





Quantifying the Environmental Footprint of Curating Datasets with LLMs

Sarah Lang¹ , Wishyut Pitawanik¹, Pascal Belouin¹ , Emma Sevink², Jesse Olszynko-Gryn¹ , Alfred Freeborn¹ , Etienne Benson¹

¹ Max Planck Institute for the History of Science, Berlin, Germany

² Freie Universität Berlin, Berlin, Germany

Abstract

This study evaluates the environmental trade-offs of using large language models to curate cross-collection oral-history datasets in the *Commoning Oral Histories of Knowledge* (CORAL) project. Manual screening of 2,606 interviews was benchmarked against a workflow that tested four instruction-tuned LLMs and two prompt designs. Environmental impact was approximated using token-based inputs to EcoLogits, although implementing such assessments remains non-trivial. Ultimately, we conclude that the environmental impact of our project's use case could be considered moderate compared to common academic activities such as traveling to conferences. However, such impacts should be monitored closely, as they may vary significantly across different research setups and are likely to scale with larger datasets and broader adoption of LLMs in the field. Finally, the paper urges sufficiency-oriented practices and transparent carbon reporting in Computational Humanities research.

Keywords: environmental footprint, environmental sciences, oral history, Large Language Models, digital humanities, critical digital humanities

Sarah Lang, Wishyut Pitawanik, Pascal Belouin, Emma Sevink, Jesse Olszynko-Gryn, Alfred Freeborn and Etienne Benson. "Quantifying the Environmental Footprint of Curating Datasets with LLMs." *Working Paper*, 2025. <https://doi.org/10.5281/zenodo.17902822>.

© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

1 Introduction

Recent scholarship and activism – both within and beyond the digital humanities – have begun to engage critically with the environmental implications of artificial intelligence (AI), as well as with the complicity of academic research in adopting such technologies [15]. In the digital humanities, a number of initiatives have emerged in response to the ecological footprint of computational tools. Debates around this issue remain ongoing and, at times, polarised. Some argue for the wholesale rejection of AI due to ethical and ecological risks, while others advocate more moderate, exploratory approaches. Many digital humanities practitioners continue to experiment with AI-driven tools to test their potential value. Although discourse on the environmental costs of AI is expanding, one of the primary challenges lies in the difficulty of generating precise and reliable measurements of these impacts. Tools such as EcoLogits, CodeCarbon, CallToChange, and the AI Energy Score seek to address this by visualising energy consumption, thereby making its implications more accessible to researchers. This paper contributes to debates surrounding the environmental impact of the adoption of AI technologies in humanities research by evaluating the effectiveness and environmental costs of using large language models (LLMs) to identify relevant cross disciplinary boundaries in relevant oral history collections. We assess the qualitative value and environmental cost of LLM use in our specific scenario, thereby enabling a more informed judgment about their justification in scholarly workflows.

A wealth of historical and humanities-related resources is available online, for example through digital libraries, that scholars can use for their research. Yet, while much material has been digitised, researchers often face difficulties navigating these ever-more abundant collections. Scholars are often required to manually collect and enrich data to create datasets that reflect their unique needs, yet much material still remains underused or buried in vast digital corpora. Advancing digital humanities scholarship entails more than ensuring access to and the preservation of digitised materials; it also requires the development of effective entry points and analytical tools that can facilitate meaningful engagement with the existing resources. This is particularly the case for oral history, where researchers frequently encounter challenges such as non-standard metadata, diverse access protocols, and highly variable data formats. These factors complicate discovery, comparison, and analysis, underscoring the necessity of tailored infrastructures and methodologies to support scholars in deriving insights from such heterogeneous collections.

A significant challenge in the historical study of science, technology, medicine, and the environment is that, although interviews held across various institutions represent a valuable source base, they are frequently underutilised due to inconsistencies in access policies, metadata standards, and user interfaces. As a result, many potentially significant narratives remain overlooked. The *Commoning Oral Histories of Knowledge* (CORAL) project responds to this need by supporting the creation of personalized, curated oral history datasets, drawing across existing collections.¹ Such datasets tailored to specific research interests are vital not only for advancing computational analysis techniques but also for traditional scholarship engaging with the growing body of digitised sources. CORAL is not merely an aggregator for oral history interviews. Rather, it provides a

¹ <https://www.mpiwg-berlin.mpg.de/research/projects/coral-commoning-oral-histories-knowledge>
Developed by the Department on Knowledge Systems and Collective Life at the Max Planck Institute for the History of Science (MPIWG), CORAL aims to address this issue by providing a digital platform for cross-institutional discovery and thematic analysis of oral history interviews. <https://coral.mpiwg-berlin.mpg.de> While the platform does not host or provide direct access to interview recordings or transcripts – users must access these through the holding institutions – it enables the identification of relevant materials and enhances the discoverability of oral history collections. CORAL builds directly on the earlier *Commoning Biomedicine* (ComBio) project, which was developed by the MPIWG’s independent research group “Practices of Validation in the Biomedical Sciences,” led by Lara Keuck between 2021 and 2024. ComBio aimed to centralise access to disparate oral history resources in the biomedical sciences and currently provides searchable metadata for 1,637 records drawn from 15 different collections. <https://combio.mpiwg-berlin.mpg.de>. This working paper was written in July 2025 and represents the state of research at the time of writing.

workflow to help scholars craft custom collections and curate datasets for further analysis, digital or traditional, tailored to their unique research aims. To make this process more efficient, we have experimented with using LLMs in selected cases. While CORAL is designed to address a wide range of themes, in the case of the *Storying the Earth and Environmental Sciences* (SEES) sub-collection we discuss in this article, we are specifically focusing on environment-related sources. Such topics are often subsumed under a wide range of synonyms and abstract terms, which complicates traditional filtering methods like simple, targeted keyword searches. LLMs may offer a means of improving the efficiency of discovery and retrieval of relevant materials in such contexts.

It is important to note that LLMs do not always add value and thus, should only be employed selectively, especially given their environmental impacts. For topics suited to keyword searches, where relevant terms are neither obscured by umbrella concepts nor require complex semantic interpretation, LLM-based approaches may be unnecessary. LLMs might even introduce errors, thus not only failing to add value but actually subtracting it. Our present research thus involves a comparative evaluation of this LLM-supported workflow in terms of its environmental impact and usefulness in our research. By measuring energy consumption and comparing performance against conventional methods, we aim to determine whether such approaches offer a sustainable and effective means of supporting our research practices. This paper exemplifies an approach that balances the use of advanced computational technologies to enhance humanities research with critical reflection on our own practices [12, cf. p. vii–xii, p. 1–7]. As digital humanities scholars have noted, we must ‘turn the microscope on ourselves [13, p. 73] from time to time to understand how these tools shape our work and whether their use is justified – accepting that, in some cases, LLMs may not offer sufficient benefits to warrant their environmental costs.’²

2 Environmental Impact and Sustainability of Large Language Models

The *Digital Humanities and the Climate Crisis Manifesto*, authored by a transnational collective of scholars [3], recognises that the field of digital humanities both participates in and perpetuates the global climate crisis through its practices, emphasising that the digital is inherently material.³ Prendergrass and colleagues [29] critically examine digital preservation infrastructures and their environmental impacts, observing that the term *sustainability* in this context often refers more to workforce continuity and financial viability than to environmental concerns. However, these infrastructures do have an environmental footprint, which can be partially mitigated through technological adaptations such as using greener hosting providers. The paper ultimately calls for environmentally conscious archival practices to be recognised as parts of a professional ethics. Baillot contends that the apparent efficiency gains achieved through AI in the Global North result in corresponding losses of time and resources in the Global South [2]: Our AI usage obscures its asymmetric global distribution of labour and environmental burden, effectively reinforcing existing colonial structures of power. By externalising the material and human costs of digital technologies, AI development contributes to the persistence of exploitative global hierarchies, wherein the benefits accrued in one region are made possible by systemic extractions from another.

The training and deployment of LLMs entail considerable energy consumption, a demand that

² In many research contexts within the humanities, smaller, task-specific machine learning models may, in fact, be more appropriate. Such models are not only more stable and easier to fine-tune, but they also consume considerably fewer computational resources. This approach may be more compatible with a commitment to sustainability, not only in ecological terms but also in relation to financial and technical feasibility: Large models necessitate access to high-performance computing infrastructure, which many institutions cannot afford or maintain. A use-what-is-necessary approach corresponds with both ecological sufficiency and institutional resource constraints.

³ Additional institutional efforts are underway to promote transparency and responsible computing. The ‘Greening DH’ working group of the German Digital Humanities association *Digital Humanities im deutschsprachigen Raum* (DHD), for example, compiles best practices and empirical data to foster sustainability in digital scholarship [4]. Baillot also assesses the costs of digital access to text [1]. A related area of research is Digital Environmental Humanities [34].

is expected to increase significantly in the absence of regulatory measures [10].⁴ Although some argue that the environmental impact of occasional individual use is negligible – particularly when compared to emissions generated by practices such as international conference travel – this comparison should not be taken as justification for uncritical or excessive use.⁵ Taken together, existing contributions to this debate suggest that machine learning and generative AI are not inherently unsustainable. Sustainability depends on how models are trained and deployed. Nevertheless, many scholars advocate for mandatory carbon disclosure, ideally within a centralised emissions repository, and for comprehensive life cycle assessments of environmental impact.⁶ While energy use dominates discussions of sustainability in computational research, calculating it remains complex [6, 7, 11, 14, 16, 18, 19, 21, 22, 23, 24, 27, 30, 31, 33].

