Environmental Data Science (2025), 4: e51, 1–42
doi:10.1017/eds.2025.10024

CAMBRIDGE
UNIVERSITY PRESS

SURVEY PAPER

Language models for the analysis of and interaction with climate change documents

Elena Volkanovska

Institute of Linguistics and Literary Studies, TU Darmstadt, Germany
Email: elena.volkanovska@tu-darmstadt.de

Received: 22 November 2024; **Revised:** 31 August 2025; **Accepted:** 09 October 2025

Keywords: climate change; language models; text classification; text summarization; question-answering

Abstract

Language models (LMs) have attracted the attention of researchers from the natural language processing (NLP) and machine learning (ML) communities working in specialized domains, including climate change. NLP and ML practitioners have been making efforts to reap the benefits of LMs of various sizes, including large language models, in order to both simplify and accelerate the processing of large collections of text data, and in doing so, help climate change stakeholders to gain a better understanding of past and current climate-related developments, thereby staying on top of both ongoing changes and increasing amounts of data. This paper presents a brief history of language models and ties LMs' beginnings to them becoming an emerging technology for analysing and interacting with texts in the specialized domain of climate change. The paper reviews existing domain-specific LMs and systems based on general-purpose large language models for analysing climate change data, with special attention being paid to the LMs' and LM-based systems' functionalities, intended use and audience, architecture, the data used in their development, the applied evaluation methods, and their accessibility. The paper concludes with a brief overview of potential avenues for future research vis-à-vis the advantages and disadvantages of deploying LMs and LM-based solutions in a high-stakes scenario such as climate change research. For the convenience of readers, explanations of specialized terms used in NLP and ML are provided.

Impact Statement

This survey paper reviews the ways in which the natural language processing (NLP) and machine learning (ML) communities have been developing and utilizing language models (LMs) of various sizes to help process or interact with large collections of climate change data. The work provides an overview of the design and motivations underpinning LMs and LM-based systems intended for the climate change domain and would benefit researchers interested in the intersection between NLP, ML, and climate change documents used as data. The study simultaneously paints a broad picture of existing climate-change-related LMs and LM-based systems, and a detailed description of their respective building blocks, equipping researchers with information about existing tools and hopefully providing pointers for future research directions.

1. Introduction

The past few decades have witnessed an increase in the amount of text data on the topic of climate change (CC) that has become publicly available. The data can comprise publications dedicated explicitly to the

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

topic of climate change; this category encompasses, for example, reports published by stakeholders working exclusively in the climate change domain, such as the Intergovernmental Panel on Climate Change (IPCC), as well as scientific publications on the topic of climate change, alongside blogs and other web content published on this topic. At the same time, climate-relevant information can also be incorporated within reports concerning a different topic, such as corporate financial reports. This is especially common in cases when companies are required to disclose climate risks their operations might entail. The scale of data growth has been captured in various studies analysing climate change: one such example is the study by Korte et al. (2023), who analyse data published in OpenAlex (Priem et al., 2022), a digital database for scientific works, and observe a tenfold increase in the number of scientific papers labelled with the tag "climate change" published annually between 2001 and 2021: from 4,145 to 41,093 papers. The sheer volume of data highlights the need for an efficient and reliable method of processing and analysing climate-related texts.

One of the tasks of natural language processing (NLP), a branch at the intersection of linguistics and machine learning (ML), is to facilitate the analysis of large collections of data. Among other things, NLP- and ML-supported text-mining methods allow for processing large corpora at scale by, for example, unveiling the topics of the documents comprising the collection, or by classifying documents into predefined categories. The introduction of the Transformer deep learning architecture (Vaswani et al., 2017),[1] coupled with ever-increasing compute resources, ushered in the era of Transformer-based language models (LMs) for various NLP tasks, including tasks for processing climate-related texts. The public release of ChatGPT[2] at the end of 2022 and its rising popularity have shifted the focus from text classification tasks to text generation and manipulation tasks, the latter describing instances where language models generate answers to user-provided questions or summarize a longer text.

NLP and ML methods designed for climate change text data have been published in the context of domain-agnostic ML and NLP conferences and journals, where the goal is to showcase novel tools, methodologies, frameworks, data collection, or data annotation approaches, and where researchers opt to demonstrate what their systems can perform using data sourced from the CC domain. The past few years have also seen the establishment of (1) initiatives focusing solely on research at the intersection of artificial intelligence (AI) and ML[3] on the one hand, and climate change on the other, and (2) initiatives for developing socially-responsible NLP tools, where the goal is to motivate researchers to explore how ML and AI can help tackle issues of relevance to society, especially in fields that have been overlooked in NLP research. An example of (1) would be events and publication opportunities presented by organizations such as Climate Change AI (CCAI),[4] which specifically promote research at the intersection of ML and CC. Some examples of research published in the context described in (2) include the NLP for Social Good initiative,[5] the International Conference on Environmental Design and Health,[6] as well as the Climate NLP workshops hosted by the Association of Computational Linguistics (ACL).

Against the backdrop of a growing body of research showcasing how LMs are integrated in climate change research, the primary objective of this survey is to provide an overview of LMs and LM-based systems that have been developed to assist with text-based tasks in the climate change domain. In this context, the term *language models* is used to refer to systems that are trained to predict the next word or a missing word in a text; these are discussed in more detail in Section 3. The utilization of LMs in the CC

[1] Transformer is a type of neural network architecture that has led to considerable advancement in NLP due to the capacity to model relationships between words even across very long sentences, and to weigh the importance of different words or tokens relative to each other (Raschka, 2024).

[2] https://chatgpt.com/.

[3] "ML" and "AI" are not used interchangeably, and it should be pointed out that ML is a subset of AI. AI is the general ability of computers to emulate human thought and perform tasks in a real-world environment, while ML describes the technologies and algorithms allowing computers to identify patterns, make decisions, and improve their performance (Engineering, 2023; Raschka, 2024).

[4] https://www.climatechange.ai/.

[5] https://nlp4sg.vercel.app/.

[6] https://iced.eap.gr/.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

domain is a vast topic encompassing multiple disciplines (Rolnick et al., 2022; Lu, 2024).[7] The goal is to zero in on approaches to developing LMs intended to classify sections of natural texts in pre-determined categories, to aid analysis tasks where the input and the output are natural language text, or to answer questions by using collections of text documents as repositories of relevant information. One possible exemption from this "rule" is LM-powered question-answering systems capable of retrieving information from a Google search result or from tables, such as the system described in Section 5.1.2. However, both the system's input and output are natural language text. Web-based-only services for analysing climate data, such as the conversational assistant for climate question-answering ClimateQ&A developed by Ekimetrics[8] or Climind.Copilot, a conversational engine for climate-specific scenarios hosted on the platform Climind,[9] will not be comprehensively described, but might be mentioned in Section 6.

The objectives of the survey are laid out in detail in Section 2, followed by an explanation of its structure and intended audience in Sections 2.1 and 2.2, respectively. Sections 3 and 3.1 provide a brief history of language modelling and some considerations relevant to the survey, while Sections 4, 5, and 6 are dedicated to existing LMs and LM-based systems for the climate change domain. The survey ends with a summary and recommendations for future work in Section 7. A quick overview of all reviewed LMs and LM-based systems is available in the three tables comprising Appendix A.

2. Survey objectives

Reviews of language models can be conducted from multiple angles, including domain-specificity. Some survey and evaluation studies that are exclusively interested in domain-specific LMs have been performed for the legal field (Sun, 2023; Wehnert, 2023), code generation (Chen et al., 2021), education (Kasneci et al., 2023), and healthcare (Yang et al., 2023), to name just a few. In the climate change context, efforts have been made to assess how well general-purpose LMs can answer climate-related questions (Bulian et al., 2023), or whether LMs have the potential to assist in monitoring innovations in climate technology (Toetzke et al., 2023). Climate change has also been named a topic that is yet to harness the benefits of LLM-based toolstacks (Kaur et al., 2024), alongside domains such as *fitness* and *well-being*.

Surveys focusing on general-purpose LMs describe models and groups of models (model families) popular in the AI community, zeroing in on their features, functionalities, and limitations, alongside development techniques, popular datasets for pretraining, fine-tuning, and LLM evaluation metrics and benchmarks (Minaee et al., 2024). Zhao et al. (2023) aim to introduce terminology that distinguishes between models of various sizes, using the term "pretrained language models" (PLM) to refer to models developed before the onset of generative pretrained transformers (GPTs), and "large language models" (LLMs) to refer to models developed after the advent of GPTs.[10] The survey authors perform a comprehensive examination of four major aspects of LLM development: pretraining, adaptation tuning,[11] utilization,[12] and capability evaluation. Within this framework, in addition to describing datasets for pretraining and fine-tuning LLMs, model architectures and training processes, LLM utilization is presented from two aspects: *methods of utilization*, which include types of prompt engineering and planning for complex task-solving, and *functional applications*, which focus on the integration of LLMs in solutions for various tasks, ranging from classic NLP tasks to applications employing autonomous LLM-based agents. Hadi et al. (2023) focus on the implementation of ChatGPT-based solutions in the medical, education, finance, and engineering applications; the authors also dedicate a section on the

[7] At the time of writing, the latter paper is awaiting peer review.

[8] https://huggingface.co/spaces/Ekimetrics/climate-question-answering.

[9] https://www.climind.co.

[10] *GPT* is used to refer to present-day large language models developed to generate content, which can be unimodal (texts or images only) or multimodal (texts, images, videos etc.).

[11] *Adaptation tuning* is a technique for fine-tuning LLMs on specific topics or tasks.

[12] In their survey, *utilization* refers to contexts for using an LLM for downstream tasks.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

impact the training and deployment of LLMs have on the environment, alongside a set of measures that are being undertaken to promote sustainable development.

This survey paper aims to merge the functionality-, architecture-, feature-, and utilization-based descriptions found in reviews of generic models, and apply them toward a systematic review of domain-specific LMs and LM-powered systems, which have been developed exclusively for working with text data from the climate change domain. The objectives of the survey include: taking stock of existing tools for processing climate-related documents, identifying the type of tasks, text analysis and text annotation they enable, as well as providing a brief description of the technical approach behind these tools, the data that has been used in their development, and whether the resulting LM or LM-based system is publicly available or proprietary.[13] Some LMs are available both as open-source/open-weight models and paid-for service, and these instances are flagged up in the paper.

Given the rapid developments in the LM ecosystem, with new models being released every day, this survey should be perceived as a *snapshot* of the landscape at the time of writing. While substantial effort has been made to offer a survey that is as comprehensive as possible, there always exists a chance of relevant research not being included in the selection of papers. For this reason, the selection of LMs for climate change presented in the paper also "lives" as a GitHub project,[14] with the hope that the research community will show interest in complementing the selection with additional resources.

2.1. Survey structure

The survey starts with a high-level classification of climate-change-relevant LMs and LM-based systems, distinguishing between *domain-specific* language models, which have been built for climate change applications (Section 4), and systems relying on *general-purpose* language models, which are integrated as a component of a system developed for climate change applications (Section 5).

Within each of the two sections, LMs and LM-based systems are further classified based on the task for which they have been developed, namely: question-answering (Sections 4.1 and 5.1), question-answering and scoring (Section 5.2), text summarization (Section 4.2), text classification (Section 4.3), and text classification and text generation (Section 4.4). Figure 1 provides an overview of the LMs and LM-based systems discussed in this article.

For each LM and LM-based system and whenever information is available, the survey includes: (1) intended use and audience of the model / system, (2) model / system architecture, training, and data; (3) evaluation and results; and (4) access, transparency, and engagement. Information reported under item (4) includes details about the accessibility of the model, whether its development has been documented in terms of data acquisition, data usage, and code, and the number of all-time downloads for open-source/open-weight models, when possible. All but one of the LMs and LM-based systems presented in this study have been developed to be primarily used with English language data. Efforts to make an LM accessible in a language other than English are mentioned under item (4).

2.2. Potential audience

This survey paper is aimed at climate change researchers and practitioners interested in learning more about and/or integrating LMs in their line of work and who possess some degree of understanding about

[13] A "publicly available" model is one that anyone can access; publicly available models are sometimes *open-source*, which means that the model architecture, the data and the code used in their development are accessible, or *open-weight*, which means that only the trained weights of the model can be downloaded and used, but the data and the code used in its development are not available. A "proprietary" model is accessed as a paid-for service, usually available through a web interface, an application programming interface (API), or both. It needs to be borne in mind that an open-source or an open-weight model is not automatically accessible to everyone, since running and task-adapting LLMs of certain sizes requires expensive infrastructure or funds to access such infrastructure and can be difficult on consumer-grade equipment.

[14] https://github.com/volkanovska/Language-models-for-climate-change-texts.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

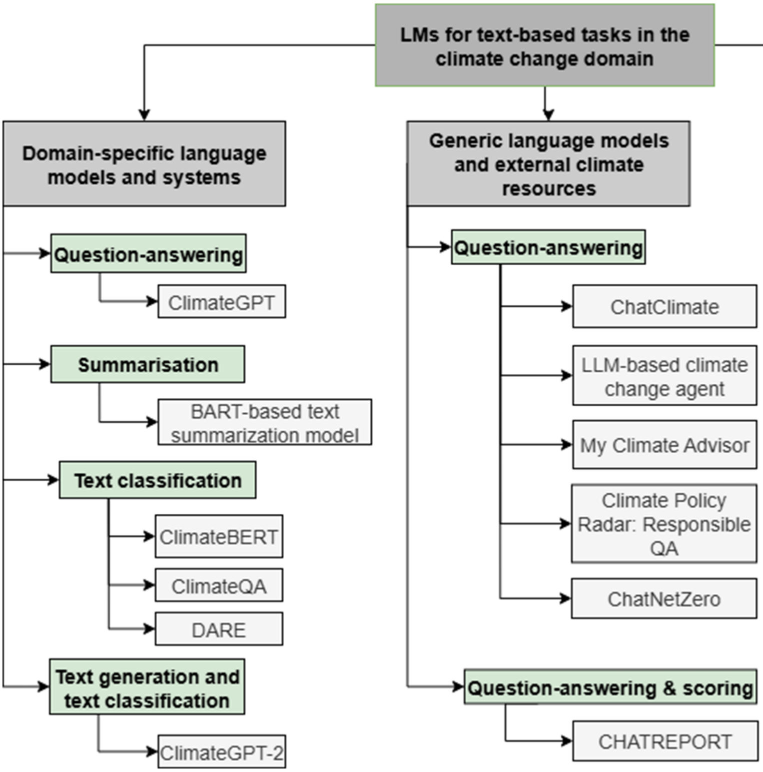
LMs for text-based tasks in the climate change domain
Domain-specific language models and systems
Question-answering
ClimateGPT
Summarisation
BART-based text summarization model
Text classification
ClimateBERT
ClimateQA
DARE
Text generation and text classification
ClimateGPT-2
Generic language models and external climate resources
Question-answering
ChatClimate
LLM-based climate change agent
My Climate Advisor
Climate Policy Radar: Responsible QA
ChatNetZero
Question-answering & scoring
CHATREPORT
Other LMs and resources
CC and other topics
Fact checking and claim detection
Benchmarking datasets
Other models and projects
INDUS
Project Gaia

Figure 1. *Classification of LMs and LM-based systems described in this survey. LMs and LM-based systems that are not given specific names by the paper authors are referred to descriptively.*

the underlying principles of language models. Explanations of technical concepts and terms typical for the LM domain are provided either as footnotes or in the running text.

3. Brief overview of language model development

As mentioned previously, the term *language model* (LM) denotes a system that has been trained on token[15] prediction tasks. The most common training approaches are next-token prediction, where an LM is tasked with predicting the next token given a starting sequence, or masked-token prediction, where an LM is trained to fill in missing words in a text.[16] As outlined by Bender et al. (2021), the idea of LMs was first proposed by Shannon (1949) and deployed in the 1980s in systems for automatic speech recognition, machine translation, or document classification. Initially, language modelling was done by using n-grams and large collections of data. The term *n-gram* stands for an “n” number of items, such as characters, sub-word units, or words, in a continuous sequence. N-gram models were superseded by pretrained representations of word distributions, known as *word vectors* or *word embeddings*, where an artificial neural network, which is a type of ML algorithm, is given large collections of text as input, from which it generates numerical representations for words in the collections. The introduction of the Transformer neural networks (Vaswani et al., 2017) in language modelling, alongside the increased availability of computing power and data, led to improved performance of models on popular benchmarks.

[15] A token can be a character, a subword unit, a word, or a punctuation mark.

[16] Commonly referred to as next-token prediction and masked language modelling.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Scaling up model size further led to the development of large language models (LLMs), which can process information in a zero-shot setting. *Zero-shot* refers to a setting where an LM is tested on a task not explicitly included in its training data. A tendency developed to design models that can process instructions given in a natural language and perform the desired task without being previously trained on the task or the domain in question. A common way of allowing the intended audience to "communicate" with these models is through a chatbot interface, where a user can "probe" the model for a certain output. Incremental improvements in benchmarks measuring various model functionalities led to LLMs becoming employed in domains that considerably differ from the domains for which language modelling was originally developed; one of these "new" domains is climate change. Prior to taking a deeper dive in the topic, I point out several considerations that should be taken into account when discussing LMs and LM-based systems.

3.1. Important considerations

The first consideration is a comment about the terminology commonly employed to describe LMs and LM-based systems. LMs are frequently described as systems possessing "knowledge" and "capabilities", as well as having the ability to "understand" language or given tasks. While discussing the terminological choices authors make when describing LMs and LM-based systems is beyond the scope of this study, it is important to underscore that LMs do not hold knowledge or understand language or the world the same way humans do. For more information on the anthropomorphization of descriptions of technical systems, see Inie et al. (2024). In the scope of the survey, the terminological choices made by the developers of the reviewed systems are mostly preserved when relaying the technical description of a system, with de-anthropomorphized terminology being adopted when applicable.

A comment needs to be made about the naming conventions for LMs, especially those that are hosted on the Hugging Face Hub (HFH).[17] Models included in the survey and that are hosted on HFH will be accompanied by their HFH model identifier. The model identifier contains the name of the organization, research entity, or user that developed the model and the name of the LM itself. From the model identifier *google-bert/bert-large-uncased* it can be deduced that the entity that developed the model is *Google*, and that the unique name of the model is *bert-large-uncased.* This is done to assist findability and reusability of LMs.

When describing LMs, a distinction is made between large language models (LLMs) and language models that would not be considered "large" by current standards, referred to as pretrained language models (PLMs), in line with the terminology proposed by Zhao et al. (2023). In the scope of this study, LM size will be considered in terms of the number of trainable parameters and the amount of data used to pretrain the model.[18] Data size is usually, though not always, reported as the number of tokens.

