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

Anna Steinberg[1], Laia Domenech Burin[2], Ailin Liu[1], Malte Schierholz[1], Emily Kormanyos[3], Andreas Dimmelmeier[2], Lisa Reichenbach[3], Maurice Fehr[3]

[1]LMU München (Munich Center for Machine Learning)
[2]LMU München
[3]Deutsche Bundesbank

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))

Background

- 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

Background

- 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))

Scopes

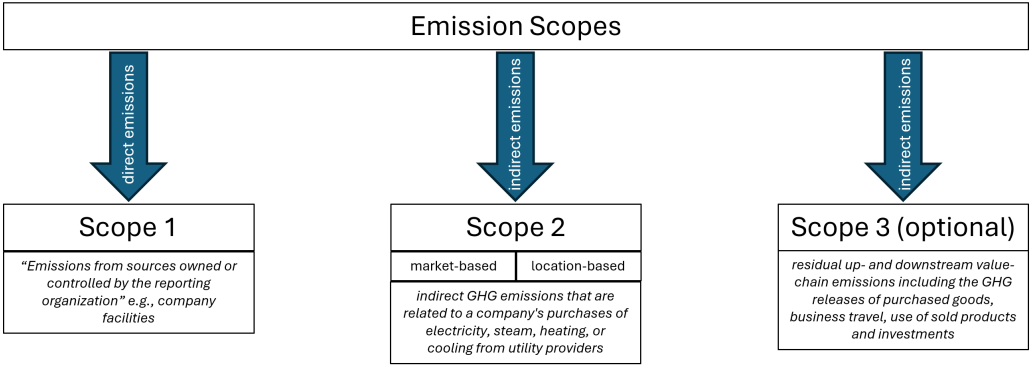


Figure: Visual illustration of GHGP guidance. Based on World Resource Institute (WRI) and World Business Council for Sustainable Development (WBCSD), 2011

Example: Matching table

17 | CO_2 emissions from energy consumption (in 1,000 t)*

GRI 305-1/-2

	2016	2017	2018	2019	2020
CO_2 direct (Scope 1)	1,056	1,192	1,247	1,239	1,027
CO_2 indirect (Scope 2) - market-based	1,882	1,763	1,687	1,276	1,035
CO_2 indirect (Scope 2) - location-based	2,141	2,041	1,985	1,706	1,492
Total - market-based	**2,938**	**2,955**	**2,934**	**2,516**	2,062
Total - location-based	**3,197**	**3,233**	**3,232**	**2,946**	2,519

* 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

Example: Wrong table

18 | Specific CO_2 emissions (in kg/vehicle)*
GRI 305-1/-2

	2016	2017	2018	2019	2020
Cars – CO_2 direct (Scope 1)	245	250	267	279	326
Cars – CO_2 indirect (Scope 2) – market-based**	611	565	562	431	426
Total – Cars – Scope 1 & 2	**856**	**815**	**829**	**711**	752
Trucks*** – CO_2 direct (Scope 1)	746	663	629	676	742
Trucks*** – CO_2 indirect (Scope 2) – market-based**	1,286	1,084	933	834	954
Total – Trucks – Scope 1 & 2	**2,032**	**1,747**	**1,561**	**1,510**	1,696
Vans – CO_2 direct (Scope 1)	372	340	355	346	333
Vans – CO_2 indirect (Scope 2) – market-based**	201	157	196	160	147
Total – Vans – Scope 1 & 2	**573**	**497**	**551**	**506**	479
Buses – CO_2 direct (Scope 1)	1,408	1,177	977	1,083	1,471
Buses – CO_2 indirect (Scope 2) – market-based**	1.421	1,059	948	911	1,245
Total – Buses – Scope 1 & 2	**2,829**	**2,236**	**1,924**	**1,994**	2,716

* Excluding CO_2 from liquid fuels
** Since 2016, the "market-based" and "location-based" accounting approaches have been implemented in accordance with GHG Protocol Scope 2 Guidance.
*** Reman scopes have no longer been taken into account in the Trucks division since 2020.

