Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CarbonCrediBERT: Expert-Guided Few-Shot and Selective Classification of Carbon Credit Disclosures

Hang Thanh Bui

hangtbngoaigiao@gmail.com

University of New South Wales Canberra at ADFA https://orcid.org/0000-0001-6851-7717

Eric Lee

Vercentra

Megan C Evans

University of New South Wales Canberra https://orcid.org/0000-0001-6763-310X

Research Article

Keywords: carbon credits, corporate disclosure, expert-guided learning, few-shot classification, selective classification, hard negatives, ASX100

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11679188/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

CarbonCrediBERT: Expert-Guided Few-Shot and Selective Classification of Carbon Credit Disclosures

Hang Thanh Bui[1,2*], Eric Lee[3] and Megan C. Evans[1,4]

[1]School of Business, The University of New South Wales Canberra, Canberra, Australia.
[2]The Australia Institute, Canberra, Australia.
[3]Vercentra, Melbourne, Victoria, Australia.
[4]Institute for Climate Risk and Response, The University of New South Wales, Kensington, New South Wales, Australia.

*Corresponding author(s). E-mail(s): hangthanhbui@unsw.edu.au;

Abstract

Corporate reports increasingly discuss carbon credits in relation to net-zero targets, offsetting claims, and climate transition plans, particularly as climate-related financial disclosure requirements expand. Analysing these disclosures at scale remains difficult because terms such as *credit*, *offset*, and *net* also occur in unrelated financial, accounting, and operational contexts. Expert-labelled data are also costly to produce and often contain substantial class imbalance. We propose CarbonCrediBERT, an expert-guided natural language processing framework that classifies paragraphs according to carbon credit relevance, sentiment, and reported credit use, while referring uncertain cases for human review. CarbonCrediBERT combines an expert-defined taxonomy, hard-negative examples drawn from closely related but non-relevant passages, and classification methods suited to small expert-labelled datasets. Using annual, climate, sustainability, and ESG reports published by ASX100 companies in 2024–2025, the framework produced 8,299 keyword-matched sentence records and a balanced candidate pool of 15,026 paragraphs. We evaluated ClimateBERT and DistilRoBERTa fine-tuning and SetFit–MiniLM few-shot learning using three task-specific datasets of 100 expert-labelled paragraphs each. Five-fold development cross-validation selected fine-tuned ClimateBERT for carbon credit relevance, achieving 93.8% accuracy, and SetFit–MiniLM for sentiment and reported credit use. These results show that broad climate-language detection

alone is insufficient for identifying carbon credit disclosures and that no single modelling approach performs best across all three tasks. They also highlight the continuing importance of expert judgement, particularly for sentiment and ambiguous statements about credit use. Future work will expand the domain corpus and expert-labelled evidence and evaluate transfer across companies, reporting years, and jurisdictions, establishing a pathway towards a reusable CarbonCrediBERT foundation model for carbon credit disclosure analysis.

Keywords: carbon credits, corporate disclosure, expert-guided learning, few-shot classification, selective classification, hard negatives, ASX100

1 Introduction

Corporate climate disclosure is moving from a predominantly voluntary reporting regime towards mandatory financial-market reporting and assurance. The Task Force on Climate-related Financial Disclosures (TCFD) helped establish an investor-oriented structure covering governance, strategy, risk management, metrics and targets [1]. The European Union subsequently introduced the Corporate Sustainability Reporting Directive and the European Sustainability Reporting Standards, while Australia began phasing in mandatory climate-related financial disclosure from 1 January 2025 [2–4]. As climate reporting expands, investors, regulators, and researchers must sift through lengthy and varied corporate reports to identify information relevant to their decisions. The time and expertise required make systematic comparison across companies difficult, motivating the use of automated text analysis. Recent research demonstrates the potential of fine-tuned language models to identify climate-related content in financial and non-financial disclosures [5]. Carbon credit-related analysis requires distinguishing statements about credit use from broader discussions of climate targets and emissions reductions, which do not necessarily involve carbon credits. For example, a company might state that it reduced emissions by switching to renewable electricity. This describes an emissions-reduction measure, whereas a statement that the company purchased and retired carbon credits to compensate for its remaining emissions explicitly reports credit use.

Carbon credit claims are a subset of climate-related disclosure in general, but the widespread use of carbon credits by corporates and ongoing concerns regarding their credibility [6, 7] warrant close scrutiny. Carbon credits can be acquired, generated, held, sold, retired, surrendered for compliance, or used to support a voluntary emissions reduction claim. Their interpretation can therefore depend on accounting treatment, compliance obligations, registry records, certification standards, project attributes, and the identity of the party using the credit. Yet studies report inconsistent recognition, measurement, presentation, and disclosure of carbon credits and certified emission reductions [8–10]. Evidence from 100 ASX-listed entities shows that carbon credits and offsets are increasingly visible in corporate reports, while important details about their acquisition, type, location, accounting treatment, and use are

often missing [11]. Consequently, a reference to carbon credits does not necessarily reveal what a company has done, intends to do, or claims to have achieved.

This ambiguity creates a practical problem for investors, policymakers, and other report users. Existing studies combine keyword searches with manual review to examine corporate carbon credit disclosures [10–12]. Relevant information is not always expressed using the terms *carbon credit* or *carbon offset*. Projects that generate credits may also be discussed under broader descriptions such as *natural climate solutions* or *nature-based solutions* [13]. However, these activities do not necessarily involve credits, while words such as *credit* and *offset* also occur in unrelated financial and accounting discussions. Expanding the search vocabulary can therefore retrieve additional relevant passages, but it also introduces passages that require expert interpretation to establish their relevance. Even when a passage concerns carbon credits, further interpretation may be needed to distinguish planned use from activity reported as having occurred [10, 11]. These judgements are difficult to automate through keyword rules alone and time-consuming to perform manually across companies and reporting years.

Climate-domain natural language processing (NLP) offers a starting point for learning these contextual distinctions. Existing models address climate relevance, disclosure quality, sentiment, and net-zero commitments [14–17]. ClimateBERT, for example, uses expert-labelled paragraphs to adapt models to climate-related classification tasks [14]. Carbon credit analysis requires similarly informed supervision, but with a more specific definition of relevance. A model must distinguish passages about credits from broader climate discussions, including those that contain familiar terms but do not concern credits. Such difficult non-relevant examples are known as *hard negatives*. In research on hard-negative learning, which motivates their inclusion [18], experts determine which passages fall outside the intended category. Including these examples in training and evaluation helps assess whether a classifier distinguishes carbon credit meaning from merely similar wording.

Learning these distinctions presents two further challenges. First, specialist annotation is costly, so methods must make effective use of limited labelled data. Few-shot approaches such as SetFit are designed for this setting [19], but their suitability for carbon credit tasks requires comparison with conventional fine-tuning. Second, classification errors remain possible even after expert-guided training. Selective classification addresses this problem by allowing a model to withhold a prediction when its confidence falls below a chosen threshold, referring the passage for human review [20]. Model confidence does not itself establish correctness, however, so its usefulness for identifying errors must also be evaluated. These established approaches provide useful components, but leave a task-specific question about how to combine expert interpretation, learning from limited labels, and selective review for carbon credit disclosure analysis.

We address these challenges through CarbonCrediBERT, an expert-guided framework for classifying paragraphs by carbon credit relevance, sentiment, and reported credit use. The study uses annual, climate, sustainability, and environmental, social and governance (ESG) reports from ASX100 companies published in 2024–2025. Experts inform the search taxonomy, define the three classification tasks, and label examples that distinguish relevant disclosures from superficially similar language.

Using these expert-labelled datasets, we compare climate-specific and general-purpose Transformer models with a few-shot learning approach, then assess whether confidence thresholds can identify predictions suitable for automatic acceptance. CarbonCrediBERT's contribution is the design and pilot evaluation of this connected process, examining both what models can learn from limited expert knowledge and when their predictions still require human review. The research addresses the following questions:

RQ1: Which NLP training strategies perform best for carbon credit disclosure classification when only a small expert-labelled dataset is available?

RQ2: When can the framework accept a prediction automatically, and when is the evidence too limited to avoid expert review?

RQ3: How are carbon credit-related claims distributed across ASX100 climate, sustainability, and ESG reports, and which taxonomy categories change between the 2024 and 2025 publication-year corpora?

Contributions.

This paper contributes to research on carbon credit disclosure and its automated analysis in five ways:

- **A carbon credit-related analytical framework.** Building on climate-text classification and net-zero target detection [14, 17], we develop CarbonCrediBERT as a conceptual and computational framework for three expert-defined tasks: carbon credit relevance, sentiment, and reported credit use. The framework connects the identification of relevant passages with the interpretation of their content and the assessment of when model predictions require expert review.
- **An expert-guided taxonomy and reproducible corpus workflow.** We build on keyword-assisted reviews of corporate carbon credit disclosures [10, 11] by linking an expert-defined taxonomy, paragraph extraction, source metadata, and expert labels within a reproducible ASX100 workflow. Drawing on the concept of hard negatives [18], the annotation design explicitly identifies financial, transition, and project-method passages that resemble carbon credit disclosures but are judged non-relevant by experts. This provides examples for learning and evaluating contextual distinctions rather than treating keyword matches as relevance labels.
- **Comparative evidence on learning from limited expert labels.** Prior research demonstrates the potential of ClimateBERT fine-tuning for climate-text detection [5], while SetFit offers an alternative designed for small labelled datasets [19]. We compare ClimateBERT and general-purpose DistilRoBERTa full fine-tuning with SetFit–MiniLM using 100 expert-labelled paragraphs per task. A common five-fold development procedure supports task-specific model selection, followed by refitting and fixed-test evaluation. This comparison examines whether the benefits of climate-domain pretraining or few-shot learning differ across relevance, sentiment, and credit use.
- **An empirical assessment of when to refer predictions for review.** Drawing on selective classification [20], we evaluate whether confidence thresholds can identify carbon credit classifications suitable for automatic acceptance. Thresholds are

selected using development predictions and assessed on the fixed test sets, measuring both the proportion of predictions accepted and the errors among them. The contribution is evidence about the usefulness and limitations of confidence-based referral under scarce expert supervision, rather than an assumption that confident predictions are reliable.

- **Descriptive evidence on ASX100 disclosure language.** Complementing studies of carbon credit reporting in Australia and New Zealand [10, 11], we compare keyword families, expert-defined taxonomy categories, and report characteristics across the 2024 and 2025 publication-year corpora. This analysis describes which forms of carbon credit-related language are prominent and how their distribution differs between the available collections, while distinguishing references to projects or methods from statements of credit use.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 presents the conceptual framework and evaluation design. Section 4 describes the corpus and reports the controlled benchmark, selective-classification results, and taxonomy-guided publication-year analysis. Sections 5–8 discuss limitations, implications, future work, and conclusions.

2 Related Work

CarbonCrediBERT intersects three research streams, which are: (i) corporate climate and emissions disclosure and its capital-market effects; (ii) carbon credit accounting and corporate claims; and (iii) computational text analysis with limited expert labels. Each stream addresses part of the problem, but none provides an integrated method for finding, interpreting, and selectively reviewing carbon credit-related evidence in corporate reports.

2.1 From climate disclosure to carbon credit evidence

The climate-related disclosure literature establishes that corporate climate information can affect risk assessment, valuation, cost of capital, and audit effort [21]. It also identifies persistent variation in disclosure constructs, measurement, transparency, and comparability [22, 23]. Australia's transition to mandatory climate-related reporting makes these concerns more immediate. Climate-related risks may be material to financial statements even when their numerical effects remain uncertain, and entities vary in their preparedness for reporting and assurance [2–4, 24]. This literature highlights the need for methods that can analyse large collections of corporate reports and help reviewers locate relevant information more efficiently. However, it generally examines climate or emissions disclosure broadly rather than carbon credit disclosures specifically.

