SWBL KM - Research Paper

Digital Humanism or Empty Promises? A Longitudinal Analysis of Ethical Reporting in Model Cards

Zerda Polat

Student ID: 12323111

Subject Area: Information Business

Studienkennzahl: UJ 033 561

Supervisor: Sabou, Reka Marta, Univ.Prof., Ph.D

Date of Submission: 7. January 2026

Department of Information Systems & Operations Management,
Vienna University of Economics and Business,
Welthandelsplatz 1, 1020 Vienna, Austria

Abstract

The governance framework for artificial intelligence (AI) systems frequently lags behind rigorous standards for Digital Humanism, which hinders algorithmic transparency and potentially enables unethical or biased deployment of systems. Currently, Model Cards (an instrument for standardized ethical reporting) have attracted considerable attention. Yet scarce evidence exists regarding the semantic quality of these cards. In this study, we empirically analyze the model documentation in the field and examine the content of ethical considerations within a curated dataset of 160 models. We find that the coverage of ethical risks is selective. While major tech actors prioritize social bias, they systematically omit ecological impacts. Thus, our study reveals the model of partial transparency facing current documentation practices and recommendations.

Keywords: Model Cards, Machine Learning Documentation, Digital Humanism, Algorithmic Transparency, AI Ethics

1 Introduction

Ethical documentation serves as the primary resource to understand and evaluate opaque algorithmic systems when adopting them in critical domains. Machine-learned (ML) models increasingly are scrutinized as components of social infrastructure and would benefit from rigorous transparency [7, 2]. Stakeholders resort to the documentation to answer questions such as what social biases are present, what limitations to expect, and what environmental and safety impacts to consider once the model is deployed at scale.

Nevertheless, ML models, which get shared as open-weights repositories (where their trained parameters known as "weights" are made publicly accessible) are often inconsistently documented. The same happens with proprietary services. Serious issues related to the opacity of ML models documentation have been observed in various applications, notably in discriminatory hiring, biometric surveillance, and excessive energy consumption, leading to broader concerns about their alignment with Digital Humanism principles and social justice [13].

As a reaction, significant efforts towards documenting models and datasets have been proposed. The Model Cards framework proposed by Mitchell et al. [7] has become a standard, alongside similar initiatives such as Datasheets for Datasets [4] and frameworks for internal algorithmic auditing [9]. The list also includes the popular model hosting site HuggingFace, which has adopted the term model card. Yet, it is largely unknown how the semantic quality of these proposals has impacted the practice of documenting ML models and their environmental impact.

In this work, we systematically study ML model documentation in the field. While past work has shown limited documentation during model development [2], we focus on the semantic content of ethical reporting for reusable ML models and how they meet the ethical requirements within the model cards template proposed by Mitchell et al. [7]. Specifically, this study addresses the following research questions:

- **RQ1:** To what extent and how is information about ethical consideration covered in current Model Card based documentations?
- **RQ2:** To what extent is information about ethical consideration published?
- **RQ3:** What types of ethical considerations are typically discussed?
- **RQ4:** Are there any temporal patterns in the way information about

ethical considerations is provided?

To this end, we analyze a curated dataset of 160 models, released between 2018 and 2025. A comparative analysis demonstrates that, when Generative AI was introduced, data scientists adopted documentation approaches that prioritized safety and misuse concerns over process transparency. They also showed less deliberate consideration of environmental costs compared to social biases.

Our work makes the following contributions to understanding and supporting ML model documentation practice:

(1) Delineation of the current practice of public model cards highlighting a clear gap between the social transparency provided by corporate actors and the ecological opacity maintained by the same developers;

(2) A typology of ethical considerations, based on the model cards proposal, which can be adopted by model auditors as a framework for detecting selective ethics;

(3) A longitudinal analysis, covering the "Pre-Generative" and "Generative AI" eras, to support researchers to understand and critique the regression in documentation completeness during the recent industry expansion.

