

# ClimXtract: An open-source data extraction pipeline for company-level greenhouse gas emissions

Anna Steinberg<sup>1</sup>, Laia Domenech Burin<sup>2</sup>, Ailin Liu<sup>1</sup>, Malte Schierholz<sup>1</sup>, Emily Kormanyos<sup>3</sup>,  
Andreas Dimmelmeier<sup>2</sup>, Lisa Reichenbach<sup>3</sup>, Maurice Fehr<sup>3</sup>



<sup>1</sup>LMU München (Munich Center for Machine Learning)

<sup>2</sup>LMU München

<sup>3</sup>Deutsche Bundesbank

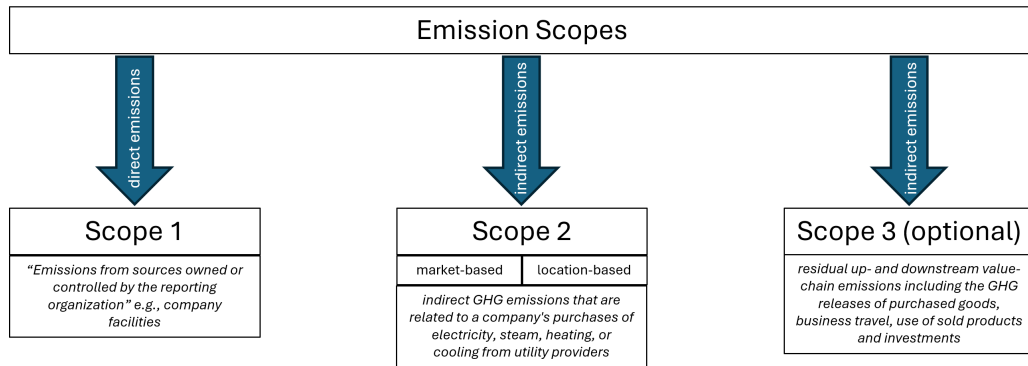
All views expressed in this presentation are personal views of the authors and do not necessarily reflect the views of Deutsche Bundesbank or the Eurosystem.

March 26, 2025

- Consistent and reliable data on company-level **greenhouse gas (GHG) emissions** are indispensable for the governance of the climate crisis (Network for Greening the Financial System (2024))

- Consistent and reliable data on company-level **greenhouse gas (GHG) emissions** are indispensable for the governance of the climate crisis (Network for Greening the Financial System (2024))
- Companies communicate their GHG emissions usually through **not standardised and unstructured** reports (PDF format), uploaded to individual company websites instead of a central repository

- Consistent and reliable data on company-level **greenhouse gas (GHG) emissions** are indispensable for the governance of the climate crisis (Network for Greening the Financial System (2024))
- Companies communicate their GHG emissions usually through **not standardised and unstructured** reports (PDF format), uploaded to individual company websites instead of a central repository
- **Third-party data providers** gather data from multiple sources through **non-transparent** methods leading to high discrepancy between different providers (Busch et al. (2022))



**Figure:** Visual illustration of GHGP guidance. Based on World Resource Institute (WRI) and World Business Council for Sustainable Development (WBCSD), 2011

## 17 | CO<sub>2</sub> emissions from energy consumption (in 1,000 t)\*

GRI 305-1/-2

	2016	2017	2018	2019	2020
CO <sub>2</sub> direct (Scope 1)	1,056	1,192	1,247	1,239	1,027
CO <sub>2</sub> indirect (Scope 2) - market-based	1,882	1,763	1,687	1,276	1,035
CO <sub>2</sub> indirect (Scope 2) - location-based	2,141	2,041	1,985	1,706	1,492
<b>Total - market-based</b>	<b>2,938</b>	<b>2,955</b>	<b>2,934</b>	<b>2,516</b>	<b>2,062</b>
<b>Total - location-based</b>	<b>3,197</b>	<b>3,233</b>	<b>3,232</b>	<b>2,946</b>	<b>2,519</b>

\* Since 2016, the "market-based" and "location-based" accounting approaches have been implemented in accordance with GHG Protocol Scope 2 Guidance. Since then, the market-based approach has been the standard accounting approach. The historical data for 2006–2015 were calculated using a method similar to the location-based approach.