Carbon credit research shows why this narrower domain matters. Credits may be held as assets, generated for sale, or acquired to meet compliance obligations or support voluntary climate claims. Studies document substantial variation in their recognition, measurement, presentation, and disclosure [8, 9]. Corporate claims introduce further distinctions between present use and future intended use, offsetting and non-offsetting

use, and net-zero and carbon-neutral claims [25]. Australian evidence shows that corporate reports may mention carbon credits without explaining whether they were purchased, generated, held, sold, retired, surrendered, or used for compliance or a public claim [11]. These distinctions matter under the Australian Sustainability Reporting Standard (AASB S2) , which requires entities to disclose their planned use of carbon credits to offset greenhouse gas emissions to achieve any net emissions target, in addition to [26]:

- the extent to which, and how, achieving any net greenhouse gas emissions target relies on the use of carbon credits;
- which third-party scheme(s) will verify or certify the carbon credits;
- the type of carbon credit, including whether it is nature-based, technological, and whether the offset is achieved via emissions reduction or removal; and
- any other factors necessary to understand the credibility and integrity of the carbon credits the entity plans to use.

AASB S2 is modeled on the International Sustainability Standards Board (ISSB) Climate standard (IFRS S2), so disclosure requirements for carbon credit use should be broadly consistent across jurisdictions that have adopted the ISSB standards for climate-related reporting. Identifying a reference to carbon credit use in corporate reports is only a first step; understanding the disclosure requires interpreting what the company plans to do and distinguishing those plans from actions it reports having completed.

The quality of the underlying credits adds another reason for careful interpretation. Research questions both their environmental integrity and their role in corporate mitigation strategies [12, 27, 28]. In a systematic assessment of 2,346 projects, covering approximately one-fifth of global carbon credit volume issued at the time, Probst et al. estimated that less than 16% of the credits issued to the assessed projects represented achieved emissions reductions [6]. Results varied substantially across the project types examined. These findings demonstrate why the volume of credits purchased or promised cannot, by itself, establish their climate benefit. They also reinforce the importance of disclosures that identify the credits involved and explain the basis for claims about their credibility and integrity.

Despite recent advances in standardisation, detecting carbon credit-related information in corporate reports may prove challenging due to the broad range of synomyms that can be used to describe carbon credit use, including nature-based solutions, carbon neutral [25], carbon dioxide removal (CDR), and Beyond Value Chain Mitigation (BVCM) [29]. Existing research on carbon credit disclosure has extracted text using a small number of keywords [10–12], but the diversity of carbon credit-related terminology and the incentive to obscure or selectively disclose carbon credit use [30] suggest that this approach may overlook important information. The use of a broader expert-defined taxonomy to extract relevant text for training, as adopted by Bingler and colleagues to train ClimateBERT_CTI [16], could assist in identifying more subtle references to carbon credit use in corporate reports.

These studies show that locating carbon credit terminology does not, by itself, establish what a company is claiming. Similar wording may describe a general climate

initiative, an intention to use credits, or a completed purchase or retirement. Conversely, a relevant passage may discuss credits without using the exact terms *carbon credit* or *carbon offset*. The challenge for automated analysis is therefore to recognise relevance beyond particular keywords and interpret what the surrounding text says about credit use. This requires expert judgement to inform the distinctions that models are trained and evaluated to recognise.

2.2 Distinguishing carbon credit evidence from similar language

The climate-domain NLP literature demonstrates that transformer models can identify climate-relevant text and analyse disclosure quality, climate commitments, net-zero targets, climate risk, and sentiment [5, 14–17, 31]. Machine-learning studies also show that Australian corporate reports can be analysed computationally [32]. These advances provide useful representations and methods, but their target concepts remain broader than carbon credit relevance and use. Detecting a net-zero statement, for example, does not necessarily show that carbon credits are being used to contribute to the target, or that the reporting entity has acquired, retired, or surrendered any units.

Rule-based methods provide a transparent way to retrieve candidate passages using a domain taxonomy [33]. Their weakness is contextual ambiguity. A paragraph containing *credit*, *offset*, *net*, *neutral*, or *market* may concern finance, taxation, operations, or a generic transition target, rather being related to carbon credits. Such passages become useful *hard negatives* which resemble relevant passages of text but fail the expert definition of carbon credit-relevance [18]. This distinction is also important in research on greenwashing and contested climate communication, where vague or selective claims cannot be assessed without identifiable supporting evidence [30, 34–38]. CarbonCrediBERT does not explicitly identify greenwashing risks, however, it identifies the evidence needed before such a judgement can be made.

2.3 Learning from limited expert labels

The next challenge is learning from limited expert-labelled data. Carbon-credit labels require specialist knowledge and are costly to produce, making it important to identify methods that can learn effectively from small samples. Few-shot methods are designed for this setting. SetFit, for example, adapts a compact language model using pairs of examples with matching or different labels, then trains a lightweight classifier [19]. Alternative approaches use a pretrained model as a fixed feature extractor, leaving its parameters unchanged, or update the model through full fine-tuning [39]. Research on climate-text detection has demonstrated the potential of fine-tuning both climate-specific ClimateBERT and general-purpose BERT [5]. These findings motivate a comparison for carbon-credit disclosure, where the distinctions are more specific and expert-labelled examples are scarce. This study therefore examines whether few-shot learning or these alternative training strategies are more effective for carbon-credit relevance, sentiment, and reported use.

Selective classification concerns whether a trained model should provide a prediction or leave the decision to a human reviewer. Instead of assigning a label to every

paragraph, the model can withhold predictions that fall below a chosen confidence threshold [20]. Active learning addresses a different question: which additional examples should be labelled to improve the model? Recent research combines model-guided selection with human and large language model annotation [40], while ActiveLLM uses language models to select useful examples when few labels are initially available [41].

Selecting the examples with the lowest model confidence does not necessarily make the best use of expert time. Settles and Craven show that uncertainty-based selection can favour unusual examples that provide little information about the wider dataset [42]. Research on imbalanced text datasets also shows the value of considering class representation alongside informative and varied examples [43]. In corporate reports, a practical concern is that low confidence could reflect text fragmented during PDF extraction rather than a meaningful ambiguity about carbon credits. Similarly, labelling many passages with nearly identical transition language may add little new information.

This literature illustrates testing additional criteria for selecting future annotation samples. Checking extraction quality would help ensure that experts receive readable passages, while limiting repeated wording would allow their review to cover a wider range of disclosures. Targeted sampling could also seek examples of infrequent labels, such as negative carbon credit sentiment. Cases where keyword matches and model predictions disagree could be examined to identify further relevant passages or difficult non-relevant examples. These are proposed directions for future evaluation. The present study tests confidence-based referral using a fixed expert-labelled sample; it does not evaluate repeated rounds of annotation and retraining.

2.4 Sentiment of carbon credit-related information

Companies may selectively disclose positive or generalised information to provide a misleading impression of transparency and accountability [30, 44]. Research shows that companies may choose to present climate change as an opportunity [16], but mandatory climate disclosure requirements have been found to reduce the presence of positive sentiment in corporate reports [16, 45].

Carbon credit-related information in corporate reports may be framed positively as supporting climate objectives, negatively as a financial burden or a source of concern about environmental integrity, or neutrally through routine reporting of purchases, holdings, or expenditure. These are contextual distinctions: reporting a cost does not necessarily express negative sentiment, just as describing a purchase does not necessarily express approval. Recent research shows the sentiment of media reporting on carbon credits has an effect on green stock returns [46] , hence corporate reports may adopt more neutral framing of carbon credit-related information following events such as the 2022–2023 integrity scandals [47]

ClimateBERT provides a precedent for distinguishing risk, opportunity, and neutral reporting in climate-related text [14, 16]. CarbonCrediBERT focuses on the carbon credit-related component of a paragraph, rather than any general climate-related text. The resulting labels describe whether carbon credit-related information is framed as positive, negative or neutral, without inferring the environmental performance of a company or the validity of its claims.

Table 1 Comparison of CarbonCrediBERT with related literature. A tick indicates that the criterion is central to the study design; a cross indicates that it is absent or not the main focus.

Author(s)	Transformer	Reports	Carbon credits	Paragraphs	ASX	Hard negatives	Selective review
Webersinke et al. [14]	✓	✓	×	×	×	×	×
Bingler et al. [15]	✓	✓	×	×	×	×	×
Bingler et al. [16]	✓	✓	×	✓	×	×	×
Schimanski et al. [17]	✓	✓	×	×	×	×	×
Moreno and Caminero [33]	×	✓	×	×	×	×	×
Harker et al. [32]	×	✓	×	×	✓	×	×
Trouwloon et al. [25]	×	✓	✓	✓	×	×	×
Lee et al. [11]	×	✓	✓	×	✓	×	×
Rouzegar and Makrehchi [40]	✓	×	×	×	×	×	✓
Bayer et al. [41]	✓	×	×	×	×	×	✓
This paper	✓	✓	✓	✓	✓	✓	✓

2.5 Research gap

The research gap concerns how these interpretive distinctions can be translated into automated analysis when expert-labelled examples are limited. Existing studies provide methods for examining corporate carbon-credit claims, classifying climate-related language, learning from small samples, and referring uncertain predictions for review. CarbonCrediBERT brings these approaches together in a framework that (i) defines carbon-credit relevance, sentiment, and reported use through expert-labelled paragraphs; (ii) links taxonomy-based retrieval to contextual interpretation, including difficult non-relevant examples; (iii) compares climate-specific and general-purpose models under the same limited expert supervision; and (iv) evaluates whether model confidence can identify predictions suitable for automatic acceptance. Its contribution is the design and pilot evaluation of this connected process for carbon credit disclosures in ASX reports. Table 1 shows that existing studies provide important components, whereas the present study brings those components together around carbon credit evidence in ASX reports.

3 Methodology

To address the research gap discussed in Section 2, we propose CarbonCrediBERT as an expert-guided framework for locating and analysing carbon credit disclosures across large collections of corporate reports. The framework connects automated report collection and paragraph extraction with an expert-defined taxonomy, expert labelling, and the comparison of models for relevance, sentiment, and reported credit use. It also evaluates whether model confidence can help determine which predictions may be accepted automatically and which should be referred for human review.

CarbonCrediBERT organises these requirements into five modules, shown in Figure 1. Module 1 establishes the report corpus and retains the source information needed to interpret every extracted passage. Module 2 converts the reports

into complete paragraph evidence, applies the expert taxonomy, and identifies relevant passages. Module 3 uses expert knowledge to implement three paragraph-level tasks: carbon credit relevance (Phase 1), sentiment (Phase 2A), and carbon credit use (Phase 2B). Module 4 compares training strategies that address limited expert-labelled data rather than assuming that one pretrained model will suit every task. Module 5 selects and evaluates the task-specific models on data kept separate from training and tests whether a confidence threshold has been met to accept a prediction automatically, or refer it for expert review.

The five modules separate two decisions that are often conflated in automated disclosure analysis as follows: (i) what a carbon credit-relevant disclosure means and (ii) how a model should learn that meaning. Hence, CarbonCrediBERT requires experts who define the taxonomy, task boundaries, and labels. The candidate models learn from those same definitions and are compared under a common evaluation procedure. The name CarbonCrediBERT therefore refers to the complete conceptual framework, while "BERT" denotes the use of contextual Transformer representations within its modelling module. The methodology of each module is presented in the following subsections.

