting inconsistencies or low-resolution content, may affect both text extraction and image rendering.

An additional limitation arises from the page-level processing strategy, which may disrupt the global document context. Because each page is analyzed independently, relationships spanning multiple pages (e.g., tables split across pages or contextual information introduced earlier in the document) may not be fully captured. While this limitation is partially mitigated by preserving local textual context and by post-processing steps designed to maintain traceability, some cross-page dependencies may remain difficult to reconstruct.

These limitations are mitigated through a combination of deterministic prompting, rule-based filtering, anomaly detection, and subsequent expert validation, which together help ensure the reliability and usability of the extracted data.

To enhance transparency, the code used for the screening and extraction modules is available in an open-source Python repository: https://github.com/charaf-bejjit/ai_lci_report_harvesting. In addition, key statistics derived from the extracted data

are presented through an interactive Streamlit dashboard: https://lci-dashboard-charaf.streamlit.app.