NetZeroFacts: Two-Stage Emission Information Extraction from Company Reports

Marco Wrzalik[1], **Florian Faust**[2], **Simon Sieber**[2], **Adrian Ulges**[1]
[1]RheinMain University of Applied Sciences, Wiesbaden / Germany
[2]Sociovestix Labs, Kaiserslautern / Germany
{marco.wrzalik,adrian.ulges}@hs-rm.de, {florian.faust,simon.sieber}@sociovestix.com

Abstract

We address the challenge of efficiently extracting structured emission information, specifically emission goals, from company reports. Leveraging the potential of Large Language Models (LLMs), we propose a two-stage pipeline that first filters and retrieves potentially relevant passages and then extracts structured information from them using a generative model. We contribute an annotated dataset covering over 14,000 text passages, from which we extracted 739 expert annotated facts. On this dataset, we investigate the accuracy, efficiency and limitations of LLM-based emission information extraction, evaluate different retrieval techniques, and estimate efficiency gains for human analysts by using the proposed pipeline. Our research demonstrates the promise of LLM technology in addressing the intricate task of sustainable emission data extraction from company reports.

1. Introduction

To address the climate crisis - probably the most central and difficult challenge of our time - economies have to cope with massive industrial restructuring. The focus is on “Net Zero”, i.e. achieving at least a neutral net balance of emitted greenhouse gases (especially CO_2) as quickly as possible. To reach this goal, policies such as the EU’s *green taxonomy* are targeted at redirecting budget flows into sustainability-oriented businesses. To qualitatively capture the environmental impact of organizations, also referred to as "carbon accounting", analysts have to examine business reports and manually locate and extract the relevant data.

However, a complete and efficient screening remains an open issue: Companies often publish a variety of reports, such as quarterly, annual, sustainability and stewardship reports, which all may include aspects of emission reporting. To have a comprehensive understanding of a companies’ net zero commitment, it is important to consider all of these reports. Analysts faced with the task of gathering net zero data need to browse large amounts of text for relevant information and extract it in a structured way. The expenses incurred are significant, in particular due to the need to carry out the screening process (1) for multiple companies, (2) in different industrial sectors, and (3) at regular intervals.

To increase the efficiency of this process, we address the question whether AI technology can support analysts with extracting structured greenhouse gas emission information from business reports. Specifically, the focus of this work is on extracting *emission goals*, i.e. the intention by companies and investment portfolios to reduce carbon emissions relatively or absolutely in certain timeframes and across certain sectors/scopes. For example, given the statement “*From 2019-2025, we will reduce the carbon footprints of our investments by 29 per cent*”, we extract a fact with formal attributes (reduction-percentage=*29%*, target year=*2025*, reference year=*2019*, sector=*None*). Note that we are interested in extracting information in a *structured*, pre-defined format to store into a knowledge base. Such information extraction tasks have been studied intensely (Yang et al., 2022). However, what makes our particular task challenging is that substantial, binding goals must be discriminated from ’soft,’ vague statements and self-praise, and that external requirements induced upon a company must be distinguished from self-imposed goals. For example, the statement “In order to achieve carbon neutrality by 2050, the Energy-Climate law provides for the reduction of fossil fuels consumption by 40% by 2030 ” does not contain a self-imposed goal.

To address the above challenges, Large Language Models (LLMs) have recently appeared as a promising technology. These models show the remarkable ability to generate human-like text and perform a wide range of natural language understanding and generation tasks, serving as domain-agnostic problem solvers. The focus of our work is to investigate LLM technology for the auto-extraction of emission goals. We suggest a two-stage pipeline following the well-known retrieval-augmented generation (RAG paradigm) (Lewis et al., 2020), in which first potentially relevant passages are identified (*filtering/retrieval*) and then structured information is extracted by reasoning over these passages (referred to as *reading/extraction*). Our contributions are the following:

1. Although prior work on text classification for climate statements and emission targets exists

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 70–84

(see Section 2), there is – to our knowledge – no public datasets covering end-to-end climate goal extraction. Therefore, we have conducted our own annotation effort, resulting in a dataset we coin *NetZeroFacts*. The dataset is based on $>$ 14,000 passages from climate-related business reports. We make it available for research purposes upon request.

2. We evaluate our pipeline through a case study on the *NetZeroFacts* dataset, and assess the overall accuracy of LLMs in three experiments: (1) We study the extraction stage, including – besides quantitative results – an analysis of error cases, (2) regarding the retrieval stage, we benchmark LLM-based retrieval against various keyword baselines and challenge the necessity of LLMs (given their high computational cost), and (3) we conduct an end-to-end evaluation, in which we identify accuracy bottlenecks and assess overall the extent to which LLMs can increase analysts' efficiency in practice.

2. Related Work

Emission Screening: The basis for emission ratings are annual reports and sustainability reports, codes of conduct, or controversial publications by the press and NGOs. These sources contain facts both in plain text and tabular form, from which analysists extract KPIs or other assessable statements (Is there a code of conduct? Has the company set emission goals? Are emissions even reported? etc.). Manual fact extraction comes with considerable manual effort, and tool support in practice has so far been limited to a coarse-grained document classification of report types (in order to filter out irrelevant documents) and keyword search, which could be problematic since reports from different sources differ in form and vocabulary (e.g. "CO2 emission" vs. "carbon release"). More advanced commercial tools such as *Intelligent Tagging*[1] can identify entities and indicators, but do not use Large Language Model (LLM) support yet.

Large Language Models (LLMs): The latest generation of large (>1 billion parameters), instruction-tuned LLMs – such as OpenAI's GPTs (OpenAI, 2023) or open-source alternatives such as Llama (Touvron et al., 2023) – learn to generate text on large-scale datasets. Since the quality of results has been shown to improve drastically with model and training data scale (Wei et al., 2022), a variety of large-scale models has been trained on increasing datasets recently – see Zhao et al. (Zhao et al., 2023) for a recent in-depth overview of the model landscape. Many models have been fine-tuned to follow instructions by a human conversation partner using reinforcement learning techniques (Ouyang et al., 2022), and can thus serve as general-purpose task solvers.

Since LLMs have led to significant progress across virtually any text understanding task, they can be useful for *both* stages of our pipeline, namely *retrieving* potentially relevant passages and *extracting* facts from them. We outline research in both areas in the following.