3.1 Module 1: Report collection and source information

The first module establishes the document base from which all later evidence is drawn. The corpus is constructed in three stages. First, annual reports and associated PDF disclosures are collected from ASX-linked announcement sources. Second, the yearly disclosure set is filtered to the ASX100 subset. Third, dedicated climate, sustainability, and ESG announcement matches are identified for the same collection years. Each retained document is linked to its company, report type, source location, and publication metadata so that extracted paragraphs can be traced to their original report context.

Publication year is retained as an observable announcement attribute and is not treated as an accounting or sustainability reporting period. ASX entities can use the Australian financial year, a calendar or global reporting cycle, and, in defined circumstances, a non-standard financial period of up to 18 months [48]. Dedicated climate and sustainability reports may also cover a different period from the annual financial report. The workflow therefore does not infer fiscal year from titles or extracted text. Reporting-period standardisation remains a separate metadata limitation that requires information within the report beyond the publication date.

3.2 Module 2: Paragraph evidence, taxonomy, and hard negatives

The second module converts the source reports into evidence units that can be retrieved, interpreted, and modelled. After PDF collection, the module extracts and analyses report text using a carbon credit-focused keyword taxonomy. The taxonomy covers seven topic families, as shown in Table 2: (1) general carbon credit and offset language, (2) Australian compliance terms, (3) voluntary-market terms, (4) carbon credit synonyms and decarbonisation language, (5) initiatives (signalling engagement),

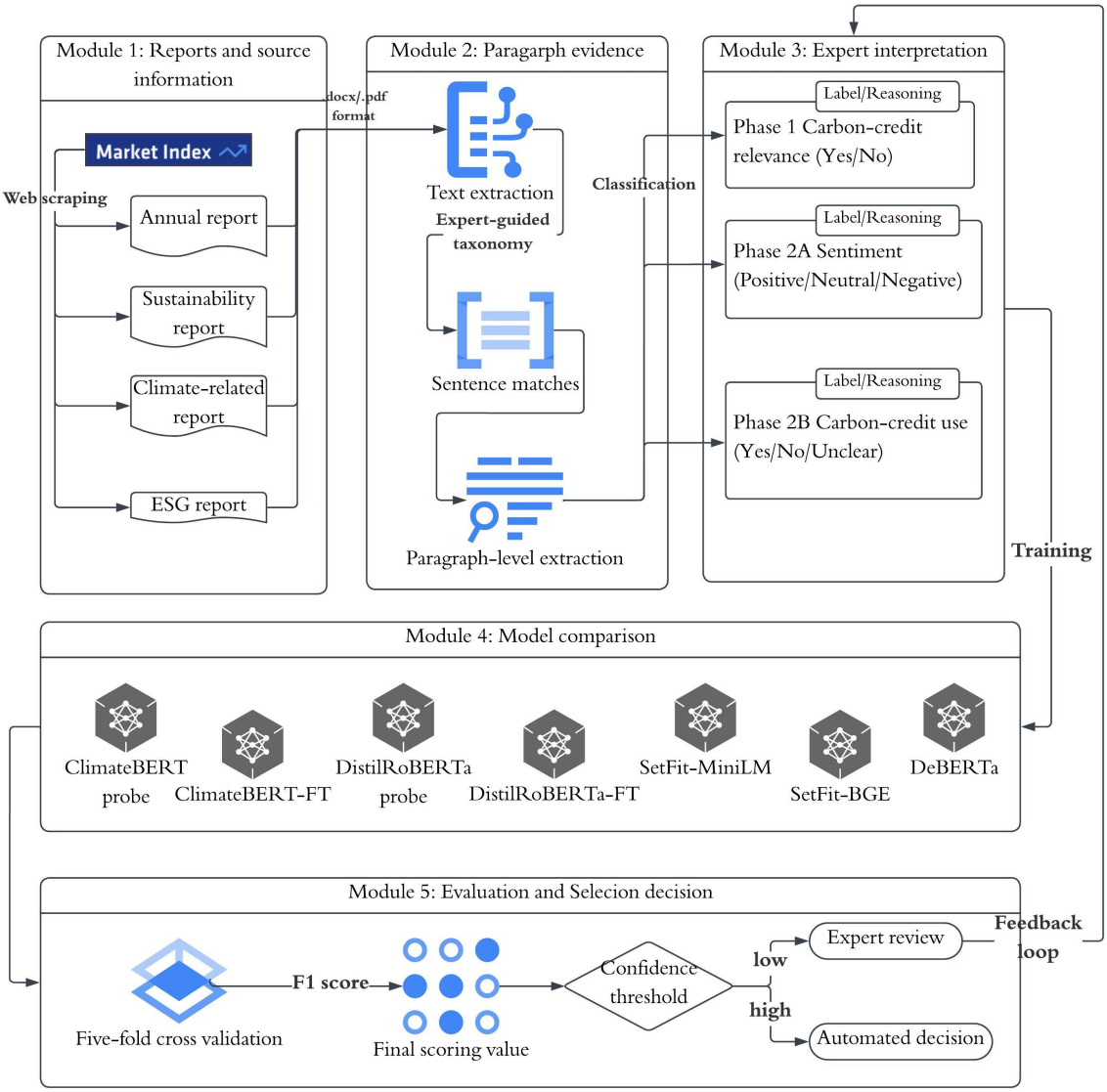


Fig. 1 CarbonCrediBERT conceptual framework

(6) international carbon offset methods, and (7) Australian carbon offset methods. *Sentence-level* extraction records report metadata, report type, matched keyword, and keyword family. These keyword matches are treated as *candidate evidence* rather than labels, because the same terms can indicate a carbon credit-relevance claim in one paragraph and ordinary finance, governance, or operational-setting language in another. The module will then extract longer candidate passages as a *paragraph level* from all available ASX100 annual reports and dedicated climate, sustainability, and

ESG reports. Each paragraph retains company, source URL, page number, source document type (e.g PDF), report subtype (annual, climate, sustainability, ESG), collection year, and publication year.

The resulting design separates three data layers with different purposes. The full report text preserves the source corpus and remains available for future domain-adaptive pretraining. The sentence-level keyword matches provide a descriptive view of carbon credit and offset terminology across the collected reports. These matches indicate where particular terms occur, rather than confirming that a company uses carbon credits. The paragraph candidates provide the units for expert annotation and model training, retaining the surrounding text needed to interpret ambiguous expressions. This separation allows the taxonomy to identify potentially relevant material while leaving the assessment of carbon-credit relevance to expert judgement and the models trained on those labels.

Table 2 CarbonCrediBERT carbon credit taxonomy used for paragraph extraction and descriptive analysis. Counts report keyword-family matches in the balanced all-ASX100 paragraph extract and a paragraph can match more than one family.

Topics	Terms	Matches
General carbon credit and offset terms	Carbon credit; Carbon offset; Offset; Carbon market	1879
Australia/compliance	ACCU; Australian Carbon Credit Unit; Safeguard mechanism	723
Voluntary	Verified carbon unit; VCU; Voluntary carbon market; VCM; Gold Standard; Verra	75
Synonyms for carbon credits	Net zero; carbon removal; CDR; carbon neutral; climate solution; retire	3728
Initiatives (signalling engagement)	Climate Active; Science Based Targets; SBTi; VCMI; ICVCM; Oxford Principles	392
Carbon offset methods – international	Afforestation; Avoided deforestation; Cookstoves; CCS; REDD; Reforestation; Wind power	603
Carbon offset methods – Australia	Human induced; Forest; Regeneration; Restoration; regrowth; forestry; Landfill; Savanna; Soil	2171

The expert-guided taxonomy is a transparent retrieval and interpretation layer, not a trained classifier. It records why a passage entered the candidate pool and helps identify recurring cases in which familiar vocabulary has a different meaning, such as *credit risk*, a general net-zero target, or a forest project with no apparent carbon crediting relationship. Treating expert-guided taxonomy matches as evidence rather than labels allows the supervised models to learn when those terms do and do not indicate a substantive carbon credit disclosure.

Hard-negative construction.

For Stage 1, a hard negative is a paragraph retrieved by the taxonomy that an expert determines is not related to carbon credits. As shown in Equation 1, let $T(x) = 1$ indicate that paragraph x is retrieved by at least one term in the broad carbon credit

taxonomy, and let $Y(x) = 1$ indicate that the expert label is carbon credit related. The taxonomy-induced hard-negative set is

$$\mathcal{H}_{\text{tax}} = \{x : T(x) = 1 \wedge Y(x) = 0\}. \tag{1}$$

By comparison, a non-retrieved or comparatively easy negative satisfies $T(x) = 0$ and $Y(x) = 0$. This definition follows the machine-learning use of *hard negative* for a non-target instance that is difficult to distinguish from a target instance because it shares salient features with that target [18]. In this study, the shared features are carbon credit-adjacent words, concepts, and reporting contexts rather than visual or embedding proximity alone.

Four recurrent hard-negative families arise in corporate reports. First, finance and accounting polysemy includes expressions such as *credit risk*, *tax offset*, and *net interest*. Second, generic climate-transition language includes net-zero targets, emissions reductions, and transition roadmaps that do not connect those commitments to credits or offsets. Third, method-only language refers to forests, soil, savanna burning, landfill gas, carbon capture and storage (CCS), or nature-based projects without linking the method to credit issuance, eligibility, ACCUs, retirement, surrender, cancellation, or offset use. Fourth, signalling language names an initiative or certification without providing substantive evidence about a carbon credit instrument or transaction.

The definition is task specific. A paragraph that substantively discusses carbon credit markets but states that a company does not use credits remains *carbon credit related* at Stage 1 and may receive *no* at Stage 2B. Similarly, negative sentiment towards offsets is a Stage 2A label, not a hard negative. Hard negatives are therefore not adverse opinions or statements of non-use; they are passages that fall outside the Stage 1 target class despite resembling it.

3.3 Module 3: Expert interpretation and task structure

The third module defines the decisions that the models are asked to learn. CarbonCrediBERT uses paragraphs as the primary annotation unit because carbon credit and offset claims often depend on relationships across adjacent sentences. For example, one sentence may introduce a net-zero target, another may describe residual emissions, and a later sentence may specify whether credits or removals will be used. Paragraphs therefore preserve enough context for annotators and models to judge whether carbon credit use is explicit, absent, conditional, or merely implied.

The annotation workflow contains one relevance gate and two downstream tasks:

1. **Stage 1: carbon credit relevance.** A binary label records whether the paragraph is related to carbon credits, offsets, removals, carbon neutrality, carbon-credit accounting, credit-related emissions claims, or compliance mechanisms such as ACCUs and the Safeguard Mechanism. A general statement about climate change or net zero is not sufficient unless the paragraph connects it to a carbon credit instrument.
2. **Stage 2A: carbon credit sentiment.** Relevant paragraphs are labelled *positive*, *negative*, or *neutral*. The label reflects the meaning of the whole paragraph, but the

evaluative tone must concern the carbon credit or offset component rather than an unrelated climate initiative or project type.

3. **Stage 2B: carbon credit use.** Relevant paragraphs are labelled *yes*, *no*, or *unclear*. *Yes* indicates actual, required, or planned use by the relevant entity, including retirement, surrender, compliance, certification, or customer and product offsetting. *No* indicates explicit non-use or discussion without attributing use to the reporting entity. *Unclear* is used when the paragraph makes reference to carbon credits or related instruments but does not identify the actor, timing, or commitment clearly enough for a firm decision.

This staged structure matters because a company can discuss carbon markets or projects without using credits itself, and it can mention credits without explaining their role in a corporate claim [11, 25].

