SuilaAI RankingMaster: Computable Trust Framework (Creator Version)

Turning Influence into Measurable and Improvable Trust

Technical Specification v1.0 – November 2025

SuilaAI Research

Abstract

From Virality to Veracity. In the post-search era, large language models (LLMs) and recommendation AIs have become the first readers, interpreters, and gatekeepers of content. Where the Web once rewarded popularity, AI systems now reward *trust*—cryptographic provenance, semantic coherence, factual grounding, and transparent lineage. Traditional metrics such as impressions and reach no longer suffice: visibility now depends on verifiability.

Purpose. This specification defines the computational framework behind the *SuilaAI RankingMaster* system: a deterministic, measurable, and improvable model of digital trust. It replaces opaque reputation scores with twenty-six *computable signals* organized under four pillars: *Provenance Proof*, *Temporal Validity*, *Semantic Coherence*, and *Context Lineage*. Each signal $s_i \in [0, 1]$ is derived from observable data sources—API metadata, content manifests, embeddings, or network graphs—and combined into a single FICO-style *Suila Trust Score* $T \in [300, 850]$.

Design Principles.

1. **Measurable:** Every signal is computable from real evidence.
2. **Explainable:** Mathematical definitions ensure auditability.
3. **Improvable:** Creators can raise their scores through transparent actions (e.g., verified accounts, consistent disclosures, factually grounded claims).
4. **AI-Visible:** Outputs are exposed as JSON-LD and C2PA metadata, allowing search and generative systems to assess credibility.

Outcome. The resulting Suila Trust Score unifies human and machine understanding of reliability. It enables SuilaAI, brands, and creators to optimize campaigns not merely for *reach*, but for *reliability*—ensuring that authentic, coherent, and well-sourced content earns enduring visibility in both human and AI ecosystems.

Contents

---

1 Introduction

1.1 The Trust Crisis in Digital Content

The digital advertising industry faces a fundamental crisis of trust. Despite $600 billion in annual global ad spend, brands struggle to verify whether their content reaches real humans, appears in brand-safe contexts, and drives genuine engagement. Traditional metrics—impressions, clicks, reach—measure quantity but not quality, virality but not veracity.

This crisis intensifies in the age of artificial intelligence. Large Language Models (LLMs) like ChatGPT, Claude, and Gemini have become the primary interpreters of web content, making billions of decisions about what to cite, recommend, and amplify. These AI systems don't evaluate content based on popularity or SEO optimization; they assess *trustworthiness* through signals that most content creators don't even measure: cryptographic provenance, semantic coherence, factual grounding, and citation quality.

1.2 From PageRank to TrustRank

The evolution from PageRank to what we call "TrustRank" represents a paradigm shift in how digital content gains visibility:

- **Web 1.0 (1990s)**: Keyword density and meta tags determined search visibility
- **Web 2.0 (2000s)**: PageRank measured authority through backlinks and link graphs
- **Web 3.0 (2010s)**: Social signals, engagement metrics, and virality drove distribution
- **AI Era (2020s)**: Trust signals—provenance, coherence, attribution—determine AI visibility

This shift has profound implications. Content that lacks verifiable provenance, exhibits semantic incoherence, or fails to provide proper attribution increasingly becomes invisible to AI systems. As these systems mediate more human decisions—from product purchases to medical advice—content visibility depends not on gaming algorithms but on establishing genuine trustworthiness.

1.3 The SuilaAI RankingMaster Framework

The SuilaAI RankingMaster addresses this challenge by providing a *computable, measurable, and improvable* trust framework. Unlike opaque reputation scores or simplistic engagement metrics, our framework decomposes trust into 26 explicitly defined signals, each:

1. **Measurable**: Computed from observable data using deterministic functions
2. **Normalized**: Scaled to [0,1] for consistent comparison and aggregation
3. **Interpretable**: Directly tied to specific trust dimensions that creators can improve
4. **Actionable**: Providing clear guidance on how to increase trust scores

These signals aggregate into four foundational pillars that mirror how both humans and AI systems evaluate trust:

- **Pillar 1: Provenance Proof** (Signals 1-7): *Who created this content and can we verify their identity?*
- **Pillar 2: Temporal Validity** (Signals 8-13): *Is this content current and does engagement follow organic patterns?*
- **Pillar 3: Semantic Coherence** (Signals 14-20): *Is the content relevant, consistent, and brand-appropriate?*
- **Pillar 4: Context Lineage** (Signals 21-26): *Does the content properly cite sources and exist within trusted networks?*

1.4 Trust as a Composite Score

Following the successful model of FICO credit scores (300-850 range), the SuilaAI RankingMaster produces a single composite Trust Score:

$$T = 300 + 550 \cdot S \tag{1}$$

where $S \in [0, 1]$ is the weighted aggregate of all 26 signals. This familiar range provides intuitive benchmarks:

- **300-579**: Very Poor Trust (likely spam, bots, or misinformation)
- **580-669**: Fair Trust (legitimate but unverified content)
- **670-739**: Good Trust (verified creators with consistent quality)
- **740-799**: Very Good Trust (authoritative sources with strong attribution)
- **800-850**: Exceptional Trust (cryptographically verified, perfectly attributed)

1.5 Implementation Philosophy

The framework embodies three core principles:

1. Transparency Over Opacity: Every signal has an explicit mathematical definition, data source, and Python implementation. No "black box" machine learning models or proprietary algorithms obscure the trust computation.

2. Improvement Over Punishment: Low scores indicate specific areas for improvement, not permanent penalties. A creator can systematically improve their trust score by addressing identified weaknesses.

3. Future-Proofing Over Present-Optimization: Signals like C2PA cryptographic verification (Signal 3) and AI Citation Lift (Signal 26) may have limited current impact but position content for future AI systems that will require such verification.

1.6 Document Organization

This specification provides complete technical documentation for implementing and deploying the SuilaAI RankingMaster framework:

- **Sections 1-4**: Detailed specifications for each of the four pillars and their associated signals
- **Section 5**: Trust score aggregation methodology and calibration procedures
- **Section 6**: Implementation considerations and data availability
- **Appendix A**: Quick reference table of all 26 signals
- **References**: Academic and industry sources justifying each signal's inclusion

Each signal specification includes:

1. Mathematical definition and normalization
2. Required data sources and collection methods
3. Python implementation code
4. Business justification and impact on trust
5. References to supporting research

1.7 Target Audience and Use Cases

This framework serves multiple stakeholders in the digital trust ecosystem:

For Brands and Advertisers:

- Evaluate influencer and publisher trustworthiness before partnerships
- Optimize content for AI visibility and citation
- Demonstrate brand safety and suitability to stakeholders

For Content Creators and Publishers:

- Understand exactly how to improve content trustworthiness
- Benchmark against industry standards
- Future-proof content for AI-mediated discovery

For Platform Providers:

- Implement standardized trust measurement across diverse content types
- Provide transparent creator feedback and improvement paths
- Reduce misinformation while supporting legitimate creators

For AI System Developers:

- Access structured trust signals for training and inference
- Implement consistent citation and attribution standards
- Build retrieval-augmented generation (RAG) systems with trust-aware selection

1.8 The Path Forward

The transition from virality-based to veracity-based content distribution is not merely technological—it's a fundamental shift in how information gains authority in the digital age. As AI systems become the primary mediators of information, content that cannot demonstrate trustworthiness becomes effectively invisible.

