International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 10, Issue 4, pp 412-417, April-2026
https://doi.org/10.47001/IRJIET/2026.104056

A Trust-Aware Framework for Scoring and Routing in Multimodal Generative AI Systems

Sujeet Sharma

Technical Architect, Hearst USA. E-mail: sujeetsharma1980@gmail.com

Abstract **- Generative AI now sits in the critical path of modern content delivery. A single user request can trigger a summary from a large language model, a synthetic thumbnail, an audio narration, and an automated transcript, each with its own failure modes. In practice, these failures do not remain isolated. A weak retrieval result can surface as an unsupported summary claim, influence a generated headline, and then shape recommendation and search behavior. This paper presents a trust-aware framework that scores each generated artifact before delivery and uses that score to guide routing decisions such as publish, escalate, revise, or block. The proposed framework combines evidence support, source credibility, provenance completeness, disclosure quality, explanation utility, editorial review, calibration, metadata validity, and policy risk into a single composite score. We describe a platform-agnostic architecture, show how the same controls can be applied across text, image, audio, and transcription pathways, and report scenario-based evaluation results on a controlled synthetic workload. Compared with a baseline retrieval-augmented pipeline, the trust-aware configuration reduced unsupported claims, substantially improved provenance coverage and search metadata validity, and increased latency by a manageable amount. We also discuss operational tradeoffs, limitations, and deployment considerations for publishers and other organizations that deliver AI-generated content at scale.**

Keywords: Trust-aware AI, Multimodal Generative AI, Content Delivery, Provenance, Source Credibility, Explainability, Editorial Review, Retrieval-Augmented Generation, SEO Integrity, Human Oversight.

I. INTRODUCTION

Generative AI has moved from experimental tooling to production infrastructure in less than a few years. In many content systems, one request now triggers multiple model calls: a summary from a language model, a thumbnail from an image generator, a voice track from a text-to-speech system, and a transcript from automatic speech recognition. Each artifact may be plausible on its own, yet the user experiences them as one coherent product. That is precisely where trust problems emerge.

Early safeguards for generative systems focused mostly on model-level quality: reducing hallucinations, filtering unsafe prompts, or adding disclosure labels. Those steps help, but they do not solve a larger systems problem. A fabricated visual, a loosely grounded summary, or an aggressive search title can silently propagate through the rest of the pipeline unless some shared mechanism detects that the overall artifact is not trustworthy enough to deliver without additional review.

This paper argues that trust should be handled as a routing signal, not merely as a disclosure afterthought. The question is not only whether a model output is acceptable in isolation, but whether the system should be confident enough to publish it automatically, publish it with explanation, or hold it back for editorial review.

We make five concrete contributions. First, we define a composite trust score that captures multiple dimensions of reliability rather than relying on a single confidence signal. Second, we show how that score can drive routing decisions across modalities. Third, we describe a practical system architecture that separates content generation from trust evaluation and governance. Fourth, we extend the same logic to cross-modal pathways where independent outputs must remain semantically aligned. Finally, we evaluate the framework on a controlled workload and discuss how it changes quality, latency, and editorial workload.

The rest of the paper is organized as follows. Section II outlines the requirements and threat model. Section III presents the architecture. Section IV describes the cross-modal pathways. Section V defines the trust score and supporting metrics. Sections VI and VII cover operational and evaluation results. Sections VIII through X discuss implications, limitations, and future directions.

II. SYSTEM REQUIREMENTS AND THREAT MODEL

A. Functional Requirements

A trust-aware content pipeline must satisfy several requirements that differ from those of a conventional content

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 4, pp 412-417, April-2026

https://doi.org/10.47001/IRJIET/2026.104056

management system. First, generated outputs should remain tied to retrievable evidence wherever possible. Summaries, captions, and metadata should not rely solely on model priors when supporting material exists [1], [2].

Second, provenance should be treated explicitly rather than as a binary yes-or-no attribute. During implementation, we found it useful to distinguish among verified assets with intact provenance, partially verified assets with incomplete chains or damaged manifests, and unknown assets with no provenance information [3].

Third, consistency across delivery surfaces is essential. The same story may appear as an on-site summary, a newsletter blurb, a push notification, a narrated audio segment, and a generated thumbnail. Inconsistencies between these forms can damage trust even when each individual artifact appears superficially acceptable.