The initial expert sample was prepared from the balanced all-ASX100 paragraph extract. The process first retains every paragraph with at least one carbon credit-taxonomy match and then samples an equal number of paragraphs without a match from the same report collection. This gives experts both likely carbon credit passages and ordinary financial, governance, and operational language. The resulting extract contains 15,026 rows: 7,513 taxonomy-matched paragraphs and 7,513 comparison paragraphs. It spans 7,769 rows from the 2024 collection year and 7,257 from 2025. Annual reports contribute 7,174 rows, while climate, sustainability, and ESG reports contribute 7,852. By subtype, the extract contains 7,174 annual-report rows, 2,123 climate-report rows, 5,096 sustainability-report rows, and 633 ESG-report rows. It covers 63 of the 71 taxonomy keywords observed in the available report corpus.

Before expert review, the balanced extract is summarised by year, document type, report subtype, keyword family, and keyword coverage. This check shows whether the sample represents the wider ASX100 collection rather than only its most frequent terms. The expert-labelled collection contains three task-specific batches of 100 paragraphs: carbon credit relevance (Stage 1), carbon credit-relevance sentiment (Stage 2A), and carbon credit use (Stage 2B). These are 300 task instances, not 300 unique paragraphs moving through one complete sequence. Because the batches were prepared separately and overlap only partly, the two downstream tasks are evaluated independently. The present study therefore does not report an end-to-end score from relevance through sentiment and use.

3.4 Module 4: Model comparison and training strategy

The fourth module examines which learning strategy makes best use of the small expert-labelled sample. The central question is whether carbon credit classification benefits most from prior exposure to climate language, from general-purpose language representations adapted directly to the task, or from a few-shot method that learns through comparisons among labelled examples. The primary comparison therefore includes ClimateBERT full fine-tuning, generic DistilRoBERTa full fine-tuning, and SetFit–MiniLM. These candidates represent three distinct learning strategies while remaining compact enough to train repeatedly during cross-validation. Four additional configurations are retained as diagnostics, but they do not determine the selected task

model. Table 3 demonstrates the primary and diagnostic configurations alongside the taxonomy's role in candidate retrieval.

ClimateBERT: climate-domain transfer.

ClimateBERT is the closest existing domain baseline because it was pretrained on climate-related text [14]. It is included to test whether broad climate-language knowledge transfers to the narrower distinctions required for carbon credit relevance, sentiment, and use. The frozen probe starts from `climatebert/distilroberta-base-climate-f`, leaves its 81.7 million encoder parameters unchanged, converts each paragraph into a fixed representation, and fits a class-balanced logistic-regression classifier. This probe measures what the existing representation already contains. The primary ClimateBERT candidate updates the full encoder and classification head, allowing the representation to adapt to the expert-defined carbon credit task. It is trained for up to four epochs with learning rate 2×10^{-5}, batch size four, class-weighted cross-entropy, maximum length 384, and seed 42. Mean held-out macro F1 across the five folds selects the epoch, with the earlier epoch used to break an exact tie.

DistilRoBERTa: matched general-purpose control.

The generic comparison begins from `distilbert/distilroberta-base`, the base architecture from which the climate checkpoint was adapted [49]. It is chosen because its architecture and parameter count are close to ClimateBERT, making the contrast more informative than a comparison with an unrelated large model. The frozen probe uses the same representation and logistic-regression procedure as the ClimateBERT probe, while the fully fine-tuned candidate uses the same optimisation choices, token limit, and random seed as ClimateBERT. Differences between these matched candidates therefore provide evidence about the contribution of climate-domain pretraining rather than model scale or training conditions.

SetFit–MiniLM: few-shot contrastive learning.

SetFit is included because it was developed for classification with limited labelled data and does not require prompts or generated examples [19]. It learns from pairs of paragraphs: examples with the same label are brought closer in the representation space, while examples with different labels are pushed apart. A lightweight class-balanced classifier is then fitted to those representations. This procedure makes repeated use of the relationships among the expert-labelled paragraphs without treating synthetic text as new expert evidence. The primary candidate starts from `sentence-transformers/all-MiniLM-L6-v2`, a compact 22.7-million-parameter encoder. Twenty contrastive iterations generate 2,560 pairs within each 64-paragraph cross-validation training fold and 3,200 pairs for the final 80-paragraph fit. Training uses one epoch, batch size 16, learning rate 2×10^{-5}, maximum length 384, and seed 42. Held-out fold paragraphs and test paragraphs are never used to construct training pairs.

Secondary architecture diagnostics.

SetFit with `BAAI/bge-small-en-v1.5` tests whether the SetFit result depends on the MiniLM representation rather than on the contrastive procedure itself [50]. DeBERTa-v3-small provides a comparison with a different pretrained architecture and replaced-token detection objective [51]. It uses the same four-epoch search applied to the fully fine-tuned DistilRoBERTa models. Despite its name, the implemented DeBERTa classifier contains approximately 141.9 million parameters because of its large vocabulary embeddings. Together with the two frozen probes, these secondary models help interpret the primary comparison; they are not used to choose the final task model and therefore do not expand the model-selection search after the test results are known.

Table 3 Controlled model configurations and the taxonomy evidence layer. Parameter counts are measured from the implemented task models.

Component	Initial representation	Update from expert labels	Parameters	Role
Taxonomy layer	carbon credit terms and exclusions	No fitted update	0	Candidate retrieval and interpretation context
ClimateBERT probe	ClimateBERT climate-domain encoder	Frozen encoder; balanced linear head	81.7M loaded; $\leq$2,307 fitted	Secondary representation diagnostic
ClimateBERT-FT	ClimateBERT climate-domain encoder	Full encoder and classification head	82.3M	Primary climate-domain candidate
DistilRoBERTa probe	Generic DistilRoBERTa encoder	Frozen encoder; balanced linear head	81.5M loaded; $\leq$2,307 fitted	Secondary representation diagnostic
DistilRoBERTa-FT	Generic DistilRoBERTa encoder	Full encoder and classification head	82.1M	Primary generic comparison
CarbonCrediBERT-SetFit-MiniLM	MiniLM sentence encoder	Contrastive body plus balanced logistic head	22.7M	Primary few-shot candidate
CarbonCrediBERT-SetFit-BGE	BGE-small sentence encoder	Contrastive body plus balanced logistic head	33.4M	Secondary sentence-encoder diagnostic
DeBERTa-v3-small	DeBERTa-v3 replaced-token pretraining	Full encoder and classification head	141.9M	Secondary larger-architecture diagnostic

3.5 Module 5: Evaluation and selective decision

The fifth module separates model development from final evaluation and then examines whether confidence can support selective referral. Each task contains 100 expert-labelled paragraphs: 80 are used for model development and 20 form a fixed test set. Five-fold cross-validation divides the 80 development paragraphs into five groups of 16. In turn, a model is trained on 64 paragraphs and predicts the 16 it has not

seen. Repeating this process gives every development paragraph one prediction from a model that was not trained on that paragraph. These are called *out-of-fold* (OOF) predictions. They provide a more informative development estimate than choosing a model from a single 16-paragraph validation split.

The folds preserve the class composition as far as the small samples allow. Stage 1 is divided across relevant passages, taxonomy-retrieved look-alike negatives, and ordinary negatives; sentiment and carbon credit use are divided by their final expert labels. Source identifiers and normalised text are checked to prevent overlap between the development and test sets.

The same five folds are used for ClimateBERT full fine-tuning, generic DistilRoBERTa full fine-tuning, and SetFit–MiniLM. For the two fully fine-tuned models, the average macro F1 across the five held-out groups selects a common training duration of one to four epochs. The OOF predictions from all 80 development paragraphs are then combined. The strategy with the highest OOF macro F1 is selected for each task, with OOF accuracy used only to break a tie. After the model strategy, training duration, and confidence rule are fixed, the selected model is retrained on all 80 development paragraphs and evaluated once on the unchanged 20-paragraph test set. The four models retained as secondary diagnostics do not influence this selection.

The test set is excluded from cross-validation, training choices, SetFit pair construction, confidence-threshold selection, and final refitting. The same 20 records appeared in an earlier exploratory analysis, so they are described as a fixed held-out test for the present procedure rather than as a previously unseen external benchmark. Their labels are not used for any decision in the revised analysis. Independent confirmation will require a new test set drawn from different issuers, reports, or reporting years. Training records retain the exact development paragraphs used for each final fit so that the results can be reproduced.

Stage 1 relevance, Stage 2A sentiment, and Stage 2B carbon credit use are evaluated separately. This avoids averaging a gate with conditional downstream tasks and avoids implying an end-to-end result from task-specific samples. The present split is paragraph-disjoint but is not yet an issuer-holdout, report-family-holdout, or temporal validation design. Those stricter tests require a larger expert-labelled sample.

The evaluation reports accuracy, macro precision, macro recall, macro F1, weighted F1, results for each class, confusion matrices, and training cost. OOF accuracy is the proportion of the 80 development paragraphs predicted correctly when each paragraph is assessed by a model that did not train on it. OOF macro F1 first balances precision and recall within each class and then gives every class equal weight. Macro F1 is emphasised because overall accuracy can appear reasonable even when the model fails on a rare class, as occurs for negative sentiment. Fold-level scores show how much the result changes with the particular paragraphs held out. Confidence intervals are estimated from 5,000 stratified resamples of each selected model's 20-paragraph fixed test. They describe uncertainty for that fit and test partition, but not variation from other random seeds, expert samples, companies, or reports.

Context-window audit.

The candidate models process at most 384 tokens from each paragraph. Before introducing a longer-context model, the study checks the text omitted after token 384. A long-context comparison is warranted only if the omitted text contains evidence needed for a label, if at least 5% of expert-labelled paragraphs are affected, or if at least 1% of the wider candidate pool contains a direct carbon credit term only in the omitted text. This audit tests whether a longer model would address observed information loss rather than assuming that a larger input window must improve classification [52].

Referring uncertain predictions.

Selective classification tests whether a model can withhold an automatic label when its confidence is insufficient and refer that paragraph for expert review [20]. *Coverage* is the proportion of predictions accepted automatically. *Accepted-case error*, called risk in the selective-classification literature, is the proportion of accepted predictions that are wrong. The aim is not to force a decision for every paragraph, but to determine whether accepting fewer predictions produces a defensibly lower error rate.

For each selected task model, the 80 OOF predictions are ranked by confidence. A candidate threshold must accept at least four paragraphs and keep the observed accepted-case error at or below 0.20. Because an apparently low error can arise by chance in a small sample, the upper end of a one-sided 90% binomial interval must also remain at or below 0.20. The threshold is selected from development data, fixed before the final model is trained on all 80 paragraphs, and applied unchanged to the test set. If no threshold meets these conditions, the evidence does not support automatic acceptance for that task. Because the same development predictions inform both model selection and threshold selection, this analysis is exploratory; a separate calibration sample would be required before the rule could be applied to new reports.

Computational profiling.

The technical profile records model parameters, wall-clock benchmark time, and operational energy and CO_2e estimated with CodeCarbon 3.2.9 [53] for Australia on an Apple M4 system with 10 CPU cores and 16 GB RAM. The instrumented architecture screen covers one fit per task and configuration using 64 training paragraphs. It is retained as a per-recipe comparison and does not represent the cumulative footprint of five-fold selection and final refitting. Tracking begins immediately before checkpoint loading and fitting and ends after validation and test probabilities are produced. Because privileged macOS `powermetrics` access was unavailable, CodeCarbon used its CPU-load fallback; the values are comparative operational estimates within one machine and software environment, not direct electrical measurements or lifecycle emissions.

4 Corpus and Experimental Results

Table 4 summarises the weakly supervised sentence layer produced from the 2024–2025 ASX100 report collection. The extraction yields 8,299 keyword-matched sentence rows across the two publication-year corpora. Category 4, which groups net-zero and

carbon credit-adjacent terminology, dominates the distribution, whereas voluntary-market terminology remains comparatively sparse.