Figure: Source: Daimler report 2022

## 18 | Specific CO<sub>2</sub> emissions (in kg/vehicle)\*

GRI 305-1/-2

	2016	2017	2018	2019	2020
Cars – CO <sub>2</sub> direct (Scope 1)	245	250	267	279	326
Cars – CO <sub>2</sub> indirect (Scope 2) – market-based**	611	565	562	431	426
<b>Total – Cars – Scope 1 &amp; 2</b>	<b>856</b>	<b>815</b>	<b>829</b>	<b>711</b>	<b>752</b>
Trucks*** – CO <sub>2</sub> direct (Scope 1)	746	663	629	676	742
Trucks*** – CO <sub>2</sub> indirect (Scope 2) – market-based**	1,286	1,084	933	834	954
<b>Total – Trucks – Scope 1 &amp; 2</b>	<b>2,032</b>	<b>1,747</b>	<b>1,561</b>	<b>1,510</b>	<b>1,696</b>
Vans – CO <sub>2</sub> direct (Scope 1)	372	340	355	346	333
Vans – CO <sub>2</sub> indirect (Scope 2) – market-based**	201	157	196	160	147
<b>Total – Vans – Scope 1 &amp; 2</b>	<b>573</b>	<b>497</b>	<b>551</b>	<b>506</b>	<b>479</b>
Buses – CO <sub>2</sub> direct (Scope 1)	1,408	1,177	977	1,083	1,471
Buses – CO <sub>2</sub> indirect (Scope 2) – market-based**	1,421	1,059	948	911	1,245
<b>Total – Buses – Scope 1 &amp; 2</b>	<b>2,829</b>	<b>2,236</b>	<b>1,924</b>	<b>1,994</b>	<b>2,716</b>

\* Excluding CO<sub>2</sub> from liquid fuels

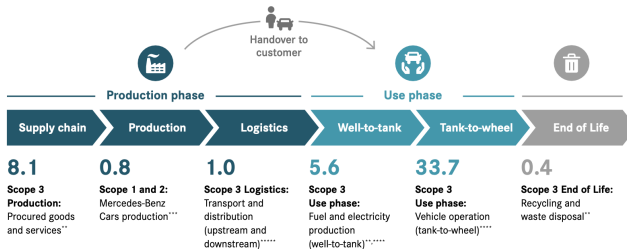
\*\* Since 2016, the "market-based" and "location-based" accounting approaches have been implemented in accordance with GHG Protocol Scope 2 Guidance.

\*\*\* Remain scopes have no longer been taken into account in the Trucks division since 2020.

Figure: Source: Daimler report 2022

## 9 | Scope 1, 2 and selected Scope 3 CO<sub>2</sub> emissions in tons per vehicle Mercedes-Benz Cars (2020)

GRI 305-3



\* For calculation basis see appendix ■ How we calculate and document our CO<sub>2</sub> emissions and ■ Scope 3 emissions Mercedes-Benz Cars

\*\* See life cycle assessment of vehicles

\*\*\* See ■ key figures environment

\*\*\*\* Driving emissions of Mercedes-Benz Cars fleet (EU, China, USA and RoW) standardized, mileage: 200,000 km, for data basis see chapter ■ Climate protection: Our CO<sub>2</sub> emissions – in all of our fleets

\*\*\*\*\* Forecast value

Figure: Source: Daimler report 2022



- Deutsche Bundesbank gathered **sustainability and annual reports** of companies acting on the European and global level

