A Review of Arabic Post-Training Datasets and Their Limitations

Mohammed Alkhowaiter[1*†] **Norah Alshahrani**[2,*] **Saied Alshahrani**[3,*]
Reem I. Masoud[4,7,*] **Alaa Alzahrani**[5,*] **Deema Alnuhait**[6,*]
Emad A. Alghamdi[8] **Khalid Almubarak**[8]
[1]Refine AI [2]ASAS AI [3]University of Bisha
[4]University College London [5]King Salman Global Academy for Arabic
[6]University of Illinois at Urbana-Champaign [7]King Abdulaziz University
[8]HUMAIN

Abstract

Post-training has emerged as a crucial technique for aligning pre-trained Large Language Models (LLMs) with human instructions, significantly enhancing their performance across a wide range of tasks. Central to this process is the quality and diversity of post-training datasets. This paper presents a review of publicly available Arabic post-training datasets on the Hugging Face Hub, organized along four key dimensions: (1) LLM Capabilities (e.g., Question Answering, Translation, Reasoning, Summarization, Dialogue, Code Generation, and Function Calling); (2) Steerability (e.g., Persona and System Prompts); (3) Alignment (e.g., Cultural, Safety, Ethics, and Fairness); and (4) Robustness. Each dataset is rigorously evaluated based on popularity, practical adoption, recency and maintenance, documentation and annotation quality, licensing transparency, and scientific contribution. Our review revealed critical gaps in the development of Arabic post-training datasets, including limited task diversity, inconsistent or missing documentation and annotation, and low adoption across the community. Finally, the paper discusses the implications of these gaps on the progress of Arabic-centric LLMs and applications while providing concrete recommendations for future efforts in Arabic post-training dataset development.

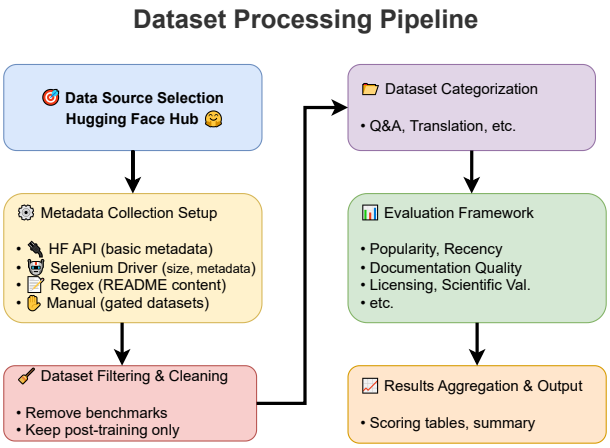


Figure 1: General Processing Pipeline for Arabic Post-Training Dataset Collection, Filtering, and Evaluation.

1 Introduction

Recent years there has been a growing interest in building high-quality post-training datasets to steer and enhance the capabilities of Large Language Models (LLMs). The nature of post-training has evolved alongside advancements in AI models. Although post-training still occurs after pre-training on large text corpora, its focus has shifted. Previously, post-training often involved task-specific fine-tuning, such as sentiment analysis, topic classification, or image classification, with models like BERT (Devlin et al., 2019). Today, it has expanded into a broader and more general concept.

This shift became clear with the emergence of capabilities of LLMs, as highlighted by Brown et al. (2020), which demonstrated strong performance on various tasks through zero-shot or few-shot prompting, even without explicit task-specific training. These capabilities were further advanced by works like Ouyang et al. (2022), which aligned models to better follow user intent, enabling more engaging and coherent interactions in dialogue formats to utilize these capabilities. This trend has also extended to other languages, such as Arabic, which has witnessed significant growth through several Arabic-centric LLMs, aimed at enhancing and improving post-training datasets.

A variety of strategies have been utilized to develop post-training datasets tailored to Arabic-centric LLMs. For the JAIS models (Sengupta et al., 2023), instruction tuning was performed using a mix of English and Arabic datasets. The Arabic portion was primarily composed of trans-

*Contributed equally; contributions varied by focus.
†Corresponding author: mohammed@refineai.dev.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 323–337
November 8-9, 2025

lated adaptations of widely adopted English post-training resources, including those from Wang et al. (2022); Taori et al. (2023); Conover et al. (2023), along with template-based instruction datasets such as Muennighoff et al. (2023). In addition to these translated datasets, two original datasets—NativeQA-Ar and SafetyQA-Ar—were specifically developed to incorporate culturally and contextually relevant content for the United Arab Emirates and the wider Arab region.

Huang et al. (2024) introduced an Arabic-centric LLMs, dubbed AceGPT, by continuing pre-training from Llama 2 (Touvron et al., 2023). In the post-training phase, their primary focus was on localizing instructions and preference data. They generated synthetic Arabic data by prompting GPT-4 model directly in Arabic, which resulted in more culturally nuanced responses compared to prompts in English. Additionally, they incorporated well-known datasets, such as Alpaca, Evol-Instruct, and Code-Alpaca, into their Supervised Fine-tuning (SFT) mixture and generated corresponding Arabic versions using GPT-4 (Achiam et al., 2023). ALLaM series of models (Bari et al., 2024) were post-trained on datasets collected from public and proprietary sources, covering a diverse range of topics, including education, history, Arabic linguistics, politics, and religion. Additionally, their post-training dataset underwent multiple filtering steps to ensure high quality. A more recent methodology proposed by Fanar et al. (2025) introduced a synthetic data generation pipeline aimed at enriching post-training datasets with culturally contextualized content. Despite these significant efforts, publicly available Arabic post-training datasets remain considerably behind those of many other languages. Even the Arabic-centric LLMs developed to date still struggle to compete closely with known LLMs, whether open-source ones, like DeepSeek and Qwen, or proprietary models, like ChatGPT, Claude, and Gemini, according to the Open Arabic LLM Leaderboard by El Filali et al. (2025).

