Climate-ModernBERT: Revisiting Corpus Composition for Domain-Adaptive Continued Pretraining

Anonymous ACL submission

Abstract

Natural Language Processing (NLP) in the climate domain requires models to process heterogeneous text sources, including scientific literature, policy disclosures, and synthetic reports. However, how to effectively combine diverse domain corpora during continued pretraining (CPT) remains underexplored. We introduce CLIMATE-MODERNBERT, a family of climate-adapted encoder models obtained through continued pretraining of ModernBERT-Base on three climate corpora: academic climate text, climate-filtered web data, and synthetic climate documents. We systematically compare joint continued pretraining on corpus mixtures with parameter-space merging of independently specialized checkpoints. Across nine climate NLP benchmarks, our best model achieves 76.3 average F_1, improving significantly over a vanilla ModernBERT baseline by 2.8 points. Within the climate NLP setting, the results show that academic climate corpora provide the strongest adaptation signal among the evaluated sources, while parameter-space merging improves over joint multi-source training and better preserves complementary information from heterogeneous climate corpora. We release all CLIMATE-MODERNBERT variants and training checkpoints[1] to support future research in climate NLP and domain-adaptive pretraining.

1 Introduction

The rapid escalation of the global climate crisis increases the need for advanced NLP tools that can parse, synthesize, and verify large corpora of environmental literature and policy disclosures (Hershcovich et al., 2022). Various efforts in the NLP community aim to support the automatic analysis of climate-related text through tasks such as climate topic detection (Varini et al., 2020; Yu et al., 2024b), climate risk classification (Bingler et al., 2024), and question answering over climate knowledge (Spokoyny et al., 2023). These tools enable practitioners in the climate domain to extract insights from rapidly growing collections of climate reports, scientific publications, and regulatory corpus.

Yet building strong NLP models for this domain remains challenging in practice. Climate text resources are highly heterogeneous (Shi et al., 2023), spanning informal web discourse, scientific literature, and unstructured policy documents, while evaluation benchmarks cover diverse NLP tasks across short and long documents with specialized terminology and concepts. Climate practitioners must therefore operate over noisy and stylistically inconsistent corpora, making it difficult to train a single domain-adapted model that performs reliably across diverse downstream applications. Recent advances in encoder architectures provide a promising direction for addressing these challenges. ModernBERT (Warner et al., 2025) introduces a compact yet powerful encoder model with strong performance across natural language understanding tasks. Compared with large generative decoders, encoder-only models offer a favorable balance between performance and efficiency, making them suitable for large-scale climate analysis pipelines. They support non-generative tasks such as classification, retrieval, and information extraction, and serve as key components in retrieval systems and retrieval-augmented generation pipelines over large document collections (Lewis et al., 2020). These properties make them well-suited for climate applications requiring long-document processing and nuanced semantic understanding across heterogeneous text sources.

Evidence from other scientific domains further supports the benefits of adapting ModernBERT through domain-specific pretraining. Models such as Bioclinical ModernBERT (Sounack et al., 2025),

[1] Anonymized Github for review: https://anonymous.4open.science/r/ClimateModernBERT-25BC/

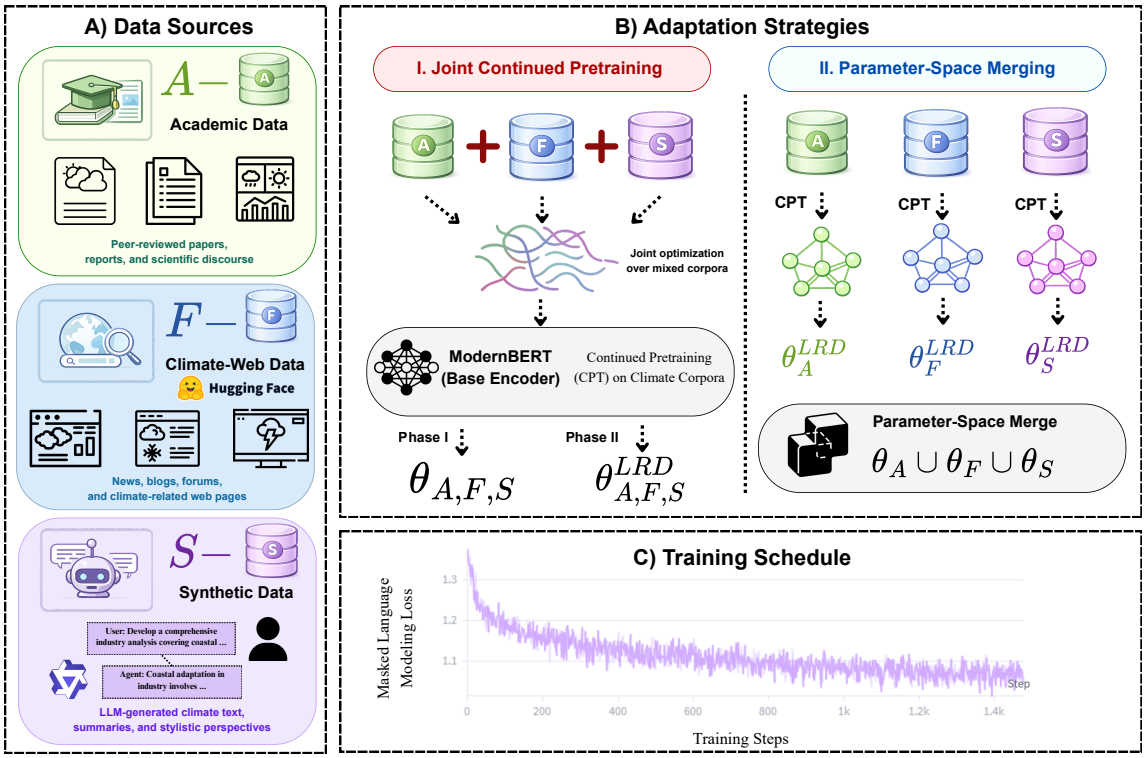

Figure 1: Overview of the CLIMATE-MODERNBERT adaptation framework. (A) Three primary climate-domain corpora used for continued pretraining: academic climate text ($\mathcal{A}$), climate-filtered web data ($\mathcal{F}$), and synthetic climate text ($\mathcal{S}$). (B) Two adaptation strategies studied in this work: joint continued pretraining on mixed corpora and parameter-space merging of independently specialized checkpoints. (C) Following the ModernBERT training schedule, consisting of context extension followed by learning-rate decay specialization. Context-extension-only checkpoints are additionally evaluated as an ablation.

Clinical ModernBERT (Lee et al., 2025), and Legal ModernBERT (Stammbach et al., 2026) show that continued pretraining on large-scale domain corpora consistently improves performance on specialized downstream tasks. These results suggest that a similar strategy could benefit climate NLP systems. However, the role of corpus composition in domain adaptation remains poorly understood, despite growing evidence that different in-domain data sources can lead to substantially different downstream outcomes (Chalkidis et al., 2020; Longpre et al., 2024).

As such, an important question remains largely underexplored in climate NLP: **how should heterogeneous text sources be combined during continued pretraining to adapt modern encoder architectures effectively?** We study two complementary aspects of this problem: (1) the composition of the training corpus and the mechanism used to integrate multiple sources, and (2) either through joint continued pretraining or parameter-space merging of independently specialized checkpoints. More broadly, this case study provides empirical insights into corpus composition and adaptation strategies for climate-domain encoder adaptation, while motivating future investigations in other specialized NLP domains.

We introduce **CLIMATE-MODERNBERT**, a family of climate-adapted encoder models obtained through ModernBERT's continued pretraining procedure on heterogeneous climate corpora. Figure 1 provides an overview of the corpora, adaptation strategies, and two-stage training schedule. Our contributions are threefold:

- We introduce CLIMATE-MODERNBERT, a suite of climate-adapted ModernBERT encoders trained on three primary climate corpora: academic climate text, climate-filtered web data, and synthetic climate documents.
- We conduct a large-scale empirical study of corpus composition for climate-domain encoder adaptation, evaluating 21 climate-adapted model variants across nine climate NLP benchmarks, yielding over 810 fine-tuning runs.
- We compare joint continued pretraining and parameter-space merging under matched data coverage, showing that curated academic cli-

mate corpora provide the strongest adaptation signal and that parameter-space merging outperforms joint multi-source training.

2 Related Work

2.1 Encoder-only Models in Climate NLP

Encoder-based language models remain a core component of climate NLP systems, supporting tasks such as climate-risk assessment (Diggelmann et al., 2020), sustainability disclosure analysis (Bingler et al., 2024), information retrieval (Schimanski et al., 2024), and scientific literature mining (Manivannan et al., 2024). Domain-adapted encoders such as ClimateBERT (Webersinke et al., 2021) demonstrate strong performance across climate-specific classification and retrieval tasks and remain widely adopted in climate NLP pipelines. Recent evaluations further show that fine-tuned encoders, including ClimateBERT-based models, often match or exceed the zero-shot performance of larger decoder-based models on supervised climate NLP tasks when sufficient labeled data are available (Trajanov et al., 2023; Bucher and Martini, 2024).

