Towards Climate Awareness in NLP Research

Daniel Hershcovich
Department of Computer Science
University of Copenhagen
dh@di.ku.dk

Nicolas Webersinke   Mathias Kraus
FAU Erlangen-Nuremberg
{nicolas.webersinke,
mathias.kraus}@fau.de

Julia Anna Bingler
ETH Zurich
binglerj@ethz.ch

Markus Leippold
University of Zurich
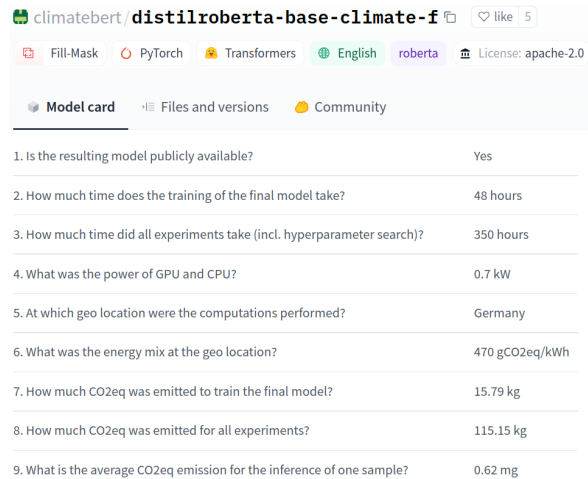markus.leippold@bf.uzh.ch

Abstract

The climate impact of AI, and NLP research in particular, has become a serious issue given the enormous amount of energy that is increasingly being used for training and running computational models. Consequently, increasing focus is placed on efficient NLP. However, this important initiative lacks simple guidelines that would allow for systematic climate reporting of NLP research. We argue that this deficiency is one of the reasons why very few publications in NLP report key figures that would allow a more thorough examination of environmental impact, and present a quantitative survey to demonstrate this. As a remedy, we propose a climate performance model card with the primary purpose of being practically usable with only limited information about experiments and the underlying computer hardware. We describe why this step is essential to increase awareness about the environmental impact of NLP research and, thereby, paving the way for more thorough discussions.[1]

1 Introduction

As Artificial Intelligence (AI), and specifically Natural Language Processing (NLP), scale up to require more computational resources and thereby more energy, there is an increasing focus on efficiency and sustainability (Strubell et al., 2019; Schwartz et al., 2020). For example, training a single BERT base model (Devlin et al., 2019) requires as much energy as a trans-American flight (Strubell et al., 2019). While newer models are arguably more efficient (Fedus et al., 2021; Borgeaud et al., 2022; So et al., 2022), they are also an order of magnitude larger, raising environmental concerns (Bender et al., 2021). The problem will only worsen with time, as compute requirements double every 10 months (Sevilla et al., 2022).

Figure 1: Example usage of our proposed climate performance model card (§5), on the Hugging Face Hub.[2]

This problem has been recognized by the NLP community, and a group of NLP researchers has recently proposed a policy document[3] of recommendations for efficient NLP, aiming to minimize the greenhouse gas (GHG) emissions[4] resulting from experiments done as part of the research. This proposal is part of a research stream aiming towards *Green NLP* and *Green AI* (Schwartz et al., 2020), which refers to "AI research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent."

While the branding of NLP and AI research as *green* has raised some awareness of the environmental impact, the large majority of NLP researchers are still not aware of their environmental impact resulting from training and running of large computational models. This also explains why a

---

[1] We provide a Jupyter notebook with the code used to conduct our survey, as well as model card templates in LaTeX and Markdown, at https://github.com/danielhers/climate-awareness-nlp.

[2] https://huggingface.co/climatebert/distilroberta-base-climate-f
[3] https://www.aclweb.org/portal/content/efficient-nlp-policy-document
[4] GHG, CO_2, and carbon are used interchangeably in this paper. CO_2eq (or CO_2e), i.e., carbon dioxide equivalent translates GHG other than CO_2 into CO_2 equivalents based on the global warming potential (Brander and Davis, 2012).

research stream in a similar direction (see §2.1), in which software tools are proposed to measure carbon footprint while training models (Lacoste et al., 2019; Henderson et al., 2020; Anthony et al., 2020; Lottick et al., 2019), have not been adopted by the community to a large extent (see §3). However, we claim that climate awareness is essential enough to be promoted in mainstream NLP (rather than only as a niche field) and that positive impact must be an inherent part of the discussion (Rolnick et al., 2019; Stede and Patz, 2021). Ideally environmental impact should always be taken into consideration, when deciding on which experiments to carry out.

We aim to simplify climate performance reporting in NLP while at the same time increasing awareness to its intricacies. Our contributions are:

- We conduct a survey of environmental impact statements in NLP literature published in the past six years (§3). This survey is conducted across five dimensions that directly influence the environmental impact.

- We delineate the different notions of "efficiency" common in the literature, proposing a taxonomy to facilitate transparent reporting, and identify ten simple dimensions across which researchers can describe the environmental impact resulted by their research (§4).

- We propose a climate performance model card (§5) with the main purpose of being practically usable with only limited information about experiments and the underlying computer hardware (see Figure 1).

2 Background

2.1 Automating Reporting

Several tools automate measurement and reporting of energy usage and emissions in ML. Lacoste et al. (2019) introduced a simple online calculator[5] to estimate the amount of carbon emissions produced by training ML models. It can estimate the carbon footprint of GPU compute by manually specifying hardware type, hours used, cloud provider, and region. Henderson et al. (2020) presented a Python package[6] for consistent, easy, and more accurate reporting of energy, compute, and carbon impacts

of ML systems by estimating them and generating standardized "Carbon Impact Statements." Anthony et al. (2020) proposed a Python package[7] that also has predictive capabilities, and allows proactive and intervention-driven reduction of carbon emissions. Model training can be stopped, at the user's discretion, if the predicted environmental cost is exceeded. Schmidt et al. (2022) actively maintain a Python package[8] that, besides estimating impact and generating reports, shows developers how they can lessen emissions by optimizing their code or by using cloud infrastructure in geographical regions with renewable energy sources. Bannour et al. (2021) surveyed and evaluated these tools and others for an NLP task, finding substantial variation in the reported measures due to different assumptions they make. In summary, automated tools facilitate reporting, but they do not substitute awareness and should not be trusted blindly.