A key reason behind this gap is that Arabic still underrepresented in post-training efforts (Guellil et al., 2021) even though it is a native language of over 400 million speakers across 22 countries, and its position as the fourth most used language on the Internet (Boudad et al., 2018). This underrepresentation is largely due to limited publication of Arabic post-training dataset. Moreover, the Arabic language has rich morphology, non-concatenative word formation, complex syntactic structures, and significant diglossia between Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA), which introduce additional layers of ambiguity (Darwish, 2014). Given Arabic's linguistic complexity, cultural richness, and global relevance (Bakalla, 2023; Versteegh, 2014), it is essential to rethink how post-training resources are developed for the language.

This paper surveys existing Arabic post-training datasets, identifies critical gaps, addresses challenges, and offers recommendations, all to guide future Arabic post-training dataset development. We list our main contributions as the following:

- We systematically reviewed publicly open Arabic datasets used for post-training and alignment of Arabic-centric language models.
- We developed tools[1] to automatically extract Arabic post-training datasets from the Hugging Face Hub and evaluate each dataset across six dimensions: documentation, popularity, adoption, recency and maintenance, licensing transparency, and scientific value.
- We identified critical gaps in Arabic post-training dataset development and offered recommendations to improve transparency, cultural relevance, and downstream usability.

2 Methodology

We exclusively collected Arabic post-training datasets' metadata from the *Hugging Face Hub*, as it represents the most comprehensive and widely-adopted machine learning platform utilized by researchers, developers, and organizations worldwide. While we initially attempted to diversify our sources by including platforms such as *GitHub* and *Kaggle*, the number of datasets with sufficient metadata and standardized formatting was negligible compared to Hugging Face Hub's extensive collection. Additionally, GitHub and Kaggle datasets often lack the structured metadata tags and consistent documentation standards essential for our automated collection methodology. Therefore, we focused solely on the *Hugging Face Hub* as our primary source to ensure data quality, consistency, and comprehensive coverage of available Arabic post-training datasets. Our dataset collection and evaluation pipeline is shown in Figure 1.

[1] www.github.com/refineaidev/mind-the-gap.

2.1 Experimental Setup

We utilized the Hugging Face Hub Python library to automatically collect the following metadata for each dataset: *Dataset ID (dataset name)*, *Number of Likes*, *Number of Downloads*, *Last Modified Date*, *Name of License*, *ArXiv Papers*, and *Number of Models that have used this dataset*. We further employed the Selenium Python library to automate the collection of additional metadata not provided by the Hugging Face Hub Python library, including *Size of Downloaded Files*, *Size of Parquet Files*, and *Number of Rows*.

2.2 Metadata Collection

We employed four distinct approaches to gather metadata for Arabic post-training datasets: 1) *automatic collection* of metadata using the Hugging Face Hub Python library, leveraging the platform's metadata tags; 2) *automated collection* of metadata using the Selenium Python library, extracting information from the dataset's statistics widget (located on the right side of the dataset card); 3) *regular expression search* for specific metadata within README.md files of datasets, such as *ACL Papers*, again utilizing the Hugging Face Hub Python library; and 4) *manual collection* of metadata for gated datasets, which are private datasets requiring access requests, making automatic and automated collection approaches infeasible. We also manually removed benchmark datasets to ensure our collection exclusively contained post-training datasets.

2.3 Evaluations of Datasets

We evaluated Arabic post-training datasets across 12 task categories, mapped to four dimensions: (1) LLM Capabilities (e.g., *Q&A*, *Translation*, *Reasoning and Multi-Step Thinking*, *Summarization*, *Dialogue*, *Code Generation*, and *Function Calling*); (2) Steerability (e.g., *Persona* and *System Prompt*); (3) Alignment (e.g., *Cultural Alignment*, *Safety, Ethics, and Fairness*); and (4) Robustness. The selection of the 12 task categories was informed by two criteria: (1) alignment with established taxonomies in prior research, like Chen et al. (2025); Minaee et al. (2024), and (2) representation of distinct, functionally coherent areas relevant to LLM evaluation and dataset availability. Specifically, we synthesized insights from Minaee et al. (2024), who provide a broad survey of LLM capabilities across general NLP domains. This combined perspective ensured that our categories address both specialized applications, such as *Code Generation*, and general-purpose tasks, such as *Summarization*.

Each dataset was assessed using framework comprising six evaluation criteria: documentation and annotation quality, popularity, practical adoption, recency and active maintenance, licensing transparency, and scientific contribution. Each criterion utilizes a structured scoring system designed for simplicity, consistency, and reproducibility.

To illustrate our methodology, Table 1 presents an example of evaluation criteria and scoring rubrics used to assess documentation and annotation quality across datasets. We deliberately employed straightforward rubrics to ensure simplicity, efficiency, and effectiveness in our evaluation process. The remaining set of evaluation criteria and corresponding scoring systems for all assessment dimensions is provided in Appendix A (Table 4), offering full transparency in our methodology and enabling reproducibility of our findings.

3 Analysis and Results

We analyzed 366 datasets across 12 Natural Language Processing (NLP) domains, summarized in Table 3. Due to unbalanced group sizes and small sample sizes in certain domain categories, we present only descriptive statistics to avoid Type I and Type II errors associated with insufficient statistical power and unequal groups (Field, 2017). The remainder of this section will first cover the descriptive statistics of the collected datasets, followed by the evaluation results for those datasets.