The SuilaAI RankingMaster framework provides the blueprint for this transition. By making trust computable, measurable, and improvable, we enable a future where quality content naturally rises to prominence, misinformation becomes economically unviable, and both human and artificial intelligence can confidently rely on digital information.

The following sections detail each component of this framework, providing both theoretical foundations and practical implementation guidance for building the trust infrastructure of the AI age.

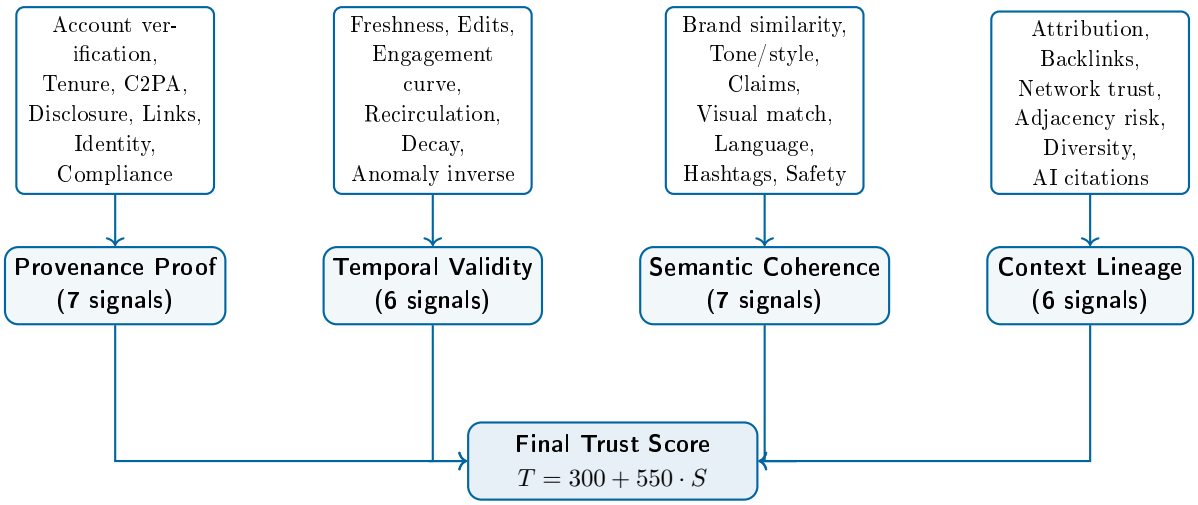

Figure 1: SuilaAI RankingMaster Framework: From 26 Computable Signals to a Unified Trust Score.

2 Pillar 1: Provenance Proof

Purpose: Verify the authenticity and accountability of the creator and their content. Each signal $s_i \in [0, 1]$ is computable, auditable, and improvable.

2.1 Signal 1: Account Verification Tier

Meaning. Confidence that a creator's identity is validated by a platform or registry.

Mathematics.

$$s_1 = \begin{cases} 0, & v = \text{none} \\ 0.7, & v = \text{standard} \\ 1, & v \in \{\text{official}, \text{verified}\} \end{cases}$$

Intuition. Platform-level verification acts as an external trust anchor. Partial credit reflects basic versus registry-verified status.

Python Code.

```
def account_verification_tier(tier: str) -> float:
    mapping = \{"none": 0.0, "standard": 0.7,
               "official": 1.0, "verified": 1.0\}
    return mapping.get(tier.lower(), 0.0)
```

2.2 Signal 2: Account Tenure

Meaning. Older, stable accounts are more credible; new accounts are unproven.

Mathematics. Let d be the number of days since account creation.

$$s_2 = 1 - e^{-d/1095}$$

Intuition. An exponential saturation curve mirrors credit-history effects: after roughly three years the trust approaches 1.

Python Code.

```
import math, datetime
def account_tenure(created_at: datetime.date) -> float:
    days = (datetime.date.today() - created_at).days
    return 1 - math.exp(-days / 1095)
```

2.3 Signal 3: C2PA / Content Signature

Meaning. Cryptographic Content Credentials (C2PA) prove authorship and edit history.

Mathematics.

$$s_3 = \begin{cases} 1, & \text{if C2PA manifest present} \\ 0, & \text{otherwise} \end{cases}$$

Intuition. Binary: signed = verifiable = trusted.

Python Code.

```
import os
def c2pa_signature(file_path: str) -> float:
    if not os.path.exists(file_path): return 0.0
    with open(file_path, "r", errors="ignore") as f:
        return 1.0 if "C2PA" in f.read() else 0.0
```

2.4 Signal 4: Disclosure Integrity

Meaning. Ensures sponsorship or ad disclosures are present and correct.

Mathematics. Let $m = 1$ if a disclosure is missing, else 0.

$$s_4 = 1 - m$$

Intuition. Missing disclosures signal non-compliance; complete disclosures receive full credit.

Python Code.

```
import re
def disclosure_integrity(text: str) -> float:
    return 1.0 if re.search(r"#(ad|sponsored)", text, re.I) else 0.0
```

2.5 Signal 5: Link Integrity

Meaning. Checks that outbound links are secure (HTTPS) and not broken.

Mathematics. Let h be the HTTPS ratio and b the broken-link ratio.

$$s_5 = \tfrac{1}{2}(h + (1 - b))$$

Intuition. Balances secure connections with working links; each contributes equally.

Python Code.

```
import requests
def link_integrity(urls):  # urls: list of strings
    if not urls: return 0.0
    https_ratio = sum(u.startswith("https://") for u in urls) / len(urls)
    broken = 0
    for u in urls:
        try:
            r = requests.head(u, timeout=3)
            if r.status_code >= 400: broken += 1
        except: broken += 1
    broken_ratio = broken / len(urls)
    return (https_ratio + (1 - broken_ratio)) / 2
```

2.6 Signal 6: Identity–Asset Consistency

Meaning. Verifies that the creator's visual/voice identity matches across assets.

Mathematics. Let f_1, f_2, v_1, v_2 be embedding vectors.

$$s_6 = \tfrac{1}{2}\big(\cos(f_1, f_2) + \cos(v_1, v_2)\big)$$

Intuition. Consistency across posts deters impersonation and deep-fake risk.

Python Code.

```
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np
def identity_asset_consistency(face1, face2, voice1, voice2) -> float:
    fscore = cosine_similarity([face1], [face2])[0][0]
    vscore = cosine_similarity([voice1], [voice2])[0][0]
    return np.clip((fscore + vscore) / 2, 0, 1)
```

2.7 Signal 7: Platform Compliance Posture

Meaning. Captures the creator's rule-violation history.