Tools for estimating the energy footprint of custom-trained models are now available (see overview in appendix A), but many widely-used AI platforms, such as ChatGPT, do not offer transparent data on energy usage. Even much more privacy-conscious infrastructure providers like the *Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen* (GWDG), which support high-performance computing for researchers, still lack integrated mechanisms to track or report energy consumption.⁷

Beyond energy, other forms of environmental cost remain insufficiently addressed. As Kate Crawford has highlighted, water usage is a significant but often overlooked factor [10, 20]. Many data centres are situated in desert regions where substantial water resources are diverted for cooling purposes – resources that might otherwise serve essential human needs. This raises serious ethical questions about the prioritisation of computational infrastructure over the welfare of local populations. Electronic waste (e-waste) constitutes an additional concern. The continuous advancement of AI systems necessitates ongoing hardware upgrades, leading to the disposal of older equipment. A significant proportion of this e-waste is exported to countries in the Global South, frequently without proper recycling or regulatory oversight. Thus, the environmental impact of

⁴ Kate Crawford’s seminal *Atlas of AI* underscores the material and ecological costs embedded in AI systems [9]. This concern also highlighted in the now well-known ‘Stochastic Parrots’ paper, particularly in relation to LLMs [5].

⁵ It is nonetheless worth noting that even environmentalists concur that typical everyday interactions with LLMs by individual users generate only minimal environmental cost during inference: Andy Masley, in a widely circulated 2025 blog post [26], argues that focusing climate concerns on chatbot usage of individuals is misguided, as it distracts from the structural causes of emissions. He calls for evidence-based prioritisation in climate strategy, contending that public and activist attention directed toward climate anxiety around ChatGPT diverts focus from more impactful policy-level and systemic issues. Masley supports his argument with calculations showing that while emissions from LLM usage scale with the number of queries, they remain small compared to everyday electricity and water consumption in the Global North. He advocates redirecting attention toward higher-impact behaviours such as eating meat, air travel, and structural interventions. Even the energy-intensive training phase, he notes, becomes negligible when averaged across billions of uses. In his estimate, individual ChatGPT interactions consume less than 4 Wh. However, this steady demand may drive AI companies to invest substantial resources in model training, which remains far more energy-intensive.

⁶ Patterson et al. argue that machine learning papers requiring significant computational resources should, where possible, make their energy consumption explicit, and that CO₂ emissions should be a key evaluation metric—covering both training and inference [28]. Strubell et al. similarly contend that comparing models through cost-benefit analyses, including accuracy and environmental impact, would be beneficial [33]. Luccioni and Hernandez-Garcia present a broader survey of emissions across 95 models [22]. Overall, factors such as model architecture, data centre location, and infrastructure choices play a substantial role in determining environmental outcomes, and researchers should pay close attention to these variables, which leads Patterson et al. [27] to argue that widespread adoption of best practices could significantly reduce emissions, given the large differences resulting from design and deployment choices. To assess this, full life cycle assessments (LCA) of AI services would be ideal [21]. Discussing the dilemma of sustainable scaling in AI, Desroches et al. confirm that large models consume far more energy than traditional ones, reinforcing the point that architectural decisions matter [11]. We can also draw on existing work on making AI more sustainable [8, 17, 24, 35].

⁷ The *Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen* (GWDG) functions as a service organisation jointly operated by the University of Göttingen and the Max Planck Society. It serves as a central data and IT service centre, offering infrastructure and support to research institutions and universities. Among its core responsibilities is the operation and maintenance of high-performance computing systems for academic use.

high-performance computing extends well beyond the immediate sites of technological production and use. Even areas in Digital Humanities not directly related to AI, such as the research data preservation or the long-term archiving of digital cultural heritage, are a factor in these systemic issues [29]. Researchers must therefore consider the full material footprint of AI technologies [21]. The large-scale operation of LLMs, especially when avoidable, contributes to broader patterns of ecological degradation. However, efficiency in prompting and measured API use can mitigate both financial and environmental costs. Since energy usage, as calculated by measurement tools available today, seems to correlate closely with token volume, optimising inputs and outputs offers a practical way to reduce impact in larger-scale applications.

A key concern going forward is how exactly to measure LLM energy consumption, which remains methodologically difficult to capture accurately. Reliable data on energy use per conversation, message, or token is still scarce. Toolkits such as CodeCarbon, EcoLogits, and Call-to-Change offer way to track energy consumption at runtime.⁸ These software libraries already include features to track environmental metrics related to inference-time token usage. These tools may become increasingly relevant for institutions such as libraries or digital humanities centres conducting large-scale projects. However, they often exclude training-related costs, which are better captured by aggregate metrics such as the AI Energy Score. It is important to emphasise that most environmental impact estimates for large language model usage, such as those derived using the EcoLogits Calculator [32], are based on publicly available data and generalised assumptions.⁹ These assumptions typically include average figures for energy consumption per token, model size, and hardware efficiency. As a result, such estimates offer a valuable point of comparison between different models or workflows, but they should not be interpreted as precise measurements of real-world emissions. In practice, actual energy usage and associated carbon emissions can vary considerably. Factors such as GPU performance, hardware architecture, thermal conditions, model optimisation strategies, and load-balancing across distributed systems all play a role in shaping the environmental footprint of an LLM task. Furthermore, the source of electricity powering the compute infrastructure, whether derived from renewable or non-renewable sources, can have a significant impact on carbon output, yet is often not reflected in estimation tools. Thus, while tools like EcoLogits provide a useful and accessible framework for environmental benchmarking, they remain approximations. They are most effective for comparative analysis, helping researchers identify relatively more or less efficient models or configurations. Their outputs should be interpreted with caution and contextualised by an awareness of the broader system-level variables that influence emissions.

While there is ample work on measuring the environmental impact of LLMs in Computer Science contexts [6, 7, 11, 14, 16, 18, 19, 21, 22, 23, 24, 27, 30, 31, 33], practical implementations are still rare in the Digital Humanities, with many researchers still not taking advantage of available resources.¹⁰ This may be due to a lack of preliminary work to guide the process. With this publication, we aim to address that gap by applying these tools to our research setup and documenting the relevant context and potential difficulties scholars may encounter when using them in their own work.

⁸ We provide an overview with the results of our survey of relevant tools in appendix A.

⁹ The following overview draws on information from available calculator tools and specifically the disclaimer from <https://llmemissions.com/>, a carbon calculator based on [31].

¹⁰ The CorDeep project, for instance, has introduced an interface feature allowing users to generate an environmental impact statement during inference (of its non-LLM ML application) displayed before users agree to run the analysis, with the primary aim of raising awareness about the environmental costs of deploying machine learning. It supports the principle of sustainability by reducing redundant efforts – such as the need for researchers to retrain models independently – by offering the trained model as a web service: <https://cordeep.mpiwg-berlin.mpg.de>

3 Workflow Development and Model Evaluation in CORAL

The initial phase of the CORAL project focused on surveying existing oral history collection platforms to create a comparative overview of available materials. This involved compiling structured data for each collection in a shared spreadsheet accessible to technical staff. Following this groundwork, researchers established inclusion criteria to guide the selection of relevant interviews. For instance, in the *Storying the Earth and Environmental Sciences* (SEES) subproject, we initially focused on interviews from academic experts on the history of earth and environmental sciences. In this first round of analysis, interviews centered on policy, activism, or experiential environmental knowledge were excluded. The intent was to ensure a coherent and manageable dataset for researchers working within a clearly defined domain.

However, in practice, the implementation of these criteria proved complex. As many interviews fell into a grey area between relevance and irrelevance, it would have been challenging to detect relevant interviews solely using simple keyword searches. Once criteria were set, we assessed which interviews from identified collections met these standards. Manual review began with the Voices oral history collection of the US National Oceanic and Atmospheric Administration.¹¹ This process involved examining each sub-collection, reading abstracts, and, where necessary, consulting transcripts or researching the interviewee's background. Three broad outcomes emerged: 1) fully relevant sub-collections; 2) small, partially relevant sub-collections manageable for full manual review; and 3) large, mixed-content collections that were too labour-intensive to screen manually. This latter category raised questions about scalability and led to the exploration of large language models (LLMs) as a filtering tool which would assess interview relevance and provide brief justifications, with a human expert reviewing all positively identified interviews prior to final inclusion. This approach was intended to reduce the tedium of manually screening large datasets, not to replace human judgement. In this case study, only 16% (411) of 2,606 interviews in the collections were found to meet inclusion criteria, highlighting the scale of irrelevant material and the potential labour savings offered by automated filtering.

To that end, the team tested a variety of prompts, instructing the LLM to categorise responses as 'relevant' or 'irrelevant' based on specified criteria, including keyword density and thematic cues. Despite significant effort to optimise prompts through systematic rephrasing, practical testing initially showed limited improvement through adapted prompts. Given the review structure, the project prioritised recall over precision; false positives (0;1) were acceptable as they would still undergo human review, whereas false negatives (1;0) risked the exclusion of relevant material without further verification. A formal evaluation followed, comparing LLM classifications against the manually reviewed NOAA dataset.