Passage Retrieval: To identify passages containing relevant facts, the predominant industry solution remains keyword search employing word occurrence statistics such as BM25 (Robertson and Zaragoza, 2009), which has proven an effective, cheap strategy for many use cases. However, more recently, LM-based models have been shown to yield improved results. These can be trained on labeled data (e.g., (Karpukhin et al., 2020)) or in a self-supervised fashion, with adjacent text segments treated as positive sample pairs (Neelakantan et al., 2022)), and encode both queries and passages into vector representations called *embeddings*. By comparing queries' and passages' embeddings via nearest neighbor search, this dense representation-based retrieval becomes a powerful alternative to traditional retrieval methods. In our study, we will compare and discuss both fundamental approaches – keyword search and embeddings search – for identifying emission goals.

Information Extraction: For more than five years, LMs have been the go-to approach for the extraction of facts from sentences and short documents. Early LM-based approaches add a so-called *head* component on top of a pre-trained LM, and fine-tune the resulting model to the targeted extraction task given a limited number of annotated training sentences. This way, models can be tailored to specific extraction tasks (e.g., (Gao et al., 2019; Eberts and Ulges, 2019)). With the aforementioned development of instruction-tuned LLMs as general-purpose problem solvers, it seems that a quality comparable to specialized extraction heads can be reached by prompting a system and requesting it to yield a structured output (Jiao et al., 2023; Zhang et al.; Gao et al., 2023). Since this comes without the need for fine-tuning, prompting instruction-tuned models appears to be the predominant approach today, and we follow this line of work.

Climate Fact Extraction: While information extraction has been applied to various domains (such as medical texts (Rasmy et al., 2020)) and target structures (such as arguments graphs (Lawrence and Reed, 2019)), the extraction of *climate-related* information has been studied rather scarcely. Stammbach et al. (Stammbach et al., 2023) formu-

[1] https://www.lseg.com/en/data-analytics/products/intelligent-tagging-text-analytics

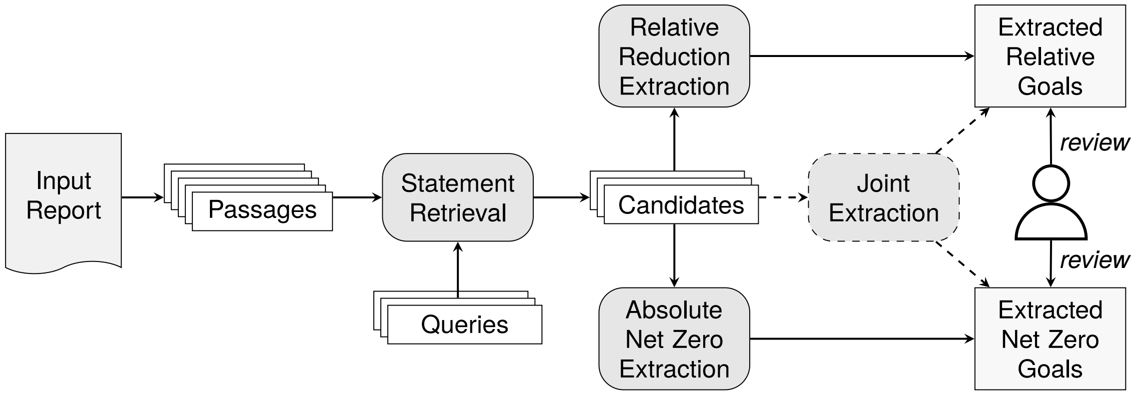

Figure 1: Approach Overview. A given report is broken down into passages, which are then filtered by statement retrieval. The resulting candidates are processed by extractors, which identify relative and net zero targets either in a joint processing step or separately.

late the detection of broader environmental statements as a binary classification problem (classifying a high vs low priotitization of environmental issues). ClimateBERT (Webersinke et al., 2021) follows the seminal BERT model (Devlin et al., 2019), combining a self-supervised masked LM pretraining on domain-specific text with a supervised fine-tuning of dedicated head models. On three climate-related text classification tasks, improvements over domain-agnostic pretraining are demonstrated. ClimateBERT-netzero (Schimanski et al., 2023) contributes a classification model and dataset for emission goal extraction. All these works primarily address text classification tasks with fine-tuned LMs, while we target a complete extraction pipeline (including retrieval *and* the extraction of structured information) and employ large-scale instruction-tuned LMs. The only other work we are aware of investigating these models specifically for climate-related text is ChatClimate (Ashraf Vaghefi et al., 2023), which – similar to our approach – investigates LLMs coupled with a climate-related text corpus. This work, however, adresses the answering of broad, climate-related questions, and not the bulk extraction of structured facts.

3. Approach

Our proposed method for extracting climate goals from given reports is targeted at two types of goals:

- A *net zero goal* expresses that a company wants to reach (at least) carbon neutrality. It comes with a target year, and optionally a subdivision of the company or the company operations.
- A *relative goal* expresses that a company wants to reduce its emissions by a certain percentage. It comes with a target year, reduction rate, reference year, and optionally a subdivision of the company or the company operations.

Goals are expressed in passages of text inside a report, consisting of at least one sentence up to a paragraph. Each report can contain multiple relevant passages, and each passage can state multiple different goals. For example, the text *"We commit to a target of carbon neutrality in own operations and own scope 1 and 2 GHG emissions reduced by at least 80% by 2030 compared with baseline year 2019."* contains

1. a *net zero goal* (`target year=2030, subdivision=own operations`)
2. and a *relative goal* (`target=80%, target year=2030, reference year=2019, subdivision=scope 1 and 2`)

Our approach towards extracting these goals is divided into two stages: First, a retrieval stage acts as a filter, limiting the amount of text to be processed and reducing false positives. Second, given the retrieved passages, we extract goals of both types. The result of the extraction is a list of structured facts, each with the aforementioned set of information fields. Figure 1 gives an overview of the approach.

3.1. Pre-processing

Our approach operates on the basis of plain text passages. Starting with PDF reports, we first extract the textual contents of each page using *Apache Tika*[2]. Next, we split the textual content of each page into sentences using the Python library *SoMaJo* (Proisl and Uhrig, 2016). The resulting sequence of sentences is used to generate overlapping passages: Each passage consists of three sentences, with subsequent passages sharing one

[2] https://tika.apache.org/

sentence. In other words, we use a sliding window of three sentences and shift this window by two positions to take the next passage. The resulting overlap reduces the risk that a passage is split in such a way that some information is missing from the target goal.

