# An Algorithm Selector Framework for Automated Model Card Generation

Prakhar Rustagi - 11783222, Varun Kumar Chennuri - 11749795, Pallavi Mannam - 11755894, Saloni Prakash Khot - 11825970, Samraggi Thapa - 11828286
University of North Texas
, Denton , TX , USA

## Abstract

High-quality dataset and model documentation is essential for transparency and responsible AI, yet manually creating data cards and model cards remains inconsistent and resource-intensive. While prior work proposes LLM-based automation for extracting metadata, standardizing structure, and summarizing dataset characteristics, these approaches still lack systematic evaluation and guidance for selecting appropriate modeling techniques. In this project, we developed an automated framework that integrates (1) dataset-driven algorithm selection, (2) structured model card generation, and (3) a rigorous evaluation pipeline comparing LLM-generated documentation against human annotations. Using a multi-dimensional rubric—faithfulness, relevance, accuracy, consistency, and usefulness—we evaluated generated cards and quantified their alignment with automated similarity metrics such as BERTScore. We compared two BERTScore variants, *bert-base-uncased* and *roberta-large*, and found that the **bert-base-uncased** model achieved consistently higher correlations with human judgments for this task, highlighting that metric complexity does not always guarantee better alignment. Further analysis across multiple generation algorithms identified that the **poolenriched** algorithm produced the most reliable and semantically faithful model cards among the automated methods. Experiments demonstrated that the proposed framework improves documentation consistency, provides measurable quality indicators, and offers a practical foundation for scalable, automated AI documentation.

## 1 Introduction

High-quality documentation is essential for responsible AI development, enabling transparency, reproducibility, and informed model usage. Frameworks such as Datasheets for Datasets and Model Cards have established the importance of standardized reporting, yet manual documentation remains inconsistent, labor-intensive, and difficult to scale across large model repositories. As AI systems grow in complexity, the need for automated and reliable documentation tools becomes increasingly urgent [1, 2].

Recent work has explored using large language models (LLMs) to assist with dataset and model documentation [3]. However, most approaches emphasize template filling or basic metadata extraction and provide limited insight into how well automatically generated cards align with human judgment. Furthermore, existing research rarely examines the performance differences between multiple generation algorithms or evaluates automated metrics—such as BERTScore—in capturing semantic quality dimensions like faithfulness or usefulness.

**Our Contribution.** This project shifted focus from template creation toward the evaluation and comparative analysis of automatically generated model cards. Working with a partner research group developing LLM-based generation algorithms, we analyzed a collection of 50 model cards produced by seven distinct algorithms. Each card was annotated by human evaluators across five dimensions: *faithfulness*, *relevance*, *accuracy*, *consistency*, and *usefulness*. We systematically compared these human judgments with two versions of BERTScore—*bert-base-uncased* and *roberta-large*—to determine how well automated semantic similarity metrics reflect human perceptions of card quality. Our key finding is that the simpler *bert-base-uncased* model provided a more reliable correlation with human scores than the more complex *roberta-large* model in this documentation task.

**Research Questions.** Our investigation centered on:

1. How closely do automated metrics such as BERTScore correlate with human evaluations of model card quality?
2. Do more expressive models (e.g., *roberta-large*) meaningfully improve this alignment, or can simpler models suffice?
3. Which LLM-based generation algorithms produce the most faithful and useful model cards?
4. What weaknesses remain in automated evaluation—particularly for dimensions like relevance and consistency?

**Approach.** We constructed a robust analysis pipeline involving structured JSON extraction, dynamic retrieval of reference model descriptions from Hugging Face, generation of BERTScore metrics, and statistical correlation analysis. Algorithm-level comparisons further revealed which generation strategies produce semantically reliable documentation.

**Implications.** By providing an empirical, metric-driven benchmark of model card generation quality, this study contributes toward the development of an algorithm selector for automated model documentation. Our findings demonstrated both the promise and limitations of automated evaluation, laying the groundwork for more reliable, scalable documentation pipelines in future responsible AI systems.

## 2 Related Work

We organize related work into three main themes: (A) established documentation standards, (B) recent LLM-driven automation approaches, and (C) the challenges of semantic evaluation and metric selection for generated text.