Analysis of classification discrepancies identified several recurring issues. The LLMs often failed to recognise academic or scientific credentials unless explicitly stated. It also struggled with adjacent disciplines, such as engineering or applied environmental roles, misclassifying them as irrelevant despite their academic dimensions. In some cases, the model offered no rationale for its decisions. Misclassifications also arose from overreliance on surface features, such as titles (e.g., "Dr.") or environmental terminology, which led to false positives among consultants, journalists, or legal professionals outside the project's academic focus. Conversely, obliquely phrased academic affiliations often resulted in false negatives. In some instances, the model generated inaccurate academic attributions based on keyword inference rather than verifiable evidence. These outcomes demonstrated the model's sensitivity to phrasing. Misjudgments frequently stemmed from conflating environmental language with academic expertise, especially in interviews referencing aspirations, informal learning, or non-research roles. The ongoing task is to refine the prompt to improve recall and precision, with attention to maintaining systematic records of prompt versions

¹¹ NOAA (National Oceanic and Atmospheric Administration) Voices Oral History Archives: <https://www.climate.gov/maps-data/dataset/noaa-voices-oral-history-archives>.

and their outcomes to prevent duplication. Crucially, the usefulness of specific prompts to the task at hand is closely tied to the clarity of inclusion criteria, which must be precisely formulated for each collection. Cases falling into the ‘grey zone’ posed challenges both for human reviewers and the LLM.¹²

4 Prompt Design, Model Selection, and Evaluation Methodology

To evaluate the potential of LLMs in classifying oral history interviews by relevance, we employed models offered by the *Gesellschaft für wissenschaftliche Datenverarbeitung* mbH Göttingen (GWDG).¹³ All models were accessed via the infrastructure provided by GWDG and no architectural modifications were made. Initially, the following pre-trained instruction-tuned LLMs were tested: llama-3.1-sauerkrautlm-70b-instruct, mistral-large-instruct, meta-llama-3.1-8b-instruct and llama-4-scout-17b-16e-instruct. Each model received a structured prompt consisting of: (1) a short topic description, (2) the extractive summary of the interview transcript, and (3) an instruction to produce a JSON-formatted response containing a binary classification (‘relevant’ or ‘irrelevant’) and a justification for the decision. To improve reproducibility and reduce stochastic variation, all models were run with a temperature setting of 0. Each transcript was evaluated using a single prompt instance. Our study then focused on two LLMs capable of structured output using the Instructor module¹⁴, llama-3.1-sauerkrautlm-70b-instruct¹⁵ and mistral-large-instruct.¹⁶ mistral-large-instruct took significantly longer to process transcripts compared to the llama variant.

Prompt Variants Two types of prompts were tested (see appendix B). The basic prompt offered concise instructions asking whether an interview reflected an academic perspective within the environmental sciences. The detailed prompt extended this with examples, a list of subfields (e.g., ecology, climatology, conservation), and clarification on relevant academic profiles.¹⁷ If an official summary was available, it was inserted before the key sentences to provide additional context.

Dataset and Task The dataset comprised 2,606 oral history interviews, each stored in PDF format. Of these, 2,491 transcripts were successfully processed and evaluated. A subset of 115 interviews was excluded due to missing or unprocessable transcripts. Among the evaluated transcripts, 411 were manually labelled by a researcher as relevant to the defined inclusion criteria.

¹² The manual classification of interviews was the most labour-intensive component of the workflow, with significant variation in difficulty depending on the collection’s structure. The breakdown of labour for this phase of CORAL is as follows:

- Researching relevant oral history collections (including legal and contact information): approximately 15 hours by the current student assistant, but building on earlier work.
- Drafting and testing LLM prompts: approximately 8 hours.
- Manual classification of NOAA interviews (n=2,455): approximately 50 hours.
- Comparative analysis of LLM vs. human classification: approximately 15 hours.
- Workflow development, coordination, and reporting: approximately 10 hours.

¹³ <https://docs.hpc.gwdg.de/services/chat-ai/models/index.html>

¹⁴ <https://python.useinstructor.com/>

¹⁵ <https://huggingface.co/VAG0solutions/Llama-3.1-SauerkrautLM-70b-Instruct>

¹⁶ <https://huggingface.co/mistralai/Mistral-Large-Instruct-2407>. Other models, including meta-llama-3.1-8b-instruct, were tested but ultimately excluded due to incompatibility with our requirement of generating structured output.

¹⁷ Future work could explore further improving the detailed prompt, for instance by explicitly requesting information on academic background and relevance to environmental science, as our results suggest that increased specificity and clarity in prompt wording significantly improve its effectiveness.

Large language models (LLMs) were selected over other machine learning classifiers for several reasons. First, LLMs offer strong generalisability to new topics without the need for retraining or model-specific tuning. Second, the LLMs we used do not retain training data internally, thereby reducing privacy concerns. Third, and crucially for this application, LLMs are capable of producing interpretable outputs in the form of natural-language justifications for each classification decision – an affordance typically unavailable in standard machine learning approaches.

The task of identifying relevant interviews within oral history collections was framed as a binary classification problem characterised by a highly imbalanced label distribution. Of the 2,606 interviews examined, only 411 (16%) were deemed relevant. The primary objective was to maximise the detection of these relevant cases; thus, the evaluation prioritised recall over other metrics, with the F1 score considered secondary.

Preprocessing and Summarisation Interview transcripts were extracted using `pypdf`¹⁸. We initially considered three summarisation strategies for inputs which exceed the 125k token limit: A first-N-token heuristic, where extracted content is truncated to fit the token limit (125k tokens), starting from the beginning of the transcript, keyword frequency (term frequency) and TF-IDF-based sentence extraction.¹⁹ For input selection, we finally adopted the first-n-token heuristic method, feeding the first portion of each transcript up to the token limit. This was based on the assumption that interviewees typically introduce their professional background at the beginning. More elaborate summarisation strategies, such as keyword frequency or TF-IDF-based extraction, were explored but ultimately abandoned.²⁰ Only five of 2,491 interviews (0.2%) exceeded the LLM context window, making such methods unnecessary for the vast majority of transcripts.

Evaluation Setup and Metrics Each model was prompted once per transcript using temperature set to 0 to minimise randomness from the LLM. Responses were returned in JSON format, including a binary classification decision (relevant or irrelevant) and a justification. If an API or formatting error occurred, the prompt was retried up to five times.²¹ After five failures, the transcript was excluded from the evaluation. Models were evaluated based on standard metrics like accuracy, precision, recall and F1 score. Since our primary objective was to identify as many relevant interviews as possible, we attempted to optimise our prompts for recall. This reflects the underlying curation strategy, wherein false positives (i.e., irrelevant interviews flagged as relevant) could be manually reviewed, but false negatives risked permanent exclusion from the curated dataset. Ongoing efforts include testing summarisation methods on the few interviews exceeding the context

¹⁸ <https://pypi.org/project/PyPDF2/>

¹⁹ The token limit for both Sauerkraut and Mistral is 128k, but we set it to 125k to maintain a 3k buffer for Instructor and other components.

²⁰ Both extractive techniques involved sentence ranking based on token frequency weights, a method with roots in early automatic summarisation research from the 1950s and 1970s. In our experimental implementation, an extractive summary was generated using a sentence-ranking algorithm based on term frequency, reminiscent of early summarisation techniques [25]. The summarisation procedure involved the following steps:

1. Tokenisation and linguistic annotation using a SpaCy NLP model.
2. Removal of stop words, punctuation, numeric characters, and whitespace tokens.
3. Computation of term frequency for the remaining tokens, normalised by the maximum frequency observed.
4. Ranking of sentence importance by summing token frequencies within each sentence.
5. Iterative selection of top-ranked sentences, in descending order of importance, until the LLM token limit was reached. Sentences were then reordered to match their original sequence.

²¹ An example of an error message illustrating the types of issues encountered is: “Generated extractive summary for Record ID 5804 (Person_Name) is empty.” or “None.”

window and developing a more targeted prompt that further improves recall while maintaining an acceptable F1 score.

Each trial combination was evaluated three times for consistency. For each model-prompt combination, we report the mean and standard error of each metric. Future work may also explore ensemble voting strategies or chunk-based processing for longer transcripts, e.g. sliding-window methods. Furthermore, some justifications produced by the models were uninformative, such as generic references to the interview subject without elaboration on their relevance. Additionally, we did not explicitly instruct the models to identify and state the interviewee’s academic background or research contributions, which may have limited the usefulness of the generated justifications for human reviewers.

Results and Discussion Our results are summarised in table 1: The detailed prompt consistently improved recall across both models, though at the cost of lower precision. For example, `llama-3.1-sauerkrautlm-70b-instruct` with the detailed prompt correctly identified 97% of relevant interviews but had only a 51% precision rate, meaning that researchers would need to manually review a large number of false positives. The best balance was achieved by `mistral-large-instruct` with the detailed prompt, which yielded high recall (96%) and moderate precision (61%).²² Responses generated using detailed prompts were also longer on average, which may have implications for inference-time energy use. Prompts could potentially be shortened, or justifications removed entirely, to reduce token count. However, this would also defeat the purpose of using a technology capable of delivering justifications, a key motivation why we chose to use this technology over others alternatives.

Model	Prompt	Evaluated	Recall	Precision	F1 Score	Avg. Tokens
llama-70b	basic	2459	0.93	0.60	0.73	69.3
llama-70b	detailed	2457	0.97	0.51	0.67	78.3
mistral	basic	2452	0.79	0.69	0.74	80.2
mistral	detailed	2450	0.96	0.61	0.74	101.0

Table 1: Model performance across prompts (more details in appendix C). These values represent the averages of three runs with very little variation between them.

5 Implementing Environmental Impact Assessments in Digital Humanities

To estimate the environmental impact of our workflow, we evaluated the available tools discussed in appendix A with regard to their usability in estimating the carbon footprint of our LLM usage scenarios. *Ecologits* is a promising solution for emissions estimation during inference, with an API and calculator interface.²³ However, its coverage of models hosted by the GWDG infrastructure is still limited. *CodeCarbon*, while detailed, tracks only local machine emissions and is, thus, not applicable to our current workflow. Other tools, such as *LLMCarbon* [14] and the `llmemissions.com` calculator based on [31], offer theoretical estimates based on model size and usage but do not track live resource consumption.