3.2. Statement Retrieval

Our statement retrieval (see Figure 2) serves as a filter for passages that contain climate goals. We use an information retrieval approach, i.e. queries are defined to express the information need for emission statements, and passages are ranked according to the relevance to these queries. Specifically, we explore two query types:

1. **Search by Question**: These are hand-crafted natural language queries that specifically ask for details to climate goals such as "*By what year do they expect to be carbon neutral?*". For this query modality, we have created a set of 14 questions.

2. **Search by Example**: These are example sentences or short passages that express one or more climate goals such as "*We are committed to carbon neutrality by 2050 with our investments*". We have collected 131 examples from held-out reports for this query modality.

Note that both methods use a pool of *multiple* queries $q_1, ..., q_n$ and that these pools can be refined iteratively with feedback.

Given a query q_i and a corpus of passages $d_1, ..., d_m$, a retriever model computes scores $s_{i,j} = score(q_i, d_j)$ which estimate the relevance of the passage. We explore two retriever models:

1. **Keyword Search** relies on *Elasticsearch*[3], a renowned industry standard search engine built on *Apache Lucene*[4]. Precisely, BM25 (Robertson and Zaragoza, 2009) is employed, a common relevance scoring technique based on keyword matching that adjusts each match based on the uniqueness of the word.

2. **Embedding Search** uses nearest neighbor search on LLM embeddings. We specifically adopt the OpenAI embedding model `text-embedding-ada-002`, which, according to the BEIR retrieval benchmark (Thakur et al., 2021), is the highest performing model currently available from OpenAI. Note that embedding search is more costly compared to keyword search, since it requires an LLM forward pass for each passage in the corpus.

[3] https://www.elastic.co/
[4] https://lucene.apache.org/

Both retriever models – given a query q_i – yield a ranked list of top results with scores s_{ij}. Given a passage d_j, these scores are fused across the queries using *score fusion* to obtain a single relevance score s^*_j indicating whether the passage contains a relevant fact (as illustrated in Figure 2). Given the passage's scores resulting from n different queries as $s_{1j}, s_{2j}, \ldots, s_{nj}$, we explore three score fusion techniques:

- **Max-Pooling**: Adopts the maximum score for a passage across all queries:

$$s^*_j = \max(s_{1j}, s_{2j}, \ldots, s_{nj}) \tag{1}$$

- **Sum**: The fused score for a passage is the sum of the scores across all queries:

$$s^*_j = \sum_{i=1}^{n} s_{ij} \tag{2}$$

- **Sum with Min-Max Normalization**: Each score is min-max normalized within its query's ranking: Let $s_i^{\min} := min_j\, s_{ij}$ and $s_i^{\max} := max_j\, s_{ij}$ be the minimum and maximum scores calculated for the i-th query, respectively. The fused score is calculated as:

$$s^*_j = \sum_{i=1}^{n} \frac{s_{ij} - s_i^{\min}}{s_i^{\max} - s_i^{\min}} \tag{3}$$

If a passage d_j is not retrieved by a query q_i, we set $s_{ij} = 0$.

3.3. Information Extraction

We feed all passages (ranked by the retriever) up to a certain cut-off rank to the extraction model. As described in the beginning of Section 3, we are interested in extracting two types of emission goals from passages, namely *net zero* goals vs *relative* goals. Both goals come with several attributes, such as a target year and (in case of relative goals) a reduction rate.

We tackle the extraction of emission targets in a two-stage process that relies heavily on few-shot prompting, using an instruction-tuned LLM (Ouyang et al., 2022). Specifically, we use the OpenAI model `gpt-3.5-turbo`. In this context, "few-shot" refers to the inclusion of a limited set of examples with correct answers, which serve as a pseudo-history accessible to the LLM.

1. In the first *filtering* stage, the LLM is asked *whether* the input passage describes at least one goal. The prompt instructs the model to respond with either "true" or "false", which is demonstrated in few-shot examples.

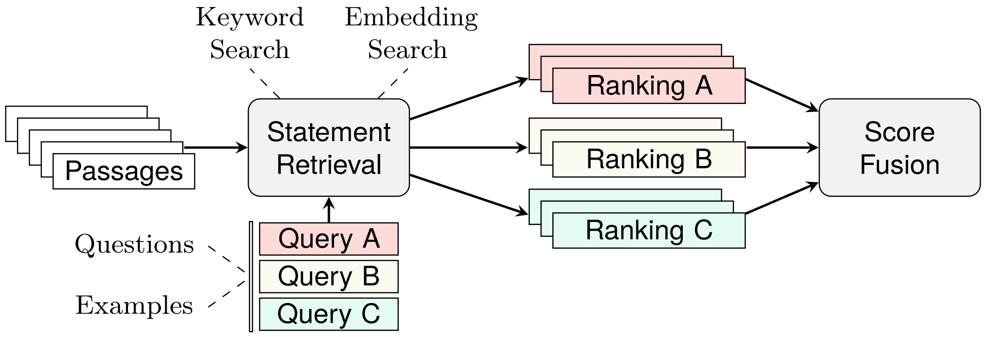

Figure 2: Retrieval is performed separately for each query or example. The resulting rankings are fused afterwards using conventional score fusing techniques.

2. If the LLM's response is positive (i.e., it starts with "true" or "yes"), the system proceeds to the actual extraction phase, which utilizes another prompt with three manually defined few-shot examples. Both the prompt and the examples instruct the LLM to produce output in the form of JSON objects containing only the fields relevant to the particular extractor. After successful parsing, these JSON objects represent our final extraction results.

In prior experiments, we found this two-stage process to outperform a single-stage extraction. This is also in line with OpenAI's public prompt engineering guidelines (Ope, 2024), which recommend to split tasks into simpler subtasks.

Note that the above two-stage process can either be executed for both types of goals at once, or separately. We explore both options:

- **Separate** prompting runs two separate two-stage processes, one for netzero goals and one for relative goals, resulting in four prompts per text passage.
- **Joint** prompting runs a single two-stage process: the filtering prompt responds positively if *either* type of goal is declared, and correspondingly the *extraction* prompt covers both goal types, resulting in two prompts per passage.

Each of the above six prompts was optimized independently from the others in a manual process of about 10 iterations, each including a small-scale inspection of a few responses but no quantitative benchmarking. Public prompt engineering guidelines were consulted in the process.

4. The NetZeroFacts Dataset

In this section, we introduce the *NetZeroFacts* dataset, which is based on real-world business reports known to contain emission statements. These were chunked into passages following our preprocessing as described in Section 3.1, and annotated by domain experts according to the criteria layed out in Appendix A. To evaluate not only end-to-end performance of our pipeline but also the individual steps, namely retrieval and extraction, the dataset consists of three partitions. We share our dataset, including all its partitions, upon request for research purposes.