2.2 Continued Pretraining for Domain Adaptation

Continued pretraining, also known as domain-adaptive pretraining, is a widely adopted approach for adapting pretrained language models to specialized domains. Prior work demonstrates consistent gains from CPT in biomedical (Sounack et al., 2025), clinical (Lee et al., 2025), legal (Stammbach et al., 2026), and spatial (Singh et al., 2025) settings. Existing studies, however, primarily focus on the benefits of domain adaptation itself or the impact of factors such as training duration, optimization, and data selection (Chen et al., 2025; Shi et al., 2025). This leaves an open question of how multiple data sources should be combined during continued pretraining (Chen et al., 2025). This question is particularly challenging in climate NLP, where training data span scientific literature, policy documents, web discourse, and synthetic text with substantially different distributions.

2.3 Parameter-Space Model Merging

When multiple data sources are available, the standard approach is to combine them during training through joint continued pretraining. An alternative is to adapt separate checkpoints for each source and combine them afterward in parameter space. Early work shows that simple weight averaging can improve generalization when models share a common initialization (Wortsman et al., 2022). Subsequent methods, including Task Arithmetic (Ilharco et al., 2023), TIES-Merging (Yadav et al., 2023), and DARE (Yu et al., 2024a), introduce mechanisms for combining parameter updates from independently trained models, typically in multi-task learning settings where each model is fine-tuned on a distinct downstream task. Their use as a strategy for integrating heterogeneous domain corpora during continued pretraining remains underexplored. We investigate this setting in climate NLP by comparing parameter-space merging against joint continued pretraining under a matched data budget.

Corpus	Format	#Docs	Tokens
$\mathcal{A}$ (Academic)	XML / CSV / PDF	∼5 M	∼1.28 B
$\mathcal{F}$ (Web)	Parquet	2.59 M	∼5 B
$\mathcal{S}$ (Synthetic)	JSONL	∼20 K	∼0.14 B
Total	—	—	∼6.42 B

Table 1: High-level composition of the climate pretraining corpus across the three sources $\mathcal{A}$, $\mathcal{F}$, and $\mathcal{S}$.

3 CLIMATE-MODERNBERTS

3.1 Data Curation

High-quality and diverse training corpora are essential for effective language modeling (Longpre et al., 2024), particularly in climate NLP, where language spans scientific, policy, and public discourse. To study how corpus composition influences continued pretraining, we construct a 6.42B-token climate corpus from three primary sources: (1) academic climate data, (2) climate-filtered web data, and (3) synthetic climate text. Table 1 summarizes the high-level composition.

Academic Climate Data ($\mathcal{A}$) The academic corpus aggregates four complementary sources of climate-related text: (i) peer-reviewed journal articles spanning climate science, earth systems, and energy economics; (ii) the *ClimateNews* archive containing climate-related news articles from 2000–2022; (iii) climate-focused arXiv preprints sampled from categories including `cs`, `physics.ao-ph`, `econ`, and `q-fin.RM`; and (iv) further included reports from the Intergovernmental Panel on Climate Change (IPCC)[2] and climate handbooks (Pachauri et al., 2014). We treat these sources as a unified

[2] https://www.ipcc.ch/data/

academic corpus due to their substantial overlap in vocabulary, citation conventions, and discourse structure. Document-level layout (e.g., sections, paragraphs, and captions) is preserved whenever possible to support long-context training with ModernBERT's 8,192-token context window. After applying deduplication and decontamination preprocessing pipeline (details are described in Appendix A), we arrive at a corpus of 1.28B tokens. A detailed per-source breakdown is reported in Appendix B.

Web Data ($\mathcal{F}$) The web corpus is derived from FineWeb-Edu (Penedo et al., 2024), a large-scale high-quality web dataset containing diverse educational and informational content. We extract climate-relevant documents using a two-stage filtering pipeline. We apply a high-recall keyword filter built from 166 climate-related terms followed by deduplication (detailed in Appendix C), yielding 2.59M candidate documents.

Synthetic Climate Data ($\mathcal{S}$) To improve coverage of underrepresented climate subdomains (Marwala et al., 2023), we generate synthetic climate text conditioned on seed excerpts drawn from six in-domain sources: five academic journal subsets from $\mathcal{A}$ (*Climate Risk Management*, *Environmental Science*, *Renewable & Sustainable Energy Reviews*, *Urban Climate*, and *Weather and Climate Extremes*) together with the *ClimateNews* archive sampled across 22 years. For each seed document, we extract an 800-character passage and prompt `Qwen3.5-122B-A10B` (Qwen Team, 2026) to generate text in three climate communication perspectives inspired by Schäfer and Painter (2021): *public awareness*, *industry perspective*, and *environmental journalism*. Generation details, prompt templates, and seed statistics are provided in Appendix D.

3.2 Training Schedule

We adapt ModernBERT-base (Warner et al., 2025) following its original continued pretraining procedure. Specifically, we apply context extension (CX) followed by learning-rate decay (LRD) specialization, where the resulting CX+LRD checkpoints serve as the final climate-adapted models used throughout our main experiments. We additionally evaluate CX-only checkpoints as a case study to isolate the contribution of LRD specialization.

Initialization. Following prior domain-adaptive ModernBERT studies (Sounack et al., 2025; Lee et al., 2025), we initialize from the public ModernBERT-base context-extension stage checkpoint. This checkpoint preserves the stable optimization landscape, which facilitates continued training on new domain data. In contrast, resuming from a post-LRD checkpoint would require reintroducing warm-up (Ash and Adams, 2020) to recover from the near-zero learning rate reached during decay (Hägele et al., 2024).

Context Extension. We retain the context-extension checkpoints as an ablation to examine the contribution of LRD specialization. CX training uses a constant learning rate of $3e-4$, global batch size of 576, sequence length of 8,192, and MLM masking rate of 30%. We optimize with Decoupled StableAdamW (Wortsman et al., 2023) in BF16 mixed precision for three epochs.

LRD specialization. Following the original ModernBERT training procedure, we further specialize each CX checkpoint using the $1-\sqrt{t}$ learning-rate decay schedule (Warner et al., 2025). We use an initial learning rate of $3e-4$ and a final learning-rate factor $\alpha_f = 1e{-}3$ for three additional epochs, while keeping all other hyperparameters unchanged. The resulting CX+LRD checkpoints serve as the final climate-adapted models and are used for all subsequent corpus-composition and model-merging experiments described in §4.1.

Model Architecture and Computing. Every variant uses the ModernBERT-base architecture (150M parameters, 22 transformer layers, hidden size 768, 12 attention heads, and a 50,368-token vocabulary), with RoPE positional embedding and alternating 128-token-window local and global attention with a global block every three layers (Wu et al., 2025). Training runs on 4× NVIDIA A100 GPUs using the MosaicML Composer framework (Warner et al., 2025). Checkpoints are saved every 3,000 steps during CX training and every 1,000 steps during LRD specialization. Final checkpoints are converted to the HuggingFace Transformers FlexBERT format for release. Complete hyperparameter and architecture specifications are deferred to Appendix E.

4 Evaluation Setting

We evaluate all climate-domain variants against two complementary baselines: the ModernBERT-

base stable-phase checkpoint used for initialization and the original ClimateBERT model (Webersinke et al., 2021), a widely adopted climate-domain encoder. The ModernBERT-base comparison isolates the effect of climate-domain continued pretraining from the original model capabilities, while ClimateBERT provides context against an established climate-adapted encoder. We fine-tune all models under the same downstream protocol to ensure a controlled comparison. We do not adopt the final publicly released ModernBERT checkpoint as the primary baseline, as it has already undergone full learning-rate decay on general-domain data, which would confound assessment of domain adaptation effects; for completeness, we still fine-tune that post-LRD checkpoint (results in Appendix G).

4.1 Evaluated Models

Our training design studies two aspects of climate adaptation: (i) the composition of the climate pretraining corpus and (ii) the contribution of LRD specialization through a CX-only ablation. Let the set of domain corpora introduced in §3.1 be $\mathcal{D} = \{\mathcal{A}, \mathcal{S}, \mathcal{F}\}$, where $\mathcal{A}$ denotes the academic climate corpus, $\mathcal{S}$ the synthetic climate corpus, and $\mathcal{F}$ the climate-filtered FineWeb-Edu subset.

For any non-empty subset $\mathcal{X} \subseteq \mathcal{D}$, we denote by $\theta_{\mathcal{X}}^{\text{LRD}}$ the final climate-adapted checkpoint obtained after CX followed by LRD specialization on the union of corpora in $\mathcal{X}$. We additionally denote by $\theta_{\mathcal{X}}$ the corresponding CX-only checkpoint used for ablation analysis. We further denote by θ_{base} the stable-phase ModernBERT-base initialization from which all models are derived and against which all comparisons are made.

