LLM-Assisted Resource-Aware Weakly Supervised Multimodal Extraction of Climate-Risk Signals from ESG Documents

Panhapin Theang[1] and Nimol Thuon[1,2⋆]

[1] Université Paris Cité, 75013 Paris, France
[2] University of Science and Technology of China, 230052 Hefei, China
panhapin.theang@etu.u-paris.fr,nimol.thuon@gmail.com

Abstract. Climate-risk evidence in ESG reports is often distributed across narrative text, KPI tables, targets, and governance disclosures, making block-level extraction difficult under limited annotation. We propose a resource-aware framework that combines rule-based weak supervision, selective multimodal teacher refinement, and compact student models. ESG PDFs are parsed into a report–page–block corpus, and initial labels are generated using keyword and heuristic rules. Qwen2.5-VL reviews only 1,000 ambiguous blocks using the page image and block text; accepted outputs are merged with a manually corrected gold subset. We compare TF-IDF and DistilBERT under weak-only, weak+gold, weak+LLM, and weak+LLM+ gold supervision using a company-level gold-only test set. TF-IDF is a strong low-cost baseline, whereas DistilBERT benefits most from improved supervision. The best weak+LLM +gold DistilBERT model reaches 0.8694 macro-F1 on 102 gold test blocks, compared with 0.8459 for TF-IDF. The results indicate that multimodal LLMs are most useful as selective supervision refiners, while compact models remain suitable for efficient deployment.

Keywords: ESG document analysis · climate-risk extraction · weak supervision · multimodal VLM · resource-aware learning

1 Introduction

Climate-risk disclosure is increasingly important in ESG and sustainability reporting. Frameworks such as TCFD, IFRS S2, and the Greenhouse Gas Protocol encourage companies to report governance, strategy, risk management, emissions, and targets [4, 1, 10, 6, 8]. However, relevant evidence is rarely presented in a standardized form. It is distributed across long reports containing paragraphs, tables, KPI summaries, targets, and governance sections, making climate-risk extraction a document-intelligence problem rather than a conventional document-level classification task.

A single ESG page may contain emissions values, physical-risk statements, and governance disclosures. Flattening such pages into plain text can remove

⋆ Corresponding author

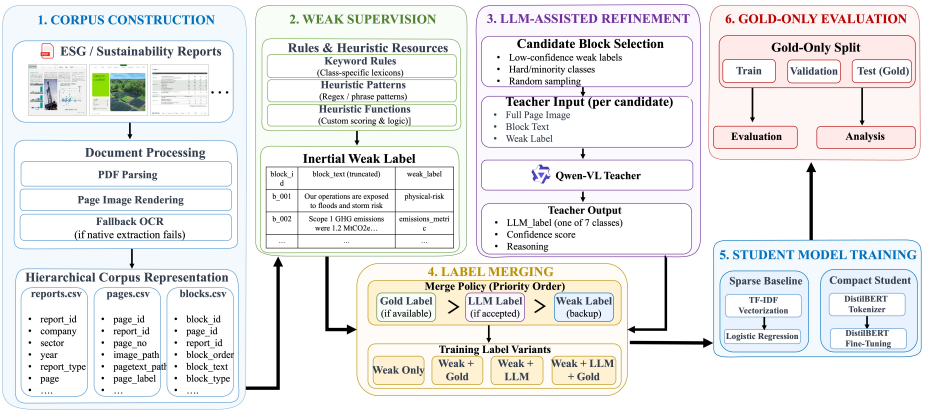

Fig. 1. Proposed framework. Weak labels are selectively refined by a multimodal teacher, merged with corrected gold labels, and used to train efficient student models evaluated on a gold-only split.

useful visual and layout context, while fully manual block-level annotation is expensive. Prior document models show the value of combining text, layout, and visual information [18, 17, 7], but deploying large multimodal models across complete report collections remains costly.

We therefore investigate whether a multimodal LLM can be used selectively during supervision construction rather than during inference. As shown in Fig. 1, the pipeline parses ESG reports into report, page, and block units; assigns weak labels using rules; asks a vision-language teacher to review ambiguous blocks; and trains sparse or compact student models on the resulting labels. This follows weak-supervision principles in which scalable but noisy labeling functions are improved using limited high-quality supervision [13, 11, 12, 15]. The task contains seven classes: physical risk, transition risk, emissions metric, target commitment, adaptation/mitigation action, governance/risk management, and non-climate. We compare TF-IDF with logistic regression and DistilBERT under four supervision settings. The contributions of this paper are as follows:

1. We present a practical end-to-end framework for block-level climate-risk signal extraction from ESG documents, combining corpus construction, weak supervision, multimodal teacher refinement, and efficient student training.
2. We construct a hierarchical supervision resource at the report, page, and block levels, supporting weak labels, corrected gold labels, and LLM-assisted label refinement.
3. We provide a controlled comparison of sparse, compact transformer, and LLM-assisted student settings under a company-level gold-only evaluation protocol.
4. We provide pilot evidence that selective multimodal teacher assistance is most effective as a supervision refinement mechanism, enabling a compact

student to improve over strong sparse baselines while preserving resource-aware deployment.