Another distinction relevant to this survey is the difference between an LM and an LM-supported system. In most instances, the latter integrates an LM with a retrieval-augmented generation (RAG) component. The retrieval component has a database of preprocessed documents, and its task is to find passages from texts that are most relevant to a user's query. These passages are passed to the LM, which should generate an answer grounded in the information contained in the retrieved passages.

The increased popularity of artificial neural networks and the associated rise in computational needs prompted researchers to pay more attention to the energy costs connected to developing and training language models (Strubell et al., 2020). The emergence and overarching use of large language models on a broad range of tasks has motivated proposals about the ways in which the cost and the environmental impact of LM development should be accounted for, especially for generative models. Luccioni et al. (2023) estimate the carbon footprint of a model titled "BigScience Large Open-science Open-access Multilingual Language Model" (BLOOM) at three levels, including the final training process, all other processes ranging

[17] HFH is a platform for sharing LMs, datasets, and demo applications: https://huggingface.co/docs/hub/en/index.

[18] "Trainable" or "learnable" parameters are values that can be adjusted during an LM training process. The higher the number of trainable parameters, the more compute power is needed to deploy the LM. In 2018, Devlin et al. (2019) released *google-bert/bert-large-uncased,* a model with 340 million (M) parameters. GPT-3 (Brown et al., 2020), released in 2020, has 175 billion (B) parameters.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

from equipment manufacturing to energy-based operational consumption, and finally, the energy consumption and carbon emissions at deployment, i.e., inference time. Meanwhile, Moro et al. (2023) propose a method of measuring the environmental impact from a task-based perspective (summarization). While the topic of environmental impact is not central to this survey, whenever the authors of the papers reviewed include information about the work's carbon footprint or any associated data, this is mentioned in the subsection *access, transparency, and engagement.* Information about the GPU usage in developing domain-specific models is also summarized in Table 7.

Last but not least, LMs exist either as stand-alone models or within a family of models. The latter entails a group of models that have a common denominator (for example, their foundation model), but differ in terms of size or intended use (Webersinke et al., 2022; Thulke et al., 2024). Stand-alone, or single models, are models that do not belong to a larger group of related LMs.

4. Domain-specific language models and systems

The LMs and LM-based systems included in this section have been developed for the following tasks: answering climate-change-related questions, summarization, classifying text in various climate-relevant categories, or both text generation and classification.

Two large model families are described: ClimateGPT (Thulke et al., 2024), a family of five LLMs for CC question-answering (QA), and ClimateBERT (Webersinke et al., 2022), a family of thirteen domain-adapted and fine-tuned PLMs for various climate-related text classification tasks. The models ClimateGPT-2 (Vaghefi et al., 2022) (text classification and text generation), DARE (Xiang and Fujii, 2023) (text classification), ClimateQA (Luccioni et al., 2020) (text classification), and a BART-based model for summarizing climate-related political press releases (Dickson, 2023) (text summarization) are single models.

This section concludes with a brief comparative summary of the domain-specific models discussed in detail, and a table focusing on GPU use for the models' development.

4.1. Question-answering

4.1.1. ClimateGPT: a family of LLMs and an LLM-based system for climate change question-answering

Models of the **ClimateGPT** family have been developed and made available through a collaboration of several research institutions, including the Endowment for Climate Change Intelligence (ECI)[19] and ErasmusAI (Thulke et al., 2024).[20] The former is a decentralized foundation that delivers AI solutions for climate change, while the latter is a platform that leverages LLMs and generative AI to "address complex global challenges, such as climate change" (Erasmus.AI, 2024). The **intended use** of the five LLMs pretrained and instruction-tuned on climate-relevant data is question-answering: they are to serve as "personal climate experts" that are "breaking down questions and concepts to the level of expertise of the user" (Thulke et al., 2024, p.12), while the **intended audience** includes decision-makers, scientists, policymakers, and journalists involved in climate discussions. The ClimateGPT model family comprises the models: ClimateGPT-FSG-7B, ClimateGPT-FSC-7B, ClimateGPT-70B, ClimateGPT-13B, and ClimateGPT-7B.[21] In addition to individual models, an LM-based system is available on the website of ECI as two versions: ClimateGPT,[22] available to researchers contingent upon their application for access through ECI's website, and ClimateGPT+,[23] an enterprise version aimed at businesses and organizations that would like to use the models with their own data.

Model architecture: Models of the ClimateGPT family are decoder-only Transformer models. Such models are also known as **generative pretrained models** (GPTs). They are trained with the objective of predicting the next word—token—in a sequence and they are mainly used for text completion tasks

[19] https://www.eci.io/.

[20] https://erasmus.ai/.

[21] HFH model identifiers, in the same order as the LM names: *eci-io/climategpt-7b-fsg*, *eci-io/climategpt-7b-fsc*, *eci-io/climategpt-70b*, *eci-io/climategpt-13b*, *eci-io/climategpt-7b*.

[22] https://www.eci.io/climategpt.

[23] https://www.eci.io/climategptplus.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

(Raschka, 2024). In the design, Thulke et al. (2024) closely follow the architecture of Meta's Llama-2 (Touvron et al., 2023).[24] Details about the model components, including normalization techniques for stable and efficient model learning, positional embeddings, and the activation function, are available in the model's extensive technical report (Thulke et al., 2024).

Training and data: The models are pretrained with the next-token prediction objective on a large climate-relevant corpus, and are instruction fine-tuned for the downstream task of question answering. Two domain-specific pretraining methods are applied: from-scratch pretraining (FSPT) and continued pretraining (CPT). In FSPT, model training starts from randomly initiated weights, and the model relies on domain-specific data only. In CPT, training starts from a general-purpose foundation model, which already holds an existing knowledge base from having been trained on general data, and is then adapted to a specific domain using domain-relevant information.[25]

Pretraining is done with **two corpora**: a corpus of 300 billion (B) tokens curated for content on the topics of climate, humanitarian issues, and science, and a 4.2B token corpus of hand-picked climate-change data. The 300B corpus contains news, various publications, modern books, patents, Wikipedia data, policy and finance data, and science. The 4.2B corpus contains news on extreme weather events, over 500 pages of technical documentation on game-changing breakthroughs in the areas of energy, CC, food security, health etc., a breakdown of the United Nations' (UN) 17 Sustainable Development Goals, CC news, CC specific corpora including documents published by the World Bank, IPCC, the United States Government, and other international development organizations, treaty organizations, non-governmental organizations, and national state governments, as well as academic research in climate. The **300B** corpus is used for the **from-scratch pretraining** of ClimateGPT-FSG-7B. The **300B** and the **4.2B corpus** are used for the **from-scratch pretraining** of ClimateGPT-FSC-7B. Finally, the **4.2B** corpus is used for the **continued pretraining** of ClimateGPT-7B, ClimateGPT-13B, and ClimateGPT-70B, which are the models built on top of Llama-2.

Instruction fine-tuning (IFT) is done with 271,525 IFT training samples (TSs); of these, 106,269 TSs are from general-domain datasets, and 165,256 TSs are from climate-specific datasets. The **general-domain TS set** comprises the datasets Databricks Dolly[26] (Conover et al., 2023), OpenAssistant Conversations 1 (OASST-1)[27] (Köpf et al., 2023), as well as FLAN v2 and CoT (as described in (Wang et al., 2023), an internal set of prompt-completion pairs, referred to as **AppTek General**, and 9,846 TSs formulated as question-and-answer pairs from the StackExchange communities for earth science, sustainability, and economics.

The **climate-specific TS set** contains demonstrations,[28] grounded expert demonstrations, and grounded non-expert demonstrations. The 1,332 TSs dubbed "demonstrations" are created by interviewing a small team of senior climate experts on foundational concepts in an expert's field of expertise, current CC trends, expected developments, pivotal findings and research papers, key arguments, and scenarios in which an LLM would be of use to an expert. Another 7,254 TSs are collected from grounded expert demonstrations and 146,871 TSs from grounded non-expert demonstrations. The *grounded expert demonstrations* include CC questions and structured answers provided by nine climate scientists at graduate or PhD level, as well as synthetically generated question-answer pairs corrected by the same group of experts; while *non-expert demonstrations* contain ideas for prompt and completion pairs, where the completion contains in-text citations to the relevant sources, collected from external annotators.[29]

[24] *Llama* stands for *Large Language Models Meta AI.*

[25] Since the CPT models are built on top of Llama-2 as a foundation model, see Touvron et al. (2023) for data comprising the pretraining corpus of the foundation model.

[26] 15,000 human-generated prompt-response pairs designed for LLM instruction tuning, available at https://huggingface.co/datasets/databricks/databricks-dolly-15k.

[27] A dataset of over 150,000 human-generated datasets for training dialogue agents, available at https://huggingface.co/datasets/OpenAssistant/oasst1.

[28] Instruction-completion pairs.

[29] The authors also intended to use synthetically generated demonstrations, where an existing general-purpose LLM would be prompted with few-shot examples and a document, but decided not to use the data for IFT as they observed lack of consistent improvements in perfomance on automatic benchmarks.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Protocols for safe text generation are observed by generating completions for each prompt in the dataset using a safe model, namely Llama-2-Chat-70B (Touvron et al., 2023). Many of the responses are then manually checked, appended to the Do-Not-Answer Dataset, which is a collection of instructions that responsible models should not follow and the appropriate responses (Wang et al., 2024),[30] and this augmented resource is then added to the IFT dataset.

Improving retrieval: Retrieval augmented generation (RAG) is used to counter model errors and overcome limitations posed by the knowledge cut-off date, that is, enable models to access documents published after their training had been completed. *RAG* is not part of a model's architecture and is a component that can be added to an LLM-powered text generation system, where a user's prompt triggers a retrieval system to query a knowledge base of documents, retrieve the ones that have the highest similarity to the query, and generate a response using the retrieved documents. Documents that are stored in the database used in RAG include IPCC reports, the Potsdam Papers, documents from the Earth4All process, and 73 other non-specified open-access documents.[31]

Evaluation and results: Two types of evaluation are implemented: automatic and human. For *automatic evaluation* the authors use both domain-specific and general-purpose datasets. The **climate-specific evaluation datasets** include ClimaBench (Spokoyny et al., 2023),[32] a collection of open-source climate-related datasets allowing systematic evaluation of model performance across various classification tasks. ClimaBench includes the following datasets: ClimateStance[33] and ClimateEng[34] (Vaid et al., 2022), Climate-FEVER[35] (Diggelmann et al., 2020), ClimaText[36] (Varini et al., 2020), and CDP-QA[37] (Spokoyny et al., 2023). In addition to the evaluation datasets of the ClimaBench collection, the authors use the datasets Pira 2.0 MCQ[38] (Pirozelli et al., 2024) and Exeter Misinformation[39] (Coan et al., 2021). The **general-domain evaluation datasets** include HellaSwag[40] (Zellers et al., 2019), PIQA[41] (Bisk et al., 2020), OpenBookQA[42] (Mihaylov et al., 2018), WinoGrande[43] (Sakaguchi et al., 2021), and the MMLU dataset[44] (Hendrycks et al., 2020).

The performance of ClimateGPT models is compared against the performance of same-size models of the Llama-2-Chat family, as well as against 7 other general-purpose foundation models in the 3-13B size range, namely: Stability-3B, Pythia-6.9B, Falcon-7B, Mistral-7B, Llama-2-7B, Jais-13B, Jais-13B-

[30] https://github.com/Libr-AI/do-not-answer?tab=readme-ov-file.

[31] No additional information is provided about the Potsdam Papers; it is possible they encompass publications by the Potsdam Institute for Climate Impact Research (PIK).

[32] https://huggingface.co/datasets/iceberg-nlp/climabench.

[33] The dataset contains climate change tweets expressing three stances: supporting climate change prevention, opposing climate change prevention, and an ambiguous stance; the model is expected to predict the right label.

[34] A dataset of climate change tweets classified in the topics *disaster*, *ocean/water*, *agriculture/forestry*, *politics*, and *general*; the model should predict the right topic.

[35] A fact-verification dataset of climate-related claims collected online and paired with passages containing evidence for each claim. The evidence passages are extracted from Wikipedia. Claim-evidence pairs are labelled by human annotators as: SUPPORTS, REFUTES, and NOT_ENOUGH_INFO, which is also the default label. Each claim is paired with five sentences. The label DISPUTED is used in instances where there is both supporting and refuting evidence.

[36] A dataset for a binary classification task, where the model should predict whether a sentence is climate change-related or not. It entails 123,000 sentences extracted from Wikipedia, from annual regulatory filings made by listed companies to the U.S. Securities and Exchange Commission, and from the Climate-FEVER dataset.

[37] A dataset containing question-answer pairs from the questionnaires of the Carbon Disclosure Project (CDP), which is a not-for-profit charity supporting companies, cities, states, and regions, in their efforts to measure and manage their climate change risks and opportunities.

[38] A collection of scientific abstracts and reports from the United Nations on the topics of climate change, Brazilian coast, and ocean. The model's task is to select the correct answer to a question from a set of five candidates.

[39] A dataset for a classification task where a model should discern whether a given text contains a contrarian claim about climate change.

[40] The dataset contains multiple-choice questions where a model should predict the next event in grounded situations.

[41] A dataset of binary choice questions that can be answered by understanding real-world object interactions in physical scenarios.

[42] A dataset of multiple-choice elementary-level science questions.

[43] A dataset of fill-in-the-blank tasks for ambiguous pronouns.

[44] A dataset for evaluation of world knowledge in several fields, including STEM, humanities, social sciences etc.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 1. *Performance comparison of different ClimateGPT models on climate-specific and general benchmarks (weighted averages of accuracy)*

Model	CC benchmarks	Gen. benchmarks
Stability-3B	62.8	61.5
Pythia-6.9B	50.8	53.1
Falcon-7B	48.3	59.3
Mistral-7B	**73.7**	**70.1**
Llama-2-7B	62.6	63.9
Jais-13B	54.4	58.3
Jais–13B-Chat	65.3	63.1
Llama-2-Chat-7B	68.5	62.9
Llama-2-Chat-13B	71.4	66.4
Llama–2-Chat–70B	**77.0**	**68.6**
ClimateGPT-7B	77.1	65.1
ClimateGPT-13B	**78.0**	68.8
ClimateGPT-70B	77.2	**73.7**
ClimateGPT-FSG-7B	46.2	48.8
ClimateGPT-FSC-7B	42.1	49.9

Chat.[45] On the **climate benchmarks**, the three CPT models of the ClimateGPT family outperform the 7B, 13B, and 70B models of the Llama-2-Chat family, as well as the 7 other general-purpose foundation models. ClimateGPT FSPT models perform worse than the general-purpose foundation models, the same-size Llama models, and the Llama-based ClimateGPT models. On the **general-domain benchmarks**, the ClimateGPT CPT models outperform their respective counterparts (in terms of parameters) from the Llama-2-Chat family; however, both ClimateGPT-13B and ClimateGPT-7B lag behind Mistral-7B (Jiang et al., 2023). The results of the automatic evaluation, given as weighted averages, are summarized in Table 1.

The authors also conduct **human evaluation**, where seven climate change students at master, PhD, and post-doc level provide feedback on the output of ClimateGPT-70B, ClimateGPT-7B, and ClimateGPT-FSC-7B by ranking them against each other on a series of items, including qualifying the goodness of each answer on a scale in the range of -2 to +2, where -2 is the lowest, 0 is average, and 2 is the highest score. ClimateGPT-70B receives an average rank of 1.0 and the lowest number of hallucinations[46] among the models (2 instances of hallucination; the highest number is 5). ClimateGPT-70B and ClimateGPT-7B, both CPT models built on top of Llama-2, perform better than the FSPT model, ClimateGPT-FSC-7B.

Access, transparency, and engagement: The five models are published on the HFH.[47] In addition, a QA chat system powered by ClimateGPT models can be accessed through two websites: ECI's[48] and Erasmus.AI's.[49] Both require registration. On ECI's website, the chatbot is available in two versions: ClimateGPT and ClimateGPT+. The latter is an enterprise version aimed at businesses and organizations that would like to use the models with their own data. According to ECI, a portion of the revenue generated from ClimateGPT+ will be reinvested into ECI, thereby honouring their commitment to open-access

[45] HFH model identifiers, in the same order as the LM names: *stabilityai/stablelm-3b-4e1t*, *EleutherAI/pythia-6.9b*, *tiiuae/falcon-7b*, *mistralai/Mistral-7B-v0.1*, *meta-llama/Llama-2-7b-hf*, *core42/jais-13b.*

[46] A type of error identified in the output of generative models, when they either fail to render information from a reference document correctly, or when their output cannot be verified from the source (Havlik and Pias, 2024).

[47] https://huggingface.co/eci-io.

[48] https://www.eci.io/climategpt.

[49] https://climategpt.ai/.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

climate AI and funding future endeavours in this field. ClimateGPT is made available in 21 other languages[50] using a cascaded machine translation approach.[51]

Transparency in terms of listing the data used to train and evaluate the models is preserved. Not all items comprising the pretraining data, the fine-tuning data, and the instruction fine-tuning data are publicly available; this means that information about the data cannot be obtained beyond the comprehensive description provided in the paper. In terms of evaluation, the authors make available the relevant Python script that, at the time of writing, plugs in all datasets for automatic evaluation, except for Climate-FEVER.[52] In addition, a model card and a sustainability scorecard are provided, with information about the hardware and software used in the training process, as well as the models' carbon footprint.

In terms of engagement, at the time of reviewing,[53] ClimateGPT-7B had the highest number of all-time downloads (18151), followed by ClimateGPT-70B (3760), ClimateGPT-13B (670), ClimateGPT-7B-FSG (468), and ClimateGPT-7B-FSC (291).

4.2. Summarization

4.2.1. BART-based model for summarizing climate-related political press releases

The publicly available model, whose HFH identifier is *z-dickson/bart-large-cnn-climate-change-summarization*, has the **intended use** of serving as a text summarization model that takes as input a political text in the domain of climate change, environment, or energy, and generates as output a summary of it. The model should detect the primary issue in the text and include it in the generated summary, alongside the position of the political party giving the press release, and a general summary of one to two sentences. The **intended audience** is not specified, but can be inferred as researchers interested in political parties' stance on climate change policies (Dickson and Hobolt, 2024).

Model architecture, training, and data: The model is a fine-tuned version of *Facebook/bart-large-cnn* (Lewis et al., 2020), which has been trained on the summarization dataset CNN/Daily Mail (Nallapati et al., 2016). The dataset is publicly available[54] and contains multi-sentence summaries of approximately 300,000 news articles published by CNN and the Daily Mail. The underlying model BART (which stands for Bidirectional Auto-Regressive Transformer), is an encoder-decoder Transformer model. As per Dickson (2023), the model *Facebook/bart-large-cnn* was fine-tuned on 7,000 press release/summary pairs from 66 political parties in 12 countries.[55] In Dickson and Hobolt (2024), the number of press release/summary pairs is reported at 6,000. As per Dickson and Hobolt (2024), the fine-tuning data was generated by prompting GPT-3.5 for automatic summaries. However, the associated Hugging Face Model Card specifies the use of a more recent model, GPT-4, for summary generation. The generated summaries are then qualitatively examined and slightly modified as necessary.