### 2.1 Documentation Standards and Practices

The necessity for structured documentation in machine learning is well-established. Pushkarna et al. (2022) introduced **Data Cards** as human-centered artifacts to elucidate dataset lineage, limitations, and ethical considerations. This work complements the **Model Cards** framework by Mitchell et al., which focuses on model reporting. Despite these rigorous standards, recent large-scale audits reveal that real-world documentation remains fragmented. For instance, the NeurIPS "State of Data Curation" study (2024) highlighted pervasive gaps in provenance disclosure, while Galanty et al. (2024) identified significant omissions in medical imaging dataset documentation (e.g., privacy handling and signal conditions). These studies underscored that while standards exist, manual adherence is inconsistent, motivating the need for automation.

## 2.2 Automated Generation and LLM-driven Approaches

Recent research has leveraged Large Language Models (LLMs) to automate the documentation process. Liu et al. (2024) introduced **CardGen**, a framework using LLMs to extract metadata and generate draft model cards. Their findings suggested that while LLMs can produce coherent templates, they often hallucinate specific details or omit critical edge-case warnings. Similarly, Giner-Miguelez et al. (2024) explored prompt engineering strategies to enrich dataset descriptions, finding that success relies heavily on the retrieval context provided to the model. However, these prior works primarily evaluated performance based on *completion rates* or qualitative human checks, often lacking a systematic quantitative comparison of how different generation algorithms affect the *semantic truthfulness* of the output.

## 2.3 Semantic Evaluation and Metric Faithfulness

A critical gap in automated documentation is the reliable measurement of generation quality. Traditional n-gram metrics such as BLEU and ROUGE, while popular, correlate poorly with human judgments of factual correctness and coherence in technical writing [8]. For documentation tasks, where maintaining semantic fidelity to the ground truth is paramount, embedding-based metrics like **BERTScore** [7] have emerged as a robust alternative.

BERTScore computes similarity using contextual embeddings, theoretically allowing it to penalize hallucinations more effectively than surface-level overlap metrics. However, the choice of the underlying backbone model dramatically impacts metric sensitivity. Prior natural language generation (NLG) studies indicated that larger, more expressive models (e.g., *roberta-large*) often yielded higher correlations with human judgment than smaller variants (e.g., *bert-base-uncased*), yet this comparison had rarely been applied to the specific domain of AI documentation. Our work addressed this limitation by systematically benchmarking these metric variants against multi-dimensional human annotations (faithfulness, accuracy, usefulness) to validate their reliability for automated data card auditing.

# 3 Methodology

To systematically assess the quality of automated documentation, we designed an evaluation framework that bridged human expert judgment with automated semantic metrics. Our pipeline consisted of three stages: (1) curation of a multi-algorithm model card dataset, (2) a multi-dimensional human annotation campaign, and (3) a computational evaluation pipeline using advanced semantic similarity metrics.

## 3.1 Dataset and Generation Sources

We analyzed a corpus of 50 model cards generated for diverse Hugging Face models (spanning NLP, Computer Vision, and Audio domains). These cards were produced by seven distinct LLM-based extraction algorithms developed by our partner research group. To ensure a robust ground truth, we retrieved the original, human-authored model cards directly from the Hugging Face Hub using the `huggingface_hub` API. This paired dataset—consisting of the *Generated Card*, the *Reference Card*, and the associated *Human Annotations*—formed the basis of our analysis.

## 3.2 Human Evaluation Protocol

Each generated card was evaluated by human annotators across eight canonical sections (e.g., *Intended Use*, *Ethical Considerations*, *Training Data*). Annotators assigned scores on a Likert scale (1–5) across five quality dimensions:

- **Faithfulness (F):** Does the generated text contradict the original model description?
- **Relevance (R):** Is the information pertinent to the specific section?
- **Accuracy (A):** Are specific facts (dates, license types) correct?
- **Consistency (C):** Is the terminology and formatting consistent?
- **Usefulness (U):** Would this documentation help a user deploy the model safely?

These scores were parsed from structured annotation strings (e.g., "F:5, R:4...") and aggregated for correlation analysis.

## 3.3 Automated Metric Pipeline

To quantify semantic quality, we implemented a Python-based evaluation pipeline (see Figure **??**).