Joint-training variants. We train seven joint-training variants corresponding to different non-empty subsets of

$$\mathcal{D}: \begin{array}{c} \{\mathcal{A}\},\ \{\mathcal{S}\},\ \{\mathcal{F}\}, \\ \{\mathcal{A},\mathcal{S}\},\ \{\mathcal{A},\mathcal{F}\},\ \{\mathcal{S},\mathcal{F}\},\ \{\mathcal{A},\mathcal{S},\mathcal{F}\}. \end{array}$$

For each subset $\mathcal{X}$, we train the final LRD-specialized checkpoint $\theta_{\mathcal{X}}^{\text{LRD}}$ used in our main experiments. The corresponding CX-only checkpoint $\theta_{\mathcal{X}}$ is retained for ablation analysis to evaluate the contribution of LRD specialization.

Model-merging variants. We further study parameter-space merging over independently trained checkpoints. Specifically, we take the three single-source LRD-specialized checkpoints $\theta_{\{c\}}^{\text{LRD}}$ with $c \in \{\mathcal{A}, \mathcal{S}, \mathcal{F}\}$ and combine their parameter deltas relative to θ_{base} using four merging strategies: simple weight averaging (θ_{SOUP}), Task Arithmetic with scaling factors $\lambda \in \{0.5, 1.0\}$ ($\theta_{\text{TA}(\lambda)}$), TIES-Merging with drop ratios $d \in \{0.5, 0.7\}$ ($\theta_{\text{TIES}(d)}$), and DARE-TIES with the same drop ratios ($\theta_{\text{DARE}(d)}$). This trajectory yields seven merged checkpoints whose training data correspond to the full corpus union $\{\mathcal{A}, \mathcal{S}, \mathcal{F}\}$, enabling direct comparison between joint training and post-hoc parameter merging under matched data coverage.

Evaluation budget. In total, we evaluate 21 climate-adapted checkpoints against θ_{base} on the downstream climate tasks described below. To support the paired-seed ablation analysis of synthetic data, the four checkpoints in the $\{\mathcal{A}\}/\{\mathcal{A}, \mathcal{S}\}$ comparison pair (with and without LRD specialization) are fine-tuned with n=10 random seeds, whereas all remaining variants use n=3 seeds.

4.2 Downstream Tasks

We evaluate all models on nine climate NLP benchmarks. These include six single-label classification tasks: **Climate Detection** (Det.) (Bingler et al., 2024), **Climate Specificity** (Spec.) (Bingler et al., 2024), **Commitments & Actions** (Comm.) (Bingler et al., 2024), **Climate Sentiment** (Sent.) (Bingler et al., 2024), **Net Zero & Reduction** (NetZ.) (Schimanski et al., 2023a), and **TCFD Recommendations** (TCFD) (Bingler et al., 2024). We further evaluate two multi-label classification tasks, **WFB Nature** (WFB) (Schimanski et al., 2023b) and **WX-ImpactBench** (WXI) (Yu et al., 2025), as well as one retrieval benchmark, **ClimRetrieve** (Retr.) (Schimanski et al., 2024), formulated as binary relevance classification. Acronyms in parentheses are used throughout the results section (§5). Taken together, these benchmarks cover a broad range of climate NLP settings, including topical detection, disclosure analysis, climate commitments, sentiment analysis, information retrieval, and environmental impact assessment. The underlying datasets span corporate annual reports, sustainability disclosures, national policy pledges, and historical news articles, capturing both structured disclosure language and narrative climate communication (Calamai et al., 2025). Detailed dataset statistics, label definitions, and data provenance are provided in Appendix H.

Fine-tuning protocol. For each ($\theta_{\mathcal{X}}$, task) pair, we jointly fine-tune all encoder parameters together

Model	Retr.	Comm.	Det.	Spec.	Sent.	NetZ.	TCFD	WFB	WXI	Avg.
ModernBERT-base										
θ_{BASE}	$86.7_{\pm 1.9}$	$65.9_{\pm 1.3}$	$94.5_{\pm 0.3}$	$67.3_{\pm 0.7}$	$71.8_{\pm 4.1}$	$98.7_{\pm 0.6}$	$60.1_{\pm 0.7}$	$96.0_{\pm 0.4}$	$20.4_{\pm 5.0}$	$73.5_{\pm 1.7}$
Original RoBERTa model										
ClimateBERT	$82.4_{\pm 2.6}$	$70.9_{\pm 0.0}$	$97.5_{\pm 0.0}$	$67.8_{\pm 0.0}$	$76.6_{\pm 0.0}$	$99.1_{\pm 0.0}$	$51.4_{\pm 0.1}$	$92.5_{\pm 0.0}$	$3.3_{\pm 4.6}$	$72.1_{\pm 0.8}$
Climate-ModernBERT: CX + LRD specialization — $\theta_{\mathcal{X}}^{\text{LRD}}$										
$\{\mathcal{A}\}$	$81.1_{\pm 4.1}$	$68.7_{\pm 2.1}$	$93.6_{\pm 0.2}$	$67.3_{\pm 1.1}$	$77.1_{\pm 0.7}$	$98.0_{\pm 1.2}$	$61.3_{\pm 0.5}$	$96.9_{\pm 0.3}$	$25.2_{\pm 3.2}$	$74.4_{\pm 2.4}$
$\{\mathcal{S}\}$	$84.5_{\pm 2.1}$	$66.4_{\pm 3.4}$	$93.8_{\pm 0.2}$	$69.4_{\pm 0.9}$	$74.8_{\pm 2.7}$	$99.0_{\pm 0.1}$	$61.1_{\pm 0.2}$	$96.5_{\pm 0.3}$	$25.0_{\pm 2.4}$	$74.5_{\pm 1.4}$
$\{\mathcal{F}\}$	$85.0_{\pm 2.5}$	$67.4_{\pm 1.6}$	$95.8_{\pm 0.0}$	$68.7_{\pm 1.9}$	$77.5_{\pm 0.4}$	$98.7_{\pm 0.0}$	$60.3_{\pm 1.7}$	$96.8_{\pm 0.2}$	$20.3_{\pm 4.4}$	$74.5_{\pm 1.4}$
$\{\mathcal{A}, \mathcal{S}\}$	$83.2_{\pm 4.1}$	$66.1_{\pm 1.0}$	$93.2_{\pm 0.8}$	$69.5_{\pm 1.0}$	$\mathbf{78.1}_{\pm 0.5}$	$98.9_{\pm 0.2}$	$61.3_{\pm 1.0}$	$96.2_{\pm 0.4}$	$24.3_{\pm 4.9}$	$74.5_{\pm 1.5}$
$\{\mathcal{A}, \mathcal{F}\}$	$83.5_{\pm 4.4}$	$69.5_{\pm 0.6}$	$95.5_{\pm 0.6}$	$70.2_{\pm 0.7}$	$76.0_{\pm 0.4}$	$99.1_{\pm 0.1}$	$63.0_{\pm 2.5}$	$97.0_{\pm 0.5}$	$18.5_{\pm 5.0}$	$74.7_{\pm 1.6}$
$\{\mathcal{S}, \mathcal{F}\}$	$85.3_{\pm 2.2}$	$62.2_{\pm 2.3}$	$\mathbf{95.7}_{\pm 0.2}$	$71.0_{\pm 0.8}$	$75.5_{\pm 0.6}$	$97.9_{\pm 0.3}$	$59.4_{\pm 2.2}$	$96.9_{\pm 0.5}$	$23.4_{\pm 3.7}$	$74.1_{\pm 1.4}$
$\{\mathcal{A}, \mathcal{S}, \mathcal{F}\}$	$\mathbf{86.8}_{\pm 1.4}$	$60.0_{\pm 2.3}$	$95.7_{\pm 0.0}$	$\mathbf{71.8}_{\pm 1.4}$	$76.0_{\pm 1.2}$	$98.6_{\pm 0.4}$	$\mathbf{62.8}_{\pm 2.2}$	$\mathbf{97.5}_{\pm 0.2}$	$23.9_{\pm 5.9}$	$74.8_{\pm 1.7}$
Ablation: Context Extension-only checkpoints — $\theta_{\mathcal{X}}$										
$\{\mathcal{A}\}$	$85.5_{\pm 1.0}$	$\mathbf{71.8}_{\pm 1.1}$	$93.2_{\pm 0.3}$	$68.9_{\pm 1.6}$	$77.4_{\pm 1.0}$	$98.5_{\pm 0.8}$	$59.0_{\pm 0.9}$	$97.0_{\pm 0.2}$	$26.1_{\pm 3.2}$	$\mathbf{75.3}_{\pm 1.1}$
$\{\mathcal{S}\}$	$86.0_{\pm 1.8}$	$65.2_{\pm 1.0}$	$93.5_{\pm 0.8}$	$68.0_{\pm 0.6}$	$71.8_{\pm 4.1}$	$98.7_{\pm 0.4}$	$60.0_{\pm 0.7}$	$97.2_{\pm 0.2}$	$20.0_{\pm 2.1}$	$73.4_{\pm 1.3}$
$\{\mathcal{F}\}$	$85.4_{\pm 1.0}$	$64.8_{\pm 3.4}$	$94.5_{\pm 0.3}$	$68.0_{\pm 1.5}$	$76.3_{\pm 0.5}$	$\mathbf{99.2}_{\pm 0.1}$	$59.1_{\pm 1.0}$	$97.2_{\pm 0.1}$	$23.4_{\pm 5.1}$	$74.1_{\pm 1.4}$
$\{\mathcal{A}, \mathcal{S}\}$	$86.1_{\pm 1.9}$	$67.9_{\pm 1.4}$	$94.0_{\pm 0.6}$	$67.6_{\pm 0.9}$	$77.0_{\pm 2.0}$	$98.2_{\pm 0.9}$	$61.4_{\pm 0.5}$	$96.6_{\pm 0.5}$	$24.8_{\pm 3.4}$	$74.8_{\pm 1.3}$
$\{\mathcal{A}, \mathcal{F}\}$	$86.0_{\pm 0.8}$	$65.9_{\pm 0.7}$	$93.8_{\pm 0.4}$	$67.2_{\pm 1.1}$	$76.8_{\pm 1.0}$	$99.1_{\pm 0.2}$	$59.7_{\pm 1.2}$	$96.5_{\pm 0.3}$	$24.1_{\pm 3.2}$	$74.3_{\pm 1.0}$
$\{\mathcal{S}, \mathcal{F}\}$	$85.4_{\pm 1.0}$	$64.8_{\pm 3.4}$	$94.5_{\pm 0.3}$	$68.0_{\pm 1.5}$	$76.3_{\pm 0.5}$	$\mathbf{99.2}_{\pm 0.1}$	$59.1_{\pm 1.0}$	$97.2_{\pm 0.1}$	$23.4_{\pm 5.1}$	$74.2_{\pm 1.4}$
$\{\mathcal{A}, \mathcal{S}, \mathcal{F}\}$	$86.5_{\pm 1.0}$	$67.4_{\pm 1.3}$	$95.2_{\pm 0.1}$	$70.3_{\pm 0.2}$	$75.2_{\pm 0.6}$	$99.1_{\pm 0.1}$	$59.2_{\pm 1.1}$	$96.5_{\pm 0.9}$	$18.5_{\pm 5.3}$	$74.1_{\pm 1.2}$