Evaluation and results: The model is not evaluated against a baseline and an F1 score is not reported. It is mentioned by Dickson and Hobolt (2024) that the model output has been qualitatively checked. Dickson (2023) points out that while the model is capable of identifying the primary issue in the political text, it does not include the name of the political party in the response. It identifies the author of the text as "the party" and summarizes the position as such (Dickson, 2023). In terms of **access, transparency, and engagement**, the model is publicly available; the fine-tuning dataset is not. Its total number of downloads at the time of reviewing is 7456.

[50] Arabic, Dutch, German, Indonesian, Lithuanian, Portuguese, Thai, Bengali, Finnish, Greek, Japanese, Pashto, Russian, Turkish, Chinese (simplified), French, Hebrew, Korean, Persian, Spanish, and Vietnamese.

[51] A multi-step machine translation process, where the user input is translated into English, used to prompt the LLM of interest, and the LLM-generated text is translated from English back into the language of the original prompt.

[52] Evaluation script available at: Hugging Face (https://huggingface.co/datasets/eci-io/climate-evaluation) and GitHub (https://github.com/eci-io/climategpt-evaluation).

[53] August 2025.

[54] https://huggingface.co/datasets/abisee/cnn_dailymail.

[55] Italy, Sweden, Switzerland, The Netherlands, Germany, Denmark, Spain, United Kingdom, Austria, Poland, Ireland, France.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

4.3. Text classification

4.3.1. ClimateQA

The **intended audience and use** of ClimateQA's (Luccioni et al., 2020) are analysts combing through financial reports for climate change-related risks and liabilities. The model classifies paragraphs of a report as potential answers to a pre-determined set of questions. Users of ClimateQA are not expected to have substantial technical knowledge, as the model is integrated in a web application hosted on the Microsoft Azure cloud solution; it is intended for them to interact with the model by uploading PDF files for analysis and downloading an output file, in which passages pertaining to a set of 14 questions proposed by the Task Force on Climate-Related Financial Disclosures (TCFD) are highlighted. Processing time is between 5 and 15 minutes per report.

Model architecture, training, and data: ClimateQA is based on the architecture of RoBERTa-base, a foundation model with 125M parameters.[56] *RoBERTa* stands for "robustly optimized BERT approach"; it is a language model seen as an improvement of the original Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019), mostly in terms of being able to handle more data and use better optimization strategies during training.

The foundation model is both pretrained on unlabelled data and fine-tuned on labelled data. For the former, the authors scrape 2249 publicly available financial reports from the databases of the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR)[57] and the Global Reporting Initiative (GRI).[58] EDGAR has been created by the US Securities and Exchange Commission to enable corporate filings by entities who are required to submit such forms under US legislation, while GRI is an international, independent organization that "helps businesses, companies, and other organizations understand and communicate their impact on issues such as climate change, human rights, and corruption" (Global Reporting Initiative, 2024). The collected reports span a timeline of 10 years and are from publicly-traded companies in the sectors: agriculture, food, and forests; energy; banks; transportation; insurance; and materials and buildings.

The classification task is grounded by the set of 14 TCFD questions. Passages seen as appropriate answers to the TCFD questions are labelled by a team of sustainability analysts. This results in a dataset of question-answer pairs (QA dataset), with the question being one of the TCFD questions, and the answer the portion of the text labelled as an answer by the human expert. Negative examples are generated by pairing the remaining sentences with the questions. The labelled dataset has a train set of 15,000 negative and 1,500 positive examples, a development set of 7,500 negative and 750 positive examples, and a test set of 1,200 negative and 400 positive examples.

Evaluation and results: The model is evaluated by comparing the difference in performance between the F1 scores ClimateQA achieves on the validation and the test QA data splits. They find that on average, the F1 score achieved on the test data split is 9.7% lower than the F1 score achieved on the validation data split. There are substantial differences in the model's performance on data from different sectors, with the best performance being observed in the sector *Energy* (F1 score on validation dataset higher by 4.4% relative to that on test dataset), and the worst in the sector *Materials and Buildings* (24.2% validation-test F1 score difference). The type of question also affects performance, with less-frequent questions, and questions whose answers could be more diverse proving to pose the biggest challenge to the model. The difference between the validation and the test score for each question is also provided.

Access, transparency, and engagement: The model is not publicly available, and at the time of writing, it is also not accessible through the dedicated web application. The labelled dataset generated for this study is not publicly available. While an official model card is not provided, the authors do elaborate on the choice of a foundation model vis-à-vis training time and energy efficiency. It is indicated that the model RoBERTa-base is selected over RoBERTa-large[59] because in the training experiments, the

[56] FacebookAI/roberta-base.

[57] https://www.sec.gov/edgar.shtml.

[58] https://www.globalreporting.org/.

[59] FacebookAI/roberta-large.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

former's training time was less than 5 hours on a 12GB GPU, as opposed to the latter, which needs almost 12 hours. The large model showed minor improvements in the F1 scores achieved on the validation and test datasets. It is pointed out that longer training times translate into higher energy costs, which is another factor in favour of choosing a smaller model.

4.3.2. ClimateBERT: a family of LMs for climate-relevant text classification

What started as four models pretrained on CC data (Webersinke et al., 2022) grew into a Hugging Face repository which, at the time of writing, hosts four foundation models and nine fine-tuned classification models for tasks involving CC texts, as well as 7 fine-tuning datasets. Information about models belonging to this family is also available on a designated website.[60] Unlike ClimateGPT, this family of models is based on a PLM and mainly focuses on text classification tasks.[61]

Intended use and audience: Models of the ClimateBERT family are intended for researchers using NLP models to process CC texts, and are geared towards climate-related classification tasks on paragraphs and sentences. Models supporting paragraph-level classification have been developed to classify paragraphs as: (1) climate-related or not, (2) expressing a sentiment of opportunity, risk, or a neutral one, (3) being specific or not specific about climate-related actions and intentions, (4) being about climate commitments and actions or not, and (5) being related to climate-relevant recommendation categories in the context of financial disclosure or not. The sentence-level classification models can detect (1) whether a sentence is an environmental claim or not, (2) whether it contains content on transition and physical climate risks or not, (3) whether it is about renewable energy or not, and (4) whether it is connected to net zero and reduction targets or not.

Model architecture, training, and data: The initial ClimateBERT models, ClimateBERT$_F$, ClimateBERT$_S$, ClimateBERT$_D$, and ClimateBERT$_{D+S}$,[62] were developed by conducting domain-adaptive pretraining of the model DistilRoBERTa-base (Sanh et al., 2020), a distilled version of the model RoBERTa (Liu et al., 2019).[63] *Knowledge distillation* is a method where a smaller and simpler model is trained to mimic the behaviour of a larger, more complex model. The goal is to develop a faster, more efficient model that possesses similar functionalities to the larger model.

A large corpus of texts considered representative of general- and domain-specific climate-related language, containing a total of 2,046,523 paragraphs, is compiled and used for domain-adaptive pretraining. Of the total number of paragraphs, 1,025,412 (50.1% of all paragraphs) are news articles on the topics of climate politics, climate actions, flood, and droughts retrieved from Refinitiv Workspace and climate-related news articles crawled from the web, 530,819 (25.9%) are abstracts of scientific papers published mainly between 2000 and 2019 and retrieved from the Web of Science, and 490,292 (23.9%) are texts from corporate climate and sustainability reports of over 600 companies published between 2015 and 2020, retrieved from Refinitiv Workspace and from the respective companies' websites. The authors report the average length of paragraphs in each category: news-related paragraphs have a mean length of 56 words, abstracts 218 words, and corporate report paragraphs 65 words. As the size of the pretraining corpus is not expressed in number of word-based tokens, using the average paragraph length for each category and the number of paragraphs, I calculate an estimation of the training corpus size expressed as number of words: the *news* corpus would have ca. 57 million words, the *abstracts* corpus would have ca. 116 million words, and corporate reports would account for ca. 32 million words; in total, the domain-specific pretraining corpus size is estimated at ca. 205 million words.

[60] Found at: www.chatclimate.ai. The additional LM-supported system for analysing CC texts, hosted on the same website, is discussed in Section 5.1.1.

[61] While the foundation models of the ClimateBERT family are much smaller in size in terms of trainable parameters relative to the models of the ClimateGPT family, the four foundation models are sometimes referred to as large language models.

[62] HFH identifiers, in the same order: *climatebert/distilroberta-base-climate-f*, *climatebert/distilroberta-base-climate-s*, *climatebert/distilroberta-base-climate-d*, *climatebert/distilroberta-base-climate-d-s*.

[63] https://huggingface.co/docs/transformers/en/model_doc/roberta.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

The complete corpus is used in **domain-adaptive pretraining** of the model ClimateBERT$_F$. The models ClimateBERT$_S$, ClimateBERT$_D$, and ClimateBERT$_{D+S}$ are trained on different portions of the corpus: ClimateBERT$_S$ is trained on 70% of corpus samples most similar to samples of the planned text classification task, ClimateBERT$_D$ is trained on 70% of corpus samples most diverse to the samples of the planned text classification task, and ClimateBERT$_{D+S}$ is trained on 70% of samples that have the highest composite score, calculated by using the similarity and diversity metrics applied to the previous two cases and summing over the scaled values. In addition, the vocabulary of DistilRoBERTa-base is extended by adding the 235 most common tokens from the pretraining corpus to the model's tokenizer, thereby allowing the model to learn representations of frequently occurring terms in climate-related texts, such as *CO_2*, *emissions*, *temperature*, *environmental*, *soil* etc. The four foundation models are available for download from the HFH.[64] The models will be referred to as ClimateBERT-DAPT (domain-adaptive pretrained) models hereinafter.

Finally, the four DAPT models are **fine-tuned** on three tasks: (1) text classification, where hand-selected paragraphs from companies' annual or sustainability reports are labelled as climate-related or not, (2) sentiment analysis, where climate-related paragraphs are labelled as expressing a negative sentiment of *risk*, positive sentiment of *opportunity*, or a *neutral* sentiment (sentiment analysis), and (3) fact-checking, for which the Climate-FEVER dataset is used (Diggelmann et al., 2020).

Later works make use of, update, or create new fine-tuned versions of ClimateBERT$_F$ for paragraph-level classification tasks (Bingler et al., 2022b, 2024) and sentence-level classification tasks (Stammbach et al., 2022; Deng et al., 2023; Schimanski et al., 2023a), resulting in five and four new task-specific models, respectively. The tasks, models, and the datasets used to train the model on paragraph- and sentence-level classification tasks are given below. All datasets have a train and a test split, and some have a development split, too. Train and development splits are used during model training and optimization, and the test split is used for model evaluation.

Paragraph-level text classification. Details about the dataset creation and annotation procedure for the training and testing data mentioned in items 1 to 5 are available in Webersinke et al. (2022) (1 and 2), Bingler et al. (2024) (1, 2, 3, and 4), and Bingler et al. (2022b, 2024) (5). At the time of writing, all models and datasets for paragraph-level text classification are available for download.

1. Classification of paragraphs as climate-related or not: *climatebert/distilroberta-base-climate-detector.* The dataset consists of hand-selected paragraphs from companies' annual or sustainability reports and annotated with *yes* if they relate to climate, or *no* if they do not.[65]
2. Classification of climate-related paragraphs into the classes OPPORTUNITY, NEUTRAL, or RISK: *climatebert/distilroberta-base-climate-sentiment.* The sentiment analysis dataset was developed by annotating paragraphs annotated with *yes* in dataset (1) as expressing a *neutral* sentiment, a sentiment of *opportunity* or a sentiment of *risk.*[66]
3. Classification of paragraphs as either SPECIFIC or NON-SPECIFIC. A paragraph is *specific* if it contains details about climate-related performance, action, or tangible and verifiable targets, and *non-specific* if these features are missing: *climatebert/distilroberta-base-climate-specificity.* The manually annotated paragraphs have been extracted from corporate reports.[67]
4. Classification of paragraphs as being or not being about climate commitments and actions: *climatebert/distilroberta-base-climate-commitment.* The paragraphs are extracted from companies' annual reports.[68]
5. Classification of climate-related paragraphs into four CC-related categories as defined by the Task Force on Climate-Related Financial Disclosures (TCFD),[69] namely: *governance*, *strategy*, *risk*

[64] https://huggingface.co/climatebert.
[65] https://huggingface.co/datasets/climatebert/climate_detection.
[66] https://huggingface.co/datasets/climatebert/climate_sentiment.
[67] https://huggingface.co/datasets/climatebert/climate_specificity.
[68] https://huggingface.co/datasets/climatebert/climate_commitments_actions.
[69] Currently disbanded; for more information visit https://www.fsb-tcfd.org/.

management, and *metrics and targets*: *climatebert/distilroberta-base-climate-tcfd.* The dataset consists of paragraphs extracted from reports issued by companies supporting TCFD reporting guidelines and annotated as related to one of the four categories or not.[70]

Sentence-level text classification. Details about the annotation and dataset design are available in Stammbach et al. (2022) (item 1), Deng et al. (2023) (2 and 3), and Schimanski et al. (2023a) (4). The datasets of items 1 and 4 are publicly available at the time of writing.

1. Classification of sentences as environmental claims or not. An *environmental claim* refers to the practice of suggesting or creating an impression that a service or a product is either environmentally friendly or not as damaging to the environment as competing goods:[71] *climatebert/environmental-claims.* The dataset contains sentences extracted from companies' sustainability reports, earning calls, and annual reports.[72]
2. Classification of sentences as being related to transition or to physical climate risks or not: *climatebert/transition-physical.* The dataset contains sentences extracted from earnings conference call transcripts and manually annotated as related to the topics of transition and physical climate exposure or not.
3. Classification of sentences as being related to renewable energy or not: *climatebert/renewable.* The dataset contains the sentences of dataset (2) marked as related to transition, which are further annotated as relating to renewable energy or not.
4. Classification of sentences as either being connected to emission net-zero or reduction targets or not: *climatebert/netzero-reduction.* The sentences have been collected and annotated in collaboration with the Net Zero Tracker project (Lang et al., 2023), which collects data from companies and governments and attempts to measure how serious they are about cutting their net emissions to zero. Sentences from previous climate-related projects are also added.[73]

Models of the ClimateBERT-DAPT family have been fine-tuned in other contexts - for example, Garrido-Merchán et al. (2023) fine-tune a model of this family to predict whether a sentence is climate-related or not using the dataset ClimaText (see Section 4.1.1 for dataset details). While the authors do show that a ClimateBERT-DAPT model that has been additionally fine-tuned on climate-relevant sentences outperforms the baseline ClimateBERT-DAPT model, it is not clear which model of the ClimateBERT-DAPT group has been fine-tuned and the resulting fine-tuned model has not been published.

Evaluation and results: The foundation models ClimateBERT$_F$, ClimateBERT$_S$, ClimateBERT$_D$, and ClimateBERT$_{D+S}$ are compared against DistilRoBERTa on the downstream tasks of (1) text classification (classify paragraphs as climate-related or not), (2) sentiment analysis (classify climate-related paragraphs as expressing risk, opportunity, or being neutral), and (3) fact-checking. All models outperform the baseline in terms of F1 score, with the highest F1 scores achieved by ClimateBERT$_S$ in task (1), ClimateBERT$_F$ in task (2), and ClimateBERT$_{D+S}$ in task (3). The results are originally reported in Webersinke et al. (2022) and aggregated in Table 2.

The performance on paragraph-level classification tasks is reported in the context of ClimateBERT$_{CTI}$, an NLP methodology for calculating a "cheap talk index". The index aims to measure the level of superficiality in companies' climate commitments (Bingler et al., 2024). There are four baseline scores obtained from four machine learning and deep learning models: a Least Absolute Shrinkage and Selection

[70] https://huggingface.co/datasets/climatebert/tcfd_recommendations.

[71] This definition has been obtained from an official document published by the European Commission titled "Guidance on the implementation/application of Directive 2005/29/EC on unfair commercial practices", available at https://enterprise.gov.ie/en/legislation/legislation-files/european-commission-guidance-on-the-implementation-application-of-the-unfair-commercial-practices-directive.pdf.

[72] https://huggingface.co/datasets/climatebert/environmental_claims.

[73] https://huggingface.co/datasets/climatebert/netzero_reduction_data.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 2. *ClimateBERT baselines and performance (loss / F1). Validation means that loss has been calculated for the validation dataset*

Task: token prediction	Baseline (model + validation loss)	ClimateBERT$_F$	ClimateBERT$_S$	ClimateBERT$_D$	ClimateBERT$_{D+S}$
Predict randomly masked tokens in domain-specific texts	DistilRoBERTa: 2.238	1.157	1.205	1.204	1.203
Task: paragraph classification	**Baseline (model + F1 score)**	**ClimateBERT$_F$**	**ClimateBERT$_S$**	**ClimateBERT$_D$**	**ClimateBERT$_{D+S}$**
climate-related or not	DistilRoBERTa: 0.986	0.989	0.991	0.988	0.988
sentiment analysis	DistilRoBERTa: 0.825	0.838	0.836	0.835	0.834
fact-checking	DistilRoBERTa: 0.748	0.755	0.753	0.752	0.757

Table 3. *Paragraph classification tasks: baseline models and F1 scores versus ClimateBERT*$_F$

Classification: paragraph	Baseline models	Baseline F1 scores	ClimateBERT$_F$
climatebert/distilroberta-base-climate-detector	LASSO[a] Naïve Bayes SVM + BoW SVM + ELMo	0.86 to 0.89 0.87 0.87 0.89	0.97
climatebert/distilroberta-base-climate-sentiment	LASSO Naïve Bayes SVM + BoW SVM + ELMo	0.3 to 0.62 0.72 0.72 0.75	0.8
climatebert/distilroberta-base-climate-specificity	LASSO Naïve Bayes SVM + BoW SVM + ELMo	0.55 to 0.67 0.75 0.75 0.76	0.77
climatebert/distilroberta-base-climate-commitment	LASSO Naïve Bayes SVM + BoW SVM + ELMo	0.62 to 0.69 0.75 0.76 0.79	0.81

Note: Details on the baseline models' training can be found in Bingler et al. (2022a, 2024).

[a]LASSO stands for Least Absolute Shrinkage and Selection Operator, a type of machine learning model.