The environmental and legal implications of deploying large language models are playing an increasingly important role in research policy and institutional decision-making. The *Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen* (GWDG) serves as a pivotal infrastructure partner for our project in this regard, as it ensures that data remains within German jurisdiction,

²² See also tables 3 and 4 in Appendix C.

²³ <https://ecologits.ai/latest/>

thereby maintaining compliance with privacy standards by avoiding data transfer to the United States.²⁴ However, it is for this reason we encounter a limitation: The available libraries for calculating the environmental – or more precisely, energy – costs of LLMs are typically wrapper libraries for specific providers, such as OpenAI’s ChatGPT. These tools function automatically only when the corresponding model from that provider is used. Due to legal regulations, however, we cannot send our data to the US and must instead use a local LLM cloud service provided by GWDG.²⁵ While this is preferable in terms of privacy and data protection, it prevents us from using the more easily accessible API-based measurement tools.²⁶ According to our research, most of these libraries estimate environmental impact based primarily on the number of output tokens generated. Fortunately, EcoLogits offers a web-based tool (EcoLogits Calculator [32]), which in expert mode allows users to input their own values manually, including details such as the country in which the server is located. Since we have access to token-level output data from our own analyses, we decided to bypass the API-based methods, which would likely return inaccurate results based on US-specific energy data, and instead use the calculator to input our own values. Still, the outcome will necessarily remain an estimate.²⁷ Furthermore, our focus is solely on inference-related costs. In our view, understanding the cost per query and how it scales across a dataset alongside an approximate comparison to everyday energy consumption is sufficient. While these estimates are not exact, they still provide a meaningful sense of the overall scale of resource use.

Thus, due to the architecture of our current workflow – specifically the reliance on remote API access via GWDG infrastructure – we were unable to directly integrate common carbon footprint estimation tools such as CodeCarbon or EcoLogits’ runtime tracking modules. These tools generally require local access to hardware-level energy metrics, which is not available in our setup. In

²⁴ On the privacy aspects, see <https://info.gwdg.de/news/en/gwdg-llm-service-generative-ai-for-science/> and the models available in this service: <https://docs.hpc.gwdg.de/services/chat-ai/models/index.html>.

²⁵ The GWDG team is currently preparing a paper on its AI energy usage which will provide fuller context. However, for now, we thank Julian Kunkel from GWDG for this response which gives context on the estimated per-request energy consumption under realistic server conditions at GWDG: The precise energy cost per inference request depends heavily on context, including hardware configuration, concurrency, and usage patterns. Nonetheless, a rough approximation can be made based on representative values and observed system behaviour: A typical server with four GPUs—commonly used for running large language models—draws approximately 2 kW. At an electricity rate of €0.30 per kWh, this corresponds to €0.60 per hour in energy costs, or 7,200 kilojoules (kJ) of energy consumed per hour. From GWDG’s measurements, average execution time per request is around 10 seconds for models such as LLaMA 70B or 8B, even though the median is only 1–2 seconds. For the purpose of this estimate, we use the more conservative mean value. On a given server, either one instance of a 70B model or four parallel instances of an 8B model may be hosted. Assuming single-threaded inference (parallelism = 1), the system can process roughly 360 requests per hour. This results in an estimated energy cost of approximately 20 kJ per request for the 70B model and about 5 kJ per request for the 8B model. However, under full utilisation, the model may handle up to 64 concurrent requests. In such a configuration, while response latency per request roughly doubles (already accounted for in the 10-second average), the throughput increases significantly. In practice, this yields a potential efficiency gain by a factor of up to 64 relative to purely sequential execution but all depends heavily on the actual usage load. However, we believe the GWDG information is not particularly useful in this case for our purposes without additional data. More precisely, EcoLogits already incorporates most, if not all, energy consumption assumptions, including system architecture and concurrency factors. Since we do not have visibility into the concurrency aspects of our tasks – unless we ran them directly ourselves – these GWDG figures lack the necessary context for meaningful integration into our calculations. We wanted to include them anyway for full transparency into our research outcomes.

²⁶ In the near future we intend to test this on a dedicated machine, which would enable us to run these calculations locally during model inference. This would allow for more precise measurements, as we would no longer be reliant on estimates derived after the fact. This approach mirrors the method used by the aforementioned CorDeep team, who were able to provide more accurate measurements by running their models locally. Furthermore, our cloud LLM service provider, GWDG, is currently preparing to implement an energy reporting system, which would enable systematic measurement and annual reporting of the infrastructure’s energy use. While this system is not yet operational, we are in communication with GWDG and may be able to obtain additional data prior to review, either directly from them or through running our process locally.

²⁷ To acknowledge the conjectural nature of many of these calculations, the EcoLogits Calculator [32] uses the fitting term ‘guesstimating’ in its *Methodology* tab.

response, we adapted our workflow to record the number of output tokens generated by the model during each interaction. This figure can then be entered into tools such as the EcoLogits Calculator to yield an approximate environmental impact estimate.²⁸ These environmental assessments will be integrated into the next phase of evaluation to ensure that efficiency and sustainability remain central to model selection and deployment decisions.

While this strategy cannot offer precise runtime measurements, it provides a consistent and replicable way to approximate the carbon cost of model inference. We consider this approach both valid and instructive.²⁹ It offers a feasible alternative for researchers using hosted models and APIs, where direct measurement of energy consumption is not possible. Indeed, given the broader difficulties associated with measuring emissions from LLMs – such as variability across hardware, cloud infrastructure, and regional energy mixes – most tools available today are only capable of providing estimates. In this context, using token-based proxies for emissions is a reasonable and informative compromise. The rationale behind our implementation of these ‘guesstimate’ calculations is explained in appendix D. We recommend this simplified strategy to others working with LLMs via APIs. It is straightforward to implement and offers baseline environmental indicators that can guide more sustainable research practices.

Activity	Activity Equivalent of Sum of Runs	5× Equivalent
Walking (km)	1710–1974 km	8552–9871 km
Running (km)	1140–1316 km	5701–6581 km
Electric vehicle travel (km)	218–253 km	1092–1266 km
Streaming (hours)	389–449 h	1946–2246 h
Flight (percentage)	1.39–1.65%	6.95–8.25%

Table 2: Environmental equivalents of the total impact and corresponding 5× scaled values (calculated based on the data in C, table 5 and the methodology in appendix D)

All in all, the results displayed in appendix C and summarized in table 2 amount to approximately 250 kilometres driven in an electric vehicle. Even assuming we ran the whole workflow around five times as often, the total would equate to about 1200 kilometres – hardly an intolerable impact for a medium-sized project.³⁰ Of course, scaling this up to tens of thousands of documents would change the picture, but we already covered more than 2,000 documents here. While it is important to monitor such figures – and to acknowledge that many additional runs were, strictly speaking, unnecessary and aimed primarily at reaching publication standard – the overall environmental cost at this scale appears acceptable if the process meaningfully supports the project’s goals. The overall environmental cost at this scale appears acceptable if the process meaningfully supports the project’s goals. There seems little reason not to proceed – mindfully – aside from the broader ethical concerns such practices may raise.

²⁸ <https://huggingface.co/spaces/genai-impact/ecologits-calculator>

²⁹ We anticipate acquiring a more powerful local machine in the near future, which would allow us to rerun a subset of the models locally and generate actual runtime-based figures. If this hardware becomes available before the peer review process concludes, we intend to include these updated measurements in the final version of the paper. For now, all environmental figures reported are derived from the EcoLogits Calculator and should be interpreted as estimates.

³⁰ We mention this number as we assume it corresponds to the actual number of times we ran the workflow from its initial development to support researchers. This accounts for a relatively small number of runs during which we gradually improved the prompts and setup to achieve usable results—followed by many more iterations aimed at refining the workflow for publication.

6 Conclusion: Estimating the Environmental Costs of Computational Humanities Work

Token-based estimations are not only simple to implement; they also address the limitation that most fully automated tools require the exact model used to be supported. If a different model is used, the resulting figures would be approximations regardless. Thus, token-based estimation may offer a low-effort, viable, and sustainable entry point for integrating environmental impact assessments into computational humanities work.

However, this estimation framework also raises a broader issue about the environmental costs of optimisation and, ultimately, for Computational Humanities as a field: Bringing model performance to a publishable standard for venues such as the Computational Humanities Research Conference requires extensive testing, parameter tuning, and re-evaluation, all of which consume considerably more computational resources than would be required to produce ‘good enough’ outputs for everyday scholar support. This tension highlights the need to reconsider what constitutes sufficient model performance in a research context where the tool’s primary purpose is practical support, not optimisation for its own sake. It also points to a potential role for conferences and journals in addressing the sustainability costs of current evaluation norms, particularly the expectation that researchers repeatedly rerun large models to achieve marginal improvements in accuracy, precision, or recall. It suggests a tension between scholarly publishing practices and the original practical purpose of our tool: to assist researchers in navigating large oral history corpora. In our case, the original aim was to reduce the manual labour of filtering large oral history collections. Once the tool met that need in a functional way, the bulk of its utility had already been achieved. Further refinement was driven largely by the requirements of publishing this work, which introduces additional computational and environmental costs. From an environmental perspective, workflows that produce usable results without extensive optimisation are far more efficient. The marginal gains required for conference-ready evaluation may not justify the environmental cost, especially when the tool is not intended as a general-purpose benchmark or production model.

On one hand, we believe it is necessary to use our tools thoroughly in order to critique them responsibly. Without using them to their full potential, any critique would risk being superficial or misinformed. On the other hand, that very process of deep engagement entails a level of computational expense can reproduce the environmental concerns we are attempting to analyse. It places us in the uncomfortable position of contributing to the very phenomenon we are critiquing. Yet this predicament underscores the importance of developing evaluation standards and scholarly practices that account not only for precision and reproducibility but also for sustainability.