The remainder of this paper is structured as follows:

- Section 2 reviews the related work regarding the model documentation frameworks and the challenges of "Red AI".
- Section 3 describes the methodlogoy, which included the data colleciton, stratification strategy, as well as the content analysis framework.
- Section 4 presents the results of the empirical analysis.
- Section 5 discusses the implications of the findings.
- Finally, Section 6 concludes the study.

The artifacts created in this study including the coded dataset, list of assessed model cards, and analysis scripts, are shared as supplementary materials alongside the paper to support future investigation on improving ML documentation can be found in the GitHub repo.

2 Related Work

2.1 Definitions and peculiarities of the problem

Performance of interest may depend on hidden context, not given explicitly in the form of global accuracy. A typical example is facial recognition error rates that may vary radically with the demographic subgroup. Another example is the patterns of toxicity detection that may change with context, depending on the dialect, cultural background, or specific terminology used.

Often the cause of failure is hidden, not known a priori, making the auditing task more complicated. Changes in the hidden context can induce more or less radical changes in the target reliability, which is generally known as bias [7]. An effective documentation framework should be able to track such changes and to quickly expose them.

To validate this approach, Mitchell et al. [7] proposed the Model Cards framework to benchmarks such as the CelebA dataset, demonstrating that while models achieved high global accuracy, disaggregated reporting revealed error rates as high as 30% in specific subgroups. Thus, an ideal documentation system should be able to:

(1) quickly adapt to subgroup performance;

(2) be robust to aggregate metrics;

(3) recognize and treat hidden failure modes.

To this end, Mitchell et al. [7] propose nine distinct sections to ensure accountability:

- **Model Details:** Lists basic metadata about the model, including the release date, version, architecture type, license, responsible parties, and citation information.
- **Intended Use:** Delineates the primary use cases and target users, as well as scenarios that are out-of-scope but easily confused with or highly related to the primary task.
- **Factors:** Records how demographic or phenotypic groups, alongside instrumental and environmental factors, influence model performance.
- **Metrics:** Covers the measurement of performance, including the specific thresholds, confidence intervals, and benchmarks utilized.
- **Evaluation Data:** Describes the datasets used to quantitatively eval-

uate the model, including justifications for dataset selection and any preprocessing procedures.

- **Training Data:** Details the provenance of the dataset used for training. When proprietary information cannot be disclosed, it should provide basic distributional statistics over groups.
- **Quantitative Analyses:** Illustrates performance through disaggregated evaluation with respect to the identified factors and their intersections.
- **Ethical Considerations:** Discusses the ethical considerations taken during development, such as the use of sensitive data, foreseen risks (e.g., environmental cost, bias), and mitigation strategies.
- **Caveats and Recommendations:** Lists additional concerns that are not covered in previous sections.

2.2 Types of documentation standards

Although previous conceptual frameworks and guidance for documentation have included the user perspective, key contextual and goal-related distinctions have not been fully discussed. To close this gap, the literature normally distinguishes between two kinds of documentation standards:

(1) output-focused (Model Cards)

(2) lifecycle-focused (FactSheets).

The framework by Arnold et al. [1], FactSheets, relies on theories of supplier conformity and highlights specific elements of the data lineage that support auditing processes.

Unlike Model Cards, which focus heavily on the final model output, FactSheets argue for documenting the entire lifecycle, including data provenance, cleaning, and testing methodologies. Although useful for specific safety-critical audits, this model shares a common limitation with the broader landscape of ethical AI guidelines [6];[5]: it relies on voluntary adherence by developers. This may lead to a lack of usability for real end-users.

2.3 Systems for handling documentation practice

Prior work has tried to create categories of documentation to define transparency needs, but, as discussed in Bhat et al. [2], practice often deviates from theory. Bhat et al. distinguish two approaches in available repositories:

(1) technical metadata recording

(2) qualitative ethical discussion.

Current evaluation studies provide insight into how developers prioritize these approaches. Bhat et al. [2] show in their experiments that qualitative ethical discussion handles transparency worse than analogous technical metadata techniques, which is likely due to the "Traceability Gap."