Operator (LASSO), a Naïve Bayes classifier, a Support Vector Machine (SVM) with Bag-of-Words (BoW), and an SVM with Embeddings from Language Model (ELMo). Based on the reported F1 scores, ClimateBERT$_{CTI}$ outperforms all baseline models on the four tasks. Detailed evaluation information is available in Table 3 and is based on the results reported in Bingler et al. (2022a, 2024).

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 4. *Sentence classification tasks: baseline models and F1 scores versus fine-tuned ClimateBERT$_F$ evaluated on the test data split*

Classification: sentence	Baseline models	Baseline F1 scores	ClimateBERT$_F$
climatebert/environmental-claims	tf-idf SVM	0.82	0.87
	DistilBERT	0.85	
	RoBERTabase	0.86	
	RoBERTalarge	0.91	
climatebert/transition-physical	n/a	n/a	0.97
climatebert/renewable	n/a	n/a	0.96
climatebert/netzero-reduction	DistilRoBERTa	0.94	0.96
	RoBERTa-base	0.96	
	GPT-3.5-turbo	0.92	

Note: There were many baseline models as points of comparison for climatebert/environmental-claims; this table only includes those with comparable performance. For more details see Stammbach et al. (2022).

In terms of the fine-tuned models for sentence-level classification, *climatebert/environmental-claims* is compared against SVM-based models and three Transformer models. On the train dataset, *climatebert/environmental-claims* is outperformed by both RoBERTa-base and RoBERTa-large; on the development dataset, it either outperforms or is on par with other Transformer-based models, and on the test dataset, it is outperformed only by RoBERTa-large. For *climatebert/renewable* and *climatebert/transition-physical*, Deng et al. (2023) report an F1 score of 0.96 and 0.97, respectively. Finally, Schimanski et al. (2023a) report an F1 score of 0.962 for *climatebert/netzero-reduction*, which is higher than the F1 scores of DistilRoBERTa and GPT-3.5-turbo, and very close to the score of RoBERTa-base, 0.958. These results are summarized in Table 4.

Access, transparency, and engagement: The four ClimateBERT-DAPT and the nine fine-tuned models, as well as most of the datasets used in fine-tuning the models on downstream tasks, are available on HFH at the time of writing.[74] Transparency in terms of data collection, usage, and evaluation approach is also preserved. Webersinke et al. (2022) include information about the carbon footprint of developing the ClimateBERT-DAPT models, accompanied by a climate performance model card.

In terms of engagement, of the ClimateBERT-DAPT models, ClimateBERT$_F$ has the highest number of downloads (over 200 thousand); of the models for paragraph-level text classification, the model *climatebert/distilroberta-base-climate-detector* has the highest number of downloads (nearly 1.4 million), while *climatebert/environmental-claims* is the most popular LM among the sentence-level classification models (over 58 thousand downloads). Fine-tuned models for paragraph classification seem to be more popular than sentence classification models. Table 5 gives information about the all-time downloads of all models of the ClimateBERT family.

4.3.3. DARE

The BERT-based distillation and reinforcement ensemble model (DARE) designed by Xiang and Fujii (2023) should show how a lighter, more efficient version of purely BERT-based models can be developed for text classification tasks in a low-data setting. The model's **intended audience and use** are stakeholders interested in using a language model to analyse ambiguities of CC-related information for sentiment analysis (risk, opportunity, neutral) and fact-checking, and do so with limited access to compute power.

[74] https://huggingface.co/climatebert.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 5. *Number of all-time downloads for 13 models of the ClimateBERT family, in descending order for each LM type*

LM type/task	LM name / HFH identifier	No. of downloads
Foundation	ClimateBERT$_F$	**200437**
	ClimateBERT$_S$	3612
	ClimateBERT$_{D+S}$	2776
	ClimateBERT$_D$	1270
Paragraph classification	*climatebert/distilroberta-base-climate-detector*	**1389207**
	climatebert/distilroberta-base-climate-sentiment	476501
	climatebert/distilroberta-base-climate-specificity	298495
	climatebert/distilroberta-base-climate-commitment	189761
	climatebert/distilroberta-base-climate-tcfd	92608
Sentence classification	*climatebert/environmental-claims*	**58673**
	climatebert/netzero-reduction	53921
	climatebert/renewable	22169
	climatebert/transition-physical	20840

Note: The count of downloads was retrieved in August 2025.

Model architecture, training, and data: In their choice of architecture, the DARE developers attempt to address two concerns in model engineering and usage: (1) training and deployment of large language models are resource-intensive in terms of compute power, and (2) developing fine-tuned models is a data-intensive process. Concern (1) is addressed by combining knowledge distillation and domain adaptation. As explained in Section 4.3.2, *knowledge distillation* is a method where a smaller and simpler model is trained to mimic the behavior of a larger, more complex model. For DARE, the "teacher" model is an encoder-only 12-layer BERTbase model, while the "student" model is a Bi-LSTM-Attention model.[75] *Domain adaptation* is the use of domain-specific text to adapt an existing model to a new target domain. To address domain-specific data scarcity, Xiang and Fujii (2023) propose a new data augmentation strategy, where a component titled *Generator-Reinforcer Selector collaboration network* is used to replace nouns, verbs, adjectives, and adverbs in a sentence with suitable candidates. The Generator proposes sentences with replaced words, while the Reinforced Selector chooses samples that truly augment the data, rather than add noise to it.

The **data** used for domain-adaptive pretraining comprises scientific literature related to climate change and health (Berrang-Ford et al., 2021), published between 1 January 2013 and 9 April 2020, obtained from the Web of Science Core Collection and Scopus. While Xiang and Fujii (2023) do not provide a pointer to the data, it seems that part of it is publicly available as supplementary materials to work done by Berrang-Ford et al. (2021).[76] For sentiment analysis, the authors create a dataset of 1220 hand-selected paragraphs extracted from the Web of Science records and another 1000 paragraphs from Scopus records, and, similarly to the sentiment analysis dataset used in fine-tuning *climatebert/distilroberta-base-climate-sentiment*, annotate the paragraphs as OPPORTUNITY (positive sentiment), RISK (negative sentiment), and NEUTRAL, using the annotation software Prodigy.[77] Finally, these datasets are augmented with the Generator-Reinforcer Selector data augmentation strategy described above, which doubles the number of records from 40,671 to 80,750. The latter is split in a train set of 60,560 records and a development set of

[75] A machine learning architecture used for sequence modelling tasks combining two architectures: A bidirectional long short-term memory (Bi-LSTM) neural network, which processes a sequence in both forward and backward direction, and an attention mechanism, which allows the model to focus on specific parts of the sequence.

[76] The data is available as an Excel document with information about 16078 academic documents, available at https://zenodo.org/records/4972515.

[77] https://prodi.gy/.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 6. *Performance comparison of different models on sentiment analysis (F1) and fact-checking (macro F1)*

Model	F1, sentiment	Macro F1, fact-checking
BERTbase	0.931	**0.791**
BERTbase, domain pretrained	**0.955**	n/a
TinyBERT	0.870	n/a
RoBERTa	n/a	0.712
DistillRoBERTa	n/a	0.704
DistilBERT	0.899	n/a
ClimateBERT	0.875	0.729
CCLA+Max-Pooling	0.829	n/a
Bi-LSTM-Attention	0.719	n/a
POS-Bi-LSTM-Attention	0.882	n/a
DARE	0.894	0.788

Note: The type of F1 score for the sentiment analysis task is not specified. This is a simplified representation of the results; for a detailed overview of various experimental setups, see Xiang and Fujii (2023).

20,190 records. The authors do not clarify what the term "records" refers to and what the average number of tokens per record is; therefore, it is not possible to estimate the number of tokens in the training corpus. Neither the sentiment-annotated data nor the augmented data are publicly available.

Evaluation and results: Evaluation is conducted on two downstream tasks: sentiment analysis, using the sentiment analysis dataset created as part of the study, and fact-checking, using the Climate-FEVER dataset. The model's performance is compared against eight other models: BERTbase and domain-pretrained BERTbase (Devlin et al., 2019), TinyBERT (Jiao et al., 2020), DistilBERT (Sanh et al., 2020), ClimateBERT (Webersinke et al., 2022), a coordinated Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) attention + Max Pooling model (Zhang et al., 2019), a Bi-LSTM-Attention model, and a POS-Bi-LSTM-Attention model. The latter two incorporate the same type of model as the *student* model used in the knowledge distillation process described above (a Bi-LSTM-Attention model); in the latter one, *POS* stands for part-of-speech, as the authors incorporate POS vectors, which are believed to strengthen the sentiment connection and features in the sentiment analysis task.

On the sentiment analysis task, DARE's accuracy and F1 score are lower than those of BERTbase, the domain-pretrained BERTbase model, and DistilBERT. DARE comes second best in the fact-checking task, being outperformed by BERTbase. The authors point out that although DARE is outperformed by BERTbase and DistilBERT, the model still performs comparatively well, with a 50.65x and 12.66x inference speed-up respectively, thereby offering substantial acceleration. Finally, the authors conduct an ablation study and show that all building blocks of the DARE model positively contribute to its performance. Table 6 provides an overview of DARE's performance relative to other language models.

Access, transparency, and engagement: The model is not available either as a public or a proprietary model. The sentiment analysis dataset and the augmented data have not been published. The authors give a detailed report on the model parameter settings and mention that all experiments are conducted on a single Nvidia 16 GB V100 GPU. The speed-up times are also recorded, which is of relevance given that one of the study's central goals is to provide a light-weight model.

4.4. Text classification and text generation

4.4.1. ClimateGPT-2

The model's **intended audience** is policymakers, researchers, and climate activists, and its **intended use** is to help them analyse large and complex climate-change-related collections of texts, by making use of the model's two main functionalities: claim (text) generation and fact-checking.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Model architecture, training, and data: Vaghefi et al. (2022) initialize this model with weights from GPT-2 (Radford et al., 2019). GPT-2 belongs to a family of generative pretrained Transformer-based decoder-only models trained on general-domain corpora. The family contains models with 117M, 355M, and 1.5B parameters. While Vaghefi et al. (2022) mention that GPT-2 comes in different sizes, the size of the GPT-2 model that has been used in the study is not specified.

The dataset for domain-adaptive pretraining consists of 360,233 abstracts from papers on climate change published by well-known scientists in the CC domain, whose names have been retrieved from a list of CC scientists curated by Reuters.[78] Once domain-adaptive pretraining is completed, the model is fine-tuned on two tasks: (1) fact-checking, using the Climate-FEVER dataset (Diggelmann et al., 2020) (see Section 4.1 for dataset information) and (2) text generation, where the model is tasked with generating CC texts prompted with (a) a title only and (b) a title accompanied by a list of keywords.

Evaluation and results: The model's performance is evaluated against that of GPT-2. In the fact-checking task, where the model should correctly "decide" if an evidence sentence supports or refutes a claim, an F1 score of 0.72 is reported, which is higher than GPT-2's baseline score of 0.67. For text generation, Vaghefi et al. (2022) report improved performance, reflected in a lower validation loss. The authors also report that the model generates semantically coherent sentences and that the initial three sentences of the model's output are related to the title and/or the keywords.

Access, transparency, and engagement: The model is available in a public GitHub repository.[79] The collection of abstracts used to conduct the domain-specific pretraining is not available. In terms of model evaluation, the authors do not elaborate what metrics for semantic coherence and relatendess of the generated sentences to the keywords and title are being applied. The authors mention the GPU infrastructure and time required for model training, but do not provide a model scorecard.

4.5. Domain-specific LMs: summary

Section 4 gave an overview of several domain-specific models available for analysing climate change texts. Two groups of target audiences can be distinguished: researchers, scientists, climate activists, and journalists on the one hand (ClimateGPT, ClimateGPT-2, DARE, BART-based summarization model), and financial and corporate sustainability analysts on the other (ClimateBERT, ClimateQA). The model architecture and design choices reflect the model's intended use: text-generation and question-answering models mostly rely on decoder-only architecture (ClimateGPT, ClimateGPT-2), classification models use encoder-only architecture (ClimateBERT, ClimateQA, DARE), while text summarization models use encoder-decoder architecture (BART-based summarization model).

The degree of improvement achieved by techniques for continued pretraining and/or fine-tuning can be observed by comparing the performance of target models to that of baseline models: in ClimateGPT, a considerable improvement is observed for ClimateGPT-7B on climate-change tasks (+8.6 relative to a same-size Llama-2 model). ClimateBERT's best-performing model is ClimateBERT$_F$ across next-token prediction and text classification tasks. DARE's slightly lower performance compared to its baseline model is compensated for by faster inference time. Finally, Climate-GPT2 sees an improvement of 0.05 against its baseline model.

Since the model design techniques of this section are data- and compute-intensive, an interesting point of comparison is the use of GPUs for the models discussed in the section. From Table 7 it is evident that ClimateGPT's design is the most resource-intensive, which is understandable given the size of the model and the pretraining dataset. The other models are considerably smaller and require fewer resources. The pretraining data for these models is mostly unavailable, while most of the training datasets for ClimateBERT's classification models can be downloaded from the dedicated HFH repository (see Section 4.3.2 for more details).

[78] https://www.reuters.com/investigates/special-report/climate-change-scientists-list/.

[79] https://github.com/saeedashraf/climateGPT-2.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 7. *Model, model size, GPU type and time (in hours or days) for model training*

Model	Size	GPU	Time
ClimateGPT-7B	7B	HPC cluster with 31 nodes, each with 8 Nvidia H100-SXM GPUs	157 GPU hours
ClimateGPT-13B	13B		301 GPU hours
ClimateGPT-70B	70B		2182 GPU hours
ClimateGPT-FSC-7B	7B		14,131 GPU hours
ClimateGPT-FSG-7B	7B		14,288 GPU hours
ClimateGPT (total) (Thulke et al., 2024)	see above		31,059 (training) + 3685 (experiments) GPU hours.
ClimateBERT$_F$ + 3 fine-tuned classification models (Webersinke et al., 2022)	82.4M	2 x Nvidia RTX A5000 GPU	350 hours (all experiments), 48 hours (final models)
climatebert/distilroberta-base-climate-detector (Webersinke et al., 2022; Bingler et al., 2024)	82.3M	1 x Nvidia RTX A5000 GPU	n/a, possibly entailed in ClimateBERT
climatebert/distilroberta-base-climate-sentiment (Webersinke et al., 2022; Bingler et al., 2024)	82.3M		
climatebert/distilroberta-base-climate-specificity (Bingler et al., 2024)	82.3M		
climatebert/distilroberta-base-climate-commitment (Bingler et al., 2024)	82.3M		
climatebert/distilroberta-base-climate-tcfd (Bingler et al., 2022b, 2024)	82.3M	n/a	not able to infer information
climatebert/environmental-claims (Stammbach et al., 2022)	82.3M	1 x Nvidia GeForce GTX 1080 Ti GPU	60 hours
climatebert/transition-physical (Deng et al., 2023)	82.3M	n/a	n/a
climatebert/renewable (Deng et al., 2023)	82.3M		
climatebert/netzero-reduction (Schimanski et al., 2023a)	82.3M		
ClimateGPT-2 (Vaghefi et al., 2022)	124M	2 x Nvidia Gforce 2800 GPUs	3 days
DARE (Xiang and Fujii, 2023)	7.26M	Nvidia 16 GB V100	n/a
ClimateQA* (Luccioni et al., 2020)	125M	12 GB GPU	ca. 5 hours

Note: GPU hours used only when explicitly stated. Size of ClimateBERT$_F$ and its related models reported as stated in HFH repositories; *size of ClimateQA inferred from its base model.

5. General language models and external climate-relevant resources

The advent of more powerful LLMs trained on larger amounts of diversified data has led researchers to turn to knowledge-guided NLP (Ignat et al., 2024). This approach does not involve creating LMs from scratch, or continued pretraining or fine-tuning of existing open-source models, as was the case with models discussed in Section 4. The focus here shifts on providing an existing LLM with an external source of domain-specific information; this source continuously updates the LLM's data to prevent it from generating incorrect or outdated information (Vaghefi et al., 2023).

This section describes five LLM-based question-answering systems and one question-answering and scoring system: ChatClimate (Vaghefi et al., 2023), an LLM agent relying both on a database and a Google search functionality (Kraus et al., 2023), My Climate Advisor (Nguyen et al., 2024), a system for responsible question-answering developed by Climate Policy Radar (Juhasz et al., 2024), ChatNetZero (Hsu et al., 2024), and CHATREPORT (Ni et al., 2023).

5.1. Question-answering

5.1.1. ChatClimate

Built on top of GPT-3.5 Turbo and GPT-4, ChatClimate is **intended** to answer climate questions by drawing on factual information (Vaghefi et al., 2023). The system's **intended audience** includes decision-makers and anyone interested in obtaining trustworthy information on climate change.

System architecture and data: The decoder-only models GPT-3.5 Turbo and GPT-4, which power the application, remain unchanged. To improve the quality of the generated answers, the authors rely on (1) external long-term memory and (2) prompts. External long-term memory is built by parsing PDF files of IPCC reports of the sixth assessment cycle into JSON files.[80] The extracted text is then chunked into smaller sizes using LangChain,[81] a framework for building LLM-based applications, and embedded with *text-embedding-ada-002*, an embedding model developed by OpenAI.[82] The vectors are then stored into a vector database from which they can be retrieved.

Using prompts, the QA tool is guided whether to make use of: (1) GPT-4 only, (2) IPCC AR6 reports only, an approach dubbed *ChatClimate*, which is configured to exclude the LLM's in-house knowledge, and (3) IPCC AR6 reports and in-house GPT-4 knowledge, an approach dubbed *Hybrid ChatClimate.*

Evaluation and results: A set of 13 questions with five levels of difficulty is used to evaluate the system. The difficulty levels and the number of questions per level are: very low (level 1)/1 question, low (2)/1 question, medium (3)/4 questions, high (4)/4 questions, and very high (5)/3 questions.[83] The system's answers are evaluated by experts, who give a score of 1 to 5 to the answers generated by each augmented retrieval method, with 1 being the lowest and 5 the highest score. It is observed that ChatClimate, which is instructed not to take into account the in-house knowledge of the model, tends to hallucinate[84] less than GPT-4 and Hybrid ChatClimate. Each system version's answer to questions Q3

[80] Later iterations of the experiment include reports from the World Meteorological Organization (WMO), alongside more IPCC reports. The IPCC (a) and WMO (b) reports used in the study include: (a): Summary for Policymakers from each of the Working Groups (I, II, III): 3 PDF files; All chapters (Working Group I: Chapters 1–12, Working Group II: Chapters 1–18, Cross-Chapters 1–12, Working Group III: Chapters 1–17) and technical summaries from each of the three working groups (https://www.ipcc.ch/report/ar6/wg1, https://www.ipcc.ch/report/ar6/wg2, https://www.ipcc.ch/report/ar6/wg3): 3 PDF files; The IPCC Synthesis Report 2023: 1 PDF file. (b): 2022 State of Climate Services: Energy (WMO No. 130): 1 PDF file; WMO Global Annual to Decadal Climate Update 2023–2027: 1 PDF file; State of the Climate in Asia 2022 (WMO No. 1321): 1 PDF file; State of the Climate in South-West Pacific 2021 (WMO No. 1302): 1 PDF file; State of the Climate in Africa 2021 (WMO No. 1300): 1 PDF file.

