AdaQE-CG: Adaptive Query Expansion for Web-Scale Generative AI Model and Data Card Generation

Haoxuan Zhang
haoxuanzhang@my.unt.edu
University of North Texas
Denton, TX, USA

Ruochi Li
rli14@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Zhenni Liang
zhenniliang@my.unt.edu
University of North Texas
Denton, TX, USA

Mehri Sattari
mehrisattari@my.unt.edu
University of North Texas
Denton, TX, USA

Phat Vo
phatvo@my.unt.edu
University of North Texas
Denton, TX, USA

Collin Qu
collinqu@gmail.com
Bellevue High School
Bellevue, WA, USA

Ting Xiao
ting.xiao@unt.edu
University of North Texas
Denton, TX, USA

Junhua Ding
junhua.ding@unt.edu
University of North Texas
Denton, TX, USA

Yang Zhang*
yang.zhang@unt.edu
University of North Texas
Denton, TX, USA

Haihua Chen*
haihua.chen@unt.edu
University of North Texas
Denton, TX, USA

Abstract

Transparent and standardized documentation is essential for building trustworthy generative AI (GAI) systems. However, current automated model and data card generation methods still face three key challenges: **(i) Static templates.** Most systems rely on fixed query templates that cannot adapt to diverse paper structures or evolving documentation requirements. **(ii) Information scarcity.** Web-scale repositories such as Hugging Face often provide incomplete or inconsistent metadata, resulting in missing or noisy information. **(iii) Lack of benchmarks.** The absence of standardized datasets and evaluation protocols prevents fair and reproducible assessment of documentation quality. To address these challenges, we propose **AdaQE-CG**, an *Adaptive Query Expansion for Card Generation* framework that integrates dynamic information extraction with cross-card knowledge transfer. The **Intra-Paper Extraction via Context-Aware Query Expansion (IPE-QE)** module iteratively refines extraction queries to capture richer and more complete information from scientific papers and repositories. The **Inter-Card Completion using the MetaGAI Pool (ICC-MP)** module enriches missing fields by transferring semantically relevant content from similar cards within a curated dataset. In addition, we construct **MetaGAI-Bench**, the first large-scale, expert-annotated benchmark for evaluating GAI documentation. Comprehensive experiments across five quality dimensions demonstrate that AdaQE-CG significantly outperforms existing approaches, surpasses human-authored data cards, and approaches human-level quality for model cards. Code, prompts, and data are publicly available at: https://github.com/haoxuan-unt2024/AdaQE-CG.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Natural language processing**; • **Applied computing** → **Document management and text processing**.

Keywords

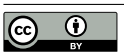
Web-scale Document, Generative AI, Model Card, Data Card, Large Language Model, Query Expansion

ACM Reference Format:
Haoxuan Zhang, Ruochi Li, Zhenni Liang, Mehri Sattari, Phat Vo, Collin Qu, Ting Xiao, Junhua Ding, Yang Zhang, and Haihua Chen. 2026. AdaQE-CG: Adaptive Query Expansion for Web-Scale Generative AI Model and Data Card Generation. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792685

1 Introduction

As artificial intelligence (AI) systems proliferate across the web at an unprecedented scale, the need for transparent and standardized documentation has become increasingly critical. Model and data cards, which are fundamentally web-based data artifacts, have emerged as key transparency mechanisms that provide structured, machine-readable documentation for models and datasets underpinning AI systems [30, 36]. In the era of generative AI (GAI), where large language models (LLMs), multimodal architectures, and web-scale GAI platforms exhibit complex behaviors and opaque data dependencies, these web-native documentation frameworks are

*Corresponding authors.



WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792685

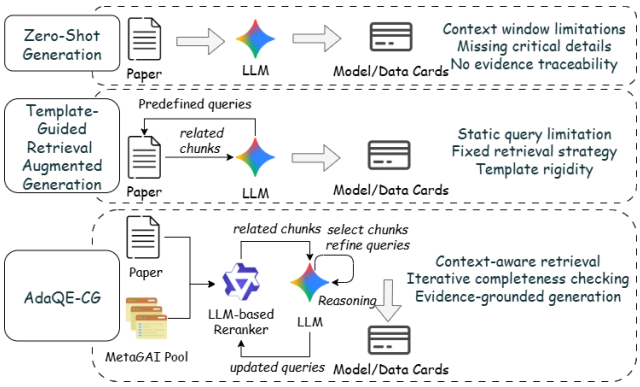


Figure 1: A comparison of card generation paradigms: (a) Zero-Shot Generation (top), (b) Template-Guided RAG (middle), and (c) AdaQE-CG (bottom).

indispensable for conveying information about training corpora, performance characteristics, and potential biases [16, 29, 41].

Beyond promoting transparency, model and data cards constitute a critical layer of web infrastructure for enabling responsible, reproducible, and scalable GAI development—particularly in high-stakes domains such as healthcare, finance, and law, where provenance, fairness, and ethical accountability are essential [14, 37]. As web-scale GAI ecosystems continue to expand, characterized by emergent model behaviors, intricate data supply chains, and dynamic deployment contexts, model and data cards provide the minimal yet indispensable accountability substrate that supports development, evaluation, governance, and interoperability across the global AI landscape.

Model and data card generation also forms the foundation for a broad spectrum of critical applications that advance transparency, accountability, and governance in AI systems. These web-based documentation artifacts enable longitudinal tracking of model and dataset evolution, supporting systematic analyses of changes in architecture, data composition, and performance over time [23, 25, 46]. They also power public repositories and registries that facilitate cross-model and cross-dataset comparisons of capabilities, biases, licenses, and usage constraints, thereby promoting interoperability and reproducibility within the broader AI ecosystem [5, 17, 44]. Moreover, model and data cards are instrumental in benchmarking and governance, providing structured metadata for automated compliance auditing against ethical principles, data consent requirements, and security standards [28, 35]. Beyond compliance, they serve as key enablers of risk and fairness auditing, helping researchers and practitioners identify, quantify, and mitigate potential harms in AI deployment contexts [27]. Collectively, these applications highlight the indispensable role of model and data cards as the connective infrastructure that bridges AI research, practice, and policy across web-scale generative systems.

However, model and data card generation is a non-trivial task. Manual card creation faces significant scalability challenges, as current human-generated documentation suffers from inconsistencies, incompleteness, and a heavy reliance on developers' subjective interpretations of what should be reported [25, 26, 46]. Automatic generation of model and data cards promises consistency and scalability, but faces several challenges, as shown in Figure 1. Zero-shot generation synthesizes complete documentation in a single inference pass but encounters three critical deficiencies: context window constraints that truncate lengthy papers, missing critical details due to a lack of systematic extraction, and the absence of evidence traceability that prevents source verification. Template-guided RAG methods [26] employ predefined question templates yet face distinct limitations: static query constraints preventing adaptation to diverse paper structures, fixed retrieval strategies failing to accommodate domain-specific patterns, and template rigidity creating schema mismatches with evolving documentation requirements. Beyond paradigm-specific constraints, both approaches confront shared challenges: (1) the absence of ground truth makes it difficult to objectively assess completeness and correctness, (2) algorithms that rely solely on information from academic papers or web-based data sources such as Hugging Face or GitHub rarely achieve optimal performance due to information scarcity in these sources, (3) the risk of hallucinations when LLMs summarize lengthy documentation beyond their context windows, and (4) inconsistent templates and schema variants across data sources.

To address these limitations, we propose **AdaQE-CG**, an **Ada**ptive **Q**uery **E**xpansion for **C**ard **G**eneration framework that tackles both static query constraints and information scarcity. The first module, Intra-Paper Extraction via Context-Aware Query Expansion (IPE-QE) dynamically adapts extraction strategies by iteratively refining queries based on identified information gaps, overcoming static template limitations. The second module, Inter-Card Completion using the MetaGAI Pool (ICC-MP) enriches incomplete fields by transferring knowledge from architecturally and semantically similar cards in the curated MetaGAI Pool. This hybrid approach balances automation efficiency with documentation quality while maintaining provenance transparency.