Their empirical analysis showed that while technical metadata is almost always present, qualitative sections are frequently treated as noise. Specifically, they found that the "Ethical Considerations" section was missing or empty in a statistically significant portion of the sample. This implies that unless developers are explicitly nudged, they prioritize technical reproducibility over ethical transparency.

2.4 Base risks for handling environmental accountability

Researchers in the HCI and ML communities have proposed frameworks for user-centered design in documentation. Prominent examples inlcude Datasheets for Datasets [4], which standardizes the documentation of data provenance, and FactSheets [1], which captures the linange of such AI services though a declaration of conformity by the supplier.

This body of literature focuses mainly on *who* the documentation is provided to (the stakeholder) and *why* they require it (specific verification goals). However, while these are important elements in understanding the social context of a model, little attention seems to be paid to *where* or *when* users require documentation regarding ecological impact. These questions relate to the physical environment in which a model is expected to operate.

In the real world, risks may often be ecological, e.g., only particular types of carbon footprints may change with scale, with early studies highlighting that a single training run can emit as much CO2 as five cars over their lifetimes [11]. This reflects a growing dominance of what Schwartz et al. [10] calls "Red AI". These systems buy incremental performance improvements at the cost of massive power consumption.

The landscape of documentation requirements has been fundamentally altered by the advent of Foundation Models (2023–2025). Bommasani et al. [3] note that the move towards large-scale generative models has rendered the supply chain more opaque.

Furthermore, Weidinger et al. [12] argue that Large Language Models (LLMs) introduce specific risks, specifically highlighting environmental impact. They note that traditional documentation often fails to capture carbon footprint benchmarks. Motivated by these evolving risks, we seek to address whether current documentation practices have adapted to capture the ecological costs inherent to Generative AI.

3 Methodology Framework

3.1 Data Overview

The Hugging Face Hub serves as the primary source for this data. It is an open online repository that provides metadata for machine learning models, based on actual developers' uploads.

We curated a dataset consisting of 160 model cards, which were all released from the time period 2018 to 2025. In order to organize the metadata, we used an Excel spreadsheet as a structured data collection tool. A specific rubric was used to serve as a blueprint for filling out this sheet, ensuring that key identifiers (including the model name, release date, license, etc.,) were captured consistently for every entry.

3.2 Data Processing and Filtering (RQ1)

We also developed a custom python script to analyse the documentation content alltogether. During the data cleaning, 2 models were excluded due to missing temporal metadata (release date), thus resulting in a final sample size of 158 models for analysis instead of the initial 160.

The Python script identified the key metadata fields for this study namely: Q1 (Creator), Q3b (Date), Q20-Q22 (Ethical Considerations), and Model Task (Hugging Face).

An algorithm to distinguish between meaningful documentation and empty placeholders was implemented. We came to the conclusion to use a binary score of 1 (Documented) or 0 (Undocumented), so we could assign it to each model based on the following logic:

- **Length Filter:** Entries with fewer than 3 characters were automatically scored as 0.
- **Negative Phrase Filtering:** We defined a list of "exclusion phrases" (e.g., *"N/A", "None", "Not explicitly discussed", "No information avail-*

able”). If the documentation contained only these phrases, it was reclassified as Undocumented (0). This approach ensures that compliance is measured by its semantic content, rather than by its file existence.

3.3 Stratification Strategy

To answer our research questions regarding the influence of creator identity and time, we stratified the dataset along three dimensions.

1. Creator Identity (Big Tech vs. Independent): As shown in Table 1, the raw data for ‘Q1 - Model Creator’ consists of the specific organization or username associated with the repository (e.g. Google, Meta). We wanted to analyze the influence of Major Tech against Independent creators on documentation quality. This is why those two distinction were created.