Table 2: Per-task F_1 scores (%, mean $\pm$ standard deviation across three random seeds) for jointly trained climate-adapted variants. LRD-specialized checkpoints represent our primary climate-adapted models following the ModernBERT two-stage training procedure, while context extension-only checkpoints are reported as an ablation to isolate the contribution of LRD specialization. Binary tasks and ClimRetrieve report positive-class F_1; classification tasks report macro-F_1. Best results for each task are shown in **bold**.

with a task-specific classification head. Following Sounack et al. (2025), all variants use a shared hyperparameter configuration: learning rate 4e-5, per-device batch size 32 with 2 gradient accumulation steps (effective batch size 64), weight decay 0.01, and up to 10 training epochs with early stopping based on validation F_1. Training is performed in BF16 using the fused AdamW optimizer. We report test-set performance using F_1-based metrics. Binary classification tasks and ClimRetrieve are evaluated using positive-class F_1, while multi-class and multi-label tasks are evaluated using macro-F_1. The same fine-tuning configuration is applied to ClimateBERT.

5 Experimental Results

§5.1 examines how corpus composition affects climate adaptation and reports a CX-only ablation to analyze the contribution of LRD specialization, including a paired-seed ablation of the synthetic corpus. §5.2 compares parameter-space merging against joint training under the same data budget and uses drop-one merges to attribute per-corpus contributions.

5.1 Joint Continued Pretraining

Academic corpora provide the strongest CPT recipe, yet broader mixtures can hurt performance. As shown in Table 2, climate adaptation generally improves aggregate performance relative to θ_{BASE}, despite task-specific trade-offs. All adapted variants obtain average F_1 scores at or above the baseline, with the highest average achieved by $\theta_{\{\mathcal{A}\}}$ (75.3). Improvements are concentrated on register-sensitive tasks, including *Sentiment* (+6.3), *Commitments* (+5.9), *WXImpact-Bench* (+7.5), and *Specificity* (+4.2), while already saturated benchmarks such as *Detection* (+1.2) and *NetZero* (+0.5) shift only marginally, consistent with observations from Calamai et al. (2025).

Notably, adding broader data sources to $\mathcal{A}$ does not improve performance. The average CX-only F_1 declines from 75.3 ($\theta_{\{\mathcal{A}\}}$) to 74.8 ($\theta_{\{\mathcal{A},\mathcal{S}\}}$), and further drops to 74.1 on $\theta_{\{\mathcal{A},\mathcal{S},\mathcal{F}\}}$. This effect is particularly pronounced on *Commitments & Actions*, where adding synthetic and web corpora to the academic training set lowers performance from 71.8 to 59.5, suggesting interference from out-of-register sources (Kurfalı et al., 2025). The overall pattern

Task	CX-only ($\theta_{\mathcal{X}}$) Δ	p_t	p_w	LRD specialization ($\theta_{\mathcal{X}}^{\text{LRD}}$) Δ	p_t	p_w
Retrieval	↑ 0.6	.31	.26	↑ 2.2	.66	.77
Commitments	↓ 3.9	< **.001**	**.002**	↓ 2.6	**.016**	**.037**
Detection	↑ 0.8	**.012**	**.020**	↓ 0.4	.18	.38
Specificity	↓ 1.3	.07	.13	↑ 2.2	**.001**	**.002**
Sentiment	↓ 0.4	.43	.56	↑ 0.9	**.007**	**.006**
NetZero	↓ 0.3	.51	.63	↑ 0.9	.07	.08
TCFD	↑ 2.4	< **.001**	**.002**	↓ 0.0	.96	.92
WFB	↓ 0.4	.06	.08	↓ 0.8	**.001**	**.004**
WXImpact	↓ 1.3	.25	.16	↓ 0.9	.49	.63

Table 3: Paired-seed synthetic-data ablation (n=10). Δ denotes the mean effect of adding $\mathcal{S}$; p_t and p_w are paired-t and Wilcoxon signed-rank p-values. Bold indicates significance (α=0.05) under both tests.

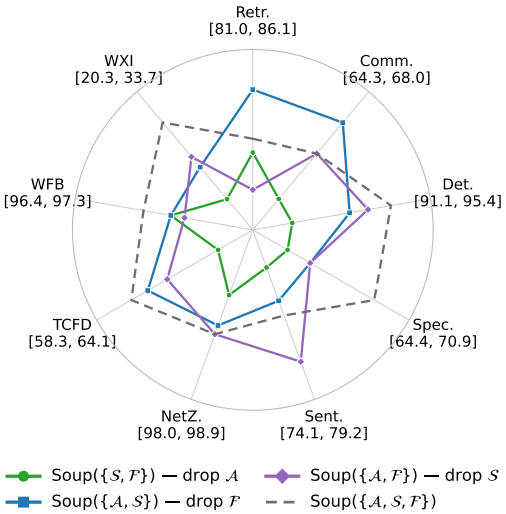

Figure 2: Per-task F_1 of the three drop-one Soups against the full Soup($\{\mathcal{A}, \mathcal{S}, \mathcal{F}\}$) baseline (dashed grey). Axes are scaled per task to the $[\min, \max]$ range across the four variants.

further indicates that adaptation benefits are driven by source–task alignment, consistent with observations from domain-adaptive pretraining studies (Gururangan et al., 2020). Different data combinations excel on different benchmarks. Considering the best-performing configuration across the evaluated adaptation settings, $\theta_{\{\mathcal{A}\}}$ achieves the highest scores on *Commitments* and *WXImpactBench*.

Synthetic climate text exhibits task-dependent effects. To isolate the effect of synthetic data, we compare $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{A},\mathcal{S}}$ under both the final LRD-specialized models and the CX-only ablation setting using n=10 shared fine-tuning seeds. As shown in Table 3, the effect of $\mathcal{S}$ is strongly task dependent. Adding synthetic data significantly improves several framework-oriented benchmarks, including *TCFD* (+2.4, $p_t < .001$) and *Specificity* (+2.2, $p_t = .001$), while consistently degrading *Commitments & Actions* across both training regimes (−3.9, $p_t < .001$). In contrast, retrieval performance remains statistically unchanged. Overall, these results suggest that synthetic climate text benefits taxonomy- and framework-driven classification tasks, but transfers less effectively to tasks requiring finer-grained discourse and commitment understanding.

5.2 Parameter-space Merging

We evaluate model merging both as a data-attribution method and as an alternative to joint training under matched data coverage. Table 4 compares the three single-source LRD-specialized checkpoints, their jointly trained counterpart, and seven merged variants, while Figure 2 presents drop-one Soup ablations relative to the full $\theta_{\text{SOUP}}(\{\mathcal{A}, \mathcal{S}, \mathcal{F}\})$ model. We additionally evaluate model merging using CX-only checkpoints ($\theta_{\mathcal{X}}$). These results are reported in Appendix F.