In a nutshell, our contributions are summarized as follows:

- **MetaGAI: Web-Scale Dataset and High-Quality Benchmark.** We construct the MetaGAI-Dataset of 6,481 data cards and 123,013 model cards across four GAI modalities, and establish a high-quality MetaGAI-Bench of 1,200 expert-annotated cards with strong inter-annotator agreement.
- **Comprehensive Empirical Statistical Analysis.** We introduce the Weighted Card Completeness Index (WCCI) to quantify documentation quality. Our analysis reveals that data cards significantly outperform model cards, with critical gaps in responsible AI fields. We identify strong correlations between completeness and popularity through systematic correlation analysis.
- **AdaQE-CG: Adaptive Query Expansion Framework.** We propose a novel hybrid-module architecture combining IPE-QE for dynamic extraction and ICC-MP for cross-card knowledge transfer, and validate its effectiveness through comprehensive ablation studies.
- **Rigorous Evaluation Framework and Empirical Validation.** We employ a multi-perspective assessment combining LLM-as-a-Judge and human evaluation across five quality dimensions (*Faithfulness, Relevance, Accuracy, Consistency, Usefulness*). Our AdaQE-CG framework outperforms

the human baseline for data cards and achieves human-level performance for model cards in LLM evaluation. In human evaluation, our framework consistently ranks second among all methods.

2 Related Work

Web-scale semantic integration transforms heterogeneous documentation into standardized, machine-interpretable knowledge. For example, an analysis of 2 million models on Hugging Face (a large web repository) revealed declining documentation quality as model cards become templated and auto-generated [23]. Horwitz et al. [19] further mapped over 400,000 model relationships to infer undocumented attributes, underscoring the need for scalable semantic synthesis. To improve coverage and consistency, Liu et al. [26] proposed CardGen, building a 4.8k model and 1.4k data card corpus. Beyond repositories, large-scale efforts such as AutoSchemaKG [3], TEXT2DB [21], structured scientific extraction using LLMs [11], and retrieval-augmented generation for multi-document and multimodal reasoning [26, 45], further demonstrate the trend toward progressive web-scale synthesis of structured knowledge.

Model and data card generation is also critical for responsible AI as accountability and traceability become major concerns in web-scale GAI applications [22, 31, 33]. A lack of visibility into how models are trained, evaluated, and deployed often obscures their limitations and societal risks, underscoring the need for transparent documentation [12, 20, 32]. For models, Model Cards [30] were proposed to disclose evaluations across demographics and usage conditions, and subsequent research expanded their scope with explainability principles [34], consumer-style labels [40], complementary card families [1, 42], and toolkits for tracking and reporting model information [2]. More recently, this line of work has evolved toward regulatory compliance and system-level governance [7], as well as end-to-end transparency and trust [43]. For data, Datasheets [13], Data Statements [4], and Data Nutrition Labels [18] set baseline practices for recording provenance, consent, and fitness-for-use. These foundations were succeeded by templated Data Cards and accompanying guidance for documenting crowd-sourced data [36]. Building on this line of work, recent studies have explored automated generation of dataset documentation [26] and the design of machine-readable open datasheets [38, 39] to improve completeness, consistency, and discoverability in large-scale repositories. Despite these advances, prior methods remain largely static, relying on predefined templates or constrained multi-source extraction.

3 Preliminary

Model cards and data cards provide structured, machine-interpretable documentation for AI models and datasets. Following Mitchell et al. [30] and Pushkarna et al. [36], we define a semantic card as $C = \{f_1, f_2, \ldots, f_m\}$ where m is the total number of fields, and each field $f_i = (k_i, v_i)$ comprises (1) a field name k_i from the card taxonomy and (2) a field value v_i containing natural language descriptions or structured data.

Given document chunks $\mathcal{D} = \{c_1, c_2, \ldots, c_n\}$ from a scientific paper and its web repository page (where n is the total number of chunks), together with a **MetaGAI Pool** $\mathcal{M} = \{C_1, C_2, \ldots, C_N\}$, which is a curated collection of N high-quality cards filtered from the MetaGAI-Dataset by completeness (τ_{wcci}) and popularity (τ_{pop}) criteria, the goal is to generate a complete card C by extracting and synthesizing information across these sources.

Task Definition: The automated card generation task is formulated as two sequential modules:

Module 1: Intra-Paper Extraction via Context-Aware Query Expansion (IPE-QE). For each target field f_{id}, iteratively refine queries q_r (round r) to extract information from chunks $\mathcal{D}$, terminating when information gain Δ_r falls below the threshold ϵ or completeness is reached. This produces an initial card C':

$$C'[f_{id}] = A_{r^*}, \quad r^* = \min\{r : \Delta_r \leq \epsilon \vee \text{IsComplete}(A_r)\} \tag{1}$$

where A_r represents the generated answer for round r, computed as $A_r = \text{LLM}(q_r, \text{Retrieve}(\mathcal{D}, q_r))$, and q_r is generated based on previous answers and query history $\mathcal{Q}$.

Module 2: Inter-Card Completion using the MetaGAI Pool (ICC-MP). Given the initial card C' from Module 1, identify incomplete fields $\mathcal{F}_{inc} = \{f_i \in C' : v_i = \emptyset\}$ and retrieve similar cards from the MetaGAI Pool $\mathcal{M}$ via TF-IDF (Term Frequency–Inverse Document Frequency) tag matching (threshold α) and semantic reranking (top-k). The enriched card C is produced by:

$$C[k_i] = \text{Synthesize}(\mathcal{V}_i, C'), \quad \mathcal{V}_i = \{v_j : C_j[k_i] \in \mathcal{M}_{sim}\} \tag{2}$$

where $\mathcal{M}_{sim} \subseteq \mathcal{M}$ is the set of top-k most relevant cards, $\mathcal{V}_i$ represents the collection of field values for field name k_i from similar cards, and C represents the final enriched card. This module is applied only when $\mathcal{F}_{inc} \neq \emptyset$ and similar cards exist in $\mathcal{M}$; otherwise, $C = C'$. Detailed algorithms for both modules are presented in Section 5.

4 MetaGAI Dataset and Benchmark

4.1 MetaGAI-Dataset Construction

We construct the MetaGAI-Dataset from Hugging Face[1], a prominent web-based platform hosting over 2 million open-source AI models and datasets. While the platform provides highly credible human-authored descriptions, these resources exhibit substantial heterogeneity and data sparsity, which pose significant challenges for semantic web applications. The descriptions lack standardized schemas and machine-interpretable semantics, and numerous fields remain unfilled, impeding automated knowledge integration.

To address these challenges, we implement a three-stage semantic data pipeline. First, we leverage the Hugging Face Hub API[2] to systematically retrieve models and datasets across four GAI modalities: Multimodal, Computer Vision, Natural Language Processing, and Audio. Second, we establish publication provenance by filtering resources with associated scientific papers through metadata tags containing DOI or arXiv identifiers, enabling paper-to-card semantic alignment. Third, we employ Gemini-2.5 Flash-Lite [10] to transform unstructured web descriptions into machine-interpretable JSON-formatted semantic cards following our standardized taxonomy (Table 4 in Appendix A.1). Our semantic annotation process generates structured cards where each field is augmented with confidence scores $\{0.25, 0.5, 0.75, 1.0\}$ to support quality assessment

[1]Hugging Face: https://huggingface.co
[2]Hugging Face Hub API: https://huggingface.co/docs/huggingface_hub/

and downstream completeness analysis. The resulting MetaGAI-Dataset comprises 6,481 data cards and 123,013 model cards with machine-interpretable semantics, collected on August 20, 2025.

4.2 MetaGAI-Bench Construction

To evaluate the performance of the card generation algorithm, we randomly sampled 600 model cards and 600 data cards from the MetaGAI-Dataset for manual annotation to establish a high-quality benchmark. Our annotation team comprised six members: two Ph.D. students in Information Science and four master's students in Data Science, all with expertise in GAI. The team was divided into three groups. Annotators reviewed original Hugging Face content, associated papers, and generated structured cards to complete their annotations.

The annotation process followed two stages. In the first stage, each group independently annotated 50 model cards and 50 data cards. We then assessed intra-group consistency to ensure annotation quality. Traditional inter-annotator agreement metrics like Cohen's kappa [9] are unsuitable for generative tasks, as identical meanings can be expressed through different wording without definitive correct answers. Therefore, we employed BERTScore [47] to calculate semantic similarity for each card field between annotators, treating annotations as consistent when semantic similarity exceeded 0.8. This approach yielded average kappa-equivalent scores of 0.939 for model cards and 0.806 for data cards, both indicating high agreement. After resolving inconsistencies within each group, the second stage began, in which each group completed the remaining annotations.