Acknowledgements

References

- [1] Baillot, Anne. *From Handwriting to Footprinting: Text and Heritage in the Age of Climate Crisis*. Cambridge: Cambridge University Press, 2023.
- [2] Baillot, Anne. “Why human civilization can’t afford AI.” Internal seminar, Working Group “AG Klimaschutz”, Centre Marc Bloch, Berlin. Apr. 2025. URL: <https://hal.science/hal-05053884v1>.
- [3] Baillot, Anne et al. “Digital Humanities and the Climate Crisis: A Manifesto.” <https://dhc-barnard.github.io/dhclimate/>. Online document. 2021.
- [4] Baillot, Anne et al. “Empfehlungen der AG Greening DH zum ressourcenschonenden Umgang mit Forschungsdaten.” Zenodo. Finale Version vom 30.4.2025, nach Open Peer Review. Apr. 2025. DOI: 10.5281/zenodo.15288095. URL: <https://doi.org/10.5281/zenodo.15288095>.
- [5] Bender, Emily M. et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: <https://doi.org/10.1145/3442188.3445922>.
- [6] Berthelot, Adrien et al. “Estimating the environmental impact of Generative-AI services using an LCA-based methodology.” In: *Procedia CIRP* 122 (2024), pp. 707–712. DOI: 10.1016/j.procir.2024.01.098.
- [7] Bouza, Lucía, Bugeau, Aurélie, and Lannelongue, Louis. “How to estimate carbon footprint when training deep learning models? A guide and review.” In: *Environmental Research Communications* 5., no. 11 (2023), p. 115014. DOI: 10.1088/2515-7620/acf81b.
- [8] Chauhan, Dipti, Bahad, Pritika, and Jain, Jay Kumar. “Sustainable AI: Environmental Implications, Challenges, and Opportunities.” In: *Explainable AI (XAI) for Sustainable Development*. 1st. Book chapter. Chapman and Hall/CRC, Taylor & Francis, 2024.
- [9] Crawford, Kate. *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021.
- [10] Crawford, Kate. “Generative AI’s environmental costs are soaring – and mostly secret.” In: *Nature* 626., no. 8000 (Feb. 2024), p. 693. DOI: 10.1038/d41586-024-00478-x. URL: <https://www.nature.com/articles/d41586-024-00478-x>.
- [11] Desroches, Clément et al. “Exploring the sustainable scaling of AI dilemma: A projective study of corporations’ AI environmental impacts.” Preprint. 2025. arXiv: 2501.14334 [cs.CY]. URL: <https://arxiv.org/abs/2501.14334>.
- [12] Dobson, James E. *Critical Digital Humanities: The Search for a Methodology*. Champaign, IL: University of Illinois Press, 2019.
- [13] Eichmann-Kalwara, Nickoal, Jorgensen, J., and Weingart, Scott B. “Representation at Digital Humanities Conferences (2000-2015).” In: *Bodies of Information: Intersectional Feminism and Digital Humanities*, ed. by Jacqueline Wernimont and Elizabeth Losh. 1st ed. Minneapolis, Minnesota: University of Minnesota Press, 2018.
- [14] Faiz, Ahmad et al. “LLMCarbon: Modeling the End-To-End Carbon Footprint of Large Language Models.” Preprint. 2023. arXiv: 2309.14393. URL: <https://doi.org/10.48550/arXiv.2309.14393>.

- [15] Guest, Olivia et al. “Against the Uncritical Adoption of ‘AI’ Technologies in Academia.” Zenodo. Sept. 2025. DOI: 10.5281/zenodo.17065099. URL: <https://doi.org/10.5281/zenodo.17065099>.
- [16] Henderson, Peter et al. “Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning.” Preprint. 2022. arXiv: 2002.05651. URL: <https://doi.org/10.48550/arXiv.2002.05651>.
- [17] Kaack, Lynn H. et al. “Aligning artificial intelligence with climate change mitigation.” In: *Nature Climate Change* 12 (2022), pp. 518–527. DOI: 10.1038/s41558-022-01377-7. URL: <https://doi.org/10.1038/s41558-022-01377-7>.
- [18] Lacoste, Alexandre et al. “Quantifying the Carbon Emissions of Machine Learning.” Preprint. 2019. arXiv: 1910.09700. URL: <https://doi.org/10.48550/arXiv.1910.09700>.
- [19] Lannelongue, Louis, Grealey, Jack, and Inouye, Michael. “Green Algorithms: Quantifying the carbon footprint of computation.” Preprint. 2020. arXiv: 2007.07610. URL: <https://doi.org/10.48550/arXiv.2007.07610>.
- [20] Li, Pengfei et al. “Making AI Less ‘Thirsty’: Uncovering and Addressing the Secret Water Footprint of AI Models.” In: *Communications of the ACM* 68., no. 7 (July 2025), pp. 54–61. DOI: 10.1145/3724499. URL: <https://doi.org/10.1145/3724499>.
- [21] Ligozat, Anne-Laure et al. “Unraveling the Hidden Environmental Impacts of AI Solutions: Life Cycle Assessment of AI Solutions.” In: *Sustainability* 14., no. 9 (2022), p. 5172. DOI: 10.3390/su14095172.
- [22] Luccioni, Alexandra Sasha and Hernandez-Garcia, Alex. “Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning.” Preprint. 2023. arXiv: 2302.08476. URL: <https://arxiv.org/abs/2302.08476>.
- [23] Luccioni, Alexandra Sasha, Jernite, Yacine, and Strubell, Emma. “Power Hungry Processing: Watts Driving the Cost of AI Deployment?” In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*. 2024, pp. 85–99. DOI: 10.1145/3630106.3658542.
- [24] Luccioni, Alexandra Sasha, Viguier, Sylvain, and Ligozat, Anne-Laure. “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model.” In: *Journal of Machine Learning Research* 24 (2023), pp. 1–15. URL: <https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf>.
- [25] Luhn, H. P. “The Automatic Creation of Literature Abstracts.” In: *IBM Journal of Research and Development* 2., no. 2 (1958), pp. 159–165. DOI: 10.1147/rd.22.0159.
- [26] Masley, Andy. “Why using ChatGPT is not bad for the environment – a cheat sheet.” <https://andymasley.substack.com/p/a-cheat-sheet-for-conversations-about>. Substack blog post. Apr. 2025.
- [27] Patterson, David et al. “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.” In: *Computer* 55., no. 7 (July 2022), pp. 18–28. DOI: 10.1109/MC.2022.3148714.
- [28] Patterson, David A. et al. “Carbon Emissions and Large Neural Network Training.” In: *CoRR abs/2104.10350* (2021). URL: <https://arxiv.org/abs/2104.10350>.
- [29] Pendergrass, Keith et al. “Toward Environmentally Sustainable Digital Preservation.” In: *The American Archivist* 82., no. 1 (June 2019), pp. 165–206. DOI: 10.17723/0360-9081-82.1.165. URL: <https://doi.org/10.17723/0360-9081-82.1.165>.

- [30] Poddar, Soham et al. “Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models.” In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2025. URL: <https://aclanthology.org/2025.naacl-long.632.pdf>.
- [31] Samsi, Siddharth et al. “From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference.” In: *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2023, pp. 1–9. DOI: 10.1109/HPEC58204.2023.10244409.
- [32] Samuel Rincé, Adrien Banse and Defour, Valentin. “EcoLogits Calculator.” <https://huggingface.co/spaces/genai-impact/ecologits-calculator>. 2025.
- [33] Strubell, Emma, Ganesh, Ananya, and McCallum, Andrew. “Energy and Policy Considerations for Deep Learning in NLP.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3645–3650. URL: <https://aclanthology.org/P19-1355>.
- [34] Charles Travis et al., edited by. *Routledge Handbook of the Digital Environmental Humanities*. Routledge, 2023.
- [35] Wu, Carole-Jean et al. “Sustainable AI: Environmental Implications, Challenges and Opportunities.” In: *Proceedings of the 5th MLSys Conference*. Santa Clara, CA, USA, 2022. arXiv: 2111.00364. URL: <https://doi.org/10.48550/arXiv.2111.00364>.

A Survey of Available Environmental Impact Estimation Tools

As a variety of approaches and tools have emerged to estimate the energy costs associated with LLMs, we want to provide a brief overview of those we found most relevant:

CodeCarbon is a Python package that estimates carbon emissions based on the power consumption of CPU, GPU, and RAM. It operates by either measuring power usage in real time or approximating it based on hardware specifications.³¹ CodeCarbon then correlates these values with the carbon intensity of the electricity mix from the local grid or a specified cloud provider. When exact carbon intensity data is not available, it relies on estimates derived from national or regional electricity profiles. It tracks emissions from local machines and is thus not applicable to remote-hosted models but would be suitable if LLMs were locally deployed.

EcoLogits concentrates on quantifying the environmental footprint of generative AI systems, particularly during the inference phase.³² It enables users to assess the environmental costs of interacting with large language models via API calls, covering major providers such as OpenAI and Anthropic. Grounded in life-cycle assessment methodologies, EcoLogits accounts not only for electricity used during the *usage phase* but also for environmental impacts associated with hardware production, transfer, and infrastructure in the *embodied phase*.³³

The **EcoLogits Calculator** [32] offers a visual interface for estimating the footprint of individual interactions with generative AI, building on the underlying EcoLogits Python library.³⁴ It presents metrics across multiple dimensions from electricity consumption, greenhouse gas emissions, abiotic resource depletion to primary energy use. To assist with interpretation, it contextualises results by comparing them to familiar activities such as walking, driving an electric vehicle, streaming video, or taking a transatlantic flight between Paris and New York.