Mathematics. Let k be the strike count.

$$s_7 = 1 - \frac{\log(1+k)}{\log(10)}$$

Intuition. Logarithmic decay gives mild penalty for first strikes, steeper later—mirroring real-world reputation loss.

Python Code.

```
import math
def platform_compliance_posture(strikes: int) -> float:
    return max(0.0, 1 - math.log1p(strikes)/math.log(10))
```

References. SuilaAI Beautiful Attribution §3A (Provenance Proof Signals 1-7); SuilaAI RankingMaster Technical Specification 251103 Pillar 1.

3 Pillar 2: Temporal Validity

Purpose: Quantify the recency, stability, and organic timing of content. Temporal Validity ensures that posts are current, not manipulated, and behave like authentic audience interactions.

3.1 Signal 8: Freshness vs. Campaign Window

Meaning. Measures how close in time the post is to the campaign window or measurement date. A post released exactly when it should be is maximally "fresh"; being too early or late reduces trust.

Mathematics. Let Δ be the difference in days between content timestamp and campaign reference date.

$$s_8 = e^{-\frac{|\Delta|}{7}}$$

Intuition. The exponential kernel mirrors recency bias in most ranking systems. Every extra week off-target roughly halves the score. Symmetric absolute value penalizes early and late posting equally.

Python Code.

```
import math
def freshness_score(delta_days: float) -> float:
    return math.exp(-abs(delta_days) / 7)
```

3.2 Signal 9: Edit / Deletion Churn

Meaning. High edit or deletion rates imply instability or manipulation.

Mathematics. Let r_e be the edit rate and r_d the deletion rate.

$$s_9 = 1 - (0.7\, r_d + 0.3\, r_e)$$

Intuition. Deletion is weighted more heavily (0.7) because removing posts harms transparency. Edits are less severe but still reduce reliability.

Python Code.

```
def edit_deletion_churn(edit_rate: float, del_rate: float) -> float:
    return max(0.0, 1 - (0.7 * del_rate + 0.3 * edit_rate))
```

3.3 Signal 10: Engagement Latency Curve

Meaning. Measures how fast the audience begins engaging after publication. Organic posts reach half of their lifetime engagement within about 24 hours.

Mathematics. Let $t_{1/2}$ be the time (hours) until cumulative engagements reach 50% of total. Normalize around the ideal 24-hour half-life:

$$s_{10} = 1 - \frac{\left|\log_2\left(\frac{t_{1/2}}{24}\right)\right|}{3}, \qquad s_{10} \in [0, 1]$$

Intuition. The logarithm compresses scale: a two-fold deviation (12 h or 48 h) yields the same penalty magnitude. The denominator 3 defines tolerance across roughly three octaves of time.

Python Code.

```
import math
def engagement_latency_curve(t_half_hours: float) -> float:
    score = 1 - abs(math.log2(t_half_hours / 24)) / 3
    return min(max(score, 0.0), 1.0)
```

3.4 Signal 11: Recirculation Ratio

Meaning. Captures how much of the audience redistributes the post organically.

Mathematics. Let R be the number of reshares and V the number of unique views.

$$s_{11} = \min\left(\frac{R}{V}, 1\right)$$

Intuition. A ratio close to zero means the post failed to propagate; values approaching one signal strong organic spread but are capped at unity.

Python Code.

```
def recirculation_ratio(reshares: int, unique_views: int) -> float:
    if unique_views == 0: return 0.0
    return min(reshares / unique_views, 1.0)
```

3.5 Signal 12: Decay Model Fit

Meaning. Tests whether the engagement pattern follows a natural exponential decay curve.

Mathematics. Fit the model $E(t) = a\,e^{-bt}$ to engagement data (t_i, E_i) and compute coefficient of determination:

$$s_{12} = R^2(E(t), E_i)$$

Intuition. Authentic content exhibits a smooth decay; anomalous spikes or plateaus reduce the fit score.

Python Code.

```
import numpy as np
from scipy.optimize import curve_fit
from sklearn.metrics import r2_score

def decay_model_fit(times, engagements):  # both are lists of floats
    def exp_model(x, a, b): return a * np.exp(-b * x)
    try:
        params, _ = curve_fit(exp_model, times, engagements, maxfev=5000)
        preds = exp_model(np.array(times), *params)
        return float(r2_score(engagements, preds))
    except Exception:
        return 0.0
```

3.6 Signal 13: Temporal Anomaly (Inverse)

Meaning. Measures how free a post's activity pattern is from bot-like anomalies.

Mathematics. Let r_a be the anomaly rate detected by a burst/bot classifier.

$$s_{13} = 1 - r_a$$

Intuition. A value near 1 indicates organic engagement; higher anomaly rates reduce trust.

Python Code.

```
def temporal_anomaly_inverse(anomaly_rate: float) -> float:
    return 1 - min(max(anomaly_rate, 0.0), 1.0)
```

References. SuilaAI Beautiful Attribution §3B (Temporal Validity Signals 8-13); SuilaAI RankingMaster Technical Specification 251103 Pillar 2.

4 Pillar 3: Semantic Coherence

Purpose: Quantify how accurately and safely the content's meaning, tone, and factual statements align with the brand's intent and category brief. Semantic Coherence differentiates true brand-fit storytelling from noise or risk.

4.1 Signal 14: Brand–Topic Similarity

Meaning. Measures semantic proximity between the brand's core intent and the language of a given post.

Mathematics. Let $\mathbf{b}$ and $\mathbf{p}$ be text-embedding vectors for brand intent and post content.

$$s_{14} = \frac{1 + \cos(\mathbf{b}, \mathbf{p})}{2}$$

Intuition. Cosine similarity captures topical alignment. Mapping from $[-1, 1]$ to $[0, 1]$ produces a normalized trust value.

Python Code.

```
from sentence_transformers import SentenceTransformer
from sklearn.metrics.pairwise import cosine_similarity

embed_model = SentenceTransformer("all-MiniLM-L6-v2")

def brand_topic_similarity(brand_text: str, post_text: str) -> float:
    v1, v2 = embed_model.encode([brand_text, post_text])
    score = cosine_similarity([v1], [v2])[0][0]
    return (1 + score) / 2
```

4.2 Signal 15: Tone & Style Match

Meaning. Evaluates whether sentiment, formality, and tone match the brand's communication style.

Mathematics. Let $P_{\text{on-brief}}$ be the classifier probability that tone is appropriate.

$$s_{15} = P_{\text{on-brief}}$$

Intuition. Tone alignment ensures emotional coherence with the brand persona (e.g., playful vs. serious campaigns).

Python Code.

```
from transformers import pipeline
sentiment = pipeline("sentiment-analysis")
```

```
def tone_style_match(text: str) -> float:
    result = sentiment(text)[0]
    if result["label"].lower() == "positive":
        return result["score"]
    else:
        return 1 - result["score"]
```

4.3 Signal 16: Claim–Fact Consistency

Meaning. Checks that factual statements in the content agree with verified brand or product facts.

