ARC AI Framework Protocol (May 21, 2025)

Applying the ARC Framework (Accurate, Reliable, Contextual) for Human-AI

Contextualization: A Model for Assuring AI Integrity

Ed Woods

ARC AI Framework Initiative

Aligned to APA 7th Edition Standards and ARC Tier-1 & Tier-2 Validation

Abstract

As artificial intelligence (AI) systems increasingly mediate knowledge work, decision-making, and strategic operations, the integrity of human-AI contextualization—how humans frame, interpret, and act on AI outputs—has emerged as a critical ethical and epistemic concern. This paper presents the ARC Framework—Accurate, Reliable, Contextual—as a structured protocol for auditing and enhancing the trustworthiness of AI-generated intelligence. Developed under the ARC AI Framework Initiative, a research and governance effort focused on responsible AI deployment, ARC is grounded in peer-reviewed research and aligned with institutional standards. It mitigates known risks such as hallucination, bias propagation, and context drift. Tier 1 embeds rigor through role definition, source validation, and citation integrity; Tier 2 guides users through analysis, reframing, and context-aligned commitment.

In this 2025 revision, ARC is formalized as an epistemic protocol—a reproducible system with defined checkpoints for validation, citation, and judgment. It integrates Retrieval-Augmented Generation (RAG) as a Tier-1 engine for source-grounded generation and redefines the "A" in the ARC Loop as Analyze/Audit, uniting interpretive critique with forensic traceability. Together,

these developments position ARC as a full-spectrum, governance-ready protocol for responsible, context-aware AI deployment.

Keywords: ARC Framework, Epistemic Trust, Human-AI Collaboration, Contextual Intelligence, Tier-1 Validation & Rigor, Tier-2 Decision & Refinement, AI Governance

Introduction

The rapid integration of AI systems across sectors such as education, governance, climate modeling, and healthcare has raised urgent concerns about the trustworthiness of machine-generated outputs (Floridi & Cowls, 2022). Many generative AI models, including large language models, are prone to hallucinated facts, fragile logic chains, and contextual misalignment (Raji et al., 2020). These failures are not mere technical glitches—they represent deeper epistemological vulnerabilities in how knowledge is produced, framed, and interpreted. To address these challenges, the ARC Framework has evolved into a structured, human-centered protocol designed to ensure AI output integrity across three foundational dimensions: Accuracy, Reliability, and Contextuality (Woods, 2025).

More than a conceptual model, ARC functions as a repeatable epistemic protocol, integrating role-specific prompting, citation validation, and traceable reasoning steps. Recent industry consensus supports this evolution.

As the Forbes Technology Council (2023) notes, "contextualization is the key to unlocking generative AI's potential," reinforcing ARC's foundational principle that relevance and meaning must be constructed through intentional, human-guided framing.

Clarifying ARC's Core Structure: Pillars vs. Tiers

The ARC Framework is built upon three epistemic pillars: **Accuracy**, **Reliability**, and **Contextuality**. These are not tied to a specific tier or stage; rather, they form the evaluative foundation used to assess AI outputs at all levels. Think of them as the "why" behind ARC—why we audit AI outputs, why we structure human-AI interactions, and why integrity matters.

To apply these pillars in practice, ARC introduces two tiers:

- **Tier-1: Validation & Rigor** – Ensuring AI responses are verifiable, sourced, and role-appropriate using structured protocols.

- **Tier-2: Decision & Refinement** – Guiding users through analysis, re-framing, and context-aligned commitment.

Protocol Structure

Protocol Step	ARC Pillar Aligned	Function
Act As	Contextuality	Domain-specific role prompting
Reference	Reliability	Retrieval of peer-reviewed, timestamped sources (via RAG)
Cite Sources	Accuracy	Inline or footnoted traceable claims for auditability

This protocol reinforces ARC as a **thinking system** and **validation engine**, guiding AI toward structured epistemic output.

Contextualization in the Scholarly Landscape

ARC's emphasis on contextualization aligns with a growing body of interdisciplinary research across human-computer interaction (HCI), AI education, systems theory, and epistemology. Eguchi et al. (2021) underscore the value of culturally responsive, context-aware AI education. Hirsch et al. (2023) explore contextual framing in HCI-based cultural heritage systems. Murphy and Largacha-Martinez (2023) argue that contextualization is essential for aligning AI systems with real-world organizational complexity. Addas (2020) warns that neglecting context often results in systemic failure. Nardi (1996), in her foundational work on activity theory, establishes that context is not peripheral—it is constitutive of interface meaning.