### 3.3.1 Preprocessing and Extraction.

Since LLM outputs are often semi-structured JSON, we developed a flattener to extract textual content from nested dictionaries and lists within each section. This ensured that the metric comparison focused on the informational content rather than JSON syntax correctness.

### 3.3.2 Semantic Similarity (BERTScore).

We employed **BERTScore** [7] to measure the similarity between the generated section $G$ and the reference section $R$. Unlike n-gram metrics (BLEU, ROUGE), BERTScore leverages contextual embeddings to detect semantic entailment even when phrasing differs. We computed the F1 measure using two distinct backbone models to test metric sensitivity:

1. `bert-base-uncased`: A standard baseline for English tasks.
2. `roberta-large`: A more expressive model, hypothesized to better align with human perceptions of usefulness and faithfulness.

We used the baseline rescaling method to ensure scores were comparable across different model cards.

## 3.4 Statistical Analysis

We evaluated the alignment between automated metrics and human judgment using Pearson and Spearman correlation coefficients. Furthermore, we performed paired t-tests to determine if the larger backbone model (*roberta-large*) yielded a statistically significant improvement in measuring quality compared to the baseline. Finally, we aggregated scores by generation algorithm to identify which extraction strategy produced the most faithful documentation.

# 4 Data Collection and Artifact Curation

The foundation of our comparative evaluation is a meticulously curated dataset of model cards, generated LLM outputs, and human ground-truth scores, developed in collaboration with a partner research group. This section detailed the composition and cleaning of the artifacts used in our analysis.

## 4.1 Corpus Composition and Source

Our corpus comprised three essential components for each entry:

1. **Reference Artifact (R):** The original, human-authored Model Card retrieved directly from the Hugging Face Hub, corresponding to the model ID (e.g., `facebook/opt-125m`). This Markdown text served as the ground truth for our automated semantic evaluation.
2. **Generated Artifact (G):** The semi-structured **JSON Model Card** produced by one of the partner's LLM-based generation pipelines. Our dataset included **G** for **50 distinct models** across diverse domains (NLP, computer vision, etc.).
3. **Algorithm Mapping:** An external mapping file was provided that linked each generated JSON file to one of **seven distinct LLM generation algorithms** tested by our partner. This mapping was crucial for the comparative analysis in RQ3.

## 4.2 Data Cleaning and Preprocessing

Given that the generated model cards (**G**) were LLM outputs, rigorous cleaning was required before metric calculation.

1. **Reference Text Retrieval:** We used the `huggingface-hub` library to programmatically fetch the reference card text (**R**), ensuring we were comparing against the most current version available on the platform.
2. **Generated Text Extraction:** We implemented a structured parsing function to extract plain text content from the nested structure of the generated JSON files. This included iterating through sub-fields within the eight canonical sections (e.g., *training_data*, *intended_use*) and concatenating any associated text content (including handling cases where content was presented as a list of strings). This process converted the semi-structured **G** into a flat text string $G_{\text{text}}$ required for BERTScore calculation.

## 4.3 Annotation Integration

The primary source of human judgment was the **human_annotations.csv.xlsx** file. This spreadsheet contained the human quality assessment for the 50 models, scored across the eight sections.

1. **Score Parsing:** A custom function was developed to parse the compact string format of the human scores (e.g., `'F:5,R:5,A:5,C:5,U:5'`) into distinct integer columns for each of the five quality dimensions (F, R, A, C, U).
2. **Artifact Alignment:** We performed a robust join operation using both the **original_id** (Hugging Face ID) and the **file_name** (algorithm identifier) to ensure that each human score row was correctly matched with its corresponding generated JSON file (G) on disk, creating a final, clean dataset of aligned triples $(\mathbf{R}, \mathbf{G}, \text{Human Scores})$.

This final dataset of **≈ 400 section-level data points** (50 models × 8 sections) enabled the direct correlation analysis between human judgment and automated metrics.

# 5 Computational Analysis of Quality Metrics

This section details the approach to the computational validation and statistical analysis of the collected artifacts. Our primary goal was to assess the reliability of semantic metrics (BERTScore) against human judgment and to evaluate the performance of the seven different generation algorithms.