Table 1: Representative Examples of Raw Ethical Documentation (Excerpt from Dataset)

Q2 - Model Type	**Model Task (Hugging Face)**	**Q3b - Model Date**	**Q1 - Model Creator**	**Q20-Q22 - Ethical Considerations**
Llama-3.2-1B	Text Summarization	2024-09-25	Meta	“Testing conducted to date has not covered... Llama 3.2’s potential outputs cannot be predicted...”
stable-diffusion-v-1-4-original	Generating images from text prompts	2022-11-09	Not available	“The model should not be used to intentionally create or disseminate images that create hostile or alienating environments...”
StreetCLIP	Zero-Shot Image Classification	2024-02-19	Not available	“Even if the training data used for this model could be characterized as fairly neutral, this model can have biased predictions.”
t5-base	Text-to-text generation	2024-02-14	Not available	“Carbon emissions can be estimated using the Machine Learning Impact calculator... Hardware Type: More information needed...”

This stratification was performed programmatically using a keyword-matching algorithm. We defined a list of “Big Tech” keywords based on dominant market presence, including: *‘google’, ‘meta’, ‘facebook’, ‘microsoft’, ‘nvidia’, ‘openai’, ‘amazon’, ‘apple’, ‘ibm’, ‘stabilityai’, ‘deepmind’, ‘salesforce’, ‘adobe’, ‘intel’*.

We isolated the column "Q1 - Model Creator" to identify the creator identity. Although this field often contained unstructured data, ranging from corporate email addresses (e.g., vision-team@salesforce.com) to repository URLs (e.g., https://github.com/meta-llama/llama-models/issues), we did not scan the entire dataset. Instead, we applied a targeted search within this column only. Any creator string containing one of the defined strings was labeled as "Big Tech". All other creators (individual researcher), that did not trigger the filter were classified as "Independent" creators.

2. Temporal Era (Pre-GenAI vs. GenAI) (RQ3): We parted the timeline into two distinct eras to observe industry shifts. Models released between 2018 and 2022 were labeled *Pre-GenAI*, while those released from 2023 to 2025 were labeled *GenAI Era*. This split allows us to isolate the impact of the recent generative AI boom on documentation standards.

3. Model Modality (Domain) (RQ4): Models were categorized into domains (e.g., NLP, Computer Vision) based on their specific Hugging Face task tags. To standardize the diverse range of task names found in the metadata, we implemented a Python classification function that mapped task strings to high-level domains based on keyword presence.

The categorization logic operated as follows:

- **Multimodal:** Assigned if the task string contained cross-modal terms such as *'text-to-image'*, *'image-to-text'*, or *'multimodal'*.
- **NLP:** Assigned if the task included text-processing keywords like *'text'*, *'translation'*, *'summarization'*, *'question'*, or *'fill-mask'*.
- **Computer Vision:** Assigned if the task included visual processing terms such as *'image'*, *'video'*, *'vision'*, *'detection'*, or *'segmentation'*.
- **Audio:** Assigned for tasks containing *'audio'* or *'speech'*.

Any task not matching these predefined keywords was categorized as 'Other'. This programmatic approach ensured that functionally similar models (e.g., *'object-detection'* and *'image-classification'*) were consistently grouped under the same analytical domain.

3.4 Keyword and Theme Analysis (RQ2)

To categorize the content of the ethical reporting, we employed a dictionary-based content analysis. We grounded our theme selection in the established "Ethics Guidelines for Trustworthy AI" identified by Jobin et al. [6].

We operationalized these concepts into four distinct coding categories:

- **Bias & Fairness:** Following the original *Model Cards* proposal by Mitchell et al. [7]. This theme captures terms related to representational harm (e.g., *bias, fairness, gender, race, inequality*).
- **Safety & Misuse:** Drawing on the risk taxonomies defined by Weidinger et al. [12] and the misuse concerns highlighted by Bhat et al. [2], this category captures disclosures regarding unintended use, robustness, and toxicity (e.g., *safety, abuse, malicious, hallucination, attack*).