CallToChange is another Python library designed to log and estimate carbon emissions associated with API calls to large language models.³⁵ It often requires as little as a single line of code to operate. Its core functionality is analysing LLM calls to compute an estimated CO₂-equivalent emission, taking into account factors such as model size, token length, cloud provider or geographic location, and hardware type.

While CodeCarbon provides system-level measurements and EcoLogits focuses on provider-specific model footprints, CallToChange targets emissions at the level of user interactions, making it especially useful for companies deploying AI services at scale. Collectively, these tools offer complementary perspectives on the environmental footprint of LLM use: CodeCarbon for system-wide energy profiling, EcoLogits for life-cycle-based analysis of provider services, and CallToChange for granular, interaction-level accounting.

The **AI Energy Score** initiative responds to a gap in sustainability reporting in light of the growing environmental footprint of generative AI technologies, responding to a lack of consistent benchmarks for assessing the energy and material demands of individual models across a range of tasks.³⁶ The AI Energy Score addresses this by proposing a systematic framework for evaluating energy efficiency, extending beyond direct emissions to incorporate broader environmental metrics such as water consumption, depletion of critical materials, and the production of electronic waste.³⁷

³¹ <https://codecarbon.io/>

³² <https://ecologits.ai/>

³³ While EcoLogits provides an API and GUI for estimating emissions, limitations for the presented project include incompatibility with some GWDG-hosted models.

³⁴ <https://huggingface.co/spaces/genai-impact/ecologits-calculator>

³⁵ <https://calltochange-theta.vercel.app/>

³⁶ <https://huggingface.github.io/AIEnergyScore/>

³⁷ At the core of the framework is a rating system that assigns models a score ranging from one to five stars. This score is based on GPU water consumption during inference across specific machine learning tasks. The framework currently encompasses ten standardised tasks, including text generation and summarisation, for which custom datasets have been developed to ensure consistency and reproducibility. Performance and environmental data are compiled into a public

B Prompt Variations

B.1 Basic prompt

The following are sentences extracted from an interview transcript with {person_name}. Based *only* on these sentences, please determine if the full interview is likely relevant to the topic: ‘Academic perspective from a profession within the environmental sciences’. For the evaluation of relevance especially consider the questions: Does this interview portray an academic perspective from a profession within the environmental sciences? Is this interview held with an academic expert within the field of the environmental sciences research?

The official interview summary (if it exists) is added, followed by the extracted sentences from the interview.

B.2 Detailed prompt

The following are sentences extracted from an interview transcript with {person_name}.

Based **only** on these sentences, please determine if the full interview is likely relevant to the topic: ‘Academic perspective from a profession within the environmental sciences’. When evaluating the interview, consider that environmental science contains various sub-disciplines such as, but not limited to: environmentalism, ecology, geography, geology, geoscience, climatology, meteorology, oceanography, hydrology, toxicology, biodiversity, agriculture, atmospheric sciences, conservation and more. For the evaluation of relevance especially consider the questions: Does this interview portray an academic perspective from a profession within the environmental sciences? Is this interview held with an academic expert within the field of the environmental sciences research? Note that the interviewee’s educational background does not necessarily have to be in environmental sciences, but their work and insights should be relevant to the field.

(*) Please rule out interviews that do not present the perspective of academic experts, even if they are involved in environmental sciences. For example, a conservationist without an academic background should not be considered relevant, while an academic regardless of their field discussing their research or work in conservation or oceanography would be relevant.

(**) Please rule out interviews that do not present the perspective of academic experts, or those that focus on environmental movements, activism, policy, or on practical forms of knowing the environment. The official interview summary (if it exists) is added, followed by the extracted sentences.

In both cases, if an official summary of the interview existed, we added a line ‘Additional context related to this interview: {official_summary}’ before the inputting the transcript sentences.

During the iterative development of the prompting strategy, two specific sentences—marked here as (*) and (**)—were identified as contributing to a significant decline in recall. Though initially intended to clarify the topic and reinforce inclusion criteria, these sentences caused the language model to adopt an overly restrictive filtering behaviour, excluding a substantial number of interviews that were judged relevant by domain experts. Sentence (*) served as a topic clarification, while sentence (**) was derived from scholar-authored instructions provided to the LLM to guide

leaderboard, which is updated twice per year. In addition to total electricity usage and associated carbon emissions, the framework reports on water consumption and other forms of resource use.

its classification decisions. In practice, both sentences led the model to apply more rigid criteria than intended, thus compromising its effectiveness in identifying relevant material.

Given that the overarching goal was to maximise recall while maintaining an acceptable F1 score, these two sentences were removed from the prompt. Commenting them out improved the model’s ability to identify borderline or implicitly relevant interviews, which were often overlooked under the stricter instructions. This underscores the sensitivity of LLM output to prompt phrasing and the importance of empirical testing in prompt design.

C Results Tables

In this section, we present two tables: the first ‘master table’ is the full data for all runs, whereas the second one with mean +/- standard error represents the aggregated results.

C.1 Model Performance Metrics

Table 3: Model Performance Metrics

Model	Prompt	Evaluated	Errors	Relevant	Irrelevant	Accuracy	Precision	Recall	F1	Avg Tokens (min-max)	Sum Tokens	Interviews
llama-3.1-sauerkrautlm-70b-instruct	basic	2459	32	403	2056	0.89	0.60	0.94	0.73	69.37 (17-315)	170583	2491
	basic	2459	32	403	2056	0.88	0.59	0.93	0.73	69.15 (17-303)	170033	2491
	basic	2459	32	403	2056	0.89	0.60	0.93	0.73	69.49 (17-344)	170869	2491
llama-3.1-sauerkrautlm-70b-instruct	detailed	2457	34	402	2055	0.84	0.50	0.97	0.66	78.37 (18-387)	192551	2491
	detailed	2459	32	403	2056	0.84	0.50	0.97	0.66	77.97 (18-417)	191720	2491
	detailed	2455	36	402	2053	0.85	0.52	0.97	0.67	78.54 (17-342)	192810	2491
mistral-large-instruct	basic	2453	38	399	2054	0.91	0.69	0.80	0.74	80.39 (15-472)	197187	2491
	basic	2453	38	399	2054	0.91	0.69	0.80	0.74	80.56 (15-472)	197615	2491
	basic	2451	40	398	2053	0.91	0.68	0.79	0.73	79.64 (16-338)	195193	2491
mistral-large-instruct	detailed	2451	40	397	2054	0.89	0.61	0.97	0.75	101.65 (18-470)	249154	2491
	detailed	2447	44	398	2049	0.89	0.60	0.96	0.74	101.03 (18-595)	247216	2491
	detailed	2454	37	401	2053	0.89	0.61	0.94	0.74	100.19 (15-552)	245857	2491

Table 4: Metrics with mean +/- standard error

Model	Prompt	Evaluated	Errors	Relevant	Irrelevant	Accuracy	Precision	Recall	F1	Avg Tokens (min-max)	Sum Tokens	Interviews
llama-3.1-sauerkrautlm-70b-instruct	basic	2459.00 ± 0.00	32.00 ± 0.00	403.00 ± 0.00	2056.00 ± 0.00	0.89 ± 0.00	0.60 ± 0.00	0.93 ± 0.00	0.73 ± 0.00	69.34 ± 0.10 (17.00 ± 0.00-320.67 ± 12.17)	170495.00 ± 245.31	2491.00 ± 0.00
	detailed	2457.00 ± 1.15	34.00 ± 1.15	402.33 ± 0.33	2054.67 ± 0.88	0.84 ± 0.00	0.51 ± 0.01	0.97 ± 0.00	0.67 ± 0.00	78.29 ± 0.17 (17.67 ± 0.33-382.00 ± 21.79)	192360.33 ± 328.78	2491.00 ± 0.00
mistral-large-instruct	basic	2452.33 ± 0.67	38.67 ± 0.67	398.67 ± 0.33	2053.67 ± 0.33	0.91 ± 0.00	0.69 ± 0.00	0.79 ± 0.00	0.74 ± 0.00	80.19 ± 0.28 (15.33 ± 0.33-427.33 ± 44.67)	196665.00 ± 746.30	2491.00 ± 0.00
	detailed	2450.67 ± 2.03	40.33 ± 2.03	398.67 ± 1.20	2052.00 ± 1.53	0.89 ± 0.00	0.61 ± 0.00	0.96 ± 0.01	0.74 ± 0.00	100.96 ± 0.43 (17.00 ± 1.00-539.00 ± 36.67)	247409.00 ± 956.64	2491.00 ± 0.00