Dropping $\mathcal{A}$ hurts more than dropping either other corpus. Figure 2 provides a complementary perspective on corpus importance through drop-one simple weight averaging ablations. Across nearly all benchmarks, removing the academic corpus (SOUP($\{\mathcal{S}, \mathcal{F}\}$), green) results in the largest performance degradation, reducing average F_1 by 4.0 points relative to the full SOUP($\{\mathcal{A}, \mathcal{S}, \mathcal{F}\}$). In contrast, excluding either the synthetic or web corpus leads to substantially smaller average degradations (−1.5), and in some cases even improves performance on individual benchmarks. These results provide independent evidence for the findings in §5.1: academic climate data contributes the strongest and most broadly transferable adaptation signal, whereas the benefits of synthetic and web data are comparatively task-dependent.

Interestingly, this importance is not reflected by task-vector magnitude alone. Although the FineWeb-specialized checkpoint produces the largest parameter update ($\|\boldsymbol{\delta}_F\| \approx 43$ per layer, compared with approximately 16 for $\mathcal{A}$ and $\mathcal{S}$), removing $\mathcal{A}$ causes by far the largest performance degradation. This observation indicates that larger parameter updates do not necessarily translate into greater downstream utility, highlighting an interesting direction for future work on interpreting adaptation dynamics in continued pretraining.

Model	Retr.	Comm.	Det.	Spec.	Sent.	NetZ.	TCFD	WFB	WXI	Avg.
Joint training on the union — $\theta^{\text{LRD}}_{\{\mathcal{A},\mathcal{S},\mathcal{F}\}}$										
$\{\mathcal{A},\mathcal{S},\mathcal{F}\}$	**86.8**$_{\pm 1.4}$	60.0$_{\pm 2.3}$	95.7$_{\pm 0.0}$	**71.8**$_{\pm 1.4}$	76.0$_{\pm 1.2}$	98.6$_{\pm 0.4}$	62.8$_{\pm 2.2}$	**97.5**$_{\pm 0.2}$	23.9$_{\pm 5.9}$	74.8$_{\pm 1.7}$
Parameter-space merging of the three single-source components										
θ_{SOUP}	83.6$_{\pm 1.6}$	66.5$_{\pm 3.7}$	95.4$_{\pm 0.6}$	70.9$_{\pm 2.1}$	76.7$_{\pm 1.4}$	98.9$_{\pm 0.0}$	**64.1**$_{\pm 2.2}$	97.3$_{\pm 0.1}$	**33.7**$_{\pm 5.3}$	**76.3**$_{\pm 1.9}$
$\theta_{\text{TA}(0.5)}$	80.9$_{\pm 5.7}$	67.5$_{\pm 1.8}$	94.7$_{\pm 0.4}$	67.9$_{\pm 0.2}$	71.0$_{\pm 0.2}$	99.1$_{\pm 0.0}$	60.3$_{\pm 2.4}$	95.8$_{\pm 0.5}$	24.8$_{\pm 6.0}$	73.6$_{\pm 1.9}$
$\theta_{\text{TA}(1.0)}$	85.0$_{\pm 0.4}$	71.2$_{\pm 1.2}$	94.9$_{\pm 0.5}$	68.5$_{\pm 0.7}$	77.0$_{\pm 3.1}$	**99.2**$_{\pm 0.1}$	61.3$_{\pm 0.7}$	97.1$_{\pm 0.2}$	27.2$_{\pm 5.1}$	75.7$_{\pm 1.3}$
$\theta_{\text{TIES}(0.5)}$	86.1$_{\pm 1.4}$	67.6$_{\pm 2.5}$	93.6$_{\pm 0.6}$	68.9$_{\pm 0.9}$	**77.1**$_{\pm 0.7}$	99.0$_{\pm 0.1}$	61.4$_{\pm 0.6}$	96.5$_{\pm 1.2}$	28.5$_{\pm 1.7}$	75.4$_{\pm 1.1}$
$\theta_{\text{TIES}(0.7)}$	85.1$_{\pm 2.8}$	**71.9**$_{\pm 2.7}$	94.9$_{\pm 0.1}$	70.7$_{\pm 0.1}$	75.6$_{\pm 1.6}$	99.0$_{\pm 0.2}$	59.7$_{\pm 1.1}$	96.7$_{\pm 1.3}$	26.9$_{\pm 3.8}$	75.6$_{\pm 1.5}$
$\theta_{\text{DARE}(0.5)}$	86.3$_{\pm 0.9}$	67.2$_{\pm 0.4}$	94.8$_{\pm 0.3}$	69.2$_{\pm 0.2}$	74.3$_{\pm 0.1}$	98.9$_{\pm 0.2}$	57.8$_{\pm 1.3}$	97.1$_{\pm 0.2}$	26.9$_{\pm 2.3}$	74.7$_{\pm 0.7}$
$\theta_{\text{DARE}(0.7)}$	85.5$_{\pm 0.9}$	63.3$_{\pm 1.1}$	94.7$_{\pm 0.1}$	67.6$_{\pm 0.3}$	74.1$_{\pm 2.0}$	98.9$_{\pm 0.2}$	61.6$_{\pm 1.0}$	97.0$_{\pm 0.3}$	26.2$_{\pm 6.2}$	74.3$_{\pm 1.3}$

Table 4: Single-source LRD components, joint training on the corpus union, and seven parameter-space merges the same effective corpus, reporting per-task F1 (%, mean ± standard deviation across three random seeds).

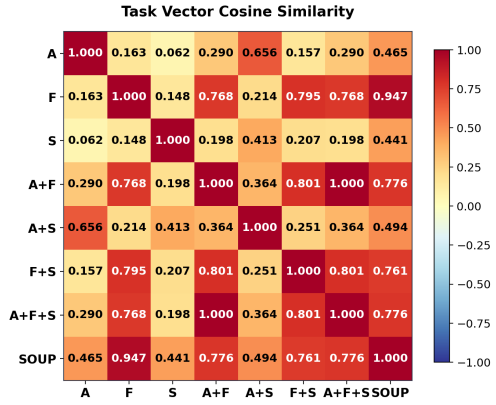

Figure 3: Cosine similarity between source task vectors.

Merging surpasses both the strongest single source and joint training on the same data. Across the seven merged models in Table 4, five outperform the strongest single-source LRD checkpoint ($\theta^{\text{LRD}}_{\{\mathcal{S}\}}$, 74.5), with uniform weight averaging performing best overall. θ_{SOUP} achieves 76.3 average F_1, improving by +1.8 over the best individual source and by +1.7 over joint training on the same effective corpus (74.6). The gains are most pronounced on *Commitments* (+7.0) and *WXImpactBench* (+9.3), while joint training retains only narrow advantages on *Detection*, *Specificity*, and *Retrieval*. Although weight averaging is known to improve performance in supervised fine-tuning settings (Wortsman et al., 2022), evidence in the context of multi-source continued pretraining remains limited.

To better understand this behavior, we analyze the task vectors $\boldsymbol{\delta}_i = \theta_{\text{LRD},i} - \theta_{base}$ of the three source-specific checkpoints. Figure 3 shows that the resulting parameter updates are nearly orthogonal, with pairwise cosine similarities ranging from 0.06 to 0.18 between $\mathcal{A}$, $\mathcal{F}$, and $\mathcal{S}$. Such weak alignment indicates that each corpus contributes largely distinct parameter updates, consistent with observations from task-vector analyses in fine-tuning (Ilharco et al., 2023). The complementarity of these updates provides a plausible explanation for why parameter-space merging can recover information from multiple sources more effectively than joint optimization.

6 Conclusion

We introduce CLIMATE-MODERNBERT, a collection of climate-adapted encoder models obtained through ModernBERT's continued pretraining procedure with LRD specialization of ModernBERT-Base on academic climate corpora, climate-filtered FineWeb-Edu web data, and synthetic climate text. Across 21 adapted variants and nine climate NLP benchmarks, we reveal three main findings: *(1)* academic climate data provides the strongest adaptation signal, while broader corpus mixtures often fail to improve and can degrade performance; *(2)* synthetic climate text exhibits task-dependent effects, benefiting taxonomy-driven tasks yet hurting performance on pragmatically grounded benchmarks; and *(3)* simple weight averaging of independently specialized checkpoints achieves the best overall performance (76.3 average F_1), outperforming both single-source models and joint training on the union of the full corpora.

These results highlight the importance of corpus composition in domain-adaptive pretraining and suggest that parameter-space merging is a simple and effective alternative to joint multi-source training. We release all CLIMATE-MODERNBERT variants, training checkpoints, and data-processing pipelines to support future research in climate NLP and domain-adaptive encoder pretraining.