- **Privacy & PII:** Reflecting the global consensus on data protection found in Jobin et al.'s meta-analysis [6], this theme focuses on data provenance and consent terminology (e.g., *privacy, PII, GDPR, consent, anonymization*).
- **Environmental Impact:** Addressing the "Green AI" gap identified by Schwartz et al. [10] and the carbon footprint concerns raised by Strubell et al. [11], we added this theme to capture often-overlooked ecological metrics (e.g., *carbon, CO2, emission, energy, footprint*).

We created a custom dictionary for each theme. The algorithm iterates through the documentation text and tags a model if at least one keyword from the respective category is present. To account for non-standard terminology, we manually expanded the keyword lists (e.g., mapping *"power consumption"* to the Environmental theme).

3.5 Measurement and Analysis

To allow for fair comparisons between groups of unequal sizes (e.g., 110 NLP models vs. 24 Computer Vision models), we report findings as **Adherence Rates (Percentages)** rather than raw counts. This normalization ensures that a domain with higher upload volume does not artificially appear more transparent. This is where we analyzed the distribution of ethical themes across these groups to identify their possible priorities (e.g., comparing the frequency of "Safety" vs. "Environmental" tags in Corporate vs. Independent models).

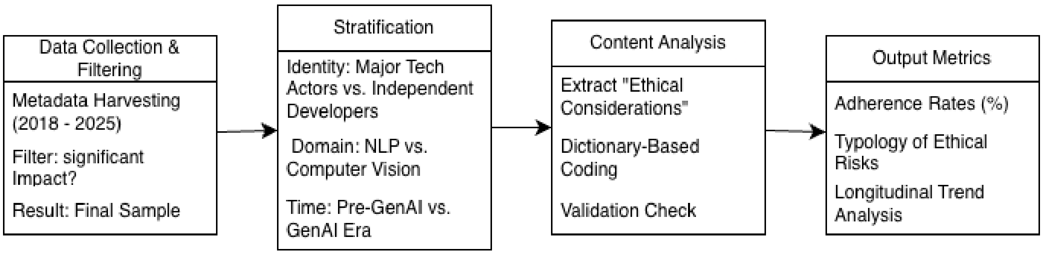
Data Collection & Filtering
Metadata Harvesting (2018 - 2025)
Filter: significant Impact?
Result: Final Sample
Stratification
Identity: Major Tech Actors vs. Independent Developers
Domain: NLP vs. Computer Vision
Time: Pre-GenAI vs. GenAI Era
Content Analysis
Extract "Ethical Considerations"
Dictionary-Based Coding
Validation Check
Output Metrics
Adherence Rates (%)
Typology of Ethical Risks
Longitudinal Trend Analysis

Figure 1: **Methodological Framework Overview.** The workflow illustrates the data harvesting process from the Hugging Face Hub, the stratification logic (by Creator, Domain, and Era), and the dictionary-based content analysis used to derive ethical adoption rates.

3.6 Threats to Validity

Finally, it is imperative that the scope of this framework is defined. Our content analysis identifies the presence of information rather than the technical correctness of reported metrics. Furthermore, we excluded linked external papers to focus on primary documentation within the repository. While the dataset of 160 models is robust, results are limited to English-language documentation. Nuanced discussions using non-standard terminology might be overlooked despite iterative dictionary expansion.

Declaration of AI Use. During the preparation of this work, we utilized Large Language Models (Gemini) to assist with the following: LaTeX typesetting, code generation for the analysis outputs and stylistic refinement and the initial planing stage to ensure that academic standards are upkept. The conceptualization, data analysis, and interpretation of findings remain the sole work of the authors.

4 Results

The developers' difficulty in maintaining balanced documentation was a common feature of many repositories and, whilst the study had been designed expecting that this might be the case, the results often highlighted significant selective reporting, with both Major Tech and Independent actors being mistaken in their prioritization of ethical risks. We shall contrast the results between all 158 analyzed models (derived from the initial 160), particularly with regard to establishing the difference between the corporate-backed disclosures and the independent disclosures, and comparing this to the difference between the Pre-Generative and Generative AI eras.