NetZeroFacts-Small is based on 222 reports by different asset owner companies reporting sustainability and financial aspects (sustainability, annual, and integrated reports). The dataset's passages have been annotated by climate rating analysts during their daily sustainability rating activities, resulting in 270 passages annotated with a total of 317 climate goals. The purpose of the dataset is to evaluate the extraction step in-depth on a small-scale set of relevant passages.

NetZeroFacts-Big serves to evaluate the extraction of facts on a dense corpus of (widely irrelevant) text. It contains $13,950$ passages covering the complete content of 16 reports disjoint from the reports used for NetZeroFacts-Small.

To annotate the dataset, we applied extraction (using *separate* prompts, see Section 3.3) *densely* to all passages, resulting in 1250 climate goal facts belonging to 619 passages. The extracted facts were manually validated by an expert, resulting in a set of 422 positively validated facts in 289 passages. This dataset includes all passages, the automatically extracted facts, and the expert validation for each fact.

NetZeroFacts-Retrieval To evaluate the retrieval step, what matters is *whether* a passage contains at least one climate goal. Thus, we extend *NetZeroFacts-Big* to contain relevance labels. Relevant passages include those labeled positively by the expert *NetZeroFacts-Big*. However, since these include only passages for which LLM extraction was successful, and since our extractor may miss some climate goal facts in other passages, we also annotate additional passages for relevance using a top-15 pooled annotation of our best-performing keyword and embedding retrievals, focusing on those passages for which no facts have

Table 1: Extraction Recall and Precision on *NetZeroFacts-Small*. LLM-based extraction discovers 74.8% of known facts (left) and also yields new, unknown facts, at a precision of 71.3% (right). P is the number of positives, TP true positives, FP false positives.

Goal Type	**Annotated**	**Extracted**	**Recall**	**P**	**TP**	**FP**	**Precision**
Net Zero	295	221	0.749	358	269	89	0.751
Relative	22	16	0.727	136	83	53	0.610
Total	317	237	0.748	494	352	142	0.713

been extracted. In the resulting pool, 21 more passages were annotated as relevant. Note that such pooling is common practice in corpora for which dense annotation of relevance ratings is infeasible.

5. Experiments

In our experiment on the NetZeroFacts dataset, we investigate the individual system components and the overall end-to-end pipeline, and assess the level to which LLMs can improve the process of extracting CO_2 reduction targets from heterogeneous corporate documents:

- **Extraction Evaluation**: We first focus on the extraction step, and conduct two experiments: (a) a detailed evaluation on a small-scale set of relevant passages (*NetZeroFacts-Small*), and (b) a precision-oriented evaluation in which extraction is applied densely over all reports in *NetZeroFacts-Big*.
- **Retrieval Evaluation**: Retrieval as a pre-filtering is a key step to avoid a dense extraction – which would come with substantial computational cost and response delay in application. Therefore, we explore the different retrieval models proposed in Section 3 and assess their quality based on recall measures.
- **End-to-end Evaluation**: Finally, we assess the performance of our end-to-end pipeline, which includes the best-performing retrieval setting and the two variants of the extraction component.

5.1. Extraction Evaluation

We evaluate the extraction component in two experiments: First, we apply extraction on passages known to contain emission targets (NetZeroFacts-Small), second on the large but sparse dataset (NetZeroFacts-Big). This subsection's experiments focus on *separate* prompts (we will present a comparison of both prompting variants in the end-to-end evaluation in Section 5.3).

Detailed Evaluation (NetZeroFacts-Small): Our first evaluation on *NetZeroFacts-Small* gives us an assessment of the recall and discovery capabilities of extraction, and allows us to inspect challenge cases and common errors in-depth.

We ran extraction on *NetZeroFacts-Small*'s 270 passages, after which the correctness of the extracted facts was revised manually by an expert. Thereby, an extraction only counts as correct if <u>all</u> its fields are extracted correctly. Extractions that did not satisfy this strict criterion are counted as false positives. On the dataset, 237 extracted facts had previously been extracted by analysts in daily operations. Out of these, 74.8% have been extracted by the LLM (Table 1, left). Also, our LLM-based extraction managed to yield new facts undiscovered in the daily operations, which were again revised by the analyst. Table 1 (right) shows that 352 correct facts were discovered in total (including 115 *new* facts), at a precision of 71.3%. This indicates our pipeline's potential to increase the coverage of extraction.

An in-depth inspection revealed that most extraction mistakes fall into the following categories (ordered by descending frequency):

- relative goals and net zero goal are misclassified (69$\times$)
- the fact is missing altogether (74$\times$)
- the fact is incorrect (46$\times$)
- one goal refers to a target year of another goal in the same passage (27$\times$).

Dense Evaluation (NetZeroFacts-Big): While the passages in the last experiment were prefiltered to contain known emission targets, in a real-world scenario, the extractor is also faced with many irrelevant passages. Therefore, we performed the extraction densely for *all* 13,950 passages passages in *NetZeroFacts-Big*, resulting in 1,198 extracted fact candidates expressed in 657 passages.

These were manually revised, and the precision of the facts is reported in Table 2. We observe a significant drop in precision ($<$ 40%) compared to the previous experiment, which indicates that the LLM extracts a substantial amount of false positives from non-relevant passages. This is another motivation for pre-filtering candidate passages with a retrieval step, which will be investigated in the next section.

Table 2: Dense extraction results on *NetZeroFacts-Big* indicate a lower precision, showing that extraction tends to produce false positives on irrelevant facts.

Goal Type	Extracted	Precision
Net Zero	453	0.411
Relative	745	0.317
Total	1198	0.352

5.2. Retrieval Evaluation

We evaluate retrieval on the *NetZeroFacts-Retrieval* dataset, which features the $13,950$ passages from *NetZeroFacts-Big* with 310 positive relevance ratings. Thereby, we test the four retrieval variants outlined in Section 3.2: Using either keyword or embedding-based scoring, and using either questions or examples as queries. For each report, we rank all its passages and employ the Recall@k metric, which indicates how many of the relevant passages the analyst would discover when inspecting the top-k passages. These metrics are averaged over all reports.

