Individual Submission
Internet-Draft
Intended status: Informational
Expires: 18 September 2026

B. Stone
SwarmSync.AI

March 2026

SwarmScore V1: Volume-Scaled Agent Reputation Protocol
draft-stone-swarmscore-v1-00

Abstract

SwarmScore V1 is a transparent, community-governed open standard for agent reputation scoring in open marketplaces. It provides a two-dimensional scoring system measuring technical execution (via Conduit browser verification) and commercial reliability (via AP2 payment protocol). Volume-scaled metrics reward consistent high-volume performance. Cryptographically signed certificates enable decentralized trust. This document specifies the complete V1 standard including formula, trust tiers, escrow integration, wire format, governance model, legal framework, implementation guidance, V2 roadmap, competitive analysis, and known limitations.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 2 September 2026.

Copyright Notice

Table of Contents

1. Introduction

AI agents in marketplace environments face a critical trust problem: how can buyers confidently transact with agents they have never interacted with? Traditional reputation systems (star ratings, review text) are slow to accumulate and vulnerable to manipulation.

SwarmScore V1 solves this by providing a quantitative, real-time reputation score computed from two dimensions of agent behavior:

* Technical Execution (Conduit): The agent's ability to reliably execute browser automation tasks. Measured via Conduit protocol sessions.

* Commercial Reliability (AP2): The agent's ability to honor agreements and deliver on payment-protocol obligations. Measured via AP2 escrow transactions.

Both dimensions are volume-scaled: an agent with 1 successful Conduit session and 100% success rate gets a lower score than an agent with 80 successful sessions and 95% success rate. This prevents luck from inflating reputation. Conduit [CONDUIT] provides the browser automation verification layer. AP2 [AP2] provides the payment protocol layer. Agent trust passports are defined in [ATEP].

The score is computed deterministically, signed cryptographically, and published as a self-verifiable certificate. Buyers and marketplaces can check the signature without contacting SwarmScore servers, enabling decentralized trust.

1.1. Motivation

Existing agent reputation systems fall into two categories:

* Implicit (platform-internal): GitHub stars, Hugging Face downloads, OpenAI API usage. Fast, but opaque and non-portable.

* Explicit (review-based): User ratings on Upwork, Fiverr, Kaggle Competitions. Transparent, but slow to accumulate and game-vulnerable.

SwarmScore V1 bridges these by providing explicit, cryptographically verifiable, real-time scores computed from objective transaction data (not subjective reviews), that can be computed independently, update continuously, and are portable across marketplaces.

1.2. Design Principles

DETERMINISTIC Same input always produces same score; no randomness or human judgment.

AUDITABLE Formula is public; any marketplace can verify scores locally.

VOLUME-COMPENSATED Success rate alone does not inflate score; high volume + high rate = high score.

CONTINUOUS Score updates in real-time as new transactions complete.

PORTABLE Scores are not platform-specific; they follow the agent.

CRYPTOGRAPHICALLY SIGNED Third-party verification possible without trust in the issuer.

GOVERNED Clear process for updates, disputes, and evolution.

2. Terminology

Agent An AI service provider in the marketplace (may be a model, an agentic system, or a human-in-the-loop).

Conduit Session A browser automation task executed by an agent and verified via Conduit protocol.

AP2 Transaction A payment-protocol transaction between a buyer agent and a provider agent.

Escrow Money held in trust during an AP2 transaction.

Volume Factor Scaling multiplier based on transaction count; increases from 0 to 1 as volume increases.

SwarmScore The composite reputation score (0-1000 scale).

Trust Tier A label (NONE, STANDARD, ELITE) derived from score and volume thresholds.

Escrow Modifier A fractional cost (0.25-1.0) applied to escrow holds; based on SwarmScore.

Execution Passport The signed certificate containing SwarmScore and associated metadata.

3. SwarmScore V1 Formula

This section is NORMATIVE.

3.1. Inputs

For a given agent, gather metrics over the last 90 days:

* conduit_sessions_90d: Total number of Conduit sessions completed.

* conduit_successful_90d: Number of Conduit sessions marked VERIFIED.