Table 4 Weakly supervised sentence counts from the 2024–2025 ASX100 keyword extraction.

Category	Total	2024	2025
Category 1: General	2846	1465	1381
Category 2: Australia/compliance	931	445	486
Category 3: Voluntary	84	40	44
Category 4: Synonyms for carbon credits	4196	2227	1969
Category 5: Initiatives (signalling engagement)	242	144	98
Total	8299	4321	3978

At the keyword level, Table 5 lists illustrative high-frequency matches, including *Net zero* (3,205 sentence rows), *Offset** (1,232 sentence rows), *Carbon credit* (829 sentence rows), and *Safeguard mechanism* (422 sentence rows). This distribution identifies areas of strong lexical coverage but also anticipates important failure modes. In particular, the broad *Offset** rule is exposed to false positives from accounting and finance language, while the voluntary-market group is too sparse to support robust modelling without targeted sampling or additional data collection.

Table 5 High-frequency keyword counts in the 2024–2025 sentence layer.

Keyword	Category	Total	Interpretation
Net zero	Category 4	3205	Dominant transition and claims phrase
Offset*	Category 1	1232	Broad term with potential false positives
Carbon credit	Category 1	829	Core carbon credit indicator
Safeguard mechanism	Category 2	422	Australia-specific compliance signal
ACCU	Category 2	339	Australian unit-level signal
Climate Active	Category 5	212	Initiative and certification signal

The paragraph-level CarbonCrediBERT extract expands the earlier sentence-level view into an expert-review pool. Table 6 shows this pool, its report subtypes, and the three task-specific expert-review batches. The all-ASX100 balanced extract contains 15,026 paragraph rows, split evenly between 7,513 paragraphs with at least one carbon credit-taxonomy keyword match and 7,513 taxonomy-unmatched comparison paragraphs sampled from the same broad report collection. This balance is designed to expose the model to both carbon credit and offset language and ordinary financial report language that may contain terms such as *credit*, *offset*, or *net*. The extract contains 7,769 rows from the 2024 collection year and 7,257 rows from the 2025 collection year. It also preserves report-type diversity: 7,174 rows come from annual reports,

while 7,852 rows come from climate, sustainability, or ESG reports. By report subtype, the extract contains 7,174 annual-report rows, 2,123 climate-report rows, 5,096 sustainability-report rows, and 633 ESG-report rows.

Table 6 CarbonCrediBERT balanced all-ASX100 paragraph extract and expert-review package.

File or subset	Rows	Purpose
Balanced paragraph extract	15026	Full all-ASX100 expert-review pool
Keyword-match paragraphs	7513	Paragraphs with at least one taxonomy keyword match
Taxonomy-unmatched comparison paragraphs	7513	Balanced comparison examples from report text
Stage 1 pilot batch	100	Binary carbon credit relevance review
Stage 2A pilot batch	100	Sentiment review for carbon credit-relevant candidates
Stage 2B pilot batch	100	Carbon credit-use review for carbon credit-relevant candidates
Annual report subtype	7174	Annual-report paragraph rows
Climate report subtype	2123	Dedicated climate-report paragraph rows
Sustainability report subtype	5096	Sustainability-report paragraph rows
ESG report subtype	633	ESG-report paragraph rows

4.1 Expert-label composition and cross-validated model selection

The three tasks differ in class support and in the semantic relation required for a decision. Relevance is usually tied to observable carbon credit instrument, market, claim, or transaction language, although taxonomy-induced hard negatives share much of the same vocabulary. Sentiment is narrower than general climate sentiment because evaluative tone must attach to the carbon credit-related component of the paragraph. Carbon credit use requires entity and temporal attribution: the classifier must distinguish the reporting entity's use from customer, product, project, or market discussion and separate present action from future, conditional, or referential statements.

All 100 expert-labelled records in each task are retained. Table 7 shows the development and fixed-test composition. Each development paragraph is held out once across five-fold cross validation (five-fold CV) of 16. The sentiment task is especially imbalanced: only six negative paragraphs are available for development, giving one or two negative examples per held-out fold, and the fixed test contains one. These counts make macro and class-level measures essential.

Table 8 reports the primary model-selection analysis. The OOF columns summarise predictions made during five-fold cross-validation: each development paragraph is predicted by a model trained on the other four folds. ClimateBERT full fine-tuning leads relevance with OOF macro F1 of 0.923, ahead of generic DistilRoBERTa full fine-tuning (0.870) and SetFit–MiniLM (0.857). SetFit leads sentiment (0.548) and carbon credit use (0.595), although its fold scores vary substantially for both downstream tasks. The result favours a different training strategy for different questions rather than one encoder for all three decisions.

Table 7 Class composition of the expert-labelled development and fixed-test sets.

Task	Development with five-fold CV	Fixed test
Stage 1 relevance	$n = 80$: related 23; not related 57	$n = 20$: related 5; not related 15
Stage 2A sentiment	$n = 80$: negative 6; neutral 40; positive 34	$n = 20$: negative 1; neutral 10; positive 9
Stage 2B credit use	$n = 80$: no 23; unclear 32; yes 25	$n = 20$: no 7; unclear 8; yes 5

After selection, ClimateBERT is retrained for four epochs on all 80 relevance development paragraphs. SetFit is retrained on all 80 sentiment and use development paragraphs using the fixed contrastive recipe. On the 20-paragraph fixed tests, the selected models obtain macro F1 of 0.785 for relevance, 0.513 for sentiment, and 0.744 for use. The intervals remain wide because each test contains only 20 paragraphs.

Table 8 Five-fold development selection and fixed-test evaluation. Only the selected strategy for each task is retrained on all 80 development paragraphs and evaluated on the fixed test. Intervals are percentile intervals from 5,000 stratified test resamples.

Task	Strategy	OOF accuracy	OOF macro F1	Final epoch	Test accuracy	Test macro F1 [95% interval]
Stage 1 relevance	**ClimateBERT-FT**	**0.938**	**0.923**	4	0.850	0.785 [0.531, 0.937]
	DistilRoBERTa-FT	0.888	0.870	4	–	–
	SetFit–MiniLM	0.888	0.857	fixed	–	–
Stage 2A sentiment	ClimateBERT-FT	0.738	0.511	4	–	–
	DistilRoBERTa-FT	0.725	0.502	4	–	–
	SetFit–MiniLM	0.675	**0.548**	fixed	0.750	0.513 [0.376, 0.616]
Stage 2B credit use	ClimateBERT-FT	0.488	0.478	4	–	–
	DistilRoBERTa-FT	0.413	0.407	4	–	–
	SetFit–MiniLM	**0.600**	**0.595**	fixed	0.750	0.744 [0.555, 0.907]

Note: OOF means out-of-fold. OOF accuracy is the proportion of development paragraphs classified correctly when each paragraph is predicted by a model that did not train on it. OOF macro F1 gives each class equal weight and therefore reveals errors in rare classes more clearly than accuracy alone. "Fixed" indicates SetFit's pre-specified one-epoch training procedure; a dash indicates that the strategy was not selected for final test evaluation.

The class-level results reveal why accuracy alone is misleading. The selected relevance model correctly rejects 14 of 15 non-relevant paragraphs and retrieves three

of five relevant paragraphs. It rejects all eight ordinary negatives and six of seven taxonomy-induced hard negatives. The one hard-negative error is assigned carbon credit relevance with high confidence, while both missed relevant paragraphs are also high-confidence errors. In this small sample, a high predicted probability therefore does not mean that the model has resolved the difficult distinction correctly.

The selected sentiment model identifies seven of ten neutral and eight of nine positive paragraphs but predicts the only negative paragraph as positive. The task remains under-supported for negative sentiment despite 0.750 overall accuracy. The selected carbon credit-use model retrieves three of seven *no*, seven of eight *unclear*, and all five *yes* paragraphs. Its test macro F1 is higher than its cross-validated development estimate, but the model still struggles to distinguish an explicit statement of non-use from a passage in which use is simply unclear.

4.2 Selective classification

Figure 2 presents the development evidence used to select the training duration and examine whether prediction confidence distinguishes correct classifications from errors. Mean held-out macro F1 selects epoch 4 for both fully fine-tuned candidates on every task. Panel (a) averages the five fold scores to choose an epoch. Table 8, by contrast, combines all 80 OOF predictions and calculates macro F1 once to choose a model strategy. Because macro F1 is not a simple average, the two values need not be identical. Panel (b) also explains why confidence alone is an unreliable guide in this sample: some incorrect predictions receive probabilities close to 1.0.

Meanwhile, Figure 3 shows the relationship between the share of predictions accepted automatically and the error rate among those accepted predictions. Only relevance passes the development screen. Its confidence threshold of 0.589 accepts all 80 cross-validated development predictions, including five errors. This corresponds to a development error rate of 0.063 and a one-sided 90% upper bound of 0.113. When applied unchanged to the fixed test, the rule accepts 19 of 20 paragraphs and makes three errors, an accepted-case error rate of 0.158. The one referred paragraph is not one of the errors. Confidence therefore does not yet reduce the number of incorrect relevance decisions in a practically useful way.

For sentiment, a threshold can accept 50 of 80 development predictions with an observed error rate of 0.20, but the upper confidence bound is too high for the rule to pass. For carbon credit use, no threshold accepting at least four paragraphs reaches even the observed error target. The available development evidence therefore does not support automatic acceptance for either downstream task. Table 9 demonstrates which task models pass the development screen and the resulting automatic coverage and accepted-case error on development and test data.

This exercise shows how confidence-based referral could be tested, but it does not establish a safe operating threshold. Every development paragraph is predicted by a model that did not train on it; however, the same 80 predictions are used both to choose the model and to search for a threshold, and their probabilities have not been calibrated on independent data. The reported rule is therefore evidence from this pilot, not a guarantee about future reports. A larger, separate calibration sample is needed before automatic acceptance can be justified.

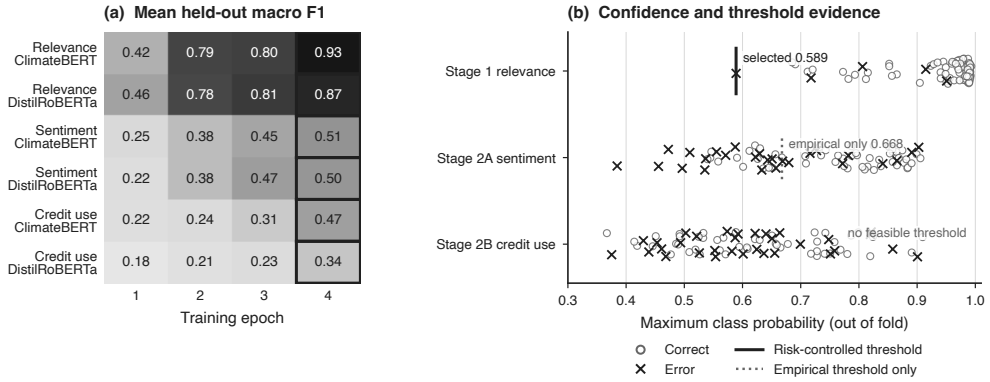


Fig. 2 Training-duration and confidence-threshold diagnostics. Panel (a) reports mean held-out macro F1 at epochs 1–4 for the two fully fine-tuned candidates; outlined cells mark the selected epoch. SetFit is omitted because its contrastive training duration is fixed. Panel (b) shows the maximum predicted probability for each OOF prediction. Crosses identify errors, the solid line marks the Stage 1 threshold that passes the development screen, and the dotted line marks the Stage 2A threshold that does not.

Table 9 Confidence-based referral screen for each selected task model. A rule passes the development screen only if it accepts at least four paragraphs and the one-sided 90% upper bound on accepted-case error is no greater than 0.20.