These perspectives affirm ARC's central claim: context is not an annotation—it is a condition for epistemic trust. ARC operationalizes this by embedding context into its protocol structure through Tier-1 role prompting, Tier-2 reframing, and the full Act-As/Analyze/Audit cycle. In this light, contextualization is not just an academic value—it is a structural function that determines whether AI can serve as a trustworthy partner in knowledge work.

Pillars in Practice

1. Accuracy: Verifiable, Fact-Based Outputs

Accuracy in AI is not simply about avoiding errors—it is about aligning machine output with domain-valid knowledge (Barocas, Hardt, & Narayanan, 2019). The ARC Framework mandates source-backed responses, requiring that all factual claims be explicitly tied to verifiable, timestamped references.

For instance, in climate science or medical diagnostics, even minor inaccuracies can lead to cascading failures (NIST, 2023). The ARC Framework mandates source-backed responses, requiring that all factual claims be explicitly tied to verifiable, timestamped references.

Example: "According to the IPCC Sixth Assessment Report (2022), carbon emissions must peak by 2025 to remain within 1.5°C warming targets."

Such specificity transcends vague generalizations, grounding AI claims in reproducible knowledge

2. Reliability: Traceable Reasoning and Institutional Standards

Ribeiro, Singh, and Guestrin (2016) argue that users must understand why an AI system produced a given output, not just what it produced. This aligns with ARC's call for "reasoned transparency," requiring that logic chains and institutional sources be traceable and explainable (Woods, 2025).

The U.S. National Institute of Standards and Technology's (NIST) AI Risk Management Framework explicitly states that "traceability, explainability, and transparency are foundational to responsible AI" (NIST, 2023, p. 14).

3. Contextuality: Contextual: Human-in-the-Loop Framing and Relevance

Context is the operating environment in which AI outputs are interpreted and made meaningful. Building on the work of Kahneman (2011), who demonstrates that human reasoning is inherently context-bound, as well as Dervin's (1998) theory of sense-making and Schön's (1983) reflective practitioner model, the ARC Framework asserts that context is the invisible structure shaping meaning. Without intentional framing, AI can produce technically accurate but practically irrelevant responses—a form of epistemic drift.

To mitigate this, ARC mandates user-defined roles, goals, and domains (e.g., "Act as a medical ethicist advising on gene editing policy") to ensure that outputs remain relevant, tone-

appropriate, and decision-aligned. This emphasis on Contextual Intelligence shifts curation power back to the human, restoring interpretive authority where it belongs and making the user an active participant in meaning-making, not a passive recipient of machine-generated logic.

Tier-1: ARC as a Protocolic Foundation for Validation and Rigor

Tier-1 is where the ARC Protocol transitions from epistemic principle to operational system. While the pillars of Accuracy, Reliability, and Contextuality define the values of trustworthy AI, Tier-1 establishes the procedural mechanisms that enforce them. Through structured role prompting, Retrieval-Augmented Generation (RAG), and citation-layered output validation, Tier-1 transforms AI from a generative engine into a source-aligned reasoning system. This tier ensures that AI-generated responses meet institutional standards for scholarly integrity and are suitable for high-stakes, domain-sensitive applications.

The Tier-1 protocol consists of three interdependent components that ensure AI outputs meet scholarly and operational standards:

- Act As: The AI must adopt a clearly defined expert role to contextualize its output (e.g., "Act as a climate policy advisor"). This aligns the generation process with domain-specific expectations and interpretive frameworks

- Reference: All claims must be grounded in authoritative, verifiable sources—preferably peer-reviewed publications, institutional reports, or timestamped documents relevant to the user's domain.

- Cite Sources: Every factual assertion must include traceable, properly formatted citations (e.g., APA 7th edition or discipline-specific style) to support transparency, reproducibility, and independent auditability.