4.3 Statistics and Analysis

To quantify documentation quality across heterogeneous card structures, we propose the Weighted Card Completeness Index (WCCI), a novel metric built upon three fundamental principles: interpretability, confidence-weighted evaluation, and uniform field weighting. The detailed formula and field-level analysis by task category are presented in Appendices A.2 and A.3, respectively. To examine factors associated with card completeness, we computed Spearman correlations between WCCI scores and metadata tags, as shown in Table 5 in Appendix A.2.

Popularity metrics demonstrate significant positive associations with documentation completeness for both types of cards, with model cards exhibiting particularly strong correlations compared to data cards. These positive correlations align with prior findings that more popular artifacts on Hugging Face consistently exhibit higher documentation quality [25, 46].

Task-specific patterns reveal an important paradox. Specialized tasks such as text-to-speech ($\rho = 0.211$) and multimodal synthesis ($\rho = 0.132$) show positive correlations with WCCI scores. However, text generation models show a significant negative correlation ($\rho = -0.153$), despite being the most common task category ($n = 97{,}689$). This finding indicates that high-volume domains have lower documentation quality, creating gaps where standardized practices are most needed [25, 46]. For data cards, NLP-centric tasks (zero-shot classification: $\rho = 0.203$, text classification: $\rho = 0.201$) show moderate positive correlations, reflecting more established documentation practices in these research areas.

For model cards, infrastructure and licensing choices show strong associations with documentation completeness. Apache-2.0 license ($\rho = 0.374$) and PyTorch framework ($\rho = 0.386$) have the strongest positive correlations, suggesting that open-source ecosystems and popular frameworks promote better documentation through community standards. For data cards, crowdsourced annotation shows a strong negative correlation with documentation completeness ($\rho = -0.384$), likely reflecting coordination challenges and a lack of centralized oversight.

5 Methodology

We propose AdaQE-CG, a hybrid-module framework that implements the pipeline defined in Section 3 to generate complete semantic cards from academic papers and web repository metadata. As illustrated in Figure 2, Module 1 (IPE-QE) iteratively extracts field values v_i from document chunks $\mathcal{D}$ using dynamic query refinement and LLM-based reranking to produce the initial card C'. Module 2 (ICC-MP) is conditionally applied when $\mathcal{F}_{inc} \neq \emptyset$ and similar cards exist in the MetaGAI Pool $\mathcal{M}$, enriching incomplete fields through two-phase architectural and semantic matching to yield the final card C. This hybrid pipeline ensures deep extraction from source documents while enabling broad completion through cross-card knowledge transfer when applicable.

5.1 Module 1: Intra-Paper Extraction via Context-Aware Query Expansion

Module 1 implements the extraction process formalized in Equation 1. For each target field, we iteratively refine queries to extract field values from document chunks, terminating when information gain falls below the threshold or completeness is achieved.

5.1.1 Document Preprocessing and Chunking. To construct the document set $\mathcal{D}$, we parse scientific papers from PDF to Markdown using Nougat [6]. Each section of the paper forms an independent chunk c_i, while Hugging Face metadata constitutes a separate chunk. This section-level granularity balances information completeness with retrieval precision, ensuring each chunk $c_i \in \mathcal{D}$ contains semantically coherent content for effective extraction.

5.1.2 LLM-Based Reranking for Retrieval. We employ an LLM-based reranker [48] to identify relevant chunks for each query. This approach enables contextual understanding of semantic relevance and effective handling of complex, multifaceted queries. The reranker produces ranked document chunks and filters them to retain only relevant information for answer generation.

5.1.3 Dynamic Multi-Round Query Expansion. As formalized in Algorithm 1, for each target field with its initial query, we generate an initial answer using the top-ranked relevant chunks from the reranker, then employ two adaptive mechanisms: *context-aware query refinement* and *adaptive stopping criterion*. After each round, the LLM evaluates answer completeness and generates refined queries by analyzing information gaps in the previous answer while leveraging accumulated query history to avoid redundant extraction. The *ComputeGain* function assigns scores (0-3) based on completeness, quality, new information, and utility criteria, while *IsComplete* evaluates whether the query generator indicates completion (returns "COMPLETE"), signaling that all required information

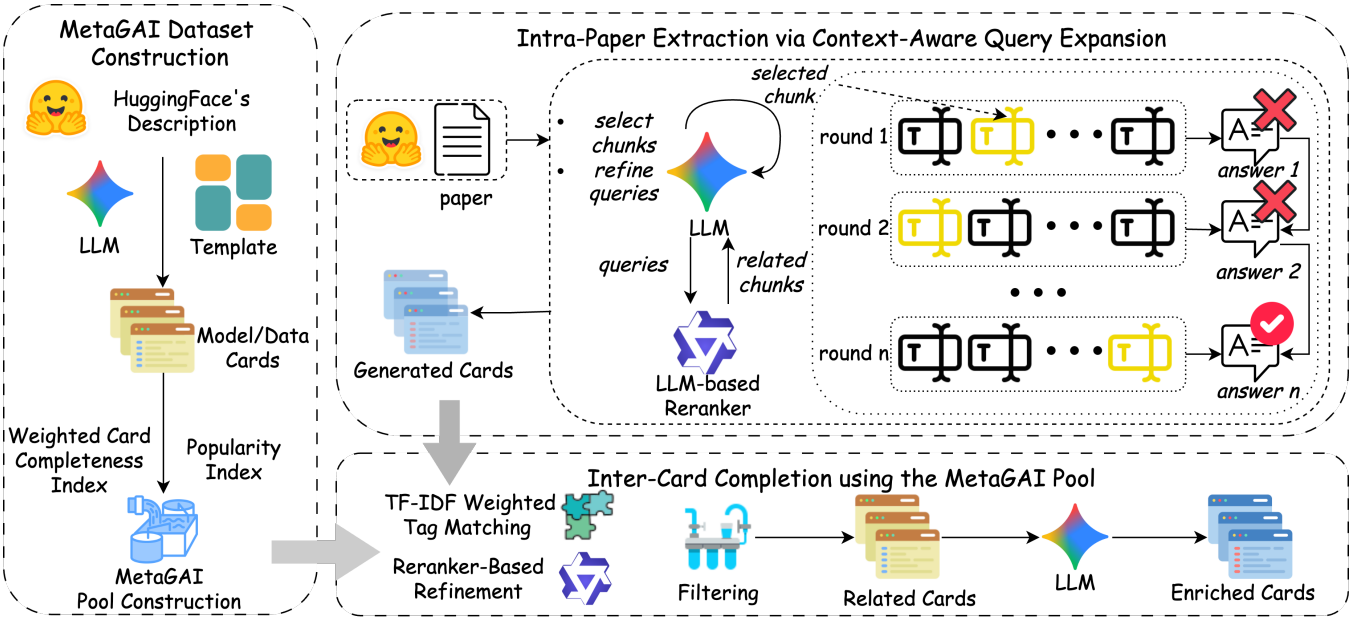


Figure 2: AdaQE-CG: Hybrid-module pipeline for automated card generation. Module 1 (IPE-QE): iteratively extracts field values from document chunks via context-aware query expansion to produce the initial card. Module 2 (ICC-MP): conditionally enriches incomplete fields using the MetaGAI Pool when similar cards are available, producing the final card.

has been extracted. The expansion terminates when: (1) the answer is complete, (2) information gain is insufficient ($\leq \epsilon$) for two consecutive rounds, or (3) maximum rounds R_{max} is reached, producing initial card C' with extracted field values for all target fields.

5.2 Module 2: Inter-Card Completion using the MetaGAI Pool

Module 2 implements the conditional enrichment process formalized in Equation 2. Given the initial card from Module 1, we identify incomplete fields and enrich them using similar cards from the MetaGAI Pool when available. If no incomplete fields exist or no similar cards are found, the initial card is returned as the final output.

The MetaGAI Pool $\mathcal{M}$ is pre-constructed from the MetaGAI-Dataset using dual quality criteria to ensure reliable knowledge sources: (1) *Completeness Filtering* retains cards meeting WCCI thresholds (top 30% for data cards and top 10% for model cards) to ensure comprehensive documentation quality; (2) *Popularity Filtering* prioritizes cards exceeding minimum download thresholds, reflecting community validation and practical utility [25, 46]. Our correlation analysis confirms this relationship in the MetaGAI-Dataset (Table 5). This dual-criterion approach yields a curated pool combining thorough documentation with proven practical value.