[81] https://github.com/langchain-ai/langchain.

[82] As mentioned before, "embedding" is a process of converting data, such as text, into numerical representations i.e. vectors. The vectors can then be used to find paragraphs of the text that are similar to a query using a similarity metric.

[83] An example of an easy question is: *What does overshoot mean?* A question on a level 5 difficulty is: *Will glaciers in Scotland melt?*

[84] A type of error identified in the output of generative LMs, which refers to generated content that is unsensical or unfaithful to the provided source content (Ji et al., 2023).

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 8. *Average response accuracy score for each system on a total of 11 questions (responses to questions 1 and 2 are not available)*

Difficulty level	No. of questions	Hybrid ChatClimate	ChatClimate	GPT-4
1	1	5	5	2
2	1	4	4	4
3	2	5	4	3
4	4	4.25	4.25	2.5
5	3	3.67	4	2

to Q13 is available in the supplementary materials to the paper.[85] To facilitate comparability between the three models, I summarized the contents of the evaluation report in Table 8. The authors acknowledge that a more comprehensive evaluation is needed to evaluate the system's responses, alongside a fact-checking step.

Access and transparency: ChatClimate is available as a web service,[86] where users can choose between GPT-3.5 Turbo and GPT-4, used in two modes: stand-alone and hybrid. In terms of transparency, at the time of writing, there is no official disclosure about the data that powered the training of OpenAI's GPT-3.5 Turbo and GPT-4 models. The documents used to curb the models' hallucination and to overcome the limitation of outdated data are publicly available as PDF files; the parsed data that has been used for system development is not accessible. The code to reproduce the ChatClimate system is published on GitHub.[87]

5.1.2. LLM CC agent: a prototype

LLM agents are systems that are built on top of an LLM, which acts as an "agent" performing various tasks, such as data analysis or web browsing. Kraus et al. (2023) present a prototype system whose **intended use** is to serve as a question-answering system that has access to recent and precise information on the topic of climate change, and the **intended audience** is perceived to be organizations, institutions, and companies interested in obtaining such information.

System architecture and data: The backbone of the LLM agent is OpenAI's model *text-davinci-003*, which is a decoder-only model. The system is built in the LangChain environment and utilizes a ReAct framework (Yao et al., 2022).[88] The resulting LLM agent should be able to interact with structured data, such as data stored in tables (pandas dataframes in Python), and, if necessary, conduct a simple Google search and retrieve additional information.

The system is guided with the help of a prompt to (1) read the question asked by the user, (2) think about what it should do, (3) describe what type of action the LLM agent should undertake, (4) specify the input to the action, and (5) present the result of the action. The five sections in the prompt are called: Question, Thought, Action, Action Input, and Observation. Finally, the system is requested to perform a Google search only if it cannot find relevant information in the data from Climate Watch, an online platform aggregating various climate-change-related datasets.[89]

Evaluation and results: The LLM agent is asked two questions (Q1 & Q2): Q1, *What is the average emission of Italy between 2010 and 2015?*, whose answer requires the use of one data source, and Q2,

[85] Available at: https://www.nature.com/articles/s43247-023-01084-x.

[86] https://www.chatclimate.ai/.

[87] https://github.com/saeedashraf/chatipcc.

[88] *ReAct* stands for Reasoning and Acting and it is a way of integrating "reasoning" capabilities with "acting" or "decision-making" capabilities in an LLM. The authors use the term "capability" to describe a model's functionality for multi-step processing, where a model can break up a task into smaller, consecutive tasks in an attempt to generate a more accurate response.

[89] https://www.climatewatchdata.org/.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Which European country has the most ambitious net zero plans? How did the emissions of this country develop over the last 10 years? Remember only to include single countries., whose answer requires the use of combined data sources. The LLM agent is reported to have successfully retrieved the relevant data and calculated the emissions of CO_2 in Italy between 2010 and 2015; for Q2, the LLM agent is reported to have successfully retrieved the additional data by performing a Google search. No additional evaluation steps are reported.

Access and transparency: The LLM that serves as a backbone to the tool is OpenAI's *text-davinci-003*, which at the moment of writing is a deprecated model. Data used as a source for the system to generate an answer to the user question is provided by Climate Watch and is described as publicly available. It is not clear whether the tool that allows the system to browse through the data obtained from Climate Watch is open-source. The other library used to augment the system's functionalities, Google Serper,[90] is a proprietary API which offers new users a limited number of free queries. The prototype presented in the paper is not publicly available. The authors express awareness of the carbon footprint of LLMs, both in the context of training and inference, and point out that further research must ensure that the benefits of using LLMs are not outweighed by the environmental cost.

5.1.3. My Climate Advisor

Nguyen et al. (2024) develop a RAG-based question-answering system whose **intended use and audience** is to assist farmers and farm advisors gain access to information from scientific papers, grey literature, and climate projection data. The focus is on Australia as a geographical region, and the goal is to help farmers improve their resilience to the effects of climate change.

System architecture and data: The RAG database is built by drawing on three sources of information. The first addresses general-purpose agriculture questions. It is a corpus of 1.36 million articles obtained from the Semantic Scholar Open Research Corpus (S2ORC) using the labels "Agricultural and Food Sciences" and "Environmental Science". The second source addresses climate adaptation issues, and the dataset for it is a 126,000-article corpus obtained from the top 100 highest-impact agriculture journals and from the scientific publisher Elsevier. Finally, the third source targets climate issues specific to Australia and relies on an expert-curated corpus of climate risk information, books, and industry reports, with a total of 28 documents. The corpus size is reported in gigabytes, number of documents, and number of chunks, where each chunk is approximately 400 tokens long. As per this information, the corpus might have approximately 12.3B tokens.[91] The RAG database is used with the model Llama-3-8B as a backbone.

Evaluation and results: The output of the system is compared against 11 models across seven criteria: context, structure, language, specificity, comprehensiveness, accuracy, and citation. The models against which the system is compared are: GPT-4 Turbo, Llama-3-70B, Claude 3 Opus, Gemini 1.5 Pro, Llama-3-8B, Claude 3 Haiku, Mistral-7B, Gemini 1.0 Pro, GPT-3.5 Turbo, Llama-3-70B+RAG, and Mistral-7B +RAG. The evaluation is performed by two human experts, a climate scientist and an agronomist, who score the systems' answers to a set of 15 questions about the Australian climate change impacts and adaptation. The questions had been developed in consultation with climate risk and adaptation experts. The scores range between 0 and 4, and an average score is calculated from the scores on the seven criteria. The results reveal that GPT-4 Turbo has the highest score on 6 of 8 categories, and that Llama-3-8B + RAG scores higher only on the criterion *citation.* However, the inter-annotator agreement is rather low, and preference for GPT-4 Turbo and Llama-3-70B is observed.

Access and transparency: The dataset used for the database is not publicly available at the time of writing. The developed system is also not publicly available at the time of writing, but the authors promise to publish it in the future.

[90] https://serpapi.com/.

[91] Based on the choice of tool (https://crates.io/crates/text-splitter), these are likely subword-based rather than word-based tokens.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

5.1.4. Climate policy radar: responsible question-answering

Juhasz et al. (2024) present a RAG-supported system, ultimately to be hosted by Climate Policy Radar (CPR), whose **intended use** is to improve accessibility to information contained in documents on climate law and policy. The **intended audience** includes policymakers, analysts, academics, and any professionals who read climate policy and law documents as part of their job.

System architecture and data: The authors emphasize the importance of evaluation and user experience (UX) considerations in the system design. The system developed in this study has four components, of which the first and fourth components serve as guardrails against malicious content entered either by the user or generated by the LM. Sandwiched between them are an information retrieval component and an answer synthesis component, both of which are thoroughly evaluated. Models used as a backbone in the generation experiment include Llama-3.1-70B Instruct, Llama-3.1-8B Instruct, Gemini 1.5 Flash, and ClimateGPT-7B.

The dataset comprises a sample of 550 documents sourced from CPR's database of national laws and policies and from submissions from the United Nations Framework Convention on Climate Change (UNFCCC). The sample of texts is distributed equally across World Bank Regions. For the evaluation portion that focuses on adherence to CPR's policy standards for generation of content, the sample is complemented with documents published by the International Energy Agency (IEA), the International Atomic Energy Agency (IAEA), the Organization for Security and Co-operation in Europe (OSCE), and the World Meteorological Organization (WMO).

Evaluation and results: This study focuses on creating an evaluation pipeline that integrates human annotations and LLMs-as-a-judge. The comprehensive evaluation is done along two main tracks: (1) retrieval of the most relevant passages from the source documents, and (2) generation of answers that meet a set of criteria, which include: alignment with the CPR generation policy, faithfulness,[92] formatting, and system-response.

For the evaluation under (1), Juhasz et al. (2024) have human annotators mark passages as relevant or not to 194 synthetic questions and use the annotations to deploy an automated solution using LLM-as-a-judge with GPT-4o. The human evaluators pinpoint several problems when assessing retrieved passages as relevant, namely: a passage that "signalled" the possibility of a relevant passage nearby would be marked as relevant; imprecise language in the passage makes it difficult to assess its level of usefulness; and document metadata is at times necessary to respond to a query. A separate human-annotated dataset is created to measure the degree to which the models can generate an answer in line with CPR's generation policy. To this end, 16 domain experts from several national governments and international organizations, including the United Nations (UN), the International Renewable Energy Agency (IRENA) and WMO, participate in a 3-week annotation sprint to annotate generated data related to 800 documents.

The system's generation is also assessed across three prompt templates: a basic task explanation, a prompt steering the system towards an "educative" response, and a Chain-of-Thought (CoT) prompt. These are populated with non-adversarial queries, which are sourced from user interviews, and adversarial queries, whose purpose is to "nudge" the system towards generating an answer that violates the guardrails or the prompt instructions.

The aggregated results across the prompt types and evaluation levels show that a basic prompt seems to work best for faithfulness and adherence to CPR policy, while a prompt for an educative response results in the system observing the formatting requirements. The models seem to successfully identify adversarial queries, as 6.4% to 15% of the no-response cases are related to this type of query. It is found that violations of the CPR generation policy are concerning, especially given that the end-use of such systems is a user-facing scenario. In addition, during the annotation sprint, it is found that policy violations correlate with violations of faithfulness and that violations of formatting, which include missing or non-existing citations, coincide with hallucinations.

[92] A type of hallucination error where the model diverges from the provided context.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

In terms of **access and transparency**, the prompts, the methodology, and the evaluation datasets are publicly available.[93]

5.1.5. ChatNetZero

Hsu et al. (2024) acknowledge that LLM-powered chatbots, such as Google's Gemini (Team et al., 2023) or OpenAI's ChatGPT (OpenAI, 2022), have become the first point of contact for many when conducting initial research on a topic. While the intended audience of ChatNetZero is not narrowly defined, the **intended use** of the system is to serve as a question-answering platform for climate policy-specific information, with a special focus on net-zero texts. The authors collaborate with experts on two occasions: (1) to develop the dataset that serves as a basis for the RAG component, and (2) to evaluate ChatNetZero's output.

System architecture and data: The database is built from the following documents: a report titled "Integrity Matters: Net Zero Commitments by Businesses, Financial Institutions, Cities and Regions" by the United Nations High-Level Expert Group (HLEG) (HLEG, 2022), the Net Zero Tracker database[94] and the Net Zero Stocktake reports (Net Zero Tracker, 2022, 2023), and the Corporate Climate Responsibility Monitor Reports published by the NewClimate Institute[95] (New Climate Institute, 2022, 2023). In addition to the database, ChatNetZero has a module for query processing, as well as anti-hallucination, reference, and enhanced analytical capabilities modules. *Query processing* involves a pre-processing step for identification of all *actors*[96] in the query, and ensuring that the retrieved passages from the database refer to the identified actor. The *anti-hallucination* module is used to post-process the LLM+RAG output and verify that each generated sentence can be traced to the original passage from the dataset.[97] The *reference* module adds a reference based on the text passage's ID. Finally, the *enhanced analytical capabilities* module restructures tabular source data into natural sentences for easier retrieval. The authors also include the system prompt sent to GPT-4 Turbo, which seems to be the backbone LLM.

Evaluation is performed (1) by checking the factuality of generated answers and (2) by having ten climate scientists and policy experts score answers to 12 questions on a scale of 1 to 5 across three dimensions: quality, factual accuracy, and relevance. To be considered factually accurate in the sense of (1), the system's response would have to contain the exact factual information from the reference material, including figures. On (1), the system outperforms ChatClimate (Vaghefi et al., 2023), GPT-4 Turbo, Gemini 1.0 Ultra, and Coral with Web Search (Cohere, 2023). However, in the evaluation by experts, ChatNetZero receives the lowest ranking among the LLMs, which is thought to be a consequence of its answers being shorter in length compared to those of the other models.

Access and transparency: The system is available online at the time of writing.[98] The generated answers that were used to conduct the evaluation are also available online.[99]

5.2. Question-answering and scoring

5.2.1. CHATREPORT

Ni et al. (2023) develop a system whose **intended use** is assisting the analysis of sustainability reports by calculating (1) a report's conformity score (on a scale from 0 to 100) to the reporting guidelines developed by the Task Force on Climate-Related Financial Disclosures (TCFD) and (2) an option for user-defined analysis with question-answering. Its **intended audience** includes policymakers, investors, and the general public.

[93] https://huggingface.co/ClimatePolicyRadar.

[94] https://zerotracker.net/.

[95] https://newclimate.org/.

[96] Possibly named entities.

[97] This is done by embedding each generated sentence and comparing it against selected chunks of the RAG module to verify the sentence's origin.

[98] https://chatnetzero.ai/.

[99] https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/VZYKID.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Ni et al. (2023) use OpenAI's model *text-embedding-ada-002* to retrieve text embeddings, and ChatGPT and GPT-4 as backbone models for summarization and question-answering tasks. CHATREPORT's **architecture** includes the following elements: chunking and embedding a target report, generating a summary of the target report grounded in TCFD questions, calculating a TCFD adherence score, and a question-answering module. To combat hallucinations, the authors make the model's answer traceable (similarly to Thulke et al. (2024) and Hsu et al. (2024)) by assigning numbers to the source texts. Domain experts are utilized in an iterative process to craft prompts for summarization and question-answering: when a model generates an output based on a prompt, an expert provides feedback on the output. This feedback is then integrated into the prompt.

The system's success in retrieving the correct text passage for an answer and not hallucinating in the generated text is *evaluated* by (1) sampling 10 sustainability reports with 110 question-answer pairs, and (2) having two different annotators label the system's answers as containing hallucinations or not. Hallucination is defined in terms of content, where all generated content needs to be traceable to the source data, and source, where the model needs to retrieve the correct references (dubbed "honesty").[100] ChatGPT outperforms GPT-4 on this task with an "honesty" rate of 86.63%, as opposed to 51.5%. However, the inter-annotator agreement, measured as Cohen's Kappa score, between the two annotators on this task is 0.54 for ChatGPT and 0.21 for GPT-4, indicating that identifying hallucination is not an easy feat. The authors believe that the higher inter-annotator agreement for ChatGPT might indicate that it is easier to recognize this phenomenon in ChatGPT's answers.

In terms of **access and transparency**, the annotated data and code are published on GitHub,[101] while the system itself is available in a web interface. The usage of the system is also explained in a video available on YouTube.[102] The collection of corporate sustainability reports has not been published, and there is no information about the total cost of the API calls for this experiment.

5.3. General-purpose LMs: Summary

Section 5 discussed several studies where out-of-the-box LLMs are used with external sources in a question-answering scenario. The overarching intended use of these models is to reduce the time researchers, decision-makers, and policy analysts need to analyse large collections of data. An interesting outlier is the system My Climate Advisor, which specifically targets farmers from a given geographic region (Australia).

There are various techniques that the studies use to overcome data deficiencies and hallucinations in LLMs: from web browsing and access to databases with data structured in tables (LLM CC agent), to custom-made databases (ChatClimate, My Climate Advisor, Climate Policy Radar's system, ChatNetZero), as well as processing a target document in real time and using it as a source for generating answers (CHATREPORT). ChatClimate and the LLM CC agent focus on providing answers to general CC questions; the remaining QA systems seem to narrow down their domain to agriculture and climate change (My Climate Advisor), climate policies and laws (Climate Policy Radar), net zero carbon emissions (ChatNetZero), and corporate adherence to TCFD reporting guidelines (CHATREPORT).

There is a great variety of evaluation approaches to assess the quality of the generated answers. In some instances, experts score a model's output based on their own knowledge; in others, they cross-check if the model's output is grounded in a passage of the documents the model is expected to use. In addition to having experts assess models' output, Climate Policy Radar also creates annotated datasets and uses these as resources in an LLM-as-a-judge scenario. The number of experts can be two (My Climate Advisor, CHATREPORT), ten (ChatNetZero), or sixteen (Climate Policy Radar). Scoring methods differ across studies, with scores ranging from −2 to 2 (ClimateGPT), 0 to 4 (My Climate Advisor), 1 to 5 (ChatClimate), to annotating whether the generated answer contains hallucinations or not (CHATREPORT).

[100] A third annotator is involved to decide on disagreements.

[101] https://github.com/EdisonNi-hku/chatreport.