First, we discuss the effectiveness of the statement retrieval using hand-crafted questions. Results are presented in the upper part of Table 3. The embedding-based retrieval variants outperform their keyword-based counterparts significantly. Furthermore, the data indicate the impact of the score fusion method and the need for its careful selection, whith a min-max-normalized sum fusion working best.

The bottom part of Table 3 shows the retrieval results when using sample *passages* as queries, which significantly improves the performance of the keyword-based retrieval approach. Again, the combination of sum fusion and min-max normalization appears most effective, while the max-pooling method significantly lags behind. A possible reason for this observation is the tendency of keyword searches to assign higher scores to longer queries. Given the different lengths of the sample passages, the longer examples are predisposed to receive higher scores, potentially leading to their dominance in a max-pooling fusion. In contrast, the embedding-centric search has an intrinsic normalization within the $[-1, 1]$ interval, making max-pooling the superior choice. However, it is noteworthy that – with example-based queries – the embedding-based approach performs much worse compared to the keyword-based search. To summarize, the sample-based retrieval method exhibits commendable performance, achieving a 95.2% recall rate for positive passages within the top 100 ranks. To do so, a keyword-based approach suffices.

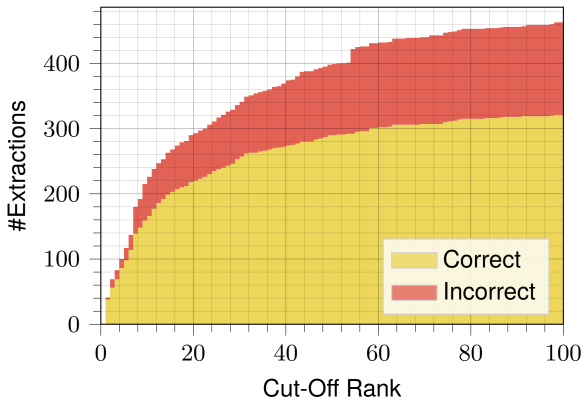

Figure 3: **End-to-end evaluation**: The number of extracted facts yield by our pipeline (*joint* prompting was used for extraction).

5.3. End-to-end Evaluation

Finally, we evaluate the entire pipeline of retrieval and extraction. We focus on the best-performing retrieval setup (keyword search with examples as queries) and evaluate extraction both with separate prompts or joint prompts (cmp. Section 3.3). For both extraction methods, an expert inspected the top-100 extracted facts (according to the associated passages' retrieval score).

Table 4 gives a comparison of both prompting methods. Joint prompting clearly outperforms separate prompts, which may be due to two reasons: First, we found the separate prompts to yield many false positives in which goal types were confused (e.g., together with a net zero goal, a relative goal with `target_rate=100%` would be extracted). Obviously, offering the LLM both goal types in the same prompt improves disambiguation between the types. Second, it should also be mentioned that separate and joint prompts were optimized independently (and ad-hoc), such that the joint prompt might per se be better suited. We share all prompts in Appendix B, and also make the prompts available with the *NetZeroFacts* dataset.

Figure 3 plots the number of facts extracted end-to-end, plotted against the cut-off rank (i.e., the number of passages per document fed to the extraction step). We observe that the correct facts flatten out at Rank 50, which yields 90% of recall compared to Rank 100. This indicates that manually reviewing only relatively few facts per report may suffice, and that the majority of facts to revise is correct.

Table 3: Evaluation measures for the retrieval stage. The best results are highlighted in boldface. Underscores indicate insignificant differences ($p \leq 0.05$) to the best result, according to a paired Student's t-test.

Search by	Method	Norm.	Fusion	Recall@10	Recall@20	Recall@50	Recall@100
Questions	Keyword	-	max	0.211	0.347	0.535	0.750
		-	sum	0.256	0.420	0.615	0.766
		min-max	sum	0.230	0.381	0.598	0.787
	Embedding	-	max	0.323	0.473	0.732	0.834
		-	sum	<u>0.352</u>	0.535	0.727	0.837
		min-max	sum	0.342	0.527	0.735	0.847
Examples	Keyword	-	max	0.280	0.414	0.618	0.817
		-	sum	<u>0.385</u>	<u>0.595</u>	0.853	<u>0.949</u>
		min-max	sum	<u>0.379</u>	**0.612**	**0.886**	**0.952**
	Embedding	-	max	<u>0.354</u>	0.510	0.728	0.851
		-	sum	<u>0.392</u>	0.526	0.737	0.858
		min-max	sum	**0.407**	0.543	0.778	0.869

Table 4: Performance metrics for joint and separate fact extraction.

Prompt	Goal Type	Extracted@100	P@5	P@15	P@50	P@100
Separate Prompt	Net Zero	402	0.445	0.452	0.478	0.440
	Relative	541	0.517	0.460	0.412	0.381
	Total	943	0.483	0.456	0.441	0.406
Joint Prompt	Net Zero	193	0.910	0.922	0.874	0.870
	Relative	258	0.774	0.653	0.631	0.585
	Total	451	0.838	0.757	0.727	0.693

6. Discussion

Workflow Efficiency In practice, analysts currently search with a list of keywords and manually inspect the detected keywords' textual contexts. Compared to this, our pipeline offers the following benefits: (1) our retrieval operates with a larger set of sample queries over which we pool, yielding a better prioritization to begin with, (2) analysts can skip passages for which extraction has yield no results (i.e., the extractor acts as an additional filter), and (3) while analysts must read and digest each detected passage so far, they only have to check pre-filled facts when using our approach. This leads to significant speed-ups in the extraction process: While processing one report with the current workflow takes 91 minutes on average (estimated on the 222 base documents from which *NetZeroFacts-Small* was constructed), we found the inspection of LLM results up to Rank 100 to take ≈ 15 minutes per report (which corresponds to a $6\times$ speedup).

Accuracy and Bottlenecks: According to our results, LLM processing is not sufficient for dark processing without expert cross-checking yet. However, we found many results to be *partially correct*, or semantically correct but formatted inadequately. True error cases for extraction are often tied with complicated passages in which multiple goals coincide (see Appendices C+D for examples). Accuracy could definitely be improved significantly with more thorough postprocessing, and future research could investigate domain-specific fine-tuning.

When it comes to accuracy bottlenecks, we consider the extraction step the limiting factor towards a fully automated extraction. Retrieval appears to be sufficiently accurate (with a recall@100 of $> 95\%$). Accordingly, we have refrained from fine-tuning task-specific LM-based retrieval models. Also, we found a well-tuned keyword-based approach employing example passages as queries to outperform LLM-based embeddings, which is good news from a cost perspective.