2.2 Greenwashing

While branding NLP and AI research as *green* increases awareness of the environmental impact, there is a risk that the current framing, which exclusively addresses efficiency, will be perceived as the solution to the problem. Of course, we attribute benevolent motives to the authors of the proposed policy document. Nevertheless, we would like to avoid a situation analogous to a common phenomenon in the financial field, where companies brand themselves as *green* or *sustainable* for branding or financial reasons, without implementing proportional measures in practice to mitigate the negative impact on the environment (Delmas and Burbano, 2011). This malpractice is analogous to *greenwashing*. While this is a general term, one aspect of greenwashing is "a claim suggesting that a product is green based on a narrow set of attributes without attention to other important environmental issues" (TerraChoice, 2010).[9] Our motivation is in line with the EU Commission's initiative to "require companies to substanti-

---

[5]https://mlco2.github.io/impact/
[6]https://github.com/Breakend/experiment-impact-tracker

[7]https://github.com/lfwa/carbontracker
[8]https://codecarbon.io

[9]Paper, for example, is not necessarily environmentally preferable just because it comes from a sustainably harvested forest. Other important environmental issues in the paper-making process, such as greenhouse gas emissions or chlorine use in bleaching, may be equally important. Other examples are energy, utilities, and gasoline corporations that advertise about the benefits of new sources of energy while some are drilling into unexplored areas to source oil and thus destroying natural habitats and losing biodiversity, disguising the imbued hidden impacts (de Freitas Netto et al., 2020).

ate claims they make about the environmental footprint of their products/services by using standard methods for quantifying them."[10] While Schwartz et al. (2020) certainly do not argue that efficiency is *sufficient* for sustainability, this notion, which is potentially implied by the *green* branding, is misleading and even harmful: regardless of the extent of reduction, resources are still consumed, and GHGs are still emitted, among other negative effects. The efficiency mindset aims, at best, to prolong the duration of this situation. However, scaling up the performance of AI to satisfy the increasing demands from consumers risks ignoring the externalities incurred. Concepts such as reciprocity with the environment, which are central in some indigenous worldviews (Kimmerer, 2013), are absent from the discourse.

2.3 Carbon Offsetting

A common perception is that *carbon neutrality* can be achieved by compensating for emissions by financial contributions, a practice referred to as *carbon offsets*. This approach is problematic and controversial: the level of carbon prices required to achieve climate goals is highly debated (Hyams and Fawcett, 2013). The Intergovernmental Panel on Climate Change (IPCC) and various international organizations like the International Energy Agency (IEA) clearly state that mitigation activities are essential. Compensation activities will be necessary for hard-to-abate-sectors, once all other technological solutions have been implemented, and where mitigation is not (yet) feasible.[11] Moreover, economic dynamic efficiency requires investments in decarbonization technologies to keep the climate targets within reach. Compensation activities, especially in the afforestation area, delay the needed investments. This delay might exacerbate the likelihood of crossing climate tipping points and/or yields to a disorderly transition to a decarbonized economy (European Systemic Risk Board, 2016).

3 Survey of Climate Discussion in NLP

The issue of environmental impact is more general and not limited to NLP, but relevant to the entire field of AI: Schwartz et al. (2020) surveyed pa-
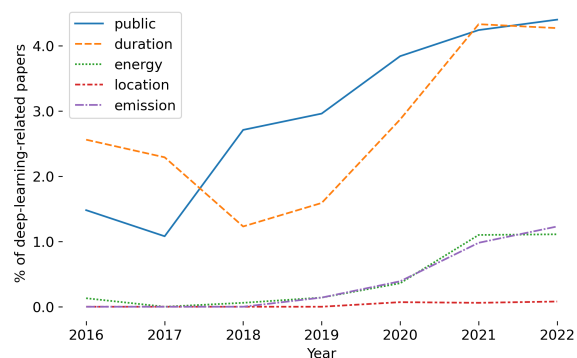

Figure 2: Development of proportions of deep-learning-related *ACL papers discussing *public* model weights, *duration* of model training or optimization, *energy* consumption, *location* where computations where performed, and *emission* of GHG. While climate awareness is on the rise, it is still low overall. The numbers were calculated by counting pattern matches for papers in the ACL Anthology.

pers from ACL, NeurIPS, and CVP. They noted whether authors claim their main contribution to improving accuracy or some related measure, an improvement to efficiency, both, or other. In all the conferences they considered, a large majority of the papers target accuracy. However, we claim that the issue is more complex, and it is not sufficient to consider only the "main contribution." Every paper should ideally have a positive impact or provide sufficient information to discuss meaningful options to reduce and mitigate negative impacts.

3.1 Quantitative Analysis

We analyze the statistics of papers in *ACL venues from 2016–2022 by downloading them from the ACL Anthology.[12] However, instead of focusing on the main contribution, we look for any discussions on climate-related issues. We identify five dimensions in our study sample and create a regular expression pattern to match text for each (see Appendix A). These dimensions are public model weights, duration of model training or optimization, energy consumption, the location where computations are performed, and GHG emissions. If the pattern matches the text of a paper at least once,[13] we consider that paper as discussing the corresponding category. We derive the proportions of papers by dividing the number of papers discussing a cate-

---

Dimension	public	duration	energy	location	emission
Proportion (%)	13	28	0	3	0

Table 1: Proportions along dimensions from Figure 2 in a manual annotated sample from EMNLP 2021.

gory by the number of deep learning-related papers. We only consider deep learning-related papers, as for these papers, climate-related issues are of much higher relevance than for those using other approaches (Strubell et al., 2019).

Figure 2 shows our findings. In general, researchers discuss climate-related issues more and more in their work. For instance, the proportion of papers that publish their model weights has almost quadrupled from about 1% in 2017 to more than 4% in 2022. We also find an increase in the proportion of papers that provide information on emissions or energy consumption. Nevertheless, the proportion for these categories remains low.

