MetaGAI: A Large-Scale and High-Quality Benchmark for Generative AI Model and Data Card Generation

Anonymous ACL submission

Abstract

The rapid proliferation of Generative AI necessitates rigorous documentation standards for transparency and governance. However, manual creation of Model and Data Cards is not scalable, while automated approaches lack large-scale, high-fidelity benchmarks for systematic evaluation. We introduce MetaGAI, a comprehensive benchmark comprising 2,541 verified document triplets constructed through semantic triangulation of academic papers, GitHub repositories, and Hugging Face artifacts. Unlike prior single-source datasets, MetaGAI employs a multi-agent framework with specialized Retriever, Generator, and Editor agents, validated through four-dimensional human-in-the-loop assessment, including human evaluation of editor-refined ground truth. We establish a robust evaluation protocol combining automated metrics with validated LLM-as-a-Judge frameworks. Extensive analysis reveals that sparse Mixture-of-Experts architectures achieve superior cost-quality efficiency, while a fundamental trade-off exists between faithfulness and completeness. MetaGAI provides a foundational testbed for benchmarking, training, and analyzing automated Model and Data Card generation methods at scale. Our data and code are available at https://anonymous.4open.science/r/MetaGAI-DBB4.

1 Introduction

The rapid proliferation of Generative AI (GenAI) has fundamentally transformed machine learning deployment, with more than two million models and datasets now hosted on community platforms such as Hugging Face (Wolf et al., 2020; Horwitz et al., 2025). As these systems transition from research artifacts to critical infrastructure, documentation standards have evolved from foundational Model Cards (Mitchell et al., 2019) and Data Cards (Pushkarna et al., 2022) to comprehensive System Cards (OpenAI et al., 2024; Comanici et al., 2025; OpenAI et al., 2025), enabling longitudinal model tracking, compliance auditing, and risk assessment in high-stakes domains (Longpre et al., 2024; Castaño et al., 2024). Without high-quality documentation, the AI ecosystem lacks the interoperability required to benchmark capabilities or trace data provenance across complex supply chains (Rahman et al., 2025).

However, a significant bottleneck impedes widespread adoption. The GenAI ecosystem is increasingly driven by the "long tail" of open-source contributions, comprising thousands of models on community platforms (Wolf et al., 2020; Horwitz et al., 2025). Unlike well-resourced industry laboratories, developers in this ecosystem often lack the capacity to maintain rigorous documentation. Manual card creation suffers from severe scalability constraints, characterized by pervasive incompleteness, inconsistency, and subjectivity (Yang et al., 2024; Liang et al., 2024), leading to a transparency crisis where many research papers and repositories lack structured metadata necessary for reproducibility (Olmo et al., 2025). Automated documentation generation has emerged as a critical necessity (Liu et al., 2024a). To address this challenge, current approaches face significant hurdles: zero-shot methods often hallucinate details when summarizing lengthy documents, while retrieval-augmented strategies struggle to align diverse paper structures with rigid schemas. Progress is stalled by the lack of large-scale, high-quality benchmarks that can objectively measure automated generation accuracy against verified ground truth.

To fill this gap, we introduce **MetaGAI**, a large-scale benchmark designed to systematically evaluate automated Model and Data Card generation. Unlike prior datasets treating documentation as simple summarization, MetaGAI formulates it as a complex multi-source information generation task, mirroring real-world requirements for verifying scientific claims against implementation details.

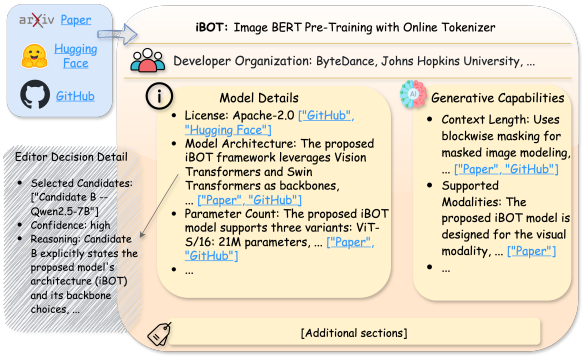


Figure 1: MetaGAI Benchmark Construction Example. Automated GenAI card generation for the iBOT model (Zhou et al., 2022) demonstrating Multi-Source Triangulation combining architectural concepts from Papers, hyperparameters from GitHub, and licensing data from Hugging Face, with Editor-Based Synthesis to produce high-fidelity ground truth.

As illustrated in Figure 1, creating complete documentation cards requires triangulating evidence from heterogeneous sources. Taking the iBOT model (Zhou et al., 2022) as an example, architectural concepts are derived from the academic Paper, implementation details are extracted from the GitHub repository, and deployment constraints are verified via Hugging Face. Our construction pipeline utilizes a multi-agent framework comprising specialized Retriever, Generator, and Editor agents to synthesize these signals into verified ground truth. We implement rigorous human-in-the-loop validation across four dimensions: (1) retrieval strategy validation through domain expert annotation, (2) generator divergence analysis validating ensemble diversity, (3) benchmark ground truth quality assessment through hybrid human-LLM evaluation of editor-refined ground truth, and (4) editor architecture selection through pairwise comparisons, establishing a rigorous standard for evaluating automated systems.

In a nutshell, the key contributions of this study are summarized as follows:

- We construct MetaGAI, the largest high-quality benchmark for GenAI documentation, comprising 2,541 verified triplets with rigorous human-in-the-loop validation across four dimensions.
- We propose a robust evaluation framework combining granular automated metrics with a validated LLM-as-a-Judge protocol.
- We provide extensive empirical analysis revealing that sparse Mixture-of-Experts (MoE) architectures offer superior cost-quality efficiency, though a systematic trade-off persists between faithfulness and completeness in generating metadata.

2 Related Work

Model Documentation. The lack of standardized documentation for trained machine learning models limits transparency, reproducibility, and responsible use in Natural Language Processing (NLP). Model Cards were introduced to provide structured summaries of trained models, including intended use, evaluation settings, performance, and limitations (Mitchell et al., 2019). Dataset-level documentation was developed in parallel. Data Cards focus on data provenance, collection processes, representativeness, and ethical considerations (Pushkarna et al., 2022). Together, Model Cards and Data Cards form a documentation framework that addresses both model-level and data-level sources of uncertainty. Subsequent work extended the model documentation to interactive and machine-readable formats. Interactive Model Cards were shown to better support user understanding of model behavior (Crisan et al., 2022), while linked and semantic representations of documentation were proposed to improve traceability and reuse (Donald et al., 2023). Empirical analysis of a large number of Model Cards indicates substantial variation in documentation quality, with evaluation and limitation sections often missing or incomplete (Liang et al., 2024).

Automatic Model and Data Card Generation. To reduce the manual effort required for documentation, recent work has explored automatic generation of Model and Data Cards. Liu et al. (2024a) introduced CARDBENCH, a benchmark of human-written cards, together with CARDGEN, an LLM-based system that generates structured documentation by retrieving information from model repositories and associated papers. Their results show that automatic generation can improve documentation completeness. However, existing benchmarks remain limited in scale and coverage, which constrains systematic evaluation of generation methods. To address this limitation, we introduce MetaGAI, a large-scale benchmark for GenAI Model and Data Card generation.

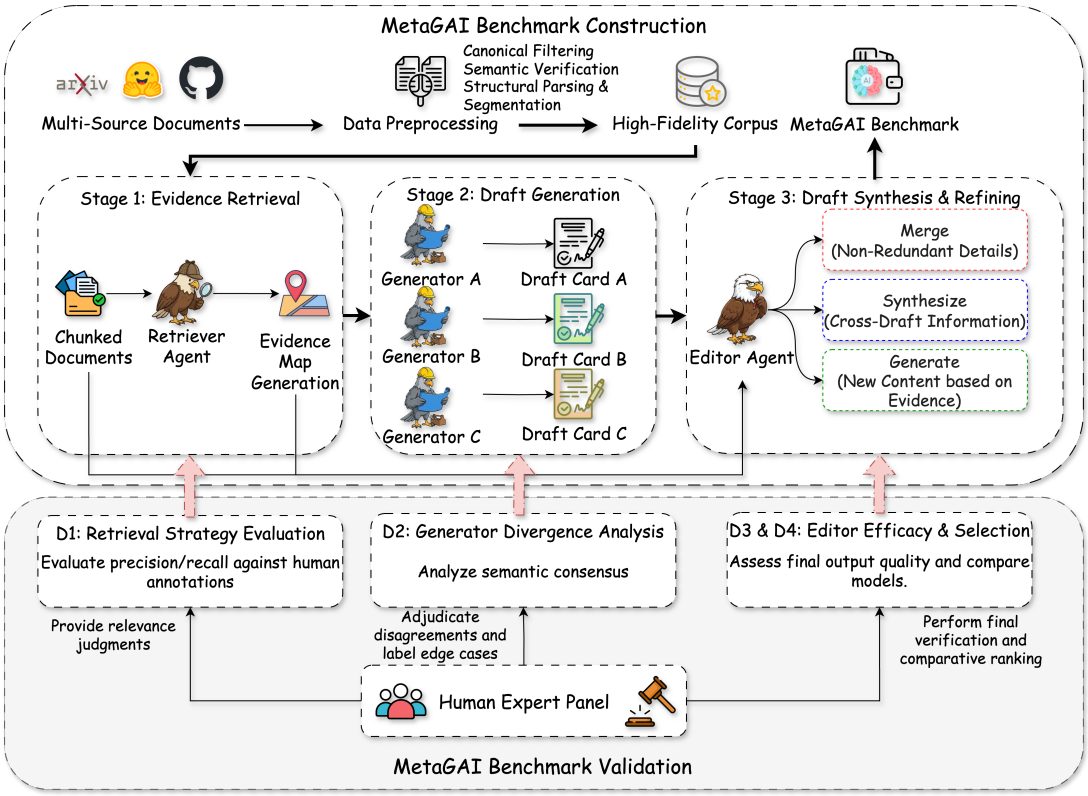


Figure 2: **MetaGAI Benchmark Construction and Validation Framework.** The pipeline integrates multi-source document preprocessing, a multi-agent generation framework (Evidence Retrieval, Draft Generation, Draft Synthesis and Refining), and a four-dimensional validation protocol (D1-D4) incorporating human expert adjudication.