Future Research: One direction of future work could be to investigate NetZero classifiers based on fine-tuned LMs (Schimanski et al., 2023) for retrieval, another one to incorporate analysts' feedback when correcting LLM results. Note that both steps of our pipeline can take such feedback into account: In retrieval, passages can be used as example-based queries. In extraction, similar or particularly challenging passages can be included as few-shot examples (see, e.g., (Zhao et al., 2021)). Our *NetZeroFacts* dataset offers a good basis to explore these issues further.

7. Bibliographical References

2024. OpenAI docs prompt engineering.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara colesanti Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *CoRR*, abs/1909.07755.

Jun Gao, Huan Zhao, Yice Zhang, Wei Wang, Changlong Yu, and Ruifeng Xu. 2023. Benchmarking large language models with augmented instructions for fine-grained information extraction.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical*

Methods in Natural Language Processing, pages 15745–15756, Singapore. Association for Computational Linguistics.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *ArXiv*, abs/2110.12010.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19).

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *ACL*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A. Annotation Criteria

Annotations for *NetZeroFacts* were collected by domain experts deciding whether given text passages express a (net zero, or relative) goal according to the following criteria. Particularly, the expressed goals ...

- **must be measurable**: The mere description of climate-relevant activities without setting a reduction goal is labeled as irrelevant (such as *"We are working on infrastructure to make our facilities more energy efficient"*).
- **must be self-imposed**, meaning that constraints imposed through laws or other actors rather than the report's authors themselves do not count as goals (such as *"The Clean Shipping Act calls for the elimination of carbon emissions by 2024."* or *"70% of our customers have set the goal to reach carbon neutrality by 2024"*).
- **must directly address a CO_2 metric**: Statements addressing an indirect effect on emissions are not considered emission goals (such as *"We announce our commitment to phase out of coal by 2030"*).
- **must not report past emission statistics**: Obviously, the reporting of past emission statistics (such as *"We have reduced our absolute scope 1 and 2 emissions by 3% in 2022"*) does not count as a goal.
- **may not match the given schema**: Rarely, we discovered goals expressing an *absolute* reduction, but *not* to net zero (such as *"By 2025, we will reduce GHG emissions by 700 tons"*). While these cases were so rare that we did not measure them in the extraction + end-to-end benchmarks, we labeled them as relevant in the retrieval evaluation.

B. Prompts

For the sake of transparency, we are sharing our prompts with the community. All six prompts used in our extraction evaluation are listed below.

B.1. Net Zero Goal Filter

```
System: You are an information
    extraction tool for net zero climate
     goals. I will provide you with
    statements taken from asset manager
    reports. You will determine whether
    the given statements indicate an
    commitment to achieving net zero or
    carbon neutrality. Relative
    reductions in greenhouse gas
    emissions are excluded. Refraining
    from fossil fuels entirely does
    count as becoming carbon neutral.
    Your answers are 'True' if the
    statement contains an absolute net
    zero goal and 'False' in any other
    situation.
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
     (AOA) where a large number of the
    worlds biggest investor s commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: True
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI:  False
Human: In our climate targets published
    in 2019, we are committed to exiting
     from investments in thermal coal by
     2025. We are also committed to
    excluding oil exploration from our
    investments by 2030.
AI: False
Human: <input>
```

B.2. Relative Goal Filter

```
System: You are an information
    extraction tool for net zero climate
     goals. I will provide you with
    statements taken from asset manager
    reports. You will need to determine
    whether the statements given express
     a percentage of reduction in
    greenhouse gas emissions or not.
    Absolute zero emission targets or
    goals to achieve net zero or become
    carbon neutral do not matter. Your
    answers will be 'True' if the
    statement contains a relative
    reduction target expressed as a
    percentage, and 'False' otherwise.
    Only answer 'True' if a concrete
    relative reduction rate is given.
    Achieving net-zero does not count a
    relative reduction.
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
     (AOA) where a large number of the
    worlds biggest investors commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: False
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI: True
Human: In our climate targets published
    in 2019, we are committed to exiting
     from investments in thermal coal by
     2025. We are also committed to
    excluding oil exploration from our
    investments by 2030.
AI: False
Human: <input>
```

B.3. Net Zero Goal Extraction Prompt

```
System: You are an information
    extraction agent for climate goals.
    I will provide you with statements
    taken from asset manager reports
    that may contain one or more goals
    concerning net zero or neutral
    carbon emissions. For each goal you
    wil extract the target year until
    which the goal should be achieved.
    In addition, a goal can be dedicated
     to a certain sub-division, meaning
    that the reduction applies only to
    this area. Extract all of these
    goals in a list of json objects.
    Ignore all relative goals that state
     a specificreduction percentage.
    Only consider goals regarding
    absolute carbon neutrality or net
    zero emissions. If the target year
    or the sub-division is not given,
    fill the respective field with 'null
    '. Examples are provided below.
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI: []
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
     (AOA) where a large number of the
    worlds biggest investors commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: [
  {
    "target_year": "2050",
    "division_subset": null
  }
]
Human: The company has published its
    commitment to decarbonisation,
    setting stringent objectives: to
    reduce the intensity of its CO2
    emissions to 50% below those of 2007
     by 2030, and to be carbon- neutral
    by 2050, forecasting virtually zero
    emissions in Europe by 2030. These
    goals have been recognised as being
    based on science in accordance with
    the Science Based Targets initiative
AI: [
  {
    "target_year": "2050",
    "division_subset": null
  },
  {
    "target_year": "2030",
    "division_subset": "Europe"
  }
]
Human: <input>
```