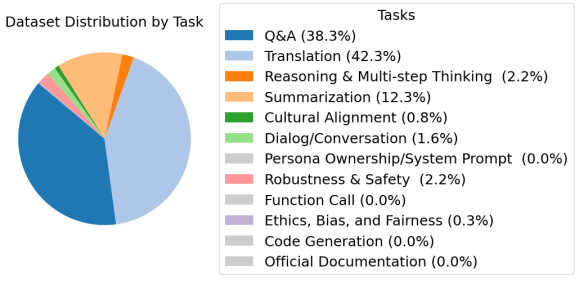


Figure 2: Distribution of datasets across tasks. Labels include the percentage of datasets in each task. Tasks with no datasets are shown for the sake of completeness.

3.1 Dataset Results

As shown in Figure 2 and detailed in Appendix B, the distribution of the datasets is highly skewed towards specific tasks. For example, *Translation* and *Question Answering (Q&A)* dominate, comprising 42.3% and 38.3% of the datasets, respec-

Table 1: An detailed example of the evaluation criteria and scoring system used for evaluating documentation and annotation quality. The remaining evaluation criteria and scoring rubrics are provided in Appendix A (Table 4).

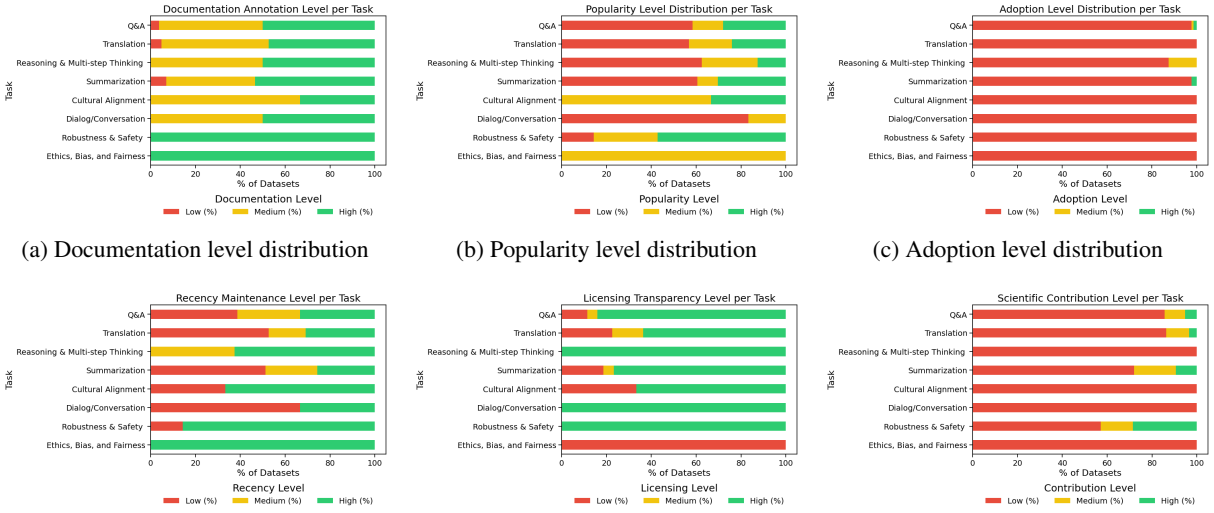
Evaluation	Criteria	Avg. Score	Quality Level
Documentation	• Dataset card explains the usage of dataset • Dataset card states the license clearly • Dataset card shows examples of dataset • Dataset card includes or cites a paper • Dataset card describes the datasets • Dataset card states the authors or maintainers	$4 \leq \text{score} \leq 6$ $2 \leq \text{score} < 4$ $\text{score} < 2$	High Medium Low
Annotation	• Metadata tags specify a task • Metadata tags specify a language • Metadata tags state a size • Metadata tags state a license • Metadata tags include dataset source • Metadata tags include configurations	$4 \leq \text{score} \leq 6$ $2 \leq \text{score} < 4$ $\text{score} < 2$	High Medium Low

tively. *Summarization* adds another 12.3%, while the remaining six tasks account for fewer than 30 datasets combined. Notably, *Function Call*, *Persona Ownership*, *Code Generation*, and *Official Documentation* have no datasets (zero datasets), revealing major gaps in current publicly available Arabic post-training resources.

3.2 Automated Evaluation Results

We present our findings from the automated evaluation of the collected datasets, focusing on their documentation and annotation quality, popularity, practical adoption, recency maintenance, licensing transparency, and scientific contribution, with detailed results shown in Appendix C.

- **Documentation Quality:** Documentation standards show mixed results across tasks. Figure 3a demonstrates that specialized domains like *Ethics, Bias, and Fairness* and *Robustness & Safety* achieve excellent documentation quality (100% high-quality scores). Still, these domains contain only 9 datasets in total, which may not adequately represent the broader landscape and could limit their applicability to diverse research contexts.
- **Popularity:** Dataset popularity varies significantly across tasks. Figure 3b shows that traditional NLP tasks, like *Q&A*, *Translation*, and *Summarization*, include many widely-used datasets with strong community adoption. In contrast, tasks such as *Dialog/Conversation* and *Ethics, Bias, and Fairness* are dominated by low-popularity and medium-popularity datasets, reflecting either niche applications or limited awareness in the broader community.
- **Community Adoption:** Figure 3c reveals consistently low adoption rates across all task categories, indicating limited reuse and citation of existing datasets. This pattern suggests that researchers may be creating new datasets rather than building upon existing work, potentially leading to fragmented efforts and reduced cumulative progress in the field.
- **Dataset Maintenance:** Maintenance practices vary considerably, highlighting inconsistent update schedules across the ecosystem. Figure 3d shows that newer research areas like *Robustness & Safety* and *Ethics, Bias, and Fairness* maintain current datasets, while established tasks such as *Summarization* and *Translation* contain many outdated resources that lack regular maintenance cycles.
- **Licensing Transparency:** Licensing practices show positive trends toward open accessibility. Figure 3e demonstrates that most Arabic datasets provide clear licensing information, with many adopting permissive licenses like Apache-2.0. This transparency facilitates both academic research and commercial applications, supporting broader utilization of Arabic post-training datasets.
- **Scientific Contribution:** Research integration remains limited across the dataset landscape. Figure 3f indicates that most datasets lack formal scientific validation through peer-reviewed publications or DOI assignment. This gap suggests that many datasets represent individual contributions rather than systematically validated research contributions.