3 MetaGAI Benchmark Construction

Figure 2 illustrates the MetaGAI benchmark construction and validation pipeline, comprising three sequential stages: (1) evidence retrieval from multi-source documents, (2) multi-generator draft synthesis, and (3) editor-based consolidation and refinement. This architecture extends beyond traditional single-model approaches (Liu et al., 2024a) by incorporating ensemble generation and rigorous human-in-the-loop validation to ensure high-fidelity card generation for GenAI ecosystems.

3.1 Task Definition

We formulate **MetaGAI** as a structured generation task mapping unstructured scientific papers $\mathcal{P}$ to standardized cards $\mathcal{C}$. The benchmark embodies a deliberate asymmetry between ground truth construction and evaluation input. The ground truth $\mathcal{C}_{GT}$ is derived from a document triplet $(\mathcal{P}, \mathcal{G}, \mathcal{H})$, where $\mathcal{G}$ denotes the GitHub repository and $\mathcal{H}$ the Hugging Face artifact, to capture the implementation details frequently omitted in manuscripts (Mitchell et al., 2019; Pushkarna et al., 2022). The evaluation objective, however, is restricted to paper input $\mathcal{P}$ alone, reflecting the most prevalent real-world deployment scenario and allowing scalable, fair, and difficulty-appropriate evaluation across artifacts.

We adopt and extend the taxonomies of Mitchell et al. (2019) and Pushkarna et al. (2022) with GenAI-specific fields (Appendix A, Table 4). Each card $\mathcal{C}$ comprises M field-value pairs $\{(k_i, v_i)\}_{i=1}^{M}$, where k_i is a schema attribute and v_i its content. The objective is to learn a mapping f_θ predicting $\hat{\mathcal{C}}$ from paper input alone:

$$\hat{\mathcal{C}} = f_\theta(\mathcal{P}) \tag{1}$$

minimizing divergence from the multi-source reference $\mathcal{C}_{GT}(\mathcal{P}, \mathcal{G}, \mathcal{H})$, thereby quantifying how much documentation quality can be recovered under paper-only constraints.

3.2 Data Acquisition and Filtering

Corpus Construction. We construct the benchmark through systematic triangulation of arXiv, GitHub, and Hugging Face metadata to capture implementation details absent from published papers. Unlike previous work that aggregates existing documentation (Liu et al., 2024a), our work constructs a

new corpus from scratch, with ground truth verified via cross-source semantic consistency. An initial corpus of **15,727** candidates is canonically filtered to identify unique paper–model–dataset linkages, resulting in **4,068** entries.

Semantic Verification. To eliminate spurious citations and ensure cross-source alignment, we employed Qwen3-30B-A3B-Instruct (Yang et al., 2025) for automated semantic consistency verification across the three sources (Prompt F.1). To validate this automated approach, two domain experts independently performed a binary correspondence judgment on a pilot sample of 100 triplets (50 model, 50 dataset): each annotator was asked solely whether the Paper, GitHub repository, and Hugging Face artifact referred to the same entity, without generating or evaluating any card content. This objective correspondence task yielded perfect inter-annotator agreement, confirming that source-level entity matching is unambiguous when triplets are well-formed and validating the reliability of the automated filtering procedure. The full filtering pipeline produced a final high-fidelity corpus of **2,541** verified triplets (Appendix B). Detailed provenance analysis appears in Appendix G.1.

3.3 Benchmark Generation Pipeline

We developed an automated pipeline synthesizing high-fidelity cards through document preprocessing and a multi-agent framework.

3.3.1 Pre-processing

Raw PDF documents were converted to structured Markdown using OLMoCR-2, a state-of-the-art OCR model optimized for academic layouts (Poznanski et al., 2025). To accommodate context window constraints, the converted papers and README files were segmented into 1024-token chunks with 10% overlap between adjacent chunks to preserve semantic continuity. Chunk boundaries are always extended to the nearest sentence boundary to avoid splitting semantically coherent passages, ensuring accurate retrieval of implementation details dispersed throughout documents.

3.3.2 Multi-Agent Framework

We employ a multi-agent architecture (Algorithm 1) to maximize factual grounding and minimize hallucinations (Du et al., 2024). Where traditional approaches rely on single-model generation following retrieval (Liu et al., 2024a), our framework introduces ensemble diversity and editor-based cross-validation. The framework processes full context $\mathcal{P} \cup \mathcal{S}$, where $\mathcal{S} = \{\mathcal{G}, \mathcal{H}\}$ denotes supplementary documentation, through three specialized agents:

$$v_i = \text{Retriever}(\mathcal{P} \cup \mathcal{S}, k_i) \quad (2)$$

$$\tilde{\mathcal{C}} = \text{Generator}(\{(k_i, v_i)\}_{i=1}^{M}) \quad (3)$$

$$\hat{\mathcal{C}} = \text{Editor}(\mathcal{P} \cup \mathcal{S}, \tilde{\mathcal{C}}) \quad (4)$$

Retriever Agent. This agent aligns unstructured text with schema fields through exhaustive chunk-level classification. Each document segment receives a relevance score (0–4) along with matched sub-fields selected from a predefined ontology (Prompt F.2). Comparative analysis (Section 4.1) established generative models' superiority over discriminative rerankers for complex schema alignment, informing our selection of Qwen3-30B-A3B-Instruct as the retrieval backbone. This generative reasoning approach contrasts with embedding-based similarity matching, enabling more nuanced interpretation of implicit metadata requirements.

Generator Agent. To mitigate architectural bias and capture diverse interpretations, we employ ensemble generation (Wang et al., 2023) with three architecturally distinct LLMs (OLMo-3-7B, Llama-3.1-8B, Qwen2.5-7B). Each model independently synthesizes the draft content, evidence quotations, and confidence scores. A strict "Subject Focus" constraint (Prompt F.3) ensures agents synthesize information about the proposed model or dataset only, excluding baselines and prior work. This multi-model design addresses single-architecture limitations in capturing the full evidence spectrum across complex GenAI systems.

Editor Agent. A "Chief Editor" consolidates candidate drafts using a larger, cross-family model to avoid self-enhancement bias (Zheng et al., 2023). The editor validates semantic alignment between drafts and raw evidence, filters attribution errors, and merges non-redundant details into concise final entries (Prompt F.4). This explicit cross-validation stage against original sources provides stronger hallucination mitigation than single-pass generation, which is particularly critical for technical specifications where factual precision is essential. Concrete examples appear in Appendix H.

Model	P@1	R@5	F1@5
Discriminative Rerankers			
BGE-Reranker-v2-m3	0.135	0.114	0.101
Qwen3-Reranker-8B	0.115	0.137	0.091
Generative LLMs			
Llama-3.1-8B-Instruct	0.615	0.406	0.326
Qwen2.5-7B-Instruct	0.538	0.282	0.232
Qwen3-30B-A3B-Instruct	**0.635**	**0.469**	**0.342**

Table 1: Comparative validation of retrieval models (D1). Best results in bold.

System	Ann-1	Ann-2	Llama	GPT-OSS	Qwen
Bench_GPT-OSS	**4.475**	3.669	4.101	**3.019**	**4.303**
Bench_Mistral	4.355	**3.799**	**4.634**	2.897	3.990
Raw Baseline	3.709	2.966	3.130	2.951	3.447

Table 2: Mean Likert scores (1–5) for benchmark ground-truth card quality across two domain annotators and three LLM judges on a 100-card stratified sample. **Bold** denotes the best score per column among editor-refined systems.

4 MetaGAI Validation and Analysis

To rigorously validate the benchmark construction pipeline, we conduct human-in-the-loop experimentation in four critical dimensions, substantially expanding beyond prior validation efforts (Liu et al., 2024a) through systematic component-wise evaluation. Our validation framework examines retrieval strategies (**D1**), quantifies semantic variation between ensemble generators (**D2**), measures quality improvements from editor-based consolidation (**D3**), and compares performance between editor architectures (**D4**).

4.1 Retrieval Strategy Validation (D1)

We compared five retrieval models across two paradigms on a human-verified gold standard of 276 unanimously agreed chunk-field decisions (Appendix D.3). As shown in Table 1, discriminative rerankers achieve P@1 < 0.14, indicating that semantic similarity alone is insufficient for schema mapping. Qwen3-30B-A3B-Instruct achieves the best performance (P@1: 0.635, F1@5: 0.342), establishing it as the retrieval backbone.

4.2 Generator Divergence Analysis (D2)

We analyzed semantic consensus within the generator ensemble (OLMo-3-7B, Llama-3.1-8B, Qwen2.5-7B) by computing pairwise BERTScore (Zhang et al., 2020) similarity across all 2,541 samples. Fields clustering in the high-consensus zone ($\mu > 0.8$), such as *Ethical Considerations* ($\mu = 0.848$), predominantly reflect evidence scarcity rather than genuine agreement (Appendix G.4). Conversely, substantial portions of both card types exhibit pronounced divergence ($\sigma > 0.4$), confirming that architecturally distinct models generate complementary details from information-rich contexts and validating our ensemble design. Full mean-variance analysis appears in Appendix G.3.