While the ARC Protocol emphasizes explicit, verifiable citation wherever possible, it also acknowledges the current architectural limitations of large language models. In some cases, references may be inferred through recognizable synthesis patterns rather than formally cited sources. As LLMs continue to evolve toward greater traceability and epistemic accountability, ARC is structured to both accommodate and guide this evolution. Until then, active human oversight remains indispensable to assess the credibility, relevance, and contextual fit of all referenced material.

This protocol is not merely a formatting standard—it is a trust mechanism that enables users to evaluate credibility, challenge assumptions, and ensure contextual alignment. Each of the ARC pillars—Accuracy, Reliability, and Contextuality—is operationalized through this three-part structure, establishing a reproducible foundation for human-AI sensemaking and responsible decision framing.

ARC's emphasis on traceability, verifiability, and contextual oversight aligns directly with the U.S. National Institute of Standards and Technology's AI Risk Management Framework (NIST AI RMF), which identifies these features as essential to trustworthy AI. This compatibility positions ARC not only as a scholarly protocol but as an actionable governance model for NIST-aligned deployment.

To demonstrate this interoperability, Figure 1 illustrates how ARC's Tier-1 validation protocol maps to institutional principles outlined in the NIST AI RMF. This crosswalk reinforces ARC's legitimacy as a protocol for compliant, sector-spanning AI systems.

Moreover, ARC's Tier-1 methodology integrates Retrieval-Augmented Generation (RAG) to ground AI outputs in verifiable source material. This advances ARC from philosophical preference to a formalized epistemic system, enabling source-traceable reasoning through structured pipeline checkpoints.

Tier-1 RAG Workflow

1. **Intent Parsing** – Interprets role/domain from user prompt

2. **Query Construction** – Designs retrievable semantic vectors

3. **Source Retrieval** – Prioritizes high-authority, timestamped documents

4. **Relevance Filtering** – Applies epistemic integrity filters (recency, source trust)

5. **Context Injection** – Embeds concise excerpts into model context

6. **Output Generation** – Uses structured synthesis with citations

7. **Citation Layer** – Attaches verifiable anchors to every claim

This transforms ARC Tier-1 from a checklist into an operational **protocol of trust**.

Tier-2: Contextual Intelligence Tier — From Verification to Judgment

If Tier-1 establishes credibility, Tier-2 restores responsibility. While Tier-1 validates AI outputs through structured rigor, Tier-2 shifts the focus to critical human engagement—guiding users to interrogate assumptions, reframe insights, and commit to context-aligned action. This is where

human discernment reclaims center stage, transforming passive interaction into intentional contextualization and epistemic accountability.

- **Analyze/Audit – Surface Blind Spots and Validate Provenance**

  The Analyze/Audit step serves a dual function. "Analyze" prompts users to examine what is missing, misaligned, or unconsciously accepted in AI-generated content—echoing Kahneman and Tversky's (1984) framing effect theory. "Audit" adds forensic rigor: tracing the output's logic, source provenance, and contextual alignment. This step ensures that decisions are not merely intuitive, but epistemically grounded and verifiably sound.

- **↺ Reframe – Shift Perspective and Expand Possibility**

  Reframing challenges users to re-interpret AI-generated insights through alternative lenses—behavioral, ethical, strategic, or theoretical. This reflects Schön's (1983) reflective practitioner model and Weick's (1995) dynamic sensemaking framework. ARC asks: *How would this recommendation shift if viewed through a different stakeholder, discipline, or equity lens?*

- **Commit – Make Context-Aligned, Owned Decisions**

  Commitment marks the point where analysis yields action. Informed by Dervin's (1998) theory of sense-making, ARC positions commitment as the moment where knowledge gaps are bridged through deliberate, situated choices. It compels the user to ask: *Given the trade-offs, risks, and contextual obligations, what decision am I prepared to own and implement?*

Tier-2 does not replace AI with human judgment—it embeds human agency as the final architect of meaning, responsibility, and execution.

ARC Protocol Risk Registry: Documenting Epistemic Failures and Mitigation Paths

The ARC Protocol includes a structured Risk Registry that documents real-world anomalies, breakdowns, and epistemic failures observed during AI system interaction. Each entry is classified by ARC Tier, Pillar, Failure Type, and Mitigation Strategy—providing a feedback mechanism for refining both prompt engineering and model trust boundaries. This registry transforms ARC from a validation framework into a dynamic epistemic system, capable of learning from failure in alignment with risk-informed governance protocols such as NIST AI RMF and ISO 31000.