[102] https://www.youtube.com/watch?v=Q5AzaKzPE4Mt=15sab_channel=JingweiNi.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table 9. *Relative performance of question-answering (QA) systems (LLM+RAG)*

QA system	Compared against	Backbone model(s)	Better than	Worse than
ClimateGPT-FSC-7B (Thulke et al., 2024)	ClimateGPT-70B, ClimateGPT-7B	trained from scratch	None	ClimateGPT-70B, ClimateGPT-7B
ClimateGPT-7B (Thulke et al., 2024)	ClimateGPT-FSC-7B, ClimateGPT-70B	Llama-2-7B	ClimateGPT-FSC-7B	ClimateGPT-70B
ClimateGPT-70B (Thulke et al., 2024)	ClimateGPT-FSC-7B, ClimateGPT-7B	Llama-2-70B	ClimateGPT-FSC-7B, ClimateGPT-7B	None
ChatClimate (Vaghefi et al., 2023)	GPT-4, Hybrid ChatClimate	GPT-3.5 Turbo, GPT-4	GPT-4, Hybrid ChatClimate	None
LLM CC agent: a prototype (Kraus et al., 2023)	None, human evaluation (scoring) of generated answers	text-davinci-003 (OpenAI)	n/a	n/a
My Climate Advisor: Llama3-8B + RAG (Nguyen et al., 2024)	GPT-4 Turbo, Llama-3-70B, Claude 3 Opus, Gemini 1.5 Pro, Llama-3-8B, Claude 3 Haiku, Mistral-7B, Gemini 1.0 Pro, GPT-3.5 Turbo, Llama-3-70B + RAG, Mistral-7B + RAG	Llama–3–8B	Llama3-8B, Mistral-7B+RAG, Claude 3 Haiku, Mistral-7B, Llama-3-70B+RAG, Gemini 1.0 Pro, GPT-3.5 Turbo, Gemini 1.5 Pro[103]	GPT-4 Turbo, Llama-3-70B, Claude 3 Opus
ChatNetZero (Hsu et al., 2024)	ChatClimate, GPT-4, Gemini, Coral	GPT-4 Turbo	None	ChatClimate, GPT-4, Gemini, Coral
CHATREPORT (Ni et al., 2023)	RAG systems compared against each other	ChatGPT, GPT-4	ChatGPT+RAG outperforms GPT-4+RAG	n/a

Note: The table serves only as an overview of the scope of comparison*:* as there is no unified procedure for human evaluation, comparability between different QA systems is limited to the individual projects.

Climate Policy Radar's evaluation strategy (Juhasz et al., 2024) seems to have the most comprehensive design, simultaneously placing importance on factuality, user experience, and adherence to predetermined standards for generated texts.

While efforts to manually evaluate models' output provide some insights into the quality of their work, the lack of standardized evaluation guidelines, alongside the inherent subjectivity this type of evaluation entails, renders any comparison across different systems impossible. Nevertheless, to give an overview of the models these systems were compared against, and how they performed in this comparison, Table 9 summarizes the relative performance of QA systems (as judged by human evaluators) reported in each study. The table also includes the domain-specific QA system using ClimateGPT (Thulke et al., 2024).

6. Other relevant models and research topics

Sections 4 and 5 presented domain-specific LMs and systems powered by general-purpose LLMs designed for climate change tasks. This section should serve as a catch-all and briefly mention other relevant LMs and LM-powered applications for analysing climate change texts. In addition to briefly mentioning other relevant models, systems that focus specifically on fact-checking and models that have been developed for scientific tasks and deployed in a climate change context are also mentioned.

6.1. Additional relevant projects

With some LMs and LM-based systems, climate change is analyzed in the context of broader topics, such as environmental, social, and governance (ESG) information (ESGBert by Mehra et al. (2022)), climate health effects (CliMedBERT by Jalalzadeh Fard et al. (2022)), financial information in combination with

[103] Equal performance.

ESG information (FinBERT by Huang et al. (2023)), as well as a set of LMs that can classify environmental, social, and governance information individually (Schimanski et al., 2024b), models that classify sentences as related or not related to the topics of water, forest, biodiversity, and nature (Schimanski et al., 2023b), a RAG-based system that detects sustainable development goals in environmental reports (Garigliotti, 2024), as well as using AI to assess the environmental impact of a company as reported in company reports (Colesanti Senni et al., 2024).

6.2. Fact checking and claim detection

Environmental claim detection was briefly mentioned in Section 4.3 and in the overview of work done by Stammbach et al. (2022). Other works that are along the lines of environmental claim detection and fact checking include the EnClaim BERT-based classifier (Saha et al., 2024), and Climinator (Leippold et al., 2025), which is an LLM-supported system for automated fact-checking of climate change claims.

6.3. Benchmarking datasets

There are some efforts to improve and bring structure to the landscape of evaluation datasets. In addition to ClimaBench (Spokoyny et al., 2023), discussed in Section 4.1.1, other relevant work in this area that has not been mentioned in this survey yet includes that by Schimanski et al. (2024a), who released a dataset of 8.5K question-source-answer pairs, as well as Kurfali et al. (2025), who aggregated climate-relevant benchmarks for NLP research.

6.4. Other models and projects

INDUS Bhattacharjee et al. (2024) present a family of models trained on a large scientific corpus and applied, among other downstream tasks, to the task of named entity recognition (NER) for climate-specific named entities. This study is significant because it presents the first NER dataset developed exclusively for scientific literature on climate change.

Project Gaia This is an LLM-powered application developed by the Bank for International Settlements (BIS), together with the Bank of Spain, the Deutsche Bundesbank, and the European Central Bank.[104] It should assist analysts of climate-related risks in the financial system to automatically extract climate-related indicators from publicly available corporate reports.

7. Conclusions and future work

7.1. General summary

This paper reviews research done at the intersection of language models and climate change, with an emphasis on approaches to developing LMs and LM-based systems for text-based processing and analysis of climate data. It summarizes studies on the deployment of language models for climate change use-cases by analysing LMs and LM-based systems at four levels: (1) intended use and audience, (2) architecture, training, and data, (3) evaluation and results, and (4) access, transparency, and engagement. The study presents 22 LMs adapted for climate change data, 6 LM-based systems for analysing climate change documents, and several other possibly relevant LMs, LM blueprints, and LM-based projects. Appendix A provides a summary of the main findings presented in Sections 4, 5, and 6 (Tables A1, A2 and A3 in Appendix A).

The two **main functionalities** of the presented LMs and LM-based systems are: (1) text generation, in the form of question-answering, summarization, and generation of texts, and (2) text classification, where the model or the model-based system is expected to assign the correct label from a limited number of labels to a paragraph or a sentence (for example, if the paragraph or a sentence is about climate change or not), or to classify a paragraph as a potential answer to a question. Twelve of the LMs and LM-based

[104] https://www.bis.org/about/bisih/topics/suptech_regtech/gaia.htm.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

systems are dedicated to tasks under (1), fifteen to (2), and one model (ClimateGPT-2) has been developed for both task groups.

A large portion of the **data** used to enhance a model's memory with CC-relevant information stems from domain-relevant news, scientific publications, and publicly available reports published by institutions such as the Intergovernmental Panel on Climate Change (IPCC) and the World Meteorological Organization (WMO). It is noticeable that the developers of ClimateGPT, ClimateBERT, and ChatClimate have made a substantial effort to report on the data used in creating the LM or LM-based system, and in most instances provide comprehensive descriptions of the corpora used in domain-adaptive pretraining efforts.

A substantial portion of the presented LMs and LM-based systems have been developed for more narrowly defined **downstream classification tasks** concerning the analysis of sustainability, financial, and annual reports issued by companies. Meanwhile, models of the ClimateGPT family are expected to be able to answer questions from a broader span of topics; this might be the reason why models of this family have a wide range of professional profiles listed as their **intended audience.** There is a substantial overlap, implicit or explicit, in the target audiences for reviewed models, with decision- and policymakers being frequently described as professionals who would benefit from NLP-supported data processing. Researchers and anyone analysing corporate reports are also an important target group. An interesting outlier to this trend is farmers and farming consultants, the target audience of My Climate Advisor (Nguyen et al., 2024).

There also seems to be awareness of the importance of *human evaluation* when developing these systems, with many of the LMs and LM-based systems having undergone some form of human evaluation, either by experts giving scores on a model's output, or by experts analysing the model's performance and its implications within a broader task. However, manual evaluation does not automatically translate into comparability between models and systems. As mentioned in Section 5.3, the number of experts evaluating models' output can range between two and sixteen. The studies apply various evaluation approaches and scores. It has been noted by Nguyen et al. (2024) that inter-annotator agreement can also be a problem, and can occur even when the annotators had been involved in designing the evaluation study. For these reasons, reports on models' performance based on human evaluation should only be interpreted in the context in which the human evaluation took place. In the future, it would be helpful if information about the annotators' field of expertise and the inter-annotator agreement were consistently reported.

In terms of accessibility, 24 of the LMs can be downloaded either from the HFH or GitLab; 11 can (also) be accessed through a web interface, and only 3 are inaccessible at the time of writing. The number of all-time downloads enables comparison of the popularity of models within a model family. Although this number is dynamic, a clear trend of preferences can be noticed. For models of the ClimateGPT family, ClimateGPT-7B has by far the highest number of all-time downloads, possibly disclosing a preference for smaller models that might be more accessible to the research community. In the ClimateBERT family of models, the foundation model pretrained on all data, ClimateBERT$_F$, is clearly the most popular one, presumably because the wide range of data it has been exposed to allows for better adaptation to downstream tasks. In terms of paragraph classification, there is an obvious interest to detect whether a paragraph is climate-related or not, with *climatebert/distilroberta-base-climate-detector* topping the charts. Finally, for sentence-level annotations, the model trained on classifying sentences as environmental claims or not appears to be the most popular one.

The tables in Appendix A provide a chance to explore the timeline in which these LMs and LM-based models were published. All LMs and LM-based systems presented in this paper have been published between 2020 and 2025. Text classification was a heavily researched task between 2020 and 2023, while text generation tasks, including question-answering and summarization, have gained more prominence between 2022 and 2025.

7.2. Findings and future research

This section elaborates on important findings resulting from the survey and highlights possible research directions that could help to address some of the challenges encountered during the research. These

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

aspects are discussed across four topics: data transparency; evaluation and comparability; intended use and accessibility; and lifecycle, uptake, and carbon footprint.

Data transparency. When comparing the LMs described in Sections 4 and 5, inconsistencies were noticed in the way in which the size of datasets used for model pretraining, fine-tuning, and RAG-based augmentation is reported. For example, Thulke et al. (2024) follow current conventions and express data size in number of tokens; Webersinke et al. (2022) report the number of paragraphs and average number of tokens per paragraph and per data source, which allows for calculating an approximate corpus size; Xiang and Fujii (2023) express data size in records and paragraphs, but do not specify what *record* stands for; in Vaghefi et al. (2022), data size is expressed as the number of abstracts. In the future, it would be helpful if developers of climate-adapted LMs followed a uniform method of reporting data size, which would allow for more streamlined comparison of the size of corpora used for domain adaptation.

Large text corpora are the backbone of the LMs discussed in this paper. However, limited information is provided on the contents of corpora used for pretraining and continuous training, and the steps that have been taken to ensure the quality of the text collections. In most instances, not even a basic statistical description of the corpora is provided. While it is understandable that copyright and intellectual property considerations restrict the publication of corpora, there have been initiatives that propose ways of providing information about training data without granting access to or publishing it. One such example is the platform “What’s In My Big Data” (WIMBD), proposed by Elazar et al. (2023),[105] which offers a set of analysis steps that allow for descriptive corpus information on three high-level categories: data statistics,[106] data quality,[107] and community- and society-relevant measurements.[108] In the future, it would be helpful if the community adopted a method of providing more transparent data description, as long as that does not infringe upon intellectual property rights.

Some LMs and LM-based systems, such as ClimateGPT, ChatClimate, and ClimateQA, obtain data from and/or process documents in a PDF format. In some instances, such as ClimateGPT, the tool used to parse PDF documents is mentioned; however, this is not always revealed. The studies do not always touch upon the challenges of parsing PDF documents and the possibility of noise being introduced in the data. While PDF extraction tools have seen substantial improvement, it would be beneficial to reflect on the challenges of extracting data from PDF files and how these have been addressed.

Evaluation and comparability. Benchmarks are a popular way of *evaluating* progress in LM development and are “framed as foundational milestones on the path towards flexible and generalizable AI systems” (Raji et al., 2021, p.1). Although they undoubtedly offer many benefits, such as comparability and the possibility to measure incremental improvements, benchmarks have shortcomings, too, of which the most relevant one is that they are datasets offering a limited testing context for a task. They cannot comprehensively address all possible uses an LM-powered application might have in a real-world scenario. Another problematic aspect of benchmarks is data contamination, where models are evaluated on tasks that have been included in the training data. It was mentioned in Section 5.3 that the current state of human evaluation lacks standardized guidelines. The research community would benefit from human evaluation procedures that are transparent, standardized, and which offer clear guidelines on addressing the issue of subjectivity when scoring LMs’ output. Ongoing efforts to make the output of large language models reliable and trustworthy would be of interest to researchers designing tools with a real-world deployment scenario in mind.

Creating a platform where the performance of climate-related LLMs for text generation would be compared, or adding these models to an already existing platform, such as the LM Arena (Zheng et al., 2023),[109] allowing side-by-side comparison between domain-specific climate-related

[105] https://github.com/allenai/wimbd.

[106] Summary statistics and internet domain distribution.

[107] Most- and least-common n-grams, duplicates, document length distribution.

[108] Benchmark contamination and personally identifiable information.

[109] Previously known as *chatbot arena.* Available at https://lmarena.ai/.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

LMs or between a domain-specific LM and a general-purpose LM, would allow interested stakeholders a more hands-on context to test the models and a faster feedback loop for LM developers.

Intended use and accessibility. While generating answers to user-based questions is undoubtedly a helpful functionality that might reduce the time needed to conduct research, LMs and LM-based systems are far from flawless and can generate errors that could hamper rather than aid understanding, leading to poor research outcomes. Potential users should be educated about these limitations, especially in the case of systems hosted on websites and made available to audiences who might not be familiar with the inner workings of LMs or lack domain expertise to recognize errors in a model's output. Unaccounted-for errors could lead to perpetuated biases or to incorrect information echoing across multiple channels. Climate change is a high-stakes scenario, and claims for increased research efficiency should not be treated as viable gains if they come at the cost of accuracy.

Training and fine-tuning LLMs, as well as creating LLM-based tools, is a resource-intensive process (Strubell et al., 2020; Bender et al., 2021; Luccioni and Hernandez-Garcia, 2023; Oliver et al., 2024). To this end, it would be helpful if future development of technical solutions involving large language models were guided by surveys revealing the needs of the tools' target group.[110] Researchers in the future might consider (1) defining real-world needs in the target domain and (2) implementing more comprehensive evaluation of existing resources, rather than making the training of ever-larger LLMs the go-to solution for every challenge (Wiggers, 2024).

Almost all LMs and LM-based systems presented in this paper have been developed for and made accessible in the English language. The only deliberate exception to this is ClimateGPT models, which can be used in 21 other languages, a functionality that has been implemented with cascaded machine translation. Making LMs and LM-based systems accessible in languages other than English could potentially pave the way to a more inclusive LM and LM-based system development, although it is highly unlikely that another language could become as dominant in LM development as the English language in the near future. It also needs to be pointed out that machine translation systems in a specialized domain could generate unintended errors that could then propagate through the question-answering functionality, which is why any multilingual solution based on machine translation would benefit from comprehensive evaluation prior to deployment in a real-world setting.

Lifecycle, uptake, and carbon footprint. The six LLM-based systems presented in Section 5 make use of out-of-the-box, proprietary models. Using LLMs through an API significantly reduces development complexities and allows researchers who do not have access to sufficient compute power to deploy LLMs in their applications. However, there are many associated risks with using proprietary LLMs, some major ones being uncertainties regarding data protection and treatment of sensitive information, and the risk of the proprietary model becoming unavailable due to an outage or deprecation. For example, the LLM-based system built by Kraus et al. (2023) is using OpenAI's model *text-davinci-003*, which has since been deprecated. Future research and development should take the deprecation risk into account when planning a system's lifecycle and focus on systems of a modular built, where an LLM of a newer generation can be seamlessly evaluated and integrated.

It would also be beneficial if a more reliable method of reporting on a model's uptake by the community existed. Measuring the engagement with a model in all-time downloads should not be interpreted as a measurement of a model's popularity. At most, this nugget of information could be useful in gauging which model in a family of models attracts more attention, and if that could be interpreted as a reflection of the interest of practitioners utilizing LMs in the climate change context. Perhaps a better way of reporting *uptake* would be to calculate the average number of downloads per day for each model; however, this cannot be done with high accuracy using information that is publicly available at the moment of writing.

Finally, given the popularity of question-answering systems, future research would benefit from consistent reporting of the CO_2 emissions of both model development and inference. A good example

[110] This has been touched upon by Thulke et al. (2024), who in interviews invite experts to consider scenarios in which an LLM would be of use to them.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

of this, albeit not as comprehensive as the reporting proposed by Luccioni et al. (2023), is given by Thulke et al. (2024): both a *Model card* and a *Sustainability scorecard* for their models are provided, the latter of which contains information about average inference emissions per sample. Along these lines, it is encouraging to see that some practitioners opt for less computationally intensive models, such as DistilRoBERTa-base instead of BERT for ClimateBERT models (Webersinke et al., 2022) and the teacher-student design applied in Xiang and Fujii (2023). In the future, it would be beneficial to examine to what degree models such as the encoder-only TinyBERT (Jiao et al., 2020) and DistilBERT (Sanh et al., 2020), or the decoder-only models of Hugging Face's SmolLM collection[111] could be adapted for climate-change-relevant research.

Benefits for CC research. As this survey shows, the development of climate-change specific LMs and LM-based systems is a dynamic research field that could benefit a wide range of stakeholders involved in climate change research or in processing data for the purpose of assisting climate change research. The common goal of these systems is to improve access to climate-relevant knowledge, which should ultimately accelerate the adjustment of climate policies and adaptation efforts to a changing environment, as well as assist in the fast detection of emerging problematic areas that need to be addressed in mitigation efforts. If used responsibly and within their limitations, these systems could provide a promising starting point for expert-led research.

Acknowledgements. I would like to thank Dr. Sabine Bartsch, the Principal Investigator of the InsightsNet project, who encouraged the conceptualization and writing of the paper. I sincerely appreciate the insightful comments on the technical sections of the paper provided by Dr. Andreas Hamm of the German Aerospace Centre, and the equally meticulous review of the paper's structure, conducted by Dr. Katharina Herget of the Institute of Linguistics and Literary Studies at the Technical University of Darmstadt, both of which have significantly contributed to improving the quality of the manuscript. The valuable comments and constructive suggestions provided by the anonymous reviewers have also greatly improved the final manuscript. All content, interpretations, and any errors or omissions in the manuscript remain my responsibility.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.10024.

Author contribution. The survey paper has been written by a single author, who conceptualized the project, designed the methodology, conducted the analysis, and wrote the paper.

Competing interests. The author declares none.

Data availability statement. This is a survey paper, and no new datasets were generated or analysed in this study. The reviewed studies and some of the data supporting these studies are available in their respective repositories, as indicated in the references and in the manuscript. Information about the models reviewed in the paper is included in a GitHub repository: https://github.com/volkanovska/Language-models-for-climate-change-texts.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. The research presented in this paper was conducted within the research project InsightsNet (https://insightsnet.org/), funded by the Federal Ministry of Education and Research (BMBF) under grant no. 01UG2130A. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Bender EM, Gebru T, McMillan-Major A and Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. https://doi.org/10.1145/3442188.3445922.