2 Related Work

Research on ESG analytics has expanded rapidly, particularly in document classification, retrieval, question answering, and long-context report understanding. Domain-specific models such as ClimateBERT demonstrate the benefit of adapting language representations to climate terminology and disclosure patterns [16, 3]. Recent ESG benchmarks further reflect growing interest in large-language-model understanding of sustainability reports [5, 19].

Visually structured ESG reports present challenges beyond plain-text classification because paragraphs, tables, KPI boxes, captions, and headings often appear on the same page. Relevant climate-risk information may therefore depend on both textual and visual context. LayoutLM, LayoutLMv2, and LayoutLMv3 demonstrate that jointly modeling text, layout, and visual information improves visually rich document understanding [18, 17, 7].

Weak supervision reduces annotation costs by generating labels from noisy rules and heuristics. Data programming combines such signals into scalable training resources [13, 11], while later work extends this approach through multi-task supervision and automated labeling-signal discovery [12, 15]. Building on these directions, our framework combines rule-based weak labels with a small corrected gold subset and selective multimodal refinement, where the teacher reviews only uncertain or difficult blocks.

3 Corpus and Supervision

3.1 Corpus Construction

The corpus contains 51 publicly available ESG, sustainability, climate, impact, and selected annual reports with substantial environmental disclosure content. Each PDF is organized hierarchically at the report, page, and block levels. Native PDF parsing is used as the primary text-extraction method because many ESG reports contain embedded text layers. Each page is also rendered as an image, while OCR is used only when native text extraction fails or when a page is image-only.

The final training set contains 63,932 blocks. The manually corrected validation and test sets contain 115 and 102 blocks, respectively. To reduce leakage caused by repeated corporate terminology and document templates, all reports and blocks from the same company are assigned to a single split. Table 1 summarizes the corpus size and supervision resources used in the experiments.

As shown in Table 1, manual correction is limited to relatively small validation and test subsets, whereas weak supervision is applied to the substantially larger training collection. Multimodal review is further restricted to 1,000 selected candidates.

Table 1. Summary of the constructed ESG climate-risk corpus and supervision resources.

Item	Count / Description
ESG-related PDF reports	51
Document type	ESG-related corporate reports
Corpus organization	Report–page–block hierarchy
Training blocks	63,932
Gold validation blocks	115
Gold test blocks	102
Target classes	7
LLM-reviewed candidate blocks	1,000
Primary text extraction	Native PDF parsing
Visual context	Full-page rendered images
Fallback extraction	OCR only when native extraction fails
Evaluation protocol	Company-level gold-only split

3.2 Label Space and Weak Supervision

The task contains seven classes: physical risk, transition risk, emissions metric, target commitment, adaptation/mitigation action, governance/risk management, and non-climate. The gold test set contains 11, 10, 12, 18, 21, 10, and 20 instances, respectively. Because the class distribution is uneven, macro-F1 is the primary metric.

Weak labels are generated using class-specific keywords and phrase patterns. Emissions terms such as *Scope 1–3* and CO_2e indicate emissions metrics; target years and phrases such as *net zero* indicate commitments; and terms such as flood, drought, wildfire, and extreme heat indicate physical risk. Similar rules are used for the remaining classes. Although scalable, these rules may be unreliable when terms are ambiguous or context-dependent.

3.3 Gold and LLM-Assisted Refinement

The manually corrected subset provides reliable validation and test labels while also contributing a limited amount of high-quality supervision during training. The multimodal teacher reviews 1,000 candidate blocks selected from low-confidence weak labels and difficult categories, particularly physical risk, transition risk, and adaptation/mitigation action. For each candidate, the teacher receives the full-page image, the target block text, and the current weak label, and returns a refined class together with a confidence score.

Teacher outputs are not treated as ground truth. Accepted predictions are merged according to the following priority rule:

$$\text{gold label} > \text{accepted LLM label} > \text{weak label}. \tag{1}$$

This process produces four supervision regimes: weak-only, weak+gold, weak+LLM, and weak+LLM+gold.