4.3 Benchmark Ground Truth Quality Assessment

A benchmark is only as trustworthy as its ground truth: evaluation results are meaningful only if the reference cards are of sufficiently high quality that observed effect sizes exceed the noise floor. To directly address this concern, we conducted a controlled human quality study assessing the final editor-refined cards, which constitute the MetaGAI ground truth, against an unedited generation baseline. We compare three outputs on a stratified random sample of 100 cards (50 models, 50 dataset): (1) **Raw Baseline**, the highest-token-count unedited generator draft; (2) **Bench_Mistral**, the output of the Mistral-3-14B-Instruct editor; and (3) **Bench_GPT-OSS**, the output of the GPT-OSS-20B editor. The complete annotation protocol, judge configuration, and inter-annotator agreement are reported in Appendix D.4.

Editor Efficacy (D3). Table 2 reports the mean Likert scores (1–5) in two domain annotators and three LLM judges. Editor-refined ground-truth cards achieve mean human scores of 4.41 (Bench_GPT-OSS: Ann-1 4.48, Ann-2 3.67; Bench_Mistral: Ann-1 4.36, Ann-2 3.80), substantially above the Raw Baseline (3.71 / 2.97). This 15–20% absolute gain on a 5-point scale is well above any noise threshold that would render small effect sizes uninterpretable, confirming that editor consolidation reliably elevates ground-truth quality across all independent evaluators.

Editor Architecture Selection (D4). Bench_GPT-OSS and Bench_Mistral achieve near-equivalent quality: Bench_GPT-OSS leads marginally under Ann-1 (4.48 vs. 4.36), while Bench_Mistral scores highest among LLM judges via Llama-3.3-70B-Instruct (4.63 vs. 4.10); average annotator scores are indistinguishable (4.07

vs. 4.08). Given this parity, we incorporate both architectures and average their outputs as the final benchmark ground truth to ensure architectural neutrality.

5 Experiments

5.1 Experimental Setup

We evaluate all models with zero-shot, generating cards $\hat{\mathcal{C}}$ exclusively from paper text $\mathcal{P}$, consistent with the task definition in Section 3.1. This reflects realistic deployment constraints, as many published artifacts lack repositories or Hugging Face pages entirely (Liang et al., 2024; Yang et al., 2024). To balance statistical rigor with computational feasibility, automated metrics are calculated on the **full benchmark** ($N = 2{,}541$), while LLM-as-a-Judge evaluations use a **stratified random sample** of 500 entries (250 each).

5.2 Baselines

We examine models categorized by access mode and architectural paradigm. Preliminary experiments with models below 14B parameters reveal consistent failures in producing structured outputs; therefore, we restrict evaluation to architectures with 20B+ parameters.

Open-Weight Models. We evaluate two architectural classes. **Dense Models** employ standard transformer architectures (20B–32B parameters), including Mistral-Small-3.2-24B-Instruct, Gemma-3-27B-IT (Team et al., 2025), and Qwen3-32B (Yang et al., 2025). **MoE Models** utilize Mixture-of-Experts architectures providing high parameter counts with efficient active parameter usage, including GPT-OSS-20B (OpenAI et al., 2025), NVIDIA-Nemotron-3-Nano-30B-A3B (NVIDIA et al., 2025), and Qwen3-30B-A3B-Instruct (Yang et al., 2025). **Closed-Source Models.** We include proprietary models accessed via API: GPT-5.1 series (Mini/Nano)[1] and Gemini-2.5 series (Flash/Flash-Lite) (Comanici et al., 2025).

5.3 Evaluation Metrics

We employ a dual-layered evaluation combining quantitative structural metrics with qualitative expert judgment.

Automated Metrics. We measure recall and semantic alignment through three metrics:

- **Completeness:** Quantifies field-level recall. Let $\mathcal{K}(\mathcal{C})$ denote the set of populated keys in card $\mathcal{C}$:

$$\text{Completeness} = \frac{|\mathcal{K}(\hat{\mathcal{C}}) \cap \mathcal{K}(\mathcal{C}_{GT})|}{|\mathcal{K}(\mathcal{C}_{GT})|} \quad (5)$$

- **Semantic Similarity:** We report **ROUGE-L** (Lin, 2004) for lexical overlap and **BERTScore (F1)** (Zhang et al., 2020) for semantic alignment.

LLM-as-a-Judge Evaluation. To capture nuances beyond n-gram matching, we implement an ensemble framework using GPT-OSS-120B (OpenAI et al., 2025), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), and Qwen3-235B-A22B-2507 (Yang et al., 2025). Following the protocol in Appendix D (Prompt F.5), judges evaluate the generated fields on a 1–5 Likert scale. Scores are averaged among judges to mitigate bias in the single-model (agreement analysis in Appendix E).

Cost Efficiency. To assess economic feasibility at scale, we introduce a Cost Index that estimates the inference cost per card generation task. The index normalizes costs across models by standardizing token consumption to 1M input tokens and 0.2M output tokens, enabling direct price comparison independent of model-specific tokenization efficiency. Pricing is derived from standardized rates on OpenRouter[2], facilitating economic comparison between proprietary and open-weight models.

5.4 Experimental Results

We establish a comprehensive evaluation protocol combining automated metrics in 2,541 entries with LLM-as-a-Judge evaluation in 500 samples. Beyond quality assessment, we introduce cost-efficiency analysis, revealing that sparse MoE models achieve optimal cost-quality trade-offs, while traditional lexical metrics inversely correlate with semantic quality. Table 3 presents comprehensive evaluation results in all baseline models. We analyze performance through architectural efficiency, evaluation metric validity, and cognitive limitations.

5.4.1 Architectural Efficiency and Economic Viability

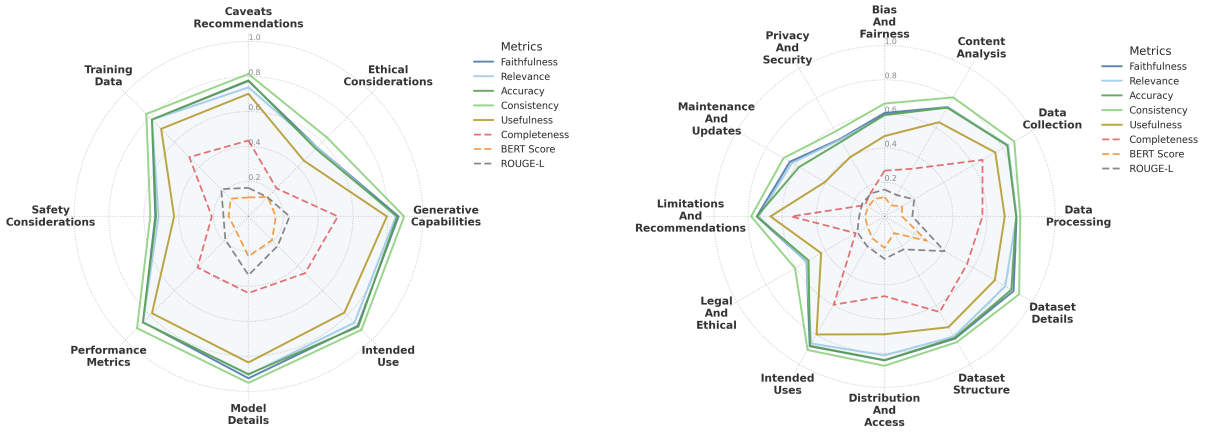
MoE Superiority with Scale Threshold. Sparse MoE architectures demonstrate superior parameter efficiency, but only at sufficient scale. Within

[1] https://openai.com/index/gpt-5-1/

[2] https://openrouter.ai/

Model	Faithfulness		Relevance		Accuracy		Consistency		Usefulness		Qual. Avg		Comp.		BScore		RL		Cost
	D	M	D	M	D	M	D	M	D	M	D	M	D	M	D	M	D	M	
Dense Models																			
Mistral-Small-24B	3.39	3.75	3.28	3.59	3.33	3.70	3.64	3.97	2.67	3.14	3.26	3.63	.280	.327	**.193**	.194	**.270**	.255	0.10
Gemma-3-27B	3.81	3.97	3.76	3.94	3.76	3.96	4.03	4.23	3.26	3.58	3.72	3.94	.464	.734	.151	.151	.214	.203	0.07
Qwen3-32B	3.40	3.72	3.29	3.58	3.32	3.69	3.66	3.94	2.65	3.10	3.26	3.60	.386	.513	.184	**.198**	.267	**.265**	0.13
MoE Models																			
GPT-OSS-20B	3.18	3.49	3.09	3.36	3.11	3.45	3.45	3.74	2.49	2.87	3.06	3.38	.394	.492	.146	.133	.235	.220	**0.06**
Nemotron-Nano-30B-A3B	3.45	3.80	3.41	3.76	3.39	3.77	3.71	4.03	2.88	3.35	3.37	3.74	.557	.644	.124	.133	.199	.206	0.11
Qwen3-30B-A3B-Instruct	**4.33**	**4.50**	**4.36**	**4.57**	**4.31**	**4.51**	**4.46**	**4.67**	**4.06**	**4.49**	**4.30**	**4.55**	**.702**	**.786**	.169	.174	.246	.243	0.15
Closed-Source Models																			
GPT-5-Mini	4.14	4.32	4.09	4.26	4.10	4.32	4.25	4.43	3.74	4.07	4.06	4.28	.556	.216	.102	.117	.185	.207	0.65
GPT-5-Nano	3.18	3.38	3.06	3.18	3.10	3.32	3.46	3.57	2.53	2.73	3.06	3.23	.383	.127	.113	.149	.188	.228	0.13
Gemini-2.5-Flash	3.88	4.08	3.80	3.97	3.82	4.06	4.04	4.19	3.27	3.69	3.76	4.00	.443	.181	.160	.170	.241	.246	0.80
Gemini-2.5-Flash-Lite	3.93	4.10	3.88	4.03	3.87	4.08	4.12	4.25	3.39	3.69	3.83	4.03	.511	.194	.140	.144	.207	.208	0.18

Table 3: **MetaGAI Benchmark Comprehensive Results.** Performance across Data Card (D) and Model Card (M) generation. Completeness, BERTScore, and ROUGE-L are reported on the full test set (N=2,541). Qual. Avg (Qualitative Average) is evaluated on a sample of 500 entries. **Bold**/Underline indicate best/second-best performance.