B.4. Relative Goal Extraction Prompt

```
System: You are an information
    extraction agent for climate goals.
    I will provide you with statements
    taken from asset manager reports
    that contain one or more goals
    concerning relative reductions in
    carbon emissions and you will
    extract these goals including
    specific facts: Each goal must state
     a concrete reduction percentage.
    Optionally, each goal has a target
    year until which the goal should be
    achieved, and a reference year to
    which the relative reduction is
    compared. In addition, a goal can be
     dedicated to a certain sub-division,
     meaning that the reduction applies
    only to this area. Extract all of
    these goals and their in a list of
    json objects. Ignore all absolute
    net zero or carbon neutrality goals,
     only considerrelative reductions
    with a reduction percentage. If a
    reference year or the sub-division
    is not given, fill the respective
    field with 'null'.Examples are
    provided below.
```

```
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI: [
  {
    "relative_target_value": "29 per
        cent",
    "target_year": "2025",
    "reference_year": "2019",
    "division_subset": null
  }
]
Human: Therefore, together with the Net-
    Zero Asset Owner Alliance, we have
    spec- ified concrete CO2 reduction
    targets for the different parts of
    the real estate portfolio for the
    2019 to the end of 2024 period. For
    our Danish commercial prop- erties,
    the target is to reduce our CO2
    emissions by 39 per cent, and for
    our Danish residential properties,
    the target is 35 per cent. For the
    international part of the property
    portfolio, we are following the
    recommenda- tions of the EU-backed
    CRREM (Carbon Risk Real Estate
    Monitor) initiative, which is based
    on the Paris Agree- ments target of
    keeping global temperature increases
     below 1.5 degrees Celsius this
    century. In this context, the
    reduction target is 20 per cent for
    our German residential properties.
AI: [
 {
    "relative_target_value": "20 per
        cent",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "German
        residential properties"
  },
  {
    "relative_target_value": "39 per
        cent",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        commercial prop- erties"
  },
  {
    "relative_target_value": "35 per
        cent",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        residential properties"
  }
]
Human: <input>
```

B.5. Joint Goal Filter

```
System: You are an information
    extraction tool for climate goals
    that classifies whether a given text
     contains a statement about the
    commitment to a goal regarding
    carbon emissions. I will present to
    you passages from asset managers'
    reports. You will determine whether
    the given text contains a commitment
     to either a specific relative
    reduction in carbon emissions or to
    achieving net zero or carbon
    neutrality. Ignore any vague
    statements; a target is only a
    target if it states by when the
    target is to be achieved. For
    relative emission reductions, a
    specific percentage reduction must
    be stated. Ignore goals of third
    parties. Your answers are only 'True'
     if the statement contains such a
    climate target of the asset manager
    and 'False' if it does not.
Human: A standout feature of the
    GreenTech Solutions Factory is its
    pledge to function entirely on
    renewable energy sources, aiming for
     net-zero emissions across its
    operations.
AI: False
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
     (AOA) where a large number of the
    worlds biggest investors commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: True
Human: Net zero emissions means
    achieving a balance between
    greenhouse gas (GHG) emissions
    produced and the amount removed from
     the atmosphere, consistent with
    limiting global warming to 1.5C and
    neutralising the impact of any
    residual emissions by permanently
    removing an equivalent amount of
    carbon dioxide (CO2). For BTPS this
    will mean reducing the portfolio's
    emissions through changing
    investments and investing in
    technologies which reduce emissions.
AI: False
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI: True
Human: In our climate targets published
    in 2019, we are committed to exiting
     from investments in thermal coal by
     2025. We are also committed to
    excluding oil exploration from our
    investments by 2030.
AI: False
Human: <input>
```

B.6. Joint Goal Extraction

```
System: As an information extractor
    specialized in climate goals, your
    task is to analyze reports from
    asset managers. You are going to
    extract specific goals related to
    reducing carbon emissions or
    achieving net-zero carbon emissions,
     whether through their investments
    or policies within their own company.
     You'll respond with JSON objects
    detailing these goals.
 There are two types of targets:
   1. Relative Reduction (goal_type: '
       relativeReduction'): Specifies a
       percentage reduction with a target
        and base year.
   2. Net Zero (goal_type: 'netZero'):
       Specifies a commitment to carbon
       neutrality with a target year.
A 100 per cent relative reduction is
    also classified as netZero and not
    as relativeReduction. In addition, a
     goal can be dedicated to a certain
    sub-division, meaning that the goal
    applies only to this area, such as '
    energy consumption', 'fossil fuels'
    or emissions in a certain scope. If
    no target year or subdivision is
    specified, use 'null'. However, a
    relative reduction goal MUST specify
     a concrete reduction percentage;
    otherwise it is not a relative goal.
     Do not extract goals of third
    parties. Return an empty list if no
    targets are found. Ensure that the
    JSON objects are valid.
Human: The Albert Jackson Processing
    Centre will operate with net-zero
    emissions.
AI: []
Human: We aim to reduce CO2 emissions by
     39% for Danish commercial
    properties and 35% for residential
    by 2024, from a 2019 baseline. For
    German residential properties, the
    target is a 20% reduction by 2024,
    aligning with CRREM's guidelines.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "39%",
```

```
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        commercial properties"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "35%",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        residential properties"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "20%",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "German
        residential properties"
  }
]
Human: Our commitment includes reducing
    CO2 emission intensity by 50% from
    2007 levels by 2030, and 100% by
    2050.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "50%",
    "target_year": "2030",
    "reference_year": "2007",
    "division_subset": null
  },
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": null
  }
]
Human: From 2019 to 2025, we will cut
    our investments' carbon footprint by
     29%.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "29%",
    "target_year": "2025",
    "reference_year": "2019",
    "division_subset": "investments"
  }
]
Human: Joining the Net-Zero Asset Owner
    Alliance, we commit to net carbon
    neutrality for investments by 2050.
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": "investments"
  }
]
Human: <input>
```

C. Exemplary Extraction Showcase

Based on the joint prompt extractor, we present some extractions, some of which reflect remarkable capabilities of the underlying LLM.

C.1. Successful Single-Target Extraction

```
Human: This means that with the same
    construction measures, by 2040 a
    CO² intensity of 2.41 kg CO²eq/m 2a
    will be reached, which corresponds
    to a reduction of 88% compared with
    the 2019 reference year. The
    targeted objectives are thus largely
     already being achieved. In order to
     plan the specific measures covering
     the three action areas -energy,
    heat and building shell renovations -
    holistically and co-ordinate them
    with each other, Swiss Prime Site
    integrated the requirements of the
    CO² reduction pathway into the
    property strategy for each
    individual property.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "88%",
    "target_year": "2040",
    "reference_year": "2019",
    "division_subset": null
  }
]
```