Limitations

Our study has several limitations. To the best of our knowledge and recent literature (Calamai et al., 2025), current climate NLP benchmarks remain primarily designed around sentence- or passage-level classification tasks over corporate disclosures, sustainability reports, and policy documents. Although ModernBERT is designed for long-context processing, existing benchmark suites provide limited opportunities to evaluate long-document understanding, cross-section reasoning, or information integration across extended contexts beyond ClimRetrieve. Consequently, the potential benefits of long-context architectures cannot yet be fully assessed within the current climate NLP evaluation landscape. We therefore view the development of standardized long-context and document-centric climate NLP benchmarks as an important direction for future research.

Furthermore, CLIMATE-MODERNBERT collection is trained exclusively on English climate text and adapted from a single encoder family, ModernBERT-Base. Our analysis focuses on climate NLP and evaluates adaptation strategies using climate-specific corpora and benchmarks. Therefore, the observed effects of corpus composition and parameter-space merging should be interpreted as findings within this domain rather than universal principles of domain adaptation. Extending this analysis to other specialized domains, such as biomedical or legal NLP, remains an important direction for future work. More broadly, our analysis is restricted to continued pretraining. We do not examine instruction-tuned encoder variants (Clavié et al., 2025), whose representations and adaptation dynamics may differ substantially from base models. Understanding how corpus composition and parameter-space merging interact with instruction tuning is thus an important direction for future work.

Ethical Considerations

Our academic climate corpus is compiled from climate-related scientific publications, reports, and other copyrighted sources. As climate NLP increasingly supports decision-making and policy analysis, we respect licensing and intellectual property restrictions by releasing only trained model checkpoints and data-processing pipelines where permitted, rather than the raw text data.

Climate change is a societal challenge that requires collaboration among governments, industry, researchers, and citizens (Reid and Toffel, 2009). We hope that climate-adapted language models can support information retrieval, document analysis, and evidence synthesis across large climate-related document collections.

References

Jordan Ash and Ryan P Adams. 2020. On warm-starting neural network training. *Advances in neural information processing systems*, 33:3884–3894.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv preprint arXiv:2406.08660*.

Tom Calamai, Oana Balalau, and Fabian Suchanek. 2025. Benchmarking the benchmarks: Reproducing climate-related nlp tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17967–18009.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 2898–2904.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Wayne Xin Zhao, Zhicheng Dou, Jiaxin Mao, and 1 others. 2025. Towards effective and efficient continual pre-training of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5779–5795.

Benjamin Clavié, Nathan Cooper, and Benjamin Warner. 2025. It is all in the [mask]: Simple instruction-tuning enables bert-like masked language models as generative classifiers. *Natural Language Processing Journal*, 11:100150.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8342–8360.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna B Allal, Leandro Von Werra, and Martin Jaggi. 2024. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in nlp research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*.

Murathan Kurfalı, Shorouq Zahra, Joakim Nivre, and Gabriele Messori. 2025. Climateeval: A comprehensive benchmark for nlp tasks related to climate change. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 194–207.

Simon A Lee, Anthony Wu, and Jeffrey N Chiang. 2025. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and 1 others. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. 2024. Climaqa: An automated evaluation framework for climate question answering models. *arXiv preprint arXiv:2410.16701*.

Tshilidzi Marwala, Eleonore Fournier-Tombs, and Serge Stinckwich. 2023. The use of synthetic data to train ai models: Opportunities and risks for sustainable development. *arXiv preprint arXiv:2309.00652*.

NVIDIA. 2024. Nemo curator: Gpu-accelerated data curation for training ai models.

Rajendra K Pachauri, Myles R Allen, Vicente R Barros, John Broome, Wolfgang Cramer, Renate Christ, John A Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, and 1 others. 2014. *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Ipcc.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

BiChen Rao and Erkang Zhu. 2016. Searching web data using minhash lsh. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2257–2258.

Erin M Reid and Michael W Toffel. 2009. Responding to public and private politics: Corporate disclosure of climate change strategies. *Strategic management journal*, 30(11):1157–1178.

Zacharias Sautner, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang. 2023. Firm-level climate change exposure. *The Journal of Finance*, 78(3):1449–1498.

Mike S Schäfer and James Painter. 2021. Climate journalism in a changing media ecosystem: Assessing the production of climate change-related news around the world. *Wiley Interdisciplinary Reviews: Climate Change*, 12(1):e675.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023a. Climatebert-netzero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. Climretrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524.

Tobias Schimanski, Chiara Colesanti Senni, Glen Gostlow, Jingwei Ni, Tingyu Yu, and Markus Leippold. 2023b. Exploring nature: Datasets and models for analyzing nature-related disclosures. *arXiv preprint arXiv:2312.17337*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Amrendra Singh, Maulik Shah, Dharshan Sampath, Javis AI Team, and 1 others. 2025. Spatial modernbert: Spatial-aware transformer for table and key-value extraction in financial documents at scale. *arXiv preprint arXiv:2507.08865*.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards answering climate questionnaires from unstructured climate reports. *arXiv preprint arXiv:2301.04253*.

Dominik Stammbach, Kylie Zhang, Patty Liu, Nimra Nadeem, Inyoung Cheong, Lucia Zheng, and Peter Henderson. 2026. Legal retrieval for public defenders. *Preprint*, arXiv:2601.14348.

Dimitar Trajanov, Gorgi Lazarev, Ljubomir Chitkushev, and Irena Vodenska. 2023. Comparing the performance of chatgpt and state-of-the-art climate nlp models on climate-related text classification tasks. In *E3S Web of Conferences*, volume 436, page 02004. EDP Sciences.

Francesco S Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. *arXiv preprint arXiv:2012.00483*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. 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 (Volume 1: Long Papers)*, pages 2526–2547.

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

Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. 2023. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36:10271–10298.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carber, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2025. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024a. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*.

Yinan Yu, Samuel Scheidegger, Jasmine Elliott, and Åsa Löfgren. 2024b. climatebug: A data-driven framework for analyzing bank reporting through a climate lens. *Expert Systems with Applications*, 239:122162.

Yongan Yu, Qingchen Hu, Xianda Du, Jiayin Wang, Fengran Mo, and Renée Sieber. 2025. Wximpactbench: A disruptive weather impact understanding benchmark for evaluating large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4016–4035.

A Data Preprocessing Pipeline

Both corpora ($\mathcal{A}$, $\mathcal{S}$) pass through a shared deduplication and decontamination pipeline implemented with NeMo Curator (NVIDIA, 2024). Raw documents are read from JSONL and normalized: Unicode characters are reformatted to a canonical encoding, repeated whitespace is collapsed while paragraph breaks are preserved, and URLs are stripped. The climate-filtered web data ($\mathcal{F}$) is preprocessed with MinHash–LSH (Rao and Zhu, 2016) configured with 24-character n-grams, 20 hash bands, and 13 MinHash signatures per band, retaining pairs that share approximately 80% of their tokens at a tractable computational cost.

B Academic Climate Data: Sources and Statistics

The academic climate corpus combines four streams of in-domain text: peer-reviewed journal articles, year-aggregated climate news, climate-focused arXiv preprints, and a small set of climate handbooks. Table 5 reports a category-level summary with the format, number of source titles, approximate document count, and raw token contribution (before deduplication) from each stream.

Acquisition and preprocessing. We parse journal articles in structured XML directly, retain section labels and captions, and exclude reference lists, author affiliations, and copyright boilerplate. The *ClimateNews* archive is distributed as twenty-three annual CSV shards (one per year, 2000–2022), each of which we filter for climate relevance using the keyword list of Appendix C before tokenization. arXiv preprints are downloaded as PDF and converted to text with a layout-aware extractor that preserves section hierarchy and detects figure/table captions; we restrict the pool to climate-relevant categories (`physics.ao-ph`, `econ.GN`, `q-fin.RM`, and climate-tagged `cs` papers). The climate handbooks are processed with the same PDF pipeline. Across all four streams we apply boilerplate stripping, English-only language identification, and a 200-token minimum-length filter before documents enter the deduplication pipeline.

Licensing and ethics. Peer-reviewed articles are accessed under publisher licenses held by our institution; arXiv preprints are released by their authors under permissive arXiv terms; the news shards and climate handbooks were collected and processed for non-commercial research use. The corpus is used solely for research on domain-adaptive pretraining and is not redistributed; only model checkpoints, the keyword list, and aggregate per-source statistics are released. The corpus contains only text already in the public or licensed scientific record and does not specifically target personally identifiable information.

C Web Data: Keyword List for FineWeb-Edu Filtering

The web-data pipeline retains a FineWeb-Edu document if any of the case-insensitive terms below appears in the document (for full list please refer to our github repository). The terms are extracted from the climate concept in General Multilingual Environment Thesaurus (GEMET)[3] and Sautner et al. (2023). For deduplication we used the NeMo Curator (NVIDIA, 2024) deduplication pipeline.