(a) Model Card Fields

(b) Data Card Fields

Figure 3: **Field-Level Performance Patterns Averaged Across All Baselines.** Evaluation metrics (colored lines) across card fields (axes). Performance is strong on signal-rich fields (*Model Details*) but degrades on abstract categories (*Ethical Considerations*), revealing systematic generation difficulty when documentation is sparse.

the Qwen family, MoE-based Qwen3-30B-A3B-Instruct (4.55 Model Card quality) outperforms dense Qwen3-32B (3.60) by 0.95 points despite comparable parameter counts. Given shared pre-training lineages, this gap validates that sparse activation enables more effective information synthesis. However, this advantage requires approximately 30B parameters: the smallest MoE (GPT-OSS-20B) achieves the lowest performance (3.06/3.38), underperforming comparable dense models. This establishes a minimum viable scale threshold for MoE benefits in complex generation tasks.

Open-Weight Cost Dominance. Qwen3-30B-A3B-Instruct occupies the Pareto-optimal position on the cost-quality frontier. It delivers state-of-the-art performance (4.55 quality score) at the lowest normalized cost (Cost Index: 0.15), outperforming GPT-5-Mini (4.28 quality, Cost Index: 0.65) by 4.3× in cost efficiency. Gemini-2.5-Flash presents the least favorable value proposition: inferior quality (4.00) at the highest normalized cost (Cost Index: 0.80). Detailed analysis in Appendix G.6 confirms that optimized open-weight MoE architectures provide economically superior solutions for production-scale card generation.

5.4.2 Evaluation Metric Validity

Lexical Similarity Metrics Invalidation. Traditional content matching metrics exhibit an in-

verse correlation with semantic quality. ROUGE-L demonstrates this paradox: Mistral-Small-24B achieves the highest ROUGE-L (0.270) through naive verbatim copying, yet produces one of the lowest quality (3.26), while Qwen3-30B-A3B-Instruct yields lower ROUGE-L (0.246) through abstractive synthesis but superior quality (4.30). BERTScore similarly fails as a quality discriminator, compressing all models into a narrow 0.10–0.20 range despite 1.5-point quality differences (3.06–4.55). These metrics penalize good abstractive synthesis that necessarily diverges from the source text.

Completeness-Quality Orthogonality. The structural coverage and semantic quality represent independent dimensions. On Model Cards, Gemma-3-27B achieves high Completeness (0.734, second-best) yet scores only 3.94 in quality, substantially below Qwen3-30B-A3B-Instruct (4.55) despite comparable Completeness (0.786). Similarly, on Data Cards, Nemotron-Nano-30B-A3B achieves 0.557 Completeness yet scores only 3.37 in quality, far below Qwen3-30B-A3B-Instruct (4.30) despite the latter's higher Completeness (0.702). This decoupling reveals that high field coverage does not guarantee semantic quality without refinement. Figure 3 illustrates field-level performance patterns: averaged across all baselines, models achieve greater than 0.7 Completeness in explicit fields (*Model Details*, *Performance Metrics*) containing abundant signals, but drop below 0.5 in abstract categories (*Ethical Considerations*, *Maintenance*) requiring inference from sparse contexts. Appendix G.4 shows that ground truth information density explains 31% of Completeness variance, confirming source sparsity as the primary driver of generation difficulty. These findings validate our multi-metric framework that combines automated structural diagnostics with LLM-as-a-Judge semantic evaluation.

5.4.3 Systematic Cognitive Limitations

Universal Data Card Difficulty. All models show consistent performance degradation on Data Cards versus Model Cards (average gap: 0.2–0.4 points), from weakest (GPT-OSS-20B: 3.06 vs 3.38) to strongest (Qwen3-30B-A3B-Instruct: 4.30 vs 4.55). Data Card schemas demand high-granularity lifecycle documentation (privacy protocols, security measures, maintenance plans), whereas papers treat datasets as ancillary artifacts with sparse experimental descriptions. This systematic scarcity imposes the need to synthesize complete profiles from limited signals, in which current models show consistent degradation in completeness and quality metrics.

Faithfulness-Completeness Trade-off. Qwen3-30B-A3B-Instruct achieves near-perfect Faithfulness, avoiding hallucinations, yet exhibits only 0.786 Completeness, systematically omitting 21% of ground truth fields. This precision-recall imbalance aligns with documented long-context retrieval limitations (Liu et al., 2024b), where models struggle to access dispersed information. To disentangle whether this gap stems from source sparsity or model inferential capacity, we conducted an ablation study (Appendix G.2) providing the model with full multi-source context during inference. The results show negligible improvement ($\Delta < 0.02$ among reliable judges), confirming that the completeness bottleneck is attributable to long-context inferential limitations rather than restricted access to evidence. Linguistically, outputs show high abstraction with hedging phrases and cross-references contrasting with the ground truth's concrete specifications. Appendix G.5 quantifies these divergences through Log-Odds Ratio analysis, revealing systematic narrative bias.

6 Conclusion

We introduce MetaGAI, a large-scale benchmark comprising 2,541 validated entries constructed through multi-source triangulation and a multi-agent framework. Unlike prior work that aggregates existing documentation, our approach constructs ground truth through rigorous semantic verification and four-dimensional human-in-the-loop validation, addressing incompleteness and inconsistency in human-authored cards. Our experiments reveal that sparse MoE architectures achieve optimal cost-quality performance, while traditional lexical metrics inversely correlate with semantic quality. A fundamental faithfulness-completeness trade-off emerges where models systematically omit fields despite maintaining high factual accuracy. MetaGAI provides a critical infrastructure to advance transparency and reproducibility in GenAI documentation on a production scale.

Limitations

While MetaGAI provides a strong baseline for automated Model and Data Card generation, it remains limited in scope. Our current framework relies on text-only generation, which overlooks important information embedded in figures, tables, and other non-textual modalities. In addition, MetaGAI treats document generation as isolated paper-level tasks, without modeling the complex dependencies among papers, models, and datasets in the broader GenAI ecosystem. Addressing multimodal content and capturing ecosystem-level relationships through structured or graph-based representations are promising directions for future work. Beyond these technical limitations, automated documentation also introduces potential risks. Errors or omissions in generated cards may propagate misleading signals about model capabilities, data provenance, or licensing conditions, especially when such artifacts are reused at scale. Without careful validation and human oversight, these inaccuracies could undermine transparency efforts or lead to misplaced trust. Addressing multimodal understanding, ecosystem-level modeling, and robust verification mechanisms are important directions for future work.

Ethics Statement

We introduce MetaGAI to improve transparency and reproducibility in GenAI through automated documentation. We address the following ethical considerations:

Data Provenance and Licensing. Our benchmark triangulates public data from arXiv, GitHub, and Hugging Face, strictly adhering to each platform's terms of use. Source materials are used exclusively for scientific research and card generation. No private information beyond publicly associated author names is collected.

Human Evaluation. The validation involved volunteer Ph.D. students in NLP and machine learning. All evaluators were informed about the nature of the task, with a reasonable workload and no exposure to harmful content.

Risks of Automated Documentation. We acknowledge the inherent risks of hallucination and omission in LLM-based documentation generation and address them at multiple levels. At the system level, the Generator and Editor Agents produce, for every field, supporting evidence spans, confidence scores, and explicit source attribution, with the Editor verifying attribution against raw source chunks across three independently generated candidates. At the evaluation level, our faithfulness–completeness analysis directly quantifies these failure modes, identifying which fields and models are most susceptible and require prioritized human verification. At the deployment level, generated cards should be treated as structured preliminary drafts subject to human oversight, not as authoritative replacements for expert review.

Environmental Considerations. Although the benchmark construction used large-scale LLMs, our cost-quality analysis demonstrates that sparse MoE architectures achieve high-quality generation with substantially lower computational requirements, helping reduce energy consumption in large-scale documentation efforts.

References

Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. 2024. Analyzing the evolution and maintenance of ml models on hugging face. In *Proceedings of the 21st International Conference on Mining Software Repositories*, MSR '24, page 607–618, New York, NY, USA. Association for Computing Machinery.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2025. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

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

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 427–439, New York, NY, USA. Association for Computing Machinery.

Andy Donald, Apostolos Galanopoulos, Edward Curry, Emir Muñoz, Ihsan Ullah, M. A. Waskow, Maciej Dabrowski, and Manan Kalra. 2023. Towards a semantic approach for linked dataspace, model and data cards. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1468–1473, New York, NY, USA. Association for Computing Machinery.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xiaojie Guo and Qiming Hu. 2022. Low-light image enhancement via breaking down the darkness. *Int. J. Comput. Vision*, 131(1):48–66.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. 2025. We should chart an atlas of all the world's models. *Preprint*, arXiv:2503.10633.

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):744–753.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona Diab. 2024a. 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)*, pages 1975–1997, Mexico City, Mexico. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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

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

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

NVIDIA, :, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, et al. 2025. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *Preprint*, arXiv:2512.20848.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, et al. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *Preprint*, arXiv:2502.18443.

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*, FAccT '22, page 1776–1826, New York, NY, USA. Association for Computing Machinery.

Mohammad Shahedur Rahman, Peng Gao, and Yuede Ji. 2025. Hugginggraph: Understanding the supply chain of llm ecosystem. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 5997–6005, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, et al. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xinyu Yang, Weixin Liang, and James Zou. 2024. Navigating dataset documentations in AI: A large-scale analysis of dataset cards on huggingface. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2022. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*.