Berrang-Ford L, Sietsma AJ, Callaghan M, Minx JC, Scheelbeek PF, Haddaway NR, Haines A and Dangour AD (2021) Systematic mapping of global research on climate and health: A machine learning review. *The Lancet Planetary Health 5*(8), e514–e525. https://doi.org/10.1016/S2542-5196(21)00179-0.

Bhattacharjee B, Trivedi A, Muraoka M, Ramasubramanian M, Udagawa T, Gurung I, Pantha N, Zhang R, Dandala B, Ramachandran R, Maskey M, Bugbee K, Little MM, Fancher E, Gerasimov I, Mehrabian A, Sanders L, Costes SV, Blanco-Cuaresma S, Lockhart K, Allen T, Grezes F, Ansdell M, Accomazzi A, El-Kurdi Y, Wertheimer D, Pfitzmann B, Berrospi Ramis C, Dolfi M, De Lima RT, Vagenas P, Mukkavilli SK, Staar PWJ, Vahidinia S, McGranaghan R and Lee

[111] https://huggingface.co/blog/smollm.

TJ (2024) INDUS: Effective and efficient language models for scientific applications. In Dernoncourt F, Preoţiuc-Pietro D and Shimorina (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Miami: Association for Computational Linguistics, pp. 98–112. https://doi.org/10.18653/v1/2024.emnlp-industry.9

Bingler J, **Kraus M**, **Leippold M and Webersinke N** (2022a) *How Cheap Talk in Climate Disclosures Relates to Climate Initiatives, Corporate Emissions, and Reputation Risk.* Swiss Finance Institute Research Paper No. 22-01. https://doi.org/10.2139/ssrn.4000708.

Bingler JA, **Kraus M**, **Leippold M and Webersinke N** (2022b) Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters 47*, 102776. https://doi.org/10.1016/j.frl.2022.102776.

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

Bisk Y, **Zellers R**, **Gao J**, **Choi Y**, et al. (2020) PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34), pp. 7432–7439. https://doi.org/10.1609/aaai.v34i05.6239.

Brown TB, **Mann B**, **Ryder N**, **Subbiah M**, **Kaplan J**, **Dhariwal P**, **Neelakantan A**, **Shyam P**, **Sastry G**, **Askell A**, **Agarwal S**, **Herbert-Voss A**, **Krueger G**, **Henighan T**, **Child R**, **Ramesh A**, **Ziegler DM**, **Wu J**, **Winter C**, **Hesse C**, **Chen M**, **Sigler E**, **Litwin M**, **Gray S**, **Chess B**, **Clark J**, **Berner C**, **McCandlish S**, **Radford A**, **Sutskever I and Amodei D** (2020) Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc. https://doi.org/10.5555/3495724.3495883.

Bulian J, **Schäfer MS**, **Amini A**, **Lam H**, **Ciaramita M**, **Gaiarin B**, **Huebscher MC**, **Buck C**, **Mede N**, **Leippold M**, et al. (2023) Assessing Large Language Models on Climate Information. *arXiv preprint.* https://doi.org/10.48550/arXiv.2310.02932.

Chen M, **Tworek J**, **Jun H**, **Yuan Q**, **de Oliveira Pinto HP**, **Kaplan J**, **Edwards H**, **Burda Y**, **Joseph N**, **Brockman G**, **Ray A**, **Puri R**, **Krueger G**, **Petrov M**, **Khlaaf H**, **Sastry G**, **Mishkin P**, **Chan B**, **Gray S**, **Ryder N**, **Pavlov M**, **Power A**, **Kaiser L**, **Bavarian M**, **Winter C**, **Tillet P**, **Such FP**, **Cummings D**, **Plappert M**, **Chantzis F**, **Barnes E**, **Herbert-Voss A**, **Guss WH**, **Nichol A**, **Paino A**, **Tezak N**, **Tang J**, **Babuschkin I**, **Balaji S**, **Jain S**, **Saunders W**, **Hesse C**, **Carr AN**, **Leike J**, **Achiam J**, **Misra V**, **Morikawa E**, **Radford A**, **Knight M**, **Brundage M**, **Murati M**, **Mayer K**, **Welinder P**, **McGrew B**, **Amodei D**, **McCandlish S**, **Sutskever I and Zaremba W** (2021) Evaluating Large Language Models Trained on Code. https://doi.org/10.48550/arXiv.2107.03374.

Coan TG, **Boussalis C**, **Cook J and Nanko MO** (2021) Computer-assisted classification of contrarian claims about climate change. *Scientific Reports 11*(1), 22320. https://doi.org/10.1038/s41598-021-01714-4.

Cohere (2023) *Introducing Coral, the Knowledge Assistant for Enterprises.* Available at: https://cohere.com/blog/introducing-coral-the-knowledge-assistant-for-enterprises (accessed 11th Aug 2025).

Colesanti Senni C, **Vaghefi S**, **Schimanski T**, **Manekar T and Leippold M** (2024) Using AI to assess the decision-usefulness of corporates' nature-related disclosures. *Swiss Finance Institute Research Paper* No. 24–90. https://doi.org/10.2139/ssrn.4860331.

Conover M, **Hayes M**, **Mathur A**, **Xie J**, **Wan J**, **Shah S**, **Ghodsi A**, **Wendell P**, **Zaharia M and Xin R** (2023) Free Dolly: Introducing the world's first truly open instruction-tuned LLM. *Company Blog of Databricks.* Available at: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm (accessed 10th Aug 2024).

Deng M, **Leippold M**, **Wagner AF and Wang Q** (2023) War and Policy: Investor Expectations on the Net-Zero Transition. *Swiss Finance Institute Research Paper* No. 22–29. https://doi.org/10.2139/ssrn.4080181.

Devlin J, **Chang M-W**, **Lee K and Toutanova K** (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein J, Doran C and Solorio T (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis: Association for Computational Linguistics, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Dickson ZP (2023) *BART for Climate Change Summarization.* Available at: https://huggingface.co/z-dickson/bart-large-cnn-climate-change-summarization (accessed 10th July 2024).

Dickson ZP and Hobolt SB (2024) Going against the grain: Climate change as a wedge issue for the radical right. *Comparative Political Studies 58*(8), 1733–1759. https://doi.org/10.1177/00104140241271297.

Diggelmann T, **Boyd-Graber J**, **Bulian, J.**, **Ciaramita, M.**, **and Leippold, M.** (2020). Climate-Fever: A Dataset for Verification of Real-World Climate Claims. *arXiv preprint.* https://doi.org/10.48550/arXiv.2012.00614.

Elazar Y, **Bhagia A**, **Magnusson I**, **Ravichander A**, **Schwenk D**, **Suhr A**, **Walsh P**, **Groeneveld D**, **Soldaini L**, **Singh S**, et al. (2023) What's in My Big Data? *arXiv preprint.* https://doi.org/10.48550/arXiv.2310.20707.

Engineering C (2023) *Artificial Intelligence vs. Machine Learning: What's the Difference?* Available at: https://ai.engineering.columbia.edu/ai-vs-machine-learning/ (accessed 09th Aug 2024).

Erasmus.AI (2024). *Erasmus.ai – Superhuman insight.* Available at: https://erasmus.ai/ (accessed 12th Aug 2024).

Garigliotti D (2024) SDG target detection in environmental reports using retrieval-augmented generation with LLMs. In Stammbach D, Ni J, Schimanski T, Dutia K, Singh A, Bingler J, Christiaen C, Kushwaha N, Muccione V, Vaghefi SA and Leippold M (eds.), *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Bangkok: Association for Computational Linguistics, pp. 241–250. https://doi.org/10.18653/v1/2024.climatenlp-1.19.

Garrido-Merchán EC, **González-Barthe C and Vaca MC** (2023) Fine-Tuning ClimateBert Transformer with ClimaText for the Disclosure Analysis of Climate-Related Financial Risks. *arXiv preprint.* https://doi.org/10.48550/arXiv.2303.13373.

Global Reporting Initiative (2024). *GRI Standards.* Available at: https://www.globalreporting.org/ (accessed 14th Aug 2024).

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Hadi MU, **Qureshi R**, **Shah A**, **Irfan M**, **Zafar A**, **Shaikh MB**, **Akhtar N**, **Wu J**, **Mirjalili S**, et al. (2023) A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv.* https://doi.org/10.36227/techrxiv.23589741.v1.

Havlik D and Pias MR (2024) Common errors in generative AI systems used for knowledge extraction in the climate action domain. *Open Research Europe 4*, 221. https://doi.org/10.12688/openreseurope.17258.1.

Hendrycks D, **Burns C**, **Basart S**, **Zou A**, **Mazeika M**, **Song D and Steinhardt J** (2020) Measuring Massive Multitask Language Understanding. *arXiv preprint.* https://doi.org/10.48550/arXiv.2009.03300.

HLEG, U. N. (2022) Integrity Matters: Net Zero Commitments by Businesses, Financial Institutions, Cities and Regions. *Report From the United Nations' High-Level Expert Group on the Net Zero Emissions Commitments of Non-State Entities. United Nations.* Available at: https://www.un.org/sites/un2.un.org/files/high-level_expert_group_n7b.pdf.

Hsu A, **Laney M**, **Zhang J**, **Manya D and Farczadi L** (2024) Evaluating ChatNetZero, an LLM-chatbot to demystify climate pledges. In Stammbach D, Ni J, Schimanski T, Dutia K, Singh A, Bingler J, Christiaen C, Kushwaha N, Muccione V, Vaghefi SA, and Leippold M (eds.), *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Bangkok: Association for Computational Linguistics, pp. 82–92. https://doi.org/10.18653/v1/2024.climatenlp-1.6.

Huang AH, **Wang H and Yang Y** (2023) FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research 40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832.

Ignat O, **Jin Z**, **Abzaliev A**, **Biester L**, **Castro S**, **Deng N**, **Gao X**, **Gunal AE**, **He J**, **Kazemi A**, **Khalifa M**, **Koh N**, **Lee A**, **Liu S**, **Min DJ**, **Mori S**, **Nwatu JC**, **Perez-Rosas V**, **Shen S**, **Wang Z**, **Wu W and Mihalcea R** (2024) Has it all been solved? Open NLP research questions not solved by large language models. In Calzolari N, Kan M-Y, Hoste V, Lenci A, Sakti S and Xue N (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino: ELRA and ICCL, pp. 8050–8094. Available at: https://aclanthology.org/2024.lrec-main.708.pdf.

Inie N, **Druga S**, **Zukerman P and Bender EM** (2024) From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* pp. 2322–2347. https://doi.org/10.1145/3630106.3659040.

Jalalzadeh Fard B, **Hasan S and Bell J** (2022) CliMedBERT: A Pre-trained Language Model for Climate and Health-related Text. *arXiv preprints.* https://doi.org/10.48550/arXiv.2212.00689.

Ji Z, **Lee N**, **Frieske R**, **Yu T**, **Su D**, **Xu Y**, **Ishii E**, **Bang YJ**, **Madotto A and Fung P** (2023) Survey of hallucination in natural language generation. *ACM Computing Surveys 55*(12). https://doi.org/10.1145/3571730.

Jiang AQ, **Sablayrolles A**, **Mensch A**, **Bamford C**, **Chaplot DS**, **Casas DDI**, **Bressand F**, **Lengyel G**, **Lample G**, **Saulnier L**, et al. (2023) Mistral 7B. *arXiv preprint.* https://doi.org/10.48550/arXiv.2310.06825.

Jiao X, **Yin Y**, **Shang L**, **Jiang X**, **Chen X**, **Li L**, **Wang F and Liu Q** (2020) TinyBERT: Distilling BERT for natural language understanding. In Cohn T, He Y and Liu Y (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020.* Online. Association for Computational Linguistics, pp. 4163–4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372.

Juhasz M., Dutia, K., Franks, H., Delahunty, C., Mills, P. F., and Pim, H. (2024). Responsible Retrieval Augmented Generation for Climate Decision Making from Documents. *arXiv preprint.* https://doi.org/10.48550/arXiv.2410.23902.

Kasneci E, **Seßler K**, **Küchemann S**, **Bannert M**, **Dementieva D**, **Fischer F**, **Gasser U**, **Groh G**, **Günnemann S**, **Hüllermeier E**, et al. (2023) ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kaur P, **Kashyap GS**, **Kumar A**, **Nafis MT**, **Kumar S and Shokeen V** (2024) From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility. *arXiv preprint.* https://doi.org/10.48550/arXiv.2402.16142.

Köpf A, **Kilcher Y**, **von Rütte D**, **Anagnostidis S**, **Tam Z-R**, **Stevens K**, **Barhoum A**, **Duc NM**, **Stanley O**, **Nagyfi R**, **ES S**, **Suri S**, **Glushkov D**, **Dantuluri A**, **Maguire A**, **Schuhmann C**, **Nguyen H, and Mattick A** (2023) OpenAssistant conversations – Democratizing large language model alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA*. Curran Associates Inc.. https://dl.acm.org/doi/10.5555/3666122.3668186.

Korte JW, **Bartsch S**, **Beckmann R**, **El Baff R**, **Hamm A and Hecking T** (2023) From causes to consequences, from chat to crisis. The different climate changes of science and Wikipedia. *Environmental Science & Policy 148*, 103553. https://doi.org/10.1016/j.envsci.2023.103553.

Kraus M, **Bingler JA**, **Leippold M**, **Schimanski T**, **Senni CC**, **Stammbach D**, **Vaghefi SA and Webersinke N** (2023) Enhancing large language models with climate resources. *Swiss Finance Institute Research Paper No. 23–99*. https://doi.org/10.2139/ssrn.4407205.

Kurfali M, **Zahra S**, **Nivre J and Messori G** (2025) ClimateEval: A comprehensive benchmark for NLP tasks related to climate change. In Dutia K, Henderson P, Leippold M, Manning C, Morio G, Muccione V, Ni J, Schimanski T, Stammbach D, Singh A, Su AR and Vaghefi SA (eds), *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 194–207. https://doi.org/10.18653/v1/2025.climatenlp-1.13

Lang J, **Hyslop C**, **Yeo ZY**, **Black R**, **Hale T**, **Chalkley P**, **Hans F**, **Hay N**, **Höhne N**, **Hsu A**, **Kuramochi T**, **Mooldijk S and Smith S** (2023) Net Zero Tracker. Available at: https://zerotracker.net/ (accessed 10th Aug 2024).

Leippold M, **Vaghefi SA**, **Stammbach D**, **Muccione V**, **Bingler J**, **Ni J**, **Senni CC**, **Wekhof T**, **Schimanski T**, **Gostlow G**, et al. (2025) Automated fact-checking of climate claims with large language models. *npj Climate Action 4*(1), 17. https://doi.org/10.1038/s44168-025-00215-8.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Lewis M, **Liu Y**, **Goyal N**, **Ghazvininejad M**, **Mohamed A**, **Levy O**, **Stoyanov V and Zettlemoyer L** (2020) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky D, Chai J, Schluter N and Tetreault J (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703.

Liu Y, **Ott M**, **Goyal N**, **Du J**, **Joshi M**, **Chen D**, **Levy O**, **Lewis M**, **Zettlemoyer L and Stoyanov V** (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint.* https://doi.org/10.48550/arXiv.1907.11692.

Lu L (2024) In-depth analysis of artificial intelligence for climate change mitigation. *Preprint 2024*, 2024020022. https://doi.org/10.20944/preprints202402.0022.v1.

Luccioni, A. S. and Hernandez-Garcia, A. (2023). Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning. *arXiv preprint.* doi: https://doi.org/10.48550/arXiv.2302.08476.

Luccioni A, **Baylor E and Duchene N** (2020) Analyzing Sustainability Reports Using Natural Language Processing. *arXiv preprint.* https://doi.org/10.48550/arXiv.2011.08073.

Luccioni, A. S., **Viguier, S.**, **and Ligozat, A.-L.** (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, *24*(253):1–15. Available at: https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf.

Mehra S, **Louka R and Zhang Y** (2022) ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices. *arXiv preprint.* https://doi.org/10.48550/arXiv.2203.16788.

Mihaylov T, **Clark P**, **Khot T and Sabharwal A** (2018) Can a suit of armor conduct electricity? A new dataset for open book question answering. In Riloff E, Chiang D, Hockenmaier J and Tsujii J (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, pp. 2381–2391. https://doi.org/10.18653/v1/D18-1260

Minaee S, **Mikolov T**, **Nikzad N**, **Chenaghlu M**, **Socher R**, **Amatriain X and Gao J** (2024) Large Language Models: A Survey. *arXiv preprint.* https://doi.org/10.48550/arXiv.2402.06196.

Moro G, **Ragazzi L and Valgimigli L** (2023) Carburacy: Summarization Models Tuning and Comparison in Eco-Sustainable Regimes with a Novel Carbon-Aware Accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, pp. 14417–14425. https://doi.org/10.1609/aaai.v37i12.26686.

Nallapati R, **Zhou B**, **dos Santos C**, **Gulçehre Ç and Xiang B** (2016) Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Riezler S and Goldberg Y (eds), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, pp. 280–290. https://doi.org/10.18653/v1/K16-1028

Net Zero Tracker (2022). Net Zero Stocktake 2022: Assessing the Status and Trends of Net Zero Target Setting Across Countries, Sub-National Governments and Companies. Technical Report, Net Zero Tracker. Available at: https://newclimate.org/resources/publications/net-zero-stocktake-2022.

Net Zero Tracker (2023) Net Zero Stocktake 2023: Assessing the Status and Trends of Net Zero Target Setting across Countries, Sub-National Governments and Companies. Technical Report, Net Zero Tracker. Available at: https://newclimate.org/resources/publications/net-zero-stocktake-2023.

New Climate Institute (2022) Corporate Climate Responsibility Monitor 2022: Assessing the Transparency and Integrity of Companies' Emission Reduction and Net-Zero Targets. Technical Report, New Climate Institute. Available at: https://newclimate.org/resources/publications/corporate-climate-responsibility-monitor-2022.

New Climate Institute (2023) *Corporate Climate Responsibility Monitor 2023: Assessing the Transparency and Integrity of Companies' Emission Reduction and Net-Xero Targets.* Technical Report, New Climate Institute. Available at: https://newclimate.org/resources/publications/corporate-climate-responsibility-monitor-2023.