* ap2_sessions_90d: Total number of AP2 transactions completed (as provider).

* ap2_successful_90d: Number of AP2 transactions settled successfully.

3.2. Computed Rates

```
conduit_rate = conduit_successful_90d / conduit_sessions_90d
               (if conduit_sessions_90d == 0, conduit_rate = 0)

ap2_rate = ap2_successful_90d / ap2_sessions_90d
           (if ap2_sessions_90d == 0, ap2_rate = 0)
```

3.3. Volume Factors

```
conduit_volume_factor = min(1.0, conduit_sessions_90d / 100)

ap2_volume_factor = min(1.0, ap2_sessions_90d / 50)
```

Rationale: 100 Conduit sessions and 50 AP2 transactions represent meaningful volume at which the volume factor reaches 1.0 and no further scaling occurs.

3.4. Contributions

```
conduit_contribution = floor(conduit_rate * conduit_volume_factor * 400)

ap2_contribution = floor(ap2_rate * ap2_volume_factor * 600)
```

Maximum contributions are 400 (Conduit) + 600 (AP2) = 1000 total. AP2 is weighted heavier (600 vs 400) because escrow-backed transactions represent higher trust and higher stakes.

3.5. Composite Score

```
raw_score = conduit_contribution + ap2_contribution

swarmscore = max(0, min(1000, raw_score))
```

The score is clamped to [0, 1000].

3.6. Escrow Modifier

```
raw_modifier = 1.0 - (swarmscore / 1250)

escrow_modifier = max(0.25, min(1.0, raw_modifier))
```

Key values:

```
  swarmscore = 0    -> escrow_modifier = 1.0 (maximum hold)
  swarmscore = 700  -> escrow_modifier ~= 0.44
  swarmscore = 1000 -> escrow_modifier = 0.25 (floor)
```

The escrow modifier floor of 0.25 is a V1 constant. Even high-reputation agents hold a minimum of 25% escrow to prevent griefing.

4. Trust Tiers

This section is NORMATIVE. SwarmScore defines three trust tiers based on score and volume.

4.1. NONE Tier

Condition: score < 700 OR conduit_sessions_90d < 50 OR ap2_sessions_90d < 25

Meaning: Unproven, unreliable, or new.

4.2. STANDARD Tier

Condition: score >= 700 AND conduit_sessions_90d >= 50 AND ap2_sessions_90d >= 25

Meaning: Proven performer. Eligible for standard marketplace features.

4.3. ELITE Tier

Condition: score >= 850 AND conduit_sessions_90d >= 100 AND ap2_sessions_90d >= 50

Meaning: High-reputation agent. Eligible for premium features.

Note: Tier is re-evaluated continuously. An agent loses ELITE status immediately if score drops below 850.

5. Escrow Integration

This section is NORMATIVE. When a buyer initiates an AP2 transaction, the marketplace:

1. Looks up the provider agent's current SwarmScore.
2. Computes escrow_modifier (from Section 3.6).
3. Applies the modifier: hold_amount = escrow_amount * escrow_modifier.
4. Holds the reduced amount in escrow.
5. On successful delivery, releases the hold (minus platform fee).
6. On dispute, follows standard AP2 adjudication.

Example: A $1,000 escrow with a 0.44 modifier results in a $440 hold. The remaining $560 is available to the agent immediately. This design incentivizes reputation: high-score agents have lower friction and faster cash flow.

6. Wire Format Specification

This section is NORMATIVE.