(a) Documentation level distribution

(b) Popularity level distribution

(c) Adoption level distribution

(d) Recency maintenance per task

(e) Licensing transparency per task

(f) Scientific contribution level per task

Figure 3: Overview of dataset quality across tasks. The subfigures present quality indicators including documentation, popularity, adoption, recency, licensing transparency, and scientific contribution. While the full taxonomy includes 12 tasks, we report results for the 9 tasks with available datasets. *Persona & System Prompts*, and *Function Call*, *Code Generation*, and *Official Documentation* are excluded as no datasets were available for those tasks.

4 Identified Gaps

Our analysis has identified several critical gaps that could significantly hinder Arabic NLP research and applications (as per Table 3). Potential gaps and limitations include the following:

- **Limited Arabic Post-Training Data:** Small coverage of Arabic post-training datasets leads to slow advancement in Arabic-centric LLMs, and hence their applications. There are almost no Arabic datasets available for key post-training tasks such as *Function Call*, *Persona Ownership*, *Code Generation*, and *Official Documentation*. Undoubtedly, this scarcity significantly hampers the development of sophisticated Arabic large language models that can perform complex tasks.

- **Poor Dataset Documentation:** Poor documentation and annotation of datasets leads to invisible and inaccessible resources within the Arabic NLP community. As shown in Table 2, many valuable datasets remain uncategorized and difficult to discover, creating barriers for researchers who could benefit from existing work. This lack of proper documentation surely prevents the efficient reuse and building upon previous efforts in the field.

- **Low Community Engagement:** Low popularity of Arabic datasets reflects how the Arabic NLP community remains small and sometimes discouraging to new contributors. This limited engagement raises research ethical issues, including failure to cite others' work and not giving proper credit to dataset creators.

Table 2: Total of Arabic datasets categorized under the 12 selected tasks, compared to uncategorized datasets.

Dataset Type	**Total**
Categorized Datasets (for the 12 tasks)	366
Uncategorized Datasets	341

- **Limited Open-Source Integration:** Limited adoption of Arabic datasets in training open-source models and public Hugging Face spaces restricts the broader accessibility of Arabic NLP applications. One possible reason for this limited integration is the lack of computational resources available to researchers and practitioners working with Arabic language models. This creates a barrier that prevents the wider deployment and testing of Arabic NLP solutions in real-world applications.

- **Lack of Dataset Maintenance:** Lack of recency and maintenance characterizes the majority of Arabic datasets, with most open-source resources rarely receiving updates or maintenance for periods exceeding 12 months. This stagnation means that datasets become outdated and potentially less relevant to current research needs. The absence of regular

updates suggests a lack of sustained community support and ongoing development efforts.

- **Weak Scientific Standards:** Weak scientific contribution characterizes most Arabic datasets, with almost all datasets not being released as part of peer-reviewed research papers or having DOI identifiers. The majority represent individual contributions rather than rigorous academic work, which typically results in lower quality standards. This pattern reflects poorly on the overall quality of Arabic datasets, as those released with research papers or DOIs tend to demonstrate higher quality and more thorough validation.

5 Case Study: Safety and Cultural Alignment

Safety and cultural alignment datasets are crucial for developing responsible, culturally sensitive NLP systems. However, our findings reveal significant gaps in both areas. As shown in Figures 2 and 4, *Cultural Alignment* accounts for less than 1% of all surveyed datasets, while *Robustness & Safety* includes only 8 datasets, with substantial variation in size and coverage. Both categories show consistently low adoption rates, and *Cultural Alignment* additionally exhibits limited scientific contribution (Figure 3), suggesting underutilization despite the relatively strong popularity of some datasets.

This underrepresentation is especially concerning given the importance of cultural sensitivity and safety in Arabic-speaking contexts, where linguistic, societal, and religious norms differ greatly from dominant English-based benchmarks. The lack of culturally aware and safety-focused datasets increases the risk of deploying misaligned or even harmful NLP systems, like LLMs. To address these blind spots, we strongly recommend prioritizing the development of high-quality datasets tailored to Arabic cultural contexts and safety concerns, ensuring that future models are not only technically robust but also ethically and socially aligned.

6 Recommendations and Future Directions

The findings of this review highlight the strategic importance of post-training datasets for advancing Arabic-centric LLMs. While the existing resources on Hugging Face Hub provide a starting point, they fall short in coverage, documentation quality, cultural alignment, and scientific rigor. To address these limitations and accelerate the development of Arabic LLMs, we offer the following forward-looking recommendations, structured around priority domains, practical dataset creation strategies, and principles for collaborative research.

6.1 High-Priority Domains for Future Post-Training Datasets

This subsection outlines specific domains in Arabic post-training that are currently underrepresented or entirely missing, yet are crucial for building capable, safe, and culturally aligned Arabic LLMs. These domains should be prioritized in future post-training dataset development initiatives due to their strategic importance and lack of coverage.