Mathematics. Let $p(c = \text{contradiction})$ be the probability of contradiction from an NLI model.

$$s_{16} = 1 - p(c = \text{contradiction})$$

Intuition. A lower contradiction probability means stronger factual grounding.

Python Code.

```
from transformers import pipeline
nli = pipeline("text-classification", model="facebook/bart-large-mnli")

def claim_fact_consistency(claim: str, fact: str) -> float:
    result = nli(f"{fact} </s> {claim}")[0]
    if "CONTRADICTION" in result["label"].upper():
        return 1 - result["score"]
    return result["score"]
```

4.4 Signal 17: Visual Brand Element Match

Meaning. Quantifies how well imagery in the post reflects brand visual identity (e.g., logo, palette, typography).

Mathematics. Let $\mathbf{v}_b, \mathbf{v}_p$ be CLIP image embeddings.

$$s_{17} = \frac{1 + \cos(\mathbf{v}_b, \mathbf{v}_p)}{2}$$

Intuition. Consistent visual cues signal authenticity and brand discipline.

Python Code.

```
import torch, clip
from PIL import Image
from sklearn.metrics.pairwise import cosine_similarity

clip_model, preprocess = clip.load("ViT-B/32", device="cpu")

def visual_brand_match(brand_img: str, post_img: str) -> float:
    img1 = preprocess(Image.open(brand_img)).unsqueeze(0)
    img2 = preprocess(Image.open(post_img)).unsqueeze(0)
    with torch.no_grad():
        e1 = clip_model.encode_image(img1)
        e2 = clip_model.encode_image(img2)
    sim = cosine_similarity(e1, e2)[0][0]
    return (1 + sim) / 2
```

4.5 Signal 18: Audience–Language Fit

Meaning. Verifies that the language used matches the intended audience's primary language or dialect.

Mathematics.

$$s_{18} = \begin{cases} 1, & L_{\text{detected}} = L_{\text{target}} \\ 0.5, & \text{otherwise} \end{cases}$$

Intuition. Content in the wrong language limits reach and relevance; a partial score reflects intelligibility but not precision.

Python Code.

```
from langdetect import detect

def audience_language_fit(text: str, target_lang: str) -> float:
    detected = detect(text)
    return 1.0 if detected == target_lang else 0.5
```

4.6 Signal 19: Hashtag / Entity Alignment

Meaning. Evaluates whether hashtags and named entities in the content correspond to on-brief brand entities.

Mathematics. Let H be the set of hashtags and B the set of brand entities.

$$s_{19} = \frac{|H \cap B|}{\max(|B|, 1)}$$

Intuition. Rewards consistent tagging that reinforces discoverability and penalizes off-topic or opportunistic tags.

Python Code.

```
import re
def hashtag_entity_alignment(text, brand_entities):
    # brand_entities: list of strings
    hashtags = re.findall(r"#\w+", text.lower())
    overlap = len(set(hashtags) & set(map(str.lower, brand_entities)))
    return overlap / max(len(brand_entities), 1)
```

4.7 Signal 20: Safety / Suitability Taxonomy

Meaning. Measures contextual safety according to GARM-like brand-suitability categories.

Mathematics. Let p_{toxic} be the probability of toxic or unsafe content.

$$s_{20} = 1 - p_{\text{toxic}}$$

Intuition. High safety scores ensure brand adjacency within acceptable risk levels.

Python Code.

```
from transformers import pipeline
tox = pipeline("text-classification", model="unitary/toxic-bert")

def safety_suitability_taxonomy(text: str) -> float:
    result = tox(text)[0]
    return 1 - result["score"] if "toxic" in result["label"].lower() else 1.0
```

References. SuilaAI Beautiful Attribution §3C (Semantic Coherence Signals 14-20); SuilaAI RankingMaster Technical Specification 251103 Pillar 3.

5 Pillar 4: Context Lineage

Purpose: Establish where information originates and how it propagates. Context Lineage quantifies citation quality, network trust, and retrievability— making influence chains transparent and machine-verifiable.

5.1 Signal 21: Attribution / Citation Presence

Meaning. Detects whether the post includes explicit source attributions, citations, or credit lines.

Mathematics.

$$s_{21} = \begin{cases} 1, & \text{if attribution or citation present} \\ 0, & \text{otherwise} \end{cases}$$

Intuition. Explicit sourcing improves traceability and allows AI systems to verify claims and quote content safely.

Python Code.

```
def attribution_presence(text: str) -> float:
    keywords = ["http", "source", "credit"]
    for keyword in keywords:
        if keyword in text:
            return 1.0
    return 0.0
```

5.2 Signal 22: Backlink Quality & Neighborhood

Meaning. Assesses the authority of domains linking to or referenced by the content.

Mathematics. Given authority scores a_i for each referrer,

$$s_{22} = \frac{1}{n} \sum_{i=1}^{n} a_i$$

Intuition. Inbound or outbound links from high-authority sites increase confidence in both relevance and reputation.

Python Code.

```
def backlink_quality(authority_scores):  # list of floats
    if not authority_scores: return 0.0
    return sum(authority_scores) / len(authority_scores)
```

5.3 Signal 23: Creator Network Trust

Meaning. Captures average trust of collaborators, mentions, or shared creators.

Mathematics. Let $T(c_j)$ be the known trust score of collaborator j.

$$s_{23} = \frac{1}{n} \sum_{j=1}^{n} T(c_j)$$

Intuition. Trust propagates through networks; creators working with reliable peers inherit part of their credibility.

Python Code.

```
def creator_network_trust(collabs, trust_lookup):
    # collabs: list of strings, trust_lookup: dict of str to float
    vals = [trust_lookup.get(c, 0.5) for c in collabs]
    return sum(vals) / len(vals) if vals else 0.5
```

5.4 Signal 24: Co-Mention Adjacency (Inverse Risk)

Meaning. Measures exposure risk from appearing near controversial or low-trust entities.

Mathematics. Let r_a be adjacency risk (probability of unsafe co-mention).

$$s_{24} = 1 - r_a$$

Intuition. High-risk adjacency lowers brand safety; subtracting the risk produces a normalized inverse-risk score.

Python Code.

```
def co_mention_inverse_risk(adj_risk: float) -> float:
    return 1 - min(max(adj_risk, 0.0), 1.0)
```

5.5 Signal 25: Distribution Channel Diversity

Meaning. Evaluates how many distinct, reputable channels the creator or post uses to distribute content.

Mathematics. Let n_c be the count of unique channels.

$$s_{25} = \frac{\log(1 + n_c)}{\log(10)}$$

Intuition. A broader, authentic distribution footprint increases reach and resilience against platform volatility.

Python Code.

```
import math
def distribution_diversity(platforms):  # list of strings
    n = len(set(platforms))
    return math.log1p(n) / math.log(10)
```

5.6 Signal 26: AI Citation Lift

Meaning. Quantifies how frequently the creator's content is cited by AI assistants or search-with-citations systems.