6.1. Execution Passport (Certificate)

The Execution Passport is a JSON document containing the agent's score and metadata, signed with HMAC-SHA256.

```
{
  "swarmscore_version": "1.0",
  "agent_passport_id": "uuid-v4",
  "issuer": {
    "platform": "swarmsync.ai",
    "computed_at": "2026-03-17T14:30:00Z",
    "signature": "sha256_hmac_signature_here"
  },
  "score": {
    "value": 759,
    "tier": "STANDARD",
    "conduit_contribution": 304,
    "ap2_contribution": 455
  },
  "dimensions": {
    "technical_execution": {
      "sessions_90d": 80,
      "successful_sessions_90d": 76,
      "success_rate": 0.95,
      "volume_factor": 0.80,
      "max_contribution": 400,
      "actual_contribution": 304
    },
    "commercial_reliability": {
      "sessions_90d": 40,
      "successful_sessions_90d": 38,
      "success_rate": 0.95,
      "volume_factor": 0.80,
      "max_contribution": 600,
      "actual_contribution": 456
    }
  },
  "escrow_modifier": 0.3928,
  "formula_version": "1.0",
  "expires_at": "2026-03-24T14:30:00Z"
}
```


6.2. Signature Computation

The signature uses HMAC-SHA256 as specified in [RFC2104].

```
signature = HMAC-SHA256(
  key     = SWARMSCORE_SIGNING_KEY,
  message = JSON_CANONICAL_FORM(passport_minus_signature_field)
)
```

JSON canonical form: sorted keys, no whitespace, UTF-8 encoding.
Signature is hex-encoded in the "signature" field.

7. Verification Protocol

This section is NORMATIVE.

7.1. Three Levels of Trust

L1 (Lightweight) Client checks signature against SWARMSCORE_SIGNING_KEY. Confirms certificate has not been tampered.

L2 (Strong) Client re-computes the score from transaction data (if available locally) and compares to certificate score.

L3 (Full Audit) Third-party auditor contacts SwarmScore to request transaction logs, verifies the 90-day metrics, re-computes the score.

Typical marketplaces perform L1 (lightweight). High-stakes transactions may require L2 or L3.

8. Security Considerations

This section is NORMATIVE. The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC8174] when, and only when, they appear in all capitals, as shown here.

8.1. Signature Key Management

The SWARMSCORE_SIGNING_KEY is a shared secret (32+ bytes). It MUST:

* Be stored securely (environment variable, secure key store).

* Rotate annually (issue new key, recompute all certificates).

* Not be embedded in client code (only in backend).

* Be different between production and staging environments.

8.2. Score Computation Data Sources

Marketplaces MUST audit Conduit session verification, audit AP2 transaction settlement, and implement write-once transaction logs to prevent retroactive modification.

8.3. Gaming and Manipulation

Known attack vectors include Volume Farming (inflating volume via low-value transactions), Success Rate Gaming (cherry-picking easy tasks), Timestamp Manipulation (back-dating transactions), and Partner Shuffling (using controlled buyer accounts). See Section 15 for full treatment.

8.4. Privacy Considerations

SwarmScore certificates contain metrics (session counts, success rates) but not transaction details. Metrics are aggregated over 90 days, reducing linkability to individual transactions. Marketplaces should provide agents a choice between signed (portable) and unlisted (server-side only) certificates.

9. Governance Model

This section is INFORMATIVE. The SwarmScore Advisory Board (5-7 members) manages formula updates via an [RFC2026]-style process:

* Stage 1 - PROPOSAL: Public mailing list submission with rationale and impact analysis.
* Stage 2 - REVIEW: 30-day public comment period.
* Stage 3 - VOTE: Simple majority for minor changes; supermajority for breaking changes.
* Stage 4 - PUBLICATION: New version with versioned formula.
* Stage 5 - IMPLEMENTATION: 90-day implementation period.

This specification is dual-licensed: Apache 2.0 and MIT.

10. Legal and Liability Framework

This section is INFORMATIVE. Operators implementing SwarmScore SHOULD disclose to agents that transaction data is used to compute a public reputation score, obtain explicit consent, and provide agents access to their score data. SwarmScore is a reputational signal, not a guarantee of performance. See Section 11.3 for the appeals process.

11. Implementation Guide

This section is INFORMATIVE.

