LLM-driven Open Information Extraction of Business Data from Documents on Climate Change

Samuele Baroni[1], Chiara Ferri[1] and Fabio Celli[1,*]

[1]Research & Development, Maggioli SpA, via Bornaccino 101, 47822 Santarcangelo di Romagna

Abstract

Climate change adaptation projects frequently face challenges with poor financial sustainability, a problem rooted in factors like non-viable business models and high investment uncertainty. This paper presents the preliminary results of an Open Information Extraction task, based on Large Language Models, to obtain structured business-related data from documents about climate change. We define a specific schema centered around the Business Model Canvas and we test different models and prompts. To ensure trustworthiness and mitigate risks of cultural hacking, we specifically focus on the performance of a European model, Mistral, compared to similar commercial US models. We evaluate the quality of the information extracted against two test sets, one manually compiled by three human annotators, and one filled with random words from the documents. Our robust evaluation framework includes soft matching with cosine similarity, coverage and sparsity, and hard matching with traditional Precision, Recall, and F1 measure.

Keywords

Information retrieval and access, Open Information Extraction, Large Language Models, climate change

1. Introduction and Related Work

The issue of poor financial sustainability characterizes most climate change adaptation initiatives and projects, despite their proven socio-economic value [1]. To address this issue, it is necessary to analyze the root causes of this poor financial sustainability, like the absence of viable business models, difficulties in identifying co-benefits, and the high uncertainty of climate investments. Crucially, digital libraries and online repositories contain a vast amount of unstructured data about climate change in the form of text, tables, and charts within scientific papers and project reports [2]. Previous approaches exploited Open Information Extraction (OIE) and Knowledge Graphs (KG) to extract this information and structure it [3], but the rise of Large Language Models (LLMs) introduces both opportunities and challenges for data extraction and evaluation [4].

The approach of building KGs from climate publications and scientific texts involves the previous identification of key entities (e.g., projects, tools, events, etc.) and the relationships between them [5]. LLMs proved to be helpful in this approach. Systematic evaluation of entity extraction with Llama-3.3-70B achieves state-of-the-art performance in Named Entity Recognition (F1 measure: 0.378), but analysis reveals critical challenges in technical relation extraction and emerging concept linking, with 26.4% of unlinkable entities, especially if the goal is to move beyond extracting only technical variables to also capture societal resilience and historical impact records of climate change

Building KGs from publications about climate change requires first identifying a domain-specific ontology with key entities, such as persons, organizations, projects, tools, events, and their underlying relationships [5]. While LLMs have become integral to this process, systematic evaluation of Llama-3.3-70B reveals nuanced performance: although it achieves state-of-the-art results in Named Entity Recognition (F1: 0.378), significant hurdles remain. Specifically, the model struggles with technical relation extraction and linking emerging concepts, leaving 26.4% of entities unlinkable. These limitations

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.

fabio.celli@maggioli.it (F. Celli)

https://github.com/facells/fabio-celli-publications (F. Celli)

0000-0002-7309-5886 (F. Celli)

are particularly evident when attempting to move beyond technical variables to capture broader themes like societal resilience and historical climate impacts [6]. Previous attempts to build trusted KGs to mitigate climate change denial, highlight the need for specialized Named Entity Recognition models and the annotation of domain-specific relations [7]. However, creating a large, expert-annotated corpus from complex climate science and policy documents is often prohibitively expensive and time-consuming [8].

On the contrary, OIE aims to distill structured representations (often subject-predicate-object triples) of information from unstructured text, without requiring a predefined domain-specific ontology. Information Extraction systems are increasingly shifting away from specialized solutions toward generalizable methods. These systems must now convert free-form text into structured knowledge across diverse, unseen data types, a transition largely driven by the practical impossibility of gathering training data for every specific information need [9]. The scientific literature shows a growing trend in using OIE techniques to manage the vast and complex corpus of climate change documents [10], but also highlights significant challenges inherent to the domain [11]. For instance, while OIE is designed to be domain-independent, its performance often declines when applied to scientific and policy documents compared to general web or encyclopedic content. Information extraction is further complicated by non-standardized document formats, inconsistent structures, and rich non-textual elements like figures, which remain difficult to process accurately. However, LLMs have shown significant promise in OIE tasks, particularly in few-shot settings. This is especially valuable for the low-resource climate domain, where extensive manual annotation is often unfeasible [12]. The evaluation of OIE systems necessitates more nuanced metrics than the classical Precision, Recall, and F1. These traditional measures are primarily designed for exacting string overlaps, making them insufficient for the broader requirements of OIE extraction.