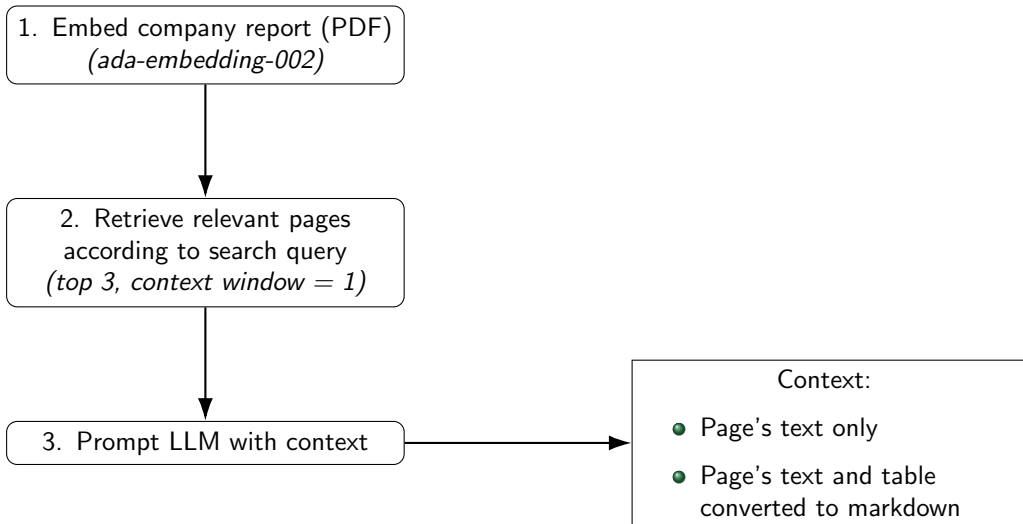
- Deutsche Bundesbank gathered **sustainability and annual reports** of companies acting on the European and global level
- We focus on a **random sample of 124 reports**

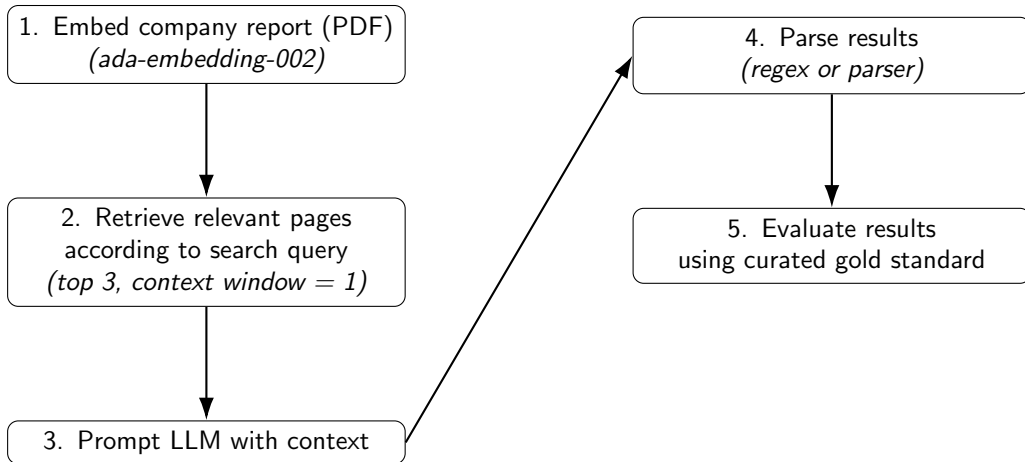
- Deutsche Bundesbank gathered **sustainability and annual reports** of companies acting on the European and global level
- We focus on a **random sample of 124 reports**
- We focus on metric *Scope* according to GHGP Protocol and extract from **text, table or graphic**

- Deutsche Bundesbank gathered **sustainability and annual reports** of companies acting on the European and global level
- We focus on a **random sample of 124 reports**
- We focus on metric *Scope* according to GHGP Protocol and extract from **text, table or graphic**

Goal: **open-source data extraction pipeline:**

1. Find reports  $\Rightarrow$  2. Extract emissions data  $\Rightarrow$  3. Save in database





Extract key pieces of information from this sustainability report.  
If a particular piece of information is not present, output 'Not specified'.