4.1 Extent of Coverage (RQ1)

To address RQ1, we first examined the general presence of ethical documentation. Of 158 models identified, 137 met inclusion criteria (86.71%), while the remaining 21 either lacked ethical considerations sections or were left undocumented.

In Table 2, we can clearly observe the extent of coverage and adoption organized by creator type. The Global Rate (86.71%) represents the average adoption rate across the entire dataset. This aggregate metric was calculated by dividing the total number of documented models (137) by the final sample size (158).

The script also performed a conditional aggregation using a boolean filter "is big tech", to calculate separate adherence for each subgroup. We operationalized the concept of “Institutional Resources” by using the corporate keyword list as a proxy: creators flagged as Big Tech’ (e.g., Meta, Google) were assumed to possess centralized compliance teams, whereas those failing the keyword filter were classified as Community-driven’.

When we compare this against specific groups ,we can see a resource divide: Major Tech actors, utilize their institutional resources to achieve a higher adoption rate of 94.74%, while Independent creators, often community-driven, show a lower adoption rate of 84.17%.

Table 2: Extent of Coverage and Adoption by Creator Type (RQ1)

Category	**Count / %**	**Context**
Total Models	158	-
Documented	137	86.71% Global Rate
Undocumented	21	13.29%
	Adoption by Creator	
Major Tech Actors	94.74%	High institutional resources
Independent	84.17%	Community-driven

4.2 Domain Analysis (RQ4)

Regarding the RQ4, the documentation adherence for different model domains is given in Table 3.

To derive these categories, we implemented a programmatic keyword analysis of the raw task metadata. A custom classification function mapped the Hugging Face task tags into standardized groups for example:

NLP: Tags containing terms such as "translation", "summarization", or "fill-mask" were aggregated under "NLP".

Computer Vision: Those referencing "detection", "segmentation", or "vision" were coded as "Computer Vision".

This ensured that functionally similar models were analyzed together regardless of minor variations in their naming conventions.

The results show similar measures for correctly identifying hazards in text-heavy domains; NLP models achieved an 86.36% adoption rate, closely tracking the global average. However, Computer Vision documentation exhibited lower adherence at 83.33%.

Table 3: Documentation Adherence by Model Domain (RQ4)

Domain	Adoption Rate	Sample (n)
Multimodal	100.00%	5
Audio	100.00%	2
Other	88.24%	17
NLP (Text)	86.36%	110
Computer Vision	83.33%	24

4.3 Typology of Ethical Themes (RQ2)

To answer RQ2 (Types of Ethical Considerations), we analyzed the specific content of the documented models. The mean scores recorded for all models are given in Table 4.

As shown in the table, the most prevalent category is Bias & Fairness, which appeared in 65.0% of the documented models. This was followed by Safety & Misuse at 48.9%, indicating that developers primarily associate "ethical reporting" with immediate representational harms and toxicity risks.

In contrast, we observe a sharp decline in adherence for the remaining categories. Privacy & PII considerations were found in only 10.2% of the sample, while Environmental Impact was the least reported theme, appearing in just 6.6% of cases. This hierarchy suggests that while social risks are frequently acknowledged and reported, the same cannot be said about ecological costs and data provenance.

Table 4: Typology of Ethical Considerations (RQ2)

Theme	% of Models	Description
Bias & Fairness	65.0%	Gender, race, representation
Safety & Misuse	48.9%	Toxicity, jailbreaking, deepfakes
Privacy & PII	10.2%	Consent, surveillance
Environmental	6.6%	Carbon footprint, energy

4.4 Comparative Analysis (RQ3 & Deep Dive)

Finally, to address RQ3 (Temporal Patterns), we performed a cross-case analysis to test the distinguishability of one epoch or creator from the other (Table 6).

In the data we can clearly observe two distinct divergences:

1. Temporal Shift (Safety):

Comparing the Pre-Generative Era and the Generative AI Era, we observed a distinct rise in safety reporting. Rather than a random variation in quality, the data shows a clear trend: adoption rates for "Safety and Misuse" increased substantially from 33.3% in the older models to 52.7% in the modern Generative era.