C.2. Multiple Goals from Bullet Point List

```
Human: Sustainability Report 2022 |
    Environment 24 Our climate roadmap
    Our climate roadmap -what we are
    aiming for Founding member of the
    Net Zero Asset Managers initiative
    Addressing our own emissions
    Addressing the emissions of our
    lending activities (from 2020 levels)
     Addressing our supply chain Net
    zero across our activities (scopes 1,
     2 and 3) -in line with fi duciary
    duties •Achieve net-zero scope 1 and
     2 emissions •Reduce our own energy
    consumption by 15% from 2020 levels
    •Offset historical emissions from
    own operations back to 2000 •Net-
    zero GHG emissions by our key
    vendors Commitment to net zero by
    2050 and founding member of the Net-
    Zero Banking Alliance (the NZBA) and
     the Glasgow Financial Alliance for
    Net Zero (GFANZ) 2021 by 2025 by
    2030 by 2035 by 2050 •Reduce
    emissions intensity associated with
    UBS lending to: -residential real
    estate by 42% -commercial real
    estate by 44% -power generation by
    49% -cement by 15% •Reduce absolute
    fi nanced emissions associated with
    UBS lending to: -fossil fuels by 71%
     2020 2022Advisory vote on the
    climate roadmap passed by
    shareholders at the Annual General
    Meeting 24
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": null
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "42%",
    "target_year": "2030",
    "reference_year": "2020",
    "division_subset": "residential real
        estate"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "44%",
    "target_year": "2030",
    "reference_year": "2020",
    "division_subset": "commercial real
       estate"
  },
  <...>
]
```

C.3. Translating "halve" to 50%

```
Human: We aim to halve our carbon
    footprint this decade from 2018
    levels and will report our progress,
     transparently. To achieve that, we
    are working on transforming every
    aspect of our business -from
    switching our trucks to alternative
    fuels and our factories to renewable
     electricity, to improving circular
    packaging solutions that reduce
    plastic and can be more easily
    recycled or reused. These
    transformations are not just good
    for the climate: becoming a business
     that puts regeneration at the core
    of our company reduces our costs and
     makes us more resilient -and this
    is good for Nestlé, too.
AI: [
```

```
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "50%",
    "target_year": "2030",
    "reference_year": "2018",
    "division_subset": null
  }
]
```

C.4. Successful Interpretation of Table Data

```
Human: These metrics and targets are
    shown in the table on the right.
    STRATEGY PILLAR 2026 MILESTONE
    TARGETS 2030 ASPIRATIONS 2030 SBTi
    COMMITMENTS 2050 TARGET METRIC 2022
    BASELINE END 2026 TARGETS ENERGY %
    reduction in Scopes 1&2 CO2e
    emissions 212.3k TONNES 22%
    REDUCTION 70% of energy to come from
     renewables -46.2% in Scopes 1&2
    emissions vs 2019 baseline Net-Zero
    emissions in our value chain
    MATERIALS % volume free from new oil-
     extraction raw materials 26% 60%
    100% -33% in Scope 3 emissions vs
    2019 baseline WATER % of water to be
     recycled 23% 33% INCREASE IN
    RECYCLING RATE WASTE No waste to
    landfill 1.96k TONNES ZERO TO
    LANDFILL All effluent to meet ZDHC
    limits 92% 100% PEOPLE % employees
    in units with Great Place to Work
    certification 86% 88% 90% % of
    females in Senior Leadership
    positions 21% 30% 40% 12 Coats Group
     plc Sustainability Report 2022
    STRATEGY OVERVIEW CLIMATE REPORT
    WATER WASTE PEOPLE MATERIALS ABOUT
    ENERGY MANAGING SUSTAINABILITY
    PERFORMANCE SUMMARY
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "22%",
    "target_year": "2026",
    "reference_year": "2022",
    "division_subset": "energy"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "46.2%",
    "target_year": "2030",
    "reference_year": "2019",
    "division_subset": "energy"
  },
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": "energy"
  },
  <...>
]
```

D. Exemplary Error Cases

We usually encounter similar types of errors during extraction. Some examples of typical errors are listed below.

D.1. Confusion with Non-Climate Goals

```
Human: Performance Report // Climate
    protection targets in harmony with
    the Paris Agreement and net zero
    emissions by 2050 // Proportion of
    women in top management 33% by 2025
    and 50% by 2030 // Sustainability
    targets as part of the variable
    compensation of the Board of
    Management
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": null
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "33%",
    "target_year": "2025",
    "reference_year": null,
    "division_subset": "proportion of
       women in top management"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "50%",
    "target_year": "2030",
    "reference_year": null,
    "division_subset": "proportion of
       women in top management"
  },
  {
    "goal_type": "null",
    "target_year": null,
    "division_subset": "sustainability
       targets as part of the variable
       compensation of the Board of
       Management"
  }
]
```

D.2. Additional Irrelevant Fact

```
Human: Reducing impacts from production
    Net zero Operations Downstream
    Upstream Promoting environmentally
    preferable products Improving
    transit management and supporting
    mixed-use buildings Improving energy
     efficiency of products
    Transitioning to renewable
    electricity Building resilient
    supply chain management and
    procurement Investing in digital
    solutions for real estate management
     Reducing impact from maintenance
    and service Climate action In 2021,
    we were the first global E&E company
     to declare a commitment to achieve
    net zero by 2040, with mid-term 2030
     targets covering scope 1, 2, and 3
    emissions against a 2020 baseline
    aligned to a 1.5-degree pathway.
    These were formally approved in 2022
     by the Science Based Targets
    initiative (SBTi), a multi-
    stakeholder climate initiative. We
    have also committed to shorter-term
    targets: fully transitioning to
    renewable electricity by 2025 when
    joining RE100 in 2021, the global
    environmental initiative led by the
    Climate Group in partnership with
    CDP; and imple- menting the ISO
    50001 management system for all
    major production sites by 2025.
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2040",
    "division_subset": null
  },
  <...>
  {
    "goal_type": null,
    "target_year": "2025",
    "division_subset": "ISO 50001
       management system"
  }
]
```

D.3. Reporting Instead of Goals

```
Human: Canada Post Corporation 2022
    Sustainability Report 7 Introduction
     Governance Environment Social
    Transparency 2022 ESG highlights
    Environment Cl imate action •One of
    the first 150 companies in the world
     (one of only four Canadian
    companies) with a net-zero target
    approved by the SBTi •6% decrease in
     scope 1 and 2 emissions in 2022
```

```
    versus 2019 •12% decrease in scope 1
     and 2 emissions intensity (kt CO 2
    e/billion dollars revenue) versus
    2019 <...>
AI: [
  {
    "goal_type": "netZero",
    "target_year": null,
    "division_subset": null
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "6%",
    "target_year": null,
    "reference_year": "2019",
    "division_subset": "scope 1 and 2
        emissions"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "12%",
    "target_year": null,
    "reference_year": "2019",
    "division_subset": "scope 1 and 2
        emissions intensity"
  },
]
```