3.2 Manual Annotation

To complement our automatic pattern-based search approach, we also manually annotate a random sample of 100 papers from EMNLP 2021[14] for the same five dimensions as before. Table 1 shows the proportions. Borderline cases are counted as "reported" for an optimistic estimate. This leads to the proportions of papers publishing model weights (13%) and the duration of model training or optimization (28%) being much higher than with our automatic approach, which cannot judge borderline cases and is thus more restrictive. Still, these proportions are at a low level. The proportion of papers reporting on the location, energy consumed and GHG emitted are in line with the results from our pattern-based search.

3.3 Qualitative Survey

We examine article contents to ensure precision and to elaborate on existing practices. We review papers reporting on at least one dimension according to our pattern-based search or our manual analysis. Interestingly, many papers provide information in the context of reproducibility, publishing code but not necessarily model weights and reporting computation time only for specific steps.

As examples for specific papers that go beyond what is usually expected in terms of reporting, Anderson and Gómez-Rodríguez (2021) evaluate both

accuracy and efficiency in dependency parsers, finding that different approaches are preferable depending on whether accuracy, training time or inference time are prioritized. Lakim et al. (2022) provide a detailed holistic assessment of the carbon footprint of an Arabic language model, considering the entire project, including data storage, researcher travel, training and deployment.

Our findings highlight the need to raise awareness of climate-related issues further and find a simple but effective way to report them transparently. Besides awareness and facilitation, incentives to address these issues could be a complementary approach. However, in the rest of this paper we focus on the former "intrinsic" motivation factors, leaving "extrinsic" motivation factors to future work.

4 Towards Actionable Awareness

Efficiency (alongside accuracy) has been one of the main objectives in NLP (and computer science in general) long before its environmental aspects have been widely considered. In general, it refers to the amount of resources consumed (input) in order to achieve a given goal, such as a specific computation or accuracy in a task (output). Different definitions of efficiency correspond to different concepts of input and output. It is crucial to (1) understand the different concepts, (2) be aware of their differences and consequently their climate impact, and (3) converge towards a set of efficiency measures that will be applied for comparable climate performance evaluation in NLP research.

4.1 Related Work in NLP and AI

Strubell et al. (2019) quantify the financial and environmental cost of various NLP models, exposing substantial costs from model development and not just final model training. They recommend reporting training time and hyperparameter sensitivity, and prioritizing efficient hardware and algorithms.

Schwartz et al. (2020) compare several efficiency measures, focusing on input or resource consumption: CO_2eq emission, electricity usage, elapsed real time, number of parameters, and FPO (floating-point operations). They suggest FPO as a concrete, reliable measure for climate-related efficiency that does not depend on the underlying hardware, local electricity infrastructure, or algorithmic details. They suggest measuring efficiency as a trade-off between performance and training set size to enable comparisons with small training budgets.

---

[14]https://2021.emnlp.org/

Henderson et al. (2020) show that FPOs "are not adequate on their own to measure energy or even runtime efficiency." They recommend reporting various key figures, providing an automatic tool.

Alongside improvements in measurement methods, AI computations increasingly utilize cloud infrastructures, hindering transparency. Dodge et al. (2022) provide a framework to measure carbon intensity in cloud instances, finding that data center region and time of day play significant roles.

Finally, Wu et al. (2022) highlight the role of system hardware development in AI environmental impact, encouraging a holistic perspective.

4.2 Adopting Principles from Finance

The Greenhouse Gas Protocol[15] is a widely used reporting framework for corporates. However, this standard does not foresee, so far, an explicit ICT (information and communications technology) component. We build on the general principles of the GHG Protocol (relevance, completeness, consistency, transparency, and accuracy) to propose principles for improving climate-related performance reporting of AI. While the Greenhouse Gas Protocol focuses on GHG emissions, we propose a more general framework corresponding to the different concepts of efficiency. We, therefore, replace the term GHG emissions with the term *climate-related performance assessments*.[16]

Relevance Ensure the climate-related performance assessment appropriately reflects the climate-related performance of training, evaluation and deployment, and serves the decision-making needs of users—both internal and external to the research group. Consider both factors inherent to the model (e.g., number of parameters) and model-external factors (e.g., energy mix).

Completeness Account for and report on all relevant climate-related performance assessment items, using standardized model cards (see §5) to ensure accessibility to relevant information. Disclose and justify any specific exclusions or missing information, and explain which data input would be required to provide it. State

how you will deal with the missing information in the future to reduce information gaps.

Consistency Use consistent methodologies to make meaningful comparisons of reported emissions over time. Transparently document any changes to the data, inventory boundary, methods, or other relevant factors in the time series. Use readily-available emission calculation tools to ease comparison with other models. If you decide not to use available tools, explain why you deviate from available tools and report your assumptions about the energy mix, the conversion factors, and further assumptions required to calculate model-related emissions.

Transparency Address all relevant issues factually and coherently to allow reproducible measurement of climate-related performance by independent researchers. Disclose any relevant assumptions and refer to the accounting and calculation methodologies and data sources used.

Accuracy of reporting Achieve sufficient accuracy of the quantification of climate-related performance to enable users to make decisions with reasonable assurance as to the integrity of the reported information. Ensure that you report on the climate-related performance, even if you are in doubt about the accuracy. If in doubt, state the level of confidence.

4.3 Actions Towards Improvement

Reporting climate-related performance is not a goal on its own. Instead, it should be a means to raise awareness and translate it into actionable climate-related performance improvements when training and deploying a model. In addition, climate-aware model performance evaluations should ensure that downstream users of the technology can use the model in a climate-constrained future. Researchers should aim for climate-resilient NLP and algorithms to unlock long-term positive impacts. How to future-proof AI and NLP models should become an essential consideration in setting up any project.