Registry Function & Purpose

The ARC Risk Registry serves three primary functions:

1. **Capture**: Record epistemic anomalies during AI generation or post-prompt review

2. **Classify**: Organize by Tier (1 or 2), Pillar (A/R/C), and failure mode (e.g., hallucination, lexical ambiguity, context loss)

3. **Mitigate**: Recommend prompt reforms, retrieval improvements, or model filtering strategies

Risk ID	Tier	Pillar	Failure Type	Example Prompt	Mitigation Path
ARC-R-101	Tier 1	Accuracy	Lexical Ambiguity	"Does this contain the word 'and'?"	Use boundary regex in input design
ARC-C-203	Tier 2	Contextual	Overgeneralized Framing	"Summarize ethics of AI"	Add domain/role: 'As a bioethicist'

Registry as Governance Mechanism

The Risk Registry supports traceability, reproducibility, and institutional compliance by maintaining a transparent log of known system risks and their epistemic consequences. It also serves as a pedagogical tool—enabling ARC users to learn from previous failures and improve future interactions.

The Curation Continuum: Intelligence as a Living Process

AI use is not a one-time transaction—it is a recursive cycle of interpretation, refinement, and epistemic contextualization. The ARC Protocol formalizes this reality through the Curation Continuum, a model informed by systems thinking, organizational learning, and iterative design. Within this framework, curation is not an afterthought; it is a sustained epistemic practice that safeguards human-AI collaboration from drift, distortion, and decontextualization.

With the integration of Retrieval-Augmented Generation (RAG) and Tier-1 validation checkpoints, ARC evolves beyond a scholarly framework into a living epistemic infrastructure—capable of governing AI reasoning pipelines in applied contexts. Designed to be both rigorous and adaptable, the ARC Protocol supports domain-specific governance systems while remaining anchored in critical reasoning, source traceability, and contextual alignment.

Scholarly Foundations:

ARC's epistemic structure draws from multiple theoretical traditions. Nonaka and Takeuchi's (1995) model of knowledge creation describes a continuous spiral between tacit and explicit understanding—an idea central to ARC's looped reasoning and user engagement structure. Bawden and Robinson (2020) emphasize that effective digital knowledge work depends on sustained, intentional curation, directly aligning with ARC's Curation Continuum. Snowden's

(2007) Cynefin Framework reinforces the need for iterative, context-sensitive sensemaking within complex systems, mirroring ARC's emphasis on role framing, domain fit, and adaptive interpretation.

ARC is not a filter applied after the fact. It is a recursive engagement system—a way to ensure that decisions evolve as context shifts, goals adapt, and new information emerges.

The Curation Continuum reinforces three essential truths:

- Context must be refreshed: Prior assumptions may no longer apply as conditions change.

- Reliability must be recalibrated: New data, models, or sources may shift what is considered valid.

- Accuracy must be re-verified: Outputs must be checked against updated references and constraints.

This approach protects against static reliance on AI and fosters epistemic agility. With ARC, the user doesn't just manage AI outputs—they curate an evolving understanding, staying aligned with reality through intentional, iterative stewardship.

ARC-Validated Benefits:

- Formalizes what it means to "do ARC right"

- Encourages iterative learning

- Supports open epistemic governance

ARC Certification & Stewardship: Building Capacity for Responsible AI Practice

To institutionalize adoption and ensure cross-sector alignment, the ARC Framework Protocol includes a formal **Certification and Stewardship Pathway**, launching in 2025. This program provides a scalable model for education, public service, and private organizations to build capacity in responsible AI usage, based on validated ARC principles.

ARC Certification enables individuals and institutions to formalize their understanding and application of the ARC Protocol. The program includes tiered credentials based on depth of engagement—from foundational literacy to protocol-based deployment—and centers around the ARC pillars of Accuracy, Reliability, and Contextuality.

ARC Certification Structure

Certification Tier	Audience	Focus
ARC-1: Foundations	Educators, librarians, general users	ARC pillars, Tier-1 prompting, and source verification
ARC-2: Applied Practice	Curriculum designers, civic orgs	Tier-2 analysis, risk registry usage, contextual reframing
ARC-3: Steward Level	Institutional leads, facilitators	Governance, prompt auditing, protocol training and policy integration

Each tier includes guided modules, practical use case assignments, and a final assessment that reflects real-world application of ARC principles. Completion grants an official ARC certification badge, signaling epistemic responsibility and contextual fluency.