Keywords `climate change, global warming, greenhouse gas, GHG, carbon dioxide, CO₂, methane, nitrous oxide, carbon budget, net zero, decarbonization, renewable energy, carbon capture, carbon sequestration, negative emissions, emissions trading, carbon pricing, carbon neutrality, integrated assessment model, climate scenario, climate modelling, IPCC, El Niño, ENSO, radiative forcing, radiative imbalance, arctic amplification, ocean heat content, sea level rise, permafrost, cryosphere, monsoon, flood risk, storm surge, extreme weather, heatwave, drought, climate hazard, loss and damage, loss and damage fund, 1.5°C, 2°C, Paris Agreement, global stocktake, nationally determined contribution, NDC, just transition, transition pathway, climate litigation, climate governance, climate finance, climate justice.`

D Synthetic Climate Data: Seed Pool, Prompts, and Generation

Table 6 reports the seed pool used for synthetic generation. The five academic seed pools are full subsets of the corresponding journals in $\mathcal{A}$ (§3.1); the *ClimateNews* pool is sampled at 200 documents per annual shard. For each seed document, we extract the first 800 characters of cleaned body text and substitute it into one of three style templates below. Each (seed, style) pair is sent to `Qwen3.5-122B-A10B` once; the generator is decoded with temperature 0.6, top-p 0.95, and a bud-

[3] https://www.eionet.europa.eu/gemet/en/themes/

Source / Category	Fmt	Srcs	Docs	Raw
Peer-reviewed journals				
Climate / Earth-science journals *e.g., Climate Risk Mgmt, Urban Climate, Env. Sci. & Policy, Weather & Climate Extremes, Atmospheric Env., J. of Climate, Sci. Total Env., Earth-Science Reviews, Global Env. Change.*	XML	∼25 titles	∼30K	∼1.70B
Energy / resource / env. economics *e.g., Energy Economics, Energy Policy, Ecological Economics, Resources Policy, Resource & Energy Econ., Renewable & Sustainable Energy Reviews.*	XML	∼12 titles	∼15K	∼0.80B
Finance / macro / accounting (climate-relevant) *e.g., J. Banking & Finance, J. Financial Economics, J. Public Econ., J. Corporate Finance, Accounting Org. & Society.*	XML	∼45 titles	∼40K	∼1.40B
Other in-domain sources				
ClimateNews archive (2000–2022)	CSV	23 shards	∼5M articles	∼0.60B
Climate arXiv preprints (physics, econ, cs)	PDF→text	—	>1K	∼0.12B
Climate handbooks	PDF→text	7 vols	7 books	∼0.025B
Total	—	—	—	∼4.65B

Table 5: Per-category composition of the academic climate corpus $\mathcal{A}$. `Srcs`: number of unique source titles; `Docs`: number of source documents; `Raw`: token count before deduplication.

Category	Seeds
Climate Risk Management	654
Environmental Science	3,077
Renewable & Sustainable Energy Reviews	1,028
Urban Climate	541
Weather and Climate Extremes	685
ClimateNews (2000, 2005, 2017, 2022)	4×200
Total	**6,785**

Table 6: Seed pool used to condition synthetic generation. Each seed yields three outputs (one per style); the resulting raw synthetic corpus before deduplication contains approximately 20K passages.

Task	θ_{BASE} (stable-phase)	ModernBERT (post-LRD)
Retr.	$86.7_{\pm 1.9}$	$83.1_{\pm 1.5}$
Comm.	$65.9_{\pm 1.3}$	$60.6_{\pm 1.5}$
Det.	$94.5_{\pm 0.3}$	$94.4_{\pm 0.4}$
Spec.	$67.3_{\pm 0.7}$	$67.4_{\pm 2.0}$
Sent.	$71.8_{\pm 4.1}$	$74.9_{\pm 2.5}$
NetZ.	$98.7_{\pm 0.6}$	$98.9_{\pm 0.3}$
TCFD	$60.1_{\pm 0.7}$	$57.9_{\pm 0.3}$
WFB	$96.0_{\pm 0.4}$	$96.8_{\pm 0.1}$
WXI	$20.4_{\pm 5.0}$	$20.2_{\pm 0.7}$
Avg.	$73.5_{\pm 1.7}$	$72.7_{\pm 1.0}$

Table 7: Per-task F_1 comparison between the pre-LRD stable-phase ModernBERT-base (θ_{BASE}) used as initialization throughout the paper and the final publicly released (post-LRD) ModernBERT-base checkpoint. Both models are fine-tuned on the nine climate benchmarks under the same protocol (mean ± standard deviation across three random seeds).

get of 1,024 new tokens. All generations share the same system prompt: *"You are an expert climate researcher and journalist capable of generating high-quality, factual content about climate change based on real research excerpts."* The `ClimateNews` category reuses the same three templates with the lead-in reworded to "climate news excerpt" / "climate news piece".

Style 1 — Public awareness.

> *Based on this climate research excerpt: "<EXTRACT>".*
> *Create an accessible article about climate change that: (i) **Explains Complex Concepts**—break scientific terms into everyday language; (ii) **Local Relevance**—connect global climate issues to local community impacts; (iii) **Practical Actions**—suggest concrete steps individuals can take; (iv) **Human Stories**—include relatable examples and potential human impacts. Write in an engaging, informative tone suitable for general public awareness; focus on making climate science understandable and actionable.*

Style 2 — Industry perspective.

> *Using this climate-related research as context: "<EXTRACT>".*
> *Develop a comprehensive industry analysis covering: (i) **Sector Impact**—how different industries are affected by or contributing to climate change; (ii) **Innovation & Technology**—emerging solutions and technological adaptations; (iii) **Business Strategy**—corporate responses and sustainable business models; (iv) **Investment Trends**—green finance and ESG considerations. Write*

Training Configuration		Model Architecture	
Context Extension		**ModernBERT-Base**	
Data	Full climate corpus (§3.1)	Parameters	150M
Epochs	3	Layers	22
Learning rate	3×10^{-4} (constant)	Hidden size	768
Warm-up	0 epochs	Intermediate size	1,152
Global batch size	576	Attention heads	12 (head dim 64)
Per-GPU batch size	72	Vocabulary size	50,368
Microbatch size	12	Position encoding	RoPE (base 160K)
Sequence length	8,192	Attention pattern	Alternating local/global
MLM masking rate	30%	Sliding window	128
Optimizer	StableAdamW	Global attention	Every 3 layers
β_1, β_2	0.9, 0.98	Optimizations	FlashAttention, unpadding, `torch.compile`
ϵ	10^{-6}		
Weight decay	10^{-5}	**Compute & Checkpointing**	
Precision	BF16 (AMP)	Hardware	$4\times$ NVIDIA A100
LRD Specialization		Framework	MosaicML Composer
Data	Variant-specific subset	Mixed precision	BF16
Epochs	3	Gradient checkpointing	Enabled
LR schedule	$1 - \sqrt{t}$ decay	Checkpoint (Context Extension)	Every 3,000 batches
Initial LR	3×10^{-4}	Checkpoint (LRD Specialization)	Every 1,000 batches
Final LR factor (α_f)	0.001	Release format	HF Transformers FlexBERT
Other settings	Same as context entension phase		

Table 8: Training configuration, architecture, and compute details for all CLIMATE-MODERNBERT variants.

from a business and industry perspective, highlighting opportunities and challenges.

Style 3 — Environmental journalism.

Drawing from this climate research piece: "<EXTRACT>".
Create an in-depth environmental journalism article that: (i) ***Investigative Depth****—explore underlying causes and systemic issues; (ii)* ***Ecosystem Impact****—detail effects on biodiversity, habitats, and natural systems; (iii)* ***Climate Justice****—address equity and vulnerable-population concerns; (iv)* ***Solution Spotlight****—highlight successful interventions and best practices. Write in an investigative journalism style that combines factual reporting with compelling narrative.*

E Training Setup: Full Configuration

Table 8 reports the complete hyperparameter and architecture specification for every continued-pretraining run. All variants differs only in the data subset (see Table 2) and the learning-rate schedule. Final Composer-format checkpoints are converted to the Hugging Face Transformers *FlexBERT* implementation for release.

F Context Extension-only Model Merging Analysis

To further analyze the effect of the LRD specialization stage on model merging, we additionally evaluate merging strategies using context-extension-only checkpoints ($\theta_{\mathcal{X}}$). Unlike the main experiments, which use final CX+LRD checkpoints following the ModernBERT training procedure, this analysis isolates whether the observed benefits of parameter-space merging depend on the final LRD adaptation stage. We evaluate three merging strategies: Linear merging, F-half merging, and Norm-balanced merging. Results are shown in Table 10.

Overall, CX-only merging achieves competitive performance across climate NLP benchmarks, suggesting that the effectiveness of model merging is not exclusively dependent on LRD specialization. However, the LRD-specialized checkpoints remain our primary setting because they represent the final adapted models following the established ModernBERT training pipeline.