The overall process of integrating these considerations would use enhanced transparency to unlock actionable awareness. Reporting on climate-related model performance should put researchers in a position to reflect on their setup and take immediate action when training the next model. To support

---

[15]https://ghgprotocol.org/

[16]While the primary focus is about eventual GHG emissions in our case as well, by addressing *climate-related performance*, we shift the focus from their direct measurement to a more holistic viewpoint.

this reflection for the researchers, the following proposes our climate performance model card.

5 Climate Performance Model Cards

Since 2020, NeurIPS requires all research papers to submit broader impact statements (Castelvecchi et al., 2021; Gibney, 2020). NLP conferences followed suit and introduced optional ethical and impact statements, starting with ACL in 2021.[17] Leins et al. (2020) discuss what an ethics assessment for ACL should look like but focus solely on political and societal issues. Tucker et al. (2020) analyze the implications of improved data efficiency in AI but only discuss the societal aspect of access in research and industry, leaving environmental issues unexplored. Mitchell et al. (2019) introduced model cards to increase transparency about data use in AI, similarly due to societal issues. We propose extending impact statements and model cards to include information about the climate-related performance of the development and training of the model, improvements compared to alternative solutions, measures undertaken to mitigate negative impact, and importantly, about the expected climate-related performance of reusing the model for research and deployment.[18]

Our proposed model card also includes any *positive* impact on the environment. A large direct negative impact does not rule out net positive impact due to contribution to downstream environmental efforts. While net impact cannot be measured objectively, since it depends on priorities and projections on the future use of the technology, we can set a framework for discussing this complex issue, providing researchers with the best practices to inform future researchers and practitioners.

Table 2 shows our proposed sustainability model card, structured into a minimum card and an extended card. The minimum card contains very basic information about the distribution of the model, its purpose for the community, and roughly the computational work that has been put into the optimization of the models. The extended card then includes the energy mix to compute the CO_2eq emissions. In total, our sustainability model card contains eleven elements:

Minimum card	
Information	**Unit**
1. Is the resulting model publicly available?	Yes/No
2. How much time does the training of the final model take?	Time
3. How much time did all experiments take (incl. hyperparameter search)?	Time
4. What was the power of GPU and CPU?	Watt
5. At which geo location were the computations performed?	Location

Extended card	
6. What was the energy mix at the geo location?	gCO$_2$eq/kWh
7. How much CO_2eq was emitted to train the final model?	kg
8. How much CO_2eq was emitted for all experiments?	kg
9. What is the average CO_2eq emission for the inference of one sample?	kg
10. Which positive environmental impact can be expected from this work?	Notes
11. Comments	Notes

Table 2: Proposed climate performance model card.

1. **Publicly available artefacts.** In recent years, NLP researchers often make their final model available for the public. This trend came up to increase transparency and reprehensibility, yet, at the same time, it avoids the necessity to train frequently used models multiple times across the community (Wolf et al., 2020). Thus, by publishing model (weights), computational resources and thereby CO_2eq emissions can be reduced.

2. **Duration—training of final model.** This field denotes the time it took to train the final model (in minutes/hours/days/weeks). In case, there are multiple final models, this field asks for the training time of the model which has been trained the longest.

3. **Duration of all computations.** The duration of all computations required to produce the results of the research project is strongly correlated with the CO_2eq emissions. Thus, we want to motivate NLP researchers to vary model types and hyperparameters reasonably. While determining the beginning of a project and deciding what counts as an experiment are in many cases difficult and subjective, we claim that an estimate of this quantity, along with a transparent confidence margin, is better than leaving it unreported.

4. **Power of hardware.** Besides the duration of training, the power of the main hardware is a driving factor for CO_2eq emissions. Depending on the implementation, the majority of energy is consumed by CPUs or GPUs. We ask researchers to report the power in watts of the main hardware being used to optimize the model. For the sake of simplicity, we ask to specify the peak power of the hardware, for which the sum of the thermal design power (TDP) of the individual hardware components is a reasonable proxy. The manufacturers provide this information e.g. on their website.[19] We want to underline again, that this model card's objective is not to have the most precise information but rather to have a rough estimate about the power.

5. **Geographical location.** The energy mix (the CO_2eq emissions per watt consumed) depends on the geographical location. Thus, it is important to report where the model was trained.

6. **Energy mix at geographical location.** To compute the exact CO_2eq emissions, the energy mix at the geographical location is required. Organizations such as the International Energy Agency (IEA)[20] report these numbers.

7. **CO_2eq emissions of the final model.** This field describes an estimation for the emitted CO_2eq. Given the time for the computation (see item 3), the power, and the energy mix, the total CO_2eq emissions for the research can be calculated by

$$\text{ComputationTime (hours)} \times \text{Power (kW)} \times \text{EnergyMix (gCO}_2\text{eq/kWh)} = \text{gCO}_2\text{eq.}$$

Although awareness of the factors that affect CO_2eq emissions is important, we recommend using automated tools for the actual calculation (see §2.1).

8. **Total CO_2eq emissions.** Similar to the previous item, this field describes the total CO_2eq emitted during the training of all models. The calculation is equivalent to item 8.

9. **CO_2eq emissions for inference.** Given that a model might be deployed in the future, the expected CO_2eq emissions in use of the model can be of value. To assure comparison between models, we ask the authors to report the average CO_2eq emission for the inference of one sample. For a dataset of n samples, it can be calculated by

$$1/n \times \text{InferenceTime (hours)} \times \text{Power (kW)} \times \text{EnergyMix(gCO}_2\text{eq/kWh)} = \text{gCO}_2\text{eq.}$$

10. **Positive environmental impact.** NLP technologies begin to mature to the point where they could have an even broader impact and support to address major problems such as climate change. In this field, authors can state the expected positive impact resulting from their research. In case that the underlying work is not likely to have a direct positive impact, authors can also categorize their work into "fundamental theories", "building block tools", "applicable tools", or "deployed applications" (Jin et al., 2021), and discuss why their work could set the basis for future work with a positive environmental impact.

11. **Comments.** The objective of this climate performance model card is to collect the most relevant information about the computational resources, energy consumed, and CO_2eq emitted that were the result of the conducted research. Comments can include information about whether a number is likely over- or underestimated. In addition, this field can be used to provide the reader with indications of possible improvements in terms of energy consumption and CO_2eq emissions.