For each incomplete field, we employ a two-phase retrieval pipeline followed by category-aware synthesis, as formalized in Algorithm 2:

Phase 1 - TF-IDF Tag Matching: We compute TF-IDF weights for metadata tags such as task type and model architecture to identify architecturally similar cards. This structural filtering efficiently narrows the search space by retaining only cards exceeding

Algorithm 1 Intra-Paper Extraction via Context-Aware Query Expansion (IPE-QE)

Require:
$\mathcal{D} = \{c_1, c_2, \ldots, c_n\}$: Document chunks
$\{f_1, \ldots, f_m\}$: Target fields where $f_i = (k_i, v_i)$
q_0: Initial queries
R_{max}, ϵ: Maximum rounds (default: 10) and gain threshold (default: 1)
Ensure:
$C' = \{f_1, \ldots, f_m\}$: Initial card
$C' \leftarrow \emptyset$
for each field f_{id} **do**
  $Q \leftarrow \{q_0^{(id)}\}$, $r \leftarrow 0$, stall_count $\leftarrow 0$
  while $r < R_{max}$ **do**
    $A_r \leftarrow \text{LLM}(q_r, \text{RerankerRetrieve}(\mathcal{D}, q_r))$
    if IsComplete$(A_r, q_0^{(id)}, f_{id})$ **then**
      break ▷ Success: all information extracted
    end if
    if $r > 0$ **and** ComputeGain$(A_{r-1}, A_r, q_0^{(id)}, f_{id}) \leq \epsilon$ **then**
      stall_count $\leftarrow$ stall_count + 1
    else
      stall_count $\leftarrow 0$
    end if
    if stall_count ≥ 2 **then**
      break ▷ Stalled: insufficient progress
    end if
    $q_{r+1} \leftarrow \text{RefineQuery}(q_0^{(id)}, A_r, Q, f_{id})$
    $Q \leftarrow Q \cup \{q_{r+1}\}$, $r \leftarrow r + 1$
  end while
  $C'[f_{id}] \leftarrow A_r$
end for
return C'

a weighted overlap threshold α, focusing on models with similar architectures or datasets with comparable characteristics. *Phase 2 - Semantic Reranking:* An LLM-based reranker evaluates fine-grained semantic similarity between the target card and filtered candidates from Phase 1. This phase selects the top-k most semantically relevant cards, capturing nuanced similarities beyond surface-level tag matching. The two-phase design balances computational efficiency with retrieval accuracy.

Category-Aware Synthesis: We collect field values from retrieved similar cards and synthesize appropriate content based on field categorization. Fields are classified into three categories: (1) *Shared Properties* such as intended use and limitations that can be reasonably inferred from architecturally similar cards, (2) *Unique Properties* such as model developers and contact details that are artifact-specific and must never be transferred, and (3) *General Information* such as ethical considerations that are broadly applicable across similar contexts. The LLM evaluates applicability for each category and synthesizes content that integrates relevant information while preserving artifact-specific accuracy.

Algorithm 2 Inter-Card Completion using the MetaGAI Pool (ICC-MP)

Require:
$C' = \{f_1, \ldots, f_m\}$: Initial card where $f_i = (k_i, v_i)$
$\mathcal{M}$: MetaGAI Pool (pre-filtered by τ_{wcci} and τ_{pop})
k, α: Top-k parameter (default: 10) and tag overlap threshold (default: 0.5)
Ensure:
C: Final card (enriched if similar cards found, otherwise $C = C'$)
$\mathcal{F}_{inc} \leftarrow \{f_i \in C' : v_i = \emptyset\}$
if $\mathcal{F}_{inc} = \emptyset$ **then**
  return C'
end if
$C \leftarrow C'$
for each field $f_i = (k_i, v_i) \in \mathcal{F}_{inc}$ **do**
  $\mathcal{M}_{tag} \leftarrow \{C_j \in \mathcal{M} : \text{TagOverlap}(C', C_j) > \alpha\}$
  if $\mathcal{M}_{tag} = \emptyset$ **then**
    continue
  end if
  $\mathcal{M}_{sim} \leftarrow \text{RerankTopK}(C', \mathcal{M}_{tag}, k)$
  $\mathcal{V}_i \leftarrow \{C_j[k_i] : C_j \in \mathcal{M}_{sim} \wedge C_j[k_i] \neq \emptyset\}$
  if $\mathcal{V}_i \neq \emptyset$ **then**
    $C[k_i] \leftarrow \text{Synthesize}(\mathcal{V}_i, C')$
  end if
end for
return C

6 Experiments and Results

In this section, we investigate three research questions to guide our experimental evaluation:

RQ1: How can web-scale information be effectively integrated to create comprehensive model and data cards for GAI?
RQ2: How can we overcome static template limitations and leverage web-scale data to achieve high-quality card generation?
RQ3: What evaluation frameworks can effectively assess the quality of generated cards across multiple quality dimensions?

Table 1: Evaluation metrics for assessing the generated model and data cards. Each metric is scored on a 1-5 scale by domain experts.

Metric	Definition
Faithfulness (F)	Accurately reflects information from source materials without introducing unsupported claims or omitting key points
Relevance (R)	Content focused on the specific category being evaluated, avoiding unrelated or off-topic information
Accuracy (A)	Statements are factually correct based on available references and can be directly verified
Consistency (C)	Information is internally consistent within the card, with no contradictions or logical gaps
Usefulness (U)	Provides clear, practical, and helpful information for users or researchers who want to understand or use the model or dataset

6.1 Experimental Setup

6.1.1 Baselines. To evaluate the effectiveness of our method, we compare against two automated baselines and human-annotated cards as a reference standard, representing different paradigms in card generation.

Zero-Shot Generation. This baseline employs a direct synthesis approach where the LLM generates complete model cards from source papers and web repository pages in a single inference pass without retrieval augmentation. This method represents an end-to-end generation paradigm that processes entire documents holistically to produce comprehensive model documentation in one step.

Template-Guided RAG (CardGen). Following the established CardGen methodology [26], this baseline implements a structured query extraction approach using predefined question templates to systematically retrieve relevant information before generation. This method provides consistent structured extraction by employing a fixed set of queries designed to capture essential model card components across different paper types. Detailed hyperparameter configurations and implementation details for all models are provided in Appendix A.4.

Human-Annotated. A detailed description of the human annotation process is provided in Section 4.2.

6.1.2 Evaluation Metrics. To comprehensively assess the quality of generated cards, we employ a five-dimensional evaluation framework. We adopt three established metrics from Liu et al. [26], *Faithfulness*, *Relevance*, and *Accuracy*, which evaluate factual correctness and source alignment. However, these metrics cannot capture the structural coherence and practical utility requirements of multi-field documentation. Therefore, we introduce **two complementary metrics:** *Consistency* measures internal coherence across independently generated fields to detect contradictions, while *Usefulness* evaluates whether content provides actionable guidance for deployment decisions, addressing a critical gap in existing frameworks that assess correctness without considering operational value. Each metric is scored from 1 (poor) to 5 (excellent), and detailed definitions for all five metrics are provided in Table 1.

6.1.3 Evaluation Method. We employ an LLM-as-a-Judge evaluation framework to assess generated card quality. This approach provides scalable, cost-effective assessment while maintaining high agreement with human judgments [15, 24], and is particularly well-suited for evaluating open-ended generation tasks where traditional metrics fail to capture nuanced qualities [8]. To mitigate potential biases inherent in any single large language model [15], we utilize two distinct models as judges: GPT-5-nano and Gemini-2.5 Flash-Lite, and detailed evaluation procedures are provided in Appendix A.5.

6.2 Experimental Results

Table 2 presents a comprehensive evaluation of AdaQE-CG relative to two baseline methods and the human-annotated MetaGAI-Bench across five quality dimensions, evaluated independently by two LLM judges.

6.2.1 ***Effective Integration of Web-Scale Information (RQ1).*** AdaQE-CG effectively integrates heterogeneous web-scale information sources—scientific papers, repository metadata, and the MetaGAI Pool—to generate comprehensive documentation. For data cards, AdaQE-CG achieves first-place rankings across all dimensions for both judges (average scores: 4.47 GPT, 4.48 Gemini), substantially outperforming all baselines. For model cards, AdaQE-CG excels in Relevance (rank 1, scores: 4.79 GPT, 4.65 Gemini) and Usefulness (rank 1, scores: 4.02 GPT, 4.24 Gemini), with competitive performance on other metrics.