G Comparison Against the Post-LRD Public ModernBERT Checkpoint

All climate-adapted variants in the main paper are initialized from the pre-LRD stable-phase ModernBERT-base checkpoint, denoted θ_{BASE}. For completeness, we also fine-tune the final publicly released ModernBERT-base checkpoint — the post-LRD release that downstream practitioners typically use — on the same nine climate benchmarks under the identical fine-tuning protocol described in §4.2. Table 7 reports per-task F_1 scores side-by-

Task	Type	#C	#Train	#Test	Labels
Single-label classification					
Climate Detection (Bingler et al., 2024)	Binary	2	1,170	400	*climate/not*: passage is related to climate change or environmental topics.
Climate Specificity (Bingler et al., 2024)	Binary	2	900	320	*specific/non-specific*: contains concrete, firm-specific commitments or measurable targets vs. vague phrasing.
Commitments & Actions (Bingler et al., 2024)	Binary	2	900	320	*yes/no*: states a forward-looking pledge or describes a realized climate action.
Climate Sentiment (Bingler et al., 2024)	Multi-class	3	900	320	*risk/neutral/opportunity*: evaluative stance toward climate change in corporate or policy discourse.
Net Zero & Reduction (Schimanski et al., 2023a)	Multi-class	3	2,753	344	*net-zero/reduction/none*: paragraph contains a net-zero target, a reduction target, or no target.
TCFD Recommendations (Bingler et al., 2024)	Multi-class	4	1,170	400	*governance/strategy/risk/metrics-and-targets*: TCFD disclosure pillar.
Multi-label classification					
WFB Nature (Schimanski et al., 2023b)	Multi-label	4	1,760	220	*water/forest/biodiversity/nature*: TNFD-aligned nature dimensions co-occurring in the passage.
WXImpactBench (Yu et al., 2025)	Multi-label	6	970	208	*infrastructural/political/financial/ecological/agricultural/human-health*: societal impact categories of weather events.
Retrieval (binary relevance)					
ClimRetrieve (Schimanski et al., 2024)	Retrieval	2	6,800	850	*relevant/not*: passage answers the climate question; relevance binarized at threshold ≥ 1 on the original ordinal scale.

Table 9: Overview of downstream evaluation tasks. *#C* denotes the number of target classes or labels; *#Train* and *#Test* report fine-tuning split sizes; the *Labels* column lists the label set followed by a short definition.

side with θ_{BASE}.

The post-LRD checkpoint averages 72.7 F_1, 0.8 points below θ_{BASE} (73.5). Per-task, the largest regressions are on *Retrieval* (-3.6), *Commitments & Actions* (-5.3), and *TCFD* (-2.2); modest gains appear on *Sentiment* ($+3.1$), *WFB* ($+0.8$), *NetZero* ($+0.2$), and *Specificity* ($+0.1$); *Detection* (-0.1) and *WXImpactBench* (-0.2) are essentially unchanged.

H Downstream Tasks: Detailed Descriptions

This appendix expands on the task list summarized in §4.2. Table 9 reports the type, number of classes, fine-tuning split sizes, and label semantics for each of the nine benchmarks; the paragraphs below describe each task's label space, the linguistic phenomenon it targets, and, where relevant, the construction pipeline and annotation provenance of the underlying dataset.

Single-label classification. **Climate Detection** (Bingler et al., 2024) is a binary task that determines whether a passage is climate-related, requiring the model to distinguish domain-relevant content from superficially similar but topically unrelated text. **Climate Specificity** (Bingler et al., 2024) classifies climate-related statements as either *specific* or *vague*, where specific statements contain concrete commitments or measurable targets; this distinction is central to assessing the substantiveness of corporate disclosures and to detecting greenwashing. **Climate Commitments and Actions** (Bingler et al., 2024) categorizes text into forward-looking pledges versus descriptions of implemented measures, operationalizing the gap between corporate rhetoric and realized climate action. **Climate Sentiment** (Bingler et al., 2024) assigns one of three labels — *risk*, *opportunity*, or *neutral* — to climate-related paragraphs, capturing the evaluative stance toward climate change in corporate and policy discourse. **Net Zero and Reduction Targets** (Schimanski et al., 2023a) classifies text into *net-zero target*, *reduction target*, or *no target*, based on an expert-annotated dataset of $\sim$3.5K samples drawn from corporate reports and national pledges. **TCFD Recommendations** (Bingler et al., 2024) assigns text to the disclosure categories of the Task Force on Climate-related Financial Disclosures, covering governance, strategy, risk management, and metrics and targets; the multi-class formulation requires the model to differentiate structurally similar but thematically distinct disclosure dimensions.

Model	Retr.	Comm.	Det.	Spec.	Sent.	NetZ.	TCFD	WFB	WXI	Avg.
LRD-specialized merging ($\theta_{\mathcal{X}}^{\text{LRD}}$)										
θ_{SOUP}	$83.6_{\pm 1.6}$	$66.5_{\pm 3.7}$	$95.4_{\pm 0.6}$	$70.9_{\pm 2.1}$	$76.7_{\pm 1.4}$	$98.9_{\pm 0.0}$	$\mathbf{64.1}_{\pm 2.2}$	$\mathbf{97.3}_{\pm 0.1}$	$\mathbf{33.7}_{\pm 5.3}$	$\mathbf{76.3}_{\pm 1.9}$
$\theta_{\text{TA}(0.5)}$	$80.9_{\pm 5.7}$	$67.5_{\pm 1.8}$	$94.7_{\pm 0.4}$	$67.9_{\pm 0.2}$	$71.0_{\pm 0.2}$	$99.1_{\pm 0.0}$	$60.3_{\pm 2.4}$	$95.8_{\pm 0.5}$	$24.8_{\pm 6.0}$	$73.6_{\pm 1.9}$
θ_{NORM}	$84.3_{\pm 1.8}$	$67.3_{\pm 1.7}$	$94.1_{\pm 1.1}$	$67.5_{\pm 1.0}$	$78.0_{\pm 1.5}$	$98.2_{\pm 0.1}$	$61.4_{\pm 0.7}$	$96.4_{\pm 0.3}$	$23.8_{\pm 3.4}$	$74.6_{\pm 1.3}$
CX-only merging ($\theta_{\mathcal{X}}$) – Ablation										
$\theta_{\text{SOUP}}^{\text{CX}}$	$84.5_{\pm 1.6}$	$66.5_{\pm 2.6}$	$94.1_{\pm 0.3}$	$69.6_{\pm 0.7}$	$78.2_{\pm 2.8}$	$98.3_{\pm 0.1}$	$61.3_{\pm 1.5}$	$96.5_{\pm 0.2}$	$22.5_{\pm 8.3}$	$74.2_{\pm 1.5}$
$\theta_{\text{TA}(0.5)}^{\text{CX}}$	$86.4_{\pm 0.8}$	$67.1_{\pm 1.6}$	$94.7_{\pm 1.1}$	$68.2_{\pm 1.0}$	$78.6_{\pm 3.1}$	$98.1_{\pm 0.1}$	$61.0_{\pm 1.0}$	$96.3_{\pm 0.2}$	$20.9_{\pm 7.1}$	$74.6_{\pm 1.8}$
$\theta_{\text{NORM}}^{\text{CX}}$	$86.5_{\pm 0.4}$	$70.1_{\pm 1.7}$	$94.4_{\pm 1.1}$	$68.8_{\pm 1.0}$	$77.8_{\pm 1.5}$	$98.3_{\pm 0.5}$	$61.2_{\pm 0.7}$	$96.5_{\pm 0.5}$	$26.8_{\pm 2.5}$	$75.9_{\pm 1.2}$

Table 10: Model merging comparison between LRD-specialized and context-extension-only (CX-only) checkpoints. All values report F_1 scores (%, mean $\pm$ standard deviation across three random seeds). LRD-specialized merging represents the primary adaptation setting following the ModernBERT training procedure, while CX-only merging is reported as an ablation to investigate whether effective parameter-space merging depends on LRD specialization.

Multi-label classification. Water, Forest, Biodiversity, and Nature (WFB Nature) (Schimanski et al., 2023b) is grounded in the Taskforce on Nature-related Financial Disclosures framework, where each text sample may be labeled across four nature dimensions — water, forest, biodiversity, and a general nature category — requiring the model to capture co-occurring thematic content in corporate environmental disclosures. **WXImpact-Bench** (Yu et al., 2025) evaluates understanding of disruptive weather impacts on society across six impact categories: infrastructural, political, financial, ecological, agricultural, and human health. The dataset is constructed from historical newspaper articles through a four-stage pipeline of keyword extraction, event categorization, topic-aware article selection, and expert annotation.

Retrieval. ClimRetrieve (Schimanski et al., 2024) benchmarks information retrieval from corporate climate disclosures. The dataset pairs 16 detailed climate-related questions with 30 sustainability reports, yielding over 8.5K question-passage pairs annotated at multiple relevance levels. We follow the report-level setup of Schimanski et al. (2024) and cast the task as binary relevance classification by collapsing the original ordinal relevance scale at threshold ≥ 1, so a single architecture and metric are shared with the classification tasks; the split sizes in Table 9 count $(\text{query}, \text{passage}, \text{relevance})$ triples after the $80/10/10$ stratified split.