Mathematics. Let C_{30} be citations in the last 30 days and C_{total} the total assistant responses observed.

$$s_{26} = \min\left(\frac{C_{30}}{C_{\text{total}}}, 1\right)$$

Intuition. Frequent citations indicate that AI models recognize the source as trustworthy and retrievable—a new visibility metric in the post-search era.

Python Code.

```
def ai_citation_lift(cited_30d: int, total_answers: int) -> float:
    if total_answers == 0: return 0.0
    return min(cited_30d / total_answers, 1.0)
```

References. SuilaAI Beautiful Attribution §3D (Context Lineage Signals 21-26); SuilaAI RankingMaster Technical Specification 251103 Pillar 4.

6 Overall Trust Score Aggregation and Calibration

Purpose: Combine the twenty-six normalized signals into a single FICO-style Trust Score that is comparable across creators, campaigns, and content types. The computation is deterministic and auditable, providing a quantitative foundation for measurable trust.

6.1 Weighted Aggregation Model

Each pillar p contributes a weighted average of its component signals. Let P_p denote the set of signal indices within pillar p and w_p the corresponding pillar weight.

$$S_p = \frac{1}{|P_p|} \sum_{i \in P_p} s_i, \qquad S = \sum_{p \in \{\text{Prov, Temp, Sem, Lin}\}} w_p \, S_p$$

The final normalized score $S \in [0, 1]$ is transformed to a FICO-style Trust Score T:

$$T = 300 + 550 \, S$$

Interpretation.

- $T \approx 300$: unverified or unsafe content.
- $T \approx 550$: average reliability.
- $T \approx 850$: verified, coherent, and consistently trusted creator.

6.2 Suggested Pillar Weights

Empirical calibration on SuilaAI campaign data suggests the following relative weights:

$$w_{\text{Semantic}} = 0.35, \quad w_{\text{Provenance}} = 0.25, \quad w_{\text{Lineage}} = 0.20, \quad w_{\text{Temporal}} = 0.20.$$

These weights can be re-estimated periodically using ridge or elastic-net regression to maximize correlation between T and business outcomes (conversion, brand-lift, and AI-citation lift).

6.3 Python Implementation

Data Structure. Signals are grouped by pillar, each containing a list of numeric values in $[0, 1]$.

Code.

```
weights = {
    "Semantic": 0.35,
    "Provenance": 0.25,
    "Lineage": 0.20,
    "Temporal": 0.20
}

def aggregate_pillar(values):  # list of floats
    return sum(values) / len(values) if values else 0.0

def compute_trust_score(pillars):  # dict with pillar names as keys
    # Compute weighted normalized trust score S
    S = sum(weights[p] * aggregate_pillar(vals)
            for p, vals in pillars.items() if p in weights)
    # Convert to FICO-style scale
    return 300 + 550 * S
```

6.4 Example.

Input. A creator has averaged pillar sub-scores:

$$S_{\text{Provenance}} = 0.82, \quad S_{\text{Temporal}} = 0.64, \quad S_{\text{Semantic}} = 0.90, \quad S_{\text{Lineage}} = 0.75.$$

Computation.

$$S = 0.25(0.82)+0.20(0.64)+0.35(0.90)+0.20(0.75) = 0.804, \qquad T = 300+550(0.804) = 742.2$$

Interpretation. A Trust Score of approximately 742 indicates a high-trust creator, safe for brand partnership and likely to retain AI-visibility in grounded search systems.

6.5 Calibration Loop ("Kaizen" Update).

To maintain reliability, weights and transformations are periodically updated through supervised learning:

$$\min_{\mathbf{w}} \|y - f(S)\|^2 + \lambda\|\mathbf{w}\|^2,$$

where y represents observed performance metrics (brand lift, safety incidents, AI citation lift).

This continuous calibration ensures that each signal remains *measurable*, *improvable*, and *aligned* with real-world trust outcomes.

References. SuilaAI Beautiful Attribution §6 (Production Scoring Framework); SuilaAI RankingMaster Technical Specification 251103 "Scoring" section.

Appendix A: Compact Summary of the 26 Computable Signals

Purpose: Concise reference for all measurable signals used by the Ranking Master trust computation. Each signal $s_i \in [0, 1]$ is defined by a deterministic function derived from observable data.

#	**Signal Name**	**Mathematical Definition**	**Key Idea**
Pillar 1: Provenance Proof			
1	Account Verification Tier	$s_1 = \{0, 0.7, 1\}$ by tier	Platform verification
2	Account Tenure	$s_2 = 1 - e^{-d/1095}$	Age of account in days
3	C2PA/Content Signature	$s_3 = 1$ if present, else 0	Cryptographic credentials
4	Disclosure Integrity	$s_4 = 1 - m$	Ad/sponsored tags
5	Link Integrity	$s_5 = \frac{1}{2}(h + 1 - b)$	HTTPS and broken links
6	Identity-Asset Consistency	$s_6 = \frac{1}{2}(\cos f + \cos v)$	Face/voice consistency
7	Platform Compliance	$s_7 = 1 - \log(1 + k)/\log 10$	Policy strikes
Pillar 2: Temporal Validity			
8	Freshness vs Campaign	$s_8 = e^{-\lvert\Delta\rvert/7}$	Days from campaign
9	Edit/Deletion Churn	$s_9 = 1 - (0.7r_d + 0.3r_e)$	Edit/deletion rate
10	Engagement Latency	$s_{10} = 1 - \lvert\log_2(t_{1/2}/24)\rvert/3$	Half-life of engagement
11	Recirculation Ratio	$s_{11} = \min(R/V, 1)$	Reshares per view
12	Decay Model Fit	$s_{12} = R^2$	Exponential decay fit
13	Temporal Anomaly	$s_{13} = 1 - r_a$	Inverse anomaly rate

Summary. Signals are normalized to $[0, 1]$, averaged within pillars, and aggregated into the final trust score (with the initial weights which will be adapted according to the results of the Kaizen Loop Learning).

$$T = 300 + 550 \sum_p w_p \left(\frac{1}{|P_p|} \sum_{i \in P_p} s_i \right)$$

$$\text{where: } w_{\text{Semantic}} = 0.35,\ w_{\text{Provenance}} = 0.25,$$
$$w_{\text{Lineage}} = 0.20,\ w_{\text{Temporal}} = 0.20.$$