Fourth, transparency should be proportional to risk. A low-risk evergreen summary may need only a concise disclosure, while high-sensitivity content should surface source evidence, provenance state, and review status more explicitly [4]. Fifth, personalization features should rely only on consented signals. Finally, search metadata must be validated before publication, not only after indexing problems are discovered.

B. Threat Model

The framework addresses a broader set of risks than factual hallucination alone. Table 1 summarizes representative threats, likely manifestations, corresponding mitigations, and residual risks that remain even after controls are applied.

Table 1: Representative threats and mitigations in trust-aware delivery pipelines

Threat	Example manifestation	Framework mitigation	Residual risk
Factual confabulation	Model invents developments in a breaking story	Retrieval grounding, evidence threshold, escalation	Verification lag in fast-moving news
Provenance loss	Image credentials are stripped during reformatting	Manifest validation, three-state provenance classification	Legacy assets without retroactive manifests
SEO inflation	Generated title overstates the article's conclusions	Semantic similarity checks, metadata linting	Search engines may still rewrite snippets
Transcription drift	ASR error becomes an indexed quote	Confidence thresholds, diarization, correction queue	Persistent speaker ambiguity in difficult audio

III. SYSTEM ARCHITECTURE

The proposed architecture separates five responsibilities: ingestion and verification, evidence and provenance management, generation, trust scoring, and routing and delivery. This separation turned out to be important in practice. In early prototypes, generation and governance logic were tightly coupled, making the system difficult to tune [5].

The ingestion layer handles source collection, deduplication, rights classification, freshness tracking, and basic privacy checks. Small improvements at this stage had outsized downstream benefits because noisy or duplicated inputs increased the frequency of uncertain outputs and review escalations.

The knowledge layer maintains the retrieval index, source credibility registry, and provenance store. We found that provenance is much easier to preserve at ingestion time than to reconstruct later. The layer also stores audit records linking each published artifact to the evidence and controls that shaped it.

The generation layer wraps model calls with grounding and policy checks. Rather than treating model output as self-justifying, the framework requires explicit evidence support where applicable and validates the outputs against downstream constraints such as disclosure rules and metadata policies.

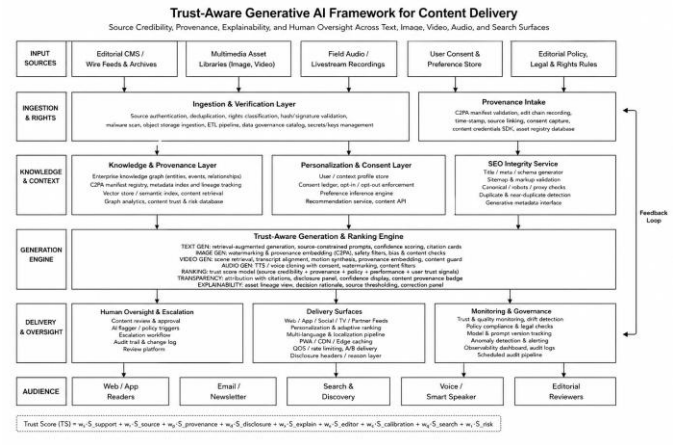

Figure 1: Trust-aware reference architecture for multimodal generative AI systems

The trust scoring and routing layer acts as the operational gate. It determines whether an artifact can be published

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 4, pp 412-417, April-2026
https://doi.org/10.47001/IRJIET/2026.104056

automatically, published with disclosure, revised, or escalated for review. Post-publication feedback, including corrections and drift signals, can then feed back into the system to support recalibration.

IV. CROSS-MODAL TRUST PATHWAYS

A central design goal of the framework is to apply consistent trust controls across modalities without pretending that all modalities fail in the same way. Text generation errors are often semantic and discourse-level; image generation errors are frequently contextual or compositional; transcription errors can be highly localized but still damaging once indexed and reused. A shared scoring framework therefore needs modality-specific checks at the signal level while retaining a common routing policy at the decision level [6],[7].

In the text-to-image path, the image prompt is derived from verified or high-confidence source material rather than unconstrained user text. The resulting image is then compared with the story representation through a cross-modal alignment check and assigned a provenance state before publication. In the text-to-audio path, the narration script is grounded in the same evidence used for summary generation, which reduces the chance that the audio track introduces unsupported emphasis or claims.