A Definition of Model and Data Card

The definitions of Model card and Data card for GenAI are shown in Table 4.

B Dataset Characteristics

Figure 4 (Top) illustrates the temporal distribution of our corpus, which shows an increase from 141 triplets in 2022 to 931 in 2025. The domain composition is dominated by Computer Vision (`cs.CV`) and Computational Linguistics (`cs.CL`), reflecting our filtering criteria targeting papers with verified GitHub and Hugging Face artifacts in generative modeling. The word count analysis (Bottom) reveals distinct content characteristics across sources: academic papers provide a comprehensive context (median ≈10k words), while GitHub and Hugging Face README files offer concise implementation details (median <1k words). This complementary information structure empirically justifies our multi-source triangulation approach.

C MetaGAI Benchmark Generation Algorithm

The Algorithm 1 formalizes the three-stage MetaGAI generation pipeline. For each schema key $k_i \in \mathcal{K}$, the Retriever Agent maps segmented source chunks $\mathcal{X}$ to relevant evidence spans, which are passed to the Generator Agent to synthesize an independent structured draft. The Editor Agent then consolidates each draft by cross-validating against the full multi-source context $\mathcal{P} \cup \mathcal{S}$, verifying attribution, and merging complementary content to produce the final card $\hat{\mathcal{C}}$.

Algorithm 1 MetaGAI Benchmark Generation Pipeline

Require: $\mathcal{P} \cup \mathcal{S}$: Full Context (Paper, GitHub, Hugging Face) $\mathcal{X}$: Segmented chunks of $\mathcal{P} \cup \mathcal{S}$ $\mathcal{K}$: Schema Keys $\{k_i\}_{i=1}^{M}$
Ensure: $\hat{\mathcal{C}}$: Final Card

Stage 1: Retriever Agent (Evidence Mapping)
$\mathcal{E} \leftarrow \emptyset$
for each $k_i \in \mathcal{K}$ **do**
  $v_{evidence} \leftarrow \text{Retriever}(\mathcal{X}, k_i)$ ▷ Extract factual content
  $\mathcal{E}[k_i] \leftarrow v_{evidence}$
end for

Stage 2: Generator Agent (Draft Synthesis)
$\tilde{\mathcal{C}} \leftarrow \emptyset$
for each $k_i \in \mathcal{K}$ **do**
  $\tilde{v}_i \leftarrow \text{Generator}(\mathcal{E}[k_i], \text{Prompt}_{\text{Gen}})$ ▷ Synthesize structured draft
  $\tilde{\mathcal{C}}[k_i] \leftarrow \tilde{v}_i$
end for

Stage 3: Editor Agent (Consolidation)
for each $k_i \in \mathcal{K}$ **do**
  ▷ Editor verifies draft against full context
  $\hat{v}_i \leftarrow \text{Editor}(\mathcal{P} \cup \mathcal{S}, \tilde{\mathcal{C}}[k_i], \text{Prompt}_{\text{Edit}})$
  $\hat{\mathcal{C}}[k_i] \leftarrow \hat{v}_i$
end for
return $\hat{\mathcal{C}}$

D MetaGAI Quality Validation Protocols

To rigorously assess generated card quality, we employ a comprehensive evaluation framework. This framework adopts three established metrics (*Faithfulness*, *Relevance*, and *Accuracy*) from previous work (Liu et al., 2024a), and introduces two additional metrics (*Consistency* and *Usefulness*) to address the structural and practical requirements of technical documentation.

D.1 Metric Definitions

Faithfulness (F). Measures the degree to which generated content is grounded in provided source

Card	Field	Description
Model Card	Model Details	Model architecture, parameter count, developer organization, license, release date, and other model metadata
	Intended Use	Primary applications, target audience, out-of-scope uses, domain restrictions, and other intended uses
	Generative Capabilities	Context length, inference latency, multilingual support, supported modalities, and other capabilities
	Safety Considerations	Safety alignment methods, red teaming results, jailbreak resistance, harm reduction strategies, and other safety analysis
	Training Data	Training corpus, data mixture, filtering pipeline, data provenance, and other training data information
	Performance Metrics	Benchmark results, accuracy metrics, reasoning scores, safety scores, robustness metrics, and other quantitative metrics
	Ethical Considerations	Environmental impact, intellectual property, dual-use risks, societal implications, and other ethical issues
	Caveats & Recommendations	Hardware requirements, deployment guidelines, operational constraints, model limitations, and other recommendations
Data Card	Dataset Details	Dataset name, version identifier, creators and curators, funding, license, text language, and other dataset metadata
	Dataset Structure	Instance count, field schema, data splits, dataset size, and other structural details
	Data Collection	Collection methodology, data sources, collection timeframe, consent process, and other collection details
	Data Processing	Preprocessing steps, cleaning procedures, labeling process, filtering criteria, deduplication, and other processing steps
	Intended Uses	Primary tasks, intended use cases, prohibited uses, commercial restrictions, and other usage information
	Bias & Fairness	Demographic distribution, geographic coverage, social bias analysis, fairness assessment, and other bias or fairness information
	Privacy & Security	Personally identifiable information (PII), anonymization methods, data security protocols, confidentiality measures, and other privacy measures
	Content Analysis	Content types, toxicity analysis, misinformation risks, offensive language, and other content risks
	Legal & Ethical	Copyright status, terms of use, ethical review, compliance requirements, and other legal details
	Maintenance & Updates	Maintenance plan, update frequency, versioning policy, deprecation plan, and other maintenance information
	Distribution & Access	Access mechanism, download instructions, repository link, citation requirements, and other access details
	Limitations & Recommendations	Known limitations, usage guidelines, quality caveats, and other recommendations

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

materials (paper, GitHub, Hugging Face), ensuring absence of unsupported claims or hallucinations.

Relevance (R). Evaluates the information density and pertinence to the target field. High-scoring content strictly adheres to field definitions, avoiding redundancy, verbosity, or out-of-scope information.

Accuracy (A). Assesses factual correctness of technical details (numerical values, licenses, entity names). Unlike Faithfulness, which verifies source alignment, Accuracy validates objective correctness against domain knowledge.

Consistency (C). Evaluates internal logical coherence across sentences and fields within a card, detecting contradictions that compromise documentation integrity.

Usefulness (U). Measures practical value for downstream users. Useful content provides specific, actionable insights that support deployment decisions, rather than generic descriptions.

D.2 Scoring Rubric

Each metric is evaluated on a 1–5 Likert scale. Table 5 presents the scoring criteria applied by both human evaluators and LLM judges.

D.3 Retrieval Strategy Validation: Experimental Details

Ten randomly sampled entries were segmented and evaluated by two domain experts. From 6,116 chunk-field decisions, we retained 276 chunks achieving unanimous consensus as ground truth. We compared BGE-Reranker-v2-m3 (Chen et al., 2025), Qwen3-Reranker-8B (Zhang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct, and Qwen3-30B-A3B-Instruct (Yang et al., 2025). We report Preci-

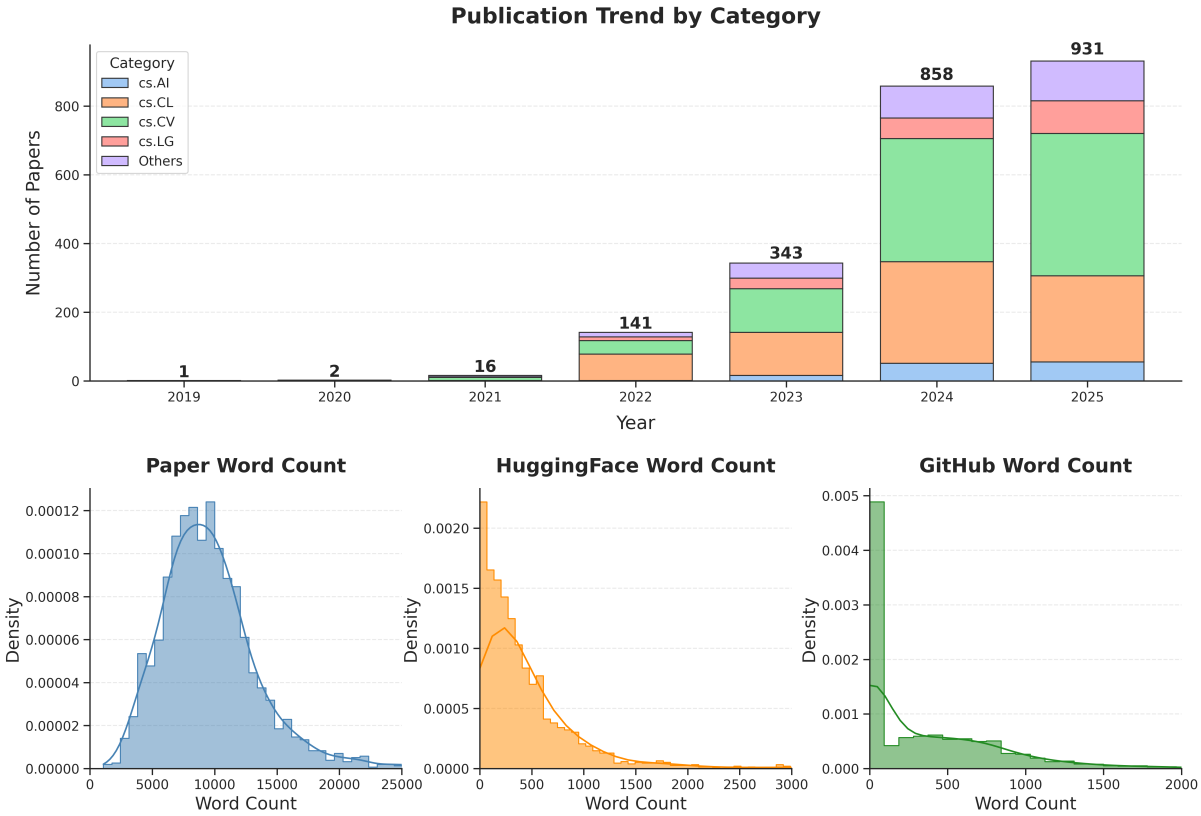


Figure 4: **Dataset Overview.** *Top:* Publication trends of the 2,541 collected triplets (2019–2025), categorized by primary arXiv domain. *Bottom:* Word count distributions across three data sources (Academic Papers, Hugging Face, GitHub), illustrating complementary information granularities.

sion at rank 1 (P@1) to measure top-result accuracy, alongside Recall and F1 at rank 5 (R@5, F1@5) to assess broader retrieval coverage. The results are shown in Table 1.