The differential performance across card types reveals the information integration challenges. Data card information follows more standardized documentation patterns, enabling systematic extraction, while model cards require nuanced interpretation of experimental details not always explicitly documented. Nevertheless, AdaQE-CG's automated approach achieves comparable or superior scores compared with human-annotated on most dimensions, validating multi-source web-scale integration.

6.2.2 ***Overcoming Static Template Limitations (RQ2).*** Comparing AdaQE-CG and CardGen directly addresses RQ2. CardGen performs well for model cards (average rank 2.2-2.4) but poorly on data cards (rank 4), revealing a fundamental limitation of static templates: predefined questions capture standardized model properties but fail to accommodate diverse dataset characteristics.

AdaQE-CG's dynamic query expansion achieves dominant data card performance (rank 1 across all dimensions) while maintaining competitive model card results. Strong Relevance scores (4.79-4.85 for GPT) demonstrate effective information identification while filtering off-topic content. The performance gap on data cards (0.24-0.30 points) quantifies adaptive extraction's value over static templates for heterogeneous web documentation.

6.2.3 ***Multi-Dimensional Evaluation Framework (RQ3).*** Our five-dimensional framework effectively assesses card quality, with each dimension capturing distinct aspects, as evidenced by varying inter-judge correlations (ρ=0.118-0.331, r=0.141-0.386).

Faithfulness and Accuracy evaluate content grounding. AdaQE-CG achieves first-rank performance on both metrics for data cards. For model cards, results show judge-dependent variation, with strong performance under Gemini evaluation (rank 1) but mixed results under GPT (Faithfulness rank 2, Accuracy rank 4). **Relevance** demonstrates strong discriminative power. AdaQE-CG's top rankings validate that this metric effectively captures whether iterative query expansion identifies pertinent information while filtering irrelevant content. **Consistency** reveals methodological tradeoffs. Zero-shot generation occasionally achieves top rankings, suggesting single-pass generation produces more uniform content, though potentially sacrificing completeness. Lower inter-judge correlation (ρ=0.118) indicates greater subjectivity. **Usefulness** assesses practical value beyond correctness. AdaQE-CG's top rankings demonstrate that dynamically retrieved content provides more actionable guidance than template-based or direct generation approaches.

6.3 Human Evaluation

To complement the LLM-as-a-Judge assessments, we conducted an independent human evaluation study. We sampled 50 model cards and 50 data cards from MetaGAI-Bench. Six Data Science master's students, divided into two independent groups, evaluated all 100 cards across five quality dimensions using the 1-5 scoring rubric in Table 1. Evaluators were selected from a different pool than benchmark annotators to prevent recognition bias, and results are presented in Table 3.

We computed inter-group agreement using Spearman and Pearson correlation coefficients to assess evaluation consistency. Correlation coefficients vary by dimension: Faithfulness shows the highest agreement (ρ=0.509, r=0.467), followed by Usefulness (ρ=0.446, r=0.453), while Relevance exhibits the lowest (ρ=0.204, r=0.231). These moderate but statistically significant correlations (all $p < 0.001$) indicate reasonable consistency, though variation reveals some quality aspects are more subjectively interpreted. Notably, human correlation values (ρ=0.204-0.509) exceed LLM judge correlations (ρ=0.118-0.331), suggesting more aligned human evaluation criteria.

6.3.1 ***Human Validation of Web-Scale Information Integration (RQ1).*** Human evaluation validates that AdaQE-CG produces high-quality documentation approaching expert standards. AdaQE-CG consistently ranks second across card types and evaluator groups (model cards: rank 2.4/2.0; data cards: rank 2.0/2.0), substantially outperforming automated baselines while remaining competitive with human-annotated cards.

Score gaps reveal remaining quality differences. For model cards, human-annotated achieve 3.53-3.66 versus AdaQE-CG's 3.35-3.44 (a gap of 0.18-0.22); for data cards, gaps widen to 0.40-0.85 points (human-annotated: 3.22-3.95 vs. AdaQE-CG: 2.42-3.71). This suggests current methods are less effective at capturing the explanatory depth and contextual appropriateness that human evaluators prioritize.

6.3.2 ***Human Perception of Dynamic vs. Static Approaches (RQ2).*** Human evaluation confirms that dynamic query expansion overcomes static template limitations. AdaQE-CG outperforms CardGen across most quality dimensions, with notable advantages in Accuracy and Usefulness.

Performance gaps are more pronounced for data cards than model cards, mirroring LLM evaluation patterns. This convergent evidence supports the claim that dynamic adaptation is essential for diverse documentation structures where standardized templates

Table 2: Performance evaluation of LLM-as-a-Judge for model card and data card generation across quality dimensions. GPT = GPT-5-nano; Gemini = Gemini-2.5 Flash-Lite. Values are shown as Rank (Score). Correlation coefficients between GPT and Gemini: ρ = Spearman correlation, r = Pearson correlation. Significance levels: * $p < 0.001$. Within each cell, bold indicates better performance (lower rank or higher score), underline indicates the second-best performance.

Card Type	Method	Faithfulness ρ=0.215*, r=0.243*		Relevance ρ=0.228*, r=0.375*		Accuracy ρ=0.204*, r=0.218*		Consistency ρ=0.118*, r=0.141*		Usefulness ρ=0.331*, r=0.386*		Average	
		GPT	Gemini	GPT	Gemini	GPT	Gemini	GPT	Gemini	GPT	Gemini	GPT	Gemini
Model	Zero-shot	4.0 (4.02)	3.0 (4.33)	4.0 (4.64)	3.0 (4.50)	3.0 (4.07)	3.0 (4.33)	**1.0 (4.75)**	**1.0 (4.66)**	4.0 (3.72)	3.0 (4.05)	3.2 (4.24)	2.6 (4.37)
Model	CardGen [26]	3.0 (4.07)	2.0 (4.36)	2.0 (4.74)	2.0 (4.62)	2.0 (4.08)	2.0 (4.36)	3.0 (4.73)	3.0 (4.65)	2.0 (3.97)	2.0 (4.19)	2.4 (4.32)	2.2 (4.43)
Model	Human-Annotated	**1.0 (4.09)**	4.0 (4.26)	3.0 (4.66)	4.0 (4.49)	**1.0 (4.11)**	4.0 (4.26)	2.0 (4.75)	4.0 (4.62)	3.0 (3.83)	4.0 (4.04)	**2.0 (4.29)**	4.0 (4.33)
Model	AdaQE-CG	2.0 (4.08)	**1.0 (4.37)**	**1.0 (4.79)**	**1.0 (4.65)**	4.0 (4.05)	**1.0 (4.37)**	4.0 (4.73)	2.0 (4.65)	**1.0 (4.02)**	**1.0 (4.24)**	2.4 (4.33)	**1.2 (4.46)**
Data	Zero-shot	3.0 (4.10)	2.0 (4.27)	3.0 (4.62)	2.0 (4.57)	2.0 (4.19)	2.0 (4.27)	2.0 (4.75)	2.0 (4.67)	3.0 (3.83)	2.0 (4.08)	2.6 (4.30)	2.0 (4.37)
Data	CardGen [26]	4.0 (4.01)	4.0 (4.07)	4.0 (4.57)	4.0 (4.39)	4.0 (4.07)	4.0 (4.07)	4.0 (4.67)	4.0 (4.47)	4.0 (3.82)	4.0 (3.93)	4.0 (4.23)	4.0 (4.18)
Data	Human-Annotated	2.0 (4.12)	3.0 (4.19)	2.0 (4.64)	3.0 (4.51)	3.0 (4.18)	3.0 (4.19)	3.0 (4.73)	3.0 (4.60)	2.0 (3.86)	3.0 (4.05)	2.4 (4.31)	3.0 (4.31)
Data	AdaQE-CG	**1.0 (4.26)**	**1.0 (4.37)**	**1.0 (4.85)**	**1.0 (4.71)**	**1.0 (4.24)**	**1.0 (4.36)**	**1.0 (4.81)**	**1.0 (4.69)**	**1.0 (4.18)**	**1.0 (4.28)**	**1.0 (4.47)**	**1.0 (4.48)**

Table 3: Human evaluation of model card and data card generation across quality dimensions. Group 1 and Group 2 represent two independent sets of human evaluators assessing the same algorithms and card types. Values shown as Rank (Score). Correlation coefficients between Group 1 and Group 2: ρ = Spearman correlation, r = Pearson correlation. Significance levels: * $p < 0.001$. Within each metric, bold indicates best performance (lowest rank or highest score), underline indicates the second-best performance.