For text-to-video workflows, scene templates can be anchored to story timelines and evidence spans, which helps prevent narrative drift and synthetic reenactments that exceed the source material. In the audio or video-to-text direction, low-confidence transcript spans are withheld from indexing and routed to correction queues. This proved especially important because even brief transcription errors could later surface as apparently authoritative evidence in other parts of the pipeline.

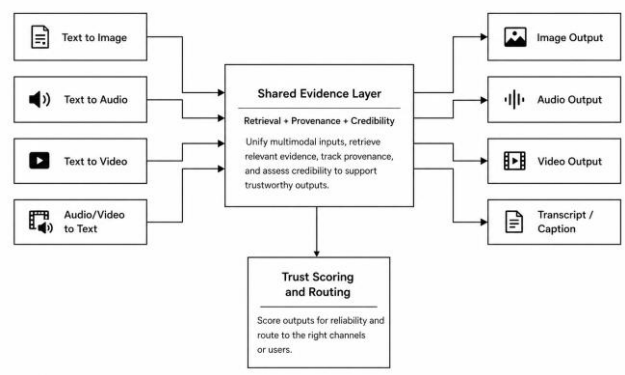

Figure 2: Cross-modal pathways with shared evidence, trust scoring, and routing

A repeated pattern during testing was that many quality issues only became visible when outputs were compared across modalities. A generated image might look plausible until it is evaluated against the summary, or a transcript might seem acceptable until its key quotes are compared against the source audio. Shared trust evaluation therefore improved overall coherence more than isolated modality checks alone [9], [10].

V. TRUST SCORING MODEL

Each candidate artifact receives a composite trust score before any delivery decision is made. The score is designed to summarize the most operationally relevant trust signals without collapsing them into a black box. In practice, editors and system owners still need to understand why an item scored poorly, so each component is stored separately even when a single score is used for routing [8].

T(i) = w1*S_support + w2*S_source + w3*S_provenance + w4*S_disclosure + w5*S_explain + w6*S_editor + w7*S_calibration + w8*S_search - w9*S_risk

Here, S_support measures evidence coverage, S_source captures source credibility, S_provenance reflects provenance completeness, S_disclosure measures policy-compliant disclosure, S_explain represents explanation quality, S_editor encodes editorial review outcome, S_calibration measures confidence reliability, S_search captures metadata validity, and S_risk aggregates policy and manipulation risk.

SCS(s) = v1*rep_score + v2*author_verified + v3*corroboration + v4*prov_completeness + v5*freshness - v6*correction_risk

XQS(i) = u1*evidence_coverage + u2*audit_trace + u3*review_clarity + u4*editor_use - u5*opacity

Weight calibration is handled through a hybrid process. Expert judgment provides initial weights, and empirical calibration against editorial decisions adjusts those values within bounded ranges. We found that purely learned weights were unstable across scenario shifts, while purely expert-defined weights were easier to interpret but slower to adapt. A hybrid approach gave more stable routing behavior.

The final routing decision uses the trust score together with relevance, risk, and latency. This prevents highly relevant content from bypassing a minimum trust floor simply because it is likely to attract engagement.

VI. OPERATIONAL CONSIDERATIONS

Introducing trust-aware controls adds both computational cost and latency. The most visible contributors are model inference, moderation and validation checks, provenance processing, and human review for borderline cases. In our

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 4, pp 412-417, April-2026
https://doi.org/10.47001/IRJIET/2026.104056

tests, the additional delay was noticeable but acceptable for most non-breaking workflows.

A useful operational insight was that better grounding reduced the number of items needing manual review. In other words, trust controls increase overhead upstream but can reduce editorial effort downstream. This matters because a system that improves quality at the cost of unbounded review load is not viable in production.

Latency-sensitive scenarios still require careful tuning. In a breaking-news context, it may be reasonable to allow a constrained fast path with temporary limitations on enrichment, provided the system records that decision and supports rapid correction if later checks fail. Trust-aware delivery therefore is not synonymous with maximum strictness; it is about applying controls deliberately and transparently [11],[12].