Figure: Source: Daimler report 2022

Example: Wrong graphic

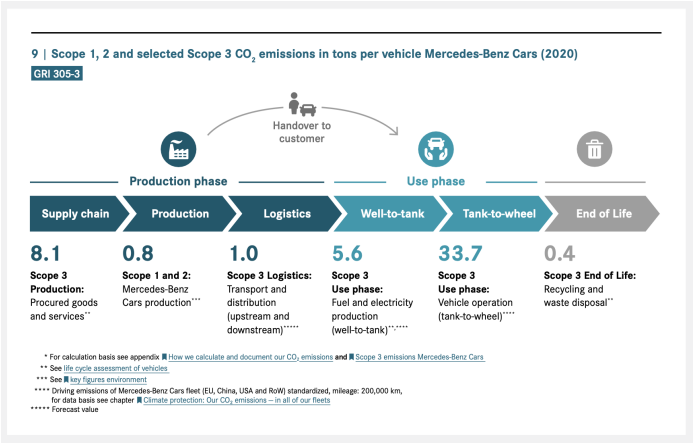


Figure: Source: Daimler report 2022

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

Setup

- 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**

Setup

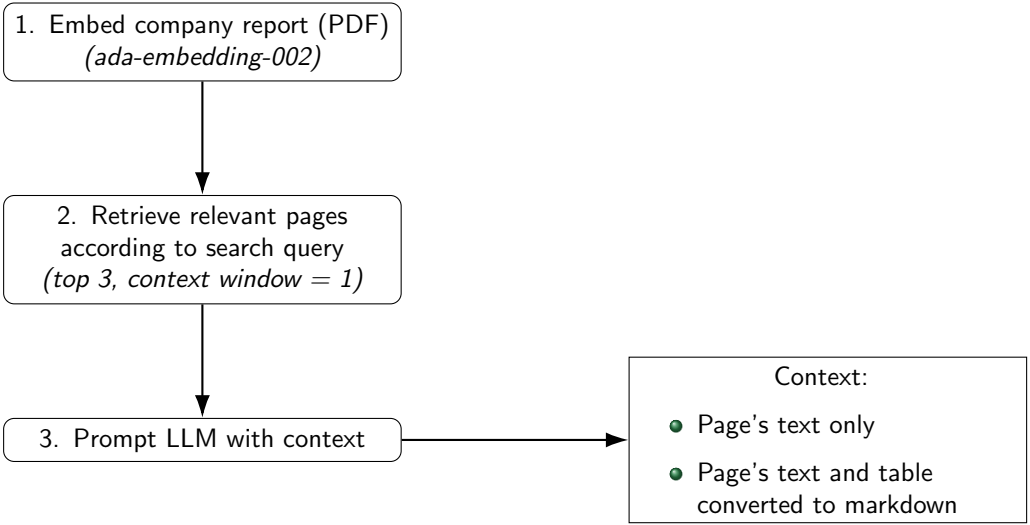
- Deutsche Bundesbank gathered **sustainability and annual reports** of companies acting on the European and global level
- We focus on a **random sample of 124 reports**
- We focus on metric *Scope* according to GHGP Protocol and extract from **text, table or graphic**

Setup

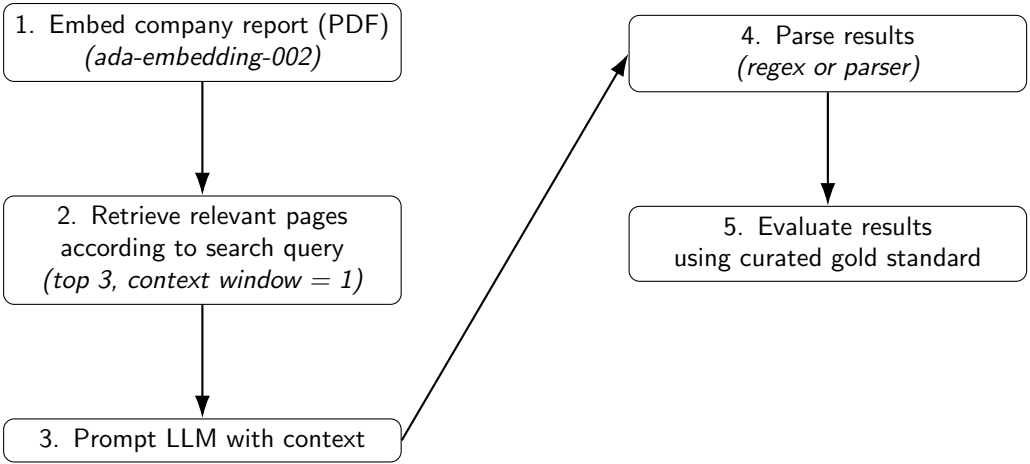
- Deutsche Bundesbank gathered **sustainability and annual reports** of companies acting on the European and global level
- We focus on a **random sample of 124 reports**
- We focus on metric *Scope* according to GHGP Protocol and extract from **text, table or graphic**

Goal: **open-source data extraction pipeline:**
1. Find reports ⇒ 2. Extract emissions data ⇒ 3. Save in database

Data extraction pipeline



Data extraction pipeline (cont.)



Prompt types: QA prompt

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

Prompt types: Structured prompt

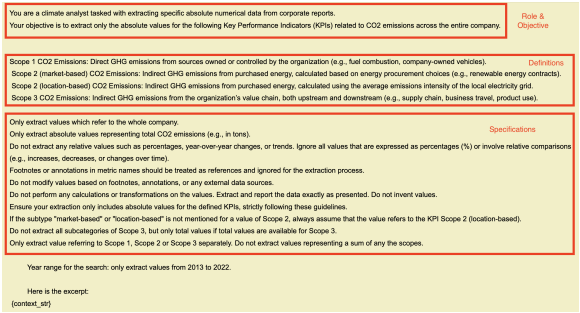


Figure: Structured prompt with role and objective, KPI definitions and specifications

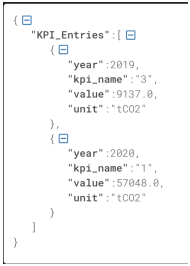


Figure: Format instructions for JSON object output

Experiments

- 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“)

Experiments

- 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“ $\Rightarrow$ Results presented for “gpt-4o“

Experiments

- 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“ $\Rightarrow$ 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

Information extraction metrics: Results

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

Information extraction metrics: Results

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

Information extraction metrics: Results

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

Information extraction metrics: Results

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

Information extraction metrics: Results

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

Custom metrics on report level

Report-level aggregates

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

Custom metrics on report level

Report-level aggregates

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

Custom metrics on report level

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

Custom metrics on report level

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

Custom metrics on report level

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)

Custom metrics on report level: Results

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	38
Total number of reports	124	124	124	124

Custom metrics on report level: Results

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	38
Total number of reports	124	124	124	124

Custom metrics on report level: Results

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	38
Total number of reports	124	124	124	124

Custom metrics on report level: Results

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	38
Total number of reports	124	124	124	124

Custom metrics on report level: Results

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	**38**
Total number of reports	124	124	124	124

Conclusions

- 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

Conclusions

- 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

Conclusions

- 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

Next steps

- 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

Next steps

- 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

Next steps

- 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 ...

References

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'.