- *Reasoning and Multi-Step Thinking:* Datasets supporting logical reasoning, problem-solving, and chain-of-thought prompting are vital for advanced LLM capabilities.
- *Summarization:* While moderately covered, many existing datasets lack consistency in documentation, linguistic variety, and practical relevance to real-world use cases.
- *Cultural Alignment:* Data that reflects nuanced Arab world values, norms, and social constructs is crucial for building culturally sensitive NLP systems and applications.
- *Dialog/Conversation:* This domain suffers from very limited coverage and low-quality documentation and annotation. Rich, dialect-sensitive dialogue datasets are essential for improving conversational fluency and natural interaction in Arabic-centric LLMs.
- *Persona and System Prompting:* Needed for conversational agents to maintain consistent behavior and alignment across interactions.
- *Robustness & Safety:* Despite its importance for responsible AI development, the availability of high-quality Arabic post-training datasets in this domain remains limited.
- *Function Calling:* Essential for tool-augmented NLP and API-connected LLMs, yet currently nonexistent in public Arabic post-training resources.
- *Ethics, Bias, and Fairness:* Arabic datasets in this area are extremely limited, despite growing ethical concerns in global LLM adoption, development, and deployment.
- *Code Generation:* There are currently no open Arabic datasets supporting code generation.

Table 3: Summary of Arabic Post-training Dataset Coverage and Key Identified Gaps

Category	Coverage	Key Gaps
Question Answering (Q&A)	Strong (140 datasets)	Lacks community adoption & scientific validation
Translation	Strong (155 datasets)	Lacks community adoption & needs maintenance
Reasoning & Multi-Step Thinking	Very limited (8 datasets)	Needs significant scale expansion
Summarization	Moderate (45 datasets)	Lacks community adoption & scientific rigor
Cultural Alignment	Critically limited (3 datasets)	Needs culturally nuanced datasets
Dialog/Conversation	Very limited (6 datasets)	Lacks popularity & needs maintenance
Persona/Ownership/System Prompt	**No datasets**	Requires development
Robustness & Safety	Limited (8 datasets)	Needs broader coverage & adoption
Function Call	**No datasets**	Requires development
Ethics, Bias, and Fairness	Critically limited (1 dataset)	Needs coverage & licensing transparency
Code Generation	**No datasets**	Requires development
Official Documentation	**No datasets**	Requires development

- *Official Documentation:* This domain is completely absent from current post-training resources, although critical for building capable LLMs that can handle policies, manuals, formal content, or structured instructions.

6.2 Practical Guidelines for Building Arabic Post-Training Datasets

This subsection focuses on practical and scalable methods for creating Arabic post-training datasets. These guidelines are intended for researchers and developers, who aim to build new resources and address domain-specific gaps. The listed methods are grounded in existing tools, community collaboration, and modern data generation strategies.

Dialectal Dialogue Collection Capturing authentic spoken Arabic from various dialect regions is essential. We recommend collecting spontaneous conversations from native speakers across the Arab world, followed by accurate transcription that preserves dialectal features.

Collaborative Annotation Platforms A crowdsourced annotation platform can empower native speakers to label data along cultural and contextual dimensions. By providing well-defined annotation guidelines, especially on culturally sensitive topics, the platform can produce high-quality datasets with rich sociocultural nuance.

Human–LLM Hybrid Annotation Large language models can be leveraged to perform initial annotations, which are then verified or refined by human annotators. This semi-automated approach balances efficiency with quality assurance and reduces manual annotation overhead.

Synthetic Data Generation Arabic-capable LLMs can be prompted to generate new post-training data for underrepresented tasks. Although synthetic data offers scalability, rigorous validation is necessary to ensure linguistic correctness, cultural appropriateness, and task alignment.

6.3 Recommendations for Future Research and Collaboration

This final subsection presents high-level, strategic guidance for the broader research community. These recommendations emphasize principles like authenticity, cultural representation, and open collaboration. They are intended to shape future initiatives and encourage ethical, inclusive, and sustainable development of Arabic post-training datasets.

- **Prioritize Missing Domains:** Direct funding, research, and community efforts toward domains with little to no coverage in Arabic (e.g., *Function Calling* and *Code Generation*).
- **Promote Authenticity over Translation:** Native Arabic content should be favored to avoid loss of context, nuance, or cultural misalignment present in translated material. While translated datasets can serve as a temporary bridge to address data scarcity, they fundamentally compromise the linguistic and cultural integrity essential for powerful Arabic LLMs. Native Arabic content preserves cultural subtleties, idiomatic expressions, and the language's unique morphological complexity that translation inevitably distorts. In culturally sensitive domains—including religious discourse, legal frameworks, and social interactions—native content ensures terminological accuracy and cultural appropriateness that

directly impacts model performance and user acceptance. Thus, we recommend prioritizing investment in native Arabic dataset creation as a sustainable strategy for developing LLMs that authentically serve Arabic-speaking communities rather than imposing linguistic patterns from other language contexts.

- **Incorporate Cultural Context:** Datasets should reflect ethical, religious, and societal views, values, and cultures of the Arab world to ensure cultural robustness in AI outputs.
- **Broaden Linguistic Representation:** Both Modern Standard Arabic (MSA) and regional Dialectal Arabic (DA) should be represented in future dataset development to support real-world use cases across the Arab region.
- **Foster Open Collaboration and Transparency:** Dataset creators are encouraged to share licensing details, evaluation metrics, and use-case documentation to increase reproducibility, transparency, and adoption.
- **Investigate Dataset-Performance Relationships:** Future research should investigate relationships between our categorized dataset characteristics and actual model performance. Such studies could leverage our framework to conduct controlled experiments across task categories, establishing empirical relationships between dataset quality metrics and model effectiveness. This would provide valuable guidance for dataset creators and model developers in the Arabic NLP community.