6 Discussion

AI and NLP research are behind in incorporating sustainability discourse in the discussion. In the field of finance, an increasing amount of companies worldwide are soon required to state their environmental and broader sustainability-related impacts and/or commitments in their annual reports,

---

[19]See, for instance, https://www.nvidia.com/en-us/data-center/a100/#specifications or https://ark.intel.com/content/www/us/en/ark.html. Alternatively, users can run nvidia-smi on the command line if using an NVIDIA GPU.

[20]See https://www.iea.org/countries.

mostly following the recommendations laid out by the Task Force on Climate-related Financial Disclosures (TCFD; Financial Stability Board, 2017).[21]

Responsibility and accountability. Significant differences exist between annual reports and research papers: companies are increasingly asked to take responsibility for their actions and are held accountable to their commitments by stakeholders, while researchers can shake off responsibility by transferring it to practitioners who use technology based on their research. Researchers are thus never held responsible for committing to reducing negative environmental impact unless they choose to submit their work to specific workshops or conference tracks on sustainable and efficient NLP. However, there are no best practices on what they can do to help those who are responsible for committing to sustainability—what information is necessary for accurate reporting and informed decision making?

Extrapolation to indirect impact. The quantification of indirect impact during reuse and deployment of artifacts developed in research is complex and can only be estimated. We, therefore, expect that this discussion in environmental impact statements will be more abstract and harder to assess. As a framework, we propose borrowing the notion of scopes from corporate GHG accounting (Patchell, 2018), where scopes 1, 2 and 3 correspond, respectively, to direct emissions (not applicable to NLP research); indirect emissions from operations, e.g., due to energy consumption (very common in NLP research); and indirect emissions upstream or downstream the value chain. For our case, we suggest the following scopes:

1. Emissions generated during experiments for the paper itself, usually electricity consumption-related.
2. Impact on other researchers and practitioners in reducing emissions using the technology.
3. The use of the technology for reducing emissions or other positive impact.

Note that these correspond, respectively, to scopes 2, 3 and 3 in the GHG Protocol mentioned above.

Multi-objective optimization. Performance should not only be assessed in terms of output, but also inputs required to obtain a certain outcome. Based on this principle, performance evaluation

Standard	**Model performance** (i.e., model output accuracy)
Emerging	**Climate-related performance** (i.e., CO_2eq emissions generated by training, deploying and using the model)
Future	**Climate-related efficiency performance** (i.e., marginal accuracy improvements relative to marginal input requirements)

Table 3: Extended model performance evaluation.

should be based on both model performance and climate performance (cf. Table 3). This can take the form of explicitly introducing climate performance into the objective function for optimization (Puvis de Chavannes et al., 2021) and in benchmarking (Ma et al., 2021).

Positive impact. NLP is relevant in several aspects to the UN sustainable development goals (Vinuesa et al., 2020; Conforti et al., 2020; Swarnakar and Modi, 2021). Jin et al. (2021) defined a framework for the social impacts of NLP, of which environmental impacts are a special case. They define an impact stack consisting of four stages, from (1) fundamental theory to (2) building block tools and (3) applicable tools, and finally to (4) deployed applications. Furthermore, they identify questions related to sustainable development goals for which NLP is relevant. They categorize Green NLP as relevant only to the particular goal of "mitigating problems brought by NLP," by minimizing direct impact as part of technology development. However, we claim that Green NLP must be viewed more broadly. For example, Rolnick et al. (2019) discuss how machine learning can be used to tackle climate change, listing several fields with identified potential. For NLP, they mention the impact on the future of cities, on crisis management, individual action (understanding personal footprints, facilitating behavior change), informing policy for collective decision-making, education, and finance. Stede and Patz (2021) note that the topic of climate change has received little attention in the NLP community and propose applying NLP to analyze the climate change discourse, predict its evolution and respond accordingly. Indeed, NLP is increasingly being used to analyze sustainability reports and environmental claims, facilitating enforcement of reporting requirements (Luccioni et al., 2020; Bingler et al., 2021; Stammbach et al., 2022).

---

[21]For instance, in the United Kingdom a new legislation will require firms to disclose climate-related financial information, with rules set to come into force from April 2022.

7 Recommendations

As pointed out by Schwartz et al. (2020), a comparison between research and researchers from various locations and with various prerequisites can be difficult. Therefore, we want to point out rules of thumb that, in our opinion, should be followed by the authors of papers, as well as from reviewers who assess the quality thereof.

Do increase transparency. With our climate performance model card, we aim to provide guidelines that give concrete ideas on how to report energy consumption and CO_2eq emissions. Our model card, on purpose, still allows for flexibility so that authors can change it to their respective setup. In case of high CPU usage, the authors can simplify their calculation of energy consumption by only looking at the CPU power; in the case of the GPU, it can simply be based on the GPU. Our main goal is transparency for users and increased awareness for modelers and researchers. Hence, transparency is to be weighted over accuracy.

Do use the model cards to enable research institutions and practitioners to report on their climate performance and GHG emissions. An increasing number of first-moving research labs and institutes have started to account for their GHG emissions from direct energy use and flying, and intend to include their ICT emissions (e.g., ETH Zurich, 2021; UZH Zurich, 2021). However, harmonized approaches are still lacking. Use the model cards to road-test how far they could support your institutions' GHG and climate impact reporting.

Do not use our model card for assessing research quality. The value of research is often only clear months or years after publication. Thus, the ratio between emitted CO_2eq and contribution to the NLP community cannot be measured accurately. Additionally, the emitted CO_2eq depends on the hardware used for the computations. Researchers working with less energy-efficient hardware would have a disadvantage if the emitted CO_2eq were being used for assessing the quality. However, considering the energy efficiency of model performance might indirectly reduce a Global North–South bias, given that access to computational power is not evenly distributed across the World. Hence, targeting energy efficiency and reducing the computational power required to train and run models might mitigate some concerns on the inequality of research opportunities.