However, this increased focus on safety coincided with a general decline in documentation coverage. We identified a regression in completeness: while the Pre-GenAI era (2018-2022) maintained a 90.00% adoption rate, this fell to 85.94% in the GenAI era (2023-2025) as shown in Table 5.

Table 5: Temporal Evolution of Documentation Completeness (RQ3)

Era	**Adoption Rate**
Pre-GenAI (2018-22)	90.00%
GenAI Era (2023-25)	85.94%

To quantify this shift, the analysis script separated the dataset into two timelines (Pre-2023 vs. Post-2023) and calculated the percentage of models in each group that triggered the “Safety and Misuse” keyword filter.

2. Creator Divergence (Environmental):

While Independent models contained environmental disclosures in 8.9% of cases, institutional disclosures for Big Tech remained at 0.0%.

This statistic was generated by cross-referencing the "Environmental” theme against the creator label,” revealing a complete absence of ecological keywords in the Corporate sample.

However, Big Tech prioritized other areas: each Major Tech model had a higher score for identifying Bias (77.8%) than the independent category. This suggests a split in priorities, where corporate actors focus heavily on social risks while ignoring ecological costs.

Table 6: Cross-Case Comparison: Temporal Shifts and Creator Priorities (RQ3 & Deep Dive)

Theme	By Era (Temporal)		By Creator (Institutional)	
	Pre-GenAI	GenAI Era	Big Tech	Independent
	(2018-22)	(2023-25)	(Corp)	(Open Source)
Bias & Fairness	63.0%	**65.5%**	**77.8%**	60.4%
Safety & Misuse	33.3%	**52.7%**	**58.3%**	45.5%
Privacy & PII	7.4%	10.9%	16.7%	7.9%
Environmental	7.4%	6.4%	**0.0%**	**8.9%**

5 Discussion

The mixed-methods approach described in Section 3 was designed to evaluate the semantic quality of machine learning documentation, without reducing the complex sociotechnical landscape to binary compliance checks. In the analysis of the 160 models, we have seen some of the possibilities afforded by this method.

Firstly, the content analysis can extract a detailed reconstruction of the developers' conceptualisation of ethics. Our investigation of the selective reporting, defined by the stark discrepancy between high adherence to social fairness metrics (77.8% in Big Tech) and the near-total omission of environmental responsibility (0.0%), provides us with interesting detail on the interaction between such concepts.

It is clear that the design of documentation frameworks requires a careful balance: templates should be designed to encourage holistic transparency, while taking care not to lead developers into a "compliance checklist" mentality that prioritizes brand safety over invisible costs.

Secondly, the longitudinal analysis can produce a quantitative result on whether the industry is progressing towards the standards of Digital Humanism, even if we cannot evaluate such alignment directly.

The time series study found that developers exhibited significant regression in documentation completeness between the Pre-Generative Era and the GenAI Era (falling from 90% to 86%). Quantity ratings alone might suggest a robust ecosystem, but the semantic data tell us more: that the rapid expansion of Generative AI achieves its aim of rapid deployment at the cost of the documentation standards seen in earlier phases.

Our findings reflect two realities, corporate and independent; but further, they illustrate two philosophical attitudes. The major actors achieve high adoption rates for social bias sections, but at the cost of imposing a predetermined, risk-averse conceptual framework onto the documentation, excluding environmental impact.

The analysis of the "Modality Bias" observed in the lower adherence rates for Computer Vision (83.33%) compared to NLP (86.36%) suggests a reactive attitude in which the key risks are determined by the media cycle (e.g., text toxicity) rather than the inherent risks of the technology (e.g., computer vision surveillance).

5.1 Comparison with other approaches

A useful point of comparison is the original proposal by Mitchell et al. [7], involving nine distinct sections to ensure accountability. As previously discussed, this approach raises issues of selective adoption. Indeed, Bhat et al. [2] investigate the documentation practice and find technical metadata to be present while qualitative ethical sections are treated as noise.