C.2 Ecological Impacts and Equivalents Table

Table 5: Ecological impact numbers and equivalents

	Model	Prompt	kWh	GWP	ADPE	PE	Equivalents
meta-llama/Meta-Llama-3.1-70B-Instruct Usage Embodied		basic	1.93 – 2.35 1.93 – 2.35	1.3 – 1.57 1.26 – 1.53	3.16e-06 – 3.2e-06 1.7e-07 – 2.1e-07	17.46 – 21.1 16.9 – 20.54	walk: 89.08 – 107.68 km run: 59.39 – 71.79 km ev: 11.36 – 13.81 km stream: 20.28 – 24.51 h flight: 0.07 – 0.09 %
			–	0.04 – 0.04	2.99e-06 – 3e-06	0.56 – 0.56	
mistralai/Mistral-Large-Instruct-2407 Usage Embodied		detailed	4.64 – 5.25 4.64 – 5.25	3.12 – 3.51 3.02 – 3.41	7.1e-06 – 7.15e-06 4.1e-07 – 4.6e-07	41.86 – 47.18 40.61 – 45.93	walk: 213.57 – 240.73 km run: 142.38 – 160.49 km ev: 27.31 – 30.88 km stream: 48.62 – 54.79 h flight: 0.18 – 0.2 %
			–	0.1 – 0.1	6.69e-06 – 6.69e-06	1.25 – 1.25	
mistralai/Mistral-Large-Instruct-2407 Usage Embodied		basic	3.67 – 4.16 3.67 – 4.16	2.47 – 2.78 2.39 – 2.7	5.62e-06 – 5.66e-06 3.2e-07 – 3.7e-07	33.13 – 37.34 32.14 – 36.35	walk: 169.02 – 190.52 km run: 112.68 – 127.01 km ev: 21.61 – 24.44 km stream: 38.48 – 43.36 h flight: 0.14 – 0.16 %
			–	0.08 – 0.08	5.29e-06 – 5.3e-06	0.99 – 0.99	
mistralai/Mistral-Large-Instruct-2407 Usage Embodied		detailed	4.61 – 5.21 4.61 – 5.21	3.09 – 3.48 2.99 – 3.39	7.04e-06 – 7.1e-06 4e-07 – 4.6e-07	41.53 – 46.82 40.29 – 45.57	walk: 211.91 – 238.86 km run: 141.27 – 159.24 km ev: 27.09 – 30.64 km stream: 48.24 – 54.36 h flight: 0.17 – 0.2 %
			–	0.1 – 0.1	6.64e-06 – 6.64e-06	1.24 – 1.24	
continued on next page							

	Model	Prompt	kWh	GWP	ADPE	PE	Equivalents
meta-llama/Meta-Llama-3.1-70B-Instruct Usage Embodied		basic	1.93 – 2.34 1.93 – 2.34	1.3 – 1.57 1.25 – 1.52	3.15e-06 – 3.19e-06 1.7e-07 – 2.1e-07	17.4 – 21.04 16.84 – 20.48	walk: 88.79 – 107.33 km run: 59.2 – 71.55 km ev: 11.33 – 13.77 km stream: 20.22 – 24.43 h flight: 0.07 – 0.09 %
			–	0.04 – 0.04	2.99e-06 – 2.99e-06	0.56 – 0.56	
meta-llama/Meta-Llama-3.1-70B-Instruct Usage Embodied		detailed	2.18 – 2.65 2.18 – 2.65	1.47 – 1.77 1.42 – 1.72	3.57e-06 – 3.61e-06 1.9e-07 – 2.3e-07	19.71 – 23.82 19.07 – 23.19	walk: 100.55 – 121.55 km run: 67.04 – 81.03 km ev: 12.83 – 15.59 km stream: 22.89 – 27.66 h flight: 0.08 – 0.1 %
			–	0.05 – 0.05	3.38e-06 – 3.38e-06	0.63 – 0.63	
meta-llama/Meta-Llama-3.1-70B-Instruct Usage Embodied		detailed	2.17 – 2.64 2.17 – 2.64	1.46 – 1.77 1.41 – 1.72	3.56e-06 – 3.6e-06 1.9e-07 – 2.3e-07	19.62 – 23.72 18.99 – 23.09	walk: 100.12 – 121.02 km run: 66.75 – 80.68 km ev: 12.77 – 15.53 km stream: 22.79 – 27.54 h flight: 0.08 – 0.1 %
			–	0.05 – 0.05	3.37e-06 – 3.37e-06	0.63 – 0.63	
meta-llama/Meta-Llama-3.1-70B-Instruct Usage Embodied		basic	1.94 – 2.35 1.94 – 2.35	1.3 – 1.57 1.26 – 1.53	3.17e-06 – 3.21e-06 1.7e-07 – 2.1e-07	17.49 – 21.14 16.93 – 20.58	walk: 89.23 – 107.86 km run: 59.49 – 71.91 km ev: 11.38 – 13.84 km stream: 20.32 – 24.55 h flight: 0.07 – 0.09 %
			–	0.04 – 0.04	3e-06 – 3e-06	0.56 – 0.56	
meta-llama/Meta-Llama-3.1-70B-Instruct Usage		detailed	2.18 – 2.65 2.18 – 2.65	1.47 – 1.78 1.42 – 1.73	3.58e-06 – 3.62e-06 1.9e-07 – 2.3e-07	19.74 – 23.85 19.1 – 23.22	continued on next page

	Model	Prompt	kWh	GWP	ADPE	PE	Equivalents
Embodied			–	0.05 – 0.05	3.38e-06 – 3.39e-06	0.63 – 0.63	walk: 100.69 – 121.71 km run: 67.13 – 81.14 km ev: 12.84 – 15.61 km stream: 22.92 – 27.7 h flight: 0.08 – 0.1 %
Usage	mistralai/Mistral-Large-Instruct-2407	basic	3.68 – 4.16	2.47 – 2.79	5.63e-06 – 5.67e-06	33.2 – 37.42	walk: 169.39 – 190.93 km run: 112.93 – 127.29 km ev: 21.66 – 24.5 km stream: 38.56 – 43.46 h flight: 0.14 – 0.16 %
			3.68 – 4.16	2.39 – 2.71	3.2e-07 – 3.7e-07	32.21 – 36.43	
Embodied			–	0.08 – 0.08	5.31e-06 – 5.31e-06	0.99 – 0.99	
Usage	mistralai/Mistral-Large-Instruct-2407	detailed	4.58 – 5.18	3.08 – 3.47	7e-06 – 7.06e-06	41.31 – 46.56	walk: 210.74 – 237.55 km run: 140.5 – 158.36 km ev: 26.94 – 30.48 km stream: 47.97 – 54.06 h flight: 0.17 – 0.2 %
			4.58 – 5.18	2.98 – 3.37	4e-07 – 4.6e-07	40.07 – 45.32	
Embodied			–	0.1 – 0.1	6.6e-06 – 6.6e-06	1.24 – 1.24	
Usage	mistralai/Mistral-Large-Instruct-2407	basic	3.64 – 4.11	2.44 – 2.75	5.56e-06 – 5.6e-06	32.79 – 36.96	walk: 167.31 – 188.59 km run: 111.54 – 125.73 km ev: 21.39 – 24.2 km stream: 38.09 – 42.92 h flight: 0.14 – 0.16 %
			3.64 – 4.11	2.36 – 2.67	3.2e-07 – 3.6e-07	31.81 – 35.98	
Embodied			–	0.08 – 0.08	5.24e-06 – 5.24e-06	0.98 – 0.98	

D Methodology for Estimating Environmental Impact Following the Ecologits Framework

To approximate the environmental impact of our model runs, we reproduced the calculations implemented by the Ecologits Calculator, drawing upon the methodological explanations provided in its documentation.³⁸ In order to conduct these estimates, we developed a custom wrapper that emulated the internal structure of the Ecologits wrapper. This allowed us to input our own model parameters while selecting from the pre-defined models listed in the tool.

For the purposes of our estimations, we matched our usage to the closest available configurations. Specifically, we selected MistralAI / mistral large instruct, i.e. mistralai/Mistral_large_instruct-2407. In the case of the Sauerkraut variant of Meta’s LLaMA 3.1 70B model, developed by the German company Vago Solutions, we mapped it to the most comparable option in the tool: Meta’s LLaMA 3.1 70B. Although the Sauerkraut model is distinct from Meta’s original, this approximation was necessary given the available presets in the Ecologits framework.

Our analysis relies on various environmental metrics derived from three principal impact indicators provided by Ecologits: Primary Energy (PE), Total Energy (in kWh), and Global Warming Potential (GWP, measured in kgCO₂eq). These are used to generate intuitive equivalencies: To convey energy consumption in familiar physical terms (as in the EcoLogits Calculator [32]), we translated Primary Energy values (given in megajoules, MJ) into estimated distances for walking and running.³⁹ After converting MJ to kilojoules (kJ) by a factor of 1,000, we applied energy expenditure rates:

- **Walking:** 196 kJ/km, such that the walking distance is calculated as: Distance (km) =

³⁸ See ‘Methodology’ in: <https://huggingface.co/spaces/genai-impact/ecologits-calculator>.

³⁹ Our environmental impact estimations are aligned with the methodological assumptions outlined in the ‘Methodology’ tab of the Ecologits Calculator [32]. The following conversion parameters were used in the derivation of equivalent activity metrics: For physical activity equivalents, energy expenditures are based on average values associated with movement at specific speeds. **Walking** is calculated at 196 kJ/km, corresponding to a speed of 3 km/h, while **running** is assumed at 294 kJ/km, associated with a pace of 10 km/h. **Electric vehicle (EV) distance** is based on an average consumption rate of 0.17 kWh per kilometre and used to estimate how far a standard electric vehicle would travel on the same energy budget as that consumed by a given model inference run. **Streaming time equivalence** is based on the global warming potential (GWP) of the request, with 1 kgCO₂eq considered equivalent to 15.6 hours of video streaming. For comparisons with **air travel**, the calculator estimates that a return flight between Paris and New York City emits 1,770 kgCO₂eq per passenger. Assuming an average passenger load of 100 per flight, the emissions per flight are scaled accordingly. Our implementation of the ‘flight percentage’ calculation differs from the version used on the Ecologits Calculator web interface (<https://huggingface.co/spaces/genai-impact/ecologits-calculator>), which relates the flight emissions to the question: “What if 1% of the planet does this request every day for 1 year?” In this model, the impact of a single request is scaled by the factor $0.01 \times 8 \text{ billion people} \times 365 \text{ days}$, yielding a hypothetical global-use scenario. We instead based our values on the explanation provided under the ‘Methodology’ tab, specifically under the section on the number of Paris to New York City return flights, where it is stated:

We compare the GHG emissions (scaled) of the request and of a return flight Paris ↔ New York City. From impactco2.fr (<https://impactco2.fr/outils/comparateur?value=1&comparisons=&equivalent=avion-pny>) we consider that a return flight Paris → New York City → Paris for one passenger emits 1,770 kgCO₂eq and we consider an overall average load of 100 passengers per flight. We divide the scaled GHG emissions by this value to get the equivalent number of return flights.