VII. EVALUATION

We evaluated the framework on a synthetic but structurally realistic workload built from approximately 1,200 articles spanning general news, business, technology, and evergreen explainers. Each run varied retrieval depth, evidence coverage, provenance availability, and metadata generation conditions. Five independently seeded runs were used to observe whether the trust-aware configuration remained stable across small distribution shifts.

The baseline system used retrieval-augmented generation without explicit trust scoring, provenance-aware routing, or metadata validation. The trust-aware version added the composite score, provenance classification, confidence-aware transcript handling, and pre-publication metadata checks. Although the workload is simulated, it was designed to reproduce the kinds of boundary cases that frequently expose weaknesses in production pipelines [13].

Table 2: Scenario-based evaluation results for the baseline and trust-aware configurations

Metric	Baseline	Trust-aware	Change
Unsupported claim rate	6.8%	2.1%	-69%
Provenance coverage	18%	94%	+422%
Search metadata validity	71%	96%	+35%
p95 latency	1.55 s	1.92 s	+24%

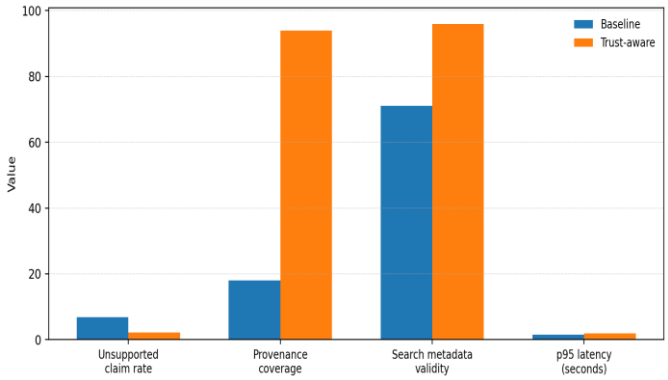

Figure 3: Scenario-based evaluation: baseline vs. trust-aware configuration

The trust-aware configuration consistently reduced unsupported claims and improved provenance and metadata quality. The most meaningful quality gains came from combining evidence grounding with routing, rather than from any single check in isolation. When we disabled the routing logic but kept the other controls, several borderline artifacts still reached publication because no decision layer consolidated the signals [14].

Residual failures were concentrated in low-evidence scenarios. When the retrieval layer returned sparse or ambiguous material, trust scoring often behaved correctly by escalating the output, but the baseline system tended to publish weakly supported content. This suggests that the main value of the framework is not merely to boost average quality, but to manage uncertainty more explicitly when evidence is poor [15], [16].

One of the clearest lessons from this work is that trust cannot be solved at the model level alone. Better models matter, but system-level controls determine how uncertainty is surfaced, how provenance is preserved, and when human review is invoked. The practical question is not how to build a perfect model, but how to build a system that behaves responsibly when models are imperfect.

Disclosure is another area where systems design matters. Labels are often treated as the visible end state of trust, yet poorly framed disclosures can reduce confidence without adding understanding. In our view, disclosure works best when it reflects underlying process information: what evidence supported the output, whether a human reviewed it, and what provenance status applies [17],[18].

Finally, provenance remains constrained by ecosystem adoption. Even a well-designed system cannot retroactively recover missing provenance for large legacy libraries. For that reason, the framework uses graded provenance states rather than a brittle verified-or-unverified split.

ISSN (online): 2581-3048
Volume 10, Issue 4, pp 412-417, April-2026
https://doi.org/10.47001/IRJIET/2026.104056

The evaluation is based on a synthetic workload rather than a live newsroom deployment. Although the scenarios were chosen to resemble realistic content flows, they cannot capture the full messiness of production behavior, audience interaction, or organizational workflow [19].

The recommended weight ranges for the trust score may not transfer directly across domains. Video-first publishers, enterprise assistants, and archival systems may require different thresholds and review policies. Human reviewer capacity also was not modeled in detail, which is an important practical consideration once escalation rates rise.

Finally, policy and platform conditions evolve continuously. Search guidance, disclosure expectations, and model behavior can change faster than a static trust policy. The framework therefore should be interpreted as a control structure that requires ongoing calibration rather than as a one-time configuration.