Our findings extend this by revealing that even when ethical sections are present, they are "Ethically Cherry-Picked." The purely voluntary adherence method may therefore only be appropriate to cases in which the developer has intrinsic motivation to be transparent.

One advantage of the current voluntary method is that it is quick to administer, allowing for rapid deployment in open-source ecosystems. However, our findings provide further evidence of this "Traceability Gap," [2] where critical metrics like carbon emissions are technically measurable but semantically absent from the public documentation.

A third alternative, which is worthy of further exploration, is to gather data via automated extraction. Drawing parallels with the software engineering context (CI/CD pipelines) and the model training context suggest that methods could be adopted for automatically logging carbon metrics.

This is explained further in the work by Kothandapani et al. [8], who propose using model cards as building blocks for automated compliance checks, effectively moving away from manual regulatory reporting.

6 Conclusions

Traditional governance frameworks for artificial intelligence systems frequently lag behind rigorous standards for Digital Humanism. In this article we have presented findings based on a mixed-methods analysis of 158 models.

Our analysis aims to characterise the conceptual structures developers bring to bear in rendering a Model Card in their corporate context. We found that the content describing ethical risks presents a selective coverage of sociotechnical aspects. Specifically, we revealed that corporate actors prioritize social bias mitigation while systematically omitting ecological costs. Furthermore, we identified a "GenAI Regression," where the rush to deploy generative models has led to a prioritization of safety warnings over process transparency.

More generally, we believe that this area is underexplored and needs much more research, such as the further development of automated approaches to carbon tracking, or the application of rigorous third-party auditing to challenge the "negotiated truths" of corporate documentation.

References

[1] Matthew Arnold, Rachel Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John Richards, Kush Varshney, and Jason Tsay. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, PP:1–1, 09 2019.

[2] Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L.C. Guo. Aspirations and practice of ml model documentation: Moving the needle with nudging and traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, and Antoine Bosselut et al. On the opportunities and risks of foundation models, 2022.

[4] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64, 03 2018.

[5] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30, 03 2020.

[6] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1, 09 2019.

[7] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Raji, and Timnit Gebru. Model cards for model reporting. pages 220–229, 01 2019.

[8] Hariharan Pappil Kothandapani. Ai-driven regulatory compliance: Transforming financial oversight through large language models and automation. 3:12–24, 01 2025.

[9] Inioluwa Raji, Andrew Smart, Rebecca White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, 01 2020.

[10] Roy Schwartz, Jesse Dodge, Noah Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63:54–63, 11 2020.

[11] Emma Strubell, Ananya Ganesh, and Andrew Mccallum. Energy and policy considerations for deep learning in nlp. pages 3645–3650, 01 2019.

[12] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Hendricks, and Iason Gabriel. Ethical and social risks of harm from language models, 12 2021.

[13] H. Werthner, Allison Stanger, Viola Schiaffonati, Peter Knees, Lynda Hardman, and Carlo Ghezzi. Digital humanism: The time is now. *Computer*, 56:138–142, 01 2023.

List of Aids Used

In accordance with the course guidelines, Table 7 details the specific application of AI and software tools in the creation of this research paper.

Table 7: Declaration of AI and Software Tools Used

Tool Used	**Type of Use**	**Sections**	**Documentation (e.g., prompt strategy)**
Gemini	Code Generation	Methodology, Results	Used to generate Python scripts for data processing and LaTeX code for table visualizations.
Gemini	Stylistic Refinement	All	Used to check grammar, tone consistency ("too formal/serious"), and sentence logic throughout the manuscript.
Gemini	Research Planning	Introduction, Methodology	Used in the initial phase to map out the paper structure and validate the relevance and fit of selected references.
DeepL	Translation & Synonyms	All	Used to translate concepts from German to English and find appropriate synonyms.
diagrams.net	Visualization	Methodology	Used to manually create the workflow diagram (Figure 1) visualizing the research methodlogoy