Do not report your voluntary financial climate protection contributions as emission offsetting. While emission offsetting used to be hailed as an efficient way to reduce global greenhouse gas emissions, this notion had to be revised with updated climate science consensus, at the latest with the IPCC's Special Report on Global Warming of 1.5C from 2018 (Masson-Delmotte et al., 2018). Related to this aspect, do not communicate relative (efficiency-related) improvements as absolute climate-related performance improvements.

Do not use this model card to assess net climate-related impacts. AI as an enabler for higher-order effects, for example, for climate-neutral economies and societies, is an important topic, which is, however, not in our scope. Instead, our approach aims to increase transparency about *every model's first order effects*, be the model designed for societal change (or any other higher-order effect) or not.

8 Conclusion

We argued that branding efficient methods in NLP as *green* or *sustainable* is insufficient and that due to the importance of the issue, climate awareness must be promoted in mainstream NLP rather than only in niche areas. We conducted a survey of climate discussion in NLP papers and found that climate-related issues are increasingly being discussed but are still uncommon. We proposed actionable measures to increase climate awareness based on experience from the finance domain and finally proposed a model card focusing on reporting climate performance transparently, which we encourage NLP researchers to use in any paper.

While our discussion, survey, and recommendations are aimed towards the NLP community, much is applicable to other AI fields. Indeed, specific recommendations have been made for machine learning (Henderson et al., 2020; Patterson et al., 2022) and medical image analysis (Selvan et al., 2022), for example. Concurrently, Kaack et al. (2022) propose a system-level roadmap addressing both GHG emissions and use of AI for climate change mitigation holistically. Our focus on NLP enabled us to be more specific about relevant modeling components in our model card, as they are commonly used in NLP work. Furthermore, framing our arguments within the discourse initiated in the NLP community allowed us to address the specific points raised in this discussion so far, and highlight specific avenues for positive impact.

9 Limitations

While climate awareness is necessary for including environmental considerations in decisions made during NLP research work, it is not sufficient for behavior change, namely, concrete actions by researchers and practitioners to reduce their negative impact and potentially contribute positively: as evident in various other societal issues, values do not necessarily determine behavior (Boström, 2020). Instead, climate-responsible behavior must also become "the new normal" for it to be mainstream. Awareness is only the first step in reaching that goal (Lockie, 2022).

Furthermore, the climate awareness model card we propose requires less precise details than existing reporting tools (Lacoste et al., 2019; Henderson et al., 2020; Anthony et al., 2020; Schmidt et al., 2022), which could limit its usefulness for informed decision making. However, as we claim in the paper, quantifying uncertainty may mitigate over-reliance on this information, which would otherwise possibly simply not have been reported at all.

Finally, if climate reporting becomes mandatory in NLP, it can actually be used for greenwashing if it entails financial or other incentives and if there is no control mechanism to check for honesty of the researchers. This is analogous to the situation in the corporate world, and can possibly be counteracted similarly, e.g., using ClimateBert.

References

Mark Anderson and Carlos Gómez-Rodríguez. 2021. A modest Pareto optimisation analysis of dependency parsers in 2021. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 119–130, Online. Association for Computational Linguistics.

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. ArXiv:2007.03051.

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, pages 610–623.

Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2021. Cheap talk and cherry-picking: What climateBert has to say on corporate climate risk disclosures. *Available at SSRN*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens.

Magnus Boström. 2020. The social life of mass and excess consumption. *Environmental Sociology*, 6(3):268–278.

Matthew Brander and G Davis. 2012. Greenhouse gases, CO2, CO2e, and carbon: What do all these terms mean. *Econometrica, White Papers*.

Davide Castelvecchi et al. 2021. Prestigious AI meeting takes steps to improve ethics of research. *Nature*, 589(7840):12–13.

Costanza Conforti, Stephanie Hirmer, Dai Morgan, Marco Basaldella, and Yau Ben Or. 2020. Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8427–8444, Online. Association for Computational Linguistics.

Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32(1):1–12.

Magali A Delmas and Vanessa Cuerel Burbano. 2011. The drivers of greenwashing. *California management review*, 54(1):64–87.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the carbon intensity of AI in cloud instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1877–1894.

ETH Zurich. 2021. ETH Zurich Sustainability Report 2019/2020. Technical report, ETH Zurich.

European Systemic Risk Board. 2016. *Too late, too sudden : transition to a low-carbon economy and systemic risk*. European Systemic Risk Board.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

Financial Stability Board. 2017. Task force on climate-related financial disclosures. *Final Report: Recommendations of the Task Force on Climate-Related Financial Disclosures*.

Elizabeth Gibney. 2020. The battle for ethical AI at the world's biggest machine-learning conference. *Nature*, 577(7791):609–610.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Keith Hyams and Tina Fawcett. 2013. The ethics of carbon offsetting. *WIREs Climate Change*, 4(2):91–98.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, pages 1–10.

Robin Wall Kimmerer. 2013. *Braiding sweetgrass: Indigenous wisdom, scientific knowledge and the teachings of plants*. Milkweed Editions.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*.

François Lagunas, Ella Charlaix, Victor Sanh, and Alexander M Rush. 2021. Block pruning for faster transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10619–10629.

Imad Lakim, Ebtesam Almazrouei, Ibrahim Abualhaol, Merouane Debbah, and Julien Launay. 2022. A holistic assessment of the carbon footprint of noor, a very large Arabic language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 84–94, virtual+Dublin. Association for Computational Linguistics.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Stewart Lockie. 2022. Mainstreaming climate change sociology.

Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability. *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*.

Sasha Luccioni, Emi Baylor, and Nicolas Duchene. 2020. Analyzing sustainability reports using natural language processing. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. In *Advances in Neural Information Processing Systems*, volume 34, pages 10351–10367. Curran Associates, Inc.

Valérie Masson-Delmotte, Panmao Zhai, Hans-Otto Pörtner, Debra Roberts, Jim Skea, Priyadarshi R Shukla, Anna Pirani, W Moufouma-Okia, C Péan, R Pidcock, et al. 2018. Global warming of 1.5 c. *An IPCC Special Report on the impacts of global warming of*, 1(5).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Jerry Patchell. 2018. Can the implications of the ghg protocol's scope 3 standard be realized? *Journal of Cleaner Production*, 185:941–958.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Hung Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeffrey Dean. 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.