Card Type	Method	Faithfulness ρ=0.509*, r=0.467*		Relevance ρ=0.204*, r=0.231*		Accuracy ρ=0.358*, r=0.366*		Consistency ρ=0.310*, r=0.319*		Usefulness ρ=0.446*, r=0.453*		Average	
		Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
Model	Zero-shot	4.0 (2.73)	4.0 (2.94)	4.0 (4.12)	4.0 (3.01)	4.0 (2.69)	4.0 (2.85)	4.0 (3.70)	4.0 (3.17)	4.0 (2.68)	4.0 (2.79)	4.0 (3.18)	4.0 (2.95)
Model	CardGen [26]	3.0 (3.29)	3.0 (3.21)	2.0 (4.53)	3.0 (3.33)	3.0 (3.28)	3.0 (3.17)	2.0 (4.19)	3.0 (3.39)	3.0 (3.26)	3.0 (3.09)	2.6 (3.71)	3.0 (3.24)
Model	Human-Annotated	**1.0 (3.69)**	**1.0 (3.53)**	**1.0 (4.66)**	**1.0 (3.62)**	**1.0 (3.66)**	**1.0 (3.43)**	**1.0 (4.32)**	**1.0 (3.59)**	**1.0 (3.55)**	**1.0 (3.35)**	**1.0 (3.98)**	**1.0 (3.50)**
Model	AdaQE-CG	2.0 (3.47)	2.0 (3.31)	3.0 (4.38)	2.0 (3.43)	2.0 (3.44)	2.0 (3.22)	3.0 (4.05)	2.0 (3.45)	2.0 (3.35)	2.0 (3.17)	2.4 (3.74)	2.0 (3.32)
Data	Zero-shot	4.0 (2.94)	4.0 (2.76)	4.0 (3.44)	4.0 (2.41)	4.0 (3.01)	4.0 (2.13)	4.0 (2.95)	4.0 (2.20)	4.0 (2.50)	3.0 (2.23)	4.0 (2.97)	3.8 (2.35)
Data	CardGen [26]	3.0 (3.08)	3.0 (2.98)	3.0 (3.50)	3.0 (2.61)	3.0 (3.08)	3.0 (2.25)	3.0 (3.13)	3.0 (2.53)	3.0 (2.69)	4.0 (2.21)	3.0 (3.10)	3.2 (2.52)
Data	Human-Annotated	**1.0 (3.79)**	**1.0 (4.01)**	**1.0 (3.95)**	**1.0 (3.92)**	**1.0 (3.57)**	**1.0 (3.57)**	**1.0 (3.43)**	**1.0 (3.56)**	**1.0 (3.22)**	**1.0 (3.45)**	**1.0 (3.59)**	**1.0 (3.70)**
Data	AdaQE-CG	2.0 (3.39)	2.0 (3.15)	2.0 (3.71)	2.0 (2.83)	2.0 (3.31)	2.0 (2.60)	2.0 (3.32)	2.0 (2.70)	2.0 (2.95)	2.0 (2.42)	2.0 (3.34)	2.0 (2.74)

fail, producing more actionable, contextually appropriate documentation.

6.3.3 ***Cross-Validation of Multi-Dimensional Evaluation Framework (RQ3).*** Human and LLM-based evaluation reveal complementary strengths. Human evaluators maintain higher inter-rater consistency (ρ=0.204-0.509) than LLM judges (ρ=0.118-0.331), but score systematically differently from each other: humans rate human-annotated cards higher, while scoring automated methods conservatively compared to LLMs' compressed distributions.

LLM judges prioritize syntactic correctness and factual accuracy, occasionally ranking AdaQE-CG first. Human evaluators apply holistic criteria encompassing explanatory depth, audience appropriateness, and practical utility. Convergence in relative rankings validates the framework's ability to distinguish method quality, while divergence in absolute scores reveals complementary evaluation perspectives.

7 Conclusion

This work addresses the critical challenge of automating semantic documentation for GAI systems. We introduce the MetaGAI-Dataset, comprising 6,481 data cards and 123,013 model cards, along with a high-quality human-annotated MetaGAI-Bench of 600 data cards and 600 model cards. Additionally, we propose the AdaQE-CG framework that combines dynamic query expansion with cross-card knowledge transfer. Our analysis reveals significant documentation quality disparities, with data cards achieving substantially higher completeness than model cards and critical deficiencies in safety and ethical considerations. The proposed AdaQE-CG demonstrates that integrating dynamic retrieval with cross-card knowledge transfer enables scalable, high-quality semantic card generation, achieving first-place rankings across all quality dimensions for data cards and competitive performance for model cards in both LLM-based and human evaluations. Limitations include degraded performance for novel GAI architectures lacking reference cards and the inability to extract visual information from figures and tables. Future work will address multimodal knowledge extraction from figures and tables, unified data-model card mapping, and targeted methods for enhancing ethical documentation.

References

[1] David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. 2022. Prescriptive and descriptive approaches to machine-learning transparency. In *CHI conference on human factors in computing systems extended abstracts*. 1–9.

[2] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[3] Jiaxin Bai, Wei Fan, Qi Hu, Qing Zong, Chunyang Li, Hong Ting Tsang, Hongyu Luo, Yauwai Yim, Haoyu Huang, Xiao Zhou, et al. 2025. AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora. *arXiv preprint arXiv:2505.23628* (2025).

[4] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[5] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track. *Advances in Neural Information Processing Systems* 37 (2024), 53626–53648.

[6] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418* (2023).

[7] Danilo Brajovic, Niclas Renner, Vincent Philipp Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klaeb, Janika Kutz, Jens Neuhuettler, and Marco F Huber. 2023. Model reporting for certifiable ai: A proposal from merging eu regulation into ai development. *arXiv preprint arXiv:2307.11525* (2023).

[8] Cheng-Han Chiang and Hung-Yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15607–15631.

[9] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[11] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications* 15, 1 (2024), 1418.

[12] Roxana Daneshjou, Mary P Smith, Mary D Sun, Veronica Rotemberg, and James Zou. 2021. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA dermatology* 157, 11 (2021), 1362–1369.

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[14] Stephen Gilbert, Rasmus Adler, Taras Holoyad, and Eva Weicken. 2025. Could transparent model cards with layered accessible information drive trust and safety in health AI? *npj Digital Medicine* 8, 1 (2025), 124.

[15] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).

[16] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *CoRR* (2024).

[17] Carolina AM Heming, Mohamed Abdalla, Shahram Mohanna, Monish Ahluwalia, Linglin Zhang, Hari Trivedi, MinJae Woo, Benjamin Fine, Judy Wawira Gichoya, Leo Anthony Celi, et al. 2023. Benchmarking bias: Expanding clinical AI model card to incorporate bias reporting of social and non-social factors. *arXiv preprint arXiv:2311.12560* (2023).

[18] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data protection and privacy* 12, 12 (2020), 1.

[19] Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. 2025. Charting and Navigating Hugging Face's Model Atlas. *arXiv e-prints* (2025), arXiv–2503.

[20] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[21] Yizhu Jiao, Sha Li, Sizhe Zhou, Heng Ji, and Jiawei Han. 2024. TEXT2DB: Integration-Aware Information Extraction with Large Language Model Agents. In *Findings of the Association for Computational Linguistics ACL 2024*. 185–205.

[22] Joshua A Kroll. 2021. Outlining traceability: A principle for operationalizing accountability in computing systems. In *Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency*. 758–771.

[23] Benjamin Laufer, Hamidah Oderinwale, and Jon Kleinberg. 2025. Anatomy of a Machine Learning Ecosystem: 2 Million Models on Hugging Face. *arXiv preprint arXiv:2508.06811* (2025).

[24] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).

[25] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. *Nature Machine Intelligence* 6, 7 (2024), 744–753.

[26] Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona Diab. 2024. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1975–1997.

[27] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence* 6, 8 (2024), 975–987.

[28] Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Katy Gero, Sandy Pentland, and Jad Kabbara. 2024. Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them? *arXiv preprint arXiv:2404.12691* (2024).