The field of OIE has increasingly recognized the limitations of these metrics in capturing semantic equivalence, especially when dealing with the outputs of modern, generative, or LLM-based OIE systems. Evaluation metrics for OIE systems are task-dependent, aligning with three main tasks [13]: Open Relation Triplet Extraction (ORTE); Open Relation Span Extraction (ORSE) and Open Relation Clustering (ORC). ORTE is noted for its efficiency, but struggles with overlapping relations. ORSE extracts patterns rather than triplets, and handles overlapping relations, but suffers from error propagation and higher computational cost. ORC, focused on flexibility and generalization, primarily faces issues like the lack of explicit relation labels and scalability problems. While ORTE and ORSE are typically supervised – allowing hard matching measured with Precision, Recall, F1 score – ORC is unsupervised and necessitates other metrics. Before the advent of LLMs. literature proposed metrics for coreference resolution, such as B^3 [14] or cluster purity metrics such as V-measure [15]. More recently, Cosine Similarity has been proposed for evaluating the semantics of the information extracted [16], typically comparing it to the embeddings of gold-standard annotations, to perform soft matching. This metric is particularly suitable in the context of transformer models and LLMs, but also requires to be integrated by measures of coverage.

2. Scope of this work

This paper presents a preliminary attempt to exploit OIE to extract business-related information from climate change documents, in order to build a structured dataset for the analysis of poor financial sustainability in the domain of climate change projects. To do so, we (i) developed an open source prototype for LLM-based OIE named CliminvestAgent[1]; (ii) defined a schema for information extraction that includes the business model canvas, and metadata; (iii) collected a small test set of documents about projects on climate change, and (iv) prepared a human-compiled gold standard manually extracted from the test set by three annotators. Due to the LLM-based nature of the system, prioritizing models developed within the European Union is essential to proactively prevent instances of cultural hacking

[1]The system is freely available as an online Colab Notebook, released under Creative Commons license at https://colab.research.google.com/drive/1ZztbfsNuCEh78GbCOiKrVqzZPdDmTAH3?usp=sharing CliminvestAgent allows users to replicate the experiments with their own models from Huggingface. It requires the user to input their own API key.

[17], the misalignment of information and culture associated with foreign-developed LLMs. This security measure is particularly vital given the sensitivity of climate change, a topic frequently targeted by denialist narratives.

This paper seeks to address two Research Questions. RQ1: Which prompt design proves most effective for the open extraction of business-related information in the climate change domain? RQ2: Is a European model, such as Mistral, competitive in this task when compared against similar commercial models developed in the US? To answer these research questions, we performed experiments with LLMs to extract jsonl files from the selected documents, and compared them against the human-compiled gold standard. We evaluated the quality of the information extracted with different steps and metrics. First we checked jsonl conformance, discarding all the result that are not well formed, Then we computed different metrics: execution time in seconds (T); Cosine Similairity (CS), which measures the semantic distance between the information extracted, the human gold standard and the random test set; coverage (Cover), defined as

$$coverage = \frac{N_{\text{extracted}}}{N_{\text{required}}}$$

where $N_{\text{extracted}}$ is the total number of rows extracted by the model and N_{required} is the total number of required rows in the gold standard; sparsity (Spar), defined as the average proportion of missing values across all columns, or

$$sparsity = \frac{1}{N_C} \sum_{j=1}^{N_C} \left(\frac{M_j}{N_R} \right)$$

where N_R is the total number of rows, N_C the total number of columns and M_j is the count of missing values in column j. We also extracted a one-hot encoded table from the jsonl in order to compute precision (Prec), recall (Rec) and F1-measure (F1). In the next section we describe the experimental settings and report the results of the experiments.