7 Conclusion

In this paper, we conducted the first systematic survey of publicly available Arabic post-training datasets hosted on the Hugging Face Hub, with a focus on evaluating their quality, coverage, licensing transparency, and scientific contribution, across 12 key LLM capabilities. Our findings reveal several critical gaps, most notably the near absence of datasets in high-impact domains, such as *Function Calling*, *Code Generation*, *Ethical Alignment*, and *Official Documentation*. Despite the growing importance of post-training in aligning LLMs with human intent, Arabic remains substantially underrepresented in this space. Many existing datasets suffer from limited documentation, outdated maintenance, and low practical adoption. These shortcomings hinder the advancement of robust, culturally aligned, and ethically grounded Arabic LLMs.

We proposed a set of high-priority domains that require urgent dataset development and provided practical, scalable guidelines for building Arabic post-training resources through community collaboration, hybrid human–LLM annotation, and synthetic data generation. Additionally, we outlined strategic recommendations for promoting native content, cultural awareness, and linguistic diversity in future dataset creation efforts. Lastly, we release two open-source demo versions of our dataset collection and evaluation tools to the Arabic NLP research community. The introduction of these tools will facilitate standardized evaluation practices as well as reproducible research. In the near future, we aim to publicly share production versions with detailed documentation to ensure broad accessibility and adoption across research institutions.

Limitations

While this study provides the first structured review of Arabic post-training datasets, it is subject to several limitations. First, this review covers only datasets openly available on Hugging Face Hub, omitting any private or gated resources.

Second, our collection and evaluation rely heavily on metadata and Dataset Cards (README) documentation, which may not always accurately reflect the actual quality or usability of the datasets. Some datasets may be underdocumented despite being high-quality in practice, and others may appear polished but lack effective downstream utility.

Third, this study does not assess how the reviewed datasets directly impact model performance. While our review provides essential infrastructure for dataset discovery, examining correlations between dataset characteristics and model effectiveness would require extensive computational resources and standardized benchmarking protocols beyond this study's scope. As such, the current study did not examine the relationship between the reviewed datasets and model performance.

Ethical Considerations

While this study does not collect new data or generate text, analyzing public Arabic datasets raises ethical concerns, including unclear licensing, cultural bias, and dual-use risks. We encourage transparent licensing, inclusive annotations, and responsible governance in future dataset development.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Muhammad Hasan Bakalla. 2023. *Arabic Culture Through Its Language and Literature*. Routledge, Abingdon, UK.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. ALLaM: Large Language Models for Arabic and English. *Preprint*, arXiv:2407.15390.

Naaima Boudad, Rdouan Faizi, Rachid Oulad Haj Thami, and Raddouane Chiheb. 2018. Sentiment Analysis in Arabic: A Review of the Literature. *Ain Shams Engineering Journal*, 9(4):2479–2490.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Meng Chen, Philip Arthur, Qianyu Feng, Cong Duy Vu Hoang, Yu-Heng Hong, Mahdi Kazemi Moghaddam, Omid Nezami, Duc Thien Nguyen, Gioacchino Tangari, Duy Vu, Thanh Vu, Mark Johnson, Krishnaram Kenthapadi, Don Dharmasiri, Long Duong, and Yuan-Fang Li. 2025. Mastering the craft of data synthesis for CodeLLMs. 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)*, pages 12484–12500, Albuquerque, New Mexico. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

Kareem Darwish. 2014. Arabizi Detection and Conversion to Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 217–224, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Preprint*, arXiv:1810.04805.

Ali El Filali, Manel ALOUI, Tarique Husaain, Ahmed Alzubaidi, Basma El Amel Boussaha, Ruxandra Cojocaru, Clémentine Fourrier, Nathan Habib, and 1 others. 2025. Open Arabic LLM Leaderboard 2.

Fanar, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An Arabic-Centric Multimodal Generative AI Platform. *arXiv preprint arXiv:2501.13944*.

Andy Field. 2017. *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications, London, UK.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic Natural Language Processing: An Overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, Localizing Large Language Models in Arabic. *arXiv preprint arXiv:2309.12053*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj,

Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *arXiv preprint arXiv:2308.16149*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Kees Versteegh. 2014. *The Arabic Language*, 2 edition. Edinburgh University Press, Edinburgh, UK.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, and 21 others. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. *arXiv preprint arXiv:2204.07705*.

A Evaluation Criteria

Appendix A presents a comprehensive scoring rubric for evaluating Arabic datasets across five key categories: Popularity, Adoption, Recency and Maintenance, Licensing Transparency, and Scientific Contribution, as shown in Table 4. Each category includes specific criteria and is scored based on defined numerical thresholds, which are then mapped to qualitative levels—High, Medium, or Low. For example, Popularity is measured by the number of likes and downloads, with a dataset considered highly popular if it receives a total of 200 or more. Adoption reflects how widely the dataset is used across models and spaces, while Recency and Maintenance assess how recently the dataset has been updated, rewarding more actively maintained resources.

Licensing Transparency evaluates whether the dataset includes a clear license, with high scores given to those that explicitly state a recognized license. In contrast, datasets marked as "unknown," "other," or "none" receive lower scores. The Scientific Contribution category assesses the dataset's presence in the academic field, based on references to or arXiv papers and the inclusion of DOI objects. This rubric offers a structured framework for evaluating dataset quality and academic relevance, making it easier to compare datasets and identify those best suited for research and development in Arabic NLP.

Table 4: Scoring rubric for evaluating Arabic datasets based on popularity, adoption, recency and maintenance, licensing transparency, and scientific contribution. Each criterion is scored individually and mapped to a qualitative level (High, Medium, or Low). The documentation criteria and scoring rubric are previously displayed in Table 1.