Lucas Høyberg Puvis de Chavannes, Mads Guldborg Kjeldgaard Kongsbak, Timmie Rantzau, and Leon Derczynski. 2021. Hyperparameter power impact in transformer language model training. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 96–118, Virtual. Association for Computational Linguistics.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2019. Tackling climate change with machine learning.

Victor Schmidt, Goyal-Kamal, Benoit Courty, Boris Feld, SabAmine, Franklin Zhao, Aditya Joshi, Sasha Luccioni, Mathilde Léval, Alexis Bogroff, Niko Laskaris, LiamConnell, Ziyao Wang, Armin Catovic, Douglas Blank, Michał Stęchły, alencon, Amine Saboni, JPW, MinervaBooks, Hugues de Lavoreille, Connor McCarthy, Jake Tae, Sébastien Tourbier, and kraktus. 2022. mlco2/codecarbon: v2.0.0a3.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM (CACM)*, 63(12):54–63.

Raghavendra Selvan, Nikhil Bhagwat, Lasse F Wolff Anthony, Benjamin Kanding, and Erik B Dam. 2022. Carbon footprint of selecting and training deep learning models for medical image analysis. *arXiv preprint arXiv:2203.02202*.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning.

David R. So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V. Le. 2022. Primer: Searching for efficient transformers for language modeling.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims. *Available at SSRN*.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Pradip Swarnakar and Ashutosh Modi. 2021. NLP for climate policy: Creating a knowledge platform for holistic and effective climate action. *arXiv preprint arXiv:2105.05621*.

TerraChoice. 2010. The sins of greenwashing: home and family edition. *Underwriters Laboratories*.

Aaron D Tucker, Markus Anderljung, and Allan Dafoe. 2020. Social and governance implications of improved data efficiency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 378–384.

UZH Zurich. 2021. UZH Zurich Sustainability Report 2019/2020. Technical report, UZH Zurich.

Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the sustainable development goals. *Nature communications*, 11(1):1–10.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2022. Sustainable ai: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813.

A Patterns for Quantitative Analysis

The following are the regular expression patterns applied to identify papers according to the dimensions described in §3.1.

Public model weights:

```
(((model|weight) (will be|is) ?|(
   models|weights) (will be|are)
   ?) (public|available|upload|
   made available|made public|
   provided (at|under|on)))|((
   publish|upload) [a-zA-Z0-9,
   ]{0,20}(model(s)?|weight(s)?))
   |(make [a-zA-Z0-9, ]{0,20}(
   model(s)?|weight(s)?) (
   available|public))|(provide [a
   -zA-Z0-9, ]{0,20}(model(s)?|
   weight(s)?) (at|under|on))
```

Duration of model training or optimization:

```
(((pre(-)?)?train(ing|ed)?|
   optimize|optimization|(fine(-)
   ?)?tun(e|ed|ing)) ([a-zA-Z0-9,
    ]{0,20})(for|took|take(s)?)
   ([a-zA-Z0-9, ]{0,20})(seconds|
   minute|hour|day|week|month)+)|
   hours of computation
```

Energy consumption:

```
(energy|power|electricity) (
   consumption|usage)|(is|of|at)
   [1-9]{1}[0-9]{2,5} (watt(s)?|(
   k)?w) | pue
```

Location where computations are performed:

```
((data ?center|(a|the) cloud|(
   virtual|gpu) machine|computer
   cluster|hpc) (is )?(at|in) )|(
   cloud|azure|google|aws)([a-zA-
   Z0-9, ]{0,20})region
```

GHG emission:

```
(co2(e|eq)?|ghg|carbon) (
   footprint|emission(s)?|emitted
   |offset(ting)?)
```

The patterns were applied to the full paper text (including abstract, main contents and appendices, ignoring capitalization) for deep-learning-related papers identified by matching the following pattern:

```
deep learning|neural network|lstm
   |recurrent neural network|rnn|
   transformer|mlp|convolutional
   neural network|cnn|gpt
```

ClimateBert	
1. Model publicly available?	Yes
2. Time to train final model	8 hours
3. Time for all experiments	288 hours
4. Power of GPU and CPU	0.7 kW
5. Location for computations	Germany
6. Energy mix at location	470 gCO$_2$eq/kWh
7. CO$_2$eq for final model	2.63 kg
8. CO$_2$eq for all experiments	94.75 kg
9. Average CO$_2$eq for inference per sample	0.62 mg

Table 4: Climate performance model card for ClimateBert (Webersinke et al., 2021).

B Example Model Card

Table 4 provides an example climate performance model card according to the guidelines proposed in this paper. The model is ClimateBert, a language model which was finetuned on climate-related text (Webersinke et al., 2021). The same information is provided on Hugging Face, illustrated in Figure 1. Further information about each field is provided in the following:

1. All weights of the final model are publicly available on `https://huggingface.co/climatebert`. The paper proposes a fine-tuned language model on climate-related text. Thus, the proposed models are specific to a field and not task agnostic.

2. The duration for optimizing the final model was around 8 hours. Note, that the paper proposes four final models but this field should only mention the optimization time for one model.

3. In total, we estimate the duration for all computations to be 12 days (=288 hours). This estimation is likely pessimistic, i.e., the duration for all computations was likely lower. However, we want to point out again that this model card values transparency over accuracy.

4. The main hardware used for training were 2 x NVIDIA RTX A5000 with each GPU taking 230 watts. We add another 120 watts for the remaining hardware which would not be required by our model card.

5. The models were all trained on servers in Germany.

6. The energy mix is roughly 470 gCO$_2$eq/kWh.[22]

---

[22]According to `umweltbundesamt.de/publikati`

7. Calculating

$$8 \text{ hours} \times 0.7 \text{ kW} \times 470 \text{ gCO}_2\text{eq/kWh}$$

leads to 2.63kg CO_2eq emissons.

