Quantified Reliability Assessment Rubric

This rubric provides a detailed, scorable framework for assessing content reliability, expanding upon Table 4 of the "A Multi-Layered Framework for the Systematic Assessment of Content Reliability." Each criterion is scored on a 0-5 scale, with specific, quantified descriptors for each level.

I. Provenance & Source (Weight: High)

This category assesses the origin of the information, focusing on the credibility of the author and publisher.

Criterion	5 - Excellent	4 - Good	3 - Fair	2 - Poor	1 - Very Poor	0 - Unacceptable
Author Authority	Author is a widely recognized expert in the specific field, with a strong, relevant publication record and affiliation with a top-tier institution.	Author has verifiable credentials and publications in the relevant field, with a clear institutional affiliation.	Author has verifiable credentials, but they are in a general area, not the specific topic. Or, they are an established journalist from a reputable outlet.	Author is identifiable, but credentials are not directly relevant or cannot be fully verified.	Author is anonymous, uses a pseudonym with no established reputation, or has known/disclosed conflicts of interest that are not managed.	Author is known to have a history of producing inaccurate information or has significant, undisclosed conflicts of interest.
Publisher Authority	Published in a top-tier, peer-reviewed academic journal or by a major news organization with a strong international reputation for factual reporting and corrections.	Published in a respected, peer-reviewed journal, a reputable trade publication, or a well-regarded news source with clear editorial standards.	Published on a platform with some editorial oversight (e.g., a university website, established organization) but not formal peer review or rigorous	Self-published, or published by a platform with no clear editorial oversight or a known political/ideological agenda.	Published by a source known for propaganda, misinformation, or predatory practices (e.g., a predatory academic journal).	The source is actively attempting to deceive the user by impersonating a legitimate publisher.

			journalistic standards.			
Citation Quality	>90% of citations are to relevant, high-authority primary or secondary sources. All cited material is represented accurately.	70-90% of citations are to high-quality, relevant sources. Any minor inaccuracies in representation do not affect the core argument.	50-70% of citations are to credible sources, but may rely heavily on secondary or tertiary sources.	30-50% of citations are to credible sources. Some sources are low-quality, irrelevant, or misrepresented.	<30% of citations are to credible sources. Widespread use of biased sources, quote mining, or irrelevant material.	Citations are fabricated, lead to dead links, or are used to support claims completely unrelated to the source content.

II. Argument & Evidence (Weight: High)

This category analyzes the internal logic of the content, the quality of its evidence, and its rhetorical style.

Criterion	**5 - Excellent**	**4 - Good**	**3 - Fair**	**2 - Poor**	**1 - Very Poor**	**0 - Unacceptable**
Logical Coherence	Argument is well-structured, logical, and internally consistent. No logical fallacies are detected.	The core argument is sound, but may contain 1-2 minor informal fallacies that do not undermine the central thesis.	The argument has a clear structure, but contains at least one significant logical fallacy or several minor ones that weaken its validity.	The argument contains multiple significant logical fallacies (e.g., Straw Man, Ad Hominem) that demonstrate poor reasoning.	The argument is fundamentally built upon one or more major formal fallacies (e.g., Circular Reasoning) and is logically invalid.	The content lacks a discernible logical structure and consists primarily of incoherent or contradictory claims.
Evidentiary Integrity	All major claims are supported by	Most claims are supported by	Claims are supported by evidence, but it is of	Claims are primarily supported by	Claims are unsupported by evidence, are	Claims are supported by fabricated or

	high-quality evidence (e.g., systematic reviews, meta-analyses, large datasets from reputable sources).	high-quality evidence, with some reliance on individual studies or credible secondary reports.	mixed quality (e.g., expert opinion, small-scale studies, tertiary sources).	low-quality evidence (e.g., anecdotes, personal testimony, appeal to ignorance).	inherently unfalsifiable, or rely on evidence that has been widely debunked.	deliberately misinterpreted evidence.
Contextual Completeness	Presents a comprehensive view, explicitly acknowledging and accurately representing key counterarguments and contradictory evidence.	Acknowledges the existence of counterarguments but may not engage with them in depth. No significant omission of contradictory data is found.	Presents a one-sided argument but does not distort facts. Key counterarguments or contradictory datasets are largely ignored.	Systematically omits readily available evidence that contradicts the central thesis (cherry-picking).	Presents a distorted view by misrepresenting the opposing viewpoint to create a false impression of consensus.	Falsely claims that no counterarguments or contradictory evidence exists.
Rhetorical Style	Language is consistently neutral, precise, and objective. The tone is analytical and focused on evidence (System 2).	Language is mostly objective, with minimal use of evaluative or emotional terms.	Language is generally neutral, but with noticeable instances of persuasive or slightly biased framing.	Language contains a significant amount of emotionally charged, judgmental, or manipulative terms (System 1 appeal).	Language is overwhelmingly emotional, inflammatory, and designed to provoke outrage, fear, or other strong reactions.	Language is dehumanizing, incendiary, or constitutes hate speech.

III. Multimedia Forensics (Weight: Varies by content)

This category assesses the authenticity and transparency of non-textual content.

Criterion	**5 - Excellent**	**4 - Good**	**3 - Fair**	**2 - Poor**	**1 - Very Poor**	**0 - Unacceptab**

						le
Authenticity & Integrity	Passes all forensic checks. Metadata is consistent and complete. No indicators of manipulation.	Minor inconsistencies detected (e.g., missing metadata, common compression artifacts) that are likely non-malicious.	Some forensic flags are present (e.g., signs of re-saving, minor lighting mismatches) that warrant caution but are not conclusive proof of manipulation.	Multiple forensic flags suggest digital manipulation (e.g., inconsistent shadows, splicing artifacts).	Strong and clear evidence of manipulation from multiple forensic methods. Video contains subliminal frames.	The media is a known fake or is being used in a completely false context with clear intent to deceive.
AI Generation Transparency	Content passes high-confidence AI detection models OR is clearly and verifiably labeled as AI-generated via a strong method like C2PA Content Credentials.	AI detection is inconclusive, potentially due to post-processing or novel generation methods. No claims are made about its origin.	Content is presented as authentic but is flagged by AI detectors with low-to-medium confidence.	Content is presented as authentic and is flagged as AI-generated by multiple detectors with high confidence.	Content is detected as AI-generated and also shows signs of malicious manipulation (e.g., a deepfake of a person saying something harmful).	Content is part of a documented, large-scale, malicious AI-driven disinformation campaign.
Proactive Trust Signals	Contains valid, verifiable C2PA Content Credentials from a reputable source that	Contains other verifiable trust signals, such as a known, benign watermark from a	No proactive trust signals are present.	The media contains signals intended to mislead, such as a forged C2PA manifest or a watermark	The media's trust signals have been deliberately stripped or altered to obscure its origin.	The media contains malicious code or tracking signals embedded within its structure.

	logs the media's entire lifecycle.	reputable organization.		from a disreputable source.		