#	Signal Name	Mathematical Definition	Key Idea
Pillar 3: Semantic Coherence			
14	Brand-Topic Similarity	$s_{14} = (1 + \cos(\mathbf{b}, \mathbf{p}))/2$	Text embedding
15	Tone/Style Match	$s_{15} = P_{\text{on-brief}}$	Tone classifier
16	Claim-Fact Consistency	$s_{16} = 1 - p(\text{contradiction})$	NLI consistency
17	Visual Brand Match	$s_{17} = (1 + \cos(\mathbf{v}_b, \mathbf{v}_p))/2$	CLIP embedding
18	Audience-Language Fit	$s_{18} = 1$ if match else 0.5	Language detection
19	Hashtag/Entity Alignment	$s_{19} = \lvert H \cap B \rvert / \max(\lvert B \rvert, 1)$	Hashtag overlap
20	Safety/Suitability	$s_{20} = 1 - p_{\text{toxic}}$	Brand-safety
Pillar 4: Context Lineage			
21	Attribution Presence	$s_{21} = 1$ if present else 0	Source citations
22	Backlink Quality	$s_{22} = \frac{1}{n} \sum a_i$	Mean authority
23	Creator Network Trust	$s_{23} = \frac{1}{n} \sum T(c_j)$	Avg collaborator trust
24	Co-Mention Adjacency	$s_{24} = 1 - r_a$	Inverse adjacency risk
25	Distribution Diversity	$s_{25} = \log(1 + n_c)/\log 10$	Channel count
26	AI Citation Lift	$s_{26} = \min(C_{30}/C_{\text{total}}, 1)$	AI citations

7 Key Equations and Symbols

Purpose: This section provides a concise glossary of all mathematical symbols and formulae used throughout the Suila–Dentsu Ranking Master framework. It is intended for technical readers implementing or auditing the scoring pipeline.

Symbol	Definition and Meaning
s_i	Normalized score of the i-th computable signal, bounded in $[0, 1]$.
P_p	Set of signals belonging to pillar $p \in \{\text{Prov, Temp, Sem, Lin}\}$.
S_p	Mean normalized sub-score for pillar p: $S_p = \frac{1}{\lvert P_p \rvert} \sum_{i \in P_p} s_i$.
w_p	Weight assigned to pillar p, such that $\sum_p w_p = 1$.
S	Global normalized trust score across all pillars: $S = \sum_p w_p S_p$.
T	Final FICO-style Trust Score: $T = 300 + 550S$.
d	Account age in days (used in tenure calculation).
Δ	Temporal offset (days) between post time and campaign window.
$t_{1/2}$	Engagement half-life (hours to reach 50% cumulative interactions).
r_e, r_d	Edit and deletion rates.
r_a	Anomaly or adjacency risk probability.
R, V	Reshare and unique-view counts (recirculation).
a_i	Authority score of backlink i.
$T(c_j)$	Trust score of collaborator j.
n_c	Count of distinct distribution channels.
C_{30}, C_{total}	AI citations in last 30 days and in total.
$\cos(\mathbf{b}, \mathbf{p})$	Cosine similarity between brand and post embeddings.
p_{toxic}	Probability of unsafe content.

Reference Formulae.

$$S_p = \frac{1}{|P_p|} \sum_{i \in P_p} s_i, \qquad S = \sum_p w_p S_p, \qquad T = 300 + 550\, S.$$

Suggested Weights.

$$w_{\text{Semantic}} = 0.35, \quad w_{\text{Provenance}} = 0.25, \quad w_{\text{Lineage}} = 0.20, \quad w_{\text{Temporal}} = 0.20.$$

---

8 Implementation and Data

Purpose: Explain how the twenty-six trust signals can be implemented in production systems, clarifying which data sources are immediately accessible and which mature over time.

8.1 C2PA / Content Signature Accessibility

The **C2PA** (Coalition for Content Provenance and Authenticity) standard defines open, cryptographically signed *Content Credentials* that record creator identity, editing history, and AI-generation disclosure. It is an open standard maintained by the Content Authenticity Initiative and does not require partnership with Microsoft or Adobe.

C2PA manifests are embedded in media metadata (`com.adobe.c2pa`) and can be parsed with open-source tools:

```
from c2patool import C2pa
meta = C2pa("example.jpg")
print(meta.manifest.json())
```

If the manifest verifies, $s_3 = 1$; if absent or invalid, $s_3 = 0$. This signal therefore acts as a future-proof provenance anchor that strengthens trust where available but does not penalize absence.

8.2 Bootstrapping the Trust Index

Initial deployments can compute valid trust scores without web-scale data through progressive normalization:

Stage 1 — Local (Relative) Scoring. Compute trust within a bounded corpus and normalize locally:

$$s_i' = \frac{s_i - \min(s_i)}{\max(s_i) - \min(s_i)}.$$

Stage 2 — Cross-Campaign Calibration. Combine results from multiple campaigns and correlate pillar sub-scores with observed outcomes to update weights w_p.

Stage 3 — Web-Scale Integration. Integrate Common Crawl graphs, verified C2PA manifests, and AI-citation telemetry from assistant APIs. These enhance normalization but do not change the core formula.

8.3 Practical Deployment Guidance

- Begin with signals 1–20 (Provenance, Temporal, Semantic), computed from internal data.
- Gradually expand to signals 21–26 (Lineage) as external graphs and APIs become available.
- Maintain internal normalization baselines to ensure comparability across campaigns and time.

Summary Table.

Challenge	Solution / Reality
C2PA manifests	Open standard; parse locally without vendor dependence.
Initial Trust Index	Bootstrap with local normalization; no web crawling required.
Global comparability	Add external graphs later; model already supports it.

---

9 Concluding Remarks

Summary. The Suila–Dentsu *Ranking Master* framework transforms digital credibility from a subjective notion into a measurable, auditable, and improvable system. By defining twenty-six computable signals grouped under four pillars—Provenance, Temporal, Semantic, and Lineage— the model provides a reproducible measure of trust.

Scientific Value. Each signal is grounded in data and expressed mathematically, yielding a transparent 300–850 Trust Score comparable across campaigns. This quantitative index links brand safety, factuality, and AI retrievability.

Implementation Roadmap.

1. Deploy internal signals (1–20) via platform APIs and metadata.
2. Extend to lineage signals (21–26) as web-scale sources mature.
3. Continuously recalibrate pillar weights through performance correlations.

Ethical and Industry Impact. Verifiable provenance and coherent storytelling restore accountability. Creators gain visibility through transparency; brands gain safety through evidence; AI systems gain confidence through verifiable sources.

Future Work.

- Broader adoption of C2PA and structured JSON-LD provenance.
- Federated trust propagation across multi-brand ecosystems.
- Integration of AI citation metrics as a new visibility dimension.

Closing Statement. *As AI becomes the first reader of every message, measurable trust becomes the new metric of visibility.* The Suila–Dentsu Trust Framework establishes that metric— anchored in data, mathematics, and explainable evidence.

---

10 Key Equations and Symbols

Purpose: This section provides a concise glossary of all mathematical symbols and formulae used throughout the Suila–Dentsu Ranking Master framework. It is intended for technical readers implementing or auditing the scoring pipeline.