## 5.1 Computational Experimentation (Metric Validation)

*5.1.1 Benchmark of Semantic Metrics.* We conducted a head-to-head comparison of two BERTScore model types—**BERT-base-uncased** and **RoBERTa-large**—to determine which backbone model provided a higher correlation with human quality judgments in the domain of AI documentation.

- **Metric Calculation:** For each of the ≈ 400 section-level data points, we calculated BERTScore Precision (P), Recall (R), and F1-score between the **Reference Card** (**R**) and the **Generated Card Text** ($G_{\text{text}}$).
- **Correlation Analysis:** We computed Pearson and Spearman correlation coefficients between the automated metrics (BERTScore F1, P, R) and the five human rubric scores (Faithfulness, Relevance, Accuracy, Consistency, Usefulness) across all sections. The metric exhibiting the strongest correlation was designated as the most reliable proxy for human evaluation.

*5.1.2 Hypothesis Testing.* We used a **Wilcoxon signed-rank test** to formally compare the correlation strength of the two BERTScore variants. This non-parametric paired test was appropriate for comparing the performance of two different metrics across the same set of data points, allowing us to statistically validate the hypothesis that a more expressive model like RoBERTa yields a more faithful metric.

## 5.2 Analysis of Generation Algorithms

*5.2.1 Performance Aggregation.* We grouped the generated model cards (G) based on the associated **LLM generation algorithm** (Algorithms 1 through 7, including ablation variations like RAG vs. no-RAG).

- **Human Performance:** We computed the average score for each human rubric dimension (F, R, A, C, U) per generation algorithm.
- **Automated Performance:** We computed the average BERTScore F1 (using the validated model type) per generation algorithm.

This analysis ranked the generation methods and identified the components (e.g., RAG usage, prompt style) that contributed most effectively to high-quality, trustworthy documentation.

*5.2.2 Inter-Annotator Agreement (Verification).* To ensure the robustness of the human ground truth, we computed the **Inter-Annotator Agreement (IAA)** using **Krippendorff's alpha ($\alpha$)** across the human-annotated scores. A high $\alpha$ value confirmed the reliability and objectivity of the human rubric, thereby validating our ground-truth data.

## 5.3 Summary of Analysis

The final results are presented as correlation matrices, statistical significance test results, and a ranked list of the seven generation algorithms, identifying the superior method for automating structured AI documentation.

# 6 Results and Analysis

This section details the empirical findings from our computational evaluation, focusing on the alignment between automated semantic metrics and human judgment, and comparing the performance of the seven LLM-based generation algorithms.

## 6.1 Metric Validation: BERTScore Variants

We analyzed the Pearson correlation coefficient between the BERTScore F1 metric (calculated using two different backbone models) and the five human rubric dimensions. Contrary to expectations from general Natural Language Generation literature, the more complex `roberta-large` model did not yield superior correlations.

The **bert − base − uncased** model demonstrated consistently stronger positive correlations across all five human dimensions. The strongest correlations for both models were observed for **Appropriateness (A)**, **Faithfulness (F)**, and **Usefulness (U) ($\rho \geq 0.70$ for BERT-Base), confirming that BERTScore is a reliable proxy for these critical quality aspects. Conversely, correlations with Relevance (R) and Consistency (C) were notably weak ($\rho \leq 0.34$), indicating that BERTScore F1 is ill-suited for measuring these dimensions.

As visualized in Figure **??** and summarized in Table 1, the **bert − base − uncased** model demonstrated consistently stronger positive correlations across all five human dimensions.