3. Method and Experiments

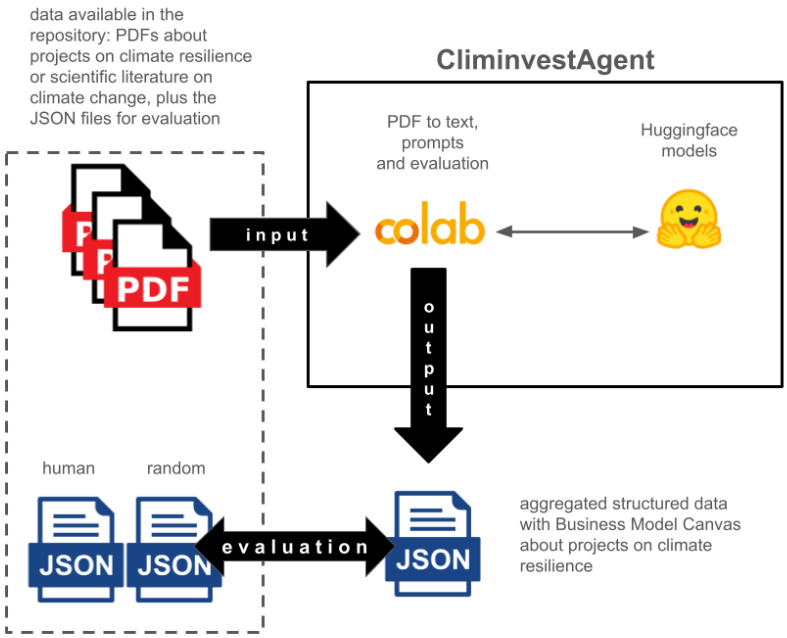

Figure 1: Schema of the system and experimental settings.

The pipeline of the experimental settings is depicted in Figure 1. To facilitate manual annotation, we chose a test set of three climate change documents totaling ≈ 25000 words. The first is a scientific review of adaptation and mitigation projects; the second is a checklist for implementing solar energy mitigation data; and the third describes tools produced within a urban climate resilience project.

We defined a schema for the extraction of data related to business processes. Our data extraction schema is designed to capture all essential components of a project's Business Model Canvas (BMC) [18], specifically applied to climate change adaptation and resilience initiatives. The structure begins with core project metadata (project, country, year, target events), then details the nine elements of the BMC, including key partners, resources, activities, customer segments, value propositions, customer relationships, channels, cost structure, and revenue streams. Beyond the traditional BMC, the schema integrates crucial financial and resilience metrics, such as the discount rate cost effectiveness (percentage of financial advantage given by the project's results) and a list of performance indicators to monitor (e.g., sea level rise, electricity consumption, etc.). Finally, the schema includes descriptive fields for project context and financial strategy: main project result, keywords, project summary, sustainability strategy, and concrete actions to reach bankability. The schema is then turned into a json structure.

We asked three annotators to manually extract information from the selected documents and fill in the schema. We used a broad definition of "project", comprising also descriptions of projects' results, such as tools and replication guidelines. Then we compared the three annotations and selected as official test set the one with the highest average pairwise cosine similarity: 0.801. This is the human annotated test set. Using the same schema, we created a second test set by filling the fields with random words extracted from the documents, and random null values. We used this test set as noise to compute the baseline. In practice, measuring the CS distance from both human-compiled and the random test sets, we are able to measure the quantity of information extracted with respect to noise. To compute the pairwise cosine similarity we used a multilingual transformer, sentence-BERT paraphrase-multilingual-MiniLM-L12-v2 [19], for the presence of Spanish text in some documents.