Our version, however, does *not* apply the global scaling step. We compare the GHG emissions from a single execution of our workflow with the 1,770 kgCO₂eq emitted by a return flight for *one passenger*. We *do not* scale this further to estimate the number of flights that would result if the request were executed by 1% of the global population daily for a year. This is because our use case is not intended to be run repeatedly or widely deployed. The process is designed to be executed once (or a small number of times, once sufficient output quality is achieved) in order to support scholars in filtering and analysing data – not as a continually repeated step in an automated pipeline.

These parameters, drawn directly from the Ecologits Calculator’s own documentation, inform all derived environmental equivalencies presented in our analysis. However, it must be acknowledged that many assessments of the ecological impact of specific activities, such as estimates of the environmental impact of an intercontinental flight per passenger, are themselves controversial.

$$\frac{\text{PE (MJ)} \times 1000}{196}$$

- **Running:** 294 kJ/km, with: Distance (km) = $\frac{\text{PE (MJ)} \times 1000}{294}$

Total energy consumption, reported in kilowatt-hours (kWh), was used to estimate how far an electric vehicle could travel on the same energy (Electric Vehicle (EV) Distance). Assuming an average consumption rate of 0.17 kWh/km:

$$\text{EV Distance (km)} = \frac{\text{Total Energy (kWh)}}{0.17}$$

To relate GWP to a commonly understood activity, we used the equivalence of 1 kgCO₂eq = 15.6 hours of video streaming. Thus:

$$\text{Streaming Time (hours)} = \text{GWP (kgCO}_2\text{eq)} \times 15.6$$

For contextualising emissions, we also compared the model's GWP to the emissions from a round-trip flight between Paris and New York City, which is estimated at 1,770 kgCO₂eq:

$$\text{Flight Percentage} = \left(\frac{\text{GWP (kgCO}_2\text{eq)}}{1770} \right) \times 100$$

The use of megajoules (MJ) to represent Primary Energy values aligns with conventions in Life Cycle Assessment (LCA) methodology.⁴⁰ MJ is a standard unit in environmental impact reporting, particularly for energy-related metrics, ensuring consistency across datasets and studies.⁴¹ The Ecologits framework further sources electricity mix data from a file named `electricity_mixes.csv`, where Primary Energy factors are provided in MJ instead of kWh. To maintain fidelity to this source, all energy impact computations are retained in MJ. The total Primary Energy is calculated by multiplying the model's energy consumption (in kWh) by the relevant conversion factor (in MJ/kWh), producing a final output in MJ. We present all computed equivalence values in a summary table (table 2), based on a single run of the workflow and additionally scaled by a factor of five. This multiplier is intended to more accurately reflect the number of times we ran the full pipeline during development and output refinement, and to demonstrate how energy consumption scales with repeated use, in this case when optimising performance for publication.

D.1 Code for generating energy equivalents

```
import csv
from ecologits.tracers.utils import llm_impacts
from ecologits.utils.range_value import RangeValue

# Define the model mappings
model_mapping = {
    "llama-3.1-sauerkrautlm-70b-instruct": ("huggingface_hub", "meta-llama/Meta-Llama-3.1-70B-Instruct"),
    "mistral-large-instruct": ("huggingface_hub", "mistralai/Mistral-Large-Instruct-2407")
}

# Define the headers for the CSV file
```

⁴⁰ To aid interpretation, note that energy (measured in joules) and power (measured in watts) are related but distinct concepts. One watt is equivalent to one joule per second (1 W = 1 J/s). A watt-hour (Wh) denotes the amount of energy used over time, and 1 kilowatt-hour (kWh) is equivalent to 1,000 Wh. In this framework, power refers to the rate at which energy is consumed or generated, while energy refers to the total quantity used.

⁴¹ For more information, specifically relating to relevant units, see https://green-forum.ec.europa.eu/green-business/environmental-footprint-methods/life-cycle-assessment-ef-methods_en and for the EU Environmental Impact Assessment (EIA) Directive, see https://environment.ec.europa.eu/law-and-governance/environmental-assessments/environmental-impact-assessment_en.

```

headers = [
    "model_name", "prompt_type",
    "total_energy_min (kWh)", "total_energy_max (kWh)",
    "total_gwp_min (kgCO2eq)", "total_gwp_max (kgCO2eq)",
    "total_adpe_min (kgSbeq)", "total_adpe_max (kgSbeq)",
    "total_pe_min (MJ)", "total_pe_max (MJ)",
    "usage_energy_min (kWh)", "usage_energy_max (kWh)",
    "usage_gwp_min (kgCO2eq)", "usage_gwp_max (kgCO2eq)",
    "usage_adpe_min (kgSbeq)", "usage_adpe_max (kgSbeq)",
    "usage_pe_min (MJ)", "usage_pe_max (MJ)",
    "embodied_gwp_min (kgCO2eq)", "embodied_gwp_max (kgCO2eq)",
    "embodied_adpe_min (kgSbeq)", "embodied_adpe_max (kgSbeq)",
    "embodied_pe_min (MJ)", "embodied_pe_max (MJ)",
    "equivalent_walking_km_min", "equivalent_walking_km_max",
    "equivalent_running_km_min", "equivalent_running_km_max",
    "equivalent_ev_km_min", "equivalent_ev_km_max",
    "equivalent_streaming_hours_min", "equivalent_streaming_hours_max",
    "gwp_as_percent_of_flight_min", "gwp_as_percent_of_flight_max"
]

def custom_round(value):
    """Rounds the value to 2 decimal places, with more precision for small numbers."""
    if abs(value) < 0.0001:
        return round(value, 8)
    if abs(value) < 0.01:
        return round(value, 4)
    return round(value, 2)

def flatten_results(model_name, prompt_type, impacts):
    """Flattens the nested impact results into a single dictionary."""
    row = {"model_name": model_name, "prompt_type": prompt_type}

    def process_value(prefix, impact_name, value):
        if isinstance(value, RangeValue):
            row[f"{prefix}_{impact_name}_min"] = custom_round(value.min)
            row[f"{prefix}_{impact_name}_max"] = custom_round(value.max)
        else:
            row[f"{prefix}_{impact_name}_min"] = custom_round(value)
            row[f"{prefix}_{impact_name}_max"] = custom_round(value)

    if impacts.energy:
        process_value("total", "energy", impacts.energy.value)
    if impacts.gwp:
        process_value("total", "gwp", impacts.gwp.value)
    if impacts.adpe:
        process_value("total", "adpe", impacts.adpe.value)
    if impacts.pe:
        process_value("total", "pe", impacts.pe.value)

    if impacts.usage:
        process_value("usage", "energy", impacts.usage.energy.value)
        process_value("usage", "gwp", impacts.usage.gwp.value)
        process_value("usage", "adpe", impacts.usage.adpe.value)
        process_value("usage", "pe", impacts.usage.pe.value)

    if impacts.embodied:
        process_value("embodied", "gwp", impacts.embodied.gwp.value)
        process_value("embodied", "adpe", impacts.embodied.adpe.value)
        process_value("embodied", "pe", impacts.embodied.pe.value)

    # Add estimations

```

```

if impacts.pe:
    pe_min_kj = impacts.pe.value.min * 1000
    pe_max_kj = impacts.pe.value.max * 1000
    row["equivalent_walking_km_min"] = custom_round(pe_min_kj / 196)
    row["equivalent_walking_km_max"] = custom_round(pe_max_kj / 196)
    row["equivalent_running_km_min"] = custom_round(pe_min_kj / 294)
    row["equivalent_running_km_max"] = custom_round(pe_max_kj / 294)

if impacts.energy:
    row["equivalent_ev_km_min"] = custom_round(impacts.energy.value.min / 0.17)
    row["equivalent_ev_km_max"] = custom_round(impacts.energy.value.max / 0.17)

if impacts.gwp:
    row["equivalent_streaming_hours_min"] = custom_round(impacts.gwp.value.min * 15.6)
    row["equivalent_streaming_hours_max"] = custom_round(impacts.gwp.value.max * 15.6)
    row["gwp_as_percent_of_flight_min"] = custom_round((impacts.gwp.value.min / 1770) * 100)
    row["gwp_as_percent_of_flight_max"] = custom_round((impacts.gwp.value.max / 1770) * 100)

# Adjust headers to match the keys in the row dictionary
final_row = {}
for header in headers:
    key_name = header.split(" ")[0]
    final_row[header] = row.get(key_name)

return final_row

with open("final_impact_report.csv", "w", newline="") as csvfile:
    writer = csv.DictWriter(csvfile, fieldnames=headers)
    writer.writeheader()

with open("llm_summary.csv", "r") as summary_file:
    reader = csv.DictReader(summary_file)
    for row in reader:
        model_key = row["model_name"]
        if model_key in model_mapping:
            provider, model_name = model_mapping[model_key]

            impact_results = llm_impacts(
                provider=provider,
                model_name=model_name,
                output_token_count=int(float(row["sum_comp_tokens"])),
                request_latency=float(row["elapsed_without_sleep"]),
                electricity_mix_zone="DEU",
            )

            if impact_results and not impact_results.has_errors:
                flat_row = flatten_results(model_name, row["prompt_type"], impact_results)
                writer.writerow(flat_row)

print("Final impact report saved to final_impact_report.csv")

```