4 Proposed Framework

The proposed framework combines scalable weak supervision with selective multimodal label refinement and compact student training. The main design principle is resource-aware: large multimodal models are used only during supervision construction, while final prediction is performed by lightweight student models. This design follows the broader goal of reducing deployment-time cost while still benefiting from stronger models during data construction.

4.1 LLM-Assisted Multimodal Refinement

The multimodal teacher is applied only to a restricted candidate set rather than to the full corpus. Candidate blocks are selected from difficult or ambiguous cases, including low-confidence weak labels and hard classes such as `physical_risk`, `transition_risk`, and `adaptation/mitigation_action`. For each candidate, the teacher receives the full page image as visual context, the target block text as the local semantic unit, and the current weak label as a reference. This design allows the teacher to use layout and visual context while still focusing on the intended block. We use Qwen2.5-VL as the multimodal teacher because recent vision-language models have shown strong ability to process combined image-text inputs and document-like visual content [2].

The teacher output is not treated as perfect supervision. The teacher-label analysis in Section 6.2 shows that it changes a large fraction of weak labels, providing a rich but potentially noisy alternative signal. For this reason, LLM labels are most effective when merged with corrected gold labels rather than used alone.

4.2 Student Models

We evaluate two student model families. The first is a sparse lexical baseline using TF-IDF vectorization followed by logistic regression, a standard and efficient baseline for text classification implemented using scikit-learn [9]. This model is computationally efficient and strong when label distinctions are dominated by repeated lexical cues. The second is a compact transformer student based on DistilBERT [14]. DistilBERT provides contextual representation while remaining lightweight enough for resource-aware training and deployment.

The central question is whether improved supervision from gold correction and LLM-assisted refinement enables a compact student model to improve over a strong sparse baseline without requiring a large multimodal model at inference time.

5 Experimental Setup

5.1 Task and Data Splits

We evaluate multiclass block-level classification over the seven labels defined in Section 3.2. Each block receives exactly one label. All models are evaluated

Table 2. Experimental settings for student-model training.

ID	Student model	Label source	Input
E1	TF-IDF + LR	Weak	Text
E2	TF-IDF + LR	Weak + gold	Text
E3	TF-IDF + LR	Weak + LLM	Text
E4	TF-IDF + LR	Weak + LLM + gold	Text
E5	DistilBERT	Weak	Text
E6	DistilBERT	Weak + gold	Text
E7	DistilBERT	Weak + LLM	Text
E8	DistilBERT	Weak + LLM + gold	Text

on the same company-level gold-only test set. The company-level split prevents leakage from repeated corporate phrasing, because all reports and blocks from the same company are assigned to the same split.

Validation and test sets contain only corrected gold labels. Training labels vary by experiment and may use weak labels, weak+gold labels, weak+LLM labels, or weak+LLM+gold labels. This protocol ensures that all model comparisons are made against the same clean reference standard.

5.2 Models and Training Settings

Table 2 summarizes the eight student-model experiments obtained by combining two model families with four supervision regimes. The LLM-assisted teacher uses multimodal input during supervision construction, but all student models are trained and evaluated using block text. This design isolates the effect of supervision quality from the cost of deploying a multimodal model at inference time.

As indicated in Table 2, the input remains fixed across all experiments; only the student architecture and training-label source vary. For the sparse baseline, we use TF-IDF vectorization with logistic regression and class-balanced training. For the compact transformer, we fine-tune DistilBERT with cross-entropy supervision. We use a maximum sequence length of 256, batch size 16, AdamW optimization, and three training epochs. Validation-based model selection is performed using macro-F1. In our final DistilBERT runs, three epochs provided better generalization than longer schedules.

5.3 Evaluation Metrics

We report accuracy, macro-F1, and weighted-F1. Accuracy measures overall correctness, weighted-F1 reflects class-frequency-weighted behavior, and macro-F1 emphasizes balanced performance across classes.

Table 3. Results on the company-level pilot gold-only test set.