Task	Selected strategy	Passes screen	Threshold	Development coverage	Development error	Test coverage	Test error
Stage 1	ClimateBERT-FT	Yes	0.589	1.00	0.063	0.95	0.158
Stage 2A	SetFit–MiniLM	No	–	0.00	–	0.00	–
Stage 2B	SetFit–MiniLM	No	–	0.00	–	0.00	–

4.3 Computational resource and operational estimates

Figure 4 separates parameter count from the cost of the implemented training recipe. The ClimateBERT and generic DistilRoBERTa probes each load approximately 82 million encoder parameters but fit at most 2,307 head parameters. Their fully fine-tuned counterparts update approximately 82 million parameters. The MiniLM and BGE-small SetFit encoders contain 22.7 and 33.4 million parameters, whereas the DeBERTa-v3-small classification model contains 141.9 million because its vocabulary embeddings dominate the total despite its six-layer encoder. Across the three task heads, the instrumented CPU runs require 0.23 minutes for each frozen probe, 2.44 minutes for ClimateBERT-FT, 2.47 minutes for DistilRoBERTa-FT, 12.30 minutes for SetFit-MiniLM, 48.80 minutes for SetFit-BGE, and 4.21 minutes for DeBERTa. The corresponding CodeCarbon CPU-load energy estimates are 0.057 Wh for the ClimateBERT probe, 0.055 Wh for the DistilRoBERTa probe, 0.530 Wh for ClimateBERT-FT, 0.556 Wh for DistilRoBERTa-FT, 3.229 Wh for SetFit-MiniLM, 10.887 Wh for SetFit-BGE, and 1.101 Wh for DeBERTa. These are workload-specific

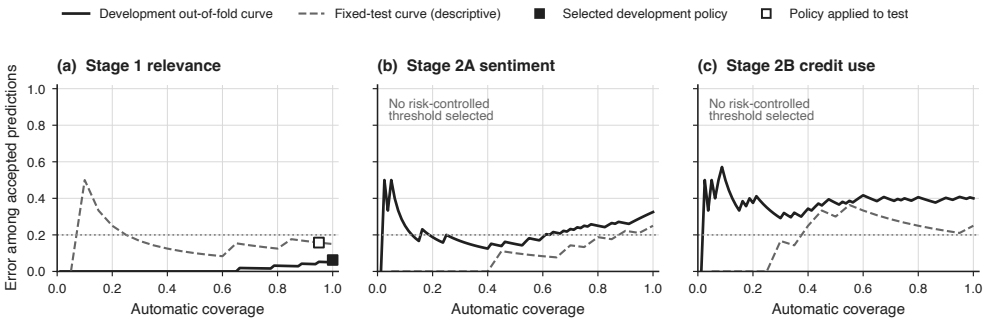


Fig. 3 Accepted-case error as the proportion of automatically accepted predictions increases. Solid lines use the combined OOF development predictions; dashed test curves are descriptive and do not select thresholds. Filled and open squares show the relevance rule on development and test data, respectively. The dotted line marks the target accepted-case error of 0.20. Neither downstream task produces a threshold that passes the development screen.

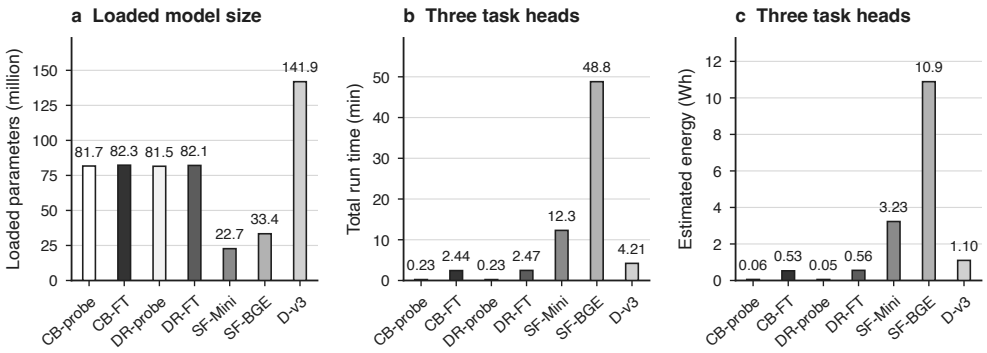


Fig. 4 Implemented configuration resource profile. CB denotes ClimateBERT, DR denotes DistilRoBERTa, SF indicates SetFit, D stands for DeBERTa. Panel (a) reports total loaded model parameters; only the linear heads, at most 2,307 parameters, are fitted for the frozen probes. Panels (b)–(c) report total instrumented CPU time and estimated energy for one run of each of the three task heads. Each SetFit configuration expands a 64-paragraph training split into 2,560 contrastive pairs, so a smaller encoder does not necessarily imply lower realised training cost.

measurements, in particular, they do not establish an intrinsic runtime ordering between sentence encoders.

These measurements are operational estimates from one Apple M4 CPU environment, not lifecycle carbon footprints. CodeCarbon used a CPU-load fallback because direct powermetrics access was unavailable. The estimates do not include upstream checkpoint pretraining, data preparation, report processing, expert work, most model-development runs, storage, networking, cooling, or embodied hardware emissions. The primary protocol produced 45 initial fold fits, 12 additional Transformer fits needed to reproduce predictions at the selected epoch, and three final fits. Saved training timers total 91.6 minutes across these 60 fits, but CodeCarbon did not measure the complete process from beginning to end. The practical lesson is that parameter count alone does not describe the cost of a few-shot training procedure.

4.4 Benchmark

The benchmark shows that model choice depends on the task. ClimateBERT full fine-tuning performs best for relevance, whereas SetFit–MiniLM is selected for sentiment and carbon credit use. The confidence analysis is less encouraging. Although relevance passes the development screen, its threshold refers only one test paragraph and does not prevent any error. Neither downstream task passes the screen, so the present evidence does not support automatic acceptance for sentiment or carbon credit use.

The errors also show which evidence is missing. Relevance needs more examples of both look-alike negative passages and relevant paraphrases that contain no taxonomy term. Negative carbon credit sentiment requires deliberate collection because seven examples cannot support a reliable three-class model. Carbon credit use needs more passages that distinguish explicit non-use from uncertainty and company action from customer offers, project generation, or general market discussion. A separate calibration sample is also needed before confidence is used to accept predictions automatically. Any improvement from these additions must be measured on new companies or reporting years rather than on the test set reported here.

4.5 Taxonomy-guided comparison of carbon credit disclosure language

To examine how carbon credit language differs between the two publication-year corpora, we compare the seven expert-defined categories in Table 2. Unlike topics inferred automatically from statistical patterns, these categories have a consistent meaning and can be traced to specific keyword rules.

The analytical corpus contains 23,566 overlapping text chunks from 131 reports representing 62 companies. The 2024 corpus contributes 67 reports and the 2025 corpus contributes 64. These are publication-year collections rather than a balanced panel of identical issuers, and publication year is not interpreted as the accounting or sustainability reporting period. Accordingly, the comparison describes changes in the composition and intensity of taxonomy-aligned language in the available corpora; it does not estimate within-company change or a causal reporting trend.

The comparison uses exact contextual regular-expression matches so that every aggregate remains traceable to the taxonomy. Repeated matches within a chunk are log scaled to limit the influence of boilerplate and keyword lists. For chunk c, category t, and publication year y, the measures are

$$
\begin{aligned}
x_{c,t,y} &= \sum_{k \in K_t} \log(1 + n_{c,k,y}), \\
S_{t,y} &= \sum_{c \in C_y} x_{c,t,y}, \qquad P_{t,y} = \frac{S_{t,y}}{\sum_{j \in \mathcal{T}} S_{j,y}}, \\
R^{chunk}_{t,y} &= 1000 \frac{S_{t,y}}{|C_y|}, \qquad R^{report}_{t,y} = \frac{S_{t,y}}{|D_y|}.
\end{aligned}
\tag{2}
$$

Table 10 Publication-year comparison of exact taxonomy-category evidence. Δ_{chunk} is the change in score per 1,000 chunks; Δ_{report} is the change in score per report.

Taxonomy category	2024 share	2025 share	Δ pp	Δ_{chunk}	Δ_{report}
General	23.5%	22.7%	-0.8	-4.8	-4.1
Australia/compliance	8.0%	7.9%	-0.1	-1.1	-1.3
Voluntary	0.7%	0.9%	+0.2	+0.4	0.0
Synonyms for carbon credits	37.4%	36.6%	-0.8	-6.2	-6.3
Initiatives (signalling engagement)	6.0%	4.6%	-1.4	-4.8	-1.6
Carbon offset methods – international	6.2%	5.9%	-0.2	-1.3	-1.1
Carbon offset methods – Australia	18.1%	21.4%	+3.3	+7.9	-1.4

where K_t is the keyword set for category t, $n_{c,k,y}$ is the number of context-valid matches, C_y is the set of chunks, D_y is the set of reports, and $\mathcal{T}$ is the seven-category taxonomy. A chunk may contribute to more than one category. $P_{t,y}$ measures within-corpus composition, while $R_{t,y}^{chunk}$ and $R_{t,y}^{report}$ normalise evidence by text volume and report count. All 2025–2024 differences are descriptive rather than inferential estimates.

Table 10 combines the compositional comparison with changes under the two alternative denominators.

Australian carbon-offset methods have the largest compositional increase (+3.3 percentage points), and the increase persists after text-volume normalisation (+7.9) but not per report (-1.4). Signalling initiatives have the largest share decrease (-1.4 percentage points), while voluntary-market terminology remains the least prevalent category. General carbon credit and offset terminology and broader carbon credit-adjacent synonyms together account for approximately 60% of exact evidence in both corpora.

The Australian-methods category includes terms such as *forest*, *landfill*, *savanna*, *soil*, and *CCS*, which may be hard negatives when no crediting relationship is stated. Its increase therefore indicates greater salience of method-related language, not greater corporate reliance on carbon credits. Paragraph-level classification and expert review remain necessary to establish credit relevance and use.

5 Limitations

The purpose of this initial benchmark is to determine what can be learned from a small expert-labelled sample and where that evidence remains insufficient. Its results should therefore be read as a controlled pilot.

The most immediate limitation is the amount and composition of the labelled evidence. Each task contains 80 training paragraphs and 20 test paragraphs, so one changed test prediction moves accuracy by five percentage points. Negative sentiment is represented by only seven paragraphs in total, and the selected model misses the only negative example in the test set.

The model comparison is correspondingly provisional. Cross-validation reduces dependence on one small validation split, but each candidate is evaluated with one specified training recipe and random seed. The fixed test is excluded from the revised model-selection procedure, although its labels had been inspected during earlier exploratory work. The pilot's result shows that ClimateBERT is useful for relevance in this sample and SetFit is useful for the downstream tasks.

The corpus also limits what the models can learn. Candidate paragraphs are found through the taxonomy, so relevant paraphrases that contain none of its terms may be missed. Conversely, only a small number of the many possible hard negative passages appear in the expert sample. PDF extraction can break tables, headings, footnotes, and multi-page arguments, while a paragraph may not contain all the evidence needed to interpret a disclosure.

Publication year also cannot be treated as the reporting period because Australian companies use different financial and sustainability reporting cycles, hence manual checks would be required to confirm the reporting period of corporate reports.

Therefore, the limitations demonstrate two needs. More expert labels are required to represent rare and ambiguous meanings, while a broader domain corpus is required if CarbonCrediBERT is to learn carbon credit language that extends beyond the wording of a small ASX100 sample. These two needs frame the discussion and the foundation-model proposal in Section 7.