Model	Prompt	Trial	T	CS.n	CS.h	Cover	Spar	ΔCS	Prec	Rec	F1
Claude 3.5 Sonnett	P1	1	36	0.790	0.799	0.095	0.125	0.009	0.0000	0.0000	0.0000
		2	37	0.764	0.783	0.095	0.175	0.019	0.0678	0.0317	0.0432
		3	36	0.774	0.785	0.143	0.083	0.011	0.0143	0.0079	0.0102
	P2	1	55	0.732	0.747	0.238	0.190	0.015	0.0000	0.0000	0.0000
		2	53	0.729	0.746	0.238	0.130	0.017	0.0084	0.0159	0.0110
		3	48	0.761	0.734	0.190	0.113	-0.027	0.0000	0.0000	0.0000
GPT oss	P1	1	23	0.823	0.760	0.429	0.805	-0.063	0.0000	0.0000	0.0000
		2	21	0.864	0.815	0.333	0.333	-0.049	0.0000	0.0000	0.0000
		3	21	0.840	0.804	0.381	0.778	-0.036	0.0000	0.0000	0.0000
	P2	1	21	0.836	0.817	0.286	0.300	-0.019	0.0000	0.0000	0.0000
		2	22	0.820	0.804	0.190	0.263	-0.016	0.0000	0.0000	0.0000
		3	18	0.765	0.776	0.190	0.300	0.011	0.0000	0.0000	0.0000
Mistral 3 medium	P1	1	52	0.830	0.807	0.143	0.417	-0.023	0.0101	0.0079	0.0089
		2	52	0.830	0.807	0.095	0.375	-0.023	0.0103	0.0079	0.0090
		3	42	0.830	0.807	0.143	0.400	-0.023	0.0123	0.0079	0.0097
	P2	1	51	0.726	0.746	0.333	0.221	**0.020**	0.0168	0.0317	0.0220
		2	49	0.737	0.762	0.286	0.217	**0.025**	0.0140	0.0238	0.0176
		3	45	0.723	0.744	0.286	0.217	**0.021**	0.0127	0.0238	0.0165

Table 1
Results of the experiments measured with soft matching: CS.n (cosine similarity to noise), CS.h (cosine similarity to human-compiled data). Coverage and Sparsity were used to measure the quantity of generated lines and the frequency of missing values, respectively. ΔCS is the improvement over the baseline. Results measured with hard matching: Prec (precision), Rec (recall), F-measure (F1). Only trials resulting in well-formatted outputs were considered for this evaluation.

We developed two distinct prompts for our evaluation. Prompt 1 was manually authored using few-shot techniques and basic field descriptions. Prompt 2 was optimized using Gemini 2.0, incorporating role-play prompting and rigorous Markdown formatting instructions. To prevent overfitting given the

small size of the dataset, we utilized synthetic few-shot examples that do not appear in the test data. Due to concerns regarding cultural hacking, we selected Mistral 3 Medium, a Mixture-of-Experts model with 195 billion parameters, as our target LLMs. We compared its performances against two models of comparable size and capabilities: GPT oss 120 billion, an open weight reasoning model [20], and Claude 3.5 Sonnet, a 175 billion parameter reasoning model. We performed three runs per model.

All the data used in our experimental setting can be downloaded from the notebook environment of the CliminvestAgent. The results of the experiments, reported in Table 1, show at least three interesting findings:

1. The optimized prompt (Prompt 2) helps extracting meaningful, well formatted information. In particular with Mistral 3 medium.
2. Mistral shows good results consistency over more trials.
3. The overall performance of Mistral 3 medium in business-oriented climate change OIE is competitive against LLMs of similar size.

Answering RQ1, an automatically optimized prompt is effective to extract meaningful information. Regarding RQ2, Mistral shows a performance comparable to LLMs of similar size, and consistency across multiple trials. The comparison between Cosine Similarity and with F1-measure reveals that Mistral is consistent also when measured with hard matching.

4. Conclusion and Future

In this paper, we introduced CliminvestAgent, a novel, open-source prototype that successfully bridges the gap between unstructured climate change documentation and machine-readable financial analysis. By defining a specific Open Information Extraction schema centered around the Business Model Canvas, we demonstrated the feasibility of extracting high-value, business-related data with the use of Large Language Models.

Crucially, our work moves beyond standard evaluation, proposing a robust framework incorporating specialized metrics: Cosine Similarity, coverage, and data sparsity. The findings from our investigation into various models (Mistral, Claude 3.5, and GPT OSS 120b) and prompting strategies provide critical insights into model selection and prompt design efficacy for generating reliable structured output, demonstrating that Mistral can be employed for this task. This fact has the potential for mitigating the risks of cultural hacking for the European Union in this sensitive domain. Still, coverage is poor, and sparsity, despite reduced by the optimized prompt, indicates that some information we are searching for was not present in the initial data. Future research will focus performing OIE on a large set of climate-change related documents, beside end-to-end dashboards for business intelligence, to create a comprehensive, automated tool for climate-related project investment decision-making.