[29] Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why AI Is WEIRD and Shouldn't Be This Way: Towards AI for Everyone, with Everyone, by Everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28657–28670.

[30] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[31] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2024. Accountability in artificial intelligence: What it is and how it works. *Ai & Society* 39, 4 (2024), 1871–1882.

[32] Trishan Panch, Heather Mattie, and Rifat Atun. 2019. Artificial intelligence and algorithmic bias: implications for health systems. *Journal of global health* 9, 2 (2019), 020318.

[33] Emmanouil Papagiannidis, Patrick Mikalef, and Kieran Conboy. 2025. Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems* 34, 2 (2025), 101885.

[34] P Jonathon Phillips, P Jonathon Phillips, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, and Mark A Przybocki. 2021. Four principles of explainable artificial intelligence. (2021).

[35] Tim Puhlfürß, Julia Butzke, and Walid Maalej. 2025. Model Cards Revisited: Bridging the Gap Between Theory and Practice for Ethical AI Requirements. *arXiv preprint arXiv:2507.06014* (2025).

[36] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1776–1826.

[37] Shaina Raza, Rizwan Qureshi, Anam Zahid, Safiullah Kamawal, Ferhat Sadak, Joseph Fioresi, Muhammaed Saeed, Ranjan Sapkota, Aditya Jain, Anas Zafar, et al. 2025. Who is responsible? the data, models, users or regulations? a comprehensive survey on responsible generative ai for a sustainable future. *arXiv preprint arXiv:2502.08650* (2025).

[38] Anthony Cintron Roman, Jennifer Wortman Vaughan, Valerie See, Steph Ballard, Jehu Torres, Caleb Robinson, and Juan M Lavista Ferres. 2023. Open datasheets: Machine-readable documentation for open datasets and responsible ai assessments. *arXiv preprint arXiv:2312.06153* (2023).

[39] Marco Rondina, Antonio Vetrò, and Juan Carlos De Martin. 2023. Completeness of datasets documentation on ML/AI repositories: An empirical investigation. In *EPIA Conference on Artificial Intelligence*. Springer, 79–91.

[40] Christin Seifert, Stefanie Scherzinger, and Lena Wiese. 2019. Towards generating consumer labels for machine learning models. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 173–179.

[41] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine* 46, 2 (2025), e70002.

[42] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 850–861.

[43] Huzaifa Sidhpurwala, Emily Fox, Garth Mollett, Florencio Cano Gabarda, and Roman Zhukov. 2025. Blueprints of Trust: AI System Cards for End to End Transparency and Governance. *arXiv preprint arXiv:2509.20394* (2025).

[44] Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh Chawla. 2024. BenchmarkCards: Standardized Documentation for Large Language Model Benchmarks. *arXiv preprint arXiv:2410.12974* (2024).

[45] Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. 2025. VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6088–6109.

[46] Xinyu Yang, Weixin Liang, and James Zou. 2024. Navigating dataset documentations in AI: A large-scale analysis of dataset cards on hugging face. *arXiv preprint arXiv:2401.13822* (2024).
[47] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.
[48] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

A Appendix

A.1 Definition of Model and Data Card

The definitions of model and data cards for GAI are shown in Table 4.

A.2 Weighted Card Completeness Index

The WCCI quantifies documentation quality through content availability and confidence levels. The metric incorporates three design principles: *interpretability* for cross-artifact comparison; *confidence-weighted evaluation* reflecting information reliability; and *uniform field weighting* preventing systematic bias. For each field i with content c_i and confidence level conf_i, the completeness score is: 0.0 for missing fields, 1.0 for non-applicable fields, and $w(\text{conf}_i) \in \{0.25, 0.5, 0.75, 1.0\}$ for confidence levels {low, medium, high, and certain}. The overall WCCI is:

$$\text{WCCI} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \text{completeness\_score}(i) \tag{3}$$

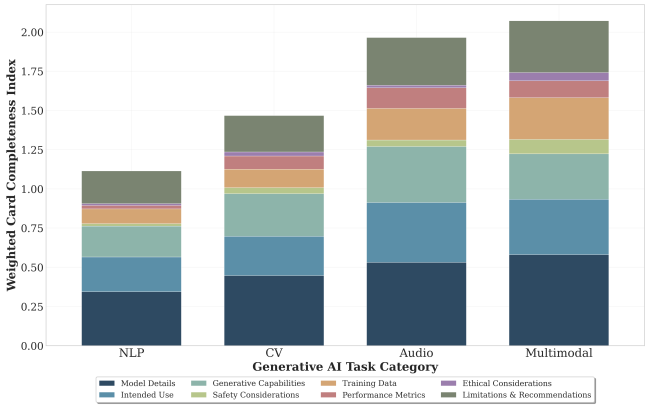
Table 5 shows correlations between WCCI scores and metadata tags.

A.3 Field-Level WCCI Analysis by Task Category

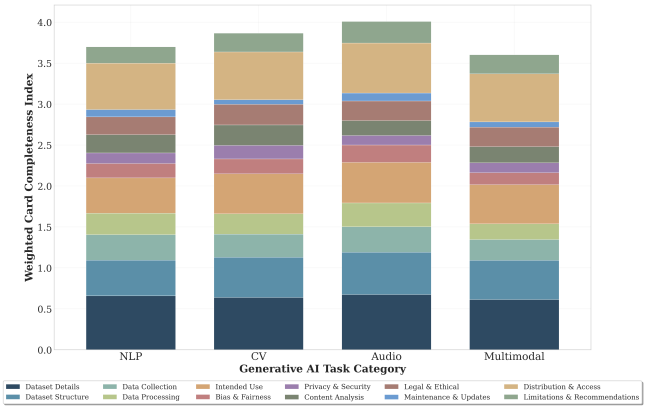
MetaGAI-Dataset exhibits pronounced documentation disparities across artifact types and modalities (Figure 3). Data cards substantially outperform model cards. Within model cards, performance varies by modality, with Multimodal and Audio exceeding CV and NLP. Critically, NLP demonstrates the lowest completeness despite representing 88% of model cards. Field-level decomposition reveals that model cards contain negligible contributions from Safety Considerations, Ethical Considerations, and Performance Metrics. Data cards achieve higher completeness primarily through comprehensive technical documentation (Dataset Details, Dataset Structure, Data Collection), while responsible AI dimensions (Privacy & Security, Legal & Ethical, Limitations & Recommendations) remain underrepresented across all modalities. This pattern indicates systematic prioritization of technical specifications over responsible AI documentation in both artifact types [25, 46].

A.4 Model Details

We employ Gemini-2.5 Flash-Lite [10] as the large language model for all card generation tasks. For fair comparison, both baseline methods (Zero-Shot and CardGen) also use Gemini-2.5 Flash-Lite as their underlying language model. We configure the model with temperature = 0.2, top-p sampling = 0.9, and maximum output tokens = 8,192. This configuration is consistently applied across all methods to ensure comparable evaluation conditions. For retrieval reranking, we employ the Qwen3-Reranker-4B model [48], configured with a maximum sequence length of 8,192 tokens, to process retrieved document chunks and score their relevance to queries.



(a) Model Cards



(b) Data Cards

Figure 3: Field-level WCCI decomposition across generative AI task categories. (a) Model cards show substantially lower completeness with minimal Safety and Ethical Considerations. (b) Data cards achieve higher completeness dominated by technical specifications, with responsible AI dimensions contributing smaller segments.

A.5 Evaluation Details

For LLM-as-a-Judge evaluation, we employ two distinct large language models: GPT-5-nano (version gpt-5-nano-2025-08-07 from OpenAI) and Gemini-2.5 Flash-Lite [10]. Each judge LLM independently assesses outputs from all algorithms on five metrics: Faithfulness, Relevance, Accuracy, Consistency, and Usefulness. To ensure robustness, we implement three bias mitigation strategies: (1) algorithm anonymization by removing all identifiers, (2) five independent evaluation rounds with randomized presentation orders, and (3) dual-judge assessment using two distinct LLMs. Final scores are computed as the average across both judges and all five rounds, yielding 10 independent assessments per evaluation instance.