Model	Labels	Acc.	Macro-F1	W-F1
TF-IDF	Weak	0.8431	0.8459	0.8466
TF-IDF	Weak+gold	0.8431	0.8459	0.8466
TF-IDF	Weak+LLM	0.8333	0.8414	0.8405
TF-IDF	Weak+LLM+gold	0.8431	0.8459	0.8466
DistilBERT	Weak	0.8333	0.8396	0.8372
DistilBERT	Weak+gold	0.8431	0.8478	0.8456
DistilBERT	Weak+LLM	0.8431	0.8426	0.8461
DistilBERT	**Weak+LLM+gold**	**0.8627**	**0.8694**	**0.8665**

6 Results and Analysis

6.1 Main Results

Table 3 reports the main results. TF-IDF is a strong sparse baseline, achieving 0.8459 macro-F1 with weak-only supervision. Adding corrected or LLM-refined labels does not improve the sparse model, suggesting that its performance is dominated by stable lexical regularities. DistilBERT trained only on weak labels performs below TF-IDF, but improves once corrected gold labels are introduced. The best result is obtained by DistilBERT trained with weak+LLM+gold labels, reaching 0.8694 macro-F1, 0.8627 accuracy, and 0.8665 weighted-F1.

The comparison in Table 3 supports the paper's central hypothesis. LLM labels alone do not improve DistilBERT and may introduce noise when used without correction. However, combining weak, LLM-assisted, and gold supervision yields the strongest compact student. Because the test set contains only 102 corrected blocks, the improvement should be interpreted as pilot evidence rather than a definitive benchmark.

6.2 Teacher and Per-Class Analysis

The teacher returned valid labels for all 1,000 reviewed candidates and changed 601 weak labels, corresponding to a 60.1% change rate. This indicates that the teacher contributes a substantially different supervision signal but also confirms that its predictions should not be accepted blindly. The teacher was applied to approximately 1.6% of the 63,932 training blocks, limiting multimodal computation to a small selected subset.

Table 4 complements the aggregate results in Table 3 by showing the per-class behavior of the best student. Emissions metric, target commitment, and governance/risk management achieve high F1 because they often contain distinctive units, dates, or board-level terminology. Physical risk achieves perfect recall but lower precision, indicating over-prediction. Adaptation/mitigation action has lower recall and may be confused with targets, transition-risk language, or general sustainability actions.

These findings support a resource-aware division of labor. TF-IDF provides a low-cost reference, Qwen2.5-VL reviews selected ambiguous labels during corpus

Table 4. Per-class results for the best weak+LLM+gold DistilBERT model.

Class	P	R	F1	N
Physical risk	0.5789	1.0000	0.7333	11
Transition risk	0.8182	0.9000	0.8571	10
Emissions metric	1.0000	0.9167	0.9565	12
Target commitment	1.0000	0.8333	0.9091	18
Adapt./mitig. action	0.9333	0.6667	0.7778	21
Governance/risk mgmt.	1.0000	0.9000	0.9474	10
Non-climate	0.8636	0.9500	0.9048	20

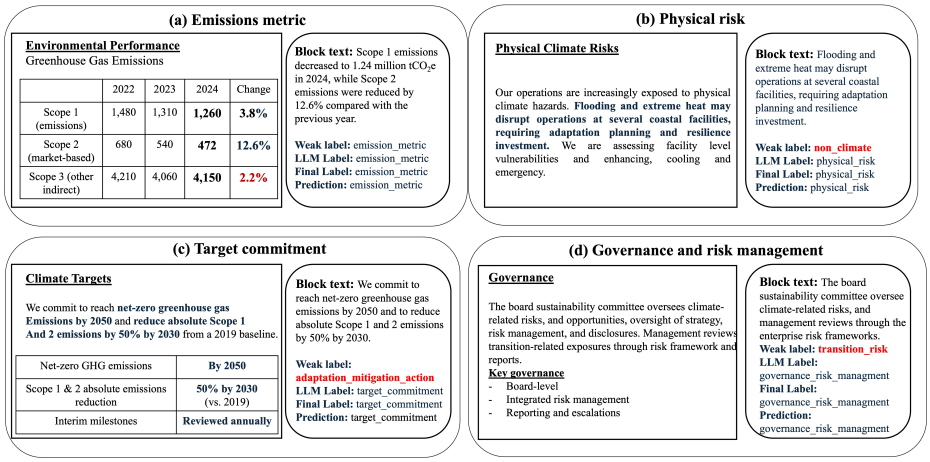

Fig. 2. Qualitative examples of LLM-assisted label refinement. Each panel shows an ESG page snippet, a short target block excerpt, and the label transition from weak supervision to LLM refinement, final merged label, and student prediction. Examples (b)–(d) illustrate cases where the multimodal teacher corrects noisy or ambiguous weak labels.

construction, and DistilBERT performs deployment-time prediction. The large multimodal model is therefore excluded from routine inference.