Use the following format:

0. What is the title

1. What are the Scope 1 emissions in 2013

2. What are the Scope 1 emissions in 2014

3. What are the Scope 1 emissions in 2015

4. What are the Scope 1 emissions in 2016

5. What are the Scope 1 emissions in 2017

6. What are the Scope 1 emissions in 2018

7. What are the Scope 1 emissions in 2019

8. What are the Scope 1 emissions in 2020

9. What are the Scope 1 emissions in 2021

10. What are the Scope 1 emissions in 2022

11. What are the Scope 2 (market-based) emissions in 2013

12. What are the Scope 2 (market-based) emissions in 2014

13. What are the Scope 2 (market-based) emissions in 2015

14. What are the Scope 2 (market-based) emissions in 2016

15. What are the Scope 2 (market-based) emissions in 2017

16. What are the Scope 2 (market-based) emissions in 2018

17. What are the Scope 2 (market-based) emissions in 2019

18. What are the Scope 2 (market-based) emissions in 2020

19. What are the Scope 2 (market-based) emissions in 2021

20. What are the Scope 2 (market-based) emissions in 2022

21. What are the Scope 2 (location-based) emissions in 2013

<p>You are a climate analyst tasked with extracting specific absolute numerical data from corporate reports. Your objective is to extract only the absolute values for the following Key Performance Indicators (KPIs) related to CO2 emissions across the entire company.</p>	<p>Role &amp; Objective</p>
<p>Scope 1 CO2 Emissions: Direct GHG emissions from sources owned or controlled by the organization (e.g., fuel combustion, company-owned vehicles). Scope 2 (market-based) CO2 Emissions: Indirect GHG emissions from purchased energy, calculated based on energy procurement choices (e.g., renewable energy contracts). Scope 2 (location-based) CO2 Emissions: Indirect GHG emissions from purchased energy, calculated using the average emissions intensity of the local electricity grid. Scope 3 CO2 Emissions: Indirect GHG emissions from the organization's value chain, both upstream and downstream (e.g., supply chain, business travel, product use).</p>	<p>Definitions</p>
<p>Only extract values which refer to the whole company. Only extract absolute values representing total CO2 emissions (e.g., in tons). Do not extract any relative values such as percentages, year-over-year changes, or trends. Ignore all values that are expressed as percentages (%) or involve relative comparisons (e.g., increases, decreases, or changes over time). Footnotes or annotations in metric names should be treated as references and ignored for the extraction process. Do not modify values based on footnotes, annotations, or any external data sources. Do not perform any calculations or transformations on the values. Extract and report the data exactly as presented. Do not invent values. Ensure your extraction only includes absolute values for the defined KPIs, strictly following these guidelines. If the subtype "market-based" or "location-based" is not mentioned for a value of Scope 2, always assume that the value refers to the KPI Scope 2 (location-based). Do not extract all subcategories of Scope 3, but only total values if total values are available for Scope 3. Only extract value referring to Scope 1, Scope 2 or Scope 3 separately. Do not extract values representing a sum of any the scopes.</p>	<p>Specifications</p>
<p>Year range for the search: only extract values from 2013 to 2022.</p>	
<p>Here is the excerpt: {context_str}</p>	

```
{  
  "KPI_Entries": [  
    {  
      "year": 2019,  
      "kpi_name": "3",  
      "value": 9137.0,  
      "unit": "tCO2"  
    },  
    {  
      "year": 2020,  
      "kpi_name": "1",  
      "value": 57048.0,  
      "unit": "tCO2"  
    }  
  ]  
}
```

Figure: Structured prompt with role and objective, KPI definitions and specifications

Figure: Format instructions for JSON object output



- We perform experiments varying
  - the **input mode** (“text” or “text+table”): only page’s text or page’s text+page’s table converted to markdown
  - as well as the **prompt type** (“qa\_prompt” or “structured\_prompt”)

- We perform experiments varying
  - the **input mode** (“text” or “text+table”): only page’s text or page’s text+page’s table converted to markdown
  - as well as the **prompt type** (“qa\_prompt” or “structured\_prompt”)
- Preliminary tests on subsample showed **best performance of “gpt-4o”** compared to “gpt-4o-mini” or “gpt-4-turbo” ⇒ Results presented for “gpt-4o”