Acknowledgments

This research was supported by the European Commission, grant 10121294: Bankable by Design, Continuous, and Predictive Climate Adaptation Investments with Co-Benefits - CLIMINVEST. Artificial intelligence has been used for minor editing and text improvement.

Conflict of Interest

The authors declare no conflict of interest in the content of this paper.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the

content as needed and take(s) full responsibility for the publication's content.

References

[1] R. Rojas, L. Feyen, P. Watkiss, Climate change and river floods in the european union: Socio-economic consequences and the costs and benefits of adaptation, Global Environmental Change 23 (2013) 1737–1751.

[2] L. C. Pouchard, M. L. Branstetter, R. B. Cook, R. Devarakonda, J. Green, G. Palanisamy, P. Alexander, N. F. Noy, A linked science investigation: enhancing climate change data discovery with semantic technologies, Earth science informatics 6 (2013) 175–185.

[3] Z. Hong, L. Ward, K. Chard, B. Blaiszik, I. Foster, Challenges and advances in information extraction from scientific literature: a review, JOM 73 (2021) 3383–3400.

[4] S. Marchesin, G. Silvello, O. Alonso, Large language models and data quality for knowledge graphs, Information Processing & Management 62 (2025) 104281.

[5] C. Zhu, M. Chen, C. Fan, G. Cheng, Y. Zhang, Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 4732–4740.

[6] H. Pan, M. Adamu, Q. Zhang, E. Dragut, L. J. Latecki, Climateie: A dataset for climate science information extraction, in: Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), 2025, pp. 76–98.

[7] A. Poleksić, S. Martinčić-Ipšić, Towards dataset for extracting relations in the climate-change domain, in: Proceedings of the 3rd International workshop one knowledge graph generation from text (TEXT2KG) and Data Quality meets Machine Learning and Knowledge Graphs (DQMLKG) co-located with the Extended Semantic Web Conference (ESWC 2024), Heraklion: CEUR, 2024.

[8] Z. Jin, C. Zhang, Z. Hu, J. Yu, R. Ma, Q. Chen, X. Liao, Y. Zhang, Cycleoie: A low-resource training framework for open information extraction, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 3372–3390.

[9] L. Guo, J. Sanz-Cruzado, R. Mccreadie, Selecting the Right Experts: Generalizing Information Extraction for Unseen Scenarios via Task-Aware Expert Weighting, Technical Report, 2025.

[10] G. Blanchy, L. Albrecht, J. Koestel, S. Garré, Potential of natural language processing for metadata extraction from environmental scientific publications, Soil 9 (2023) 155–168.

[11] O. Alonso, R. Baeza-Yates, Information Retrieval: Advanced Topics and Techniques, ACM, 2024.

[12] N. Gandhi, T. Corringham, E. Strubell, Challenges in end-to-end policy extraction from climate action plans, in: Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), 2024, pp. 156–167.

[13] L. Pai, W. Gao, W. Dong, L. Ai, Z. Gong, S. Huang, L. Zongsheng, E. Hoque, J. Hirschberg, Y. Zhang, A survey on open information extraction from rule-based model to large language model, Findings of the association for computational linguistics: EMNLP 2024 (2024) 9586–9608.

[14] B. Lin, R. Shah, R. Frederking, A. Gershman, Cone: Metrics for automatic evaluation of named entity co-reference resolution, in: Proceedings of the 2010 Named Entities Workshop, 2010, pp. 136–144.

[15] A. Rosenberg, J. Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure, in: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 410–420.

[16] W. Léchelle, P. Langlais, Revisiting the task of scoring open ie relations, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[17] F. Celli, A. Samal, Monitoring historical cultural hacking in large language models (2025).

[18] A. Murray, V. Scuotto, The business model canvas, Symphonya. Emerging Issues in Management (2015) 94–109.

[19] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks,

in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[20] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, et al., gpt-oss-120b & gpt-oss-20b model card, 2025.