Evaluation	**Criteria**	**Score**	**Total Score**	**Level**
Popularity	Dataset's Number of Likes	Number of Likes	$200 \leq$ Score $100 \leq$ Score < 200 Score < 100	High Medium Low
	Dataset's Number of Downloads	Number of Downloads		
Adoption	Number of Used Models	Number of Models	$50 \leq$ Score $20 \leq$ Score < 50 Score < 20	High Medium Low
	Number of Used Spaces	Number of Spaces		
Recency & Maintenance	Dataset's Last Modified Date	Last Modified – Collection Date	Score ≤ 6 Mo 6 Mo $<$ Score ≤ 12 Mo Score > 12 Mo	High Medium Low
Licensing Transparency	Dataset card states the license	License Name	Known license 'unknown'/'other' 'none'	High Medium Low
	Metadata tags state the license	License Name		
Scientific Contribution	Dataset card includes ACL Papers	ACL Papers	$3 \leq$ Score $1 \leq$ Score < 3 Score $= 0$	High Medium Low
	Metadata tags include ArXiv Papers	ArXiv Papers		
	Metadata tags include a DOI Object	DOI Object		

B Dataset Characteristics by Task

This appendix provides a comprehensive overview of dataset characteristics and quality across Arabic post-training tasks. Table 5 summarizes key statistics for each task category, including the number of datasets, average Hugging Face likes, downloads, model usage, and citation counts in ACL and ArXiv papers. These metrics offer insight into dataset visibility, reuse, and scholarly contribution.

Figure 4 complements this summary by illustrating the range of dataset sizes per task on a logarithmic scale. This visualization reveals substantial variation both across and within tasks, with some datasets ranging from a few dozen to over 10 billion rows. Given this high variance, we emphasize range-based visualizations rather than relying solely on averages when assessing dataset scale.

Table 5: Values represent means with standard deviations in parentheses. For each task category, the table reports the number of datasets (n), mean number of Hugging Face likes and downloads, average count of model implementations, and mean number of ACL and ArXiv papers citing the dataset. For tasks with n = 1, standard deviations are not applicable and are indicated by (-). For tasks with n = 0, all values are indicated by (-) as no data is available.

Task	**n**	**Likes**	**Downloads**	**Models**	**ACL Papers**	**ArXiv Papers**
Q&A	140	10.6 (43.9)	1285 (8288)	3.1 (19.7)	0.22 (0.61)	0.27 (0.45)
Translation	155	9 (20.5)	721 (1805)	1 (5.1)	0.16 (0.52)	0.21 (0.41)
Reasoning & Multi-Step Thinking	8	10 (11.6)	105 (112)	3.5 (7.2)	0 (0)	0 (0)
Summarization	45	9.9 (22.9)	2826 (13931)	3 (12.6)	0.33 (0.71)	0.24 (0.43)
Cultural Alignment	3	19.7 (27.4)	171 (59)	1.7 (1.5)	0 (0)	0.33 (0.58)
Dialog/Conversation	6	1.8 (2.3)	47 (42)	0.2 (0.4)	0 (0)	0.17 (0.41)
Persona Ownership/System Prompt	0	- (-)	- (-)	- (-)	0 (-)	0 (-)
Robustness & Safety	8	4.9 (8.9)	253 (167)	1.4 (2.7)	0.75 (1.04)	0.62 (0.52)
Function Call	0	- (-)	- (-)	- (-)	0 (-)	0 (-)
Ethics, Bias, and Fairness	1	16 (-)	176 (-)	0 (-)	0 (-)	0 (-)
Code Generation	0	- (-)	- (-)	- (-)	0 (-)	0 (-)
Official Documentation	0	- (-)	- (-)	- (-)	0 (-)	0 (-)