ARC Stewardship complements certification by creating a dynamic, open-source community of practice. Participants are invited to co-develop use cases, share evolving risks, and contribute improvements to the protocol structure. ARC stewards serve as local facilitators in libraries, schools, and civic institutions, helping ensure the protocol remains inclusive, adaptable, and grounded in public benefit.

This certification and stewardship pathway positions ARC not just as a technical protocol, but as an epistemic infrastructure for education, governance, and ethical AI deployment.

ARC-Validated Benefits

- Establishes a common standard for AI contextualization and output integrity

- Enables credentialed teaching, governance, and risk-aware use of AI tools

- Promotes iterative learning and collaborative refinement

- Supports local and institutional trust-building through documented practice

Conclusion: Toward Epistemic Integrity in Human-AI Collaboration

The ARC Protocol for Human-AI Contextualization provides a rigorously structured, philosophically grounded system for epistemic alignment in AI use. It is not a checklist—it is a repeatable, role-sensitive protocol for curating, validating, and contextualizing AI-generated intelligence. In contrast to generalized collaboration frameworks, ARC explicitly positions the human as the epistemic anchor: responsible for framing prompts, verifying claims, and interpreting meaning in light of evolving ethical, cultural, and domain-specific norms.

While industry discourse is increasingly recognizing the importance of context in generative AI (Forbes Technology Council, 2023), ARC offers the operational structure to implement it—

through Tier-1 validation mechanisms and Tier-2 decision loops. Grounded in scholarship, validated through field use, and adaptable across sectors, ARC equips human agents not only to interpret AI outputs, but to own them—framing, challenging, and committing to decisions with both clarity and care.

NIST AI RMF Principle	ARC Tier-1 Implementation
Traceability	*Cite Sources:* All factual claims must be traceable to timestamped, domain-valid sources.
Explainability	*Reference + Role Clarity:* Outputs reflect declared expertise and rational justification.
Transparency	*Act As:* Outputs explicitly state AI's role, assumptions, and intended function.
Contextual Awareness	*Contextuality Pillar:* Human defines role, goal, and domain, ensuring relevance and alignment.

Note. ARC's Tier-1 validation protocol operationalizes core NIST AI RMF principles, bridging academic rigor and institutional governance. This alignment positions ARC as a protocol for both critical evaluation and real-world deployment, fully interoperable with compliance frameworks such as NIST AI RMF 1.0 (2023).

Protocol Licensing and Commercial Deployments

The ARC Framework Protocol is published under a Creative Commons Attribution–No Derivatives 4.0 International (CC BY-ND 4.0) license to support open access, scholarly reuse, and epistemic transparency. However, specific commercial implementations of the ARC protocol are governed separately under proprietary terms. **Reef Flow™** Systems represents the first narrative-based, AI-integrated product built on the ARC epistemic structure. Designed for small-to-medium businesses and civic organizations, Reef Flow™ operates outside the Creative Commons license as a licensed commercial deployment. While it extends the ARC framework into applied pedagogical environments, all Reef Flow™ implementations are required to uphold Tier-1 validation standards and Tier-2 contextual alignment, preserving the protocol's core commitments to accuracy, reliability, and contextual integrity.

Selected References (APA 7)

Addas, S. (2020). Contextualization in human-computer interaction. Transactions on Human-Computer Interaction, 2(4).

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. https://fairmlbook.org

Bawden, D., & Robinson, L. (2020). Introduction to information science (2nd ed.). Facet Publishing.

Dervin, B. (1998). Sense-making theory and practice. Journal of Knowledge Management, 2(2), 36–43.

Eguchi, A., et al. (2021). Contextualizing AI education for K-12. PMC.

Floridi, L., & Cowls, J. (2022). A unified framework for AI in society. Harvard Data Science
Review, 4(1).