A natural next step is deployment in a real production environment with live editorial and audience feedback. A field study would make it possible to measure not only quality and latency, but also how trust-aware routing affects reviewer workload, correction rates, and user perception over time.

Additional work is also needed on adaptive weighting. Static weights are easier to govern, but some contexts may benefit from carefully bounded online adjustment using editorial outcomes and drift signals. Another promising direction is interoperability: if provenance and credibility registries could be shared across organizations, trust-aware systems would not need to rebuild the same evidence structures independently.

VIII. CONCLUSION

This paper presented a trust-aware framework for scoring and routing multimodal generative AI outputs. The central idea is straightforward: trust should influence whether and how content is delivered, not merely how it is labeled after the fact. By combining evidence support, source credibility, provenance, explanation quality, metadata validation, and risk into a single operational layer, the framework provides a practical way to manage uncertainty across text, image, audio, and transcription pathways.

Our results suggest that meaningful improvements in content reliability are achievable without making the system operationally impractical. The larger contribution, however, is architectural. Treating trust as a systems property creates a clearer path for governance than relying on model confidence or post hoc disclosure alone.

REFERENCES

[1] N. Newman, Journalism, Media and Technology Trends and Predictions 2026. *Oxford, U.K.: Reuters Institute*, 2026. doi:10.60625/risj-ps1d-np11.

[2] F. M. Simon, R. K. Nielsen, and R. Fletcher, Generative AI and News Report 2025. *Oxford, U.K.: Reuters Institute*, 2025. doi:10.60625/risj-5bjv-yt69.

[3] B. Toff and F. M. Simon, "The dilemma of AI disclosure for audience trust in news," *Int. J. Press/Politics*, 2025. doi: 10.1177/19401612241308697

[4] A.Nanz, A. Binder, and J. Matthes, "AI in the newsroom," *Journalism Studies.* doi: 10.1080/1461670X.2025.2547301

[5] K. P. Venkatraj *et al.*, "Understanding AI disclosure needs for news production," *in Proc. ACM Int. Conf. Mobile and Ubiquitous Multimedia*, 2025.

[6] C. Trattner *et al.*, "C2PA provenance labels increase trust in news platforms," doi: 2025.10.31219/osf.io/pdhaz_v1

[7] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework, AI 100-1, 2023.

[8] National Institute of Standards and Technology, Reducing Risks Posed by Synthetic Content, AI 100-4, 2024.

[9] Coalition for Content Provenance and Authenticity, *Technical Specification* v2.3, 2025.

[10] Q. V. Liao and J. Wortman Vaughan, "AI transparency in the age of LLMs," *Harvard Data Science Review*, 2024.doi: 10.48550/arXiv.2306.01941

[11] S. Wang *et al.*, "Trustworthy recommender systems," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 4, 2024.

[12] V. Resendez et al., "More than justifications," *Journalism Studies*, vol. 26, no. 11, 2025.

[13] C. Si *et al.*, "LLMs help humans verify truthfulness except when convincingly wrong," *in Proc. NAACL-HLT*, 2024.doi: 10.18653/v1/2024.naacl-long.81

[14] O. Schilke and M. Reimann, "The transparency dilemma: how AI disclosure erodes trust," *Organizational Behavior and Human Decision Processes*, vol. 188, 2025.doi: 10.1016/j.obhdp.2025.104405

[15] T. J. Thomson, R. J. Thomas, and P. Matich, "Generative visual AI in news organizations," *Digital Journalism*, 2024. doi: 10.1080/21670811.2024.2331769

[16] Google Search Central, "Guidance on generative AI content," Dec. 2025.

[17] Google Search Central, "Google Search Essentials," accessed Mar. 2026.

ISSN (online): 2581-3048
Volume 10, Issue 4, pp 412-417, April-2026
https://doi.org/10.47001/IRJIET/2026.104056

[18] A.Radford *et al.*, "Robust speech recognition via large-scale weak supervision," *in Proc. Int. Conf. Machine Learning*, 2023.doi: 10.48550/arXiv.2212.04356

[19] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022.

Citation of this Article:

Sujeet Sharma. (2026). A Trust-Aware Framework for Scoring and Routing in Multimodal Generative AI Systems. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(4), 412-417. Article DOI https://doi.org/10.47001/IRJIET/2026.104056

****