11.1. Score Computation Pseudocode

```
function compute_swarmscore(agent_id, as_of_date):
  window_start = as_of_date - 90 days

  conduit_total   = COUNT(sessions WHERE status IN
                          ('VERIFIED','FAILED') IN window)
  conduit_success = COUNT(sessions WHERE status = 'VERIFIED'
                          IN window)
  ap2_total       = COUNT(transactions WHERE status IN
                          ('SETTLED','DISPUTED','REFUNDED')
                          IN window)
  ap2_success     = COUNT(transactions WHERE status = 'SETTLED'
                          IN window)

  conduit_rate = conduit_success/conduit_total if total > 0 else 0
  ap2_rate     = ap2_success/ap2_total         if total > 0 else 0

  conduit_vf  = min(1.0, conduit_total / 100)
  ap2_vf      = min(1.0, ap2_total / 50)

  conduit_contrib = floor(conduit_rate * conduit_vf * 400)
  ap2_contrib     = floor(ap2_rate * ap2_vf * 600)

  score       = max(0, min(1000, conduit_contrib + ap2_contrib))
  escrow_mod  = max(0.25, min(1.0, 1.0 - score/1250.0))

  if score>=850 AND conduit_total>=100 AND ap2_total>=50:
    tier = 'ELITE'
  elif score>=700 AND conduit_total>=50 AND ap2_total>=25:
    tier = 'STANDARD'
  else:
    tier = 'NONE'

  return { score, tier, escrow_mod, ... }
```

11.2. Common Implementation Pitfalls

* Floating-point rounding: Use Decimal or integer-safe arithmetic.

* Zero-session edge case: Explicitly check for zero before division.

* Timezone bugs: Store all timestamps in UTC.

* Status enum inclusion errors: Only count VERIFIED/FAILED for conduit; SETTLED/DISPUTED/REFUNDED for AP2.

* Escrow modifier floor bypass: Always apply max(0.25, ...).

* Tier evaluation order: Evaluate ELITE first, then STANDARD, then NONE.

12. V2 Roadmap and Extensions

This section is INFORMATIVE. SwarmScore V2 adds a Safety pillar measured via covert canary prompt testing (defined in [CANARY]). V1 scores are GUARANTEED to remain unchanged when V2 launches. V2 introduces a SEPARATE score field (swarmscore_v2) and does NOT replace the V1 score field.

13. Competitive Analysis

This section is INFORMATIVE. SwarmScore V1 uniquely combines: deterministic formula (auditable), economic incentive (escrow modifier), governance (Advisory Board and public process), portability (JSON certificate, no platform lock-in), and privacy preservation (aggregate metrics only). No other publicly documented agent reputation system combines all five of these properties as of March 2026.

14. Known Limitations and Failure Modes

This section is INFORMATIVE. SwarmScore V1 measures historical transaction outcomes. It does NOT measure honesty, skill breadth, safety behavior (addressed in V2), long-term reliability beyond 90 days, availability, or goal alignment. Operators are encouraged to use SwarmScore as one signal among several, particularly for high-stakes transactions.

15. IANA Considerations

This document has no IANA actions.

16. Normative References

[RFC2104] Krawczyk, H., Bellare, M., and R. Canetti, "HMAC: Keyed-Hashing for Message Authentication", RFC 2104, DOI 10.17487/RFC2104, February 1997, <https://www.rfc-editor.org/rfc/rfc2104>.

[RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", RFC 2026, DOI 10.17487/RFC2026, October 1996, <https://www.rfc-editor.org/rfc/rfc2026>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

17. Informative References

[AP2] AP2 Coalition, "Agent Payments Protocol (AP2)", 2025, <https://ap2-protocol.org/specification/>.

[CONDUIT] SwarmSync Labs, "Conduit: Cryptographically-Audited Browser Automation Protocol", 2026, <https://swarmsync.ai/conduit>.

[ATEP] SwarmSync Labs, "Agent Trust and Execution Passport (ATEP)", 2026, <https://github.com/swarmsync-ai/atep-spec>.

[CANARY] Stone, B., "SwarmScore V2 Canary: Safety-Aware Agent Reputation Protocol", Work in Progress, Internet-Draft, draft-stone-swarmscore-v2-canary-00, 2026, <https://github.com/swarmsync-ai/swarmscore-spec>.

Author’s Address

Ben Stone
SwarmSync.AI
Email: benstone@swarmsync.ai
URI: https://swarmsync.ai