D.4 Human Annotation Protocol for Benchmark Quality Assessment

Annotators and Sample. Two Ph.D. students specializing in NLP (Ann-1, Ann-2) independently annotated a stratified random sample of 100 cards (50 Model Cards, 50 Data Cards). The cards were presented in random order under blind conditions.

Evaluation Instrument. Each card was scored on a 1–5 Likert scale in five quality dimensions defined in Appendix D: Faithfulness, Relevance, Accuracy, Consistency, and Usefulness. Annotators were provided with detailed rubrics and calibration examples prior to the main annotation task.

LLM Judge Configuration. Three LLM judges –Llama-3.3-70B-Instruct (Grattafiori et al., 2024), GPT-OSS-120B (OpenAI et al., 2025), and Qwen3-235B-A22B-2507 (Yang et al., 2025) – evaluated the same outputs under randomized presentation order. Each judge was queried three times per instance with independently shuffled orderings; scores were averaged across runs to mitigate position bias.

Inter-Annotator and Human-LLM Agreement. Table 6 reports Pearson correlation coefficients at the instance level (per card $\times$ system pair). The human inter-annotator agreement is moderate ($r = 0.44$, $p < 0.001$), consistent with the documented subjectivity of documentation quality judgments. Among LLM judges, Qwen3-235B and Llama-3.3-70B achieve a strong and significant alignment with Ann-1 ($r = 0.70$ and $r = 0.67$; both $p < 0.001$) and moderate significant alignment with Ann-2 ($r = 0.37$ and $r = 0.49$; both $p < 0.001$). GPT-OSS-120B does not achieve a significant correlation with Ann-1 ($r = 0.09$, $p = 0.054$, n.s.) and shows only weak agreement with Ann-2 ($r = 0.19$, $p < 0.001$), motivating the exclusion of any single-judge design and supporting our ensemble evaluation strategy.

Metric	Score 1 (Poor)	Score 3 (Acceptable)	Score 5 (Excellent)
Faithfulness	Contains major hallucinations; claims unsupported by sources.	Mostly supported with minor extrapolations.	Fully grounded in source text; no unsupported claims.
Relevance	Largely off-topic, redundant, or contains significant noise.	Generally on-topic with some redundant details.	Concise and strictly focused on field definition.
Accuracy	Contains critical factual errors.	Core facts correct; minor imprecisions present.	Factually precise; all technical details correct.
Consistency	Contains direct logical contradictions.	Generally consistent; minor ambiguities present.	Logically coherent; no internal contradictions.
Usefulness	Vague or generic; provides no practical value.	Provides basic information but lacks depth.	Rich, actionable insights supporting deployment.

Table 5: **Evaluation Scoring Rubric.** Criteria for assessing generated metadata cards on a 1–5 Likert scale across five quality dimensions.

Target Judge	Ann-1	Ann-2
Ann-2	0.439^*	—
Llama-3.3-70B	0.675^*	0.493^*
GPT-OSS-120B	0.087 (n.s.)	0.191^*
Qwen3-235B	0.703^*	0.367^*

Table 6: Instance-level Pearson correlations between domain annotators and LLM judges. $^*p < 0.001$; n.s. $p \geq 0.05$.

Judge Model	Llama	GPT	Qwen
Llama-3.3-70B-Instruct	1.000	0.101^*	0.540^*
GPT-OSS-120B	0.101^*	1.000	0.226^*
Qwen3-235B-A22B-2507	0.540^*	0.226^*	1.000

Table 7: **Inter-Judge Correlation Matrix.** Pearson correlations computed on 500 samples. All correlations are statistically significant ($^*p < 0.001$). **Llama**: Llama-3.3-70B-Instruct, **GPT**: GPT-OSS-120B, **Qwen**: Qwen3-235B-A22B-2507.

E Inter-Judge Agreement Analysis

To validate the robustness of our LLM-as-a-Judge framework, we computed Pearson correlation coefficients between the three judge models in the stratified evaluation set ($N = 500$). Table 7 reveals substantial variation between judges. While Qwen3-235B-A22B-2507 and Llama-3.3-70B-Instruct demonstrate moderate alignment ($\rho = 0.540$), GPT-OSS-120B exhibits markedly different scoring patterns (correlations ≤ 0.226). All observed correlations achieve statistical significance ($p < 0.001$), indicating that the divergence of the score reflects systematic architectural differences in quality assessment rather than random measurement errors. This architectural diversity justifies our ensemble averaging strategy: aggregating judgments across models with distinct evaluation perspectives mitigates single-model biases and provides more robust quality estimates than any individual judge.

F Prompt Templates

F.1 Semantic Verification Prompt

Semantic Verification Prompt

System Instruction
You are an expert research evaluator specializing in [dataset/model] identification.

Task Description
Determine whether a research paper introduction and two README files (GitHub and Hugging Face) describe the **same** [dataset/model].

Evaluation Criteria

1. **Existence**: Does the paper explicitly introduce a specific named entity?
2. **Consistency**: Do BOTH the GitHub README and Hugging Face README reference the SAME entity as the paper?
3. **Alignment**: Compare names, domain, size metrics, methodology, and unique features.

Input Data
Paper Content: <*Paper text*>
GitHub README: <*GitHub README text*>
Hugging Face README: <*HF README text*>

Output Format
RELATED: [Yes/No]
CONFIDENCE: [High/Medium/Low]
EXPLANATION: [2-3 sentences citing specific evidence]

F.2 Retriever Agent Prompt

Retriever Agent Prompt

System Instruction
You are a metadata classification assistant. Output valid JSON only. STRICTLY select sub-fields from the provided ontology.

Task Description
Classify the provided text chunk into predefined metadata fields for [Dataset/Model] documentation.

Evaluation Criteria

1. **Relevance**: Assign a score (0–4) indicating how well the chunk describes the field.
2. **Sub-field Selection**: Select 1–5 sub-fields **strictly** from the provided "Selectable Sub-fields" list. Do not extract raw text.

Input Data
Source: *<[PAPER] / [GITHUB] / [HUGGINGFACE]>*
Content: *<Text chunk content>*

Output Format

```
{ "classifications": [ {
"field": "field_name",
"relevance": 3,
"matched_sub_fields": ["sub_field1"] } ] }
```

F.3 Generator Agent Prompt

Generator Agent Prompt

System Instruction
You are an expert AI Researcher. Generate specific metadata sections for a [Model/Dataset] Card. Output valid JSON only.

Critical Constraint: Subject Focus

- **Focus**: Solely on the entity **introduced** in this paper.
- **Ignore**: Baselines, pre-training models, or comparisons.

Input Data
Target: *<Field Name>* (e.g., *Training Data*) | **Context**: *<Retrieved chunks>*

Output Requirements (Per Sub-field)

1. **Content**: Summarized factual answer.
2. **Evidence Quote**: Direct verbatim quote supporting the answer.
3. **Confidence**: [low, medium, high, certain].
4. **Source**: Provenance (e.g., "Paper+GitHub").

Output Format

```
{ "sub_field": {
"content": "...",
"evidence_quote": "...",
"confidence": "high",
"source": "Paper" } }
```

F.4 Editor Agent Prompt

Editor Agent Prompt

System Instruction
You are the Chief Editor. Consolidate 3 candidate drafts into ONE **concise** entry describing the **proposed** entity.

Task Description
Review three independent drafts (Candidates A, B, C) against the source ground truth. Identify the correct **proposed** entity, filter out hallucinations, and merge valid details.

Evaluation Logic

1. **Identify Proposed Entity**: Describe ONLY the entity introduced in THIS paper. Discard candidates describing baselines.
2. **Verify Evidence**: Check attribution.
3. **Conciseness**: Remove fluff; use direct facts (1–3 sentences).

Input Data
Target: *<Field Name>* | **Ground Truth**: *<Raw chunks>*
Candidates: *<Draft A>*, *<Draft B>*, *<Draft C>*

Output Format

```
{
"selected_candidates": ["Candidate B"],
"final_content": "...",
"final_evidence": "...",
"reasoning": "..." }
```

F.5 Evaluation Judge Prompt

Evaluation Judge Prompt

System Instruction
You are an expert evaluator for Generative AI documentation. Your task is to compare the generated metadata card against the provided triangulated sources (Paper, GitHub, Hugging Face).

Task Description
Evaluate the candidate text based on the five dimensions below. Assign a score from 1 (Poor) to 5 (Excellent) for each dimension.

Evaluation Metrics

- **Faithfulness (F)**
- **Relevance (R)**
- **Accuracy (A)**
- **Consistency (C)**
- **Usefulness (U)**

Analysis Considerations
Classify content into: Accurate facts (verifiable), Vague facts, Logical reasoning, Illogical reasoning, or Acceptable inferences.