6 Discussion

This study examined how expert knowledge and natural language processing can help identify and interpret carbon credit disclosures across large collections of corporate reports. The challenge is not simply to locate references to credits or offsets, but to distinguish relevant disclosures from similar language used in other contexts and interpret what companies report about them. CarbonCrediBERT addresses this problem by connecting an expert-defined taxonomy, paragraph-level interpretation, and models trained on limited expert-labelled data. Its contribution is a framework that brings these elements together for three distinct questions: whether a passage concerns carbon credits, how those credits are presented, and whether the passage reports their use. It also examines when model predictions may be accepted automatically and when human review remains necessary.

The findings highlight why these questions should not be treated as interchangeable classification tasks. The fine-tuned ClimateBERT model was selected for carbon-credit relevance, whereas SetFit–MiniLM was selected for sentiment and reported use. Relevance requires distinguishing carbon credit discussion from unrelated financial language and broader climate statements. Sentiment requires determining whether the credits themselves are presented positively, negatively, or neutrally, rather than assigning the general tone of the paragraph. Interpreting reported use also requires attention to who is acting and whether a statement describes an action, an intention, or a conditional possibility. These differences help explain the need to evaluate training strategies separately for each task.

Taken together, these findings point to the next stage of CarbonCrediBERT. The proposed framework defines the questions, supplies expert examples, and provides a benchmark for comparing models. A carbon credit pretrained model would address a different limitation by learning the language of instruments, registries, methodologies, compliance schemes, accounting treatments, and corporate claims from a much larger unlabelled corpus before it sees the small task-specific samples. Expert labels would still define relevance, sentiment, and use, but they would no longer have to teach the model both the wider domain and the classification task from a few examples.

Furthermore, the descriptive comparison in Section 4 provides an Australian reason to pursue this work carefully. Method-related language becomes more prominent in the 2025 publication-year corpus, but terms such as *forest*, *soil*, *landfill*, *savanna*, and *CCS* do not by themselves establish credit generation or use. At the same time, Australia's compulsory sustainability-reporting regime is intended to improve the quality, consistency, and comparability of climate-related financial information [26, 54]. Scalable analysis can help users locate and compare the evidence behind corporate statements, but it must preserve the distinction between mentioning a method and reporting a carbon credit action.

The Australian context also raises a more specific question about AI sovereignty. The National AI Plan supports domestic capability, targeted high-value models, high-quality data, documentation, and human oversight, and notes that locally developed models can reflect Australian context and reduce some data-security risks [55]. A CarbonCrediBERT model developed with Australian reporting concepts, transparent evaluation, and locally governed data could contribute to this capability. It could support scrutiny of ACCUs, the Safeguard Mechanism, and Australian sustainability reports without relying entirely on opaque external services. It could not guarantee the integrity of a corporate report or establish AI sovereignty on its own. Those responsibilities remain with preparers, directors, assurance providers, standard setters, and regulators. CarbonCrediBERT's role would be to help people locate, organise, and compare the relevant evidence.

The resource results reinforce the need for proportionate development. Smaller encoders are not automatically cheaper when their training procedure creates thousands of paragraph pairs, and training a new model from random initialisation would carry much greater data and computational demands. The next step should therefore be continued pretraining of a strong existing encoder, accompanied by transparent reporting of data sources, compute, energy, and evaluation. Section 7 develops this proposal.

7 Future Work

The proposed CarbonCrediBERT foundation model follows directly from the two main limitations of the present framework. The expert-labelled sample is too small to represent rare meanings reliably, while the ASX100 report corpus is too narrow to teach a reusable model the carbon credit domain. Here, a *foundation model* means a pretrained language model that can be adapted to several carbon credit

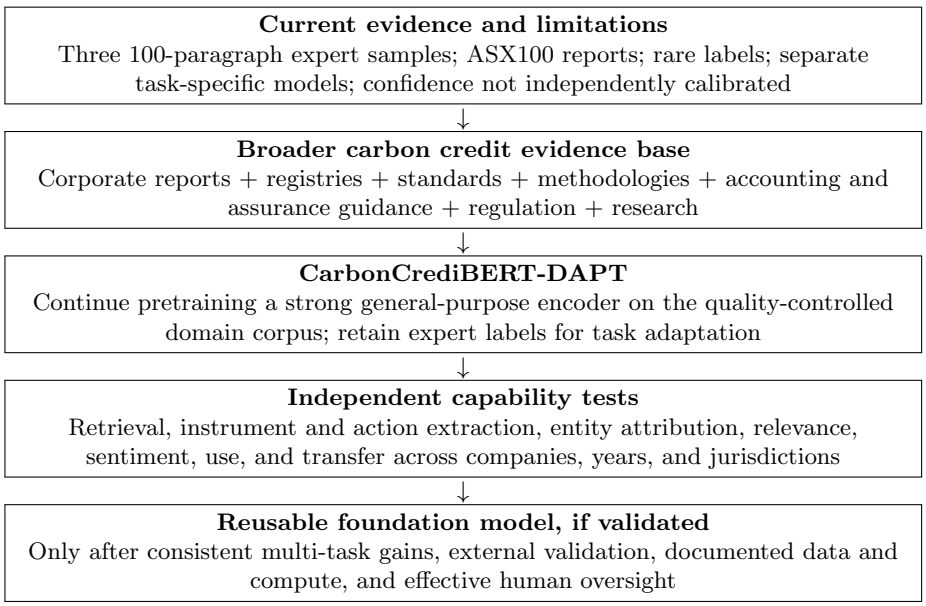


Fig. 5 Proposed pathway from the present CarbonCrediBERT framework to a possible carbon credit foundation model. Each stage addresses an observed limitation of the current study and creates evidence required for the next. CarbonCrediBERT-DAPT is the proposed intermediate pretrained model. Foundation-model status depends on broad, independently validated reuse.

tasks, rather than a separate classifier built for one label. The proposed intermediate model is called CarbonCrediBERT-DAPT, where DAPT means domain-adaptive pretraining. Figure 5 outlines the proposed progression from the current framework to a reusable carbon-credit model, linking the pilot's limitations to broader data collection, continued pretraining, and independent evaluation.

The first change is to broaden the evidence from which the model learns. Corporate reports remain important because they contain the claims presented to investors, but they describe only one side of the domain. The pretraining corpus should also include Australian and international registries and retirement records, compliance and voluntary-market standards, project methodologies, accounting and assurance guidance, legislation, regulatory decisions, and peer-reviewed research. These sources would expose the model to the relationships missing from broader climate-relevance text of which instruments exist, who can issue or retire them, how projects generate units, how units are accounted for, and how their use supports compliance or corporate claims. Every source should retain its document type, jurisdiction, date, and licence, and the corpus should be deduplicated and checked for extraction quality.

The next stage, CarbonCrediBERT-DAPT, would continue pretraining a strong general-purpose encoder on this corpus. This is a more proportionate experiment than training from random initialisation: it preserves general language knowledge while testing whether carbon credit text adds useful domain knowledge [56, 57]. ClimateBERT provides an important comparator because it was created through continued pretraining on a large climate corpus [14]. A matched experiment should compare the

generic encoder, ClimateBERT, CarbonCrediBERT-DAPT, and each model after the same task-specific adaptation.

Expert evidence remains essential in this design. More paragraphs are needed for negative sentiment, explicit non-use, future or conditional reliance on carbon credits , customer offsetting, generation of carbon credits from projects, and relevant paraphrases that the taxonomy does not retrieve. Future samples should use the same paragraphs for all three labels so that the complete relevance-to-sentiment-and-use sequence can be evaluated. Model-guided selection should be compared with random sampling under the same expert time budget, and new issuer- or time-based test sets should remain untouched until model choices are complete.

The proposed model must also demonstrate reuse beyond the three current classifiers. Evaluation should include evidence retrieval from long reports, extraction of instruments and actions, identification of the acting entity, distinction between compliance and voluntary markets, and transfer across companies, report types, years, and jurisdictions. Confidence should be calibrated on separate data before any prediction is accepted automatically. Evaluated outputs should retain the complete paragraph and its source so that each classification remains connected to the evidence on which it is based.

For Australia, this work could support a locally governed capability for analysing disclosures under Australian terminology and reporting institutions. The training corpus could give explicit coverage to ACCUs, the Safeguard Mechanism, Australian Accounting Standard Board (AASB) sustainability disclosure requirements, Australian registries, and local reporting cycles while retaining international material needed for comparison. Its contribution to sovereign AI capability would come from transparent data documentation, reproducible evaluation, human oversight, and the option to run the model within Australian infrastructure.

8 Conclusion

Carbon credit disclosure cannot be interpreted reliably by counting terms such as *credit*, *offset*, or *net zero*. The meaning depends on the instrument, the action, the organisation involved, and the purpose of the statement. CarbonCrediBERT addresses this problem by combining expert definitions with complete paragraph evidence and by separating carbon credit-relevance from the distinct questions of sentiment and reported use. The study provides 8,299 keyword-matched sentence records, a balanced review pool of 15,026 paragraphs, and 100 expert-labelled paragraphs for each task.

The experiment does not identify one model that is best for each downstream tasks. ClimateBERT is selected for carbon credit relevance and records test macro F1 of 0.785, while SetFit–MiniLM is selected for sentiment and carbon credit use, with test macro F1 of 0.513 and 0.744. More important than the ranking is what the errors reveal. The sentiment model misses the only negative test paragraph, and confidence does not reliably separate correct from incorrect predictions. The present evidence therefore supports expert-guided analysis and cautious referral, not unattended classification of corporate reports.

The study also establishes a reasoned path beyond the pilot. More expert-labelled examples are needed for passages that use similar language without reporting a carbon-credit action, such as purchasing, generating, holding, selling, retiring or surrendering credits. A broader corpus is needed to teach a reusable model the relationships among credits, registries, projects, accounting treatments, compliance obligations, and corporate claims. CarbonCrediBERT-DAPT would test whether continued pretraining on that evidence improves several tasks beyond the current framework. Only consistent gains across independent tasks, companies, reporting years, and jurisdictions would justify describing the resulting model as a carbon credit foundation model. In Australia, such a model could strengthen locally governed analytical capability and make disclosure evidence easier to examine, but responsibility for the integrity of corporate reporting would remain with people and institutions.