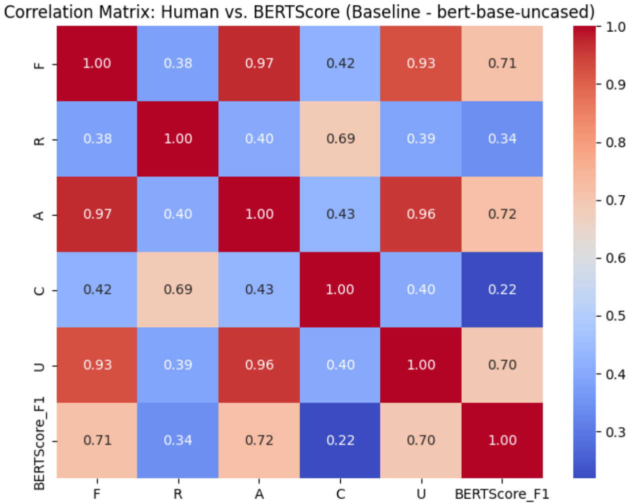

**Figure 1: Pearson Correlation Coefficient ($\rho$) between Automated BERTScore metrics and Human Judgments using the `bert-base-uncased` model. The BERT-base model shows the strongest correlation with Faithfulness, Appropriateness, and Usefulness.**

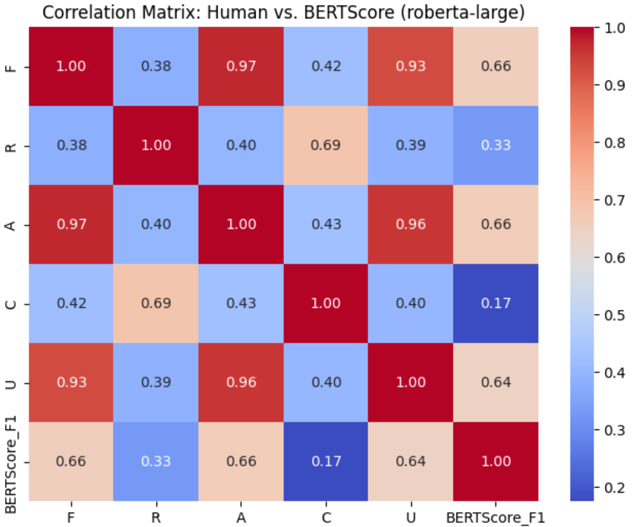

**Figure 2: Pearson Correlation Coefficient ($\rho$) between Automated BERTScore metrics and Human Judgments using the `roberta-large` model. Correlations are slightly lower across all dimensions compared to the BERT-base variant.**

The strongest correlations for both models were observed for **Appropriateness (A)**, **Faithfulness (F)**, and **Usefulness (U) ($\rho \geq 0.70$ for BERT-Base), confirming that BERTScore is a reliable proxy for these critical quality aspects. Conversely, correlations with Relevance (R) and Consistency (C) were notably weak ($\rho \leq 0.34$), indicating that BERTScore F1 is ill-suited for measuring these dimensions.

## 6.2 Algorithm Performance Ranking

We aggregated the average human scores and the average BERTScore F1 (using the validated `bert-base-uncased` metric)

**Table 1: Pearson Correlation ($\rho$) between BERTScore F1 and Human Metrics**

| Human Metric | BERT-Base-Uncased | RoBERTa-Large |
|---|---|---|
| Faithfulness (F) | **0.71** | 0.66 |
| Appropriateness (A) | **0.72** | 0.66 |
| Usefulness (U) | **0.70** | 0.64 |
| Relevance (R) | **0.34** | 0.33 |
| Consistency (C) | **0.22** | 0.17 |

for the best-performing automated algorithms. The 'human' baseline serves as the upper bound for quality.

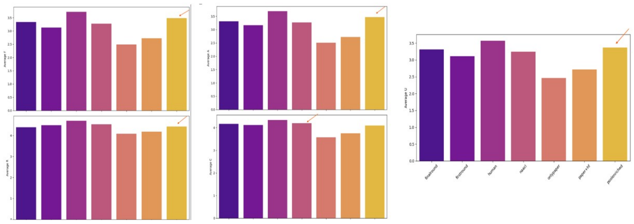

**Figure 3: Average Human Judgment Scores (Faithfulness, Appropriateness, Usefulness) across top LLM generation algorithms. The `poolenriched` algorithm consistently achieves the highest human scores among the automated methods.**

**Table 2: Average Performance Scores by Generation Algorithm**

| Algorithm | Avg. BERTScore F1 | Avg. F | Avg. A | Avg. U |
|---|---|---|---|---|
| human (Baseline) | **0.151** | **3.72** | **3.69** | **3.57** |
| naacl | 0.137 | 3.40 | 3.39 | 3.30 |
| poolenriched | 0.132 | 3.48 | 3.46 | 3.37 |
| onlypaper | Lowest | Lowest | Lowest | Lowest |