Input Data
`Sources: <Context>; Target Field: <Field Name>; Candidate: <Text>`

Output Format

```
{"scores": {"F": 5, "R": 5, "A": 5, "C": 5,
"U": 5}, "reasoning": "..."}
```

Source Category	Bench_Mistral		Bench_GPT-OSS	
	Model	Data	Model	Data
Paper Only	65.9%	61.9%	57.1%	45.1%
Multi-Source	22.4%	26.7%	22.2%	30.7%
GitHub Only	3.5%	3.3%	3.4%	3.9%
Not Provided	8.2%	8.1%	17.4%	20.3%

Table 8: **Information Provenance Distribution.** Percentage of metadata fields derived from single vs. multiple sources across the two benchmark subsets.

	Bench_Mistral	Bench_GPT-OSS
Consolidation Strategy		
Best Match	80.0%	94.9%
Merge	20.0%	5.1%
Winning Candidate Source		
Qwen2.5-7B	64.6%	63.8%
Olmo3-7B	27.8%	30.3%
Llama-3.1-8B	7.6%	5.9%

Table 9: **Editor Agent Dynamics.** *Top:* The logic used by the Editor to finalize content. *Bottom:* The distribution of Generator drafts selected as the final ground truth.

G Comprehensive Analysis of Benchmark Characteristics and Baseline Behaviors

To understand the MetaGAI benchmark's inherent challenges and model behaviors, we conduct a granular analysis covering information provenance, information sparsity impact, comparative lexical quality, and cost-efficiency trade-offs.

G.1 Information Provenance and Editor Dynamics

A central design principle of MetaGAI is that high-fidelity card generation requires multi-source triangulation. We validate this through provenance analysis of final ground-truth fields.

Source Distribution. Table 8 demonstrates that while academic papers provide the majority of information (45–66%), a substantial proportion of metadata fields (22–31%) requires multi-source triangulation across paper, GitHub, and Hugging Face artifacts. Data Cards exhibit higher reliance on multi-source integration (27–31%) compared to Model Cards (22%), confirming that dataset documentation is typically fragmented across theoretical descriptions in papers and implementation details in repositories, necessitating cross-source synthesis.

Editor Strategy. Table 9 reveals different consolidation patterns. The Best Match strategy dominates (80–95%), while the Merge strategy contributes 5–20% depending on the editor architecture. This demonstrates that achieving comprehensive ground truth for certain fields necessitates synthesizing complementary information across multiple generator drafts rather than selecting a single best candidate.

G.2 Ablation Study: Source Completeness vs. Model Inferential Limits

The faithfulness–completeness trade-off raises a fundamental question: does the systematic gap in field coverage arise from the model's inferential capacity, or is it an artifact of restricting inference to paper-only context while the ground truth was constructed from all three sources? To disentangle these factors, we evaluate Qwen3-30B-A3B-Instruct (the top-performing model on the main benchmark) under two conditions on the identical set of 100 cards used in Section 4.3:

- **Paper-only**: standard evaluation protocol, conditioned solely on paper text P, consistent with the main benchmark conditions.
- **All-source**: full multi-source context $P \cup S$, where S comprises the GitHub README and Hugging Face artifact, constituting the same evidence available during ground-truth construction.

Results and Implications. Table 10 presents the mean Likert scores under both conditions alongside the editor-refined benchmarks for reference. Providing a full multi-source context yields only marginal and inconsistent variation: Ann-1 shifts by $+0.018$, Ann-2 by -0.021, and the Qwen judge by -0.078, with no judge registering a meaningful improvement attributable to the additional context. Crucially, the all-source condition continues to underperform both Bench_GPT-OSS and Bench_Mistral under domain annotators (4.376 vs. 4.475 / 4.355), despite enjoying equivalent evidential access during inference.

These results indicate that the completeness bottleneck is not an artifact of source sparsity in the evaluation setting, but is instead attributable to the model's long-context inferential capacity. In particular, the model struggles to locate, integrate, and populate sparse or implicitly stated metadata fields across heterogeneous multi-document inputs, consistent with documented limitations in dispersed-

	Ann-1	Ann-2	Llama	GPT-OSS	Qwen
Editor-Refined Benchmarks (Reference)					
Bench_GPT-OSS	4.475	3.669	4.101	3.019	4.303
Bench_Mistral	4.355	3.799	4.634	2.897	3.990
Qwen3-30B-A3B-Instruct (Ablation)					
Paper-only	4.358	3.456	3.770	3.061	4.304
All-source	4.376	3.435	3.946	2.726	4.226
Δ	$+0.018$	-0.021	$+0.176$	-0.335	-0.078
Raw Baseline	3.709	2.966	3.130	2.951	3.447

Table 10: Ablation on source context. Qwen3-30B-A3B-Instruct is evaluated under paper-only and full multi-source conditions on the same 100-card test set used in Section 4.3. Editor benchmarks and Raw Baseline are shown for reference. The near-zero Δ across reliable judges localises the completeness bottleneck to model inferential capacity rather than source sparsity.

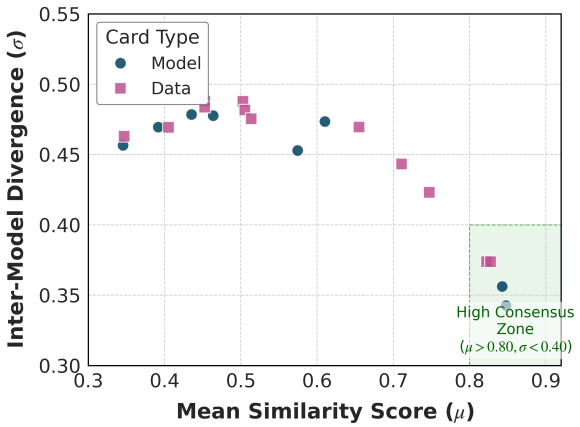


Figure 5: Mean-Variance analysis of semantic alignment. Mean BERTScore similarity (μ, X-axis) versus variance (standard deviation σ, Y-axis) across card fields. High-consensus zones reflect evidence scarcity with converged placeholder responses.

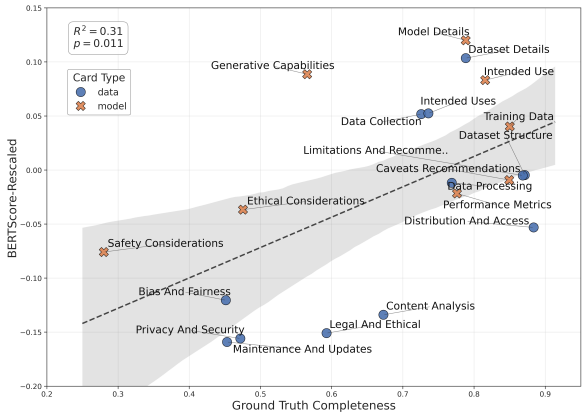


Figure 6: **Information Sparsity Impact on Generation Difficulty.** The relationship ($R^2 = 0.31, p = 0.011$, OLS linear regression) between ground truth information density (X-axis, measured by completeness) and generation difficulty (Y-axis, measured by BERTScore as a diagnostic indicator). Fields with high sparsity (lower completeness) result in significantly lower generation success, highlighting the challenge of inferring abstract metadata from sparse signals.

information retrieval from extended contexts. Consequently, the paper-only evaluation protocol constitutes a faithful and reproducible measure of model capability, rather than an artificially constrained one.

G.3 Generator Divergence: Full Mean-Variance Analysis

Figure 5 presents the full mean-variance distribution across all fields of cards. Fields such as *Ethical Considerations* ($\mu = 0.848$) cluster in the high-consensus zone, predominantly reflecting evidence scarcity rather than genuine agreement: when documentation is absent from source materials, generators converge on similar placeholder responses (Appendix G.4). Fields with pronounced divergence ($\sigma > 0.4$) confirm that architecturally distinct generators surface complementary details from information-rich contexts, validating the ensemble design.

G.4 Impact of Information Sparsity

To explain the field-level performance stratification observed in Figure 3, we analyze the relationship between the density of the source text information and the generation performance using BERTScore as a diagnostic indicator of the difficulty of generation.

We hypothesize that the difficulty of generating abstract fields (*Safety Considerations*, *Bias And Fairness*) stems primarily from information sparsity in ground truth sources. High-density fields such as *Model Details* provide abundant explicit signals, enabling high semantic alignment. Conversely, abstract fields contain sparse or implicit

information, challenging models to infer unstated constraints without generating hallucinations.

Figure 6 demonstrates a statistically significant positive relationship ($R^2 = 0.31, p = 0.011$) between the completeness of the ground truth (information density) and semantic similarity (BERTScore), indicating that 31% of the variance of BERTScore is explained by the completeness of the source. This diagnostic analysis confirms that a primary bottleneck for current LLMs is the inferential capacity required to populate sparse metadata fields from weak textual signals, rather than simple retrieval failure.

G.5 Comparative Lexical Analysis

We perform lexical analysis using the Log-Odds Ratio (Monroe et al., 2008) to assess the differences in linguistic quality between the generated content and ground truth. We focus on *Model Details* and *Dataset Details* fields, which require complex generation of open-ended, high-density technical information.

Figure 7 reveals a systematic linguistic divergence. Baseline models (red) exhibit narrative-oriented patterns, frequently employing hedging language (*suggests*, *implying*) that summarizes experimental narratives rather than generating factual specifications. In contrast, MetaGAI ground truth (blue) uniquely captures high-value technical entities (*CDLA*, *OpenRAIL*) and precise licensing terms. This shows that our pipeline successfully generates concrete implementation specifications (artifact metadata) rather than paper narratives.

G.6 Cost-Efficiency Analysis

To further evaluate the economic feasibility of deploying these models at the scale of millions of papers, we analyze the trade-off between generation quality and inference cost. We define a Cost Index normalized by standardizing token consumption (1M input / 0.2M output) based on OpenRouter pricing.