References

[1] Task Force on Climate-related Financial Disclosures: Recommendations of the task force on climate-related financial disclosures. Final Report (2017)

[2] Locke, N.: International Developments in Mandatory Sustainability Disclosure and What It Means for Australian Corporate Disclosure. Australian Journal of Corporate Law **39**(1), 1–26 (2024). Forthcoming

[3] Baird, R.: The big ask of australia's new climate disclosure laws: Will they improve information asymmetry, quality and consistency and drive down emissions? Commercial Law Quarterly **39**(2), 4–11 (2025)

[4] Simnett, R., Tan, Y.H., You, J., Zhou, S.: Australian listed entities' preparedness for mandatory reporting and assurance of climate-related disclosures. Abacus **62**(2), 337–391 (2026) https://doi.org/10.1111/abac.12363

[5] Garrido-Merchan, E.C., Gonzalez-Barthe, C., Coronado-Vaca, M.: Fine-tuning ClimateBERT transformer with ClimaText for the disclosure analysis of climate-related issues in corporates' financial and non-financial reports. Neural Computing and Applications **38**, 12 (2026) https://doi.org/10.1007/s00521-025-11820-3

[6] Probst, B.S., Toetzke, M., Kontoleon, A., Diaz Anadon, L., Minx, J.C., Haya, B.K., Schneider, L., Trotter, P.A., West, T.A.P., Gill-Wiehl, A., Hoffmann, V.H.: Systematic assessment of the achieved emission reductions of carbon crediting projects. Nature Communications **15**, 9562 (2024) https://doi.org/10.1038/s41467-024-53645-z

[7] Macintosh, A., Trencher, G., Probst, B., Barley, S., Cullenward, D., West, T.A.P., Butler, D., Rockstrom, J.: Carbon credits are failing to help with climate change: Here's why. Nature **646**, 543–546 (2025) https://doi.org/10.1038/d41586-025-03313-z

[8] Kumar, P., Firoz, M.: Accounting for certified emission reductions (CERs) in

india: An analysis of the disclosure and reporting practices within the financial statements. Meditari Accountancy Research **28**(2), 365–389 (2020) https://doi.org/10.1108/MEDAR-01-2019-0428

[9] Salih, L.G.: Decarbonization and the obstacles to carbon credit accounting disclosure in financial statement reports: The case of UAE. Asian Journal of Accounting Research **9**(2), 169–180 (2024) https://doi.org/10.1108/AJAR-04-2023-0128

[10] McGuinness Institute: Analysing carbon offset information disclosed in 2021–2024 annual reports and sustainability reports of nzsx-listed companies. Technical Report 2025/07, McGuinness Institute, Wellington (2025). Accessed 8 October 2026. https://www.mcguinnessinstitute.org/wp-content/uploads/2026/01/WP-2025-07-20260114-4pm-Carbon-Offsets-FINAL.pdf

[11] Lee, E., Tan, E., Tingey, J.: Disclosures of carbon credits in Australia [Working Paper]. Technical report (2025). https://doi.org/10.2139/ssrn.5134021 . https://ssrn.com/abstract=5134021

[12] Stolz, N., Probst, B.S.: The negligible role of carbon offsetting in corporate climate strategies. Nature Communications **16**, 7963 (2025) https://doi.org/10.1038/s41467-025-62970-w

[13] Hartmann, T., Carbone, G., Bhatia, V.: Here's how natural climate solutions and carbon finance can tackle nature loss and climate change. World Economic Forum. Accessed 9 October 2026 (2022). https://www.weforum.org/stories/climate-action/carbon-finance-natural-climate-solutions-climate-change/

[14] Webersinke, N., Kraus, M., Bingler, J.A., Leippold, M.: ClimateBERT: A pretrained language model for climate-related text. In: Proceedings of the AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges (2022). https://www.climatechange.ai/papers/aaaifss2022/12

[15] 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 **47**, 102776 (2022) https://doi.org/10.1016/j.frl.2022.102776

[16] Bingler, J.A., Kraus, M., Leippold, M., Webersinke, N.: How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. Journal of Banking & Finance **164**, 107191 (2024) https://doi.org/10.1016/j.jbankfin.2024.107191

[17] Schimanski, T., Bingler, J., Kraus, M., Hyslop, C., Leippold, M.: ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15745–15756. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.975

. https://aclanthology.org/2023.emnlp-main.975/

[18] Robinson, J., Chuang, C.-Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. Advances in Neural Information Processing Systems **34**, 1–28 (2021)

[19] Tunstall, L., Reimers, N., Jo, U.E.S., Bates, L., Korat, D., Wasserblat, M., Pereg, O.: Efficient few-shot learning without prompts. In: NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing (2022). https://doi.org/10.48550/arXiv.2209.11055

[20] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., ??? (2017)

[21] Alshahrani, F., Eulaiwi, B., Duong, L., Taylor, G.: Climate change disclosure performance and audit fees: evidence from australia. Sustainability Accounting, Management and Policy Journal **16**(5), 1208–1242 (2024) https://doi.org/10.1108/SAMPJ-07-2023-0509

[22] Borghei, Z.: Carbon disclosure: A systematic literature review. Accounting & Finance **61**, 5255–5280 (2021) https://doi.org/10.1111/acfi.12757

[23] Wang, Q.: Financial effects of carbon risk and carbon disclosure: A review. Accounting & Finance **63**, 4175–4219 (2023) https://doi.org/10.1111/acfi.13090

[24] Li, A., Michaelides, M., Rose, M., Garg, M.: Climate-related risk and financial statements: Implications for regulators, preparers, auditors and users. Australian Accounting Review **29**, 599–605 (2019) https://doi.org/10.1111/auar.12296

[25] Trouwloon, D., Streck, C., Chagas, T., Martinus, G.: Understanding the use of carbon credits by companies: A review of the defining elements of corporate climate claims. Global Challenges **7**, 2200158 (2023) https://doi.org/10.1002/gch2.202200158

[26] Australian Accounting Standards Board: AASB S2 Climate-related Disclosures. As amended December 2025. Accessed 18 August 2026 (2025). https://standards.aasb.gov.au/aasb-s2-dec-2025

[27] Cullenward, D., Badgley, G., Chay, F.: Carbon offsets are incompatible with the paris agreement. One Earth **6**, 1085–1088 (2023) https://doi.org/10.1016/j.oneear.2023.08.014

[28] Kreibich, N., Hermwille, L.: Caught in between: Credibility and feasibility of the voluntary carbon market post-2020. Climate Policy **21**, 939–957 (2021) https://doi.org/10.1080/14693062.2021.1948384

[29] Science Based Targets initiative: Corporate Net-Zero Standard version 1.3.1. Technical report, Science Based Targets initiative (2026). https://sciencebasedtargets.org/corporate-net-zero-v1

[30] Lyon, T.P., Maxwell, J.W.: Greenwash: Corporate environmental disclosure under threat of audit. Journal of Economics & Management Strategy **20**(1), 3–41 (2011) https://doi.org/10.1111/j.1530-9134.2010.00282.x

[31] Anoop, V.S., Krishnan, T.K.A., Daud, A., Banjar, A., Bukhari, A.: Climate change sentiment analysis using domain specific bidirectional encoder representations from transformers. IEEE Access **12**, 114912–114922 (2024) https://doi.org/10.1109/ACCESS.2024.3441310

[32] Harker, C., Hassall, M., Lant, P., Rybak, N., Dargusch, P.: What can machine learning teach us about australian climate risk disclosures? Sustainability **14**, 10000 (2022) https://doi.org/10.3390/su141610000

[33] Moreno, A.I., Caminero, T.: Application of text mining to the analysis of climate-related disclosures. International Review of Financial Analysis **83**, 102307 (2022) https://doi.org/10.1016/j.irfa.2022.102307

[34] Freitas Netto, S.V., Sobral, M.F.F., Ribeiro, A.R.B., Soares, G.R.d.L.: Concepts and forms of greenwashing: A systematic review. Environmental Sciences Europe **32**, 19 (2020) https://doi.org/10.1186/s12302-020-0300-3

[35] Montgomery, A.W., Lyon, T.P., Barg, J.: No end in sight? a greenwash review and research agenda. Organization & Environment **37**, 221–256 (2024) https://doi.org/10.1177/10860266231168905

[36] Boussalis, C., Coan, T.G.: Text-mining the signals of climate change doubt. Global Environmental Change **36**, 89–100 (2016) https://doi.org/10.1016/j.gloenvcha.2015.12.001

[37] Coan, T.G., Boussalis, C., Cook, J., Nanko, M.O.: Computer-assisted classification of contrarian claims about climate change. Scientific Reports **11**, 22320 (2021) https://doi.org/10.1038/s41598-021-01714-4

[38] Oppong-Tawiah, D., Webster, J.: Corporate sustainability communication as fake news: Firms' greenwashing on twitter. Sustainability **15**, 6683 (2023) https://doi.org/10.3390/su15086683

[39] Peters, M.E., Ruder, S., Smith, N.A.: To tune or not to tune? adapting pretrained representations to diverse tasks. In: Augenstein, I., Gella, S., Ruder, S., Kann, K., Can, B., Welbl, J., Conneau, A., Ren, X., Rei, M. (eds.) Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pp. 7–14. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-4302 . https://aclanthology.org/W19-4302/

[40] Rouzegar, H., Makrehchi, M.: Enhancing text classification through LLM-driven active learning and human annotation. In: Henning, S., Stede, M. (eds.) Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII), pp. 98–111. Association for Computational Linguistics, St. Julians, Malta (2024). https://doi.org/10.18653/v1/2024.law-1.10 . https://aclanthology.org/2024.law-1.10/

[41] Bayer, M., Lutz, J., Reuter, C.: ActiveLLM: Large language model-based active learning for textual few-shot scenarios. Transactions of the Association for Computational Linguistics **14**, 1–22 (2026) https://doi.org/10.1162/TACL.a.63

[42] Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 1070–1079. Association for Computational Linguistics, ??? (2008). https://aclanthology.org/D08-1112/

[43] Fairstein, Y., Kalinsky, O., Karnin, Z., Kushilevitz, G., Libov, A., Tolmach, S.: Class balancing for efficient active learning in imbalanced datasets. In: Proceedings of the 18th Linguistic Annotation Workshop, pp. 77–86. Association for Computational Linguistics, ??? (2024). https://doi.org/10.18653/v1/2024.law-1.8

[44] Marquis, C., Toffel, M.W., Zhou, Y.: Scrutiny, norms, and selective disclosure: A global study of greenwashing. Organization Science **27**, 483–504 (2016) https://doi.org/10.1287/orsc.2015.1039

[45] Grewal, J., Richardson, G., Wang, J.: Effects of mandatory carbon reporting on greenwashing. The Accounting Review **101**, 263–293 (2026) https://doi.org/10.2308/TAR-2023-0247

[46] Rogmann, J., Schreiber, S.: Carbon credit sentiments and green energy stocks. Applied Economics **57**, 7671–7691 (2025) https://doi.org/10.1080/00036846.2024.2393891

[47] Panizza, U., Tripoli, F., Mauro, B.W.: An autopsy of the voluntary carbon market. Technical report, Centre for Economic Policy Research (2026). https://cepr.org/publications/dp21788

[48] Australian Securities and Investments Commission: Change a company financial year. Accessed 23 July 2026 (2026). https://www.asic.gov.au/for-business-and-companies/companies/changes-to-company-details/change-a-company-financial-year/

[49] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint (2019)

[50] Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.-Y.: C-Pack: Packed

Resources for General Chinese Embeddings. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 641–649 (2024). https://doi.org/10.1145/3626772.3657878

[51] He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with Disentangled Attention (2021). https://arxiv.org/abs/2006.03654

[52] Warner, B., Chaffin, A., Clavie, B., Weller, O., Hallstrom, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Adams, G.T., Howard, J., Poli, I.: Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pp. 2526–2547 (2025). https://doi.org/10.18653/v1/2025.acl-long.127

[53] CodeCarbon: CodeCarbon Documentation. Accessed 24 July 2026. https://docs.codecarbon.io/latest/

[54] Australian Securities and Investments Commission: ASIC issues early observations on sustainability reporting ahead of 30 June 2026. Accessed 18 August 2026 (2026). https://www.asic.gov.au/about-asic/news-centre/news-items/asic-issues-early-observations-on-sustainability-reporting-ahead-of-30-june-2026

[55] Australian Government Department of Industry, Science and Resources: National AI Plan. Accessed 18 August 2026 (2025). https://www.industry.gov.au/publications/national-ai-plan

[56] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.740 . https://aclanthology.org/2020.acl-main.740/

[57] Najem-Meyer, S., Kaplan, F., Romanello, M.: Don't stop pretraining! efficiently building specialised language models in resource-constrained settings. In: Kazantseva, A., Szpakowicz, S., Degaetano-Ortlieb, S., Bizzoni, Y., Pagel, J. (eds.) Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025), pp. 252–260. Association for Computational Linguistics, Albuquerque, New Mexico (2025). https://doi.org/10.18653/v1/2025.latechclfl-1.22 . https://aclanthology.org/2025.latechclfl-1.22/