- We perform experiments varying
  - the **input mode** (“text” or “text+table”): only page’s text or page’s text+page’s table converted to markdown
  - as well as the **prompt type** (“qa\_prompt” or “structured\_prompt”)
- Preliminary tests on subsample showed **best performance of “gpt-4o”** compared to “gpt-4o-mini” or “gpt-4-turbo” ⇒ Results presented for “gpt-4o”
- We evaluate our pipeline output **against curated gold standard** containing ground truth values for any scope-year combination using
  - standard information extraction metrics, e.g. precision, on row level
  - custom metrics on report-level

Metric	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
True positives	326	331	310	<b>350</b>
False positives	<b>277</b>	289	508	542
True negatives	<b>4327</b>	4321	4255	4225
False negatives	86	79	50	<b>43</b>
Precision	<b>0.54</b>	0.53	0.38	0.39
Recall	0.79	0.81	0.86	<b>0.89</b>
F1 Score	<b>0.64</b>	<b>0.64</b>	0.53	0.54

Metric	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
True positives	326	331	310	<b>350</b>
False positives	<b>277</b>	289	508	542
True negatives	<b>4327</b>	4321	4255	4225
False negatives	86	79	50	<b>43</b>
Precision	<b>0.54</b>	0.53	0.38	0.39
Recall	0.79	0.81	0.86	<b>0.89</b>
F1 Score	<b>0.64</b>	<b>0.64</b>	0.53	0.54

Metric	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
True positives	326	331	310	<b>350</b>
False positives	<b>277</b>	289	508	542
True negatives	<b>4327</b>	4321	4255	4225
False negatives	86	79	50	<b>43</b>
Precision	<b>0.54</b>	0.53	0.38	0.39
Recall	0.79	0.81	0.86	<b>0.89</b>
F1 Score	<b>0.64</b>	<b>0.64</b>	0.53	0.54

Metric	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
True positives	326	331	310	<b>350</b>
False positives	<b>277</b>	289	508	542
True negatives	<b>4327</b>	4321	4255	4225
False negatives	86	79	50	<b>43</b>
Precision	<b>0.54</b>	0.53	0.38	0.39
Recall	0.79	0.81	0.86	<b>0.89</b>
F1 Score	<b>0.64</b>	<b>0.64</b>	0.53	0.54

Metric	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
True positives	326	331	310	<b>350</b>
False positives	<b>277</b>	289	508	542
True negatives	<b>4327</b>	4321	4255	4225
False negatives	86	79	50	<b>43</b>
Precision	<b>0.54</b>	0.53	0.38	0.39
Recall	0.79	0.81	0.86	<b>0.89</b>
F1 Score	<b>0.64</b>	<b>0.64</b>	0.53	0.54



## Report-level aggregates

- ① Correct result: All correct values (irrespective of unit)

## Report-level aggregates

- ① Correct result: All correct values (irrespective of unit)
- ② Correct result: No CO2 emissions found

## Report-level aggregates

- ① Correct result: All correct values (irrespective of unit)
- ② Correct result: No CO2 emissions found
- ③ Retrieval failure: Incomplete text passed to LLM

## Report-level aggregates

- ① Correct result: All correct values (irrespective of unit)
- ② Correct result: No CO2 emissions found
- ③ Retrieval failure: Incomplete text passed to LLM
- ④ LLM extracts at least 1 wrong value
  - ① no ground truth available
  - ② ground truth on different page

## Report-level aggregates

- ① Correct result: All correct values (irrespective of unit)
- ② Correct result: No CO2 emissions found
- ③ Retrieval failure: Incomplete text passed to LLM
- ④ LLM extracts at least 1 wrong value
  - ① no ground truth available
  - ② ground truth on different page
- ⑤ LLM extracts 0 or more correct values (non-NA)