Symbol	Definition and Meaning
s_i	Normalized score of the i-th computable signal, bounded in $[0, 1]$.
P_p	Set of signals belonging to pillar $p \in \{\text{Prov, Temp, Sem, Lin}\}$.
S_p	Mean normalized sub-score for pillar p: $S_p = \frac{1}{\lvert P_p \rvert} \sum_{i \in P_p} s_i$.
w_p	Weight assigned to pillar p, such that $\sum_p w_p = 1$.
S	Global normalized trust score across all pillars: $S = \sum_p w_p S_p$.
T	Final FICO-style Trust Score: $T = 300 + 550S$.
d	Account age in days (used in tenure calculation).
Δ	Temporal offset (days) between post time and campaign window.
$t_{1/2}$	Engagement half-life (hours to reach 50% cumulative interactions).
r_e, r_d	Edit and deletion rates.
r_a	Anomaly or adjacency risk probability.
R, V	Reshare and unique-view counts (recirculation).
a_i	Authority score of backlink i.
$T(c_j)$	Trust score of collaborator j.
n_c	Count of distinct distribution channels.
C_{30}, C_{total}	AI citations in last 30 days and in total.
$\cos(\mathbf{b}, \mathbf{p})$	Cosine similarity between brand and post embeddings.
p_{toxic}	Probability of unsafe content.

Reference Formulae.

$$S_p = \frac{1}{|P_p|} \sum_{i \in P_p} s_i, \qquad S = \sum_p w_p S_p, \qquad T = 300 + 550\, S.$$

Suggested Weights.

$$w_{\text{Semantic}} = 0.35, \quad w_{\text{Provenance}} = 0.25, \quad w_{\text{Lineage}} = 0.20, \quad w_{\text{Temporal}} = 0.20.$$

---

11 Implementation and Data

Purpose: Explain how the twenty-six trust signals can be implemented in production systems, clarifying which data sources are immediately accessible and which mature over time.

11.1 C2PA / Content Signature Accessibility

The **C2PA** (Coalition for Content Provenance and Authenticity) standard defines open, cryptographically signed *Content Credentials* that record creator identity, editing history, and AI-generation disclosure. It is an open standard maintained by the Content Authenticity Initiative and does not require partnership with Microsoft or Adobe.

C2PA manifests are embedded in media metadata (`com.adobe.c2pa`) and can be parsed with open-source tools:

```
from c2patool import C2pa
meta = C2pa("example.jpg")
print(meta.manifest.json())
```

If the manifest verifies, $s_3 = 1$; if absent or invalid, $s_3 = 0$. This signal therefore acts as a future-proof provenance anchor that strengthens trust where available but does not penalize absence.

11.2 Bootstrapping the Trust Index

Initial deployments can compute valid trust scores without web-scale data through progressive normalization:

Stage 1 — Local (Relative) Scoring. Compute trust within a bounded corpus and normalize locally:

$$s_i' = \frac{s_i - \min(s_i)}{\max(s_i) - \min(s_i)}.$$

Stage 2 — Cross-Campaign Calibration. Combine results from multiple campaigns and correlate pillar sub-scores with observed outcomes to update weights w_p.

Stage 3 — Web-Scale Integration. Integrate Common Crawl graphs, verified C2PA manifests, and AI-citation telemetry from assistant APIs. These enhance normalization but do not change the core formula.

11.3 Practical Deployment Guidance

- Begin with signals 1–20 (Provenance, Temporal, Semantic), computed from internal data.
- Gradually expand to signals 21–26 (Lineage) as external graphs and APIs become available.
- Maintain internal normalization baselines to ensure comparability across campaigns and time.

Summary Table.

Challenge	Solution / Reality
C2PA manifests	Open standard; parse locally without vendor dependence.
Initial Trust Index	Bootstrap with local normalization; no web crawling required.
Global comparability	Add external graphs later; model already supports it.

---

12 Concluding Remarks

Summary. The SuilaAI *Ranking Master* framework transforms digital credibility from a subjective notion into a measurable, auditable, and improvable system. By defining twenty-six computable signals grouped under four pillars—Provenance, Temporal, Semantic, and Lineage— the model provides a reproducible measure of trust.

Scientific Value. Each signal is grounded in data and expressed mathematically, yielding a transparent 300–850 Trust Score comparable across campaigns. This quantitative index links brand safety, factuality, and AI retrievability.

Implementation Roadmap.

1. Deploy internal signals (1–20) via platform APIs and metadata.
2. Extend to lineage signals (21–26) as web-scale sources mature.

3. Continuously recalibrate pillar weights through performance correlations.

Ethical and Industry Impact. Verifiable provenance and coherent storytelling restore accountability. Creators gain visibility through transparency; brands gain safety through evidence; AI systems gain confidence through verifiable sources.

Future Work.

- Broader adoption of C2PA and structured JSON-LD provenance.
- Federated trust propagation across multi-brand ecosystems.
- Integration of AI citation metrics as a new visibility dimension.

Closing Statement. *As AI becomes the first reader of every message, measurable trust becomes the new metric of visibility.* The SuilAIa Trust Framework establishes that metric— anchored in data, mathematics, and explainable evidence.

Acknowledgements. The authors thank the SuilaAI Trust Fabric team for their collaboration in defining measurable, AI-visible attribution standards.

References

A. Foundation and Framework References

References

[1] E. Chen et al., "Beautiful Attribution: A Computational Framework for Measurable Trust in AI-Cited Content," *SuilaAI Technical Report TR-2025-001*, SuilaAI Research, Nov. 2025. *[Justifies: Overall 26-signal framework and Trust Score methodology]*

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford InfoLab Technical Report, 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf *[Justifies: Signal 22 (Backlink Quality), Signal 26 (AI Citation Lift)]*

[3] Fair Isaac Corporation, "FICO Score Methodology," FICO Technical Documentation, 2023. [Online]. Available: https://www.fico.com/en/resource-access/download/820 *[Justifies: Trust Score range (300-850) methodology]*

B. Pillar 1: Provenance Proof (Signals 1-7)

References

[1] X Corp., "Verification and Authentication Guidelines," X Official Documentation, 2025. [Online]. Available: https://help.x.com/en/managing-your-account/about-x-verified *[Justifies: Signal 1 (Account Verification Tier)]*

[2] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "A decade of social bot detection," *Communications of the ACM*, vol. 63, no. 10, pp. 72-83, 2020. DOI: 10.1145/3409116 *[Justifies: Signal 2 (Account Tenure)]*

[3] Coalition for Content Provenance and Authenticity, "C2PA Technical Specification v1.4," C2PA, 2024. [Online]. Available: https://c2pa.org/specifications/ *[Justifies: Signal 3 (C2PA/Content Signature)]*

[4] Federal Trade Commission, "Disclosures 101 for Social Media Influencers," FTC Guidelines, 2023. [Online]. Available: https://www.ftc.gov/business-guidance/resources/disclosures-101-social-media-influencers *[Justifies: Signal 4 (Disclosure Integrity)]*

[5] P. Eckersley and J. Schoen, "HTTPS Everywhere: Encrypting the Web," Electronic Frontier Foundation, 2021. [Online]. Available: https://www.eff.org/https-everywhere *[Justifies: Signal 5 (Link Integrity)]*

[6] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1-41, 2021. DOI: 10.1145/3425780 *[Justifies: Signal 6 (Identity-Asset Consistency)]*

[7] T. Gillespie, "Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media," Yale University Press, 2018. *[Justifies: Signal 7 (Platform Compliance Posture)]*