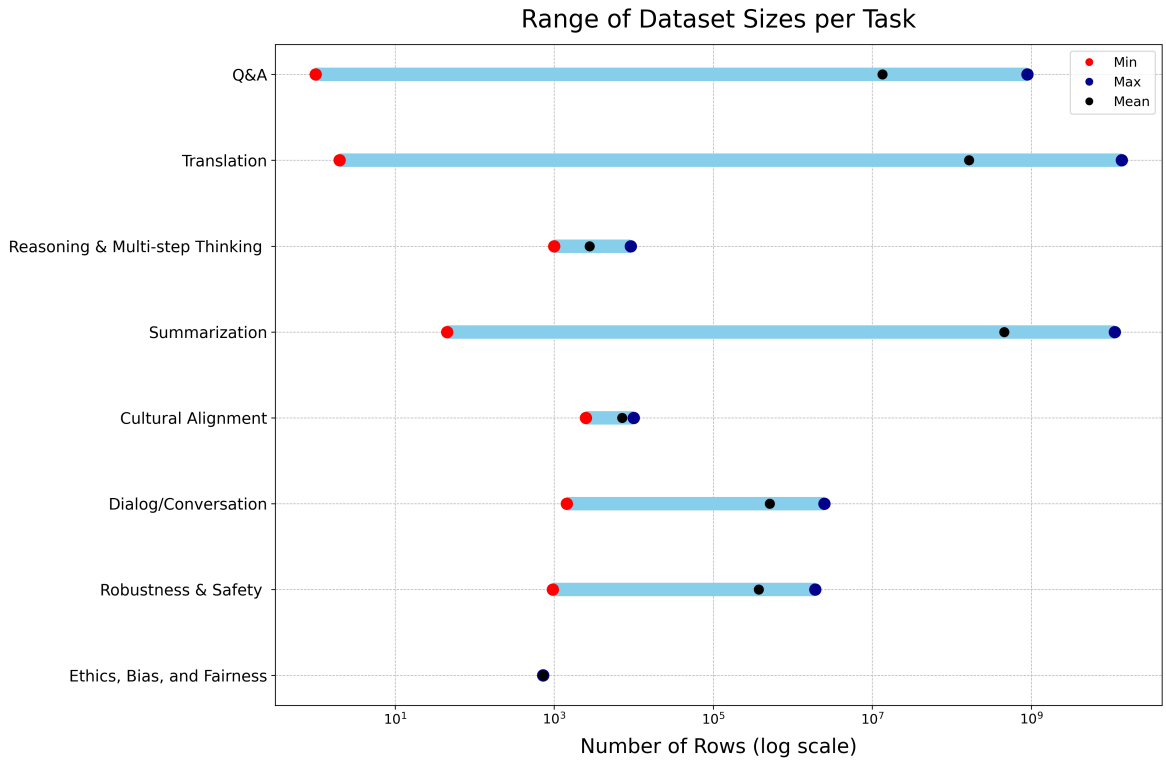


Figure 4: Range of dataset sizes per task (log scale). Each horizontal bar represents the minimum and maximum number of rows for datasets within a task, with red, blue, and black points denoting the minimum, maximum, and mean sizes, respectively. The wide variation in size highlights disparities in dataset availability and scale across post-training tasks. Although there are 12 tasks, here we only present the size of datasets with available data (n= 9). This figure reveals that dataset sizes vary dramatically not only across tasks but also within the same task category. Some tasks, such as Summarization and Translation, contain datasets ranging from a few dozen rows to over 10 billion. This high variance makes aggregate measures like the mean misleading; therefore, we emphasize range-based visualizations over summary statistics when discussing dataset scale.

C Quality Score Proportions By Task

This appendix presents a task-level summary of dataset quality scores across six evaluation dimensions. Table 6 reports the proportion of datasets rated as low, medium, or high for each criterion: documentation and annotation quality, popularity, adoption, recency and maintenance, licensing transparency, and scientific contribution. These scores reflect both the strengths and limitations of available Arabic post-training datasets and provide a quantitative basis for identifying quality gaps across task categories. Missing values are also reported to ensure transparency in coverage and support reproducibility.

Table 6: Dataset quality levels across tasks and evaluation dimensions. The **Missing** column refers to the number of datasets with missing scores for the specified level type. For example, in the *Robustness & Safety* task, 2 datasets lack documentation level, and 1 lacks all evaluation scores. Tasks with no datasets are marked with (–).

Task	# Datasets	Missing	Level Type	Low (%)	Medium (%)	High (%)
Q&A	140	8	documentation_annotation_level	3.79	46.21	50.00
			popularity_level	58.33	13.64	28.03
			adoption_level	97.73	0.76	1.52
			recency_maintenance_level	38.64	28.03	33.33
			licensing_transparency_level	11.36	4.55	84.09
			scientific_contribution_level	85.61	9.09	5.30
Translation	155	9	documentation_annotation_level	4.79	47.95	47.26
			popularity_level	56.85	19.18	23.97
			adoption_level	100.00	0.00	0.00
			recency_maintenance_level	52.74	16.44	30.82
			licensing_transparency_level	22.60	13.70	63.70
			scientific_contribution_level	86.30	10.27	3.42
Reasoning & Multi-Step Thinking	8	0	documentation_annotation_level	0.00	50.00	50.00
			popularity_level	62.50	25.00	12.50
			adoption_level	87.50	12.50	0.00
			recency_maintenance_level	0.00	37.50	62.50
			licensing_transparency_level	0.00	0.00	100.00
			scientific_contribution_level	100.00	0.00	0.00
Summarization	45	2	documentation_annotation_level	6.98	39.53	53.49
			popularity_level	60.47	9.30	30.23
			adoption_level	97.67	0.00	2.33
			recency_maintenance_level	51.16	23.26	25.58
			licensing_transparency_level	18.60	4.65	76.74
			scientific_contribution_level	72.09	18.60	9.30
Cultural Alignment	3	0	documentation_annotation_level	0.00	66.67	33.33
			popularity_level	0.00	66.67	33.33
			adoption_level	100.00	0.00	0.00
			recency_maintenance_level	33.33	0.00	66.67
			licensing_transparency_level	33.33	0.00	66.67
			scientific_contribution_level	100.00	0.00	0.00
Dialog/Conversation	6	0	documentation_annotation_level	0.00	50.00	50.00
			popularity_level	83.33	16.67	0.00
			adoption_level	100.00	0.00	0.00
			recency_maintenance_level	66.67	0.00	33.33
			licensing_transparency_level	0.00	0.00	100.00
			scientific_contribution_level	100.00	0.00	0.00
Robustness & Safety	8	2	documentation_annotation_level	0.00	0.00	100.00
		1	popularity_level	14.29	28.57	57.14
			adoption_level	100.00	0.00	0.00
			recency_maintenance_level	14.29	0.00	85.71
			licensing_transparency_level	0.00	0.00	100.00
			scientific_contribution_level	57.14	14.29	28.57
Ethics, Bias, and Fairness	1	0	documentation_annotation_level	0.00	0.00	100.00
			popularity_level	0.00	100.00	0.00
			adoption_level	100.00	0.00	0.00
			recency_maintenance_level	0.00	0.00	100.00
			licensing_transparency_level	100.00	0.00	0.00
			scientific_contribution_level	100.00	0.00	0.00
Persona Ownership/System Prompt	0	-	No data available		-	
Function Call	0	-	No data available		-	
Code Generation	0	-	No data available		-	
Official Documentation	0	-	No data available		-	