Result type	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
Correct: All values extracted	9	8	12	<u>13</u>
Correct: None found	54	54	37	38
Retrieval failure	2	2	2	2
At least 1 wrong value	16	16	34	33
0 or more correct values (non-NA)	43	44	39	38
Total number of reports	124	124	124	124

Result type	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
Correct: All values extracted	9	8	12	13
Correct: None found	<b><u>54</u></b>	<b><u>54</u></b>	37	38
Retrieval failure	2	2	2	2
At least 1 wrong value	16	16	34	33
0 or more correct values (non-NA)	43	44	39	38
Total number of reports	124	124	124	124

Result type	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
Correct: All values extracted	9	8	12	13
Correct: None found	54	54	37	38
Retrieval failure	<u>2</u>	<u>2</u>	<u>2</u>	<u>2</u>
At least 1 wrong value	16	16	34	33
0 or more correct values (non-NA)	43	44	39	38
Total number of reports	124	124	124	124



Result type	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
Correct: All values extracted	9	8	12	13
Correct: None found	54	54	37	38
Retrieval failure	2	2	2	2
At least 1 wrong value	<b><u>16</u></b>	<b><u>16</u></b>	34	33
0 or more correct values (non-NA)	43	44	39	38
Total number of reports	124	124	124	124

Result type	qa_prompt text	qa_prompt text+table	structured_prompt text	structured_prompt text+table
Correct: All values extracted	9	8	12	13
Correct: None found	54	54	37	38
Retrieval failure	2	2	2	2
At least 1 wrong value	16	16	34	33
0 or more correct values (non-NA)	43	44	39	<b><u>38</u></b>
Total number of reports	124	124	124	124

- Well-known trade-off between recall and precision apparent for this task too: adding page's table as additional context increases recall, but retrieved values are incorrect

- Well-known trade-off between recall and precision apparent for this task too: adding page's table as additional context increases recall, but retrieved values are incorrect
- Specifications given in structured prompt which serve as correctness criteria seem to be at least partially ignored, e.g. LLM converts values from "kton" to "t" in structured prompt

- Well-known trade-off between recall and precision apparent for this task too: adding page's table as additional context increases recall, but retrieved values are incorrect
- Specifications given in structured prompt which serve as correctness criteria seem to be at least partially ignored, e.g. LLM converts values from "kton" to "t" in structured prompt
- Distinguishing between reports with reported emissions and those without reported emissions:
  - results suggest already good performance for reports without any reported values (no hallucinations)
  - for reports with reported values the pipeline shows insufficient performance: the majority of reports are only partially correctly extracted

- Targeted prompt engineering based on in-depth error analysis using custom metrics
  - Refine specifications using human annotation guidelines
  - Few-shot learning
  - Chain-of-thought prompting with step to review output

- Targeted prompt engineering based on in-depth error analysis using custom metrics
  - Refine specifications using human annotation guidelines
  - Few-shot learning
  - Chain-of-thought prompting with step to review output
- Experiments with other models for failed reports
  - Models trained on self-reflection (e.g. Self-RAG) to mitigate retrieval failures
  - Models with reasoning capabilities (GPT 3o-mini) to address false positives and false negatives

- Targeted prompt engineering based on in-depth error analysis using custom metrics
  - Refine specifications using human annotation guidelines
  - Few-shot learning
  - Chain-of-thought prompting with step to review output
- Experiments with other models for failed reports
  - Models trained on self-reflection (e.g. Self-RAG) to mitigate retrieval failures
  - Models with reasoning capabilities (GPT 3o-mini) to address false positives and false negatives
- Adapt evaluation strategy to incorporate multiple ground truths values (same value on different pages or values given in different units)



Thanks for your attention!

CONTACT INFO:  
ANNA STEINBERG  
SODA LAB, LMU MUNICH  
ANNA.STEINBERG@LMU.DE

Ready for questions ...

Busch, T., Johnson, M. & Pioch, T. (2022), 'Corporate carbon performance data: Quo vadis?', *Journal of Industrial Ecology* **26**(1), 350–363.

Network for Greening the Financial System (2024), 'Information note: Improving greenhouse gas emissions data'.