8. Calculating

$$288 \text{ hours} \times 0.7 \text{ kW} \times 470 \text{ gCO}_2\text{eq/kWh}$$

leads to 94.75kg CO_2eq emissons.

9. A pass of 100,000 samples through the proposed model took 0.187 hours on the same server (using a batch size of 512). We then calculate

$$\frac{0.187}{100,000} \text{ hours} \times 0.7 \text{ kW} \times 470 \text{ gCO}_2\text{eq/kWh}$$
$$= 0.62 \text{ mgCO}_2\text{eq}$$

as the emission for the inference of one sample.

Positive impact. The proposed language model on its own does not directly have a positive environmental impact. However, it can be used to train more accurate NLP models on climate-related downstream tasks. For instance, question-answering systems for climate-related topics or greenwashing detectors could benefit from this pre-trained language model. This work can therefore be categorized as a "building block tools" following Jin et al. (2021), as it supports the training of NLP models in the field of climate change and, thereby, have a positive environmental impact in the future.

Possible improvements. Block pruning is a method which drops a large number of attention heads in transformer models while only decreasing model performance slightly (Lagunas et al., 2021). Thus, the number of weights after block-pruning is decreased considerably which, in turn, decreases the CO_2eq emissions. Very likely, this method would show the same effect on the proposed ClimateBert model.

C Timeline of Emissions in NLP

Figure 3 shows the computational power that was put into the development of the major NLP models (Sevilla et al., 2022). With few exceptions, the training compute for NLP models has steadily increased over the past decade. Although progress

onen/entwicklung-der-spezifischen-kohlen
dioxid-7.

has also been made in terms of more energy efficient hardware (e.g., 19.5 GFLOPS/watt in a 2013 GTX Titan to 168.3 GFLOPs/watt in a 2021 RTX A6000), the increase in terms of required FLOPs is substantially larger. For example, going from GPT (in 2018) to GPT-3 175B (in 2020), the training compute increase from 1.1E19 to 3.14E23 FLOPs—an increase by a factor larger than 25,000.

D GHG Protocol Information Requirements for Companies

Whilst the GHG Protocol does not provide an ICT sector tool, it provides emission factors by fuel source to calculate GHG emissions based on the energy consumption. The emission factors reflect the scientific climate consensus, based on the report of the Intergovernmental Panel on Climate Changes' latest Assessment Report—IPCC's AR5.[23]

In terms of specific information to be disclosed, the GHG protocol guidance states several items relevant to NLP and ML research,[24] which serve to build our model card approach. The items that can be used for our approach are presented in Figure 4.

Furthermore, the GHG Protocols' Appendix A provides a guidance on accounting for indirect emissions from purchased electricity. This would be an important source of information for AI-related GHG accounting.

---

[23]https://www.ipcc.ch/report/ar5/syr/.
Note that AR6 will be released in late 2022 or early 2023:
https://www.ipcc.ch/ar6-syr/.
[24]https://ghgprotocol.org/corporate-st
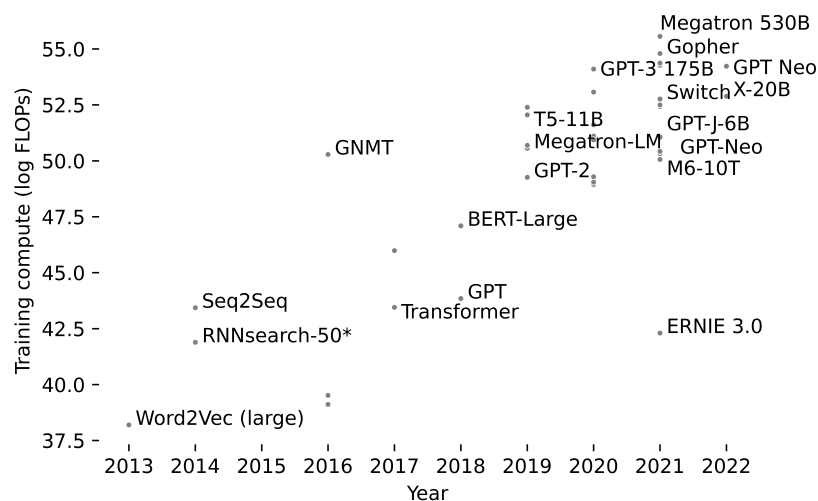andard

Figure 3: Development of floating point operations required to train NLP models. Note the log-scale.

- DESCRIPTION OF THE COMPANY AND INVENTORY BOUNDARY
  - An outline of the organizational boundaries chosen, including the chosen consolidation approach
  - An outline of the operational boundaries chosen, and if scope 3 is included, a list specifying which types of activities are covered.

- INFORMATION ON EMISSIONS
  - Total scope 1 and 2 emissions independent of any GHG trades such as sales, purchases, transfers, or banking of allowances.
  - Emissions data separately for each scope.
  - Methodologies used to calculate or measure emissions, providing a reference or link to any calculation tools used.
  - Any specific exclusions of sources, facilities, and / or operations.

- INFORMATION ON EMISSIONS AND PERFORMANCE
  - Emissions data from relevant scope 3 emissions activities for which reliable data can be obtained.
  - Emissions data further subdivided, where this aids transparency, by business units/facilities, country, source types, and activity types
  - Relevant ratio performance indicators (e.g. emissions per kilowatt-hour generated, tonne of material production, or sales).
  - An outline of any GHG management/reduction programs or strategies.
  - An outline of any external assurance provided and a copy of any verification statement, if applicable, of the reported emissions data.
  - Information on the quality of the inventory (e.g., information on the causes and magnitude of uncertainties in emission estimates) and an outline of policies in place to improve inventory quality.

- INFORMATION ON OFFSETS
  - Information on offsets that have been purchased or developed outside the inventory boundary, subdivided by GHG storage/removals and emissions reduction projects. Specify if the offsets are verified/certified and/or approved by an external GHG program.

Figure 4: Extract from the GHG Protocol Corporate Standard on which our climate performance reporting recommendations are based.