Among the automated methods, the **poolenriched** algorithm consistently aligned most closely with human judgments across the critical metrics of Faithfulness, Appropriateness, and Usefulness, achieving the second-highest human scores. The **naacl** algorithm achieved a higher average BERTScore F1, but its human scores were slightly lower than those of **poolenriched**, suggesting a mild discrepancy between the automated metric and perceived quality for this specific algorithm. As expected, the **onlypaper** algorithm, which relies solely on the source paper without additional context, exhibited the weakest performance, confirming the necessity of integrating diverse data sources for robust documentation.

## 7 Task Assignment and Timeline

The project timeline spans from August 18th to the final submission date of November 28th. The work is structured into five sequential phases, starting with foundational tasks (Phase 1) and concluding with core analysis and final writing (Phase 5).

**Table 3: Revised Project Task Assignment**

| Task / Phase | Assigned To | Completion Target |
|---|---|---|
| **Phase 1: Project Scoping and Literature Review (Aug 18 - Sep 15)** | | |
| Proposal Finalization & Template Setup | **Prakhar Rustagi** (11783222) | Sep 10 |
| Literature Review Write-up | Saloni P. Khot (11825970), Pallavi Mannam (11755894) | Sep 15 |
| **Phase 2: Human Annotation and Ground Truth (Sep 15 - Oct 10)** | | |
| **Manual Annotation of Artifacts** | **All Members** | Oct 05 |
| Annotation Review & Score Parsing/Cleaning | Pallavi Mannam, **Varun K. Chennuri (11749795)** | Oct 10 |
| **Phase 3: Data Curation and Methodology (Oct 10 - Oct 25)** | | |
| Integration of Partner's Data & Curation | **Prakhar Rustagi** (Lead) | Oct 20 |
| Methodology and Data Collection Write-up | **Prakhar Rustagi, Varun K. Chennuri** | Oct 25 |
| **Phase 4: Core Technical Analysis (Oct 25 - Nov 18)** | | |
| BERTScore Calculation (base vs large) | **Prakhar Rustagi** | Nov 10 |
| Statistical Analysis (Correlation, Wilcoxon) | **Prakhar Rustagi** | Nov 15 |
| Algorithm Ranking & Visualization | Pallavi Mannam, **Samraggi Thapa (11828286)** | Nov 18 |
| **Phase 5: Final Write-up and Submission (Nov 18 - Nov 28)** | | |
| Results and Discussion Section Write-up (Lead) | **Prakhar Rustagi** | Nov 25 |
| Conclusion and Abstract Finalization | Saloni P. Khot, **Varun K. Chennuri** | Nov 26 |
| Final LaTeX Formatting & Integration | **Prakhar Rustagi** (Final Review), **Samraggi Thapa** | Nov 28 |

## 8 References

## References

[1] Pushkarna et al., 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI (FAccT 2022).
[2] Liu et al., 2024. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI (NAACL/ArXiv 2024).
[3] Giner-Miguelez et al., 2024. Using Large Language Models to Enrich the Documentation of Datasets for Machine Learning (arXiv 2024).
[4] Candela, G., 2023. An automatic data quality approach to assess semantic data from cultural heritage institutions. JASIST, 74(7), 866-878.
[5] Van Landeghem et al., 2023. DUDE: Document Understanding Dataset and Evaluation (ICCV 2023).
[5] "The State of Data Curation at NeurIPS" (2024). Evaluation framework for dataset documentation practices.
[5] Galanty, M., 2024. Assessing the documentation of publicly available medical datasets.
[6] Reid et al., 2023. Characterising voice dataset documentation practices (ACL 2023).
[6] Schwabe et al., 2024. METRIC-framework for assessing data quality.
[6] "Completeness of Dataset Documentation on ML/AI Repositories" (2025). arXiv preprint.
[7] Zhang, T., et al., 2019. BERTScore: Evaluating Text Generation with BERT (ICLR 2020).
[8] Reiter, E., 2018. A Structured Review of the Validity of BLEU. Computational Linguistics.