6.3 Qualitative Analysis

Figure 2 provides qualitative evidence for the trends reported in Tables 3 and 4. The examples include both stable cases, where the weak label and LLM label agree, and correction cases, where the teacher changes an ambiguous weak label into a more appropriate final label. For example, physical-risk language involving flooding and extreme heat may be missed by weak rules, while target commitments and governance statements may be confused with broader action or transition-risk language. These examples help explain why LLM-assisted refinement can improve the training signal for the compact student.

7 Discussion and Conclusion

The experiments yield three main findings. First, Table 3 shows that weak supervision alone does not enable DistilBERT to outperform the strong TF-IDF baseline. Second, corrected gold labels benefit DistilBERT more than TF-IDF, indicating that contextual models are more sensitive to supervision quality. Third, Table 3 and Fig. 2 show that multimodal teacher assistance is most effective when combined with corrected gold labels rather than used alone. These findings support using large multimodal models for selective supervision refinement while deploying compact students for efficient inference.

This study is limited by the small gold-only test set of 102 corrected blocks, the limited range of companies and sectors, and the use of text-based student models despite the teacher's multimodal page context. Future work should expand the corpus and gold evaluation set and investigate compact layout-aware students.

In conclusion, this paper presented an LLM-assisted, resource-aware, weakly supervised framework for block-level climate-risk extraction from ESG documents. The framework combines native PDF processing, rule-based weak supervision, selective multimodal refinement, and compact student training under a company-level gold-only protocol. TF-IDF remains a strong sparse baseline, but combining corrected gold labels with selective LLM refinement enables DistilBERT to achieve the best overall performance. This provides a practical strategy for ESG monitoring: use multimodal models selectively to improve supervision and compact models for scalable deployment.

References

1. Baboukardos, D., Seretis, E., Slack, R., Tsalavoutas, Y., Tsoligkas, F.: Companies' readiness to adopt ifrs s2 climate-related disclosures (2022)
2. Bai, S., Chen, K., Liu, X., Wang, J., et al.: Qwen2.5-vl technical report (2025)
3. Bingler, J.A., Kraus, M., Leippold, M., Webersinke, N.: Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. Finance Research Letters (2022). https://doi.org/10.1016/j.frl.2022.102776
4. Eccles, R.G., Krzus, M.P.: Implementing the task force on climate-related financial disclosures recommendations: An assessment of corporate readiness. Schmalenbach Business Review **71**(2), 287–293 (2019). https://doi.org/10.1007/s41464-018-0060-4
5. He, C., Zhou, X., Wu, Y., Yu, X., Zhang, Y., et al.: Esgenius: Benchmarking llms on environmental, social, and governance (esg) and sustainability knowledge. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 14623–14664 (2025). https://doi.org/10.18653/v1/2025.emnlp-main.739
6. Hettler, M., Graf-Vlachy, L.: Corporate scope 3 carbon emission reporting as an enabler of supply chain decarbonization: A systematic review and comprehensive research agenda. Tech. Rep. 2 (2024). https://doi.org/10.1002/bse.3486
7. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of

the 30th ACM International Conference on Multimedia. p. 4083–4091. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548112
8. Palea, V., Drogo, F.: Carbon emissions and the cost of debt in the eurozone: The role of public policies, climate-related disclosure and corporate governance. Business Strategy and the Environment **29**(8), 2953–2972 (2020). https://doi.org/10.1002/bse.2550
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**(null), 2825–2830 (2011). https://doi.org/10.5555/1953048.2078195
10. Protocol, G.G.: Greenhouse gas protocol. Tech. rep. (2011)
11. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: rapid training data creation with weak supervision. Proc. VLDB Endow. **11**(3), 269–282 (2017). https://doi.org/10.14778/3157794.3157797
12. Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., Ré, C.: Training complex models with multi-task weak supervision. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 4763–4771 (2019)
13. Ratner, A., Sa, C.D., Wu, S., Selsam, D., Ré, C.: Data programming: creating large training sets, quickly. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 3574–3582. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
14. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter (2019)
15. Varma, P., Ré, C.: Snuba: automating weak supervision to label training data. p. 223–236. VLDB Endowment (2018). https://doi.org/10.14778/3291264.3291268
16. Webersinke, N., Kraus, M., Bingler, J.A., Leippold, M.: ClimateBERT: A pretrained language model for climate-related text (2021)
17. Xu, Y., Xu, Y., Lv, T., et al.: LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2579–2591. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.201
18. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 1192–1200. KDD '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3403172
19. Zhang, L., Zhou, X., He, C., Wang, D., Wu, Y., Xu, H., Liu, W., Miao, C.: Mmesgbench: Pioneering multimodal understanding and complex reasoning benchmark for esg tasks (2025). https://doi.org/10.1145/3746027.3758225