Figure 8 visualizes this cost-quality landscape. The analysis reveals a distinct Pareto frontier dominated by open-weight architectures. Specifically, the sparse MoE model Qwen3-30B-A3B-Instruct achieves the highest qualitative scores while maintaining one of the lowest costs per task. In contrast, proprietary models such as Gemini-2.5-Flash and GPT-5-Mini, while capable, sit far to the right of the efficient frontier, incurring significantly higher costs (up to 5×) without a proportional gain in generation fidelity. This disparity suggests that, for structured generation tasks, specialized open-weight models offer a far superior return on investment.

H Case Study

To demonstrate the efficacy of the MetaGAI pipeline in complex retrieval and synthesis scenarios, we present two qualitative case studies illustrating how multi-source triangulation and multi-agent architecture mitigate common failure modes including information omission and incomplete evidence coverage.

H.1 Multi-Source Triangulation

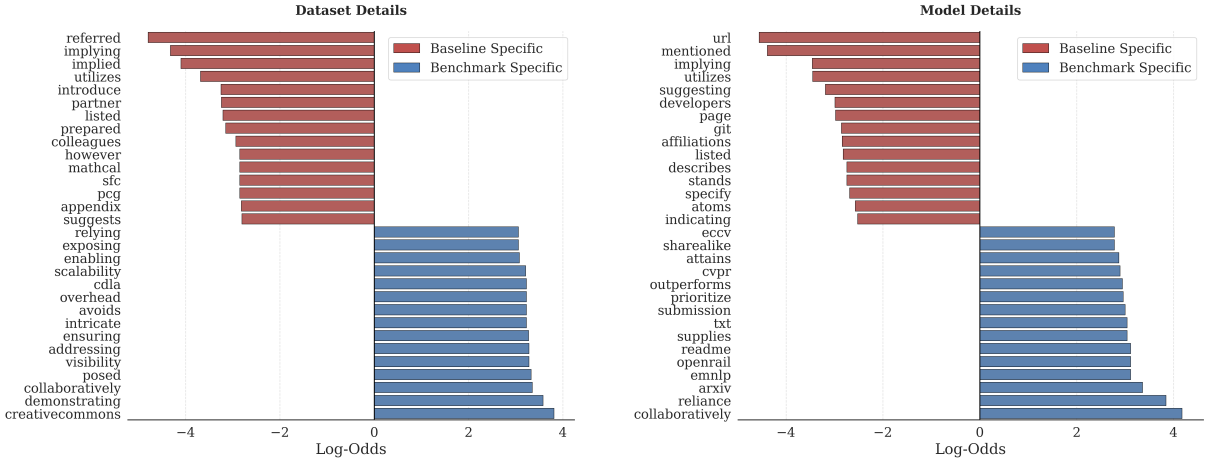
We analyze the *Model Architecture* generation for the paper "Low-light Image Enhancement via Breaking Down the Darkness" (Guo and Hu, 2022). As illustrated in Figure 9, high-fidelity output requires fusing distinct information modalities: the paper defines macro-level architectural topology, GitHub specifies implementation hyper parameters typically omitted from manuscripts, and Hugging Face validates entity alignment and deployment availability.

This case validates the pipeline's ability to resolve the granularity gap between conceptual descriptions and implementation specifications. While the paper provides the architectural framework, the GitHub repository provides the concrete hyper parameters necessary for reproducibility. The Editor Agent (GPT-OSS-20B) correctly identified that Candidate B integrated evidence across all three sources, selecting it over simpler drafts that merely paraphrased the abstract.

H.2 Editor-Driven Information Synthesis

While retrieval ensures evidence access, individual Generator Agents frequently exhibit narrow focus, generating information from specific paper sections while overlooking complementary content. The Editor Agent provides critical synthesis capabilities to resolve this incompleteness.

Continuing with the same paper, we analyze the *Robustness Metrics* field for the "Bread" model. As shown in Figure 10, the two Generator Agents produced factually accurate but incomplete drafts: Candidate A (Olmo3-7B) generated only comparative benchmark results, while Candidate B (Qwen2.5-7B) focused exclusively on the findings of the ablation study. The Editor Agent (GPT-OSS-

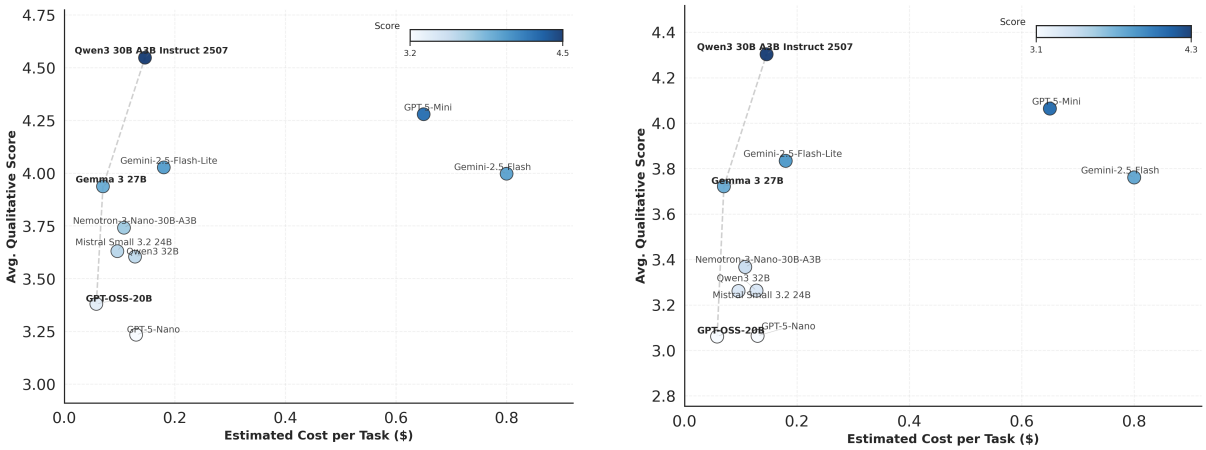


(a) Dataset Details Divergence

(b) Model Details Divergence

Figure 7: **Lexical Divergence (Log-Odds Ratio). Left (Red):** Words over-represented in Baselines, indicating narrative bias and format hallucinations. **Right (Blue):** Words specific to MetaGAI, highlighting the capture of high-value entities and technical specifications.

20B) recognized that both candidates captured orthogonal robustness dimensions: external performance validation versus internal architectural stability. Rather than selecting a single candidate, the editor merged non-redundant information to generate comprehensive coverage addressing both comparative benchmarking and component ablation analysis. This synthesis capability demonstrates a critical advantage of our multi-agent architecture over single-model generation approaches, which typically exhibit selection bias toward particular evidence types.



(a) Model Card Generation

(b) Data Card Generation

Figure 8: **Cost-Efficiency Analysis.** The Pareto frontier (dashed grey line) highlights models that offer optimal quality for a given cost. Note the significant gap between the efficient open-weight frontier and closed-source proprietary models.

Triangulation Logic for "Bread" Architecture

[Source] Paper
*"Figure 4 shows the overall architecture... comprises an illumination adjustment net (**IAN**), an adaptive noise suppression net (**ANSN**), and a color adaption net (**CAN**)."*

[Source] GitHub
(Code Implementation)

```
class IAN(nn.Module):
layers = [Conv3x3] * 3
# 3 down-sampling
# 3 up-sampling
out = torch.sigmoid(x)
```

[Source] Hugging Face
(Metadata Verification)

```
Task: Image Restoration
Tags: [PyTorch, bread]
```

Confirms model identity and availability.

MetaGAI Editor Decision (GPT-OSS-20B)

Selected Candidate: Qwen2.5-7B (Draft B)
Source Types Verified: Paper, GitHub, Hugging Face
Reasoning: *"Candidate B accurately captures the proposed Bread architecture, whereas the other candidates lack relevant details."*

Final Generated Content:
Bread is a three-stage CNN pipeline:

1. **Illumination Adjustment Network (IAN)** predicts a relative illumination map to brighten the luminance.
2. **Adaptive Noise Suppression Network (ANSN)** denoises the brightened luminance using multiple suppression strengths fused by a Noise Fusion Module.
3. **Color Adaption Network (CAN)** refines chrominance guided by the denoised luminance.

Synthesis: Each sub-network follows a symmetric encoder–decoder with 3 down-sampling and 3 up-sampling layers, 3×3 convolutions, and a sigmoid output (except ANSN).

Figure 9: **Multi-Source Triangulation.** The Editor Agent synthesizes the high-level topology from the Paper, specific layer depths and activation functions from GitHub, and validates alignment via Hugging Face, resulting in a complete specification.

Synthesis Logic for "Bread" Robustness

Candidate A (Olmo3-7B)
(Focus: External Benchmarks)
*"Bread achieves the lowest **NIQE and LOE scores** on the DICM, NPE, and VV datasets, outperforming all competing methods."*
✗ *Missing internal analysis.*

Candidate B (Qwen2.5-7B)
(Focus: Internal Ablation)
*"**Ablation studies** show that removing key modules such as the noise fusion module causes marked drops in PSNR/SSIM."*
✗ *Missing comparative stats.*

MetaGAI Editor Decision (GPT-OSS-20B)

Action: Merge Non-Redundant Information
Reasoning: *"Both candidates describe valid, orthogonal aspects of robustness. Candidate A covers external metrics (NIQE/LOE), while Candidate B covers internal validity (Ablation). Merging creates a holistic summary."*

Final Generated Content:

- **External Performance:** Bread achieves the lowest NIQE and LOE scores on the DICM, NPE, and VV datasets.
- **Internal Stability:** Ablation studies confirm framework robustness; removing the noise fusion module causes marked drops in PSNR/SSIM.

Figure 10: **Resolving Incompleteness via Synthesis.** The Editor Agent detects that Candidates A and B focus on orthogonal aspects of robustness (External Benchmarks vs. Internal Ablation). Instead of selecting a single winner, the Editor merges them to generate a comprehensive entry.