Floridi, L. (2022). Artificial intelligence and its limits: An epistemological framework.
Philosophy & Technology, 35(3), 1–23. https://doi.org/10.1007/s13347-022-00532-1

Forbes Technology Council. (2023, August 30). Contextualization: The key to unlocking
generative AI's potential.
https://www.forbes.com/sites/forbestechcouncil/2023/08/30/contextualization-the-key-to-unlocking-generative-ais-potential/

Hirsch, L., et al. (2023). Human-computer interaction for contextualizing cultural heritage.
Applied Sciences, 14(17).

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Mittelstadt, B. D. (2023). AI as an epistemic technology. Science and Engineering Ethics, 29(1),
1–18. https://doi.org/10.1007/s11948-023-00451-3

Murphy, J. W., & Largacha-Martinez, C. (2023). Contextualizing AI. In Contextualizing AI in
Practice (pp. 19–35). Springer.

Nardi, B. A. (1996). Context and consciousness: Activity theory and HCI. MIT Press.

National Institute of Standards and Technology (NIST). (2023). *AI Risk Management Framework
(AI RMF 1.0).* U.S. Department of Commerce.
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Nonaka, I., & Takeuchi, H. (1995). The knowledge-creating company. Oxford University Press.

Raji, I. D., & Buolamwini, J. (2020). Closing the AI accountability gap: Defining an end-to-end

    framework for internal algorithmic auditing. Proceedings of the 2020 Conference on

    Fairness, Accountability, and Transparency, 33–44.

    https://doi.org/10.1145/3351095.3372873

Raji, I. D., et al. (2020). Closing the AI accountability gap. FAccT 2020 Proceedings.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Explaining the predictions of any classifier.

    KDD '16 Proceedings.

Schön, D. A. (1983). The reflective practitioner. Basic Books.

Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. AIS

    Transactions on Human-Computer Interaction, 12(3), 109–124.

    https://doi.org/10.17705/1thci.00131

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective

    sociotechnical approach. Proceedings of the 2020 CHI Conference on Human Factors in

    Computing Systems, 1–12. https://doi.org/10.1145/3313831.3376592

Weick, K. E. (1995). Sensemaking in organizations. Sage Publications.

Woods, E. (2025). *Applying the ARC Framework (Accurate, Reliable, Contextual) for Human-AI*

    Contextualization: A Model for Assuring AI Integrity [Scholarly paper]. ARC Initiative

ARR as a Participatory Epistemic Practice

Unlike traditional risk management systems that rely solely on institutional oversight or static

taxonomies, the ARC Risk Registry (ARR) is designed as a living, user-powered feedback

mechanism. It enables ARC users—including students, educators, researchers, and civic

practitioners—to actively log, classify, and share epistemic anomalies they encounter during human-AI interaction. Each submission contributes to a growing, structured body of knowledge that refines the ARC Protocol's real-world applicability. In this way, ARR functions not only as a diagnostic tool but as a participatory epistemic commons—inviting communities of practice to co-curate reliability, interpretive authority, and context-aware governance.

ARR as a Participatory Epistemic Practice

Also known in ARC-aligned learning environments as "The Driftline"

Unlike traditional risk management systems that rely solely on institutional oversight or static taxonomies, the ARC Risk Registry (ARR) is designed as a living, user-powered feedback mechanism. It enables ARC users—including students, educators, researchers, and civic practitioners—to actively log, classify, and share epistemic anomalies they encounter during human-AI interaction. Each submission contributes to a growing, structured body of knowledge that refines the ARC Protocol's real-world applicability.

Within ARC-enabled platforms such as Reef Flow™, this participatory layer is colloquially referred to as The Driftline—a term drawn from oceanic ecology, where floating fragments collect at the convergence of shifting currents. In this context, Driftline entries represent surfaced misalignments, logic snags, and context drifts that, if left unexamined, could erode trust or distort action. By capturing these anomalies early, ARC users learn to reframe uncertainty not as failure, but as an invitation to collective sensemaking.

In this way, the ARR functions not only as a diagnostic tool but as a participatory epistemic commons—inviting communities of practice to co-curate reliability, interpretive authority, and context-aware governance.

Supplemental References for Risk Registry Alignment

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)*, 220–229. https://doi.org/10.1145/3287560.3287596

National Institute of Standards and Technology. (2023). *AI Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Partnership on AI. (2023). *AI Incident Database*. https://incidentdatabase.ai

License Statement