Table 4: Definitions of Model and Data Card for Generative AI

Card	Field	Description
Model Card	Model Details	Information about the model developer, architecture, size, training methodology, modalities, version, license, and contact details
	Intended Use	Primary applications, target users, supported languages/domains, out-of-scope uses, and age restrictions
	Generative Capabilities	Generation quality, content types, length limitations, consistency, latency, and customization options
	Safety Considerations	Content safety measures, bias analysis, fairness metrics, red team testing, jailbreaking resistance, and child safety
	Training Data	Training corpus details, data filtering processes, demographic representation, language coverage, consent/privacy, and evaluation datasets
	Performance Metrics	Generation quality metrics, safety metrics, factual accuracy, bias metrics, cultural sensitivity, and robustness measures
	Ethical Considerations	Dual-use risks, misinformation potential, intellectual property concerns, economic/environmental impact, cultural appropriation, privacy, and consent issues
	Caveats & Recommendations	Known limitations, deployment recommendations, monitoring requirements, and user guidelines
Data Card	Dataset Details	Dataset name, version, creators/curators, funding, type, text language, license, and related resources
	Dataset Structure	Instances, fields, missing information, relationships, splits, and size statistics
	Data Collection	Collection process, data sources, timeframe, ethical review, consent process, and data validation
	Data Processing	Preprocessing steps, cleaning procedures, labeling process, quality control, filtering criteria, and deduplication
	Intended Use	Primary tasks, suitable/unsuitable applications, research applications, commercial applications, and prohibited uses
	Bias & Fairness	Demographic representation, geographic/temporal coverage, known biases, bias mitigation, and fairness considerations
	Privacy & Security	Personally identifiable information, sensitive information, privacy protection measures, data security, anonymization/pseudonymization, and retention/deletion policies
	Content Analysis	Content types, harmful content identification, content moderation, toxicity analysis, misinformation risks, and cultural sensitivity
	Legal & Ethical	Copyright considerations, terms of use, ethical guidelines, compliance requirements, subject rights, and institutional review
	Maintenance & Updates	Maintenance plan, update frequency, versioning, error reporting, community contribution, and deprecation plan
	Distribution & Access	Access mechanism, distribution format, download instructions, API access, access restrictions, and citation requirements
	Limitations & Recommendations	Known limitations, recommended uses, usage guidelines, performance considerations, environmental impact, and future work

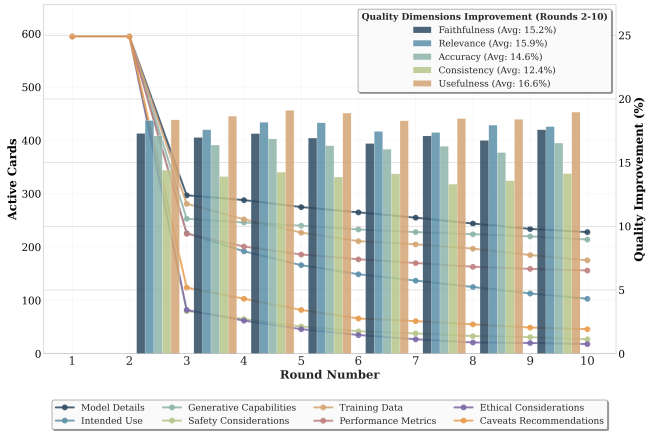
A.6 Ablation Study

A.6.1 MetaGAI Pool Contribution. ICC-MP successfully enhanced 52.7% of model card fields and 56.5% of data card fields (Figure 5). Of 243 model cards processed by ICC-MP, 128 (52.7%) showed improvement, 60 (24.7%) showed no change, and 55 (22.6%) showed tied performance. Data cards demonstrated a stronger impact with 134 of 237 cards (56.5%) improved, 94 (39.7%) showing no change, and 9 (3.8%) tied, validating cross-card knowledge transfer when source documents lack information.

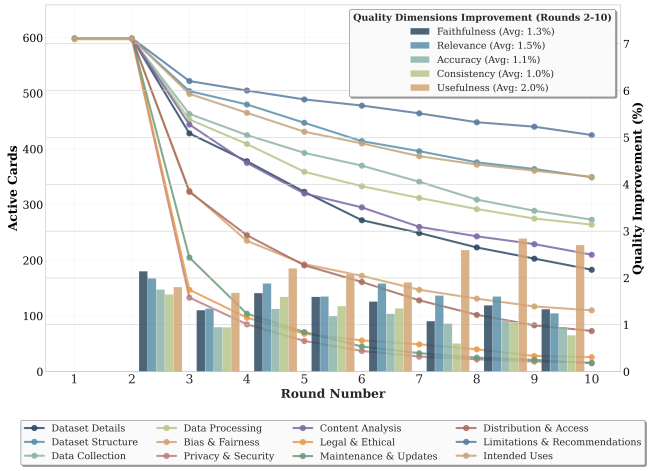
A.6.2 Effectiveness of Multi-Turn Query Expansion. Using LLM-as-a-Judge (Gemini-2.5 Flash-Lite), we compared consecutive rounds (R_{i-1} vs. R_i) across five quality dimensions. Figure 4 reveals: (1) progressive convergence—active cards decline from 600 (Round 2) to 100-430 (Round 10); (2) sustained quality gains—model cards achieve 12-16% average improvements per dimension while data cards show 1-2% gains; (3) field-specific behavior—Model Details and Dataset Structure maintain higher activity (430-520 cards at Round 10), while responsible AI fields converge rapidly to under 200 cards.

A.7 Case Study of Generated Model and Data Cards

To demonstrate AdaQE-CG's practical effectiveness, we present partial cards for *all-hands/openhands-lm-32b-v0.1-ep3* (model) and *MushanW/GLOBE_V3* (data). IPE-QE extracted core information through iterative query refinement, while ICC-MP enriched incomplete fields using similar artifacts. Enrichments marked by yellow background and [enriched by ICC-MP] demonstrate successful cross-card knowledge transfer, validating AdaQE-CG's systematic extraction and conditional enrichment capabilities.



(a) Model Cards



(b) Data Cards

Figure 4: Query expansion effectiveness across rounds 2-10. Lines (left y-axis) show active cards undergoing query expansion per field; bars (right y-axis) track cumulative quality improvements across five dimensions.

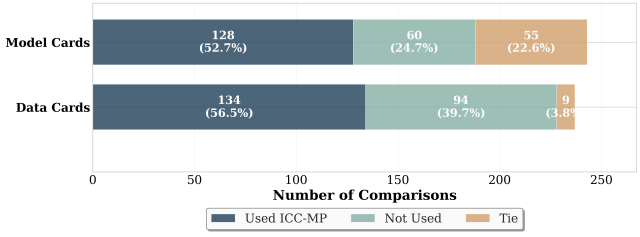


Figure 5: Quality improvement contribution of ICC-MP. Blue bars show cases where ICC-MP improved quality scores over IPE-QE-only cards, green bars show no improvement, and orange bars indicate tied scores.

Table 5: Spearman Correlations between WCCI Scores and Metadata Tags. * indicates $p < 0.001$.

Card Type	Category	Metadata Tags	Correlation (ρ)	Frequency (n)
Model Cards	Popularity	likes	0.408*	123,013
		downloads	0.285*	123,013
	Task	text-to-speech	0.211*	3,145
		text-to-audio	0.173*	2,781
		image-text-to-text	0.132*	2,872
		text-to-image	0.105*	1,223
		text-generation	-0.153*	97,689
	License	apache-2.0	0.374*	12,520
		cc-by-nc-4.0	0.222*	3,445
	Framework	pytorch	0.386*	16,750
Data Cards	Popularity	downloads	0.257*	6,481
		likes	0.182*	6,481
	Task	zero-shot classification	0.203*	680
		text-classification	0.201*	1,220
		table-QA	0.195*	717
		question-answer	0.178*	3,261
		multiple-choice	0.164*	665
	Annotation	crowdsourced	-0.384*	1,389

Model Card: all-hands/openhands-lm-32b-v0.1-ep3

Model Details
Developer: OpenHands
Architecture: 32B Qwen2.5-Coder-Instruct.
License: **MIT** [enriched by ICC-MP]
...

Intended Use
Primary Applications: Agent scaffold for general-purpose prompting in software engineering tasks
Out of Scope: **Best suited for solving GitHub issues ...** [enriched by ICC-MP]
...
[Additional sections]

Data Card: MushanW/GLOBE_V3

Dataset Details
Dataset Name: GLOBE_V3
Version: **V3 version...** [enriched by ICC-MP]
...

Dataset Structure
Instances: The dataset includes utterances from 23,519 speakers.
Fields: Detailed metadata is available for all speakers ...
...
[Additional sections]