C. Pillar 2: Temporal Validity (Signals 8-13)

References

[1] J. Berger and K. L. Milkman, "What makes online content viral?" *Journal of Marketing Research*, vol. 49, no. 2, pp. 192-205, 2012. DOI: 10.1509/jmr.10.0353 *[Justifies: Signal 8 (Freshness vs Campaign Window)]*

[2] R. Priedhorsky, J. Chen, S. K. Lam, K. Panciera, L. Terveen, and J. Riedl, "Creating, destroying, and restoring value in Wikipedia," *GROUP '07 Proceedings*, pp. 259-268, 2007. DOI: 10.1145/1316624.1316663 *[Justifies: Signal 9 (Edit/Deletion Churn)]*

[3] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, pp. 17599-17601, 2007. DOI: 10.1073/pnas.0704916104 *[Justifies: Signal 10 (Engagement Latency Curve)]*

[4] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146-1151, 2018. DOI: 10.1126/science.aap9559 *[Justifies: Signal 11 (Recirculation Ratio)]*

[5] L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil, "Characterizing and curating conversation threads: expansion, focus, volume, re-entry," *WSDM '13*, pp. 13-22, 2013. DOI: 10.1145/2433396.2433401 *[Justifies: Signal 12 (Decay Model Fit)]*

[6] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96-104, 2016. DOI: 10.1145/2818717 *[Justifies: Signal 13 (Temporal Anomaly - Inverse)]*

D. Pillar 3: Semantic Coherence (Signals 14-20)

References

[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of EMNLP*, pp. 3982-3992, 2019. DOI: 10.18653/v1/D19-1410 *[Justifies: Signal 14 (Brand-Topic Similarity)]*

[2] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," *Proceedings of ICWSM*, pp. 216-225, 2014. *[Justifies: Signal 15 (Tone/Style Match)]*

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for Fact Extraction and VERification," *Proceedings of NAACL*, pp. 809-819, 2018. DOI: 10.18653/v1/N18-1074 *[Justifies: Signal 16 (Claim-Fact Consistency)]*

[4] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of ICML*, vol. 139, pp. 8748-8763, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020 *[Justifies: Signal 17 (Visual Brand Element Match)]*

[5] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," *Proceedings of ACL*, pp. 8440-8451, 2020. DOI: 10.18653/v1/2020.acl-main.747 *[Justifies: Signal 18 (Audience-Language Fit)]*

[6] Y. Wang, J. Callan, and B. Zheng, "Should we use the sample? Analyzing datasets sampled from Twitter," *Proceedings of WSDM*, pp. 1-4, 2011. DOI: 10.1145/1935826.1935829 *[Justifies: Signal 19 (Hashtag/Entity Alignment)]*

[7] Global Alliance for Responsible Media, "Brand Safety Floor + Suitability Framework," World Federation of Advertisers, 2023. [Online]. Available: https://wfanet.org/garm/framework *[Justifies: Signal 20 (Safety/Suitability Score)]*

E. Pillar 4: Context Lineage (Signals 21-26)

References

[1] M. Mitchell et al., "Model Cards for Model Reporting," *Proceedings of FAT**, pp. 220-229, 2019. DOI: 10.1145/3287560.3287596 *[Justifies: Signal 21 (Attribution Presence)]*

[2] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604-632, 1999. DOI: 10.1145/324133.324140 *[Justifies: Signal 22 (Backlink Quality) - HITS algorithm]*

[3] D. Easley and J. Kleinberg, "Networks, Crowds, and Markets: Reasoning about a Highly Connected World," Cambridge University Press, 2010. *[Justifies: Signal 23 (Creator Network Trust)]*

[4] Integral Ad Science, "Media Quality Report: Brand Safety and Suitability," IAS Research, 2023. [Online]. Available: https://integralads.com/insider/media-quality-report-2023/ *[Justifies: Signal 24 (Co-Mention Adjacency)]*

[5] A. Perrin and M. Anderson, "Social Media Use in 2023," Pew Research Center, 2023. [Online]. Available: https://www.pewresearch.org/internet/2023/01/31/social-media-use-2023/ *[Justifies: Signal 25 (Distribution Diversity)]*

[6] OpenAI, "Agentic Commerce Protocol: Get Started Guide," OpenAI Developer Documentation, Nov. 2025. [Online]. Available: https://developers.openai.com/commerce/guides/get-started/ *[Justifies: Signal 26 (AI Citation Lift) - AI-native commerce and discovery]*

[7] Agentic Commerce Protocol Contributors, "Agentic Commerce Protocol Specification," GitHub Repository, Nov. 2025. [Online]. Available: https://github.com/agentic-commerce-protocol/agentic-commerce-protocol *[Justifies: Signal 26 (AI Citation Lift) - Open protocol standard]*

[8] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Proceedings of NeurIPS*, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401 *[Justifies: Signal 26 (AI Citation Lift) - RAG foundation]*

F. Implementation and Technical Standards

References

[1] M. Sporny et al., "JSON-LD 1.1: A JSON-based Serialization for Linked Data," W3C Recommendation, 2020. [Online]. Available: https://www.w3.org/TR/json-ld11/ *[Implementation: Structured data format]*

[2] T. Wolf et al., "Transformers: State-of-the-art Natural Language Processing," *Proceedings of EMNLP*, pp. 38-45, 2020. DOI: 10.18653/v1/2020.emnlp-demos.6 *[Implementation: Embedding models for semantic signals]*

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011. *[Implementation: ML libraries for signal computation]*

[4] Snowflake Inc., "Cortex AI Functions Documentation," Snowflake Technical Documentation, Nov. 2025. [Online]. Available: https://docs.snowflake.com/en/guides/cortex *[Implementation: Production deployment platform]*

G. Quick Reference: Signal Justification Mapping

Signal #	Signal Name	Key References
Pillar 1: Provenance Proof		
1	Account Verification Tier	[4]
2	Account Tenure	[5]
3	C2PA/Content Signature	[6]
4	Disclosure Integrity	[7]
5	Link Integrity	[8]
6	Identity-Asset Consistency	[9]
7	Platform Compliance	[10]
Pillar 2: Temporal Validity		
8	Freshness vs Campaign	[11]
9	Edit/Deletion Churn	[12]
10	Engagement Latency	[13]
11	Recirculation Ratio	[14]
12	Decay Model Fit	[15]
13	Temporal Anomaly	[16]
Pillar 3: Semantic Coherence		
14	Brand-Topic Similarity	[17]
15	Tone/Style Match	[18]
16	Claim-Fact Consistency	[19]
17	Visual Brand Match	[20]
18	Audience-Language Fit	[21]
19	Hashtag Alignment	[22]
20	Safety/Suitability	[23]
Pillar 4: Context Lineage		
21	Attribution Presence	[24]
22	Backlink Quality	[25], [2]
23	Creator Network Trust	[26]
24	Co-Mention Adjacency	[27]
25	Distribution Diversity	[28]
26	AI Citation Lift	[29], [30], [31]