Nguyen V, **Karimi S**, **Hallgren W**, **Harkin A and Prakash M** (2024) My Climate Advisor: An application of NLP in climate adaptation for agriculture. In Stammbach D, Ni J, Schimanski T, Dutia K, Singh A, Bingler J, Christiaen C, Kushwaha N, Muccione V, Vaghefi SA and Leippold M (eds.), *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Bangkok: Association for Computational Linguistics, pp. 27–45. https://doi.org/10.18653/v1/2024.climatenlp-1.3

Ni J, **Bingler J**, **Colesanti-Senni C**, **Kraus M**, **Gostlow G**, **Schimanski T**, **Stammbach D**, **Ashraf Vaghefi S**, **Wang Q**, **Webersinke N**, **Wekhof T**, **Yu T and Leippold M** (2023) CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In Feng Y and Lefever E (eds.) *Proceedings of the 2023 conference on empirical methods in natural language processing: System demonstrations*. Singapore: Association for Computational Linguistics, pp. 21–51. https://doi.org/10.18653/v1/2023.emnlp-demo.3

Oliver RY, **Chapman M**, **Emery N**, **Gillespie L**, **Gownaris N**, **Leiker S**, **Nisi AC**, **Ayers D**, **Breckheimer I**, **Blondin H**, et al. (2024) Opening a conversation on responsible environmental data science in the age of large language models. *Environmental Data Science 3*:e14. https://doi.org/10.1017/eds.2024.12.

OpenAI (2022) *Introducing ChatGPT.* Available at: https://openai.com/index/chatgpt/ (accessed 11th Aug 2025).

Pirozelli P, **José MM**, **Silveira I**, **Nakasato F**, **Peres SM**, **Brandão AA**, **Costa AH and Cozman FG** (2024) Benchmarks for Pirá 2.0, a reading comprehension dataset about the ocean, the Brazilian coast, and climate change. *Data Intelligence 6*(1), 29–63. https://doi.org/10.1162/dint_a_00245.

Priem J, **Piwowar H and Orr R** (2022) *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts.* https://doi.org/10.48550/arXiv.2205.01833.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Radford A, **Wu J**, **Child R**, **Luan D**, **Amodei D**, **Sutskever I**, et al. (2019) Language models are unsupervised multitask learners. *Technical Report, OpenAI 1*(8), 9. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Raji D, **Denton E**, **Bender EM**, **Hanna A and Paullada A** (2021) AI and the everything in the whole wide world benchmark. In Vanschoren J and Yeung S (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. *1*. Available at: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Raschka S (2024) *Build a Large Language Model (from Scratch)*. Simon and Schuster

Rolnick D, **Donti PL**, **Kaack LH**, **Kochanski K**, **Lacoste A**, **Sankaran K**, **Ross AS**, **Milojevic-Dupont N**, **Jaques N**, **Waldman-Brown A**, et al. (2022) Tackling climate change with machine learning. *ACM Computing Surveys (CSUR) 55*(2), 1–96. https://doi.org/10.1145/3485128.

Saha D, **Sinha M and Dasgupta T** (2024) EnClaim: A style augmented transformer architecture for environmental claim detection. In Stammbach D, Ni J, Schimanski T, Dutia K, Singh A, Bingler J, Christiaen C, Kushwaha N, Muccione VA Vaghefi S and Leippold M (eds.), *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Bangkok: Association for Computational Linguistics, pp.123–132. https://doi.org/10.18653/v1/2024.climatenlp-1.9.

Sakaguchi K, **Bras RL**, **Bhagavatula C and Choi Y** (2021) WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM 64*(9), 99–106. https://doi.org/10.1145/3474381.

Sanh V, **Debut L**, **Chaumond J and Wolf T** (2020) DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint.* https://doi.org/10.48550/arXiv.1910.01108.

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

Schimanski T, **Colesanti Senni C**, **Gostlow G**, **Ni J**, **Yu T and Leippold M** (2023b). Exploring nature: Datasets and models for analyzing nature-related disclosures. *Swiss Finance Institute Research Paper* No. 24–95. Available at SSRN: https://ssrn.com/abstract=4665715.

Schimanski T, **Ni J**, **Martín RS**, **Ranger N and Leippold M** (2024a) ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In Al-Onaizan Y, Bansal M and Chen Y-N (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami: Association for Computational Linguistics, pp. 17509–17524. https://doi.org/10.18653/v1/2024.emnlp-main.969

Schimanski T, **Reding A**, **Reding N**, **Bingler J**, **Kraus M and Leippold M** (2024b) Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters 61*, 104979. https://doi.org/10.2139/ssrn.4622514.

Shannon CE (1949) *The Mathematical Theory of Communication*. Urbana: University of Illinois Press

Spokoyny D, **Laud T**, **Corringham T and Berg-Kirkpatrick T** (2023) Towards Answering Climate Questionnaires from Unstructured Climate Reports. *arXiv preprint.* https://doi.org/10.48550/arXiv.2301.04253.

Stammbach D, **Webersinke N**, **Bingler JA**, **Kraus M and Leippold M** (2022) A Dataset for Detecting Real-World Environmental Claims. *arXiv preprint.* https://doi.org/10.48550/arXiv.2209.00507.

Strubell E, **Ganesh A and McCallum A** (2020) Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 13693–13696. https://doi.org/10.1609/aaai.v34i09.7123.

Sun Z (2023) A Short Survey of Viewing Large Language Models in Legal Aspect. *arXiv preprint.* https://doi.org/10.48550/arXiv.2303.09136.

Team G, **Anil R**, **Borgeaud S**, **Alayrac J-B**, **Yu J**, **Soricut R**, **Schalkwyk J**, **Dai AM**, **Hauth A**, **Millican K**, et al. (2023) Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint.* https://doi.org/10.48550/arXiv.2312.11805.

Thulke D, **Gao Y**, **Pelser P**, **Brune R**, **Jalota R**, **Fok F**, **Ramos M**, **van Wyk I**, **Nasir A**, **Goldstein H**, et al. (2024) ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change. *arXiv preprint.* https://doi.org/10.48550/arXiv.2401.09646.

Toetzke M, **Probst B and Feuerriegel S** (2023) Leveraging large language models to monitor climate technology innovation. *Environmental Research Letters 18*(9), 091004. https://doi.org/10.1088/1748-9326/acf233.

Touvron H, **Martin L**, **Stone K**, **Albert P**, **Almahairi A**, **Babaei Y**, **Bashlykov N**, **Batra S**, **Bhargava P**, **Bhosale S**, et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint.* https://doi.org/10.48550/arXiv.2307.09288.

Vaghefi S, **Muccione V**, **Huggel C**, **Khashehchi H and Leippold M** (2022) Deep climate change: A dataset and adaptive domain pre-trained language models for climate change related tasks. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*. Available at: https://www.climatechange.ai/papers/neurips2022/27.

Vaghefi SA, **Stammbach D**, **Muccione V**, **Bingler J**, **Ni J**, **Kraus M**, **Allen S**, **Colesanti-Senni C**, **Wekhof T**, **Schimanski T**, et al. (2023) ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment 4*(1), 480. https://doi.org/10.1038/s43247-023-01084-x.

Vaid R, **Pant K and Shrivastava M** (2022) Towards fine-grained classification of climate change related social media text. In Louvan S, Madotto A and Madureira B (eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Dublin: Association for Computational Linguistics, pp. 434–443. https://doi.org/10.18653/v1/2022.acl-srw.35

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Varini FS, **Boyd-Graber J**, **Ciaramita M and Leippold M** (2020) ClimaText: A dataset for climate change topic detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*. https://doi.org/10.5167/uzh-191999.

Vaswani A, **Shazeer N**, **Parmar N**, **Uszkoreit J**, **Jones L**, **Gomez AN**, **Kaiser LU and Polosukhin I** (2017 Attention is all you need. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R, editors, *Advances in Neural Information Processing Systems*, Volume *30*. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang Y, **Ivison H**, **Dasigi P**, **Hessel J**, **Khot T**, **Chandu KR**, **Wadden D**, **MacMillan K**, **Smith NA**, **Beltagy I and Hajishirzi H** (2023) How far can camels go? Exploring the state of instruction tuning on open resources. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA*. Curran Associates Inc. https://doi.org/10.5555/3666122.3669390.

Wang Y, **Li H**, **Han X**, **Nakov P and Baldwin T** (2024) Do-not-answer: Evaluating safeguards in LLMs. In Graham Y, Purver M (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics: St. Julian's, Malta, pp. 896–911. Available at: https://aclanthology.org/2024.findings-eacl.61/.

Webersinke N, **Kraus M**, **Bingler J and Leippold M** (2022) ClimateBERT: A pretrained language model for climate-related text. *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*. https://doi.org/10.48550/arXiv.2110.12010.

Wehnert, S. (2023). Justifiable Artificial Intelligence: Engineering Large Language Models for Legal Applications. *arXiv preprint.* https://doi.org/10.48550/arXiv.2311.15716.

Wiggers K (2024) Women in AI: Anna Korhonen Studies the Intersection Between Linguistics and AI. Available at: https://techcrunch.com/2024/04/21/women-in-ai-anna-korhonen-studies-the-intersection-between-linguistics-and-ai/. (accessed 05th July 2024).

Xiang K and Fujii A (2023) DARE: Distill and reinforce ensemble neural networks for climate-domain processing. *Entropy 25*(4), 643. https://doi.org/10.3390/e25040643.

Yang R, **Tan TF**, **Lu W**, **Thirunavukarasu AJ**, **Ting DSW and Liu N** (2023) Large language models in health care: Development, applications, and challenges. *Health Care Science 2*(4), 255–263. https://doi.org/10.1002/hcs2.61.

Yao S, **Zhao J**, **Yu D**, **Du N**, **Shafran I**, **Narasimhan K and Cao Y** (2022) ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint.* https://doi.org/10.48550/arXiv.2210.03629.

Zellers R, **Holtzman A**, **Bisk Y**, **Farhadi A and Choi Y** (2019) HellaSwag: Can a machine really finish your sentence? In Korhonen A, Traum D and Màrquez L (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 4791–4800. https://doi.org/10.18653/v1/P19-1472

Zhang Y, **Zheng J**, **Jiang Y**, **Huang G and Chen R** (2019) A text sentiment classification modeling method based on coordinated CNN-LSTM-attention model. *Chinese Journal of Electronics 28*(1), 120–126. https://doi.org/10.1049/cje.2018.11.004.

Zhao WX, **Zhou K**, **Li J**, **Tang T**, **Wang X**, **Hou Y**, **Min Y**, **Zhang B**, **Zhang J**, **Dong Z**, et al. (2023) A Survey of Large Language Models. *arXiv preprint.* https://doi.org/10.48550/arXiv.2303.18223.

Zheng L, **Chiang W-L**, **Sheng Y**, **Zhuang S**, **Wu Z**, **Zhuang Y**, **Lin Z**, **Li Z**, **Li D**, **Xing EP**, **Zhang H**, **Gonzalez JE and Stoica I** (2023) Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA*. Curran Associates Inc. https://doi.org/10.5555/3666122.3668142.

A. Appendix: summary tables for survey findings

The Appendix contains Tables A1, A2, and A3, which summarize key information from the survey about the LMs and LM-based systems regarding the task, intended audience, method of evaluation, the domain of CC data, whether an LM or an LM-based system is publicly available or not, and the year of publication. The tables provide an additional view on the LMs and LM-based systems described in the paper: the two families of domain-specific language models, ClimateGPT and ClimateBERT, are presented in Table A1; the single domain-specific LMs are presented in Table A2, and the systems for climate change analysis built with generic LMs and climate-relevant resources are provided in Table A3.

The column "Task" is coloured *blue* for models that perform text-generation tasks, such as question-answering and text summarization, and *orange* for models that perform text classification tasks. The color *yellow* is used for models that have been envisaged to perform both tasks (text generation and text classification).

The column "Year of publication" entails information about the year an LM or an LM-based system is published. For LMs hosted on the HFH, the year of publication has been obtained by looking at the repository's history. If a model has two publication years, it means that it has either been updated in a different year or that the paper corresponding to the model has a different year of publication. The latter is true in cases where a paper is published as a preprint before its publication in a journal or in conference proceedings.

Cite this article: Volkanovska E (2025). Language models for the analysis of and interaction with climate change documents. *Environmental Data Science*, 4: e51. doi:10.1017/eds.2025.10024

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table A1. *Summary table for families of language models*

Fam	Name of model / HFH model identifier	Task	Audience	Evaluation		CC data domain	Model publicly available	Year of publication
				Automatic	Manual[a]			
ClimateGPT	ClimateGPT-FSG-7B (Thulke et al., 2024)	Question-answering	Decision-makers, scientists, policymakers, journalists involved in climate discussions	Yes	No	Climate, humanitarian issues, science (PT[b]); QA pairs from earth science, sustainability, economics, expert-generated instruction, and completion pairs (IFT[c])	Yes	2023–2024
	ClimateGPT-FSC-7B (Thulke et al., 2024)			Yes	Yes	Climate, humanitarian issues, science (PT); extreme weather events, climate change reports, technical reports on breakthroughs, academic CC research (PT); QA pairs from earth science, sustainability, economics, expert-generated instruction and completion pairs (IFT)		
	ClimateGPT-70B (Thulke et al., 2024)			Yes	Yes	Extreme weather events, climate change reports, technical reports on breakthroughs, academic CC research (CPT[d]); QA pairs from earth science, sustainability, economics, expert-generated instruction and completion pairs (IFT)		
	ClimateGPT-13B (Thulke et al., 2024)			Yes	No			
	ClimateGPT-7B (Thulke et al., 2024)			Yes	Yes			
ClimateBERT	ClimateBERT$_F$ (Webersinke et al., 2022)	Text classification	Researchers and stakeholders interested in using NLP models for processing climate-change texts and researchers and stakeholders analyzing corporate annual or sustainability reports	Yes	No	Paragraphs from corporate sustainability reports, abstracts of scientific papers, and news related to climate change, climate actions, climate politics, flood, droughts (complete dataset)	Yes	2021, 2022
	ClimateBERT$_S$ (Webersinke et al., 2022)			Yes	No	70% of the complete dataset most similar to task-related data		2021
	ClimateBERT$_D$ (Webersinke et al., 2022)			Yes	No	70% of the complete dataset least similar to task-related data		2021
	ClimateBERT$_{D+S}$ (Webersinke et al., 2022)			Yes	No	70% of complete dataset with highest similarity - diversity composite score		2021
	climatebert/distilroberta-base-climate-detector (Webersinke et al., 2022; Bingler et al., 2024)	Text classification (paragraph)		Yes	Yes[e]	Paragraphs from corporate reports (annual or sustainability)	Yes	2021
	climatebert/distilroberta- base-climate-sentiment (Webersinke et al., 2022; Bingler et al., 2024)			Yes	Yes[e]	Paragraphs from corporate reports (annual or sustainability)	Yes	2021
	climatebert/distilroberta-base-climate-specificity (Bingler et al., 2024)			Yes	Yes[e]	Paragraphs from corporate reports (annual or sustainability)	Yes	2021

Continued

Table A1. *Continued*

Fam	Name of model / HFH model identifier	Task	Audience	Evaluation		CC data domain	Model publicly available	Year of publication
				Automatic	Manual[a]			
	climatebert/distilroberta-base-climate- commitment (Bingler et al., 2024)	Text classification (sentence)		Yes	Yes[e]	Paragraphs from corporate reports (annual or sustainability)	Yes	2021
	climatebert/distilroberta-base-climate-tcfd (Bingler et al., 2022b, 2024)			Yes	Yes[e]	Paragraphs from corporate reports (annual or sustainability)	Yes	2021
	climatebert/environmental-claims (Stammbach et al., .2022)			Yes	Yes[e]	Sentences from sustainability reports, earnings calls, and annual reports	Yes	2022
	climatebert/transtion-physical (Deng et al., 2023)			Yes	Not clear	Sentences from earnings conference call transcripts	Yes	2023
	climatebert/renewable (Deng et al., 2023)			Yes	Not clear	Sentences from earnings conference call transcripts marked as related to transition	Yes	2023
	climatebert/netzero-reduction (Schimanski et al., 2023a)		Public and private actors assessing sustainability commitments	Yes	Yes	Sentences from Net Zero Tracker project and from previous climate-related projects	Yes	2023

[a] Also entails *manual error analysis.*

[b] Pretraining (data).

[c] Instruction fine-tuning (data).

[d] Continued pretraining (data).

[e] Manual evaluation conducted as part of another evaluation task.

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

https://doi.org/10.1017/eds.2025.10024 Published online by Cambridge University Press

Table A2. *Summary table for single domain-specific LMs*

Name of model / HFH model identifier	Task	Audience	Evaluation		CC data domain	Model publicly available	Year of publication
			Automatic	Manual[a]			
DARE (Xiang and Fujii, 2023)	Text classification	Stakeholders using NLP to study the dynamics of ambiguous climate information	Yes	No	Scientific literature related to climate change and health	No	2023
ClimateGPT-2 (Vaghefi et al., 2022)	Text generation and text classification	Policymakers, researchers, climate activists	Yes	No	Abstracts from papers on climate change	Yes	2022
ClimateQA (Luccioni et al., 2020)	Text classification	Analysts trying to detect climate- change-related risks and liabilities in financial reports	Yes	Yes	Financial and sustainability reports	No	2020
BART-based model for summarizing climate-related political press releases (Dickson, 2023; Dickson and Hobolt, 2024)	Summarization	Not specified	No	Yes	Summaries of press releases of political parties	Yes	2023

[a] Also entails *manual error analysis.*

Table A3. Summary table for systems using generic LMs

Name of LM-based system using a generic language model	Task	Audience	Evaluation		CC data domain (external data)	Model publicly available	Year of publication
			Automatic	Manual[a]			
ChatClimate (Vaghefi et al., 2023)	Question-answering	Decision-makers	No	Yes	IPCC reports and prompt-guided retrieval	Yes	2023
LLM CC agent: a protype (Kraus et al., 2023)	Question-answering	Organizations, institutions, companies	No	Yes	Climate Watch tables and Google search	No	2023
My Climate Advisor (Nguyen et al., 2024)	Question-answering	Farmers, farm advisors	No	Yes	Scientific papers on agriculture, food and environmental science, climate adaptation, region-specific climate risk information, books and industry reports	No	2024
Climate Policy Radar: Responsible question-answering (Juhasz et al., 2024)	Question-answering	Policymakers, analysts, academics	Yes	Yes[b]	550 documents from Climate Policy Radar's database of national laws and policies, UNFCCC submissions, documents from IEA, IAEA, OSCE, WMO	Yes[c]	2024
ChatNetZero (Hsu et al., 2024)	Question-answering	General public	No	Yes	Data from the United Nations High-Level Experts Group, Net Zero Tracker database, New Climate Institute	Yes	2024
CHATREPORT (Ni et al., 2023)	Question-answering and scoring	Policymakers, investors, general public	No	Yes	Relies only on the report that is the target of analysis	Yes	2023

[a] Also entails *manual error analysis.*

[b] Large-scale effort to annotate data to be used in LLM-as-a-judge scenario.

[c] CPR have a web-interface for CC QA.